MATHEMATICAL MULTI-LOCUS APPROACHES TO LOCALIZING COMPLEX HUMAN TRAIT GENES

Josephine Hoh and Jurg Ott

Statistical analysis methods for gene mapping originated in counting recombinant and non-recombinant offspring, but have now progressed to sophisticated approaches for the mapping of complex trait genes. Here, we outline new statistical methods that capture the simultaneous effects of multiple gene loci and thereby achieve a more global view of gene action and interaction than is possible by traditional gene-by-gene analysis. We aim to show that the work of statisticians goes far beyond the running of computer programs.

RECOMBINATION FRACTION
The proportion of offspring that receives a recombinant haplotype from a parent, or the probability that recombination occurs between two loci.

BACKCROSS
Originally, backcross referred to the mating of an offspring with one of its parents, in which the offspring is heterozygous, with the parent being homozygous for one of the alleles in the offspring's genotype. Nowadays, backcross simply refers to a mating between individuals with those two genotypes.

Laboratory of Statistical Genetics, Rockefeller University, New York 10021, USA.
Correspondence to J.O.
e-mail: ott@rockefeller.edu
doi:10.1038/nrg1155

> "It was the math that gave us the edge, not the machines".
>
> *Craig Venter* (REF. 1)

Mendelian disease traits are generally rare and tend to occur in populations at frequencies of less than 0.05%. Common heritable disease (or susceptibility) traits, such as diabetes, schizophrenia and obesity, represent a more severe burden on human populations, with frequencies of 1% or more. Although these traits cluster among relatives, they do not show Mendelian modes of transmission and are generally thought to be under the influence of multiple, and possibly interacting, genes. Localizing such genes is of high importance for pharmaceutical companies because the genes that contribute to these traits could identify drug targets that are difficult to find otherwise. Adverse drug events also tend to have a strong genetic basis. For these reasons, statistical methodological research has shifted from finding genes that underlie Mendelian disorders to those that contribute to complex traits.

In many experimental organisms, finding genetic linkage between two neighbouring loci amounts to counting a sufficient number of suitable offspring. For example, the so-called *F2* generation in a double-BACKCROSS allows researchers to count recombinant and non-recombinant animals. Two loci are said to be genetically linked if the proportion of recombinant animals (the RECOMBINATION FRACTION) is significantly less than 50% — a small recombination fraction is indicative of a short physical distance between the two loci investigated. One of the loci might be a Mendelian disease locus with unknown position in the genome, whereas the other locus is a genetic marker with known position. Therefore, significant linkage allows the approximate localization of the disease locus in the vicinity of the marker locus. If the result is inconclusive, the researcher will simply perform more crosses, so there is no need for sophisticated statistics in this case.

In human genetics, however, investigators are plagued with factors that cannot be controlled by experimental design, such as missing observations, small family size, and various other problems that render data analysis more complicated than for experimental animals. However, the principle is the same — a small estimated recombination fraction between disease and marker loci provides an approximate localization of the disease gene. In the case of complex traits, difficulties are greatly compounded by the fact that an individual can be affected because of the effects of susceptibility genes that lie on different chromosomes. Disease might occur only if a particular combination (pattern) of genotypes is present at different susceptibility loci, and not as a result of a single disease gene alone. So, each single susceptibility gene will have only a small effect and cannot easily be detected by methods that search for one gene at a time.

Box 1 | **Likelihood and lod score**

For family data with partially missing observations, it is not possible or is at least very inefficient to estimate recombination fractions by simple inspection of the data. However, it is still possible to compute the likelihood — that is, the probability of the occurrence (or the plausibility) — of the observed data when specific values of parameters such as the recombination fraction are assumed. Such a parameter could be estimated by trial and error — that is, assume many different parameter values and inspect the likelihoods that result from them. The parameter value that makes the data most plausible (with the largest likelihood) is taken to be the best estimate. The associated likelihood ratio (the largest likelihood divided by the likelihood when the recombination fraction is 50%) is known as the odds for linkage and its logarithm is known as the (maximum) lod score.

Such complexity might seem daunting, but statisticians have been developing appropriate analysis methods that can capture contributions from multiple susceptibility loci and provide evidence for the localization of disease genes on human chromosomes. Such localizations are only approximate and accurate to within several megabases, but they represent a starting point and guide future molecular-genetic research.

Here, we discuss the specific properties and difficulties that statisticians face when dealing with complex human traits. We do not discuss haplotype analyses — that is, investigations of haplotype frequencies and their differences in cases and controls. These methods (including specialized approaches such as cladistic analysis[2]) deserve a separate treatise and represent important multi-locus alternatives to the approaches discussed below.

In general, the term 'multipoint analysis' refers to the joint analysis of multiple neighbouring marker loci, with the purpose of localizing one disease locus independently of other disease loci that might exist elsewhere in the genome. To avoid confusion with this classical technique, we use multi-locus methods to refer to approaches that are specifically designed to find multiple disease loci, possibly on different chromosomes. It is these methods that are reviewed here. Although analogous approaches in experimental organisms have recently been reviewed[3], our outline is the first review of the mathematical aspects of multi-locus approaches in human genetics.

LIKELIHOOD ANALYSES
A statistical method that calculates the probability of the observed data under varying hypotheses, to estimate model parameters that best explain the observed data and determine the relative strengths of alternative hypotheses.

LOD SCORE
The logarithm of the likelihood ratio (odds) for genetic linkage versus no linkage at a given value of the recombination fraction.

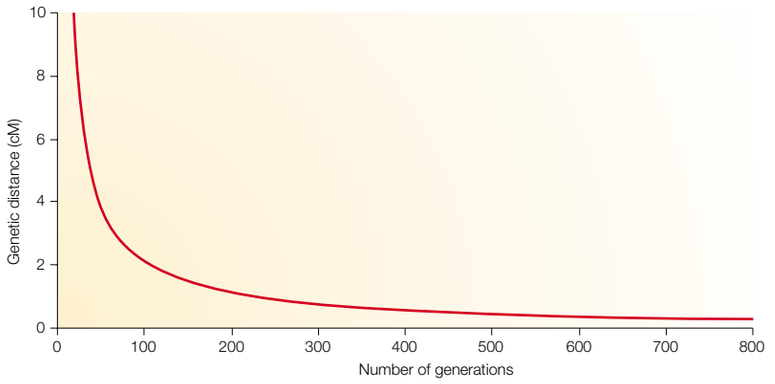

Figure 1 | **Expected decay of linkage disequilibrium around a recessive disease locus in successive generations after an initial mutation.** Predicted length in centimorgans (cMs) of the region around a disease locus in which no crossover is expected to occur within the given number of generations after the initial occurrence of a disease mutation. A chromosome length of 200 cM is assumed and the disease is rare and recessive. Reproduced with permission from REF. 14.

