Association mapping: where we've been, where we're going

Dahlia M Nielsen[†] *and Dmitri Zaykin*

This paper provides a review of recent work in the area of marker-phenotype association studies, specifically as used for localizing – or mapping – genes affecting a trait of interest. We describe the basis of association mapping and discuss a number of the commonly used techniques. We have also included references to various papers that have evaluated the use of these methods.

Expert Rev. Mol. Diagn. 1(3), (2001)

CONTENTS

[†]*Author for correspondence*
North Carolina State University,
Program in Statistical Genetics,
Department of Statistics,
Raleigh, NC 27695-7566, USA
Tel.:+1 919 515 2586
Fax: +1 919 515 7315
dahlia@statgen.ncsu.edu

KEYWORDS:
association mapping, case-control, genetic mapping, linkage disequilibrium, marker-phenotype association, transmission/disequilibrium

Genetic mapping based on linkage analysis relies on the principle that alleles at loci close together on a chromosome tend to be inherited together, as the probability of a recombination event occurring between them is related to the distance between them. Unlinked linked loci segregate independently. To estimate distance between loci, recombination events between them are counted as alleles are transmitted through pedigrees. An unknown gene affecting the trait of interest is mapped by discovering marker loci whose alleles are segregating with the trait of interest.

Association-based mapping relies on a similar principle, although with a different scope. These methods measure the degree to which alleles at a marker locus are associated with the trait of interest at a population level. The degree to which this occurs depends on the strength of linkage disequilibrium (LD) between the marker and the locus that affects the trait. If locus **A** has alleles *A* and *a* and locus **B** has alleles *B* and *b*, LD between alleles *A* and *B* (D_{AB}) can be written as:

$$(1)\ D_{AB} = P_{AB} - P_A P_B$$

Here, P_A is the probability of receiving an *A* allele (P_B for the *B* allele) and P_{AB} is the probability of receiving both an *A* and a *B* allele together from one parent. In other words, this quantity measures the difference between the rate alleles at different loci are actually inherited together *versus* the rate they would be inherited together just by chance. In this case of two alleles per locus, there are four LD terms (one for each pair of alleles), but:

$$D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}.$$

A common misperception about LD is that it refers to both linkage and LD. This has, unfortunately, been propagated in the recent human association mapping literature, leading to some inconsistency and confusion. The term was originally defined in the field of population genetics as far back as the early 1960s [1], taking on a form equivalent to EQUATION 1. By examining this equation (and other equations used to describe LD), it is clear that values regarding physical linkage, such as recombination rate, are absent. It is desirable to maintain the definition of LD in terms of a quantifiable statistical measure for which an explicit equation can be written. As it is not clear how this can be achieved in a general manner if physical linkage is also included in the measure, we maintain the traditional definition. This is consistent with current literature discussing measures [2] or estimates of LD [3]. It is also consistent with some of the earlier work on statistical tests of both linkage and LD [4].

We note that the term 'association' is often used interchangeably with or in place of the term 'LD'. Unfortunately, the term association can be ambiguous. It may refer to association between alleles at two loci (implying not only LD, but possibly other higher-order measures as well) or it may refer to association between the marker and a phenotype (which may or may not involve LD along with other factors). In this paper we attempt to be consistent and unambiguous, using the term association only when we mean marker-phenotype association and LD when we mean the specific quantity defined in EQUATION 1.

Another common misperception about LD is that it can be created by physical linkage. LD is actually created by evolutionary forces, such as mutation, migration, population admixture and founder effects and can exist between loci on different chromosomes. Once LD is created, recombination between loci does cause it to decay over time. Hence, once it has been created, LD tends to be maintained in the presence of tight linkage. If D_{AB}^{0} is the original amount of LD in a population, after G generations of random mating, the current level of LD is expected to be:

$$(2)\ D_{AB}^{G} = D_{AB}^{0}\ (1-\theta)^{G}$$

Here, θ is the recombination rate between the two loci. Therefore, the larger the recombination rate, the faster the decay. By looking at a population at generation G, it is expected that loci that exhibit stronger LD are closer together than those with weaker LD. This is the basis of association mapping. Following the LD pattern along a map should indicate the closest locus to the one being compared.

While in theory this is a reasonable strategy, in reality there are a number of caveats. Foremost is the fact that this strategy relies upon a reasonably restricted population history. Specifically, an event must have occurred to create an initial level of LD in the population, but since that event sufficient time has passed so that LD has decayed with recombination. In addition, no ongoing events are occurring in the population to maintain or create LD anew. A recent admixture event can create genome-wide LD, which is not conducive to inferring proximity between loci until many generations of random mating have occurred. If the admixture is ongoing, genome-wide LD patterns may continue to exist. In addition, mutational events that give rise to new alleles can occur at very different times for different loci. LD may have completely decayed between two adjacent loci, yet be newly created between these and a more distance locus. Selection may also maintain LD over time between distant loci.

Even if population history is such that LD is expected to behave as desired overall (EQUATION 2), large variations due to the stochastic nature of population history exist for any population of finite size [5,6]. Therefore, it is not unlikely that LD may differ substantially from its expected value, even in an ideal population. Of course, LD generally must be estimated from a population sample, so that these estimates also include sampling error.

