Genetic and Genomic Discovery Using Family Studies

Ingrid B. Borecki, PhD; Michael A. Province, PhD

Genetic studies traditionally have been performed on sets of related individuals, that is, families. Mendel’s early studies in sweet peas (\textit{Pisum sativum}) on the inheritance patterns of discrete traits from parents with specific mating types to offspring has shed light on the basic mechanisms of inheritance, including the fundamental laws of segregation of discrete factors (genes) from parents to offspring and the cosegregation of genes that are closely located on a chromosome (linkage). The distribution of traits within families exhibited mathematical segregation ratios in offspring from known mating types. These expected segregation ratios have been used as an important discovery tool in the study of human diseases in pedigrees, providing evidence for a multitude of single-gene disorders. Furthermore, in some cases, trait cosegregation with genetic markers with known positions provides mapping information that enables localization and, ultimately, identification of the relevant causative gene.

Pedigree studies have been used fruitfully to identify genes influencing a wide range of monogenic, highly penetrant traits of biomedical importance, including a variety of inborn errors of metabolism and other genetic diseases (eg, cystic fibrosis, Duchenne muscular dystrophy, Huntington disease). These have been documented extensively in \textit{Mendelian Inheritance in Man} (V.A. McKusick; \url{http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM}). In general, complex traits, such as coronary heart disease and its risk factors, are distinguished from these conditions in that (1) they are relatively common; (2) although they cluster in families, they do not demonstrate clean mendelian segregation patterns, suggesting the possibility of multiple underlying genes; (3) they are often influenced by >1 underlying pathway where several defects or mutations might contribute to phenotypic variation; (4) the marginal effect of any single gene on a relevant clinical end point such as atherosclerosis is likely to be small; and (5) alternative mechanisms or pathways may lead to a particular clinical outcome, that is, there may be genetic heterogeneity. These properties produce many challenges to the dissection of the genetic architecture of complex traits, some of which are advantageously met with family studies.

Family studies have several favorable features for gene discovery. Studies of extended pedigrees, or even nuclear families, are likely to represent a more homogeneous and limited set of causative genes and pathways. These features enhance statistical power for gene discovery. This approach has allowed discovery of novel loci and pathways; in a recent example, ascertainment of families with early coronary artery disease and apparent mendelian segregation led to the identification of a novel associated mutation in LRP6 (low-density lipoprotein receptor–related protein 6) (Mani et al, 2007). Clinical characteristics common to family members also can be used to reduce heterogeneity by defining subgroups of families for analysis (eg, early-onset breast cancer; Hall et al, 1990), or, in the domain of cardiovascular disease, families with maturity-onset diabetes of youth (Bowden et al, 1992) or familial combined hyperlipidemia (Badzioch et al, 2004). Analysis of trait segregation in carefully characterized pedigrees and demonstration of linkage with known genetic markers remain particularly robust and fruitful approaches to gene discovery.

Another favorable feature of family studies in contrast to studies of unrelated individuals rests in the issue of controls. The analysis of phenotypes among family members is controlled to some extent for both genetic background and environmental exposures. Because family members share a predictable proportion of their genes identical-by-descent, the background genetic variation is controlled to some extent as a function of the degree of relationship (or kinship coefficient), which can be modeled as a polygenic component. In the extreme, monozygotic twins have a strong control for genotype, leaving trait variation to epigenetic phenomena, environmental modifiers, or interactions. Similarly, (close) family members also tend to have more homogeneous environmental exposures, living in similar geographic locations with similar socioeconomic status, and perhaps even similar health-related habits such as diet, smoking, alcohol consumption, and habitual physical activity. Although these factors are not as strongly controlled as they might be in animal models, studies of families reduce residual noise variance, thereby enhancing power to detect relevant trait determinants.

