CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Elsevier - Publisher Connector

Am. J. Hum. Genet. 75:35–43, 2004

Linkage Disequilibrium Mapping via Cladistic Analysis of Single-Nucleotide Polymorphism Haplotypes

Caroline Durrant,[1] Krina T. Zondervan,[1] Lon R. Cardon,[1] Sarah Hunt,[2] Panos Deloukas,[2] and Andrew P. Morris[1]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford; and [2]Wellcome Trust Sanger Institute, Hinxton, United Kingdom

We present a novel approach to disease-gene mapping via cladistic analysis of single-nucleotide polymorphism (SNP) haplotypes obtained from large-scale, population-based association studies, applicable to whole-genome screens, candidate-gene studies, or fine-scale mapping. Clades of haplotypes are tested for association with disease, exploiting the expected similarity of chromosomes with recent shared ancestry in the region flanking the disease gene. The method is developed in a logistic-regression framework and can easily incorporate covariates such as environmental risk factors or additional unlinked loci to allow for population structure. To evaluate the power of this approach to detect disease-marker association, we have developed a simulation algorithm to generate high-density SNP data with short-range linkage disequilibrium based on empirical patterns of haplotype diversity. The results of the simulation study highlight substantial gains in power over single-locus tests for a wide range of disease models, despite overcorrection for multiple testing.

Introduction

Disease-marker association studies of samples of unrelated affected cases and unaffected controls have been widely recognized as having the potential to map genetic polymorphisms contributing to complex traits, provided that the variant is not extremely rare (Risch and Merikangas 1996; Zondervan and Cardon 2004). With the publication of the SNP map of the human genome (International SNP Map Working Group 2001; International Human Genome Sequence Consortium 2001) and improvements in the efficiency of high-throughput genotyping technology, genomewide screens of high-density marker panels are becoming increasingly feasible for large sample sizes. The success of this approach to gene mapping now depends on the availability of powerful statistical analysis techniques.

The key concept underlying any analysis of disease-marker association studies is linkage disequilibrium (LD), the nonrandom assortment of alleles at loci within populations of unrelated individuals, generated as a result of their shared ancestry. Consider a disease arising as a result of relatively recent mutations at proximal loci within the same gene. Figure 1 illustrates an example of a genealogical tree used to represent the ancestry of a sample of chromosomes at the disease gene. A pair of *disease* chromosomes carrying the *same* mutation are expected to share a more recent common ancestor at the disease gene than a pair of chromosomes carrying *different* mutations. Moreover, the most recent common ancestor (MRCA) at the disease gene of mutation-free *normal* chromosomes is expected to be more ancient than the founders for any specific disease mutation event.

At the instant a specific disease mutation occurs, it is carried on a single founding haplotype and is in complete LD with alleles at any other SNP. Over subsequent generations, recombination will break down the founder haplotype, weakening LD with the disease mutation. However, with high-density maps of markers, the probability of recombination between the disease gene and neighboring SNPs is small. Thus, the founder haplotype is expected to be preserved in the vicinity of the disease gene on chromosomes carrying the mutation. A mismatch of alleles within the preserved region can occur only as a result of marker mutation.

The same representation could be applied to normal chromosomes. However, recombination is expected to have broken down LD in normal chromosomes even in the region directly flanking the disease gene, because their MRCA is more ancient than for disease chromosomes. Consequently, a sample of disease chromosomes is expected to display excess sharing of the founder SNP haplotype(s) over normal chromosomes, with the excess decaying with distance from the disease gene. This representation assumes low-risk alleles to be ancient and thus precludes recent mutations with protective effects, for example.

Received January 27, 2004; accepted for publication April 21, 2004; electronically published May 13, 2004.

Address for correspondence and reprints: Dr. Andrew Morris, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, United Kingdom. E-mail: amorris@well.ox.ac.uk

© 2004 by The American Society of Human Genetics. All rights reserved.
0002-9297/2004/7501-0006$15.00

Unfortunately, this simple model of LD is unrealistic for marker association with complex diseases. Environmental factors, dominance, polygenic effects, and epistasis will affect the relative frequencies of *sporadic* normal chromosomes carried by affected cases and *nonpenetrant* disease chromosomes carried by unaffected controls, introducing substantial noise in the relationship between disease phenotype and genotype. Further, *false-positive* signals of disease-marker association can occur at an increased rate as a result of population substructure that is not accounted for in the ascertainment process. The challenge for the analysis of disease-marker association studies is to develop methodology that can efficiently detect LD resulting from the common ancestry of specific disease mutations in a complex genetic setting and can differentiate between it and SNP haplotype sharing due to background patterns of association generated by the underlying demographic structure.

In this article, we present a novel approach to disease-gene mapping via cladistic analysis of SNP haplotypes obtained from large-scale population-based association studies. Large genomic regions are treated as *sliding windows* of SNPs, with separate analyses performed within each window. SNP haplotype diversity is quantified in terms of the proportion of marker matches within the window. Such a metric is consistent with haplotype diversity driven by marker mutation, in the absence of recombination. Hence, a window can be thought of as corresponding to a *haplotype block,* with high levels of LD between SNPs maintained by minimal ancestral recombination (Daly et al. 2001; Goldstein 2001; Gabriel et al. 2002).

