European Journal of Human Genetics (2009) 17, 644 – 650

www.nature.com/ejhg

ARTICLE

A composite-likelihood approach for identifying polymorphisms that are potentially directly associated with disease

Joanna M Biernacka*,[1] and Heather J Cordell[2]

[1]*Division of Biostatistics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA;* [2]*Institute of Human Genetics, Newcastle University, Newcastle upon Tyne, UK*

If a linkage signal can be fully accounted for by the association of a particular polymorphism with the disease, this polymorphism may be the sole causal variant in the region. On the other hand, if the linkage signal exceeds that explained by the association, different or additional directly associated loci must exist in the region. Several methods have been proposed for testing the hypothesis that association with a particular candidate single-nucleotide polymorphism (SNP) can explain an observed linkage signal. When several candidate SNPs exist, all of the existing methods test the hypothesis for each candidate SNP separately, by fitting the appropriate model for each individual candidate SNP. Here we propose a method that combines analyses of two or more candidate SNPs using a composite-likelihood approach. We use simulations to demonstrate that the proposed method can lead to substantial power increases over the earlier single SNP analyses.
European Journal of Human Genetics (2009) **17,** 644–650; doi:10.1038/ejhg.2008.242; published online 17 December 2008

Keywords: fine mapping; association; linkage

Introduction

Genetic mapping studies often reveal a region of linkage containing a number of disease-associated polymorphisms. Although, historically, linkage analysis was not as successful at identifying gene-harboring regions for complex traits as had perhaps been hoped, some 'candidate' regions were originally identified or implicated by linkage analysis (eg the *NOD2/CARD15* gene in Crohn's disease,[1] and the location of the complement factor *H* gene in age-related macular degeneration[2]). Other loci show significant evidence of linkage in addition to association (eg human leucocyte antigen (HLA) in type 1 diabetes[3]). Full understanding of the disease predisposing genetic variations is still not known for many of these regions. As pointed out by Clerget-Darpoux and Elston[4] one of the advantages of family-based studies is that they can provide different and complementary information from case/control association studies, namely information regarding patterns of allele sharing among different types of relative. This information can allow one to distinguish between different underlying models that are equally consistent with an observed association. The method we propose here exploits this kind of extra information that is available from family data.

A marker in a disease-linked region may be associated with the disease because it is 'causal' and thus has a direct influence on disease susceptibility. Alternatively, a marker may be indirectly associated with disease as a result of being in linkage disequilibrium (LD) with a causal polymorphism. Several methods have been proposed that can help distinguish polymorphisms that may be directly

*Correspondence: Dr JM Biernacka, Division of Biostatistics, Department of Health Sciences Research, Mayo Clinic, Harwick 7, 200 First Street SW, Rochester, MN 55905, USA.
Tel: +1 507 538 5274; Fax: +1 507 284 9542;
E-mail: biernacka.joanna@mayo.edu
Received 18 June 2008; revised 10 October 2008; accepted 13 November 2008; published online 17 December 2008

associated with a trait from those that are indirectly associated because of LD. One approach is to test whether association of the candidate polymorphisms with the disease can fully explain the observed linkage signal. If a particular variant is the only causal polymorphism in the region, then association with this variant should be able to explain all the linkage in the region. On the other hand, if the variant is not the causal polymorphism, or is not the only causal polymorphism in the region, evidence of linkage should exceed that explained by the association with this variant.

A method for testing whether a candidate single-nucleotide polymorphism (SNP) can fully explain an observed linkage signal proposed by Li *et al*[5] has received much attention.[6] This method tests the null hypothesis that a particular variant can explain all of the linkage in the region *versus* the alternative that it cannot. Li *et al*[5] model the likelihood of the marker data conditional on the trait data for a sample of affected sib pairs (ASPs), with disease SNP penetrances and disease locus-candidate SNP haplotype frequencies as parameters. Assume that we are interested in testing the null hypothesis for a candidate SNP A. Li *et al*[5] base inference on the likelihood

