ARTICLE

Searching for Genotype-Phenotype Structure: Using Hierarchical Log-Linear Models in Crohn Disease

Juliet M. Chapman,[1] Clive M. Onnie,[2] Natalie J. Prescott,[2] Sheila A. Fisher,[2] John C. Mansfield,[3] Christopher G. Mathew,[2] Cathryn M. Lewis,[2] Claudio J. Verzilli,[1] and John C. Whittaker[1,*]

There has been considerable recent success in the detection of gene-disease associations. We consider here the development of tools that facilitate the more detailed characterization of the effect of a genetic variant on disease. We replace the simplistic classification of individuals according to a single binary disease indicator with classification according to a number of subphenotypes. This more accurately reflects the underlying biological complexity of the disease process, but it poses additional analytical difficulties. Notably, the subphenotypes that make up a particular disease are typically highly associated, and it becomes difficult to distinguish which genes might be causing which subphenotypes. Such problems arise in many complex diseases. Here, we concentrate on an application to Crohn disease (CD). We consider this problem as one of model selection based upon log-linear models, fitted in a Bayesian framework via reversible-jump Metropolis-Hastings approach. We evaluate the performance of our suggested approach with a simple simulation study and then apply the method to a real data example in CD, revealing a sparse disease structure. Most notably, the associated NOD2.908G→R mutation appears to be directly related to more severe disease behaviors, whereas the other two associated *NOD2* variants, 1007L→FS and 702R→W, are more generally related to disease in the small bowel (ileum and jejenum). The ATG16L1.300T→A variant appears to be directly associated with only disease of the small bowel.

Introduction

Many diseases are phenotypically complex, being divided into a number of possibly overlapping disease classes or subsets. We refer to these subsets as "subphenotypes" of the overall phenotype. We consider below the example of Crohn disease (CD; inflammatory bowel disease [MIM 266600]), which has a number of clinical types or behaviors and can also occur at a number of different locations. Treatment of disease as a single affected or unaffected categorization ignores this complexity, and though this is reasonable in the first phase of discovery of genetic associations, it is important in subsequent studies to advance our understanding of how associated genes are related to particular subphenotypes of the overall disease and how these subphenotypes are related to one another.

The main problem in analysis is that subphenotypes of a particular disease are very often highly associated with one another, so that an individual with one subphenotype is much more likely to have another subphenotype, compared to the population as a whole. This means that it can be particularly difficult to localize which gene is causing which subphenotype by the usual univariate approaches, because a gene causing an effect on one phenotype can appear to be having an effect on other correlated phenotypes as well.

Our aim is to deduce which genes are *directly* related to which subphenotypes and which subphenotypes are *directly* influencing one another. In order to do this, we need to define both *direct* and *indirect* associations. An association between two variables is *indirect* if the relationship is *entirely* mediated (or confounded) by the effect of other variables. A *direct* association is therefore one that is not *entirely* influenced (or confounded) by another variable and one that, given the data, represents some true direct relationship between two (or more) variables. Notice that this is equivalent to the concepts of conditional independence and dependence in the literature on undirected graphical models: two variables (or nodes) in a graph are dependent if directly linked by an edge or are conditionally independent if other nodes are present on all paths between them.

Within this paper, we suggest modeling the data jointly by using a Poisson log-linear model,[1,2] which defines a model for the cell counts of a contingency table. Within the following section, we explain briefly how this model can be defined in terms of a series of interaction parameters that equate to direct relationships between variables. We can infer which direct relationships are important or unimportant by determining which interaction parameters are necessary and which can be dropped from the model. Notice that not all log-linear models can be represented as an undirected graph; in particular, a graph corresponding to a model containing only first-order interactions between, say, three variables may be represented as a complete clique, which would wrongly imply the presence of interactions of higher order (three-way in the example here).[3]

Although most multifactorial diseases have moderate numbers of subphenotypes and known associations are relatively rare, the model spaces we expect to encounter are huge because there are very many possible interactions (one for each cell of the contingency table, in fact) and, therefore, there are even more possible models. For this

[1]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK; [2]Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London SE1 9RT, UK; [3]Faculty of Medical Sciences, University of Newcastle, Newcastle upon Tyne NE1 3BZ, UK
*Correspondence: john.whittaker@lshtm.ac.uk
DOI 10.1016/j.ajhg.2008.12.015. ©2009 by The American Society of Human Genetics. All rights reserved.

reason, we focus upon a Bayesian-model averaging scheme. This not only allows us to search the model space more efficiently, but it also means that we are able to include prior information that downweights models with many high order interaction terms, in order to prevent overfitting. Within the Material and Methods section and the Appendix, we give details of the Poisson log-linear model[1] and the reversible-jump Metropolis-Hastings algorithm (RJMH)[4] that we implement. Then, we investigate the efficiency of this approach, compared to usual univariate analyses, through simulations and then apply this method to an example in which we examine the relationships between CD subphenotypes and a number of well-known associated loci. Our simulations show that the RJMH approach can distinguish well between direct and indirect associations and, thus, has advantages over the usual univariate analysis. The application to CD helps to clarify which genes are directly associated to which subphenotypes.

Material and Methods

The Model: Poisson Log-Linear Model

Our aim is to model all variables, i.e., genotypes and subphenotypes, within a single model; therefore, there is no need for a notational distinction between genotypes and subphenotypes. If we assume that we have G genotypes, $X_1, \ldots, X_G$, and S subphenotypes, $Y_1, \ldots, Y_S$, we can pool these into one set of $\mathrm{p} = (G + S)$ variables, $Z_1, \ldots, Z_P$, in which the first G variables are genotypes and the last S variables are subphenotypes. We say that the p^{th} variable has L_p levels (for $\mathrm{p} = 1, \ldots, \mathrm{P}$), and for ease of notation, we label these levels from 0 to $(L_p - 1)$. A genotype may, for example, have $L = 3$ different levels, labeled as 0, 1, 2 (we do not assume Hardy-Weinberg equilibrium).

The data can be displayed within a P-dimensional contingency table, with each dimension corresponding to a different variable. The number of cells within this table is equal to I, the product of the number of levels across all variables ($I = \prod_{p=1}^{P} L_p$). Each cell corresponds to a unique realization of (Z_1, .., Z_P), and we can define cell i as (z_{1i}, …, z_{Pi}).

