Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application

Rita M. Cantor,[1,*] Kenneth Lange,[1,2] and Janet S. Sinsheimer[1,2]

Genome-wide association studies (GWAS) have rapidly become a standard method for disease gene discovery. A substantial number of recent GWAS indicate that for most disorders, only a few common variants are implicated and the associated SNPs explain only a small fraction of the genetic risk. This review is written from the viewpoint that findings from the GWAS provide preliminary genetic information that is available for additional analysis by statistical procedures that accumulate evidence, and that these secondary analyses are very likely to provide valuable information that will help prioritize the strongest constellations of results. We review and discuss three analytic methods to combine preliminary GWAS statistics to identify genes, alleles, and pathways for deeper investigations. Meta-analysis seeks to pool information from multiple GWAS to increase the chances of finding true positives among the false positives and provides a way to combine associations across GWAS, even when the original data are unavailable. Testing for epistasis within a single GWAS study can identify the stronger results that are revealed when genes interact. Pathway analysis of GWAS results is used to prioritize genes and pathways within a biological context. Following a GWAS, association results can be assigned to pathways and tested in aggregate with computational tools and pathway databases. Reviews of published methods with recommendations for their application are provided within the framework for each approach.

Introduction

Genome-wide association studies (GWAS) have rapidly become a standard method for disease gene discovery. When they were first conceived, it was thought that GWAS would provide an effective and unbiased approach to revealing the risk alleles for genetically complex non-Mendelian disorders. The premise of the GWAS design is that extensive common variation in the human genome, as exhibited by SNPs with frequencies greater than 1%, is responsible for the risk of most genetically complex disorders. A controversy ensued as to whether much of this risk could be explained by the older and more common gene variants that GWAS were designed to detect or whether much of the risk was likely to be caused by multiple rare variants with frequencies less than 1%. That is, because success in GWAS is contingent upon being able to statistically detect the association of an SNP that is in linkage disequilibrium with a predisposing gene variant, that variant would have to be at a sufficient frequency in order to detect it. It was expected that testing the genome with dense SNPs that capture the linkage disequilibrium in case and control samples would produce results that explain much of the risk for each disorder if the "common disease, common gene" hypothesis were the rule. Thus, in addition to their focus on revealing the biological contributions to complex traits and disorders, the results of GWAS also provide substantive information regarding the extent of the contributions made by common variants to complex traits and disorders.

GWAS require three essential elements: (1) sufficiently large study samples from populations that effectively provide genetic information regarding the research question, (2) polymorphic alleles that can be inexpensively and efficiently genotyped and cover the whole genome adequately, and (3) analytic methods that are statistically powerful and can be employed to identify the genetic associations in an unbiased fashion. Each of these three elements was developed for GWAS in just a few years, and we briefly summarize some of the specific accomplishments. Regarding the first element, a substantial number of large study samples were formed through productive collaborations. In many cases, the resulting sample sizes provide sufficient statistical power to identify relatively small associations of common variants. The GWAS show that for discrete phenotypes, most of the detectable odds ratios are between 1.1 and 1.3.[1] Thus, most of the significantly associated SNPs are in linkage disequilibrium with predisposing variants that increase the carriers' disease risks by between 10% and 30% over the risk in noncarriers. It is likely that there are many more common variants that have not been detected by GWAS because they raise the risk by smaller values, perhaps as low as 1%. However, these undetected associations might be important, because they may be effective in elucidating the biologic basis of the disorders and suggesting treatments. Regarding the second element, substantial numbers of SNPs throughout the genome were identified by the HapMap Project and placed in easily accessible databases. Product development by technology companies used these SNPs and produced accessible tools for high-throughput genotyping. These tools have evolved to provide more dense coverage of the genome at increasingly affordable costs. Third, and perhaps most important in the context of this review, a substantial number of analytic challenges have been successfully addressed. The most critical may be the thorny problem of coordination of SNP genotypes among study samples. This problem occurred with the recognition

[1]Department of Human Genetics, [2]Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA
*Correspondence: rcantor@mednet.ucla.edu
DOI 10.1016/j.ajhg.2009.11.017.

that large samples were needed and that the individual study samples that contributed to them were genotyped on different and rapidly evolving microarray platforms. Although this was problematic, investigators also viewed it as a potential strength, because it provided the opportunity to increase the number of SNPs tested for association through imputation. Imputation is the use of additional information to predict missing values in a sample. The problem sparked intensive research efforts and the creation of appropriate software that exploited linkage disequilibrium among SNPs and accurately imputed large numbers of "missing" genotypes.[2–4]

Initially, it was anticipated that GWAS would reveal a substantial number of statistically significant associations for each disorder, requiring their prioritization for follow-up studies. Saccone and colleagues provide an extensive list of criteria to consider[5] and a method to apply them. A genomic information network generates a score representing the accumulation of evidence supporting the biological relevance of each associated SNP. The inverse of the overall score is used to weight the SNPs' overall association p value. The score includes factors such as (1) occurring in a gene that is part of an implicated pathway, (2) playing a known functional role, (3) support from comparative genomics, and (4) linkage information. Feasibility is shown by applying the approach to identify genes for nicotine dependence.

The substantial number of recent published GWAS indicate that for most disorders, only a few common variants are implicated and the associated SNPs explain only a small fraction of the genetic risk,[1] even if they are considered in aggregate. The small estimated effect sizes have been disappointing to some, and many current investigations are now focused on pursuing sequencing studies to identify the contributions of rare variants to the same disorders. Rare variants, those that appear in less than 1% of the population, may even be private mutations that only appear in a few individuals or families. Mutations with these low frequencies have been seen in many Mendelian disorders. For example, cystic fibrosis (CF [MIM 219700]), which occurs with an incidence of 1/3000 in individuals of European descent, has had at least 1500 variants, most of which are autosomal recessive and fully penetrant.

Large-scale sequencing, which was once prohibitively expensive, is more financially feasible today and is likely to become a common option in the future. It will provide additional information regarding the genetic etiology of complex disorders and will perhaps blur the boundary between investigations of common and rare variants. The rare variants, which will only be seen in a small fraction of those who are affected, are expected to raise the risk in carriers at a substantially higher rate than common variants, although this expectation remains to be investigated. Sequencing studies in complex disorders will also be conducted with the caveat that rare causal variants may be even more difficult to identify and interpret. If such variants are not found in coding regions with the possibility of disrupting protein production, they may remain unsequenced, or their importance may go overlooked. As with common variants in GWAS, rare-variant analyses will present a large number of statistical challenges that are likely to lead to the development of interesting and useful methods that reveal important results.

Recently, one review and three commentaries describing and evaluating the success of GWAS were published in a single issue of the *New England Journal of Medicine*.[1,6–8] Although focused primarily on whether GWAS met their expectations, these salient publications each provide some support for the position that GWAS are only the first step in the gene identification process. More specifically, the informative review by Hardy and Singleton defines and explains many of the important GWAS concepts.[8] It also concludes that if the genetics of complex diseases is comparable to a jigsaw puzzle, we have put the edges and corners in place and now have a framework to pursue the genetic etiology of complex disorders. Kraft and Hunter[6] advocate meta-analysis, the first approach reviewed herein, to provide more stable risk estimates that can be translated into clinical information. In his commentary, David Goldstein[1] explains his significant concerns about the lack of detection of variants with sufficient effect sizes to explain the estimated heritabilities and relative risks of complex disorders. Although he advocates the detection and analysis of rare variants, he also indicates that these differences may be explained by large interaction effects, thus providing support for applying the second method we review. In his commentary, Joel Hirschhorn[7] responds to the small effect sizes of many of the nearly 250 associated variants that have been detected by pointing out that we are making progress in detecting the true targets of these analyses, the biologic pathways, thus advocating the third approach that we review.

One can view the findings from GWAS as providing preliminary genetic information available for additional analysis by statistical procedures that accumulate evidence. These secondary analyses are very likely to provide valuable information that will help us prioritize the strongest constellations of results. Here we review and discuss some of the analytic methods that are currently being applied to combine preliminary GWAS association statistics to identify genes, alleles, and pathways for deeper investigations. We anticipate that some of these methods designed to aggregate GWAS results will also be adapted to identify the contributions of rare variants to complex disorders.

Thus, one can view single-SNP GWAS analyses as a preliminary step in the gene identification process, where methods that help prioritize the more important results should be applied, and this review is written with that focus. Here we review and discuss three current methodologies that combine results to reveal larger effects in order to prioritize them for future studies. The first, meta-analysis, seeks to pool information from multiple GWAS to increase the chances of finding true positives among the false

positives. It provides a way to combine associations across GWAS, even when the original data are unavailable. We also discuss a related approach, Bayesian hierarchical modeling, which allows the incorporation of other results through a prior function.

The second approach searches for epistasis within a single GWAS study in order to identify stronger results that are revealed when genes interact. We discuss the nature and parameterization of interaction models, and, because it is impractical to consider all possible pairwise interactions, we consider only those that include at least one significant result. Current approaches of model selection by lasso penalized ordinary linear regression and logistic regression are explained in greater detail.

The third approach prioritizes alleles and genes by using information from known pathways. Complex disorders may result from the accumulation of the effects of genetic variants within pathways. The pathway model is similar to one that is polygenic; however, the genes and their variants are drawn from only a few specific pathways. Following a GWAS, association results can be assigned to pathways with the appropriate computational tools and pathway databases. Those who are developing statistical methods to implicate pathways face many challenges, because this is the least developed approach. For all three methods, we enumerate guidelines that should be followed when they are applied to GWAS data. Directions for future research are also discussed.

