Statistical methods in genetics

Giovanni Montana

Submitted: 4th May 2006; Received (in revised form): 13th July 2006

Abstract

In recent years, a very large variety of statistical methodologies, at various levels of complexity, have been put forward to analyse genotype data and detect genetic variations that may be responsible for increasing the susceptibility to disease. This review provides a concise account of a number of selected statistical methods for population-based association mapping, from single-marker tests of association to multi-marker data mining techniques for gene–gene interaction detection.

Keywords: *association studies; linkage disequilibrium; gene mapping; data mining; epistasis*

INTRODUCTION

Statistical genetics is an area at the convergence of genetics and quantitative analysis. Over the last few years it has experienced a drastic shift of paradigm, from a mostly theoretical subject with little room for empirical evidence to a heavily data-oriented discipline where the existence of large repositories of genetic data allows researchers to generate and explore new scientific hypotheses.

With the advent of relatively cost-efficient high-throughput genotyping technology (with recent proposals from the US National Institutes of Health to decrease these costs further to 0.001 USD per genotype) it is now feasible to investigate the aetiology of complex diseases, the biological processes through which DNA is inherited and the evolutionary histories of human populations. From a medical perspective, advances in the design and analysis of pharmacogenetics studies, i.e. studies in which genetic variability is correlated to drug response, may ultimately lead to the development of a 'personalised medicine' approach to healthcare. Each of these areas of investigation requires, of course, specialised inferential and computational techniques.

This review of statistical methods in genetics is confined to *association mapping*: a powerful methodology that is believed will help understand the genetic basis of human diseases and other phenotypes of interest. Rather than attempting a broad coverage of association mapping methods, the exposition is narrowed to include only data analysis approaches for case-control studies or for situations when only diseased individuals are available. The aim of this review is to embark the reader on a non-technical tour around a number of selected statistical methods currently used for gene mapping.

PRINCIPLES OF ASSOCIATION MAPPING

The distinctive feature of a case-control design is that the subjects included in the sample are randomly selected from a given population by their disease status, retrospectively. The genetic make-ups of individuals belonging to the two groups, cases and controls, are compared in the hope that their differences, in some narrow regions of the genome, may offer a causal explanation for the disease status. Among different types of genetic markers, single-nucleotide polymorphisms (SNPs) play a central role for mapping complex diseases. Across the human genome, there are at least 10 million SNPs with frequency >1% which are thought to account for around 90% of human genetic variation [1].

A fundamental notion in association mapping is that of linkage disequilibrium (LD) between a genetic marker and the locus that affects the trait under study. LD captures a deviation from probabilistic independence among alleles or genetic markers. For instance, LD between two alleles, say A and B, can be quantified by measuring the difference between p_{AB}, the probability of observing haplotype

Giovanni Montana, Department of Mathematics, Imperial College London, Huxley Building Room 532, 180 Queen's Gate, London SW7 2AZ, UK. Tel, +44 (0) 207-594-8577; E-mail, g.montana@imperial.ac.uk

Giovanni Montana is a Lecturer of Statistics in the Mathematics Department, Imperial College London, UK. He is interested in data mining and statistical pattern recognition.

AB (i.e. the linear arrangement of the two alleles on the same chromosome, inherited as a unit), and the product $p_A p_B$, where p_A and p_B are the probabilities of observing alleles *A* and *B*, respectively. Haplotypes, however, are not directly available in most cases and their frequencies must be inferred probabilistically from genotype data.

Inferential methods based on variants of the expectation-minimization (EM) algorithm, an iterative technique for obtaining maximum likelihood estimates in missing-data models, are popular choices for obtaining sample haplotype frequencies [2, 3]. The EM algorithm's accuracy for estimating haplotype frequencies, under a variety of simulation designs and as a function of allele frequencies as well as many other factors, has been documented [4]. Recent developments exploit the observation that, over short regions, haplotypes in a population tend to cluster into groups, and this clustering tends to vary along the chromosome. It has been noted that the resulting patterns of genetic variation can be described well by hidden Markov models, and parameter estimates have been carried out by an EM algorithm in order to infer haplotypic phase as well as missing genotype data [5].

Alternatively, a measure of composite genotypic disequilibrium can be computed directly from two-locus genotypic data [6]; under the assumption of random mating, it corresponds to the aforementioned allelic LD measure. A number of other common LD coefficients and their properties have been studied both analytically and via simulations [7, 8]. Once LD is created by a number of evolutionary forces, it is subjected to recombination events taking place between loci, which cause it to decay with time.

The essential idea is that a marker in strong LD with a disease locus is expected to be located nearby. But how is LD used to map a gene? After all, when one of the loci of interest is the gene that is being mapped, we have no information about its allele or genotype frequencies. Association mapping techniques attempt to detect LD *indirectly*, by measuring the association between a candidate marker and the phenotype of interest, provided there is a rich pattern of LD between some of the typed markers and the real, unobserved causal variant. How dense this map of markers should be and what the distribution of LD looks like in modern human populations are crucial issues being extensively explored [1, 9]. Notably, the HapMap project is enabling the characterisation of genome-wide patterns of LD in several populations [10].

