CHAPTER 15.

Family-based designs

Christopher I. Amos and Christoph Lange

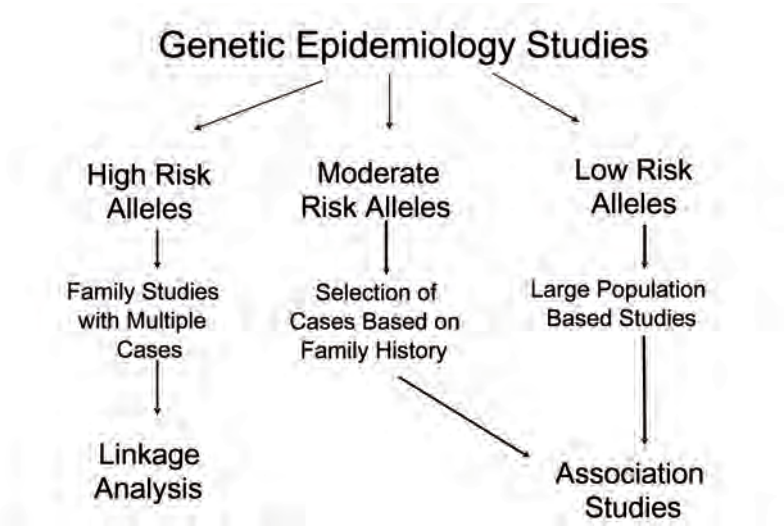
Summary

Family-based designs are used for a variety of reasons in genetic epidemiology, including the initial estimation of the strength of genetic effects for a disease, genetic linkage analysis by which genetic causes can be sublocalized to chromosomal regions, as well as to perform association studies that are not confounded by ethnic background. This chapter describes some of the approaches that are followed in the initial characterizing of genetic components of disease and family-based designs for association analysis and linkage with genetic markers.

Family studies of phenotypes

To obtain an initial assessment of the genetic contributions to disease, and determine which subsequent approach is most likely to be effective, a variety of family-based designs are employed (Figure 15.1). The heritability of a disease indicates the proportion of the covariation in risk for disease that can be attributed to genetic factors. Heritability in the narrow sense excludes covariation due to gene–environment interactions, while in the broad sense includes all genetic contributions to disease. Heritability estimation can be performed using either data from population-based twin registries or from family studies that include different types of relatives. The study of twin registries allows investigators to contrast the similarity in disease among monozygotic twins, who share all their genetic material in common, versus dizygous twins. While the study of twin registries can provide important insights concerning the contribution of genetic factors to disease, twin studies have limitations. For the study of rare diseases, such as cancer, very large collections of twins must be followed for many years. Second, important assumptions that the monozygotic and dizygotic environments are similar are difficult to evaluate. Despite some concerns and weaknesses of this design, twin studies have indicated strong genetic components of risk for the most common cancers (1), as well as many autoimmune conditions (2). Since the development and maintenance of twin registries is beyond the scope of most epidemiologists, this design is not discussed here (see (3,4) for several comprehensive resources).

Figure 15.1. Designs for genetic epidemiological studies to identify genetic factors for diseases



An alternate measure that is more often used to characterize the genetic contribution to disease is the recurrence risk to a class of relatives. For example, for genetically influenced diseases, a monozygous twin who shares all genes in common with their cotwin (a zeroth-degree relative) should have a higher risk of developing disease if their cotwin also has the disease, compared with dizygous twins, siblings or a parent or child who shares only half their genes in common. Each of these pairs of relatives is called a first-degree relative. Similarly, second-degree relatives (half-siblings, avuncular pairs, and grandchild-grandparent pairs) should show even lower risks for disease given that one of the pair members has the disease compared with first- or zeroth-degree relatives. Evidence that there are genetic contributions to disease is found by observing the relative risk for disease either among different classes of relatives, or if population-based estimates are available, by forming the relative recurrence risk by contrasting the risk to relatives of a certain type to the risk in the general population. The easiest such relative risk to estimate is the risk to cosiblings. Relative recurrence risks (RRR) to siblings for cancers range from 2–2.5 for most common epithelial cancers (5), but are much higher for selected cancers, such as non-medullary thyroid cancer (RRR = 15.6), Hodgkin's disease (RRR = 6.5), testicular cancer (RRR = 6.6), ovarian cancer (RRR = 4.9) and renal cancer (RRR = 4.7). Relative recurrence risks are much higher for some cancers if multiple relatives are affected and also higher for relatives of earlier onset cancers.

Contrasting recurrence risks for other types of relative combinations can provide initial insights into whether or not there are recessive or dominant effects for a disease, and the number of genetic factors that are likely to be important in disease causation (6,7). If there are recessive effects influencing disease causation, then risk to monozygous twins who share all their genetic material in common will be much greater than risks to dizygous pairs of siblings. In turn these risks will be higher than the risks to offspring or parents, because parent-offspring pairs never share two alleles in common, while siblings share on average one quarter of the time, and sibling pairs share both alleles in common. Thus, if a disease includes a recessive effect, then a co-sib of an affected individual is also more likely to have two deleterious alleles and be affected than a parent would be (see (8) for a detailed description concerning the estimation of allele frequencies in co-sibs and other relatives). The fall-off of the recurrence can also be used to provide insights into the number of loci that may influence a disease (6).

The probability that an individual becomes affected given that they carry a particular genotype is called the penetrance. The penetrance of disease can depend upon genotype(s) at one or more loci, as well as environmental factors. The genotypic risk ratio is the ratio of the penetrance given that an individual has one particular genotype compared to the risk for disease given another genotype. The relative recurrence risk depends upon the genotypic risk ratio and the population prevalence of the genetic factor (9). The genotypic recurrence risk determines the power of association studies, as described in more detail below.

Standard epidemiological approaches have been modified and further developed for characterizing evidence that genetic factors influence a disease. The most straightforward approach that epidemiologists will initially apply seeks to identify the odds ratio of a disease with family history for the disease in relatives. In this approach, the epidemiologist asks cases and controls to delineate the occurrence of disease among relatives of each type (e.g. siblings and parents). Then, usual

epidemiological approaches such as logistic regression can be used to obtain an odds ratio that a case reports a family history of disease compared with a control. While this approach is similar to other analytical approaches commonly used in epidemiology, a historical cohort approach is preferred by genetic epidemiologists for many reasons. In the historical cohort approach, case and control participants are asked to provide medical history information on selected relatives, such as first-degree relatives. The data that must be collected for each relative includes either the age at disease onset(s), the current age if alive and unaffected, or the age at death if deceased without disease. The relatives of the cases and controls then form a historical cohort with the follow-up period extending from birth until either disease onset or last age (death or current age). There are many advantages of this approach over the case–control design (10). In the historical cohort design, both absolute and relative risks can be obtained. As this is a cohort design, multiple disease endpoints can be studied. In the case–control design, because cases and controls are typically selected not to have multiple diseases, it becomes impossible to evaluate whether one disease, such as rheumatoid arthritis, is associated with an increase in relatives for another disease, such as systemic lupus erythematosis. In addition, according to both the two-stage and multistage models of carcinogenesis, individuals with inherited susceptibility to disease should have higher hazards ratios for cancer(s) at earlier ages compared with older ages. Case–control designs that are typically matched on age have difficulty estimating differential risks for disease according to age, but for the historical cohort design, variation in disease risk according to age can be readily estimated. Because the relatives in a family are correlated, tests of the relative risks for disease are biased unless a variance correction is introduced to allow for this correlation. The Huber-White variance correction procedure can be applied and is readily available in standard analytical packages such as SAS, STATA or R. As an example, the occurrence of rheumatoid arthritis in relatives of cases compared to controls and the occurrence of other autoimmune conditions was studied (11). The results showed that aside from rheumatoid arthritis, which was more frequent in case relatives compared to control relatives, other autoimmune conditions occurred more frequently in relatives of controls.

When a candidate mutation has been studied in a family, an approach to estimate the penetrance specific to that mutation is the kin-cohort approach (12–14). This method takes advantage of the extensive data on family members that can be obtained using the historical cohort approach discussed above, but also allows the penetrance to be estimated specifically from the mutation. Among those probands who are found to have a rare mutation, about 50% carry the mutation, while nearly none of the relatives of probands not carrying the mutation are carriers. By contrasting the age-specific risk in relatives of carriers versus relatives of non-carriers, one can derive an estimate of the penetrance associated with the mutation being studied. An issue in applying this method is how to correct for the selection of probands based upon their being affected when there is risk for disease, not only from the mutation being studied, but also from other loci (14,15). Using these methods, the risk associated with carriage of breast cancer 1 (*BRCA1*) and *BRCA2* mutations could be estimated from the population-based Washington Ashkenazi Study, since the prevalence of mutations in this population was sufficiently high. Using the kin-cohort approach, the risk for breast cancer due to carriage of either of the three common mutations in Ashkenazim was 56% to age 70, which is considerably less than had been estimated previously from the study of families ascertained through multiple affected relatives (16). This variation in risk according to the sampling design likely reflected the incomplete ascertainment correction provided by earlier studies of families that included many affected relatives. Previous approaches to ascertainment correction in family studies derived for linkage analysis conditioned only on the specific measured genetic factors (e.g. *BRCA1* and *BRCA2*) and failed to allow for effects from unmeasured lower-penetrant loci. A more recent alternative approach to the kin-cohort method adapts segregation analysis to incorporate effects from a known measured genetic factor, such as *BRCA1* and *BRCA2*, as well as residual risk from unmeasured genetic factors (17). Application of these methods has yielded penetrance estimates similar to those given by the kin-cohort approach.

Of concern when performing genetic epidemiological studies in which a case or control is interviewed about the occurrence of disease in relatives is the reliability of the reporting by such subjects. Numerous studies have shown that for some common epithelial cancers, such as breast, colon, prostate and lung, reporting of disease in relatives is acceptably

accurate (18,19). For cancers of the internal organs or common metastatic sites, such as ovarian, liver and brain cancers, reliability of reporting is extremely poor (20). Studies of these cancers would entail obtaining medical records to verify reporting by the case or control. Reporting of autoimmune diseases also shows variable reliability, with rheumatoid arthritis, for example often being confused with other types of arthritis. Reports of rheumatoid arthritis in relatives were confirmed using medical records and reports from multiple relatives (11).