Historical background

In 1902, Garrod[4] interpreted the clustering of abnormalities in sibships as the likely result of Mendelian inheritance. After successful gene-mapping experiments in the fruitfly by Morgan and Sturtevant, early approaches to localizing a human-disease locus, based on its linkage to a genetic marker locus, were carried out by tedious LIKELIHOOD ANALYSES (BOX 1) of X-linked traits. The introduction of LOD SCORES and their tabulation for two-generation human families of various sizes and parental phenotypes[5] opened this field to a wide spectrum of researchers. The formulation of likelihoods for large pedigrees[6], and the development of computer algorithms for the automated calculation of likelihoods and lod scores[7], set the stage for worldwide efforts to localize genes that are responsible for Mendelian disorders. Among the first genes to be found using this approach were those that are mutated in familial hypercholesterolaemia[8,9], Huntington's disease[10] and cystic fibrosis[11]. Once the loci are localized, it still typically takes several years for the molecular characterization (identification) of the gene to be accomplished. Elucidating the mechanisms of disease, and perhaps the cures, generally still takes many years of additional research.

Association studies

Gene-mapping studies are based on the random occurrence of cross-overs in meiosis and are of two types: genetic linkage analysis and association analysis. Here, we focus on the essence of association analysis, as the relationship between these two techniques has recently been reviewed[12].

Association studies rely on the fact that alleles at loci that surround a disease locus tend to segregate together. In the absence of crossing-over, the disease chromosome (the chromosome on which the mutated disease locus lies), and all the alleles at other loci that happen to be on that chromosome, would be transmitted as a block (known as a haplotype) to the descendants of the given individual. However, owing to the occurrence of crossing-over, the region around the mutated allele that will be transmitted as a block to the next generation tends to shrink in successive generations[13]. FIGURE 1 shows the length of the region around the disease locus where no crossing-over is expected to occur (assuming that cross-overs occur independently of each other) for a rare recessive disease that is caused by a mutation at a single locus[14]. For example, for a disease-causing mutation that occurs in the middle of a chromosome that is 200 cM long (~200 Mb), the region around the disease locus that will be unaffected by crossing-over would be

reduced to 1 cM after 200 generations. Each locus within this region that receives a copy of the ancestral mutation will tend to inherit the same set of alleles (haplotype), coupled with the ancestral mutation. Individuals who do not inherit the disease mutation are expected to show a random assortment of alleles at the various loci throughout the given chromosome. If a genetic trait is at least partly due to the particular mutation considered here, we expect that individuals affected with the trait tend to share a different set of alleles at loci around the mutated locus than do unaffected individuals. This is referred to as linkage disequilibrium (LD). If a genetic marker is in LD with a disease susceptibility locus, it would be expected that genotype and allele frequencies at the marker loci will differ between case and control individuals. Investigations of such differences are known as association studies.

Recurring disease and marker mutations will impact negatively on the expected extent of LD between marker and disease loci. By contrast to microsatellite markers, a particular set of markers, the single-nucleotide polymorphisms (SNPs), is known to mutate very rarely, which renders them extremely useful for association studies. Importantly, SNPs are very abundant — several million of them are known, whereas the number of useful microsatellites is in the order of 1,000. In many studies, the effects of genetic linkage can be detected over distances of 10 to 20 Mb. On the other hand, investigations of association for pairs of SNPs have shown that moderately strong LD in a United States population of north-European descent typically extends only 60 kb from a SNP[15]. So, it might be expected that the current range of LD around a disease locus is of a similar magnitude, which, given the size of the genome, requires 10,000s of SNPs for genome-wide association studies[16].

Multi-locus methods

As there are many susceptibility genes, each gene by itself might have a rather small effect. Therefore, investigating associations between marker genotypes and disease phenotypes for only one marker at a time without considering other (unlinked) markers will probably capture only a small proportion of the total combined effect of all disease genes. It is for this reason that statisticians are developing analysis methods that allow the joint analysis of multiple SNPs in their associations to multiple disease loci.

The number of genetic marker loci (input variables) is potentially very large and, when combined with the generally much smaller number of observations, creates a statistical problem that has been referred to as the 'curse of dimensionality'[17], as it precludes a classical joint analysis of all variables. In genomic screens, this problem has traditionally been circumvented by the analysis of one input variable at a time. To do justice to multigenic complex traits, it has been proposed that a two-step approach should be used[18]. First, a small number of important or influential markers is selected. Second, based on the selected subset, interactions between markers and dependent variables are modelled. Note that the second step can also be accomplished using classical statistical analysis methods. Several approaches discussed below focus on step one, whereas others combine both steps one and two.

Early approaches. Attempts to analyse more than one disease locus at a time go back at least 70 years. In the first instance, analysis considered two-locus (also known as digenic) inheritance[19]. For example, Hogben[20] postulated that pairs of unlinked loci caused disease and, for given parental phenotypes, analytically derived the expected proportion of affected offspring, which he then compared with the corresponding proportion observed in specific diseases. Although these efforts were more directed towards elucidating the mode of inheritance than gene mapping, they were among the first to consider multiple disease loci simultaneously.

More recently, MacLean *et al.*[21] argued that two loci, the joint action of which confers susceptibility to disease, should be expected to show correlated lod scores, which is not expected of loci that are not associated with disease. So, these researchers proposed an analysis procedure that incorporated the correlation between maximum lod scores obtained at different genomic regions; however, the statistical properties of this procedure remain unknown. A subsequent refinement of this approach has successfully shown that two loci interact to increase susceptibility to diabetes[22].

