Chapter 2

Finding the Epistasis Needles in the Genome-Wide Haystack

Marylyn D. Ritchie

Abstract

Genome-wide association studies (GWAS) have dominated the field of human genetics for the past 10 years. This study design allows for an unbiased, dense exploration of the genome and provides researchers with a vast array of SNPs to look for association with their trait or disease of interest. GWAS has been referred to as finding needles in a haystack and while many of these "needles," or SNPs associating with disease, have been identified, there is still a great deal of heritability yet to be explained. The missing or phantom heritability is due, at least in part, to epistasis or gene–gene interactions, which have not been extensively explored in GWAS. Part of the challenge for epistasis analysis in GWAS is the sheer magnitude of the search and the computational complexity associated with it. An exhaustive search for epistasis models is not computationally feasible; thus, alternate approaches must be considered. In this chapter, these approaches will be reviewed briefly, and the incorporation of biological knowledge to guide this process will be further expanded upon. Real biological data examples where this approach has yielded successful identification of epistasis will also be provided. Epistasis has been known to be important since the early 1900s; however, its prevalence in mainstream research has been somewhat overshadowed by molecular technology advances. Due to the increasing evidence of epistasis in complex traits, it continues to emerge as a likely explanation for missing heritability.

Key words Epistasis, Prior knowledge, Missing heritability, Filtering, Enrichment, Pathways

1 Introduction

The search for the missing heritability [1, 2] in genome-wide association studies (GWAS) has become an important focus for the human genetics community – especially as larger and larger sample sizes have resulted in even smaller effect sizes to be identified. The National Human Genome Research Institute (NHGRI) GWAS catalog was developed to store all of the GWAS results in a central database. A few years ago, NHGRI looked at the distribution of GWAS-associated SNPs and found a majority were associated with small effects sizes (odds ratios less than 1.4) [3]. In January 2014, we evaluated the GWAS catalog to see if the trend had changed, and unfortunately, due to increasing sample sizes,

Jason H. Moore and Scott M. Williams (eds.), *Epistasis: Methods and Protocols*, Methods in Molecular Biology, vol. 1253, DOI 10.1007/978-1-4939-2155-3_2,

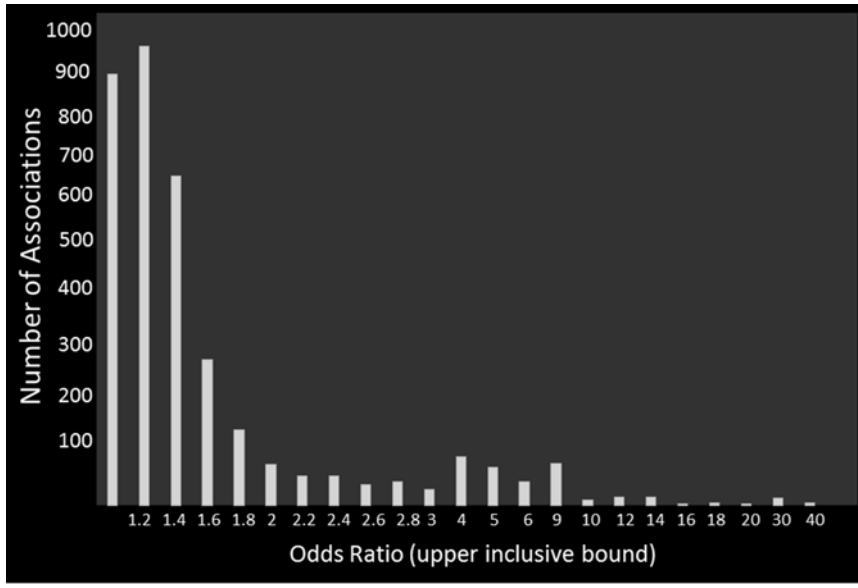

Fig. 1 Distribution of odds ratios (effect size) from the NHGRI GWAS catalog as of January 2014

the effect sizes identified have become even smaller. Figure 1 shows the distribution of these effects; the majority have odds ratios less than 1.2. This leads to much discussion regarding the missing or phantom heritability. Lander and colleagues explain that much of the missing heritability could be due to genetic interactions [4]. This opinion has been shared and emphasized by others in the literature for several years [5–11]. As such, it is believed by many that epistasis is important and should be explored in the context of GWAS; however, specific applications of epistasis analysis in GWAS have been much fewer than single variant main effects analyses. The computational burden of exploring gene–gene interactions in the wealth of data generated in GWAS, along with small to moderate sample sizes, has led to epistasis being an afterthought, rather than a primary focus of GWAS analyses.

