

Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans

Heather J. Cordell*

University of Cambridge, Department of Medical Genetics, JDRF/WT Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research (CIMR), Wellcome Trust/MRC Building, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2XY, UK

Received July 8, 2002; Accepted July 27, 2002

Epistasis, the interaction between genes, is a topic of current interest in molecular and quantitative genetics. A large amount of research has been devoted to the detection and investigation of epistatic interactions. However, there has been much confusion in the literature over definitions and interpretations of epistasis. In this review, we provide a historical background to the study of epistatic interaction effects and point out the differences between a number of commonly used definitions of epistasis. A brief survey of some methods for detecting epistasis in humans is given. We note that the degree to which statistical tests of epistasis can elucidate underlying biological interactions may be more limited than previously assumed.

Epistasis, defined generally as the interaction between different genes, has become a hot topic in complex disease genetics in recent years. For complex traits such as diabetes, asthma, hypertension and multiple sclerosis, the search for susceptibility loci has, to date, been less successful than for simple Mendelian disorders. This is probably due to complicating factors such as an increased number of contributing loci and susceptibility alleles, incomplete penetrance, and contributing environmental effects. The presence of epistasis is a particular cause for concern, since, if the effect of one locus is altered or masked by effects at another locus, power to detect the first locus is likely to be reduced and elucidation of the joint effects at the two loci will be hindered by their interaction. If more than two loci are involved, the situation is likely to be further complicated by the possibility of complex multiway interactions among some or all of the contributing loci.

In the literature, discussion of epistasis has been considerably confused by differing (usually unstated) definitions and assumptions and by the use of the same terminology to apply to quite different statistical and biological concepts. In essence, epistasis refers to departure from 'independence' of the effects of different genetic loci in the way that they combine to cause disease. However, what is meant by 'independence' is not always precisely stated, and indeed may vary, particularly between the definitions assumed by biologists, epidemiologists, statisticians, and human and quantitative geneticists.

DEFINITIONS OF EPISTASIS

The term 'epistatic' was first used in 1909 by Bateson (1) to describe a masking effect whereby a variant or allele at one

Table 1. Example of phenotypes (e.g. hair colour) obtained from different genotypes at two loci interacting epistatically, under Bateson's (1909) definition of epistasis

Genotype at locus B	Genotype at locus G		
	<i>g/g</i>	<i>g/G</i>	<i>G/G</i>
<i>b/b</i>	White	Grey	Grey
<i>b/B</i>	Black	Grey	Grey
<i>B/B</i>	Black	Grey	Grey

locus (denoted at that time as an 'allelomorphic pair') prevents the variant at another locus from manifesting its effect. This was seen as an extension of the concept of dominance for alleles within the same allelomorphic pair i.e. at a single locus. Suppose we have two loci, B and G, that influence a trait such as hair colour in mice. Locus B has two possible alleles, *B* or *b*, and locus G has two possible alleles, *G* or *g*. The possible phenotypic outcomes (white, black or grey hair) given genotype are as shown in Table 1. We see that, regardless of genotype at locus B, individuals with any copies of the *G* allele have grey hair, i.e., at locus G, allele *G* is dominant to allele *g*, effectively masking any 'effect' of allele *g*. We also see that if the genotype at locus G is *g/g* then an individual with any copies of the *B* allele has black hair, so that at locus B, allele *B* is dominant to *b*. If, however, the genotype at locus G is not *g/g* then the effect at locus B is not observable, since individuals with any copies of the *G* allele have grey hair regardless of genotype at locus B, i.e. the effect at locus B is masked by that of locus G: locus G is said to be epistatic to locus B (or, more specifically, allele *G* at locus G is epistatic to allele *B* at locus B).

Table 2. Example of penetrance table for two loci interacting epistatically in a general sense

Genotype at locus A	Genotype at locus B		
	<i>b/b</i>	<i>b/B</i>	<i>B/B</i>
<i>a/a</i>	0	0	0
<i>a/A</i>	0	1	1
<i>A/A</i>	0	1	1

This definition of epistasis is similar to the concept most often used by a biologist or biochemist when investigating biological interaction between proteins. In fact, what is meant by biological interaction is not always precisely defined (2), but usually corresponds to a situation in which the qualitative nature of the mechanism of action of a factor is affected by the presence or absence of the other (3). There are, however, some problems with this definition, particularly when applied to binary traits. In human genetics, the phenotype of interest is often qualitative and usually dichotomous, indicating presence or absence of disease. Mathematical models for the joint action of two or more loci usually focus on the penetrance, the probability of developing disease given genotype. Suppose that we have two loci, A and B, that influence a binary trait. Locus A has two possible alleles, *A* or *a*, and locus B has two possible alleles, *B* or *b*. Suppose that a predisposing allele is required at both loci in order to exhibit the trait, i.e. one or more copies of both allele *A* and allele *B* are required. Then, when the effects of both loci are considered, we obtain the penetrance table shown in Table 2. In this table, the effect of allele *A* can only be observed when allele *B* is also present: without the presence of *B*, the effect of *A* is not observable. The effect at locus A would appear to be 'masked' by that at locus B. By analogy with the example in Table 1, we might say that locus B is epistatic to locus A, since when the genotype *b/b* is present at locus B, the effect of the alleles at locus A is not observable. However, we might equally well say that locus A is epistatic to locus B, since when the genotype *a/a* is present at locus A, the effect of the alleles at locus B is not observable. This leads to a situation that is not precisely analogous to that described by Bateson (1). In Bateson's (1) definition, it is clear that if factor B is epistatic to factor A, we do not expect factor A to also be epistatic to factor B. This is illustrated by the lack of symmetry in Table 1. In Table 2, the symmetry between the effects at loci A and B means that we cannot say that one of the loci is epistatic to the other. Nevertheless this type of penetrance table has been interpreted as representing a more general form of epistasis between the loci (4), albeit of a rather different nature from that originally implied by the term.

