GENOME-WIDE ASSOCIATION STUDIES

Detecting gene–gene interactions that underlie human diseases

Heather J. Cordell

Abstract | Following the identification of several disease-associated polymorphisms by genome-wide association (GWA) analysis, interest is now focusing on the detection of effects that, owing to their interaction with other genetic or environmental factors, might not be identified by using standard single-locus tests. In addition to increasing the power to detect associations, it is hoped that detecting interactions between loci will allow us to elucidate the biological and biochemical pathways that underpin disease. Here I provide a critical survey of the methods and related software packages currently used to detect the interactions between genetic loci that contribute to human genetic disease. I also discuss the difficulties in determining the biological relevance of statistical interactions.

Data mining
The process of extracting hidden patterns and potentially useful information from large amounts of data.

Machine learning
The ability of a program to learn from experience, that is, to modify its execution on the basis of newly acquired information. A major focus of machine-learning research is to automatically produce models (rules and patterns) from data.

Bayesian model selection
A statistical approach for selecting models by incorporating both prior distributions for parameters of the models and the observed experimental data.

Institute of Human Genetics, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK.
e-mail: heather.cordell@ncl.ac.uk
doi:10.1038/nrg2579
Published online 12 May 2009

The search for genetic factors that influence common complex traits and the characterization of the effects of those factors is both a goal and a challenge for modern geneticists. In recent years, the field has been revolutionized by the success of genome-wide association (GWA) studies[1–5]. Most of these studies have used a single-locus analysis strategy, in which each variant is tested individually for association with a specific phenotype. However, a reason that is often cited for the lack of success in genetic studies of complex disease[6,7] is the existence of interactions between loci. If a genetic factor functions primarily through a complex mechanism that involves multiple other genes and, possibly, environmental factors, the effect might be missed if the gene is examined in isolation without allowing for its potential interactions with these other unknown factors. For this reason, several methods and software packages[8–15] have been developed that consider the statistical interactions between loci when analysing the data from genetic association studies. Although in some cases the motivation for such analyses is to increase the power to detect effects[16], in other cases the motivation has been to detect statistical interactions between loci that are informative about the biological and biochemical pathways that underpin the disease[7]. We return to this complex issue of biological interpretation of statistical interaction later in the article.

The purpose of this Review is to provide a survey of the methods and related software packages that are currently being used to detect the interactions between the genetic loci that contribute to human genetic disease. Although the focus is on human genetics, many of the concepts and approaches are strongly related to methods used in animal and plant genetics. I begin by describing what is meant by statistical interaction and by setting up the definitions and notation for the following sections. I then explain how one might test for interaction between two or more known genetic factors and how one might address the slightly different question of testing for association with a single factor while allowing for interaction with other factors. In practice, one rarely wishes to test for interactions that occur only between known factors, unless perhaps to replicate a previous finding or to test a specific biological hypothesis. It is more common to search for interactions or for loci that might interact, given genotype data at potentially many sites (for example, from a GWA analysis or from a more focused candidate gene study). I continue the article by outlining different methods and software packages that search for such interactions, ranging from simple exhaustive searches to data-mining and machine-learning approaches to Bayesian model selection approaches. Throughout these sections I use the analysis of a publicly available genome-wide data set on Crohn's disease from the Wellcome Trust Case Control Consortium (WTCCC) as a recurring example[1]. I conclude the article with a section discussing the biological interpretation of results found from such statistical interaction analyses.

There is a long history of the investigation of interactions in genetics, ranging from classical quantitative genetic studies of inbred plant and animal populations[17–19] to evolutionary genetic studies[20] and, finally, to linkage and association studies in outbred human populations. In this article, I focus primarily on human genetic association studies; readers are referred to REFS 20–25 for a

Box 1 | **Statistical models of interaction**

Linear, multiple and logistic regression

Statistical interaction can best be described in relation to a linear model that describes the relationship between an outcome variable and some predictor variable or variables. In linear regression, we model a quantitative outcome y as a function of a predictor variable x using the regression equation $y = mx + c$. Here the regression coefficient m corresponds to the slope of the best-fit line and the regression coefficient c corresponds to the intercept. We use the values of pairs of data points (x, y) (for example, if x and y are, respectively, measurements of height and weight in different individuals) to estimate m and c, such that the line $y = mx + c$ fits the observed data as closely as possible.

In multiple regression, we extend this idea to include several different predictor variables using an equation such as $y = m_1x_1 + m_2x_2 + m_3x_3 + c$. Here we are implicitly assuming that there is a linear relationship between each of the predictor variables x_1, x_2 and x_3 and the outcome variable y, so that for each unit increase in x_1, y is expected to increase by m_1 (and similarly for x_2 and x_3).

In logistic regression, rather than modelling a quantitative outcome y, we model the log odds $\ln(p/(1-p))$ (in which p is the probability of having a disease). For example, we might propose the model $\ln(p/(1-p)) = \alpha + \beta x_B + \gamma x_C + i x_B x_C$, in which x_B and x_C are measured binary indicator variables that represent the presence or absence of genetic exposures at loci B and C respectively, β and γ are regression coefficients that represent the main effects of exposures at B and C, and the coefficient i represents an interaction term[16] (a term that is required in addition to the linear terms for B and C).

Testing for interaction

Tests of interaction correspond to testing whether the regression coefficients that represent interaction terms in the above mathematical formula equal zero or not. In the logistic regression example above, this would correspond to a one degree of freedom test of $i = 0$. In the saturated genotype model described in Supplementary information S1 (box), it would correspond to a four degrees of freedom test of $i_{11} = i_{12} = i_{21} = i_{22} = 0$. Tests of association (for example, at a given locus C) while allowing for interaction (for example, with another locus B) correspond to comparing a linear model in which the main effects of B, C and their interactions are included with a model in which all the terms (main or interaction) that involve locus C are removed. For example, if modelling the log odds as $\ln(p/(1-p)) = \alpha + \beta x_B + \gamma x_C + i x_B x_C$, then the test of association at C allowing for interaction with B corresponds to a two degrees of freedom test of $\gamma = i = 0$. This is in contrast to the one degree of freedom pure interaction test of $i = 0$. One could also construct a pairwise test of the joint effects at both loci, including interactions, by comparing a model in which the main effects of loci B, C and their interactions are included with a model in which only the baseline intercept α is included. This gives a three degrees of freedom test of association allowing for interaction if a binary or allelic code is used, or an eight degrees of freedom test[52] if a saturated genotype model (Supplementary information S1 (box)) is used. Tests with fewer degrees of freedom could be used by prior grouping of the two-locus genotypes according to certain prespecified classification schemes[15,29].

discussion of interactions in the context of evolutionary genetics or in human genetic linkage analysis.

Definition of statistical interaction

Interaction as departure from a linear model. The most common statistical definition of interaction relies on the concept of a linear model that describes the relationship between an outcome variable and a predictor variable or variables. We propose a particular model for how we believe the predictors might relate to the outcome and we use data (measurements of the relevant variables from a number of individuals) to determine how well the model fits our observed data and to compare the fit of different models. Arguably the most well-known form of this type of analysis is simple linear or least squares regression[26], in which we relate an observed quantitative outcome y (for example, weight) to a predictor variable x (for example, height) using a 'best fit' line or regression equation $y = mx + c$. More generally, we might use multiple regression[26] to include several different predictor variables (for example, x_1, x_2 and x_3, to represent height, age and gender).

From a statistical point of view, interaction represents departure from a linear model that describes how two or more predictors predict a phenotypic outcome (BOX 1). For a disease outcome and case–control data, rather than modelling a quantitative trait y, the usual approach is to model the expected log odds of disease as a linear function of the relevant predictor variables[26,27]. Using genotype data, we can evaluate the likelihood of the data under this model and use maximum likelihood or other methods to estimate the regression coefficients and test hypotheses, such as the hypothesis that the interaction term (i in the mathematical formula in BOX 1) equals zero.

Supplementary information S1 (box) describes some specific models that follow this general formula, including the saturated genotype model. Although this model provides the best possible fit to the data, it includes many parameters. We can make parameter restrictions to generate fewer degrees of freedom and thus increase power. Although written in terms of nine or fewer regression parameters, the models in Supplementary information S1 (box) represent an infinite number of different models, depending on the values taken by the regression parameters. There has been some interest in categorizing these models[28–30] to aid mathematical or biological interpretation. As discussed below, biological interpretation is usually easiest when the penetrance values all equal either zero or one, leading to a clear relationship between the genotype and phenotype; however, this situation is unlikely for complex genetic diseases.

Marginal effects. An important issue in genetic studies is whether there are factors that display interaction effects without displaying marginal effects[6,31]. Factors that display interaction effects without displaying marginal effects will be missed in a single-locus analysis, as they do not lead to any marginal correlation between the genotype and phenotype when each locus is considered individually. It is not clear in practice how often this might occur, as many models that include an interaction term even in the absence of main effects (α and β in the mathematical formula in BOX 1) lead to substantial marginal effects, that is, they show correlations between the genotype and phenotype that are detectable in a single-locus analysis. Thus, although one may derive mathematical models (sets of specific values for the regression coefficients) that lead to single-locus models without marginal effects[6], it remains to be seen whether such models represent common underlying scenarios — and thus a potentially serious problem — in complex genetic diseases.

For simplicity, I have concentrated here on defining interaction in relation to two genetic factors (two-locus interactions). In practice, however, for complex diseases we might also expect three-locus, four-locus and even higher-level interactions. Mathematically, such higher-level interactions are simple extensions to the two-locus models described earlier. The problem with these models

Maximum likelihood
A statistical approach that is used to make inferences about the combination of parameter values that gives the greatest probability of obtaining the observed data.

Saturated
A term for a statistical model that is as full as possible (saturated) with parameters. Such a model is sometimes useful as it serves as a benchmark to quantify how well a simpler model (one with fewer parameters) fits the data.

is that they contain many parameters, and extremely large data sets would be required to accurately estimate these parameters. Interpreting the resulting parameter estimates is also complicated, except perhaps in some simple cases; for example, when risk alleles at all loci are required to alter disease risk (that is, when only the full multi-locus interaction term differs from zero).

Testing for interaction between known factors

Regression models. For two or more known or hypothetical genetic factors that influence disease risk, arguably the most natural way to test for statistical interaction on the log odds scale is to fit a logistic regression model that includes the main effects and relevant interaction terms and then to test whether the interaction terms equal zero. A similar approach can be used for quantitative phenotypes, in which case linear rather than logistic regression is used. These analyses can be performed in almost any statistical analysis package after construction of the required genotype variables. Alternatively, the '--epistasis' option in the whole-genome analysis package PLINK[12] provides a logistic regression test for interaction that assumes an allelic model for both the main effects and the interactions.

A more powerful approach in case–control studies is to use a case-only analysis[32–34]. Case-only analysis exploits the fact that, under certain conditions, an interaction term in the logistic regression equation corresponds to the dependency or the correlation between the relevant predictor variables within the population of cases. A case-only test of interaction can therefore be performed by testing the null hypothesis that there is no correlation between alleles or genotypes at the two loci in a sample that is restricted to cases alone. This test can easily be performed using a simple χ^2 test of independence between genotypes (a four degrees of freedom test) or alleles (a one degree of freedom test), or using logistic or multinomial regression in any statistical analysis package.

The main problem with the case-only test is its requirement that the genotype variables are not correlated in the general population. It is this assumption, rather than the design *per se*, that provides the increased power compared with case–control analysis. The case-only test is therefore unsuitable for loci that are either closely linked or show correlation for another reason (for example, if certain genotype combinations are related to viability). In contrast to epidemiological studies of environmental factors, in which correlation and confounding between variables is common, in genetic studies the assumption of independence between unlinked genetic factors seems reasonable. One could use a two-stage procedure to test first for correlation between the loci in the general population and then use the outcome to determine whether to perform a case-only or case–control interaction test. However, this procedure has potential bias[35].

A preferable approach is to incorporate the case-only and case–control estimators into a single test. Zhao *et al.*[36] proposed a test based on the difference in inter-locus allelic association between cases and controls, an idea originally suggested by Hoh and Ott[37]. The '--fast-epistasis' option in PLINK[12] performs a similar test. Zhao *et al.*[36] found that their test had greater power than a four degrees of freedom logistic regression test of gene–gene interaction. However, this increase in power might be largely due to the lower number of degrees of freedom in their allelic test compared with a genotypic test. Mukherjee and Chatterjee[35,38] proposed an empirical Bayes procedure that uses a weighted average of the case–control and case-only estimators of the interaction. This approach exploits the gene–gene independence assumption and thus the power of case-only analysis, and additionally incorporates controls, allowing the estimation of main effects. Routines that implement this procedure are available for Microsoft Office Excel and MATLAB.