In this chapter, we discuss some potential approaches to make epistasis analysis more computationally tractable in a GWAS dataset. A number of alternative approaches are described, but the primary focus is 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 GWAS study can be done in several different ways and the knowledge can be extracted from a variety of database sources. These approaches will be discussed, and some successful applications will be described. Incorporating biological knowledge is likely to be fruitful in the search for epistasis in large-scale genomic studies of the current state-of-the-art and into the future.

2 The Scope of the Problem

The ultimate goal of any disease gene discovery project is to identify as much of the genomic variation as possible that is relevant to the phenotype (disease or trait) being studied. As molecular 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 GWAS studies and the ability to integrate results across experiments have simply not kept pace with this flood of genotyping and now sequencing data. It is a central fallacy that simply generating more data and collecting more samples/individuals will solve the problem. Instead, it is this tsunami of data that has made distinguishing true scientific discoveries from the thousands or millions of false discoveries even more challenging. The ultimate success of our monumental investment in data generation will depend largely on the development and use of innovative analytic approaches and intelligent study designs that allow for the detection of gene–gene and gene–environment interactions.

A major hurdle in discovering epistasis, however, is the variable selection problem. Exhaustively evaluating all of the possible combinations of SNPs is not computationally feasible. For example, testing all two-SNP models in genome-wide data including one million SNPs generates 5.00×10^{11} possible two-SNP models; this requires extensive computing resources and produces many statistically significant results. If we consider going beyond pairwise models, one million SNPs generate 1.7×10^{17} three-SNP models, 4.2×10^{22} four-SNP models, 8.3×10^{27} five-SNP models, 1.4×10^{33} six-SNP models, and so on (calculated based on (n/m), where n is the number of SNPs and m is the number of variables in the model). This creates an enormous computing challenge as well as a multiple testing correction issue. 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-SNP models [12]. Steffens et al. 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 [13]. These strategies, as well as many others that have been developed in the past few years, provide capability to perform an exhaustive pairwise GWIA [14–16]. However, beyond pairwise interactions, exhaustive searching is not tractable.

3 Methods for Data Reduction

It is clear that while the goal of GWAS is to survey the entire genome in an unbiased way, this type of approach simply does not work in the search for epistasis, especially beyond pairwise interactions. A number of filtering approaches have been suggested to reduce the computational burden. First, using statistical evidence of single-SNP effects to prioritize SNPs can be promising and has been shown to have high power [14, 17]. 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). For example, in a 500,000 SNP GWAS analysis, one might use a threshold of $p < 1 \times 10^{-5}$ based on a chi-square 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. Filtering SNPs based on the strength of independent main effects can identify SNP combinations among loci with small to moderate main effects, such as two 2-SNP models identified for Amyotrophic lateral sclerosis (ALS) [18] or multiple sclerosis (MS) [19]. If, however, the genetic variants that are important for disease risk have effects only through their interactions with other genes, this filtering by main effects approach would potentially miss these types of discoveries.

The second approach is to use intrinsic knowledge extracted from the dataset to filter the list of SNPs to test for interactions. Data reduction algorithms that explicitly assess the quality of an SNP in its relationship to the clinical outcome are an alternative to pure 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 [20, 21]. These approaches use a nearest neighbor approach to assess SNP quality to detect attributes associated with disease. In this case, nearest neighbors are individuals in the dataset who are genetically similar at the many SNPs across 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 computational complexity issues as well as multiple comparisons issues [22].

Third, the use of extrinsic biological knowledge to filter SNPs and then evaluate multi-marker combinations based on biological criteria has been suggested [23–24]. If we filter variables using biological information extrinsic to our dataset—i.e., only examine interactions between SNPs in a common pathway or with a common structure or function based on the literature or information in databases—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 analysis is conditioned on 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 [22].

Each of these strategies imposes a specific bias into the analysis, and no one strategy will be optimal in all cases (*see* **Note 1**). Each of these has advantages and disadvantages with known biases and limitations. While all of the proposed approaches for filtering have clear strengths and limitations, we propose that filtering based on extrinsic biological knowledge will 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 burdens. By limiting the search space, we limit the number of statistical tests, multiple comparison burden, as well as computational complexity. The remainder of this chapter will focus on approaches being developed for using biological knowledge to prioritize the search for missing heritability in the epistasis domain and provide some examples where these strategies have been implemented.

4 Methods for Incorporating Biological Knowledge

The incorporation of prior knowledge into GWAS has been proposed and many new tools have been developed to allow for this type of analysis (*see* **Note 2**). 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 with suggestions for how they could guide the search for epistasis in large, genome-wide datasets.

