Annals of
human genetics

doi: 10.1111/j.1469-1809.2010.00630.x

# Using Biological Knowledge to Uncover the Mystery in the Search for Epistasis in Genome-Wide Association Studies

Marylyn D. Ritchie*

*Departments of Molecular Physiology & Biophysics and Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232-0700, USA*

## Summary

The search for the missing heritability in genome-wide association studies (GWAS) has become an important focus for the human genetics community. One suspected location of these genetic effects is in gene–gene interactions, or epistasis. The computational burden of exploring gene–gene interactions in the wealth of data generated in GWAS, along with small to moderate sample sizes, have led to epistasis being an afterthought, rather than a primary focus of GWAS analyses. In this review, I discuss some potential approaches to filter a GWAS dataset to a smaller, more manageable dataset where searching for epistasis is considerably more feasible. I describe a number of alternative approaches, but primarily focus on the use of prior biological knowledge from databases in the public domain to guide the search for epistasis. The manner in which prior knowledge is incorporated into a GWA study can be many and these data can be extracted from a variety of database sources. I discuss a number of these approaches and propose that a comprehensive approach will likely be most fruitful for searching for epistasis in large-scale genomic studies of the current state-of-the-art and into the future.

Keywords: Epistasis, prior knowledge, pathways, protein–protein interactions, gene–gene interactions

## Introduction

The identification and characterization of susceptibility genes for common complex human disease is a difficult challenge. The usual approach of focusing a study on just one or a few candidate genes limits our ability to identify novel genetic effects associated with disease. In addition, many susceptibility genes may exhibit effects that are partially or solely dependent on interactions with other genes and/or the environment. Genome-wide association studies (GWAS) have been proposed as a solution to these problems; however, the analysis of GWAS data is problematic because we must separate the one or few true, but modest, signals from the extensive background noise. GWAS analyses must embrace abundant clinical and environmental data available to complement the rich genotypic data with the ultimate goal of revealing the genetic and environmental factors important for disease risk.

The ultimate goal of any disease gene discovery project is to identify all the genomic variations relevant to the phenotype being studied. As technology has advanced, the field has gone from very coarse genomic examination embodied in cytogenetic analyses, to higher resolution linkage analyses, and now to very high-resolution association analyses. Methodological advances in the analysis of large-scale or GWA studies and the ability to integrate results across experiments have simply not kept pace with this flood of genotyping data. It is a central fallacy that simply collecting enough data will solve the problem of elucidating disease susceptibility loci. Instead, it is this wealth of data that has made distinguishing true scientific discoveries from the thousands of false discoveries even more challenging. The growing disparity in developing data collection versus data analysis methods mandates a more concerted effort to develop the necessary analytical tools to successfully interpret the genotypic data and thus ultimately improve the prevention, diagnosis, and treatment of common disease. The ultimate utility of our monumental investment in data generation will depend largely on the development of innovative analytical strategies and study designs that allow for the detection of gene–gene and gene–environment interactions.

While GWAS have been extremely successful (Hindorff et al., 2009) there is clearly evidence that there is a large proportion of the genetic heritability for common, complex disease that has yet to be uncovered. Maher (2008) explains the possibilities for where the missing heritability may be hiding; one of which is "underground networks," where it is the

*Corresponding author: Marylyn D. Ritchie, PhD, Departments of Molecular Physiology & Biophysics and Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University, 519 Light Hall, Nashville, Tennessee 37232-0700. Tel: 615-343-5851; Fax: 615-343-8619; E-mail: ritchie@chgr.mc.vanderbilt.edu

interactions between genes in a more complex network that explains a larger proportion of the heritability (Maher, 2008). However, teasing gene–gene interaction networks apart from the many false positive loci in a GWA study is incredibly difficult. Cantor et al. (2010) review prioritization of GWAS results and identify epistasis and pathway analysis as two potential areas for deeper investigation in the quest for the missing heritability. This manuscript will focus on the combination of prior biological knowledge and pathway information for the detection of epistasis (gene–gene interactions) in GWAS data.

Epistasis was first described by Bateson (1909) as the effect of one gene masking (or literally *standing upon*) the effect of another. The Bateson view of epistasis has also been described as *biological epistasis* (Moore & Williams, 2005), where variation in the physical interaction of biomolecules affects a phenotype (Moore, 2003). From a statistical perspective, epistasis was also observed as multiallelic segregation patterns by Fisher (1918) who mathematically described the phenomenon as deviation from additivity in a linear model of genotypes. Statistical epistasis and biological epistasis eventually converge as scientific understanding progresses. For example, Bridges (1919) discovered statistical epistasis in *Drosophila* eye color. These alleles influence a common set of biochemical pathways controlling eye pigmentation that was elucidated many years later (Lloyd et al., 1998).

Most rare Mendelian genetic disorders, such as cystic fibrosis, are influenced by the effects of a single gene (although epistasis is being discovered in Mendelian disorders as described below). However, common diseases, such as multiple sclerosis (MS), breast cancer, or diabetes, are likely influenced by more than one gene, some of which may be associated with disease risk primarily through nonlinear interactions (Ritchie et al., 2001; Moore & Williams, 2002). The possibility of complex interactions makes the detection and characterization of genes associated with common, complex disease difficult. Templeton (2000) documents that gene–gene interactions are commonly found when properly investigated. Based on recent research, epistasis is not merely a theoretical argument. Epistasis has been identified as a component of complex phenotypes in a number of studies (Ming & Muenke, 2002). For example, Mendelian disorders such as retinitis pigmentosa (Kajiwara et al., 1994), Hirschsprung disease (Auricchio et al., 1999), juvenile-onset glaucoma (Vincent et al., 2002), familial amyloid polyneuropathy (Soares et al., 2005), and cystic fibrosis (Dipple & McCabe, 2000a,b) are documented examples of epistasis where modifier genes interact with Mendelian inherited main-effect genes. More compelling are studies in model organisms where there is both biological and statistical evidence for epistasis. Three arthritis loci have been identified in a quantitative trait locus (QTL) in mice that exhibit epistatic interactions (Johannesson et al., 2005a,b). Epistatic effects have also been documented in a number of other phenotypes in mice including obesity (Warden et al., 2004) and fluctuating asymmetry of tooth size and shape (Leamy et al., 2005). Similarly, other model organisms such as *Saccharomyces cerevisiae* have documented epistasis associated with quantitative traits such as metabolism (Segre et al., 2005). These model organism studies provide additional evidence that epistasis detected *via* statistical and computational techniques may be relevant biologically. This is something that is not possible to assess easily in human genetic studies (Moore & Williams, 2005).

