Genetic Location of Heritable Traits Through Association Studies: A Review

D. Garcia, J. Cañon and S. Dunner*

Laboratorio de Genética. Facultad de Veterinaria, Universidad Complutense de Madrid. 28040 Madrid, Spain

Abstract: In the last few years association and linkage disequilibrium studies have come to play an important role in the search for the location of genes underlying certain traits, since linkage analyses provide less accurate estimations of the positions of the genes as the complexity and rareness of the traits increase, partly due to the difficulty of getting large and informative enough samples. These approaches have been proven to be able to narrow the distance between the expected site of the locus and the nearest marker, and to reduce sample requirements in terms of size and structure when compared to those needed for linkage studies to obtain evidence for a gene's involvement. On the other hand, the lack of robustness with respect to population history and structure makes them still a subject of constant research. The kind of sample, the analysis to perform, the approach of seeking association with a particular marker versus conducting a complete scan of a wide part of the genome, the type and number of markers used, the nature of the trait (discrete or continuous), and the underlying model of the disequilibrium are some of the different factors needed to be taken into account when considering association studies. Any of them would by itself justify an individual review, but it is our intention to provide an overall perspective of the different approaches available at the current time.

INTRODUCTION

One of the main concerns of genetics is the full understanding of the behaviour of genes. With the advancement of molecular techniques there has been a growing interest in characterizing the actual genes underlying a certain trait and the particular place where these genes lie throughout the whole genome. Traditionally, linkage analysis has been used as an important tool to find these locations, first with biochemical and protein markers [1], and later with molecular ones such as Restriction Fragment Length Polymorphisms (RFLPs) [2,3], and then Short Tandem Repeats (STRs), of which the most known are microsatellites, [4-7] and, recently, Single Nucleotide Polymorphisms (SNPs) [8,9]. However, as the traits studied become more rare and complex, the need for sufficient dense marker sets and a high number of informative individuals may make linkage a somewhat limited and even unsuccessful procedure [10-13]. To deal with this obstacle, in the last few years more attention has been paid to association and linkage disequilibrium (sometimes wrongly used as synonymous concepts) studies, since these approaches have proven to be more powerful for genes of small to modest effects [14,15], reduce sample requirements in terms of size and structure when compared to those needed for linkage studies to obtain similar significance and narrow the distance between the expected site of the locus and the nearest marker (see Table **1** in [16] for a comparison of mapping distances obtained by linkage and by association procedures).

On the other hand, the lack of robustness with respect to population structure makes them still a subject of constant research. In the present article we intend to review the existing strategies that have been developed within the last years in association analysis. After describing the underlying theoretical basis for association and linkage disequilibrium (LD), we will look at the methods for detecting association between markers and discrete traits (affected/non-affected type), first with samples of individuals drawn randomly from the population, and secondly with the use of family information (e.g.: family trios -father/ mother/affected child). Then we will explore the possibilities of genome scans using association approaches. Finally we will review the existing extensions to quantitative traits.

LINKAGE, ASSOCIATION AND LINKAGE DISEQUILIBRIA: THEORETICAL AND BIOLOGICAL BACKGROUND

The underlying basis of linkage is physical. At the genetic level, two genes are said to be linked when the proportions in which a parent produces recombinant and nonrecombinant haplotypes are different. In particular, they will be completely linked when a parent *only* segregates non-recombinant haplotypes. A necessary condition for linkage to happen is that the two genes are located on the same chromosome. It is not a sufficient condition, though, as there may be no linkage detected between two genes on the same

*Address correspondence to this author at the Laboratorio de Genética. Facultad de Veterinaria, Universidad Complutense de Madrid. 28040 Madrid. Spain; Tel: (91) 394 3765; Fax: (91)394 3772; E-mail: dunner@eucmax.sim.ucm.es

1389-2029/02 $35.00+.00

chromosome if they are situated in opposite extremes of it, since if sufficiently far apart they segregate independently.

How linked two genes are is measured by the *recombination fraction*, θ, defined as the probability that a parent produces a recombinant gamete. It ranges from 0 (complete linkage) to 1/2 (absence of linkage). The reason for having 1/2 as the highest value is fairly simple: let *A* and *B* denote two unlinked loci and A_1, A_2, B_1, B_2 some of their alleles. Then if an individual has an $A_1B_1|A_2B_2$ genotype, the probability of transmission of any of the four haplotypes A_1B_1, A_2B_2, A_1B_2 and A_2B_1 should be equal (and obviously equal to 1/4). But the only way in which $\theta \cdot 1/2 = P(A_1B_1) = P(A_1B_1) = P(A_1B_2) = A_2B_1 = (1-\theta)\cdot 1/2$ is with θ being equal to1/2.

Gametic disequilibrium, or gametic phase disequilibrium [17] is more a statistical rather than a physical concept, although it may indeed have physical implications. Two genes are said to be in gametic disequilibrium if the frequencies of their possible haploid genotypes do not equal the products of the frequencies of their respective alleles. For instance, if we consider loci *A* and *B* from the previous paragraph, let p_1, q_1, p_2 and q_2 be the frequencies of A_1, A_2, B_1 and B_2 alleles, respectively, and P_{11}, P_{12}, P_{21} and P_{22} those of A_1B_1, A_1B_2, A_2B_1 and A_2B_2 haplotypes, respectively. Then *A* and *B* would be in an equilibrium state if $P_{11} = p_1p_2$, $P_{12} = p_1q_2$, $P_{21} = q_1p_2$ and $P_{22} = q_1q_2$, and we say that they are in gametic disequilibrium in any other case.

In order to measure the possible disequilibrium between these two genes, the *disequilibrium rate* is defined as

$$D = P_{11}P_{22} - P_{12}P_{21},$$

or, equivalently, as

$$D = P_{11} - p_1p_2.$$

It is easy to see that after an initial disequilibrium D_0 in a certain population for which we suppose random mating and the same recombination rate between males and females, θ, the disequilibrium rate after *t* generations will have descended to (see, for instance, [18])

$$D_t = (1-\theta)^t D_0,$$

so the higher the recombination rate is, the faster the disequilibrium disappears.

It is not unusual to find (e.g. [8,19]) the terms "association" (or "gametic disequilibrium") and "linkage disequilibrium" used in a synonymous way, although strictly speaking they do not apply to the same concepts, being actually the second a particular case of the first. The noun "association" refers to the lack of randomness in the frequencies of the haplotypes for a certain couple of loci, or, equivalently, to the non-independence of the alleles of one locus with respect to the other at the time of determining haplotypes. It *is*, indeed a synonym for "gametic disequilibrium". When this disequilibrium is produced by the physical proximity of the two genes, i.e., they show association because they are linked, then we can properly talk of "linkage disequilibrium". This distinction should not be underrated, since as we will soon see, there are several other situations which can lead to a situation of gametic disequilibrium, the so called *spurious associations*.

Some of the various factors which can eventually produce an association are the following:

- *Founder effect*: an external individual from a different population introduces through his offspring a previously non-existing allele to the one under study, which will be therefore associated with the allelic configuration of the loci surrounding the one of interest.

- *Mutations*: this phenomenon is analogous to the previous one, except for the fact that the new allele is introduced in the population not by an external individual by means of reproduction, but due to the mutation of an existing allele in one of the population members.

- *Sudden changes in population size*: events like a bottleneck effect can induce the appearance of associations, since the individuals surviving the process do not necessarily represent the former population, with the subsequent changes in allelic and haplotypic frequencies.

- *Natural selection*: the pressure of selection towards favouring phenotypes favours in its turn specific allelic combinations, thus creating disequilibrium.

- *Genetic drift*: when dealing with finite populations, the effect of random genetic drift is capital, since it strongly affects the frequencies of the alleles.

- *Population structure*: probably the most hard to handle sources of spurious associations are admixed populations and population stratification. These are two frequent events that, if not dealt with care, easily mask the results by creating associations between unlinked loci which may be confounded with true linkage disequilibrium. The use of populations which result from the admixture of several others is therefore, in general, not recommended to perform association analyses. However, when admixing is well understood, we can take advantage of the fact that the admixture of populations itself generates disequilibria in order to detect linked genes. Briscoe *et al*. [20] proved that if we cross into an F_1 two populations for which the marker loci considered differ significantly in their allele frequencies and these loci are in Hardy-Weinberg and linkage disequilibria, then in the F_2 obtained by crossing the two F_1 the only disequilibrium that will remain will be that due to physical linkage, whereas the spurious one created in the original admixture procedure will have disappeared in the crossing process.

DETECTION OF DISEQUILIBRIUM AT THE POPULATION LEVEL

The common feature of all the procedures in this section is that they all are carried out by extracting random samples

of individuals from the population, whereas the other main group of methods takes as sample units nuclear families, consisting in one or the two parents and one or more members of their offspring. We will take a look first at the "classic" procedures, which are nothing but the plain application of standard statistical tests to this particular case, to move next to the more somewhat "sophisticated" tests specifically developed for association analysis.

Classic Procedures

These procedures basically look for association from the population by means of the analysis of contingency tables with chi-square statistics. We will not deal with quantitative traits here (deferred to last section), so we will assume we are looking for association of a marker with a dichotomous trait. For the sake of ease of notation, we will use disease-like terminology, hence we will denote *D* as the allele of the disease locus responsible for the affected status, which is the one we will be trying to associate with one of the alleles of the marker, and + as the allele responsible for the non-affected status. Let *n* be the number of individuals sampled. If the haplotypes for the marker (denoted *M*, with alleles M_1 and M_2) and disease loci were known, then we would display them as in (Table **1**).

Table 1. Contingency Table for Haplotype-Known Chi-Square Analysis

	D	**+**	**Totals**
M_1	n_{M_1D}	n_{M_1+}	n_{M_1}
M_2	n_{M_2D}	n_{M_2+}	n_{M_2}
Totals	n_D	n_+	2*n*

The usual Chi-square test would be applied to test for the null hypothesis of independence between the marker and the disease locus. In the case of not having enough observations in any of the cells (five or more per cell are required), Fisher's exact test or approximative methods should then be applied. Recent examples of the application of this method can be found in [21,22].

An equivalent way of looking at the problem which arrives at the same statistic is by performing a Chi-square goodness of fit test, which compares the observed values of haplotypic frequencies with those expected under the null hypothesis of nonexistence of disequilibrium. The data would then be presented as in (Table **2**).

Nevertheless, in most practical situations the actual frequencies of the haplotypes of the marker and the disease locus are not available, because it is seldom possible to ascertain the phase of the genotypes. In most cases we also cannot determine the genotype for the disease locus.

To overcome this problem, when this occurs the usual strategy is to sample from two groups in the population: the affected individuals, called *cases*, and the unaffected ones, called *controls*. We then look for associations with one of the genotypes of the marker, in which case we would display the data in a contingency table such as (Table **3**).

Table 3. Case-Control Haplotypic Contingency Table

	M_1M_1	M_1M_2	M_2M_2	**Totals**
Case	$n^A_{M_1M_1}$	$n^A_{M_1M_2}$	$n^A_{M_2M_2}$	n_A
Control	$n^U_{M_1M_1}$	$n^U_{M_1M_2}$	$n^U_{M_2M_2}$	n_U
Totals	$n_{M_1M_1}$	$n_{M_1M_2}$	$n_{M_2M_2}$	2*n*

Alternatively, we can try to find an association with the presence of a particular allele of the marker, instead of with the genotypes, in which case we would have something like (Table **4**).

Table 4. Case-Control Allelic Contingency Table

	M_1	M_2	**Totals**
Case	$n^A_{M_1}$	$n^A_{M_2}$	n_A
Control	$n^U_{M_1}$	$n^U_{M_2}$	n_U
Totals	n_{M_1}	n_{M_2}	2*n*

In either case the procedure to follow would be exactly identical to the previous ones: compute the Chi-square statistic and study the significance of the table to test the independence of rows and columns.