In what regions of the genome do we look for disease-bearing genes? In *candidate-gene* approaches, it is assumed that prior biological hypotheses about plausible locations of the candidate gene have been previously obtained, and therefore the search is localised to those regions of interest. *Genome-wide studies*, on the other hand, screen the entire genome, thus enabling a more comprehensive search for genetic risk factors. These studies will soon be less expensive, and therefore more routinely employed. From a statistical and computational standpoint, genome-wide explorations introduce non-trivial challenges due, among other causes, to the very large amount of markers to be included in the analysis compared to the usually smaller sample sizes; these issues will be considered further later.

Another question generating much discussion, and fuelling the development of new analytical methods, is whether complex diseases are caused by a single common variant or many variants having small effects. The *common disease/common variant* hypothesis states that the genetic risk for common diseases will often be due to disease producing alleles found at relatively high frequencies [11–13]. So far, evidence in its favour has been limited. It is plausible to assume that common diseases are expected to be controlled by more complex genetic mechanisms characterised by the joint action of several genes, with each gene having only a small marginal effect, perhaps because natural selection has removed the genes having larger effects. In this scenario, groups of markers should be tested jointly for association, which can be done in two main ways: by grouping markers together in multi-locus genotypes so that the basic unit of statistical analysis is still the individual, or via haplotypes, thus effectively doubling the sample size. We next review a number of selected techniques, starting from the simplest case of single-marker analysis.

SINGLE-MARKER ANALYSES

Common tests and other likelihood-based methods

Suppose we are investigating the effects of biallelic markers, e.g. SNPs, on disease. In a case-control setting, counts of either the two alleles or the three genotypes at a locus in the two groups, affected and controls, are compared. If there is a difference

in frequencies between the two samples, there is evidence that the marker is in LD with the gene affecting the disease susceptibility. Since allelic and genotypic distributions in the samples can be arranged in a standard Fisherian contingency table of dimension 2×2 and 2×3, respectively, a number of well-known statistical tests exist to assess the null hypothesis of no association [14].

A simple test for independence is the Pearson's chi-squared test statistic, at both the allelic and genotypic level. However, it has been noted that this test is not robust to departures from Hardy–Weinberg equilibrium (HWE) in control subjects [15]. HME implies the statistical independence of the two alleles at a locus so that, for instance, the proportions of genotypes *AA*, *Aa* and *aa* are p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$, respectively. The utility of single-marker LD testing that uses a case-control study design with either diallelic or multiallelic markers is discussed in [16]. In this work, under the alternative hypothesis of unequal marker allele frequencies between cases and controls, the asymptotic distribution of the chi-squared test is expressed as a function of G^2, a genetic distance measure, which depends on the population history; using a simple deterministic population genetic model accounting for a single mutation and ignoring genetic drift, the value of G^2 can be computed and the power of the test obtained under various disease models and population histories. A more robust test statistic is Cochran–Armitage (CA) trend test, a method of directing chi-squared tests towards narrow alternatives [17]. This test should be used at the genotypic level when HWE fails to hold for both cases and controls [15, 18].

Fine localisation of a disease–susceptibility locus can be accomplished by investigating deviations from HWE among affected individuals alone [19, 20]. Hybrid tests have also been suggested, for instance a test statistic obtained as a weighted average between the CA trend test statistic and the difference between test statistics based on HWE computed in cases and controls [21, 22]. Departures from HWE can also serve as a quality check on the data, as experience suggests that gross deviations from HWE often indicate genotyping errors.

Alternatively, one can explicitly model the penetrance of the disease, that is the conditional probability that a randomly selected individual in the population possesses the disease, given the data. In the logistic regression (LR) formulation, the logit transform of the penetrance parameter is modelled as a linear combination of the marker data. Then, by asymptotic results of maximum likelihood estimators, inferences can be based on standard Walk, likelihood ratio and score methods [14]. In particular, the score statistic in this case corresponds to the CA trend test. An obvious advantage of this formulation is that covariates can be easily added into the model. In situations where the sample size is large compared with the number of parameters, as well as in matched case-control studies (where cases and controls are paired according to some control variates), improved inference can be achieved by using conditional maximum likelihood, in a generalisation of Fisher's exact test for 2×2 tables [14]. Recent developments that allow efficient approximate conditional inference for LR models include Monte Carlo methods and saddle-point approximations [23].

Several statistical methods for association mapping, including LR as well as other generalised linear models, require the specification of a genetic model of inheritance. For instance, in a CA test, or score statistics from logistic regression, an additive model can be imposed by giving genotype weights 0, 1 and 2, depending on the number of copies of the minor allele. Forcing a specific genetic model provides a powerful means of detecting association when the hypothesised model is close to the true underlying genetic mechanism, but may also lead to very low power when the true model is different [18, 24]. Methods that do not require the specification of a genetic model are usually recommended [25].

Methods to correct for population stratification and cryptic relatedness

The approaches presented so far rely on two fundamental assumptions: first, the population under study must be genetically homogeneous, i.e. there should be no population stratification; second, all subjects in the samples must represent statistically independent units drawn from that population. Tests of association that do not protect against departures from these two assumptions may have inflated type I error rates.

If the target population does consist of several subpopulations, spurious associations at a candidate marker may occur if the disease prevalence differs between subpopulations, i.e. when the subpopulation proportions are different among cases and controls and if allele frequencies at that marker vary between subpopulations. When the population

is indeed heterogeneous, family-based association studies are generally more powerful than case-control studies, and tests that rely on the transmission of alleles from parents to off-springs are usually adopted [26]. However, these study designs present other drawbacks, most notably the difficulty of collecting DNA from relatives of affected individuals, especially for late-onset diseases, thus mitigating against the recruitment of large samples. Research in the area of case-control studies has actively addressed these issues and several alternatives are available.