As epistasis is believed to play an important role in the genesis of complex disease, analysis strategies for detecting epistasis in large-scale data are increasingly important. A major hurdle in discovering epistasis, however, is the variable selection problem. Exhaustively evaluating all two-marker models in whole-genome data is a computational and statistical challenge, as processing the 5.00e11 possible two-marker models from a set of 1 million SNPs requires extensive computing resources and produces a plethora of statistically significant results with limited biological interpretability. That said, there are analytic strategies that have been developed or adapted specifically for this purpose. Some have argued that epistasis is unlikely to contribute in a significant way to the missing heritability, because in their opinion, epistatic effects may explain even less of the heritability than the independent single locus effects. They also comment that there are few, if any, convincing replicated gene–gene interaction models. However, other researchers present the contradictory point of view and provide evidence that epistasis is likely to exist and may have even larger effect sizes than the main-effect counterparts (Eichler et al., 2010). One of the challenges in identifying convincing epistasis models from large-scale genomic studies is the difficultly in replication. How should "replication" be defined for epistasis models? Some argue in support of a conservative Bonferroni corrected *p*-value cut-off for an association describing the same SNP, with the same direction of effect, in the same race/ethnicity; the other extreme is to allow for replication of a particular pathway or genes to satisfy the replication requirement.

## Analytic Approaches

Many groups are considering approaches to epistasis in GWAS. The most straightforward approach is an exhaustive search through all of the combinations of genes using some analytic approach. However, in GWAS, this is an intractable computation issue. For example, if one considers 500, 000 SNPs, that leads to 500, 000 single-locus tests, $1 \times 10^{11}$ two-locus tests, $2 \times 10^{16}$ three-locus tests, and so on (calculated based on $\binom{n}{m}$), where $n$ is the number of SNPs and $m$ is the number of variables in the model). It has been shown that using the parallel multifactor dimensionality reduction approach (pMDR), it is possible to scan through an exhaustive search of possible two-locus combinations in a 500K GWAS dataset (Bush et al., 2006). Steffens et al. (2010) also demonstrate a genome-wide interaction analysis (GWIA) and the strategies for data compression, specific data representations, interleaved data organization, and parallelization of the analysis on a multiprocessor system. These strategies provide capability to perform an exhaustive pair wise GWIA. In fact, even standard logistic regression can be implemented in such a way that two-locus models can be evaluated at a genome-wide level (Evans et al., 2006). However, beyond pair wise interactions, exhaustive searching is not practical.

To deal with the challenge of detecting interactions, much research is underway for improved statistical and computational methodologies. Many researchers are exploring variations and modifications of logistic regression such as logic regression (Kooperberg et al., 2001), penalized logistic regression (Zhu & Hastie, 2004), classification/regression trees (CART) and multivariate adaptive regression splines (MARS) (Cook et al., 2004). Additional studies are being conducted in data mining and machine learning research, including data reduction and pattern recognition approaches. Data reduction involves a collapsing or mapping of the data to a lower dimensional space. Examples of data reduction approaches include the combinatorial partitioning method (CPM) (Nelson et al., 2001), restricted partition method (RPM) (Culverhouse et al., 2004), set association (Wille et al., 2003), and multifactor dimensionality reduction (MDR) (Ritchie et al., 2001; Hahn et al., 2003; Ritchie et al., 2003). Pattern recognition on the other hand, involves extracting patterns from the data to discriminate between groups using the full dimensionality of the data. Examples of pattern recognition methods include cluster analysis (Kaufman & Rousseeuw 1990), support vector machines (SVM) (Cristianini & Shawe-Taylor, 2000), self-organizing maps (SOM) (Hastie et al., 2001), and neural networks (NN) (Ripley, 1996; Motsinger-Reif et al., 2008a,b,c).

## Using Filters to Magnify the Epistatic Genes

It is clear that while the goal of GWAS is to survey the entire genome in an unbiased way, this type of approach simply cannot work for the search for epistasis, especially beyond pair wise interactions. Because of the sheer magnitude of possible combinations of SNPs, alternative strategies to reduce the search space and address the variable selection problem are warranted. A number of such filtering approaches have been suggested. First, using statistical evidence of single-locus effects to prioritize SNPs can be promising. Filtering SNPs based on the strength of independent main effects, evaluating interactions only between SNPs that meet a certain effect size threshold, can certainly identify SNP combinations among loci with small to moderate main effects, such as two 2-SNP models identified for Amyotrophic lateral sclerosis (ALS) (Sha et al., 2009). The second approach is to use intrinsic knowledge extracted from the dataset to filter the list of SNPs to test for interactions. Third, the use of extrinsic biological knowledge to filter SNPs and then evaluate multimarker combinations based on biological criteria (Carlson et al., 2004) has been suggested. Each of these strategies imposes a specific bias into the analysis, and no one strategy will be optimal in all cases. Each of these has its own advantages and disadvantages with known biases and limitations. The next section will discuss each of these filtering approaches briefly.

### Filter: Statistical Evidence for Single-Locus Effects

The search for interactions among SNPs (or genes) that exhibit statistically significant main effects has been proposed as a possible strategy to filter a GWAS set down to a more manageable dataset for exploring epistasis (Evans et al., 2006). This approach follows from the hierarchical model building principles of the general linear model whereby interaction terms are tested only after all main-effect terms are deemed statistically significant (as some predefined *p*-value threshold). As shown in simulation, this approach will certainly work well assuming that the genes involved in the epistasis model *do exhibit* statistically detectable main effects that allow them to exceed the significance threshold established for the study. For example, in a 500K GWAS analysis, one might use a threshold of $p < 10^{-5}$ based on an Armitage trend test. As such, it is expected that there would be approximately five SNPs significant by chance alone; presumably additional SNPs will be significant because some of those will be true effects for that particular dataset. If the SNPs that are important for the epistatic model are not among those top hits, the interactions will not be tested. If we select or filter variables based on their main effects, we bias the analysis using statistical information, and assume that relevant interactions occur only between markers that independently have some effect on the phenotype alone. Several studies have proposed complex theoretical penetrance models that influence the trait only through the interaction of two or more genetic variants (Frankel & Schork, 1996; Culverhouse et al., 2002; Moore et al., 2004), and filtering based on main effects would potentially miss these types of discoveries.