Another point is that LD – as written in EQUATION 1 – is bounded by the allele frequencies at the loci involved. The maximum amount of LD possible between a locus and its close neighbor may be much smaller than that between the locus and a more distant site. In this case, it is quite possible that the more distant loci exhibit larger values of LD than the closer pair.

In spite of these potential problems, there is a lot of optimism regarding the use of LD in mapping genes and many methods have been developed for this purpose. Some of these methods have addressed the problems inherrent in using LD alone as a measure of distance by designing tests sensitive to recombination in addition to LD. Others attempt to model population structure in their tests, so that the tests are less sensitive to some of the problems that may arise.

Measures of LD

The difficulty in estimating LD is that usually only genotypes, not haplotypes, are available for examination, yet the haplotype frequency, P_{AB}, is necessary in order to estimate LD. One approach to get around this problem is based on the expectation-maximization (EM) algorithm. This method finds the haplotype frequencies that maximize the sample likelihood. Consider the following simple example. Suppose we observe a sample of six individuals typed at two diallelic markers. The first three individuals are double homozygotes, *AA/BB*, contributing six gametes of the type *AB* to the sample counts. The next two are *aa/bb*, contributing four gametes of the type *ab*. The last individual, however, is a double heterozygote, *Aa/Bb* and there are two possible pairs of haplotypes this individual may have inherited; *AB* with *ab*, or *Ab* with *aB*. Taking into account the first part of the sample, the EM algorithm assigns an extremely low probability to the second (*Ab* with *aB*) arrangement because these types of gametes have not been observed among unambiguous individuals. To make such calculations, it is necessary to assume Hardy-Weinberg equilibrium (HWE), which implies that two-marker genotypes do not deviate from the products of haplotype frequencies. Although the EM algorithm is an iterative technique, Weir and Cockerham showed that two-marker haplotype frequencies can be obtained by looking at the final iteration, when the next and the previous likelihoods are evaluated at the same frequencies [7]. The resulting equation can be solved explicitly, revealing that there are biologically plausible situations when no real-valued roots exist. This argues against blind application of the EM algorithm. Weir and Cockerham further suggest routine use of a composite genotypic disequilibrium method that does not assume HWE and that provides an unbiased estimate of LD under HWE.

Once an estimate of D_{AB} is obtained, there is a problem of its dependency on the allele frequencies, P_A and P_B. Many measures that attempt to standardize LD have been proposed. Devlin and Risch give an overview of five of these measures in the context of fine mapping [2]: the correlation coefficient Δ, Lewontin's D', the robust formulation of the population attributable risk δ, Yule's Q and Kaplan and Weir's proportional difference d under the assumption of initial complete disequilibrium between disease and marker loci, concluding that δ is the superior measure followed closely by D'. Guo confirmed these findings by calculating expected values of these measures [8], in contrast to the simulation approach of Devlin and Risch [2]. Morton *et al.* derive a measure of LD on a basis of population genetics argument [9]. Their measure, ρ, arises as the association probability and is algebraically equivalent to $|D'|$.

Certainly, the ranking of the measures reflects the specific goal of relating LD with the recombination probability. A different question, such as the optimal density of the markers, may result in a different ranking. Here Δ may score well, as it is the most related to the χ^2 statistic of the single-marker association test.

LD as a tool for mapping

As mentioned above, to estimate LD directly, it is necessary to understand something about both loci being examined (such as haplotype and allele frequencies). When one of the loci of interest is an unknown gene that we would like to map, alleles or genotypes at that gene are generally not available for examination. In this case, it is still expected that LD can be useful for indicating proximity between the putative gene and the marker locus being examined. However, the degree of LD between the gene and a marker must be determined indirectly. In association mapping, this is done by substituting phenotypes for genotypes at the gene. Instead of estimating correlations between alleles at both loci, relationships between the marker alleles and the phenotype are examined.

Examining LD indirectly by use of substituting phenotype for genotypes in this way has consequences for mapping. In doing this, the manner in which the gene acts on phenotype becomes confounded with actual LD [10,11]. This is intuitively obvious by considering the size of effect the gene has; genes with large effects on the phenotype should be much easier to locate than those with weak effects. This confounding of genetic effects and LD can create some unexpected patterns in marker-phenotype associations, which can cause added difficulties in association mapping studies [10,11].

Case-control tests for discrete traits

A very straightforward and popular method for detecting marker-trait associations for discrete traits – those in which an individual can be classified as either expressing the trait or not expressing the trait – is the case-control test. A common discrete phenotype of interest is whether or not an individual is affected with a disease under study (affected individuals are the cases, unaffected individuals are the controls). In the case-control test, individuals from both phenotypic categories are sampled in somewhat similar proportions. They are then genotyped for the markers or candidate loci of interest. Allele or genotype frequencies from the two groups are then compared to see if they differ.

Case-control tests are sensitive to overall population LD between the marker and a locus affecting the trait and to the effects of the gene. As was previously discussed, LD can exist in a population between loci that are quite far apart or even on different chromosomes. Due to this, strong marker-trait association is not necessarily evidence for proximity between the marker and a gene affecting the phenotype. In addition, it is possible to detect association even if there is no genetic component to the phenotype. This can occur when heterogeneity within the population exists. If the population to be sampled consists of several subpopulations and the prevalence of the trait in question differs between these subpopulations, so-called 'spurious associations' may be detected. In this case, individuals from the subpopulation with the higher prevalence of the trait are more likely to be selected to be cases, whereas subpopulations with lower prevalences are preferentially selected to be controls. If marker allele frequencies also happen to differ between subpopulations, these differences will be seen between the cases and controls. Different prevalences of the phenotype among subpopulations can result from nongenetic environmental factors, so that evidence of marker-phenotype association does not necessarily imply LD, let alone proximity.

