Editorial

Are We Ready for Genome-wide Association Studies?

Duncan C. Thomas

University of Southern California, Los Angeles, California

The tension between hypothesis-driven and exploratory research crosses scientific disciplines (1) but is particularly well illustrated in the current excitement about genome-wide association (GWA) studies. Standard linkage analysis has the potential to localize major susceptibility genes to within a few million base pairs using as few as 300 microsatellite markers for a genome-wide scan. However, it has increasingly been recognized that linkage analysis may not be powerful enough to detect genes involved in ''complex diseases'' like cancer, which are caused by multiple genes and multiple environmental factors, interacting in complicated ways. Thus, molecular epidemiologists are turning to candidate-gene association studies or studies of entire candidate pathways, driven by specific biological hypotheses, as an alternative approach. Now comes the prospect, first seriously proposed a decade ago by Risch and Merikangas (2), of testing virtually all ~10 million common single nucleotide polymorphisms (SNP) in the human genome for associations with a given disease, either directly or by linkage disequilibrium with other SNPs. Recent developments in ultra-high-volume genotyping chip technology now make the study of as many as 500,000 SNPs commercially viable. Coupled with the extensive haplotype tagging SNP information being catalogued by the HapMap project (3), it now seems that that this density may be sufficient to permit indirect tests of association with the majority of all common SNPs (4), although this fundamental assumption has recently been questioned (5). Does this development mark the end of pathway-driven research? I suggest that it is possible to marry the hypothesis-driven and exploratory approaches in a way that will make better use of this novel but expensive technology.

Background

The first GWA study was published in 2002 (6), using an early 100 K version of the technology. Within the last year, several others have been published (7-10), and many other studies have been launched or proposed. Time will tell whether these early reports represent true positives, but the simultaneous publication in *Science* of two confirmatory studies (11, 12) for the Complement Factor H association with age-related macular degeneration along with the GWA scan has sparked great enthusiasm, and several reports have seemed subsequently confirming the association. However, this finding remains to be confirmed in population-based studies, which may be less vulnerable to selection bias and less weighed towards advanced cases.

Is the time really ripe for wholesale adoption of the GWA approach? The cost of the genotyping technology is bound to keep falling over the next few years, and many study design and analysis issues remain to be resolved. Yet, investigators may feel that if they do not hop on this bandwagon before everyone else does, they will be left out. But there simply are not enough resources to fund all eager investigators, even all those with already established and well-characterized cohorts or case-control samples for which the major expense would be the genotyping costs.

So how many GWA studies can the scientific community afford and how should they be prioritized relative to hypothesis-driven studies? By any standard, these will be expensive. Supposing a typical study might require at least 1,000 cases and 1,000 controls and at current genotyping costs of approximately US$1,000 per sample for a 500 K chip, the genotyping alone will cost more than US$2 M per study. DNA pooling offers the potential to dramatically reduce the genotyping cost, but substantial technical difficulties remain, and the power and false-positive rates of the approach relative to individual genotyping remain uncertain (13-17).

Are multiple GWA studies really needed? Replication of entire scans is not a good use of limited resources, except perhaps for protection against false negatives; thus, there is little need for multiple studies of the same condition, but there will be many investigators well poised to propose such studies for any given disease. Competition for the best proposal(s) is a sensible approach, but rather than using standing disease-oriented study sections, in which GWA proposals would have to compete against lower-cost, hypothesis-driven proposals for the same disease, GWA proposals for a broad range of diseases should be evaluated against each other. At this early stage of the methodology, this would help prioritize the available funds among the diseases for which GWA studies are likely to be the most informative and would also help refine the methods by ensuring some uniformity of standards against which they would be judged.