If we have extra information, such as the inheritance mode and penetrances of the disease gene, it is possible in some cases to obtain the haplotype frequencies for the marker and disease locus, in which case we would be able to apply more accurate tests than the case-control ones.

For a deeper insight into this classical Chi-square tests, the reading of [18,19] is recommended.

Table 2. Contingency Table of Observed Vs. Expected Values for Haplotype-Known Chi-Square Analysis

	M_1D	M_1D	M_2+	M_2+	**Total**
Observed values	n_{M_1D}	n_{M_1D}	n_{M_2+}	n_{M_2+}	2*n*
Expected values	$2n\,\hat{p}_{M_1}\hat{p}_D$	$2n\,\hat{p}_{M_1}\hat{p}_D$	$2n\,\hat{p}_{M_2}\hat{p}_+$	$2n\,\hat{p}_{M_2}\hat{p}_+$	2*n*

Specific Procedures

We will take a look here at the various approaches that have been proposed by several authors to look for association in samples drawn at the population level. This section could have been entitled "Analyses with likelihoods and alternative procedures", since it is mostly by devising and testing the appropriate likelihoods that new methods for population-based LD analysis are being developed. The other major trend in population-based analysis is the use of Chi-square tests applied to case-control sampling structures, supported by some authors (e.g. [12,23,24]), in spite of the awareness of the high susceptibility and lack of robustness of these methods with respect to population stratification. In general, these methods provide good results when knowledge of the population structure is sufficient to assure an ascertainment scheme which properly matches cases and controls to the existing strata, or which focuses on particular subgroups [13]. Obviously, the best situation is that in which the population under study has not gone under any artificial, spurious association-causing event. Then, simply comparing the frequencies of alleles in affected and nonaffected individuals turns to be a good enough procedure. An example of such a case is the pioneering paper in LD mapping of Hästbacka *et al*. (1992) [15]. They mapped the diastrophic dysplasia locus using the information provided by the disequilibrium detected in the Finnish population to apply Luria-Delbrück equations, originally developed for bacterial mutation studies, and adapted in this work to obtain recombinational distances from LD measures. Finally, we will take a look at other alternative approaches that have been proposed.

One of the most known and applied (see, for instance, [26,27]) of the likelihood ratio (LR) tests is the one proposed by Terwilliger [28]. His study was motivated by the necessity of getting more power than that obtained until then from the Chi-square tests when the marker locus is multiallelic (meaning with that more than two alleles), which is the case with, for example, microsatellites. The problem comes from the fact that, for a marker with m alleles, a $2\times m$-dimensional contingency table must be analysed, which provides a statistic distributed as a Chi-square with $m - 1$ degrees of freedom. Therefore, the higher the number of alleles, the more power is lost, since the increase in degrees of freedom comes from the loss of information of dealing with a higher number of alleles -remember that the underlying assumption up to now is that we are trying to detect an association of the disease allele in the disease locus with one of the alleles in the marker, so it should be easier to determine if one out of two alleles is associated than if one out of ten is.

The solution proposed by Terwilliger consists of a LR test in which a single parameter is introduced, which measures the proportion of excess of the potentially associated allele in the affected individuals with respect to the whole population. Suppose the disease allele D was associated when introduced in the population to allele i of a nearby marker. As generations pass by, recombinations make it possible to find other marker alleles in the same haplotype than D, but there will still be an excess of i alleles coupled with D. To quantify this excess, a parameter λ is defined as the proportion of increase of allele i in the affected chromosomes with respect to the population frequency of the rest of the alleles. For example, if allele i is found with probability 1/3 in the whole population and with probability 1/2 in the affected individuals, then.

$$\lambda = \frac{1/2 - 1/3}{2/3} = \frac{1}{4}.$$

According to this model, $P(i|D) = q_i = p_i + \lambda(1- p_i)$, and for each allele $j\neq i$, $q_j = p_j - \lambda p_j$, where p_k is the frequency in the population of allele k of the marker locus, for $k = 1,\ldots,m$, being m the number of alleles.

Furthermore, as $p_i = q_i p_D + r_i(1- p_D)$, where $r_i = P(i|+)$, r_i can be calculated as $r_i = p_i - \lambda(1- p_i)p_D/(1- p_D)$, and analogously, for each $j\neq i$, $r_j = p_j + \lambda p_j p_D/(1- p_D)$, hence the model is completely determined.

Therefore, given a sample in which there are X_j individuals with haplotype $j|D$ and Y_j with haplotype $j|+$ for each $j = 1, ..., m$, Terwilliger proposes the following as the likelihood function:

$$L_i(\lambda) = \sum_{j=1}^{m} q_j^{X_j} r_j^{Y_j},$$

where subindex i in the likelihood refers to the fact that marker allele i is assumed to be associated to D.

However, to avoid making assumptions on which one is the associated allele, each $L_i(\lambda)$ can be weighted by the frequency of each allele i –understood as the probability of allele i being the one initially associated to D–, so the following joint likelihood is proposed:

$$L(\lambda) = \sum_{i=1}^{m} p_i L_i(\lambda).$$

And the likelihood ratio test takes the usual shape:

$$\Lambda = -2\ln\left(\frac{L(\lambda=0)}{L(\lambda_{ML})}\right)$$

whose distribution is claimed [28] to be 1/2 of a $\mathbf{X}^2_1$, and being λ_{ML} the maximum likelihood (ML) estimator of λ, with $\lambda = 0$ as null hypothesis, since that is the case in which disequilibrium does not exist.

In order to assess the fitness of the Chi-square distribution to the proposed statistic under the null hypothesis some simulation studies were performed, which also helped to compare the power of this test with that of the usual contingency table-based ones. After varying the values of λ, p_i, or m, the main conclusions were that the higher the

number of alleles, the more conservative the Chi-square assumption appeared to be, and that the power of this test was generally higher than that of the independence test. In addition, the Chi-square test power decreases rapidly as the number of marker alleles increases, while the proposed test is much more stable [28].

The test can be easily generalised to the case of more than one allele of the same marker associated to the disease allele, although Terwilliger suggests that it is not worth to consider more than two associated alleles.

Another extension can be made to analyse simultaneously multiple loci. It is based in modelling the decay of the parameter λ as a function of θ, the recombination rate. Since the physical situation of the markers is known, then for a given point inside the markers map, the recombination rate of this point with each of the markers is a known constant, so the likelihood can be maximised for the parameters of the function converting θ into λ and for the physical position along the marker map. The value of the likelihood in the ML estimator of the parameters can then be compared in the usual way with the value under the null hypothesis (λ=0) and form the LR statistic.

Multipoint likelihoods are in general very useful tools to perform fine mapping of traits, because they are usually dependent on and thus take advantage of parameters related to physical position along the chromosome. Consequently, the maximisation of the likelihood provides ML estimations and confidence intervals of positions or recombination rates of the disease locus. The likelihood can also be plotted against map position to obtain a graphical overview of its variation and its maximums. In particular, Terwilliger applied his method to the data related to cystic fibrosis (CF) reported by Kerem *et al*. [29], and obtained an estimation of the position of the CF gene only 0.16 cM away from its predicted position according to [29].

In spite of these seemingly good results, Terwilliger's method has a certain number of weak points. For example, Devlin *et al*. [30] noted the lack of treatment of evolutionary variances in this method. Is conservativeness and departure of the assumed asymptotic distribution have recently been addressed, and alternative models based on mixture distributions and posterior weights on the likelihood have been proposed [31]. We believe that another arguable aspect can be the assumption of independence between marker loci; that is, the fact that association of a certain allele to the disease allele does not induce an increase of the probability of some other neighbouring locus' alleles of being also associated. This condition is necessary to compute the multipoint likelihood in the way it is done, but it is not clear that it holds easily. Apart from that, we think that another point of discussion arises about the appropriateness of the parameter λ from the way it is defined: λ is dependent on the associated allele, so it does not seem rather natural to compute the whole likelihood for a given marker with just one only λ. Since as all the probabilities involved depend on it, the conceptual gain we obtain from weighting the partial likelihoods of each allele, thus taking all of them into account, is in some sense lost when we calculate the probabilities in each partial likelihood based upon the same parameter. It would seem reasonable, therefore, to consider an *m*-dimensional, *m* being the number of alleles of the marker locus, λ-vector, and compute the likelihood of allele *i* with its corresponding λ_i, to finally obtain the whole likelihood for the locus depending on *m* parameters. These ideas, however, are still under study and deserve further investigation.

Another interesting likelihood method was the one proposed by Xiong and Guo [32]. Their approach is somewhat more elaborated that Terwilliger's. They consider a more general situation in which factors such as population growth or the fact that the marker alleles frequencies vary with time are taken into account. They also note that the use of microsatellites needs to model the recombination rate. In this respect, they assume a stepwise mutation model, in which a certain allele only mutates to its immediately superior or immediately inferior (in size) allelic state. The population assumptions include nonoverlapping generations, random mating and nonexistence of population substructures, although on the other hand their modelling of the growth is not limited to exponentially growing populations, as other authors do [25,33].

Based on the same multinomial modelling of the sample allele counts as Terwilliger, for a given time *t*, the probability of obtaining the observed allele counts for the disease chromosomes ($k_{1_d},\ldots,k_{m_d}$) would be

$$f\left[k_{1_d},\ldots,k_{md} \mid P(t)\right] = \frac{k_d!}{\prod_{i=1}^{m} k_{i_d}} \prod_{i=1}^{m} p_{i_d}(t)^{k_{i_d}}$$

where $P(t) = \left\{p_{id}(t)^{k_{id}}\right\}_{i=1,\ldots,m}$ are the marker allele frequencies from the disease chromosomes at time *t*, which follow a Wright-Fisher population genetics model [32]. Taking expectations in the above equation over the generations, the unconditional sampling distribution is obtained, which is proportional to what will be considered as the likelihood function at maximisation effects:

$$L(\theta) = E\left(\prod_{i=1}^{m} p_{i_d}(t)^{k_{i_d}}\right),$$

which is a function of θ, the recombination rate, since each single $E\left(p_{id}(t)\right)$ is a function of θ.

This is not an easy expectation to compute, so approximation methods are needed. Initially, a Monte Carlo (MC) estimator of $L(\theta)$ was proposed [33], but Xiong and Guo [32] found it to be rather unsatisfactory. They proposed an alternate procedure consisting in the approximation of $L(\theta)$ by means of the first and second order Taylor series expansions of the marker allele frequencies $p_{i_d}(t)$, $i = 1,\ldots,m$. The evaluation of the first and second moments $p_{i_d}(t)$ of requires solving a system of differential equations which eventually leads to the desired values, with dependencies on θ, the disease allele frequency, the disease allele mutation rate, the marker alleles mutation rates and the initial values of $p_{i_d}(t)$.

The method is extended to perform a multipoint analysis, building what they called a "composite likelihood", since, as Terwilliger's, it assumes independence between allele marker frequencies at different loci to allow the formulation of the whole likelihood as the product of each marker's. Another particular feature of this method is that it includes as particular cases those in [25,28].

The method was applied to genes responsible of four genetic diseases whose position was already located: CF, Huntington's Disease (HD), Friedreich's ataxia and progressive myoclonus epilepsy. The general conclusions were that a first order approximation of the marker allele frequencies in the disease chromosomes is good enough, although second order approximations provide narrower support intervals. Also, its performance is generally better that that of the rest of methods to which it was compared, with errors in the estimation of the location ranging from approximately 10 kb, performing a multipoint analysis in the HD example, to 75 kb -multipoint analysis for the CF gene [32].