Case-control tests for quantititive traits

Often the phenotype of interest in a mapping study is quantitative in nature. Without seperating individuals into discrete categories, the usual case-control test cannot be performed. Instead, analysis of variance (ANOVA) can be performed, with phenotype as the dependent variable and genotypes as the independent variables [12]. Other variables of interest, such as race, sex or age can also be included in the model. Since the number of genotypes at a locus is a function of the number of alleles squared, as the number of alleles grows, the number of genotypes becomes very large. This could lead to a problem with sparse data. Page and Amos discuss an allele-based ANOVA test, along with proposing various sampling strategies to try to increase power [13]. These sampling strategies involve genotyping individuals only from the top and bottom tails of the phenotype distribution. They compare the different ANOVA methods to one another and to various transmission/disequilibrium tests (TDT) [14]. Nielsen and Weir discuss regression of phenotype on genotype and ANOVA in the context of a general genetic model [10].

TDT for discrete traits

To get around the problem of spurious associations that can arise when using case-control tests, methods using family-based controls have been proposed. The most popular of these are the TDT-type tests. These methods are sensitive to non-zero LD and physical linkage between a marker and a gene affecting susceptibility. The original TDT was proposed by Spielman *et al.* for mapping disease susceptibility loci [4]. In this test, affected individuals and their parents are sampled and genotyped. By selecting affected offspring only, susceptibility alleles at the gene should be enriched among the sample of offspring. If there is LD between the gene and the marker in the parent generation, certain marker/susceptibility haplotypes will be more common than expected by chance. With no free recombination to separate these alleles as haplotypes are transmitted from parents to affected offspring, the marker alleles in LD with the susceptibility alleles will be transmitted more frequently to the affected offspring. The TDT assesses the transmission rates in the sample *versus* what would be expected by chance. Since homozygous parents can only transmit one allele type, these parents do not add to the power of the test and in most TDT, are not used. Spielman and Ewens extended the test to consider multiple alleles at the marker [15].

The TDT was originally proposed as a test of linkage in the presence of association; that is, if there is evidence that the phenotype is correlated with certain marker alleles, the TDT can be used to test if there is also linkage in the genetic region. In this case, the TDT can be performed using all affected offspring in a nuclear family. However, if there is linkage between the marker and a gene affecting phenotype, performing the TDT with multiple offspring as a test of LD is not valid. This is because in the presence of linkage, transmissions to siblings are not independent. As a test of LD in the presence of linkage or as a test for linkage and LD, only one affected offspring per nuclear family can be used. As this leads to a loss of information if multiple offspring are available, Martin *et al.* proposed a valid test of linkage and LD using all affected offspring in a nuclear family [16].

As an allele-based test, the TDT is sensitive mainly to additive genetic effects [10]. Schaid proposed a series of tests based on log-linear models that are genotype-based tests [17,18]. While these tests have more degrees of freedom than the allele-based tests, they are sensitive to the full genetic effect of the locus. Depending on how the gene acts to affect phenotype, the tradeoff between degrees of freedom and sensitivity to nonadditive effects may be advantageous.

TDT using sib data

While the TDT-type tests have the advantage of being insensitive to population structure, they do require parents of affected individuals to be collected. For some situations, such as in late-onset diseases, it may be difficult or impossible to collect parents of affected individuals. Several tests have been proposed to avoid this problem, using discordant sibling pairs – pairs composed of one affected and one unaffected sibling – in place of parental data. In these tests, the number of alleles seen in the marker genotype of the affected sibling is contrasted to the number seen in the unaffected sibling. The S-TDT calculates the mean and variance of the given marker allele for each family, then sums over families to get an overall mean and variance [19]. These values are then combined in a test statistic that is approximately normally distributed. Spielman and Ewens show how this test can be combined with the usual TDT if both sibships and trios are available [4,19]. Boehnke and Langefeld propose a similar test, in which allele counts among affected and unaffected siblings are recorded in a 2xK contingency table (where K is the number of alleles at the marker) and a homogeneity test is performed *via* permutation [20]. Curtis also proposed a sibling-based test of association, which uses a likelihood ratio statistic [21]. Monks *et al.* performed a comparative study of these tests [22].

The DAT, S-TDT and Curtis' test are valid tests of linkage and LD only if a single affected and a single unaffected individual are chosen from a sibship [19–21]. If larger sibships are available, subsamples must be chosen to perform these tests and information from the discarded siblings is lost. Curtis recommended a sampling procedure that involves selecting the discordant sibling pair who exhibit maximally different marker genotypes [21]. To avoid the problem of subsampling, Hovarth and Laird proposed the sibship disequilibrium test (SDT) [23], which is a valid test of linkage and LD using all discordant sibpairs within a sibship. The test compares the average number of alleles in affected siblings in a sibship *versus* the average for unaffected siblings in the sibship. The test statistic is evaluated using a nonparametric sign test.