Finally, on the technical side, family data allow a deeper level of genotyping quality control than is possible in studies of unrelated individuals. High rates of mendelian inconsistencies (see Glossary in the online-only Data Supplement) or markers that show significant deviations from Hardy-Weinberg equilibrium (see Glossary in the online-only Data Supplement) are likely to represent a more homogeneous and limited set of causative genes and pathways. These features enhance statistical power for gene discovery. This approach has allowed discovery of novel loci and pathways; in a recent example, ascertainment of families with early coronary artery disease and apparent mendelian segregation led to the identification of a novel associated mutation in LRP6 (low-density lipoprotein receptor–related protein 6) (Mani et al, 2007). Clinical characteristics common to family members also can be used to reduce heterogeneity by defining subgroups of families for analysis (eg, early-onset breast cancer; Hall et al, 1990), or, in the domain of cardiovascular disease, families with maturity-onset diabetes of youth (Bowden et al, 1992) or familial combined hyperlipidemia (Badzioch et al, 2004). Analysis of trait segregation in carefully characterized pedigrees and demonstration of linkage with known genetic markers remain particularly robust and fruitful approaches to gene discovery.

Another favorable feature of family studies in contrast to studies of unrelated individuals rests in the issue of controls. The analysis of phenotypes among family members is controlled to some extent for both genetic background and environmental exposures. Because family members share a predictable proportion of their genes identical-by-descent, the background genetic variation is controlled to some extent as a function of the degree of relationship (or kinship coefficient), which can be modeled as a polygenic component. In the extreme, monozygotic twins have a strong control for genotype, leaving trait variation to epigenetic phenomena, environmental modifiers, or interactions. Similarly, (close) family members also tend to have more homogeneous environmental exposures, living in similar geographic locations with similar socioeconomic status, and perhaps even similar health-related habits such as diet, smoking, alcohol consumption, and habitual physical activity. Although these factors are not as strongly controlled as they might be in animal models, studies of families reduce residual noise variance, thereby enhancing power to detect relevant trait determinants. 

Finally, on the technical side, family data allow a deeper level of genotyping quality control than is possible in studies of unrelated individuals. High rates of mendelian inconsistencies (see Glossary in the online-only Data Supplement) or markers that show significant deviations from Hardy-Weinberg equilibrium (see Glossary in the online-only Data Supplement) are likely to represent a more homogeneous and limited set of causative genes and pathways. These features enhance statistical power for gene discovery. This approach has allowed discovery of novel loci and pathways; in a recent example, ascertainment of families with early coronary artery disease and apparent mendelian segregation led to the identification of a novel associated mutation in LRP6 (low-density lipoprotein receptor–related protein 6) (Mani et al, 2007). Clinical characteristics common to family members also can be used to reduce heterogeneity by defining subgroups of families for analysis (eg, early-onset breast cancer; Hall et al, 1990), or, in the domain of cardiovascular disease, families with maturity-onset diabetes of youth (Bowden et al, 1992) or familial combined hyperlipidemia (Badzioch et al, 2004). Analysis of trait segregation in carefully characterized pedigrees and demonstration of linkage with known genetic markers remain particularly robust and fruitful approaches to gene discovery.

Another favorable feature of family studies in contrast to studies of unrelated individuals rests in the issue of controls. The analysis of phenotypes among family members is controlled to some extent for both genetic background and environmental exposures. Because family members share a predictable proportion of their genes identical-by-descent, the background genetic variation is controlled to some extent as a function of the degree of relationship (or kinship coefficient), which can be modeled as a polygenic component. In the extreme, monozygotic twins have a strong control for genotype, leaving trait variation to epigenetic phenomena, environmental modifiers, or interactions. Similarly, (close) family members also tend to have more homogeneous environmental exposures, living in similar geographic locations with similar socioeconomic status, and perhaps even similar health-related habits such as diet, smoking, alcohol consumption, and habitual physical activity. Although these factors are not as strongly controlled as they might be in animal models, studies of families reduce residual noise variance, thereby enhancing power to detect relevant trait determinants.

Finally, on the technical side, family data allow a deeper level of genotyping quality control than is possible in studies of unrelated individuals. High rates of mendelian inconsistencies (see Glossary in the online-only Data Supplement) or markers that show significant deviations from Hardy-Weinberg equilibrium (see Glossary in the online-only Data Supplement) are likely to represent a more homogeneous and limited set of causative genes and pathways. These features enhance statistical power for gene discovery. This approach has allowed discovery of novel loci and pathways; in a recent example, ascertainment of families with early coronary artery disease and apparent mendelian segregation led to the identification of a novel associated mutation in LRP6 (low-density lipoprotein receptor–related protein 6) (Mani et al, 2007). Clinical characteristics common to family members also can be used to reduce heterogeneity by defining subgroups of families for analysis (eg, early-onset breast cancer; Hall et al, 1990), or, in the domain of cardiovascular disease, families with maturity-onset diabetes of youth (Bowden et al, 1992) or familial combined hyperlipidemia (Badzioch et al, 2004). Analysis of trait segregation in carefully characterized pedigrees and demonstration of linkage with known genetic markers remain particularly robust and fruitful approaches to gene discovery.
Supplement) can be signs of genotype error, sample mixups, and quality problems.

