Statistical tools for linkage analysis and genetic association studies

Paola Forabosco, Mario Falchi and Marcella Devoto†

Genetic mapping by linkage analysis has been an invaluable tool in the positional strategy to identify the molecular basis of many rare Mendelian disorders. With the attention of the scientific and medical community shifting towards the analysis of more common, complex traits, it has become necessary to develop new approaches that take into account the complexity of the genetic basis of these disorders and their possible interaction with other, nongenetic factors. Linkage disequilibrium studies are now becoming increasingly popular thanks to the advent of genotyping platforms that allow genome-wide searching for association between hundreds of thousands of random polymorphisms and disease phenotypes in large samples of unrelated individuals. Moreover, the definition of the disease phenotype itself is being reconsidered to include quantitative traits that may better define the underlying biologic mechanisms for many pathologic conditions. This article will review classic and new approaches to genetic mapping by linkage and association analysis and discuss the directions this field is likely to take in the near future.

Expert Rev. Mol. Diagn. 5(5), 781–796 (2005)

CONTENTS

†Author for correspondence
Nemours Children's Clinic, Department of Biomedical Research, Wilmington, DE, USA
Tel.: +1 302 651 6838
Fax: +1 302 651 6895
mdevoto@nemours.org

KEYWORDS:
association studies, gene mapping, linkage analysis, quantitative trait loci, statistical genetics

The past 20 years have seen the elucidation of the molecular basis of an ever-increasing number of genetic disorders become possible thanks to the parallel advances in the development of biomolecular technology available in the laboratory and of the statistical tools necessary for the analysis of these data. The 1990s have been characterized by a plethora of linkage (family-based) studies that have taken advantage of the introduction of the extremely polymorphic microsatellite markers and the implementation in statistical genetic software of sophisticated analytical tools. These have made feasible the likelihood analysis of complex pedigree structures, and thus the mapping of the many genes underlying Mendelian disease in such pedigrees. The so-called model-based methods that have been used for these studies are based on the analysis of recombination between the hypothetical disease locus and random genetic markers with known location in the human genome, and require specification of the mode of inheritance at the disease locus. However, the same statistical tools have proven less successful in identifying the genes responsible for common complex traits such as obesity, schizophrenia or diabetes. Other approaches that do not require specification of a genetic model for the disease, such as model-free linkage analysis based on allele sharing among relatives, may be more appropriate for identifying the genes underlying susceptibility to a complex genetic disease. A debate has ensued on which methods have more power and are more appropriate under different circumstances.

Success in linkage mapping depends primarily on the magnitude of the effect of the locus involved in the trait under investigation, and low-penetrant genetic variants underlying complex traits may be too weak to be detected by linkage studies. For low-penetrant genetic variants, association approaches have been shown to have increased power compared with linkage methods – much smaller sample sizes would be required to detect association than to detect linkage. Association study is a widely accepted important complement to linkage analysis in refining the location of disease genes in regions

10.1586/14737159.5.5.781

ISSN 1473-7159

previously identified by linkage (fine-scale mapping). With the discovery of massive numbers of genetic markers and the development of better statistical tools, association studies have received renewed attention and the potential use of these strategies as a tool for identifying more elusive genes involved in complex diseases in the whole genome is currently under debate.

More recently, attention has grown around the study of quantitative traits that underlie or are used to define many common conditions such as obesity and hypertension. Direct genetic analysis of the quantitative traits may provide increased power for the identification of susceptibility genes for these disorders. This article will review these methods and instruments of genetic analysis, and the directions that this field is likely to take in the near future.

Linkage analysis of binary traits

Model-based linkage analysis

The principal method of analysis for disease gene mapping of Mendelian binary traits in pedigrees is the so-called lod-score method [1,2]. The general idea underlying linkage analysis is to correlate the segregation of the disease and of random genetic markers with known location in the human genome in selected families. The basis of the lod-score method is the analysis of recombination, or the phenomenon by which alleles at different loci are separated at meiosis as a result of an odd number of crossovers (FIGURE 1). Since the probability of a crossover occurring between two loci is a function of the distance between them, it is possible to estimate the distance between two loci by means of the frequency at which recombination is observed between them. If this frequency, known as the recombination fraction and often indicated as θ, is small, then the two loci must be closely located in the genome. In contrast, when two loci are far apart on the same chromosome, or on different chromosomes, they will segregate independently and θ will be equal to 50%. By using maps of markers with known location in the human genome, we can thus identify the location of the disease gene of interest.

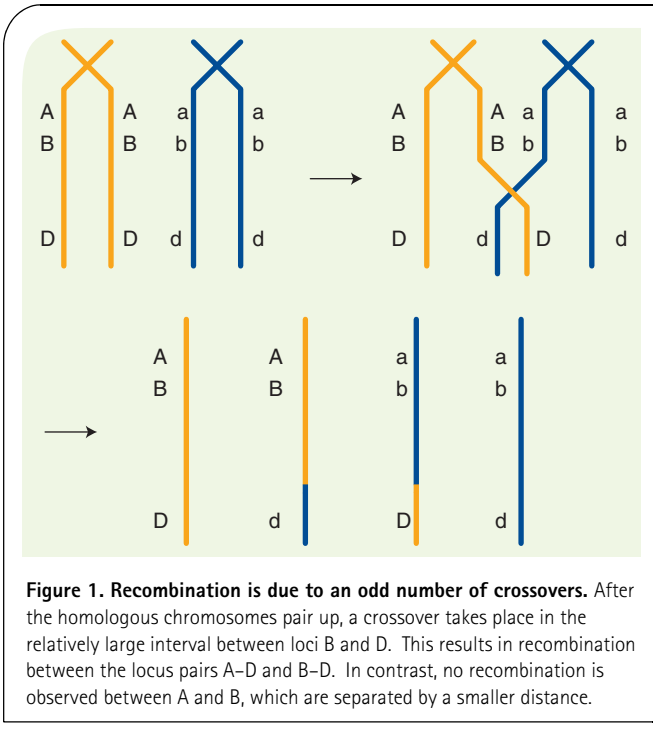

Figure 1. Recombination is due to an odd number of crossovers. After the homologous chromosomes pair up, a crossover takes place in the relatively large interval between loci B and D. This results in recombination between the locus pairs A–D and B–D. In contrast, no recombination is observed between A and B, which are separated by a smaller distance.

In classic linkage analysis, the recombination fraction θ is estimated by likelihood-based statistical methods [3–5]. This method of linkage analysis is also known as model-based, or parametric, as it requires definition of a genetic model for the disease. The likelihood is a measure of the plausibility of the observations, which depends on assumptions about the genetic model parameters, and on the value of the recombination fraction. Specifically, one must define the so-called penetrance, or the probability of having the disease phenotype given an underlying genotype at the disease locus. Other parameters that need to be specified include the disease and marker allele frequencies, but studies have shown that while the power of the linkage test is sensitive to the degree of dominance, and slightly sensitive to the actual values of penetrance, it is not greatly affected by errors in the disease gene frequency [6]. On the other hand, model-based linkage analysis can accommodate complications derived from genetic heterogeneity (when the same clinical phenotype is caused by mutations in different independent genes), incomplete penetrance (when the probability of manifesting the disease is <100%, even in individuals with the susceptible genotypes) and presence of phenocopies (when even within the same families there may be individuals affected by the same disease due to different, possibly nongenetic, causes) [1]. Different algorithms have been proposed and are implemented in statistical genetic software to calculate the likelihood of the observed pedigrees as a function of the recombination fraction between the disease and the marker loci, and of the parameters that define the mode of inheritance of the disease in the same pedigrees (TABLE 1) [3–5].

Once the likelihood is defined, in classic linkage analysis, this is evaluated for different values of the recombination fraction varying between 0 (complete linkage) and 50% (independence, or no linkage). The value of the recombination fraction that maximizes the likelihood is taken as its best estimate, and its significance is measured by the maximum lod-score, defined as the $\log_{10}$ of the ratio between the maximum likelihood and the likelihood at θ = 50%. Significant linkage at a given θ is declared when the corresponding lod-score is greater than or equal to 3. This corresponds to a very stringent asymptotic p-value of 0.0001. Historically, this was chosen to compensate for the small *a priori* probability of linkage between two loci taken at random. In more recent studies, a lod-score of 3.3 has been shown to provide a genome-wide significance level of 0.05, when accounting for the multiple tests that are performed when linkage to a disease locus throughout the whole genome is sought [7].

Following the introduction of the DNA polymorphisms [8], linkage analysis by the lod-score method has been used to map the genes responsible for hundreds of relatively rare Mendelian disorders. As of April 2005, Online Mendelian Inheritance in Man (OMIM; a catalog of human genes and Mendelian disorders [201]) gene map lists 1583 loci that have been mapped by

Table 1. Most commonly used software for linkage and association analysis.