However, as the authors themselves mention, there are still some obstacles to overcome in the quest for a completely general and nonrestricted procedure: the treatment of locus heterogeneity, allele frequencies in the normal population, population substructure, incomplete penetrance, phenocopies, or nonrarity of the disease are open fields for further development of their technique. Rannala and Slatkin [34] also pointed out the need for modelling the frequency of the mutant chromosome, since Xiong and Guo [32] assumed it as constant through time, which in most cases is not valid.

Rannala and Slatkin's [34] approach is slightly different. They proposed a method with somewhat wider objectives: the estimation of either the recombination rate with a certain diallelic marker locus, the mutation rate or the age of the allele that was supposedly introduced in the population some generations ago.

They calculate the likelihood function by first obtaining the distribution of coalescence times for a sample of chromosomes descended from a nonrecurrent mutant ancestor and modelling the process of recombination and mutation, and base the likelihood on conditional probabilities of the number of MA_1 chromosomes after a certain j^{th} coalescence event conditioned on the number of MA_1 chromosomes after previous coalescence events, M being the disease allele and A_1 the marker allele supposedly associated with M. The resulting function to evaluate is:

$$P(Y_0 \mid \theta_1)=\sum_{Y_i=0}^{i}\sum_{Y_{i-1}=0}^{i-1}\cdots\sum_{Y_1=0}^{1}\int_{t_2=0}^{t_1}\int_{t_3=0}^{t_2}\cdots\int_{t_i=0}^{t_{i-1}} P(Y_0 \mid Y_i, t, \theta_2)$$

$$\times\prod_{j=2}^{i} P(Y_j \mid Y_{j-1}, t, \theta_3)\times P(Y_1 \mid p)\times P(t \mid \theta_4)\, dt_i \cdots dt_2$$

where Y_j denotes the number of MA_1 chromosomes immediately after the j^{th} coalescence event, i is the number of chromosomes in the sample carrying M, of which Y_0 carry allele A_1 as well, t_j the waiting time until i sampled chromosomes carrying M coalesce to j–1 ancestral chromosomes. The parameters upon which the likelihood depends are $\theta_1=\{u, v, t_1, i, f, \xi\}$, with $\theta_2=\{u, v, i\}$, $\theta_3=\{\{u, v, t_1, i\}$ and $\theta_4=\{ t_1, i, f, \xi\}$, where u is the probability of transition (by recombination or mutation) of MA_1 to MA_2 in an interval dt, which involves c, the recombination rate, μ, the mutation rate from A_1 to A_2, and p, the frequency of A_1 among nonmutant chromosomes, v is the analogous for MA_2 to MA_1, which involves c, p, and ν, the mutation rate from A_2 to $A1$, t_1 the number of generations since the introduction of the mutation, f is the fraction of the population sampled ($n/2N$, being n the sample size and N the number of individuals in the whole population), and ξ is a parameter which incorporates the effects of population growth and selection in the heterozygous individuals for M.

As it can be seen, the expression of the likelihood is rather complex and has too many terms in it, so the authors suggest a Monte Carlo estimation of $P(Y_0|\theta_1)$ based on simulations of the distribution of Y_i. Then, with the knowledge of some of the parameters involved in the computation, one of them can be left unknown to be estimated and confidence intervals obtained.

The method was applied to three examples with different results. In the diastrophic dysplasia (DTD) data [25], the estimation of the recombination rate of the DTD allele with the *CSF1R* locus was rather satisfactory and in agreement with other studies [32,33], finding a distance from the marker to the disease gene only approximately 10 kb larger than the real one. However, when dealing with the estimation of the ages of the alleles in the other two examples, the results were not so positive. In spite of this, the method seems interesting for practical association mapping, as it involves a large number of parameters of interest, and future improvements by developing more exact methods of estimation may lead to better results.

Some of the criticisms that can be made to their method are similarly applicable to other ones: nonrandom mating in general, and population substructure in particular are still lacking adequate treatment. Also, when the likelihood is used to build likelihood ratio tests, either for testing a noninfinite age of the mutation, or a recombination rate being different from zero, the null hypotheses are both in the boundaries of their corresponding parametric spaces, so the convergence of the statistic to the known Chi-square distribution is not well justified. In addition, the use of Monte Carlo simulations to approximate the likelihood, together with the use of estimations of the rest of the parameters instead of their real values, introduces a source of variation which is not properly taken into account.

Graham and Thompson [35] described a similar method to that of Rannala and Slatkin, in the sense that they first simulate a coalescent genealogy for a given sample of disease alleles to then calculate the likelihood of the observed sample through Monte Carlo (MC) estimations. Their approach consists basically in performing several MC realisations of backwards in time reconstructions of the coalescent ancestry and place by simulation of disease-

bearing haplotypes (or equivalently, recombination events), mimicking the disease ancestry and what they call the "recombinant classes".

A recombinant class is defined as a subset of the sample that is descended from a given recombination event [35], this is, a subset of sample individuals identical by descent at the marker locus. In this way they form a partition of the sample with good properties for recombination count purposes.

The likelihood, as a function of the recombination rate between the disease and the marker locus would take the form:

$$L(\theta) = P_{q,\theta,\Delta}(Y) = \sum P_q(Y \mid X)\, P_{\theta,\Delta}(X)$$

where q is the vector of population allele frequencies at the marker locus -assumed constant over time-, Δ is a bivariate demographic parameter including the time in generations at which the disease allele was present as a single copy in the population and the population growth rate, as a function of time, Y is the vector of allelic counts for the marker in the disease population and X is a vector of dimension equalling the highest size of a recombinant class with $X(i)$ being the number of classes of size i. To evaluate $L(\theta)$, values of X are sampled by simulation according to $P_{\theta,\Delta}(X)$.

To account for the variability on the likelihood estimation and obtain confidence intervals for θ, a parametric bootstrap approach is taken, and realisations of Y are generated under the maximum likelihood estimate to reestimate θ for each replication. When applied to a simulated example, this method provided good results by estimating a recombination rate of 0.005 for a marker with a disease locus for which the actual one was 0.006. The authors also note the nonvalidity of the Chi-square approximation of the minus twice the loglikelihood ratio to obtain confidence intervals since it assumes independence among marker alleles, but positive correlation increases as the recombination rate decreases [35].

The method was extended to interval and multipoint mapping. The interval mapping procedure is essentially the same as the single locus one, with the likelihood depending on k recombination rates instead of on one, being k the number of loci considered. Tested on a simulated example, it seems to behave fairly good –the disease location was estimated at θ = 0.000 from its true location [35]. As desirable, the multilocus analysis provides better results than the separate analyses of flanking markers in single-marker mapping, since the confidence interval obtained is narrower than those of the single-marker analyses. The approach for multipoint mapping is slightly different, since the concept of recombinant class has to be extended to include all the markers at the same time, but it has not yet been properly investigated.

In order to compare their method with other published ones, the authors analysed the DTD data [25] and found the results to be similar to those of previous studies [25,33,34], both in point and confidence interval estimation, and in good agreement with the real situation, obtaining an estimation of the distance of the marker to the disease 15 kb larger than the real one. When interval mapping was applied to the same data, similar estimations were obtained, but it has the advantage of providing indications of to which side of the marker the disease locus is more likely to be located.

As a general comment, Graham and Thompson's approach fills some of the gaps left in other studies, such as the consideration of multiallelic markers, the use of multimarker techniques or the flexibility of the demographic model. However, perhaps it may be arguable that too much simulation is involved in the calculus -Monte Carlo sampling of the recombinant classes plus bootstrapping to get confidence bounds on the likelihood estimates. In spite of their well developed analytical computation of the probabilities of the observed disease-marker haplotypes conditional on the recombinant classes, more real-data validation may be needed, since the method was only checked on a two isolated and well historically documented populations as the Japanese and Finnish with two favourable disease situations: one simulated by the authors and the other the DTD disease.

Morris *et al*. [36,37] have applied Bayesian analyses into association mapping by likelihood procedures. Their approach considers diallelic markers, since it is designed to be used with SNPs. In [36], they set an initial location for the disease locus and determine the expected allele frequencies of the markers at the left and right of that location, conditioned on the chromosome's ancestral state at the disease locus (this is, the property of Identity By Descent -IBD- with the founder chromosome), throughout the whole candidate region by means of independent Markov processes. These frequencies are used to estimate the expected allele frequencies in affected and unaffected individuals in the population. These, in turn, are put into the likelihood in the usual way. The whole log-likelihood takes then the following expression:

$$\ln L(data \mid x, \Gamma, p, \omega)_{TOT} = \ln L(data \mid x, \Gamma, p, \omega)_L + \ln L(data \mid x, \Gamma, p, \omega)_R$$

where L and R in the subindexes refer to the markers to the left and right, respectively, of location x, Γ is a vector of model parameters, which account for recombination, mutation, and phenocopies, among other things, p is the vector of observed allele frequencies and ω is a vector of ancestral indicators -ω_i = 1 if allele 1 of marker locus i is present on the founder chromosome and 0 otherwise. The "right" and "left-side" terms look rather familiar. The right-side term, for example, is

$$\ln L\ (data \mid x, \Gamma, p, \omega)_R = \sum_{i=1}^{R} \sum_{j=1}^{2} \sum_{P=A}^{U} n_{ij}^{P} \ln \left[\varphi_{ij}^{P}\right] + C_{R,}$$

where φ_{ij}^{P} is the expected allele frequency for allele *j* at locus *i* in the population with affection phenotype *P*, *A* means affected, *U* unaffected, *R* is the number of marker loci, and C_R is a constant related to the disease frequency in the population.

The final step is to use Markov Chain Monte Carlo (MCMC) methods, in particular a Metropolis-Hastings algorithm, to obtain posterior distributions for the model parameter estimates. As an example it was tested on the CF and HD genes data [29] with good results. The mean value of the posterior distribution for the distance of the mutation to its nearest marker was 96 Kb away from its true value, although the 99% confidence interval does not include it. When the dependence between case chromosomes is modelled, the mean of the posterior distribution improves and becomes 82 Kb away from its true value, this time being included in the 99% confidence interval. However, with this extra modeling, posterior distributions for the parameters involved are flatter in general, as a consequence of an increase in the variance, which produces wider confidence intervals. The results regarding the HD gene are similar, with the mean value of the posterior distribution being 0.1 Mb away from its true value.

The advantages of performing the analysis in a Bayesian framework are clear: one deals directly with a distribution of the parameter, instead of with a single estimator, either point or interval, and prior information about model parameters can be readily incorporated into the model. The model presented is more complicated than previous ones, since it takes into account factors like phenocopies or recombination rate heterogeneity across the candidate region. However, the results, although in accordance with other studies, do not seem to improve significantly the existing methods.

In [37], the procedure is identical once the likelihood is computed. They follow, however, a different approach to calculate the likelihood. First, it is separated in the part corresponding to cases and the part corresponding to control individuals. The likelihood of the cases is then calculated by considering an arbitrary genealogical tree, conditioning the likelihood on several parameters concerning the structure of the tree and then summing over the possible tree configurations. The MCMC methods are then applied to approximate the sum, sampling from the conditional distribution with a Metropolis-Hastings algorithm. When applied to the CF gene the method provides a slightly worse mean value for the location of the disease locus than the previous one, at a distance of 0.148 Mb from the real value, though in this case, it is included in the 99% confidence interval.

Several other methods do not make an explicit use of likelihoods for detecting association. For instance, Devlin and Roeder [38] developed a method known as Genomic Control in which a case-control design is analysed for diallelic markers, assuming an additive genetic model, by means of Armitage's trend test for genotypes,

$$Y^2 = \frac{N\left[N\,(r_1 + 2r_2) - R\,(n_1 + 2n_2)\right]^2}{R\,(N - R)\left[N\,(n_1 + 2n_2) - (n_1 + 2n_2)^2\right]},$$

where, if the locus considered has alleles A_1 and A_2, (Table **5**) shows the notation used.