On the other hand, family studies have disadvantages. It is more difficult and therefore more costly to identify, recruit, and enroll entire pedigrees than it is to study unrelated individuals, especially in a mobile society such as that of the United States. If one wishes to study the extremes of a distribution, such as hypertension versus hypotension or high versus low atherosclerosis as measured by intimal-medial wall thickness or coronary artery calcification, a case-control study will definitely be simpler and cheaper and may be a more efficient design, with all other factors being equal (eg, good matching, homogeneity of environmental exposures, and control of background genetic variation).

**Approaches to Gene Mapping**

Two general strategies exist for gene discovery: linkage and association studies (see Borecki and Province, 2007, for review).\(^6\) Linkage studies exploit the cosegregation of trait loci with genetic markers within families; thus, family data are a necessity. By contrast, association studies can be performed in families but also in unrelated individuals under either a random or a case-control sampling scheme. However, association tests in family data can afford added protection against elevated type I error rates (due to hidden stratification) and improve power compared with use of data on unrelated individuals, again pointing to the utility of family-based designs.

**Linkage**

In linkage studies, we seek to identify trait loci that cosegregate with known genetic markers within families. Trait and marker loci will remain on the same gametic haplotype as a function of the distance between the 2 loci, which can be measured as the recombination frequency. Classic parametric linkage analysis explicitly models the linkage, estimating the recombination fraction under a variety of trait models (eg, dominant, additive, recessive) with appropriate penetrance. Recombination fraction under a variety of trait models (eg, dominant, additive, recessive) with appropriate penetrance. The support for linkage can be quantified as the log of the odds ratio of a linkage hypothesis compared with a null model, called the logarithm of odds (LOD) score (Morton, 1955\(^7\); Cottingham et al, 1993\(^8\)). In this approach, the effect of the locus influencing either a disease or a quantitative trait is explicitly modeled as a diallelic locus with a specific penetrance. This paradigm is quite powerful and is appropriate when good justification exists for the assumed genetic model. It has been used extensively to create the map of the human genome (among genetic markers whose genotype is equivalent to the phenotype) as well as for a number of diseases in which the mode of inheritance is well known (eg, fully penetrant, recessive inheritance for cystic fibrosis). However, in the case of complex traits, strong assumptions about a single underlying trait locus seem untenable, and alternative methods have emerged that avoid that pitfall of almost certainly incorrect assumptions about mode of inheritance.

Nonparametric methods obviate the need to characterize the trait locus by focusing on relative pairs (eg, sibs) and the correlation between the allele identity at specific locations and the similarity of their phenotypes. Thus, for a disease trait, affected sib pairs would be expected to share a greater proportion of alleles identical by descent at a marker that is linked to a trait locus than expected under the mendelian null (50% allele sharing). Likewise, the more alleles shared identical-by-descent at a linked marker locus, the more similar are the quantitative phenotypes under the alternative hypothesis of linkage; under the null, no relationship exists between the two. These general expectations have given rise to a number of statistical strategies for linkage analysis including nonparametric linkage scores (Kruglyak et al, 1996\(^9\)) and variance components models (Almasy and Blangero, 1998\(^10\); Province et al, 2003\(^11\) and Abecasis et al, 2003\(^12\)).