Program	Type of analysis	Website address	Main features
GENEHUNTER	Model-based and model-free linkage, QTL, TDT	www.fhcrc.org/labs/kruglyak/Downloads/index.html	Simple to use, it handles medium-size pedigrees and large numbers of markers; includes many types of statistical genetic analysis.
LINKAGE	Model-based linkage	http://linkage.rockefeller.edu or www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html for Fastlink version	The first software available for multipoint linkage analysis, has basically no limitation on pedigree structure but can only handle a limited number of markers
MENDEL	Model-based and model-free linkage, QTL, TDT	www.genetics.ucla.edu/software/mendel5	Performs likelihood-based analysis for a variety of genetic problems in addition to model-based and model-free linkage analysis
MERLIN	Model-based and model-free linkage, QTL	www.sph.umich.edu/csg/abecasis/Merlin	One of the fastest pedigree analysis packages, it performs several types of genetic analyses
SIMWALK2	Model-based and model-free linkage	watson.hgen.pitt.edu/register/soft_doc.html or www.genetics.ucla.edu/software/simwalk2	Analyzes any size of pedigree and very large numbers of markers by means of MCMC and simulated annealing algorithms
VITESSE	Model-based linkage	http://watson.hgen.pitt.edu/register/soft_doc.html	A modified version of Linkage that is faster and allows the analysis of a larger number of markers
ALLASS	Fine mapping by LD analysis of binary traits	http://cedar.genetics.soton.ac.uk/pub/PROGRAMS/ALLASS	Implements the Malecot model for localizing disease genes by allelic association in a set of disease and normal haplotypes
DHSMAP	Fine mapping by LD analysis of binary traits	http://galton.uchicago.edu/~mcpeek/software/dhsmap	Estimates the location of susceptibility genes from marker haplotypes or genotypes in affected individuals and controls
FBAT	Association of binary, measured and time-to-onset traits	www.biostat.harvard.edu/~fbat/fbat.htm	Well-documented software to test for association between disease phenotypes and haplotypes by utilizing family-based controls
HAPLOVIEW	Association of binary traits	www.broad.mit.edu/mpg/haploview	Performs LD and haplotype block analysis, single SNP and haplotype association tests, permutation tests and tagSNP selection
HAPLO_stats	Association of binary, ordinal and quantitative traits	http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm	S-PLUS/R routines for haplotype analysis. Provide score statistics for associations for a variety of traits, with the possible inclusion of covariates
UNPHASED	Association of binary and quantitative traits	www.mrc-bsu.cam.ac.uk/personal/frank/software/unphased	Performs association analysis of multilocus haplotypes from genotype data in trios, case/control sets and general pedigrees
TRANSMIT	Association of binary traits	www-gene.cimr.cam.ac.uk/clayton/software	Transmission disequilibrium testing for marker haplotypes based on several closely linked markers. Allows parental genotype and/or haplotype phase to be missing
QTDT	Association of binary and quantitative traits	www.sph.umich.edu/csg/abecasis/QTDT	Performs several family-based tests of LD for quantitative traits
LOKI	Model-based QTL linkage analysis	http://loki.homeunix.net	MCMC-based software primarily for segregation and linkage analysis of quantitative traits in large and complex pedigrees
SOLAR	Variance component QTL linkage analysis	www.sfbr.org/solar/index.html	Software package for quantitative linkage analysis, allows modeling multiple loci, dominance and epistatic effects. Also includes bivariate analysis of two quantitative traits

Information on these and many other softwares available for statistical genetic analysis can be found at http://linkage.rockefeller.edu/soft.
LD: Linkage disequilibrium; MCMC: Markov Chain Monte Carlo; QTL: Quantitative trait locus; SNP: Single nucleotide polymorphism; TDT: Transmission disequilibrium test.

linkage analysis in pedigrees using DNA polymorphisms, out of a total of 9216 entries [202]. Starting from their localization by genetic linkage analysis, many of the genes responsible for rare Mendelian disorders have been subsequently identified thanks to ever more powerful molecular approaches that can now also take advantage of the knowledge of the human genome sequence. In contrast, the application of the same positional approach to the identification of genes responsible for more common diseases such as schizophrenia, diabetes, hypertension and cancer has proven less successful. These are also known as complex or multifactorial diseases to indicate that multiple genetic factors as well as nongenetic ones (such as chance or environment) and their interactions contribute to increased susceptibility of developing the disease. One classic example of such a disease is Type 1 Diabetes (T1D) [9]. Several epidemiologic studies have shown that genetic factors play an important role in determining susceptibility to T1D, as indicated, for example, by the increased concordance rate in monozygotic compared with dizygotic twins. However, the same studies show that T1D is likely to arise from multiple genes in addition to environmental risk factors, as well as the interactions between them. One major candidate gene locus (*IDDM1*) for T1D susceptibility is represented by the human leukocyte antigen (HLA) region on chromosome 6. A second locus, *IDDM2*, has been identified by association to alleles of a variable number of tandem repeat (VNTR) polymorphism in the insulin gene (*INS*) on chromosome 11. However, association to HLA haplotypes and *INS* VNTR alleles only partially explains the genetic predisposition to T1D. Many other loci and candidate genes have been implicated in different studies (for an updated list, see [203]) but results are often conflicting, and no other genetic risk factor has been identified unequivocally.

In general, after several years of research into the genetic basis of common diseases, only a few loci have been mapped consistently across different studies, and even fewer genes have actually been identified following their putative localization by genetic linkage analysis.

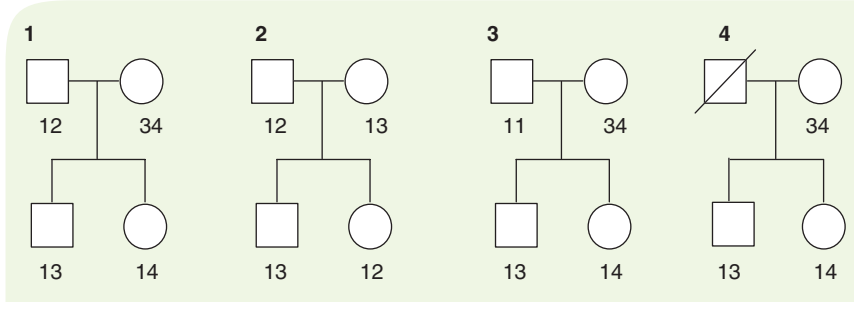

Figure 2. Identity by descent (IBD) versus identity by state (IBS). Genotypes at a marker with four different alleles are indicated below each individual. **(Pedigree 1)** The two sibs share allele 1 IBD as it has been transmitted to both of them from the father. **(Pedigree 2)** Allele 1 in the two sibs is IBS but not IBD (assuming the parents are unrelated) since it has different parental origin. **(Pedigree 3)** It is impossible to tell whether allele 1 in the two sibs is IBD or just IBS because the father is homozygous for allele 1 (and thus not informative). **(Pedigree 4)** A probability of allele 1 being IBD can be calculated based on its population frequency, and thus the probability of the father carrying 1 or 2 copies of the same allele. IBD: Identical by descent; IBS: Identical by state; Sib: Sibling.

Model-free linkage analysis

A possible limitation of the standard linkage approach by means of the lod-score method to the analysis of complex traits is the need to define a mode of inheritance for the disease. In view of the complexity underlying disease transmission, methods that do not require an explicit definition of the genetic model parameters may be more appropriate to the analysis of complex traits. These methods are known as nonparametric or model-free, in contrast to the model-based or parametric lod-score method. The classic model-free methods are based on the concept of identity by descent sharing among relatives. Two alleles observed in two different individuals are termed identical by descent (IBD) when they are in fact copies of the same allele originally present in a common ancestor to the two individuals. In contrast, two alleles can be identical by state (IBS) when they are identical from the molecular point of view, but they do not share a common origin, or their common origin cannot be unequivocally determined (FIGURE 2). Two relatives affected by the same disease are expected to share IBD alleles at the disease susceptibility loci, and at all the marker loci tightly linked to it. It is possible to predict the proportion of marker alleles that a group of affected relatives of a given degree will share IBD when there is no linkage between the disease and the marker itself [10]. A deviation from the expected IBD sharing towards increased observed IBD sharing is thus taken as an indication of linkage between the disease and the marker locus (FIGURE 3). All this can be accomplished without specification of a mode of transmission for the disease. Sibling pair (sibpair) linkage analysis was first described by LS Penrose in 1935 [11], and the affected sibpairs (ASPs) design is still the most commonly used of the model-free methods of linkage analysis. Various statistical tests with different properties depending on the underlying disease model have been proposed [12–15]. Most of these tests can accommodate a variety of situations including extension to different types of relative pairs, extended pedigrees, missing genotype information, uncertain IBD status, multipoint (i.e., including multiple marker loci) analysis and so on. Some of the software that perform model-free linkage analysis in pedigrees of various sizes and with variable number of markers are indicated in TABLE 1. A comparison of several model-free tests specific to simpler study designs such as ASP or small nuclear pedigrees and the corresponding softwares can be found in [16].

