Published in final edited form as:

Epidemiology. 2009 July ; 20(4): 508–511. doi:10.1097/EDE.0b013e3181a93b98.

Complex Diseases, Complex Genes: Keeping Pathways on the Right Track

Peter Kraft^a and Soumya Raychaudhuri^{b,c}

^a Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston MA

^b Medical and Population Genetics, Broad Institute, Cambridge, MA

^c Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Boston, MA

In his “Defense of Beanbag Genetics,” J.B.S. Haldane¹ responded to critics who challenged a marginal (ie, one-locus-at-a-time) approach to studying population genetics. Such “beanbag genetics” had been accused of being too simplistic and ignoring the contributions of multiple loci in different (and ever-changing) environmental contexts. Haldane conceded this point of principle (“beanbag genetics do not explain the physiologic interaction of genes and the interaction of genotype and environment”) but went on to argue that despite its simplifications, the marginal approach had proven itself in practice and led to important insights.¹

The arrival of massive amounts of data from genome-wide association studies (GWAS) has turned up the heat on a similar debate in the field of genetic epidemiology. On the one hand, the admittedly simple approach of averaging over genetic and environmental backgrounds and testing each marker in a GWAS for association with the studied trait marginally has proven quite successful,^{2,3} despite early concerns that by ignoring the underlying complexity this naïve approach would fail.^{4,5} On the other hand, the loci discovered to date do not come close to explaining the observed heritability for most studied traits.⁶ “Pathway analyses” acknowledge complexity by considering multiple loci simultaneously and relating them to known functional annotations. In principle, pathway analyses could lead to new discoveries missed by the simple marginal analyses.^{7–11} Moreover, successful identification of associated pathways can clarify disease pathogenesis; indeed, for some phenotypes, multiple loci identified through GWAS have been linked to common pathways.^{12–15}

However, in practice, multilocus pathway analyses present a number of challenges, some of which should be very familiar to epidemiologists.^{16,17} The paper by Breitling et al¹⁸ in this issue discusses the issue of overfitting when there is a large number of potential predictor variables. Other issues may be less familiar—the consequences of the peculiarities of genomic data and how it is annotated. Here we discuss 3 loosely defined approaches to “pathway analysis” and touch on potential pitfalls for each.

Data mining

In this approach, investigators identify a group of functionally related genes, and then apply multivariable analysis techniques to the markers in these genes.^{10,19–23} The hope may be that, although none of the individual markers were credibly associated with the studied trait, the aggregate evidence for association may be strong. Alternatively, a more flexible analysis that incorporates nonlinear relations between the coded marker genotypes and the trait may uncover additional compelling evidence for association. The paper by Lesnick et al,²⁴

discussed by Breitling et al, took this approach, analyzing single nucleotide polymorphisms (SNPs) in genes involved in the axon guidance process using a stepwise regression approach, incorporating SNP main effects and product interaction terms.

As Breitling et al point out, extreme care must be taken with data-mining techniques that search over a large set of possible models, when applied to data sets with a large number of markers. (The “axon guidance pathway” contained between 1,195 and 1,460 markers, depending on the data set analyzed.) Failing to appropriately correct for the model selection procedure can lead to drastic, downwardly biased *P*-values and overestimates of the precision of predicted trait values based on the fit model in a new data set. Breitling et al use a permutation procedure to correct for this overfitting, and report a *P*-value for association that is 35 orders of magnitude larger than that presented by Lesnick et al in their abstract.²⁴ We note that in the presence of population stratification or differential genotyping errors, a simple permutation procedure that randomizes phenotypes against genotypes may still yield downwardly biased *P*-values. Permutations should be done within strata assumed to be homogeneous with respect to ancestry or genotyping errors, if possible.

An alternative approach to assess statistical significance is to compare the observed test statistic for the pathway to the distribution of the test statistic across multiple synthetic “pathways”: sets of genes randomly drawn from the genome (and assumed not to be associated with the studied trait). Although this procedure might account for population stratification bias or genotyping errors (assuming these affect the tested pathway and the rest of the genome similarly), it comes with its own set of potential biases. In particular, this approach could be confounded by differences in characteristics between the tested and synthetic pathways, such as differences in gene size, linkage disequilibrium, etc. Ideally, the synthetic pathways should be matched to the tested pathway on these characteristics. Lesnick et al implemented this approach by randomly selecting individual SNPs from the genome and found that the observed test statistic for the axon guidance pathway was far larger than for any of the 4000 random marker sets simulated. However, by drawing markers randomly from their genome-wide data set, the authors ensured that the linkage disequilibrium patterns in the axon guidance pathway and the synthetic pathways differed greatly.