In a *genomi-control* approach [27], test statistics for the null hypothesis of no association are computed at 'null' markers in the genome, i.e. at markers unlinked to affect liability. If population structure or cryptic relatedness is present in the sample, the variability and magnitude of the test statistics at the null markers are inflated and tests computed at candidate loci can be adjusted accordingly.

A different remedy, often referred to as *structured association*, prescribes using loci unlinked to candidate genes under study to infer subpopulation membership and conduct a test of association within subpopulations. The idea is that, conditional on subpopulation, there is neither bias nor excess of variance due to population substructure. The method can be implemented as a two-step procedure, in which subpopulation proportions are estimated first and then incorporated into a test statistic [28, 29] (for instance, as covariates in a LR model), or as a unified analysis which may account for estimation uncertainty [30, 31]. Either way, the task of estimating subpopulation memberships from genotype data is essentially a clustering (or unsupervised learning) application, often addressed by using finite mixture distributions [32]. Both Bayesian and likelihood-based inferential procedures can then be employed [33–35]. In a variation of the structured association idea, the disease status is included in the clustering algorithm used for inferring the hidden population structure, leading to a supervised clustering approach [36]. The related question of what markers are particularly informative for ancestry estimation is also important and has been investigated from an information-theoretic perspective [37]. Recently, it has also been suggested that the use of LR alone, which dispenses entirely with the notion of subpopulation and is computationally faster, may be a better alternative [38, 39].

Unlike genomic control, structured association alone does not protect against cryptic relatedness, and more specialised solutions are needed. However, a recently developed theory to predict the amount of cryptic relatedness expected in random mating populations suggests that confounding effects in this situation are particularly serious only in special cases [40]. On the other hand, even moderate levels of population stratification may lead to an increased number of false positives [41], especially in large case-control studies [42] and even in well-designed studies [43] or when the population under study is believed to be homogeneous [44].

Methods to correct for multiple testing

In a single-marker analysis, a test statistic is computed at each candidate marker. When M hypothesis tests are conducted with the same significance level α, the probability of finding at least one significant result among the M tests is greater than α. To deal with this multiple-testing problem, one has to resort to some form of correction that preserves the probability of observing unusually 'large' values of a test statistic and not just at a specific locus, in the usual point-wise sense, but anywhere in the region being tested.

One way to deal with this situation is to control the family-wise error rate (FWER), i.e. the probability of making one or more type I errors among all the hypotheses when performing multiple pair-wise tests. Suppose we have obtained the p-values p_i, $i=1,\ldots,M$, from each individual test. A common one-step adjustment is the Bonferroni correction: an hypothesis i is rejected when $p_i \leq \alpha/M$. Since this correction is too conservative when M is large, leading to power loss, a number of step-wise procedures have been developed. For instance, in the Hochberg's procedure [45], the individual p-values are initially ordered; then, starting from $i=M$, all the hypothesis for $i \leq j$ are rejected once the p-value at position $j \leq \alpha/(M-j+1)$.

An alternative multiple hypothesis testing error measure is the false discovery rate (FDR) [46], which is loosely defined to be the expected proportion of false positives among all significant hypotheses. The FDR is especially appropriate for exploratory analyses in which one is interested in finding several significant results among many tests. After ordering the p-values, all the hypotheses for $i \leq j$ are rejected if $p_j \leq j\alpha/M$. As a special case, when all the hypotheses are true, this error rate equals the FWER.

One way to account for all hypotheses simultaneously is to form the product of all p-values at less than a preset threshold [47, 48]; as a variation along these lines, one can form the product of the $L < M$ most significant p-values only, which is an appropriate choice when the aim is to detect a small set of fixed effects among a large number of null effects [49].

It must be noted that there are substantial correlations among test statistic values along the map induced by LD between genetic markers. As a result of this correlation, the 'effective' number of independent tests, say $M^\star$, is expected to be smaller than M. One way to compute $M^\star$ and correct for multiple testing is via spectral decompositions of the LD matrix [50], as in the adaptive principal component test [51]. Generally, it is difficult to formally account for this serial correlation. For instance, distributions for products of p-values are only known when the tests are independent. Monte Carlo procedures are commonly used, for instance by randomly permuting phenotypic labels [52] or by using permutation sampling for fitting extreme value distributions [53]. However, many simulation-based methods become extremely time consuming when applied to large studies. Other approaches deal with this problem by sequentially decorrelating the tests, either by application of a single transformation derived from the correlation matrix [47] or by successive greedy transformations [54].

MULTI-MARKER ANALYSES

Methods for combining information from single-marker coefficients

Under simple models of evolution, ignoring population-specific history and structure, the probability of recombination is a monotonic function of genetic distance, and the degree of LD across a chromosome is expected to follow a unimodal curve with a peak at the true location of the disease mutation. Under this assumption, a strategy to combine information from single markers in a region is to fit a smooth curve to the LD coefficients computed at all markers and then look for its mode. In practice, however, the pattern of observed LD may fluctuate substantially, even erratically, across contiguous genomic regions [1]. The gene-mapping problem then becomes one of pattern recognition: the task is to look for regions with a consistent overall pattern of LD supporting the existence of a disease-associated marker.