These ideas led to additional formal developments of analysis methods that involved more than one locus at a time. For example, approaches for the simultaneous linkage analysis of two susceptibility loci were proposed and have been shown to be generally more powerful than the analysis of one locus at a time[23–25]. Whether this also holds on a genome-wide basis was investigated in a careful theoretical analysis by Dupuis and colleagues[26]. They assumed the existence of two unlinked disease loci with different modes of inheritance and evaluated three approaches to localizing one or both of the two susceptibility loci by linkage analysis. First, they tested single-locus search, which represents the traditional approach of analysing one marker at a time and looking for the most significant results. Second, they tested simultaneous search, which looks at all possible pairs of markers and picks those with the strongest effects on linkage. Finally, they tested conditional search, which is based on a clearly significant linkage result and, given that result, looks at all other loci one by one. The conditional search differs from the single-locus search in that, in the former, the finding of a new locus given an established locus can exploit interactions between these two susceptibility loci[27].

An important conclusion from the comparison of these three genome-wide search strategies was that the power of each strategy depends on how susceptibility loci interact to cause disease. To successfully localize at least one disease locus over a broad range of conditions, neither a single-locus search nor a simultaneous search seemed to be strongly preferred. Investigating all possible pairs of loci leads to a greatly increased number of statistical tests, which is referred to as a multiple-testing

Box 2 | **Multiple testing**

When markers are tested for linkage or association over the whole genome, each test results in a locus-specific, or point-wise, *p*-value, which is the probability that the test statistic exceeds a given threshold by chance. Obviously, this probability should be kept to a low level by setting an appropriate criterion for significance. In a genomic scan, the probability that one or more of the tests exceed the chosen threshold would be useful to know. This probability is the genome-wide (or experiment-wise) significance level, which for a given threshold is much higher than the pointwise *p*-value. Formulas have been developed that provide a sufficiently low value of the point-wise *p*-value for the genome-wide significance level not to exceed 0.05 (REF. 56)**. However, it is important to recognize that these formulas do not apply, for example, to searches for pairs of susceptibility loci because the number of pairs is much higher than the number of loci. In this case, for *n* marker loci, $n(n-1)/2$ tests rather than only *n* tests are done, which exacerbates the multiple testing problem. Possible solutions are computer-based test procedures, such as permutation tests that free statisticians from the need to focus on statistical measures, the theoretical properties of which can be analysed mathematically[33]. However, this is generally possible only for simple statistics and independent data. For example, in a case–control data set with large numbers of marker genotypes, randomly permuting the labels case and control breaks any association between genotypes and disease status. For each of a sufficiently large number (several hundred) of such permutation samples, an analysis is carried out as was done on the original data. The proportion of permutation samples in which any marker statistic exceeds a given threshold is then an estimate for the experiment-wise significance level that is associated with that threshold.**

problem because each statistical test carries with it the possibility of a false-positive result (BOX 2). Conditional search seems particularly useful for heterogeneous traits (when disease is due to one or the other of the two loci), but has comparatively little value when different loci interact epistatically (that is, when alleles from both loci are required for disease). These conclusions were based on the assumptions of a particular two-locus inheritance model. As is shown below, consideration of alternative inheritance models might lead to different results.

In case–control studies with multiple-marker loci that potentially have different genotype frequencies in case and control individuals, it is tempting to use the LOGISTIC REGRESSION MODEL — that is, to form a weighted sum of genotype codes, where, for example, the three genotypes at a SNP are assigned codes of 0, 1 or 2. Weights are determined in such a way that the resulting sum discriminates in the best possible way between case and control individuals by showing large values for the former and low values for the latter. Such studies have been undertaken successfully for many years. For example, in a study of the effects of genotypes at a number of genes on the occurrence of retinopathy among diabetics, *HLA-DR4* was shown to confer a significantly increased risk for this eye disease[28]. In a comparison of adopted and non-adopted children at varying risk of speech disorders, logistic-regression analysis showed that parental status with regard to speech disorder was the best predictor of whether offspring were affected, whereas other potential risk factors, such as the child's IQ, were not significantly associated with speech phenotype[29]. However, the logistic regression approach suffers from several shortcomings. For example, if the number of marker loci is larger than the number of observations, this method fails completely. This obstacle might be overcome with STEPWISE REGRESSION analysis, in which markers are added to the regression equation one after another by some suitable criteria, but statistical analysis shows that the usual stepwise-model selection methods are suboptimal[30]. Also, the regression model imposes fixed relationships between marker genotypes and phenotype (case versus control), which might not be realistic.

Early approaches also included the use of neural networks, which can be thought of as generalizations of logistic regression to nonlinear relationships[30,31]. The idea was that neuronal networks would elegantly circumvent the curse of dimensionality problem that is mentioned above. However, with the growth of genetic markers, particularly SNPs (up to several million are now available), a type of data preprocessing (two-step approach) seems to be more promising. Consequently, neuronal-network-based analysis is used with a decreasing frequency.

Sums of single-marker statistics. One approach to finding a set of important SNP markers among a potentially very large number of such loci is parameter-free and works as follows[18]. First, conventional single-locus statistics for each SNP are calculated — for example, the chi-square statistic is computed from a contingency table with two rows that correspond to cases and controls, and three columns that correspond to three SNP genotypes. Large chi-square values are indicative of association between a SNP and the trait that is being studied.

LOGISTIC REGRESSION MODEL
A statistical model for the dependency of a binomial (two-class) phenotype on a number of risk factors. The probability, *p*, for one of the two phenotype states is expressed in the form of its logit, log(p/(1 – p)), which is assumed to be predicted by the linear combination (weighted sum) of the risk factors.

STEPWISE REGRESSION
The step-by-step build-up of a regression model, which represents a dependent variable as a weighted sum (linear combination) of independent (risk) variables.

Box 3 | **Bootstrapping**

To take a bootstrap sample of a number *n* of individuals means to randomly pick *n* copies of individuals. A bootstrap sample might contain some individuals more than once and others might be missing from it, but a bootstrap sample has all the essential statistical properties of the original data set and could be regarded as a random copy of it[57]. Usually, large numbers of bootstrap samples are generated to investigate the statistical properties of some procedure.