If we ignore disease phenotype, haplotype diversity in each window is represented by means of a *cladogram,* constructed using standard hierarchical clustering techniques (Everitt 1993). In windows overlapping the region flanking the disease gene, the cladogram is expected to approximate the genealogical tree underlying the shared ancestry of case and control chromosomes. Consequently, we expect correlation between disease phenotypes and clusters in the cladogram, with excess sharing of the founder SNP haplotype(s) among the high-risk clade(s) of chromosomes.

Our method is developed in a logistic-regression framework that can be generalized to incorporate covariates, which might include potential environmental risk factors or genotypes at additional unlinked markers to control for population structure (Pritchard and Rosenberg 1999). We demonstrate the power of this approach by simulation of high-density SNP data, on the basis of empirical patterns of haplotype diversity across a 10-Mb region of chromosome 20 (Ke et al. 2004), highlighting substantial gains over single-locus tests to detect associations for a wide range of complex disease models.

Methods

Consider a sample of unrelated affected cases and unaffected controls, typed for M SNPs in a region of interest. We assume that phase-known genotype data are available, where the pair of haplotypes carried by the ith individual is denoted by $\mathbf{H}_i = \{H_{i1}, H_{i2}\}$, and the haplotype $H_{ij} = \{H_{ij[1]}, H_{ij[2]}, \ldots, H_{ij[M]}\}$. Alleles, $H_{ij[m]}$, at SNP m are coded 1 and 2, with 0 denoting missing data. The relative frequency of allele 1 at SNP m is denoted as q_m and is estimated from the available genotype data.

Cladistic Representation of Haplotype Diversity

We represent haplotype diversity by means of a *cladogram,* an example of which is presented in figure 2. The cladogram illustrates successive *partitions* of SNP haplotypes, $\mathbf{T}[h], \mathbf{T}[h-1], \ldots, \mathbf{T}[1]$. The first partition, $\mathbf{T}[h]$, consists of h *clusters,* each corresponding to a group of chromosomes carrying the same distinct SNP haplotype, represented by nodes at the foot of the cladogram. Subsequent partitions merge together increasingly diverse clusters of haplotypes. The final partition, $\mathbf{T}[1]$, at the top of the cladogram, combines all haplotypes in a single cluster.

The cladogram is constructed using simple hierarchical group averaging techniques. At each partition, clusters of haplotypes from the previous partition are merged such that the mean pairwise haplotype diversity is minimized within the new clade. We define the diversity

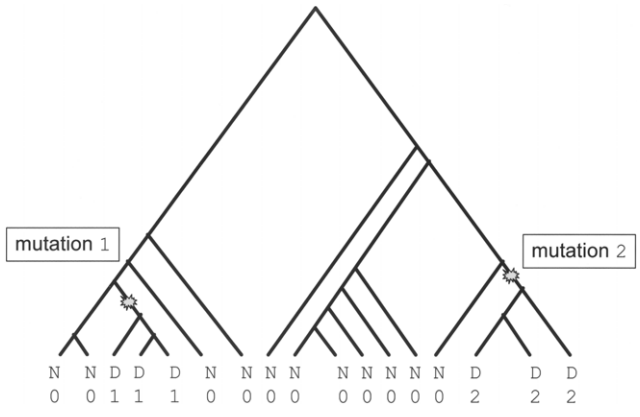

Figure 1 Example of a genealogical tree representing the shared ancestry of chromosomes at the disease gene. Disease chromosomes (D) carrying the same mutation (1 or 2), share more recent common ancestry than normal chromosomes (N) carrying no mutation (0).

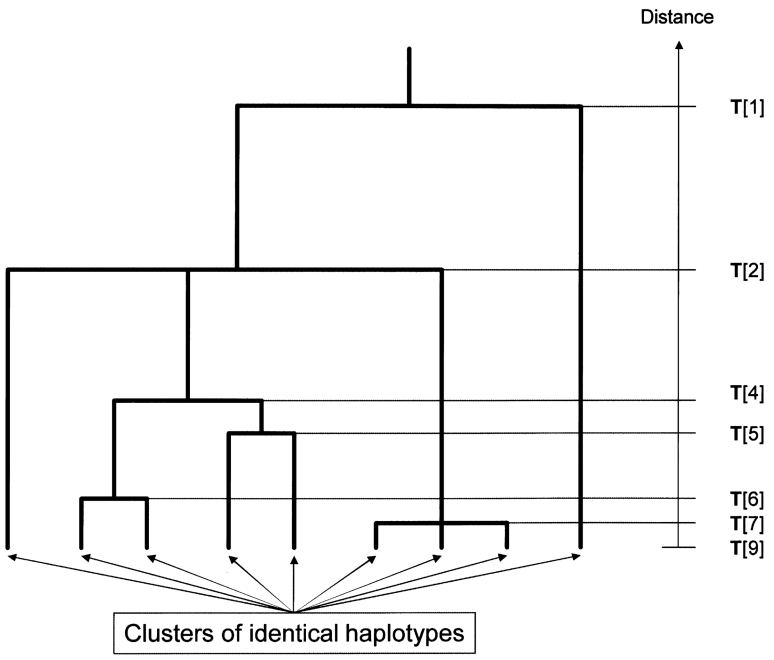

Figure 2 Example of a cladogram representing haplotype diversity within a window of SNPs. The cladogram is constructed using hierarchical group average clustering on pairwise haplotype differences, expressed in terms of the proportion of marker mismatches within the window of SNPs.