### Filter: Intrinsic Knowledge

Filter algorithms that explicitly assess the quality of a SNP in its relationship to the clinical outcome are an alternative to statistical or biological filters. A series of Relief algorithms have been explored including Relief, ReliefF, Spatially Uniform ReliefF (SURF), Tuned ReliefF (TuRF), and SURF and TURF (Greene et al., 2009). These approaches use a nearest neighbor approach to assess SNP quality to detect attributes associated with disease through interactions or independent main effects without providing a specific model for the effect. In this case, nearest neighbors are individuals in the dataset who are genetically similar at the most SNP locations throughout the genome. Relief uses a single neighbor, ReliefF uses multiple nearest neighbors, and the SURF and TURF are various extensions to the ReliefF filtering. Filtering approaches that use intrinsic properties of the data, such as these ReliefF methods, look like a promising alternative for epistasis in GWAS. According to published studies, they will be successful in removing nonfunctional SNPs while maintaining the SNP–SNP interaction models. This will effectively reduce the number of statistical tests that need to be performed, which relieves the computational complexity issues as well as the multiple comparisons issues.

### Filter: Extrinsic Biological Knowledge

If we filter variables using biological information extrinsic to our dataset—that is, only examine interactions between SNPs in a common pathway or with a common structure or function—we bias the analysis in favor of models with an established biological foundation in the literature, and novel interactions between SNPs would be missed. Furthermore, the entire analysis is conditional upon the quality of the biological information used. However, the interaction models with detectable statistical epistasis will have good evidence for biological epistasis and a high likelihood of being interpretable.

While all of the proposed approaches for filtering have clear strengths and limitations, I propose that filtering based on extrinsic biological knowledge will demonstrate to be a robust approach for the detection of epistasis in large-scale genomic analyses including GWAS as well as next-generation sequencing. While the available biological knowledge is incomplete and always evolving, it provides a framework for exploring epistasis in which models are plausible, more likely to be interpretable, and reduces the computational and statistical burden. By limiting the search space, we limit the number of statistical tests in addition to the computational complexity. The remainder of this review will focus on approaches being developed for using biological knowledge to prioritize the search for the missing heritability in the epistasis domain.

## Prior Biological Knowledge

The incorporation of prior knowledge into GWA studies has been proposed by many research groups. Several new tools have recently been developed to incorporate biological information with analytical approaches for GWAS data. Prioritizer is a Bayesian approach to incorporate multiple sources of gene inter-relationships in a global "functional gene network." This network is used to prioritize significant single-SNP results by gene function (Franke et al., 2006). Other methods use structured knowledge as a way to guide (but not restrict) variable selection for regression-based modelling. Province & Borecki (2008) propose a Bayesian resampling approach to select collections of SNPs that may have very small independent effects but function in aggregate to explain a more substantial portion of trait variance. Conti proposes a hierarchical modelling approach that uses expert knowledge ontology to search and test complex multi-SNP models. This Bayesian modelling process is flexible, allowing SNPs outside the knowledge base to be also used in models (Thomas et al., 2009).

It has been shown in a GWA study of Parkinson disease that the use of biological information to filter statistical results can be beneficial (Maraganore et al., 2005). While the nature of GWAS is inherently unbiased with respect to genes, hence the genome-wide approach, still many research groups use knowledge about genes and pathways to annotate the genome-wide associated variants. Annotating results is certainly one way that biological knowledge can be used in a GWAS design. Subsequently, prioritizing SNPs to explore in an epistasis analysis in a GWAS dataset has been proposed by many investigators.

As described earlier, it is computationally intractable to search through all possible combinations of variants in a GWAS (or sequence-based) dataset. Therefore, prioritizing variants based on some type of filter is essential. While statistical, data-driven, or knowledge-based filters could be implemented, a biological filter has inherent strengths for epistasis analysis. First, there is a tremendous amount of existing knowledge out there. Fortunately, many sophisticated database systems have been developed to store and annotate the wealth of biological knowledge in the domain. Because so much effort has been spent over the years building our extensive knowledge base of biological pathways, networks, and systems, it is conceivable that this information can mutually benefit GWAS analysis for epistasis.

Second, because the use of biological knowledge typically constructs models based on biology, the models with statistical evidence for effects are more likely to make sense. In particular for epistasis models, interpreting statistical models of interactions is extremely difficult. However, when we explicitly test models of genes that participate in the same biological pathway or network, the interpretations are much more straightforward. Furthermore, once statistically significant models are detected, developing the next steps in terms of model organisms or cell line experiments, that is, going "back to the bench," is much easier.

As with any filtering technique, there are of course disadvantages. Of primary importance is the reality that when we rely on the public domain for building the knowledge base, we are restricted to the information in the scientific domain. With that, there is inherent literature bias, such that the only knowledge that can be used for filtering is published knowledge. Often, negative findings are not published. Similarly, the literature is flooded with false positive discoveries. Little can be done to modify these biases; however, being aware of the bias is critical. With respect to false positives, it is also conceivable that by using a biased approach, there may be additional likelihood of false positives—although probably no more so than an unbiased approach. The current state of the art is relying on replication that will continue to be important. Although, what is defined as replication for epistasis is not yet fully determined and supported by the community. Next, if we limit an analysis to gene combinations prioritized based on biology, we may be prohibited from learning novel biology. However, as discussed later in this review, some groups have made some modifications to their approach that allow for some novel discoveries. Additionally, it is usually the case that these approaches all still allow for discoveries that are new relationships of particular genes or pathways to a particular disease.

## Methods for Biological Filtering

As described, there are a number of approaches by which biological knowledge can be used in GWAS analysis. Torkamani and Schork (2009) and Pedroso (2010) describe/review some of these approaches. While most of them have been utilized and published based on a single-locus test of association, nearly all of them could be used in the search for epistasis. This is certainly a rapidly growing area of research; as such it is not possible to thoroughly describe all of the recent developments. However, in the following sections, a number of approaches will be described and suggested for how they could guide the search for the missing heritability.