Other approaches. Although regression-based tests of interaction seem the most natural approach, given the definition of interaction as departure from a linear regression model, alternative approaches have been proposed. Yang *et al.*[39] proposed a method based on partitioning of χ^2 values that, similarly to REF. 36, compares inter-locus association between cases and controls. Their method was more powerful than logistic regression when the loci had no marginal effects. Recently, there has been interest in information theory or entropy-based approaches for modelling genetic interactions[40–43]. It is unclear whether this framework offers any advantage over more standard statistical methods of modelling of the same predictor variables as, in most cases, the conditional probability statements that are implied by the two approaches are equivalent[44].

Family-based studies. Here I focus on testing for interaction in the context of case–control or population-based studies. Several related methods have been proposed to test for interaction in the context of family-based association studies[45–49]. The case–pseudocontrol approach[46] offers a regression-based framework that allows interaction tests that are similar to those described here. Given the larger sample sizes that are required when testing for interaction rather than main effects[50,51], it is unclear whether investigators will have family-based cohorts of a sufficient size to provide high power to detect interactions. However, such cohorts might provide a useful resource for the replication and characterization of interaction effects that have been found using alternative methods.

Tests for association allowing for interaction

Rather than testing for interaction *per se,* many researchers are interested in allowing for interaction with other genetic or environmental factors when testing for association at a given genetic locus. The rationale is that, if the test locus influences the disease or phenotypic outcome by interacting with another factor, then allowing for this interaction should increase the power to detect the effect at the test locus. From a mathematical point of view, a test for association at a given locus C while allowing for interaction with another locus B (a joint test[16]) corresponds to comparing the fit to the observed data

Penetrance
The probability of displaying a particular phenotype (for example, succumbing to a disease) given that one has a specific genotype.

Marginal effects
The average effects (for example, penetrances) of a single variable, averaged over the possible values taken by other variables. These could be calculated for one locus of a two-locus system as the average of the two-locus penetrances, averaged over the three possible genotypes at the other locus.

Logistic regression model
A statistical model that is used when the outcome is binary. It relates the log odds of the probability of an event to a linear combination of the predictor variables.

Multinomial regression
A statistical approach, similar to logistic regression, which is used when the outcome takes one of several possible categorical values.

Confounding
A phenomenon whereby the measure of association between two variables is distorted because other variables, associated with both variables of interest, are not controlled for in the calculation.

Empirical Bayes procedure
A hierarchical model in which the hyperparameter is not a random variable but is estimated by another (often classical) method.

Information theory
A branch of applied mathematics involving the quantification of information.

Entropy
A key measure used in information theory that quantifies the uncertainty associated with a random variable. For example, a variable indicating the outcome from a toss of a coin will have less entropy than a variable indicating the outcome from a roll of a die (two versus six equally likely outcomes).

of a linear model in which the main effects of B, C and their interactions are included with a model in which all the terms (main or interaction) involving locus C are removed (BOX 1).

Theoretically, if no interaction effects exist, these joint tests will be less powerful than marginal single-locus association tests. However, if interaction effects exist, then the power of joint tests can be higher than that of single-locus approaches[52]. Kraft *et al.*[16] showed that the joint test of a genetic effect while allowing for interaction with a known environmental factor had a near optimal performance over a wide range of plausible underlying models. This test uses case–control data to test the combination of a main effect at locus C and an interaction effect. As case-only analysis provides a more powerful test for the interaction effect[32–34], Chapman and Clayton[53] proposed using a version of the joint test that combines a case–control main effect component with a case-only interaction component.

The joint test of association while allowing for interaction assumes that there is some known or hypothetical measured factor that might interact with the test locus. In the absence of a specific factor of this type, a natural approach is to average over all other potentially interacting genetic factors when performing a test at a locus. A Bayesian method for this approach in the context of GWA studies is in development[14] and a beta version of the associated Bayesian Interaction Analysis software is available in limited release from its authors on request. Rather than averaging over all possible interacting loci, Chapman and Clayton[53] proposed using the maximum value of the joint test evaluated over a predefined set of potentially modifying loci and assessing significance using a permutation argument.

I have concentrated on the issue of testing either for interaction or for association while allowing for interaction at one or two specific genetic variants of interest. Rather than testing a single variant, it is now common to have genotype data for many variants that might or might not have any prior evidence for involvement with disease. Given such data, various model selection approaches have been proposed that allow one to step through a sequence of regression models searching for significant effects, including both main effects and interactions[8–10,13,37,54–56]. These approaches are described in more detail in subsequent sections. First, I describe an approach that is feasible provided the number of main and interaction effects to be examined is not too large, namely, a simple exhaustive search.

Exhaustive search

Two-locus interactions. Given genotype data at several different loci, arguably the simplest way to search for interactions between these loci is by an exhaustive search. For example, to test all two-locus interactions, one could analyse all possible pairs of loci and perform the desired interaction test for each pair. Similarly, if testing for association while allowing for interaction, one could perform the relevant three or eight degrees of freedom test[52] (BOX 1, Supplementary information S1 (box)). Clearly, an exhaustive search of this type raises a multiple testing issue analogous to the multiple testing issue encountered in single-locus analysis of GWA studies[1]. If all the tests are independent, a Bonferroni correction is appropriate[52]; however, linkage disequilibrium between loci can induce correlation between many of the tests. When testing for association while allowing for interaction, additional correlation occurs owing to the fact that the main effect of a locus will be a component of all tests that involve that locus. Theoretically, one can use permutation[53] to assess significance while allowing for the multiplicity of and correlation between the tests performed, but, for several loci, this approach might be computationally prohibitive.

A pragmatic approach to the multiple testing issue in single-locus analysis of GWA studies is to use a stringent significance threshold (for example, $p = 5 \times 10^{-7}$) coupled with replication in an independent data set to avoid generating large numbers of false positives. Stringent significance thresholds can also be motivated by Bayesian arguments concerning the low prior probability of any given variant being associated with disease[1]. In practice, the Q–Q plot[1] has emerged as the tool of choice for visualizing the results from an entire-genome scan.

An exhaustive search of all two-locus interactions from a genome scan is time consuming but computationally feasible. Marchini *et al.*[52] quote a time of 33 hours on a 10-node cluster to perform all pairwise tests of association allowing for interaction at 300,000 loci in 1,000 cases and 1,000 controls. The PLINK[12] website quotes 24 hours to test (using the '--fast-epistasis' option) all pairwise interactions at 100,000 loci typed in 500 individuals. Given that genome-wide studies now routinely generate between 500,000 and 1,000,000 markers in 5,000 or more individuals, these times will need to be scaled upwards by several weeks or even months, but an exhaustive search of all two-locus interactions still remains feasible. In addition, as each test can be computed independently of all other tests, the entire search can be split up into several separate jobs and analysed by parallel processing facilities, if they are available.

Higher-order interactions. The problem with an exhaustive search is that it does not scale up to analyse higher-order interactions. Because the number of tests and therefore the time taken to perform the analysis increases exponentially with the order of interaction analysed, an exhaustive search of all three-way, four-way or higher-level interactions seems impractical in a genome-wide setting. For this reason, two-stage procedures have been proposed[52,57,58], in which a subset of loci that pass some single-locus significance threshold are chosen, and an exhaustive search of all two-locus interactions (or a higher order if required, perhaps conditional on significant lower-order effects[58]) is carried out on this 'filtered' subset. The obvious drawback with this approach is that loci will only be filtered into the second or subsequent stages of the testing procedure if they show a marginal association with the phenotype. Therefore, this procedure would not be expected to be useful for detecting interactions that genuinely occur in the absence of marginal effects.

Permutation
This method is often used in hypothesis testing. An empirical distribution of a test statistic is obtained by permuting the original sample many times and recalculating the value of the test statistic in each permuted data set. Each permuted sample is considered to be a sample of the population under the null hypothesis.

Multiple testing
An analysis in which multiple independent hypotheses are tested. If a large number of tests are performed, the significance level (*p* value) of any particular test must be interpreted in light of this fact, as the overall combined probability of making a type I error will increase.

Bonferroni correction
The simplest correction of individual *p* values for multiple hypothesis testing can be calculated using $p_{corrected} = 1 - (1 - p_{uncorrected})^n$, in which n is the number of hypotheses tested. This formula assumes that the hypotheses are all independent, and simplifies to $p_{corrected} = np_{uncorrected}$ when $np_{uncorrected} << 1$.

Q–Q plot
A quantile–quantile plot is a diagnostic plot that can be used to compare the distribution of observed test statistics with the distribution expected under the null hypothesis. Those tests that lie significantly above the line of equality between observed and expected quantiles are considered significant in the context of the number of tests performed.

Use of a single-locus significance threshold is not the only way to reduce the number of markers for testing. Several of the machine-learning approaches described in the next section (in particular ReliefF and random forests) could be used, as they do not require a locus to have a significant marginal effect. Biological plausibility offers an alternative strategy. Bochanovits *et al.*[59] used evidence of co-adaptation between loci in the mammalian genome to select genes for interaction testing in a human study. Emily *et al.*[60] used experimental knowledge of biological networks to reduce the number of interaction tests from 1.25×10^{11} to 7.1×10^{4} when analysing genotype data from the WTCCC[1]. In their analysis of seven disease cohorts, they found four significant interaction effects, including one of $p = 1 \times 10^{-9}$ between rs6496669 on chromosome 15 and rs434157 on chromosome 5 in Crohn's disease. An example of applying semi-exhaustive testing to this same data set using the '--fast-epistasis' and '--case-only' options in PLINK[12] is shown in FIG. 1.

Data-mining methods and related approaches

Traditional regression-based methods are often criticized[8,31,61] for their inability to deal with nonlinear models and with high-dimensional data that contain many potentially interacting predictor variables, leading to sparse contingency tables that have many empty cells. For this reason, machine-learning or data-mining methods developed in the field of computer science are sometimes preferred. The selection of predictor variables and the interactions between them that predict an outcome variable is a well-known problem in the fields of machine learning and data mining. Data-mining approaches do not fit a single prespecified model, nor do they attempt an exhaustive search, but rather they attempt to step through the space of possible models, including potentially large numbers of main effects and multiway interactions, in a computationally efficient way. Many data-mining approaches are equivalent to stepping through a particular sequence of regression models and attempting to find the model that best fits the data; the distinction that is often made between data-mining and regression models is therefore, to some extent, false. Nonlinearity is not an issue when fitting a saturated model, although it might be an issue for more restricted models. One common theme in data mining is the use of cross-validation[62] to avoid overfitting problems.

Data-mining methods typically have problems dealing with incomplete or unbalanced data sets; for example, when the number of cases and controls are unequal[63]. They also do not always deal well with correlated predictors that show colinearity. This has been addressed in the mainstream statistics literature by the introduction of penalized regression approaches[64,65] that allow large numbers of predictor variables to be included in a regression model but with many estimated regression coefficients reduced towards zero. In genetics, the use of such techniques is just starting to emerge, including penalized logistic regression[66,67] and least-angle regression[68] for identifying gene–gene interactions[69,70] in binary traits.

A good overview of several machine-learning approaches for detecting gene–gene interactions is given by McKinney *et al.*[31]. For the remainder of this section, I focus on several methods that have become popular or seem to show promise for detection of gene–gene interactions or, more precisely, for detection of genes that might interact.

Recursive partitioning approaches. Recursive partitioning approaches (BOX 2) have been used as an alternative to traditional regression methods for detecting the genetic loci and their interactions that influence a phenotypic outcome[71–73]. These approaches produce a graphical structure that resembles an upside-down tree that maps the possible values of certain predictor variables (for example, SNP genotypes) to a final expected outcome (for example, disease status). Each vertex or node of the tree represents a predictor variable and there are arcs or edges from each node leading down to 'child' nodes, in which each edge corresponds to a different possible value that could be taken by the variable in the 'parent' node. A path through the tree represents a particular combination of values taken by the predictor variables that are present within that path. Recursive partitioning approaches do not include interaction variables *per se* in the model. Rather, the trees constructed allow for interaction in the sense that each path through a tree corresponds to a particular combination of values taken by certain predictor variables, thus including the potential interactions between them. The aim of tree-based approaches therefore corresponds most closely to testing for association while allowing for interaction rather than testing for interaction *per se*. One limitation of recursive partitioning is that, because it conditions on the main effects of variables at the first stage and on the main effects conditional on previously selected variables at subsequent stages, pure interactions in the absence of main effects can be missed[74].