Reporting of disease in relatives can raise issues concerning the privacy of the relatives. The American Society of Human Genetics has issued a policy statement that indicates reporting by an individual about a relative is hearsay, and hence does not constitute a violation of privacy (21). However, an evaluation of risk associated with the collection of reported disease in relatives will require Internal Review Board review. Inadequate compliance with an approved protocol for the collection of reported medical data on relatives, led to temporary cessation of research at the University of Virginia, when a father complained that his child was being asked to report sensitive information about him as a part of a research study. In the USA, researchers involved in studies of diseases for which risk can accrue to either the patient or the researcher can obtain a certificate of confidentiality. This certificate protects the research from legal discovery.

Segregation analysis

To more precisely model the familial and genetic factors affecting disease expression, case–control studies have often been followed by segregation analyses. This is particularly useful when initial studies identify high risk associated with a family history of disease, and the disease is rare, suggesting the involvement of one or a few genetic factors having high penetrance. Segregation analyses seek to identify the relationship between an individual's genotype and the resulting phenotype. Inheritance of genetic factors results in a specific form of genotype dependence among family members. Although the genotypes at a disease locus cannot usually be determined, the inheritance of disease within families can be compared with that expected under specific genetic models. In segregation analyses, the model that most closely approximates the observed familial data is sought. The models that are evaluated by classic segregation analyses include a genetic factor, environmental effects which may be correlated among family members, and polygenic effects. These polygenic effects are a mathematical construct that corresponds to the inheritance of many independent genetic factors, each having small effects.

The classic paradigm of segregation analysis also requires scrupulous definition and attention to the ascertainment criteria. For most diseases, the occurrence of genetic susceptibility is sufficiently uncommon that random sampling would result in low power to detect genetic effects. However, most patterns of selection through affected individuals introduce biases into the genetic analyses. When the selection or ascertainment events are well characterized, these biases can often be controlled for appropriate mathematical conditioning (22). For segregation analysis, the units of observation are individuals within families, and although the modeling process is applied to individuals, it also requires information on their close relatives. Thus, the unit of sampling and analysis is the family. Summary statistics from segregation analytic studies include the gene frequency of the disease-causing locus, the penetrance for the susceptible genotypes, and the sporadic risk for the non-susceptible genotypes. During segregation analysis, the parameters describing the penetrance and the gene frequency are inferred using maximum likelihood methods. The parameters that most accurately describe the observed data are identified by computationally intensive numerical evaluations. To allow for the variable size and structure of human families, very general algorithms were developed, largely as a result of seminal works by R.C. Elston (23,24).

Genetic linkage analysis

Genetic linkage analysis has been an extremely powerful tool for identifying specific genetic factors for diseases. Linkage analysis has typically been applied for identifying novel genetic factors by using a genome-wide analysis of the co-inheritance of disease with genetic markers. Evidence in favour of linkage is typically expressed by the LOD score, which is the $\log_{10}$ ratio of the likelihood of the data assuming linkage between a modelled disease susceptibility locus and a genetic marker, to the likelihood of the data assuming no linkage of the disease susceptibility and genetic marker. To allow for the large number of tests that are indicated in a genome-wide analysis, several testing paradigms have been developed. If a Bayesian approach is adopted, a LOD score of about 3.0 leads to a 5% posterior probability of linkage assuming

the existence of a single disease locus, even when many markers are genotyped over the entire genome. An approach for sequentially combining data from multiple studies by adding LOD scores across studies has been highly effective (25). From Bayesian and sequential analytical approaches, a LOD score of 3.0 was proposed as providing a meaningful critical value for declaring strong evidence for linkage. More recently, approaches to control the overall significance of genetic studies when studying multiple markers have been adopted (26). These criteria have been criticized for being excessively conservative (27), particularly when candidate regions are of primary interest (e.g. when prior studies indicated evidence for linkage to an area). The significance testing paradigm requires the slightly higher LOD score of 3.3 to declare that a significant result has been obtained while providing a genome-wide significance of 5%.

If a simple genetic mechanism explains inheritance of disease, then a genetic model can be specified and tested for co-inheritance of disease susceptibility with genetic markers. In order for linkage studies to be informative, the families chosen for study must be able to show inheritance of a genetic factor. For uncommon diseases for which the penetrance is reduced, the affected individuals provide the majority of information about the segregation or inheritance of genetic mutations predisposing to disease. For quantitative traits, sampling through individuals with extreme phenotypes can increase the probability of sampling a genetic variant influencing the trait of interest. Sampling through extreme individuals is an effective strategy for increasing the power of a linkage study, but may only be practical if the quantitative phenotype can be assayed inexpensively. Some studies of quantitative phenotypes look at many phenotypes. Sampling through extreme individuals only increases power for a single or a few correlated phenotypes.

Linkage analyses are mainly conducted using panels of single nucleotide polymorphisms (SNPs) with a density of at least 1 marker every 500 kilobases (usually at least 6K markers), but can also be performed using microsatellite panels with a density of at least 1 marker every 10 megabases (about 350 markers). SNPs are far less informative than microsatellites, so that a much denser mapping panel is required to obtain a comparable amount of information from a genetic study using SNPs compared with one using microsatellites. Evidence for genetic linkage in a region would often be followed by finer-scale mapping to improve the information for detecting linkage and to identify any recombinant individuals. Finer maps would be employed if a microsatellite panel or relatively sparse SNP panel was used, to search for associations between the disease or trait and particular marker alleles. Standard finer mapping panels for microsatellites provide a 0.5 to 0.2 megabase interval spacing (available from Decode Genetics (decodegenetics.com) or Invitrogen Genetics). Routine genotyping platforms for the purposes of genetic linkage analysis are available from Affymetrix and Illumina, and provide results from genotyping of between 6000 and 1 000 000 genome-wide SNPs, respectively. These much finer mapping panels can improve the power to detect linkages and may provide narrower intervals for positional cloning. However, the presence of strong linkage disequilibrium (LD) among the SNPs in these platforms raises many analytical complexities that must be dealt with for accurate inferences. In particular, biases occur when families are selected through multiple disease-affected relatives if LD is not precisely modelled (28).

A wide range of genetic linkage methods are available. The diversity of methods reflects, in part, the considerable success in identifying genetic causes of disease, and the consequent value and interest in using the methods by the scientific community. Computing statistics over a large number of genetic markers in families for diseases that do not show simple inheritance patterns is computationally demanding. There are three basic approaches that are used for analysis of the genetic marker data. The Elston-Stewart algorithm (23) summarizes information about haplotypes (the set of alleles on a chromosome) sequentially in a pedigree, and is therefore efficient for statistical analysis of large families, but limited in the number of markers that can be jointly modelled (usually fewer than five markers can be considered jointly). The Lander-Green-Kruglyak (LGK) method (29) adopts a different approach that facilitates the analysis of multiple markers. The LGK model first identifies the possible inheritance patterns of genotypes within families and stores this information as inheritance vectors. Because the number of inheritance vectors increases rapidly according to the number of individuals in a family, this approach is only suitable for small- or medium-sized families, usually allowing at most 25 individuals in a family to be studied. In addition, because the method stores all possible inheritance vectors in memory, the approach requires considerable RAM to be efficient. The major advantage of the LGK approach is that computational

speed increases only linearly in the number of markers so that it is highly efficient for genome-wide analyses. In addition, the adaptations of the LGK algorithm allow haplotypes to be used as markers, thus allowing for the strong LD that can exist among tightly linked markers (30).

Analyses including many markers on large pedigrees, or analyses of pedigrees that include more than a few inbred individuals, may not be effectively performed using the Elston-Stewart or LGK algorithms. In this case, Monte-Carlo Markov Chain (MCMC) algorithms are used to approximate the likelihood of the data. MCMC methods provide tools for sampling the haplotype configurations in data (31,32). The MCMC procedure samples possible haplotypes according to the underlying probability distribution that generated the data and provides an accurate approximation to the likelihood. A major advantage of MCMC procedures is a decreased need for memory, since they do not require summing over all possible genotypes as in the Elston-Stewart algorithm, or over all possible inheritance vectors as in the LGK. One disadvantage is the complexity in storing output from analyses, since results from large numbers of realizations from the sampling of genotype configurations must be stored. MCMC methods infer the genotypes for all individuals that are specified as a part of the analytical file. Individuals with known genotypes have a limited number of potential haplotypes, but individuals who have not been genotyped can have a large number of potential genotypes and haplotypes. The probability distribution from which MCMC methods must sample can become quite large if many individuals who have not been genotyped are included in the analytical file. Therefore, it is often beneficial to remove the ungenotyped individuals from MCMC analyses, particularly those who are not affected, since they contribute little in most linkage analyses.

An issue in performing genetic analysis is whether to use model-dependent or model-free methods for linkage analysis. Model-dependent methods have higher power for linkage analysis if an approximately valid genetic model can be specified to describe the manner in which disease susceptibility at a given locus is expressed. One approach for estimating penetrance to be used in a linkage study is to first perform a segregation analysis of families that have been ascertained according to a specified sampling scheme. The approach estimates parameters for models describing the inheritance of genetic and environmental factors that most closely fit the dependence in family data. For uncommon conditions, random sampling of families would not result in an informative family; a sampling scheme is usually followed in which relatives of cases with a disease are preferentially sampled. When the families are not randomly sampled, an ascertainment correction for non-random sampling is required to obtain parameter estimates that reflect the more general population of families. To correct for the non-random sampling approach usually used, a clearly defined sampling scheme must typically be followed. Using only a binary phenotype (e.g. affection or non-affection) one may not be able to estimate all the parameters that are necessary to describe the penetrance of the genotypes of the loci influencing disease susceptibility, unless restrictive assumptions about the interactions among the loci are made.