4.1 Protein–Protein Interaction Approaches

Protein–protein interactions can be measured using mass spectrometry, immunoprecipitation, yeast two-hybrids, and affinity pull down followed by mass spectrometry [25]. As discussed in Ritchie [22], a number of protein–protein interaction databases are publicly available, including the database of interacting proteins (DIPs) [26], BioGRID (Biological General Repository for Interaction Datasets) [27], and human protein reference database (HPRD) [28]. As described by Pattin and Moore, a couple of different approaches could be used to incorporate protein–protein interaction data [25]. 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 [25]. This would reduce the number of SNPs explored for epistasis. However, it would also prevent the identification of models that include 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, rather than filtered in or out of the dataset. This would allow for novel biology to be discovered, although it would favor models with a priori evidence of support [25]. Scoring systems like this can then be used for filtering as well as for Bayesian priors for analysis.

4.2 Pathway Approaches

As discussed by Ritchie [22], the use of pathway data to look for overrepresentation of genome-wide associated hits has been done in many studies. For example, Perry et al. used Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCarta, and Gene Ontology (GO) to perform a modified gene-set enrichment analysis (GSEA) for type II diabetes [29]. In rheumatoid arthritis (RA), Beyene et al. utilized a selection of prior knowledge from c2 curated gene sets, which are obtained from online pathway databases, citations in PubMed, and domain experts [30]. For GSEA, their final set included 1,900 gene sets collected from canonical pathways, chemical and genetic perturbations, BioCarta pathways, GeneMAPP, and KEGG [30].

Another similar gene set enrichment analysis, the SNP-ratio test (SRT) [31], compares the proportion of statistically significant genes to all SNPs within genes that are part of a specific pathway. An empirical *p*-value is then calculated based on comparisons in datasets where a permutation test has been performed (i.e., the assignment of case/control status has been randomized). Approaches like this rely largely on single-SNP analysis, but then look for enrichment of sets of SNPs to report interesting findings, with the idea that sets of interacting SNPs/genes would show up in pathway enrichment tests. So while it is not epistasis evidence, they are at least considering polygenic models.

Baranzini and colleagues [18] propose a protein interaction and network-based analysis (PINBPA) for the study of a multiple sclerosis (MS) dataset. An alternative approach was explored by Askland et al., where they used exploratory visual analysis (EVA) to

perform a number of pathway-based analyses of bipolar disorder [32]. Another approach, pathway genetic load (PGL), looks for evidence of epistasis between genes confined to a single pathway [33]. This approach dramatically reduces the computational complexity of an epistasis search in GWAS data.

As discussed by Ritchie [22], assessing the statistical significance of pathways is also an important and difficult challenge. It is not enough to simply look for an overrepresentation 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 greater chance of being statistically significant, and many of the bioinformatics tools used for these types of studies are biased toward detecting large, well-defined pathways [34]. Methods to perform permutation testing in pathway analysis frameworks have been developed to provide increased power and efficiency [35]. Other approaches index pathways using Gene Ontology terms and test for overrepresentation of pathways in a list of hits from a genome-wide association study (such as ALIGATOR-Association LIst Go AnnoTatOR) and successfully identify pathways for complex traits, such as bipolar disorder [36]. It is also important to re-iterate 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 accumulated single locus associations are detected. However, it is obvious that these pathway approaches will develop hypotheses regarding potential "underground networks" [1], which would be particularly interesting to focus efforts for detection of epistasis.

4.3 Comprehensive Knowledge Approach

Perhaps the most lucrative solution involves a comprehensive knowledge-based approach that includes evidence from pathways, protein–protein interactions, prior association, gene ontology, linkage, or gene expression, etc. Because we have a limited number of known epistatic models in humans, it is currently a challenge to hypothesize what structure models will involve and what relationships between genes we should expect. 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 humans, we are merely speculating and may not include all possible types of models.

Ritchie [22] described that 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 two issues. First, if you do the straightforward type of simulation study where you preselect functional SNPs based on biological knowledge, embed them into the simulation, and 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 [37, 38]. Layers of biological machinery exist between genetic variations and the phenotypes they manifest, and imposing this extra dimension of known biological information into statistical analyses 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 BioGRID, dbSNP, NCBI gene, Gene Otology, MINT, NetPath, OregAnno, Pfam, PharmGKB, Reactome, UCSC genome browser, and the NHGRI GWAS catalog [38]. 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 analyses. Rather than annotating the independent effect of each SNP in a GWAS dataset, the Biofilter allows the explicit detection and modeling of interactions between a set of SNPs preselected by the application. 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 there is much more to be discovered [38].