Model-free methods have the advantage that they are relatively simple to use, and are particularly appealing in the analysis of complex traits as these traits are expected to be more common in the population then those caused by highly penetrant genes, and are more likely to be identified in smaller families than in large multiplex families. However, the main disadvantage over model-based methods is that they often disregard available data, for example,

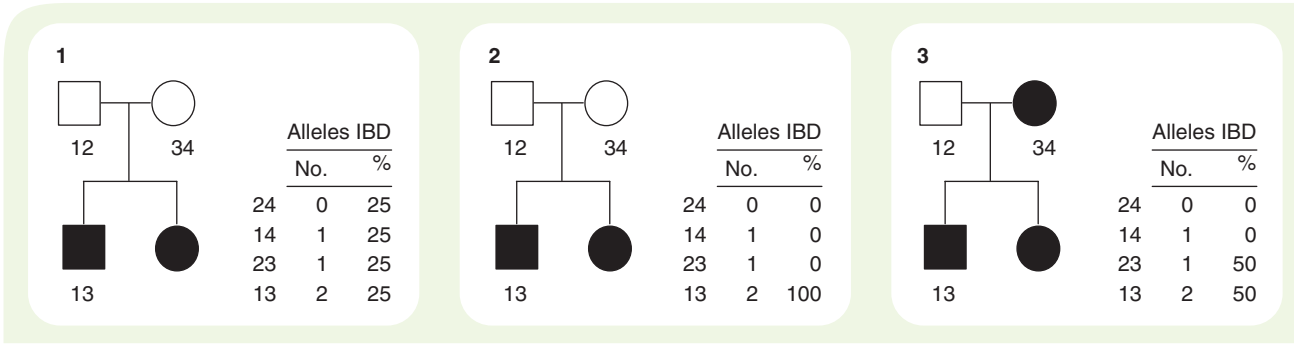

Figure 3. Alleles shared IBD by an affected sibpair. On average, 50% of alleles are shared IBD in a sample of sibpairs. Given the first affected child's genotype at a marker locus (1/3), the second affected child will show 0 (genotype 2/4), 1 (genotype 1/4 or 2/3) or 2 (genotype 1/3) alleles IBD with probabilities of 25, 50 and 25%, respectively, if marker and disease loci segregate independently **(Pedigree 1)**. The extreme cases of complete linkage between the marker and the disease loci in the case of recessive (**Pedigree 2**; both marker alleles are shared by the sibpair, average IBD sharing = 100%) and dominant transmission (**Pedigree 3**; one marker allele is shared by the sibpair, average IBD sharing = 75%) are also shown. In general, increased IBD sharing (>50% for affected sibpairs) is indicative of linkage between disease and marker loci.
IBD: Identical by descent; Sibpair: Sibling pair.

by not exploiting all the information provided by unaffected relatives, and therefore, they are usually less powerful than a correctly specified model-based linkage analysis. In fact, it has been suggested that a test based on the highest lod-score from two simple model-based analyses, one under a recessive and one under a dominant mode of inheritance at a fixed intermediate penetrance, can be at least as powerful as and often more powerful than a model-free linkage analysis under several disease gene models [17]. In any case, model-free methods are now used routinely in linkage analysis, and examples of success in the identification of genes responsible of susceptibility to complex traits are becoming available [18–21]. Notably, the case of the *CARD15*/*NOD2* gene and Crohn's disease (CD) is often cited as a successful application of the positional strategy to complex disease gene identification [19]. In this case, a clear increase in disease risk, particularly in the homozygous state, has helped single out three rare *CARD15* variants that are responsible for CD in a large region of chromosome 16, previously identified by model-free linkage analysis.

Association studies of binary traits

Association studies are a different class of model-free methods for disease gene identification. Theoretical studies have suggested that under some conditions, association-based studies would provide more power to detect genes of modest effect on disease risk (genes that confer less than twofold increase of risk in heterozygous individuals compared with homozygous individuals with the wild-type allele) compared with linkage approaches that would require unrealistically large samples to identify the same weak-effect genes [22]. The classic genetic association study is based on comparison of allele frequencies in unrelated individuals from population-based samples. However, different strategies are used in different contexts, and the authors aim to provide a general summary of the statistical approaches used in association studies and their applications. A few of the most commonly used softwares in association studies are listed in TABLE 1.

Candidate genes

Direct association studies are used to investigate the contribution to disease of specific candidate genes, for which there is evidence of a possible role in disease etiology, and/or are located in regions identified through linkage analysis. Direct association studies usually involve selection of potential susceptibility variants (e.g., nonsynonymous polymorphisms in coding and regulatory regions) [23]. The statistical approach is straightforward and is based on the classic case-control study design, where a sample of patients with a given disease (cases) and a sample of unrelated, unaffected individuals (controls), properly matched with the cases for factors that may be important in the disease etiology, are collected from the same population. Comparison of allele or genotype frequencies in the two samples can be carried out by χ^2 tests for contingency tables, Armitage's test for trend, or log-linear methods, with the expectation that a risk-conferring variant will be more common in cases than among controls. Genotype-based and allele positivity tests (performed by pooling individuals with at least one copy of the susceptibility variant – either heterozygotes or homozygotes) are suitable for the detection of susceptibility alleles that show a dominant or recessive mode of inheritance.

Tests based on Hardy–Weinberg equilibrium (HWE) – that is, based on deviation of observed genotype frequencies from those expected on the basis of the product of the frequencies of the composite alleles – have also been suggested to identify associations with functional variants [24]. Since, when collecting cases, genotypes are sampled proportionally to the rate of disease susceptibility that they confer, departures from HWE are expected in the case sample, depending on both the effects and the mode in which the alleles interact within a genotype to confer disease risk [25].

Allele-based tests are more powerful for the identification of susceptibility alleles showing a multiplicative mode of inheritance, but they assume HWE at population level [26]. When alleles act in a multiplicative manner to cause increased levels

of disease risk, deviations from HWE are not expected in the case sample, as genotypes are selected proportionally to the product of the allele frequencies. If this is the case, and if HWE holds at the population level, it implies that each subject's two chromosomes can be sampled independently from the population and chromosomes can be treated as independent units. Allele-based tests practically double the sample size compared with genotype-based tests, and are therefore more powerful [26].

Although straightforward in principle, the direct strategy would imply comprehensive analysis of a candidate gene through resequencing to search for all polymorphisms within its coding and regulatory regions, and as such is limited by our present incomplete knowledge about functional variation. Typically, association studies of candidate genes use an indirect approach, in which several neutral polymorphisms, mostly single nucleotide polymorphisms (SNPs), typed in or closely adjacent to a candidate gene, are tested for association with the disease. Association in this case may arise as a result of linkage disequilibrium (LD) between the risk-conferring variant and nearby marker alleles.

Linkage disequilibrium mapping

LD is defined as the nonrandom assortment at the population level of alleles at different loci. If alleles A and B at two loci are in LD, they are more likely to be found on the same haplotype, so that the frequency of the haplotype AB would be $p_A p_B + D$, where $p_A p_B$ is the haplotype frequency under equilibrium (the product of the population frequencies of the constituent alleles). LD between two loci can be summarized by the pairwise disequilibrium coefficient, D, calculated as the difference between the observed and expected haplotype frequency in the population. For instance, for two biallelic loci with alleles A, a and B, b, respectively, $D = p_{AB} - p_A p_B$. Two widely used measures of LD based on D are $D' = |D/Dmax|$, and $r^2 = D^2/(p_A p_a p_B p_b)$. D' tends to be favored for LD-based mapping studies as it more closely reflects the extent of recombination between the two loci. The correlation coefficient between the two loci, r^2, is often used to evaluate whether one locus can be substituted for the other (e.g., in the search for so-called tagSNPs; see below) without loss of information. Both these measures equal 0.0 when there is independence between the two loci, and 1.0 in the case of complete disequilibrium. An overview of the different measures of LD and their utility in LD mapping can be found in Devlin and Risch [27].

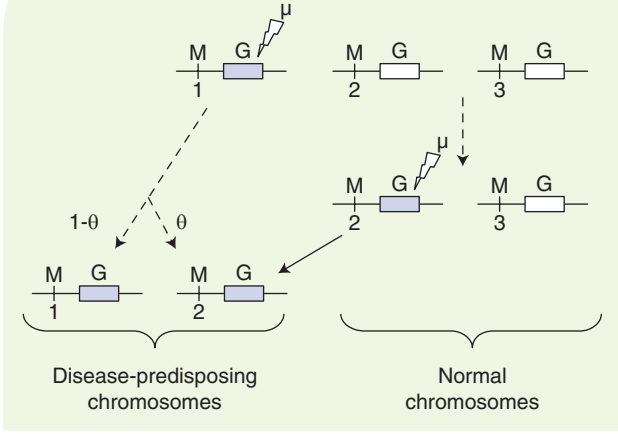

Figure 4. The process that leads to an increased frequency of a specific allele (allele 1 in this example) at a marker, M, closely linked to a gene, G, in the current population sample of disease–predisposing chromosomes, G shaded, compared with normal chromosomes. After a disease-predisposing mutation has occurred, chromosomes carrying the mutation and normal chromosomes evolve independently and will be characterized by different marker alleles. Recombination between the two loci, which occurs with probability θ, and occurrence of another disease-predisposing mutation at gene G, which occurs with a probability μ, homogenize these populations.

Of interest in LD mapping is nonindependence between alleles of tightly linked loci. If a disease-causing mutation occurred some time in the past on a given chromosome (FIGURE 4), then LD between the mutation and the alleles at other loci on that chromosome will be complete (one finds the mutation only in the presence of that specific set of marker alleles). Once LD is created, recombination between loci causes it to decay over time, and the haplotype coupled with the ancestral mutation will tend to shrink in successive generations due to recombination. As a result, of the alleles present on the ancestral chromosome, only those at marker loci tightly linked to the disease locus are expected to be in LD with the disease mutation after many generations (FIGURE 5). This phenomenon is exploited in LD mapping. If affected individuals have received a copy of the same ancestral mutation from a common ancestor, they will share a set of common alleles at loci around the mutated locus at an increased frequency compared with unaffected individuals. Since LD decays quite rapidly with distance in random-mating populations, the region of shared DNA among affected subjects in which the disease mutation must lie is expected to be narrowed significantly by historic recombination events.