between a pair of haplotypes, $H_{i_1 j_1}$ and $H_{i_2 j_2}$, by means of the distance metric

$$D_{i_1 j_1, i_2 j_2} = 1 - \frac{\sum_{m=1}^{M} s_{i_1 j_1, i_2 j_2 [m]} w_m}{\sum_{m=1}^{M} w_m},$$

where w_m is the weight assigned to similarities, $s_{i_1 j_1, i_2 j_2 [m]}$, at SNP m, given by

$$s_{i_1 j_1, i_2 j_2 [m]} = \begin{cases} 1 - q_m & \text{if } H_{i_1 j_1 [m]} = H_{i_2 j_2 [m]} = 1 \\ q_m & \text{if } H_{i_1 j_1 [m]} = H_{i_2 j_2 [m]} = 2 \\ q_m (1 - q_m) & \text{if } H_{i_1 j_1 [m]} = 0 \text{ or } H_{i_2 j_2 [m]} = 0 \\ 0 & \text{otherwise} \end{cases}.$$

Chromosomes that share rare alleles are expected to share more-recent ancestry than chromosomes sharing common alleles and thus are treated as more similar in this definition of haplotype diversity. Thus, we quantify allele sharing by the *complementary* allele frequency—that is, by $1 - q_m$, for sharing allele 1 at SNP m, and by q_m, for sharing allele 2. If allele $H_{i_1 j_1 [m]}$ is missing and allele $H_{i_2 j_2 [m]} = 1$, the probability that the two haplotypes match at SNP m is q_m, in which case the similarity score is $1 - q_m$, and, overall, $q_m (1 - q_m)$. On the other hand, if allele $H_{i_2 j_2 [m]} = 2$, the probability that the two haplotypes match at SNP m is $1 - q_m$, in which case the similarity score is q_m—again, $q_m (1 - q_m)$ overall. For the case in which both alleles are missing, $H_{i_1 j_1 [m]} = H_{i_2 j_2 [m]} = 0$, the probability that the haplotypes share allele 1 at SNP m is q_m^2, in which case the similarity score is $1 - q_m$. Conversely, the probability that the haplotypes share allele 2 at SNP m is $(1 - q_m)^2$, in which case the similarity score is q_m. Thus, overall, when $H_{i_1 j_1 [m]} = H_{i_2 j_2 [m]} = 0$, the similarity score is $q_m^2 (1 - q_m) + q_m (1 - q_m)^2 = q_m (1 - q_m)$.

The simple distance metric, $w_m = 1$ for all m, assigns equal weight to similarities at each SNP. For large genomic regions, we allow for recombination by treating the marker panel as a sliding window, $\mathcal{W}$, of SNPs,

$$w_m = \begin{cases} 1 & \text{if } m \in \mathcal{W} \\ 0 & \text{otherwise} \end{cases},$$

with separate analyses performed within each overlapping window. Under this model, windows of SNPs correspond to blocks of strong LD, maintained by minimal ancestral recombination—that is, haplotype diversity driven by marker mutation.

Logistic-Regression Model

Consider the partition of haplotypes, T[c], to c clusters in window $\mathcal{W}$. We model the disease status of individual i, denoted by $y_i = 1$ for a case and $y_i = 0$ for a control, in a logistic-regression framework. The

model is parameterized in terms of *log-odds* of disease, $\boldsymbol{\beta}^{[c]} = \{\beta_1^{[c]},\beta_2^{[c]},\ldots,\beta_c^{[c]}\}$, for each cluster. We also allow for covariates, denoted by $\mathbf{x}_i = \{x_{i1},x_{i2},\ldots,x_{iK}\}$ for individual i, with corresponding regression coefficients $\boldsymbol{\gamma}^{[c]} = \{\gamma_1^{[c]},\gamma_2^{[c]},\ldots\gamma_K^{[c]}\}$. Denoting the cluster assignment of haplotype H_{ij} in partition $\mathbf{T}[c]$ by $T[c]_{ij}$, the log likelihood of SNP haplotypes in window $\mathcal{W}$ is given by

$$\ell(\mathbf{y}|\mathbf{T}[c],\boldsymbol{\beta}^{[c]},\boldsymbol{\gamma}^{[c]},\mathbf{x},\mathcal{W}) = \sum_{i=1}^{n}[y_i \ln p_i + (1-y_i)\ln(1-p_i)] \ ,$$

where

$$p_i = \frac{\exp[\eta_i]}{1+\exp[\eta_i]} \ .$$

Under the assumption of Hardy-Weinberg equilibrium and a multiplicative disease model, the linear component, η_i, is given by

$$\eta_i = \beta^{[c]}_{T[c]_{i1}} + \beta^{[c]}_{T[c]_{i2}} + \sum_{k=1}^{K}\gamma_k^{[c]}x_{ik} \ .$$

Under the null hypothesis of no association between disease and SNPs in window $\mathcal{W}$, each haplotype has equal odds of being carried by a case or control. This is represented by a single cluster of haplotypes, given by the null partition $\mathbf{T}[1]$. Thus, a likelihood-ratio test of disease-marker association can be constructed for partition $\mathbf{T}[c]$,

$$\Lambda_{[c]} = 2\left[\ell\left(\mathbf{y}|\mathbf{T}[c],\hat{\boldsymbol{\beta}}^{[c]},\hat{\boldsymbol{\gamma}}^{[c]},\mathbf{x},\mathcal{W}\right) - \ell\left(\mathbf{y}|\mathbf{T}[1],\hat{\beta}^{[1]},\hat{\gamma}^{[1]},\mathbf{x},\mathcal{W}\right)\right] \ ,$$

having an approximate χ^2 distribution with $c-1$ df under the null hypothesis.