Conducting Meta-Analyses to Prioritize Associations

Motivation

A frequent means of prioritizing GWAS results is to combine studies via a meta-analysis. Meta-analysis is a well-established and validated statistical approach for combining evidence across any number of independent studies, each of which is designed to examine the same research hypothesis.[9–12] Rather than use the original data from these studies, which can be computationally cumbersome and logistically difficult, meta-analysis combines their results. This approach has been advocated for genetic analysis as a method of increasing power,[13–16] and it has been used extensively in reporting the findings of GWAS.[11] In June 2009, a search of PubMed with the criteria "meta-analysis AND genome-wide association study" revealed 113 articles. Meta-analyses have been used for many years in statistics, and it is impossible to cover all of the concepts in this review. Readers desiring more extensive coverage are urged to consult two excellent and comprehensive review articles addressing meta-analyses in GWAS.[15,17]

Factors to Consider

In the context of GWAS, meta-analyses combine comparable test statistics across independent studies of the same phenotype, weighting them by the confidence in the study-specific results. The definition of "comparable" is not straightforward and usually requires judgment and a justification by the investigators. Some criteria that should be considered before studies are combined are the ascertainment of the sample, the definition and measurement of the trait or disorder under analysis, and the statistics that summarize the association result. Ideally, the GWAS used in the meta-analyses were all conducted with the same ascertainment criteria on comparable populations and following the same study design, so that there is exchangeability between the individuals used in the studies.[16] Of course, in practice, this degree of homogeneity is not always possible, but it can be approached by using careful study design and cooperation between groups at the early stages of the studies.[18]

Comparability may be difficult to achieve when conducting a cumulative meta-analysis that combines the results of a new study with existing studies, because the details of the prior studies may not be available. To test for possible differences among the studies, researchers check for evidence of heterogeneity.[16,19] Two of the most popular tests are I^2 and Cochran's Q.[20] Ethnic stratification within and between studies is a source of heterogeneity that can lead to falsely impressive p values. Genomic inflation factors can be used to correct the excess of false positives as well as to test for heterogeneity.[15,18] There is a tradeoff, on one hand, between the power and knowledge to be gained from the large sample size and, on the other hand, the dilution of the effects and the difficulties in interpretation that arise when combining studies that actually examine different research hypotheses. The research hypothesis of a meta-analysis is in some sense an intersection of the research hypotheses of the individual studies. When there is too much heterogeneity, the associations can also be obscured. Heterogeneity tests can provide some insight[17]; however, it is difficult to know how strict one should be in requiring homogeneity among the studies. We recommend following the guidelines of Zintzaras and Lau,[21] who advocate using cumulative and recursive cumulative meta-analysis as well as careful examination of potential biases that could be the cause of the heterogeneity.

Inclusion of a study based on a heterogeneity analysis is not simply a statistical decision, because those SNPs that show association may not provide the complete picture. We consider three examples of meta-analyses requiring decisions regarding the tradeoff between power and information. The first meta-analysis combines studies conducted in two diverse populations. The SNPs that are associated in both samples are prioritized over SNPs that are associated in only one of the samples. Meta-analysis of studies from the same population would lead to a larger list of associations, although fewer of them will apply to other populations. In a similar way, researchers have combined studies of different but related diseases as a way to increase power and find common susceptibility genes that provide insights to all of these disorders.[22–25] Although this is a sensible approach, care must be taken to avoid overinterpreting the results of such a heterogeneous analysis, particularly negative results. Effect sizes

for the SNP may differ and, in the extreme, may lead to protection in one disease and susceptibility in another.

The second meta-analysis combines the results of an analysis that includes covariates with the results of one that does not include those covariates. The prioritized SNPs would most likely be those SNPs whose associations are largely independent of the covariates. Without recognition of this factor, information is incomplete, and valuable associations may be lost. As a third example, consider a meta-analysis combining studies that analyze a continuous trait with studies that analyze the top quartile of these trait values as cases and the bottom quartile as controls. The prioritized SNPs will be those that show a clear distinction between low and high trait values. Thus, researchers need to consider the differences in ascertainment or other aspects of the study designs that necessitated treating the phenotype differently in the two studies when interpreting their results.

Applications

Although many GWAS have used meta-analysis primarily as a means of increasing power[26] to achieve significance that exceeds a study-wide threshold,[17] the results can also be used to prioritize SNPs or genes for subsequent studies.[27–32] In a cumulative meta-analysis, a new assessment is made each time a study is conducted, which is readily adapted to multiple GWAS of the same phenotype. Cumulative meta-analysis is a form of Bayesian analysis where early studies play the role of the prior probability distribution[12,33] and the new studies contribute to the posterior probability. By combining the results of multiple studies, power is increased and the number of false positives caused by statistical fluctuation is reduced over those for a single sample. The expectation is that the true positive associations will rise to the top of the list.

As discussed previously, to obtain accurate meta-analysis results, the studies must ask the same research question and use similarly ascertained participants. The most common research question for GWAS is whether the alleles at a particular marker are associated with the disease status or trait variability. Asking the same research question in each study might appear to restrict the meta-analysis to the intersection of SNPs genotyped in all of the studies. However, imputed SNP results can be used provided that similar procedures are applied in the imputation, the reference populations remain the same, and the researchers carefully check for the effects of imputation error.[15] In addition, there are explicit approaches to incorporate the uncertainty due to SNP imputation into the meta-analysis.[15,29]

The results used in meta-analyses can be test statistics or effect sizes. If test statistics are used, a variety of ways to combine the results have been proposed. Traditional approaches include Fisher's method of combining p values,[34] and some researchers have used this method or a weighted version in GWAS.[35,36] These p value approaches have disadvantages, most notably in that they cannot provide an overall estimate of the effect size.[17] GWAS researchers have also converted the individual test statistics into z scores[31,32,36,37] and used odds ratios when the phenotype is dichotomous[38] and regression coefficients when the phenotype is continuous.[31,32,39]

Weighting can provide more power and reduce the effects of heterogeneity.[40] Meta-analyses apply the principle that, all else being equal, studies with large sample sizes inspire more confidence than studies with smaller sample sizes.[41] This is often reflected by using functions of the study sample sizes or the inverse of the variances of the estimates as weights.[15] Inverse variance weights can be estimated under a fixed-effects model or under a random-effects model.[10,12] In general, fixed-effects models assume that all genetic variation between studies is due to random error, whereas random-effects models also allow variation due to real population differences such as ethnic ancestry, study design, or phenotypic differences.[42] Random-effects models handle the possibility of heterogeneity among studies better. The range of weights tends to be less extreme than the range seen in fixed-effects models. In general, a random-effects approach results in more accurate estimates that generalize beyond the meta-analysis samples. However, the standard errors will be larger, reflecting an increase in uncertainty and making the procedure more conservative. Importantly, Ioannidis et al.[42] point out that fixed-effects models lead to overconfidence in results when there is heterogeneity between the GWAS. This overconfidence could lead to false positives. The fixed-effects approach has been used most often in GWAS meta-analysis,[29–31,38,39] possibly because (1) there are readily available tests of heterogeneity to detect this potential problem,[20] (2) accurate population estimates are not critical when the goal is to prioritize loci for follow-up, and (3) software for running the fixed-effects approach for GWAS meta-analysis is readily available (see for example METAL, a meta-analysis tool for GWAS). However, random-effects approaches have also been used[43] and should increase in popularity as more GWAS data become available across a wide spectrum of populations, phenotypes, and study designs and heterogeneity becomes more of a concern. For a more thorough discussion of weighting options as well as other nuances of GWAS meta-analysis, we again direct the reader to the review by Zeggini and Ioannidis.[17]

SNP prioritization after a meta-analysis has taken several forms. Some investigators base their decision regarding the SNPs most worthy of further study solely on the meta-analysis p values,[31,44] whereas others use predetermined decision rules to allow other prior evidence as well as the meta-analysis results to contribute to decisions.[36,37] One interesting application of this approach is to use the association results from one disease to identify candidate loci for a second disease that cosegregates with the first. For example, Fisher and colleagues[45] use the GWAS results for Crohn's disease (IBD1 [MIM 266600]) to uncover candidate loci from ulcerative colitis (IBD [MIM 266600]), and Smyth and colleagues[46] use GWAS analysis of type 1 diabetes (IDDM [MIM 222100]) as candidate loci for celiac

disease (CD [MIM 212730]) and vice versa. Although this approach is attractive, particularly in limiting the number of tests performed, it can be hard to know how to weight the evidence supporting cosegregating disease partners. Bayesian approaches provide an objective way to use this extraneous information.

Bayesian Hierarchical Models

Although the majority of meta-analyses are frequentist in their focus, Bayesian meta-analyses also have a well-established history.[41] Because Bayesian approaches hold much promise for GWAS and because they are not as well known as their frequentist analogs, we describe them here in more detail. Whittaker and colleagues[47,48] developed GWAS meta-analysis approaches that use hierarchical Bayesian models. These models are hierarchical because they assume prior distributions on the parameters. These priors in turn depend on hyperparameters drawn from some distribution. By incorporating HapMap linkage disequilibrium data as prior information, Whittaker's models allow for multimarker effects to be inferred even if these results were not part of the available data from the individual studies. Thus, they allow SNP prioritization by borrowing strength across neighboring SNPs. Because this method selects the most highly associated set of SNPs in a region, it is highly computational and poorly adapted to genome scans.

Bayesian hierarchical models can be used to generalize meta-analysis so that researchers can incorporate evidence from a wide variety of sources to create informative priors for the current study.[49,50] This evidence can include prior linkage or association results, SNP functionality, sequence conservation in multiple species, biological plausibility, or in fact anything upon which the user is willing to place a prior distribution. Both Lewinger and colleagues[49] and Chen and Witte[50] take a hierarchical approach and model the data or summary statistics of the data with parameters whose distributions are functions of additional hyperparameters. These distributions of hyperparameters depend on the locations and function of pertinent SNPs. The appeal of this approach over more computationally simple weighting schemes is that the relative importance of the additional data is not fixed a priori but is instead determined as part of the model.

The goal is to determine the posterior probability of association for each SNP and then use these posterior probabilities to prioritize the SNPs for additional study. In a fully Bayesian approach, this probability is calculated by integrating the joint probability over all possible values of the hyperparameters. In GWAS with so many hypotheses to test, this approach is computationally prohibitive. Therefore, both Chen and Witte[50] and Lewinger and colleagues[49] rely on empirical Bayes approaches, where the most probable values of the hyperparameters are calculated from the marginal likelihood of GWAS data and covariates. The probability of association is then calculated conditional on these values for the hyperparameters as well as the observed data.