Our log-linear model assumes that for cell i in 1:I, the observed cell counts n_i have a Poisson distribution with mean μ_i, and that the log of expected cell counts, $\log(\mu_i)$, is given by a linear model including the *baseline* parameter, β_0, the *main effects* of each of the variables, and interaction parameters of various order. The total number of possible parameters is always equal to the number of cells in the contingency table (I).

By dropping parameters from the *saturated* model, which contains all possible interaction parameters, we can decide which interaction parameters are important—that is, which have large posterior probabilities associated to them—for modeling the data. Those interaction parameters that are important within our model can inform us about likely direct (and indirect) associations. Two variables are deemed to be directly associated only if there exists at least one parameter including the two variables that is judged to be important. Two variables are indirectly linked if the set of important parameters include a chain of overlapping direct relationships relating the two variables.

To illustrate this, consider a data set with three binary variables, each with just two levels; 0 and 1. The contingency table has 2 × 2 × 2 = 8 cells, and the saturated model is

$$\begin{aligned}\log(\mu_i) &= \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_{12} z_{1i} z_{2i} + \beta_{13} z_{1i} z_{3i} + \beta_{23} z_{2i} z_{3i} \\ &\quad + \beta_{123} z_{1i} z_{2i} z_{3i},\end{aligned}$$

in which β_0 is the baseline parameter; β_1, β_2, and β_3 are the main effects of variables 1, 2, and 3; β_{12}, β_{13}, and β_{23} are the pairwise interactions between variables (1 and 2), (1 and 3), and (2 and 3); and β_{123} is the three-way interaction between variables (1, 2, and 3). Suppose we search across the set of all models and find evidence that the parameters $\{\beta_0, \beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{23}\}$ are nonzero but that no other parameters are needed for adequate modeling of the data. The first pairwise interaction parameter, β_{12}, tells us that variables 1 and 2 are likely to be directly associated, and β_{23} tells us that variables 2 and 3 are directly associated. Because there is no parameter containing both 1 and 3, we know that these variables are not directly associated. However, they are indirectly associated, because variable 1 is directly associated with variable 2 and variable 2 is directly associated with variable 3.

Therefore, determination of which models are the most probable tells us about direct and indirect relationships between genotypes and subphenotypes, as well as direct genotype-genotype associations and phenotype-phenotype associations. Given the complicated nature of the models considered, we implement a reversible-jump Metropolis-Hastings algorithm to search across the model space.

In fact, the algorithm that we consider is restricted to a subclass of log-linear models known as the set of *hierarchical models*. These are models that contain only parameters for which all implied parameters are also included in the model. Each parameter relates to an interaction between particular levels of one or more variables and, as such, is defined by these variable levels. All subparameters of this parameter are those parameters that are defined by a subset of these variable levels. For example, the subparameters of β_{12}, above, are β_1, β_2, and β_0. This means that β_{12} can only be included in a hierarchical model if these three parameters are also included within the model. The search algorithm is described below, and in further detail within the Appendix.

Search for Important Parameters: Reversible-Jump Metropolis-Hastings

Even with moderate numbers of variables, the space of all possible models is very large. Moreover, many of these models may have similar likelihoods. Hence, choice of a single "best" model is likely to be highly unstable, and model averaging across sampled models is preferable.[5] We adopt a Bayesian approach, which we now briefly describe. Additional details are given in the Appendix.

Denote a model by m, the corresponding set of parameters by β, and the relevant set of variables by Z. We wish to calculate the probability of the model and parameters, given the data, that is the posterior probability of the model and parameters $\mathbb{P}(m, \beta|Z) = const \times L(Z; m, \beta)\ \mathbb{P}(m, \beta)$, in which $L(Z;m, \beta)$ is the likelihood of the data given the parameter values β and the model m and $\mathbb{P}(m, \beta)$ is the joint prior of m and β. This prior distribution defines our prior beliefs about the model and its parameters. An additional benefit of the Bayesian approach is that we can include as prior information our belief that complex models with many complex, high-level interactions are unlikely, which reduces the problem of overfitting.

We use RJMH to approximate the required posterior distribution by sampling from it. The RJMH sampling scheme starts at an initial model and set of parameter values, $m^{(0)}$ and $\beta^{(0)}$. To sample the next model and set of parameters, $m^{(1)}$ and $\beta^{(1)}$, we propose a move from the current state to another model and/or set of parameter values, m^* and β^*, using a proposal function $q(m^*, \beta^*|m, \beta)$. We then accept these proposed values as the next sample with probability equal to the Metropolis-Hastings ratio:

$$MHR = \frac{L(Z; m^*, \beta^*)\mathbb{P}(m^*, \beta^*)}{L(Z; m, \beta)\mathbb{P}(m, \beta)} \times \frac{q(m, \beta \mid m^*, \beta^*)}{q(m^*, \beta^* \mid m, \beta)}.$$

If this new set of values is accepted, the proposed set is accepted as $m^{(1)}$ and $\beta^{(1)}$. Otherwise, the sample value remains equal to the current sample value, i.e., $m^{(1)} = m^{(0)}$ and $\beta^{(1)} = \beta^{(0)}$. It can be shown that this produces a sequence of samples that converge to the required posterior distribution.[4,6] More details about the scheme used are given in the Appendix.

We need to choose a prior distribution for both the parameter values and likely model distribution. As in the research by Dellaportas and Forster,[7] we choose independent normal priors for the values of each parameter included in the model, each with zero mean and precision $1/\tau^2 = 0.001$. In terms of the model prior, we expect sparse models with lower-order interaction parameters, with higher-order interactions rarely included. Therefore, we assign all parameters of a given size (1 to P) equal prior probabilities of being included in the model and allow this probability to decrease rapidly as the size of the interactions decreases, making higher-order terms less likely. The prior for a particular model is, then, formed as the product of these parameter-inclusion prior probabilities, for all parameters in the current model, again favoring sparse models. The code used for fitting the models is available from J.M.C.'s webpage (see Web Resources).