Lack of epistasis has classically been represented by penetrance tables such as Table 3 (4–6). This table represents a 'heterogeneity model' in which an individual becomes affected through possessing a predisposing genotype at either locus A or locus B. This would appear to represent a different biological phenomenon to that represented by Table 2. Table 3 is usually assumed to correspond to a situation in which the biological pathways involved in disease influenced by the two loci are at some level separate or independent (5). However, it is worth noting that if we define the 'effect' of locus B to be a

Table 3. Example of penetrance table for two loci acting together in a heterogeneity model

Genotype at locus A	Genotype at locus B		
	<i>b/b</i>	<i>b/B</i>	<i>B/B</i>
<i>a/a</i>	0	0	1
<i>a/A</i>	0	0	1
<i>A/A</i>	1	1	1

recessive disease model (so that two copies of allele *B* are required to cause disease) then having two copies of allele *A* at locus A is sufficient to 'mask' this effect, i.e., given genotype *A/A* at locus A, the effect at locus B is not observable: locus B acts differently when the genotype at locus A is *A/A* compared with when the genotype at locus A is not *A/A*. Thus the classical heterogeneity model falls within a definition interpretable as epistasis! This illustrates the inherent problems with biologically motivated definitions of epistasis, unless precise definitions of what is meant by the 'effect' of a locus (or of an allele or genotype at a locus) and 'independence of effects' are given.

The situation has been confused further by the fact that in quantitative genetics, following a paper by Fisher in 1918 (7), the term 'epistatic' has been generally used in yet another different sense from its original usage. In Fisher's 1918 definition, epistasis refers to a deviation from additivity in the effect of alleles at different loci with respect to their contribution to a quantitative phenotype. This definition is not equivalent to Bateson's 1909 definition, as was pointed out in the initial review of Fisher's 1918 paper by R.C. Punnett (8). Epistasis in the Fisher (7) sense is closer to the usual concept of statistical interaction: departure from a specific linear model describing the relationship between predictive factors (here assumed to be alleles at different genetic loci). With this definition, the choice of scale becomes important, since factors that are additive with respect to an outcome measured on one scale may not be additive, i.e. may exhibit interaction when a different, transformed scale is used (2,9).

Mathematical definitions

Mathematically, the quantitative genetic concept of epistasis may be represented (10) for two loci by the linear model

$$y = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + i_{aa}x_1x_2 + i_{ad}x_1z_2 + i_{da}z_1x_2 + i_{dd}z_1z_2,$$

where y is a quantitative phenotype and x_i and z_i are dummy variables related to the underlying genotype at locus i . For example, for a diallelic locus with alleles denoted 1 and 2, we might set $x_i = 1$ and $z_i = -0.5$ for a 1/1 homozygote, $x_i = 0$ and $z_i = 0.5$ for a heterozygote, and $x_i = -1$ and $z_i = -0.5$ for a 2/2 homozygote. The coefficients μ , a_1 , d_1 , a_2 and d_2 represent genetic parameters that may be estimated corresponding to the mean effect and additive and dominance effects at the two loci; i_{aa} , i_{da} , i_{da} and i_{dd} correspond to epistatic interaction effects. Lack of epistasis in this model implies that all

interaction coefficients are zero. If there is no epistasis then the resulting model

$$y = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2$$

is said to be additive, since, after some algebra, it may be written equivalently as

$$y_{ij} = \alpha_i + \beta_j,$$

where α_i and β_j are parameters representing the effects of the genotype i at locus 1 and genotype j at locus 2, respectively (where i and j take values 0, 1, 2 according to the number of copies of the 2 allele present at the locus).

For binary traits, similar models may be applied, but with the outcome of interest usually defined to be p_{ij} , the penetrance for genotype i at locus 1 and j at locus 2. Three models commonly considered are an additive model (5) in which p_{ij} may be written

$$p_{ij} = \alpha_i + \beta_j,$$

a multiplicative model (5,11) in which p_{ij} may be written

$$p_{ij} = \alpha_i\beta_j$$

and a heterogeneity model (4,5) in which p_{ij} may be written

$$p_{ij} = \alpha_i + \beta_j - \alpha_i\beta_j$$

The multiplicative model is usually considered to be an epistatic model in which the loci and pathways involved in disease and influenced by the two loci are not independent. The additive and heterogeneity models are usually assumed to represent non-epistatic models and to correspond to a situation in which the biological pathways are at some level separate or independent. Indeed, the classical heterogeneity model shown in Table 3, whereby an individual becomes affected through possessing a predisposing genotype at either locus 1 or locus 2, can be formulated as

$$p_{ij} = \alpha_i + \beta_j - \alpha_i\beta_j,$$

with $\alpha_0 = \alpha_1 = \beta_0 = \beta_1 = 0$ and $\alpha_2 = \beta_2 = 1$. However, the biological interpretation of the heterogeneity model when the penetrances p_{ij} are not constrained to be 0 or 1 is unclear.