Although the hierarchical Bayesian approach is quite appealing and Witte provides R code on his website, we could not find any published reports of GWAS using this approach. This omission is possibly because of its computational complexity relative to meta-analysis or because it is still difficult to decide which auxiliary covariates to use and what values to use for the priors. There is no consensus about what information should be used to create the prior support or what the appropriate forms for the prior distributions are. Both Lewinger et al.[49] and Chen and Witte[50] use gene conservation across species as well as the SNP's function as priors, but they define the priors in different ways. In addition, Lewinger et al.[49] include expression results but do not include prior linkage or association results, whereas Chen and Witte[50] use association and linkage results but not expression results.

Modeling Epistasis to Prioritize Associations

Motivation

SNPs that combine to make larger genetic effects can statistically reflect an epistatic interaction, where the alleles of one gene influence the effects of alleles of another on a trait value or risk of disease. These interactions are by definition nonlinear and thus can dramatically increase the trait or risk. Epistatic interactions that have been identified in humans include the interactions between the RET proto-oncogene (*RET* [MIM 164761]) and endothelin receptor type B (*EDNRB* [MIM 131244]) genes in Hirschsprung disease[51] (HSCR1 [MIM 142623]), the taste receptor genes taste receptor type 2 member 16 (*TAS2R16* [MIM 604867]) and taste receptor type 2 member 38 (*TAS2R38* [MIM 60775]) in nicotine dependence[52] (MIM 188890), the interleukin 4 receptor (*IL4* [MIM 147780]) variants and interleukin 13 (*IL13* [MIM 147683]) promoter variants in asthma[53] (MIM 600807), and the alpha- and beta-adrenergic receptors (*ADRA2A* [MIM 104210] and *ADRB2* [MIM 109690]) in congestive heart failure[54] (MIM 212112). Finding an interaction between SNPs increases their priority for further study.

Despite an acceptance of the importance of epistatic interactions, they have been difficult to detect. Most computers can easily handle the marginal analysis of hundreds of thousands of SNP predictors, but the assessment of all pairwise or higher-order interactions requires much greater computational resources and novel statistical approaches that are less computationally intensive. For example, likelihood ratio tests for interactions require repeated maximum likelihood estimation, whereas score tests are much less computationally intensive. However, score tests do not provide estimates of effect sizes. In the context of GWAS studies, interaction effects without prior hypotheses of the important alleles are far more difficult to detect than main effects. That is, if association studies are underpowered relative to main effects, they have almost no chance of picking up interaction effects when all possibilities are explored. Even current studies with ten of thousands of participants have trouble detecting and

confirming interactions when large numbers of SNPs are typed. The combinatorial explosion of possible interactions creates enormous difficulties in estimation, multiple testing, and overfitting. In GWAS, even for main effects, the number of predictors far exceeds the number of observations. Exhaustive examination of all pairwise interactions is possible, but for multiway interactions the task is totally impractical. We recommend making a compromise by conducting less computationally intensive score tests on all predictors and reserving more intensive likelihood ratio testing and parameter estimation for those predictors with the most significant score statistics.

Before moving on to discuss specific methods to detect epistasis, we consider how to prioritize the results of a GWAS that includes testing for interactions. Although it is not required that an SNP be involved in interactions to make it worthy of continued study, evidence of one or more interactions provides additional support. We advocate prioritizing SNPs that are most associated with a trait regardless of whether that association takes the form of a main effect or an interaction.

Statistical Methods

Interaction modeling has a long history in statistics, and it is impossible to do justice to the topic in this review. Readers wanting more extensive coverage are urged to consult two excellent comprehensive review articles.[55,56] Here we stress some recent advances in data mining that can be used to prioritize GWAS results. In dichotomous traits, logistic regression has been used to conduct association tests for GWAS; alternative tests of equality of proportions or Fisher's exact test lack the flexibility to handle multiple predictors. For continuous traits, linear regression has been used for GWAS. A major strength of regression is that it easily provides an opportunity to include interactions. Among the other advantages of regression analyses are (1) explicit parametric models; (2) stable algorithms for parameter estimation; (3) availability of likelihood ratio and F tests for both main effects and interaction effects; (4) easy incorporation of covariates such as age, sex, and ethnic origin; and (5) wide availability of reliable and well-documented software. Most of the disadvantages relate to the scale of current data sets and the excess of predictors over observations. Among the disadvantages are (1) failure of normality assumptions for quantitative traits, (2) breakdown of large sample approximations behind p values, (3) failure of search algorithms in underdetermined problems, (4) proneness to overfitting, (5) failure to deliver spare solutions, and (6) the hierarchical nature of the model selection requiring detection of main effects before detecting interaction. Before explaining how some of these defects can be addressed, it is useful to briefly summarize some of the alternative methods to detect epistasis.

The burgeoning field of data mining[57] offers many avenues to understanding interactions. Certainly, logistic regression and discriminant analysis are closely allied. In principle, any method of discriminant analysis can serve to separate cases from controls. Discriminant analysis methods such as CART (classification and regression trees) and random forests are obviously relevant to interaction modeling.[58] Other, more tailored contenders include the multifactor dimensionality reduction method,[59–61] the combinatorial partitioning method,[62] and the restricted partition method.[63] Although these tools are helpful in exploratory data analysis and excel in discriminating cases from controls, they suffer from several limitations. For example, purely combinatorial methods do not yield effect sizes or p values and are incapable of handling covariates. Depending on the algorithm employed, they can easily be overwhelmed by large numbers of predictors and the demands of cross-validation and permutation testing. In response to these criticisms, some methods are being redesigned. For instance, Lou and colleagues[52] have revised the multifactor dimensionality reduction method[64] to allow for covariate adjustment, analysis of both continuous and binary traits, and pedigree data. Their new formulation owes an intellectual debt to the FBAT (family-based association test) method.[65,66] Other innovations include introduction of entropy and conditional entropy measures of interaction[67,68]; exploitation of proximity measures between individuals[69–71]; and application of neural networks,[72] genetic programming,[73] logic regression,[74] pattern mining,[64,75] and Bayesian partitioning[76].

The need for objective evaluation of this bewildering array of methods is obvious. Cordell[55] stresses the importance of computational speed in handling massive SNP data sets. As a practical matter, statistical methods and software are inextricably intertwined. On the basis of computational speed, and presumably of ease of use, Cordell prefers the programs PLINK,[77] Random Jungle,[78] and BEAM,[76] implementing standard regression, random forests, and Bayesian partitioning, respectively. Musani and colleagues[56] are less specific in their recommendations and suggest that a combination of methods may serve consumers best. In our own research, we primarily use PLINK[77] for classical regression and Mendel[79] for lasso penalized regression.[80–82] PLINK is user friendly, with built-in data management and quality-control routines. Mendel has fewer diagnostics but excels in model selection when the number of SNPs far exceeds the number of individuals.

Before describing improvements to traditional parametric models, let us briefly mention some study design issues. In case-only designs, one looks for nonindependence of marker genotypes. Among the possible tests for departures from independence, it is worth singling out chi-square tests, entropy, and max Z scores. These tests have considerable power, but one should keep in mind their limitations. Because they do not permit covariates, they are susceptible to population stratification and linkage disequilibrium. The underlying contingency tables may be sparse, so permutation evaluation of p values is a good idea. Unfortunately, permutation testing comes at

a price of much heavier computation. At the other extreme, pedigree data also entail real challenges. Unless driven entirely by linkage disequilibrium, linkage analysis requires pedigrees.

Ascertainment may be a problem, appropriate test statistics are not always obvious, and combinatorial or linear algebra barriers rapidly mount. For these and other reasons, most geneticists are reluctant to abandon random sample or case-control study designs.

Nature and Parameterization of Epistatic Models

For the sake of this discussion, we define epistasis explicitly as deviation from linearity under a general linear model, that is $\mu_i = \delta + \beta_1 x_i + \beta_2 y_i$ versus $\mu_i = \delta + \beta_1 x_i + \beta_2 y_i + \gamma x_i y_i$, where i denotes individual, x_i and y_i are two predictors, β_1 and β_2 are their main effect sizes, δ is the intercept, and γ is the effect size for the interaction. In ordinary linear regression, trait variation is fully specified by a common variance σ^2 and a mean μ_i specific to each person i. The residuals, the deviations from the expected trait values, are assumed to be normally distributed. In logistic regression, the variance parameter σ^2 disappears, and the quantity μ_i is used to estimate the odds

$$\frac{P(z_i = 1 \mid g_i)}{P(z_i = 0 \mid g_i)} = e^{\mu_i}.$$

Suppose individual i has genotype g_i and trait value z_i. In the absence of nongenetic predictors, the most general model for a single SNP involves setting

$$\mu_i = 1_{\{g_i=1/1\}}\alpha_1 + 1_{\{g_i=1/2\}}\alpha_2 + 1_{\{g_i=2/2\}}\alpha_3.$$

Here, 1_C equals 1 when the condition C is true and 0 otherwise. The three parameters α_1, α_2, and α_3 are reduced to two in the additive model as a result of the constraint $\alpha_2 = \frac{1}{2}(\alpha_1 + \alpha_3)$. For initial screening, the additive model is adequate for two reasons. First, nature does act linearly in many cases. Second, if the minor allele frequency is low, then there will be little data to estimate the effect of the rare homozygote. Another reason for preferring the additive model is that the resulting test requires only one degree of freedom, rather than the two degrees of freedom required for the test under the general model. Thus, we recommend using the additive model for initial GWAS screening.