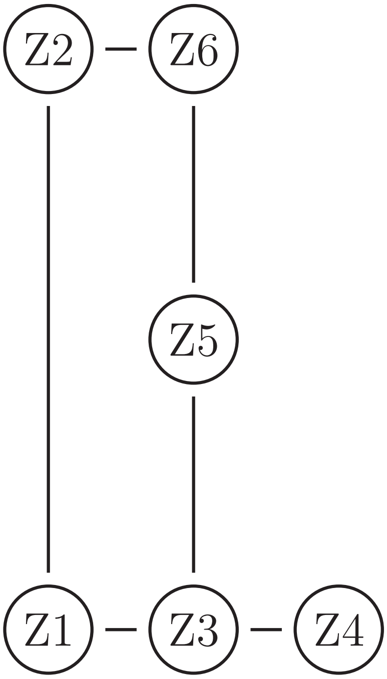

Figure 1. Interaction graph representing the true model underlying the simulated data for binary nodes Z1 to Z6
Lines between nodes represent true pairwise interactions between the two nodes.

Results

Simulation Study

The aim of this simple simulation study is to explore the performance and accuracy of the RJMH approach and, in particular, to illustrate the differences between this approach and a simple univariate analysis, which simply looks for association between any pair of variables. For simplicity, we assume that we have six binary variables, $Z_1, \ldots, Z_6$, and a single true underlying model. We assume that all log-linear main effects are equal to $\log(0.2/(1 - 0.2)) = -1.39$, which gives a frequency of 0.2 within the set of controls. We set most log-linear pairwise interactions to be 0 (i.e., no direct association) and set six of them to be nonzero; namely, those between Z_1 and Z_2 ($\log(2.72) = 1$), Z_1 and Z_3, Z_2 and Z_6 ($\log(1.5) = 0.41$), Z_3 and Z_4, Z_3 and Z_5, and Z_5 and Z_6 ($\log(2) = 0.69$). These pairwise parameters can be thought of as log of the relative risks between the two variables involved. The corresponding model for the mean expected count of cell i (on the log scale) can be written as $\log(\mu_i) = -1.39z_{1i} - 1.39z_{2i} - 1.39z_{3i} - 1.39z_{4i} - 1.39z_{5i} - 1.39z_{6i} + 1z_{1i}z_{2i} + 10.41z_{1i}z_{3i} + 0.41z_{2i}z_{6i} + 0.69z_{3i}z_{4i} + 0.69z_{3i}z_{5i} + 0.69z_{5i}z_{6i}$.

A graphical representation of this model is shown in Figure 1. Those familiar with graphical models should note that this is an interaction graph; lines between nodes represent pairwise interactions.[3]

On the basis of this model, we simulated ten data sets and ran the RJMH method upon each data set, using 30, 000 iterations, dropping the first 10, 000 as burn-in iterations and thinning by 40, leaving a sample of 1000 models. For each data set, we then calculated the posterior probability of inclusion for all pairwise interactions. Table 1 shows these posterior probabilities for each pairwise interaction, within each of the ten data sets. Those in the top six rows are those that are within the true underlying model, i.e., true positives, and those in the bottom nine rows are those that are not within the true model and should not be detected. Careful consideration of the specificity and sensitivity of different cutoff values suggests the use of 0.4 as an appropriate cutoff value for defining important parameters. Across the ten simulated data sets, this equates to a mean specificity of 1 and a mean sensitivity of 0.95. Using this cutoff value of 0.4, we can see that the RJMH approach always gets rid of untrue associations but occasionally misses out on some true associations—namely, the (1,3) interaction in samples 7 and 10, shown in bold font within the table, and the (2,6) interaction in sample 5. Notice, however, considering Table 2, that the (1,3) interaction is also missed by the usual univariate analysis in sample 10, as is the (2,6) interaction within sample 5. Within the univariate analysis, we simply carried out all possible pairwise univariate analyses, using simple score tests. Table 2 shows the results from this standard analysis. Now, if we let any pair with a p value smaller than or equal to 0.05 define a significant finding, we see that although we pick up practically all of the true interactions that the RJMH method picks up, many other false associations are also detected. These p values incorrectly deemed significant are highlighted within the table with bold font, as are the two p values that were

Table 1. Posterior Probabilities of all Pairwise Interactions for Ten Simulated Data Sets

	Posterior Probability									
Interaction	1	2	3	4	5	6	7	8	9	10
(1,2)	1	1	1	1	1	1	1	1	1	1
(1,3)	0.978	0.949	1	0.615	1	1	**0.161**	1	0.95	**0.018**
(2,6)	0.587	0.491	0.664	1	**0.040**	0.935	0.426	0.618	0.484	1
(3,4)	1	1	1	1	1	1	1	1	1	1
(3,5)	1	1	1	1	1	1	1	1	1	1
(5,6)	1	1	1	1	1	1	1	1	1	1
(1,4)	0.01	0.033	0.013	0.034	0.009	0.007	0.013	0.01	0.016	0.041
(1,5)	0.006	0.031	0.014	0.026	0.008	0.012	0.016	0.015	0.012	0.009
(1,6)	0.025	0.01	0.008	0.013	0.012	0.064	0.013	0.025	0.009	0.015
(2,3)	0.013	0.015	0.018	0.026	0.031	0.031	0.018	0.085	0.014	0.011
(2,4)	0.015	0.020	0.009	0.01	0.018	0.013	0.017	0.012	0.008	0.009
(2,5)	0.013	0.015	0.003	0.029	0.013	0.014	0.009	0.128	0.016	0.016
(3,6)	0.005	0.049	0.019	0.026	0.009	0.014	0.026	0.003	0.015	0.036
(4,5)	0.01	0.039	0.002	0.02	0.004	0.005	0.013	0.013	0.009	0.012
(4,6)	0.013	0.012	0.016	0.01	0.02	0.027	0.017	0.011	0.014	0.019

False positives and negatives are highlighted in bold type.

incorrectly deemed nonsignificant. Even if we use the unrealistically stringent Bonferroni correction and consider only those p values smaller than 0.0033 (= 0.05/15) to be significant, we observe that many of these false interactions are still detected. Notice, in particular, that with the univariate method, the (3,6) and (4,5) interactions are often incorrectly found to be significant, which makes sense if one considers the structure of the data, which indirectly links variables 3 and 6, as well as 4 and 5 (see Figure 2). This nicely demonstrates our point that the RJMH approach is able to distinguish well between direct and indirect associations, whereas standard univariate approaches are unable to discriminate easily between the two.