Sampling families and collecting information for segregation analysis can be an arduous task, and may not be fully informative about the parameters that describe the penetrance and disease allele frequencies. Therefore, investigators studying complex diseases may postulate genetic models from assumptions about the relative risks for disease that are observed from epidemiological studies. It has been shown that postulating an inaccurate genetic model for genetic linkage studies does not lead to false-positive results in a model-based linkage study. However, if multiple models are tested, there can be an inflation of the overall number of false-positive results from linkage studies because of the inherent multiple-testing problem that is introduced. A powerful approach for studying complex diseases is to evaluate the evidence for linkage, assuming simple recessive and dominant models of disease, and then to adjust the required critical value for the LOD score upwards by about 0.3 for the small multiple-testing problem so engendered (33).

If the genetic model influencing disease susceptibility cannot be inferred with any confidence, either because the genetic model appears too complex or because there is a lack of epidemiological data from which to postulate penetrance, then model-free methods are typically adopted. One approach is to set the penetrance to an artificially low level, thus restricting analysis to include only the affected subjects. With very low penetrance, unaffected individuals provide no information about their possible genotypes and so do not contribute in the linkage analysis, but this approach still makes some modelling assumptions about disease expression. An alternative approach is to evaluate

the similarity in alleles that have been inherited by common parentage (identity by descent) and test whether or not there is evidence that affected relatives share more alleles than expected identical by descent. In some cases this approach may provide a more powerful test for linkage than a model-dependent approach, particularly when multiple independent loci additively increase disease risk. Because pedigrees are usually variable in size and contain different numbers of affected relatives, a variety of different tests have been proposed and are available for testing for linkage (34,35). These tests are optimal for varying disease penetrances (which are typically unknown). As a compromise, the pairs statistic is often used, which includes all affected relatives in a pedigree and gives only moderately higher weight to families that include multiple affected relatives (29).

The joint analysis of covariates, along with genetic markers in family studies, usually has limited utility. Typically, collecting covariate information in families is difficult because data cannot be directly collected from deceased or otherwise unavailable individuals. In addition, the genetic risks that are sought in linkage analyses are often large. Some non-genetic factors, such as smoking and reproductive behaviours, can be reliably collected through proxies (when needed), are inexpensive to collect, and may have a strong effect upon risk for some diseases.

For complex diseases, a large number of families may be needed to obtain adequate power to detect linkages. Meta-analyses combining multiple studies can assist in overcoming power limitations from a single study. However, in order for meaningful results to be obtained in meta-analyses, investigators must be studying comparable classifications of the same disease. Coordination of studies by using common definitions of disease outcomes, demographic measures and covariates is necessary for the study of complex diseases. Tools for meta-analysis of both linkage and association studies are available (36,37)

Association studies using families

While parametric and non-parametric linkage analysis approaches have proved successful for mapping many disease and trait genes, in some gene mapping investigations the limited number of meioses occurring within pedigrees limit one's ability to detect, by linkage recombination, events between closely spaced (< ~1 cM) loci (38). Association studies might be used instead to map more closely spaced disease genes. These studies generally have a case–control design, where cases are recruited from a disease registry or hospital-based populations. Controls can range from the cases' family members (e.g. parents or siblings), or unrelated individuals. Genetic variants observed in cases are contrasted with those observed among controls to determine if an association exists between genes and disease.

Association studies may permit one to get closer to the disease-causing gene than allowed by linkage studies (i.e. more recombinant events over evolutionary time). This type of study can also be used to directly examine genetic variants in known candidate genes. That is, association studies can be used either in an indirect manner, as a tool for mapping genes using linkage disequilibrium, or in a direct manner, for evaluating associations with postulated causal ("candidate") genes.

The growing use of association studies is driven in part by how quickly and easily they can be undertaken, and the availability of high-density SNP genotyping technology. The SNP consortium (39) has provided sequences for 1.8 million SNPs, and at least 250 000 of these have been confirmed as polymorphic by Perlegen alone, while polymorphisms in hundreds of thousands of additional SNPs have also been verified by the SNP consortium Applera, and by many investigators and companies.

The power to detect associations using unrelated cases and unrelated controls can be increased by selection of cases that are likely to have developed the disease because of increased genetic propensity. For rare or uncommon susceptibility factors, sampling unrelated cases on the basis that they have close relatives affected by the same disease can greatly increase the power to detect associations (40). Power to detect associations can also be accomplished by seeking a homogeneous genetic etiology for the disease, which entails selecting from isolated populations and cases that show a homogeneous clinical phenotype.

Linkage disequilibrium and haplotypes

The genetic variants that cause disease arise through, for example, novel mutations or immigration of mutation carriers into a population. When a mutation initially occurs, it has a particular chromosomal location and specific neighbouring marker alleles. At this incipient point in time, the mutation is completely associated with the adjacent alleles; it is only observed when the marker alleles are also present (41). Marker alleles that were in the neighbourhood of the disease gene

when its mutation was introduced into the population will generally remain nearby over numerous generations, that is to say linkage disequilibrium. One can estimate whether particular marker alleles appear to be in disequilibrium, that is to say, are associated, with disease genes. In particular, if specific marker allele frequencies are higher among cases versus controls, this suggests linkage between the corresponding loci and a disease gene. The extent of this disequilibrium depends on the number of subsequent generations since the mutation was introduced into the population, the recombination between the disease and marker alleles, mutation rates, and selective values (e.g. epistatic components).

Alleles in linkage disequilibrium may be parts of haplotypes. Recent work indicates that there may exist discrete chromosomal regions with low haplotype diversity, termed haplotype blocks, that are separated by recombination hotspots. Information from some polymorphisms within each block may be redundant; in other words, having information on one SNP provides all the information about another if they are in strong linkage disequilibrium. The majority of the haplotypes within a block can thus be distinguished using a much smaller number of SNPs, known as haplotype tagging SNPs (htSNPs). Using such SNPs can drastically reduce the effort required to undertake large scale association studies. Instead of saturating an entire chromosomal region with genotypes in all study samples, an investigator can first screen for SNPs within a subsample of study subjects to determine the htSNPs. Then only these tagging SNPs (and possibly other promising SNPs) can be genotyped in the entire study population. Several approaches have been suggested for identifying optimal htSNPs. These include visual inspection of haplotypes, and analytic approaches that eliminate redundant markers (42–44).

Family-based association studies

The most common familial case–control designs use parents or siblings as controls. In the former, the parents themselves are not the controls, but the set of genotypes the parents could have transmitted to the case, given their own genotypes (the case's "pseudosibs"). For example, the Transmission/Disequilibrium Test (TDT) compares alleles transmitted from parents to diseased offspring with those alleles that are not transmitted (i.e. the non-diseased alleles) (45). The TDT provides a joint test of linkage and association (i.e. linkage in the presence of association or vice-versa). In doing so, when there is disequilibrium between marker and disease alleles, incorporating the additional information that the same alleles are associated across families with the TDT can provide increased power in comparison with linkage analysis. Furthermore, the use of pseudosib controls has better statistical efficiency than sibling or cousin controls (even more than population controls for a recessive gene), but the requirement that parents be available for genotyping limits its usefulness for late-onset diseases.

As with pseudosib controls, siblings are derived from the same gene pool as the cases, and thus provide another attractive source of controls for family-based studies. However, using siblings as controls can pose other difficulties. A major issue is that not every case will have an available sibling. If sibship size or other determinants of availability are associated with genotype, selection bias may result, possibly leading to false-negative or -positive results. Another issue is that controls should generally be selected from siblings who have already survived to the age at diagnosis of the case and be free of the disease. In practice, this will tend to limit control eligibility to older siblings, which can lead to confounding by factors related to year of birth, family size or birth order. Siblings are also more likely to have the same genotype as the case than are unrelated controls, thereby leading to some loss of statistical efficiency (i.e. larger sample sizes required to attain the same statistical precision).

The many successful applications of the TDT motivated the development of a large number of generalizations. The original TDT concept was extended to multiallelic marker data (45–47) and to different genetic models. In the framework of score tests for multivariate data, it has been shown that the TDT is the most powerful test under an additive mode of inheritance; alternative tests can be derived under a dominant and recessive mode of inheritance (48). (Extension to general pedigree designs and to scenarios in which parental genotypes are missing are discussed in (46,49–52)). Approaches to general pedigrees that are also valid under the null hypothesis of linkage, but no association has been developed are discussed in (53–55). Extensions to quantitative traits are described in (52,56–61).) The gamete competition model (62) provides one generalization of the TDT that can be applied to arbitrary pedigrees and extends to haplotype-based analyses. This approach has been integrated into the Mendel suite of programs (http://www.genetics.ucla.edu/software/mendel).

The family-based association tests (FBAT) approach

In this section is a review of a very general and adaptable approach to construct family-based association tests that are often referred to as the FBAT approach (60). FBATs can be applied under any mode of inheritance and in situations in which multiallelic data and/or general pedigrees are available. Various null hypotheses, and different phenotypic traits and arbitrary combinations of them (binary, quantitative, time-to-onset, repeated measurements, multivariate data, etc.), can be tested for association. FBAT can be computed for a single marker locus, haplotypes or multiple markers. The FBAT approach is built on the three key principles of the original TDT approach:

1. The FBAT statistic is a conditional test that conditions upon the parental genotype, or, as will be discussed later, equivalent information if parental data should be missing. By conditioning on the parental information, there is no need to estimate the genotype distribution of the data (e.g. the margins of the table in a case/control design) under the null hypothesis, and thereby eliminate the effects of population admixture. When parental information is missing, one can condition on the sufficient statistic for the genotype distribution in each family. For haplotype analysis, phase uncertainty will also be included in the conditioning.

2. The FBAT statistic is also computed conditional on the phenotype, which makes the approach robust against misspecification of the phenotypic assumptions that are used for the computation of the FBAT statistic.

3. Since the only random variable in the FBAT approach is the offspring genotype, whose distribution under the null hypothesis can be computed based on Mendelian transmission, Mendel's first law is the sole requirement for the validity of the approach.