Recent Initiatives

It is encouraging that several recent NIH initiatives have taken steps in this direction, in addition to European initiatives, such as the Welcome Trust Case-Control Consortium.[1] In April 2005, several Institutes pooled resources and allocated US$5.4 M to establish a cooperative group to investigate appropriate methods for design and analysis of such studies.[2] In February 2006, the National Heart, Lung and Blood Institute released a Request for Applications[3] committing US$20 M for four to six GWA studies across a range of conditions. The Center for Inherited Disease Research began receiving applications in March 2006 for genome-wide SNP genotyping[4], along the lines of their long-standing microsatellite genotyping service for linkage studies. Illumina 100 K, 300 K, or 500 K panels will be offered followed by custom genotyping up to 24 K SNPs on the second sample. On February 8, 2006, NIH announced two major initiatives aimed at providing the first stage genotyping for about 20 large case-control studies. The first of these,

Cancer Epidemiol Biomarkers Prev 2006;15(4):595–8

Requests for reprints: Duncan C. Thomas, University of Southern California 1975 Zonal Avenue KAM-110 CHP 220, 9011 HSC Los Angeles, California 90089-9023. E-mail: dthomas@usc.edu

Copyright © 2006 American Association for Cancer Research.

doi:10.1158/1055-9965.EPI-06-0146

[1] http://www.wtccc.org.uk/
[2] http://grants2.nih.gov/grants/guide/rfa-files/RFA-HL-05-011.html
[3] http://grants.nih.gov/grants/guide/rfa-files/RFA-HL-06-012.html
[4] http://www.cidr.jhmi.edu/human_gwa.html

Downloaded from http://aacrjournals.org/cebp/article-pdf/15/4/595/2264614/595.pdf by guest on 07 August 2022

known as the Genetic Association Information Network (GAIN)[5], will be funded by the private sector through the public-private partnership of the Foundation for the NIH and is expected to support at least seven studies, initially using Perlegen and Affymetrix platforms. The second initiative, known as the Genes and Environment Initiative (GEI)[6] will be publicly funded, if Congress approves the request, and will support about another dozen studies for each of 4 years. A formal Request for Applications for the GEI initiative is expected at a later date, once funding has been committed.

Both the GAIN and GEI initiatives plan to select studies by a rigorous peer review process (separately for each initiative). Applications for the GAIN initiative will be due in late April 2006, and genotyping may be under way as early as late summer 2006, a remarkably short interval compared with traditional grant cycles. In preparation for these various initiatives, Institutes have surveyed existing studies that may qualify and have considered their own priorities. But investigators who have not been privy to the planning process may find it difficult to meet the short deadline planned for at least the GAIN initiative, and it will be a challenge to organize an effective peer review process on such a short timeline with the uncertainties of several overlapping initiatives. Meanwhile, the funds available and priority given to various proposals that are currently under review through the traditional R01 mechanism could be affected by these new initiatives.

It is premature to issue rigid guidelines for the conduct of GWA studies, given the infancy of the field and the rapid evolution of the technology, design, and analysis methods, although potential applicants would benefit from some guidelines for review of proposals. In this spirit, the CIDR Access Committee has developed a set of criteria for investigators planning applications for their SNP genotyping service. The draft recommendations arose from a roundtable discussion at the October 2005 annual meeting of the International Genetic Epidemiology Society, chaired by Dan Schaid, with input from Leonid Kruglyak and David Clayton, among others. The draft criteria include such considerations as the potential societal benefit, the evidence in support of a genetic basis, plans for replication, data sharing, and localization of suggested associations, as well as a host of methodologic issues discussed below.

Methodologic Challenges

Several recent reviews (18-23) have discussed a range of methodologic challenges in the design and analysis of GWA studies in detail. Some of the issues that require careful consideration include DNA pooling versus individual genotyping; choice of genotyping platform and selection of panel of SNPs; use of multistage designs; control of, and allowance for, genotyping errors; population-based versus family-based designs and adjustment for population stratification; and multiple comparisons and criteria for claiming statistical significance.