$$\Pr(M_S, C_S|\text{ASP}),$$

where M_S denotes the marker genotypes of the sibs, and C_S denotes the candidate SNP genotypes of the sibs. Assuming there is one causal SNP in the region, the likelihood of the sibs' genotypes at a series of markers and the candidate SNP, conditional on the sibs' affection status, can be written as a function of the penetrances of the corresponding disease locus genotype and disease-SNP-candidate-SNP haplotype frequencies. By restricting these haplotype frequencies appropriately, a model corresponding to complete LD can be fit. A likelihood ratio statistic can then be constructed to test whether the candidate SNP and the 'causal SNP' are in complete LD, implying that either the candidate SNP or a polymorphism in complete LD with it may account fully for the linkage signal. Rejection of the null hypothesis leads to the conclusion that other relevant polymorphisms exist in the region.

Biernacka and Cordell[7] also considered modeling linkage and association jointly, with additional conditioning on parental genotypes. For a sample of ASPs and their parents genotyped at a set of markers plus a candidate SNP, they modeled

$$\Pr(M_S, C_S|M_P, C_P, \text{ASP}),$$

where M_P denotes the marker genotypes of the parents, M_S denotes the marker genotypes of the sibs, and C_P and C_S denote the candidate SNP genotypes of the parents and sibs, respectively. They parameterized this likelihood in terms of two relative risk parameters:

$$\text{RR}_{11} = \frac{\Pr(\text{disease}|g_D = 11)}{\Pr(\text{disease}|g_D = 22)} \quad \text{RR}_{12} = \frac{\Pr(\text{disease}|g_D = 12)}{\Pr(\text{disease}|g_D = 22)},$$

where g_D is the genotype at the disease locus, and the two LD parameters:

$$\delta_1 = \Pr(d = 1|c = 1) \quad \delta_2 = \Pr(d = 1|c = 2),$$

where d and c represent alleles on the disease SNP-candidate SNP haplotype. These LD parameters describe the conditional haplotype frequencies, that is, the probability of the high-risk allele '1' at the disease locus, given the allele at the candidate SNP on the haplotype. If allele '1' at the candidate SNP always occurs on haplotypes with allele '1' at the disease SNP, then $\delta_1 = 1$ and $\delta_2 = 0$, whereas if allele '2' at the candidate SNP always occurs on haplotypes with allele '1' at the disease SNP, $\delta_1 = 0$ and $\delta_2 = 1$. Unlike the method of Li *et al*,[5] this likelihood does not require the prespecification or estimation of marker or candidate SNP allele frequencies. Biernacka and Cordell[7] proposed using a likelihood ratio statistic to test the null hypothesis that the candidate SNP is the sole causal polymorphism, or is in complete LD with the sole causal polymorphism in the region, and therefore association with the candidate SNP can fully account for the linkage signal. In this context, 'complete LD' was defined as the situation of one-to-one correspondence between the alleles at two SNPs on a haplotype, that is $(\delta_1, \delta_2) = (1, 0)$ or $(\delta_1, \delta_2) = (0, 1)$. In terms of the widely used LD parameters D' and r^2, this definition of complete LD implies that $D' = 1$ and $r^2 = 1$. Biernacka and Cordell[7] referred to this approach as Li-cpg (cpg denotes conditional on parental genotypes). Empirical P-values for the Li-cpg likelihood ratio statistic can be estimated by simulation.[7]

In the methods of Li *et al*[5] and Biernacka and Cordell,[7] the null hypothesis that association with a given candidate SNP fully explains the observed linkage can be tested individually for each of several candidate SNPs in a region. The separate analysis of several candidate SNPs leads to repetitive estimation of the same disease SNP relative risk parameters. To improve efficiency and therefore power of these approaches, we propose combining data for all the candidate SNPs using a single model, by a composite-likelihood approach.[8]

Methods

Assume that in a small chromosomal region several candidate SNPs associated with the disease have been genotyped. For each of the candidate SNPs we can test the null hypothesis that association with this candidate SNP fully explains the observed linkage at the candidate SNP location. Thus, by separately analyzing data for two SNPs, for example, we can test the null hypotheses

H_0^1= association with the first candidate SNP fully explains the observed linkage and
H_0^2= association with the second candidate SNP fully explains the observed linkage, etc.