Epistatic modeling brings in a second SNP and more parameters. Suppose person i has genotype g_i at the first SNP and h_i at the second SNP. In the full epistatic model given by Cordell,[55]

$$\begin{aligned}\mu_i = {} & 1_{\{g_i=1/1,h_i=1/1\}}\gamma_1 + 1_{\{g_i=1/1,h_i=1/2\}}\gamma_2 + 1_{\{g_i=1/1,h_i=2/2\}}\gamma_3 \\ & + 1_{\{g_i=1/2,h_i=1/1\}}\gamma_4 + 1_{\{g_i=1/2,h_i=1/2\}}\gamma_5 + 1_{\{g_i=1/2,h_i=2/2\}}\gamma_6 \\ & + 1_{\{g_i=2/2,h_i=1/1\}}\gamma_7 + 1_{\{g_i=2/2,h_i=1/2\}}\gamma_8 + 1_{\{g_i=2/2,h_i=2/2\}}\gamma_9.\end{aligned}$$

This should be contrasted to the additive model, where the effects of the two loci are independent of one another:

$$\begin{aligned}\mu_i = {} & 1_{\{g_i=1/1\}}\alpha_1 + 1_{\{g_i=1/2\}}\alpha_2 + 1_{\{g_i=2/2\}}\alpha_3 \\ & + 1_{\{h_i=1/1\}}\beta_1 + 1_{\{h_i=1/2\}}\beta_2.\end{aligned}$$

Interaction can be tested by an F test or a likelihood ratio test. In the additive model, the term $1_{\{h_i=2/2\}}\beta_3$ is not omitted by accident. It is redundant because in its presence, μ_i is invariant under the substitution of $\alpha_j + c$ for α_j and of $\beta_k - c$ for β_k for any constant c.

With this degree of generality, testing for interactions results in a four-degrees-of-freedom test. To increase power, it is desirable to avoid the complexity of the full model.[55] If there is evidence for dominant or recessive effects, restrictions can be placed on the parameters that will reduce the degrees of freedom. In the absence of this information, assuming that the allelic effects are additive provides a reasonable compromise between genetic truth and parsimony. In that case, the interaction model with has five parameters and the additive model has four, for a single-degree-of-freedom test. Let x_i equal -1, 0, or 1 according as g_i equals 1/1, 1/2, or 2/2. Similarly, let y_i equal -1, 0, or 1 according as h_i equals 1/1, 1/2, or 2/2. In the simplified interaction model, $\mu_i = \gamma_1 + x_i\gamma_2 + y_i\gamma_3 + x_i y_i \gamma_4$. Here, interaction is tested by deciding whether $\gamma_4 = 0$. In practice, one should standardize the predictor values y_i to have mean 0 and variance 1.

In her recent review, Cordell[55] discusses tests of association allowing for interactions. In the context of the generalized linear model, assuming the allelic effects are additive, association of first SNP allowing for interaction with second SNP is tested by deciding whether $\gamma_3 = \gamma_4 = 0$. These tests will be less powerful than a single-locus test of association when there are no interactions, but they can be more powerful when there are interactions.[3]

In principle, there is nothing to prevent these models from being implemented on pedigree data as mean effects in a variance component model. In variance component models, computation time scales as the cube of the number of people within a pedigree, so screening vast number of interactions with large pedigrees would be unrealistic. Instead, we recommend testing for interactions with pedigree data in an attempt to replicate interactions detected as part of a GWAS. A beneficial side effect would be simultaneous adjustment for the polygenic background. When using a variance component model, a binary trait, such as disease affection status, can be treated as a quantitative trait that takes on only two values, zero and one. To our knowledge, no one has programmed this measured genotype model[82] with interactions, but it is clearly feasible in existing software.[79,83]

Prioritizing SNPs for Interaction Testing

The sheer number of interaction models is one of the most vexing problems plaguing GWAS. With n predictors, there are $\binom{n}{k} \approx n^k/k!$ k-way interactions. For $n = 10^6$ SNPs, this translates into nearly 5×10^{11} possible pairwise interactions. Whether testing this many potential interactions is feasible depends on computing resources. Ma and colleagues[84] calculate that it would take 5 years to conduct this many t tests where they use linear regression to estimate the parameter values and their standard errors on

their single-processor machine with 2000 individuals. Parallel computing reduces this time considerably,[3,84] but analyzing 5×10^{11} interactions with data from thousands of individuals takes weeks or months to complete.[55] As a separate issue, the problem of false positives versus power due to multiple testing is even more of a concern than it is for testing GWAS main effects.

Three strategies have been suggested to reduce the number of pairwise interactions considered. First, one can identify the top m marginal predictors and look at interactions only among them. For $m = 100$, this involves a manageable $\binom{100}{2} = 4950$ models. The second strategy is to consider all predictors paired with the top m predictors. There are now $\binom{m}{2} + m(n-m) \approx nm$ total models. For our example, $nm = 10^8$, which is a considerable increase over the number with the first strategy. The third strategy targets specific SNP pairs. These may be identified through protein-protein interactions, regulation by a common transcription factor, or participation in a common biochemical pathway.[85–87] The net cast can be wide or narrow. This approach overlaps with the approach of using molecular pathways to prioritize GWAS results that is discussed in the next section.

All three strategies involve a tradeoff between minimizing computation and multiple comparisons and maximizing power. Researchers can try all three strategies, but they should be honest about how finely they sift the data and adjust their criteria for choosing SNPs for follow-up. Strategy three may ultimately be the one of choice, but as discussed in the next section, currently it runs the risk of missing many important connections. The knowledge of networks stored in current databases is not up to the task. The more agnostic strategies one and two are perhaps safer if one is willing to pay the price of more multiple comparisons. In our own program Mendel,[79,80] the user can specify any of the three strategies.

Detecting Epistasis with Penalized Regression

A variety of alternative methods to classic regression have been proposed for detecting interactions. Cordell[55] provides a comprehensive review of Bayesian model selection approaches, penalized regression, and data-mining methods, including tree classification. Rather than repeat her comments, we focus specifically on one approach, lasso penalized regression,[80,81,88,89] that is particularly well suited to GWAS interaction detection when the number of SNPs vastly exceeds the number of subjects.

We begin by pointing out the difference between selecting interactions and formal testing of interactions. As a matter of principle, geneticists do not accept a single study as definitive, and all important findings are subject to replication. This attitude, whether justified or not in a preliminary study, places the emphasis on finding the most important SNPs for further study rather than on declaring their global significance. So instead of setting a significance level and only following up on interactions whose p values are smaller than that significance level, it is just as sensible to fix some number m, identify the m most significant associations, and then target these for follow-up with interaction studies. The value of m depends on resources. Because predictors are often correlated, via linkage disequilibrium, the m most significant predictors identified by simple linear regression may not constitute an optimal set for follow-up.

Fortunately, penalized regression is an ideal vehicle for finding a small subset of potent but weakly correlated predictors. This computational advance performs continuous model selection while avoiding some of the drawbacks of traditional forward and backward stepwise regression. It is also exceptionally quick computationally and adapts readily to allow screening of multiway interactions.

During the last 20 years, statisticians began to consider data sets where the number of predictors far exceeds the number of observations, precisely the situation in GWAS. Prior experience with shrinkage estimation suggested to statisticians and other mathematical scientists[88,90–93] that adding penalty terms to the likelihood or other objective function might stabilize parameter estimation. It soon became apparent that the form of the penalty was crucial. A ridge penalty, basically a sum of squares, was the most obvious choice for a penalty. After considerable experimentation and reflection, statisticians discovered that a sum of absolute values is more effective. This lasso penalty not only shrinks parameter estimates, it also zeros out the majority of them, thus achieving model selection. The strength of the penalty determines the number of predictors that enter a model.

Several authors have explored lasso penalized ordinary regression[81,89,94–96] in both the l_1 (least absolute deviation) and l_2 (least-squares) settings. Much of their work is largely preserved in logistic regression.[80] To move from ordinary to logistic regression, one simply substitutes the negative loglikelihood for the loss function. Logistic regression takes two to three times longer than ordinary regression. This approach to the identification of epistatic interactions is programmed and available in the Mendel software. The British celiac GWAS data[97] were analyzed for epistasis with the lasso regression module.[80] In an analysis of the 50 most significant associations, epistatic interactions between two human leukocyte antigen (HLA) SNPs and three SNPs on chromosomes 2, 3, and 8 were observed. It is particularly noteworthy that the univariate GWAS p values for these three non-HLA SNPs, considered as marginal effects, are far less impressive than their univariate p values as epistatic effects. Strong epistatic effects involving major histocompatibility complex (MHC) loci are not limited to celiac disease and may be a hallmark of diseases such as type 1 diabetes that have a strong association with MHC loci. For example, Barrett and colleagues[27] found a strong interaction between MHC loci and four non-MHC loci where the effects of non-MHC loci are

attenuated when an MHC risk allele is present. These analyses suggest that strong epistatic effects can be helpful in prioritizing SNPs for further study.

GWAS Pathway Analysis

GWAS pathway analysis (GWASPA) provides a means of integrating the results of a GWAS and the genes in a known molecular pathway to test whether the pathway is associated with the disorder. This approach is compelling, in that it addresses two important elements of post-GWAS prioritization: the selected pathway provides a biological vehicle for statistically combining GWAS association results, and an implicated pathway provides a biological interpretation. However, although GWASPA is appealing, the analytic methods are at an early stage of development, and additional factors regarding study design and statistical analysis need to be addressed. Here we (1) provide a conceptual framework and motivation for GWASPA, (2) highlight the factors to include, (3) present some online pathway resources, and (4) discuss design and statistical issues for successful GWASPA. Throughout this section, publications that identify and address analytic issues in GWASPA are reviewed, and guidelines for conducting GWASPA with the currently available tools are suggested. We do not discuss the methods that are being used to develop and establish pathways in this review, although GWAS results could also be used to support that process.

Defining Pathways and GWASPA Motivation

Although the term "pathway" is often used to describe molecular processes, its definition and application are context dependent. Pathways have been used to represent a wide range of biological processes that include cell functions, metabolic processes, biosynthesis, genetic information processing such as DNA repair, cell signaling, immune responses, features of embryo development, and factors leading to human diseases. In the context of the cell, a pathway represents a series of actions among molecules that lead to a particular endpoint or cell function. The genes that coordinate to achieve a specific task are grouped together in the same pathway to reflect that process. Those genes are sometimes referred to as gene sets and are often put on a pathway diagram that indicates the order in which the genes act and interact within the pathway. Alternative paths involving the same or different genes that lead to the same end result are also part of the pathway. There is a growing list of pathways that are at varying stages of completion, and the same genes appear in many of them. In addition, there may be ambiguity about the structure of a pathway, making pathway development an active area of investigation and the application of pathway methods difficult to conduct with certainty. Given these constraints, GWASPA, although compelling on a theoretical level, have been challenging to apply and interpret. The application of pathways is likely to identify larger genetic effects than those seen with GWAS, and a recent commentary on GWAS[7] advocates GWASPA as the next step in the process of GWAS data mining.