For large windows of SNPs, clusters of haplotypes may occur with low relative frequency in the sample of case and control chromosomes. To allow for rare haplotypes, we pool clusters in partition $\mathbf{T}[c]$ with relative sample frequencies <5%. In partition $\mathbf{T}[c]$, we denote the number of clusters with relative frequencies >5% by c^*, so that $\boldsymbol{\beta}^{[c]} = \{\beta_1^{[c]},\beta_2^{[c]},\ldots,\beta_{c^*}^{[c]},\beta_{\mathrm{P}}^{[c]}\}$, where $\beta_{\mathrm{P}}^{[c]}$ denotes the log-odds of disease for the pooled clusters of haplotypes. Under the null hypothesis of no disease-marker association, the likelihood ratio test, $\Lambda_{[c]\mathrm{P}}$, for pooled haplotypes has an approximate χ^2 distribution with c^* df.

Moving up through the cladogram, combining increasingly diverse haplotypes in fewer clusters represents a trade-off between lack of fit of the logistic-regression model and reduced degrees of freedom in the likelihood-ratio test. For each window, we select the best partition of haplotypes, $\mathbf{T}[\mathrm{MAX}]$, maximizing the evidence of disease marker association (i.e., minimizing the P value) in the likelihood-ratio test. This procedure incorporates two levels of multiple testing: (i) $t_{\mathcal{W}}$ partitions of haplotypes in the cladogram for SNPs in window $\mathcal{W}$ and (ii) N overlapping windows of SNPs. To obtain an *experimentwise* false-positive error rate of $100\alpha_E$%, the *pointwise* significance level for window $\mathcal{W}$, $100\alpha_{\mathcal{W}}$%, is adjusted by Bonferroni correction, $\alpha_{\mathcal{W}} = \alpha_E/Nt_{\mathcal{W}}$.

Software Availability

The method has been coded in the CLADHC algorithm, available as a linux executable, with accompanying documentation, on request from the corresponding author.

Simulation Study

In this section, we present details of a simulation study to evaluate the performance of the proposed cladistic analysis of SNP haplotypes as an approach to LD mapping. We generate case-control samples of high-density SNP haplotypes consistent with empirical patterns of LD across a 10-Mb region of chromosome 20, derived from 92 haplotypes obtained from CEPH pedigrees (Ke et al. 2004). We consider a range of different disease models, under the assumption of a single high-risk variant at the disease locus. The models are parameterized in terms of the disease allele frequency (DAF; in the range 0.01–0.5), and genotype relative risks (GRRs) to individuals who are heterozygous (GRR 1–3) or homozygous (GRR 1.1–40) for the high-risk variant. These models encompass a number of established gene polymorphisms with strong evidence of association with complex diseases from empirical studies, including NOD2 for Crohn disease (Hugot et al. 2001) and APOE for Alzheimer disease (Rubinsztein and Easton 1999).

For each combination of disease model and DAF, we generate 1,000 replicates of phase-known genotype data for a sample of 1,000 cases and 1,000 controls. To generate the SNP haplotypes carried by a case or control, we begin by selecting a SNP, x, at random from the 5,216 markers in the chromosome 20 panel, with minor-allele frequency approximately equal to the chosen DAF. We generate the genotype at this disease locus according to the disease model and assign one allele to each chromosome in the pair carried by that individual. The simulation for each chromosome is illustrated in figure 3 and proceeds as follows:

1. Select an empirical haplotype at random for the set of 5 SNPs $[x-2, x+2]$.
2. If the empirical haplotype carries the same allele at

Table 1

False-Positive Error Rates of Disease-Gene Detection at the 5% Experimentwise Significance Level, as a Function of DAF, Averaged over Window Size and Disease Model

Analysis Method	False-Positive Error Rate (%) at a DAF of				
	.01	.05	.1	.2	.5
T[MAX]	1.28	1.33	1.29	1.36	1.32
T[*h*]	3.03	3.03	2.85	2.94	3.13
Single locus	3.12	3.33	3.51	3.11	3.22

the disease SNP as the simulated chromosome, accept the haplotype for the set of SNPs $[x-2, x+2]$. Otherwise, return to step 1.

3. Select an empirical haplotype at random for the set of 5 SNPs $[x-1, x+3]$.
4. If the empirical haplotype and simulated chromosome match at all overlapping SNPs, accept the haplotype for the set of 5 SNPs $[x-1, x+3]$. Otherwise, return to step 3.
5. Repeat steps 3 and 4 for successive overlapping sets of 5 SNPs, $[x, x+4], [x+1, x+5], \ldots$, to the right of the disease SNP, and then $[x-3, x+1], [x-4, x], \ldots$, to the left of the disease SNP.
6. Remove disease SNP from haplotype.