### Pathway Approaches

The use of pathway data to look for over-representation of genome-wide associated hits has been done in many studies (Saccone et al., 2008; Wilke et al., 2008; Adeyemo et al., 2009; Eleftherohorinou et al., 2009; Liu et al., 2010; Zhang et al., 2010). For example, Perry et al. (2009) used Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCarta, and Gene Ontology (GO) to perform a modified gene-set enrichment analysis (GSEA) for type 2 diabetes (T2D) (Perry et al., 2009). In rheumatoid arthritis (RA), Beyene et al. (2009) utilize a selection of prior knowledge from c2 curated gene sets, which are obtained from online pathway databases, citations in PubMed, and domain experts. Their final set for GSEA also included 1900 gene sets collected from canonical pathways, chemical and genetic perturbations, BioCarta pathways, GeneMAPP, and KEGG (Beyene et al., 2009). They identified some known pathways (*HLA*) and unknown pathways (related to *CTLA4*) (Beyene et al., 2009). Similarly, De la Cruz et al. (2010) used pathway and conserved region data to identify novel pathways for Crohn's disease in the Wellcome Trust Case Control Consortium (WTCCC) GWAS. A similar GSEA , the SNP-ratio test (SRT) (O'Dushlaine et al., 2009), compares the proportion of statistically significant genes to all SNPs within genes that are part of a pathway. An empirical *p*-value is then calculated based on comparisons to ratios in datasets where a permutation test has been performed (i.e., the assignment of case/control status has been randomized). Another similar extension of GSEA, Gen-Gen (Wang et al., 2007), developed by Wang et al. in 2007 has been demonstrated useful in traits such as Alzheimer's disease (Lambert et al., 2010).

Baranzini et al. (2009) propose a protein interaction and network-based analysis (PINBPA) for the study of a MS dataset. An alternative approach was explored by Askland et al., whereby they used exploratory visual analysis (EVA) to perform a number of pathway-based analyses of bipolar disorder (Askland et al., 2009). Similarly, pathway genetic load (PGL) looks for evidence of epistasis between genes confined to a single pathway (Huebinger et al., 2010). This approach dramatically reduces the computational complexity of an epistasis search in GWAS data.

Pathways have also been used in the context of GSEA after SNP associations with eQTLs from gene expression studies (Zhong et al., 2010). Here, Zhong et al. identified 16 pathways enriched for genes corresponding to eSNPs that also show evidence of association with T2D in the WTCCC T2D GWAS and replicated in the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) study. Many of the pathways identified are proposed candidate pathways for T2D, including the calcium signalling pathway, the PPAR signalling pathway, and TGF-b signalling. Others are novel pathways for T2D including the tight junction, complement and coagulation pathway, and antigen processing and presentation pathway (Zhong et al., 2010). eQTL information can be used to filter in other ways as well (described further in "Function-Based GWAS Analysis").

Another software package developed for pathway analysis is PATH. PATH incorporates information from nine online bioinformatics resources including the National Center for Biotechnology Information (NCBI), Online Mendelian Inheritance in Man (OMIM), KEGG, University of

California, Santa Cruz, Genome Browser, Seattle SNPs, PharmGKB, Genetic Association Database, the Single Nucleotide Polymorphism database (dbSNP), and the Innate Immune Database (IIDB) (Zamar et al., 2009). These data can then be used to filter SNPs and determine which SNP–SNP interactions to test.

Assessing the significance of pathways is also an important challenge. It is not simply enough to look for an over-representation of hits in a particular pathway or set of pathways. There are reasons unrelated to the associations that can lead to this such as the selection of SNPs on the genotyping platforms or the choice of pathway annotation for analysis. Large pathways have a higher chance of being statistically significant and many of the bioinformatics tools used for these types of studies are biased toward detecting well-defined pathways (Elbers et al., 2009). Methods to perform permutation testing in a pathway analysis framework have been developed to provide increased power and efficiency (Guo et al., 2009). Other approaches index pathways using GO terms and test for over-representation of pathways in a list of hits from a genome-wide association study (ALIGATOR) and successfully identify pathways for complex traits, such as bipolar disorder (Holmans et al., 2009).

It is also important to reiterate that pathway analysis approaches, in themselves, were not developed for the purpose of detecting epistasis. These methods focus on single-SNP analyses and explore pathways where an accumulation of single-locus associations is detected. However, it is obvious that these pathway approaches will develop hypotheses regarding potential "underground networks" that would be particularly interesting to focus efforts for detection of epistasis.

## Protein–Protein Interaction Approaches

While a wealth of information lies in pathway databases and this is a rich source of data for building biologically plausible models of epistasis to test in GWAS, another source of relevant information includes protein–protein interaction databases. Protein–protein interactions can be measured using mass spectrometry, yeast two-hybrids, immunoprecipitation, and affinity pull down followed by mass spectrometry (Pattin & Moore, 2008). As reviewed by Pattin and Moore, a number of protein–protein interaction databases are publicly available including the database of interacting proteins (DIP) (Salwinski et al., 2004), BioGRID interaction database (Breitkreutz et al., 2008), and human protein reference database (HPRD) (Mishra et al., 2006) to name a few (Pattin & Moore, 2008). Pattin suggests a couple of approaches where protein–protein interaction data could be used. First, the most straightforward approach includes filtering the full SNP list by the SNPs included in the genes encoding the proteins involved in the interactions (Pattin & Moore, 2008). This would certainly limit the number of SNPs explored for epistasis. However, it would also prevent the identification of models including novel biology. An alternative and perhaps more promising approach involves developing a metric to score the relative importance of the SNPs such that the full list could be prioritized or weighted. This would allow for novel biology to be discovered, although favoring models with *a priori* evidence of support (Pattin & Moore, 2008).