The reason that additive and heterogeneity models for the penetrance are often used interchangeably is that it can be shown that these models give very similar results when used to model familial relative risks of disease (5,12). It is not clear, however, that this will hold when modelling other outcomes (e.g. the penetrances themselves in Table 3 cannot be achieved using an additive model). Note that the heterogeneity and multiplicative models can also both be expressed as additive models when transformed to different scales: the heterogeneity model by transforming to a $\log(1-p)$ scale so that

$$\log(1 - p_{ij}) = \log(1 - \alpha_i) + \log(1 - \beta_j) = A_i + B_j, \quad \text{say,}$$

and the multiplicative model by transforming to a logarithmic scale so that

$$\log p_{ij} = \log \alpha_i + \log \beta_j = A_i + B_j, \quad \text{say.}$$

In a statistical sense, therefore, all three models can be considered to signify independent additive effects on some appropriate scale. Other models that are additive on specific non-penetrance scales include an additive model for the log odds, popular with epidemiologists,

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \alpha_i + \beta_j,$$

and an additive model for liability on a probit scale (13–16), in which the loci are assumed to contribute to an underlying, unobserved, continuous trait in an additive fashion and development of disease occurs if this trait exceeds a certain threshold. Although these models are additive and therefore expressible without interaction effects on their appropriate scales, they correspond to models with interaction effects (epistasis) when transformed to the penetrance scale.

INTERPRETATION OF EPISTASIS

With so many conflicting definitions of epistasis, it is not surprising that there has been much confusion in the literature. Indeed, the term epistasis has often been used without being precisely defined, so that it is not clear which definition is being assumed in any given situation. Many authors have assumed that epistasis or interaction between loci refers to departure from additivity on the penetrance scale (5,17,18), whereas others have assumed instead that it refers to a departure from multiplicativity on the penetrance scale (19–21). Moreover, epistasis is sometimes investigated in the context of epistatic variance: the variation of a trait in a population accounted for by epistatic effects (as opposed to simple additive main effects of individual loci). The epistatic variance depends not only on the genetic model for the action of two or more loci, but also on population parameters such as multilocus genotype frequencies (22,23), in the same way that additive and dominance variances at a single locus depend not just on the model of dominance assumed but also on population genotype or allele frequencies.

Confusions of definition and terminology apart, the main problem with the interpretation of epistasis is that the word itself suggests that we are dealing with a biologically interesting phenomenon. If epistasis is detected, the assumption is that this tells us something of interest about the mechanisms and pathways involved in disease—in particular in relation to the biological interaction between implicated proteins. Indeed, the description in (5) hints strongly for a biological or causal interpretation of the models there defined. However, statistical tests of interaction are limited to testing specific hypotheses concerning precisely defined quantities. For statistical testing, we can only focus on mathematical models of epistasis such as those described (5,7,17) rather than those encoded by a rather abstract and vague notion of ‘independence’ or ‘masking’ of unspecified ‘effects’.

Unfortunately, as we have seen, there is not a precise correspondence between biological models of epistasis and those that are more statistically motivated. We should like to perform a statistical test and interpret the outcome biologically, but this is in general not permissible. Statistical interaction does not necessarily imply interaction on the biological or mechanistic level (24). A brief survey of the epidemiological literature

reveals the major difficulties that exist in inferring biological meaning from quantitative data measuring disease risk as outcome (3,25,26). The problem is that any given data pattern and statistical model can usually be obtained from a number of completely different underlying mechanisms or models for disease development (3,26,27). For instance, five very different causal mechanisms can be shown to all lead to a multiplicative model for the data used in investigating the joint effects of two risk factors (26). Only if the prior biological model can be postulated in some detail is it likely that statistical modelling of this kind will allow insight into the underlying biological mechanisms.

STATISTICAL METHODS TO DETECT EPISTASIS IN HUMANS

Since the biological interpretation hinted at in (5) is likely to be unrealistically simplistic, one might ask why we are interested in detecting epistasis in human genetic studies at all? Although the discovery of epistasis may be of limited value for elucidating the underlying biological disease process, allowing for different modes of interaction between potential disease loci can lead to improved power for detection of genetic effects. Simulation studies (12,28,29) suggest that this improvement in power may be relatively modest. Nevertheless, in analysis of real data for type 1 diabetes (12,28), type 2 diabetes (30) and inflammatory bowel disease (31), increased evidence for linkage at one locus was seen when the interaction with another locus was taken into account.

Methods for the detection of epistasis vary according to whether one is performing association or linkage analysis, and according to whether one is dealing with a quantitative or a qualitative (in particular a dichotomous) trait. For genetic association studies, standard methods for epidemiological studies may be employed, with genotypes at the various loci considered as risk factors for disease. Given case-control data genotyped at two diallelic candidate loci, for instance, we may fit the model

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + i_{aa} x_1x_2 + i_{ad} x_1z_2 + i_{da} z_1x_2 + i_{dd} z_1z_2$$

(where the genotype variables and coefficients are as defined previously) and compare it with the null model

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2$$

using standard statistical software packages for logistic regression. This provides an overall 4 degree-of-freedom (df) test for interaction, but the interaction terms could each be tested individually on 1 df by removal from the first model, if required. Note that this procedure implicitly assumes that the log odds scale is the scale of interest: to fit a null model that is additive on the penetrance scale would require use of more specialist statistical software or writing a program to maximize the likelihood directly (27). Quantitative traits can be analysed

in a similar way by use of standard multiple linear (as opposed to logistic) regression:

$$y = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + i_{aa} x_1x_2 + i_{ad} x_1z_2 + i_{da} z_1x_2 + i_{dd} z_1z_2,$$

where the quantitative phenotype y is the outcome of interest, assumed to be distributed normally given genotype.

Note that these regression procedures are actually designed for testing epistasis between loci that have been genotyped. If it is believed that these loci are not themselves the etiological variants but rather are in linkage disequilibrium (LD) with the true disease-causing variants, then epistasis between the surrogate genotyped loci is likely to be diluted compared with epistasis between the true variants, although the extent to which this occurs will depend on the magnitude of the LD.