Aspects of the demographic population history, unrelated to the disease, also play a critical role in LD structure. Although mutation and recombination have the clearest impact on LD, a range of demographic and evolutionary factors may have a significant effect on the extent and distribution of LD [28], such as admixture (the union of two or more genetically distinct populations) and migration (the introduction of new genes from one previously distinct population into another). Initially, LD might also exist between loci on different chromosomes, but as LD between unlinked markers is eroded rapidly by recombination in successive generations, only LD between nearby markers is maintained [29].

Isolated populations are often considered advantageous for association mapping, and extensive LD was demonstrated around rare variants, and thus most of the affected individuals in the current population shared a single ancestral chromosome inherited IBD from a common founder [30]. The degree to which the same will be true for more common, older variants is uncertain. In constant-sized, finite populations, the disease mutation and the associated haplotype might increase in

frequency by genetic drift (i.e., random changes in allele and haplotype frequencies from one generation to another as a result of chance events), thus also facilitating LD studies for common variants [31]. Another potential advantage of isolated populations that are recently derived from a limited pool of individuals (founder populations) is that the number of variants underlying susceptibility to a disease may be reduced compared with large outbred populations, thereby making association easier to detect. Excellent overviews on the many factors that influence LD-mapping studies can be found in [32,33].

Fine mapping by linkage disequilibrium analysis

LD mapping can be applied not only to candidate genes, but also to further examine regions of the genome that have been identified through linkage analysis. Linkage analysis typically localizes complex disease genes to relatively broad regions, often exceeding 10–20 cM, which might include hundreds of genes. If there are no natural candidate genes in the critical region, various fine-mapping methods can be used to refine the location or to estimate the position of the disease-causing mutation within the region. Collins and Morton described a likelihood-based approach that allows estimating the location of a disease gene in a high-resolution marker map, adopting the Malecot model for isolation by distance [34]. Haplotype-based fine-mapping approaches have been proposed that account for the dependence among case haplotypes and for the presence of multiple causative mutations. Many methods use a coalescence approach to reconstruct a tree structure based on mutations (rather than recombinations) for the present-day haplotypes back to their most recent common ancestor. Due to the complexity of the likelihood models, Markov Chain Monte Carlo (MCMC) methods are generally used. Some of these approaches are described in [35–39]. Other methods rely on the evaluation of haplotype sharing, as haplotypes around a causal locus will be more similar among cases than in controls. Te Meerman and coworkers observed that affected individuals will share a longer haplotype IBD than unaffected control subjects [40]. McPeek and Strahs developed a multilocus model for LD mapping, based on the decay of haplotype sharing, which is most appropriate for situations in which small regions of the ancestral haplotype around a single variant are preserved in the extant case sample [41].

Haplotype-based association analysis

Biallelic SNPs are the markers of choice for LD mapping due to their high frequency in the genome, low mutation rates and amenability to automation. Methods have also been generalized to deal with multiallelic markers, for which degrees of freedom increase dramatically with the number of alleles and problems of sparse data in contingency tables are potentially encountered [42,43].

LD analysis can be performed by testing each marker singularly, but most recent statistical approaches focus on haplotypes (i.e., combinations of alleles at multiple markers) rather than single-marker analysis. The development of methods for haplotype-based analysis is probably one of the most active and rapidly expanding areas of statistical research in genetics as a consequence of the availability of a large number of densely spaced SNPs. These methods can be more powerful as they allow the capture of the ancestral structure of the chromosome(s) flanking the susceptibility variant(s). In the direct strategy, a haplotype-based test applied to a candidate gene involves testing functional variants at different coding sites, and it might result in a more powerful test when specific combinations of the variants on the same chromosome (variants said to be in *cis*-position) have different effects on disease risk. The extent and importance of *cis*-interactions in the human genome are as yet unknown, although examples in humans of super-allele effects have been reported [44–46].

Haplotype-based methods usually require information about phase (i.e., information on which alleles at different linked loci are located on the same homologous chromosome), which is not immediately available from simple genotype data. Phase could be derived either using molecular techniques, or by typing additional relatives, but both methods are costly and often impractical, and in some cases, phase ambiguity may remain. Statistical methods have been developed to infer most likely haplotypes or all possible ones consistent with the observed data (for a detailed coverage of recent computational methods on this subject, see Niu [47]). However, it is not recommended to use statistically inferred most likely haplotypes as if they were directly observed, and phase ambiguity should be taken into account in the statistical method used. Alternatively, multilocus

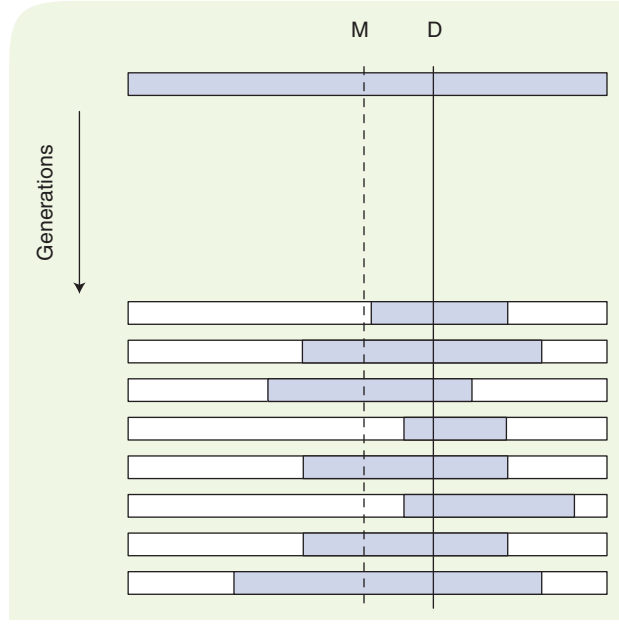

Figure 5. Ancestral haplotype carrying a disease mutation will tend to shrink in successive generations due to recombination. As a result, of the alleles present on the ancestral haplotype, only those at marker loci tightly linked to the disease locus D are expected to be in linkage disequilibrium with a disease mutation after many generations.

approaches, based on scores of allele counts at multiple loci rather than haplotypes, allow circumventing the problem of reconstructing haplotypes statistically and have been shown to provide increased power under certain circumstances [48,49].

Methods that evaluate haplotype-based association treat haplotypes as categorical variables and usually follow traditional statistical approaches. Maximum likelihood haplotype frequencies can be estimated (e.g., using the EM algorithm described in Excoffier and Slatkin [50]) separately in cases and controls and compared with those estimated in the joint sample in order to test if significant differences can be detected between the two samples by a likelihood ratio test (through a χ^2 statistic, or an exact test obtained from permutations) [51]. A limitation of this approach is that it implies reconstructing haplotypes from unphased genotype data separately for cases and controls, and thus assumes that HWE also holds in the case sample, which is true only if the causative locus (and consequently markers in strong LD with it) has an underlying multiplicative model on the genotype relative risk [26].

More general regression methods, which also provide a flexible means to include environmental effects, have recently been extended to account for uncertain haplotype estimation. Schaid and coworkers proposed a score test embedded within the classic Generalized Linear Modeling (GLM) framework for testing haplotype–phenotype association that can handle a variety of phenotypes (binary, ordered and quantitative) [52]. An alternative score test is described by Zaykin and coworkers [53]. An advantage of the score statistic is that it is quick to compute and therefore allows the estimation of empirical significance by simulations.

When many markers are considered jointly, a large number of haplotype configurations become possible, many of which are expected to be rare. For rare haplotypes, a study might have insufficient power to detect association, even if association is present. A possible solution, although not completely satisfactory, can be to set a minimum haplotype frequency threshold, and to group all rarer haplotypes together. Alternative methods group haplotypes based on specific measures of similarity among them, because similarity is likely to be determined by shared genealogy [54]. Finally, an additional active area regards the development of methods based on haplotype clustering and cladistic analysis [55–57]. A comprehensive review on current haplotype-based approaches is provided by Schaid [58].

Population stratification

Association study designs that make use of unrelated cases and controls are popular due to their efficiency and the ease of recruiting subjects. On the other hand, the case-control design has been criticized due to the potential for spurious associations due to population stratification – the existence of genetically different subgroups in the population under study. Population stratification, sometimes known as 'substructure', implies the existence of genetically different groups in the population under study. It occurs, for example, when cases and controls are not well matched ethnically or when people in the population under study have not mated randomly for several generations. Spurious association arises in a stratified sample when disease prevalence and marker allele frequencies differ among the subpopulations. The impact of erroneous association due to population stratification in case-control studies has been debated [59,60].

One approach for dealing with population stratification is to match the ethnic backgrounds of patients to controls as much as possible, although a certain amount of cryptic stratification may remain undetected. If stratification exists in a population sample, methods have been devised to adjust the association test using the information on unlinked markers (as stratification acts on the whole genome and not only locally) [61]. Other methods use unlinked markers to infer the underlying structure and then condition on the structure in the test for association [62,63].