Non-parametric curve-fitting methods embracing this idea have been initially developed for fine-mapping, thus assuming that the region under study does contain a true peak [55, 56]. However, in large exploratory scans, the possibility that the data may have a varying number of true signals, or even no signal at all, has to be taken into account. In this respect, a curve-fitting method based on Bayesian adaptive regression splines with a variable number of knots has been applied with some success [57, 58].

A different proposal consists of fitting a semi-Bayesian hierarchical model, where a pair-wise LD measure is first estimated for each locus using a first-stage model, and then spatially smoothed along the candidate region using a second-stage model that can include information on genetic or physical distances as well as haplotype structure [59].

An alternative solution for combining information from neighbouring markers consists of forming sums of single-marker test statistics and then testing the null hypothesis that none of the selected markers in each sum is associated with the disease, in what has been called the *set association* approach [60–62]. Combining genetic main effects in this way may facilitate the detection of susceptibility genes while avoiding the need to characterise detailed interaction patterns among markers. When compared with Bonferroni or FDR procedures, the sum statistics seem to show greater power [54].

Common methods for haplotypes

Rather than considering each marker individually, specific combinations of allelic variants at a series of tightly linked markers on the same chromosome, i.e. haplotypes, can be jointly tested. Incorporating information from multiple adjacent markers, haplotypes preserve the joint LD structure and more directly reflect the true polymorphisms. Therefore, in a generalisation of the single-marker analyses presented in the preceding text, haplotype frequencies between cases and controls can be compared instead of allelic and genotypic frequencies [63].

The simplest way to test whether there is an association between a haplotype and the disease status is to regard each haplotype as a distinct category, possibly lumping all rare haplotypes together into an additional class. This process is generally done in two steps: first, haplotypes

frequencies are estimated (for instance by applying the EM algorithm) so that the H distinct haplotypes compatible with the data are arranged in a $2 \times H$ contingency table; then, a standard test for association, for instance a likelihood ratio test statistic, is calculated. To deal with the inflated variance of the test statistic due to the haplotype estimation, the distribution of the test under the null can be obtained by randomly shuffling the disease status and then re-estimating haplotype frequencies [64].

Although this approach assesses overall association between haplotypes and disease, it does not provide inference on the effects of specific haplotypes or haplotype features. To address these issues, a number of tests of specific haplotype effects are based on a prospective likelihood of disease [65, 66], where the disease status is treated as an outcome, and haplotypes enter a regression model as covariates. Subjects with ambiguous haplotypes are accommodated by computing the expected value of the covariates conditional on the subject's genotypes, using inferred haplotype frequencies estimated in the pooled sample of cases and controls under the HWE assumption. Alternatively, in a retrospective likelihood approach, the distribution of haplotypes is treated as the outcome, conditional upon the case and control status; this model relaxes the requirement of HWE and has been estimated using an expectation-conditional-maximisation (ECM) algorithm [67].

Recent simulation studies compare the performance of various regression and multiple imputation approaches for the estimation and testing of genotype and haplotype effects in a case-control setting [68]. Overall, the use of haplotypes derived from phase-unknown genotype data is not always straightforward, and the value of these techniques for gene mapping is not yet clear [69, 70]. Missing data are a particular problem for haplotype analysis, especially when the data are not missing at random, as may be the case with systematic errors in genotyping assays.

In a more flexible setting, rather than comparing frequencies of entire haplotypes, one can compare the frequency of haplotype patterns, i.e. haplotypes that are allowed to contain gaps (markers that can be ignored). Gaps account for mutations, missing data, errors and recombination events that may have corrupted the ancestral haplotype shared among cases; candidate haplotype patterns can then be tested for association using a chi-squared test [71].

When either multiple markers or haplotypes are available, a generalisation of Hotelling's T^2 statistics provides another valid alternative that implicitly accounts for LD among markers and the possibility of multiple disease susceptibility loci [72]. This test statistics improves, in terms of power, upon the standard chi-squared test and has been extended to deal with haplotype blocks with multiple haplotypes [73]. Stochastic search procedures like the sequence-forward floating-selection (SFFS) algorithm can be used jointly with the T^2 test to identify markers that make the greatest contribution to disease risk [72].

Methods based on haplotype similarity and clustering

In proximity of the disease mutation, the average similarity among case haplotypes is expected to be higher than among control haplotypes because of shared ancestry. Provided that a sound similarity measure between two haplotypes can be defined, formal statistical tests can then be built to detect excess of haplotype similarity, even from multi-locus genotype data [74]. A practical analytical challenge is the possibility that not all case chromosomes have inherited disease-causing mutations from a common ancestral chromosome. Moreover, disease mutations only increase the risk of being affected, but not every subject carrying the mutations will be affected.

A different strategy, still based on a notion of haplotype similarity, is to form clusters of similar haplotypes and then test the clusters for associations with the disease rather than the individual haplotypes. The idea is that haplotypes within a cluster will contain many of the same polymorphisms, and hence, should induce similar effects on disease predisposition. Since the number of clusters is much lower than the number of all possible haplotypes, the number of degrees of freedom is reduced and statistical power is gained. The clustering approach can also be seen as a vehicle to account for evolutionary processes indirectly, without having to explicitly model them.