Sample Size and Power. To get a sense of the magnitude of the task, it is helpful to consider some rough sample size requirements. Suppose one planned to test 500,000 single-SNP associations in a single-stage case-control study and wished to control the genome-wide type I error rate at $\alpha = 5\%$ (i.e., an expected number of false positives across the entire genome of only 0.05), so that any statistically significant associations would be very likely to be true positives. A conservative Bonferroni correction would require a single-SNP significance level of $\alpha = 0.05/500{,}000 = 1 \times 10^{-7}$. Table 1 indicates the numbers of case-control pairs that would be required to attain 95% power for a range of genetic relative risks and population allele frequencies. These numbers might be reduced by about a factor of two by using a multistage design, in which only the first sample would be tested on the complete panel, with subsequent samples tested on only a subset of the most significant markers (24-26). On the other hand, testing multiple genetic models, additional SNPs or haplotypes, subgroups, or interactions would require an even stricter significance level and larger sample sizes. Thus, these sample size requirements should be taken as only rough guidelines. Methods for significance testing in GWA studies are the subject of an important research area (27-35).

Table 1. Numbers of case-control pairs required to attain a significance level of $\alpha = 1 \times 10^{-7}$ with 95% power in a single-stage study, assuming a multiplicative genetic model with the indicated relative risks per allele (i.e., homozygote relative risk being the square of the indicated relative risks)

Relative risk	MAF = 5%	MAF = 10%	MAF = 20%
1.2	28,000	15,000	8,700
1.5	5,200	2,800	1,700
2.0	1,600	870	540
2.5	830	470	300
3.0	540	310	200

Abbreviation: MAF, minor allele frequency.

Replication. It is not clear that GWA studies should be held to the same standards of replicability as candidate gene studies. The huge cost of GWA studies, and thus the likelihood that few of them will ever be undertaken for any given disease, demands that they should be designed to have very high power, even if at the cost of relatively high false positive rates. If we are going to the bother of looking for the proverbial needle in a haystack, we want to be assured that we have a good chance of finding it, if it is really there, because we are not going to have the energy, or money, to do it all over again! Among such a large number of tests, there is no guarantee that the expected modest number of true positives will rank near the top of the list (36). For example, in the Ozaki et al. (6), GWA scan of 65,671 SNPs, the one functional association was less significant than 200 spurious associations that failed to replicate. To find a true association of the observed size (relative risk = 1.6) with at least 50% power, one would have to follow up >3,400 of the most significant associations! On the other hand, because the yield of positive associations could also be very large, the cost and manpower needed to follow up on each will be considerable. Should few of these associations be replicated, the societal investment in this high-risk experiment could quickly turn sour.

Multistage sampling designs can be thought of as a form of built-in replication, but if the same epidemiologic study design and population is used in the different stages, this is really just a form of *statistical* replication (albeit a more efficient one than conducting two or more separate studies; ref. 37). True *scientific* replication involves different investigators, studying different populations, using different study designs, with potentially different strengths and weaknesses. In a recent editorial, *Nature Genetics* has made this a formal requirement for publication: ''Because meta-analysis has shown that many published associations could not be replicated, we now stipulate that the association should be observed in two independent cohorts'' (38).

An example of such an article is the recent publication of four variants associated with myocardial infarction, found to be significant in two separate case-control samples of pooled DNA, followed by replication in a third independent case-control comparison using individual genotyping (9). Replication has also been addressed in an editorial in this

[5] http://www.fnih.org/GAIN/GAIN_home.shtml
[6] http://www.genome.gov/17516707

Downloaded from http://aacrjournals.org/cebp/article-pdf/15/4/595/2264614/595.pdf by guest on 07 August 2022

journal (39), albeit in the context of candidate gene rather than GWA studies, and sparked an extensive series of follow-up commentaries (40-45). Without belaboring these points, suffice it to say that true scientific replication will be essential to make sense of the mass of associations likely to result from GWA studies, but we are hesitant to follow the lead of *Nature Genetics* by making rigid requirements for replication within a single report, particularly if attaining that goal were to come at the expense of multiple under-powered studies (45).

Data Sharing. Perhaps the best way of ensuring scientific replication is to embrace a culture of data sharing, to facilitate investigators' pursuit of alternative analyses of each others' data: generating hypotheses they can attempt to replicate in their own data or replicating their own findings in independent data. The recently announced GAIN initiative is to be commended for embracing these principles: all genotype data will be made public as soon as they are generated and checked for quality, similar to the rules for the Human Genome and HapMap projects. The phenotype data will be made available as soon as the genotype data are, through an NIH access committee. The contributing investigators will have a 9-month window for exclusive right to publish but will receive the data at the same time as anyone else.