A related method for analysis of (nuclear) family data involves a generalization of the genotype relative risk approach proposed by Schaid (32). As originally proposed, given genotype data on a trio consisting of an affected individual plus their parents, this method generates three 'pseudocontrols' consisting of the three alternative genotypes that could have been transmitted to the case, and then analyses them in a matched case-control design. Conditional logistic regression is used to fit models for the genotype relative risks. This method can be extended to fit models for genotype relative risks at two unlinked loci by generating not three but fifteen matched pseudocontrols for each case, where the genotype at the two loci for each pseudocontrol consists of one of the two-locus genotypes that could have been, but was not, transmitted to the case. Two-locus models for the genotype relative risks at the two loci are fitted using conditional logistic regression. Standard statistical software can be used to fit models that involve departure from multiplicativity in the penetrances (and hence in the genotype relative risks); more specialist software or user programming will be required for detecting epistasis defined as departure from additivity in the penetrances. A simpler approach that allows detection of epistatic (but not main) effects is to perform the original 'three-pseudocontrol' analysis (32) at one locus while considering genotype at the other locus to be a covariate that interacts with the first locus. Alternatively, one could use a 'case-only design' in which the distribution of genotypes in affected individuals only is used (33). This requires that genotypes at the loci under consideration be independent of each other in the general population, and, unlike the 'case-pseudocontrol' methods, is not robust to confounding causes of association such as population stratification. A variety of related approaches that focus on the issue of association testing but can be used to detect or allow for epistasis in family-based analysis of quantitative traits have also been proposed (34-37).

Epistasis is relatively easily incorporated into standard non-parametric (model-free) methods of linkage analysis for quantitative traits. One popular method is the variance components method, in which the phenotypic covariance between relatives is modelled in terms of variance component parameters and underlying identity-by-descent (IBD) sharing probabilities at one or more genetic loci, assuming underlying multivariate normality of the trait within pedigrees. Models that include epistatic (in the sense of departure from additive)

components of variance may be fitted and compared with models that do not contain these components using maximum-likelihood methods implemented in such programs as SOLAR (38). Another popular method of linkage analysis for quantitative traits is the Haseman–Elston method (39) and extensions (40). In this method, the squared difference or product of the phenotype values for a pair of relatives is modelled in a regression framework as a function of the underlying IBD sharing probabilities. To include epistatic interactions between loci, all that is required is to include products of the IBD sharing probabilities at different loci as predictors in the regression equation (40).

For dichotomous traits, model-free methods of linkage analysis typically focus on calculating the likelihood of the observed IBD sharing among pairs of affected relatives. Likelihoods for the IBD sharing at two or more loci can be calculated (12,28,41) under restricted models such as additive, multiplicative or heterogeneity models for the penetrances, and compared with general unrestricted models for the IBD sharing. In this way, tests for epistasis (defined as departure from any of these restricted models) can be performed for either linked (28,42) or unlinked loci. Another approach to the detection of loci that act epistatically is to consider the correlation between IBD sharing or linkage statistics at unlinked loci (30). A correlation of zero is expected under a multiplicative penetrance model, negative correlations provide evidence for a genetic heterogeneity model, and positive correlations suggest the presence of positive epistasis in the sense of departure from a multiplicative model. Tests of linkage that allow for epistasis between the test locus and a previously determined or candidate locus may also be performed (28,30). For full details of all the methods mentioned here, and software availability, the reader is directed towards the relevant references.

CONCLUSIONS

Methods of analysis that allow for or exploit the phenomenon of epistasis are clearly of growing importance in the genetic dissection of complex disease. A variety of methods exist to detect or control for the presence of epistasis. By allowing for epistatic interactions between potential disease loci, we may succeed in identifying genetic variants that might otherwise have remained undetected. In addition, identification of the most parsimonious statistical model for the joint effects of several loci, including interactions, can be useful for prediction of phenotype and for targeting of interventions. Nevertheless, we have seen that direct biological inference from the results of statistical tests is very difficult. The degree to which statistical modelling can elucidate the underlying biological mechanisms is likely to be limited, and may require prior knowledge of the underlying aetiology. The question of true biological interaction remains of paramount interest, but may ultimately be better answered via molecular, rather than statistical, investigation.

ACKNOWLEDGEMENTS

The author would like to express thanks to colleagues at Cambridge and Case Western Reserve Universities for a

number of useful discussions concerning the origins and definitions of epistasis. The author is supported by a Wellcome Trust Research Career Development Award, funded jointly by the Wellcome Trust and the Juvenile Diabetes Research Foundation.