The general FBAT statistic

The FBAT statistic assesses the association between the phenotype and the genetic locus by using a natural yardstick: the covariance between the phenotype and the Mendelian residuals. The covariance is defined by:

$$U = \Sigma T_{ij} (X_{ij} - E(X_{ij}|S_i)), \quad (1)$$

where i indexes family and j indexes non-founders in the family. The summation is over all families i and all non-founders j. The parameter T_{ij} denotes the coded trait of interest in the jth non-founder of the ith family. The corresponding genotype is given by X_{ij} which is adjusted by its expected value $E(X_{ij}|S_i)$ under the null hypothesis. Using the assumption of Mendelian transmissions from the parents to the offspring, the expected marker score $E(X_{ij}|S_i)$ is computed conditional upon the parental genotypes Si of the ith family. If parental information is missing, S_i denotes the sufficient statistic of the genetic distribution in the ith family. The adjusted genotype, $(X_{ij} - E(X_{ij}|S_i))$, can be interpreted as an Mendelian residual, measuring a potential over- or undertransmission from the parents to the offspring. In this context, it is important to note that the Mendelian residuals for families with two homozygous parents will always be zero and that such families do not contribute to the FBAT statistic. The number of families that have at least one Mendelian residual $(X_{ij} - E(X_{ij}|S_i))$, which based on S_i can be different from zero, is typically referred to as 'number of informative families.'

As discussed below, the coded phenotype T_{ij} is either centred or unadjusted, depending on the absence or presence of a phenotypic ascertainment condition. By selecting appropriate coding functions, qualitative, quantitative, time-to-onset and multivariate phenotypes are incorporated into the FBAT approach.

The basic formula (1) is applicable in virtually any scenario; the appropriate selection of the phenotypic coding function and its adjustment, and the definition of the genotypes, reflecting the underlying genetic model.

Large sample distribution of the FBAT statistic under the null hypothesis

As outlined in the discussion of the key principles of the FBAT approach, the distribution of the FBAT statistic U is computed by treating the non-founder genotype as the only random variable and both the coded phenotype, T_{ij}, and the sufficient statistic, S_i, as deterministic variables by conditioning on them. The expected value of the FBAT statistic, U, is zero by definition $(E(U) = 0)$, so to normalize U under the null hypothesis, all that is left to do is to compute the variance of U conditional upon the offspring phenotype and S_i. If the genotype and trait variable are both univariate, then

$Z = U/\sqrt{(var\ (U))}$, or equivalently,
$\chi^2_{FBAT} = U^2/var(U)$,
where

$$Var(U) = \Sigma_i \Sigma_{jj'} T_{ij} T_{ij'} cov(X_{ij}, X_{ij'}|S_i, T_{ij}, T_{ij'}) \quad (2)$$

As for the expected marker score, the covariance $cov(X_{ij}, X_{ij'}|S_i\ T_{ij}, T_{ij'})$ also conditions upon the traits and the sufficient statistics, assuming

the null hypothesis is true. Under the null hypothesis of no association and no linkage, the covariance $cov(X_{ij},X_{ij'}|S_i\ T_{ij},T_{ij'})$ does not depend on the phenotype T_{ij}, and can be computed based on independent Mendelian transmissions within a family. However, when the null hypothesis of no association in the presence of linkage is selected, the transmissions to siblings within a family are correlated (55). In this situation, the derivation of the theoretical covariance is difficult, and an empirical variance can be used to estimate *var(U)* (51).

Asymptotically, Z is normally distributed, N(0,1), and χ^2_{FBAT} follows a χ^2 distribution with one degree of freedom. When multiple alleles and/or multiple traits are tested, U is the vector and *var(U)* becomes a variance/covariance matrix. Then, the FBAT statistic is a quadratic form $U^T var(U)^- U$ and follows asymptotically a χ^2 distribution with degrees of freedom equal to the rank of *var(U)* (60,63).

When the number of families is small (e.g. in linkage studies), it is recommended either to estimate the P-value of the FBAT statistic via Monte-Carlo simulations or to use an exact test (64).

Specifying the mode of inheritance in the FBAT statistic

In the FBAT statistic, the coding of the genotype reflects the specified mode of inheritance. When testing under an additive mode of inheritance is required, X_{ij} counts the number of target alleles (i.e. 0, 1 or 2). Under a recessive model, X_{ij} is defined to be 1 for subjects who carry 2 copies of the target allele, and 0 otherwise. For multiallelic markers or haplotypes, X_{ij} is a vector whose element reflects the coded genotype for each allele/haplotype.

Coding the phenotype: Testing binary phenotypes in the FBAT approach

When the phenotype of interest is affection status, an FBAT statistic that is equivalent to the classical TDT (61), and that only incorporates information on affected subjects, can be obtained by setting $T_{ij} = 1$ for affected subjects and 0 otherwise. Unaffected subjects can be included in the FBAT statistic by defining $T_{ij} = (Y_{ij} - \mu)$, where Y_{ij} is the original 1/0 phenotype and μ is a user-defined offset parameter in the range between 0 and 1. For example, by setting $\mu = 0$, the original TDT statistic is obtained. Affected subjects ($Y_{ij} = 1$) then contributed $(1 - \mu)$ to the FBAT statistic and the unaffecteds $(1 - \mu)$. Here the FBAT statistic can be interpreted as a contrast between transmissions to affected offspring weighted by $(1- \mu)$, and unaffected offspring weighted by μ.

In samples that have been recruited without a phenotypic ascertainment condition, e.g. population samples, the optimal offset choice is the prevalence of the disorder/trait E(Y = 1) in the total population (63). Even for studies with phenotypic ascertainment conditions, this finding approximately holds (65,66). In many situations, the population prevalence of the disease/trait is unknown. Since most study designs over-sample affected subjects to maximize the genetic loading of the sample (e.g. trio-design), the population prevalence of the disease/trait cannot be estimated directly from the sample. Fortunately, the FBAT statistic achieves almost optimal power in a relatively large neighbourhood around the true population prevalence (66). In practice, rough estimates for the prevalence will be sufficient.

Handling general pedigrees and/or missing founders in the FBAT approach

The FBAT statistic is very general and can be applied to any complex pedigree as long the expected marker score, $E(X_{ij}|S_i)$, can be computed, as well as the corresponding variance/covariance structure, $cov(X_{ij},X_{ij'}|S_i\ T_{ij},T_{ij'})$, which requires the specification of the marker densities $p(X_{ij},\ |S_i\ T_{ij})$ and $p(X_{ij},X_{ij'}|S_i\ T_{ij},T_{ij'})$. For nuclear families in which both parents and one or multiple offspring are genotyped, the univariate density, $p(X_{ij}\ |S_i\ T_{ij})$, is completely defined by Mendel's law. Under the null hypothesis of no association and no linkage, the parental transmissions to all offspring are independent, $p(X_{ij},X_{ij'}|S_i\ T_{ij},T_{ij}) = p(X_{ij,}\ |S_i\ T_{ij,})^* p(X_{ij'}\ |S_i\ T_{ij'})$, and computation of the expected marker score and its variance/covariance is straightforward. In the presence of linkage, the transmissions from the parents to the offspring are not independent anymore, but rather dependent on the recombination fraction which is known. Technically, it would be possible to remove the dependence on the unknown recombination fraction by conditioning on the identity-by-descent patterns among offspring (51); however, the inclusion of this additional condition would make many families uninformative for the computation of the test statistic, and would lead to a substantial drop in statistical power. It is therefore recommended to estimate the variance/covariance structure directly by using empirical variance estimators, as discussed above.

The same ideas for the computation of the expected marker scores and their variance/covariance structure are also applicable to extended pedigrees in which the

genotypes of all founders are known (51,61). For the analysis of such data, the power of the FBAT statistic can be increased by computing the conditional marker distribution for the complete pedigree instead of splitting up the pedigree into nuclear families and analysing the data as such (51,61). For pedigrees in which founder genotypes are missing, the computation of the expected marker scores and its variance is more complex. Instead of conditioning on the parental genotypes, the distribution of the observed offspring genotypes is computed conditional on the sufficient statistics for the unobserved parental genotypes. The advantage of conditioning on the sufficient statistic here is that no assumptions about the unobserved parental genotypes are necessary. Such assumptions would make the FBAT statistic susceptible to the effects of population substructure and stratification. Although the concept of the conditioning on the sufficient statistic for the unobserved parental genotypes is very technical, the conditional distributed for the observed offspring genotypes can straightforwardly be computed using the algorithm by Rabinowitz and Laird (51). The details of the algorithm are not discussed here, and the interested reader is referred to the original paper.

Handling haplotypes and multiple markers in the FBAT approach

In candidate gene studies, and even in genome-wide association studies nowadays, closely spaced markers/SNPs are often available that characterize a gene or a well-defined region. In such scenarios, it might not be the optimal strategy to test each marker individually for association with the phenotype of interest for two reasons. First, in general, it is difficult to take the LD-structure/correlation structure between markers into account when adjusting for multiple comparisons. This often leads to adjustments for multiple comparisons that are too conservative. Second, by only testing one marker locus at a time, the available genetic information on the other marker loci is not used. Consequently, a more powerful strategy would be to test all markers that reside in a well-defined region simultaneously. Two approaches for this are available haplotype tests and multimarker tests.

Here a multiloci haplotype is defined as a set of alleles, one for each marker, that are located on the same copy of the chromosome and that are inherited from one generation to the next without recombination. There are several situations in which multiloci haplotype tests should be more powerful than single-marker tests. For example, consider the scenario in which a true disease susceptibility locus (DSL) is located in the region that is spanned by the markers, but the DSL has not been genotyped nor is in sufficiently high disequilibrium with one of the genotyped markers to be identified by a single-marker test. If the set of genotyped markers is able to capture the haplotype diversity in the region, a multiloci haplotype will exist that captures the variation at the DSL. Another scenario, in which a haplotype analysis will be more powerful than a single-marker approach, is when two or more of the observed markers have genetic effects on the phenotype of interest. On the other hand, if there is only a single DSL in the region, and its variation is sufficiently "tagged" by one of the genotyped markers, a haplotype analysis can be suboptimal.

