



## Commentary

# Implications of Small Effect Sizes of Individual Genetic Variants on the Design and Interpretation of Genetic Association Studies of Complex Diseases

John P. A. Ioannidis<sup>1,2</sup>, Thomas A. Trikalinos<sup>1,2</sup>, and Muin J. Khoury<sup>3</sup>

<sup>1</sup> Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece.

<sup>2</sup> Institute for Clinical Research and Health Policy Studies, Tufts University School of Medicine, Boston, MA.

<sup>3</sup> Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, Atlanta, GA.

Received for publication October 17, 2005; accepted for publication March 21, 2006.

Accumulated evidence from searching for candidate gene-disease associations of complex diseases can offer some insights as the field moves toward discovery-oriented approaches with massive genome-wide testing. Meta-analyses of 50 non-human lymphocyte antigen gene-disease associations with documented overall statistical significance (752 studies) show summary odds ratios with a median of 1.43 (interquartile range, 1.28–1.65). Many different biases may operate in this field, for both single studies and meta-analyses, and these biases could invalidate some of these seemingly “validated” associations. Studies with a sample size of >500 show a median odds ratio of only 1.15. The median sample size required to detect the observed summary effects in each population addressed in the 752 studies is estimated to be 3,535 (interquartile range, 1,936–9,119 for cases and controls combined). These estimates are steeply inflated in the presence of modest bias. Population heterogeneity, as well as gene-gene and gene-environment interactions, could steeply increase these estimates and may be difficult to address even by very large biobanks and observational cohorts. The one visible solution is for a large number of teams to join forces on the same research platforms. These collaborative studies ideally should be designed up front to also assess more complex gene-gene and gene-environment interactions.

association; genes; meta-analysis; odds ratio; polymorphism, genetic

Abbreviation: IQR, interquartile range.

Human genome epidemiology is rapidly changing from the investigation of single genes and gene variants to the adoption of discovery-oriented approaches that encompass searching across millions of genetic variants (1, 2). Moreover, the challenge of accumulating evidence and modeling gene-gene interactions and gene-environment interactions is becoming more tangible as more rich databases are accumulated based on collaborative case-control studies and large cohort studies and biobanks. Theoretical debates have been ongoing for some time on the exact contribution of single variants and the magnitude of expected genetic effects (3). The accumulated evidence from candidate gene-

disease association studies to date can give us some useful evidence to also guide future efforts. In this paper, we briefly review the implications of small effect sizes of individual genetic variants on the design and interpretation of genetic studies of complex diseases

## HOW LARGE ARE EFFECT SIZES OF INDIVIDUAL GENETIC VARIANTS FOR COMPLEX DISEASES?

We scrutinized an updated, comprehensive database of 122 meta-analyses of non-human lymphocyte antigen gene-disease association studies of unrelated subjects on distinct,

**TABLE 1. The five most common genes\* implicated in genetic associations†**

Rank	Gene	Associations	Publications
1	<i>ACE</i>	6	731
2	<i>MTHFR</i>	5	729
3	<i>APOE</i>	3	905
4	<i>GSTM1</i>	3	538
5	<i>DRD3</i>	3	65

\* *ACE*, angiotensin I converting enzyme 1; *MTHFR*, 5,10-methylenetetrahydrofolate reductase; *APOE*, apolipoprotein E; *GSTM1*, glutathione *S*-transferase M1; *DRD3*, dopamine receptor D3.

† The number of significant associations is based on the database of 50 meta-analyses. The number of published studies pertains to indexed papers in the Human Genome Epidemiology (HuGE) Published Literature database for the period October 1, 2000, to September 6, 2005. Human lymphocyte antigen genetic variants are not considered in either database.

nonoverlapping associations with binary outcomes, where data were available on each included study to create the pertinent 2-by-2 table (4). The a priori rules for selection of meta-analyses, studies, and genetic contrasts have been described previously (4–6). In brief, whenever information was available to create both 2-by-2 and 2-by-3 tables, we selected the former. Among the possible genetic contrasts that could result in 2-by-2 tables, we chose the one proposed by the first study on the postulated association; when it was unclear, we chose the one proposed by the meta-analysis. When this was also unclear, the order of preference was recessive model, dominant model, allele-based model. Availability of information on 2-by-2 tables ensured that all data were reanalyzed consistently across studies in a meta-analysis and that the frequency of the genetic variant of interest was precisely recorded for both the cases and controls and could be used for analysis.