Rather than using a single tree, substantial improvements in classification accuracy can result from growing an ensemble of trees. A popular ensemble tree approach is the random forests approach[75] (BOX 2), which has been used in several genetic studies[76,77]. Apart from the classification of future observations (which is not our focus of interest), the main result of a random forests analysis is a list of variable importance measures. These measure the effect of each predictor variable both individually and through multiway interactions with other predictor variables, and therefore have an advantage over a list of significance values from single-locus association testing.

Random forests provide a fast algorithm that can be applied in parallel for measuring variable importance partly because, at each split, only a small random subset of predictors is used. To allow each predictor the opportunity to enter the model and to make an accurate prediction, one must carefully choose important parameters, such as the number of trees in the forest, the number of randomly chosen SNPs analysed at each node and the number of permutations used to assess variable importance. Ideally, one would repeat the analysis several times to assess the sensitivity to the choice of these

High-dimensional data
Data that contain information on a large number of variables, albeit possibly measured in a small number of subjects or replicates.

Cross-validation
This approach involves partitioning a data set into smaller subsamples, performing an analysis in one subsample and using the other subsample to measure or validate how well the analysis has performed. To reduce variability, multiple rounds of cross-validation are often performed using different partitions of the data and the validation results are averaged over the rounds.

Overfitting
The phenomenon in which a complex model might provide a good fit to the current data set but is overfitted to the random quirks present in that particular data set and therefore cannot be generalized to future data sets in the way that a simpler model might be.

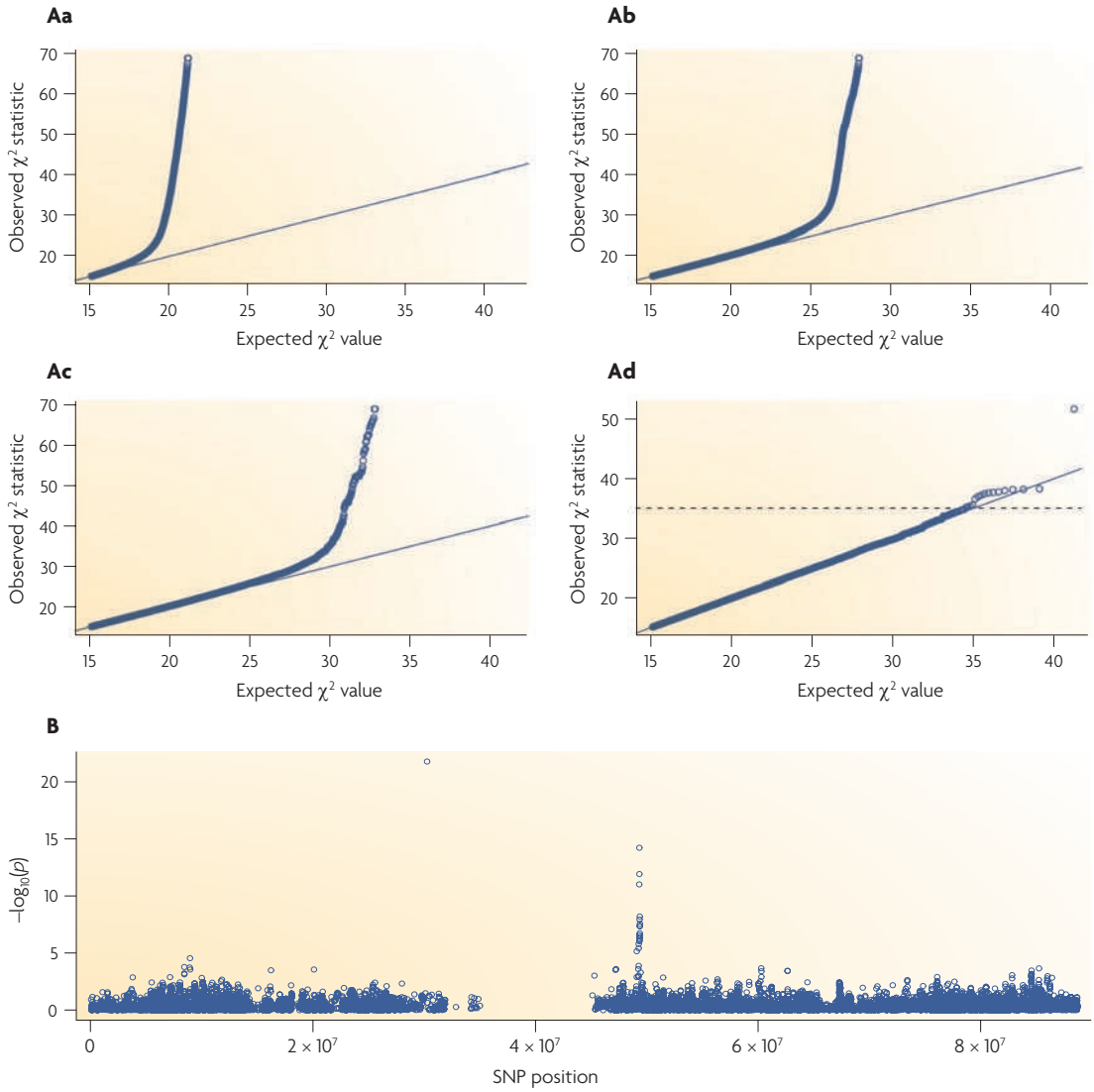
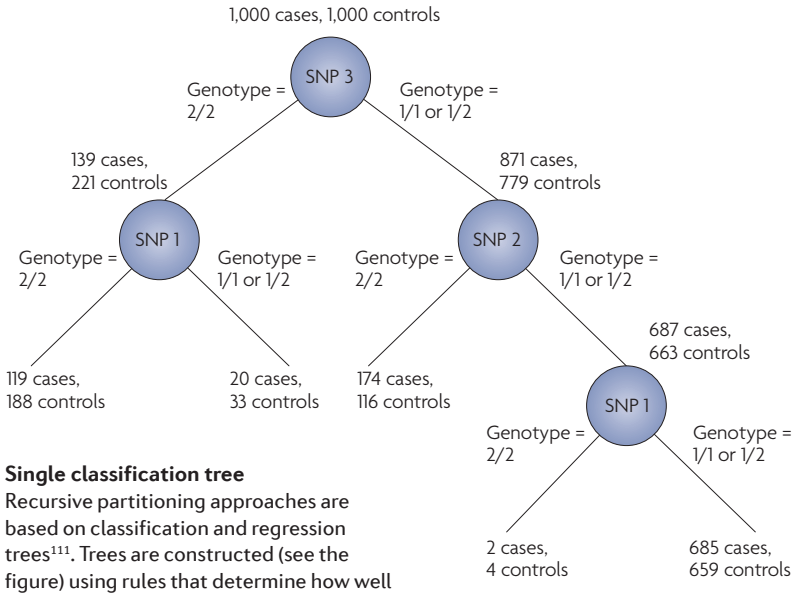


Figure 1 | **Semi-exhaustive search of pairwise interactions between 89,294 SNPs.** I used the '--fast-epistasis' and '--case-only' options in PLINK to analyse the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control samples. I used the same quality control procedures as the WTCCC to remove poor quality SNPs and samples before analysis. I additionally discarded 561 SNPs that had been analysed by WTCCC but were subsequently discarded on the basis of visual inspection of the SNP intensity cluster plots (J. Barrett, personal communication). To reduce the number of interaction tests to be performed, I selected a set of 89,294 SNPs that passed a single-locus *p* value threshold of 0.2. Analysis of the 89,294 SNPs on a single node of a computer cluster took 14 days. Unfortunately, neither SNP in the interaction detected by Emily *et al.*[60] were included in my analysis, as neither had a single-locus $p \leq 0.2$. **A** | Results from '--case-only' analysis, in which SNP pairs were discarded if they were <1 Mb apart (panel **a**), <5 Mb apart (panel **b**), and <50 Mb apart (panel **c**). The default in PLINK is to exclude tests of pairs of SNPs that are less than 1 Mb apart. Even when extreme separations of 5 Mb or 50 Mb are enforced (panels **b** and **c**), we find a large number of apparently significant results. A closer inspection showed that in many cases, these significant results are due to correlation within the sample of cases between alleles at loci on different chromosomes. Given the general departure from the expected distribution, it seems likely that these significant case-only results are artefacts rather than genuine interaction effects. Panel **d** shows a Q–Q plot of all results from the '--fast-epistasis' option with $p < -0.0001$. These results lie much closer to the expected line; only one result seems to show strong departure from the expected significance. The top-ranking results (those with $\chi^2 > 35$, as indicated by the dashed line on panel **d** are shown in Supplementary information S3 (table). Interestingly, most of the SNPs involved in the putative interactions show little single-locus significance, apart from rs4471699 on chromosome 16. This SNP was not reported as significantly associated by WTCCC[1]. **B** | Single-locus association results across chromosome 16. rs4471699 at position 30,227,808 shows the highest significance but is far removed from most of the significant results, which are situated close to nucleotide-binding oligomerization domain containing 2 (*NOD2*) (approximate position 49,297,083). Further investigation showed that this SNP had been excluded from the WTCCC analysis owing to poor genotype clustering (J. Barrett, personal communication), even though it passed the stated WTCCC exclusion criteria and was not present in the original list of additional exclusions I was given. It therefore seems likely that both the single-locus and interaction results at rs4471699 are false positives.

Box 2 | **Recursive partitioning approach**



Single classification tree

Recursive partitioning approaches are based on classification and regression trees[111]. Trees are constructed (see the figure) using rules that determine how well a split at a node (based on the values of a predictor variable such as a SNP) can differentiate observations with respect to the outcome variable (such as case–control status). A popular splitting rule is to use the variable that maximizes the reduction in a quantity known as the Gini impurity[111,112] at each node. In the figure, SNP 3 maximizes the reduction in the Gini impurity at the first node and is therefore chosen for splitting (according to the genotype at SNP 3) the original data set of 1,000 cases and 1,000 controls into two smaller data sets. Once a node is split, the same logic is applied to each child node (hence the recursive nature of the procedure). The splitting procedure stops when no further gain can be made (for example, when all terminal nodes contain only cases or only controls, or when all possible SNPs have been included in a branch) or when some preset stopping rules are met. At this stage, it is usual to prune the tree back (that is, to remove some of the later splits or branches) according to certain rules[111] to avoid overfitting and to produce a final more parsimonious model.

Ensemble approaches: random forests

Rather than using a single classification tree, substantial improvements in classification accuracy can result from growing an ensemble of trees and letting them 'vote' for the most popular outcome class, given a set of input variable values. Such ensemble approaches can be used to provide measures of variable importance, a feature that is of great interest in genetic studies and that is often lacking in machine-learning approaches. The most widely used ensemble tree approach is probably the random forests method[75]. A random forest is constructed by drawing with replacement several bootstrap samples of the same size (for example, the same number of cases and controls) from the original sample. An unpruned classification tree is grown for each bootstrap sample, but with the restriction that at each node, rather than considering all possible predictor variables, only a random subset of the possible predictor variables is considered. This procedure results in a 'forest' of trees, each of which will have been trained on a particular bootstrap sample of observations. The observations that were not used for growing a particular tree can be used as 'out-of-bag' instances to estimate the prediction error. The out-of-bag observations can also be used to estimate variable importance in different ways including through use of a permutation procedure[31,77,113].

The true model in which the important predictor variables act or interact to influence phenotype is somewhat obscured because it results from the predictions of many different classification trees, and so one might wish to follow a random forests analysis with another approach. For example, one might choose the top-ranking variables from a random forests analysis as input variables for a simple regression-based search, a standard classification and regression trees analysis or for analysis using an alternative data-mining procedure.

See REFS 31,74,113 for a good summary of the approach, the available R software (the 'randomForest', 'cforest' and 'party' libraries) and a discussion of some of the limitations of the method.

parameters. An example of applying random forests to the WTCCC Crohn's disease and control data using the Random Jungle software package[78] is shown in FIG. 2.

Multifactor Dimensionality Reduction method. A range of data-mining approaches have been used for the detection of interactions or potentially interacting variables in genetic association studies, including logic regression[79,80], genetic programming[81], neural networks[54,55] and pattern mining[82,83]. One particularly popular method is Multifactor Dimensionality Reduction (MDR)[8–10]. MDR has been used to identify potential interacting loci in several phenotypes, including breast cancer[8], type 2 diabetes[84], rheumatoid arthritis[85] and coronary artery disease[86], although to date it is unclear whether any of these identified interactions have been replicated in larger samples.