In the Li-cpg approach, genotype data from a set of markers and a single candidate SNP are used to estimate two relative risk and two LD (δ) parameters in a likelihood framework, as described above. Analysis of a second candidate SNP is parameterized in terms of the same two relative risk parameters (RR_{11} and RR_{12}), and two new LD parameters (measuring the LD between the disease locus and the new candidate SNP). Although the separate analyses test hypotheses about different candidate SNPs, the same two RR parameters are repetitively estimated by fitting likelihood models for all candidate SNPs separately. Similarly, analysis of several candidate SNPs using the approach proposed by Li *et al*[5] would also require repetitively fitting the same likelihood model to data for each candidate SNP individually. To improve inference, we propose combining data for all the candidate SNPs using a single model, by a composite-likelihood approach.[8] A composite log likelihood is formed by adding together individual component log likelihoods, each of which corresponds to a marginal or conditional event. Parameter estimates can then be obtained either by maximizing the composite log likelihood or by solving the composite score equation, where the composite score function is a sum of the possibly correlated component score functions. The main advantage of composite-likelihood methods is that they provide a substitute method of estimation when the full likelihood is difficult to calculate. In the current context, the full likelihood would be difficult to calculate because it would require specification of the full LD structure between all candidate SNPs and the unknown disease locus, and estimation of all LD parameters.

Assume that k tightly linked candidate SNPs and a set of flanking markers have been genotyped for n ASP families. Recall that in the single SNP analysis of candidate SNP j, Biernacka and Cordell[7] model the likelihood for the ith family as:

$$L_i^j = \Pr(M_{i\mathrm{S}}, C_{i\mathrm{S}}^j | M_{i\mathrm{P}}, C_{i\mathrm{P}}^j, \mathrm{ASP}),$$

where $M_{i\mathrm{P}}$, $M_{i\mathrm{S}}$, $C_{i\mathrm{P}}^j$ and $C_{i\mathrm{S}}^j$ denote the marker genotypes of the parents, marker genotypes of the sibs, and the jth candidate SNP genotypes of the parents and sibs, for the ith family, respectively. This likelihood is a function of the two relative risk parameters RR_{11} and RR_{12}, and two LD parameters δ_1^j and δ_2^j that describe the LD between the jth candidate SNP and the disease SNP (see Introduction section). In the composite-likelihood approach, we multiply these single-SNP likelihoods to get the composite-likelihood contribution for each family, and then multiply the contributions for all ASP families to get the overall composite likelihood:

$$\prod_{i=1}^{n} \prod_{j=1}^{k} L_i^j.$$

The composite likelihood is parameterized in terms of two RR parameters as well as $2k$ δ parameters, δ_i^j for $i = 1, 2$ and $j = 1, \ldots, k$.

In contrast to a full-likelihood approach for joint modeling of the effects of all candidate loci, our composite-likelihood approach does not model the LD between candidate SNPs. Given the LD between two candidate SNPs, the δ parameters could be restricted in the composite likelihood. However, this would lead to a complex requirement for haplotype reconstruction. Therefore, in the composite-likelihood approach described here, the LD parameters are not constrained other than being restricted to the interval (0, 1). The alternative approach of restricting the δ parameters according to the observed LD between candidate SNPs would result in a method similar to the haplotype extension of the Li-cpg approach proposed by Biernacka and Cordell.[7]

The single-locus hypothesis for each candidate SNP can be tested, while incorporating information from all other candidate SNPs, using the composite likelihood. For example, to test the effect of candidate SNP 1, we may consider the null hypothesis:

H_0^1: candidate SNP 1 is the sole causal variant in the region, or is in complete LD with the sole causal variant in the region; that is association with SNP 1 fully explains the observed linkage.

In terms of the parameters, this can be stated as:

H_0^1: $(\delta_1^1, \delta_2^1) = (0, 1)$ or $(\delta_1^1, \delta_2^1) = (1, 0)$; with the remaining δ parameters freely estimated, restricted to the interval [0, 1].