Fifty meta-analyses concluded with a statistically significant association ( $p < 0.05$ ) even when between-study heterogeneity was accounted for by random-effects calculations (DerSimonian and Laird model (7)). These 50 associations included a total of 752 studies. The published systematic reviews from which these 50 meta-analyses were derived examined only the specific gene variant in 39 meta-analyses, several variants of the same gene in seven meta-analyses, and variants from several genes perceived to be in the same pathway in another four meta-analyses. All studies on the 50 associations could be considered a typical comparison of cases and controls (case-control studies, cross-sectional studies, cases vs. population controls, prevalence data from cohort designs); only two meta-analyses also clearly included studies with a prospective cohort design and incident events.

Common phenotypes included cardiovascular disease outcomes ( $n = 10$ ), various cancers ( $n = 7$ ), schizophrenia ( $n = 7$ ), dementia ( $n = 4$ ), diabetes and its complications ( $n = 3$ ), and cerebrovascular outcomes ( $n = 3$ ). The five most common genes implicated in the associations are shown in table 1. It is interesting that four of these five genes are also included on the list of the five genes for which the highest number of papers appear in the published literature accord-

ing to the Human Genome Epidemiology (HuGE) Published Literature database as of September 6, 2005 (<http://www.cdc.gov/genomics/search/aboutHPLD.htm>). Postulated gene-disease associations are primarily described for the most sought-after candidate genes. Does this reflect that these genes are indeed important for many different outcomes? Does it mean that once an association has been proposed for a specific disease, bias is created and many other spurious associations of the same variant are then also reported for other diseases, or is it a manifestation of searching “under the lamp-post” until now? Probably what we see is a combination of all three factors.

Figure 1 shows the distribution of the genetic effects in the 50 meta-analyses (left panel) and in the 752 individual studies (right panel). We chose the direction of the genetic contrast in such a way that all summary odds ratios are higher than 1.00. For the meta-analyses, the median summary odds ratio is 1.43, with an interquartile range (IQR) of 1.28–1.65 and a range of 1.10–2.58. The distribution of the odds ratios in the 752 studies shows a median of 1.30 (IQR, 1.01–1.90). We should acknowledge that some of these seemingly significant gene-disease associations may not be true despite the fact that evidence of their presence comes from a considerable number of studies. In particular, for associations in the 1.1–1.3 range, even limited reporting or publication bias could produce a spurious effect (8).

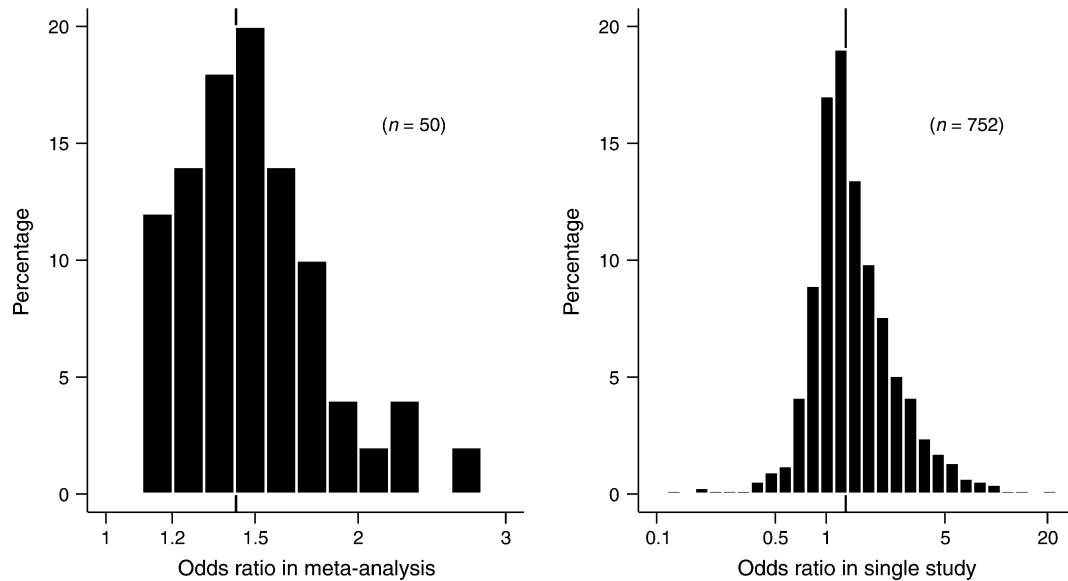
There were 168 out of 752 studies that had more than 500 participants or alleles (depending on the assessed contrasts). These “larger” studies are part of 42 meta-analyses, whereas eight meta-analyses are composed entirely of studies with a smaller sample size. The distribution of effect sizes across these 168 studies shows a median odds ratio of only 1.15 and an IQR of 1.01–1.45. Of the 42 meta-analyses with studies whose sample size exceeds 500, only 14 maintain formal statistical significance when limited to these larger studies. The median summary odds ratio for these 14 gene-disease associations is 1.45 (IQR, 1.28–1.64; range, 1.21–2.24).