There are several procedures to detect disease-linked, non-random clustering of haplotypes in localised genomic regions. An elegant one is based on a cladistic analysis: by sliding a window along the chromosome, clades of similar haplotypes in the whole genome are detected and incorporated into an LR model [75]. Under the assumption of multiplicative disease risks, the proposed model is parameterised in terms of haplotypic log

odds of disease. The method has also been extended to analyse un-phased genotypes directly [76].

A different procedure scans all markers available, one by one, and clusters haplotype segments of a specified length, centred at the marker under examination, by using a density-based clustering algorithm. A standard chi-squared test based on the contingency table derived from the numbers of case haplotypes and control haplotypes in a cluster is then used to test the null hypothesis of no association between the cluster and the disease status [77].

Another way to embed haplotypes in a metric space, so that similar haplotypes convey similar risks, is to first introduce a Voronoi tessellation structure and assign each haplotype to the nearest cluster centre [78]; the distance between a haplotype and the cluster centre, which represents the putative ancestral haplotype, is computed by counting the number of matching markers flanking the location of the causal location. As a further refinement of this approach, rather than using all possible clusters, one can restrict the search to the largest ones, and adopt a similarity measure that accounts for allele frequencies and for occasional mismatches [79]. Different procedures to group haplotypes use tree-based methods [80] and other clustering techniques [81, 82].

Methods based on population admixture

In human populations formed by relatively recent mixing of distinct ancestral groups, like African-Americans, LD extends over greater distances than in other, less heterogeneous populations. For diseases that vary in prevalence between two or more ancestral populations, this long-range LD can be exploited to search for genetic variants responsible for the ethnic difference in disease risk [83].

The fundamental observation is that, in admixed populations, markers in LD with a locus responsible for an ethnic difference in disease risk will have a greater than expected proportion of ancestry from the high-risk population. Gene mapping can be carried out by searching for narrow genomic regions that show excessive ancestry proportions from one of the constituent ancestral populations in a methodology called admixture mapping.

Population memberships at each locus, for all subjects, need to be statistically estimated from the typed markers. A commonly used probability model to describe the stochastic variation in ancestry assumes that chromosomes can be represented by blocks of common ancestry, with breakpoints between adjacent blocks occurring as a Poisson process and transitions between adjacent ancestral blocks governed by a Markov chain [84]. Several Bayesian inferential methods have been built around this model to estimate the ancestry of diseased chromosomes and detect over-represented ancestral populations [85–87]. Simulation studies and analytical computations suggest that admixture mapping has several advantages over established approaches for population-based mapping, e.g. it requires far fewer markers to search the entire genome and is less affected by allelic heterogeneity [86, 88].

Data mining methods for interaction detection

When several disease genes contribute to a trait, one may detect them jointly by modelling the complex interaction pattern among loci. However, while standard regression techniques can provide insight into the main marginal effects while controlling for potential confounders, these techniques may be inadequate for identifying gene–gene interactions given sample size limitations. For instance, in an LR, when high-order interactions are modelled, many cells in the corresponding contingency table may contain no observations, a problem that can lead to very large coefficient estimates and standard errors. More advanced methods are therefore called for.

Classification trees are non-parametric models offering advantages over typical logistic regression in that they may uncover interactions among genes even when these do not exhibit strong marginal effects [89–91]. A classification tree model is built through an iterative process known as recursive partitioning, in which the data is split into partitions: the first step finds the best split of the data into two groups, or nodes, by one of several predictor markers that captures the most information in the response's variability; succeeding steps then find the best splits of the data within each node, conditional on prior splits, which results in an easily interpreted binary tree structure. A specific adaptation of this methodology for association mapping implements a different mechanism for the selection and combination of predictors while retaining the simple tree structure [92].

A random forests model is an ensemble of individual classification trees. Such a model is grown on bootstrap samples of observations, using

a random subset of predictors to define the best split at each node. The observations left out of the bootstrap samples are used to estimate the prediction error. Random forests are highly accurate classifiers that can handle a very large number of markers and can estimate the importance of each marker as a predictor of disease status. Simulation studies under simple genetic models suggest that markers selected according to the importance measure computed with random forests perform better than markers selected according to standard Fisher's exact tests [93]. A predictive importance index for pairs of predictors has been derived, and its behaviour has been studied over a range of two-locus disease models [94]. To date, recursive partitioning methods have been applied to a number of association studies, for instance for the analysis of alcoholism and smoking [95] and ischaemic stroke [89], and are considered a useful exploratory tool in pharmacogenetics [96, 97]. It has also been found that combining clustering techniques with recursive partitioning, so that individuals are initially grouped into homogeneous genotypic groups, may lead to better predictive models [98].

Multivariate adaptive regression splines (MARS) have also proved to be potentially useful for the detection of marker interactions in association studies [89, 99]. MARS are non-parametric adaptive multiple regression models particularly useful where the data present non-linearities, complex interactions and a large number of predictors. They form basis functions, typically linear spline functions, from the original data, and use these as candidates for interaction effects. An upper limit to the maximum order of interaction can be easily set up so that, for instance, only pair-wise products of piece-wise linear functions are fitted but not higher products [100].

Several efficient strategies that look for statistical interactions based on logistic regression models and Bonferroni corrections are described in [101]. A different kind of regression model that has been recently developed to explore these interactions is the logic regression [102]. This is another adaptive regression methodology, but attempts to construct predictors as Boolean combinations of biallelic markers. In a Monte Carlo LR, a large number of logic regression models are explored using stochastic simulation methods based on Markov chains, in search for the most frequently occurring interaction patterns supported by the data. Alternatively, the model space can be explored using stochastic search mechanism based on genetic algorithms [103].