REFERENCES

- Bateson, W. (1909) *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge.
- Greenland, S. and Rothman, K.J. (1998) Concepts of interaction. In Winters, R. and O' Connor E. (eds), *Modern Epidemiology*, 2nd edn. Lippincott-Raven, Philadelphia, pp. 329–342.
- Siemiatycki, J. and Thomas, D.C. (1981) Biological models and statistical interactions: an example from multistage carcinogenesis. *Int. J. Epidemiol.*, **10**, 383–387.
- Neuman, R.J. and Rice, J.P. (1992) Two-locus models of disease. *Genet. Epidemiol.*, **9**, 347–365.
- Risch, N. (1990) Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.*, **46**, 222–228.
- Schork, N.J., Boehnke, M., Terwilliger, J.D. and Ott, J. (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am. J. Hum. Genet.*, **53**, 1127–1136.
- Fisher, R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin.*, **52**, 399–433.
- 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.
- Frankel, W.N. and Schork, N.J. (1996) Who's afraid of epistasis? *Nat. Genet.*, **14**, 371–373.
- Cockerham, C.C. and Zeng, Z.-B. (1996) Design III with marker loci. *Genetics*, **143**, 1437–1456.
- Hodge, S. (1981) Some epistatic two-locus models of disease. I. Relative risks and identity by descent distributions in affected sib pairs. *Am. J. Hum. Genet.*, **33**, 382–395.
- Cordell, H.J., Todd, J.A., Bennett, S.T., Kawagushi, Y. and Farrall, M. (1995) 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.
- Fisher, R.A. (1930) *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford, p.111.
- Pearson, K. (1900) Mathematical contributions to the theory of evolution VIII: On the inheritance of characters not capable of exact quantitative measurement. *Phil. Trans. R. Soc. Lond. A*, **195**, 79–150.
- Wright, S. (1934) An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics*, **19**, 506–536.
- Wright, S. (1934) The results of crosses between inbred strains of guinea pigs, differing in number of digits. *Genetics*, **19**, 537–551.
- Risch, N., Ghosh, S. and Todd, J.A. (1993) Statistical evaluation of multiple locus linkage data in experimental species and relevance to human studies: application to murine and human *IDDM*. *Am. J. Hum. Genet.*, **53**, 702–714.
- Tiwari H. and Elston R.C. (1998) Restrictions on components of variance for epistatic models. *Theor. Popul. Biol.*, **54**, 161–174.
- Dizier, M.H. and Clerget-Darpoux, F. (1986) Two-disease locus model: sib pair method using information on both HLA and Gm. *Genet. Epidemiol.*, **3**, 343–356.
- Morris, A. and Whittaker, J. (1999) Generalization of the extended transmission disequilibrium test to two unlinked disease loci. *Genet. Epidemiol.*, **17**(Suppl 1), S661–S666.
- Wilson, S.R. (2001) Epistasis and its possible effects on transmission disequilibrium tests. *Ann. Hum. Genet.*, **62**, 565–575.
- Cheverud, J.M. and Routman E.J. (1995) Epistasis and its contribution to genetic variance components. *Genetics*, **139**, 1455–1461.
- Culverhouse R., Suarez, B., Lin, T. and Reich, T. (2002) A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.*, **70**, 461–471.
- Witte, J. (1998) Gene–environment interaction. In Armitage, P. and Colton, T. (eds), *Encyclopedia of Biostatistics*. Wiley, Chichester, pp. 1613–1614.

25. Rothman, K.J., Greenland, S. and Walker, A.M. (1980) Concepts of interaction. *Am. J. Epidemiol.*, **112**, 467–470.
26. Thompson, W.D. (1991) Effect modification and the limits of biological inference from epidemiologic data. *J. Clin. Epidemiol.*, **44**, 221–232.
27. Cordell, H.J., Todd, J.A., Hill, N.J., Lord, C.J., Lyons, P.A., Peterson, L.B., Wicker, L.S. and Clayton, D.G. (2001) Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics*, **158**, 357–367.
28. Cordell, H.J., Wedig, G.C., Jacobs, K.B. and Elston, R.C. (2000) Multilocus linkage tests based on affected relative pairs. *Am. J. Hum. Genet.*, **66**, 1273–1286.
29. Leal, S.M. and Ott, J. (2000) Effects of stratification in the analysis of affected-sib-pair data: benefits and costs. *Am. J. Hum. Genet.*, **66**, 567–575.
30. Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I. and Kong, A. (1999) Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nat. Genet.*, **21**, 213–215.
31. Cho, J.H., Nicolae, D.L., Gold, L.H., Fields, C.T., LaBuda, M.C., Rohal, P.M., Pickles, M.R., Qin, L., Fu, Y., Mann, J.S. *et al.* (1998) Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1. *Proc. Natl Acad. Sci. USA*, **95**, 7502–7507.
32. Schaid, D.J. (1996) General score tests for associations of genetic markers with disease using cases and their parents. *Genet. Epidemiol.*, **13**, 423–449.
33. Umbach, D.M. and Weinberg, C.R. (1997) Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statist. Med.*, **16**, 1731–1743.
34. Rabinowitz, D. (1997) A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.*, **47**, 342–350.
35. George, V., Tiwari, H., Zhu, X. and Elston, R.C. (1999) A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am. J. Hum. Genet.*, **65**, 236–245.
36. Lunetta, K.L., Faraone, S.V., Biederman, J. and Laird, N. (2000) Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am. J. Hum. Genet.*, **66**, 605–614.
37. Liu, Y., Tritchler, D. and Bull, S.B. (2002) A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. *Genet. Epidemiol.*, **22**, 26–40.
38. Almasy, L. and Blangero, J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, **62**, 1198–1211.
39. Haseman, J.K. and Elston, R.C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, **2**, 3–19.
40. Elston, R.C., Buxbaum, S., Jacobs, K.B. and Olson, J.M. (2000) Haseman and Elston revisited. *Genet. Epidemiol.*, **19**, 1–17.
41. Olson, J.M. (1999) A general conditional-logistic model for affected-relative-pair linkage studies. *Am. J. Hum. Genet.*, **65**, 1760–1769.
42. Farrall, M. (1997) Affected sibpair tests for multiple linked susceptibility genes. *Genet. Epidemiol.*, **14**, 103–115.

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 effects¹⁶, 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 disease⁷. 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 example¹. 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 studies²⁰ 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 term¹⁶ (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 test³² 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 regression²⁶, 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 regression²⁶ 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 models²⁸⁻³⁰ 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 effects⁶, 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.

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).

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 PLINK¹² 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 bias³⁵.

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 Ott³⁷. The ‘-fast-epistasis’ option in PLINK¹² 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 equivalent⁴⁴.