Most of the genetic variants involved in these 50 postulated associations are relatively common. For the 752 individual studies, the median proportion for the minor genetic group (the less-frequent group according to the assumed genetic model) in the controls is 24.8 percent (IQR, 9.7–40.7 percent).

Overall, these data suggest that typical effect sizes of individual genetic variants for complex diseases pertain to odds ratios of 1.2–1.6. Some smaller effects are possible but are extremely difficult to differentiate from the potential impact of bias. Bias cannot be excluded even for the larger effects. Bias could be due to a large variety of factors. Their detailed description goes beyond the scope of this commentary but includes poor quality and design problems in single studies (9, 10), low prior probability of an association and relatively high  $p$  values (8, 11), reporting and publication biases (12, 13), and biased criteria for inclusion of studies in meta-analysis.

## IMPLICATIONS FOR SAMPLE SIZE REQUIREMENTS

One might then ask: Even if these summary effect sizes reflect the truth and if they are representative of the effect



**FIGURE 1.** Left panel: distribution of summary odds ratios based on random-effects calculations in 50 meta-analyses with formally statistically significant results for gene-disease associations of common diseases. Calculations were performed with Intercooled Stata 8.2 software (Stata Corporation, College Station, Texas). Right panel: distribution of odds ratios in the 752 studies included in these 50 meta-analyses. For both panels, the median is shown by a vertical line. For eligibility criteria for the screening of the meta-analyses, refer to Ioannidis et al. (4–6). A full list of nonoverlapping meta-analyses with binary outcomes is available in the online supplement to reference 4, and a full list of the data from the 50 included meta-analyses per study is available from the authors.