The sheer scale of GWA data sets, combined with the usual need to protect confidentiality, will pose formidable practical challenges to sharing data in a manner that will be genuinely useful. Although the GAIN plan is to post-raw genomic data on individual subjects, standards could be developed for highly detailed aggregate data tables for studies where the informed consent would preclude public posting of individual data. As a minimum, these might include summary statistics for all single-SNP associations and details of the methods used to obtain the statistical results, but more thought is needed about whether this should include haplotype associations, gene-gene or gene-environment associations, or various subgroup analyses. Rather than routinely posting such an enormous number of associations one might consider making the data available in through an online ''automatic hypothesis-testing machine'' with a limited menu of user-specified criteria that could perform the first cut at a replication analysis. Promising associations could then be followed-up with more detailed collaborative analyses by the investigators of both the discovery and replication studies before journal publication.

Interactions. Testing of interaction effects poses particular challenges in the context of GWA studies (46), first because of the enormous number of potential interactions, 2.5×10^{11} pairwise SNP $\times$ SNP interactions in a typical whole genome scan, and similarly huge numbers for even a univariate scan of SNP associations with expression of all known genes (47), and by the general lack of specific prior hypotheses for any of them, beyond a vague belief that interactions are likely to be important. There is even more doubt about the reliability of interaction reports than for main effects, if only because of the larger multiple-comparison problem, the smaller sample sizes, and the low priors for any particular interaction, not to mention the poor track record of replication (44, 48-51).

So how should such a vague belief about interactions be accommodated? In studying lung cancer, one might strongly suspect that smoking could be an important modifier (52), but would it make more sense to sample smokers or nonsmokers? Perhaps simple random sampling or oversampling the extremes of the distribution of relevant environmental factor(s) might be the best gamble, in the absence of more specific hypotheses, with the hope of testing interactions in the analysis. But then how much of the type I error rate is one willing to spend on testing interactions at the expense of power for detecting main effects? Is one perhaps better off limiting the testing of interactions to pairs of factors attaining some threshold for main effects?

Marchini et al. (53) have shown that exhaustive testing of all possible pairwise interactions can be more powerful than testing only the univariately significant ones, depending of course upon the true interaction model. For candidate gene pathways, Millstein et al. (54) have proposed a ''Focused Interaction Testing Framework'' that seems to outperform a leading alternative approach, Multifactor Dimension Reduction (55), for searching for multidimensional interactions without requiring main effects, but it remains to be seen whether these approaches will be useful for GWA studies (56).

In any event, identification of genes that mediate or modify the effects of environmental hazards, as well as gene-gene interactions, are an important priority of both the scientific community and government agencies, as reflected in the recent National Heart, Lung, and Blood Institute and GEI announcements. Indeed, the GEI initiative involves a substantial commitment to developing new environmental measurements and incorporating them into GWA studies. Ironically, however, the GEI initiative coincided with the announcement of the cancellation of the National Children's Study (57), which could have set the groundwork for prospective evaluation of gene-environment interactions at least for common early-onset conditions (58), although an adult cohort would likely be more useful for studying most cancers, while also allowing for enrollment of their offspring and subsequent generations (59).

A Unified Approach

Returning to the question posed at the outset about pathway-driven versus exploratory approaches, some attractive approaches to marrying to two are the ''weighed false discovery rate'' approach (60) and a Bayesian false-discovery rate approach (42). Essentially, the weighed false-discovery rate spends the false discovery rate nonuniformly across all associations tested by using prior knowledge (in their example, a prior linkage trace). Although Roeder et al. warn against trying many different weighting functions and choosing the one that yields the greatest number of statistically significant findings (or most satisfying set of them), hierarchical regression models (61, 62) offer a valid way of incorporating a broad range of prior genomic information (location relative to genes, putative function, evolutionary conservations, biological pathways, previous linkage or association findings, or ''-omics'' databases; refs. 39, 63) into a general framework for prioritizing associations from an initial GWA scan for follow-up in later stages of a multistage design or for replication in other studies. Current skepticism about the utility of at least presently available genomic information (43) may change as these bioinformatics tools mature.