A plausible model that motivates the application of GWASPA derives from the recognition that biological processes result from reaching successful endpoints in multiple molecular pathways and that a sufficient number of small disruptions in a sufficient fraction of those pathways may lead to a disorder. This molecular genetic model is similar to the older statistical multifactorial/threshold model (MTM), which postulates that small equal and additive variants in the genes that contribute to a biological process accumulate until a risk threshold is crossed, resulting in a disorder. The classic disorder that illustrates the MTM is the birth defect of cleft palate[98] (CL/P [MIM 119530]), where the closing of the structure that becomes the palate is contingent on a number of successfully timed molecular processes. If we postulate that the variants that accumulate are in genes contained within pathways that control palate closure, a GWASPA can be viewed as a natural extension of the MTM when SNP association data are available. A GWASPA of cleft palate would be accomplished by identifying the associated GWAS SNPs, postulating the likely pathways involved during fetal development, and connecting them analytically.

Pathway heterogeneity is an important factor, because disruptions in different pathways are likely to lead to the same disorder. Although affected individuals may share the same disrupted pathways, the mutated genes or variants within those pathways are likely to differ. GWASPA accommodate and capitalize upon this substantial degree of genetic heterogeneity. In addition, multiple GWAS are easily combined in a way that is potentially more powerful than the combinations of individual genes by meta-analyses. Genetic heterogeneity among ethnic groups, which has been a source of concern in GWAS, will not affect GWASPA if ethnic groups each contribute associations with different genes and alleles in the same pathways. This strength of GWASPA is illustrated by the analysis of the IL12/IL23 pathway in multiple ethnic groups having Crohn's disease[99] (IBD1 [MIM 266600]).

Critical Factors for a GWASPA

A successful GWASPA involves five steps that are enumerated here and discussed in greater detail subsequently. The first step is to select one or more pathways for the GWASPA. Biological insight and perhaps postanalysis examination of GWAS results drive this. The second is to select the most appropriate database (or databases) to delineate the genes in the pathway (or pathways). There is substantial variation regarding the genes in pathways named to represent the same process, and the specific genes that comprise them may be contingent upon the database used. The third is to assign the GWAS SNPs to known genes within the selected pathway, as given in the selected database. Rules for gene assignment or SNP exclusion are required because, given our current knowledge of the genome, a substantial number of the SNPs on GWAS platforms do not explicitly tag particular genes. The fourth requires a pathway scoring system that

addresses the biases inherent in the unequal distribution of SNPs among pathways resulting from differences in pathway and gene size and SNP density. The fifth is to identify a statistical approach for aggregating the GWAS results that allows one to formally test the selected pathway for association with the disorder under analysis.

Choices made for these five factors will have a substantial impact on the GWASPA outcome, and our recommendation is that each of the factors be addressed and justified for each study. Although fixed guidelines would simplify GWASPA substantially by providing analytic tools that streamline the process, they would also remove much of the creativity from the studies in this nascent stage of development, making their development and application premature. Therefore, we suggest and anticipate that GWASPA for a given trait will be tailored based on the investigators' backgrounds, biological insights making it doubtful that fixed guidelines about the choice of database, algorithm to tag genes with SNPs, or statistical test will be constructive. However, GWASPA manuscripts should be expected to justify the choices, and their results should be interpreted in a context-specific manner. The next subsections are written to help elucidate those choices.

Identifying Pathways for GWASPA

There are two general approaches to identify pathways for GWASPA. The first and most straightforward is to formulate a prior hypothesis regarding the pathways that are likely to be involved in the disorder. This approach is similar to the one used in candidate gene studies. Testing candidates has become obsolete given the unbiased nature of GWAS, but as yet, no approach to an unbiased evaluation of pathways has been proposed, and selecting "candidate pathways" remains the method of choice for prior selection. As an example, the pathways containing genes associated with inflammation are good candidates for GWASPA of autoimmune diseases. Multiple inflammation pathways can be tested for a single autoimmune disorder, or GWAS of multiple autoimmune disorders can be combined in the same GWASPA. The differences in genotyping platforms and sample sizes will have to be considered, just as they are when conducting a meta-analysis.

The second approach uses the results of the GWAS to guide the choice of candidate pathways. Tests that use this information are biased by the observed results, and although the methods of analysis are the same as those in the first option, the results will have to be interpreted within this context. That is, the GWASPA merely assess whether the accumulated data are consistent with a role for the pathway. We recommend that replication in an independent study sample be performed to implicate the pathway formally. If replication samples are available, this approach provides an appealing GWASPA study design.

Choosing Pathway Databases for GWASPA

Information regarding the structures of specific pathways is stored within web-based databases. A comprehensive listing of these pathway resources is in the Pathguide, which currently provides information for about 300 databases that are evolving rapidly in both size and number. Among those described, the Kyoto Encyclopedia of Genes and Genomes (KEGG) is easy to access and commonly used. It is being developed at Kyoto University and provides gene lists, diagrams, and pathway classification tools for many aspects of biology. Alternatively, the Gene Ontology (GO) project seeks to provide a set of structured vocabularies for specific biological domains that can be used to describe gene products in any organism. It was initiated by scientists associated with model organism databases, and each of these model organism information systems is annotating genes and gene products with GO vocabulary terms. Tools enabling curators and researchers to query and manipulate the vocabularies are provided. Analogous information is provided in the Database for Annotation, Visualization and Integrated Discovery (DAVID) and Protein ANalysis THrough Evolutionary Relationships (PANTHER). These databases were built separately, address different audiences with different research questions, and are evolving as our knowledge of pathways continues to develop. Thus, GWASPA results will be contingent upon the pathway resource used, which is clearly illustrated in a recently published study by Elbers and colleagues.[100] The authors use several currently available databases to conduct GWASPA of the same Wellcome Trust Case Control Consortium (WTCCC) data, a publicly available resource. They conclude that although the analyses can highlight the relevant gene associations, the results are likely to be biased by which database is used. The authors find that the differences in study outcome reflect the differences in the pathway information included in the resource and the way in which the information is organized. Thus, for GWASPA, we recommend using several databases for a single pathway and incorporating an algorithm for deciding whether the pathway is associated when the results are inconsistent across the databases.

Assigning SNPs to Genes in Pathways

Once the pathway and database (or databases) are selected, SNPs in the GWAS panel should be assigned to specific genes in the pathway. Current genotyping platforms select SNPs based on linkage disequilibrium patterns that support coverage of the base-pair positions across the chromosomes rather than coverage of genes in particular pathways, making assignment of some SNPs difficult. This problem has already been observed in GWAS, where strong associations have sometimes been difficult to interpret because of the location of the SNP in relation to the flanking genes. For SNPs in the coding and known regulatory regions, assignment is straightforward. For the other SNPs, rules for assignment should be made, although they are likely to be considered arbitrary. As an example, Torkamani and colleagues,[101] who are interested in illustrating the polygenic and multiple-pathway nature of complex disorders, consider the possibility of an SNP

mapping to multiple genes. They set a specific hierarchy for gene elements to help with their assignment, where their scheme is coding > intronic > 5′UTR > 3′UTR > 5′ upstream > 3′ upstream. It is difficult to assign SNPs that are 500 or more kb from a known gene, and these may have to be excluded from the analysis. We recommend constructing a formal rule for assigning SNPs to genes based on the most current research regarding this problem.

Scoring Pathways in an Unbiased Fashion

Several sources of bias that derive from the differences in pathway and gene size and SNP density must be addressed. Well-developed pathways are more likely to be included in the publicly available databases and are therefore more likely to be identified by GWASPA. Larger pathways containing greater numbers of genes, and thus larger numbers of genotyped SNPs, are expected to show more associated SNPs by chance alone. These introduce bias into GWASPA. SNP density is not consistent among genes and contributes a similar source of bias. Unless corrections are made, there is bias toward implicating the pathways that are large and well known with large genes that are more densely covered by the SNPs in the GWAS panels. This problem has been recognized by some investigators applying GWASPA. Currently, the most frequent approach is to select a single SNP with the strongest association signal from each gene. This is not optimal, because it does not remove all of the bias, and the significance of pathways that contain a few genes with multiple independent association signals may be lost. In a recent analysis, Holmans and colleagues[102] correct for linkage disequilibrium among SNPs and gene size by only counting each gene once in the analysis, regardless of how many significant associations were observed. Yu and colleagues[103] also treat genes as the units of analysis, and for each gene they provide the most significant multiple testing-adjusted p value. Wang and colleagues[104] also focus on the gene and use the SNP with the most significant permutation p value to represent it. They recognize that although only one SNP is used per gene, the size of the gene and pathway and the density of SNPs are still factors that could contribute to the type 1 error. The authors use an enrichment score based on a weighted Kolmogorov-Smirnov running sum statistic calculated for the genes in the known pathway, and to address bias, this enrichment score is normalized so that different pathways are directly comparable.

Statistical Considerations in GWASPA

A formal GWASPA includes a test statistic that reflects the aggregation of GWAS associations for the genes in the tested pathway. The observed value of the statistic is compared to its expected value under the null hypothesis of no association with the pathway and compared to its standard error. It is difficult to formulate a null distribution from which to derive these values. Permutation tests have been conducted to assess significance, and fortunately, they can also be used to adjust for many sources of bias. A choice of what to permute is sometimes not straightforward; the investigators may have to use the original genotype data, and the computational burden can become substantial. It is also possible to construct a null distribution with scores from the pathways that are deemed unlikely to be associated with the disorder, but each pathway has a different configuration, making the choices for this analysis arbitrary. Additional factors that must be considered are the choice of a cutoff value for classifying a gene as having a positive signal, the methods by which the p values are estimated and aggregated, the approach used to correct for multiple testing, and the level of significance that is set to implicate a pathway. We recommend justifying these choices in reporting and interpreting GWASPA. Methods presenting options for these choices are discussed below.

