American Journal of
EPIDEMIOLOGY

Vol. 162, No. 4
Printed in U.S.A.
DOI: 10.1093/aje/kwi200

August 15, 2005

Sponsored by the Society for Epidemiologic Research
Published by Oxford University Press

HUMAN GENOME EPIDEMIOLOGY (HuGE) COMMENTARIES

Genomic Epidemiology of Complex Disease: The Need for an Electronic Evidence-based Approach to Research Synthesis

Michael B. Bracken

From the Center for Perinatal, Pediatric, and Environmental Epidemiology, Yale University School of Medicine, New Haven, CT.

Received for publication April 29, 2005; accepted for publication May 18, 2005.

Modern microarray genotyping now permits simultaneous analysis of tens of thousands of polymorphisms, and this technology is being widely used to associate the role of genes with the etiology of complex disease. Genome-wide hypothesis-free mapping will also increasingly generate candidate genes that require further testing in association studies. At the same time, genetic effects are increasingly observed to be buffered by a wide array of biologic mechanisms that evolved to protect the genome from environmental insult and that serve to obscure observation of direct effects of polymorphisms on a disease phenotype. These two forces combine to make replication of genomic epidemiology extraordinarily difficult. Traditional research synthesis of emerging bodies of genomic epidemiology is problematic and often quickly outdated. The author proposes that electronic evidence-based methodology, perhaps modeled after that used by the Cochrane Collaboration in clinical medicine, would facilitate the systematic preparation and frequent updating of systematic reviews, which is essential for identifying valid and replicable gene-disease associations.

epidemiology; genes; genetics; genome; genome, human; meta-analysis; research; review, systematic

Abbreviations: HuGE, human genome epidemiology; SNP, single nucleotide polymorphism.

Epidemiologists are still coming to grips with the opportunities and difficulties offered by the burgeoning fields of genomics. Most research in this area is now focusing on the complex diseases that affect populations via a multifactorial genetic and environmental etiology—for example, cancer, heart disease, asthma, and diabetes—rather than disease caused by single genetic variants. Mendelian or near-Mendelian inheritance has been widely studied and is quite well understood (1). In contrast, complex diseases result from many genetic polymorphisms, across the genome, which are now being aggressively explored (2, 3). In recent years, it has become apparent that replication of observations in genomic epidemiology is increasingly difficult to achieve. This brief commentary addresses some of the reasons for this difficulty

Correspondence to Dr. Michael B. Bracken, Center for Perinatal, Pediatric, and Environmental Epidemiology, Yale University School of Medicine, One Church Street, 6th Floor, New Haven, CT 06510 (e-mail: michael.bracken@yale.edu).

and argues that electronic evidence-based systematic reviews of the extant literature are needed to provide the most valid, current evidence of associations observed in genomic epidemiology.

WHY REPLICATION IS INCREASINGLY DIFFICULT IN GENOMIC EPIDEMIOLOGY

It is estimated that, of the 3 billion base pairs in the human genome, only one in 1,200 single nucleotide polymorphisms (SNPs) varies among individuals (4). Polymorphisms can indicate a change at a single base pair or may be thousands of base pairs in size, and they are found as tandem repeats, deletions, or insertions. Haplotypes, a closely linked set of commonly evolved genes, are used to locate polymorphisms that cause disease (5). If a polymorphism occurs with less than 1 percent frequency, this occurrence is typically called a mutation. The technology for examining the structure of the genome is expanding rapidly. Microarray analysis was first presented as a concept in 1994 (6), and the first paper on the technology was published in 1995 (7). Now, the entire human genome can be placed on a single microarray (or gene chip). Microarrays permit the genome of study subjects to be examined for polymorphisms, and new high-density sequencing methods can genotype up to 500,000 SNPs.

The large volumes of genetic data being produced by genome-wide screening require new statistical methodologies that extend beyond traditional hypothesis-driven analyses of one or two candidate genes. Genome-wide screening is being conducted for large numbers of candidate genes—for example, for atopic asthma (8) and myocardial infarction (9)—but also for hypothesis-free genome-wide screening. The hypothesis-free approach must account for the simultaneous effects of multiple alleles (10, 11) while managing the statistical problems inherent in multiple comparisons (12–14). Epidemiologic studies have already taken advantage of this new strategy to identify SNPs associated with age-related macular degeneration (15).