Table 5. Table of Allelic Counts for a Genomic Control Test

	A_1 alleles			
	0	**1**	**2**	**Totals**
Case	r_0	r_1	s_2	R
Control	s_0	s_1	s_2	S
Totals	n_0	n_1	n_2	N

The innovation comes from realising and modelling how populational and genealogical phenomena, like population substructure or cryptic relatedness -this is, the fact that sample individuals, although randomly selected, may be related to some degree, this being more common amongst cases than controls. These events lead to an increase in the variance of the test statistic, modelled by a coefficient called the *variance inflation factor*, denoted by λ, which depends, among other variables, on the Wright's inbreeding coefficient, *F*. The meaning and value of *F* are context-dependent, since both matings between relatives and population substructure can vary the correlation between uniting gametes, which is what *F* measures.

The procedure then allows for the estimation of the inflation factor λ and the adjustment of the trend test for association of each locus with the disease. Certain assumptions must be made (similar mutation rates among loci under study, no strong subpopulation specific selection, and small variation of *F* across loci) in order to assure that λ is constant across loci.

For a particular marker *i*, under the hypothesis of linkage equilibrium with the disease and no subpopulation structure or cryptic relatedness, $Y_i^2 \sim X_1^2$. Allowing for extra variance, $Y_i^2/\lambda \sim X_1^2$. If the marker is associated, then Y_i^2 follows a mixture of chi-square distributions, one of them noncentered: $Y_i^2/\lambda \sim \varepsilon X_1^2(A_i) + (1-\varepsilon) X_1^2$ where ε is the prior probability of a given observation being associated, and A_i the noncentrality parameter. Specifying prior distributions for λ, ε, and A_i allows Bayesian inferences through a Gibbs sampling scheme to finally test whether marker *i* is associated or not.

The advantage of this Bayesian approach is that there is no need for a Bonferroni correction to test for association in a multilocus context. A careful choice of the priors, especially that of ε, provides better performance in many settings, while constraining the risk of false positives [38] -

actually, even when the test is performed in a non-bayesian framework, its results improve in many situations those of other tests, like the Transmission/Disequilibrium Test (TDT, [39], see next section for a description of the TDT), devised to take into account population structure [40]. To refine the location of the disease locus in a candidate gene approach, after an initial screen of the region of interest, those markers with highest posterior probabilities are investigated again with new values for ε. If necessary, a locus-by-locus scheme could be applied, with individual, locus-dependent, epsilons.

Simulations performed to compare the performance of this test versus the analogue one under a frequentist view, with its corresponding Bonferroni correction, suggest that the Type I error is small and quite stable for the Bayesian test, in contrast with the instability of the figures for the standard frequentist test. The conservative nature of both tests is also shown to increase with the number of loci analysed, to prevent a large number of false positives when a dense genome scan is performed.

Another test that seems to overcome the problems of population structure in case-control sampling schemes is the recently presented by Pritchard *et al*. [41]. They make effective use of previous methods to detect the presence of population structure in the population [42] and to differentiate the distinct subgroups, calculating the degree of belonging of any individual to each subpopulation [43]. First, substructure is tested by looking for association among a set of unlinked loci distributed throughout different chromosomes [42]. Then, if association is detected, it must be due to population structure, since the loci are chosen to be independent. Next, that structure is analysed with the method of Pritchard *et al*. [43]. They use MCMC procedures to estimate the number of subpopulations, the allele frequencies in each subpopulation and the value for each individual of the vector $q = (q_1,...,q_K)$, where K is the inferred number of populations and q_i, $i = 1, ..., K$, is the proportion of the individual's genome originated in subpopulation i.

Once the subpopulations have been determined, their method [41] is designed to test the null hypothesis of no association between alleles at a certain candidate loci and the phenotype of the trait *within* each subpopulation, instead of testing the presence of association in the whole population, as it is usually done. Furthermore, as a consequence of their procedure, since each subpopulation is assumed to be free of structures interfering with the marker-phenotype relationships, any association, if detected, would be likely due to physical proximity between the marker and the disease locus, i.e., to linkage disequilibrium.

Their procedure, named *STRAT* (STRuctured population Association Test), uses a likelihood ratio to test $H_0 \equiv$ subpopulation allele frequencies at the candidate locus are independent of phenotype vs. $H_1 \equiv$ subpopulation allele frequencies at the candidate locus depend of phenotype. The usual statistic, is employed, where P_0, P_1, and Q are the estimates of $P_0$and P_1, the population allele frequencies at the candidate locus under H_0 and H_1, and Q, the collection of vectors q representing the genetic backgrounds of the sample individuals, and C is the list of sample genotypes at the candidate locus. The probabilities of an individual bearing a certain allele are calculated as a weighted sum across the putative subpopulations of the frequencies of that allele in each subpopulations, weighted by the vector of genetic backgrounds of the individual. For example, under H_1, the probability of individual i, with phenotype $\Phi(i)$, of having allele j in its genotype would be:

$$\Lambda = \frac{P_1\left[C \mid P_1; Q \right]}{P_0\left[C \mid P_0; Q \right]},$$

$$P_1\left[c^{(i,a)} = j \mid Q, P_1, \Phi\right] = \sum_{k=1}^{K} q_k^{(i)}\, p_{kj}^{(\Phi(i))}$$

being $c^{(i,a)}$ the element of C corresponding to the a^{th} allele of i, Φ the vector of phenotypes for the individuals ($\Phi(i) \in \{0,1\}$), the proportion of genetic information of i coming from the k^{th} subpopulation, and $p_{kj}^{(\Phi(i))}$ the frequency of allele j at the candidate locus in subpopulation k among individuals with phenotype $\Phi(i)$ –under H_0, the frequencies are independent of the phenotype, so they would be denoted just as p_{kj}.

The significance of the obtained value for Λ is then approximated by a Monte Carlo simulation, in which a large number of values for C are generated under the probability model for the null hypothesis. The approximation is thus given by the proportion of times that the value of Λ for the generated Cs is higher than the observed one.

Simulation studies showed that the method performed reasonably well in different structure situations, varying from two discrete, distinct populations with the same allele associated in both of them at equal frequencies, to two admixed populations with different alleles associated in each of them at different frequencies. They indicate that in general the adjustment of the test to the theoretical significance level is fairly good, although sometimes slightly conservative. In any case it outperforms the classical χ^2 test, while obtaining similar results to those of the TDT. With respect to the power, the behaviour is notably better in "confusing" situations, i.e., those different from having the same allele associated with the same risk of disease in all the subpopulations, losing power with respect to the TDT in the latter case, but outperforming it by far when different alleles are associated in each subpopulation.

The STRAT method presents therefore several advantages in performing association mapping in structured populations. Since it does not pool the genetic information from the subpopulations, as others like the TDT do, but treats each one separately, it allows for different effects in different subpopulations. The fact that the data are collected from a case-control scheme and the non-necessity of devising complex ascertainment sampling schemes to take substructure into account are also good features of this method. However, the requirement for a large number of independent loci to test the presence of substructures and analyse them –heuristic approximations suggest more than 100 microsatellite loci, or an even larger number of biallelic

loci [41]– could make it unaffordable, due to genotyping costs in certain contexts, especially when only one or a few candidate marker loci are going to be tested. On the other hand, if there is a whole screen in which many markers are tested for association (SNPs, for example), those markers themselves can be used to detect and study the possible substructures. The novelty of the method leaves still some directions to explore, such as extending the procedure to analyse family structured samples, incorporate correlations between the ancestry of linked markers within individuals –which should improve the method [41]–, modelling the relative risks when the candidate locus is not the functional site itself, or studying in depth the errors consequence of the many simulation procedures involved.

Perhaps one of the most originals approaches, and apparently also a good one for fine mapping, is that of Lazzeroni [44]. It is based on the linkage disequilibrium measure δ, defined as

$$\delta = \frac{P(A \mid D) - P(A \mid N))}{1 - P(A \mid N))}.$$

Since it ignores confounding factors like population admixture, it was devised to be used to fine-scale map a disease for which linkage has been established to a group of markers in a certain area. After some approximations based upon Taylor series expansions, the parameter δ can be finally approximated by a polynomial on the physical location of the marker locus, x, as

$$\delta \approx, g(x) = \begin{cases} \beta_0 + \sum_{j=1}^{J} \beta_{1j} \left| x - \mu \right|^j, \text{ if } x \leq \mu \\ \beta_0 + \sum_{j=1}^{J} \beta_{2j} \left| x - \mu \right|^j, \text{ if } x \leq \mu \end{cases},$$

where J is the order of the Taylor expansions, μ is the location of the mutation site, and β_0 and β_{ij}, $i = 1,2$, $j = 1,...J$, are the coefficients of the polinomial.

The method can then be applied in steps: first, the parameter δ_i is estimated for each locus $i = 1,...,R$. The estimates have to be transformed to avoid skewness and make possible a normality assumption. Bootstrap procedures are then used to account for variability and estimate the variance-covariance matrix, and finally a generalised least squares (GLE) approach is used to fit the disequilibrium curve to the observed data and get an estimation of the location of the gene.

When tested with the CF data of Kerem *et al*. [29], the method placed the causative mutation at approximately 10 kb from its true location. In addition, all 80, 90 and 95% confidence intervals included the true location. This approach seems therefore to provide good results, but no further investigation has been made with other examples. It is also an interesting feature that it can be applied as well to family data, by adjusting the bootstrap scheme, although no comparative results have yet been produced. Further investigation should pay attention to multiallelic loci extensions, disease mutation at multiple sites and efficient algorithms to perform the whole process and be able to obtain empirical confidence intervals by replicating the estimation procedure, instead of recurring, as it was done, to somewhat inaccurate likelihood estimations.

The Problem of Population Choice

It has been agreed that in many if not all cases, populations do not match the optimal desirable conditions to perform a linkage disequilibrium study. Wright *et al*. [45] proposed up to five different ideal kinds of populations to more accomplish this, but concluded, as did Sheffield *et al*. [46], that genetically simplified isolates are more likely to properly host these studies than diverse continental ones under most assumptions. The Ashkenazi Jews [34,47], the Sardinians [48], some Taiwanese populations [49], or the Costa Ricans [27] are examples of populations used in the literature due to their isolation and historical background. But likely the most studied and quoted as a paradigmatic example of an isolated population is the Finnish (see for instance [22,25, 34,50-52]). However, recent studies seem to contradict the claim that population isolates are always the most optimal. Eaves *et al*. [53] provide data that genetic isolates like Finland and Sardinia will not prove significantly more valuable than general populations for linkage disequilibrium mapping of common variants underlying complex disease.

Instead of looking for alternatives to circumvent the problem of population structure, some have tried to take advantage of it. Briscoe *et al*. [20] showed how to detect association by means of admixed populations, and Laan and Pääbo [54] and Terwilliger *et al*. [55] used genetic drift generated linkage disequilibrium for gene mapping purposes.

DETECTION OF DISEQUILIBRIUM WITH FAMILY SAMPLES

The Use of Family Structures in Associ-ation Mapping and the Transmission/Disequili-brium Test (TDT)

It was not until recently that population-level procedures have more or less developed methods to solve the problem of population artifacts such as admixture or stratification, to avoid their influence in the analyses. We have seen how isolated populations seem to offer a possible solution, as well as how to use those effects to create a disequilibrium which can be subsequently used to map genes of interest. An alternative feasible solution to the problem of obtaining cases and controls free of the influence of the population structure was to get them, not from the whole population, but from within individual families. This solution is the one which has received most attention in the last several years. The most significant "early" tests were the AFBAC, which stands for Affected Family Based Controls [56,57], and Falk and Rubinstein's Haplotype Relative Risk (HRR) [58,59].