## Function-Based GWAS Analysis

Another group has proposed the use of genome annotation based on function to improve the ability to detect the biologically relevant signals in GWAS (Nicolae et al., 2010). Through the annotation of GWAS SNPs with information on gene expression, we may improve our ability to sift through GWAS hits to determine which are most likely to replicate in subsequent studies. In addition, information about gene expression may provide guidance in understanding the mechanisms that explain the associations discovered (Nicolae et al., 2010). Gamazon et al. describe the SNP and copy number (SCAN) database, which provides information on physical position, gene function, multilocus linkage disequilibrium, and expression quantitative trait loci (eQTL) (Gamazon et al., 2010). This database has been used to annotate single locus results from GWAS studies to look for single locus hits that occur in eQTLs. For some complex disorders, there are examples where the variants with evidence of true association with disease are most likely to be in eQTLs (Nicolae et al., 2010). A similar approach to this could be implemented to search for interactions. Rather than annotating the results using SCAN *post hoc*, perhaps the GWAS platform SNP list could be fed to SCAN to generate a list of SNPs that are in eQTLs, with particular function, and those could be the subset where interactions are evaluated. To our knowledge, this has not yet been done using the SCAN database, but it is likely that this will be performed in the future.

## Prior Association, Linkage, Expression, and Genomic Convergence Approaches

Genomic convergence has been proposed by many to guide the interpretation and prioritization of GWAS hits. Kitsios and Zintzaras (2009) implement a thorough interrogation of GWAS and genome-wide linkage studies for 19 phenotypes to look for evidence of consistent results. Convergence that was higher than expected by chance was identified for only two of 19 phenotypes. This could be due to the reality that

GWAS and Genome-wide linkage study (GWLS) are asking different questions and the study designs are not equally well suited to answer some questions related to complex traits. While convergence is suggestive evidence of genuine effects, lack of convergence is more difficult to interpret (Kitsios & Zintzaras, 2009).

## Comprehensive Approach

Perhaps the most lucrative solution involves a comprehensive knowledge-based approach that includes evidence from pathways, protein–protein interactions, prior association, linkage, or expression, etc. Because we have very few success stories of true, replicating epistatic models in humans, it is currently a challenge to hypothesize what the types of models will involve and what is the likely relationship between the genes that can be expected. We can look to the known examples of epistasis in model organisms to point us in the right direction, but until we have more examples in model organisms, we cannot know what to expect.

One of the major disadvantages of the comprehensive approach is the current inability to accurately evaluate it compared to other approaches in simulated data experiments. Unfortunately, simulation studies where biological knowledge is concerned are very difficult to perform. There are really two issues. First, if you do the straightforward type of simulation study, where you preselect functional SNPs based on biological knowledge and then embed them into the simulation, and next use that same knowledge to guide the search, the simulation is overly simplistic and really not very interesting. The second issue is that to do it right, we need to have a simulation tool whereby we can simulate pathways and networks, and then create disease models including some of the loci from these networks. This type of tool does not currently exist. So, unfortunately, while a simulation study to compare approaches would be fantastic, it is not currently feasible. Once a body of literature is published demonstrating some of the pathway and network effects, we can expect to observe in natural, biological data, we will be able to develop simulation tools to test additional novel analytic methods. After that, we may have a better detailed critique on the different approaches.

Several approaches have been developed that include a more thorough extraction of prior information from multiple sources. The Biofilter is one such system (Bush et al., 2009b). Layers of biological machinery exist between genetic variations and the phenotypes they manifest, and imposing this extra dimension of known biological information into statistical analysis may help identify relationships between genetic variants that contribute to common complex disease. The Biofilter is a database system cataloging biological information based on data from the Reactome, KEGG, GO, DIP, PFAM, Ensembl, and NetPath (Bush et al., 2009b). The strategy of Biofilter steps beyond the annotation and grouping of independent SNP effects. The Biofilter uses biological information about gene–gene relationships and gene–disease relationships to construct multi-SNP models before conducting any statistical analysis. Rather than annotating the independent effect of each SNP in a GWAS dataset, the Biofilter allows the explicit detection and modelling of interactions between a set of SNPs. In this manner, the Biofilter process provides a tool to discover significant multi-SNP models with nonsignificant main effects that have established biological plausibility. This approach has the added benefit of reducing both the computational and statistical burden of exhaustively evaluating all possible multi-SNP models. The goal of the Biofilter is to take advantage of what we know, recognizing that do not know it all (Bush et al., 2009b).

The Biofilter model generation process is gene centric, and as such, SNPs from GWAS genotyping platforms must first be assigned to genes (Bush et al., 2009a). Relationships between the genes represented by a platform can then be translated to multi-SNP models. Structured biological knowledge relevant to GWAS interaction analysis can come from various sources. We have partitioned relevant knowledge into two basic types: disease-dependent and disease-independent. *Disease-dependent* knowledge is information that relates a gene to the disease phenotype being studied, such as a previously associated SNP or a gene that is overexpressed in cases. *Disease-independent* knowledge is information that relates genes to one another, or defines collections of genes, such as a metabolic pathway or a common structural motif. These two types of information can be combined to form different classes of multi-SNP models and provide a measure of how strongly implicated a given model is based on the current available knowledge. Once these multi-SNP models are constructed, they can be evaluated using any one of the earlier mentioned analytic methods such as logistic regression, MDR, NN, etc.

Another approach for a comprehensive analysis is INTERSNP. INTERSNP is a powerful, flexible approach that implements logistic regression or log-linear models for joint analysis of multiple SNPs (Herold et al., 2009). The filtering of SNPs can be done using statistical evidence from single locus statistics, genomic evidence based on genomic location, or biologic relevance based on pathway information from KEGG (Herold et al., 2009). Approaches such as these have the greatest potential since they rely on multiple sources and types of information. This is, of course, as long as the analytic strategy is implemented in such a way that the incorporation of incorrect knowledge does not impede the ability to detect the correct models. Using prior knowledge can be an incredibly powerful tool, but we should be careful to do so in an efficient manner.

## Conclusions

The quest for the missing heritability continues with some investigators searching in the structural and architectural variation in the genome, others in the environment, and still others in "underground networks" (Maher, 2008). Because of the computational and interpretive challenges brought about by epistasis, researchers are likely to rely on a filtering approach to reduce the enormous search space for gene–gene interaction effects. If the data generated in model organisms are any indication, we can expect that in human genetic studies, gene–gene interaction effects are likely to be common, exhibit moderate to large effects, and possibly occur between genes with minimal statistically detectable main effects. If this is the case, filtering based on single locus statistics could be a detrimental path to pursue. Data driven approaches, such a ReliefF, offer a nice alternative. However, they do not aid in the interpretation of the gene–gene interaction models discovered. On the contrary, filtering based on biological knowledge overcomes this limitation. And this is the case regardless of which types of biological sources are explored including pathways, networks, protein–protein interactions, previous association, linkage, gene expression, or functional data. It is important, however, to keep in mind the major limitation of biological-based filtering approaches. We can only incorporate biological data that are known and published in the public domain. This may limit the potential to learn completely novel biology. Additionally, the gene set explored may be biased due to the expected bias we know exists in the literature. Biological knowledge is incomplete and under ongoing revision, which indicates that it is not free from error. As we continue to expand our breadth of biological knowledge, the ability to use this information to guide statistical analysis will improve. Even with these limitations, it is still highly likely that approaches such as the Biofilter will be extremely useful in the search for epistasis for years to come.