Biofilter uses biological information about gene–gene relationships to construct multi-SNP models that can then be prioritized before conducting any statistical analyses. The key idea behind Biofilter model generation is that any pathway, ontological category, protein family, experimental interaction, or other grouping of genes or proteins implies a relationship between each of those genes or proteins. Thus Biofilter provides a tool to discover significant multi-SNP models with nonsignificant main effects that have established biological plausibility. The Biofilter model generation process thus far has been protein-coding gene-centric, and as such, SNPs from GWAS genotyping platforms must first be assigned to

protein-coding regions [38, 39]. Relationships between genes represented by a genotyping platform can then be translated to multi-SNP models. If the same two genes appear together in more than one grouping, they're likely to have an important biological relationship; if they appear in multiple groups from several independent sources, then they're even more likely to be biologically related in some way, and receive a higher implication index. Biofilter has access to thousands of such groupings because of the use of multiple domain sources and can analyze all of them to identify the sets of genes or SNPs appearing together in the greatest number of groupings and the widest array of original data sources. These pairs can then be tested for significance within a research dataset, and, depending on the level of data filtering or application of an implication index cutoff, Biofilter can be used to avoid the prohibitive computational and multiple-testing burden of an exhaustive pairwise analysis. Once multi-SNP models are constructed, they can be evaluated using any relevant analytic method such as logistic regression, multifactor dimensionality reduction (MDR), Bayesian networks, etc.

We have developed the Library of Knowledge Integration (LOKI) database (Fig. 2) and integrated eleven public domain data sources. These sources are combined into one central LOKI database for use in annotation, filtering, and building models of gene–gene interactions for analysis. Biofilter and LOKI are described in detail in Pendergrass et al. [38]. Biofilter has been

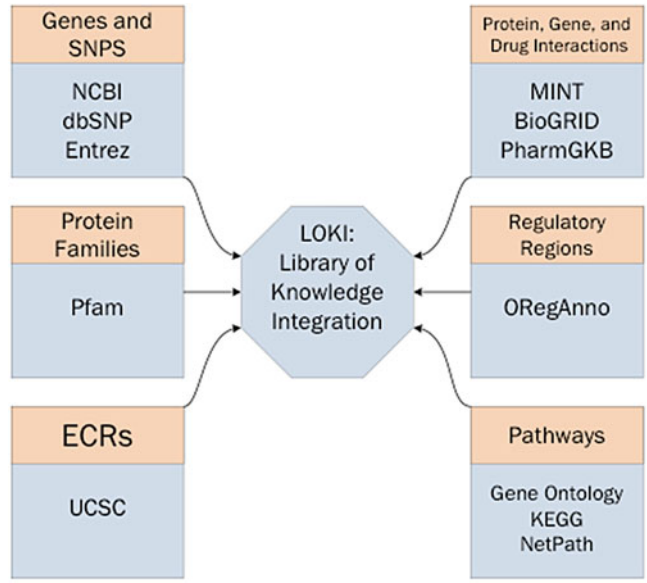

Fig. 2 The current Library of Knowledge Integration (LOKI). LOKI contains information from multiple database repositories, covering multiple domains. From [38]

applied to a number of natural, biological datasets for the discovery of gene–gene interaction models associated with complex traits including Multiple Sclerosis [40], HDL cholesterol [41], HIV Pharmacogenomics [42], and cataract status [43].

Another approach for comprehensive data integration is INTERSNP. INTERSNP is a powerful, flexible approach that implements logistic regression or log-linear models for joint analysis of multiple SNPs [16]. 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 [16]. 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 use it in an efficient manner (*see* **Note 4**).

5 Real World Example: HDL Cholesterol

Plasma concentrations of low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, triglycerides (TRI), and total cholesterol are among the most important risk factors for coronary artery disease (CAD). Lipid traits have been well studied in genome-wide association studies with between 100,000 [44] and over 188,000 individuals [45] included. While there have been over 150 loci identified for association with lipid traits, the proportion of heritability explained is still modest with ~25–30 % of the genetic variance for each lipid trait [44, 45]. Due to this missing heritability, several groups have embarked on explorations of epistasis or gene–gene interactions in lipid traits.

For example, Turner et al. looked for epistasis associated with HDL cholesterol [41] using a biological knowledge-driven filtering method. Here, the Biofilter [37, 38] was used to decrease the number of SNP–SNP models evaluated from genome-wide genotype data. Through the application of the Biofilter, eleven significant GxG models were in the discovery Biobank cohort, eight of which show evidence of replication in a second biobank cohort [41]. The strongest predictive model included a pairwise interaction between *LPL* (which modulates the incorporation of triglyceride into HDL) and *ABCA1* (which modulates the incorporation of free cholesterol into HDL) [41]. The authors required that any GxG interactions in the discovery cohort ($n = 3{,}740$ participants) showed evidence of replication in the de-identified EMRs of a second cohort ($n = 1{,}858$ participants). This resulted in replicated GxG interactions associated with variation in HDL-C, all of which have potential biological relevance.

