American Journal of Epidemiology
© The Author 2008. Published by the Johns Hopkins Bloomberg School of Public Health.
All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

Vol. 169, No. 2
DOI: 10.1093/aje/kwn351
Advance Access publication November 20, 2008

Invited Commentary

Invited Commentary: From Genome-Wide Association Studies to Gene-Environment-Wide Interaction Studies—Challenges and Opportunities

Muin J. Khoury and Sholom Wacholder

Initially submitted May 30, 2008; accepted for publication July 25, 2008.

The recent success of genome-wide association studies in finding susceptibility genes for many common diseases presents tremendous opportunities for epidemiologic studies of environmental risk factors. Analysis of gene-environment interactions, included in only a small fraction of epidemiologic studies until now, will begin to accelerate as investigators integrate analyses of genome-wide variation and environmental factors. Nevertheless, considerable methodological challenges are involved in the design and analysis of gene-environment interaction studies. The authors review these issues in the context of evolving methods for assessing interactions and discuss how the current agnostic approach to interrogating the human genome for genetic risk factors could be extended into a similar approach to gene-environment-wide interaction studies of disease occurrence in human populations.

environment; epidemiologic methods; genetics; genomics

Abbreviations: GEWIS, gene-environment-wide interaction studies; GWAS, genome-wide association studies; HuGE, human genome epidemiology.

> The breakthrough of this year has to do with humans, genomes, and genetics. But it is not about THE human genome (as if there were only one!). Instead, it is about your particular genome, or mine, and what it can tell us about our backgrounds and the quality of our futures. (1)

For human genomics research, 2007 was a banner year. Using genome-wide analytic platforms that can measure hundreds of thousands of genetic variants simultaneously, more than 100 epidemiologic studies have uncovered genetic risk factors for a wide variety of common, complex diseases (2). Human geneticists are anxious to reap the benefits of the human genome project (3) and the international HapMap project (4) by integrating genomics into health care and disease prevention. Genome-wide association studies (GWAS) have shown that an ''agnostic'' approach can interrogate the totality of the human genome and identify genetic variants associated with numerous diseases.

Certainly, the functional and clinical implications of the loci detected by using GWAS are far from clear, just as some of the rare, high-penetrance genetic variants for breast cancer, such as *BRCA1* and *BRCA2*. Biologic studies that assess the role of these variants in disease processes and risk factors may give epidemiologists clues about environmental exposures likely to be involved in human diseases (5, 6). So far, however, the odds ratios of individual genetic variants detected are small, mostly between 0.67 and 1.5 (2), and may not be useful for clinical prediction (7, 8).

A NEW ERA OF GENE-ENVIRONMENT-WIDE INTERACTION STUDIES (GEWIS)

The availability of GWAS and their rapidly declining prices, together with the emergence of collaborative epidemiologic consortia and networks (9, 10), offer major opportunities to epidemiologic researchers focused on effects of the environment, broadly defined to include behavioral, chemical, physical, and social factors (11, 12). The increasing rate of published studies focusing on gene-environment interactions pales against the exploding acceleration in published reports of

Correspondence to Dr. Muin J. Khoury, National Office of Public Health Genomics, Centers for Disease Control and Prevention, 4770 Buford Highway, MS K-89, Atlanta, GA 30341 (e-mail: muk1@cdc.gov).

Table 1. Trends in Published HuGE Articles, GWAS, and Studies Reporting on GEI, 2001–2007[a]

Year	Total HuGE Articles, no.	GWAS		GEI	
		No.	%	No.	%
2001	2,488	0	0	373	15.4
2002	3,196	0	0	444	13.9
2003	3,474	3	0.1	447	12.9
2004	4,279	0	0	518	12.1
2005	5,028	5	0.1	706	14.0
2006	5,357	12	0.2	727	13.6
2007	7,168	105	1.5	1,016	14.2

Abbreviations: GEI, gene-environment interactions; GWAS, genome-wide association studies; HuGE, human genome epidemiology.

[a] Data were derived from the HuGE Navigator (13), searched online July 10, 2008 at http://www.hugenavigator.net/.

genetic association studies. Table 1 shows the trends in published genetic association articles from 2001 to 2007, as captured by the HuGE Navigator (13), an online curated and searchable knowledge base in human genome epidemiology (HuGE), sponsored by the Human Genome Epidemiology Network (HuGENet (14)). Between 2001 and 2007, the number of total HuGE articles almost tripled and the number of reported GWAS articles rose from 0 to more than 150; the number of articles reporting on gene-environment interaction also increased, but the proportion of such articles in the total HuGE literature remained relatively flat at about 14%. We do note that undoubtedly a substantial fraction of nonsignificant tests of gene-environment interaction are unreported, leading to a distortion of the literature and too much attention to the positive reports (15). This possibility, however, is unlikely to affect the literature trends unless publication bias varies with time.