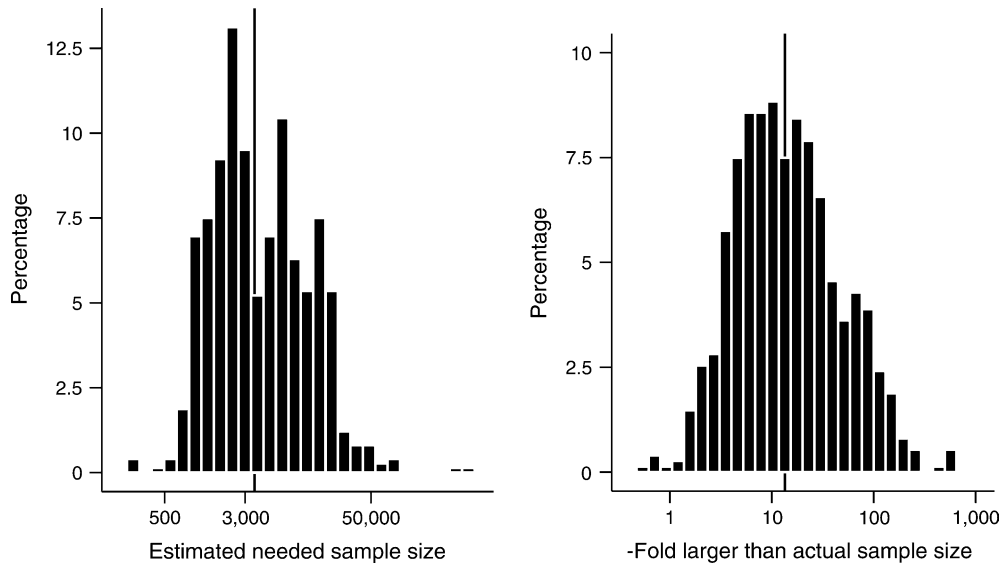
sizes for individual genetic variants associated with complex diseases, what kinds of studies are needed to document them in various population settings? Let us focus on the population settings in which these prior studies have already been conducted. For each of the 752 studies, we estimated the required sample so as to have 90 percent power to detect at  $\alpha = 0.05$  the genetic effect size seen in the respective meta-analysis, if the frequency of the genetic variants in the control group is that observed in the study. The choice of  $\alpha = 0.05$  represents a typical threshold for claiming statistical significance. Aiming for lower  $\alpha$  values (e.g., to accommodate multiple testing) would further increase the required sample size steeply. We assume the same allocation ratio between cases and controls in these hypothetical well-powered studies as the allocation ratio in the original studies. The actual allocation ratios are usually close to 1, with a median of 0.93 cases per control and an IQR of 0.53–1.15; using an allocation ratio of 1 for all calculations makes little difference overall (not shown in detail here). However, one should note that some studies understandably seem to have difficulty recruiting cases, even with the relatively small sample sizes used to date. Maintaining a reasonable allocation ratio may be a challenge if much larger samples are to be recruited, but, for now, let us assume that it can be done. Sample size calculations were implemented in Intercooled Stata 8.2 by using the *sampsi* Stata module (Stata Corporation, College Station, Texas).

Figure 2 shows that the required total sample size (cases and controls combined) can be very large. The left panel gives the distribution of the necessary sample sizes per

study, with a median of 3,535 and an IQR of 1,936–9,119 for cases and controls combined. The numbers required are much larger compared with studies conducted to date in the field. On median, 13.3-fold more subjects would have to be genotyped than in each original study conducted in each population (IQR, 5.9–31.4) (figure 2, right panel). If we try to account for even limited bias, these sample size requirements can be inflated considerably. For example, if the true odds ratios are 0.1 lower than the observed summary effect sizes (an assumption that may be quite conservative, based on the above), then the median required sample size becomes 6,244 (IQR, 2,698–35,444). If half of the observed summary effect is due to bias and half is real (e.g., for observed summary  $\log(\text{odds ratio}) = 0.46$ , the true effect is 0.23), the median required sample size becomes 14,618 (IQR, 7,791–36,435).

### CAVEATS AND LIMITATIONS

Meta-analyses in this field are becoming increasingly popular (14), but they cover only a portion of the available evidence on gene-disease associations. According to the HuGE Published Literature database, as of October 11, 2005, there were at least 17,467 published reports of original studies on human genome epidemiology, most of them ( $n = 16,267$ ) pertaining to gene-disease associations (15). It is unclear, however, whether the decision to perform and report a meta-analysis would be influenced by the postulated effect size of a significant association. Second, we acknowledge that some of the excluded, statistically nonsignificant



**FIGURE 2.** Left panel: distribution of the total sample sizes (cases and controls combined) required for 90% power to detect associations of the magnitude suggested by the summary odds ratio of a meta-analysis and the control frequency actually observed in each of the 752 studies (refer to the text for calculation details). Calculations are based on two-sided tests. Right panel: distribution of the ratio of the required (as in left panel) vs. the actual sample size used in the 752 studies. Calculations were performed with Intercooled Stata 8.2 software (Stata Corporation, College Station, Texas).

meta-analyses may have been underpowered to detect an existing, true genetic association (5). However, other aspects being equal, on average these effects are likely to be smaller than those included; thus, sample size requirements would be even larger.