A similar approach was taken by Ma et al.; they used prior knowledge from established genome-wide association study (GWAS) hits, protein–protein interactions, as well as pathway information to guide their gene–gene interaction analyses [46]. These results were further followed up through the evaluation of gene-based interaction analysis [47] as well as potential eQTLs involved in gene–gene interactions [48]. These results demonstrated that gene–gene interactions modulate complex human traits, including HDL cholesterol, and the use of prior biological knowledge can increase power to identify biologically interesting and relevant models (*see* **Note 3**).

6 Notes

1. The search space for enumerating all possible epistasis models in genome-wide datasets is computationally prohibitive; thus numerous data reduction or filtering strategies have been employed to reduce the SNP set for epistasis modeling including:
   - Statistical filtering using single-SNP statistics (such as the chi-square test).
     - Advantage: simple, unbiased with respect to the biologist.
     - Disadvantage: relies on all important genes having independent main effects.
   - Intrinsic filtering using statistical or computational data-driven approaches (such as ReliefF).
     - Advantage: unbiased with respect to the biologist; uses the data.
     - Disadvantage: complicated; models may not have biological relevance.
   - Extrinsic filtering using biological knowledge (such as Biofilter or pathway analysis).
     - Advantage: results are biologically relevant.
     - Disadvantage: limited by current state of biology; biased toward genes we know something about as a field.
2. Biological knowledge-based epistasis methods are emerging as powerful strategies for epistasis analyses.
3. Real data applications have been deemed successful finding evidence of epistasis replicating across multiple datasets.
4. Many methods for incorporation of biological knowledge into epistasis analysis exist and continue to be developed.

References

1. Maher B (2008) Personal genomes: the case of the missing heritability. Nature 456:18–21. doi:10.1038/456018a
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753. doi:10.1038/nature08494
3. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (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. doi:10.1073/pnas.0903103106
4. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109(4):1193–1198, 201119675. doi: 10.1073/pnas.1119675109
5. Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56:73–82
6. Moore JH, Williams SM (2005) Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays 27:637–646
7. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10:392–404. doi:10.1038/nrg2579
8. Templeton AR (2000) Epistasis and complex traits. Epistasis and the evolutionary process. Oxford University Press, New York, pp 41–57
9. Gibson G (1996) Epistasis and pleiotropy as natural properties of transcriptional regulation. Theor Popul Biol 49:58–89
10. Moore JH (2005) A global view of epistasis. Nat Genet 37:13–14. doi:10.1038/ng0105-13
11. McKinney BA, Pajewski NM (2011) Six degrees of epistasis: statistical network models for GWAS. Front Genet 2:109. doi:10.3389/fgene.2011.00109
12. Bush WS, Dudek SM, Ritchie MD (2006) Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. Bioinformatics 22:2173–2174
13. Steffens M, Becker T, Sander T, Fimmers R, Herold C, Holler DA, Leu C, Herms S, Cichon S, Bohn B, Gerstner T, Griebel M, Nöthen MM, Wienker TF, Baur MP (2010) Feasible and successful: genome-wide interaction analysis involving all 1.9 × 1011 pair-wise interaction tests. Hum Hered 69:268–284. doi:10.1159/000295896
14. Evans DM, Marchini J, Morris AP, Cardon LR (2006) Two-stage two-locus models in genome-wide association. PLoS Genet 2:e157. doi:10.1371/journal.pgen.0020157
15. Ueki M, Cordell HJ (2012) Improved statistics for genome-wide interaction analysis. PLoS Genet 8:e1002625. doi:10.1371/journal.pgen.1002625
16. Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T (2009) INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinform Oxf Engl 25:3275–3281. doi:10.1093/bioinformatics/btp596
17. Kooperberg C, Leblanc M (2008) Increasing the power of identifying gene x gene interactions in genome-wide association studies. Genet Epidemiol 32:255–263. doi:10.1002/gepi.20300
18. Sha Q1, Zhang Z, Schymick JC, Traynor BJ, Zhang S. Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis. BMC Med Genet. 2009 Sep 9;10:86
19. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BMJ, Kappos L, GeneMSA Consortium, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 18:2078–2090. doi:10.1093/hmg/ddp120
20. Greene CS, Penrod NM, Kiralis J, Moore JH (2009) Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions. BioData Min 2:5. doi:10.1186/1756-0381-2-5
21. Moore JH, White BC (2007) Tuning relieff for genome-wide genetic analysis. In: Moore JH, Rajapakse JC, Marchiori E (eds) Evolutionary computation, machine learning and data mining, bioinformatics. Springer, Berlin, pp 166–175
22. Ritchie MD (2011) Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. Ann Hum Genet 75:172–182. doi:10.1111/j.1469-1809.2010.00630.x
23. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452
24. Sun X, Lu Q, Mukheerjee S, Crane PK, Elston R, Ritchie MD (2014) Analysis pipeline for the

epistasis search – statistical versus biological filtering. Front Genet 5:106. doi:10.3389/fgene.2014.00106