Bootstrap samples might also be taken under some restrictive conditions. For example, if the data consist of case and control individuals with associated genotypes for a number of marker loci, independently sampling from phenotypes and genotypes generates bootstrap samples under absence of association between the trait phenotype and the genotypes of each individual. Such 'null' samples can be used to evaluate the significance level of a statistical test in much the same way as permutation samples (BOX 2)**. Indeed, bootstrapping and permutation sampling are both resampling methods, but they have different statistical properties.**

The combined effect of all markers is captured by the sum of all association statistics. The possibility of this sum being larger than expected simply by chance is determined by computing the associated *p*-value from a number of bootstrap samples (BOX 3) obtained under the null hypothesis of no association. As the sum over all markers presumably contains the effects of some disease genes but also a large amount of 'noise' from unassociated SNPs, the resulting *p*-value will generally be large (that is, not significant). The marker with the smallest TEST STATISTIC is then removed from the sum, which gives another sum with an associated *p*-value. This step is repeated for one marker at a time with decreasing *p*-values, and, as the smallest effects are being removed from the sum, will eventually contain only markers with large test statistics. The removal of markers stops once the SIGNIFICANCE LEVEL of the sum reaches a lower limit of 0.05, for example. Those markers that remain in the sum are considered to be pre-selected.

The next step is to make random copies of the original data set, with each copy consisting of a bootstrap sample obtained under association. The analysis that was carried out on the original data is then repeated in each of these replicate copies, each of which gives a set of pre-selected marker loci. The final step is then to determine for each marker in what proportion of replicates it was pre-selected. Markers that were pre-selected in more than 60% of replicates are deemed to be important for association (and are known as selected)[33] (FIG. 2).

This nested bootstrap approach was applied to a group of 779 heart disease patients who had undergone balloon ANGIOPLASTY, 342 of whom subsequently experienced CORONARY ARTERY RESTENOSIS (cases), whereas the remainder did not (controls)[34]. In the search for putative genetic factors that predispose patients to restenosis, genotypes for 94 SNPs that represented 62 candidate genes were determined. When applied to these data, the nested bootstrap procedure selected 11 out of the 94 SNP markers that represented the following 10 genes: *TNFR1*, *IL4Rα*, *TP53*, *CD14*, *APOA*, *CETP*, *TNFβ*, *CBS*, *NOS* and *MDM2*. Interestingly, the set of selected markers does not correspond exactly to the set of markers with the largest chi-square statistics. One marker (a SNP in the *TP53* gene) is not selected even though it has a larger chi-square value than some of the selected markers, presumably because its inclusion with other markers that are already in the sum was contributing less than its own chi-square value indicated. The conclusion from this analysis is that disease susceptibility loci are likely to exist in close proximity to the selected SNPs. Perhaps the 10 genes that contain these 11 SNPs are directly involved in the occurrence of restenosis and more biological work will be required to elucidate disease aetiology.

Although this approach is intuitively appealing, it does not work in a hypothesis testing framework — the researcher does not know whether results are statistically significant or not. The following modified method allows marker selection under rigid control of the experiment-wise significance level[35].

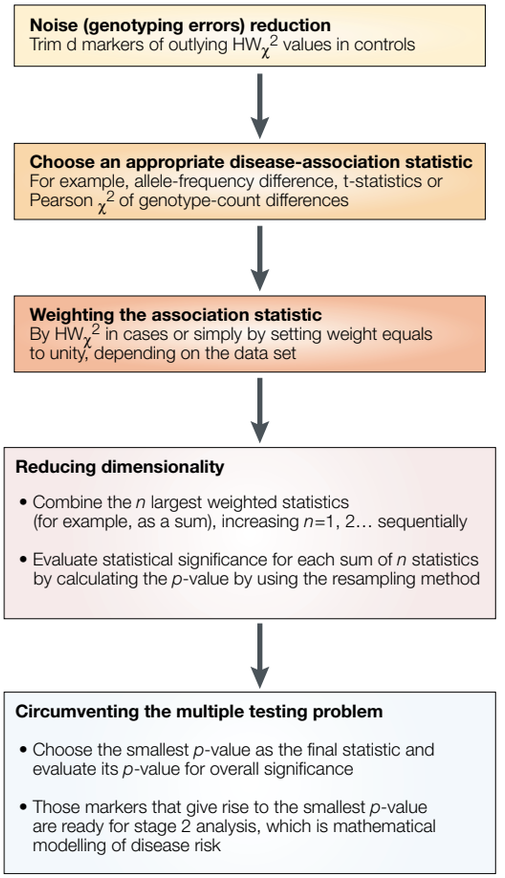

Figure 2 | **Flow chart for the set association procedure for combining association effects of multiple marker loci.** HWχ^2, Hardy–Weinberg chi squared; *p*, probability.

Consider again a single-locus test statistic for the *j*-th SNP marker. In this modified method, the single-locus test statistic is the product of two chi-squares, thereby improving the power of this statistical value. For example, this statistical value is the product of the two chi-squares — $t_i = \chi^2_{\text{Assoc}} \times \chi^2_{\text{HWD}}$ — where the first term is the chi-square that measures the difference in allele frequencies between cases and controls and the second term measures departure from the HARDY–WEINBERG EQUILIBRIUM, either in cases only, or in all individuals (depending on the data set). The basic concept is again to combine information for multiple SNPs by summing the corresponding association statistics, t_i, but in this case, sums are formed by adding important markers rather than by deleting unimportant ones. So, marker loci are first ordered according to their test statistic, $t_{(1)} \leq t_{(2)} \leq \ldots$. Then, sums are formed with increasing numbers of terms, $s_1 = t_{(1)}$, $s_2 = t_{(1)} + t_{(2)}$ and so on, up to a suitable maximum number — for example, $M = 20$, where M is a parameter of the procedure. To see whether a given sum is unusually large, its significance level is evaluated using a permutation (randomization) test. Consider the data matrix with rows that correspond to individuals and columns that correspond to variables, one of which indicates disease status

TEST STATISTIC
A statistic is any function of a random sample — in particular, of the observations in an experiment. A test statistic is a statistic that is used in a statistical test to discriminate between two competing hypotheses, the so-called null and alternative hypotheses.

SIGNIFICANCE LEVEL
The proportion of false-positive test results out of all false results — that is, results that are obtained when the effect investigated is known to be absent (see also false discovery rate).

ANGIOPLASTY
A medical procedure that is used to widen coronary arteries with a thin balloon because these blood vessels have become clogged.

CORONARY ARTERY RESTENOSIS
The re-occurrence of a narrowing or blockage of an artery at the site where angioplasty had previously been performed.