The scarcity of formal GWASPA methods has not been a deterrent to their application. The manuscripts developing and using this approach are creative and provide encouragement, although there are clear ambiguities in their findings. Lesnick and colleagues[105] demonstrate that SNPs in the axon guidance pathway collectively predispose to Parkinson's disease (PD [MIM 168600]) even though no individual SNP reaches genome-wide significance level; Dinu and colleagues[106] show that multiple complement factor pathway genes, in addition to complement factor H, are associated with the risk for developing age-related macular degeneration (ARMD1 [MIM 603075]); and Askland and colleagues[107] implicate genes mediating ion channel activity and synaptic neurotransmission in bipolar disorder (MAFD1 [MIM 125480]), illustrating the broad application of GWASPA.

Several manuscripts have made the analytic approaches to GWASPA their focus.[5,100–104,108] An early method developed by Wang and colleagues[104] uses the gene-set enrichment analysis (GSEA) algorithm to aggregate p values[109]. Their approach has been motivated by analyses of microarray data where genes in the pathway are ranked by their strongest empirical p value test statistic. The test statistics for genes within pathways are aggregated via a Kolmogorov-Smirnov-like running sum. The p value is estimated by permutations, where the phenotype labels of cases and controls are permuted a fixed number of times, and the pathway score is calculated for each permutation. A z score, which is the difference between the observed value of the score and the mean of the scores for the permutations divided by the standard deviation of the scores for the permutations, provides the overall level of significance. Using this approach, they implicated the involvement of cell adhesion molecules in autism spectrum disorder (ASD [MIM 209850]). In a more extensive analysis via the same approach, the Wnt signaling pathway was studied in type 2 diabetes (NIDDM [MIM 125853]) and did not survive a p value correction for the testing of multiple pathways.[110]

More recently, Yu and colleagues[103] presented an alternative approach to accumulating p values for GWASPA. They base their analysis on the adaptive rank truncated product statistic p value, which selects for analysis of the

k most significant genes or the set of genes with a p value less than a certain threshold. Their analysis optimizes the selection of truncation points and corrects for multiple testing by permutation analyses. Through simulation, they illustrate the important statistical property that the type 1 error rate is the expected 0.05 under different analyses for two different pathway models, thus indicating that their permutations successfully correct for the bias introduced by gene and pathway size and SNP density. The authors also provide a comprehensive analysis of their method and that of Wang[104] and indicate that their analytic approach is less computationally intensive.

In a very recent example of how GWASPA can be used to accumulate relatively weak SNP associations, Baranzini and colleagues[111] report the details of a follow-up analysis of multiple sclerosis. They begin their work with the premise that many of the SNP associations that have not been classified associated under a genome-wide level of significance are in fact true associations with effects that are too small to detect without their aggregation into pathways. Guidance in pathway construction is provided by known protein interactions and the GO database.[112] As with previous studies, one marker is selected for each gene in a pathway, and associations occurring in gene deserts are excluded. In order to identify those pathways specific to multiple sclerosis and other autoimmune, neurological, and unrelated diseases, data from the WTCCC are analyzed in a similar fashion, and their outcomes as well as pathway randomizations provide the control results for comparison. More specifically, the test statistic accumulates a score by converting the p value for each gene into a z score that is accumulated over the genes in the pathway contributing to the score. The expected value for the sum is derived from a random set of genes from the genome and their permutations, and the test statistic is a normal score. The choice of genes to include in the random set impacts score. A formal discussion of the features of their statistical approach is not provided, but a clear description of the method is given. As expected for preliminary groundbreaking studies, the authors indicate that the interpretation of the implicated pathways for multiple sclerosis is not straightforward, and their findings are somewhat ambiguous.

The statistical methods for GWASPA that have been described and applied differ mainly in the approach used to accumulate the associated SNPs. A formal analysis could clarify whether any of these is expected to be more powerful than the others. The methods involve permutation testing, which can incur a computational burden, although they address some of the sources of bias. Problems associated with testing multiple pathways are not straightforward and remain to be addressed. The most conservative approach would be a Bonferroni correction, based on the number of pathways tested, although the pathways are not necessarily independent. We recommend that each factor in the GWASPA be examined carefully before any inferences are drawn, and the most effective approach to validate the inferences would be a replication in a comparable independent study sample.

Future Directions

This review focuses on three statistical methods that can be used to prioritize GWAS results for in-depth follow-up studies and reveal associations that were not detected initially. Thus, our first suggestion for future studies is to apply these methods to the available GWAS data following our recommendations. As the analytic methods evolve and new approaches are published, the results can be reconsidered and evaluated. Additional suggestions are targeted toward individuals who are focused on developing statistical methodologies. We enumerate some of the analytic problems that need solutions below. The WTCCC data have been used to evaluate some of the methods reviewed here and may be appropriate for the new ones that are developed. The available data for these purposes may increase as studies are published. It is possible that forums and workshops could be conducted to compare and improve these methods.

Besides the already discussed challenges of defining appropriate priors for Bayesian meta-analyses, a number of challenges exist. As the number of studies increases, so can heterogeneity and errors. Thus, the development of better tests of heterogeneity and other post hoc quality-control measures are needed to improve the reliability of meta-analysis. Sequence data will increase the number and nature of the polymorphisms that need to be considered in a meta-analysis. Thus, we will need methods that allow incorporation of association results from sequence data to be combined with SNP chip data in a cumulative meta-analysis. With a few exceptions,[47,48] most meta-analyses consider the results for each SNP on its own. Thus, there is a need for highly computationally efficient methods that can model the effects of multiple SNPs in a region as well as gene-by-gene interactions.

The challenge for detection of epistasis is better understanding of the underlying biological processes rather than new statistical methods. Elucidation of metabolic and catabolic processes would help us understand the effects of the environment. When gene × environment interactions are not adequately taken into account, the residual trait variation may overwhelm the evidence for epistasis. In addition, detection of epistatic effects is handicapped by the number of potential tests. Issues regarding multiple testing can be circumvented by development of informative priors that capitalize on network relationships or evidence from realistic animal models.

A number of important analytic issues remain for those contributing to GWASPA methodology. It is the newest and least developed approach of the three discussed here. The most general issue that, if solved, would make a substantial contribution to these studies would be to develop a method to examine all possible pathways within some defined domain so that an unbiased approach could be taken in identifying the associated pathway (or

pathways). Such approaches will be computationally intensive and biologically challenging. The current methods select the most significantly associated SNP from each gene. But it may be that the pathways with genes having multiple independent associated SNPs should be weighted more heavily. GWASPA methods to establish these contrasts may be very useful in pathway prioritization. Accumulating the effects of multiple types of mutations in one analysis can make the statistical power to implicate pathways substantial, and the emerging methods of GWASPA should be based on an appeal to this complexity. Risk alleles for a given disorder may be all common or all rare, but it is most likely that the alleles will be drawn from both categories. GWASPA methods should be designed to capture both. Associated alleles in a pathway may also exhibit epistasis. GWASPA methods should be developed to include these effects. However, epistasis is tested by the formal models, and it will not be straightforward to incorporate this feature. The permutation studies that are currently being conducted require raw genotype data, which are not always available. An important methodological improvement would make the p values alone, and not raw data, the basis of analysis in GWASPA.

In conclusion, although these three statistical approaches have limitations and analytical challenges, they also provide a means to prioritize genes for bioinformatics and laboratory studies focused on identifying causal variants and their biological roles. In addition, post-GWAS studies are likely to provide a clearer picture of the true role of common variants in common complex disorders.

Acknowledgments

Support for this work was provided by the Database and Statistics Core of National Institutes of Health (NIH) grant HL28481 and NIH grants MH59490 and GM53275. We thank Jeonglim Yoon for technical assistance and Karen Reue for critical reading and comments. We also appreciate critical input from three anonymous reviewers.

Web Resources

The URLs for data presented herein are as follows:

Cross-Disorder Phenotype Group of the Psychiatric GWAS Consortium, http://pgc.unc.edu/

Database for Annotation, Visualization and Integrated Discovery (DAVID), http://david.abcc.ncifcrf.gov/

Gene Ontology Annotation (GOA) database, http://www.ebi.ac.uk/GOA

HapMap, http://hapmap.ncbi.nlm.nih.gov/

Kyoto Encyclopedia of Genes and Genomes (KEGG), http://www.genome.jp/kegg/

Mendel software, http://www.genetics.ucla.edu/software/mendel

MetaCore software, http://www.genego.com/metacore.php

METAL software, http://www.sph.umich.edu/csg/abecasis/metal/

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/

Protein ANalysis THrough Evolutionary Relationships (PANTHER), http://www.pantherdb.org

R script (Chen and Witte), http://www.epibiostat.ucsf.edu/witte_lab/gwa.htm

Wellcome Trust Case Control Consortium (WTCCC), http://www.wtccc.org.uk/ccc2/wtccc2_studies.shtml