Clearly, the next few years should be an exciting time as GWA studies get under way, and the methods for conducting them become better developed. The interpretation of the mass of data that will result can be expected to keep investigators and pundits entertained long into the future. The stakes are high. If the diseases and specific studies are chosen wisely, the methods effective, and the information shared widely, the payoff could be considerable. But careful preparation is essential, lest high-profile failures diminish the public's and the scientific community's support of such research in the future.

References

1. Goodman L. Hypothesis-limited research. Genome Res 1999;9:673–4.
2. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1616–7.
3. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. A haplotype map of the human genome. Nature 2005;437:1299–320.
4. Hinds DA, Stuve LL, Nilsen GB, et al. Whole-genome patterns of common DNA variation in three human populations. Science 2005;307:1072–9.

Downloaded from http://aacrjournals.org/cebp/article-pdf/15/4/595/2264614/595.pdf by guest on 07 August 2022

5. Terwilliger JD, Hiekkalinna T. An utter refutation of the 'Fundamental Theorem of the HapMap'. Eur J Hum Genet 2006. Epub 2006 Feb 15.
6. Ozaki K, Ohnishi Y, Iida A, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 2002;32:650–4.
7. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385–9.
8. Maraganore DM, de Andrade M, Lesnick TG, et al. High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 2005;77: 685–93.
9. Shiffman D, Ellis SG, Rowland CM, et al. Identification of four gene variants associated with myocardial infarction. Am J Hum Genet 2005;77:596–605.
10. Mah S, Nelson MR, Delisi LE, et al. Identification of the semaphorin receptor PLXNA2 as a candidate for susceptibility to schizophrenia. Mol Psychiatry 2006 Jan 10; [Epub ahead of print].
11. Edwards AO, Ritter R III, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement factor H polymorphism and age-related macular degeneration. Science 2005;308:421–4.
12. Haines JL, Hauser MA, Schmidt S, et al. Complement factor H variant increases the risk of age-related macular degeneration. Science 2005;308: 419–21.
13. Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. Ann Hum Genet 2002;66:393–405.
14. Bansal A, van den Boom D, Kammerer S, et al. Association testing by DNA pooling: an effective initial screen. Proc Natl Acad Sci U S A 2002;99: 16871–4.
15. Pfeiffer RM, Rutter JL, Gail MH, Struewing J, Gastwirth JL. Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. Genet Epidemiol 2002;22:94–102.
16. Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA Pooling: a tool for large-scale association studies. Nat Rev Genet 2002;3:862–71.
17. Zou G, Zhao H. The impacts of errors in individual genotyping and DNA pooling on association studies. Genet Epidemiol 2004;26:1–10.
18. Thomas DC, Haile RW, Duggan D. Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 2005; 77:337–45.
19. Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. Lancet 2005;366:1223–34.
20. Lawrence RW, Evans DM, Cardon LR. Prospects and pitfalls in whole genome association studies. Philos Trans R Soc Lond B Biol Sci 2005;360: 1589–95.
21. Wang WYS, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 2005;6:109–18.
22. Hirschhorn JN, Daly MJ. Genome-wide association studies for common disease and complex traits. Nat Rev Genet 2005;6:95–108.
23. Farrall M, Morris AP. Gearing up for genome-wide gene-association studies. Hum Mol Genet 2005;14 Suppl 2:R157–62.
24. Satagopan JM, Elston RC. Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 2003;25:149–57.
25. Wang H, Thomas DC, Pe'er I, Stram DO. Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 2005: In Press.
26. Lin DY. Evaluating statistical significance in two-stage genomewide association studies. Am J Hum Genet 2006;78:505–9.
27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc, ser B 1995;57: 289–300.
28. Lin DY. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 2005;21:781–7.
29. Seaman SR, Muller-Myhsok B. Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. Am J Hum Genet 2005;76:399–408.