In many cases, it is possible to reconstruct the genotypes of the parents if multiple offspring are available, though this must be performed with care to avoid bias [24]. Several methods have been proposed to reconstruct information in an unbiased manner. The reconstruction-combined (RC)-TDT was proposed by Knapp for this purpose [25]. Another procedure whereby reconstruction is performed is the likelihood-based approach of transmit [26].

Of the tests discussed so far, each uses information from two-generation pedigrees only. Martin *et al.* proposed a method whereby multi-generational pedigrees could be used in a TDT-type test without discarding individuals [27]. Their test – the pedigree disequilibrium test (PDT) – considers all discordant sibling pair comparisons and all transmissions from heterozygous parents to affected offspring within a pedigree. An average of each of these measures is taken for each pedigree and these pedigree averages then contribute to the overall test statistic. In this way, bias due to relatedness of individuals within a pedigree is eliminated. This method was developed for diallelic markers, though a series of tests comparing one allele to the rest grouped into a single category is also offered.

TDT for quantitative traits

TDT for quantitative traits have also been developed, using various strategies. Allison proposed five TDTs for quantitative traits, differing from one another in sampling and testing strategies [14]. The first test proposed (Q1) utilizes a T-test to compare the phenotypic means of randomly selected offspring inheriting the allele of interest from heterozygous parents. Tests Q2, Q3 and Q4 utilize sampling strategies, in which only offspring falling in the tails of the phenotypic distribution are included in the test. For each of these tests, a problem arises when including offspring for which both parents are heterozygous. In this case, there are two transmissions (one from each heterozygous parent) that can be included in the analysis, so that the phenotype of the offspring is included twice in the analysis. However, one assumption of these tests is that observations are independent. Allison's fifth test, Q5 attempts to circumvent this problem by use of a regression-based approach [14]. Rabinowitz independently derived an equivalent test to Allison's Q1 test [28].

Both the tests of Allison [14] and Rabinowitz [28] are valid tests of linkage and LD only when a single offspring per nuclear family is included in the test; multiple offspring can be used to test for linkage in the presence of association. George provides a regression-based test of linkage in the presence of association that uses all members of a pedigree [29]. To avoid loss of data when a test of linkage and LD is desired, Monks and Kaplan [30] extended Rabinowitz's test [28] so that the correlation structure between sibs is taken into account and all sibs can be used. This can provide a substantial increase in power if nuclear families with more than one offspring can be collected. Using a variance-components approach, Abecasis *et al.* also provide a method whereby parents and multiple offspring can be used as a test for linkage and LD [31].

Quantitative trait TDTs using sib data

As the TDT for discrete traits was extended for the case in which parents are not available, Allison *et al.* proposed several sib-based tests of linkage and LD for quantitative traits [32]. Their first method is the use of a mixed-effects ANOVA, which models sibship as a random variable and marker genotype a fixed factor; phenotype is the dependent variable. Their second approach is based on a permutation test.

Case-control tests for structured populations

While family-based TDT are able to circumvent the problems of population structure, they do come at a price, namely the need to collect and genotype family members. Another approach is to sample unrelated cases and controls, but to try to model population structure into the association test. Devlin and Roeder demonstrate the effect of population structure, including stratification and inbreeding on the variance of the usual case-control test statistic [33]. They show that if the overall population is not in HWE, the usual allele-based case-control test is not valid. The Armitage trend test can be used instead, as they show that it is better able to account for Hardy-Weinberg disequilibrium (HWD) [34]. As they mention, this is appropriate for an additive genetic model and ignores any nonadditive genetic effects. They do not mention that the genotype-based case-control test is an appropriate test in a homogeneous population, whether or not HWE holds. Schaid also discusses genotype-based tests under a log-linear model, which are also sensitive to nonadditive effects and should be resistant to HWD [17]. These types of tests assume that individuals in the population are independent, which is not generally the case in a structured population. Devlin and Roeder discuss the effects of nonindependence of individuals on the variances of the usual case-control test statistic and the Armitage trend test and show that it is inflated in this situation [33]. They propose an alternative method for testing for marker-disease susceptibility using a random sample of cases and controls. This method uses data from multiple markers outside the region of interest in order to gain information on the variance inflation due to population structure. This information can then be used to construct various tests for association. Their primary test is Bayesian, but they also provide a frequentist test.

Pritchard *et al.* offer a different solution for accommodating population structure in a case-control setting [35,36]. Their test also uses information from a number of markers outside the region of interest; these are used to infer the underlying structure of the population [35]. Once this has been done, this information is used to detect associations between the trait and marker within subpopulations [36]. The method uses a Markov-chain Monte Carlo technique to estimate the proportion of an individual's genome that originated from the contributing subpopulations, as well as the number of these subpopulations and the allele frequencies within each one. Once this has been done, these proportions are used in testing for dependencies between marker alleles and phenotype.

Both the methods of Devlin and Roeder [33] and Pritchard *et al.* [36] may offer potential power increases over the TDT-type tests for the same number of individuals genotyped.