The MDR algorithm is described in BOX 3 and in detail elsewhere[8–11,49]. Rather than testing for interaction *per se*, MDR seeks to identify combinations of loci that influence a disease outcome, possibly by interactions rather than — or in addition to — by main effects. MDR reduces the number of dimensions by converting a high-dimensional multilocus model to a one-dimensional model, thus avoiding the issues of sparse data cells and models with too many parameters that can cause problems for traditional regression-based methods. MDR classifies genotypical classes as either high risk or low risk according to the ratio of cases and controls in each class. This approach could be considered overly simplistic, and improvements that embed a more traditional regression-based approach into the cell classification step, allowing application of the method to continuous as well as binary traits and adjustment for covariates, have been proposed[87,88].

The main problem with MDR, as with other exhaustive search techniques, is that it does not scale up to allow analysis of large numbers of predictor variables (for example, many loci from a GWA study)[8,9]. If an exhaustive search for the best *n*-locus combination (within each of ten cross-validation replicates) is performed, anything more than a two-locus screen on more than a few hundred variables will be computationally prohibitive. An additional problem with early versions of the widely used Java implementation of the MDR software (but note that other software implementations exist[11,88]) is that it was not designed with genome-wide data sets in mind and thus could fail owing to memory and disc usage issues. However, these problems seem to have been addressed in the most recent version of the software.

For investigation of higher-order interactions, MDR is therefore perhaps best suited for use with small numbers of loci (up to a few hundred), which have perhaps been discovered from a candidate gene study or selected from a larger set of potential predictors using a prior processing or filtering step[40]. This step could be as simple as using a single-locus significance threshold, but that seems counter-intuitive if the goal is to detect interactions in the absence of marginal effects. Perhaps a more appealing approach would be to use a measure of variable importance that allows for possible interactions,

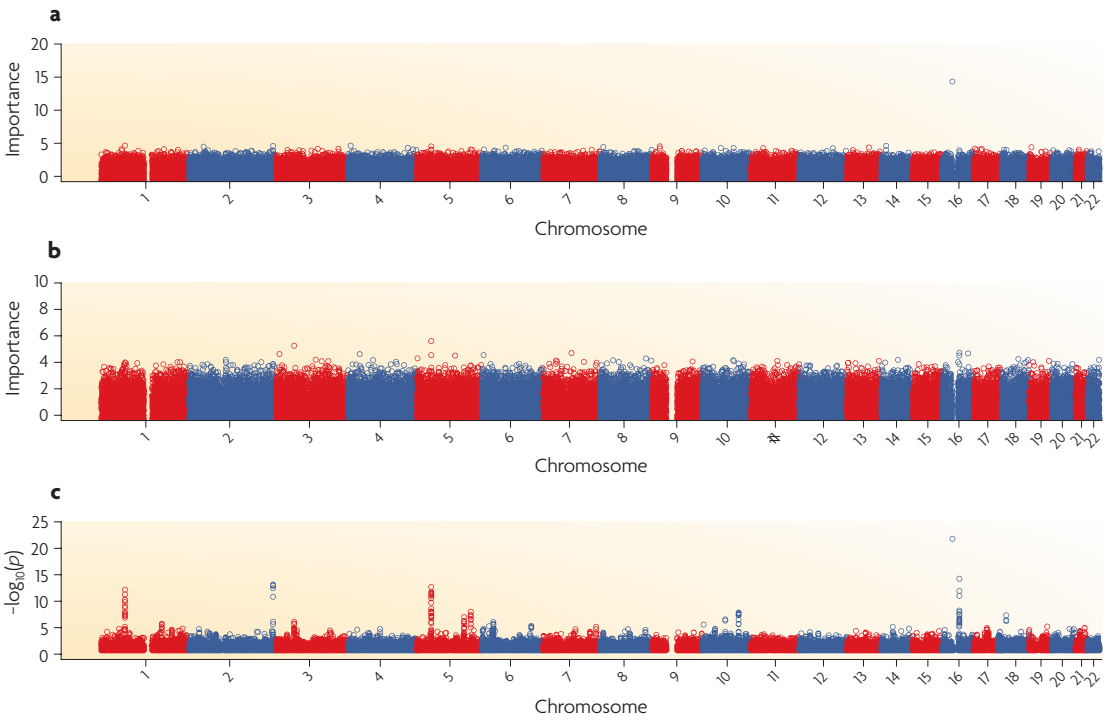


Figure 2 | **Random Jungle analysis of 89,294 SNPs.** I used the software package Random Jungle[78] to perform a random forests analysis of the 89,294 SNPs that passed a single-locus *p* value threshold of 0.2 in the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control data. As Random Jungle, in common with many other machine-learning approaches, prefers not to have missing genotype data, the missing genotypes were imputed as the single most likely values on the basis of the genotype frequencies in the case–control data set. Analysis of the 89,294 SNP set took approximately 5 hours, using 6,000 trees in the forest and $\sqrt{n} = \sqrt{89{,}294}$ randomly chosen variables at each node. **a** | Importance values from the Random Jungle analysis. These are clearly dominated by the result at rs4471699 on chromosome 16, which is likely to be a false positive. **b** | Results from Random Jungle analysis with SNP rs4471699 removed. Once this SNP is removed, the remaining SNPs are better distinguished, but it is unclear whether this analysis offers any greater insight than the single-locus analysis. **c** | Results from single-locus association analysis of all 6,113 SNPs using the trend test implemented in PLINK. In many cases, the highest ranking SNPs are in similar locations to (**b**), but with clearer significance in (**c**).

such as the variable importance measure from a random forests analysis or from one of the alternative filtering methods described below.

ReliefF, Tuned ReliefF and evaporative cooling. One promising filtering algorithm that has been proposed[40] is ReliefF[89] or its modified version, Tuned ReliefF (TuRF)[90]. This approach uses a measure of proximity between observations (individuals) — which is calculated, for example, on the basis of the genome-wide genetic similarity between individuals — to determine the nearest neighbours of each individual from within their own phenotype class and from within the opposite phenotype class. The difference in the value of each predictor variable between the pairs of neighbouring individuals, weighted negatively or positively according to whether the individuals come from the same or different phenotype classes, can be used to construct an importance measure for that variable[90]. The algorithm is simple and scalable, and should be applicable to large numbers of predictor variables and observations; an in-house C++ implementation was able to analyse 1 million loci in 200 individuals in approximately 4 minutes[90].

ReliefF and TuRF have both been implemented in the Java version of the MDR software. One problem with ReliefF is that it can be affected by large backgrounds of genetic variants that do not contribute to the phenotype[74]. This has motivated the development of an alternative approach, evaporative cooling[74,91], which can be used to combine the strengths of ReliefF with those of random forests methods[74].

An example of analysis using the Java implementation of TuRF and MDR applied to the WTCCC Crohn's disease data is shown in FIG. 3.

Bayesian model selection approaches

Bayesian model selection techniques[92] offer an alternative approach for selecting predictor variables and the interactions between them that are the best predictors of phenotype. The key difference between Bayesian model selection and simple comparisons of nested regression models using frequentist (non-Bayesian) procedures is the specification of prior distributions for the unknown regression parameters as well as for a dimension parameter in a Bayesian approach. This dimension parameter specifies how many non-zero predictors are included

Bootstrap samples
These are data sets obtained by taking a random sample of the original data, usually with replacement. One then applies the same analysis as was applied to the real data. This is repeated many times, allowing one to assess the variability in results incurred owing to random sampling.

Frequentist
A statistical approach for testing hypotheses by assessing the strength of evidence for the hypothesis provided by the data.

Box 3 | **Multifactor Dimensionality Reduction**

The Multifactor Dimensionality Reduction (MDR) method is a constructive induction algorithm[40] that proceeds as follows: the observed data is divided into ten equal parts and a model is fit to each nine-tenths of the data (the training data), and the remaining one-tenth (the test data) is used to assess model fit, thus using ten-fold cross-validation. Within each nine-tenths of the data, a set of *n* genetic factors is selected and their possible multifactor classes or cells are represented in *n* dimensional space. For example, for $n = 2$ diallelic loci, there are nine possible genotype classes or cells (Supplementary information S1 (box)). The ratio of the number of cases to the number of controls is estimated in each cell and the cell is labelled as either high risk if the case–control ratio reaches or exceeds a predetermined threshold (for example, ≥1) and low risk if it does not reach this threshold. This reduces the original *n*-dimensional model to a one-dimensional model (that is, one variable with two classes: high risk and low risk). The procedure is repeated for each possible *n*-factor combination and the combination that maximizes the case–control ratio of the high-risk group (that is, the combination that fits the current nine-tenths of the data best, giving minimum classification error among all *n*-locus models) is selected. The testing accuracy (which is equal to 1 – prediction error) of this best *n*-locus model can be estimated using the remaining test data portion of the data. The whole procedure is repeated for each of the nine-tenth-one-tenth partitions of the data, and the final best *n*-locus model is the model that maximizes the testing accuracy or, equivalently, minimizes the prediction error. The cross-validation consistency is defined as the number of cross-validation replicates (partitions) in which that same *n*-locus model was chosen as the best model (that is, the number of replicates in which it minimized classification error). The average prediction error is defined as the average of the prediction errors over the ten cross-validation test data sets. Note that the prediction error of each individual cross-validation replicate refers to the prediction error of the *n*-locus model chosen as the best model in that replicate, which will not always correspond to the final best *n*-locus model.

In practice, rather than selecting a single value of *n* in each cross-validation replicate, one might consider all possible values of *n* up to a certain maximum; for example, all single-locus genotype combinations ($n = 1$), all two-locus combinations ($n = 2$) or all three-locus combinations ($n = 3$). One thus generates a best model within each cross-validation replicate as well as a final best model (with the associated cross-validation consistency and average prediction error) for each different value of *n*. The cross-validation consistencies and average prediction errors can be used to determine the best value of *n* that gives the highest cross-validation consistency or lowest average prediction error, and thus the resulting overall best model.

Burn-in period
In Markov chain Monte Carlo analysis, a period at the start of the computation in which the values taken by the parameters are ignored when constructing the posterior distribution.

Compositional epistasis
The blocking of one allelic effect by an allele at another locus.

Statistical epistasis
The average effect of substitution of alleles at combinations of loci, with respect to the average genetic background of the population.

Functional epistasis
The molecular interactions that proteins and other genetic elements have with one another.

in the regression equation. A posterior distribution for these parameters, given the observed data, can then be calculated using Markov chain Monte Carlo (MCMC)[93] simulation techniques, in which one traverses the space of the possible models (sets of parameter values), sampling the outputs of the simulation run at intervals. Although MCMC is a flexible approach, it can require some care with respect to the choice of prior distributions, proposal schemes (determining how one moves between models) and the number of iterations required to achieve convergence.

Lunn *et al.*[56] proposed a Bayesian version of stepwise regression implemented in the software WinBUGS. This method focuses on the main effects of loci rather than interactions, but the inclusion of interaction effects is a straightforward extension. The main problem with this method is that it can deal with only a few hundred variables at most[56] and does not scale to the large numbers of predictor variables that might be encountered in a genome-wide study. However, related approaches that can deal with data sets with more dimensions have been proposed[94].

Bayesian Epistasis Association Mapping. A recently proposed MCMC approach that is specifically designed to detect interacting, as well as non-interacting, loci is Bayesian Epistasis Association Mapping[13], which is implemented in the software package BEAM. In BEAM, predictors in the form of genetic marker loci are divided into three groups: group 0 contains markers that are not associated with disease, group 1 contains markers that contribute to disease risk only by main effects and group 2 contains markers that interact to cause disease by a saturated model. Given prior distributions that describe the membership of each marker in each of the three groups and prior distributions for the values of the relevant regression coefficients given group membership, a posterior distribution for all relevant parameters can be generated using MCMC simulation. In addition to making inferences in a fully Bayesian inferential framework, one can use the results from BEAM in a frequentist hypothesis-testing framework by calculating a 'B-statistic'[13] that tests each marker or set of markers for significant association with a disease phenotype.

BEAM can handle large numbers of markers (for example, 100,000 SNPs typed in 500 cases and 500 controls[13]) although, in practice, some modification to the default parameters (namely the burn-in period, number of starting points and number of MCMC iterations) might be required to apply the method in a reasonable period of time. BEAM cannot currently handle the 500,000–1,000,000 markers that are now routinely being genotyped in genome scans of 5,000 or more individuals. In theory, BEAM can account for linkage disequilibrium between adjacent markers[13]. However, it is unclear whether linkage disequilibrium between non-adjacent markers is fully accounted for, suggesting that reducing the number of markers in the marker set might be required, not only for computational reasons, but also to ensure that the markers are in low linkage disequilibrium. An example of applying BEAM to the WTCCC Crohn's data is shown in FIG. 4.