Family-based approaches to association studies

The first attempt to tackle the problem of population stratification by means of family-based studies was undertaken by Falk and Rubinstein with their haplotype relative risk (HRR) approach [64], further developed by Terwilliger and Ott [65], and later with the transmission disequilibrium test (TDT) [66]. The rationale behind the TDT is that under the null hypothesis of the absence of both linkage and association between marker and disease loci, marker alleles will be transmitted randomly from parents to offspring. The TDT compares the frequency of transmission versus nontransmission of marker alleles to affected offspring by means of a simple χ^2 test. The strength of the TDT is that it eliminates population stratification effects completely, but is less powerful than population-based approaches when there is no population stratification. The original formulation of the TDT clearly requires the collection of DNA sample from parents of the affected individuals, which may be difficult to obtain for example for late-onset diseases. Many subsequent modifications to the TDT have been proposed, allowing for other types of relatives as controls [67], unavailable parental data [68], multiallelic markers [69], multimarker analyses [70] and families with multiple affected individuals [71]. Some methods allow the detection of association while taking into account linkage, either in nuclear families [72] or in pedigree of any size [71], while others are proposed for joint linkage and association analysis in a more powerful manner [73,74].

Genome-wide association mapping

Currently, a central issue in human genetics is whether it would be possible to use LD mapping strategies on the whole genome to identify disease genes underlying complex traits, considering the improved techniques for high-throughput identification and genotyping of polymorphisms that offer the possibility of interrogating hundreds of thousands of SNPs across the genome [75,76]. The extent and patterns of LD across the genome are crucial parameters for defining the statistical power of large-scale association studies. Strong LD in a region implies that most of the variation can be captured by a reduced number of well-chosen SNPs (tagSNPs), thus reducing the genotyping of redundant

SNPs. Large projects are currently underway (such as the International HapMap Consortium [204]) that will allow guiding the selection and spacing of SNPs that are useful in candidate genes, candidate regions, and possibly whole-genome association studies. There has been much emphasis on determining the haplotype block structure of the human genome [77,78]. A variety of definitions for haplotype blocks have been proposed, and there is an emerging literature on methods for choosing an optimal set of haplotype tagSNPs (htSNPs), in order to capture the haplotype structure of the genome [79–81]. The utility of haplotype block definition, motivated by the hypothesis that common variants play an important role in the etiology of common diseases [82] in the discovery of disease-susceptibility genes of a more general spectrum, remains uncertain [83]. An alternative approach, which is not based on haplotype block definitions, has also been proposed based on LD maps that describe the pattern of LD in a chromosomal region through a measure of disequilibrium between pairs of SNPs (LD units) [84]. LD unit maps provide a scale on which to distribute SNPs for association mapping, in a fashion analogous to the recombination maps expressed in centiMorgans and used to guide linkage studies.

Since causal variants that contribute to complex traits are likely to have modest effects, large sample sizes are crucial in association studies, although ultimately, power depends on many different factors that jointly influence the probability of success. In particular, the effect (or penetrance) of the causal variant(s) and its/their frequency(ies) in the population will affect both direct and indirect approaches; whereas, in the indirect approach, the frequency(ies) of the marker allele(s) that is/are correlated with the causal variant(s) and the extent of LD between them will also play a crucial role [75]. Large sample sizes are also important in achieving a statistically significant association, particularly in a genome-wide context, since more stringent thresholds for individual tests are needed when performing multiple comparisons using many markers. For instance, a pointwise significance of 5×10^{-8} after Bonferroni correction has been suggested in order to preserve a genome-wide Type I error rate of 0.05 [22]. This would require a prohibitively large sample and represents an over-conservative threshold, because the tests at different markers are likely to be correlated. As an alternative, permutation testing allows evaluation of how often an observed test result would occur by chance if the study was repeated a large number of times (e.g., 10,000 times) under the null hypothesis of no association. The null hypothesis of no association is obtained by shuffling case and control labels in each permuted set. Evaluation of the proportion of the permuted samples in which the test statistic exceeds a given threshold provides an estimate of the empirical significance for the test statistic observed in the actual data set [85].

Analysis of quantitative traits

Linkage analysis

The effects of multiple genes, environmental factors and their interactions may often lead to similar phenotypes, obscuring the inheritance pattern of common diseases. Several approaches have been taken in order to reduce the expected genetic heterogeneity and environmental noise, and to improve the power of linkage methods for common diseases. One approach relies on a suitable selection of the sample. For example, related individuals showing a form of the disease with increased severity are more likely to share the same genetic determinant. This study design has been successful in identifying a genetic variant involved in Alzheimer's disease [86]. Analogously, age at onset of disease information improved the ability to detect genes for breast cancer [87], as early-onset forms are expected to show minor environmental influences.

Instead of reducing the expected etiologic heterogeneity by operating a selection according to phenotypic characteristics, the same goal may be achieved by focusing the linkage studies on a genetically homogeneous population [88]. Indeed, genetically isolated populations with a small number of founders are expected to display reduced disease-influencing variants when compared with outbred populations, and the shared environment and life style are likely to enhance the signal-to-noise ratio of linkage studies. Isolated population studies have proved very valuable for the mapping and positional cloning of genes determining rare recessive Mendelian disorders (e.g., see [89]). Their value in complex disease mapping is unclear, even though the first successes are beginning to appear (e.g., see [90]).

An alternative (or integrative) approach to reducing the etiologic heterogeneity in genetic linkage studies is to focus on quantitative traits that are risk factors for the disease. The underlying quantitative phenotypes that predispose to disease development may be etiologically more homogeneous than the disease itself. For example, the pathogenesis of cardiovascular diseases can result from a combination of several etiologic factors, such as central obesity, serum lipid levels, insulin level and blood pressure, each of them under a varying degree of genetic control. Therefore, the number of susceptibility loci may be so high that it is preferable to study the intermediate phenotypes, since the disease state among affected individuals can be an insensitive indicator of different underlying processes. In other cases, the disease itself is defined on the basis of a rather arbitrary threshold set on a quantitative measurement (e.g., body mass index in the case of obesity, or bone mineral density and osteoporosis) and direct analysis of the quantitative trait itself may more objectively represent the variation observed among different individuals.

A locus underlying variation at a quantitative phenotype is known as a quantitative trait locus (QTL). In some cases, specific effects of QTL alleles on quantitative phenotypes are directly observable. For instance, markers at or near loci of known function (candidate genes) may have the potential to affect the phenotype, either directly or indirectly, through LD. This approach, named the measured genotype approach, enables assignment of phenotypic effects to specific alleles or haplotypes at the candidate loci. The measured genotype test compares by ANOVA the mean values of the quantitative trait in individuals who have either specific alleles or specific genotypes, and it has been successfully applied in human populations [91–93].

When the genetic effect of one or more QTLs is unobservable, their locations and effects are estimated through linkage analysis. Lod-score methods for quantitative traits require the specification of QTL parameters, including allele frequencies, mean effect of the genotypes and genetic variance in the population. The genetic model can be evaluated in a first step through segregation analysis (without markers) and the trait-related parameters estimated are subsequently fixed for the linkage analysis with marker data. However, if the trait model is not particularly accurate, this approach leads to an overestimation of the recombination fraction. Better results can be achieved by simultaneously modeling segregation and linkage analysis. These methods are computationally challenging, even if MCMC methods have provided new tools able to fit such complex models even in extended families [94]. Nevertheless, as previously stated for binary traits, the high parameterization of lod-score methods and the difficulties in obtaining correct parameter estimates had driven the development of various model-free methods. These methods are based upon IBD sharing among relatives, evaluated by observing the segregation pattern of polymorphic genetic markers in the pedigrees. The degree of IBD sharing among different relatives is correlated to trait similarity by either regression or variance component (VC) analysis. In 1972, Haseman and Elston (HE) developed the first regression method for quantitative trait mapping [95]. It is based on the regression of the trait-squared difference in a sibpair on their IBD sharing at a marker. If the marker is linked to the trait, small differences in the trait should be associated with high IBD sharing, and *vice versa*, and the slope of the regression line should therefore be negative. Conversely, a flat regression line would indicate no correlation between IBD sharing and trait-squared difference between the sibpairs, and thus no linkage.

Several authors have subsequently proposed modification of the original test to improve its power, using the squared sum of the trait values together with the squared difference (for a survey on revised HE methods, see [96]). In particular, Sham and coworkers extended the regression-based methods to the analysis of different types of relative pairs, as all these methods were limited to the analysis of sibpair data [97]. Compared with VC methods, regression methods are computationally less demanding and are more robust to violation of normality of the trait distribution. In addition, they are more suited to the analysis of selected samples. Selective sampling of extreme sibpairs often increases the genetic signal, especially for trait loci with low heritability (the proportion of the trait variance attributable to genetic factors), and potentially reduces the costs of the linkage study, provided that the massive phenotyping needed to identify suitable pairs would be cheaper than genotyping [98].

However, when the distribution of the trait is approximately normal, VC linkage methods [99] are more powerful than regression-based methods [100]. Indeed, the VC approach is noteworthy for its generality and flexible modeling capability, as virtually any type of effect and interaction can be easily incorporated.