Third, nondifferential measurement error in these studies may dilute the observed effect sizes, but it is more likely that selective reporting biases in favor of significant results are stronger and more than counterbalance this diluting impact. Fourth, the meta-analyzed variant may be in linkage disequilibrium with only the true culprit that has a larger odds ratio. However, the current discovery-oriented approaches generally do not necessarily target only the true, biologically important functional culprits. If anything, the 50 associations analyzed here probably have stronger functional support than the vast majority of associations that would be obtained currently through whole genome association analyses and other high-throughput approaches. Finally, our analyzed sample did not include any of the very few genetic variants that have been identified to date with a postulated odds ratio exceeding 3. The only such meta-analysis published in the time frame of our literature search (the apolipoprotein E gene (*APOE*) and Alzheimer's disease (16)) did not provide 2-by-2 tables per study. Considerations for the search of variants with very strong effects probably are different.

In the presence of genuine heterogeneity (e.g., ethnic or "racial" diversity) in the genetic effects (17), synergistic gene-gene interactions (18), and synergistic gene-environment interactions (19), the required sample sizes would easily increase further. For example, if the effect is present in only one ethnic subgroup or combination of gene(s) and environmental exposures, then analysis of an entire sample may wash out

the effect and reduce power. Misclassification is also a major concern for measurement of environmental exposures, but it can also affect genotyping. In the presence of even modest nondifferential misclassification, the required sample sizes increase steeply (20).

#### HOW DO WE MEET THE EMERGING CHALLENGES OF HUMAN GENOME EPIDEMIOLOGY?

Meeting the goals of the current research agenda in genetic association studies would probably require sample sizes in the range of several thousands to more than tens of thousands to answer the simpler questions, and sample sizes possibly in the range of 50,000–100,000 or even larger to answer questions of modest complexity. These numbers pertain to case-control studies, including in particular those nested within even larger cohorts. Even the single largest general-purpose observational cohorts and biobanks (21, 22) would be challenged to meet these numbers. Except for very common diseases, such as coronary artery disease, where a considerable fraction of the population may be suitable cases, for most diseases, the largest biobanks and cohorts may be unable to provide conclusive answers. For example, for Parkinson's disease, if the frequency of the disease in a general population cohort is 1 percent, then the cohort must include 500,000 subjects to enable enrollment of approximately 5,000 cases with the disease to conclusively answer the simple questions.

A cohort base of several million subjects may have to be recruited to answer the somewhat more complex questions. Finally, for the most common diseases, such as coronary artery disease noted above, it is unclear whether a broadly

defined phenotype would be sufficient to capture the underlying genetic complexity. The failure of single genetic variants–disease association studies of coronary artery disease to date (23, 24) may be due to the fact that such a common phenotype may reflect an array of many subphenotypes, each with a different genetic background. The pertinent subphenotypes, even if appropriately deciphered without getting lost in exploratory subgroup analyses, may have much lower prevalence and incidence rates and thus extremely high sample size requirements to delineate their risk factors. Etiologic heterogeneity with diverse subphenotypes of different genetic background in less common diseases would be even more difficult to address.

The effort to identify more complex effects should not be abandoned. Because of the small effect sizes of individual genetic variants, it may be reasonable to look for complex genotypes that operate in biologic pathways and gene–environment interactions. Yang et al. (25) have shown that the combination of a few genetic variants (10 to 20) at multiple loci, each with a modest effect size (odds ratio of about 1.5), may account for a substantial portion of the population attributable fraction for many common diseases. Of course, we do not know the form of joint effects of genetic variants (multiplicative, additive, or otherwise) nor how these variants interact with environmental factors. Conventional wisdom has taught us that looking for interactions usually requires a vast expansion of sample sizes of the original studies; however, under certain plausible biologic scenarios of more extreme interactions among genes and genes and environmental factors, there could be increased statistical power for looking for such interactions in studies designed to detect marginal effects of individual genotypes (26, 27). The search for more complex genotypes with stronger effects in such studies also makes sense in terms of the eventual application of these findings to genomic medicine. As shown by Holtzman and Marteau (28) and confirmed in the analyses presented here, individual genetic variants with weak or modest effect sizes are unlikely to be used for prediction and prevention of common diseases, whereas the combination of genetic variants, even with modest individual effect sizes, can lead to a marked increase in the ability to predict disease risks (29).