The process is repeated to generate the required number of cases and controls. By matching haplotypes in overlapping windows, the simulated chromosomes are not exact copies of those in the original sample. Instead, the 92 CEPH chromosomes are used to generate plausible haplotypes in a wider population.

For each replicate of haplotype data, we consider a sliding window of SNPs across the candidate region. For each window, we perform three association analyses:

1. single-locus allele-based analyses using Pearson's χ^2 test for 2 × 2 contingency tables, corrected for the number of SNPs;
2. haplotype-based logistic-regression analysis using the first partition of haplotypes, T[*h*], corresponding to one cluster for each distinct haplotype, corrected for the number of windows; and
3. haplotype-based logistic-regression analysis using the best partition of haplotypes, T[MAX], corrected for the number of windows and partitions of haplotypes in the cladogram.

We consider windows of size 4, 6, 8, and 10 markers and then focus on two key measures: (i) the power of each method to detect disease-marker association in windows overlapping the region flanking a disease gene and (ii) the false-positive error rates of disease-gene detection in windows that do not overlap the flanking region (i.e., null windows). The SNP panel consists of polymorphisms with an approximate uniform distribution of minor allele frequencies (MAFs). Hence, all results below are effectively averaged over marker MAF.

To generate a single replicate of SNP genotype data for a sample of 1,000 cases and 1,000 controls requires ~2 min of computing time for a dedicated Pentium 3 processor. The computing time required for the analysis of a single replicate of data using all three methods varies according to window size: <30 s for a window of size 4, <1 min for a window of size 8, and <3 min for a window of size 10.

Results

Table 1 presents the false-positive (type I) error rates of disease-gene detection in *null windows* at the 5% experimentwise significance level, for the three association analysis methods, as a function of the DAF, averaged over window size and disease model. For each analysis method, the Bonferroni correction is conservative. The best partition of haplotypes, T[MAX], is most conservative, presumably because this test requires two levels of correction: for the number of windows and for the number of partitions in the cladogram. Table 2 presents the false-positive error rates of disease-gene detection in null windows at the same experimentwise significance level, this time as a function of window size, averaged over the DAF and disease model. DAF and disease model have no apparent effect on false-positive rates (results not presented), since we focus here on null windows that do not overlap the disease gene. However, the haplotype-based tests become increasingly conservative with longer window size. Longer windows contain greater haplotype diversity, with larger numbers of degrees of freedom in the test at the first partition of haplotypes, T[*h*], and additional correction for increased levels of clustering in the cladogram for the test at the best partition of haplotypes, T[MAX].

Figure 4 presents the power to detect disease-marker association in windows of 8 markers overlapping the region flanking a disease gene, under the assumption of

Table 2

False-Positive Error Rates of Disease-Gene Detection at the 5% Experimentwise Significance Level, as a Function of Window Size (Number of Markers), Averaged over the Disease Allele Frequency and Disease Model

Analysis Method	False-Positive Error Rate (%) at a Window Size of			
	4 Markers	6 Markers	8 Markers	10 Markers
T[MAX]	1.73	1.38	1.13	1.01
T[*h*]	3.30	2.98	2.92	2.78
Single locus	3.28	3.32	3.26	3.16

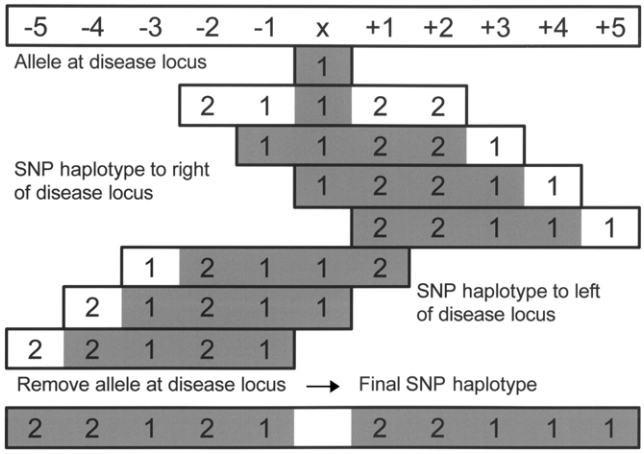

Figure 3 Generating the SNP haplotype carried by a chromosome carrying allele 1 at the disease SNP, *x*.

a 5% experimentwise significance level, with Bonferroni correction for multiple testing. Power is presented for each of the three analysis methods as a function of the disease model, for a DAF of 0.05. The power of both haplotype-based methods, **T**[MAX] and **T**[*h*], is substantially greater than the single-locus test. This trend is repeated over different DAFs (results not presented). The greatest power for all three methods is attained for the most common disease alleles, since, for common variants, the proportion of affected individuals carrying two copies of the disease allele is greatest. The best partition of haplotypes, **T**[MAX], generally has greater power to detect disease-marker association than the first partition of haplotypes, **T**[*h*]. This suggests the positive benefits of reduced degrees of freedom in a trade-off against correction for the additional levels of multiple testing.