The paucity of established gene-environment interactions to date (refer to García-Closas et al. (16) for a rare exception) in the face of substantial investment in the effort should not overly discourage epidemiologists. The ''candidate-gene'' approach to studies of genetic factors failed to find and replicate many associations, probably because genetic epidemiologists overestimated their ability to select the best candidates and because the threshold for claiming an association was too low given the low prior probability for even the best candidates (17). But just as the GWAS approach, with its broad interrogation of the genome and rigorous threshold for calling effects significant, identified common variants associated with common disease, so too are GWAS likely to help identify genetic factors that interact with environmental factors.

We have known for decades that failure to incorporate both genetic and environmental factors in a joint analysis will weaken the observed associations between a true risk factor and disease occurrence. Because the pools of susceptible and nonsusceptible persons are mixed, the observed associations tend to be shifted toward the null (18). Theoretically, if we are able to measure gene-environment interactions, we should sharpen our measurements of effects in subsets of the population and even potentially increase our statistical power in measuring such effects (19).

OBSTACLES IN ASSESSING GENE-ENVIRONMENT INTERACTION

There are obstacles on the ''environment'' side of gene-environment interaction that are not present on the ''gene'' side. Environmental epidemiology does not have the economy of scale seen in genomics, where the difference in cost between measuring a million variants and one variant is a small fraction of the average cost per participant in a case-control or cohort study that collects DNA. We may be missing important environmental determinants of disease because we do not know what to look for or because we do not know how or when to measure accurately what we do know to seek. A person's genetic makeup may be too far removed from complex physiologic or biochemical processes that could be more important risk factors for disease. Germ-line variation is static and so can be captured at any point, but variation in the timing of exposure, and the timing of subsequent risk, complicates study of environmental factors; at the same time, variation in exposure and risk over time can provide important clues about etiology. In addition, the major advances in the use of biomarkers in research and medical applications, most notably for infectious diseases, are not yet close to yielding useful measures of long-term exposure regarding diet, pharmaceuticals, and polluted air and water for the large numbers of persons needed for studies of rare diseases. Even as biomarkers continue to improve measurement of some exposures, we must also improve the accuracy of epidemiologic questionnaires, medical records, occupational records, and other proxy measurements of environmental factors.

Investigation of gene-environment interaction to learn about etiology and public health is feasible with existing data. An agnostic strategy that is implemented carelessly, however, will generate a large supply of false-positive findings and cause well-founded skepticism about claims of interactions, given the low prior probabilities of most hypotheses (15). Researchers conducting GWAS are demanding replications and requiring *P* values for significance below what we have ever thought realistic in epidemiology (20) in order to avoid false-positive findings in studying main effects of a million genetic variants. Imagine 10–30 times more tests of interaction involving genes, demographic factors, and personal and environmental exposures. Hypotheses about interaction have lower prior probabilities and tests have lower power for detecting interactions compared with tests for main effects with comparable effect size. In addition, exposures are measured with more significant misclassification than genetic variants are. Huge sample sizes are required to reach the very low *P* values for GWAS of main effect that are finding small effects. How will we decide on and achieve the enormous sample sizes needed for interactions when there are more hypotheses and lower prior probabilities of effect, and when good exposure assessment will be critical? How will we be able to distinguish and draw attention to the few interactions likely to be real from the myriad of false-positive ones?

The decades-old problem of defining interaction (21, 22) is even more prominent in the GWAS era. The statistical models we have used to declare interaction as departure from additive or multiplicative joint effects may be inadequate to describe the underlying biology of joint gene-environment effects on complex disease. The flood of new empirical data becoming available may allow us to examine both gene-gene and gene-environment interactions in new ways.

Systems biology provides novel experimental approaches to quantify molecular components of a biologic system, to assess their interactions, and to integrate such information into graphic models that may explain or predict emergent phenomena (23). However, there is still a large schism between modeling of interactions in cellular and biologic processes and our ability to use that information in observing health and disease in human populations. How can we use biologic information for defining interaction or choosing which analytic method is most useful for identifying risk factors, genetic or environmental; for describing their joint effects; and for predicting and stratifying risk? Do we look for higher-order effects only when a main genetic effect has been found? Do we try to fit a variety of models of interactions, including additive and multiplicative effects? Do we remain truly agnostic in our approach and let the data speak for themselves by using other approaches such as data mining techniques (24)? Do we continue using the multiplicative model to remove one dimension of complexity (25)? We need some analytic help to make the GEWIS efforts more productive by addressing biologic, clinical, and public health questions, not only academic abstractions!

EMERGING METHODS FOR ANALYSIS OF GEWIS

In this issue of the *Journal*, Murcray et al. (26) present a 2-step approach to evaluation of multiplicative gene-environment interaction in the context of a GEWIS. In another accompanying commentary, Chatterjee and Wacholder (27) discuss the strengths and limitations of this approach and compare it with a recently proposed (28) "1-stage" approach to gene-environment interaction. While analytic approaches to genomic data and gene-environment interaction will continue to evolve (refer to Chen et al. (29), Schwender and Ikstadt (30), Musani et al. (31), and Kraft et al. (32) for other examples), integrating analysis of genetic and environmental factors into a coherent biologic framework will be a huge challenge for epidemiology in the 21st century. Strangely enough, the current GWAS approach that took us away from biology and more toward the much-maligned "fishing expedition" in epidemiology provides more evidence about how much remains to be learned about the etiology of complex diseases. The traditional analytic tools that we have used in epidemiology have been strained by GWAS and will have to be further developed as we move from GWAS to GEWIS in the coming decades.