Similar hypotheses can be tested for the second candidate SNP, and indeed for each subsequent candidate SNP.

As the composite likelihood is constructed by multiplying nonindependent likelihood components, the resulting (composite) likelihood ratio statistic does not have a well-defined χ^2-distribution. We therefore estimate P-values for composite-likelihood ratio tests of the hypotheses described above (H_0^1, H_0^2, etc.) by simulation. This is carried out by generating a large number of datasets under the null hypothesis, calculating the test statistic for each of these datasets, and using these values of the statistic to estimate its null distribution. Assume that a set of candidate SNPs exists, and we first aim to test the null hypothesis H_0^A that SNP A is the sole causal variant, or is in complete LD with the sole causal variant, in the region. We simulate data under this null hypothesis as follows:

1. Fix SNP A genotypes of all sibs and parents at the observed values. Also fix parental genotypes at all remaining candidate SNPs and markers at the observed values.
2. For each ASP, sample the identity by descent (IBD) configuration at SNP A, given the observed SNP A genotypes of the ASP and their parents. Recall that we assume tight linkage between the candidate SNPs and

the true disease locus, so that IBD sharing at SNP A is assumed to be equal to IBD sharing at all other candidate SNPs and at the true causal SNP.

3. Next, we assign a set of candidate SNP haplotypes to each individual, and from the assigned haplotypes determine the children's alleles at the remaining candidate SNPs. The haplotypes for the families are assigned as follows: (i) Infer the probabilities of all possible haplotypes for each family, given the fixed candidate SNP genotypes, P_{hap} (recall that the fixed candidate SNP genotypes include all candidate SNP genotypes of the parents, and the test SNP genotypes of the sibs). The probabilities of the different haplotype configurations for each family may be calculated, for example, using the program ZAPLO.[9] (ii) Calculate the ASP IBD sharing distribution for each set of haplotypes, $P_{\text{IBD|hap}}$. (iii) Use these IBD sharing probabilities ($P_{\text{IBD|hap}}$) and the haplotype probabilities from ZAPLO (P_{hap}), and apply Bayes' rule to calculate the probability of each haplotype set, given the IBD status at the candidate loci and the fixed candidate SNP alleles. (iv) A set of haplotypes is then randomly selected for the family according to the conditional haplotype probabilities of each possible haplotype set as determined in step (iii).
4. Generate IBD status at the markers, conditional on the IBD status at SNP A and the intermarker distances.
5. Generate marker data for children, given the marker IBD status and the fixed parental genotypes at the markers.

This data generation process is repeated a large number of times. For each generated dataset the composite-likelihood ratio test statistic is calculated. The empirical P-value is then estimated as the proportion of test statistics that exceed the test statistic calculated from the original dataset. Note that this procedure has to be repeated for each candidate SNP that we wish to test.

Simulation study

Using simulations, we compared the new composite-likelihood approach to the Li-cpg[7] approach. Comparison of the Li-cpg with the original Li method has previously been carried out.[7] Recall that with the original Li-cpg approach, the likelihood for each candidate is maximized separately to test each candidate for complete LD with a sole causal variant in the region. To compare the power of the methods, we use them to test the same null hypotheses, that association with a particular candidate SNP can explain the observed linkage. All presented simulation results are based on 500 data replicates, with a sample size of 500 ASPs.

Data were generated for haplotypes composed of three SNPs: $A-B-C$, as well as five flanking markers. The markers were equidistant, spaced at 2.5 cm intervals, and each marker had four equally frequent alleles. Haplotype $A-B-C$ was located between the third and fourth marker, 0.2 cm from the third marker. The five markers were in linkage equilibrium with one another and with the $A-B-C$ haplotype.

The data generating models used in our simulations are shown in Table 1. For models 1 and 3–5, the third SNP (SNP C) is the causal SNP, with a multiplicative allele effect. For model 2, SNP B is the causal SNP. Under model 1 both loci A and B are in full LD with the disease locus. Therefore, association with either of these two loci should fully account for the observed linkage (H_0^A and H_0^B are both true). Under model 2, locus A is not in full LD with the disease locus, whereas locus B is in full LD, and therefore H_0^B is true, but H_0^A is not. Under model 3 neither H_0^A nor H_0^B is true, but the first locus is in lower LD with the disease locus than is the second locus. Under model 4 both loci A and B have the same level of (incomplete) LD with the causal locus. For each model, SNPs A and B were analyzed using a two-SNP composite-likelihood approach in which data from both A and B is used when testing either SNP individually.