Application to Crohn Disease

The data set consists of 1019 cases and 2757 controls, from numerous British and European sources. CD patients were recruited after ethical review and obtaining of informed consent from Guy's and St. Thomas's Hospital London, St. Mark's Hospital London, and the Royal Victoria Infirmary Newcastle, as previously described by Onnie et al.[8] and Precott et al.[9] The diagnosis of CD was made via established criteria of clinical, radiologic, and endoscopic analysis and from histology reports. Of the population controls, 1371 were obtained from the 1958 British birth cohort (National Child Development Study) and the remaining 1386 were the noninflammatory-disease controls collected at Guy's and St. Thomas's Hospital London (reported in Onnie et al.[10]), the Royal Victoria Infirmary Newcastle, and the European Collection of Cell Cultures (ECACC). A summary of this data is given in Table 3.

CD has a number of subphenotypes, and these can themselves be split into two classes: (1) *location* of disease and (2) *behavior* of disease. CD can occur at any location

Table 2. Pairwise p Values of all Pairwise Interactions for Ten Simulated Data Sets

	Pairwise p Values									
Interaction	1	2	3	4	5	6	7	8	9	10
(1,2)	8.10^{-22}	9.10^{-20}	2.10^{-16}	2.10^{-31}	1.10^{-20}	2.10^{-22}	4.10^{-21}	3.10^{-22}	1.10^{-16}	7.10^{-29}
(1,3)	5.10^{-6}	6.10^{-5}	3.10^{-7}	6.10^{-4}	8.10^{-7}	1.10^{-8}	1.10^{-2}	6.10^{-8}	8.10^{-5}	**0.27**
(2,6)	7.10^{-4}	2.10^{-3}	6.10^{-4}	3.10^{-6}	**0.073**	1.10^{-4}	3.10^{-3}	3.10^{-4}	8.10^{-4}	2.10^{-6}
(3,4)	3.10^{-25}	2.10^{-32}	8.10^{-22}	4.10^{-22}	3.10^{-34}	7.10^{-37}	7.10^{-24}	1.10^{-31}	4.10^{-26}	8.10^{-24}
(3,5)	7.10^{-29}	4.10^{-34}	7.10^{-31}	5.10^{-28}	3.10^{-19}	1.10^{-25}	3.10^{-28}	2.10^{-34}	1.10^{-28}	2.10^{-30}
(5,6)	5.10^{-23}	2.10^{-29}	2.10^{-26}	6.10^{-27}	9.10^{-26}	3.10^{-26}	3.10^{-36}	3.10^{-21}	4.10^{-38}	3.10^{-32}
(1,4)	0.80	0.86	0.18	0.69	0.11	0.059	0.26	0.37	0.07	0.24
(1,5)	0.066	0.17	**0.0046**	**0.035**	0.14	**0.028**	0.52	**0.035**	0.59	0.47
(1,6)	0.052	0.57	0.38	0.053	0.53	0.64	0.67	**0.031**	0.82	0.77
(2,3)	0.24	0.56	**0.0065**	0.96	**0.02**	0.78	0.31	**0.00022**	0.75	0.99
(2,4)	0.87	0.23	0.54	0.80	0.22	0.90	0.29	0.6	0.74	0.94
(2,5)	0.84	0.56	0.29	0.69	0.45	0.25	0.53	0.48	0.059	0.18
(3,6)	**0.0081**	**5.10^{-6}**	**0.028**	0.22	**0.02**	**0.00097**	0.11	**0.00058**	**0.0061**	0.095
(4,5)	**0.00058**	**0.047**	**0.01**	**0.0037**	**0.0092**	**0.011**	0.067	**5.10^{-6}**	**0.032**	**0.003**
(4,6)	0.46	0.18	0.63	0.83	0.12	0.66	0.079	0.37	0.1	**0.037**

False positives and negatives are highlighted in bold type.

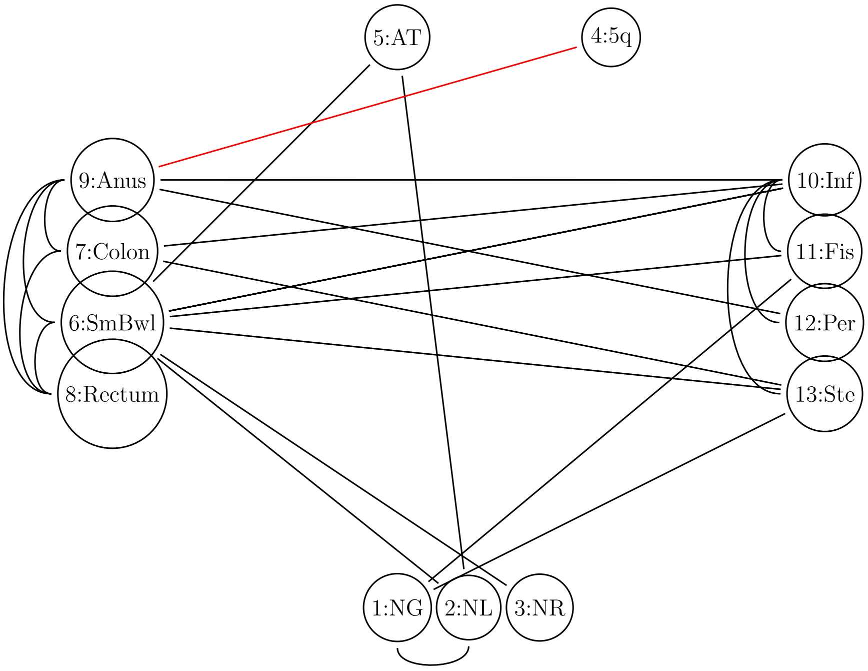

Figure 2. Interaction graph representing the output of the real Crohn disease data

Nodes represent all genotypes and all subphenotypes and lines between nodes represent marginal pairwise associations with posterior probabilities greater than 0.4. NG, NL, NR represent the 3 *NOD2* mutations; 908G > R, 1007L > FS, 702R > W, 5q represents 5q31 mutation, AT the *ATG16L1* gene and Ste, Per, Fis, Inf represent the disease behaviors; stenosing, perianal fistulating, internal fistulating and inflammatory. Notice that any clique in the graph does not imply the presence of higher order interactions, since not all log-linear models are graphical.[3]

along the gastric tract, including the ileum, jejenum, colon, rectum, and anus. For the purposes of our analysis, we will combine diseases of the ileum and jejenum and call this location the small bowel. Therefore, we classify the areas of interest into small bowel, colon, rectum, and anus, and we will denote these by binary variables, which equal 1 when an individual has disease at that place and 0 otherwise. CD can also be classed into different disease behaviors—namely, inflammatory, internal fistulating, perianal fistulating, and stenosing.