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 approach⁴⁶ 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 test¹⁶) corresponds to comparing the fit to the observed data

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 approaches⁵². 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 Clayton⁵³ 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 development¹⁴ 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 Clayton⁵³ 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 test⁵² (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 studies¹. If all the tests are independent, a Bonferroni correction is appropriate⁵²; 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 permutation⁵³ 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 disease¹. In practice, the Q-Q plot¹ 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 PLINK¹² 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 effects⁵⁸) 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 $p_{\text{corrected}} = 1 - (1 - p_{\text{uncorrected}})^n$, in which n is the number of hypotheses tested. This formula assumes that the hypotheses are all independent, and simplifies to $p_{\text{corrected}} = np_{\text{uncorrected}}$ when $np_{\text{uncorrected}} \ll 1$.

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 WTCCC¹. 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 PLINK¹² 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-validation⁶² 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 unequal⁶³. They also do not always deal well with correlated predictors that show collinearity. 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 regression⁶⁸ 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 missed⁷⁴.

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 approach⁷⁵ (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.

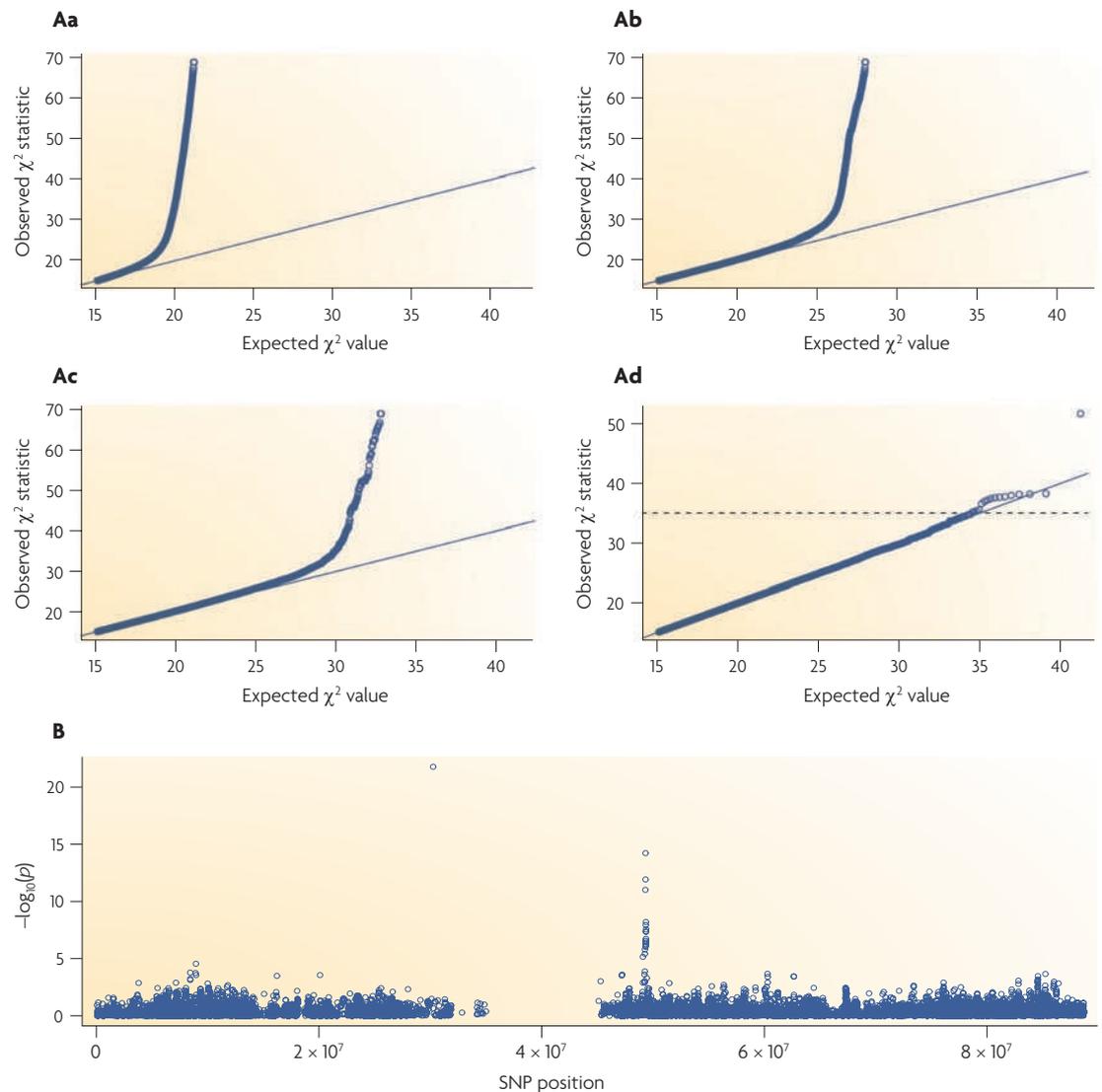
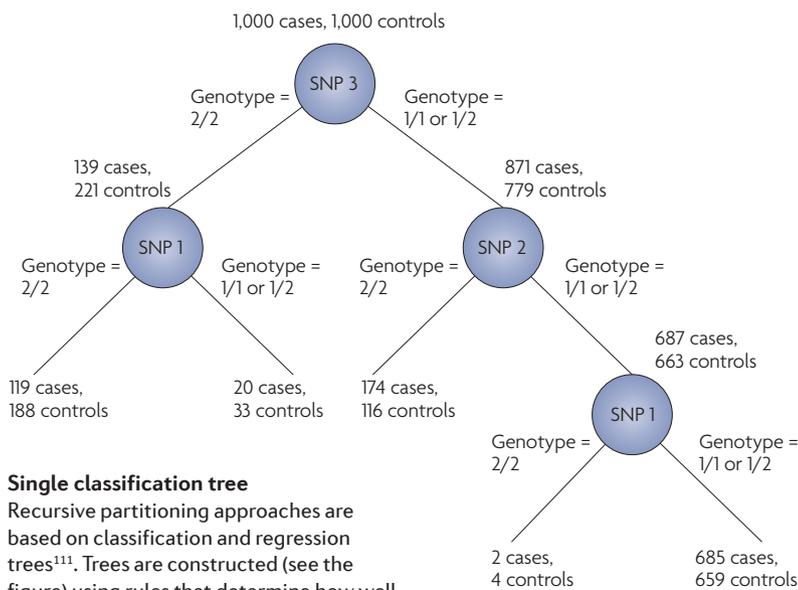