Biological interpretation

The extent to which statistical interaction implies biological or functional interaction has been extensively debated in both the genetics[19,21,95–99] and epidemiological[100–102] literature. One problem has been the inherently different nature of definitions of interaction and the use of a common term, epistasis, to encapsulate these definitions[21,95] (Supplementary information S2 (box)). In a recent review, Phillips[20] defines three different forms of epistasis — compositional epistasis, statistical epistasis and functional epistasis — that capture different concepts that are often grouped together under this single term. A unified framework, the natural and orthogonal interactions (NOIA) model, was proposed by Alvarez-Castro and Carlborg[98] for modelling both statistical and functional epistasis. However, Alvarez-Castro and Carlborg's definition of functional differs from that of Phillips. The NOIA model is actually a mathematical reparameterization of classical quantitative genetics models[19] (Supplementary information S2 (box)). The NOIA model allows the main effects to be defined with

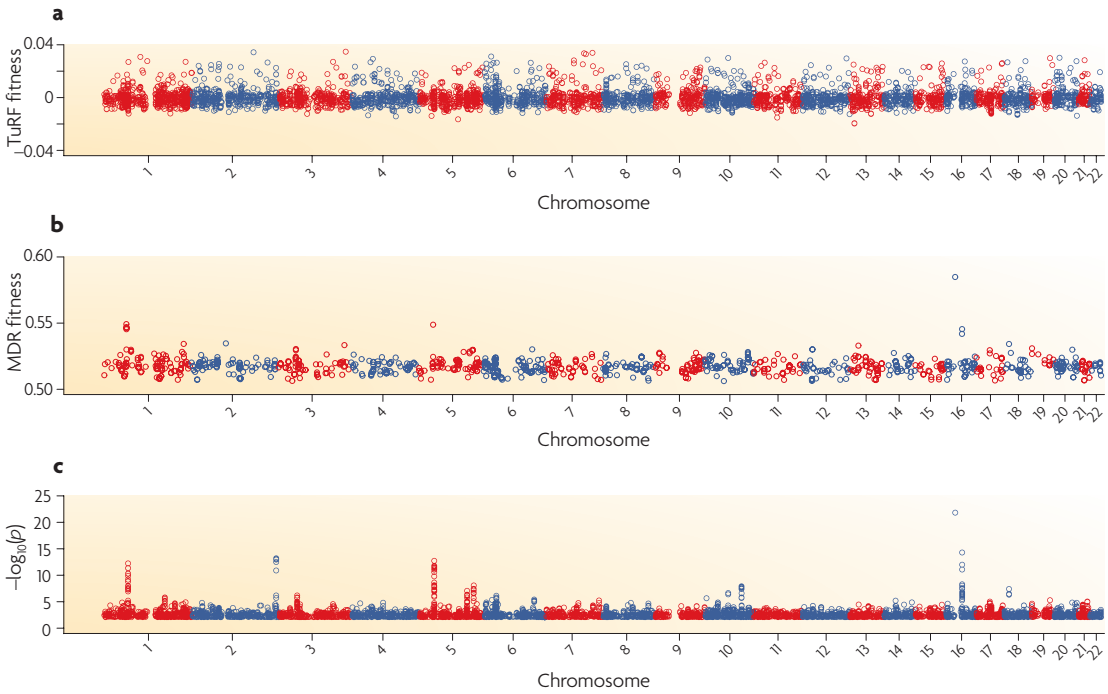


Figure 3 | **Multifactor Dimensionality Reduction (MDR) and Tuned ReliefF (TuRF) analysis of 6,113 SNPs.** I used the Java implementation of MDR to analyse 6,113 SNPs that passed a single-locus *p* value threshold of 0.01 in the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control data, with missing genotypes imputed as the single most likely values on the basis of the genotype frequencies in the case–control data set. Examination of all pairwise combinations in the entire 6,113 SNP set was computationally prohibitive but analysis using a prior filtering step with ReliefF or TuRF, which reduced the data set for MDR analysis to 1,000 SNPs, was achievable. The best single-locus model identified was rs4471699, providing a testing accuracy of 0.5852 and cross-validation consistency of 10 out of 10. The best two-locus model identified was rs4471699 and rs2076756, providing a testing accuracy of 0.5879 and cross-validation consistency of 4 out of 10. MDR, in common with the other methods investigated, has clearly been dominated by the false positive result at rs4471699. Interestingly, however, this SNP is not selected by TuRF when filtering down the set of SNPs for MDR analysis to include only 100 SNPs. Using the 100 SNP set, the best single-locus model identified was rs931058, providing a testing accuracy of 0.5114 and cross-validation consistency of 5 out of 10. The best two-locus model identified was rs931058 and rs10824773, providing a testing accuracy of 0.5205 but cross-validation consistency of only 2 out of 10. Using the 100 SNP set, it was computationally feasible to fit three-locus and four-locus models; however, the resulting best models had cross-validation consistencies as low as for the two-locus model. I also found extreme sensitivity in both TuRF and MDR to the choice of the random number seed (data not shown), suggesting that, overall, these results should be interpreted with caution. A problem with MDR is that it outputs only the best model rather than a measure of significance for all of the models or variables considered. An idea of the importance of the variables can be determined by examining the 'fitness landscape' output from the program, shown here. **a** | Fitness landscape scores from TuRF analysis of all 6,113 SNPs. **b** | Fitness landscape scores from MDR analysis using the top 1,000 out of 6,113 SNPs filtered using TuRF. **c** | Results from single-locus association analysis of all 6,113 SNPs using the trend test implemented in PLINK. It is unclear whether the fitness landscape results from TuRF (**a**) or MDR (**b**) offer any great advantage over standard single-locus analysis (**c**) with respect to determining the importance of variables.

respect to a different reference point and interaction effects to be defined with respect to different definitions of the independence of the main effects, thus allowing mapping of models between different experimental populations. As the whole issue in interaction modelling is how one defines the effect of a variable and, therefore, how one measures departure from the independence of effects (Supplementary information S2 (box)), this reparameterization does not seem to be biologically enlightening.

It may seem reasonable to assume that functional epistasis in the form of biomolecular or protein–protein interaction is a ubiquitous component of the underlying biological pathways that determine disease progression[7,103]. However, this does not mean that epistasis will be detected as a mathematical or statistical interaction[102,104], particularly if the variables that are being examined are, as in many cases, simply surrogates for the true underlying causal variants that are correlated with the causal variants because of linkage disequilibrium. The historical lack of success in genetic studies of complex disease can largely be attributed, not to ignored biological interactions[7,61,67], but to underpowered studies that surveyed only a fraction of genetic variation. The recent success of GWA studies[1–5] has shown that single-locus association analysis in sufficiently large sample collections can reliably detect modest genetic effects that are robustly replicated[105,106].

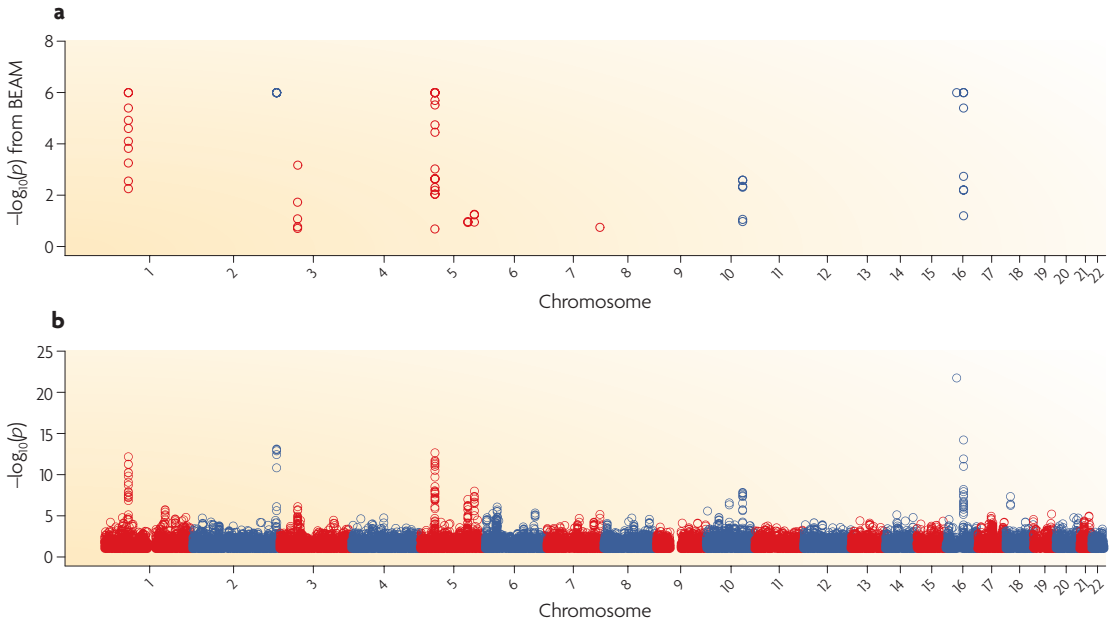


Figure 4 | **Bayesian Epistasis Association Mapping (BEAM) analysis of 47,727 SNPs.** I used BEAM to analyse a set of 47,724 SNPs that passed a single-locus *p* value threshold of 0.1 in the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control samples. Analysis of the 47,724 SNPs took 8 days (with some modification to the default settings, most notably imposing a maximum of 5×10^{-7} Markov chain Monte Carlo (MCMC) iterations[13] rather than using the default value of n^2, in which n is the number of loci). I estimated that analysis of the 89,294 SNP set passing a single-locus *p* value threshold of 0.2 with a similar number of MCMC iterations would have taken more than 5 weeks. **a** | 'B-statistic' *p* values for the 1,321 single-locus associations detected by BEAM. **b** | Results from single-locus association analysis of all 47,727 SNPs using the trend test implemented in PLINK. BEAM detects the same loci as are detected by single-locus analysis. BEAM additionally detects (with a quoted *p* value of 0.000000) four two-locus interactions, each involving an interaction of rs2532292 on chromosome 17 with a nearby SNP (either rs12150547, rs17689882, rs17650381 or rs17574824) within the same cluster. None of these SNPs shows particularly strong single-locus associations and so this putative interaction is intriguing. However, none of these pairs of SNPs showed significant (defined as a $p < 0.0001$) interaction in the PLINK '--fast-epistasis' analysis. Closer inspection of these SNPs in the control sample indicated that they are in strong linkage disequilibrium ($D' > 0.99$) with one another, suggesting that the detected interactions might correspond to marker dependencies owing to linkage disequilibrium, rather than to genuine interaction effects.

Although the extent to which biological interaction can be inferred from statistical interaction might be limited[102], some interesting recent studies[107–109] have focused on whether, given a strong prior biological model or set of models, one can use genetic or genomic data from outbred populations or inbred strains to assess the fit of the model and compare the fits of competing models. This is a more modest goal because it relies on a prior understanding or at least a strong biological hypothesis with respect to the action of the relevant predictors.

Conclusions

As we have seen, there are numerous methods and an even larger number of software implementations that allow investigators to examine or test for interaction between loci, using data that is currently generated from large-scale genotyping projects. Although the precise details of the methods differ, in many cases there are close conceptual links between the different approaches. The best way to understand these links might be provided by understanding the difference between testing for interaction versus testing for association while allowing for interaction.

From a practical point of view, probably the main difference between the methods I have described is the computational time required to implement the analysis. As data sets become larger, the development of efficient computational algorithms that can be implemented in parallel will become more important. On this note, the use of filtering approaches that allow one to preselect a subset of potentially interesting loci to input into a more computer-intensive exhaustive or stochastic search algorithm might hold promise. In my application of various methods to the WTCCC Crohn's disease data, I found that a semi-exhaustive search of two-locus interactions implemented in PLINK[12] and a random forests analysis implemented in Random Jungle[78] were the most computationally feasible of the methods examined. Bayesian Epistasis Association Mapping implemented in BEAM[13] was feasible only for a filtered data set and with some modification to the default recommended input parameter settings; it is unclear what effect, if any, this will have had on the reliability of the results. MDR was feasible for examining two-locus interactions in a filtered data set or for examining higher-level interactions in an even further reduced data set.