## Future Work

As the molecular biology technology continues to evolve and grow, analytic approaches will likewise continue to mature. There are a number of areas with respect to the hunt for epistasis where significant growth is likely to occur. In terms of using biological knowledge to guide the search for gene–gene interactions, with time we will create additional approaches that incorporate increasing amounts of knowledge and extract higher quality knowledge, thus improving the quality of our treasure map. Additionally, we will learn better ways to use this information to guide the search through the vast map of epistasis models. A second area where we can expect continued development is the exploration of additional analytic approaches. Novel methods continue to emerge (Wongseree et al., 2009; Greene et al., 2010; Steffens et al., 2010). Finally, as gene sequencing continues to become cost effective, for gene region, whole exome, and eventually whole-genome sequencing, methods for looking for epistasis in these data will become critical. It is highly likely that the use of biological knowledge will be necessary to make sense of the rare variation in the genome. It is conceivable that collapsing multivariate methods could be implemented (Li & Leal, 2008); however, it is possible that this collapsing could be performed at many levels including gene region, exon or intron, gene, gene set, pathway, network, etc. In addition, it is conceivable that much of the missing heritability can be explained by some combination of common and rare variation. In this case, we will rely on biological pathways and look for combinations of rare and common variants in particular pathways that are associated or over-represented in case groups as compared to control groups. Success can be expected as the knowledge surrounding biological pathways and statistical analytics incorporating this knowledge simultaneously mature and converge.

## References

Adeyemo, A., Gerry, N., Chen, G., Herbert, A., Doumatey, A., Huang, H., Zhou, J., Lashley, K., Chen, Y., Christman, M. & Rotimi, C. (2009) A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genet* **5**, e1000564.

Askland, K., Read, C. & 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.

Auricchio, A., Griseri, P., Carpentieri, M. L., Betsos, N., Staiano, A., Tozzi, A., Priolo, M., Thompson, H., Bocciardi, R., Romeo, G., Ballabio, A. & Ceccherini, I. (1999) Double heterozygosity for a RET substitution interfering with splicing and an EDNRB missense mutation in Hirschsprung disease. *Am J Hum Genet* **64**, 1216–1221.

Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B. M., Kappos, L., Polman, C. H., Matthews, P. M., Hauser, S. L., Gibson, R. A., Oksenberg, J. R. & Barnes, M. R. (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet* **18**, 2078–2090.

Bateson, W. (1909) *Mendel's principles of heredity*. Cambridge: Cambridge University Press.

Beyene, J., Hu, P., Hamid, J. S., Parkhomenko, E., Paterson, A. D. & Tritchler, D. (2009) Pathway-based analysis of a genome-wide case-control association study of rheumatoid arthritis. *BMC Proc* **3**(Suppl 7), S128.

Breitkreutz, B. J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bahler, J., Wood, V., Dolinski, K. & Tyers, M. (2008) The BioGRID interaction database: 2008 update. *Nucleic Acids Res* **36**, D637–D640.

Bridges, C. B. (1919) Specific modifiers of eosin eye color in *Drosophila Melanogaster*. *J Exp Zool* **28**, 337–384.

Bush, W. S., Chen, G., Torstenson, E. S. & Ritchie, M. D. (2009a) LD-spline: Mapping SNPs on genotyping platforms to genomic regions using patterns of linkage disequilibrium. *BioData Min* **2**, 7.

Bush, W. S., Dudek, S. M. & Ritchie, M. D. (2006) Parallel multifactor dimensionality reduction: A tool for the large-scale analysis of gene–gene interactions. *Bioinformatics* **22**, 2173–2174.

Bush, W. S., Dudek, S. M. & Ritchie, M. D. (2009b) Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput* **2009**, 368–379.

Cantor, R. M., Lange, K. & Sinsheimer, J. S. (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am J Hum Genet* **86**, 6–22.

Carlson, C. S., Eberle, M. A., Kruglyak, L. & Nickerson, D. A. (2004) Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446–452.

Cook, N. R., Zee, R. Y. & Ridker, P. M. (2004) Tree and spline based association analysis of gene–gene interaction models for ischemic stroke. *Stat Med* **23**, 1439–1453.

Cristianini, N. & Shawe-Taylor, J. (2000) *An introduction to support vector machines.* Cambridge: Cambridge University Press.

Culverhouse, R., Klein, T. & Shannon, W. (2004) Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* **27**, 141–152.

Culverhouse, R., Suarez, B. K., Lin, J. & Reich, T. (2002) A perspective on epistasis: Limits of models displaying no main effect. *Am J Hum Genet* **70**, 461–471.

De la Cruz, O., Wen, X., Ke, B., Song, M. & Nicolae, D. L. (2010) Gene, region and pathway level analyses in whole-genome studies. *Genet Epidemiol* **34**, 222–231.

Dipple, K. M. & McCabe, E. R. (2000a) Modifier genes convert "simple" Mendelian disorders to complex traits. *Mol Genet Metab* **71**, 43–50.

Dipple, K. M. & McCabe, E. R. (2000b) Phenotypes of patients with "simple" Mendelian disorders are complex traits: Thresholds, modifiers, and systems dynamics. *Am J Hum Genet* **66**, 1729–1735.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. & Nadeau, J. H. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446–450.

Elbers, C. C., van Eijk, K. R., Franke, L., Mulder, F., Van Der Schouw, Y. T., Wijmenga, C. & Onland-Moret, N. C. (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet Epidemiol* **33**, 419–431.

Eleftherohorinou, H., Wright, V., Hoggart, C., Hartikainen, A. L., Jarvelin, M. R., Balding, D., Coin, L. & Levin, M. (2009) Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS One* **4**, e8068.