Gene-identification studies for CD have been highly successful,[11,12] and we focus here on three of the earliest genes and regions to be associated with CD: *NOD2* (MIM 605956), the 5q31 region (MIM 606348), and *ATG16L1* (MIM 611081). Within these three loci, we model association with the variants 908G→R (rs2066845), 1007L→FS (rs2066847) and 702R→W (rs2066844) in *NOD2*,[13–15] IGR2063 (no rs number assigned) in the 5q31 region,[16,17] and *ATG16L1*.300T→A (rs2241880), all of which are strongly associated with CD in this population.[9]

Table 3. Summary of Crohn Disease Data

Minor-Allele Frequencies		
Locus	Controls (% Missing)	Cases (% Missing)
NOD2.908*G*→*R*	0.010 (17)	0.034 (3)
NOD2.1007*L*→*FS*	0.017 (31)	0.071 (4)
NOD2.702*R*→*W*	0.050 (34)	0.101 (7)
5q31	0.423 (27)	0.482 (8)
ATG16L1	0.488 (58)	0.404 (17)
Subphenotype Frequencies		
Subphenotype	Controls	Cases
small bowel	0	0.727
colon	0	0.566
rectum	0	0.215
anus	0	0.245
inflammatory	0	1
internal fistulating	0	0.215
perianal fistulating	0	0.244
stenotic	0	0.503

These eight binary subphenotype variables and five ternary genotype variables can be jointly represented within a 13-dimensional contingency table. This table has $(2^8)\times(3^5) = 62208$ cells and, therefore, parameters. Notice that, although any subphenotype can occur with any genotype and any location of disease with any disease behavior, those cells for which there is a location of disease but no disease behavior, or vice-versa, are not possible. This means that we must modify the likelihood defined above and restrict the models allowed accordingly. The Appendix gives details of the required modifications to likelihood and proposal distribution.

We apply the RJMH method to this contingency table of data and gain a sample of models drawn from the required posterior distribution. We carried out 50, 000 iterations of this procedure. We judged the procedure to have converged well before the first 10, 000 iterations, which were discarded as burn in. Posterior draws are thinned every 40 iterations, leaving a sample of 1000 models. The posterior probability for the inclusion of each parameter is then simply calculated as the proportion of these models that include that parameter. Following the suggestions of Dellaportas and Forster,[7] we set the variances of the prior and proposal normal distributions of the parameter values as $\tau^2 = 1$ and $\sigma^2 = 2$, respectively. We sampled new parameter values, keeping the model fixed with a probability of 0.25. Because some of the genotype data were missing, we also sampled new missing genotypes every 50 iterations. Varying the value of the prior precision on regression coefficients did not materially change the results. See Appendix for details.

Table 4. Posterior Probabilities of Direct Interactions within the Crohn Disease Data Set

Node	Genotype	Location	Behavior	Posterior Probability
(1,11)	*NOD2*.908*G*→*R*		internal fistulating	1
(5,6)	*ATG16L1*	small bowel		1
(8,9)		rectum, anus		1
(9,10)		anus	Inflammatory	1
(9,12)		anus	perianal fistulating	1
(7,8)		colon, rectum		1
(7,10)		colon	inflammatory	1
(6,8)		small bowel, rectum		1
(6,10)		small bowel	inflammatory	1
(6,13)		small bowel	stenosing	1
(10,12)			inflammatory, perianal fistulating	1
(10,13)			inflammatory, stenosing	1
(9,10,12)		anus	inflammatory, perianal fistulating	1
(3,6)	*NOD2*.702*R*→*W*	small bowel		0.96
(6,9)		small bowel, anus		0.92
(10,11)			inflammatory, internal fistulating	0.917
(7,13)		colon	Stenosing	0.87
(6,11)		small bowel	internal fistulating	0.87
(7,9)		colon, anus		0.811
(2,6)	*NOD2*.1007*L*→*FS*	small bowel		0.785
(1,2)	*NOD2*.908*G*→*R*, *NOD2*.1007*L*→*FS*			0.679
(1,13)	*NOD2*.908*G*→*R*		stenosing	0.603
(2,5)	*NOD2*.1007*L*→*FS*, *ATG16L1*			0.42
(4,9)	5q31	anus		0.398

Each row refers to an interaction found to have posterior probability greater than 0.4. Column 1 contains the node numbers involved within each interaction, columns 2 to 4 contain the names of genotypes, locations, and behaviors involved within each interaction, and column 5 gives the posterior probability of that interaction.

Looking at the posterior probabilities of single models, we find that the maximum posterior probability of any model in the sample is 0.005. This is indicative of the fact that the posterior model distribution is highly dispersed. For this reason, it makes sense to average over all models and simply consider the proportion of samples that include each of the interaction parameters within the model; i.e., consider the marginal probabilities of interaction terms. Table 4 shows the posterior probabilities for all pairwise interactions with a posterior probability greater than 0.4. We have dropped simple main effects from this table, because these all have a posterior probability equal to 1 and do not tell us anything more about the underlying structure of the data. Note that in our model, each genotype is represented as a variable with three levels (i.e., two nonbaseline levels) and, therefore, interactions are not defined only by the variables that interact but also by the level of the variables that are interacting. However, our focus is not upon which levels are interacting but simply upon which variables are interacting, and therefore, we suppress the variable-level information and report only interactions for the level that has the highest posterior probability.

Table 4 may be more easily represented in terms of the graph in Figure 2. As discussed above, it should be stressed, particularly to those familiar with graphical models, that this representation is closely related to interaction graphs rather than to fully fledged undirected graphs, given that each line on the graph refers to a simple pairwise interaction and any subgraph need not necessarily be saturated.