To date, few publications have incorporated interaction testing of GWA data. This is perhaps unsurprising as GWA studies have naturally focused on single-locus testing in the first instance. Curtis[110] performed pairwise tests of association at 396,591 markers using 541 subjects (cases and controls) from a genome-wide study of Parkinson's disease. He found no significant epistatic interactions, possibly because of the small sample size or because of the interaction test that was used, which might have been more powerful if it was restricted to cases alone. Gayan *et al.*[15] used the same data set to perform two-locus interaction testing using their interaction detection approach, hypothesis-free clinical cloning. This approach involves testing for association while allowing for interaction under a set of prespecified fully penetrant disease models, and the tests are performed in several different subgroups of the data, which are considered as replication groups. For the Parkinson's disease analysis, each subgroup consisted of approximately 90 cases and 90 controls, which seems a very small sample size for this kind of analysis. Unsurprisingly, little consistency between results was found when the analysis was repeated using different partitions of the data. Emily *et al.*[60] reported four significant cases of epistasis in the WTCCC data using an approach that narrows the search space on the basis of experimental knowledge of biological networks.

Given the large number of GWA studies that have recently been or are currently being performed, it is clear that, for many, genome-wide interaction testing will be the natural next step following single-locus testing. We await with interest the results of these analyses.

1. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
In this study of 17,000 individuals, many new complex trait loci were identified and key methodological and technical issues related to GWA studies were explored.
2. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
3. Frayling, T. M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
4. Plenge, R. M. *et al. TRAF1-C5* as a risk locus for rheumatoid arthritis — a genome-wide study. *N. Engl. J. Med.* **357**, 1199–1209 (2007).
5. Fellay, J. *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**, 944–947 (2007).
6. Culverhouse, R., Suarez, B. K., Lin, J. & Reich, T. A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.* **70**, 461–471 (2002).
7. Moore, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* **56**, 73–82 (2003).
8. Ritchie, M. D. *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69**, 138–147 (2001).
This was the original paper describing the popular MDR method.
9. Hahn, L. W., Ritchie, M. D. & Moore, J. H. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics* **19**, 376–382 (2003).
10. Moore, J. H. Computational analysis of gene–gene interactions using multifactor dimensionality reduction. *Expert Rev. Mol. Diagn.* **4**, 795–803 (2004).
11. Chung, Y., Lee, S. Y., Elston, R. C. & Park, T. Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions. *Bioinformatics* **23**, 71–76 (2007).
12. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
13. Zhang, Y. & Liu, J. S. Bayesian inference of epistatic interactions in case–control studies. *Nature Genet.* **39**, 1167–1173 (2007).
This paper proposed a new Bayesian approach for the detection of loci that might interact in the context of GWA studies. The related BEAM software package provides a computationally efficient implementation of the proposed algorithm.
14. Ferreira, T., Donnelly, P. & Marchini, J. Powerful Bayesian gene–gene interaction analysis. *Am. J. Hum. Genet.* **81** (Suppl.), 32 (2007).
15. Gayan, J. *et al.* A method for detecting epistasis in genome-wide studies using case–control multi-locus association analysis. *BMC Genomics* **9**, 360 (2008).
16. Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. & Gauderman, W. J. Exploiting gene–environment interaction to detect genetic associations. *Hum. Hered.* **63**, 111–119 (2007).
17. Fisher, R. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin.* **52**, 399–433 (1918).
18. Hayman, B. I. & Mather, K. The description of genetic interactions in continuous variation. *Biometrics* **11**, 69–82 (1955).
19. Zeng, Z. B., Wang, T. & Zou, W. Modeling quantitative trait loci and interpretation of models. *Genetics* **169**, 1711–1725 (2005).
This paper includes an excellent discussion of issues in the definition and interpretation of interaction in quantitative genetic studies of derived populations (inbred lines).
20. Phillips, P. C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Rev. Genet.* **9**, 855–867 (2008).
An excellent review describing the differing definitions and interpretations of epistasis.
21. Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**, 2463–2468 (2002).
22. Cordell, H. J., Todd, J. A., Bennett, S. T., Kawaguchi, Y. & Farrall, M. Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of *IDDM2* and *IDDM4* with *IDDM1* in type 1 diabetes. *Am. J. Hum. Genet.* **57**, 920–934 (1995).
23. Cox, N. J. *et al.* Loci on chromosomes 2 (*NIDDM1*) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genet.* **21**, 213–215 (1999).
24. Cordell, H. J., Wedig, G. C., Jacobs, K. B. & Elston, R. C. Multilocus linkage tests based on affected relative pairs. *Am. J. Hum. Genet.* **66**, 1273–1286 (2000).
25. Strauch, K., Fimmers, R., Baur, M. & Wienker, T. F. How to model a complex trait 2. Analysis with two disease loci. *Hum. Hered.* **56**, 200–211 (2003).
26. Armitage, P., Berry, G. & Matthews, J. N. S. *Statistical Methods in Medical Research* 4th edn (Blackwell Science, Chichester, 2002).
27. McCullagh, P. & Nelder, J. A. *Generalized Linear Models* (Chapman & Hall, London, 1989).
28. Neuman, R. J. & Rice, J. P. Two-locus models of disease. *Genet. Epidemiol.* **9**, 347–365 (1992).
29. Li, W. & Reich, J. A complete enumeration and classification of two-locus disease models. *Hum. Hered.* **50**, 334–349 (2000).
30. Hallgrimsdottir, I. B. & Yuster, D. S. A complete classification of epistatic two-locus models. *BMC Genet.* **9**, 17 (2008).
31. McKinney, B. A., Reif, D. M., Ritchie, M. D. & Moore, J. H. Machine learning for detecting gene–gene interactions: a review. *Appl. Bioinformatics* **5**, 77–88 (2006).
32. Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. *Stat. Med.* **13**, 153–162 (1994).
An important paper showing the use of case-only designs for detection of gene–environment interactions in epidemiological studies.
33. Yang, Q., Khoury, M. J., Sun, F. & Flanders, W. D. Case-only design to measure gene–gene interaction. *Epidemiology* **10**, 167–170 (1999).
34. Weinberg, C. R. & Umbach, D. M. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am. J. Epidemiol.* **152**, 197–203 (2000).
35. Mukherjee, B. *et al.* Tests for gene–environment interaction from case–control data: a novel study of type I error, power and designs. *Genet. Epidemiol.* **32**, 615–626 (2008).
36. Zhao, J., Jin, L. & Xiong, M. Test for interaction between two unlinked loci. *Am. J. Hum. Genet.* **79**, 831–845 (2006).
37. Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Rev. Genet.* **4**, 701–709 (2003).
38. Mukherjee, B. & Chatterjee, N. Exploiting gene–environment independence for analysis of case–control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64**, 685–694 (2008).
39. Yang, Y., Houle, A. M., Letendre, J. & Richter, A. *RET* Gly691Ser mutation is associated with primary vesicoureteral reflux in the French-Canadian population from Quebec. *Hum. Mutat.* **29**, 695–702 (2008).
40. Moore, J. H. *et al.* A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.* **241**, 252–261 (2006).
41. Chanda, P. *et al.* Information-theoretic metrics for visualizing gene–environment interactions. *Am. J. Hum. Genet.* **81**, 939–963 (2007).
42. Kang, G. *et al.* An entropy-based approach for testing genetic epistasis underlying complex diseases. *J. Theor. Biol.* **250**, 362–374 (2008).
43. Dong, C. *et al.* Exploration of gene–gene interaction effects using entropy-based methods. *Eur. J. Hum. Genet.* **16**, 229–235 (2008).
44. Zwick, M. An overview of reconstructability analysis. *Kybernetes* **33**, 877–905 (2004).
An excellent overview of some of the principles and techniques used in information-theory modelling of frequency and probability distributions.
45. Cordell, H. J. & Clayton, D. G. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to *HLA* in type 1 diabetes. *Am. J. Hum. Genet.* **70**, 124–141 (2002).
46. Cordell, H. J., Barratt, B. J. & Clayton, D. G. Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene–gene and gene–environment interactions and parent-of-origin effects. *Genet. Epidemiol.* **26**, 167–185 (2004).
This paper describes a regression-based framework for the analysis of family-based data that allows tests of interaction that are similar to the tests often used in case–control studies to be performed.
47. Martin, E. R., Ritchie, M. D., Hahn, L., Kang, S. & Moore, J. H. A novel method to identify gene–gene effects in nuclear families: the MDR-PDT. *Genet. Epidemiol.* **30**, 111–123 (2006).
48. Kotti, S., Bickeboller, H. & Clerget-Darpoux, F. Strategy for detecting susceptibility genes with weak or no marginal effect. *Hum. Hered.* **63**, 85–92 (2007).