Evans, D. M., Marchini, J., Morris, A. P. & Cardon, L. R. (2006) Two-stage two-locus models in genome-wide association. *PLoS Genet* **2**, e157.

Fisher, R. A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* **52**, 399–433.

Franke, L., van, B. H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M. & Wijmenga, C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* **78**, 1011–1025.

Frankel, W. N. & Schork, N. J. (1996) Who's afraid of epistasis? *Nat Genet* **14**, 371–373.

Gamazon, E. R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E. O., Nicolae, D. L., Dolan, M. E. & Cox, N. J. (2010) SCAN: SNP and copy number annotation. *Bioinformatics* **26**, 259–262.

Greene, C. S., Penrod, N. M., Kiralis, J. & Moore, J. H. (2009) Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Min* **2**, 5.

Greene, C. S., Himmelstein, D. S., Nelson, H. H., Kelsey, K. T., Williams, S. M., Andrew, A. S., Karagas, M. R. & Moore, J. H. (2010) Enabling personal genomics with an explicit test of epistasis. *Pac Symp Biocomput* **2010**, 327–336.

Guo, Y. F., Li, J., Chen, Y., Zhang, L. S. & Deng, H. W. (2009) A new permutation strategy of pathway-based approach for genome-wide association study. *BMC Bioinform* **10**, 429.

Hahn, L. W., Ritchie, M. D. & Moore, J. H. (2003) Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics* **19**, 376–382.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag.

Herold, C., Steffens, M., Brockschmidt, F. F., Baur, M. P. & Becker, T. (2009) INTERSNP: Genome-wide interaction analysis guided by a priori information. *Bioinformatics* **25**, 3275–3281.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. & Manolio, T. A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362–9367.

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. (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet* **85**, 13–24.

Huebinger, R. M., Garner, H. R. & Barber, R. C. (2010) Pathway genetic load allows simultaneous evaluation of multiple genetic associations. *Burns* **36**, 787–792.

Johannesson, M., Karlsson, J., Wernhoff, P., Nandakumar, K. S., Lindqvist, A. K., Olsson, L., Cook, A. D., Andersson, A. & Holmdahl, R. (2005a) Identification of epistasis through a partial advanced intercross reveals three arthritis loci within the Cia5 QTL in mice. *Genes Immun* **6**, 175–185.

Johannesson, M., Olsson, L. M., Lindqvist, A. K., Moller, S., Koczan, D., Wester-Rosenlof, L., Thiesen, H. J., Ibrahim, S. & Holmdahl, R. (2005b) Gene expression profiling of arthritis using a QTL chip reveals a complex gene regulation of the Cia5 region in mice. *Genes Immun* **6**, 575–583.

Kajiwara, K., Berson, E. L. & Dryja, T. P. (1994) Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science* **264**, 1604–1608.

Kaufman, L. & Rousseeuw, P. J. (1990) *Finding groups in data: An introduction to cluster analysis*. New York: Wiley-Interscience Publication.

Kitsios, G. D. & Zintzaras, E. (2009) Genomic convergence of genome-wide investigations for complex traits. *Ann Hum Genet* **73**, 514–519.

Kooperberg, C., Ruczinski, I., LeBlanc, M. L. & Hsu L (2001) Sequence analysis using logic regression. *Genet Epidemiol* **21**(Suppl 1), S626–S631.

Lambert, J. C., Grenier-Boley, B., Chouraki, V., Heath, S., Zelenika, D., Fievet, N., Hannequin, D., Pasquier, F., Hanon, O., Brice, A., Epelbaum, J., Berr, C., Dartigues, J. F., Tzourio, C., Campion, D., Lathrop, M. & Amouyel, P. (2010) Implication of the immune system in alzheimer's disease: evidence from genome-wide pathway analysis. *J Alzheimers Dis* **20**, 1107–1118.

Leamy, L. J., Workman, M. S., Routman, E. J. & Cheverud, J. M. (2005) An epistatic genetic basis for fluctuating asymmetry of tooth size and shape in mice. *Heredity* **94**, 316–325.

Li, B. & Leal, S. M. (2008) Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet* **83**, 311–321.

Liu, Y. J., Guo, Y. F., Zhang, L. S., Pei, Y. F., Yu, N., Yu, P., Papasian, C. J. & Deng, H. W. (2010) Biological pathway-based genome-wide association analysis identified the vasoactive intestinal peptide (VIP) pathway important for obesity. *Obesity (Silver Spring)* **18**, 2339–2346.

Lloyd, V., Ramaswami, M. & Kramer, H. (1998) Not just pretty eyes: Drosophila eye-colour mutations and lysosomal delivery. *Trends Cell Biol* **8**, 257–259.

Maher, B. (2008) Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21.

Maraganore, D. M., de, A. M., Lesnick, T. G., Strain, K. J., Farrer, M. J., Rocca, W. A., Pant, P. V., Frazer, K. A., Cox, D. R. & Ballinger, D. G. (2005) High-resolution whole-genome association study of parkinson disease. *Am J Hum Genet* **77**, 685–693.

Ming, J. E. & Muenke, M. (2002) Multiple hits during early embryonic development: Digenic diseases and holoprosencephaly. *Am J Hum Genet* **71**, 1017–1032.

Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K. S., Sharma, S., Chandrika, K. N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H. G., Nagini, M., Kumar, G. S., Jose, R., Deepthi, P., Mohan, S. S., Gandhi, T. K., Harsha, H. C., Deshpande, K. S., Sarker, M., Prasad, T. S. & Pandey, A. (2006) Human protein reference database—2006 update. *Nucleic Acids Res* **34**, D411–D414.

Moore, J. H. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* **56**, 73–82.

Moore, J., Hahn, L., Ritchie, M., Thornton, T. & White, B. (2004) Routine discovery of complex genetic models using genetic algorithms. *Appl Soft Comput* **4**, 79–86.

Moore, J. H. & Williams, S. M. (2002) New strategies for identifying gene–gene interactions in hypertension. *Ann Med* **34**, 88–95.

Moore, J. H. & Williams, S. M. (2005) Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. *Bioessays* **27**, 637–646.

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W. & Ritchie, M. D. (2008a) Comparison of approaches for machine-learning optimization of neural networks for detecting gene–gene interactions in genetic epidemiology. *Genet Epidemiol* **32**, 325–340.