Figure 1 | Semi-exhaustive search of pairwise interactions between 89,294 SNPs. I used the ‘--fast-epistasis’ and ‘--case-only’ options in PLINK to analyse the Wellcome Trust Case Control Consortium (WTCCC) Crohn’s disease and control samples. I used the same quality control procedures as the WTCCC to 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 WTCCC¹. **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 trees¹¹¹. 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 rules¹¹¹ 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 method⁷⁵. 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 package⁷⁸ 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 programming⁸¹, 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 cancer⁸, type 2 diabetes⁸⁴, rheumatoid arthritis⁸⁵ and coronary artery disease⁸⁶, 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 *m*-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 step⁴⁰. 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,

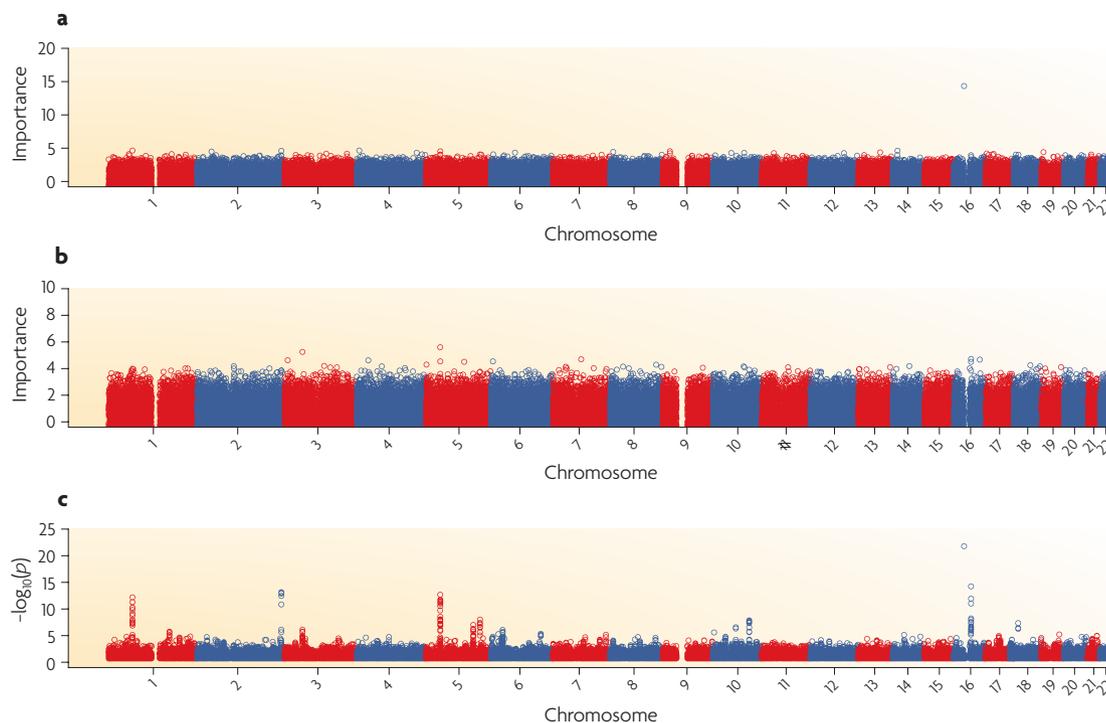


Figure 2 | Random Jungle analysis of 89,294 SNPs. I used the software package Random Jungle⁷⁸ to perform a random forests analysis of the 89,294 SNPs that passed a single-locus p value threshold of 0.2 in the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control data. As Random Jungle, 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 proposed⁴⁰ is ReliefF⁸⁹ 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 variable⁹⁰. 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 minutes⁹⁰.

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 phenotype⁷⁴. 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 methods⁷⁴.

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 techniques⁹² 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 algorithm⁴⁰ 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 $1 - \text{prediction error}$) 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.

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 most⁵⁶ 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 proposed⁹⁴.

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 Mapping¹³, 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 controls¹³) 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 markers¹³. 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, Phillips²⁰ 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 Carlborg⁹⁸ 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 models¹⁹ (Supplementary information S2 (box)). The NOIA model allows the main effects to be defined with

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.