49. Lou, X. Y. *et al.* A combinatorial approach to detecting gene–gene and gene–environment interactions in family studies. *Am. J. Hum. Genet.* **83**, 457–467 (2008).
50. Gauderman, W. J. Sample size requirements for association studies of gene–gene interaction. *Am. J. Epidemiol.* **155**, 478–484 (2002).
51. Hein, R., Beckmann, L. & Chang-Claude, J. Sample size requirements for indirect association studies of gene–environment interactions (G x E). *Genet. Epidemiol.* **32**, 235–245 (2008).
52. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.* **37**, 413–417 (2005).
This paper highlights the importance and feasibility of fitting interaction models using GWA data.
53. Chapman, J. & Clayton, D. Detecting association using epistatic information. *Genet. Epidemiol.* **31**, 894–909 (2007).
54. Motsinger, A., Lee, S., Mellick, G. & Ritchie, M. GPNN: power studies and applications of a neural network method for detecting gene–gene interactions in studies of human disease. *BMC Bioinformatics* **7**, 39 (2006).
55. Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W. & Ritchie, M. D. Comparison of approaches for machine-learning optimization of neural networks for detecting gene–gene interactions in genetic epidemiology. *Genet. Epidemiol.* **32**, 325–340 (2008).
56. Lunn, D. J., Whittaker, J. C. & Best, N. A Bayesian toolkit for genetic association studies. *Genet. Epidemiol.* **30**, 231–247 (2006).
57. Hoh, J. *et al.* Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Ann. Hum. Genet.* **64**, 413–417 (2000).
58. Millstein, J., Conti, D. V., Gilliland, F. D. & Gauderman, W. J. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am. J. Hum. Genet.* **78**, 15–27 (2006).
59. Bochdanovits, Z. *et al.* Genome-wide prediction of functional gene–gene interactions inferred from patterns of genetic differentiation in mice and men. *PLoS ONE* **3**, e1593 (2008).
60. Emily, M., Mailund, T., Schauser, L. & Schierup, M. H. Using biological networks to search for interacting loci in genomewide association studies. *Eur. J. Hum. Genet.* 11 Mar 2009 (doi: 10.1038/ejhg.2009.15).
61. Moore, J. H. & Williams, S. M. New strategies for identifying gene–gene interactions in hypertension. *Ann. Med.* **34**, 88–95 (2002).
62. Golub, G., Heath, M. & Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–224 (1979).
63. Velez, D. R. *et al.* A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol.* **31**, 306–315 (2007).
64. Copas, J. B. Regression, prediction and shrinkage. *J. Roy. Stat. Soc., Series B* **45**, 311–354 (1983).
65. Hastie, T., Tibshirani, R., & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer, New York, 2001).
66. Lee, A. & Silvapulle, M. Ridge estimation in logistic regression. *Comm. Stat. Simul. Comput.* **17**, 1231–1257 (1988).
67. Le Cessie, S. & Van Houwelingen, J. Ridge estimators in logistic regression. *Appl. Stat.* **41**, 191–201 (1992).
68. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Statist.* **32**, 407–499 (2004).
69. Park, M. Y. & Hastie, T. Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50 (2008).
70. Zhang, Z., Zhang, S., Wong, M. Y., Wareham, N. H. & Sha, Q. An ensemble learning approach jointly modelling main and interaction effects in genetic association studies. *Genet. Epidemiol.* **32**, 285–300 (2008).
71. Zhang, H. & Bonney, G. Use of classification trees for association studies. *Genet. Epidemiol.* **19**, 323–332 (2000).
72. Nelson, M. R., Kardia, S. L., Ferrell, R. E. & Sing, C. F. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* **11**, 458–470 (2001).
73. Culverhouse, R., Klein, T. & Shannon, W. Detecting epistatic interactions contributing to quantitative traits. *Genet. Epidemiol.* **27**, 141–152 (2004).
74. McKinney, B. A., Crowe, J. E., Guo, J. & Tian, D. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genet.* **5**, e1000432 (2009).
75. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
76. Lunetta, K. L., Hayward, L. B., Segal, J. & Van Eerdewegh, P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* **5**, 32 (2004).
77. Bureau, A. *et al.* Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.* **28**, 171–182 (2005).
78. Schwartz, D. F., Ziegler, A. & König, I. R. Beyond the results of genome-wide association studies. *Genet. Epidemiol.* **32**, 671 (2008).
79. Kooperberg, C., Ruczinski, I., LeBlanc, M. & Hsu, L. Sequence analysis using logic regression. *Genet. Epidemiol.* **21**, S626–S631 (2001).
80. Kooperberg, C. & Ruczinski, I. Identifying interacting SNPs using Monte Carlo logic regression. *Genet. Epidemiol.* **28**, 157–170 (2005).
81. Nunkesser, R., Bernholt, T., Schwender, H., Ickstadt, K. & Wegener, I. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics* **23**, 3280–3288 (2007).
82. Li, Z., Zheng, T., Califano, A. & Floratos, A. Pattern-based mining strategy to detect multi-locus association and gene × environment interaction. *BMC Proc.* **1** (Suppl. 1), S16 (2007).
83. Long, Q., Zhang, Q. & Ott, J. Detecting disease-associated genotype patterns. *BMC Bioinform.* **10** (Suppl. 1), S75 (2009).
84. Cho, Y. M. *et al.* Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. *Diabetologia* **47**, 549–554 (2004).
85. Julia, A. *et al.* Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction. *Genomics* **90**, 6–13 (2007).
86. Tsai, C. T. *et al.* Renin–angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene–gene interaction. *Atherosclerosis* **195**, 172–180 (2007).
87. Lee, S. Y., Chung, Y., Elston, R. C., Kim, Y. & Park, T. Log-linear model based multifactor-dimensionality reduction method to detect gene–gene interactions. *Bioinformatics* **23**, 2589–2595 (2007).
88. Lou, X. Y. *et al.* A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.* **80**, 1125–1137 (2007).
89. Robnik-Sikonja, M. & Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **53**, 23–69 (2003).
90. Moore, J. H. & White, B. C. Tuning ReliefF for genome-wide genetic analysis. *Lect. Notes Comp. Sci.* **4447**, 166–175 (2007).
91. McKinney, B. A., Reif, D. M., White, B. C., Crowe, J. & Moore, J. H. Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics* **23**, 2113–2120 (2007).
92. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* (Chapman and Hall, London, 1995).
93. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. *Markov Chain Monte Carlo in Practice* (Chapman and Hall, London, 1996).
94. Hoggart, C. J., Whittaker, J. C., De Iorio, M. & Balding, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4**, e1000130 (2008).
95. Phillips, P. C. The language of gene interaction. *Genetics* **149**, 1167–1171 (1998).
An important paper that describes the differing definitions and interpretations of epistasis used in different fields and the lack of equivalence between these definitions.
96. Moore, J. H. & Williams, S. M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* **27**, 637–646 (2005).
97. Cheverud, J. M. & Routman, E. J. Epistasis and its contribution to genetic variance components. *Genetics* **139**, 1455–1461 (1995).
98. Alvarez-Castro, J. M. & Carlborg, O. A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* **176**, 1151–1167 (2007).
99. McClay, J. L. & van den Oord, E. J. Variance component analysis of polymorphic metabolic systems. *J. Theor. Biol.* **240**, 149–159 (2006).
100. Thompson, W. D. Effect modification and the limits of biological inference from epidemiologic data. *J. Clin. Epidemiol.* **44**, 221–232 (1991).
101. Siemiatycki, J. & Thomas, D. C. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int. J. Epidemiol.* **10**, 383–387 (1981).
102. Greenland, S. Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology* **20**, 14–17 (2009).
A useful commentary on the relationship between statistical and biological interaction assessed from epidemiological studies.
103. Gibson, G. Epistasis and pleiotropy as natural properties of transcriptional regulation. *Theor. Popul. Biol.* **49**, 58–89 (1996).
104. Vanderweele, T. J. Sufficient cause interactions and statistical interactions. *Epidemiology* **20**, 6–13 (2009).
105. Todd, J. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genet.* **39**, 857–864 (2007).
106. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
107. Sepulveda, N., Paulino, C. D., Carneiro, J. & Penha-Goncalves, C. Allelic penetrance approach as a tool to model two-locus interaction in complex binary traits. *Heredity* **99**, 173–184 (2007).
108. Sepulveda, N., Paulino, C. D. & Penha-Goncalves, C. Bayesian analysis of allelic penetrance models for complex binary traits. *Comp. Stat. Data Anal.* **53**, 1271–1283 (2009).
109. Aylor, D. L. & Zeng, Z. B. From classical genetics to quantitative genetics to systems biology: modeling epistasis. *PLoS Genet.* **4**, e1000029 (2008).
110. Curtis, D. Allelic association studies of genome wide association data can reveal errors in marker position assignments. *BMC Genet.* **8**, 30 (2007).
111. Breiman, L., Freidman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees* (Chapman and Hall/CRC, New York, 1984).
112. Bastone, L., Reilly, M., Rader, D. J. & Foulkes, A. S. MDR and PRP: a comparison of methods for high-order genotype–phenotype associations. *Hum. Hered.* **58**, 82–92 (2004).
113. Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007).
This paper gives an overview of some of the strengths and limitations of random forests analysis for measuring variable importance.

Acknowledgements
Support for this work was provided by the Wellcome Trust (Grant reference 074524). I thank J. Barrett for assistance with interpretation of the WTCCC Crohn's results, and the WTCCC for making their data freely available. I also thank J. Moore for useful discussions of data-mining methods in general and MDR in particular, and K. Keen for pointing out the origins of the term epistasis.

DATABASES
OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Crohn's disease

FURTHER INFORMATION
Heather J. Cordell's homepage: http://www.staff.ncl.ac.uk/heather.cordell
BEAM: http://www.people.fas.harvard.edu/~junliu/BEAM
MDR: http://sourceforge.net/projects/mdr
Nature Reviews Genetics Series on Genome-wide association studies: http://www.nature.com/nrg/series/gwas/index.html
Nature Reviews Genetics Series on Modelling: http://www.nature.com/nrg/series/modelling/index.html
PLINK: http://pngu.mgh.harvard.edu/~purcell/plink
Random Jungle: http://randomjungle.com

SUPPLEMENTARY INFORMATION
See online article: S1 (box) | S2 (box) | S3 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Supplementary Box S1: Different models of interaction

The regression model described in Box 1 is quite general, encompassing a number of different specific cases. Suppose we consider a model of recessive effects (on the log-odds scale) at each of two diallelic interacting loci, so that the binary factors x_B and x_C correspond to indicators of homozygosity for the risk-modifying allele at each locus. The expected log-odds of disease implied by the regression formulation, given an individual's two-locus genotype combination, are shown below:

		Locus C		
	Genotype	c/c	c/C	C/C
Locus B	b/b	α	α	$\alpha+\gamma$
	b/B	α	α	$\alpha+\gamma$
	B/B	$\alpha+\beta$	$\alpha+\beta$	$\alpha+\beta+\gamma+i$

If, instead, we consider a dominant model, whereby a single allele at each locus is sufficient to modify disease risk, we obtain the expected log-odds:

		Locus C		
	Genotype	c/c	c/C	C/C
Locus B	b/b	α	$\alpha+\gamma$	$\alpha+\gamma$
	b/B	$\alpha+\beta$	$\alpha+\beta+\gamma+i$	$\alpha+\beta+\gamma+i$
	B/B	$\alpha+\beta$	$\alpha+\beta+\gamma+i$	$\alpha+\beta+\gamma+i$

The actual value of the expected log-odds in each genotype category will depend on the values of the regression parameters α, β, γ and i. For example, under the recessive model, if these parameters took values $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 1$ and $i = 3$, we would obtain log-odds values:

		Locus C		
	Genotype	c/c	c/C	C/C
Locus B	b/b	0.5	0.5	1.5
	b/B	0.5	0.5	1.5
	B/B	1	1	5

The penetrance values (probabilities of getting disease) corresponding to this model (i.e. the values of p rather than of $\ln[p/(1-p)]$) may be calculated using the identity $p = \frac{\exp(\ln[p/(1-p)])}{1+\exp(\ln[p/(1-p)])}$, and are:

		Locus C		
	Genotype	c/c	c/C	C/C
Locus B	b/b	0.62	0.62	0.82
	b/B	0.62	0.62	0.82
	B/B	0.73	0.73	0.99

Note that models with interaction effects on one scale (e.g. the penetrance scale) may correspond to models with no interaction effects on another scale (e.g. the log-odds scale). For example, if under the recessive model on the log-odds scale the regression parameters took values $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 1$ and $i = 0$, we would obtain log-odds values:

		Locus C		
	Genotype	c/c	c/C	C/C
Locus B	b/b	0.5	0.5	1.5
	b/B	0.5	0.5	1.5
	B/B	1	1	2

Here, possession of the risk genotype B/B adds a unit of 0.5 to the log-odds while posession of the risk genotype C/C adds a unit of 1.0 to the log-odds, with no additional (interaction) term required for possession of risk genotypes at both loci. The penetrance values corresponding to this model are:

		Locus C		
	Genotype	c/c	c/C	C/C
Locus B	b/b	0.62	0.62	0.82
	b/B	0.62	0.62	0.82
	B/B	0.73	0.73	0.88

Here, possession of the risk genotype B/B adds a unit of 0.11 to the penetrance while posession of the risk genotype C/C adds a unit of 0.20. However, *subtraction* of an additional -0.05 (i.e. an interaction term) is required when both risk genotypes (B/B and C/C) are possessed. This example illustrates the well-known fact that statistical interaction effects are affected by changes of scale [1]: essentially the regression parameters, including interaction terms, are defined relative to some particular scale of interest. This phenomenon has led to some confusion in terminology [2] concerning whether interaction effects

represent departure from a linear (i.e. additive) model or from a multiplicative model, with respect to the main effects of the two loci. A model that is additive on the log-odds scale will be equivalent to a model that is multiplicative on the odds scale, and so departure from either of these models may be considered as equivalent. However, this departure would not be equivalent to departure from multiplicativity on the *original* log-odds scale.

A more general 'genotype' model for the effects of two loci allows for different parameters to represent the effects of having a single copy (i.e. being heterozygous) or two copies (i.e. being homozygous) of a risk-modifying allele, as shown below:

		Locus C		
	Genotype	c/c	c/C	C/C
Locus B	b/b	α	$\alpha+\gamma_1$	$\alpha+\gamma_2$
	b/B	$\alpha+\beta_1$	$\alpha+\beta_1+\gamma_1+i_{11}$	$\alpha+\beta_1+\gamma_2+i_{12}$
	B/B	$\alpha+\beta_2$	$\alpha+\beta_2+\gamma_1+i_{21}$	$\alpha+\beta_2+\gamma_2+i_{22}$

This model includes nine different parameters: a parameter α that represents the 'baseline' log-odds for an individual who has genotypes b/b and c/c, parameters β_1 and β_2 representing the effects of replacing one or both alleles at locus B with the modifying allele B, parameters γ_1 and γ_2 representing the effects of replacing one or both alleles at locus C with the modifying allele C and four interaction parameters i_{11}, i_{12}, i_{21}, and i_{22}. This is known statistically as a 'saturated' model, which means that it is fully parameterized: nine two-locus genotype categories are modelled by nine parameters, and so these parameters may be chosen (estimated) to fit the observed nine two-locus penetrances or log-odds values precisely. No other model exists that can fit the observed penetrances any better. All other models can be considered as sub-models of ('nested' in) this most general model. Although the saturated model provides the best possible fit to the data, it includes many parameters. In statistical terms, we are usually interested in determining whether a model with fewer parameters can fit the data 'almost as well'. The 4 degree of freedom (df) test of interaction ($i_{11} = i_{12} = i_{21} = i_{22} = 0$) tests whether the interaction terms are required at all. We may also make parameter restrictions to the interaction model to generate fewer df (while retaining one or more interaction parameters) and thus increase power. The recessive and dominant models correspond to models in which certain parameters are set equal either to zero or to each other. An alternative is to assume alleles act additively within a locus, which corresponds to assuming

$\beta_2 = 2\beta_1$, $\gamma_2 = 2\gamma_1$, $i_{12} = i_{21} = 2i_{11}$ and $i_{22} = 4i_{11}$. This restriction converts the nine-parameter 'genotype' model into a four parameter 'allelic' model, $\ln[p/(1-p)] = \alpha + \beta_1 x_B + \gamma_1 x_C + i_{11} x_B x_C$, where x_B and x_C are variables taking values (0,1,2) according to the number of risk alleles at locus B and C respectively. This model contains a single interaction parameter i_{11} that may be freely estimated; a modified version of this model, that makes further restrictions on the relative magnitudes of β_1, γ_1 and i_{11}, has also been proposed [3].