25. Pattin KA, Moore JH (2008) Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. Hum Genet 124:19–29. doi:10.1007/s00439-008-0522-8
26. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res 32:D449–D451. doi:10.1093/nar/gkh086
27. Breitkreutz B-J, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V, Dolinski K, Tyers M (2008) The BioGRID interaction database: 2008 update. Nucleic Acids Res 36:D637–D640. doi:10.1093/nar/gkm1001
28. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HGM, Nagini M, Kumar GSS, Jose R, Deepthi P, Mohan SS, Gandhi TKB, Harsha HC, Deshpande KS, Sarker M, Prasad TSK, Pandey A (2006) Human protein reference database – 2006 update. Nucleic Acids Res 34:D411–D414. doi:10.1093/nar/gkj141
29. Perry JRB, McCarthy MI, Hattersley AT, Zeggini E, Wellcome Trust Case Control Consortium, Weedon MN, Frayling TM (2009) Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Diabetes 58:1463–1467. doi:10.2337/db08-1378
30. Beyene J, Hu P, Hamid JS, Parkhomenko E, Paterson AD, Tritchler D (2009) Pathway-based analysis of a genome-wide case-control association study of rheumatoid arthritis. BMC Proc 3(Suppl 7):S128
31. O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A (2009) The SNP ratio test: pathway analysis of genome-wide association datasets. Bioinform Oxf Engl 25:2762–2763. doi:10.1093/bioinformatics/btp448
32. 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. doi:10.1007/s00439-008-0600-y
33. Huebinger RM, Garner HR, Barber RC (2010) Pathway genetic load allows simultaneous evaluation of multiple genetic associations. Burns 36:787–792. doi:10.1016/j.burns.2010.02.001
34. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol 33:419–431. doi:10.1002/gepi.20395
35. 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 Bioinformatics 10:429. doi:10.1186/1471-2105-10-429
36. Holmans P, Green EK, Pahwa JS, Ferreira MAR, Purcell SM, Sklar P, Owen MJ, O'Donovan MC, 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. doi:10.1016/j.ajhg.2009.05.011
37. Bush WS, Dudek SM, Ritchie MD (2009) Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput 368–379
38. Pendergrass SA, Frase AT, Wallace JR, Wolfe D, Katiyar N, Moore C, Ritchie MD (2013) Genomic analyses with biofilter 20: knowledge driven filtering, annotation, and model development. BioData Min 6(1):25
39. Bush WS, Chen G, Torstenson ES, Ritchie MD (2009) LD-spline: mapping SNPs on genotyping platforms to genomic regions using patterns of linkage disequilibrium. BioData Min 2:7. doi:10.1186/1756-0381-2-7
40. Bush WS, McCauley JL, DeJager PL, Dudek SM, Hafler DA, Gibson RA, Matthews PM, Kappos L, Naegelin Y, Polman CH, Hauser SL, Oksenberg J, Haines JL, Ritchie MD (2011) A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility. Genes Immun 12:335–340. doi:10.1038/gene.2011.3
41. Turner SD, Berg RL, Linneman JG, Peissig PL, Crawford DC, Denny JC, Roden DM, McCarty CA, Ritchie MD, Wilke RA (2011) Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. PLoS One 6:e19586. doi:10.1371/journal.pone.0019586
42. Grady BJ, Torstenson ES, McLaren PJ, De Bakker PIW, Haas DW, Robbins GK, Gulick RM, Haubrich R, Ribaudo H, Ritchie MD (2011) Use of biological knowledge to inform the analysis of gene-gene interactions involved

in modulating virologic failure with efavirenz-containing treatment regimens in art-naïve actg clinical trials participants. Pac Symp Biocomput 2011:253–264