In the AFBAC test, for a given marker and a family, the four parental marker alleles are assigned to one of two

categories: transmitted to at least one affected offspring and not transmitted to any affected offspring. Then a standard χ^2 test is performed to look for association between the marker and disease locus. The HRR considers a sample of *n* single-child families, with the child in each family being affected. Of the 4*n* parental alleles at the marker locus, 2*n* will be transmitted to the child and 2*n* will not. The counts of the transmitted are compared to those of the non-transmitted through a χ^2 statistic.

Spielman *et al*. [39] argued that the HRR test is only valid as a test for association, and not as a test for association *and* linkage, this is, linkage disequilibrium, as it would be most desirable. In their paper, Spielman *et al*., departing from the HRR, and using some previous theoretical results about it [59], proposed a method which they called Transmission/Disequilibrium Test (TDT), which, with different adaptations and improvements has been thoroughly studied since then. One of the main features of the test is that it can be valid both as a test of linkage and as a test of association. The data come again from nuclear families with one affected child and it is devised to be used with diallelic loci, though they extended it to more than one affected child per family and gave indications on how to proceed with multiallelic loci.

In its basic form, the TDT considers *n* nuclear families, each of whose children is affected. The data from the 2*n* parents are displayed as in (Table **6**).

Table 6. Transmitted Vs. Nontransmitted Alleles Table for TDT Analysis

	Nontransmitted allele		
Transmitted allele	**M_1**	**M_2**	**Total**
M_1	*a*	*b*	*a+b*
M_2	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*2n*

To allow the test to be valid both as a test of linkage and association, it is necessary to use only data from heterozygous parents, since the values to compare are those of *b* and *c*. The statistic to use is then,

$$X^2_{td} = \frac{(b - c)^2}{(b + c)},$$

which under the null hypothesis follows a Chi-square distribution, and it tests whether there is or not linkage between the marker loci and the disease, *but* it is valid only when *D*, the disequilibrium rate, is greater than zero. Therefore, although it takes advantage of the existence of association and adopts some features like comparing allele frequencies, more proper of association studies, it has been argued that its normal use should be to detect linkage whenever an association has been found [39,60], although this should not be interpreted in general as meaning *tight* linkage [61].

A very important feature of the TDT is that it remains valid as a test of linkage under the presence of association even when population structure exists. Its analytical properties and performance were tested under some population admixture models, and these effects do not alter the TDT's properties to detect linkage and association [60]. Simulations showed that the TDT retained the nominal probability of type I error, while the contingency statistic associated to HRR and AFBAC data exceeded it –the higher the gametic disequilibrium created, the wider the gap between empirical and nominal probability–, which means that these tests provide a falsely high significance of the results, since they allow more false positives than they should. For complex diseases, the TDT seems also more powerful than heterogeneity tests involving healthy siblings, since –not as otherwise could be thought– there is no expected distortion in the transmission to nonaffected siblings [62]. However, when multiple sib families are used, the TDT loses its utility to look for association [63]. The reason is that under the hypothesis of no linkage, the transmission of marker-disease haplotypes from the heterozygous parents to their offspring is independent from one sib to another, whereas this is not so just under the hypothesis of no association.

The TDT has been extended to the case of multiallelic markers by several authors. Spielman and Ewens themselves [64] proposed a natural extension consisting in comparing for a given marker allele M_i, the number of times that it is transmitted to the affected offspring, $n_{i\cdot}$, with $n_{\cdot i}$, the number of times that it is not. For a *k*-allelic marker, the result is a statistic,

$$T_{mhet} = \frac{k - 1}{k} \sum_{i=1}^{k} \frac{(n_{i\bullet} - n_{\bullet i})}{n_{i\bullet} + n_{\bullet i} - 2n_{ii}},$$

which, according to Spielman and Ewens, approximately follows a X^2_{k-1} under the null hypothesis. Sham [65] notes that this statistic follows asymptotically a X^2_{k-1} distribution only under rather restrictive conditions -namely, the equality of the frequencies of all parental heterozygous genotypes. Martin *et al*. [63] also criticise the chi-square validity of the test for testing the hypothesis of no association or the one of no association or no linkage when there are multiple affected children in the sibships. They propose two statistics which allow testing for linkage disequilibrium and use all the affected children. Using only heterozygous parents, for two affected siblings and a biallelic marker, they define s_{11} as the number of transmissions of the first allele from the heterozygous parents to both affected children, s_{12} as the number of transmissions of the first allele to only one of the children and s_{22} as the number of transmissions of the second allele to both of them. Letting $h = s_{11} + s_{12} + s_{22}$ and $h^* = s_{11} + s_{22}$, then they define the statistic T_{sp} as $T_{sp} = (h/2h)T_{mhet}$, and letting

$$T^*_{mhet} = \frac{(s_{11} - s_{22})^2}{s_{11} + s_{22}}$$

then they define the statistic $T_{su} = T^*_{\text{mhet}}/2$. In the case of a biallelic marker, both tests are the same, but differences arise when multiallelic markers are considered. However, both statistics follow, for a *m*-allele marker and two affected siblings per nuclear family, a $X^2_{\text{m-1}}$ distribution. They show how to extend the statistics to nuclear families with more than two affected children, to sets of families with different numbers of children, and to multiallelic markers. They also provide Monte Carlo methods for power computations. When tested by simulation under different disease models, both tests are approximately equivalent, unless the sample size is smaller than 100 families and then T_{sp} is slightly more powerful. When comparing T_{sp} with T_{mhet} (taking for the latter one sibling at random from the affected sibs to assure the validity of the test as a test of association), the former was uniformly more powerful, except for the recessive model, in which both tests performed almost identically.

Previously, Sham and Curtis [66] had already implemented a multiallelic extension of the TDT, based on the notion that each marker allele is associated to a different extent with the disease. They modelled the likelihood under a logistic regression framework and obtained good results in terms of power for strong LD and recessive diseases. Based on numerical results, Spielman and Ewens [64] claim that the use of their extended TDT statistic is asymptotically equivalent to that in [66] as well as to other approaches such as Harley *et al*.'s [67] and Duffy's [68]. Kaplan *et al*. [69] proposed the application of Monte-Carlo simulation to some of these tests in order to avoid using asymptotic approximations which had been questioned elsewhere [66,70]. Power calculations were performed to compare the performance of the different tests, with different results with respect to the type of population and genetic model assumed. This way of proceeding, however, was questioned by Sham [65] in favour of the chi-square approach of his aforementioned statistic. When the available data is sparse, the use of Monte Carlo methods for the estimation of p-values become necessary, since the use of the χ^2 distribution is only valid when the number of informative –heterozygous-parents under study is large [71].

One of the biggest problems of the TDT-like statistics is that they require information on parental genotypes, which in some occasions is not always available. A clear example are the late onset diseases, for which at the time of detection in the offspring, no genotype information can be obtained from the parents, since they are for the most part deceased. For this reason, a number of efforts have been directed towards the development of tests not involving the parental generation to assess information about transmissions. Most of these use unaffected siblings. Let us review some of them.

Curtis [72] proposed a statistic to test the null hypothesis of no association or no linkage (i.e., to detect the existence of linkage disequilibrium) in minimal configuration sibship samples, that is, those with one only affected sib and one only unaffected sib. When several sibs are available, he suggested choosing at random one affected and take that unaffected bearing a maximally different genotype to that of the affected sib. His statistic is based on the values of T_{ij}, for $i \neq j$, defined as follows: each allele of the affected sib is compared with each of the unaffected one. If they are different, 1/2 is added to the value of T_{ij}, where i is the affected sib's marker allele and j the unaffected one's, while if they are equal, the comparison is ignored. When the marker is biallelic, the statistic used is

$$Z_c = \frac{T_{12} - \left(\frac{N_1}{2} + N_2\right)}{\sqrt{\frac{N_1}{4} + N_2}},$$

which is asymptotically $N(0,1)$, where N_i is the number of sibships causing an increase of i in the total count of T_{12} and T_{21} together.

Boehnke and Langefeld [73] recommend the use of the Discordance Allele Test (DAT), in which the structure studied is that of minimal configuration sibships too, with the addition that only those sibships in which sibs have different genotypes are taken into account. The comparison of both genotypes is used to build A $2\times m$ table: if the four alleles are all different, then +1 is added to each of the four corresponding cells, while if both sibs share one allele, only the two different alleles are counted in the table. Denoting by n_{ij} the total number of observations for affection status i and allele j, the usual chi-square statistic for contingency tables is used:

$$AC_2 = \sum_{i=1}^{m} \frac{(n_{1i} - n_{2i})^2}{n_{1i} + n_{2i}}.$$

Due to the correlation between the sibs genotypes, Monte-Carlo approximations applied to a permutation procedure must be employed to estimate the p-value of the test (see [74] for a detailed description of this procedure).

The Sib-TDT (S-TDT) was introduced by Spielman and Ewens [75]. For each marker allele i a random variable Y_i is defined as the number of i alleles present in the affected individuals in all sibships. The aforementioned permutation procedure is used here to calculate, for each sibship k the mean value A_{ik} and variance V_{ik} of the contribution of that family to Y_i, to be utterly able to define the normalised statistic

$$Z_i = \frac{Y_i - \sum_{k=1}^{N_s} A_{ik}}{\sqrt{\sum_{k=1}^{N_s} V_{ik}}},$$

where N_s is the number of sibships. For biallelic markers Z_1 would be the statistic to use, and the p-value would be

obtained, either from the Monte-Carlo permutation procedure if the sample is small, either from normal approximations for large samples. For multiallelic markers, the statistic

$$Z_{max} = \max_{i=1,...m} |Z_i|$$

should be applied, and its p-value approximated by the Monte-Carlo permutation approach. When compared to the original formulation of the TDT, the S-TDT requires considerably more genotyping than the TDT to achieve similar power [76], but it has, as the rest of these sibship-based statistics, the obvious advantage of not needing information from the parents. A modification was introduced to this statistic by Monks *et al*. [74] in order to be able to make asymptotic approximations. They proposed the T_{MSTDT} statistic, defined as

$$T_{MSTDT} = \frac{m - 1}{m} \sum_{i=1}^{m} Z_i^2 .$$

Its distribution under the null hypothesis of no linkage as well as under the one of no linkage or no association approximately follows a $X^2_{\text{m-1}}$. They compared by simulation the performance of the four tests, and concluded that Curtis' is uniformly less powerful than the rest, DAT and T_{MSTDT} are very close one to each other in all the models simulated and the S-TDT is more powerful than those two depending on the nature of the association between marker and disease.