30. Sabatti C, Service S, Freimer N. False discovery rate in linkage and association genome screens for complex disorders. Genetics 2003;164:829–33.
31. van den Oord EJ, Sullivan PF. False discoveries and models for gene discovery. Trends Genet 2003;19:537–42.
32. Li SS, Bigler J, Lampe JW, Potter JD, Feng Z. FDR-controlling testing procedures and sample size determination for microarrays. Stat Med 2005; 24:2267–80.
33. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Statist 2003;31:2012–35.
34. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440–5.
35. Kraft P. Efficient two-stage genome-wide association designs based on false positive report probabilities. Pac Symp Biocomput 2006;11:523–34.
36. Zaykin DV, Zhivotovsky LA. Ranks of genuine associations in whole-genome scans. Genetics 2005;171:813–23.
37. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 2006;38:209–13.
38. Anonymous. Framework for a fully powered risk engine. Nat Genet 2005; 37:1153.
39. Rebbeck TR, Martinez ME, Sellers TA, Shields PG, Wild CP, Potter JD. Genetic variation and cancer: improving the environment for publication of association studies. Cancer Epidemiol Biomarkers Prev 2004;13:1985–6.
40. Ioannidis JPA. Journals should publish all ''null'' results and should sparingly publish ''positive'' results. Cancer Epidemiol Biomarkers Prev 2006;15:186.
41. Gwinn M, Khoury MJ. Expanded publishing model for genetic association studies. Cancer Epidemiol Biomarkers Prev 2006;15:185.
42. Whittemore AS. Genetic association studies: time for a new paradigm? Cancer Epidemiol Biomarkers Prev 2005;14:1359.
43. Begg CB. Reflections on publication criteria for genetic association studies. Cancer Epidemiol Biomarkers Prev 2005;14:1364–5.
44. Pharoah PDP, Dunning AM, Ponder BAJ, Easton DF. The reliable identification of disease-gene associations. Cancer Epidemiol Biomarkers Prev 2005;14:1362.
45. Wacholder S. Publication environment and broad investigation of the genome. Cancer Epidemiol Biomarkers Prev 2005;14:1361.
46. Kraft P. Multiple comparisons in studies of gene × gene and gene × environment interaction. Am J Hum Genet 2004;74:582–4.
47. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature 2005;437:1365–9.
48. Clayton DG, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001;358:1357–60.
49. Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. Lancet 2003;361:865–72.
50. Brennan P. Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? Carcinogenesis 2002;23:381–7.
51. Matullo G, Berwick M, Vineis P. Gene-environment interactions: how many false positives? J Natl Cancer Inst 2005;97:550–1.
52. Haiman CA, Stram DO, Wilkens LR, et al. Ethnic and racial differences in the smoking-related risk of lung cancer. N Engl J Med 2006;354:333–42.
53. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005;37: 413–7.
54. 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.
55. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 2003;24:150–7.
56. Moore JH, Ritchie MD. The challenges of whole-genome approaches to common diseases. J Am Med Dir Assoc 2004;291:1642–3.
57. Kimmel CA, Collman GW, Fields N, Eskenazi B. Lessons learned for the National Children's Study from the National Institute of Environmental Health Sciences/U.S. Environmental Protection Agency Centers for Children's Environmental Health and Disease Prevention Research. Environ Health Perspect 2005;113:1414–8.
58. Collins FS. The case for a US prospective cohort study of genes and environment. Nature 2004;429:475–7.
59. Potter JD. Toward the last cohort. Cancer Epidemiol Biomark Prev 2004;13: 895–7.
60. Roeder K, Bacanu SA, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 2006;78: 243–52.
61. Hung RJ, Brennan P, Malaveille C, et al. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. Cancer Epidemiol Biomarkers Prev 2004;13: 1013–21.
62. Pan W. Incorporating biological information as a prior in an empirical Bayes approach to analyzing microarray data. Statist Appl Genet Molec Biol 2005; 4:Art. 12.
63. Rebbeck TR, Spitz M, Wu X. Assessing the function of genetic variants in candidate gene association studies. Nat Rev Genet 2004;5:589–97.

Downloaded from http://aacrjournals.org/cebp/article-pdf/15/4/595/2264614/595.pdf by guest on 07 August 2022