In VC linkage analysis, the components that model the phenotypic covariance among relatives are estimated by maximum likelihood. For a simple model in which one QTL and residual polygenic effects influence a quantitative trait, the overall phenotypic variance σ_p^2 of the trait is modeled as the sum of the phenotypic variances due to additive effects of the QTL (σ_q^2), an additive polygenic effect σ_a^2, assumed to be due to a large number of unlocalized loci acting additively, and a random environmental deviation σ_e^2. Under the VC model, we can test the null hypothesis that the additive genetic variance due to the QTL equals zero (equivalent to the hypothesis of no linkage) by comparing the likelihood of this model with that of an alternative model in which σ_q^2 is estimated. The statistical evidence for linkage is evaluated by a likelihood ratio test, presented as a lod-score [97].

The additive QTL heritability, or QTL effect size, is expressed by the ratio:

$$h_q^2 = \frac{\sigma_q^2}{\sigma_p^2}$$

and represents the QTL contribution to the total trait phenotypic variance.

Sample sizes commonly used in human linkage studies have the power to detect only QTLs with a major effect on the phenotype – usually not less than 10% of the total phenotypic variance. With underpowered data sets, when a genome-wide scan shows significant linkage to a QTL, the effect size estimate can be upwardly biased [101], especially if the true genetic effect is small. This bias should be kept in mind in result evaluation, such as when attempting to replicate a significant finding in a different sample.

VC methods can be easily extended to complex genetic models, allowing for additional sources of genetic and nongenetic variance. Shared environmental effects can be modeled by the introduction of an additional VC, indicating whether any given relative pair shares or does not share the same environment. Dominance effect as well as interaction among loci can be estimated in the VC model by specification of opportune VCs [102]. Ability to detect epistasis (or gene–gene interactions; see more later) in a QTL linkage framework is low [103]. Purcell and Sham showed that a simple additive variance component model may well be adequate for the detection of QTLs that in reality act epistatically [104].

Linkage of two correlated traits to a single chromosomal region is often taken as further support that a major gene controlling both traits exists in that region. Model-free methods for quantitative traits can be further adapted for the analysis of multivariate data [105]. Amos and coworkers compared HE and VC methods in a multivariate context [106]. They demonstrated that the power of multivariate models is high when polygenic and major gene correlation in traits were opposite in sign.

Association studies

As in the case of binary traits, QTL linkage studies have limited resolution, due to the relatively small number of observable meiotic events [107]. LD mapping provides one possible strategy

for narrowing the site of the QTL causing the phenotypic effect. Moreover, the ability to detect an association between a genetic variant and the quantitative phenotype may be higher than in linkage studies [108], particularly for traits with low heritability [109]. Many of the observations and comments already made for association studies of binary traits also apply to QTL association studies, which can also be broadly categorized into population- and family-based designs. Population-based tests make use of unrelated genotyped and phenotyped individuals. In a simple design, the population sample is stratified by marker genotype, and association between the marker and trait is inferred if there is a significant difference in trait mean between marker genotype classes.

As already mentioned in the case of the binary traits, population-based tests have been criticized for possibly inducing spurious association due to population stratification. Methods such as the TDT originally developed for binary traits [66] have been extended to quantitative trait analysis [110–112] and can be used in the presence of population stratification. These quantitative TDT-like tests examine whether subjects with different genotypes within the same family have different trait means. Since all family members belong to the same subpopulation, significant within-family differences cannot be the result of admixture and are thus indicative of association between the trait and the tested polymorphism(s). A comparison of several different LD tests for identification of loci affecting quantitative traits that use either single individuals or parent–child trios is discussed in [113]. A few of the most commonly used softwares in QTL analysis are listed in TABLE 1.

Gene–gene interactions

Gene–gene interaction (or epistasis) is expected to play a crucial role in complex trait etiology, and statistical methods designed for its detection are receiving increased attention. Most gene mapping methods described thus far are concerned with the analysis of one trait or disease locus at a time. This section will describe methods that explicitly address the simultaneous analysis of multiple unlinked loci all that affect the same trait or disease. Joint analysis of all possible disease loci can confer a clear advantage over locus-by-locus analysis when the underlying genetic model implies strong interaction effects among loci (purely epistatic models) that would have no effect by themselves (no marginal effects). If marginal effects (the effects of each single locus without considering interactions with the others) are detectable, an efficient approach would be to conduct a multistage approach that first identifies significant loci in an initial scan without considering epistasis, and then performs conditional analyses based on the identified loci to search for other interacting loci or evaluates all possible two-way interactions among them.

Statistical procedures to simultaneously analyze more then one locus have been developed for both linkage and association approaches. Specifically, methods that allow testing of two-trait-locus models have been proposed for model-based and model-free linkage analysis [114,115], but their application has been limited to date. Methods that allow the analysis of gene–gene interactions in the context of association studies are, in contrast, receiving increasing attention, although underpowered sample sizes may represent a major limitation of these approaches, as larger sample sizes are required to compensate for the increased number of hypotheses involved in multidimensional screenings.

More recent statistical methods focus on interaction models to detect disease loci with little or no marginal effects. Classic logistic regression methods are not practical when multiple interaction factors are included in the model. The combinatorial partitioning method (CPM) allows evaluation of the effects at many markers jointly, specifically for quantitative traits [116]. An extension of this method is multifactor dimensionality reduction (MDR) applicable to case-control data and discordant sibpair studies, based on pooling multilocus genotypes to reduce the dimensionality of the tests [117]. Hoh and coworkers have proposed a method termed 'set association analysis' to perform a simultaneous significance test on several sets of loci while keeping the overall Type I error in control [118]. Contributions from multiple SNPs in different genomic regions are combined by forming a sum of single-marker statistics, which results in a single test statistic with high power. In Marchini and coworkers, computationally feasible multilocus methods are proposed for large genome-wide association studies [119]. The authors showed that these approaches can be more powerful than traditional analyses, also when using a conservative correction for multiple testing, since the sharp increase in power compensates for the burden of multiple testing. Excellent reviews on methods for detection of gene–gene interactions can be found in [120–122].

Expert commentary

Over the past two decades, genetic linkage and positional cloning methods have been extremely useful in identifying many genes implicated in simple (Mendelian) diseases, whereas little success has been achieved in identifying genetic factors underlying more elusive complex traits. Complex traits are likely to involve several genes, which can be of marginal importance individually, and are characterized by extensive heterogeneity and gene–gene interactions as well as interactions of genetic with nongenetic factors.

All gene-mapping studies aim to detect correlations between the genotypes of marker loci of known genomic locations and the phenotype of interest (which can be a disease or a quantitative trait). While classic model-based approaches have been extremely powerful in the analysis of Mendelian traits, model-free methods are often preferred for their simplicity and independence from specification of disease model parameters for the analysis of complex genetic traits.

Disease-associated quantitative traits (also termed intermediate phenotypes) are increasingly being used as proxy for disease phenotypes in genetic studies of complex diseases. In general, continuous phenotype measurements are inherently more informative, objective and statistically powerful than binary

categorizations of disease status, and they avoid the problems of arbitrary dichotomization. Until recently, it was impracticable to identify the genes that are responsible for variation in continuous traits, or to directly observe the effects of their different alleles. Now, the abundance of genetic markers and sophisticated statistical approaches have made it possible to identify QTL – the regions of a chromosome or, ideally, individual sequence variants that are responsible for trait variation.

LD mapping methods are essential for fine-scale mapping of susceptibility loci, and have been suggested to be more powerful than linkage analyses in detecting weak genetic effects. Most current association mapping is indirect, with reliance on LD, and much recent methodologic work has been conducted to optimize this indirect approach, including the investigation of haplotype block structure and techniques for selecting tagging SNPs. These studies should be adequately designed to take into account the different factors that jointly influence the power of these strategies: the effect size of the susceptibility locus, the frequency of the trait-influencing allele(s), the frequency of the marker allele(s) that are correlated with the trait-influencing allele(s), and the extent of LD between the two. In the study of disease phenotypes, unless only a few mutations account for most instances of disease, the signal will be too inconsistent to find associations, although some degree of heterogeneity is tolerable and newer methods that allow for such heterogeneity by clustering of disease chromosomes have been proposed.

Stringent significance levels should be used in association studies, particularly in the case of whole-genome association studies involving testing of hundreds of thousands of markers. Replication of a significant result is critical, although nonreplication might result from true biologic differences, as when specific susceptibility variants have different frequencies in different populations.

The success of genetic mapping depends heavily on the degree of genetic homogeneity underlying a trait. Population isolates possess many advantages in this regard, and past successes with Mendelian disorders have prompted geneticists to target several population isolates for mapping genes for complex diseases. Populations with reduced genetic variation might be beneficial and have been recommended.

Ultimately, geneticists will need to turn the problem of complex disease gene identification back over to the molecular and cell biologists, as the role of a putative disease gene variant can only be confirmed after a biologically functional effect has been demonstrated.

In this article, the authors have aimed to provide a general overview on statistical genetic methods. Of course, this cannot be complete, as statistical applications in genetics are receiving a great deal of attention, and improvements in methodology are continually being proposed. Care will be required in their design, performance, analysis and interpretation.