Multiple testing

The multiple testing issues in disease association mapping are particularly complex. There is often substantial correlation between test statistic values along the map induced by LD between genetic markers. Lander and Kruglyak define pointwise and genome-wide significance levels to distinguish between probabilities of "large" values of a test statistic at a given location and anywhere in the genome scan [37]. The genome-wide significance levels relate to the notion of the family-wise error rate (FWER). For a family of hypotheses $\{H_1,\ldots,H_k\}$, the FWER is defined as:

$$\text{FWER} = \Pr(\text{reject one or more } H_i,\ i \in \text{S} \mid H_i,\ i \in \text{S are true})$$

Here, **S** is any subset of **K**.

Multiple testing methods divide into two extremes. In one group are the methods that make statements about all the individual null hypotheses $H_1,\ldots,H_L$. In the other group are the ones that are concerned with the global null hypothesis ($\cap H_i$): that some or all of the individual null hypotheses are true but without specifying which ones. It is traditionally required that tests concerned with individual hypothesis control the FWER. Westfall and Young argue that it is also desirable that the FWER is controlled in the strong sense [38]. Specifically, for the significance level α, a multiple testing procedure controls the FWER in the strong sense if $FWER \leq \alpha$ for all subsets of null hypotheses, regardless of which of them are true. The distinction between the strong and the weak control is thus only the issue in the situation of the 'partial null hypothesis', i.e., when some of the null hypotheses are false. A common method is the single-step Bonferroni technique. It rejects H_i if the p-value from testing this hypothesis, p_i, is less than or equal to α/L. Several step-wise procedures have been proposed. A method of Holm is an example of a step-down procedure [39]. It is based on the Bonferroni inequality and the min p statistic. A method of Hochberg is a further improvement and a sequential adaptation of the Bonferroni technique [40]. Hochberg's method is to order p_i's, start with $i = L$ and once $\text{p}_j \leq \alpha/(L\text{-}j\text{+}1)$, then reject all H_i for $i \leq j$. Note that the 'effective' number of tests, L^*, is expected to be smaller than L due to the positive correlation between test statistics. Lander and Botstein assume the statistic values are jointly multivariate normal and use the Ornstein-Uhlenbeck diffusion to compute L^* [41]. Westfall *et al.* discuss how PROC MULTTEST of SAS/STAT® can be used to obtain values of L^* exactly [42].

Benjamini and Hochberg propose to control the 'false discovery rate', or the expected proportion of false rejections [43]. Benjamini and Hochberg's procedure (BH) is to order p_i's and if $p_j \leq j\alpha/L$, then reject all H_i for $i \leq j$. This error rate is equivalent to the FWER when all hypotheses are true.

Clearly, the BH method controls the FWER in the weak sense. However, the gain in power to detect true effects can be substantial. The BH method is most appropriate when the number of true effects is known to be large, a moderate amount of the false-positive results can be tolerated or if a follow-up study is anticipated [44,45]. Benjamini and Hochberg state that "a desirable error rate to control may be the expected proportion of errors among the rejected hypotheses, which we term the false discovery rate (FDR)" [43]. It should be stressed, however, that the BH method controls FDR unconditionally and hence does not provide information about the proportion of false-positive results for any particular experiment. Conditional control would require knowledge of the unknown number of true null hypotheses as well as the power of the test to detect true effects.

'Combination' methods, such as of Fisher, Stouffer *et al.*, Edgington, Simes, Zaykin *et al.* are testing the global null hypothesis, $\cap H_i$ [46–50]. Once the hypothesis is rejected, it is possible only to conclude that one or more H_i's are false. This is somewhat less so with the Fisher's method, since it is disproportionally influenced by the tests resulting in small *p*-values [51].

The closure principle of Marcus *et al.* provides a bridge between global and individual tests [52]. This procedure considers all possible combination hypotheses obtained *via* the intersection of the set of individual hypotheses of interest. If an individual hypothesis and all intersections that contain it as a component are rejected by a global test, then the closure principle allows to reject the given hypothesis, strongly controlling FWER.

It has been suggested that flanking markers around significant tests must also show some significance [53]. In the context of linkage analysis, Goldin *et al.* suggested that *p*-values should be averaged across certain genetic distances [54]. Terwilliger *et al.* argued that in general, peaks tend to be wider around true positives [55]. Siegmund looked at the problem from the viewpoint of smoothing [56]. Zaykin *et al.* suggested using global tests for a similar purpose, exploiting their feature of taking a set of several moderately small *p*-values and use them to reinforce one another [50]. Thus, it is possible to produce either a more powerful test or provide a better peak definition.

Haplotype-trait association tests

As noted above, combination methods provide a way of taking into account several markers to produce a more powerful test and facilitate localization of genes of interest. This approach operates on marginal effects of single markers. It is possible, however, that using joint frequencies of alleles residing on the same haplotypes may be more successful.