Many polymorphisms have quite small, independent effects (relative risks of <1.5) with complex disease diagnoses, the phenotype, and they exert their effects principally by interacting with other polymorphisms or environmental risk factors (16–18). The effects of a polymorphism on disease causation are often further obscured by complex biologic mechanisms, some only recently discovered, which have evolved to protect or ''buffer'' the genome from environmental change (19), as well as by other epigenetic forces (20–22).

The proportion of individuals carrying a polymorphism who express the expected phenotype (usually by a specified age) varies. Genes with low penetrance pose problems in genomic epidemiology. If penetrance varies across families, estimates of penetrance from families in which a gene was first identified may be higher than in the general population. For example, the *BRCA1* allele causing breast cancer had 85 percent penetrance in the original families studied but 40–60 percent penetrance in the broader population by age 70 years (23).

Gene expression may be influenced by the parental origin of the polymorphism, or imprinting. Therefore, the insulin growth factor-2 gene (*IGF-2*) is active only if derived from the father (24). It is often not known whether parent-of-origin effects are due to imprinting or to placental or breast milk transfer of immune factors. To disentangle parent-of-origin effects, studies must collect DNA from parents, a difficult task for late-onset disease when parents may be deceased.

Polymorphisms having opposite effects on a disease may be present in the same gene. Unless the polymorphism of interest is precisely specified, studies may report different findings for the same gene. For example, in the β_2 adrenoceptor gene ($\beta 2AR$), a *Gly16* mutation increases the risk of nocturnal asthma, but another mutation at *Glu27* protects against bronchial hyperreactivity (25).

DNA sequences are predetermined at conception, but gene function is not. Environmental factors can switch gene functions on or off. Methylation is the most widely studied of these epigenetic processes (22). Epigenetic phenomena reduce specificity of polymorphism-disease associations and lower the power of genomic studies to detect real effects.

A NEED FOR ELECTRONIC EVIDENCE-BASED SYSTEMATIC REVIEWS OF GENOMIC EPIDEMIOLOGY

With so many biologic phenomena reducing the likelihood of the existence of strong single gene-disease associations, along with the enormous increase in the number of genes, SNPs, deletions or insertions, and regions of the genome (haplotypes) being explored, it is perhaps to be expected that genomic epidemiology is experiencing substantial difficulty in replicating research findings and in organizing and synthesizing the large amount of rapidly emerging data. Some specific problems in research synthesis of genomic epidemiology are highlighted below.

Publication bias

The genomic epidemiologic literature is plagued with problems of publication bias, particularly toward selectively publishing positive studies. Systematic reviews permit the formal identification and exploration of publication bias. Colhoun et al. (26) describe 19 case-control studies of angiotensin-converting enzyme gene (*ACE*) *DD* polymorphisms and coronary heart disease that show apparent publication bias in favor of positive studies, and studies with odds ratios of up to 3.0, in a series of small studies compared with the estimate of 1.1 from a much larger database. Ioannidis et al. (27) have reported on several disease areas, showing that initial studies tend to report large risk ratios for specific polymorphisms that are either much smaller or not replicated in later investigations.

Study replication

As in classical epidemiology, replication is fundamental for deciding that an observed association is likely not due to chance (28). In genomic epidemiology, replication should be exact for the alleles studied and as similar as possible in

terms of population and environment. Even exact replication is susceptible to high rates of false-positive results. In one study, simulation of random fluctuations in a whole genome scan with *no* trait-causing loci in 100 sib pairs and parents produced 22 regions significant at the $p = 0.05$ level, of which four remained significant at $p = 0.05$ in the replication (29). The effect of population differences causing heterogeneity can be explored formally within the context of a systematic review.