Data-mining techniques typically require no missing data on predictors or outcome variables. Lesnick et al implemented a form of complete-case analysis, restricting the models tested at each stage of their stepwise procedure to those with less than a certain amount of missing data. Breitling et al had difficulty replicating Lesnick et al’s procedure, and noted that small differences in tuning parameters could lead to quite different results, in both significance of the overall test for association and the markers selected in the final model. This is a disconcerting property for any analysis, and suggests that its conclusions may not be reliable. As data sharing becomes more widespread in the genetic epidemiology community,^{25,26} other investigators will be able to replicate published analyses and assess their sensitivity to different modeling assumptions, as Breitling et al have done. This should lead to more robust scientific conclusions in the long run. We are happy to note that the GWAS data used in Lesnick et al are now available to qualified researchers through dbGAP (Study Accession: phs000048.v1.p1).

Replication of analytic results in an independent data set can provide some assurance that the observed association is not a chance false positive, and provide an unbiased estimate of the prediction accuracy of the fitted model. However, applying a procedure with markedly downwardly biased *P*-values to 2 data sets and getting a “significant” *P*-value in each does not constitute replication—especially when there is little overlap between the sets of genes represented in the final models. A more compelling demonstration would be to show that the model fit in the first data set effectively predicts the trait in the second.

Gene set enrichment analysis

In the second approach to pathway analysis, investigators take a list of markers ranked by decreasing interest—typically increasing P -values from single-locus, marginal tests—and ask whether the top of this list is enriched for markers from genes in particular functional groupings (metabolic pathways, gene ontology categories, modules from gene expression data sets, etc).^{9,27–30} This approach was originally developed for analyzing gene expression data where there is a one-to-one mapping between expression level and gene, and where the fraction of genes differentially expressed across experimental conditions can be quite large.³¹

For genetic data, in contrast, the mapping from SNP marker to gene is many-to-many. Whether a SNP is assigned to a single gene or several, and how it is assigned (according to physical position or linkage disequilibrium patterns), can greatly affect results. Moreover, most of these methods require each gene be given a single association score, and many simply give each gene the maximum test statistic (or smallest P -value) of all the SNPs assigned to that gene.³² This favors genes that contain many SNPs. Thus, gene set enrichment analysis will tend to highlight any pathway that contains several large genes, and tend to miss pathways that contain only small genes. Some adjustment for multiple testing at the gene level—eg, permutation testing—is needed to account for this size bias.⁹ Of note, genes involved in axon guidance tend to be quite large.

Network-based approaches

A third approach leverages networks based on gene expression data, protein-protein interaction data, or published scientific texts.^{33–37} Most of these methods start with a set of implicated genes (eg, the “top hits” from a GWAS) and then attempt to identify nonrandom connectivity among them. One promising aspect of this approach is that it does not require the user to specify a priori a set of functional groupings to be analyzed. The Gene Relationships Across Implicated Loci (GRAIL) algorithm, for example, which builds networks using published abstracts, can identify putative relationships among genes that do not have a single cocitation (<http://www.broad.mit.edu/mpg/grail/>).

There are several pitfalls to this approach. First, the networks are not always reliable. Protein-protein interaction data can be inconsistent; correlation in gene expression may not capture important relationships; and published text is always limited by the scope of human knowledge. Second, as with gene set analysis, it is often not easy to connect markers to genes. In particular, the fact that genes of similar function physically cluster together can create spurious evidence that multiple genes in a pathway are associated with the studied trait. Finally, clear and robust statistical approaches backed by simulations need to be presented to allow these approaches to be useable. Each of the proposed methods use a different set of statistical approaches, and the complexity of networks do not lend themselves to easy calculation of P -values.

Garbage in, garbage out?

The quality of the information used in both data mining and gene set enrichment analyses affects the credibility of conclusions drawn from them.

First, in the context of genome-wide association studies, the relatively small number of common marker alleles truly associated with any given trait (perhaps several hundred at the most, of hundreds of thousands tested), and the small effect for the majority of these markers, require large sample sizes to differentiate the test-statistic distribution for associated markers from that for null markers.

Second, there are many different ways to group genes according to function, and these are of varying relevance and quality. Some metabolic pathways and cellular processes are well studied, leading to a bias in genome annotation. A gene known to be involved in apoptosis, say, is likely to be involved in other (currently unknown) processes and pathways as well. Annotations may also be mistaken. The popular Gene Ontology (GO) classifications have an associated evidence code: less than 1% of GO annotations have been confirmed experimentally.³⁸ When scanning over many gene groups for evidence of association, interpretation of results should ideally factor in prior beliefs regarding the spectrum of causal variants (limited in number and effect by observed heritability). Other important considerations are the informativeness of different groupings (some are too broad or too narrow to be of much use), and the plausibility that different functional groupings might contain multiple loci associated with the trait under study.³⁹

It is not our intent to discourage research into methods for pathway analysis or their application. There are surely limits to what can be learned from marginal analyses for GWAS data, and the approaches outlined here may be able to provide new functional insights beyond those limits. Moreover, analyses of rare variants from sequencing or copy-number studies will require that these variants be grouped into sensible categories to perform association analysis—in epidemiologic studies it may be possible to draw conclusions only about groups of rare variants, not individual rare variants. We also recognize that these analyses are typically undertaken in an effort to prioritize follow-up studies, for example additional genotyping in further samples or experimental work assessing gene function. In this context, strong control of the family-wise error rate (ie, Type I error) may not be a primary concern. But precisely because these additional studies are expensive, care must be taken to avoid biases and errors that will send researchers down blind alleys.