The potential benefit in using haplotypes is 2-fold. First, a set of alleles at single markers forming a haplotype may itself account for the variation in the trait of interest [57,58]. Second, even if a single polymorphism affects the trait, surrounding markers form haplotypes that may be in much higher LD with that polymorphism as compared to individual markers. There has been much controversy around the issue of haplotype utility. Analytical power calculations of Akey *et al.* suggest that haplotypes can significantly improve power and robustness of association mapping [59]. On the other hand, simulation studies by Long and Langley [60] and Kaplan and Morris [61], found single-marker tests at least as powerful as haplotype-based tests. A particular issue of concern is that all three studies assumed that haplotypes are observed directly, so that the haplotype frequencies are readily available. Currently, this is rarely the case. Most often only single-marker genotypes are available and only individuals that are heterozygous at no more than one marker can unambiguously be assigned a pair of haplotypes. This situation (gametic phase ambiguity) requires estimation of frequencies by missing data techniques, such as the EM algorithm, increasing the variance of the test statistic (for details of the EM in the context of haplotype frequency estimation see [62]). Future studies are needed to define optimal tests for haplotype-trait association, taking into account the balance between the increase in degrees of freedom of haplotype-based tests that results in power loss and stronger linkage disequilibria of haplotypes with functional sites that provide increase in power.

Several studies concentrated on the issue of estimating haplotype frequencies *per se*. Fallin and Schork performed simulations sampling population frequencies of haplotypes from Dirichlet and scaled normal distributions and found pronounced accuracy of the EM algorithm for inferring haplotype frequencies on average, even when the HWE assumption does not hold [63]. Stephens *et al.* devised a Markov chain Monte Carlo method that allows reconstruction of haplotypes with very large numbers of markers [64]. By building in a general mutation model they also obtained substantial reduction in error rates of assigning pairs of haplotypes to individuals.

Expert opinion

While many techniques for association mapping have been proposed along with various methods to evaluate them, until now there have been reasonably few success stories. Due to this, there is an ongoing debate as to whether association mapping is or will one day be a useful tool. Certainly, it seems obvious that mapping by association cannot be as simple as following a string of *p*-values until the smallest one is reached, indicating the position of the gene. Instead, one needs to examine regions in which evidence for association seems to be strong, hopefully taking into account the correlations between tests. This is no small task, as it implies that enough markers have been examined in a region so that an association will not be missed, and at the same time that the multiple testing problem has been sufficiently addressed. At the current time, it seems unlikely that this can be done effectively on a genome-wide scope. Instead, with the current techniques, it is probably more reasonable to limit association studies to smaller regions that have been identified through linkage analysis or because they contain biologically relevant candidate genes. Even in reduced regions there is the concern over the number of polymorphic sites that must be examined. It is here that the biggest challenge lies. This becomes more relevant and more pressing as research into new methodologies for faster and cheaper genotyping continues to occur at a rapid rate.

In a recent commentary, Weiss and Terwilliger detail many problems of the current approach [65]. While they seem to offer little in place of the techniques they criticize, they do make a number of valid points. One of these is that in order to come close to addressing the types of complex genetic problems being studied and on the scale that is desired, better experimental designs will be necessary. This is undoubtedly true. One interesting question is how much of the progress that will be made in the future will be data-driven rather than hypothesis-driven. All in all, there is a long road ahead and the future is far from clear.

Five-year view

Currently, many association studies are carried out using markers that were discovered by examining a sample (sometimes very small) of random individuals, such as those available through public databases. This leads to various power-related issues, the obvious one being that the highly polymorphic markers are not likely to be associated with the phenotype of interest. On the other hand, low-frequency 'relevant' polymorphisms – if they are even observed – can bring down statistical power considerably. Advances in computational and molecular technology will soon make it possible to obtain sequences of entire candidate genes for an individual. This will provide an entirely new way of dealing with the marker selection problem. Sets of markers will be selected in a phenotype-specific manner out of all markers in a collection of candidate genes. Markers or marker combinations that yield maximum allele or haplotype frequency differences between levels of the phenotype will be selected by data mining techniques for the consequent analysis.

Acknowledgments

This work was supported in part by NIH Grant GM 45344.

Key issues

- Association mapping can potentially be useful in determining the location of a gene affecting a trait of interest within a much smaller region than can be determined using traditional linkage analysis approaches.
- Care must be taken regarding the results of case-control association tests; significant associations for these types of tests do not necessarily indicate proximity of the marker to a gene affecting the trait. Instead, population structure, nonequilibrium conditions, or variability of disequilibrium may be causing unexpected false-positive results.
- Significant results found using tests based on family data, such as the transmission/disequilibrium type tests, can be indicators of proximity. However, care must be taken when determining whether multiple offspring from a pedigree can be included in the analysis.
- If family data are not available, some case-control tests have been developed to take population structure within the sample into consideration [36]. These tests are less likely to suffer unexpected levels of false-positive results.
- An issue of shared genealogy [33] causes dependencies among observations at the same level of response. It is not clear what the extent of this 'cryptic relatedness' in natural populations is. Methods of stratification control are applicable here.
- There is much to be gained by utilizing multiple testing procedures that take into consideration the correlations between association-based tests at nearby markers, as well as the discrete nature of the distribution of the test statistics.
- Debate over the utility of haplotyping techniques continues. Controversy exists even when haplotypes are determined experimentally, let alone when they must be inferred through statistical techniques. At the bottom of much of this debate, however, is the assumption that there is only a single polymorphic site within a gene that is contributing to phenotypic variation. In actuality, it is often found that variation in phenotype is due to combinations of polymorphisms at different sites within the gene and also in the promotor regions. For these cases, it may be necessary to examine entire haplotypes in order to dissect phenotypic variation.
- Better techniques for utilizing marker haplotype information are needed. Current haplotype/response association methods rely on assumptions, such as Hardy-Weinberg equilibrium in cases and controls and generally lack explicit biological models.
- Another unresolved issue is how to determine the density of markers needed in a region that will assure reasonable power to detect true marker-phenotype associations, but will not suffer from multiple testing issues. An important consideration in marker selection is the underlying phenotype model. It is not clear whether allele frequencies and pair-wise correlations between markers (or equivalently, two-marker haplotypes) are sufficient considerations or whether multiple-marker haplotypes with corresponding higher-order disequilibria are also important for marker selection.