The Bayesian selection of interactions method is a technique for selecting predictors in a regression model even when the number of variables entering the model is considerably larger than the sample size [104]. This model builds on a well-known Bayesian procedure for variable selection, stochastic search variable selection [105,106], which entails the specification of a hierarchical latent mixture prior and uses the posterior probabilities to identify the more promising models. Posterior distributions are estimated by Gibbs sampling. Since association mapping is essentially a variable selection problem [107], we expect that more efficient variable selection methods along these lines will be developed, especially for large genome-wide studies.

Non-parametric methods based on combinatorial arguments have also been proposed for gene–gene interaction detection, for instance the combinatorial partitioning method (CPM), originally developed for quantitative outcomes [108], and the multivariate dimensionality reduction (MDR) method for discrete outcomes in balanced case-control studies [109–111]. The rationale of the latter method is to pool multi-locus genotypes into high-risk and low-risk groups; this effectively reduces the genotype predictors to one dimension and enables interaction detection in samples of relatively small sizes. MDR may be interpreted as a special case of classification trees [112], and has received quite a lot of interest in recent years [113, 114].

CONCLUSIONS

Inevitably, several important statistical methods for association mapping have not been covered, for instance, procedures that exploit or account for the known relatedness of individuals and for quantitative traits mapping. The plethora of methods for selecting representative or 'tagging' polymorphisms in narrow genomic regions, based on patterns of LD, has not been touched upon; this task essentially entails a variable selection problem in an unsupervised setting, i.e. with no reference to the disease status, with the objective of effectively reducing the number of potential predictors in genome-wide studies [115]. Most notably, a large class of inferential methods based on approximations of the coalescent stochastic process, which explicitly describes the effects of

evolutionary forces that gave rise to the observed data, has been left out.

Despite these and other omissions, we hope to have conveyed the flavor of many statistical ideas and tools currently being employed in population-based gene mapping.

This review article highlights the main issues involved in gene mapping by linkage disequilibrium and offers a brief and non-technical overview of selected statistical methods that have been suggested in the literature. Initially, the most traditional single-marker tests of association are introduced, together with standard methodologies to correct for population structure, cryptic relatedness and multiple testing. More advanced data mining techniques involving multiple markers and haplotypes are then considered with emphasis on detecting epistatic effects.

Ackowledgement
G.M. thanks the anonymous referees for the helpful suggestions.

References

1. Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* 2005;**366**: 1223–34.
2. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995;**12**:921–27.
3. Niu T. Algorithms for inferring haplotypes. *Genet Epidemiol* 2004;**27**:334–47.
4. Fallin D, Schork NJ. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet* 2000;**67**:947–59.
5. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population structure in genetic association studies. *Genome Res* 2006;**16**:290–6.
6. Weir BS. *Genetic Data Analysis II*. Sinauer Associates, 1996.
7. Guo SW. Linkage disequilibrium measures for fine-scale mapping: a comparison. *Hum Hered* 1997;**47**:301–14.
8. Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995;**29**:311–22.
9. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;**273**:1516–7.
10. The International HapMap Consortium. The International HapMap Project. *Nature* 2003;**426**:789–96.
11. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001;**69**:124–37.
12. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant . . . or not? *Hum Mol Genet* 2002;**11**:2417–23.
13. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001;**17**:502–10.
14. Agresti A. *Categorical Data Analysis*. Wiley, 2002.
15. Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997;**53**:1253–61.
16. Chapman NH, Wijsman EM. Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am J Hum Genet* 1998;**63**:1872–85.
17. Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955;**11**:375–86.
18. Freidlin B, Zheng G, Li Z, *et al*. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered* 2002;**53**:146–52.
19. Nielsen DM, Ehm MG, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet* 1998;**63**:1531–40.
20. Jiang R, Dong J, Wang D, *et al*. Fine-scale mapping using Hardy–Weinberg disequilibrium. *Ann Hum Genet* 2001;**65**: 207–19.
21. Song K, Orloff M, Lu Q, *et al*. Fine-mapping using the weighted average method for a case-control study. *BMC Genet* 2005;**6**(Suppl 1):S67.
22. Song K, Elston RC. A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Stat Med* 2006;**25**: 105–26.
23. Corcoran C, Mehta C, Patel N, *et al*. Computational tools for exact conditional logistic regression. *Stat Med* 2001;**20**: 2723–39.
24. Schaid DJ, Kk McDonnell S, Hebbring SJ, *et al*. Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet* 2005;**76**:780–93.
25. Wang K, Sheffield CC. A constrained-likelihood approach to marker-trait association studies. *Am J Hum Genet* 2005;**77**: 768–80.
26. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;**52**:506–16.
27. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;**55**:997–1004.
28. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 2001;**60**:227–37.
29. Purcell S, Sham P. Properties of structured association approaches to detecting population stratification. *Hum Hered* 2004;**58**:93–107.
30. Hoggart CJ, Parra EJ, Shriver MD, *et al*. Control of confounding of genetic associations in stratified populations. *Am J Hum Genet* 2003;**72**:1492–504.
31. Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001;**68**:466–77.
32. McLachlan G, Peel DA. *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2000.
33. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;**155**:945–59.
34. Chikhi L, Bruford MW, Beaumont MA. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* 2001;**158**: 1347–62.