43. Pendergrass SA, Verma SS, Holzinger ER, Moore CB, Wallace J, Dudek SM, Huggins W, Kitchner T, Waudby C, Berg R, McCarty CA, Ritchie MD (2013) Next-generation analysis of cataracts: determining knowledge driven gene-gene interactions using Biofilter, and gene-environment interactions using the PhenX Toolkit. Pac Symp Biocomput 147–158
44. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, Johansen CT, Fouchier SW, Isaacs A, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Aulchenko YS, Thorleifsson G, Feitosa MF, Chambers J, Orho-Melander M, Melander O, Johnson T, Li X, Guo X, Li M, Shin Cho Y, Jin Go M, Jin Kim Y, Lee J-Y, Park T, Kim K, Sim X, Twee-Hee Ong R, Croteau-Chonka DC, Lange LA, Smith JD, Song K, Hua Zhao J, Yuan X, Luan J, Lamina C, Ziegler A, Zhang W, Zee RYL, Wright AF, Witteman JCM, Wilson JF, Willemsen G, Wichmann H-E, Whitfield JB, Waterworth DM, Wareham NJ, Waeber G, Vollenweider P, Voight BF, Vitart V, Uitterlinden AG, Uda M, Tuomilehto J, Thompson JR, Tanaka T, Surakka I, Stringham HM, Spector TD, Soranzo N, Smit JH, Sinisalo J, Silander K, Sijbrands EJG, Scuteri A, Scott J, Schlessinger D, Sanna S, Salomaa V, Saharinen J, Sabatti C, Ruokonen A, Rudan I, Rose LM, Roberts R, Rieder M, Psaty BM, Pramstaller PP, Pichler I, Perola M, Penninx BWJH, Pedersen NL, Pattaro C, Parker AN, Pare G, Oostra BA, O'Donnell CJ, Nieminen MS, Nickerson DA, Montgomery GW, Meitinger T, McPherson R, McCarthy MI, McArdle W, Masson D, Martin NG, Marroni F, Mangino M, Magnusson PKE, Lucas G, Luben R, Loos RJF, Lokki M-L, Lettre G, Langenberg C, Launer LJ, Lakatta EG, Laaksonen R, Kyvik KO, Kronenberg F, König IR, Khaw K-T, Kaprio J, Kaplan LM, Johansson A, Jarvelin M-R, Janssens ACJW, Ingelsson E, Igl W, Kees Hovingh G, Hottenga J-J, Hofman A, Hicks AA, Hengstenberg C, Heid IM, Hayward C, Havulinna AS, Hastie ND, Harris TB, Haritunians T, Hall AS, Gyllensten U, Guiducci C, Groop LC, Gonzalez E, Gieger C, Freimer NB, Ferrucci L, Erdmann J, Elliott P, Ejebe KG, Döring A, Dominiczak AF, Demissie S, Deloukas P, de Geus EJC, de Faire U, Crawford G, Collins FS, Chen YI, Caulfield MJ, Campbell H, Burtt NP, Bonnycastle LL, Boomsma DI, Boekholdt SM, Bergman RN, Barroso I, Bandinelli S, Ballantyne CM, Assimes TL, Quertermous T, Altshuler D, Seielstad M, Wong TY, Tai E-S, Feranil AB, Kuzawa CW, Adair LS, Taylor HA Jr, Borecki IB, Gabriel SB, Wilson JG, Holm H, Thorsteinsdottir U, Gudnason V, Krauss RM, Mohlke KL, Ordovas JM, Munroe PB, Kooner JS, Tall AR, Hegele RA, Kastelein JJP, Schadt EE, Rotter JI, Boerwinkle E, Strachan DP, Mooser V, Stefansson K, Reilly MP, Samani NJ, Schunkert H, Cupples LA, Sandhu MS, Ridker PM, Rader DJ, van Duijn CM, Peltonen L, Abecasis GR, Boehnke M, Kathiresan S (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713. doi:10.1038/nature09270
45. Global Lipids Genetics Consortium, Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, Beckmann JS, Bragg-Gresham JL, Chang H-Y, Demirkan A, Den Hertog HM, Do R, Donnelly LA, Ehret GB, Esko T, Feitosa MF, Ferreira T, Fischer K, Fontanillas P, Fraser RM, Freitag DF, Gurdasani D, Heikkilä K, Hyppönen E, Isaacs A, Jackson AU, Johansson A, Johnson T, Kaakinen M, Kettunen J, Kleber ME, Li X, Luan J, Lyytikäinen L-P, Magnusson PKE, Mangino M, Mihailov E, Montasser ME, Müller-Nurasyid M, Nolte IM, O'Connell JR, Palmer CD, Perola M, Petersen A-K, Sanna S, Saxena R, Service SK, Shah S, Shungin D, Sidore C, Song C, Strawbridge RJ, Surakka I, Tanaka T, Teslovich TM, Thorleifsson G, Van den Herik EG, Voight BF, Volcik KA, Waite LL, Wong A, Wu Y, Zhang W, Absher D, Asiki G, Barroso I, Been LF, Bolton JL, Bonnycastle LL, Brambilla P, Burnett MS, Cesana G, Dimitriou M, Doney ASF, Döring A, Elliott P, Epstein SE, Eyjolfsson GI, Gigante B, Goodarzi MO, Grallert H, Gravito ML, Groves CJ, Hallmans G, Hartikainen A-L, Hayward C, Hernandez D, Hicks AA, Holm H, Hung Y-J, Illig T, Jones MR, Kaleebu P, Kastelein JJP, Khaw K-T, Kim E, Klopp N, Komulainen P, Kumari M, Langenberg C, Lehtimäki T, Lin S-Y, Lindström J, Loos RJF, Mach F, McArdle WL, Meisinger C, Mitchell BD, Müller G, Nagaraja R, Narisu N, Nieminen TVM, Nsubuga RN, Olafsson I, Ong KK, Palotie A, Papamarkou T, Pomilla C, Pouta A, Rader DJ, Reilly MP, Ridker PM, Rivadeneira F, Rudan I, Ruokonen A, Samani N, Scharnagl H, Seeley J, Silander K, Stancáková A, Stirrups K, Swift AJ, Tiret L, Uitterlinden AG, van Pelt LJ, Vedantam S, Wainwright N, Wijmenga C, Wild SH, Willemsen G, Wilsgaard T, Wilson JF, Young EH, Zhao JH, Adair LS, Arveiler D, Assimes TL, Bandinelli S, Bennett F, Bochud M,