Horvath and Laird [77] also developed a sibship-based test, the sibship disequilibrium test (SDT). For a biallelic marker, with alleles 1 and 2, differences between the mean number of 1 alleles among the affected and unaffected sibs are measured for each sibship. If *b* denotes the number of times that these differences are greater than zero and *c* the number that they are less than zero, then the statistic

$$T = \frac{(b - c)^2}{b + c}$$

is computed. Under the null hypothesis of no linkage or no association it follows a X^2_1 distribution, which allows exact computation of p-values. For multiallelic markers, they extend the test by a multivariate one. In particular, for a marker with *m* alleles, the quantities d_i^j are defined as the differences between the mean number of *j* alleles among affected and unaffected sibs of sibship *i*, where *j*=1,...*m* and *i*=1,...,*N*, being *N* the number of sibships. For each *j*, where *j*=1,..*m*-1, let and *S*'= (S^1,..,S^{m-1})', where *sign*(·) takes the

$$S^j = \sum_{i=1}^{N} sign\left(d_i^j\right)$$

values –1, 0 or 1. Defining the matrix *W* as for *j*,*k*=1,...*m*-1, then the statistic to use is $T = S'W^1 S$, which, under the null hypothesis follows asymptotically a $X^2_{\text{m-1}}$ distribution. They also consider discordant sib pair data to define a class of discordant-sib-pair tests. In the biallelic marker case, for each sib pair two numbers are measured: (*i*, *j*), the number of 1 alleles in the affected and unaffected sib, respectively. In the biallelic situation, *i* and *j* take the values 0, 1 and 2. Let b_2 the number of (2,0) pairs, c_2 the number of (0,2), b_1 the number of (2,1) or (1,0) and c_1 the number of (1,2) or (0,1). Then, the class of statistics T_x, for x > 0, is defined as

$$W_{jk} = \sum_{i=1}^{N} sign\left(d_i^j\right) sign\left(d_i^k\right)$$

$$T_x = \frac{b_1 - c_1 + x(b_2 - c_2)}{\sqrt{b_1 - c_1 + x^2(b_2 - c_2)}},$$

which, under the hypothesis of no linkage or no association, follows asymptotically a $N(0,1)$ distribution. T_x generalises both SDT and Curtis' test, since T_1 is equivalent to SDT and T_2 is equivalent to Curtis' test [77]. When comparing SDT and Curtis' and S-TDT in the case of sib pairs -these two tests are equivalent in this case [77]-, the differences in power were small. When testing for disequilibrium if linkage is established, S-TDT is slightly more powerful than SDT, but when testing for linkage none is uniformly more powerful than the other [77].

Other TDT extensions were made to allow the information from larger pedigrees containing several nuclear families (which would be otherwise rejected from the study) by Martin *et al*. [78] or to allow gene-gene or gene-enviroment interactions [79]. Particularizations to the X chromosome were also performed [81].

GENOMIC SCREENS

As mentioned above, an important goal of a gene positioning study by linkage disequilibrium is to narrow the area in the chromosome to which a gene has been shown to be linked. This task first requires that a linkage analysis based genome screen has been accomplished.

However, due to the growing availability of densely genome-wide distributed molecular markers such as SNPs (see for instance [81]), the possibility of applying association studies to genome screens has lately been suggested [14, 82]. Since then, most of the investigation has been related to theoretical aspects, such as the convenience of one or another type of markers, their optimal density throughout the genome or the sample sizes necessary to obtain good power with the different statistics. Nevertheless, some practical applications have been published. For example, in [27] a whole chromosome was screened for bipolar mood disorder (BP-I), and in [21] a 2702 cM linkage disequilibrium map was generated in cattle.

Two questions have to be addressed in order to perform a successful study: an appropriate marker map density and the

possible need, depending on the method, for a correction for a high number of false positives due to the multiple comparisons that have to be made. Kruglyak [83] concluded that in a general population, useful levels of LD would not extend beyond an average distance of 3 kb, and a battery of approximately 500,000 SNPs would be necessary for whole-genome screening purposes. Durham and Feingold [84] dealt extensively with IBD probabilities in order to correct for false-positive rates in a method involving features of affected-pair analysis, a way of detecting linkage, and of linkage disequilibrium studies.

Since it appeared, the TDT has generated a considerable amount of literature. As expectable, its properties for scanning large portions of genome have been widely investigated. Risch and Merikangas [14] proposed it as an alternative in genome screens to the traditional allele sharing in affected sib pairs (ASP) tests to detect genes of small or moderate effect. Their analysis was performed under different penetrance situations for a multiplicative inheritance model of the disease, assuming, for the sake of simplification, a strong linkage between marker and disease, with recombination fraction equalling zero. The number of family units necessary to achieve an 80% power was compared in a linkage study vs. an association study performed with the TDT, with five biallelic markers in each of the 100000 genes assumed known to compose the whole genome and the same whole significance levels of α=0.05. In general the performance of the TDT was better than linkage by ASP for genes of this level of effect. Several objections were made to Risch and Merikangas study [23,24,85,86], regarding questions such as the usefulness and convenience of classic linkage analyses, the use of populational-level sampling schemes instead of family-level ones or what happens when the situation is not as optimal as having one of the markers alleles being the causative one itself. Each of these topics was subsequently discussed by Risch and Merikangas themselves [87].

The study of Risch and Merikangas [14] was broadened by considering other inheritance models [15], with similar results, in the sense of the TDT being superior to ASP tests. Pajukanta *et al*. [52] note that the lack of validity of the assumption of independence of parental transmissions for certain models. In response, Camp [88] states that the assumption of independence is not as decisive as it first seems. Knapp [89] derives, under the same models as in [15], two different approximations for the power of the TDT based on standard stochastic convergence theory. While one of them leads under the multiplicative model to the same results as in [14], the other seems much more precise, since the estimates of the real power obtained by simulation agree much better with the proposed 80%.

Another point of interest is the kind and density of the markers used for the screen. A different approach to those mentioned above was proposed by Chapman and Wijsman [90] to compare the sample sizes necessary to achieve an 80% power for biallelic and multiallelic markers in a case control context. With several assumptions –punctual introduction of the mutation, nonoverlapping generations, random sampling, no migration, genetic drift or mutations in the marker nor in the disease loci, and constant global frequencies for marker and disease loci alleles- they varied the number of generations since the introduction of the disease mutation, the markers spacing and the mode of inheritance of the disease. As expected, since the information provided by a multiallelic marker is higher than that of a biallelic one, the sample size required by a multiallelic marker analysis is much lower than that necessary to perform a biallelic marker analysis, all the rest of variables being equal. Equivalently, the density of the map to achieve a given power with a given significance level and sample size is much higher for biallelic than for multiallelic loci.

ASSOCIATION AND QTLS

Although association analyses have focused more study and attention in their application to mapping binary traits, especially simple inheritance diseases, their application to Quantitative Trait Loci (QTL) detection has also been studied, suggesting that they can usefully be applied to this problem as well. As in the binary case, there are two global strategies to approach the problem: population-level analyses or family-level studies, usually performed with family trios. Again, extensions to sibship-based tests have been made, as well as to other procedures such as the use of variance component models.

One of the first tests was proposed by Boerwinkle *et al*. [91,92]. The so called Measured-Genotype test, its two alternative forms, based on alleles (MGA test) and in genotypes (MGG test), use the measures of genotypes and phenotypes of unrelated individuals to classify them into either two (MGA) or three (MGG) groups, according to the alleles or the genotypes they bear. Then an analysis of variance (ANOVA) analysis is performed to test for significative differences between the groups.

When dealing with quantitative traits there is a wider set of options to devise a study of this kind. When facing the costs of the analysis, one of the most important decisions is to genotype every individual for which there is a phenotypic record, or restrict the analysis to those showing extreme values in the phenotypes distribution. With the latter idea in mind, Page and Amos [93] proposed two tests based on the MGA and MGG, the TMA and TMG, where the T stands for "truncated", in which the procedures are identical to those mentioned above, except that the individuals genotyped come from the extremes of the phenotypic distribution.

Luo *et al*. [94] developed a population genetics model for the distribution of linkage disequilibrium between a polymorphic marker and a QTL. This model was applied by Luo [95] to devise a procedure to detect LD using principles of analysis of variance based upon asymptotic variance results using the Expectation-Maximization (EM) algorithm by Kao and Zeng [96]. Luo and Suhai [97] finally introduced a maximum likelihood method not only to detect LD, but to allow the estimation of the amount of disequilibrium between the marker locus and the QTL out of a random sample from the population. The complexity of the model avoids an analytic solution to the likelihood equations, so their previous studies of EM estimation are applied to estimate the parameters involved in the equation. They

conclude that a sample size of at least 500 individuals and a QTL explaining at least one quarter of the phenotypic variance of the trait are necessary to provide good estimations. The application of extreme, also called selective, genotyping is also considered, although it would be useful only in the detection step, but not in the estimation one, since the final sample would not be a random one.

In the same way that occurs with discrete traits, population artifacts such as stratification or admixture can mask the detection of a QTL when population samples are used. With the precedent of the TDT in binary traits, Allison [98] presented a collection of TDT extensions for quantitative traits, termed TDT_{Q1} to TDT_{Q5}. They take the form, respectively, of a *t*-test of comparison of means, a chi-square test with selective genotyping, a *t*-test with selective genotyping, a normal statistic to test for deviations in the probability of transmission of the favourable allele in the extreme tails of the phenotypic distribution and a *F*-ratio test to compare the regressions of the phenotypic value over one model without allelic transmission information and the same one with the addition of variables related to transmission counts. Using simulations, he concluded that by selecting 20% of the extreme phenotypes it is possible to detect, at a $\alpha = 0.0001$ level, with 80% power, QTLs with a 5% effect with less than 300 observations -trios- [98], and that TDT_{Q5} performs better in general than the rest. When compared under different situations TDT_{Q1} to TDT_{Q4}, TMG, TMA, MGG, MGA and a standard case control test, Page and Amos found TMA to be in general the most powerful of the group, although the Q-TDTs proved much more efficient in situations of the presence of population events [93], quite in agreement with [99]. Rabinowitz [100] also extended the TDT to quantitative traits, maintaining the analysis unit of two parents and only one child.

A logical extension, as in the binary case, was to use information from sibship structures to avoid the problems derived from the eventual lack of data from parental generations. Allison *et al*. [101] proposed two tests, using minimal configuration sibships -i.e., two, and only two, sibs, with different genotype and phenotype-, with the possibility of using multiallelic markers. While the first procedure applies a mixed effects analysis of variance, with the sibship as random effect, the marker genotype as fixed and the phenotypic value as dependent variable, the second is a permutation test. In it, for each sibship, the difference between the observed phenotypic mean value of each allele and the expected one is computed, where the expected value is obtained by permutation of the different phenotypic values within each sibship. The statistic thus obtained can be approximated by a χ_{m-1}, where *m* is the number of alleles at the marker locus. Selective sampling, in which the selection is made based upon Mahalanobis distance computations for the sibships, seems to considerably increase the power [101].

Fulker *et al*. [102] combined linkage and association approaches using sib-pairs for QTL detection. Their procedure is based on a joint modeling of the mean allelic values and the covariance structure through a maximum likelihood approach and a biometrical model. Monks and Kaplan [103] also take into account covariances. They use the estimate of the covariance between trait values and transmission counts from the parents as the basis for their tests, since these variables should be uncorrelated under the null hypothesis of absence of association or linkage. With this approach they developed three tests: one using genotype information of both parents and all of their children, another one using genotypes of all the siblings, without parental information, and a third combining the two former –to ensure the feasibility of the analysis, only informative sibships are considered, considering as such, those who allow inferences about transmission counts from their non-genotyped parents. Other authors [104,105] have also gone into the study of variances, by exploiting variance component models. Finally, other interesting approaches are those of Deng *et al*. [106], who extend to the quantitative case the studies of Feder *et al*. [107] and Nielsen *et al*. [108], in which testing for Hardy-Weinberg equilibrium can be combined with testing for linkage disequilibrium for QTL mapping purposes, and George *et al*. [109], who, in a similar way to that in [101], proposed a multiple regression procedure, with some differences with respect to [101], such as the incorporation of covariates.

CONCLUSIONS

The main objective of this review was to provide a whole perspective of the many branches being explored in this recent, but nonetheless old -the concept of gametic disequilibrium has been known for many decades- field of investigation. The study of association procedures for gene mapping purposes has generated in the last several years a large amount of literature. We hope we have provided an insight into most of the milestones of linkage disequilibrium studies: population and family samples, biallelic and multiallelic markers, analytical and iterative results, qualitative and quantitative traits, single locus and multilocus studies. From the most conceptually simple statistics, mere applications of standard statistical test to this particular case, we have arrived to some of the most sophisticated and more up-to-date methods, in which the more they try to model reality, the more complex they grow.