Figure 5 presents the power of the best partition of haplotypes, **T**[MAX], to detect disease-marker association in windows of varying size overlapping the disease gene, under the assumption of a 5% experimentwise significance level, with Bonferroni correction for multiple testing. Power is presented as a function of the disease model, for a DAF of 0.05. For each model, a window of size 6 is most powerful, with decreasing power obtained for windows of 8 and then 10 markers. For windows consisting of >6 markers, the correction for additional partitions in the cladogram, because of increased haplotype diversity, overwhelms the extra information about LD. In general, smaller windows have greater power relative to longer windows, for higher DAFs, which is consistent with our expectation that LD will not extend as far for older alleles.

For the simulation study, a window of size 6 is optimal by averaging over the whole 10-Mb region. However, the optimal window size will vary according to the density of marker SNPs and the extent of LD across the region under investigation. For example, in regions of low LD, we expect high levels of haplotype diversity and, hence, that a small window will be optimal. One ap-

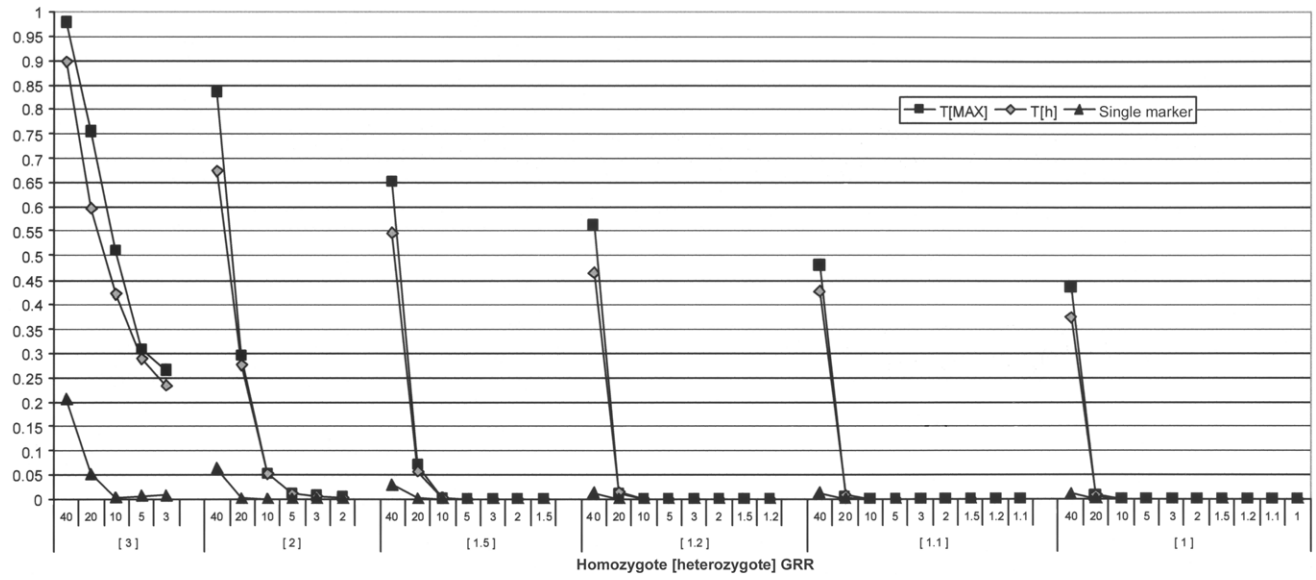

Figure 4 Power to detect disease-marker association in windows of 8 markers overlapping the region flanking a disease gene, under the assumption of a 5% experimentwise significance level, with Bonferroni correction for multiple testing. Power is presented as a function of the disease model, for a disease allele frequency of 0.05. The three analysis methods are the best partition of haplotypes, **T**[MAX]; the first partition of haplotypes, **T**[*h*]; and a single-locus test.