On the basis of Table 4 and Figure 2, we are able to make some interesting observations, not only about the direct associations that are present but also about those direct associations that are absent. We see that the *NOD2*.1007L→FS and 702R→W mutations are directly associated only to disease in the small bowel, whereas the *NOD2*.908G→R mutation seems to be directly related just to the two severe disease behaviors; internal fistulating and stenosing. A number of studies have previously found the association between NOD2 variants and ileal disease to be stronger than that between NOD2 variants and CD in general,[18,19] and this result suggests that, in fact, any general relation between NOD2 and CD may be the consequence of direct association with small-bowel or ileal disease, as well as internal fistulating and stenosing disease behaviors. We find that *ATG16L1* has a direct effect upon disease of the small bowel. This agrees with the findings of Prescott et al.,[9] who found *ATG16L1* to be associated with disease in the ileum but not with disease in the colon. When choosing a posterior probability cutoff of 0.4 to define important interactions, we find that the SNP in the 5q31 region does not appear to be directly related with any other variable (either genotype or subphenotype). However, we do see a direct association between this SNP and anal disease, with a *near important* posterior probability of 0.398, suggesting that perhaps we should not ignore this association, particularly since this has been suggested previously by Armuzzi et al.,[20] who found association only between 5q31 and perianal disease. Many locations of disease are related to other disease locations, as may be expected, given that disease locations are clearly highly correlated with one another. Similarly, all noninflammatory disease behaviors are related to inflammatory disease behavior. There is also evidence of a three-way interaction between disease in the anus, inflammatory disease, and perianal fistulating disease. In terms of direct relationships between location of disease and disease behavior, we find that disease of the small bowel is related to all disease behaviors except perianal fistulating disease, colonic disease is directly related to both inflammatory and stenosing disease, and anal disease is directly associated with inflammatory and perianal fistulating disease.

Discussion

Once genetic associations have been found, we wish to further investigate the influence of the associated genes with disease, and one important aspect of this is determining which genes are related to which subphenotypes. Standard univariate analyses are unable to do this because of high correlations between subphenotypes. Instead, we suggest the use of Bayesian model-selection procedures within the set of hierarchical log-linear models. Our simulation study shows the ability of this approach to distinguish between direct and indirect associations.

The method gives interesting insights into the probable genetic structure of CD. The direct relationships between disease location and disease behavior make biological sense. In terms of direct relationships between genes and disease subphenotypes, it appears that the effects of the *NOD2* mutations may differ, such that the G908R mutation is related to more severe disease behaviors—namely, internal fistulating and stenosing—and the 1007L→FS and 702R→W mutations appear to be directly related only to disease of the small bowel. *ATG16L1* is directly associated to disease in the small bowel, as well as to *NOD2*.1007L→FS, and the locus within the 5q31 region appears to be weakly directly related to anal disease. Although the last association is questionable, it has been detected previously.[20] However, notice that, as discussed earlier, any direct and indirect associations have to be interpreted within the context of the set of variables studied.

We have chosen to use hierarchical log-linear models. It is, of course, possible that the true model is not in this class; for example, a model in which subphenotype is due to a pure interaction between loci is nonhierarchical, because the nearest hierarchical model would include both main effects. If such models are common, a more general class of models might be preferred. Inversely, restriction to a simpler class of models would give increased power when the true model is well approximated by a member of this simpler class. For example, if the three *NOD2* loci act multiplicatively to cause disease, we would be likely to have more power to detect this interaction if we restricted ourselves to multiplicative models, rather than allowing each single locus genotype to have a different effect on the disease. However, at present, little is known about the relationship between multiple genetic factors and phenotypically complex disease, such as CD. We believe that hierarchical log-linear models represent a realistic and flexible set of models and provide an attractive compromise between parsimony and the desire to represent potential biological complexity.

Appendix

Log-Linear Models and the Poisson Likelihood

If I represents the set of all possible cells within this contingency table, n_i the observed count within cell i, and μ_i the expected count within this cell (for $i = 1, \ldots, I$), we can model the contingency table data using a Poisson likelihood, which may be written in the following form:

$$\begin{aligned} L(Z;\mu) &= \prod_{i=1}^{I} \frac{\exp(-\mu_i) \times \mu_i^{n_i}}{n_k} \\ &= const \times \prod_{i=1}^{I} \exp\left(-\mu_i\right) \times \mu_i^{n_i} \\ &= const \times \exp\left(\sum_{i=1}^{I} n_i \times \log(\mu_i) - \sum_{i=1}^{I} \mu_i \right). \end{aligned}$$

It is well known that the Poisson likelihood for cell counts gives equivalent results to the multinomial likelihood for cell probabilities but omits the need to directly normalize across all cells and is therefore a more practical model to use.[1] The model defined above is parameterized by the set of μ_I, and therefore, there is one parameter for each cell; hence, this model represents the most saturated model possible for the data, which cannot inform us about the underlying structure.

A log-linear model, however, allows us to define a model for each expected cell count (μ_i), and this model can then be used for information about the underlying structure. A log-linear model defines each cell mean in terms of a sum of "interaction" parameters present within that cell. Each "interaction" parameter corresponds to a subset of the nonbaseline variable levels, and there are an equal number of these parameters as there are cells in the contingency table. We let j index the set of all possible parameters and let β_j represent interaction parameter j, for $j = 1, \ldots, J$ (where J = I in the case of the saturated model). We also represent cell i by the vector $\mathbf{z_i} = (z_{i1}, \ldots, z_{iP}))$ and let the function $\beta_j(.)$ be equal to β_j if cell i includes parameter j and zero otherwise. The model for the log of the expected cell mean can then be written as:

$$\log(\mu_i) = \sum_{j=1}^{J} \beta_j\left(\mathbf{z}_i\right).$$

Plugging this into the Poisson likelihood, we can re-express our likelihood as:

$$L(Z;\mu) = const \times \exp\left(\sum_{i=1}^{I} n_i \times \sum_{j=1}^{J} \beta_j\left(\mathbf{z}_\mathrm{i}\right) - \sum_{i=1}^{I} \exp \times \left(\sum_{j=1}^{J} \beta_j\left(\mathbf{z}_\mathrm{i}\right) \right) \right).$$

This new parameterization allows us to make use of the flexibility of these log-linear models by dropping parameters from the saturated model, so that we can find the most parsimonious models that fit the data with the fewest degrees of complexity. This equates to our taking a subset m of 1, …, J to define our model and summing across all j in m rather than across all parameters 1, …, J. Note that there exists a correspondence between the parameters of a log-linear model and those of a (perhaps more familiar) logistic-regression model, such that the pairwise interaction between, say, Z_1 and Z_2 is equivalent to the main-effect parameter of the logistic regression of either Z_1 on Z_2 or Z_2 on Z_1. This equivalence extends to interactions of all sizes, so that a log-linear three-way interaction parameter is equivalent to a logistic pairwise interaction between two of the variables, in which the third variable is the outcome of the regression, and so on.