If the phase of the haplotype (i.e. which alleles are located on the same copy of the chromosome and are inherited jointly) is known for each subject in the study, the set of markers defining the haplotypes can be interpreted as a single marker with multiple alleles and the FBAT statistic can be computed as outlined above. However, in most applications, the phase of the haplotypes will not be known and will have to be inferred. Despite the fact that family data is available here, for which it is generally easier to determine the phase of the haplotypes than for population-based data, resolving the phase in all subjects will not be possible, especially if parents' genotypes are missing.

However, an unresolved haplotype phase in a study subject does not prevent the computation of the FBAT statistic. The same trick can be applied here as in the case for missing parental genotypes. The haplotype distribution in offspring is computed conditional upon both the parental genotypes/sufficient statistics and whether it is possible to infer the phase of the haplotypes (67). The FBAT statistic can then be calculated by assuming that the set of markers defines a multiallelic marker locus whose alleles are given by the phased haplotypes. Since this haplotype analysis approach does not make any assumptions about population parameters (e.g. haplotype frequencies, etc.), to infer haplotype phase, but conditions upon the ability/inability to reconstruct the phase, the approach maintains its robustness against population admixture and stratification. In the usual way, the FBAT statistic can either be computed for a specific target haplotype as a diallelic FBAT or as a global haplotype test based on a multiallelic FBAT. As discussed above, the presence of linkage can be accounted for by use of the empirical variance estimator.

As the numbers of markers increase, the advantages of a haplotype analysis are outweighed by characteristic disadvantages of the approach. Inferring the phase of a haplotype becomes increasingly difficult and numerically complex when the number of markers exceeds 5–10, particularly when parental information is missing or extended pedigrees are analysed. Furthermore, the assumption of non-recombination between the markers must be carefully considered. In this situation, which also applies to smaller numbers of markers, so-called multimarker FBATs can be an attractive alternative. Rather than trying to infer the underlying haplotype structure, multimarker FBATs account for the linkage disequilibrium between markers by directly estimating the variance/covariance structure between the markers. To construct a multimarker FBAT, in the FBAT statistic the univariate marker score X_{ij} is replaced by a vector X_{ij} whose elements are the genotypes for each individual marker. The vector of expected marker scores, $E(X_{ij}|S_i)$, is defined by the expected marker scores for each marker which are computed individually, conditioned upon the corresponding parental information/sufficient statistic. The linkage disequilibrium between the markers is incorporated by using the empirical estimator of $var(X_{ij})$ in the calculation of $var(U)$. The multimarker FBAT statistic is then a quadratic form which has an asymptotic χ_2 distribution, where the degrees of freedom are given by the number of markers that are linear independent. (A detailed discussion of multimarker FBATs is given in (68).) Alternative approaches are discussed in (69,70).

Complex trait analysis in the FBAT approach

Complex phenotypes are tested in the FBAT approach by selecting an appropriate coding function Tij that is selected by the user and that will depend on the trait type. The choice of the coding function should be motivated by an underlying phenotypic model, describing the phenotypes as a function of the genotypes. Since the FBAT approach conditions upon the parental genotypes and the offspring phenotype, the validity of the FBAT test will not depend upon the correct specification of the coding function, but a poor choice will affect the statistical power of the approach.

A more refined version of the phenotype affection status is the variable age-at-onset/time-to-onset. If the phenotype age-at-onset/time-to-onset contains more genetic information, such an analysis will result in greater statistical power (e.g. for childhood asthma). It can be assumed that an early onset is more related to genetic factors than is a late onset, which could be attributable to environmental factors. Various coding functions for an age-at-onset analysis are discussed in (71) and (72).

For quantitative phenotypes, standard phenotypic residuals are an obvious choice for the coding function, i.e. $T_{ij} = (Y_{ij} - \mu)$, where Y_{ij} is the original phenotype and μ is a user-defined offset parameter. For population samples (a study without any phenotypic ascertainment conditions), the optimal offset choice is the phenotypic sample mean. In such a situation, the FBAT statistic for the quantitative trait has higher statistical power than an FBAT statistic that is based on a dichotomized version of the same quantitative trait (73). To utilize this theoretical power advantage in a real data analysis, some additional work is usually required. By definition, quantitative traits contain more information and are therefore more powerful phenotypes in a statistical analysis, but they usually depend on other non-genetic factors (e.g. lung-volume measurements in asthma studies depend on age, gender and height). Such confounding variables can be probands characteristics, but they also include environmental/treatment information. For example, lung-volume measurements for asthmatics depend on smoking status/history and on treatment for asthma. An unadjusted, raw measurement of such a phenotype will be confounded by such factors and the genetic signal will be diluted, resulting in a potentially lower statistical power. For such phenotypes, it is recommended to regress the raw phenotypes on all known confounding variables and use the regression residuals as the coded phenotype in the computation of the FBAT statistic. Note that such an adjustment is study-specific and requires careful statistical model building; results might not be reproducible in other studies that do not have the same covariate information. The motivation for a within-study adjustment is to reduce the variability in the phenotype that is attributable to all other non-genetic factors. However, this requires knowledge and measurement of such variables, which are not necessarily known before the study. For such situations, efficient coding functions that do not require any covariate adjustment and that are able to achieve high power levels can be used (74).

For many complex diseases, the definition of affection status is based on a variety of phenotypes which describe and characterize the disease and its severity.

Consequently, when an association with affection status is tested for in such a situation, the aggregated and dichotomized information is assessed all at once. If now, to increase statistical power, the quantitative traits that define the disease and/or describe its severity are selected as target phenotypes instead of affection status, multiple FBATs have to be computed and the resulting multiple testing problem has to be addressed. Quantitative phenotypes for a complex disease typically are correlated and cluster together into groups (symptom groups). In asthma studies, it is standard practice to measure quantitative phenotypes that characterize such things as the lung-function of a proband (FEV1, FVC) and the atopy-reaction (number of positive skin tests, IGE-levels) (75). Depending on how well understood the disease is, symptom groups can be defined based on prior knowledge about the underlying biological pathways, clinical knowledge, or just the phenotypic correlation between the traits. A test strategy that does not incorporate this aspect of the data, but that tests all phenotypes individually and adjusts for multiple comparisons, would be optimal. Since the FBAT tests for the same symptom group will be correlated, standard adjustments for multiple testing will be too conservative here. Further, if the hypothesis is true that the phenotypes in the same symptom group are influenced by common genetic factors and/or share similar environmental confounding, it will be more powerful to assess the evidence for association for the entire symptom group at once. A multivariate method that tests all phenotypes jointly in a single test, without having to adjust for multiple comparisons, is the most desirable approach in this situation.

For the FBAT approach, such a multivariate test that examines all phenotypes simultaneously is the FBAT-GEE statistic (76). The FBAT-GEE statistic maintains the advantages of the original FBAT statistic. It is easy to compute and does not require any distributional assumptions about the phenotypes even if the selected phenotypes are of different trait types (e.g. normally distributed phenotypes, count variables, etc.).

FBAT-GEE

For each study subject it is assumed that m phenotypes have been recorded and are defined as a symptom group as described above. The vector containing all m observations for each proband is denoted by $Y_{ij} = (Y_{ij1},..., Y_{ijm})$, where Y_{ijk} is the kth phenotype for the jth offspring in the ith family. The multivariate FBAT-GEE statistic can then be obtained by defining the coding vector T_{ij},

$$T_{ij} = Y_{ij} - \hat{Y}_{ij1} = \begin{pmatrix} Y_{ij1} \\ Y_{ijk} \\ Y_{ijm} \end{pmatrix} - \begin{pmatrix} \hat{Y}_{ij1} \\ Y_{ijk} \\ \hat{Y}_{ijm} \end{pmatrix}$$

where the parameter $\hat{Y}_{ijk}$ is given either by the observed sample means for the kth trait, or by the predicted trait value based on a regression $\hat{Y}_{ijk}$ on its known covariates/confounding variables. As discussed earlier, in the situation of a multivariate trait, the univariate coding variable T_{ij} in the FBAT statistic is replaced by the vector T_{ij} and the FBAT-GEE statistic given by

$$T_{FBAT\text{-}GEE} = C^T V^{-} C.$$

Under the null hypothesis, the FBAT-GEE statistic is asymptotically χ^2-distributed with m degrees of freedom. The name of the test statistic originates from its link with the generalized estimating equation approach (77). A generalized estimating equation model can be defined by modeling the m phenotypes as a function of the genotype, using appropriate trait-dependent link-functions and a predefined variance/covariance structure. When a family-based score test is derived for this estimating equation model, the link-functions and the assumptions for the variance/covariance structure cancel out and the model-free FBAT-GEE statistic is obtained, making the multivariate FBAT-GEE statistic invariant towards distributional assumptions for the phenotype.

Pedigree-based association tests (PBATS): Bypassing the multiple comparison problem in family-based association studies

To maintain the three key properties of the original TDT approach, the FBAT statistic conditions upon the phenotype and the parental genotypes, which comes at the price that not all information about linkage and association that is contained in the data can be used. While this ensures that the robustness and the model-free character of the original approach are maintained, FBATs are in general not the most efficient test statistic. However, this extra unutilized information can be brought into play in a screening step before the computation of the FBAT statistic. The information can be used to construct an optimally informed two-stage testing strategy, or an “optimal” FBAT statistic, which has been denoted as a pedigree-based association test (PBAT). This enhances the power of the FBAT approach substantially. FBAT, with a prior screening step, can achieve power levels that are comparable to power levels that would be obtained

by a corresponding population-based analysis (78).