Motsinger-Reif, A. A., Fanelli, T. J., Davis, A. C. & Ritchie, M. D. (2008b) Power of grammatical evolution neural networks to detect gene–gene interactions in the presence of error. *BMC Res Notes* **1**, 65.

Motsinger-Reif, A. A., Reif, D. M., Fanelli, T. J. & Ritchie, M. D. (2008c) A comparison of analytical methods for genetic association studies. *Genet Epidemiol* **32**, 767–778.

Nelson, M., Kardia, S., Ferrell, R. & Sing, C. (2001) A combinatorial partitioning approach to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* **11**, 458–470.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E. & Cox, N. J. (2010) Trait-associated SNPs are more likely to be EQTLs: Annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888.

O'Dushlaine, C., Kenny, E., Heron, E. A., Segurado, R., Gill, M., Morris, D. W. & Corvin, A. (2009) The SNP ratio test: Pathway analysis of genome-wide association datasets. *Bioinformatics* **25**, 2762–2763.

Pattin, K. A. & 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.

Pedroso, I. (2010) Gaining a pathway insight into genetic association data. *Methods Mol Biol* **628**, 373–382.

Perry, J. R., McCarthy, M. I., Hattersley, A. T., Zeggini E, Weedon, M. N. & Frayling, T. M. (2009) Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes* **58**, 1463–1467.

Province, M. A. & Borecki, I. B. (2008) Gathering the gold dust: Methods for assessing the aggregate impact of small effect genes in genomic scans. *Pac Symp Biocomput* **2008**, 190–200.

Ripley, B. D. (1996) *Pattern recognition and neural networks.* Cambridge University Press.

Ritchie, M. D., Hahn, L. W. & Moore, J. H. (2003) Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* **24**, 150–157.

Ritchie, M. D., Hahn, L. W., Roodi N, Bailey, L. R., Dupont, W. D., Parl, F. F. & 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.

Saccone, S. F., Saccone, N. L., Swan, G. E., Madden, P. A., Goate, A. M., Rice, J. P. & Bierut, L. J. (2008) Systematic biological prioritization after a genome-wide association study: An application to nicotine dependence. *Bioinformatics* **24**, 1805–1811.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res* **32**, D449–D451.

Segre, D., Deluna, A., Church, G. M. & Kishony, R. (2005) Modular epistasis in yeast metabolism. *Nat Genet* **37**, 77–83.

Sha, Q., Zhang, Z., Schymick, J. C., Traynor, B. J. & Zhang, S. (2009) Genome-wide association reveals three, S. N.P.s associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis. *BMC Med Genet* **10**, 86.

Soares, M. L., Coelho, T., Sousa, A., Batalov, S., Conceicao, I., Sales-Luis, M. L., Ritchie, M. D., Williams, S. M., Nievergelt, C. M., Schork, N. J., Saraiva, M. J. & Buxbaum, J. N. (2005) Susceptibility and modifier genes in Portuguese transthyretin V30M amyloid polyneuropathy: Complexity in a single-gene disease. *Hum Mol Genet* **14**, 543–553.

Steffens, M., Becker, T., Sander, T., Fimmers, R., Herold, C., Holler, D. A., Leu, C., Herms, S., Cichon, S., Bohn, B., Gerstner, T., Griebel, M., Nothen, M. M., Wienker, T. F. & Baur, M. P. (2010) Feasible and successful: Genome-wide interaction analysis involving all 1.9 × 10 pair-wise interaction tests. *Hum Hered* **69**, 268–284.

Templeton, A. (2000) Epistasis and complex traits. In: *Epistasis and the evolutionary process* (eds. M. Wade, B. Brodie, III & J. Wolf), pp. 41–57. Oxford: Oxford University Press.

Thomas, P. D., Mi H, Swan, G. E., Lerman, C., Benowitz, N., Tyndale, R. F., Bergen, A. W. & Conti, D. V. (2009) A systems biology network model for genetic association studies of nicotine addiction and treatment. *Pharmacogenet Genomics* **19**, 538–551.

Torkamani, A. & Schork, N. J. (2009) Pathway and network analysis with high-density allelic association data. *Methods Mol Biol* **563**, 289–301.

Vincent, A. L., Billingsley, G., Buys, Y., Levin, A. V., Priston, M., Trope, G., Williams-Lyn, D. & Heon, E. (2002) Digenic inheritance of early-onset glaucoma: CYP1B1, a potential modifier gene. *Am J Hum Genet* **70**, 448–460.

Wang, K., Li, M. & Bucan, M. (2007) Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **81**, 1278–1283.

Warden, C. H., Yi, N. & Fisler, J. (2004) Epistasis among genes is a universal phenomenon in obesity: evidence from rodent models. *Nutrition* **20**, 74–77.

Wilke, R. A., Mareedu, R. K. & Moore, J. H. (2008) The pathway less traveled: Moving from candidate genes to candidate pathways in the analysis of genome-wide data from large scale pharmacogenetic association studies. *Curr Pharmacogenomics Person Med* **6**, 150–159.

Wille, A., Hoh, J. & Ott, J. (2003) Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. *Genet Epidemiol* **25**, 350–359.

Wongseree, W., Assawamakin, A., Piroonratana, T., Sinsomros, S., Limwongse, C. & Chaiyaratana, N. (2009) Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. *BMC Bioinformatics* **10**, 294.

Zamar, D., Tripp, B., Ellis, G. & Daley, D. (2009) Path: A tool to facilitate pathway-based genetic association analysis. *Bioinformatics* **25**, 2444–2446.

Zhang, L., Guo, Y. F., Liu, Y. Z., Liu, Y. J., Xiong, D. H., Liu, X. G., Wang, L., Yang, T. L., Lei, S. F., Guo Y, Yan H, Pei, Y. F., Zhang, F., Papasian, C. J., Recker, R. R. & Deng, H. W. (2010) Pathway-based genome-wide association analysis identified the importance of regulation-of-autophagy pathway for ultradistal radius BMD. *J Bone Miner Res* **25**, 1572–1580.

Zhong, H., Yang X, Kaplan, L. M., Molony, C. & Schadt, E. E. (2010) Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet* **86**, 581–591.

Zhu, J. & Hastie, T. (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5**, 427–443.

*Received*: 12 May 2010
*Accepted*: 14 October 2010