The power of nonparametric linkage has been extensively characterized. The actual effect size of a locus is a function of the allele frequencies, the penetrance, and the recombination fraction (where the latter 2 are confounded). In addition, all the usual factors affect power, including sample size, pedigree informativeness (size and variety of biological relationships), marker informativeness, the test statistic, and critical value for the statistical test. Critically, linkage analysis is inherently limited as to the smallest effect detectable when all other parameters are asymptotically at their maxima (eg, as large a sample size as could be realistically obtained) or fixed to their optimal values (eg, recombination fraction \(\theta=0\)). Risch\(^13\) (1990) explored the power characteristics of a nonparametric linkage statistic under various conditions and reported adequate power (\(\approx 80\%\)) to detect loci accounting for a minimal elevated sibling recurrence risk of \(\approx 1.5\) to \(1.7\). Similarly, simulation studies performed for the Family Heart Study with \(>3300\) subjects in 510 extended families demonstrate that, with the use of variance components linkage models, only loci influencing a minimum of \(8\%\) to \(10\%\) of the trait variation are potentially detectable even allowing for a liberal critical significance level of \(5\%\) (Figure 1). For perspective, the apolipoprotein E locus accounts for \(\approx 8\%\) of variation in total cholesterol and \(\approx 4\%\) in low-density lipoprotein cholesterol (Boerwinkle et al, 1987),\(^14\) and the e4 allele is associated with an increased risk of coronary heart disease (odds ratio \(\approx 1.25\)) compared with the e3 allele (Wilson et al, 1994).\(^15\) It is not clear whether apolipoprotein E would have been detected as an important coronary heart disease locus via linkage. Although locus discovery by linkage is a robust strategy, it is likely to be productive only for regions with a substantial effect on the trait variation, either via a single loci or a cluster of loci each with smaller effect.

**Association**

Association studies seek to directly correlate allelic variation with phenotypic variation, with the goal of statistically identifying putative genetic causes. If the relevant genetic variation is measured (eg, apolipoprotein E variants), then the power for discovery is simply a function of the effect size. However, genomewide association scans utilize panels of markers that either are anonymous and uniformly distributed or are tags for common haplotypes across the genome. Typically, these markers are common (minor allele frequency \(\geq 5\%\)) single nucleotide polymorphisms (SNPs), which serve...
well as markers but are not a catalog of all possible causal variants. Thus, these markers are not necessarily functional but may be in linkage disequilibrium with underlying causal variants. In this case, the power to identify relevant QTLs is a function of the linkage disequilibrium between the causative and measured variants.

Two general association strategies exist: simple statistical models to correlate risk genotypes with outcome (eg, contingency table analysis, logistic regression, regression, Cox proportional hazard models), which are typically applied to unrelated individuals, or family-based tests that rely on transmission patterns from parents to offspring. Quite different information is used in the latter, which seek to identify alleles with excess transmission to affected offspring compared with mendelian expectations (Spielman et al, 1993). The transmitted alleles to affected offspring form the “case” genotype, whereas the untransmitted alleles are the “control” genotype. Heterozygosity in both parents is necessary to render a particular trio fully informative for the transmission disequilibrium test, which means that, typically, some proportion of families is not used if allelic transmissions cannot be resolved. Although this leads to some sacrifice in power, the power from measured genotypes is used in the latter, which seek to identify transmission patterns from parents to offspring. Quite different information is used in the former, which seek to identify excess transmission to affected offspring compared with mendelian expectations (Spielman et al, 1993).

The transmission disequilibrium test (TDT) is equivalent to the classic disequilibrium test (TDT) is equivalent to the classic association test of both linkage and association because, in effect, the measured genotype being tested is marking transmission of a haplotype from parents to children. Thus, the TDT (and its extensions) can pick up a genetic association signal from far away (perhaps megabases) from the observed marker.

A related extension of the TDT is the Family Based Association Test (FBAT; Laird et al, 2000). Like the TDT, FBAT does not use phenotypic information, so that E[U] = 0 under the null hypothesis of no genotype-phenotype association, so that E[U] = 0 under the null. Calculating V = Var(U) = Var(S) under the null, we get the standardized

$$Z = U / \sqrt{V}$$

as approximately distributed as N(0,1), which yields the $\chi^2$ test: $\chi^2 = U \cdot V^{-1} \cdot U$ with degrees of freedom equal to the rank of V. Like the TDT, FBAT does not use phenotypic information from parents and is actually a combined test of linkage and association because it also discounts families in which transmission from parents to children is ambiguously observed (because of homozygosity in the parents). This has caused much confusion in the literature because readers assume that FBAT must be a “pure” association test (it may

Table. Transmitted Versus Untransmitted Alleles in a Parent-Offspring Trio

<table>
<thead>
<tr>
<th>Allele1</th>
<th>Allele2</th>
<th>A+B</th>
<th>C+D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A+C</td>
<td>B+D</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

See text for details.
be more properly termed FBLAT [Family Based Linkage and Association Test]).