HARDY–WEINBERG EQUILIBRIUM
A state in which the proportions of genotypes present depends only on the frequencies of alleles in the genotypes.

(case or control). The latter variable is now replaced by a random permutation of it, which results in a 'permutation sample' that represents the data without association. A large number of such permutation samples are generated and the analysis procedure is repeated for each sample. The proportion of replication samples with a particular sum that exceeds the corresponding sum seen in the observed data approximates to the *p*-value for that sum. At this point, the multiple-testing problem (BOX 2) of testing a potentially very large number of SNPs has been reduced to testing *M* sums. The crucial final step is to define the smallest of the *M* significance levels as a single test statistic, which might occur for $m \leq M$ summed marker statistics. The significance level, p_{min}, that is associated with this test statistic is then evaluated through permutation tests, and represents a single overall measure of significance for the whole procedure. The *m* SNPs, the summed test statistics of which lead to the minimum *p*-value, are selected for further analysis. As this selection is done under the controlled false-positive rate, p_{min}, any additional analyses of the selected SNPs that are of interest can then be done without incurring a penalty for additional testing.

This approach was applied to the RESTENOSIS data set[34] that had previously been analysed by the bootstrap procedure. With an overall *p*-value of 0.04, it resulted in the selection of ten SNPs that represent the following nine genes: *TP53*, *CD14*, *SERPINE1*, *APOC3*, *ITGB2*, *CBS*, *NOS*, *TNFR1* and *MDM2*. Six of the nine genes are the same as those selected by the bootstrap procedure and so are highly likely to be associated with the trait. At this point, it is unclear which of the two statistical selection approaches is more reliable. The ultimate confirmation will be provided by the elucidation of the biological pathway that leads to the disease.

Joint analysis of multiple loci. Although the above methods are a clear improvement over traditional approaches because they combine information from multiple loci, a potential drawback is that they rely on single-locus effects, often also called main effects, in contrast to interaction effects. This shortcoming is avoided in another parameter-free approach, known as the combinatorial partitioning method (CPM)[36]. It focuses on quantitative phenotypes, such as lipid levels, and represents an extension to many loci of traditional genetic analysis of variation in trait levels. For a single bi-allelic marker locus, traditional analysis compares the variability in trait values both between the three genotypes and within the three genotypes. An excess of the former over the latter represents association between the marker and the trait. The CPM extends this concept to many marker loci (SNPs). The focus of CPM is to form subsets of SNPs that contain different numbers of loci. For example, for ten SNPs numbered 1 to 10, {2, 4, 5} represents one possible such subset (here, of size 3). For a particular subset of size *m*, consider the total number of possible *m*-locus genotypes. For instance, with $m = 3$, the total number of 3-locus genotypes is $3 \times 3 \times 3 = 27$. A genotypic partition is now defined as a set of *m*-locus genotypes and the mean trait value for those individuals with *m*-locus genotypes in this partition is recorded. The object of the CPM is to find those genotypic partitions within which a trait variability is much lower than between the partitions. The loci comprised in such a set of genotypic partitions are then viewed as influencing the quantitative trait and this effectively completes step 1 of the CPM.

In the second step, the selected set of genotypic partitions is validated by a well-known technique called cross-validation. A randomly chosen portion (for example, 90%) of the observations are used for the CPM, leading to a number of optimal genotypic partitions. In the remainder of the data, the proportion of total variability within the optimal genotypic partitions is computed. This process is repeated many times and leads to a cross-validated proportion of variability that can be explained by a set of genotypic partitions. Steps 1 and 2 provide a number of sets of genotypic partitions that influence variation in quantitative trait levels. Which of these sets is the best, or indeed whether it makes sense to select a best set, will depend on the particular criteria that an investigator applies for drawing inferences about genotype–phenotype relationships (step 3 of CPM).

The CPM was applied to triglyceride levels of 188 males and 18 SNPs in six coronary heart disease candidate gene regions[36]. Genotypic partitions of size 2 (for all possible pairs of SNPs) were formed. Of the total number of 3,235,338 possible genotypic partitions, a much smaller number of 7,710 was retained after elimination of the genotypic partitions with a small effect and those that contain fewer than five individuals. Results showed that many combinations of loci are involved in triglyceride variability and that the most predictive sets of loci show non-additivity — that is, some of their effects come about through locus interactions.

A modification or extension of the CPM is the multifactor-dimensionality reduction (MDR) method[37]. As described by the authors of this approach, "with MDR, multilocus genotypes are pooled into high-risk and low-risk groups, effectively reducing the genotype predictors from *n* dimensions to one dimension. The new, one-dimensional multilocus-genotype variable is evaluated for its ability to classify and predict disease status through cross-validation and permutation testing."[37] When applied to a case–control data set with breast cancer, this approach revealed significant high-order interactions between three oestrogen metabolism genes without significant main effects.

The simple structure of case–control data lends itself to the analysis by general statistical analysis methods. As an example, logistic regression with its advantages and restrictions was mentioned above. Several additional analysis methods have been proposed, such as the tree-based association analysis[38]. This is only one of a group of RECURSIVE PARTITIONING methods[39] that are reminiscent of CLUSTER ANALYSIS. It is probably too early to tell which of the increasing number of different approaches now being tried on case–control association data will be the most successful.

RESTENOSIS
A re-narrowing or blockage of an artery at a site where angioplasty was previously done.

RECURSIVE PARTITIONING
A process for identifying complex relationships in large sets by dividing them into a hierarchy of smaller and more homogeneous subgroups on the basis of the most statistically significant indicators.

CLUSTER ANALYSIS
A mathematical algorithm that organizes a set of items according to their similarity. For example, genes can be clustered according to their similarity in pattern of expression.

Table 1 | **An example of a disease purely caused through interaction between loci**

	Epistatic disease loci			Population frequency	Penetrance	Expected number	Expected number
	Locus A	Locus B	Locus C	P(g)	(f)	of cases	of controls
Genotype at given locus							
	1/1	2/2	1/1	0.0156	1	25	0
	2/2	1/1	2/2	0.0156	1	25	0
	1/2	1/2	1/2	0.1250	0.25	50	10
Other genotype				0.8438	0	0	90
Sum				1	–	100	100

Patterns of genotypes at three purely epistatic disease loci (alleles 1 and 2) and the corresponding population frequencies (P(g)), penetrances (f), and expected numbers of cases and controls that show a given genotype pattern.