Because we restrict ourselves to the set of *hierarchical models*, all models may be defined unambiguously by the maximal set of parameters that are not subparameters of any other parameters contained in the model. These maximal parameters are known as the generators of the model, and we shall define these by G. Each hierarchical model also has a set of dual generators. These are the set of parameters that are "next up" from the generators. In other words, these are the set of parameters that are *not* contained in the model, but all subparameters of these parameters *are* included in the model. We will denote this set by D. We have decided to restrict attention to the subclass of hierarchical log-linear models, because not only is the model space smaller but the steps of the model-search approach that are proposed make more sense in this scenario and are more likely to be accepted. Both the generators and dual generators are important when this approach is used.

Reversible-Jump Metropolis-Hastings: Proposal Distributions

We have decided to consider only hierarchical models. Following the method of Dellaportas and Forster,[7] we allow three possible proposal steps. The first step is to *drop* a parameter from the model. Because we are interested only in hierarchical models, the only parameters that we can legally drop are the set known as the generators of the model, because the resulting model would otherwise contain parameters for which not all subparameters belong to the model. Therefore, so long as the set of generators is not simply the set containing α only, we will randomly propose one of the generators to drop from the model. The second step is to *add* a parameter to the model. Again, because we are interested only in hierarchical models, the only terms that we can legally add are those within a particular set known as the dual generators of the model. So long as this model is nonempty, we randomly propose one of the dual generators to add to the current model. Because we are adding a term, we also need to propose a value for this parameter. We will simply sample this value from a normal distribution with zero mean and variance σ^2, independent of both the model and the other parameter values. The final proposal type is known as the *null* move, because the model remains unchanged and we simply update the parameter values of the current model. Theoretically, this step is unnecessary for guaranteeing convergence, but it allows us to move around the space more quickly and effectively. Within this step, we simply update all parameters separately, in a random order, using a simple normal proposal with mean equal to the current value of the parameter and variance equal to σ^2. Note that ideally, we would use some form of Gibbs sampling approach, as did Dellaportas and Forster,[7] which means that we would update each parameter every time we carry out the null step. However, this can be impractically slow and time consuming for larger problems. Alternatively, we repeat the random sequence of updates a given number of times, T, in a single "null" step. Within the applications of this paper, we choose $T = 10$. When the null step is not sampled, the drop step and the add step are proposed with equal probability, so long as both are possible; otherwise, the only possible step is selected automatically.

Prior Distributions

The joint prior distribution of m and β can be written as $\mathbb{P}(m, \beta) = \mathbb{P}(m) \times \mathbb{P}(\beta|m)$.

As did Dellaportas and Forster,[7] we chose independent normal priors for each parameter included in the model, each with zero mean and precision $1/\tau^2 = 0.001$. We found that results were not very sensitive to variance in the prior precision. As an uninformative prior on the model, they chose to make all models equally likely. Although this prior at first appears uninformative, it is in fact informative on the number of terms in the model, because (for general log-linear models) it has its maximum at $\frac{|I|}{2}$ parameters, which is very large even for problems of moderate size. We expect that in reality, the models with main effects and perhaps a few pairwise interactions would be the most likely models and that higher-order interactions would become more and more unlikely the higher the order becomes. We therefore wish to choose a prior that reflects this information. A possible prior can be defined by assigning all sizes of interaction (from 1:P) with a probability so that $p(s)$ defines the probability of a parameter of size equal to s, given that all subparameters are in the model. If we let s_a define the size and order of interaction a (i.e., the number of variables included in a), then the prior for a particular model can be formed as the product of $p(s_a)$ across all parameters within the model, such that model m has a prior probability of ×

$$\mathbb{P}(m) = const \times \prod_{a \in m} p(s_a),$$

in which we choose $p(s_a) \propto e^{-s_a}$ and $const$ is the normalizing constant that ensures that the sum across all models is equal to 1. Note that this need not be calculated, because it falls out of the Metropolis-Hastings ratio.

Missing-Data Update

Although we assume that there are no missing data and no misclassification of the disease phenotypes, we allow for missing genotype data. Therefore, we need to treat this missing data as an unobserved random variable that can be updated as part of the RJMH algorithm. For simplicity, we fix the algorithm to update the missing data every 50 iterations. At each of these updates, the proposal separately updates each individual, such that new values for all missing loci in that individual are independently sampled from 0, 1, and 2 with equal probability (1/3). The Metropolis-Hastings ratio for accepting or rejecting the proposed data set for this individual is then based just upon the likelihood ratio.

Dealing with Structural Zeros

In our application, in which we have multiple classification classes, there is a small set of cells that are not possible. In usual contingency-table terms, these are referred to as "structural zeros" (distinguished from sampling zeros, in which a cell is possible but observed to have a zero count). In our specific case, these structural zeros occur because an individual with CD at a particular *location* must also have at least one *behavior* of CD and vice-versa. So it is impossible for an individual with no location of disease to have a disease behavior or an individual with no disease behavior to have a location of disease. In our case, this means that there will be $(3^5)\times((2^4 - 1) + (2^4 - 1)) = 7290$ contingency-table cells that are structural zeros. This has two implications for our RJMH method. The first is simply the need for a small adjustment to the likelihood, so as to adjust the set of all possible parameters, I, so that it no longer includes these structural zeros; therefore, within the likelihood we only sum the expected mean cell frequencies over those cells that are not structural zeros. The second adjustment is slightly more complicated. Just as sampling zeros in the frequentist framework can cause problems with fitting some parameters, structural zeros can cause problems with fitting some parameters in the Bayesian framework. The problem is simply that the maximum number of parameters that we can fit in our model is equal to the number of possible cells in the table and when some cells are not possible we find that we can no longer fit some of the parameters. Therefore, we need to restrict the set of all possible parameters to a set that is fittable given the structural zeros. If there are Q cells that are structural zeros, then we need to select Q parameters that we can drop, so that the rest of the parameters are fittable. This group of "illegal" parameters contains all parameters that contain the interaction between the highest level of all of the "behavior" variables and the highest level of one or more "location" variables, as well as all those parameters that contain the interaction between the highest level of all of the "location" variables and the highest level of one or more "behavior" variables. Note that this second point is really a problem only for very small numbers of variables and when there is no weighting of the prior against highly parameterized models. In the application that we consider, this point becomes academic.