Type I Errors and Association Tests in Families

Standard association tests also can be done in families. These approaches are generally more powerful because all subjects are informative regardless of genotype; however, a complication exists. If a standard genotype-phenotype generalized linear model holds in every family member, the residuals in this model are not independently and identically distributed, as is the case for unrelated individuals. Instead, the residual variance-covariance matrix is sparse (nonzero correlations only in family blocks). Ignoring the cluster correlation can inflate type I error, producing false inferences. The Huber-White "sandwich" estimator provides a robust variance-covariance matrix estimate for clustered sampling (Diggle et al, 1994). For S families, the sandwich variance estimator is as follows:

$$ (X' \hat{V}^{-1} X)^{-1} \left( \sum_{i=1}^{S} X_i' \hat{V}_{ii}^{-1} \hat{e}_i \hat{V}_{ii}^{-1} X_i \right) (X' \hat{V}^{-1} X)^{-1} $$

where X is fixed effects design matrix, V is variance covariance matrix, and matrices indexed by i for the ith family and the unindexed matrices are for the entire data set, so that

$$ \hat{e}_i = (y_i - X_i \hat{\beta}) $$

is the vector of ordinary least squares residuals for the ith family (ie, ignoring familial correlations), which gives an initial estimate of familial correlations. It is "sandwiched" like meat between 2 information matrices to give a more robust variance estimate.

The sandwich method allows for tests in family data without inflation of type I error arising from ignoring familial correlations (eg, Province et al, 2000) or uses all data from all subjects in all families, even those that FBAT (Laird et al, 2000) or quantitative TDT (QTDT; Abecasis et al, 2002) finds are "uninformative" because vertical transmission is ambiguous. It can also be used for qualitative phenotypes (Liu, 1998).

Another genotype-phenotype association strategy in family data is a bootstrap procedure (eg, Province et al, 2000) creating an independently and identically distributed subsample of unrelated individuals by randomly choosing 1 subject per family. However, this greatly reduces power because the effective sample size in each subsample is the number of pedigrees rather than subjects. Bootstrap theory (Efron, 1982) suggests that it is possible to get "good" parameter estimates by bootstrap sampling of entire families with replacement and averaging results across samples, which restores the effective sample size to the number of individuals and not pedigrees. Bootstrapping families preserves the dependencies between the family genotype and phenotype vectors.

With the use of SNP data from the National Heart, Lung, and Blood Institute (NHLBI) Family Heart Study (n=27,534), 1 SNP was arbitrarily designated as "causative," and a phenotype, Y, was simulated via regression $Y = \alpha + \beta \times \text{SNP} + \epsilon$ with parameters ($\alpha$, $\beta$, and $\epsilon \sim N(0, \Sigma)$, where $\Sigma$ is the family variance-covariance matrix for a polygenic trait with 40% heritability (SEGPOWER; Province et al, 2003). The simulated regression model errors are not independently and identically distributed but are correlated within families via polygenic transmission. In Figure 2, the nominal $P$ values are plotted against their ranks. Under $H_0$ (top panel), all methods give the correct null distribution, tracking the identity line. In the middle panel, $\beta$ is set to a locus-specific heritability of 5%. Under this alternative, both the sandwich and family bootstrap yield more power than either QTDT or FBAT, which eliminate "uninformative" families with ambiguous transmission. In these cases, the family-based approaches are less powerful than the analysis of individual subjects’ data.

Population stratification was generated in this simulated example by arbitrarily dividing our families into 2 equal strata, keeping $\beta=0$ for each stratum but offsetting the phenotypic means by strata-specific intercepts, $\alpha_{\text{STRATA}}$, and also swapping the minor with the major alleles of the causative locus in 1 stratum only. Analyzing the data in the usual way or using a method that does not protect against hidden stratification produces a false-positive overall significant regression because $\beta=0$ in each stratum (bottom panel, Figure 2). FBAT $P$ values almost perfectly track at the expected uniform distribution, whereas QTDT is actually slightly overly conservative. The sandwich estimator is slightly liberal but affords some protection. The family bootstrap provides almost no population stratification protection, resulting in serious type I error inflation. Thus, family-based transmission tests have the advantage in the presence of stratification.