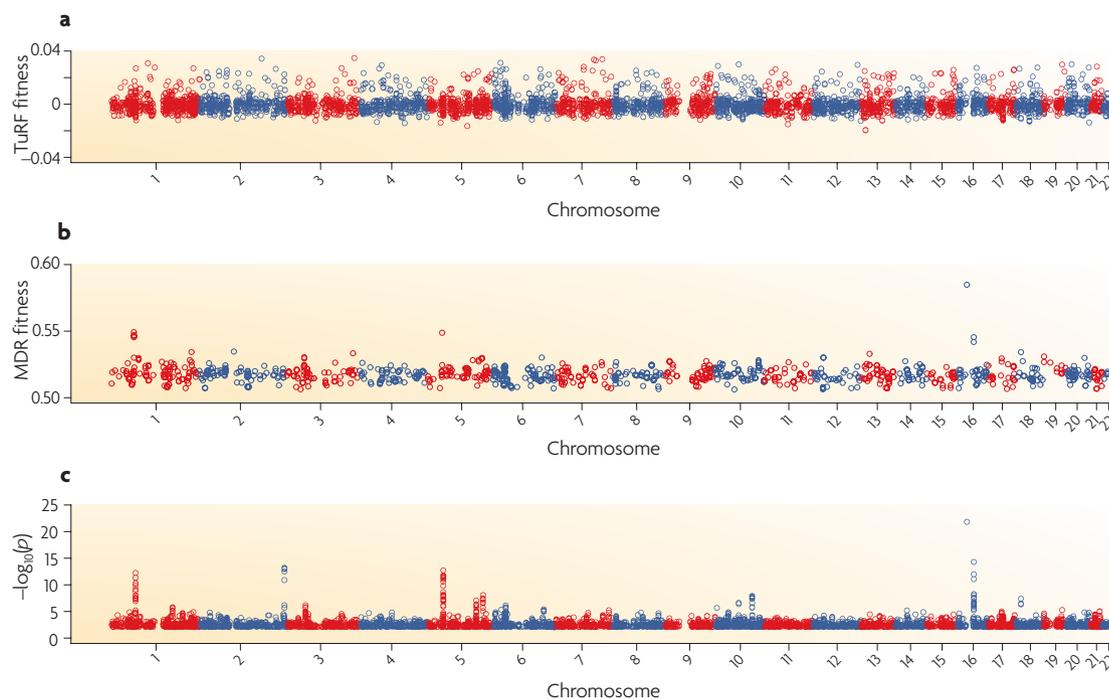


Figure 3 | Multifactor Dimensionality Reduction (MDR) and Tuned ReliefF (TuRF) analysis of 6,113 SNPs. I used the Java implementation of MDR to analyse 6,113 SNPs that passed a single-locus p value threshold of 0.01 in the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control data, 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 studies¹⁻⁵ has shown that single-locus association analysis in sufficiently large sample collections can reliably detect modest genetic effects that are robustly replicated^{105,106}.

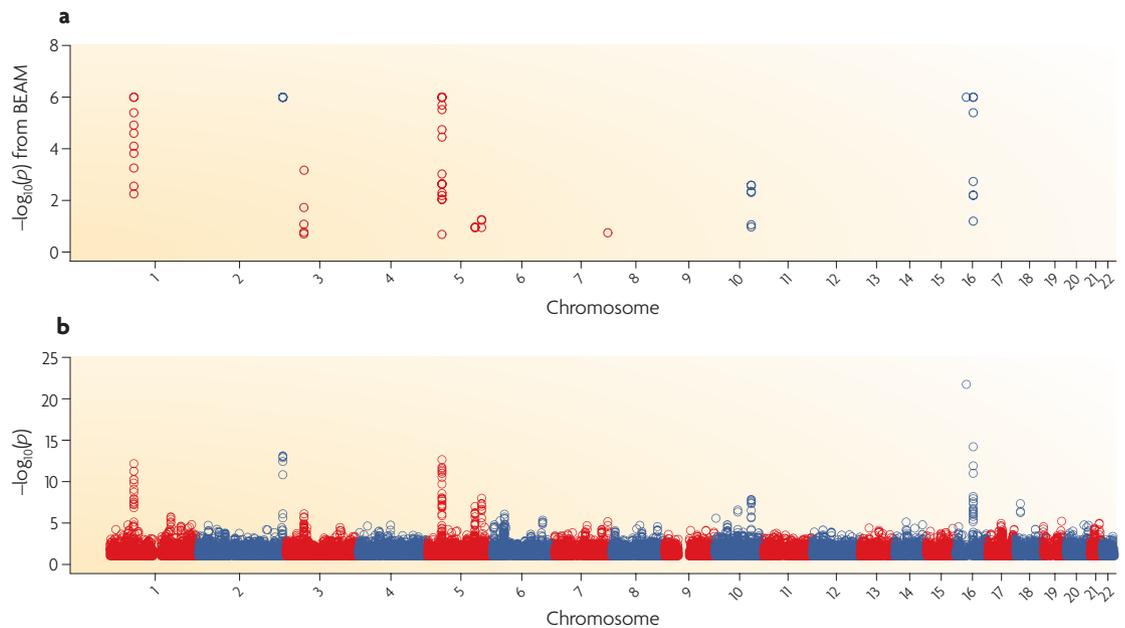


Figure 4 | Bayesian Epistasis Association Mapping (BEAM) analysis of 47,727 SNPs. I used BEAM to analyse a set of 47,724 SNPs that passed a single-locus p value threshold of 0.1 in the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control samples. Analysis of the 47,724 SNPs took 8 days (with some modification to the default settings, most notably imposing a maximum of 5×10^{-7} Markov chain Monte Carlo (MCMC) iterations¹³ 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 limited¹⁰², 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 PLINK¹² and a random forests analysis implemented in Random Jungle⁷⁸ were the most computationally feasible of the methods examined. Bayesian Epistasis Association Mapping implemented in BEAM¹³ 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. Curtis¹¹⁰ 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 5,000 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, O., 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. Kottli, 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 × 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., Schauer, 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**, 25–69 (2005).
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., Friedman, 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