Acknowledgments

The work was funded by the Wellcome Trust (076024, 081808), CORE(UK), and the Guy's and St Thomas's Charity. We acknowledge use of genotype data from the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council (grant G0000934) and The Wellcome Trust (grant 068545/Z/02) and of additional data supplied by John Todd (Juvenile Diabetes Research Foundation–Wellcome Trust Diabetes and Inflammation Laboratory, University of Oxford). We would also like to thank Julia Newton and Peter Donaldson (School of Clinical Medical Sciences, University of Newcastle) for the collection of some control individuals. Those who carried out the data collection and original analysis bear no responsibility for further analysis and interpretation of these data.

Received: October 2, 2008
Revised: December 2, 2008
Accepted: December 29, 2008
Published online: January 29, 2009

Web Resources

The URLs for data presented herein are as follows:

British National Child Development Study 1958 birth cohort, http://www.cls.ioe.ac.uk
European Collection of Cell Cultures (ECACC), http://www.ecacc.org.uk
J.M.C.'s webpage, http://homepages.lshtm.ac.uk/encdjcha/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/

References

1. Agresti, A. (1990). Categorical Data Analysis (New York: Wiley).
2. Bishop, Y., Finberg, S., and Holland, P. (1975). Discrete Multivariate Analysis: Theory and Practice (Cambridge, MA: MIT Press).
3. Darroch, J.N., Lauritzen, S.L., and Speed, T.P. (1980). Markov fields and log-linear interaction models for contingency tables. Ann. Stat. *8*, 522–539.
4. Green, P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika *82*, 711–732.
5. Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. Stat. Sci. *14*, 382–417.
6. Hastings, K.W. (1970). Monte carlo sampling methods using markov chains and their applications. Biometrika *57*, 91–109.
7. Dellaportas, P., and Forster, J. (1999). Markov chain monte carlo model determination for hierarchical graphical log-linear models. Biometrika *86*, 615–633.
8. Onnie, C.M., Fisher, S.A., Prescott, N.J., Mirza, M.M., Green, P., Sanderson, J., Forbes, A., Lewis, C.M., and Mathew, C.G. (2008). Diverse effects of the card15 and ibd5 loci on clinical phenotype in 630 patients with Crohn disease. Eur. J. Gastroenterol. Hepatol. *20*, 37–45.

9. Prescott, N.J., Fisher, S.A., Franke, A., Hampe, J., Onnie, C.M., Soars, D., Bagnall, R., Mirza, M.M., Sanderson, J., Forbes, A., et al. (2007). A nonsynonymous snp in atg16l1 predisposes to ileal Crohn disease and is independent of card15 and ibd5. Gastroenterology *132*, 1665–1671.
10. Onnie, C., Fisher, S.A., King, K., Mirza, M., Roberts, R., Forbes, A., Sanderson, J., Lewis, C.M., and Mathew, C.G. (2006). Sequence variation, linkage disequilibrium and association with Crohn disease on chromosome 5q31. Genes Immun. *7*, 359–365.
11. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn disease. Nat. Genet. *40*, 955–962.
12. Mathew, C.G. (2008). New links to the pathogenesis of crohn disease provided by genome-wide association scans. Nat. Rev. Genet. *9*, 9–14.
13. Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Czard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M., et al. (2001). Association of nod2 leucine-rich repeat variants with susceptibility to Crohn disease. Nature *411*, 599–603.
14. Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H., et al. (2001). A frameshift mutation in nod2 associated with susceptibility to Crohn disease. Nature *411*, 603–606.
15. Hampe, J., Cuthbert, A., Croucher, P.J., Mirza, M.M., Mascheretti, S., Fisher, S., Frenzel, H., King, K., Hasselmeyer, A., MacPherson, A.J., et al. (2001). Association between insertion mutation in nod2 gene and Crohn disease in german and british populations. Lancet *357*, 1925–1928.
16. Rioux, J.D., Daly, M.J., Silverberg, M.S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., et al. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to crohn disease. Nat. Genet. *29*, 223–228.
17. Mirza, M.M., Fisher, S.A., King, K., Cuthbert, A.P., Hampe, J., Sanderson, J., Mansfield, J., Donaldson, P., Macpherson, A.J.S., Forbes, A., et al. (2003). Genetic evidence for interaction of the 5q31 cytokine locus and the card15 gene in crohn disease. Am. J. Hum. Genet. *72*, 1018–1022.
18. Cuthbert, A.P., Fisher, S.A., Mirza, M.M., King, K., Hampe, J., Croucher, P.J.P., Mascheretti, S., Sanderson, J., Forbes, A., Mansfield, J., et al. (2002). The contribution of nod2 gene mutations to the risk and site of disease in inflammatory bowel disease. Gastroenterology *122*, 867–874.
19. Ahmad, T., Armuzzi, A., Bunce, M., Mulcahy-Hawes, K., Marshall, S.E., Orchard, T.R., Crawshaw, J., Large, O., de Silva, A., Cook, J.T., et al. (2002). The molecular classification of the clinical manifestations of Crohn disease. Gastroenterology *122*, 854–866.
20. Armuzzi, A., Ahmad, T., Ling, K.-L., de Silva, A., Cullen, S., van Heel, D., Orchard, T.R., Welsh, K.I., Marshall, S.E., and Jewell, D.P. (2003). Genotype-phenotype analysis of the Crohn disease susceptibility haplotype on chromosome 5q31. Gut *52*, 1133–1139.