Although linkage analysis is still in use and new advances are constantly appearing, we have seen that LD studies provide a complementary approach. Tests of association have proved to offer good power for detection and a high mapping accuracy. Their dependence on population history and structure, however, suggests that their use may be restricted to the situations for which they were devised. Population-based statistics, are in general useful when population history is known, in particular when dealing with isolated populations of medium age. Artifacts such as admixture or stratification can make these kind of studies lose power and can lead to detection of spurious associations that lead to no physical proximity at all. Nevertheless, recent studies have attempted to solve this problems and maintain the advantages of sampling directly at the population level, with promising results. When the population is a mixed one or no ascertainment scheme can be applied in order to perform a reliable population-level analysis, the alternative are the family based tests. With the TDT as their most significant and known member, the possibilities are diverse, to allow for different configurations and sampling schemes:

the parents and one child, the parents and several children, sib pairs, multiple siblings...The discussion is analogous when quantitative traits are analysed, with the added complexities, troubles and possibilities derived from the fact of dealing with a continuous phenotypic set of values. The advantages of selective genotyping, for instance, are an ongoing subject of study.

Human and outbred animal or plant populations are intricate subjects of study. Each population has its own particularities, so not all the statistics may be suitable for it: there is no such thing as a perfect test. For this reason, new methods and improvements, modifications and comparisons of existing ones are constantly being published.

The speed at which new advances are made in the molecular field, providing higher density maps of markers favour as well the use of association studies to locate genes of interest. As an example, the recent "boom" of the SNP technology is promising to soon make available the saturation of candidate gene areas at a low genotyping cost, due to the ability to of automatize of the genotyping process. Based on this fact, association approaches devised for their application on SNP information have been published. The contribution of the increasing computation power of today's computers is also an important factor to take into account, since it allows the inclusion of enormous volumes of data to perform complex analyses at a high speed.

To summarise, all these factors taken together make association and linkage disequilibrium procedures an ever increasingly important tool for effectively map traits of interest for advancing genetic knowledge.

ACKNOWLEDGEMENTS

This work was financially supported by the EC DGVI QLRT-99-30147 project. The authors would like to thank C. Carleos from Oviedo University and J. Hernandez from Roslin Institute for their useful comments and bibliografical support, as well as to the referees for their noteworthy remarks.

ABREVIATIONS

AFBAC = Affected Family Based Controls

ANOVA = Analysis of Variance

ASP = Affected Sib Pairs

CF = Cystic Fibrosis

DAT = Discordance Allele Test

DTD = Diastrophic Dysplasia

EM = Expectation-Maximization

GLE = Generalized Least Squares

HD = Huntington's Disease

HRR = Haplotype Relative Risk

IBD = Identity by Descent

LD = Linkage Disequilibrium

LR = Likelihood Ratio

MC = Monte Carlo

MCMC = Markov Chain Monte Carlo

MGA = Measured-Genotype Allele-based test

MGG = Measured-Genotype Genotype-based test

ML = Maximum Likekihood

QTL = Quantitavtive Trait Locus

RFLP = Restricted Fragment Length Polymorphism

SNP = Single Nucleotide Polymorphism

S-TDT = Sib-Transmission/Disequilibrium Test

SDT = Sibship Disequilibrium Test

SNP = Single Nucleotide Polymorphism

STR = Short Tandem Repeat

STRAT = STRuctured population Association Test

TDT = Transmission/Disequilibrium Test

TDT_{Q1} to TDT_{Q5} = Allison's Transmission/Disequilibrium Tests for Quantitative traits

TMA = Truncated Measured-genotype Allele-based test

TMG = Truncated Measured-genotype Genotype-based test

REFERENCES

[1] Mohr, J. (1964) Practical possibilities for detection of linkage in man. *Acta Genet. Basel*, **14,** 125-132.

[2] Botstein, D.; White, R.L.; Skolnick, M.; Davis, R.W. (1980) Construction of a genetic linkage map in humans using restriction length polymorphysms. *American journal of Human Genetics*, 314-331.

[3] Lander, E.S.; Botstein, D. (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121,** 185-199.

[4] Weber, J.L.; May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.*, **44,** 388-396.

[5] Weissenbach, J.; Gyapay, G.; Dib, C.; Vignal, A;; Morissette J.; Millasseau P.; Vaysseix G.; et al (1992) A

second-generation linkage map of the human genome. *Nature*, **359,** 794-801.

[6] Collins, F.S. (1995) Positional cloning moves from perditional to traditional. *Nature Genetics*, **9**, 347-350.

[7] Dib C.; Faure S.; Fizames C.; Samson D.; Drouot N.; Vignal A.; Millasseau P.; et al (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**,152-l54.

[8] Zhao, L.P.; Aragaki, C.; Hsu, L.; Quiaoit, F. (1998) Mapping of compex traits by single-nucleotide polymorphisms. *American Journal of Human Genetics*, **63,** 225-240.

[9] Ott, J.; Hoh, J. (2000) Statistical approaches to gene mapping. *American Journal of Human Genetics*, **67,** 289-294.

[10] Bohenke, M. (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *American Journal of Human Genetics*, **55,** 379-390.

[11] Lander, E.S. and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11,** 241-247.

[12] Cohen, B. (1999) Freely associating (Editorial) *Nature Genetics*, **22,** 1-2.

[13] Risch, N.; Teng, J. (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Research*, **8**, 1273-1288.

[14] Risch, N.; Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273,** 1516-1517.

[15] Camp, N. (1997) Genomewide Transmission/ Disequilibrium Testing-Consideration of the Genotipic Relative Risks at Disease Loci. *American Journal of Human Genetics*, **61,** 1424-1430.

[16] Baret, P.V.; Hill, W.G. (1997) Gametic disequilibrium mapping: potential applications in livestock. *Animal Breedeing Abstracts*, **65,** 309-318.

[17] Crow, J.F.; Kimura, M. (1970) An introduction to population genetics theory. Harper and Row Publ., New York.

[18] Hartl, D.L. and Clark, A.G. (1989) Principles of population genetics. Sinauer Associates, Sunderland, MA.

[19] Weir, B.S. (1996) Genetic data analysis II. Sinauer Associates, Sunderland, MA.

[20] Briscoe, D.; Stephens, J.C.; O'Brien, S.J. (1994) Linkage disequilibrium in admixed populations: applications in gene mapping. *The Journal of Heredity*, **85,** 59-63.

[21] Farnir, F.; Coppieters, W.; Arranz, J-J.; Berzi, P.; Cambisano, N.; Grisart, B.; Karim, L.; Marcq, F.; Moreau, L.; Mni, M.; Nezer, C.; Simon, P.; Vanmanshoven, P.; Wagenaar, D.; Georges, M. (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Research*, **10,** 220-227.

[22] Rämet, M.; Haataja, R.; Marttila, R.; Floros, J.; Hallman, M. (2000) Association between the surfactant protein A (SP-A) gene locus and respiratory-distress sindrome in the Finnish population. *American Journal of Human Genetics*, **66,** 1569-1579.

[23] Bell, D.A.; Taylor, J.A. (1997) Genetic analysis of complex diseases. *Science*, **275,** 1327-1328.

[24] Long, A.D.; Grote, M.N.; Langley, C.H. (1997) Genetic analysis of complex diseases. *Science*, **275,** 1328.

[25] Hästbacka, J.; de la Chapelle, A.; Kaitila, I.; Sistonen, P.; Weaver, A.; Lander, E. (1992) Linkage disequilibrium mapping in isolated founder populations, diastrophic dysplasia in Finland. *Nature Genetics*, **2**, 204-211.

[26] Dunner, S.; Charlier, C.; Farnir, F.; Brouwers, B, Canon, J; Georges, M. (1997) Towards interbreed IBD fine mapping of the *mh* locus: double-muscling in the *Asturiana de los valles* breed involves the same locus as in the *Belgian Blue* cattle breed. *Mammalian genome*, **8**, 430-435.

[27] Escamilla, M.A.; McInnes, L.A.; Spesny, M.; Reus, V.I.; Service, S.K.; Shimayoshi, N.; Tyler, D.J.; Silva, S.; Molina, J.; Gallegos, A.; Meza, L.; Cruz, M.L.; Batki, S.; Vinogradov, S.; Neylan, T.; Nguyen, J.B.; Fournier, E.; Araya, C.; Barondes, S.H.; Leon, P.; Sandkuijl, L.A.; Freimer, N.B. (1999) Assessing the feasibility of linkage disequilibrium methods for mapping complex traits: an initial screen for bipolar disorder loci on chromosome 18. *American Journal of Human Genetics*, **64,** 1670-1678.

[28] Terwilliger, J.D. (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polimorphic marker loci. *American Journal of Human Genetics*, **56,** 777-787.

[29] Kerem, B.; Rommens, J.M.; Buchanan, J.A.; Markiewicz, D.; Cox, T.K.; Chakravarti, A.; Buchwald, M. *et al*. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245,** 1073-1080.

[30] Devlin, B.; Risch, N.; Roeder, K. (1996) Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics*, **36**, 1-16 .

[31] Williams, N. (2000) Terwilliger association test: analysis and extension. *Proceedings of the European Mathematical Genetics Meeting 37.*

[32] Xiong, M.; Guo, S.W. (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *American Journal of Human Genetics*, **60,** 1513-1531.

[33] Kaplan, N.L.; Hill, W.G.; Weir, W.S. (1995) Likelihood methods for locating disease genes in nonequilibrium populations. *American Journal of Human Genetics*, **56,** 18-32.

[34] Rannala, B.; Slatkin, M. (1998) Likelihood analysis of disequilibrium mapping, and related problems. *American Journal of Human Genetics*, **62,** 459-473.

[35] Graham, J.; Thompson, E.A. (1998) Disequilibrium likelihoods for fine-scale mapping of a rare allele. *American Journal of Human Genetics*, **63,** 1517-1530.

[36] Morris, A.P.; Whittaker, J.C.;Balding, D.J. (2000a) Bayesian fine scale mapping of disease loci, by hidden Markov models. *American Journal of Human Genetics*, **67,** 155-169.

[37] Morris, A.P.; Balding, D.J.; Whittaker, J.C. (2000b) Fine scale association mapping of disease loci via coalescent modelling of genealogies. *Proceedings of the European Genetics Meeting 43.*

[38] Devlin, B.; Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55,** 997-1004.

[39] Spielman, R.S.; McGinnis, R.E.; Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM) *American Journal of Human Genetics*, **52,** 506-516.

[40] Bacanu, S-A; Devlin, B.; Roeder, K. (2000) The Power of Genomic Control *American Journal of Human Genetics*, **66,** 1933-1944.

[41] Pritchard, J.K.; Stephens, M.; Rosenberg, N.A.; Donelly, P. (2000a) Association Mapping in Structured Populations. *American Journal of Human Genetics*, **67,** 170-181.

[42] Pritchard, J.K.; Rosenberg, N.A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics*, **65,** 220-228.

[43] Pritchard, J.K.; Stephens, M.; Donnelly, P. (2000b) Inference of population structure using multilocus genotype data. *Genetics* 2000 155: 945-959.

[44] Lazzeroni, L.C. (1998) Linkage disequilibrium and gene Mapping: an empirical least-squares approach. *American Journal of Human Genetics*, **62** 159-170.

[45] Wright, A.F.; Carothers, A.D.; Pirastu, M. (1999) Population choice in mapping genes for complex diseases. *Nature Genetics*, **23** 397-404.

[46] Sheffield, V.C.; Stone, E.M.; Carmi, R. (1998) Use of isolated inbred human populations for identifications of disease genes. *TIG*, **14** 391-396.