References

1. Goldstein, D.B. (2009). Common genetic variation and human traits. N. Engl. J. Med. *360*, 1696–1698.
2. Halperin, E., and Eskin, E. (2004). Haplotype reconstruction from genotype data using Imperfect Phylogeny. Bioinformatics *20*, 1842–1849.
3. Marchini, J., Donnelly, P., and Cardon, L.R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat. Genet. *37*, 413–417.
4. Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. *78*, 629–644.
5. Saccone, S.F., Saccone, N.L., Swan, G.E., Madden, P.A., Goate, A.M., Rice, J.P., and Bierut, L.J. (2008). Systematic biological prioritization after a genome-wide association study: An application to nicotine dependence. Bioinformatics *24*, 1805–1811.
6. Kraft, P., and Hunter, D.J. (2009). Genetic risk prediction—are we there yet? N. Engl. J. Med. *360*, 1701–1703.
7. Hirschhorn, J.N. (2009). Genomewide association studies—illuminating biologic pathways. N. Engl. J. Med. *360*, 1699–1701.
8. Hardy, J., and Singleton, A. (2009). Genomewide association studies and human disease. N. Engl. J. Med. *360*, 1759–1768.
9. DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. Control. Clin. Trials *7*, 177–188.
10. Fleiss, J.L. (1993). The statistical basis of meta-analysis. Stat. Methods Med. Res. *2*, 121–145.
11. Yesupriya, A., Yu, W., Clyne, M., Gwinn, M., and Khoury, M.J. (2008). The continued need to synthesize the results of genetic associations across multiple studies. Genet. Med. *10*, 633–635.
12. Lau, J., Ioannidis, J.P., and Schmid, C.H. (1997). Quantitative synthesis in systematic reviews. Ann. Intern. Med. *127*, 820–826.
13. Allison, D.B., and Schork, N.J. (1997). Selected methodological issues in meiotic mapping of obesity genes in humans: Issues of power and efficiency. Behav. Genet. *27*, 401–421.
14. Ioannidis, J.P., Gwinn, M., Little, J., Higgins, J.P., Bernstein, J.L., Boffetta, P., Bondy, M., Bray, M.S., Brenchley, P.E., Buffler, P.A., et al. Human Genome Epidemiology Network and the Network of Investigator Networks. (2006). A road map for efficient and reliable human genome epidemiology. Nat. Genet. *38*, 3–5.
15. de Bakker, P.I., Ferreira, M.A., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. *17*(*R2*), R122–R128.
16. Sagoo, G.S., Little, J., Higgins, J.P., and Human Genome Epidemiology Network. (2009). Systematic reviews of genetic association studies. PLoS Med. *6*, e28.
17. Zeggini, E., and Ioannidis, J.P. (2009). Meta-analysis in genome-wide association studies. Pharmacogenomics *10*, 191–201.

18. Levy, D., Ehret, G.B., Rice, K., Verwoert, G.C., Launer, L.J., Dehghan, A., Glazer, N.L., Morrison, A.C., Johnson, A.D., Aspelund, T., et al. (2009). Genome-wide association study of blood pressure and hypertension. Nat. Genet. Published online May 10, 2009. 10.1038/ng.384.
19. Ioannidis, J.P. (2007). Non-replication and inconsistency in the genome-wide association setting. Hum. Hered. *64*, 203–213.
20. Higgins, J.P., Thompson, S.G., Deeks, J.J., and Altman, D.G. (2003). Measuring inconsistency in meta-analyses. BMJ *327*, 557–560.
21. Zintzaras, E., and Lau, J. (2008). Synthesis of genetic association studies for pertinent gene-disease associations requires appropriate methodological and statistical approaches. J. Clin. Epidemiol. *61*, 634–645.
22. Coenen, M.J., Trynka, G., Heskamp, S., Franke, B., van Diemen, C.C., Smolonska, J., van Leeuwen, M., Brouwer, E., Boezen, M.H., Postma, D.S., et al. (2009). Common and different genetic background for rheumatoid arthritis and coeliac disease. Hum. Mol. Genet. *18*, 4195–4203.
23. Craddock, N., Kendler, K., Neale, M., Nurnberger, J., Purcell, S., Rietschel, M., Perlis, R., Santangelo, S.L., Schulze, T., Smoller, J.W., Thapar, A., and Cross-Disorder Phenotype Group of the Psychiatric GWAS Consortium. (2009). Dissecting the phenotype in genome-wide association studies of psychiatric illness. Br. J. Psychiatry *195*, 97–99.
24. Lettre, G., and Rioux, J.D. (2008). Autoimmune diseases: Insights from genome-wide association studies. Hum. Mol. Genet. *17*(*R2*), R116–R121.
25. Moskvina, V., Craddock, N., Holmans, P., Nikolov, I., Pahwa, J.S., Green, E., Owen, M.J., O'Donovan, M.C., and Wellcome Trust Case Control Consortium. (2009). Gene-wide analyses of genome-wide association data sets: Evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. Mol. Psychiatry *14*, 252–260.
26. Evangelou, E., Maraganore, D.M., and Ioannidis, J.P. (2007). Meta-analysis in genome-wide association datasets: Strategies and application in Parkinson disease. PLoS ONE *2*, e196.
27. Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C., et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. Published online May 10, 2009. 10. 1038/ng.381.
28. Cooper, J.D., Smyth, D.J., Smiles, A.M., Plagnol, V., Walker, N.M., Allen, J.E., Downes, K., Barrett, J.C., Healy, B.C., Mychaleckyj, J.C., et al. (2008). Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat. Genet. *40*, 1399–1401.
29. De Jager, P.L., Jia, X., Wang, J., de Bakker, P.I., Ottoboni, L., Aggarwal, N.T., Piccio, L., Raychaudhuri, S., Tran, D., Aubin, C., , et al.International MS Genetics Consortium. (2009). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat. Genet. *41*, 776–782.
30. Kathiresan, S., Willer, C.J., Peloso, G.M., Demissie, S., Musunuru, K., Schadt, E.E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., et al. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. Nat. Genet. *41*, 56–65.
31. Kolz, M., Johnson, T., Sanna, S., Teumer, A., Vitart, V., Perola, M., Mangino, M., Albrecht, E., Wallace, C., Farrall, M., et al. EUROSPAN ConsortiumENGAGE ConsortiumPROCARDIS ConsortiumKORA StudyWTCCC. (2009). Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. PLoS Genet. *5*, e1000504.
32. Lindgren, C.M., Heid, I.M., Randall, J.C., Lamina, C., Steinthorsdottir, V., Qi, L., Speliotes, E.K., Thorleifsson, G., Willer, C.J., Herrera, B.M., et al. Wellcome Trust Case Control ConsortiumProcardis ConsortiaGiant Consortium. (2009). Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. PLoS Genet. *5*, e1000508.
33. Eddy, D.M., Hasselblad, V., and Schacter, R. (1991). Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence (New York: Academic Press).
34. Fisher, R. (1948). Combining independent tests of significance. Am. Stat. *2*, 30.
35. Timpson, N.J., Tobias, J.H., Richards, J.B., Soranzo, N., Duncan, E.L., Sims, A.M., Whittaker, P., Kumanduri, V., Zhai, G., Glaser, B., et al. (2009). Common variants in the region around Osterix are associated with bone mineral density and growth in childhood. Hum. Mol. Genet. *18*, 1510–1517.
36. Soranzo, N., Rivadeneira, F., Chinappen-Horsley, U., Malkina, I., Richards, J.B., Hammond, N., Stolk, L., Nica, A., Inouye, M., Hofman, A., et al. (2009). Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. PLoS Genet. *5*, e1000445.
37. Tanaka, T., Scheet, P., Giusti, B., Bandinelli, S., Piras, M.G., Usala, G., Lai, S., Mulas, A., Corsi, A.M., Vestrini, A., et al. (2009). Genome-wide association study of vitamin B6, vitamin B12, folate, and homocysteine blood concentrations. Am. J. Hum. Genet. *84*, 477–482.
38. Houlston, R.S., Webb, E., Broderick, P., Pittman, A.M., Di Bernardo, M.C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S., et al.Colorectal Cancer Association Study ConsortiumCoRGI ConsortiumInternational Colorectal Cancer Genetic Association Consortium. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat. Genet. *40*, 1426–1435.
39. Newton-Cheh, C., Eijgelsheim, M., Rice, K.M., de Bakker, P.I., Yin, X., Estrada, K., Bis, J.C., Marciante, K., Rivadeneira, F., Noseworthy, P.A., et al. (2009). Common variants at ten loci influence QT interval duration in the QTGEN Study. Nat. Genet. *41*, 399–406.
40. Dempfle, A., and Loesgen, S. (2004). Meta-analysis of linkage studies for complex diseases: An overview of methods and a simulation study. Ann. Hum. Genet. *68*, 69–83.
41. Egger, M., Smith, G.D., and Phillips, A.N. (1997). Meta-analysis: Principles and procedures. BMJ *315*, 1533–1537.
42. Ioannidis, J.P., Patsopoulos, N.A., and Evangelou, E. (2007). Heterogeneity in meta-analyses of genome-wide association investigations. PLoS ONE *2*, e841.
43. Ferrucci, L., Perry, J.R., Matteini, A., Perola, M., Tanaka, T., Silander, K., Rice, N., Melzer, D., Murray, A., Cluett, C., et al. (2009). Common variation in the beta-carotene 15,15′-monooxygenase 1 gene affects circulating levels of carotenoids: A genome-wide association study. Am. J. Hum. Genet. *84*, 123–133.
44. Johnson, A.D., Kavousi, M., Smith, A.V., Chen, M.H., Dehghan, A., Aspelund, T., Lin, J.P., van Duijn, C.M., Harris, T.B., Cupples, L.A., et al. (2009). Genome-wide association