References

Papers of special note have been highlighted as:
• of interest
•• of considerable interest

1 Lewontin RC, Kojima K-I. The evolutionary dynamics of complex polymorphisms. *Evolution* 14(4), 458–472 (1960).

2 Devlin B, Risch N. Linkage disequilibrium measures for fine-scale mapping. *Genomics* 29, 311–322 (1995).

• **Investigates properties of several LD measures with relation to recombination probability.**

3 Jorde LB, Watkins WS, Carlson M *et al.* Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. *Am. J. Hum. Genet.* 54, 884–898 (1994).

4 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.* 52, 506–516 (1993).

•• **Introduces the transmission/disequilibrium test for discrete traits, which provides a test for LD in the presence of linkage.**

5 Hill WG, Weir BS. Variances and covariances of squared linkage disequibria in finite populations. *Theor. Popul. Biolo* 33, 54–78 (1988).

6 Hill WG, Weir BS. Maximum likelihood estimation of gene location by linkage disequibrium. *Am. J. Hum. Genet.* 54, 705–714 (1994).

7 Weir BS, Cockerham CC. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 42, 105–111 (1979).

8 Guo SW. Linkage disequilibrium measures for fine-scale mapping: a comparison. *Human Heredity* 47, 301–314 (1997).

9 Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok P-Y, Collins A. The optimal measure of allelic association. *Proc. Natl Acad. Sci. USA* 98, 5217–5221 (2001).

10 Nielsen DM, Weir BS. A classical setting for associations between markers and loci affecting quantitative traits. *Genet. Res.* 74(3), 271–227 (1999).

11 Nielsen DM, Weir BS. Association Studies under General Disease Models. *Theor. Popul. Biol.* (2001) (In Press).

• **Discusses how the manner in which a gene acts to influence the phenotype can affect the power of association tests to detect that gene.**

12 Boerwinkle E, Charkraborty R, Sing CF. The use of measured genotype information in the analysis of quantitative phenotypes in man. *Ann. Hum. Genet.* 50, 181–194 (1986).

13 Page GP, Amos CI. Comparison of linkage-disequilibrium methods for localization of genes influencing quantitative traits in humans. *Am. J. Hum. Genet.* 64, 1194–1205 (1999).

14 Allison DB. Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* 60, 676–690 (1997).

•• **Provides several transmission/disequilibrium tests for quantitative traits. This allows tests of linkage and LD between the marker and a locus affecting the quantitative trait.**

15 Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* 59(5), 983–989 (1996).

16 Martin ER, Kaplan NL, Weir BS. Tests for linkage and association in nuclear families. *Am. J. Hum. Genet.* 61, 439–448 (1997).

17 Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. *Genet. Epidemiol.* 13, 423–449 (1996).

• **Provides a regression-based framework for performing transmission/disequilibrium-type tests in a manner that allows a genetic model plus other covariates to be included in the analysis.**

18 Schaid DJ. Likelihoods and TDT for the case-parents design. *Genet. Epidemiol.* 16, 250–260 (1999).

19 Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* 62(2), 450–458 (1998).

• **Provides a test of linkage and LD that does not require parents to be genotyped.**

20 Boehnke M, Langefeld CD. Genetic association mapping based on discordant sib pairs, the discordant alleles test (DAT). *Am. J. Hum. Genet.* 62, 950–961 (1998).

21 Curtis D. Use of siblings as controls in case-control association studies. *Ann. Hum. Genet.* 61, 319–333 (1997).

22 Monks SA, Kaplan NL, Weir BS. A comparative study of sibship tests for linkage and/or association. *Am. J. Hum. Genet.* 63, 1507–1516 (1998).

23 Hovarth SM, Laird NM. A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am. J. Hum. Genet.* 63, 1886–1897 (1998).

• **Provides a test of linkage and LD that does not require parents to be genotyped and can accomodate multiple siblings.**

24 Curtis D, Sham PC. A note on the application of the transmission disequilibrium test when a parent is missing. *Am. J. Hum. Genet.* 56, 811–812 (1995).

25 Knapp M. The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *Am. J. Hum. Genet.* 64, 861–870 (1999).

• **Describes a TDT in which missing data can be inferred in a nonbiased fashion.**

26 Clayton D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am. J. Hum. Genet.* 65, 1170–1177 (1999).

27 Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: The pedigree disequilibrium test. *Am. J. Hum. Genet.* 67, 146–154 (2000).

•• **Created a TDT that allows entire pedigrees to be used, including multiple siblings and multiple generations.**

28 Rabinowitz D. A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* 47, 342–350 (1997).

29 George V, Tiwari HK, Zhu X, Elston RC. A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am. J. Hum. Genet.* 65, 236–245 (1999).

30 Monks SA, Kaplan NL. Removing the sampling restrictions from family-based tests of association for a quantitative trait locus. *Am. J. Hum. Genet.* 66, 576–592 (2000).