ACKNOWLEDGMENTS

Author affiliations: National Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia (Muin J. Khoury); and Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland (Sholom Wacholder).

The findings in this paper reflect the viewpoints of the authors and do not necessarily reflect the views of the Department of Health and Human Services.

This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

Conflict of interest: none declared.

REFERENCES

1. Kennedy D. Breakthrough of the year [editorial]. *Science*. 2007;318(5858):1833.
2. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest*. 2008;118(5):1590–1605.
3. Collins FS, Green E, Guttmacher AE, et al. A vision for the future of genomics research. *Nature*. 2003;422(6934): 835–847.
4. International HapMap Consortium, Frazer KA, Ballinger DA, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449(7164):851–861.
5. Rothman N, Wacholder S, Caporaso NE, et al. The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochim Biophys Acta*. 2000;1471(2):C1–C10.
6. Davey Smith G, Ebrahim S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32(1): 1–22.
7. Pharoah PD, Antoniou AC, Easton DF, et al. Polygenes, risk prediction and targeted prevention of breast cancer. *N Engl J Med*. 2008;358(26):2796–2803.
8. Hunter DJ, Altshuler D, Rader DJ. Focus on research: from Darwin's finches to canaries in the coal mine—mining the genome for new biology. *N Engl J Med*. 2008;358(26): 2760–2763.
9. Kraft P, Hunter D. Integrating epidemiology and genetic associations: the challenge of gene-environment interaction. *Philos Trans R Soc Lond B Biol Sci*. 2005;360(1460): 1609–1616.
10. Seminara D, Khoury MJ, O'Brien TR, et al. The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology*. 2007;18(1):1–8.
11. Vineis P. Methodological approaches to gene-environment interactions in occupational epidemiology [electronic article]. *Occup Environ Med*. 2007;64:e3. (http://oem.bmj.com/cgi/content/extract/64/12/e3).
12. Schwartz DA. The importance of gene-environment interactions and exposure assessment in understanding human diseases. *J Expo Sci Environ Epidemiol*. 2006;16(6):474–476.
13. Yu W, Gwinn M, Clyne M, et al. A navigator for human genome epidemiology. *Nat Genet*. 2008;40(2):124–125.
14. National Office of Public Health Genomics, Centers for Disease Control and Prevention. The Human Genome Epidemiology Network (HuGENet). (http://www.cdc.gov/genomics/hugenet/default.htm). (Accessed May 8, 2008).
15. 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(4):300–310.

16. García-Closas M, Malats N, Silverman D, et al. NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet*. 2005;366(9486):649–659.
17. Wacholder S, Chanock S, García-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*. 2004; 96(6):434–442.
18. Khoury MJ, Adams MJ, Flanders WD. An epidemiologic approach to ecogenetics. *Am J Hum Genet.* 1988;42(1): 89–95.
19. Khoury MJ, Beaty TH, Hwang SJ. Detection of genotype-environment interaction in case-control studies of birth defects: how big a sample size? *Teratology.* 1995;51(5): 336–343.
20. Hunter DJ, Kraft P. Drinking from the fire hose. Statistical issues in genomewide association studies. *N Engl J Med.* 2007;357(5):436–439.
21. Hunter DJ. Gene-environment interactions in human disease. *Nat Rev Genet.* 2005;6(4):287–298.
22. Greenland S, Lash TL, Rothman KJ. Concepts of interaction. In: Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology.* 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:71–83.
23. Hood L, Heath JR, Phelps ME, et al. Systems biology and new technologies enable predictive and preventative medicine. *Science*. 2004;306(5696):640–643.
24. Onkamo P, Toivonen H. A survey of data mining methods for linkage disequilibrium mapping. *Hum Genomics*. 2006; 2(5):336–340.
25. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet.* 2005;37(4):413–417.
26. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol*. 2009;169(2):219–226.
27. Chatterjee N, Wacholder S. Invited commentary: efficient testing of gene-environment interaction. *Am J Epidemiol*. 2009;169(2):231–233.
28. Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*. 2008;64(3):685–694.
29. Chen X, Liu CT, Zhang M, et al. A forest-based approach to identifying gene and gene-gene interactions. *Proc Natl Acad Sci U S A*. 2007;104(49):19199–198203.
30. Schwender H, Ikstadt K. Identification of SNP interaction using logic regression. *Biostatistics*. 2008;9:187–198.
31. Musani SK, Shriner D, Liu N, et al. Detection of gene × gene interactions in genome wide association studies in human populations. *Hum Hered*. 2007;63(2):67–84.
32. Kraft P, Yen YC, Stram DO, et al. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*. 2007; 63(2):111–119.