Genotype patterns

The main drawback of conventional methods of gene mapping is that they rely on analysing one locus at a time. It is becoming increasingly clear that diseases most often arise as a result of rather complicated interactions between genes. A telling example is Hirschsprung disease, the non-Mendelian inheritance pattern of which seems to be due to three genes at different genomic locations[40]. Cases have been reported in which disease occurs not through any single gene, but only through interactions between multiple genes. For example, mice heterozygous for the insulin receptor or the insulin receptor substrate-1 have minor metabolic abnormalities, whereas compound heterozygous animals show marked insulin resistance[41]. Another case of interaction effects and no main effects has been described in a human family[42]: five family members with severe insulin resistance were doubly heterozygous at two unlinked loci, whereas no other family member showed this trait and these genotypes. Also, an *HLA* gene on chromosome 6 and a *KIR* gene on chromosome 19 seem to be involved in epistatic interaction: the combined presence of alleles *KIR-3DS1* and *HLA-BW4-80ILE* is associated with delayed progression to AIDS in HIV-positive individuals[43]. Additional examples of such digenic traits have recently been reviewed[44].

Traits that are due to specific genotype patterns at different loci require appropriate search methods, such as the recently developed approaches described in the previous section. Interestingly, in artificial intelligence (machine learning) and in operations research, pattern recognition (also known as data-mining) methods that were described 10–15 years ago are only now beginning to find their way into human genetics and might prove highly successful. They can only briefly be mentioned in this review. An intriguing approach was developed about 10 years ago[45] for the analysis of large databases of purchasing transactions (recorded using bar codes), in which one of the original aims was to elucidate customer preferences by detecting the patterns of articles purchased. This approach formalizes pattern recognition by defining pattern frequencies and relationships in the form of so-called association rules[45]. A specific implementation (known as the *apriori* algorithm[46]) allows for the rapid detection of patterns even in very large databases. It is freely available as a computer program (see *apriori* algorithm in online links box). Several applications of this approach have been described to search for associations between potentially large numbers of SNPs and disease status[47–49].

Even older than the *apriori* algorithm is an approach called *Logical Analysis of Data* (LAD)[50], which has been designed to "identify patterns of findings or syndromes that predict outcomes"[51]. Originally applied to problems in economics, seismology and oil exploration, LAD has recently been used to identify subsets of individuals with an increased risk of mortality after exercise electrocardiography[51]. It will be interesting to apply these more unusual methods to gene-mapping problems. Although the *apriori* algorithm seems suitable for handling extremely large data sets, LAD might be better suited for candidate genes with a limited number of genotypes because it is much more computationally intensive than the *apriori* algorithm.

As mentioned above, a theoretical investigation of different gene-mapping strategies[26] concluded that analysing all possible pairs of loci confers no advantage over locus-by-locus analysis. However, that analysis was based on inheritance models with rather strong main effects. To investigate the prospects of finding trait genes that exert weak effects by themselves, consider the extreme case of multi-locus inheritance models under which disease occurs only through interactions between loci[52,53]. TABLE 1 refers to one such purely epistatic inheritance model, in which only three genotype configurations potentially lead to disease (two of them are 100% penetrant). With allele frequencies of 0.5 at each of the three susceptibility loci, the resulting trait has a population prevalence of 6.25% and a heritability of 60%[53]. At each locus, the MARGINAL PENETRANCE is the same for each genotype, so an investigation of genotype or allele frequency differences between case and control individuals will not show anything unusual and will have no power. Interestingly, linkage analysis methods (for example, in families with two affected offspring) do have modest power for the gene mapping of such traits, which gives justification for the use of multiple analysis methods on the same data.

The only possible approach to association mapping of such an extremely complex trait is to search

MARGINAL PENETRANCE
In epistatic interactions between two loci asscoiated with disease, each with three genotypes, the nine genotype pairs might each be associated with a certain penetrance — that is, the probability that the genotype pair leads to disease. From these penetrances and the genotype frequencies, (marginal) penetrances might be computed — that is, penetrances that are associated with the genotypes at one of the two loci.

for patterns of genotypes at different loci. As TABLE 1 shows, pattern frequencies are quite different for affected and unaffected individuals at the trait loci — most patterns occur only in case or control individuals, but not in both. Although it might seem very simple, the problem is how to detect such patterns in the thousands of SNPs that are investigated. For example, with 1,000 SNPs, including disease loci *A*, *B* and *C*, the total number of different subsets of three SNPs is equal to $n = 166{,}167{,}000$. Only one of these corresponds to the three susceptibility loci. For the assumed 100 case and 100 control individuals in TABLE 1, the expected chi-square in the test for equality of pattern frequencies is equal to 166.67 (26 degrees of freedom because the total number of possible patterns is $3 \times 3 \times 3 = 27$), with an associated significance level of $p = 1.76 \times 10^{-22}$. If we test each of the n subsets of three SNPs, adjusting for multiple testing leads to a corrected significance level of $np = 3 \times 10^{-14}$, which is still highly significant. This result is realistic for the given three-locus epistatic model.

In practice, we do not know the number of susceptibility loci and would have to investigate patterns for different numbers of trait loci. On the other hand, the so-called BONFERRONI CORRECTION for multiple testing as applied above is very conservative[12]. So, it seems entirely possible to successfully search for patterns that discriminate between case and control individuals despite the very large numbers of patterns that need to be tested. Additional calculations not described here indicate that only modest increases in sample sizes are necessary to localize SNPs that are only in LD with nearby disease loci, rather than being disease causing as assumed above.