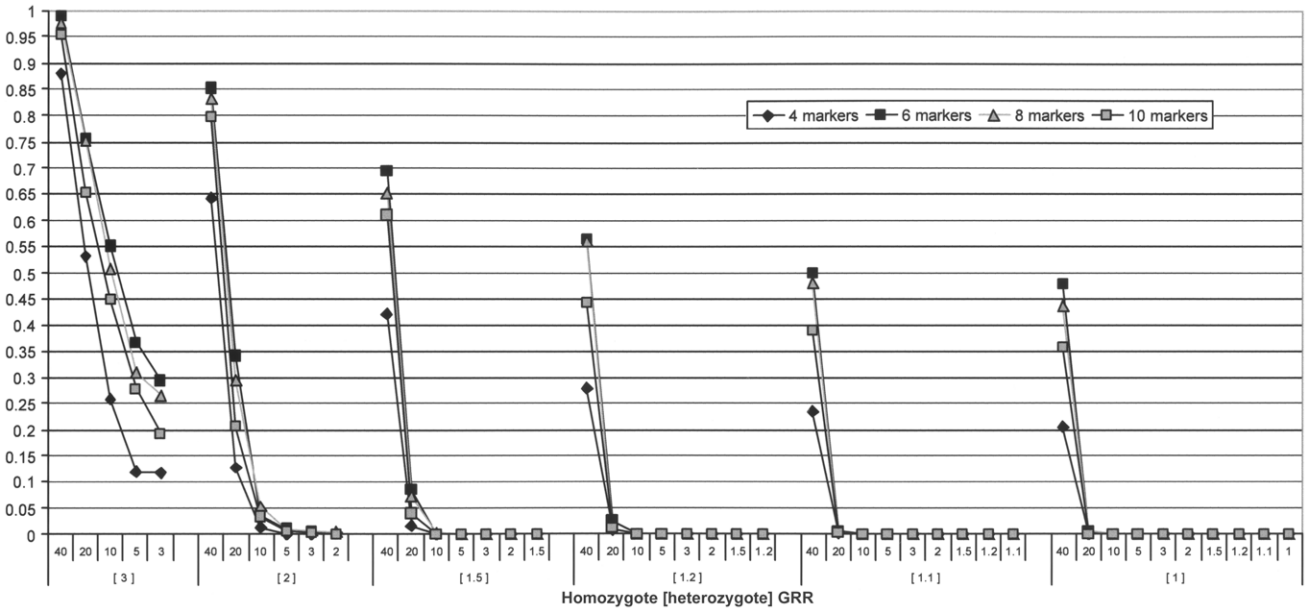

Figure 5 Power of the best partition of haplotypes, T[MAX], to detect disease-marker association in windows of varying size (numbers of markers) overlapping the disease gene, under the assumption of a 5% experimentwise significance level, with Bonferroni correction for multiple testing. Power is presented as a function of the disease model, for a disease allele frequency of 0.05.

proach to overcoming this problem would be to align windows with blocks of LD, through use of data from the International HapMap project (International HapMap Consortium 2003).

Example Application: Fine-Scale Mapping

Cystic fibrosis (CF) is the most common fully penetrant recessive disorder in white populations, with an incidence on the order of 1 case per 2,500 live births. The disease is well understood, and a single functional gene, CFTR, has been located on chromosome 7q31. It is now known that a 3-bp deletion, ΔF508, in the CFTR gene accounts for ~66% of chromosomal mutations in the same gene (Bertranpetit and Calafell 1996). Kerem et al. (1989) typed 94 case chromosomes and 92 control chromosomes at 23 diallelic markers in a 1.8-Mb region including the CFTR gene. The sample incorporates genetic heterogeneity at the CFTR locus, since only 62 of the case chromosomes carry ΔF508. Consequently, the resulting haplotype data have become a useful test set for fine-mapping methods, reviewed by Morris et al. (2002).

Figure 6 presents the distribution of $-\log_{10} P$ values across the candidate region from single-locus allele-based analyses using Pearson's χ^2 test for 2 × 2 contingency tables. The true location of ΔF508 is at 0.88 Mb, indicated by the solid vertical line. There is a region of strong LD at 0.6–0.9 Mb, although the closest marker to ΔF508 shows little evidence of association. We have analyzed the region through use of a sliding window of 6 markers and have tested for disease-marker association for the best partition of haplotypes, T[MAX], within each window. Figure 6 also presents the $-\log_{10} P$ values from these analyses, plotted against

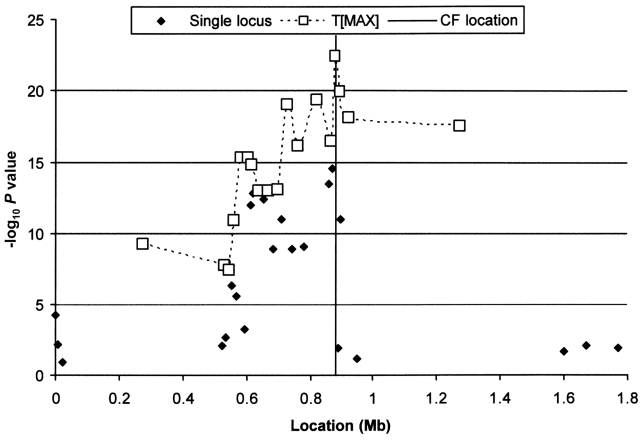

Figure 6 Distribution of single-marker association with CF across a 1.8-Mb candidate region flanking the CFTR gene (Kerem et al. 1989). The solid vertical line indicates the true location of ΔF508 at 0.88 Mb, the most common disease mutation identified in the CFTR gene. The dashed line indicates the distribution of association with CF of the best partition of haplotypes, T[MAX], using a sliding window of 6 markers across the candidate region.

Table 3

Best Partition of Haplotypes and Corresponding ORs for Cystic Fibrosis for the Data from Kerem et al. (1989)

Cluster and Haplotype	No. of Cases	No. of Controls	OR (95% CI)
1:			
212111	8	40	
212011	0	4	1 (baseline)
212112	2	0	
2:			
121121	61	3	96.8 (25.2–371.7)
121021	5	0	
3:			
112221	10	16	
112211	1	2	2.4 (0.9–6.6)
212221	0	1	
112222	0	1	
4:			
112011	0	4	1.3 (.2–7.0)
112111	2	3	
5:			
121122	1	5	
121022	0	2	.6 (.06–4.9)
121222	0	1	
Pooled[a]:			
121100	2	1	
000100	2	2	
121111	0	2	
212121	0	1	1.8 (.5–6.8)
112021	0	1	
122221	0	1	
222011	0	1	
221111	0	1	

[a] The pooled cluster includes all clusters with frequency <5%.

the location of the center of the window. All locations are highly significant, adjusting for the number of windows and partitions of haplotypes. The strongest evidence of association was obtained for the window of markers centered at 0.879 Mb, just 1 kb from the true location of ΔF508. Similar results were obtained for windows of 4, 8, and 10 markers, with a similar region identified with maximal evidence of disease-marker association.