35. Yang B, Zhao H, Kranzler HR, Gelernter J. Characterization of a likelihood based method and effects of markers informativeness in evaluation of admixture and population group assignment. *BMC Genet* 2005;**6**:50.
36. Kohler K, Bickeboller H. Case-control association tests correcting for population stratification. *Ann Hum Genet* 2006;**70**:98–115.
37. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 2003;**73**:1402–22.
38. Wang WYS, Barratt BJ, Clayton DG, *et al*. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;**6**:109–18.
39. Setakis E, Stirnadel H, Balding DJ. Logistic regression protects against population structure in genetic association studies. *Genome Res* 2006;**16**:290–6.
40. Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet* 2005;**1**:e32.
41. Koller DL, Peacock M, Lai D, *et al*. False positive rates in association studies as a function of degree of stratification. *J Bone Miner Res* 2004;**19**:1291–5.
42. Marchini J, Cardon LR, Phillips MS, *et al*. The effects of human population structure on large genetic association studies. *Nat Genet* 2004;**36**:512–7.
43. Freedman ML, Reich D, Penney KL, *et al*. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004;**36**:388–93.
44. Helgason1 A, Yngvadóttir1 B, Hrafnkelsson B, *et al*. An Icelandic example of the impact of population structure on association studies. *Nat Genet* 2005;**37**:90–95.
45. Hochberg Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 1988;**75**:800–02.
46. Benjamini Y, Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J Royal Stat Soc B Methodolog* 1995;**57**:289–300.
47. Zaykin DV, Zhivotovsky LA, Westfall PH, *et al*. Truncated product method for combining *P*-values. *Genet Epidemiol* 2002;**22**:170–85.
48. Neuhauser M, Bretz F. Adaptive designs based on the truncated product method. *BMC Med Res Methodol* 2005;**5**:30.
49. Dudbridge F, Koeleman BPC. Rank truncated product of *P*-values, with application to genomewide association scans. *Genet Epidemiol* 2003;**25**:360–66.
50. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004;**74**:765–9.
51. Wen-Chung Lee. Testing for candidate gene linkage disequilibrium using a dense array of single nucleotide polymorphisms in case-parents studies. *Epidemiology* 2002; **13**:545–51.
52. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. *Genetics* 1994;**138**:963–71.
53. Dudbridge F, Koeleman BPC. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet* 2004;**75**:424–35.
54. Wille A, Hoh J, Ott J. Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. *Genet Epidemiol* 2003;**25**:350–9.
55. Lazzeroni LC. Linkage disequilibrium and gene mapping: an empirical least-squares approach. *Am J Hum Genet* 1998; **62**:159–70.
56. Cordell HJ, Elston RC. Fieller's theorem and linkage disequilibrium mapping. *Genet Epidemiol* 1999;**17**:237–52.
57. Zhang X, Roeder K, Wallstrom G, Devlin B. Integration of association statistics over genomic regions using bayesian adaptive regression splines. *Hum Genomics* 2003;**1**:20–9.
58. Roeder K, Bacanu S, Sonpar V, Zhang X, Devlin B. Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol* 2005;**28**:207–19.
59. Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am J Hum Genet* 2003;**72**:351–63.
60. Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res* 2001;**11**:2115–19.
61. Kim S, Zhang K, Sun F. Detecting susceptibility genes in case-control studies using set association. *BMC Genet* 2003; **4**(Suppl. 1):S9.
62. Fijal BA, Kim LL, Buxbaum SG, *et al*. Predicting quantitative trait levels by modeling SNP interaction. *Genet Epidemiol* 2001;**21**(Suppl. 1):S608–13.
63. Fallin D, Cohen A, Essioux L, *et al*. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res* 2001;**11**:143–51.
64. Zhao JH, Curtis D, Sham PC. Model-free analysis and permutation tests for allelic associations. *Hum Hered* 2000;**50**: 133–9.
65. Schaid DJ, Rowland CM, Tines DE, *et al*. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002;**70**:425–34.
66. Zaykin DV, Westfall PH, Young SS, *et al*. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 2002;**53**:79–91.
67. Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003;**73**:1316–29.
68. Cordell HJ. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. *Genet Epidemiol* 2006;**30**:259–75.
69. Thomas S, Porteous D, Visscher PM. Power of direct vs. indirect haplotyping in association studies. *Genet Epidemiol* 2004;**26**:116–24.
70. Morris AP, Whittaker JC, Balding DJ. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet* 2004;**74**: 945–53.
71. Toivonen HT, Onkamo P, Vasko K, *et al*. Data mining applied to linkage disequilibrium mapping. *Am J Hum Genet* 2000;**67**:133–45.
72. Xiong M, Zhao J, Boerwinkle E. Generalized T^2 test for genome association studies. *Am J Hum Genet* 2002;**70**: 1257–68.
73. Fan R, Knapp M. Genome association studies of complex diseases by case-control designs. *Am J Hum Genet* 2003;**72**: 850–68.