Subgroup analyses

The problem of interpretation found in subgroup analysis in classical epidemiology (30) applies equally in genomic studies. If an association of a polymorphism with disease is found only in a subgroup—for example, after multiple subgroup analyses of gene, microsatellite markers, or SNPs evaluated by gender, age, or ethnic group—then the association is likely to be spurious unless supported by exact replication. Many reported gene-environment interactions are derived from subgroup analyses, and it is difficult to ascertain whether these analyses were testing a priori hypotheses; however, it seems likely that many were not. Individual patient data meta-analysis (see below) can be a particularly powerful approach to identifying subgroups at differential risk of disease.

Meta-analysis and estimating typical effect sizes

There are two principal approaches to statistically pooling data. Most commonly, summary measures of association from individual studies are weighted by their inverse variance and are analyzed to derive a ''typical'' estimate of risk by using a Mantel-Haenszel procedure (31). An example was published for the *IL-10 1082 G/G* genotype, showing increased risk for recurrent pregnancy loss (32). Exploration of effect modification from covariates may be possible with meta-regression techniques (33).

The second pooling method uses individual subject data from studies being analyzed (34). This method is preferred because it allows some control of confounding factors in the reanalysis as well as subgroup analysis, but it requires the collaboration of many investigators and their willingness to share data. Ioannidis et al. (35) used this method to summarize the protective effects of *CCR5-Δ32*, an allele found in 15 percent of Caucasians, in slowing down disease progression in individuals infected with human immunodeficiency virus.

Limitations of current databases

Several large electronic databases collect or ''bank'' genomic data for research purposes. The Environmental Genome Project is targeting SNPs in 200 ''environmentally responsive'' genes (36); the GenBank database, the National Institutes of Health's genetic sequence database, and part of the International Nucleotide Sequence Database Collaboration (37) are important repositories of SNP-level information. ALFRED is a useful resource documenting the prevalence of polymorphisms (38). Other databases describing polymorphism prevalence include the Centers for Disease Control and Prevention's Genotype Prevalence Database (39), Allele Frequencies in Worldwide Populations (40), and the National Cancer Institute's SNP500Cancer database (41). The International HapMap Project is developing a map of haplotypes with a prevalence of more than 5 percent in the human genome, based on genotyping of 1 million SNPs in 270 individuals (42). These gene banks are important for furthering research into disease associations with candidate polymorphisms, but they do not claim to be and are not a substitute for systematic reviews of epidemiologic studies of gene-disease associations.

Systematic reviews

Evidence-based medicine provides a paradigm for explicitly and systematically searching, collating, and synthesizing a complete body of evidence on a research topic (43). Systematic reviews include publication of detailed and transparent literature-searching methods, searching for possible bias in that evidence base, evaluation of study validity and heterogeneity, and consideration of the importance of effect size and precision. They have been widely adopted in clinical research as the ''gold standard'' for synthesizing and drawing conclusions from an extant body of research evidence (44). All scientific literature is amenable to systematic review, a process that does not necessarily require meta-analysis. Many systematic reviews identify a body of data not amenable to meta-analysis because of heterogeneity in study methodology or the populations studied.

In genomic epidemiology, there has been a concerted effort by the Human Genome Epidemiology (HuGE) review group, organized by the Centers for Disease Control and Prevention (45, 46), to conduct systematic reviews of gene-disease associations. Explicit criteria are being developed to assess the validity of genomic epidemiology publications (47). These criteria include recommendations for how studies may be scored for validity, data integration for calculating typical effect estimators, and reviews constructed according to common genotypes and genotyping methodology (48). A recent example, in which individual subject data were used, considered polymorphisms in the alcohol dehydrogenase gene (*ADH*) and the aldehyde dehydrogenase gene (*ALDH*) associated with the risk of head and neck cancer (49).

HuGE reviews do not always include details of analyses to identify possible publication bias, exact accounts of literature search strategies, descriptions of ''excluded'' studies, or details of actual assessments used to judge individual study validity. Importantly, HuGE reviews are not updated routinely and frequently, all of which are standard features of a modern, electronically based, systematic review. Although electronic files of HuGE publications from several journals, including the *American Journal of Epidemiology*, are accessible on the HuGE website, the reviews are not published online initially, a process that would speed up publication, lead to more uniformity in the reports, and ease updating. Furthermore, electronic publication would allow standard statistical programs for meta-analysis to be embedded within the software for creating the systematic review.