When two candidate SNPs under consideration are in very high or full LD with one another, we may expect little to be gained by combining data from these two loci in a composite-likelihood analysis. In fact, one may expect some loss in efficiency because of fitting a model with a greater number of parameters. Therefore we also considered model 5, which is similar to model 4, except that the level of LD between the two candidate loci is different. Under model 4, the two candidates have a D' of 0.52 with

Table 1 Data generating models

Model	Disease SNP	SNP $A-B-C$ haplotype frequencies[a]	SNP $A-B-C$ haplotype risks[a]	LD between SNPs		
				D'_{AB}, r^2_{AB}	D'_{AC}, r^2_{AC}	D'_{BC}, r^2_{BC}
1	A, B or C[b]	(0.70, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.30)	(0.15, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.3)	1.00, 1.00	1.00, 1.00	1.00, 1.00
2	B	(0.60, 0.00, 0.05, 0.00, 0.10, 0.00, 0.00, 0.25)	(0.15, 0.0, 0.3, 0.0, 0.15, 0.0, 0.0, 0.3)	0.74, 0.44	1.00, 0.62	1.00, 0.78
3	C	(0.35, 0.02, 0.03, 0.10, 0.10, 0.03, 0.02, 0.35)	(0.1, 0.3, 0.1, 0.3, 0.1, 0.3, 0.1, 0.3)	0.48, 0.23	0.52, 0.27	0.80, 0.64
4	C	(0.30, 0.04, 0.08, 0.08, 0.08, 0.08, 0.04, 0.30)	(0.1, 0.3, 0.1, 0.3, 0.1, 0.3, 0.1, 0.3)	0.36, 0.13	0.52, 0.27	0.52, 0.27
5	C	(0.38, 0.12, 0.00, 0.00, 0.00, 0.00, 0.12, 0.38)	(0.1, 0.3, 0.1, 0.3, 0.1, 0.3, 0.1, 0.3)	1.00, 1.00	0.52, 0.27	0.52, 0.27
6	$A-B-C$ haplotype	(0.35, 0.02, 0.03, 0.10, 0.10, 0.03, 0.02, 0.35)	(0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.3)	0.48, 0.23	0.52, 0.27	0.80, 0.64

Abbreviations: LD, linkage disequilibrium; SNP, single-nucleotide polymorphism.
[a]Haplotype frequencies and risks are shown for haplotypes (111, 112, 121, 122, 211, 212, 221, 222).
[b]Under model 1 the three SNPs are in full LD, and therefore any one of them can be considered as the 'causal' SNP.

the underlying causal locus, and with one another. Under model 5, each of the candidate loci has the same level of LD with the causal locus as in model 4 ($D' = 0.52$); however, under model 5, the two candidates are in full LD with each other (rather than being in incomplete LD with $D' = 0.52$).

An important consideration is how many candidate SNPs should be combined using the composite-likelihood approach. We expect that potential gains in power obtained by including an additional locus in the analysis may diminish as more loci are included. The introduction of a large number of LD (δ) parameters may lead to a loss of efficiency, so that eventually, subsequent addition of candidate loci may no longer lead to power improvements. Therefore, for models 1 and 3, we also carried out simulations in which we analyzed three candidate SNPs (A, B and C) using a three-SNP composite-likelihood approach, in which data from all three SNPs were used when performing the individual test at any one SNP. We also used the three-SNP composite-likelihood approach for an additional model, model 6, under which only haplotype 2–2–2 is associated with increased risk. This could represent a model with a fourth SNP, say SNP D, which is the underlying causal locus, such that the high-risk allele at D only occurs on the 2–2–2 (locus A–B–C) haplotype. In this case, association with none of SNPs A, B or C, individually, can fully explain the observed linkage, because none of them is the sole causal locus, nor is any one of them in perfect LD with SNP D. Therefore this model can be used to assess power of the composite-likelihood approach for three candidate SNPs, none of which is the true causal locus.