74. Tzeng JT, Devlin B, Wasserman L, *et al*. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet* 2003;**72**:891–902.
75. Durrant C, Zondervan KT, Cardon LR, *et al*. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am J Hum Genet* 2004;**75**:35–43.
76. Durrant C, Morris AP. Linkage disequilibrium mapping via cladistic analysis of phase-unknown genotypes and inferred haplotypes in the Genetic Analysis Workshop 14 simulated data. *BMC Genet* 2005;**6**(Suppl 1):S100.
77. Li J, Jiang T. Haplotype-based linkage disequilibrium mapping via direct data mining. *Bioinformatics* 2005;**21**: 4384–93.
78. Molitor J, Marjoram P, Thomas D. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am J Hum Genet* 2003;**73**: 1368–84.
79. Waldron ERB, Whittaker JC, Balding DJ. Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol* 2006;**30**:170–9.
80. Yu K, Xu J, Rao DC, Province M. Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. *Ann Hum Genet* 2005;**69**:577–89.
81. Yu K, Martin RB, Whittemore AS. Classifying disease chromosomes arising from multiple founders, with application to fine-scale haplotype mapping. *Genet Epidemiol* 2004;**27**:173–81.
82. Yu K, Gu CC, Province M, Xiong CJ, Rao DC. Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. *Genet Epidemiol* 2004;**27**:182–91.
83. Chakraborty R, Weiss KM. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci USA* 1988;**85**:9119–23.
84. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003;**164**:1567–87.
85. Patterson N, Hattangadi N, Lane B, *et al*. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* 2004;**74**:979–1000.
86. Montana G, Pritchard JK. Statistical tests for admixture mapping with case-control and cases-only data. *Am J Hum Genet* 2004;**75**:771–89.
87. Hoggart CJ, Shriver MD, Kittles RA, *et al*. Design and analysis of admixture mapping studies. *Am J Hum Genet* 2004;**74**:965–78.
88. McKeigue PM. Prospects for admixture mapping of complex traits. *Am J Hum Genet* 2005;**76**:1–7.
89. Cook NR, Zee RYL, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med* 2004;**23**:1439–53.
90. Pociot F, Karlsen AE, Pedersen CB, *et al*. Novel analytical methods applied to type 1 diabetes genome-scan data. *Am J Hum Genet* 2004;**74**:647–60.
91. Zhang H, Bonney G. Use of classification trees for association studies. *Genet Epidemiol* 2000;**19**:323–32.
92. Huang J, Lin A, Narasimhan B, *et al*. Tree-structured supervised learning and the genetics of hypertension. *Proc Natl Acad Sci USA* 2004;**101**:10529–54.
93. Lunetta KL, Hayward LB, Segal J, *et al*. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet* 2004;**5**:32.
94. Bureau A, Dupuis J, Falls K, *et al*. Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol* 2005;**28**:171–182.
95. Ye Y, Zhong X, Zhang H. A genome-wide tree- and forest-based association analysis of comorbidity of alcoholism and smoking. *BMC Genet* 2005;**6**(Suppl. 1):S135.
96. Young SS, Ge N. Recursive partitioning analysis of complex disease pharmacogenetic studies. I. Motivation and overview. *Pharmacogenomics* 2005;**6**:65–75.
97. Zaykin DV, Young SS. Large recursive partitioning analysis of complex disease pharmacogenetic studies. II. Statistical considerations. *Pharmacogenomics* 2005;**6**:77–89.
98. Foulkes AS, DeGruttola V, Hertogs K. Combining genotype groups and recursive partitioning: an application to hiv-1 genetics data. *J Royal Stat So C, Part 2,* 2004;**53**: 311–23.
99. York TP, Eaves L.J. Common disease analysis using multivariate adaptive regression splines (MARS): Genetic analysis workshop 12 simulated sequence data. *Genet Epidemiol* 2001;**21**(Suppl. 1):S649–54.
100. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer-Verlag, 2001.
101. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;**37**:413–7.
102. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* 2005; **28**:157–70.
103. Clark TG, De Iorio M, Griffiths RC, *et al*. Finding associations in dense genetic maps: a genetic algorithm approach. *Hum Hered* 2005;**60**:97–108.
104. Wei C, Debashis G, Trivellore ER, Kardia S. Bayesian method for finding interactions in genomic studies. Technical report, The University of Michigan Department of Biostatistics, 2004.
105. George EI, McCulloch RE. Variable selection via gibbs sampling. *J Am Stat Assoc* 1993;**88**:881–89.
106. Yi N, George V, Allison DB. Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* 2003;**164**:1129–38.
107. Devlin B, Roeder K, Wasserman L. Analysis of multilocus models of association. *Genet Epidemiol* 2003;**25**:36–47.
108. Nelson MR, Kardia SL, Ferrell RE, *et al*. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001;**11**:458–70.
109. Ritchie MD, Hahn LW, Roodi N, *et al*. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;**69**:138–47.
110. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 2003;**19**: 376–82.

111. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 2003;**24**:150–7.

112. Bastone L, Reilly M, Rader DJ, *et al*. MDR and PRP: a comparison of methods for high-order genotype–phenotype associations. *Hum Hered* 2004;**58**:82–92.

113. Coffey CS, Hebert PR, Ritchie MD, *et al*. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. *BMC Bioinformatics* 2004;**5**:49.

114. Williams SM, Ritchie MD, Phillips JA, *et al*. Multilocus analysis of hypertension: a hierarchical approach. *Hum Hered* 2004;**57**:28–38.

115. deBakker PIW, Yelensky R, Pe'er I, *et al*. Efficiency and power in genetic association studies. *Nat Genet* 2005;**37**: 1217–23.