31 Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* 66, 279–292 (2000).

•• **Provides a TDT for quantitative traits that can be applied to whole pedigrees.**

32 Allison DB, Heo M, Kaplan N, Martin ER. Sibling-based tests of linkage and association for quantitative traits. *Am. J. Hum. Genet.* 64, 1754–1764 (1999).

33 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 55, 997–1004 (1999).

•• **Discusses the effects of inbreeding and cryptic relatedness on case-control tests. Provides a framework by which case-control tests can be adjusted when these effects are present so that spurious assocations can be avoided.**

34 Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 11, 375–386 (1955).

35 Pritchard JK, Stephens M, Donnelly P. Inference of population structure using mulitlocus genotype data. *Genetics* 155, 945–959 (2000).

• **Develops a methodology whereby a random sample from a population can be examined to determine the degree of population structure and attempts to estimate various parameters describing this structure.**

36 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in

structured populations. *Am. J. Hum. Genet.* 67, 170–181 (2000).

•• **Incorporates parameter estimates describing the structure of a population** [35] **into a case-control test. This test is less susceptible to spurious associations due to population structure.**

37 Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* 11, 241–247 (1996).

• **Discusses significance levels and multiple testing adjustments in linkage scans.**

38 Westfall PH, Young SS. *Resampling-Based Multiple Testing.* Wiley, New York, USA (1993).

39 Holm S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat. 6,* 65–70 (1979).

40 Hochberg Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 75(4), 800–802 (1988).

41 Lander ES, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121, 185–199 (1989).

42 Westfall PH, Zaykin DV, Young SS. Multiple tests for genetic effects in association studies. In: *Statistical Methods in Molecular Biology.* Looney S (Ed.), (2001).

43 Benjamini Y, Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B-methodol.* 57, 289–300 (1995).

44 Weller JI, Song JZ, Heyen DW, Lewin HA, Ron M A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* 150, 1699–1706 (1998).

45 Drigalenko EI, Elston RC. False discoveries in genome scanning. *Genet. Epidemiol.* 14, 779–784 (1997).

46 Fisher RA. *Statistical Methods for Research Workers.* Oliver and Boyd, London, UK (1932).

47 Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM, Jr. *The American Soldier (Volume 1; Adjustment During Army Life).* Princeton Univ. Press, Princeton, USA, (1949).

48 Edgington ES. An additive method for combining probability values from independent experiments. *J. Psychology* 80, 351–363 (1972).

49 Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73, 751–754 (1986).

50 Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining p-values. *Genet. Epidemiol.* (2001) (In Press).

• **Discusses and compares tests for combining p-values.**

51 Rice WR. A consensus combined p -value and the family-wide significance of component tests. *Biometrics* 46, 303–308 (1990).

52 Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63, 655–660 (1976).

53 Juo SH, Beaty TH, Duffy DL *et al.* A comprehensive analysis of complex traits in problem 2A. *Genet. Epidemiol.* 14, 815–820 (1997).

54 Goldin LR, Chase GA, Wilson AF. Regional inference with averaged p -values increases the power to detect linkage. *Genet. Epidemiol.* 17, 157–164 (1999).

55 Terwilliger JD, Shannon WD, Lathrop GM *et al.* True and false positive peaks in genomewide scans: applications of length-biased sampling to genome mapping. *Am. J. Hum. Genet.* 61, 430–438 (1997).

56 Siegmund D. Is peak height sufficient? *Genet. Epidemiol.* 20, 403–408 (2001).

• **Investigates width versus height of linkage signals from the viewpoint of smoothing.**

57 Drysdale CM, McGraw DW, Stack CB *et al.* Complex promoter and coding region b2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl Acad. Sci. USA* 97, 10483–10488 (2000).

58 Joosten PHLJ, Toepoel M, Mariman ECM, Van Zoelen EJJ. Promoter haplotype combinations of the platelet-derived growth factor a-receptor gene predispose to human neural tube defects. *Nature Genet.* 27, 215–217 (2001).

59 Akey J, Jin L, Xiong M. Haplotypes *vs* single marker linkage disequilibrium tests': what do we gain? *Eur. J. Hum. Genet.* 9, 291–300 (2000).

• **Provides an analytical power study examining situations when haplotype-based tests have increased power.**

60 Long AD, Langley CH. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* 9, 720–731 (1999).

61 Kaplan N, Morris R. Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. *Genet. Epidemiol.* 20, 432–457 (2001).

• **Describes a study of spacial distributions of statistic values of single markers and haplotype-based tests for the case-control design. Relates the χ^2 noncentrality parameter to marker and disease characteristics.**

62 Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12, 921–927 (1995).

• **Describes the multilocus EM algorithm, along with sample size and LD considerations.**

63 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.* 67, 947–959 (2000).

• **Demonstrates the accuracy of the EM algorithm under a variety of scenarios, including Hardy-Weinberg disequilibrium at individual markers.**

64 Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978–989 (2001).

•• **Compares phase reconstruction methods and introduces a method with substantial reduction of errors in assigning pairs of haplotypes to individuals.**

65 Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? *Nature Genet.* 26, 151–157 (2000).

• **Discusses many of the shortcomings of current association-based mapping strategies. Provides a basis for examining current methodologies, offering a challenge to find ways to improve these techniques.**