The simulation results (Table 2) indicate that substantial gains in power for tests of the null hypothesis that association with a particular SNP can explain an observed linkage signal can be achieved by combining data from two candidate SNPs using the composite-likelihood approach described here. Under model 1, with both candidate SNPs in perfect LD with the true causal locus, correct type 1 error is obtained for analysis of each candidate SNP using the two-SNP composite likelihood. Under model 2, type 1 error of the composite-likelihood approach is again correct for SNP B, which is in full LD with the causal SNP. Meanwhile, for SNP A, substantial gains in power are observed for the composite-likelihood approach relative to the single SNP analysis. In fact, in all our simulations power was greater for both candidate SNPs with the two-SNP composite-likelihood method. A comparison of results under models 4 and 5 showed an expected trend. Recall that the level of LD between the candidate SNPs and the true causal locus is same for these two models. The models differ in the level of LD between the two candidate SNPs: Under model 4 the D' between the two candidate SNPs is 0.52, whereas under model 5 they are in full LD. As expected, the gain in power for the composite-likelihood method seen under model 5 is lower than under model 4.

Simulations with three candidate loci analyzed using the composite-likelihood method showed that whereas the type 1 error rates remained close to their nominal 5% rate, inclusion of a third locus in the analysis could lead to further power increases (see Table 2, models 1 and 3). However, as expected, under some models (eg model 6) inclusion of a third candidate SNP in the composite

Table 2 Simulation results

			Power/type 1 error[a]		
Model	*SNP tested*	*LD with disease SNP*	*Single SNP*	*Two-SNP composite*	*Three-SNP composite*
1	*A*	$D' = 1.0$, $r^2 = 1.0$	**0.044**	**0.050**	**0.054**
	B	$D' = 1.0$, $r^2 = 1.0$	—	**0.058**	**0.041**
	C	$D' = 1.0$, $r^2 = 1.0$	—	—	**0.046**
2	*A*	$D' = 0.80$, $r^2 = 0.44$	0.145	0.524	—
	B	$D' = 1.0$, $r^2 = 1.0$	**0.044**	**0.046**	—
3	*A*	$D' = 0.52$, $r^2 = 0.27$	0.644	0.916	0.994
	B	$D' = 0.80$, $r^2 = 0.64$	0.269	0.308	0.758
	C	$D' = 1.0$, $r^2 = 1.0$	—	—	**0.030**
4	*A*	$D' = 0.52$, $r^2 = 0.27$	0.564	0.714	—
	B	$D' = 0.52$, $r^2 = 0.27$	—	0.706	—
5	*A*	$D' = 0.52$, $r^2 = 0.27$	0.564	0.624	—
	B	$D' = 0.52$, $r^2 = 0.27$	—	0.640	—
6	*A*	$D' = 1.0$, $r^2 = 0.54$	0.326	0.424	0.364
	B	$D' = 1.0$, $r^2 = 0.54$	—	0.402	0.336
	C	$D' = 1.0$, $r^2 = 0.54$	—	—	0.308

Abbreviations: LD, linkage disequilibrium; SNP, single-nucleotide polymorphism.
For single-SNP analysis, redundant results are not shown. For example, under model 1, as all the SNPs have the same relationship to the disease (ie the model is symmetric), the observed type 1 error is shown only for SNP A.
For each scenario, the two-SNP composite-likelihood analyses are based on analyses SNPs A and B.
The three-SNP composite-likelihood analysis was only performed under models 1, 3 and 6.
[a]Type 1 errors are shown in bold.

likelihood could lead to power reductions relative to use of the two-SNP composite likelihood.