Five-year view

The positional approach of genome-wide mapping followed by candidate gene analysis will remain the method of choice for identifying rare, high-risk mutations, with optimal study designs based on linkage analysis in multiplex families. The search for more common, low-risk gene variants that underlie complex traits will remain of central importance in human genetics. This type of study will be based on smaller family units, such as affected sibpairs in the case of linkage analysis, trios composed of parents and one affected child in the case of LD analysis, or unrelated samples of cases and controls for population-based association studies.

Owing to the rapid increase in the availability of large numbers of genetic markers and of the entire human DNA sequence, statistical methods will mainly focus on association studies aimed at the identification of more elusive risk variants. The new data will allow testing the current arsenal of statistical approaches, and prompt for more sophisticated and powerful ones, especially for haplotype and multilocus analyses. These approaches might play an important role in the context of genome-wide association studies, which have the potential to identify genes for common diseases and quantitative traits. Two competitive strategies will be pursued: map-based and gene- or sequence-based. In the former, SNPs are chosen to comprehensively capture the common variation across the genome through LD patterns or haplotypes. Methods for selecting such SNPs and for using them efficiently will continue to be developed and refined.

Parallel to the advance in our knowledge on genomic variation, it will become feasible for association studies to examine all the variants within and around putative genes, and specifically functional variants. Statistical tests will be performed on the main effects of these variants, and haplotype methods will entail testing *cis*-interactions in addition to main effects. Functional genomic technologies involving microarrays and proteomics will add insights regarding gene function and interaction of genes, providing a powerful complement to gene-mapping statistical approaches. Additional advantages of the gene-based approach is that replication is less susceptible to erroneous findings due to genetic differences between populations and that it provides straightforward meta-analysis approaches combining data from multiple studies.

Investigating large numbers of loci leads to a greatly increased number of statistical tests, increasing the possibility of false-positive results. This remains a problematic topic requiring additional statistical genetic research. Rigorous study design, independent replication of data and careful attention to the effects of multiple testing are among the recommendations that will improve the value of association data in the future.

Assessing gene–gene and gene–environment interactions will be necessary, and will require novel statistical methods, whose properties, such as power under different scenarios, still need to be explored.

Finding a disease-associated variant will just be the beginning of applying knowledge of gene variation to human disease to assess their public health and clinical relevance. This is likely to be a difficult task as genetic studies, in contrast to classic epidemiologic studies, use confounding and ascertainment bias to help identify weak genetic effects, thus making quantification of their relevance to public health complex.

Key issues

- Statistical gene mapping allows the identification of the location of disease-susceptibility genes in the human genome in the absence of any knowledge of the function of such genes (positional strategy).
- Classic model-based methods that require specification of genetic parameters, such as penetrance, disease allele frequency, phenocopy and mutation rates, have been extremely successful in the identification of the genes responsible for rare Mendelian disorders.
- Since the inheritance pattern of common traits is usually unknown, researchers often prefer model-free methods when attempting to localize susceptibility genes for complex traits, as these methods do not require specification of a genetic model for the disease.
- With the availability of the human genome sequence and new high-throughput methods for genotyping single nucleotide polymorphisms, association studies are becoming increasingly popular. Genetic association studies offer a potentially powerful approach for detecting common alleles with modest phenotypic effects. Applications include candidate gene analysis, fine-scale mapping, linkage disequilibrium mapping and whole-genome screens.
- A possible alternative in the analysis of complex disorders is the study of quantitative phenotypes associated to disease risk. Statistical methods that allow mapping of quantitative trait loci by linkage analysis in pedigrees are available. In these cases, the classic strategy of initial localization by linkage analysis and subsequent identification by association analysis of positional candidate genes may still prove to be a valuable approach.

References

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

1 Morton NE. Sequential test for the detection of linkage. *Am. J. Hum. Genet.* 7, 277–318 (1955).

2 Ott J (Ed.). *Analysis of Human Genetic Linkage. Third Edition.* The Johns Hopkins University Press, MD, USA (1999).

•• **Still the most comprehensive textbook for classic linkage analysis, it includes references to all the relevant classic papers.**

3 Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. *Hum. Hered.* 21, 523–542 (1971).

4 Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. *Proc. Natl Acad. Sci. USA* 84, 2363–2367 (1987).

5 Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.* 58, 1323–1337 (1996).

6 Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J. Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* 42, 393–399 (1986).

7 Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* 11, 241–247 (1995).

8 Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314–331 (1980).

9 Rich SS, Concannon P. Challenges and strategies for investigating the genetic complexity of common human diseases. *Diabetes* 51(Suppl. 3), S288–S294 (2002).

10 Whittemore AS, Halpern J. Probability of gene identity by descent: computation and applications. *Biometrics* 50, 109–117 (1994).

11 Penrose LS. The detection of autosomal linkage in data that consist of pairs of brothers and sisters of unspecified parentage. *Ann. Eugen.* 6, 133–138 (1935).

12 Blackwelder WC, Elston RC. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet. Epidemiol.* 2, 85–97 (1985).

13 Risch N. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am. J. Hum. Genet.* 46, 242–253 (1990).

•• **Together with two accompanying papers, this is one of the seminal articles on the methodology of linkage analysis of complex traits.**

14 Holmans P. Asymptotic properties of affected-sib-pair linkage analysis. *Am. J. Hum. Genet.* 52, 362–374 (1993).

15 Whittemore AS, Halpern J. A class of tests for linkage using affected pedigree members. *Biometrics* 50, 118–127 (1994).

16 Davis S, Weeks DE. Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. *Am. J. Hum. Genet.* 61, 1431–1444 (1997).

17 Abreu PC, Greenberg DA, Hodge SE. Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases. *Am. J. Hum. Genet.* 65, 847–857 (1999).

18 Horikawa Y, Oda N, Cox NJ *et al.* Genetic variation in the gene encoding calpain-10 is associated with Type 2 diabetes mellitus. *Nature Genet.* 26, 163–175 (2000).

19 Hugot JP, Chamaillard M, Zouali H *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411, 599–603 (2001).

20 Van Eerdewegh P, Little RD, Dupuis J *et al.* Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* 418, 426–430 (2002).

21 Stefansson H, Sigurdsson E, Steinthorsdottir V *et al.* Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet.* 71, 877–892 (2002).

22 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 273, 1516–1517 (1996).

23 Tabor HK, Risch NJ, Myers RM. Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Rev. Genet.* 3, 391–397 (2002).

24 Nielsen DM, Ehm MG, Weir BS. Detecting marker–disease association by testing for Hardy–Weinberg disequilibrium at a marker locus. *Am. J. Hum. Genet.* 63, 1531–1540 (1998).

25 Wittke-Thompson JK, Pluzhnikov A, Cox NJ. Rational inferences about departures from Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* 76, 967–986 (2005).

26 Clayton D. Population association. In: *Handbook of Statistical Genetics.* Balding DJ, Bishop M, Cannings C (Eds), John Wiley & Sons, Ltd, Chichester, UK, 19, 519–540 (2001).

27 Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29, 311–322 (1995).

28 Jorde LB, Watkins WS, Bamshad MJ. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Mol. Genet.* 10, 2199–2207 (2001).

29 Chakraborty R, Weiss KM. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl Acad. Sci. USA* 85, 9119–9123 (1988).

30 de la Chapelle A. Disease gene mapping in isolated human populations: the example of Finland. *J. Med. Genet.* 30, 857–865 (1993).

31 Terwilliger JD, Zollner S, Laan M, Paabo S. Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum. Hered.* 48, 138–154 (1998).

32 Terwilliger JD, Weiss KM. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.* 9, 578–594 (1998).

• **Overview of the different factors influencing association studies, highlighting important points on complex etiology that should be considered for more efficient study design.**

33 Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. *Nature Rev. Genet.* 5, 89–100 (2004).

34 Collins A, Morton NE. Mapping a disease locus by allelic association. *Proc. Natl Acad. Sci. USA.* 95, 1741–1745 (1998).

35 Lam JC, Roeder K, Devlin B. Haplotype fine mapping by evolutionary trees. *Am. J. Hum. Genet.* 66, 659–673 (2000).

36 Liu JS, Sabatti C, Teng J, Keats BJ, Risch N. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* 11, 1716–1724 (2001).

37 Rannala B, Reeve JP. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am. J. Hum. Genet.* 69, 159–178 (2001).

38 Morris AP, Whittaker JC, Balding DJ. Bayesian fine-scale mapping of disease loci, by hidden Markov Models. *Am. J. Hum. Genet.* 67, 155–169 (2000).

39 Morris AP, Whittaker JC, Balding DJ. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am. J. Hum. Genet.* 74, 945–953 (2004).

40 Te Meerman GJ, Van der Meulen MA, Sandkuijl LA. Perspectives of identity by descent (IBD) mapping in founder populations. *Clin. Exp. Allergy* 25(Suppl. 2), 97–102 (1995).

41 McPeek MS, Strahs A. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* 65, 858–875 (1999).

42 Terwilliger JD. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* 56, 777–787 (1995).

43 Sham PC, Curtis D. Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann. Hum. Genet.* 59, 97–105 (1995).

44 Clark AG, Weiss KM, Nickerson DA *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* 63, 595–612 (1998).

45 Hollox EJ, Poulter M, Zvarik M *et al.* Lactase haplotype diversity in the Old World. *Am. J. Hum. Genet.* 68, 160–172 (2001).