Families Can Be More Powerful Than Unrelated Subjects for Association

The conventional wisdom is that unrelated individuals are more powerful for genetic associations than families, but several investigators are now finding the opposite to be true (eg, Krull, 2007; Wessel et al, 2007). The argument against families is that adding a nonindependent subject to a sample does not add a whole extra person’s worth of information. It only adds a fraction of information depending on the degree of familial correlation. Indeed, if genotype, phenotype, and residuals were perfectly correlated, then all data on the 2 subjects are identical and therefore redundant. However, this never actually happens in families. Even in the case of identical twins, complex phenotypes are never identical, and neither are residuals. Error variance is a critical determinant of power. The smaller the unexplained error variance, the greater is the power to estimate all model parameters. Families can be more powerful than unrelated individuals because extra information exists with which to explain variation, thus reducing error variance.

Unrelated individuals are sampled from larger family units. A sample of J sib pairs, with phenotype $Y_{ij}$ correlated to genotype $S_{ij}$, will have the siblings correlated for reasons beyond the genotype $S_j$ (on average, sibs share half their genome identical by descent, at least some of which may affect the phenotype), so the “error” in this regression model...
Figure 2. A QQ plot for various family-based tests, based on Monte Carlo simulation experiments, including 2 transmission-based (FBAT and QTDT) and 2 association-based (family-based bootstrap and sandwich estimator) tests. With the use of real SNP data from the NHLBI Family Heart Study (n=2753), one SNP was arbitrarily designated as “causative,” and a simulated phenotype, Y, was generated via regression for 1000 replications. The nominal P values reported by the test are plotted on the x axis, and their observed percentile is plotted along the y axis. Under the null hypothesis (top), if a test is valid the reported P values should run true and the line should follow the diagonal (which it does for all 4 tests). Under the alternative hypothesis (middle), lower P values should be preferred, and the higher the power, the closer to the upper left corner the line should hug. Both the sandwich estimator and bootstrap methods provide higher power because they use all of the families to test association, whereas FBAT and QTDT test transmission and throw out families for which transmission is ambiguously determined, and they use no parental phenotype information. Under the null hypothesis with population stratification (bottom), FBAT runs true, providing strong protection and giving valid P values. QTDT is overly conservative. Bootstrap provides poor protection from population stratification, whereas the sandwich estimator provides some but still shows some inflation of type I error.
on 2J subjects comes from 2 sources: (1) a variance component $\rho \Omega_j$ that is pairwise correlated in siblings by $\rho$ and (2) an independent residual $e_j$.

$$Y_{ij} = \alpha + \beta S_{ij} + [\rho \Omega_j + e_j]$$

An equal sample of 2J unrelated individuals (selecting 1 from each of a sample of twice as many sib pairs 4J) will have the same genotype-phenotype correlation [same true ($\alpha, \beta$)], but its residual will be the sum of the 2 variance components in parentheses in the first model. In the sib pairs, we can estimate the familial variance component $\rho \Omega_j$ (sandwich estimator, above), so the unexplained variance is only $e_j$. Thus, power to estimate the gene effect of interest ($\beta$) is increased in the family sample over that in the unrelated subjects when explicitly modeling the correlation among family members. Intuitively, this makes sense. Extra information is available in the family design that is unavailable in unrelated individuals, and extra information always reduces error and boosts power at the same sample sizes.

**Meta-analysis Combining Linkage/Association Results**

Increasingly, it will be useful to combine evidence from genomewide linkage, candidate gene associations, and genomewide association scans, sometimes on the same subjects. One can informally “overlay” evidence across multiple domains and qualitatively assess where they reinforce evidence for trait loci. However, it is also possible to formally integrate at least those pieces of evidence that have been characterized on the same scale (such as $P$ value or LOD score). Two basic approaches exist. Roeder et al (2006)\textsuperscript{26} use a Bayesian framework, defining priors from a linkage scan to be updated by the genomewide association scans to produce a final posterior combined $P$ value. Meta-analysis can also combine $P$ values (eg, Province, 2001).\textsuperscript{27} Both methods are easy to apply; however, caution must be exercised in combining evidence from the same subjects, which results in correlated scans under the null. Taking such correlations into account is important; otherwise, evidence from 2 different scans may accumulate as if the signals are reinforcing, when it is only the positively correlated noise reinforcing. This is another source of inflation of type I error. Province (2005)\textsuperscript{28} developed a simple correlated meta-analysis, in which one can combine multiple linkage and association scans. The idea is that the majority of a genome scan is under the null for any given phenotype, and therefore the global degree of correlations among the pairwise scans can be estimated with the use of a tetrachoric correlation matrix to correct for reinforcing noise. The correlated meta-analysis can work on many scans, in contrast to only 2, as in the Roeder method.