Discussion

We have described a composite-likelihood approach to test the hypothesis that association with a particular SNP can explain an observed linkage result, for a number of candidate SNPs. Methods for assessing whether a particular SNP may be the sole causal variant in a region tend to be formulated in terms of the null hypothesis that the candidate is the sole causal SNP in the region.[5,7,10] These methods are sometimes criticized over the fact that low power can lead to failure to reject the null hypothesis, and therefore to the false conclusion that the sole causal variant in a region has been identified.[6] Biernacka and Cordell[7] stress the fact that failure to reject the null hypothesis should not be interpreted as indicating that the sole causal variant has been definitively identified; and further discuss the fact that expecting a statistical method to enable us to make such a conclusion is not reasonable. Nevertheless, given the importance of power in these studies to detect the fact that other 'causal' variations do exist in the region, approaches that have the potential to increase this power are of great interest. Simulations demonstrate that substantial gains in power can be achieved using the composite-likelihood approach introduced in this paper, as compared to a similar approach that analyzes each candidate SNP individually.

Introduction of more SNPs to our model does not pose serious computational problems, aside from the usual difficulties related to fitting statistical models with a high number of parameters (although with the sample sizes available in most genetic studies, we anticipate that analysis of 10–20 SNPs should not pose computational problems). We also emphasize that we do not expect this method to be used with numerous SNPs in a region (eg all tag SNPs genotyped in a gene) but rather only the associated SNPs that are candidates for being the 'causal' variant in a region. Aside from rare exceptions such as associations of SNPs in the HLA region with several autoimmune disorders, there are usually only a few strongly associated SNPs in a region that are good candidates for this type of analysis.

Our composite-likelihood approach combines data from several tightly linked candidate SNPs, when assessing the potential causal role of each individual SNP. The simulation results demonstrated that, although including several candidate SNPs in a single analysis by the composite-likelihood approach can increase power, including further additional SNPs does not necessarily improve power. An interesting question to consider would be whether one could attempt to determine an optimal number or set of candidate SNPs to use in the analysis. We propose using the following model selection procedure to address this question. For a given SNP under test (SNP A, say) calculate the composite-likelihood test statistics obtained when using data from SNP A alone, when using data from SNP A plus each other candidate SNP (ie a two-locus analysis), when using data from SNP A plus each other set of two candidate SNPs (ie a three-locus analysis) and so on. Each test uses a slightly different set of information to test the same hypothesis: namely that SNP A is the only causal variant in the region. Using the simulation procedure described earlier, we may simulate a large number (eg 1000) of replicates of data under this null hypothesis, each replicate of which may be analyzed using the same sequence of tests as in the real data. For each individual test in the real data, an empirical P-value may be obtained by considering how often the observed test statistic exceeded the 1000 simulated values. Similarly, for each individual test in each simulated replicate, an empirical P-value may be obtained by considering how often the simulated test statistic exceeded the 999 other simulated values. Denote by $p\ \min_{\text{real}}$ the minimum empirical P-value observed in the sequence of tests carried out in the real data, and by $p\ \min_i$ the minimum empirical P-value observed in the sequence of tests for replicate i. An empirical P-value for the test with the strongest significance in the real data can now be obtained by observing how often (in the 1000 permuted replicates) $p\ \min_i$ is less than $p\ \min_{\text{real}}$.

The composite-likelihood approach applied here is very general, and could easily be applied to other similar problems or methods. For example, it could be used with the generally more powerful method of Li *et al*[5] implemented in the software LAMP.[11] However, difficulties in implementation may arise as a result of the need for empirical P-value estimation. As the likelihood modeled in LAMP does not condition on parental genotypes, these genotypes would not be fixed when data are generated under the null hypothesis to establish the empirical distribution of the test statistic. Therefore, the repeated generation of datasets under the null hypothesis would become much more computationally demanding. Note that in previous simulations[7] the method of Li *et al*[5] implemented in LAMP generally outperformed the Li-cpg method with a gain in power of around 5–10%, which is considerably less than the gain in power of up to 50% obtained here by use of the composite-likelihood Li-cpg approach.