meta-analysis for total serum bilirubin levels. Hum. Mol. Genet. *18*, 2700–2710.
45. Fisher, S.A., Tremelling, M., Anderson, C.A., Gwilliam, R., Bumpstead, S., Prescott, N.J., Nimmo, E.R., Massey, D., Berzuini, C., Johnson, C., et al. Wellcome Trust Case Control Consortium. (2008). Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. Nat. Genet. *40*, 710–712.
46. Smyth, D.J., Plagnol, V., Walker, N.M., Cooper, J.D., Downes, K., Yang, J.H., Howson, J.M., Stevens, H., McManus, R., Wijmenga, C., et al. (2008). Shared and distinct genetic variants in type 1 diabetes and celiac disease. N. Engl. J. Med. *359*, 2767–2777.
47. Newcombe, P.J., Verzilli, C., Casas, J.P., Hingorani, A.D., Smeeth, L., and Whittaker, J.C. (2009). Multilocus Bayesian meta-analysis of gene-disease associations. Am. J. Hum. Genet. *84*, 567–580.
48. Verzilli, C., Shah, T., Casas, J.P., Chapman, J., Sandhu, M., Debenham, S.L., Boekholdt, M.S., Khaw, K.T., Wareham, N.J., Judson, R., et al. (2008). Bayesian meta-analysis of genetic association studies with different sets of markers. Am. J. Hum. Genet. *82*, 859–872.
49. Lewinger, J.P., Conti, D.V., Baurley, J.W., Triche, T.J., and Thomas, D.C. (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genet. Epidemiol. *31*, 871–882.
50. Chen, G.K., and Witte, J.S. (2007). Enriching the analysis of genomewide association studies with hierarchical modeling. Am. J. Hum. Genet. *81*, 397–404.
51. Carrasquillo, M.M., McCallion, A.S., Puffenberger, E.G., Kashuk, C.S., Nouri, N., and Chakravarti, A. (2002). Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. Nat. Genet. *32*, 237–244.
52. Lou, X.Y., Chen, G.B., Yan, L., Ma, J.Z., Mangold, J.E., Zhu, J., Elston, R.C., and Li, M.D. (2008). A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. Am. J. Hum. Genet. *83*, 457–467.
53. Howard, T.D., Koppelman, G.H., Xu, J., Zheng, S.L., Postma, D.S., Meyers, D.A., and Bleecker, E.R. (2002). Gene-gene interaction in asthma: IL4RA and IL13 in a Dutch population with asthma. Am. J. Hum. Genet. *70*, 230–236.
54. Small, K.M., Wagoner, L.E., Levin, A.M., Kardia, S.L., and Liggett, S.B. (2002). Synergistic polymorphisms of beta1- and alpha2C-adrenergic receptors and the risk of congestive heart failure. N. Engl. J. Med. *347*, 1135–1142.
55. Cordell, H.J. (2009). Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. Nat. Rev. Genet. *10*, 392–404.
56. Musani, S.K., Shriner, D., Liu, N., Feng, R., Coffey, C.S., Yi, N., Tiwari, H.K., and Allison, D.B. (2007). Detection of gene x gene interactions in genome-wide association studies of human population data. Hum. Hered. *63*, 67–84.
57. Hastie, R., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (New York: Springer)., p. 13.
58. Strobl, C., Boulesteix, A.L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics *8*, 25.
59. Hahn, L.W., Ritchie, M.D., and Moore, J.H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics *19*, 376–382.
60. Moore, J.H., Gilbert, J.C., Tsai, C.T., Chiang, F.T., Holden, T., Barney, N., and White, B.C. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J. Theor. Biol. *241*, 252–261.
61. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., and Moore, J.H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. *69*, 138–147.
62. Nelson, M.R., Kardia, S.L., Ferrell, R.E., and Sing, C.F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res. *11*, 458–470.
63. Culverhouse, R., Klein, T., and Shannon, W. (2004). Detecting epistatic interactions contributing to quantitative traits. Genet. Epidemiol. *27*, 141–152.
64. Long, Q., Zhang, Q., and Ott, J. (2009). Detecting disease-associated genotype patterns. BMC Bioinformatics *10* (*Suppl 1*), S75.
65. Horvath, S., Xu, X., and Laird, N.M. (2001). The family based association test method: Strategies for studying general genotype–phenotype associations. Eur. J. Hum. Genet. *9*, 301–306.
66. Rabinowitz, D., and 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.
67. Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., Huang, W., and Li, Y. (2008). Exploration of gene-gene interaction effects using entropy-based methods. Eur. J. Hum. Genet. *16*, 229–235.
68. Kang, G., Yue, W., Zhang, J., Cui, Y., Zuo, Y., and Zhang, D. (2008). An entropy-based approach for testing genetic epistasis underlying complex diseases. J. Theor. Biol. *250*, 362–374.
69. McKinney, B.A., Reif, D.M., White, B.C., Crowe, J.E. Jr., and Moore, J.H. (2007). Evaporative cooling feature selection for genotypic data involving interactions. Bioinformatics *23*, 2113–2120.
70. Moore, J.H., and White, B.C. (2007). Tuning ReliefF for genome-wide genetic analysis. Lect. Notes Comput. Sci. *4447*, 166–175.
71. Robnik-Sikonja, M., and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. *53*, 23–69.
72. Motsinger, A.A., Lee, S.L., Mellick, G., and Ritchie, M.D. (2006). GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. BMC Bioinformatics *7*, 39.
73. Nunkesser, R., Bernholt, T., Schwender, H., Ickstadt, K., and Wegener, I. (2007). Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. Bioinformatics *23*, 3280–3288.
74. Kooperberg, C., and Ruczinski, I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. Genet. Epidemiol. *28*, 157–170.
75. Li, Z., Zheng, T., Califano, A., and Floratos, A. (2007). Pattern-based mining strategy to detect multi-locus association and gene x environment interaction. BMC Proc *1* (*Suppl 1*), S16.

76. Zhang, Y., and Liu, J.S. (2007). Bayesian inference of epistatic interactions in case-control studies. Nat. Genet. *39*, 1167–1173.
77. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
78. Schwartz, D.F., Ziegler, A., and Konig, I.R. (2008). Beyond the results of genomewide association studies. Genet. Epidemiol. *32*, 671.
79. Lange, K., Cantor, R., Horvath, S., Perola, M., Sabatti, C., Sinsheimer, J., and Sobel, E. (2001). Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. Am. J. Hum. Genet. *69* (*Suppl*), S504.
80. Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E.M., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics *25*, 714–721.
81. Wu, T.T., and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. Ann. Appl. Stat. *2*, 224–244.
82. Boerwinkle, E., Chakraborty, R., and Sing, C.F. (1986). The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann. Hum. Genet. *50*, 181–194.
83. Lange, K., Sinsheimer, J.S., and Sobel, E. (2005). Association testing with Mendel. Genet. Epidemiol. *29*, 36–50.
84. Ma, L., Runesha, H.B., Dvorkin, D., Garbe, J.R., and Da, Y. (2008). Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. BMC Bioinformatics *9*, 315.
85. Emily, M., Mailund, T., Hein, J., Schauser, L., and Schierup, M.H. (2009). Using biological networks to search for interacting loci in genome-wide association studies. Eur. J. Hum. Genet. *17*, 1231–1240.
86. Mechanic, L.E., Luke, B.T., Goodman, J.E., Chanock, S.J., and Harris, C.C. (2008). Polymorphism Interaction Analysis (PIA): A method for investigating complex gene-gene interactions. BMC Bioinformatics *9*, 146.
87. Pattin, K.A., and Moore, J.H. (2008). Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. Hum. Genet. *124*, 19–29.
88. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B *58*, 267–288.
89. Fu, W.J. (1998). Penalized regressions: The bridge versus the lasso. J. Comput. Graph. Stat. *7*, 397–416.
90. Chen, S.S., Donoho, D.L., and Saunders, M.A. (1998). Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. *20*, 33–61.
91. Claerbout, J.F., and Muir, F. (1973). Robust modeling with erratic data. Geophysics *38*, 826–844.
92. Santosa, F., and Symes, W.W. (1986). Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Comput. *7*, 1307–1330.
93. Taylor, H.L., Banks, S.C., and McCoy, J.F. (1979). Deconvolution with the L1 norm. Geophysics *44*, 39–52.
94. Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure Appl. Math. *57*, 1413–1457.
95. Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. Ann. Appl. Stat. *2*, 302–332.
96. Hoggart, C.J., Whittaker, J.C., De Iorio, M., and Balding, D.J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet. *4*, e1000130.
97. van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C., Bethel, G., Holmes, G.K., et al. (2007). A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet. *39*, 827–829.
98. Falconer, D.S., and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics, Fourth Edition (London: Benjamin Cummings).
99. Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J.P., Russell, R.K., Sleiman, P.M., Imielinski, M., Glessner, J., Hou, C., et al. (2009). Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am. J. Hum. Genet. *84*, 399–405.
100. Elbers, C.C., van Eijk, K.R., Franke, L., Mulder, F., van der Schouw, Y.T., Wijmenga, C., and Onland-Moret, N.C. (2009). Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet. Epidemiol. *33*, 419–431.
101. Torkamani, A., Topol, E.J., and Schork, N.J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. Genomics *92*, 265–272.
102. Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Owen, M.J., O'Donovan, M.C., Craddock, N., and Wellcome Trust Case-Control Consortium. (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am. J. Hum. Genet. *85*, 13–24.
103. Yu, K., Li, Q., Bergen, A.W., Pfeiffer, R.M., Rosenberg, P.S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009). Pathway analysis by adaptive combination of P-values. Genet. Epidemiol. *33*, 700–709.
104. Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. Am. J. Hum. Genet. *81*, 1278–1283.
105. Lesnick, T.G., Papapetropoulos, S., Mash, D.C., Ffrench-Mullen, J., Shehadeh, L., de Andrade, M., Henley, J.R., Rocca, W.A., Ahlskog, J.E., and Maraganore, D.M. (2007). A genomic pathway approach to a complex disease: Axon guidance and Parkinson disease. PLoS Genet. *3*, e98.
106. Dinu, V., Miller, P.L., and Zhao, H. (2007). Evidence for association between multiple complement pathway genes and AMD. Genet. Epidemiol. *31*, 224–237.
107. Askland, K., Read, C., and Moore, J. (2009). Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum. Genet. *125*, 63–79.
108. Chen, L., Zhang, L., Zhao, Y., Xu, L., Shang, Y., Wang, Q., Li, W., Wang, H., and Li, X. (2009). Prioritizing risk pathways: A novel association approach to searching for disease pathways fusing SNPs and pathways. Bioinformatics *25*, 237–242.
109. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for

interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

110. Perry, J.R., McCarthy, M.I., Hattersley, A.T., Zeggini, E., Weedon, M.N., Frayling, T.M., and Wellcome Trust Case Control Consortium. (2009). Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Diabetes *58*, 1463–1467.
111. Baranzini, S.E., Galwey, N.W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B.M., Kappos, L., Polman, C.H., et al. GeneMSA Consortium. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum. Mol. Genet. *18*, 2078–2090.
112. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. The Gene Ontology Consortium (2000). Gene ontology: Tool for the unification of biology. Nat. Genet. *25*, 25–29.