Analysing all pairs of SNPs

For an exhaustive search of all two-locus interactions in a case–control study, we recommend the following procedure, which is designed to find all pairs of SNPs for which a significant interaction is present in cases but not in controls, or *vice versa*. For a given pair of SNPs, construct a 3×3 contingency table that corresponds to the three genotypes at the two SNPs, one such table for cases and one for controls. Compute chi-squares with four degrees of freedom for each table and form their ratio, R, with the larger chi-square, c_L, being in the numerator. Owing to extreme SNP allele frequencies, some ratios might be large only because the smaller chi-square is close to zero. To avoid such artificial results, there is a focus on pairs of SNPs with an appreciably high c_L value — for example, $c_L > 7.78$, which corresponds to the ninetieth percentile of the chi-square distribution with four degrees of freedom. For each SNP pair that satisfies this condition, a p-value that is associated with the observed R value is computed. Owing to the conditions imposed and the fact that the SNPs might be correlated in a given group of individuals, the R values do not follow the usual F distribution under the null hypothesis of equality of SNP by SNP interactions. So, the appropriate null distribution based on five million randomly generated R values is approximated. Based on this distribution, a p-value for each pair of SNPs is computed and the SNP pairs are ordered by increasing p-value ($p_{(1)} \leq p_{(2)} \leq \ldots$). The FALSE DISCOVERY RATE (FDR) method[54] is then applied to determine which of the pairs is 'significant' at the 5% level. This is accomplished, for example, by computing $q_{(i)} = i \times 0.05/n$, where i refers to the order of p-values and n is the number of SNP pairs with $c_L > 7.78$. The pairs are classed as 'significant' when $p_{(i)} \leq q_{(i)}$. The term significant is set in quotation marks because it does not refer to the usual rate of false-positive results (that is, the proportion of results that are significant only by chance), but only to the FDR (that is, the proportion of results that are significant by chance out of all significant results)[55]. When applied to the restenosis data set[34], two pairs of SNPs (each on different chromosomes) show association in the controls but not in the cases, whereas none of the four SNPs show significantly different genotype frequencies between case and control individuals in simple chi-square tests. So, this is an example in which only interaction effects show a significant result. The previously mentioned significant sums over multiple single-SNP test statistics could be partly due to such interaction effects.

Conclusion

The rapid increase in the availability of large numbers of genetic markers, and the quest for localizing genes that underlie multi-gene disease traits, represent a challenge for statisticians to come up with statistical analysis methods that do justice to this situation. Here, we have shown the main new developments in this area. Presumably, none of these methods can be classed as the 'best'. One method might be particularly powerful for a given inheritance mode of trait, whereas another might be rather inefficient. Therefore, at least as a tool for data exploration, it will be useful to apply several different analysis methods to a given body of data.

Many of the newly developed approaches have been implemented in computer programs so that statisticians can make efficient use of them. Unfortunately, non-statisticians sometimes feel that statistical analysis is simply a matter of running a computer program — nothing could be further from the truth. Such programs belong in the hands of the specialists if truly meaningful results are to be obtained, in the same way as PCR machines belong in the hands of molecular biologists. There are probably two main reasons for the 'push-the-button' concept of statistical approaches. First, statistical analysis methods are sophisticated and often difficult to understand by 'lay' scientists, which might lead to the impression of statistics as an uncontrolled black box. Second, computer-based methods are gaining more and more importance in statistics[33], but the resulting computer programs still represent sophisticated statistical procedures, which sets them apart from programs that manipulate data or produce graphs.

BONFERRONI CORRECTION
When n statistical tests are carried out, each has the potential (probability, p, the significance level) to return a false-positive result. If tests are independent of each other, the so-called experiment-wise probability that one or more tests show a false-positive result is approximately np. So, to achieve an experiment-wise false-positive rate of p, each individual test must only be allowed a false-positive error rate of p/n, which is referred to as the Bonferroni correction.

FALSE DISCOVERY RATE
(FDR). The proportion of false-positive test results out of all positive (significant) tests (note that the FDR is conceptually different to the significance level).