[47] Cheung, V.G.; Gregg, J.P.; Gogolin-Ewens, K.J.; Bandong, J.; Stanley, C.A.; Baker, L.; Higgins, M.J.; Nowak, N.J.; Shows, T.B.; Ewens, W.J.; Nelson, S.F.; Spielman, R.S. (1998) Linkage-disequilibrium mapping without genotyping. *Nature Genetics*, **18** 225-230.

[48] Taillon-Miller, P.; Bauer-Sardiña, I.; Saccone, N.L.; Putzel, J.; Laitinen, T.; Cao, A.; Kere, J.; Pilia, G.; Rice, J.P.; Kwok, P-Y. (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genetics*, **25** 324-328.

[49] Osier, M.; Pakstis, A.J.; Kidd, J.R.; Lee, J-F.; Yin, S-J.; Ko, H-C.; Edenberg, H.J.; Lu, R-B; Kidd, K.K. (1999) Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. *American Journal of Human Genetics*, **64** 1147-1157.

[50] Chapelle, A. de la; Wright, F. (1998) Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proceedings of National Academy of Sciences***, 95**, 12416-12423.

[51] Slatkin, M.; Excoffier, L. (1996) Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity*, **76**, 377-383.

[52] Pajukanta, P.; Terwilliger, J.D.; Perola, M.; Hiekkalinna, T.; Nuotio, I.; Ellonen, P.; Parkkonen, M.; Hartiala, J.; Ylitalo, K.; Pihlajamäki, J; Porkka, K.; Laakso, M.; Viikari, J.; Ehnholm, C.; Taskinen; M-R.; Peltonen, L. (1999) Genomewide scan for familial combined hyperlipidemia genes in Finnish families, suggesting multiple susceptibility loci influencing triglyceride, cholesterol, and apoliprotein B levels. *American Journal of Human Genetics*, **64**, 1453-1463.

[53] Eaves, I.A.; Merriman, T.R.; Barber, R.A.; Nutland, A.; Tuomilehto-Wolf, E.; Tuomilehto, J.; Cucca, F.; Todd, J.A. (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **25**, 320-323.

[54] Laan, M.; Pääbo, S. (1998) Mapping genes by drift-generated linkage disequilibrium. *American Journal of Human Genetics*, **63**, 654-656.

[55] Terwilliger, J.D.; Zöllner, S.; Pääbo, S. (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: "drift mapping" in small populations with no demographic expansion. *Human Heredity*, **48,** 138-154.

[56] Thomson, G. (1988) HLA disease associations: models for insulin dependent diabetes mellitus and the study of complex human genetic disorders. *Ann. Rev. Genet*., **22,** 31-50.

[57] Thomson, G.; Robinson, W.P.; Kuhner, M.K.; Joe, S. (1989) HLA, insulin gene, and Gm associations with IDDM. *Genetic Epidemiology*, **6**, 155-160.

[58] Falk, C.T.; Rubinstein, P. (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics*, **51,** 227-233.

[59] Ott, J. (1989) Statistical properties of the haplotype relative risk. *Genetic Epidemiology*, **6**, 127-130.

[60] Ewens, W.J.; Spielman, R.S. (1995) The transmission/ disequilibrium test: history; subdivision and admixture. *American Journal of Human Genetics*, **57,** 455-464.

[61] Whittaker, J.C.; Denham, M.C.; Morris, A.P. (2000) The problems of using the Transmission/ Disequilibrium Test to infer tight linkage. *American Journal of Human Genetics*, **67,** 523-526.

[62] Benedicte, A.L.; Rønningen, K.S.; Akselsen, H.E.; Thorsby, E.; Undlien, D.E. (2000) Application and interpretation of transmission/disequilibrium tests: transmission of HLA-DQ haplotypes to unaffected siblings in 526 families with type 1 diabetes. *American Journal of Human Genetics*, **66,** 740-743.

[63] Martin, E.R.; Kaplan, N.L.; Weir, B.S. (1997) Tests for linkage and association in nuclear families. *American Journal of Human Genetics*, **61,** 439-448.

[64] Spielman, R.S.; Ewens, W.J. (1996) The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics*, **59,** 983-989.

[65] Sham, P. (1997) Transmission/disequilibrium tests for multiallelic loci (letter to the editor). *American Journal of Human Genetics*, **61,** 774-778.

[66] Sham, P.C.; Curtis, D. (1995) An extended transmission/disequilibrium test (TDT) for multiallele marker loci. *Annals of Human Genetics*, **59,** 323-336.

[67] Harley, J.B.; Moser, K.L.; Neas, B.R. (1995) Logistic transmission modelling of simulated data. *Genetic Epidemiology,* **12,** 607-612.

[68] Duffy, D.L. (1995) Screening a 2 cM genetic map for allelic association: a simulated oligogenic trait. *Genetic Epidemiology,* **12,** 595-600.

[69] Kaplan, N.L.; Martin, E.R.; Weir, B.S. (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. *American Journal of Human Genetics,* **60,** 691-702.

[70] Bickeböller, H.; Clerget-Darpoux, F. (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genetic Epidemiology,* **12,** 865-870.

[71] Schaid, D.J. (1998) Transmission disequilibrium, family controls, and great expectations. *American Journal of Human Genetics,* **63,** 935-941.

[72] Curtis, D. (1997) Use of siblings as controls in case-control association studies. *Annals of Human Genetics,* **61** 950-961.

[73] Boehnke, M.; Langefeld, C.D. (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *American Journal of Human Genetics,* **62,** 950-961.

[74] Monks, S.A.; Kaplan, N.L.; Weir, B.S. (1998) A comparative study of sibship tests of Linkage and/or association. *American Journal of Human Genetics,* **63,** 1507-1516.

[75] Spielman, R.S.; Ewens, W.J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics,* **62,** 450-458.

[76] Whittaker, J.C.; Lewis, C.M. (1999) Power comparisons of the Transmission/Disequilibrium test and Sib-Transmission/Disequilibrium test statistics. *American Journal of Human Genetics,* **65,** 578-580.

[77] Horvath, S.; Laird, N.M. (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. *American Journal of Human Genetics,* **63,** 1886-1897.

[78] Martin, E.R.; Monks, S. A.; Warren, L.L.; Kaplan, N.L. (2000) A Test for Linkage and Association in General Pedigrees: The Pedigree Disequilibrium Test *American Journal of Human Genetics,* **67,** 146-154.

[79] Lunetta, K.L.; Faraone, S.V.; Biederman, J.; Laird, N.M. (2000) Family-based tests of association and linkage that use unaffected sibs, covariates and interactions. *American Journal of Human Genetics,* **66,** 605-614.

[80] Horvath, S.; Laird, N.M; Knapp, M. (2000) The transmission/disequilibrium test for parental genotype reconstruction for X-chromosomal markers. *American Journal of Human Genetics,* **66,** 1161-1167.

[81] Wang, D. *et al* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science,* **280,** 1077-1082.

[82] Kruglyak, L. (1997) What is significant in whole-genome linkage disequilibrium studies? *American Journal of Human Genetics,* **61** 810-812.

[83] Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics,* **22,** 139-144.

[84] Durham, L.K.; Feingold, E. (1997) Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *American Journal of Human Genetics,* **61,** 830-842.

[85] Scott, W.K.; Pericak-Vance, M.A.; Haines, J.L. (1997) Genetic analysis of complex diseases. *Science,* **275,** 1327.

[86] Müller-Myhsok, B.; Abel, L. (1997) Genetic analysis of complex diseases. *Science,* **275,** 1328-1329.

[87] Risch, N.; Merikangas, K. (1997) Genetic analysis of complex diseases. *Science,* **275,** 1329-1330.

[88] Camp, N. (1999) Genomewide Transmission/ Disequilibrium Test (letter to the Editor). *American Journal of Human Genetics,* **64,** 1485-1487.

[89] Knapp, M. (1999) A note on power approximations for the transmission/disequilibrium test. *American Journal of Human Genetics,* **64,** 1177-1185.

[90] Chapman, N.H.; Wijsman, E.M. (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *American Journal of Human Genetics,* **63,** 1872-1885.

[91] Boerwinkle, E.; Chakraborty, R.; Sing,C.F. (1986) The use of measured genotype information in the analysis of quantitative phenotypes in man. *Annals of Human Genetics,* **50,** 181-194.

[92] Boerwinkle, E.; Viscikis, S.; Welsh, D.; Steinmetz, J.; Hamash, S.M.; Sing, C.F. (1987) The use of measured genotype information in the analysis of quantitative phenotypes in man. II. The role of apolipoprotein E polymorphisms in determining levels, variability, and covariability of cholesterol, betalipoprotein, and triglycerids in a sample of unrelated individuals. *Am. J. Med. Genet.,* **27,** 567-582.

[93] Page, G.P.; Amos, C.I. (1999) Comparison of linkage-disequilibrium methods for localization of genes influencing quantitative traits in humans. *American Journal of Human Genetics,* **64,** 1194-1206.

[94] Luo, Z.W.; Thompson, R.; Wooliams, J.A. (1997) A population genetics model of marker-assisted selection. *Genetics,* **146,** 1173-1183.

[95] Luo, Z.W. (1998) Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Heredity,* **80,** 198-208.

[96] Kao, C-H; Zeng, Z-B (1997) General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics,* **53,** 653-665.

[97] Luo, Z.W.; Suhai, S. (1999) Estimating Linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Genetics,* **151,** 359-371.

[98] Allison, D.B. (1997) Transmission-Disequilibrium tests for quantitative traits. *American Journal of Human Genetics,* **60,** 676-690.

[99] Long, A.D.; Langley, C.H. (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research*, **9,** 720-731.

[100] Rabinowitz, D. (1997) A transmission disequilibrium test for quantitative trait loci. *Human Heredity*, **47,** 342-350.

[101] Allison, D.B.; Heo, M.; Kaplan, N.; Martin, E.R. (1999a) Sibling-based tests of linkage and association for quantitative traits. *American Journal of Human Genetics*, **64,** 1754-1764.

[102] Fulker, D.W.; Cherney, S.S.; Sham, P.C.; Hewitt, J.K. (1999) Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics*, **64,** 259-267.

[103] Monks, S.A.; Kaplan, N.L. (2000) Removing the sample restrictions from family-based tests of association for a quantitative-trait locus. *American Journal of Human Genetics*, **665,** 576-592.

[104] Allison, D.B.; Neale, M.C.; Zannolli, R.; Schork, N.J.; Amos, C.I.; Blangero, J. (1999b) Testing the robustness of the likelihood ratio test in a variance-component quantitative-trait loci-mapping procedure. *American Journal of Human Genetics*, **65,** 531-544.

[105] Sham, P.C.; Cherny, S.S.; Purcell, S.; Hewitt, J.K. (2000) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *American Journal of Human Genetics*, **66,** 1616-1630.

[106] Deng, H-W.; Chen, W-M.; Recker, R.R. (2000) QTL fine mapping by measuring and testing for Handy-Weinberg and linkage disequilibrium at a series of linked marker loci in extreme samples of populations. *American Journal of Human Genetics*, **66,** 1027-1045.

[107] Feder, J.N.; Gnirke, A.; Thomas, W.; Tsuchihashi, Z; Ruddy, D.A.; Basava, A.; Dormishian, F. e*t al*. (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genetics*, **13,** 399-408.

[108] Nielsen, D.M.; Ehm, M.G.; Weir, B.S. (1999) Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *American Journal of Human Genetics*, **63,** 1531-1540.

[109] George, V.; Tiwari, H.K.; Zhu, X.; Elston, R.C. (1999) A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *American Journal of Human Genetics*, **65,** 234-245. .