46 Tavtigian SV, Simard J, Teng DH *et al.* A candidate prostate cancer susceptibility gene at chromosome 17p. *Nature Genet.* 27, 172–180 (2001).

47 Niu T. Algorithms for inferring haplotypes. *Genet. Epidemiol.* 27, 334–347 (2004).

48 Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* 56, 18–31 (2003).

49 Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. *Genet. Epidemiol.* 27, 415–428 (2004).

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

51 Zhao JH, Curtis D, Sham PC. Model-free analysis and permutation tests for allelic associations. *Hum. Hered.* 50, 133–139 (2000).

52 Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum.Genet.* 70, 425–434 (2002).

53 Zaykin DV, Meng Z, Ghosh SK. Interval estimation of genetic susceptibility for retrospective case-control studies. *BMC Genet.* 5, 9 (2004).

54 Tzeng JY, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.* 72, 891–902 (2003).

55 Templeton AR. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. *Genetics* 140, 403–409 (1995).

56 Seltman H, Roeder K, Devlin B. Evolutionary-based association analysis using haplotype data. *Genet. Epidemiol.* 25, 48–58 (2003).

57 Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am. J. Hum. Genet.* 75, 35–43 (2004).

58 Schaid DJ. Evaluating associations of haplotypes with traits. *Genet. Epidemiol.* 27, 348–364, (2004).

59 Morton NE, Collins A. Tests and estimates of allelic association in complex inheritance. *Proc. Natl Acad. Sci. USA* 95, 11389–11393 (1998).

60 Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol. Biomarkers Prev.* 11, 505–512 (2002).

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

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

63 Purcell S, Sham P. Properties of structured association approaches to detecting population stratification. *Hum. Hered.* 58, 93–107 (2004).

64 Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* 51, 227–233 (1987).

65 Terwilliger JD, Ott J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum. Hered.* 42, 337–346 (1992).

66 Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52, 506–516 (1993).

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

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

69 Sham PC, Curtis D. An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann. Hum. Genet.* 59, 323–336 (1995).

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

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

72 Horvath S, Xu X, Laird NM. The family based association test method: strategies for studying general genotype–phenotype associations. *Eur. J. Hum. Genet.* 9, 301–306 (2001).

73 Goring HH, Terwilliger JD. Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am. J. Hum. Genet.* 66, 1310–1327 (2000).

74 Cantor RM, Chen GK, Pajukanta P, Lange K. Association testing in a linked region using large pedigrees. *Am. J. Hum. Genet.* 76, 538–542 (2005).

75 Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* 6, 95–108 (2005).

•• **With the following paper, this is a comprehensive review on the issues around genome-wide scans for association.**

76 Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* 6, 109–118 (2005).

77 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nature Genet.* 29, 229–232 (2001).

78 Gabriel SB, Schaffner SF, Nguyen H *et al.* The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229 (2002).

79 Johnson GC, Esposito L, Barratt BJ *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* 29, 233–237 (2001).

80 Stram DO, Haiman CA, Hirschhorn JN *et al.* Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* 55, 27–36 (2003).

81 Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. *Trends Genet.* 19, 135–140 (2003).

82 Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet.* 17, 502–510 (2001).

83 Terwilliger JD, Weiss KM. Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann. Med.* 35, 532–544 (2003).

84 Maniatis N, Collins A, Xu C-F *et al.* The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA* 99, 2228–2233 (2002).

85 Dudbridge F, Koeleman BP. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am. J. Hum. Genet.* 75, 424–435 (2004).

86 Levy-Lahad E, Wijsman EM, Nemens E *et al.* A familial Alzheimer's disease locus on chromosome 1. *Science* 269, 970–973 (1995).

87 Hall JM, Lee MK, Newman B *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 250, 1684–1689 (1990).

88 Wright AF, Carothers AD, Pirastu M. Population choice in mapping genes for complex diseases. *Nature Genet.* 23, 397–404 (1999).

• **Considers in some detail the advantages and disadvantages of various populations in mapping common disease genes.**

89 Houwen RH Baharloo S, Blankenship K *et al.* Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* 8, 380–386 (1994).

90 Gianfrancesco F, Esposito T, Ombra MN *et al.* Identification of a novel gene and a common variant associated with uric acid nephrolithiasis in a Sardinian genetic isolate. *Am. J. Hum. Genet.* 72, 1479–1491 (2003).

91 Hopkinson, DA, Spencer N, Harris H. Genetical studies on human red cell acid phosphatase. *Am. J. Hum. Genet.* 16, 141–154 (1964).

92 Sing CF, Davignon J. Role of the apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. *Am. J. Hum. Genet.* 37, 268–285 (1985)

93 Boerwinkle E, Visvikis S, Welsh D, Steinmetz J, Hanash SM, Sing CF. The use of measured genotype information in the analysis of quantitative phenotypes in man. II. The role of the apolipoprotein E polymorphism in determining levels, variability, and covariability of cholesterol, β-lipoprotein, and triglycerides in a sample of unrelated individuals. *Am. J. Med. Genet.* 27, 567–582 (1987).

94 Heath SC. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.* 61, 748–760 (1997).

95 Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* 2, 3–19 (1972).

• **Describes the first regression-based method for quantitative trait linkage analysis.**

96 Feingold E. Regression-based quantitative-trait-locus mapping in the 21st century. *Am. J. Hum. Genet.* 71, 217–222 (2002).

97 Sham PC, Purcell S, Cherny SS, Abecasis GR. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am. J. Hum. Genet.* 71, 238–253 (2002).

98 Risch N, Zhang H. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* 268, 1584–1589 (1995).

99 Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* 54, 535–543 (1994).

100 Fulker DW, Cherny SS. An improved multipoint sibpair analysis of quantitative traits. *Behav. Genet.* 26, 527–532 (1996).

101 Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.* 69, 1357–1369 (2001).

102 Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* 62, 1198–1211 (1998).

• **Describes the implementation and theoretical basis for one of the most commonly used software for variance component linkage analysis**

103 Mitchell BD, Ghosh S, Schneider JL, Birznieks, G, Blangero J. Power of variance component linkage analysis to detect epistasis. *Genet. Epidemiol.* 14, 1017–1022 (1997).

104 Purcell S, Sham PC. Epistasis in quantitative trait locus linkage analysis: interaction or main effect? *Behav. Genet.* 34, 143–152 (2004).

105 Almasy L, Dyer TD, Blangero J. Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. *Genet. Epidemiol.* 14, 953–958 (1997).

106 Amos C, de Andrade M, Zhu D. Comparison of multivariate tests for genetic linkage. *Hum. Hered.* 51, 133–144 (2001).

107 Boehnke M. Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *Am. J. Hum. Genet.* 55, 379–390 (1994).

108 Sham PC, Cherny SS, Purcell S, Hewitt JK. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* 66, 1616–1630 (2000).

109 Mackay TF. The genetic architecture of quantitative traits. *Ann. Rev. Genet.* 35, 303–339 (2001).

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

• **Extends the transmission disequilibrium test (TDT) approach to quantitative trait analysis formulating five new TDT-like tests.**

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

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

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

114 Schork NJ, Boehnke M, Terwilliger JD, Ott J. Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am. J. Hum. Genet.* 53, 1127–1136 (1993).

115 Knapp M, Seuchter SA, Baur MP. Two-locus disease models with two marker loci: the power of affected-sib-pair tests. *Am. J. Hum. Genet.* 55, 1030–1041 (1994).

116 Nelson MR, Kardia SLR, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* 11, 458–470 (2001).

117 Ritchie MD, Hahn LW, Roodi N *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* 69, 138–147 (2001).

118 Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.* 11, 2115–2119 (2001).

119 Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.* 37, 413–417 (2005).

120 Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Rev. Genet.* 4, 701–709 (2003).

• **Detailed review on methods for gene–gene interaction detection.**

121 Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet.* 20, 640–647 (2004).

122 Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* 27, 637–646 (2005).

Websites

201 OMIM – Online Mendelian Inheritance in Man
www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = OMIM
(Viewed August 2005)

202 OMIM Gene map
www.ncbi.nlm.nih.gov/Omim/getmap.cgi
(Viewed August 2005)

203 T1DBase
http://t1dbase.org
(Viewed August 2005)

204 International HapMap Project
www.hapmap.org
(Viewed August 2005)

Affiliations

• *Paola Forabosco, PhD*
Research Scientist, Istituto di Genetica delle Popolazioni – CNR, Alghero, Italy;
Shardna Life Sciences, Cagliari, Italy
Tel.: +39 070 460 6114
Fax: +39 070 460 6151
pforabosco@igp.cnr.it

• *Mario Falchi, PhD*
Senior Genetic Statistician, Twin Research & Genetic Epidemiology Unit, St. Thomas' Hospital, London, UK
Tel.: +44 207 188 6738
Fax: +44 207 188 6718
mario.falchi@kcl.ac.uk

• *Marcella Devoto, PhD*
Head of Genetic Epidemiology Research Laboratory, Nemours Children's Clinic, Department of Biomedical Research, Wilmington, DE, USA;
Associate Professor of Medical Genetics, Dipartimento di Medicina Sperimentale, Universita' La Sapienza, Rome, Italy
Tel.: +1 302 651 6838
Fax: +1 302 651 6895
mdevoto@nemours.org