1. Venter, C. Presentation given at the Annual Short Course in Medical and Experimental Mammalian Genetics in Bar Harbor, July 16–27, 2001.
2. Templeton, A. R., Weiss, K. M., Nickerson, D. A., Boerwinkle, E. & Sing, C. F. Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. *Genetics* **156**, 1259–1275 (2000).
3. Doerge, R. W. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Rev. Genet.* **3**, 43–52 (2002).
   A review of analysis methods for mapping quantitative trait loci (QTLs). Many of the methods can also be applied to other biological data sets for correlating quantitative phenotypes with genotypes.
4. Garrod, A. E. The incidence of alcaptonuria: a study in chemical individuality. *Lancet* **II**, 1616–1620 (1902).
5. Morton, N. E. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277–318 (1955).
   The original paper proposing the lod score analysis for human linkage studies.
6. Elston, R. C. & Stewart, J. A general model for the analysis of pedigree data. *Hum. Hered.* **21**, 523–542 (1971).
   The landmark paper describing what is known as the Elston–Stewart algorithm for the genetic analysis of large, extended pedigree data.
7. Ott, J. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am. J. Hum. Genet.* **26**, 588–597 (1974).
8. Ott, J. *et al.* Linkage studies in a large kindred with familial hypercholesterolemia. *Am. J. Hum. Genet.* **26**, 598–603 (1974).
   The first application of the lod score method in a large human kindred allowing for age-dependent penetrance that led to identification of the gene that is responsible for familial hypercholesterolaemia.
9. Berg, K. & Heiberg, A. Linkage between familial hypercholesterolemia with xanthomatosis and the C3 polymorphism confirmed. *Cytogenet. Cell. Genet.* **22**, 621–623 (1978).
10. Gusella, J. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
11. Tsui, L. C. *et al.* Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* **230**, 1054–1057 (1985).
    This work, together with their 1989 paper in *Science*, represents the earliest triumph in genetic linkage analysis with DNA markers (restriction fragment length polymorphisms, RFLPs) followed by molecular positional cloning. It assigned the cystic fibrosis (CF) locus to the long arm of chromosome 7 (7q31) and identified the CF transmembrane regulator (*CFTR*) as the disease gene.
12. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99 (2001).
    The authors review all association studies conducted so far and discuss some crucial issues in study designs.
13. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nature Rev. Genet.* **3**, 299–309 (2002).
14. Génin, E., Todorov, A. A. and Clerget-Darpoux, F. Optimization of genome search strategies for homozygosity mapping: influence of marker spacing on power and threshold criteria for identification of candidate regions. *Ann. Hum. Genet.* **62**, 419–429 (1998).
15. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
16. Risch, N. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
17. Bellman, R. *Adaptive Control Processes: a Guided Tour* (Princeton University Press, Princeton, 1961).
18. 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).
19. Ott, J. *Analysis of Human Genetic Linkage* (Johns Hopkins University Press, Baltimore, USA, 1999).
20. Hogben, L. The genetic analysis of familial traits. II. Double gene substitutions, with special reference to hereditary dwarfism. *J. Genet.* **25**, 211–240 (1932).
21. MacLean, C. J., Sham, P. C. & Kendler, K. S. Joint linkage of multiple loci for a complex disorder. *Am. J. Hum. Genet.* **53**, 353–366 (1993).
22. 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).
23. Schork, N. J., Boehnke, M., Terwilliger, J. D. & Ott, J. Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am. J. Hum. Genet.* **53**, 1127–1136 (1993).
24. Knapp, M., Seuchter, S. A. & Baur, M. P. Two-locus disease models with two marker loci: the power of affected-sib-pair tests. *Am. J. Hum. Genet.* **55**, 1030–1041 (1994).
25. Fan, R., Floros, J. & Xiong, M. Transmission disequilibrium test of two unlinked disease loci; application to respiratory distress syndrome. *Adv. Appl. Stat.* **1**, 277–308 (2001).
26. Dupuis, J., Brown, P. O. & Siegmund, D. Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics* **140**, 843–856 (1995).
    The first rigorous theoretical work that compares single-locus search, simultaneous search and conditional search for the mapping of a trait caused by two susceptibility genes.
27. 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).
28. Cruickshanks, K. J. *et al.* Genetic marker associations with proliferative retinopathy in persons diagnosed with diabetes before 30 yr of age. *Diabetes* **41**, 879–85 (1992).
29. Felsenfeld, S. & Plomin, R. Epidemiological and offspring analyses of developmental speech disorders using data from the Colorado Adoption Project. *J. Speech Lang. Hear. Res.* **40**, 778–791 (1997).
30. Rao, C. R. & Wu, Y. in *Model Selection* (ed. Lahiri, P.) 1–57 (IMS Lecture Notes Monograph Series, Volume 38, Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2001).
31. Lucek, P. R. & Ott, J. Neural network analysis of complex traits. *Genet. Epidemiol.* **14**, 1101–1106 (1997).
32. Lucek, P., Hanke, J., Reich, J., Solla, S. A. & Ott, J. Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. *Hum. Hered.* **48**, 275–284 (1998).
33. Diaconis, P. & Efron, B. Computer-intensive methods in statistics. *Sci. Am.* **248**, 116–130 (1983).
34. Zee, R. Y. *et al.* Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. *Pharmacogenomics J.* **2**, 197–201 (2002).
35. Hoh, J., Wille, A. & Ott, J. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.* **11**, 2115–2119 (2001).
36. 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).
37. 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).
38. Zhang, H., Tsai, C. P., Yu, C. Y. & Bonney, G. Tree-based linkage and association analyses of asthma. *Genet. Epidemiol.* **21**, S317–S322 (2001).
39. Zhang, H. & Singer, B. *Recursive Partitioning in the Health Sciences* (Springer, New York, 1999).
40. Gabriel, S. B. *et al.* Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nature Genet.* **31**, 89–93 (2002).
    This paper offers an innovative method that, for the first time, provides complete genetic dissection of a multifactorial disorder.
41. Bruning, J. C. *et al.* Development of a novel polygenic model of NIDDM in mice heterozygous for IR and IRS-1 null alleles. *Cell* **88**, 561–572 (1997).
42. Savage, D. B. *et al.* Digenic inheritance of severe insulin resistance in a human pedigree. *Nature Genet.* **31**, 379–384 (2002).
43. Martin, M. P. *et al.* Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nature Genet.* **31**, 429–434 (2002).
44. Ming, J. E. & Muenke, M. Multiple hits during early embryonic development: digenic diseases and holoprosencephaly. *Am. J. Hum. Genet.* **71**, 1017–1032 (2002).
45. Agrawal, R., Imielinski, T. & Swami, A. in *Proceedings of ACM SIGMOD Conference on Management of Data* (eds Buneman, P. & Jajodia, S.) 207–216 (Association for Computing Machinery, Washington, USA, 1993).
46. Agrawal, R. & Srikant, R. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Databases [online], (cited 1 August 2003), <http://www.almaden.ibm.com/cs/people/ragrawal/papers/vldb94_rj.ps> (1994)
47. Toivonen, H. T. *et al.* Data mining applied to linkage disequilibrium mapping. *Am. J. Hum. Genet.* **67**, 133–145 (2000).
48. Flodman, P., Macula, A. J., Spence, M. A. & Torney, D. C. Preliminary implementation of new data mining techniques for the analysis of simulation data from Genetic Analysis Workshop 12: Problem 2. *Genet. Epidemiol.* **21**, S390–S395 (2001).
49. Czika, W. A. *et al.* Applying data mining techniques to the mapping of complex disease genes. *Genet. Epidemiol.* **21**, S435–S440 (2001).
50. Crama, Y., Hammer, P. L. & Ibaraki, T. Cause–effect relationships and partially defined Boolean functions. *Ann. Oper. Res.* **16**, 299–326 (1988).
51. Lauer, M. S. *et al. et al.* Use of the logical analysis of data method for assessing long-term mortality risk after exercise electrocardiography. *Circulation* **106**, 685–690 (2002).
52. Frankel, W. N. & Schork, N. J. Who's afraid of epistasis? *Nature Genet.* **14**, 371–373 (1996).
    In their comments on the two reports in the same issue of the journal, the authors predict that genetic epistasis is a common phenomenon for complex phenotypes despite only sparse evidence at the time.
53. 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).
54. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125**, 279–284 (2001).
55. Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375 (2003).
56. Lander, E. & Kruglyak, L. Genetic disseaction of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11**, 241–247 (1995).
    The authors formally address the multiple-testing problem in gene mapping and show how statistical significance can arise by chance alone due to a large number of tests performed. They provide rigorous genome-wide thresholds for testing significance based on the assumption of a dense marker map.
57. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman and Hall, New York, 1998).

Acknowledgements

This work was supported by grants from the National Institute of Mental Health.

Online links

DATABASES

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink

APOA | *APOC3* | *CBS* | *CD14* | *CETP* | *IL4α* | *KIR* | *MDM2* | *NOS* | *SERPINE1* | *TNFR1* | *TP53*

OMIM: http://www.ncbi.nlm.nih.gov/omim

Hirschsprung disease

FURTHER INFORMATION

***apriori* algorithm:**

http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html

Jurg Ott's laboratory: http://linkage.rockefeller.edu/ott

Access to this interactive links box is free online.