Importantly, online publication would permit access to a full protocol describing the scope and objectives of the planned review. At present, protocols are listed simply as titles on the HuGE website. Protocol publication explicitly documents the planned objectives of the review, including subgroups to be considered in any analysis and other criteria subject to bias while the review is being constructed. The importance of publishing protocols to avoid ex post facto manipulation of primary outcome selection has recently been demonstrated in the clinical literature (50).

Examples of electronic databases for online systematic reviews from medicine are already well established—notably, the Cochrane Library and its REVMAN and METAVIEW software for creating systematic reviews (51)—and the social sciences (Campbell Collaboration (48)). All review groups within these collaborations require frequent updating of their reviews. While the Cochrane Collaboration is currently focused on reviewing randomized clinical trials, efforts are under way to expand this focus to observational studies of similar design to genomic epidemiology. The HuGE research network, with its developed consensus guidelines (47), appears best positioned to move to full electronic creation and publication of systematic reviews of genomic epidemiology as exemplified by the Cochrane Collaboration.

SUMMARY

Candidate genes for complex disease are being discovered at an increasing rate, but replication of observed associations is proving difficult. Epigenetic phenomena are also increasingly being discovered and point to the difficulty in conducting genomic epidemiology that can identify strong gene-disease associations. Genome-wide hypothesis-free mapping will increasingly generate candidate genes that require further testing in association studies. An electronic evidence-based approach to systematically reviewing and synthesizing this rapidly emerging body of genomic epidemiology research, with the capacity for continuous updating, is essential for identifying valid and replicable gene-disease associations.

ACKNOWLEDGMENTS

This work was supported by grants AI 41040 and DA 05484 from the National Institutes of Health.

The author is grateful to Geir Jacobsen, Josephine Hoh, and several anonymous reviewers for their comments on earlier drafts of this paper.

Michael Bracken is an editor of the Cochrane Neonatal Review Group and a member of the Cochrane Methods Group.

REFERENCES

1. Risch NJ. Searching for genetic determinants in the new millennium. Nat Rev Genet 2000;405:847–56.
2. Bell J. Predicting disease using genomics. Nature 2004;429:453–6.
3. Potter JD. At the interfaces of epidemiology, genetics and genomics. Nature 2001;2:142–7.
4. Kendal WS. An exponential dispersion model for the distribution of human single nucleotide polymorphisms. Mol Biol Evol 2003;20:579–90.
5. Johnson GCL, Esposito L, Barratt BJ, et al. Haplotype tagging for the identification of common disease genes. Nat Genet 2001;29:233–7.
6. Schena M. Microarrays as toxin sensors. Pharmacogenomics J 2003;3:125–7.
7. Schena M, Shalon D, Davis RW, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467–70.
8. Kurz T, Altmueller J, Strauch K, et al. A genome-wide screen on the genetics of atopy in a multiethnic European population reveals a major atopy locus on chromosome 3q21.3. Allergy 2005;60:192–9.
9. Ozaki K, Ohnishi Y, Iida A, et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet 2002;32:650–4.
10. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005;6:95–108.
11. Wang WYS, Barratt BJ, Clayton DG, et al. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 2005;6:109–18.
12. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 2003;4:701–9.
13. Hoh J, Ott J. Genetic dissection of diseases: design and methods. Curr Opin Genet Dev 2004;14:229–32.
14. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003;19:368–75.
15. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385–9.
16. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001;358:1356–60.
17. Carlson CS, Eberle MA, Kruglyak L, et al. Mapping complex disease loci in whole-genome association studies. Nature 2004;429:446–52.
18. Khoury MJ, Millikan R, Little J, et al. The emergence of epidemiology in the genomics age. Int J Epidemiol 2004;33:936–44.
19. Newman SA, Muller GB. Epigenetic mechanisms of character origination. J Exp Zool 2000;288:304–17.
20. Jablonka E. Epigenetic epidemiology. Int J Epidemiol 2004;33:929–35.
21. Egger G, Liang G, Aparicio A, et al. Epigenetics in human disease and prospects for epigenetic therapy. Nature 2004;429:457–63.
22. Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer 2004;4:143–53.
23. Begg CB. On the use of familial aggregation in population-based case probands for calculating penetrance. J Natl Cancer Inst 2002;94:1221–6.
24. Szabo P, Tang SH, Rentsendorj A, et al. Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function. Curr Biol 2000;10:607–10.
25. Liggett SB. Polymorphisms of the β_2-adrenergic receptor. Am J Respir Crit Care Med 1997;156:S156–62.