Sun *et al*[10] described a method for testing the same hypothesis (that a candidate locus is the sole causal polymorphism in a region), which also relied on testing each candidate locus separately. In their paper they explained that the results could be used to construct confidence sets of markers that may be able to explain an observed linkage result by simply including in the confidence set all markers that are not rejected as possibly

being the sole causal variant in the region. Our Li-cpg approach can be used in a similar way, and, because the composite Li-cpg is more powerful than the original Li-cpg, more markers will be rejected and therefore fewer markers will end up in the confidence set, resulting in potentially considerably narrower intervals. Similarly, the proposed composite-likelihood approach may lead to benefits in precision and accuracy of estimates of the disease locus genotype relative risks. Properties of the relative risk estimates that can be calculated using LAMP or the Li-cpg approach, as well as using the new composite-likelihood approach, have not been thoroughly investigated, although previous work[7] suggests that these quantities may not be estimated very accurately. Nevertheless, these are important parameters, not just in the statistical model, but also of biological relevance. Although beyond the scope of this paper, further examination of their properties would be of interest.

The composite-likelihood approach may also be compared to the haplotype extension of the Li-cpg method proposed by Biernacka and Cordell,[7] keeping in mind that the two tests would normally be used for testing different hypotheses. The haplotype method tests whether a haplotype composed of the two SNPs may be in complete LD with the sole causal variant in the region, whereas the composite-likelihood approach described here is testing whether a single candidate SNP may be in complete LD with the sole causal variant. However, by properly constraining the parameters for the null likelihood of the haplotype method, the haplotype likelihood could be used to test the same single SNP hypotheses as the single SNP- and composite-likelihood approaches.

Although composite-likelihood methods have been applied to a number of genetics problems,[12–15] their use is not very widespread. Given the complex LD structures in SNP data, composite-likelihood methods may offer a relatively simple means of dealing with these often difficult to model correlations. The present application of a composite-likelihood approach has provided a demonstration of this concept.

Acknowledgements

This research was funded by the Wellcome Trust, grant reference 074524.

References

1 Hugot JP, Laurent-Puig P, Gower-Rousseau C *et al*: Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 1996; **379**: 821–823.
2 Weeks DE, Conley YP, Tsai H-J *et al*: Age-related maculopathy: a genomewide scan with continued evidence of susceptibility loci within the 1q31, 10q26, and 17q25 regions p174. *Am J Hum Genet* 2004; **75**: 174–189.
3 Davies JL, Kawaguchi Y, Bennett ST *et al*: A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* 1994; **371**: 130–135.
4 Clerget-Darpoux F, Elston RC: Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Hum Hered* 2007; **64**: 91–96.
5 Li M, Boehnke M, Abecasis GR: Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. *Am J Hum Genet* 2005; **76**: 934–949.
6 Yang Q, Biernacka JM, Chen MH *et al*: Group 4: using linkage and association to identify and model genetic effects. *Genet Epidemiol* 2007; **31** (Suppl 1): S34–S42.
7 Biernacka JM, Cordell HJ: Exploring causality via identification of SNPs or haplotypes responsible for a linkage signal. *Genet Epidemiol* 2007; **31**: 727–740.
8 Lindsay BG: Composite likelihood methods. *Contemp Math* 1988; **80**: 221–239.
9 O'Connell JR: Zero-recombinant haplotyping: applications to fine mapping using SNPs. *Genet Epidemiol* 2000; **19**: S64–S70.
10 Sun L, Cox NJ, McPeek MS: A statistical method for identification of polymorphisms that explain a linkage result. *Am J Hum Genet* 2002; **70**: 399–411.
11 Li M, Boehnke M, Abecasis GR: Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am J Hum Genet* 2006; **78**: 778–792.
12 Devlin B, Risch N, Roeder K: Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* 1996; **36**: 1–16.
13 Xiong M, Guo SW: Fine-scale genetic mapping based on linkage disequilibrium: theory and application. *Am J Hum Genet* 1997; **60**: 1513–1531.
14 Rannala B, Slatkin M: Methods for multipoint disease mapping using linkage disequilibrium. *Genet Epidemiol* 2000; **19** (Suppl 1): S71–S77.
15 Morton N, Maniatis N, Zhang W, Ennis S, Collins A: Genome scanning by composite likelihood. *Am J Hum Genet* 2007; **80**: 19–28.