In particular, in large-scale association studies, with numerous genotyped markers and multiple complex traits, the screening step/extra information can be used to guide the testing strategy with respect to minimizing the effects of multiple comparisons, model-building and phenotype selection. Discussed here is a general approach that partitions family data into two independent components corresponding to the population information, and the within family information. The population information about association, which is susceptible to population substructure, is used for the screening step, or model development, and the within-family information is used for the construction of the confirmatory FBAT statistic. The idea is similar to cross validation, except that each subject contributes information to both parts of partitioning, minimizing the variability of the genetic effect in the two subsets. For simplicity it is assumed that offspring-parent trios are given.

The distribution of the complete data is the joint distribution of the offspring phenotype, *Y*, the offspring genotype, *X*, and the parental genotype, *P* (or more generally, the sufficient statistic, *S*). Using Bayes' rule, the joint distribution can be partitioned into two independent parts:

$$P(Y, X, S) = P(X|Y,S)P(S,Y). \quad (3)$$

If the screening step (e.g. model building, hypothesis generation) uses only information on *S* and *Y*, any subsequent hypothesis testing that is based on the FBAT statistic, whose distribution is given by $P(X|S,Y)$, will be independent of the prior screening step. There are various ways to model the variables *S* and *Y* so that information about a potential association between *Y* and *X* can be obtained. In general, the appropriate model for *S* and *Y* will depend on the specific design (e.g. ascertainment conditions, trait type, etc). For example, in the situation of an unascertained population sample with a quantitative target phenotype, the population-based information about the association between the offspring genotype and phenotype can be described by the conditional mean model (73,79):

$$E(Y) = m + a*E(X|S). \quad (4)$$

The genetic effect size, *a*, can be estimated by an ordinary regression of the phenotype, *Y*, on $E(X|S)$. Note that $E(X|S)$ is computed solely based on the parental genotypes. For the uninformative families (i.e. trios with doubly homozygous parents), the actual offspring genotype, *X*, is equal to $E(X|S)$. Otherwise, if parents are informative, the offspring genotype *X* can be thought of as missing and being imputed by $E(X|S)$. Since the conditional mean model is only based on information about *Y* and *S*, under the null hypothesis all its parameter estimates will be statistically independent of the FBAT statistic. Of course, the statistical independence of the screening step/conditional mean and the FBAT statistic does not hold under the alternative hypothesis. The conditional mean model (4) can therefore be fit repeatedly for any choice of genetic model, any number of phenotypes and any number of markers. Based on the parameter estimates for the conditional mean model, the Wald test for null hypothesis of no association, H_0: $a = 0$, can be computed. Alternatively, the parameter estimates can be used to compute the conditional, predicted power of the FBAT statistic. Such conditional power calculations will also depend upon the observed parental genotypes and phenotype (73,79). It is generally recommended to use the conditional power estimates to prioritize information for the subsequent FBAT testing step (80). This basic idea can be extended to handle longitudinal and repeated measurements (FBAT-PC) (81) and multivariate data (69). There the screening step can be used to compute optimal linear combinations of traits for subsequent testing. The approach has also been adapted to scenarios in which multiple markers are tested (69). A method has been proposed to estimate the genetically relevant age range for age-at-onset data (72). This extension is particularly useful for diseases in which an early onset suggests a strong genetic component, while a late onset is mostly attributable to non-genetic/environmental effects (e.g. Alzheimer's disease or childhood asthma).

Testing strategies for large-scale association studies

Genome-wide association studies offer great potential to the field of complex disease mapping, but to translate the dramatic increase in genetic information at a genome-wide level into the identification of new disease genes (a major statistical challenge) the multiple testing problem has to be tackled. For case–control studies, multistage designs have been proposed (82,83) as a cost-efficient way of handling this problem. In each stage of the design, the number of genotyped SNPs and genome-wide significance are achieved by a joint analysis of all stages.

For family-based association studies, the concept of partitioning the association information into

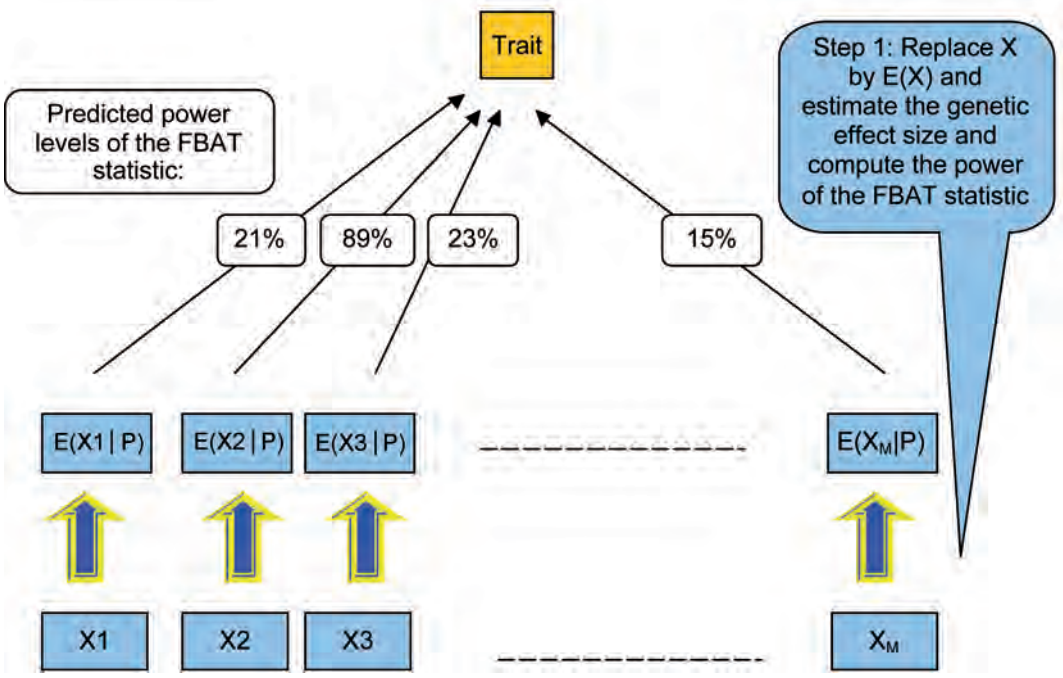
two statistically independent components is well suited to efficiently address the multiple comparison problem within one study. By using the decomposition (3), a two-stage testing strategy can be constructed that consists of two statistically independent stages, the screening step and the testing step, which can be applied to the same data set (80). This approach is illustrated in Figure 15.2.

A two-stage approach is highly effective for screening and then testing results when family controls are available for application using FBAT. In step 1, association analysis based on the conditional mean model before the FBAT testing is used to minimize the multiple testing problem. In this example, one quantitative trait and *M* SNPs are analysed. In the first step, the marker data in the offspring is assumed to be missing and imputed by the expected markers scores conditional on the parental genotypes/sufficient statistic. Based on the imputed data, the conditional mean model is fitted, and its estimates are used to compute the power of the FBAT statistic for each SNP. The power is a function of the observed parental genotypes, their frequencies, and the genetic effect size estimated from the conditional mean model. In the final step, the *K* SNPs with the highest power estimates are selected to be subsequently tested for association with the FBAT statistic at a Bonferroni-adjusted significance level of $\acute{\alpha}/K$. Since only *K* SNPs are pushed forward to the testing step, it is only necessary to adjust for *K* comparisons instead of *M*. The markers that pass this first testing step are then validated in the second step, as depicted in Figure 15.2.

In family-based designs, the screening procedure utilizes information on all families, even the non-informative ones. Assuming moderate to small genetic effect sizes, simulation studies have shown that if a true DSL, or a SNP in LD with a DSL, is included in the data set, it is sufficient to select only the highest 10 or 20 SNPs for subsequent testing to achieve high power levels. The key advantage of this testing strategy for family-based designs is that the same data set is used twice; once for the genomic screening step and once for the testing step. Thereby the effects of study heterogeneity are minimized, which can cause, in a standard two-stage design that uses different samples in each step, the failure to discover an important association. Another advantage of this approach is that it is only necessary to recruit one sample to identify SNPs/associations that achieve genome-wide significance. Replications in other studies serve the sole purpose of generalizing a significant finding to other populations.

This testing strategy has been successfully applied to a 100 000-SNP scan for obesity in the family

Figure 15.2. Using the same data set for genomic screening and testing
Step 1: Screening SNPs using conditional power estimates for the FBAT statistics. The power estimates are based on genetic effect size estimates obtained from the conditional mean model.



plates of the Framingham Heart Study. Among the top 10 SNPs from that study, as determined by estimated conditional power, there was a novel SNP whose association with body mass index (FBAT *P*-value = 0.0026) reached genome-wide significance, after having adjusted for 10 comparisons. If standard analysis methods would have been used (e.g. testing all SNPs for association and adjusting for multiple testing by the Bonferroni or Hochberg corrections), this association would have been missed. Using the same genetic model, the finding was replicated in four independent studies, including cohort, case–control and family-based samples of different ethnicities (84). Recently, the approach was extended so that all genotyped SNPs can be tested in the second stage of the testing strategy, making a decision on how many SNPs should be pushed forward to the test step redundant (78). Despite the larger number of tests in the second stage, this approach achieves power levels that are about 50% higher than in the original Van Steen approach, and that are comparable to the power levels of a population-based study with the same number of probands. The approach has been generalized so that phenotypic information on the parents can be incorporated as well (85). Extensions for case–control designs have been developed (86,87).

Other extensions of the FBAT approach include an extension to accommodate copy number variation calls (88), and an extension to allow covariate data from the parents to modify the weight assigned to transmissions of genetic information to the offspring, allowing the phenotype of the parents to influence the association analysis (89).

Figure 15.2, Step 2: The Testing Step. Select the top K SNPs with the highest power estimates for subsequent testing with the FBAT statistic. The P-value of the FBAT statistic must be smaller than ά/K to achieve genome-wide/overall significance.