Table 3 presents the best partition of haplotypes, T[MAX], in the 6-marker window of strongest association, together with the corresponding odds ratios (ORs) for CF, when the cluster with the highest frequency of controls is taken as baseline. Cluster 2 has the highest odds of CF and includes 66 case chromosomes, the majority of which carry the ΔF508 mutation. The remainder of the case chromosomes are scattered across the other five clusters and carry rarer CF mutations, each of which would be expected to have occurred on a different marker haplotype background.

Discussion

We present a novel method for disease-gene mapping using SNP haplotypes obtained from large-scale population-based association studies. In the vicinity of the disease gene, we expect that a cladistic representation of haplotype diversity, constructed using standard hierarchical clustering techniques on a simple pairwise distance metric, will approximate the ancestry of the sample. The cladogram highlights clusters of similar haplotypes that we expect to have similar contributions to the risk of disease. The method is applicable to the analysis of whole-genome screens, candidate genes, or fine-scale mapping.

We have developed a simulation algorithm to generate high-density SNP data with short-range LD based on empirical patterns of haplotype diversity. The algorithm uses relatively small samples of chromosomes typed on a high-density SNP panel to simulate plausible haplotypes in a wider population, allowing for the much larger sample sizes required for case-control studies of complex disease. The results of our simulation study suggest that cladistic analysis of haplotypes is substantially more powerful than single-locus methods. There are also gains in power attained by partitioning haplotypes according to their similarity, compared with the less parsimonious approach of allowing a different risk for each distinct haplotype, despite the need for correcting for the additional levels of multiple testing. The Bonferroni correction is conservative, so we might expect further improvements in power with more appropriate adjustments for multiple testing, such as false-discovery rates. Alternatively, the computational efficiency of the algorithm means that we could employ a permutation procedure to obtain the exact distribution of $\Lambda_{[MAX]}$ under the null hypothesis of no disease-marker association.

In our simulation study, we assume a single high-risk disease variant, which may not be entirely realistic for many complex diseases. The simulation algorithm could be adapted to allow for multiple mutations at proximal loci (i.e., within the same disease gene) or at distant loci (i.e., in different disease genes). We would expect the cladistic method to perform well in the former setting, since each disease mutation should correspond to a different clade of high-risk haplotypes. In the latter case, power to detect disease-marker association will depend on the main effects of each mutation to disease risk, since we do not model interactions between disease loci.

In our method, we assume that haplotype diversity is driven by marker mutation in the absence of recombination and thus is quantified in terms of the proportion of markers at which a pair of haplotypes are identical. To allow for recombination, we make use of sliding windows of SNPs across the marker panel, with

separate analyses performed within each window. However, to allow for recombination within a window, we could weight marker matches according to a linear or exponential function of distance from the center of the haplotype. Molitor et al. (2003*a*) discuss a number of other alternative distance metrics.

In the method presented here, we assume that we have phase-known haplotype data, which will not generally be available with current SNP genotyping technology. We can estimate the haplotype configuration of phase-unknown genotype data through use of statistical inference methods such as PHASE (Stephens et al. 2001; Stephens and Donnelly 2003). However, it is important to account for uncertainty in the haplotype reconstruction, which is an estimate and subject to error. Alternatively, we could focus on genotype diversity directly, without the need for haplotype information.

Molitor et al. (2003*b*) have proposed a promising approach to gene mapping, based on clustering haplotypes under a Bayesian partition model. They allow for missing marker data and uncertainty in the partition in a Markov chain Monte Carlo (MCMC) framework. This approach would be straightforward to adapt to phase-unknown genotype data, treating the unknown haplotypes as latent variables in the MCMC algorithm. However, MCMC methods are computationally intensive and, although this approach may be viable for analysis of candidate genes or small genomic regions, it is not clear that it could be adapted to whole-genome screens.

Acknowledgment

The authors acknowledge financial support from the Wellcome Trust.

References

Bertranpetit J, Calafell (1996) Genetic and geographical variability in cystic fibrosis: evolutionary considerations. In: Weiss K (ed) Variation in the human genome. Wiley, Chichester, England, pp 97–114

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

Everitt BS (1993) Cluster analysis, 3rd edition. Arnold, London, pp 55–89

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

Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109–111

Hugot J-P, Chamaillard M, Zouali H, Lesage S, Cezard J-P, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411: 599–603

International HapMap Consortium (2003) The international HapMap project. Nature 426:789–795

International Human Genome Sequence Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

International SNP Map Working Group (2001) A map of the human genome sequence variation contains 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris A, Bentley D, Cardon LR, Deloukas P (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet 13:577–588

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui L (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Molitor J, Marjoram P, Thomas D (2003*a*) Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. Genet Epidemiol 25:95–105

——— (2003*b*) Fine scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am J Hum Genet 73:1368–1384

Morris AP, Whittaker JC, Balding DJ (2002) Fine scale mapping of disease loci via shattered coalescent modeling of genealogies. Am J Hum Genet 70:686–707

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Rubinsztein DC, Easton DF (1999) Apolipoprotein E genetic variation and Alzheimer's disease: a meta analysis. Dement Geriatr Cogn Disord 10:199–209

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genetic data. Am J Hum Genet 73:1162–1169

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

Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5:89–100