The correlated meta-analysis is based on Fisher’s method of combining of $P$ values (1925), in which independent studies test the same hypothesis. For nonparametric linkage, because all of the negative evidence is truncated at LOD=0, the $P$ value distribution is a nonuniform discrete/continuous mixture, but this complication is overcome by interpreting LOD=0 as $P = 1/2\ln(2) \approx 0.72$ in Fisher’s formula (Province, 2001).\textsuperscript{27} For nonindependent scans (as will occur when some of the same subjects have been used to generate both linkage and association), the mixture distribution is quite complex. The $P$ values are transformed to a normal scale $Z_i = \text{probit}(p_i)$, and the basic multivariate statistics theorem is used that if $(Z_1, Z_2, \ldots, Z_k) \sim \text{N}(0, \Sigma_{ij})$, then $\Sigma (Z_i) \sim \text{N}(0, \text{SUM}(\Sigma_{ij}))$. It is only necessary to estimate $\Sigma_{ij}$ (variance-covariance matrix of scans) to apply the method. Complications include the truncation of negative linkage evidence (LOD $\leq 0$) as well as the fact that at least some genomic regions scan should be under the alternative. But since the majority of the genome is under the null, the contamination should be relatively minor. Both complications are minimized by dichotomizing the evidence at each locus around its natural balance point. Under $H_0$, we expect linkage scans to be approximately half positive and half negative (LOD $> 0$ versus LOD $\leq 0$). Similarly, for a genomewide association scan under $H_0$, we expect 50:50 $P < 0.5$ and $P \geq 0.5$. These critical points can be used to roughly dichotomize the evidence at each locus. Among K genomewide scans (linkage or association), for each pair the $2 \times 2$ table of dichotomized evidence is formed, and the underlying tetrachoric correlation is estimated to obtain estimates $\Sigma_{kk}$ to discount inflated meta-analysis evidence. The correlated meta method works well in simulation and is being used by the GeneLink consortium of genomewide linkage studies (sponsored by the NHLBI; https://genelink.nhlbi.nih.gov/index.jsp). This meta-analysis approach represents a method by which numerous lines of evidence from the analysis of family data (or from other study designs) can be combined to achieve optimal inferences about the presence and location of complex trait genes.

**Conclusions**

The advent of cost-effective technologies to interrogate the genome as a means to understand the genetic architecture of complex traits has brought substantial changes in the design of genetic studies. Family studies have long been the favored approach for genetic inquiries because they possess several advantageous features for linkage and association tests. Linkage studies only can be performed in collections of biologically related individuals, and this approach remains one of the most robust tools with which loci of moderate effect or clusters of modest effect genes might be identified. The recent change to whole genome association scans essentially has made it possible to conduct genetic studies in samples of unrelated individuals (eg, case-control studies), for which subjects are undeniably easier to recruit. Nonetheless, using family data has several advantages compared with using samples of independent subjects. First, many studies already have collected extensive characterizations of subjects in families. Prior linkage evidence from these studies can be brought to bear on the interpretation of subsequent whole genome association studies, as discussed above. Second, family-based tests of association obviate many of the difficulties of identifying properly matched controls by use of synthetic controls comprised of the untransmitted alleles from informative mating types. Moreover, population stratification is a common feature of admixed populations, such as those of the United States, and unrecognized stratification can result in elevated type I error rates, which exacerbates the already daunting problem of multiple comparisons in these whole
interactions in family studies. Dr Borecki is principal investigator of 2 R01s of family genetic studies: R01DK068336, Mapping Adiposity QTLS in the NHLBI Family Heart Study; and R01DK075681, Genetic Epidemiology of studies: R01DK068336, Mapping Adiposity QTLS in the NHLBI Family Heart Study; and R01DK075681, Genetic Epidemiology of studies: R01DK068336, Mapping Adiposity QTLS in the NHLBI Family Heart Study; and R01DK075681, Genetic Epidemiology of

References

Key Words: atherosclerosis ■ epidemiology ■ genetics ■ inheritance patterns ■ mapping ■ meta-analysis ■ statistics