Power Rank	Estimated power of the FBAT statistic	SNP	P-value FBAT statistic
1	0.92	3	0.90
2	0.89	100	0.20
3	0.85	25	0.00001
….		…	…
K	0.70	53	0.20

Software

With family-based designs, there is generally a need for special software to analyse the data. For the FBAT approach, four software packages are available. Two packages were developed by the original authors of the methods and are home-grown (PBAT, P2BAT). Despite the lack of general support for such software packages in academia, the packages have proven to be reliable and user-friendly tools. Recently, a commercial package with professional user-support and documentation has become available that is particularly suited for less statistical-oriented users and for large-scale projects. Table 15.1 shows an overview of these packages and their functions.

Discussion

Studies of families have been instrumental for describing the genetic architecture of many Mendelian and complex diseases. For initial characterization of the strength of evidence for genetic factors influencing disease risk, twin studies and evaluations of the aggregation of disease within families provide key insights. For diseases that have strong influences from genetic factors, segregation analysis followed by linkage analysis has been a highly effective strategy. When the genetic and environmental factors influence disease risk in complex ways, linkage analysis using a model-free method is a preferred strategy. For diseases with weaker genetic influences, or that result from effects of many genetic factors with each individually having a weak effect on disease risk, association studies are more successful. Family-based association studies are robust to population stratification and can have power comparable to case–control studies with unrelated cases and controls.

The area of whole-genome association scans offers great promise for the field of genetic association mapping. Most predictions agree that studies with large sample sizes are needed to identify the “needles in the haystack,” regardless of which design is used (80,83,90). It is much easier to achieve such sample sizes from existing cohorts, or from case–

Table 15.1. Software for the analysis of family-based association tests

Package	Genetic analysis capability	Phenotypic analysis capability	Special features
FBAT	Single marker, haplotype, multi-marker	Binary traits, quantitative/ multivariate traits, ranked traits, time-to-onset	X-chromosome, permutation tests
PBAT	Single marker, haplotype, multi-marker	Binary traits, quantitative traits/ multivariate, ranked traits, time-to-onset, gene-environment interaction	Covariate adjustment, Van Steen algorithm for multiple testing, X-chromosome, permutation tests
P2BAT R-implementation	Single marker, haplotype, multi-marker	Binary traits, quantitative traits/ multivariate, ranked traits, time-to-onset, gene-environment interaction	Covariate adjustment, Van Steen algorithm for multiple testing, X-chromosome, permutation tests
PBAT GoldenHelix commercial package	Single marker, haplotype, multi-marker	Binary traits, quantitative traits/ multivariate, ranked traits, time-to-onset, gene-environment interaction	Covariate adjustment, Van Steen algorithm for multiple testing, X-chromosome, permutation tests, active user-support and professional documentation

control studies, than from family samples. However the innovative use of the population information that is included in family-based data sets, combined with the robustness of the family-based association methods, can protect against both population substructures and misspecifications of the phenotypic model, creating a viable and powerful alternative to population-based studies. Further, with the ability to handle extended pedigrees with large numbers of subjects, the FBAT approach allows the continuing utilization of existing linkage studies. Recent developments to estimate and test gene–environment interaction in the FBAT approach, without any loss of robustness, are an additional advantage (91). For many complex diseases, genetic interactions with environmental exposure variables are thought to be crucial for the understanding of the disease (e.g. smoking status and/or smoking history in asthma and COPD studies) (92,93).

References

1. Lichtenstein P, Holm NV, Verkasalo PK *et al.* (2000). Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med,* 343:78–85.doi:10.1056/NEJM200007133430201 PMID:10891514

2. Hemminki K, Li X, Sundquist K, Sundquist J (2008). Familial risks for common diseases: etiologic clues and guidance to gene identification. *Mutat Res,* 658:247–258.doi:10.1016/j.mrrev.2008.01.002 PMID:18282736

3. Neale MC, Cardon LR. Methodology for genetic studies of twins and families. Dordrecth (The Netherlands): Kluwer Academic Publishers; 1992 (vol 67).

4. Emery AEH. Methodology in medical genetics. An introduction to statistical methods. Edinburgh (UK): Churchill Livingstone; 1986.

5. Hemminki K, Sundquist J, Lorenzo Bermejo J (2008). Familial risks for cancer as the basis for evidence-based clinical referral and counseling. *Oncologist*, 13:239–247. doi:10.1634/theoncologist.2007-0242 PMID:18378534

6. Risch N (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am J Hum Genet,* 46:229–241. PMID:2301393

7. Rigby AS, Voelm L, Silman AJ (1993). Epistatic modeling in rheumatoid arthritis: an application of the Risch theory. *Genet Epidemiol,* 10:311–320.doi:10.1002/gepi.1370100504 PMID:8224809

8. Khoury MJ, Beaty TH, Cohen BH. Fundamentals of genetic epidemiology. New York (NY): Oxford University Press; 1993.

9. Scott WK, Pericak-Vance MA, Haines JL (1997). Genetic analysis of complex diseases. *Science*, 275:1327–1330, author reply 1329–1330.doi:10.1126/science.275.5304.1327 PMID:9064788

10. Amos CI, Rubin LA (1995). Major gene analysis for diseases and disorders of complex etiology. *Exp Clin Immunogenet,* 12:141–155. PMID:8534501

11. Lin JP, Cash JM, Doyle SZ *et al.* (1998). Familial clustering of rheumatoid arthritis with other autoimmune diseases. *Hum Genet,* 103:475–482.doi:10.1007/s004390050853 PMID:9856493

12. Struewing JP, Hartge P, Wacholder S *et al.* (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N Engl J Med,* 336:1401–1408.doi:10.1056/NEJM199705153362001 PMID:9145676

13. Wacholder S, Hartge P, Struewing JP *et al.* (1998). The kin-cohort study for estimating penetrance. *Am J Epidemiol*, 148:623–630. doi:10.1093/aje/148.7.623 PMID:9778168

14. Gail MH, Pee D, Carroll R (1999). Kin-cohort designs for gene characterization. *J Natl Cancer Inst Monogr,* (26):55–60. PMID:10854487

15. Chatterjee N, Wacholder S (2001). A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics*, 57:245–252. doi:10.1111/j.0006-341 X.2001.00245.x PMID:11252606

16. Ford D, Easton DF, Stratton M *et al.*; The Breast Cancer Linkage Consortium (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *Am J Hum Genet*, 62:676–689.doi:10.1086/301749 PMID:9497246

17. Antoniou AC, Cunningham AP, Peto J *et al.* (2008). The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br J Cancer,* 98:1457–1466.doi:10.1038/sj.bjc.6604305 PMID:18349832

18. Ziogas A, Anton-Culver H (2003). Validation of family history data in cancer family registries. *Am J Prev Med,* 24:190–198.doi:10.1016/S0749-3797(02)00593-7 PMID:12568826

19. King TM, Tong L, Pack RJ *et al.* (2002). Accuracy of family history of cancer as reported by men with prostate cancer. *Urology*, 59:546–550.doi:10.1016/S0090-4295(01)01598-9 PMID:11927311

20. Bondy ML, Strom SS, Colopy MW *et al.* (1994). Accuracy of family history of cancer obtained through interviews with relatives of patients with childhood sarcoma. *J Clin Epidemiol,* 47:89–96.doi:10.1016/0895-4356(94)90037-X PMID:8283198

21. American Society of Human Genetics. Policy statement archives. Family history and privacy advisory; 2000. Available from URL: http://www.ashg.org/pages/statement_32000.shtml.

22. Cannings C, Thompson EA, Skolnick MH (1978). Probability functions on complex pedigrees. *Adv Appl Prob,* 10:26–61 doi:10.2307/1426718.

23. Elston RC, Stewart J (1971). A general model for the genetic analysis of pedigree data. *Hum Hered,* 21:523–542. doi:10.1159/000152448 PMID:5149961

24. Lange K, Elston RC (1975). Extensions to pedigree analysis I. Likehood calculations for simple and complex pedigrees. *Hum Hered,* 25:95–105.doi:10.1159/000152714 PMID:1150306

25. Morton NE (1955). Sequential tests for the detection of linkage. *Am J Hum Genet,* 7:277–318. PMID:13258560

26. Lander E, Kruglyak L (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet,* 11:241–247.doi:10.1038/ng1195-241 PMID:7581446

27. Witte JS, Elston RC, Schork NJ (1996). Genetic dissection of complex traits. *Nat Genet,* 12:355–356, author reply 357–358. doi:10.1038/ng0496-355 PMID:8630483

28. Huang Q, Shete S, Amos CI (2004). Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am J Hum Genet,* 75:1106–1112. doi:10.1086/426000 PMID:15492927

29. Kruglyak L, Lander ES (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet,* 57:439–454. PMID:7668271

30. Abecasis GR, Wigginton JE (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet,* 77:754–767.doi:10.1086/497345 PMID:16252236

31. Sobel E, Lange K (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet,* 58:1323–1337. PMID:8651310

32. Heath SC (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet*, 61:748–760.doi:10.1086/515506 PMID:9326339

33. Greenberg DA, Abreu PC (2001). Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. *Genet Epidemiol,* 21:299–314.doi:10.1002/gepi.1036 PMID:11754466

34. McPeek MS (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet Epidemiol,* 16:225–249.doi:10.1002/(SICI)1098-2272(1999)16:3<225::AID-GEPI1>3.0.CO;2-# PMID:10096687

35. Cordell HJ (2004). Bias toward the null hypothesis in model-free linkage analysis is highly dependent on the test statistic used. *Am J Hum Genet,* 74:1294–1302. doi:10.1086/421476 PMID:15124101

36. Etzel CJ, Guerra R (2002). Meta-analysis of genetic-linkage analysis of quantitative-trait loci. *Am J Hum Genet,* 71:56–65. doi:10.1086/341126 PMID:12037716

37. Kavvoura FK, Ioannidis JP (2008). Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum Genet,* 123:1–14.doi:10.1007/s00439-007-0445-9 PMID:18026754

38. Hästbacka J, de la Chapelle A, Kaitila I *et al.* (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet,* 2:204–211. doi:10.1038/ng1192-204 PMID:1345170

39. The SNP Consortium. Available from URL: http://snp.cshl.org/.

40. Amos CI (2007). Successful design and conduct of genome-wide association studies. *Hum Mol Genet,*16(Spec No. 2):R220–5.