Acknowledgments

P. K. was supported by NCI grant 5U01CA098233-06; S. R. was supported by an NIH Career Development Award (1K08AR055688-01A1).

References

1. Haldane JB. A defense of beanbag genetics. *Perspect Biol Med* 1964;7:343–359. [PubMed: 14161266]
2. Hunter DJ, Kraft P. Drinking from the fire hose—statistical issues in genomewide association studies. *N Engl J Med* 2007;357:436–439. [PubMed: 17634446]
3. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;118:1590–1605. [PubMed: 18451988]
4. Weiss K, Terwilliger J. How many diseases does it take to map a gene with SNPs? *Nature Genet* 2000;26:151–157. [PubMed: 11017069]
5. Shriver D, Vaughan LK, Padilla MA, et al. Problems with genome-wide association studies. *Science* 2007;316:1840–1842. [PubMed: 17600199]
6. Maher B. Personal genomes: The case of the missing heritability. *Nature* 2008;456:18–21. [PubMed: 18987709]
7. Thomas DC. The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* 2005;14:557–559. [PubMed: 15767327]
8. Chen GK, Witte JS. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet* 2007;81:397–404. [PubMed: 17668389]
9. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 2007;81:1278–1283.
10. Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, et al. Pathway analysis by adaptive combination of P-values. *Genet Epidemiol*. 2009[Epub ahead of print]

11. Ulrich CM, Nijhout HF, Reed MC. Mathematical modeling: epidemiology meets systems biology. *Cancer Epidemiol Biomarkers Prev* 2006;15:827–829. [PubMed: 16702357]
12. Kathiresan S, Willer CJ, Peloso GM, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 2009;41:56–65. [PubMed: 19060906]
13. Raychaudhuri S, Remmers EF, Lee AT, et al. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet* 2008;40:1216–1223. [PubMed: 18794853]
14. Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* 2007;39:596–604. [PubMed: 17435756]
15. Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* 2007;448:427–434. [PubMed: 17653185]
16. Robins J, Greenland S. The role of model selection in causal inference from nonexperimental data. *Am J Epidemiol* 1986;123:392–402. [PubMed: 3946386]
17. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–387. [PubMed: 8668867]
18. Breitling L, Steyerberg E, Brenner H. The novel genomic pathway approach to complex diseases: a reason for (over-)optimism? *Epidemiol* 2009;20:498–505.
19. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138–147. [PubMed: 11404819]
20. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* 2005;28:157–170. [PubMed: 15532037]
21. Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. *Hum Hered* 2003;56:83–93. [PubMed: 14614242]
22. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* 2003;4:701–709. [PubMed: 12951571]
23. Millstein J, Conti DV, Gilliland FD, Gauderman WJ. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet* 2006;78:15–27. [PubMed: 16385446]
24. Lesnick TG, Papapetropoulos S, Mash DC, et al. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet* 2007;3:e98. [PubMed: 17571925]
25. Hernan MA, Wilcox AJ. Epidemiology, data sharing, and the challenge of scientific replication. *Epidemiol* 2009;20:167–168.
26. Foster MW, Sharp RR. Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. *Nat Rev Genet* 2007;8:633–639. [PubMed: 17607307]
27. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545–15550. [PubMed: 16199517]
28. Chen Y, Zhu J, Lum PY, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature* 2008;452:429–435. [PubMed: 18344982]
29. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. *Nature* 2008;452:423–428. [PubMed: 18344981]
30. Chasman DI. On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet Epidemiol* 2008;32:658–668. [PubMed: 18481796]
31. Slonim DK. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* 2002;32(Suppl):502–508. [PubMed: 12454645]
32. Bonifaci N, Berenguer A, Díez J, et al. Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes. *BMC Med Genomics* 2008;1:62. [PubMed: 19094230]
33. Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc Natl Acad Sci U S A* 2004;101:15148–15153. [PubMed: 15471992]

34. Iossifov I, Zheng T, Baron M, Gilliam TC, Rzhetsky A. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. *Genome Res* 2008;18:1150–1162. [PubMed: 18417725]
35. Lage K, Zheng T, Baron M, Gilliam TC, Rzhetsky A. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007;25:309–316. [PubMed: 17344885]
36. Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. *Nat Genet* 2002;31:316–319. [PubMed: 12006977]
37. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;78:1011–1025. [PubMed: 16685651]
38. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. *Nat Rev Genet* 2008;9:509–515. [PubMed: 18475267]
39. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124. [PubMed: 16060722]