References

[1] Frankel, W. N. and Schork, N. J. (1996). Who's afraid of epistasis? Nat Genet *14*, 371–373.

[2] Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Molec Genet *11*, 2463–2468.

[3] Wang, K. (2008). Genetic association tests in the presence of epistasis or gene-environment interaction. Genet Epidemiol *32*, 606–614.

Supplementary Box S2: Effects – interacting, independent or otherwise

A key issue in defining and interpreting genetic interaction – epistasis – is understanding what is meant by the 'effect' of a locus, and what is meant by 'independent' effects of several loci. These concepts were first introduced in genetics by Bateson et al. [1] who described the concept of a character (phenotype) produced by the meeting of two distinct genetic factors, without using the specific terms 'interaction', 'epistasis', 'epistacy' or 'epistatic'. Subsequently, Bateson [2] used the term 'interaction' to describe this concept in the situation where one factor is not visible unless the other is also present, and the term 'epistatic' [3] [4] to describe this concept in the context of one factor preventing another from manifesting its effect. This terminology may perhaps originate from an earlier paper by Gadow [5] who used the term 'epistasis' in the context of arrested development in lizards, citing a German paper by Eimer [6] as the origin of the term.

The Batesonian concept of epistasis can be described in relation to tables such as the one shown below:

		Locus C		
	Genotype	*c*/*c*	*c*/*C*	*C*/*C*
Locus B	*b*/*b*	White	Brown	Brown
	b/*B*	Black	Brown	Brown
	B/*B*	Black	Brown	Brown

This table shows the coat colour in mice that results from a specific combination of two genetic factors. Note that here there is a clear (prior) understanding that the 'baseline' (reference point) genotype is the wild-type combination (*b*/*b*, *c*/*c*) which displays a phenotype of no colour (i.e. white), and that the effect of allele *B* at locus B is to change the color to black, while the effect of allele *C* at locus C is to change the colour to brown. Therefore, the modifying alleles at the different loci not only have different 'effects' but they also lead to different phenotype manifestations (black/brown) – meaning that which locus is operating can be determined directly by looking at the phenotype. This situation is perhaps somewhat analagous to consideration of biochemical interactions between proteins, where the function of each protein differs and has been well-established *a priori*.

Given well-defined effects such as these, the obvious question is what happens when modifiying

alleles at both loci are present. One might speculate as to what one might *expect* to happen if the alleles continued to act 'independently' – would the coat colour perhaps be mottled? In the table above we see that this does not happen; the alleles at locus C take precedence and locus C is said to be epistatic to locus B (or, more precisely, allele *C* at locus C is said to be epistatic to allele *B* at locus B).

Things became confused when Fisher [7] used the terms 'epistacy' and 'epistatic' to describe an apparently rather different concept, defined in terms of linear effects on a quantitative trait, much closer to the concept of statistical interaction described in Box 1. Indeed, R.C. Punnet pointed out this apparent difference in concept in his review of Fisher's paper [8]. Subsequently, the terms 'epistasis', 'epistacy', 'epistasy' or 'epistatic' seem to have been used more-or-less interchangably, but with potentially different implied meanings. In the quantitative genetics literature [9] (and more recently the human complex genetic disease literature) the usage seems to have mostly stemmed from Fisher's definition i.e. a statistical interaction signifying departure from linear effects with respect to prediction of a trait outcome, whereas biologists and biochemists have mostly used functional definitions closer in form to Batesonian epistasis.

The classical quantitative genetics formulation takes several different forms depending on the reference point and inbred line in question [9 10 11]; one common form is the F_∞ model shown below:

		Locus C		
	Genotype	c/c	c/C	C/C
Locus B	b/b	$\mu - a_b - a_c$	$\mu - a_b + d_c$	$\mu - a_b + a_c$
	b/B	$\mu + d_b - a_c$	$\mu + d_b + d_c + i_{dd}$	$\mu + d_b + a_c + i_{da}$
	B/B	$\mu + a_b - a_c$	$\mu + a_b + d_c + i_{ad}$	$\mu + a_b + a_c + i_{aa}$

This table shows the expected quantitative trait value for each genotype combination. In human genetics, rather than tabulating expected quantitative trait values, one might tabulate the expected log-odds or penetrance values as described in Supplementary Box S1. For simple Mendelian disorders, one would anticipate that the pentrances values should all be either 0 or 1, leading to penetrance tables such as:

		Locus C		
	Genotype	c/c	c/C	C/C
	b/b	0	0	1
Locus B	b/B	0	0	1
	B/B	1	1	1

The table above has classically been considered to represent a heterogeneity or non-epistatic model [12] (since one can aquire the disease through having the high-risk genotype at either or both loci) but note that this interpretation depends crucially on what we consider the 'effect' of each locus to be [13]. Although a 0/1 penetrance classification might seem at first sight to be similar to a categorical phenotype (as in the mouse coat colour example), in fact it is not completely equivalent since risk alleles at the two loci do not lead to different phenotype manifestations and so it is not clear which locus is actually 'causing' the phenotype; in a sense, for each cell, it is the genotype combination at both loci that 'causes' the disease. In practice, for complex diseases we do not expect to see pentrances values of 0 or 1, rather we expect a continuum of disease risks leading to penetrance tables such as:

		Locus C		
	Genotype	c/c	c/C	C/C
	b/b	0.1	0.2	0.2
Locus B	b/B	0.3	0.4	0.4
	B/B	0.3	0.4	0.4

Here, whether or not the loci 'interact' depends on what one defines the 'effect' of each locus to be. If one defines the 'effect' of a risk genotype at locus B to be the addition of a term 0.2 to the baseline pentrance, and the 'effect' of a risk genotype at locus C to be the addition of a term 0.1 to the baseline pentrance, then the loci above do not interact. If one defines the 'effect' of a risk genotype at locus B to be the multiplication of the baseline pentrance by a factor of 3, and the 'effect' of a risk genotype at locus C to be the multiplication of the baseline pentrance by a factor of 2, then the loci do interact (the non-interactive model would have values 0.6 instead of 0.4 in the table above). If one defined the 'effect' of a risk genotype at locus B to be the *conferring* of a penetrance value of 0.2 and the 'effect' of a risk genotype at locus B to be the *conferring* of a penetrance value of 0.3 then it is unclear what

the non-interactive model should be - perhaps the conferring of an average penetrance value of 0.25 instead of 0.4 in the relevant cells of the above table? Hence, depending on our definition of 'effect', and what we expect to observe if the effects operate 'independently', we may come to different conclusions concerning the presence or absence of interaction between the loci.

The relationship between linear statistical models for outcomes as observed in a population and 'effects' in terms of possible underlying biological causal mechanisms has been debated extensively in the epidemiological literature [14 15 16]. Of particular interest in this debate is the sufficient cause framework [17 18 19], in which it may be postulated that certain 'causes' of an outcome (e.g. disease) participate together in the same causal mechanism (resulting in so-called 'synergism'). Although departure from additivity with respect to a linear model defined on the absolute risk (as opposed to the log-odds) scale can, in some situations, allow one to conclude the presence of interaction or synergism in the sufficient cause sense [20], the assumptions and conditions required for this conclusion to hold are quite restrictive. It has been shown that, even if the assumptions of no unmeasured confounding and correct specification of the statistical model are met, interaction terms in statistical models do not, in fact, in general correspond to interaction or synergism in the sufficient cause sense [20].

References

[1] Bateson, W., Saunders, E. R., and Punnett, R. C. (1905). Further experiments on inheritence in sweet peas and stocks. Proc Roy Soc B *77*, 236–238.

[2] Bateson, W. (1906). The progress of genetics since the rediscovery of Mendel's papers. Prog Res Bot *1*, 368–418.

[3] Bateson, W. (1907). Facts limiting the theory of heredity. Science *26*, 649–660.

[4] Bateson, W. (1909). Mendel's principles of heredity. (Cambridge University Press).

[5] Gadow, H. (1903). Evolution of the colour-pattern and orthogenetic variation in certain Mexican species of lizards, with adaptation to their surroundings. Roy Soc Proc *72*, 109–125.

[6] Eimer, T. (1881). Untersuchungen ueber das Variiren der Mauereidechse, ein Beitrag zur Theorie von der Entwicklung aus constitutionellen Ursachen, sowie zum Darwinismus. Arch f Naturg *47*, 239–517.

[7] Fisher, R. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edin *52*, 399–433.

[8] Norton, B. and Pearson, E. S. (1976). A note on the background to and refereeing of R.A. Fisher's 1918 paper 'The correlation between relatives on the supposition of Mendelian inheritance'. Notes Rec R Soc Lond *31*, 151–162.

[9] Hayman, B. I. and Mather, K. (1955). The description of genetic interactions in continuous variation. Biometrics *11*, 69–82.

[10] Mather, K. and Jinks, J. L. (1982). Biometrical Genetics, 3rd Edition. (Chapman & Hall, London).

[11] Zeng, Z. B., Wang, T., and Zou, W. (2005). Modeling quantitative trait Loci and interpretation of models. Genetics *169*, 1711–1725.

[12] Neuman, R. J. and Rice, J. P. (1992). Two-locus models of disease. Genet Epidemiol *9*, 347–365.

[13] Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Molec Genet *11*, 2463–2468.

[14] Thompson, W. D. (1991). Effect modification and the limits of biological inference from epidemiologic data. Journal of Clinical Epidemiology *44*, 221–232.

[15] Siemiatycki, J. and Thomas, D. C. (1981). Biological models and statistical interactions: an example from multistage carcinogenesis. International Journal of Epidemiology *10*, 383–387.

[16] Greenland, S. (2009). Interactions in epidemiology: relevance, identification, and estimation. Epidemiology *20*, 14–17.

[17] Rothman, K. J. (1976). Causes. Am J Epidemiol *104*, 587–592.

[18] VanderWeele, T. J. and Robins, J. M. (2007). The identification of synergism in the sufficient-component-cause framework. Epidemiology *18*, 329–339.

[19] VanderWeele, T. J. and Robins, J. M. (2008). Empirical and counterfactual conditions for sufficient cause interactions. Biometrika *95*, 49–61.

[20] VanderWeele, T. J. (2009). Sufficient cause interactions and statistical interactions. Epidemiology *20*, 6–13.

Supplementary Table 1: Top pairwise interactions as detected from a `--fast-epistasis` analysis of the WTCCC Crohn's disease and control data using PLINK

SNP1				SNP2				Interaction test	
			Single-locus				Single-locus		
SNP	Chr	Position	*p* value	SNP	Chr	Position	*p* value	χ^2	*p* value
rs9436212	1	61643372	0.08463	rs11649428	16	76266378	0.1423	37.94	7.315e-10
rs12751992	1	63186351	0.1514	rs1601668	12	18826761	0.08044	35.10	3.144e-09
rs4677143	3	72439370	0.04577	rs8006622	14	39419113	0.05631	35.32	2.809e-09
rs1584444	4	59536174	0.02352	rs2201677	4	102215684	0.07998	38.25	6.233e-10
rs1584444	4	59536174	0.02352	rs12647454	4	102219636	0.0633	37.73	8.157e-10
rs1584444	4	59536174	0.02352	rs6532916	4	102227529	0.02714	38.18	6.493e-10
rs1584444	4	59536174	0.02352	rs10027689	4	102233096	0.03322	35.26	2.899e-09
rs668394	6	154460193	0.1078	rs10156534	9	2926864	0.1464	38.13	6.637e-10
rs511435	6	154460661	0.1116	rs10156534	9	2926864	0.1464	36.56	1.486e-09
rs509544	6	154460900	0.1049	rs10156534	9	2926864	0.1464	37.66	8.461e-10
rs524731	6	154467206	0.1094	rs10156534	9	2926864	0.1464	37.19	1.078e-09
rs7773053	6	156486474	0.1308	rs17825620	14	77254414	0.1664	36.94	1.225e-09
rs2358356	10	19538800	0.08709	rs9540533	13	65213814	0.06114	35.53	2.517e-09
rs2478836	10	30454756	0.09764	rs7217284	17	56691176	0.187	37.43	9.515e-10
rs636646	13	76492331	0.04548	rs301630	16	85211439	0.009047	37.56	8.895e-10
rs7202714	16	30085308	0.0322	rs4471699	16	30227808	1.593e-22	51.56	6.979e-13