Boehm BO, Boomsma DI, Borecki IB, Bornstein SR, Bovet P, Burnier M, Campbell H, Chakravarti A, Chambers JC, Chen Y-DI, Collins FS, Cooper RS, Danesh J, Dedoussis G, de Faire U, Feranil AB, Ferrières J, Ferrucci L, Freimer NB, Gieger C, Groop LC, Gudnason V, Gyllensten U, Hamsten A, Harris TB, Hingorani A, Hirschhorn JN, Hofman A, Hovingh GK, Hsiung CA, Humphries SE, Hunt SC, Hveem K, Iribarren C, Järvelin M-R, Jula A, Kähönen M, Kaprio J, Kesäniemi A, Kivimaki M, Kooner JS, Koudstaal PJ, Krauss RM, Kuh D, Kuusisto J, Kyvik KO, Laakso M, Lakka TA, Lind L, Lindgren CM, Martin NG, März W, McCarthy MI, McKenzie CA, Meneton P, Metspalu A, Moilanen L, Morris AD, Munroe PB, Njølstad I, Pedersen NL, Power C, Pramstaller PP, Price JF, Psaty BM, Quertermous T, Rauramaa R, Saleheen D, Salomaa V, Sanghera DK, Saramies J, Schwarz PEH, Sheu WH-H, Shuldiner AR, Siegbahn A, Spector TD, Stefansson K, Strachan DP, Tayo BO, Tremoli E, Tuomilehto J, Uusitupa M, van Duijn CM, Vollenweider P, Wallentin L, Wareham NJ, Whitfield JB, Wolffenbuttel BHR, Ordovas JM, Boerwinkle E, Palmer CNA, Thorsteinsdottir U, Chasman DI, Rotter JI, Franks PW, Ripatti S, Cupples LA, Sandhu MS, Rich SS, Boehnke M, Deloukas P, Kathiresan S, Mohlke KL, Ingelsson E, Abecasis GR (2013) Discovery and refinement of loci associated with lipid levels. Nat Genet 45:1274–1283. doi:10.1038/ng.2797

46. Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, Keinan A (2012) Knowledge-driven analysis identifies a gene-gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. PLoS Genet 8:e1002714. doi:10.1371/journal.pgen.1002714
47. Ma L, Clark AG, Keinan A (2013) Gene-based testing of interactions in association studies of quantitative traits. PLoS Genet 9:e1003321. doi:10.1371/journal.pgen.1003321
48. Ma L, Ballantyne C, Brautbar A, Keinan A (2014) Analysis of multiple association studies provides evidence of an expression QTL hub in gene-gene interaction network affecting HDL cholesterol levels. PLoS One 9:e92469. doi:10.1371/journal.pone.0092469

http://www.springer.com/978-1-4939-2154-6

Epistasis
Methods and Protocols
Moore, J.H.; Williams, S.M. (Eds.)
2015, X, 350 p. 61 illus., 17 illus. in color., Hardcover
ISBN: 978-1-4939-2154-6
A product of Humana Press