41. Jorde LB (1995). Linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet,* 56:11–14. PMID:7825565

42. Johnson GC, Esposito L, Barratt BJ *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29:233–237.doi:10.1038/ng1001-233 PMID:11586306

43. Meng Z, Zaykin DV, Xu C-F *et al.* (2003). Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet*, 73:115–130. doi:10.1086/376561 PMID:12796855

44. Stram DO, Haiman CA, Hirschhorn JN *et al.* (2003). 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.doi:10.1159/000071807 PMID:12890923

45. Spielman RS, Ewens WJ (1996). The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet*, 59:983–989. PMID:8900224

46. Sham PC, Curtis D (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet,* 59:323–336.doi:10.1111/j.1469-1809.1995.tb00751.x PMID:7486838

47. Bickeböller H, Clerget-Darpoux F (1995). Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genet Epidemiol,* 12:865–870.doi:10.1002/gepi.1370120656 PMID:8788023

48. Schaid DJ (1996). General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol,* 13:423–449.doi:10.1002/(SICI)1098-2272(1996)13:5<423::AID-GEPI1>3.0.CO;2-3 PMID:8905391

49. Spielman RS, Ewens WJ (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet*, 62:450–458.doi:10.1086/301714 PMID:9463321

50. Schaid DJ, Li HZ (1997). Genotype relative-risks and association tests for nuclear families with missing parental data. *Genet Epidemiol,* 14:1113–1118.doi:10.1002/(SICI)1098-2272(1997)14:6<1113::AID-GEPI92>3.0.CO;2-J PMID:9433633

51. Rabinowitz D, Laird N (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered,* 50:211–223. doi:10.1159/000022918 PMID:10782012

52. Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet,* 64:259–267.doi:10.1086/302193 PMID:9915965

53. Martin ER, Monks SA, Warren LL, Kaplan NL (2000). A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet,* 67:146–154.doi:10.1086/302957 PMID:10825280

54. Horvath S, Laird NM (1998). A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet,* 63:1886–1897.doi:10.1086/302137 PMID:9837840

55. Lake SL, Blacker D, Laird NM (2000). Family-based tests of association in the presence of linkage. *Am J Hum Genet,* 67:1515–1525.doi:10.1086/316895 PMID:11058432

56. Allison DB (1997). Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet,* 60:676–690. PMID:9042929

57. Abecasis GR, Cardon LR, Cookson WOC (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet,* 66:279–292.doi:10.1086/302698 PMID:10631157

58. Rabinowitz D (1997). A transmission disequilibrium test for quantitative trait loci. *Hum Hered,* 47:342–350. doi:10.1159/000154433 PMID:9391826

59. Horvath S, Xu X, Laird NM (2001). The family based association test method: strategies for studying general genotype–phenotype associations. *Eur J Hum Genet,* 9:301–306.doi:10.1038/sj.ejhg.5200625 PMID:11313775

60. Laird NM, Horvath S, Xu X (2000). Implementing a unified approach to family-based tests of association. *Genet Epidemiol,* 19 Suppl 1;S36–S42.doi:10.1002/1098-2272(2000)19:1+<::AID-GEPI6>3.0.CO;2-M PMID:11055368

61. Laird NM, Lange C (2006). Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet,* 7:385–394.doi:10.1038/nrg1839 PMID:16619052

62. Sinsheimer JS, McKenzie CA, Keavney B, Lange K (2001). SNPs and snails and puppy dogs' tails: analysis of SNP haplotype data using the gamete competition model. *Ann Hum Genet,* 65:483–490.doi:10.1046/j.1469-1809.2001.6550483.x PMID:11806856

63. Lange C, Laird NM (2002). On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genet Epidemiol,* 23:165–180.doi:10.1002/gepi.209 PMID:12214309

64. Schneiter K, Laird N, Corcoran C (2005). Exact family-based association tests for biallelic data. *Genet Epidemiol,* 29:185–194. doi:10.1002/gepi.20088 PMID:16094642

65. Whittaker JC, Lewis CM (1998). The effect of family structure on linkage tests using allelic association. *Am J Hum Genet,* 63:889–897. doi:10.1086/302008 PMID:9718338

66. Lange C, Laird NM (2002). Power calculations for a general class of family-based association tests: dichotomous traits. *Am J Hum Genet,* 71:575–584.doi:10.1086/342406 PMID:12181775

67. Horvath S, Xu X, Lake SL *et al.* (2004). Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet Epidemiol,* 26:61–69. doi:10.1002/gepi.10295 PMID:14691957

68. Rakovski CS, Xu X, Lazarus R *et al.* (2007). A new multimarker test for family-based association studies. *Genet Epidemiol,* 31:9–17.doi:10.1002/gepi.20186 PMID:17086514

69. Xu X, Rakovski C, Xu XP, Laird N (2006). An efficient family-based association test using multiple markers. *Genet Epidemiol,* 30:620–626.doi:10.1002/gepi.20174 PMID:16868964

70. Rakovski CS, Weiss ST, Laird NM, Lange C (2008). FBAT-SNP-PC: an approach for multiple markers and single trait in family-based association tests. *Hum Hered,* 66:122–126.doi:10.1159/000119111 PMID:18382091

71. Lange C, Blacker D, Laird NM (2004). Family-based association tests for survival and times-to-onset analysis. *Stat Med,* 23:179–189.doi:10.1002/sim.1707 PMID:14716720

72. Jiang H, Harrington D, Raby BA *et al.* (2006). Family-based association test for time-to-onset data with time-dependent differences between the hazard functions. *Genet Epidemiol,* 30:124–132.doi:10.1002/gepi.20132 PMID:16374805

73. Lange C, DeMeo DL, Laird NM (2002). Power and design considerations for a general class of family-based association tests: quantitative traits. *Am J Hum Genet,* 71:1330–1341.doi:10.1086/344696 PMID:12454799

74. Fardo D, Celedón JC, Raby BA *et al.* (2007). On dichotomizing phenotypes in family-based association tests: quantitative phenotypes are not always the optimal choice. *Genet Epidemiol,* 31:376–382.doi:10.1002/gepi.20218 PMID:17342772

75. DeMeo DL, Silverman EK (2003). Genetics of chronic obstructive pulmonary disease. *Semin Respir Crit Care Med,* 24:151–160. doi:10.1055/s-2003-39014 PMID:16088534

76. Lange C, DeMeo D, Silverman EK *et al.* (2003). Using the noninformative families in family-based association tests: a powerful new testing strategy. *Am J Hum Genet,* 73:801–811.doi:10.1086/378591 PMID:14502464

77. Liang KY, Zeger S (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22 doi:10.1093/biomet/73.1.13.

78. Ionita-Laza I, McQueen MB, Laird NM, Lange C (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am J Hum Genet,* 81:607–614. doi:10.1086/519748 PMID:17701906

79. Lange C, Silverman EK, Xu X *et al.* (2003). A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics*, 4:195–206.doi:10.1093/biostatistics/4.2.195 PMID:12925516

80. Van Steen K, McQueen MB, Herbert A *et al.* (2005). Genomic screening and replication using the same data set in family-based association testing. *Nat Genet,* 37:683–691. doi:10.1038/ng1582 PMID:15937480

81. Lange C, van Steen K, Andrew T *et al.* (2004). A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat Appl Genet Mol Biol,* 3:e17. PMID:16646795

82. Thomas D, Xie R, Gebregziabher M (2004). Two-Stage sampling designs for gene association studies. *Genet Epidemiol*, 27:401–414.doi:10.1002/gepi.20047 PMID:15543639

83. Hirschhorn JN, Daly MJ (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet,* 6:95–108. doi:10.1038/nrg1521 PMID:15716906

84. Herbert A, Gerry NP, McQueen MB *et al.* (2006). A common genetic variant is associated with adult and childhood obesity. *Science*, 312:279–283.doi:10.1126/science.1124779 PMID:16614226

85. Feng BJ, Goldgar DE, Corbex M (2007). Trend-TDT - a transmission/disequilibrium based association test on functional mini/microsatellites. *BMC Genet*, 8:75. doi:10.1186/1471-2156-8-75 PMID:17976242

86. Zheng G, Song K, Elston RC (2007). Adaptive two-stage analysis of genetic association in case-control designs. *Hum Hered,* 63:175–186.doi:10.1159/000099830 PMID:17310127

87. Won S, Elston RC (2008). The power of independent types of genetic information to detect association in a case-control study design. *Genet Epidemiol,* 32:731–756. doi:10.1002/gepi.20341 PMID:18481783

88. Ionita-Laza I, Perry GH, Raby BA *et al.* (2008). On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. *Genet Epidemiol,* 32:273–284.doi:10.1002/gepi.20302 PMID:18228561

89. Lu AT, Cantor RM (2007). Weighted variance FBAT: a powerful method for including covariates in FBAT analyses. *Genet Epidemiol,* 31:327–337.doi:10.1002/gepi.20213 PMID:17323371

90. Clayton DG, Walker NM, Smyth DJ *et al.* (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet,* 37:1243–1246.doi:10.1038/ng1653 PMID:16228001

91. Vansteelandt S, Lange C. A unifying approach for haplotype analysis of quantitative traits in family-based association studies: Testing and estimating gene-environment interactions with complex exposure variables. COBRA Preprint Series 2006;(Article 11). Available at URL: http://biostats.bepress.com/cobra/ps/art11.

92. Celedón JC, Lange C, Raby BA *et al.* (2004). The transforming growth factor-beta1 (TGFB1) gene is associated with chronic obstructive pulmonary disease (COPD). *Hum Mol Genet,* 13:1649–1656.doi:10.1093/hmg/ddh171 PMID:15175276

93. Demeo DL, Mariani TJ, Lange C *et al.* (2006). The SERPINE2 gene is associated with chronic obstructive pulmonary disease. *Am J Hum Genet*, 78:253–264. doi:10.1086/499828 PMID:16358219