26. Colhoun HM, McKeigue PM, Smith GD. Problems of reporting genetic associations with complex outcomes. Lancet 2003;361:865–72.
27. Ioannidis JPA, Ntzani EE, Trikalinos TA, et al. Replication validity of genetic association studies. Nat Genet 2001;29:306–9.
28. Lalouel JM, Rohrwasser A. Power and replication in case-control studies. Am J Hypertens 2002;15:201–5.
29. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 1995;11:241–7.
30. Counsell CE, Clarke MJ, Slattery J, et al. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? BMJ 1994;309:1677–81.
31. Bracken MB. Statistical methods for analysis of effects of treatment in overviews of randomized trials. In: Sinclair JC, Bracken MB, eds. Effective care of the newborn infant. New York, NY: Oxford University Press, 1992.
32. Daher S, Shulzhenko N, Morgun A, et al. Associations between cytokine gene polymorphisms and recurrent pregnancy loss. J Reprod Immunol 2003;58:69–77.
33. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002;21:1559–73.
34. Clarke MJ, Stewart LA. Obtaining individual patient data from randomized clinical trials. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care: meta-analysis in context. London, United Kingdom: BMJ Publishing Group, 2001.
35. Ioannidis JP, Rosenberg PS, Goedert JJ, et al. Effects of CCR5-Delta32, CCR2-64I, and SDF-1 3′A alleles on HIV-1 disease progression: an international meta-analysis of individual-patient data. Ann Intern Med 2001;135:782–95.
36. Environmental Genome Project: polymorphisms and genes (www.niehs.nih.gov/envgenom/snpsdb.htm). Accessed April 27, 2005.
37. GenBank database (www.psc.edu/general/software/packages/genbank/genbank.html). Accessed April 27, 2005.
38. ALFRED: The ALlele FREquency Database. A resource of gene frequency data on human populations supported by the US National Science Foundation (http://alfred.med.yale.edu/alfred/index.asp). Accessed April 20, 2005.
39. Genomics and disease prevention: Genotype Prevalence Database (www.cdc.gov/genomics/search/aboutGTP.htm). Accessed April 27, 2005.
40. Allele frequencies in worldwide populations (www.allelefrequencies.net/). Accessed April 27, 2005.
41. National Cancer Institute SNP500Cancer database (www.snp500cancer.nci.nih.gov). Accessed April 27, 2005.
42. International HapMap Project (www.hapmap.org/). Accessed July 10, 2004.
43. Sackett DL, Straus SE, Richardson SW, et al. Evidence-based medicine: how to practice and teach EMB. London, United Kingdom: Harcourt Publishers Limited, 2000.
44. Egger M, Smith GD, Altman DG. Systematic reviews in health care: meta-analysis in context. London, United Kingdom: BMJ Publishing Group, 2001.
45. Centers for Disease Control, Human Genome Epidemiology Network (http://www.cdc.gov/genomics/hugenet/). Accessed April 20, 2005.
46. Little J, Khoury MJ, Bradley L, et al. The Human Genome Project is complete. How do we develop a handle for the pump? Am J Epidemiol 2003;157:667–73.
47. Little J, Bradley L, Bray MS, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. Am J Epidemiol 2002;156:300–10.
48. The Campbell Collaboration (http://www.campbellcollaboration.org/). 2004.
49. Brennan P, Lewis S, Hashibe M, et al. Pooled analysis of alcohol dehydrogenase genotypes and head and neck cancer: a HuGE review. Am J Epidemiol 2004;159:1–16.
50. Chan AW, Hróbjartsson A, Haahr MT, et al. Empirical evidence for selective reporting of outcomes in randomized trials. Comparison of protocols to published articles. JAMA 2004;291:2457–65.
51. The Cochrane Library, Update Software Ltd (http://www.update-software.com/cochrane/). 2004.