If heterogeneity is immense across populations, it is questionable whether this predictive enterprise is feasible at all. If most genetic variance is highly defined by very “private” genetic variation interacting with highly “private” environmental exposures, then epidemiology is probably not the way to address risk factors for complex diseases. Yet, it is unclear whether anything else can take the place of epidemiologic investigation in this pursuit (30). Given the complexity of the genetics of common diseases, we should foster good a priori hypotheses regarding genes and environmental factors, innovative study designs, and strong collaborative efforts.

At a minimum, research teams working with the same disease and sets of questions should join forces in networks with common objectives. This goal is currently a major effort of the HuGENet “network of networks” initiative (31, 32). Reaching the point where our knowledge base for genetic associations for complex diseases is reliable is not easy. However, such knowledge is likely to be highly

desirable because it would allow us to explain a large proportion of the cause of most common diseases and may also lead to new therapeutic avenues and tailored preventive interventions. A road map has been recently proposed on how to reach this goal (32). It emphasizes efforts that will minimize bias in the published and unpublished literature, enhance data synthesis across diverse teams of investigators, grade the credibility of the evidence accumulated, and maintain updated field-wide synopses that summarize the evolving knowledge in a specific field in as systematic and unbiased a manner as possible. Eventually, the impact on individual and public health could be considerable.

---

## ACKNOWLEDGMENTS

Conflict of interest: none declared.

---

## REFERENCES

1. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;37:413–17.
2. Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* 2005;366:1223–34.
3. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1516–17.
4. Ioannidis JP, Ntzani EE, Trikalinos TA. ‘Racial’ differences in genetic effects for complex diseases. *Nat Genet* 2004;36:1312–18.
5. Ioannidis JP, Ntzani EE, Trikalinos TA, et al. Replication validity of genetic association studies. *Nat Genet* 2001;29:306–9.
6. Ioannidis JP, Trikalinos TA, Ntzani EE, et al. Genetic associations in large versus small studies: an empirical assessment. *Lancet* 2003;361:567–71.
7. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–88.
8. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124.
9. Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. *Lancet* 2003;361:865–72.
10. Cordell HJ, Clayton DG. Genetic association studies. *Lancet* 2005;366:1121–31.
11. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004;96:434–42.
12. Pan Z, Trikalinos TA, Kavvoura FK, et al. Local literature bias in genetic epidemiology: an empirical evaluation of the Chinese literature. *PLoS Med* 2005;2:e334.
13. Ioannidis JP. Genetic associations: false or true? *Trends Mol Med* 2003;9:135–8.
14. Lohmueller KE, Pearce CL, Pike M, et al. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003;33:177–82.
15. Little J, Houry 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.

16. Farrer LA, Cupples LA, Haines JL, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* 1997;278:1349–56.
17. Bamshad M. Genetic influences on health: does race matter? *JAMA* 2005;294:937–46.
18. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* 2002;155:478–84.
19. Hunter DJ. Gene-environment interactions in human diseases. *Nat Rev Genet* 2005;6:287–98.
20. Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev* 1999;8:1043–50.
21. Collins FS. The case for a US prospective cohort study of genes and environment. *Nature* 2004;429:475–7.
22. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. *Pharmacogenomics* 2005;6:639–46.
23. Keavney B, McKenzie C, Parish S, et al. Large-scale test of hypothesised associations between the angiotensin-converting-enzyme insertion/deletion polymorphism and myocardial infarction in about 5000 cases and 6000 controls. International Studies of Infarct Survival (ISIS) Collaborators. *Lancet* 2000;355:434–42.
24. Wheeler JG, Keavney BD, Watkins H, et al. Four paraoxonase gene polymorphisms in 11212 cases of coronary heart disease and 12786 controls: meta-analysis of 43 studies. *Lancet* 2004;363:689–95.
25. Yang Q, Khoury MJ, Friedman J, et al. How many genes underlie the occurrence of common complex diseases in the population? *Int J Epidemiol* 2005;34:1129–37.
26. Khoury MJ, Adams M, Flanders WD. An epidemiologic approach to ecogenetics. *Am J Hum Genet* 1988;42:89–95.
27. 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;51:336–43.
28. Holtzman NA, Marteau TM. Will genetics revolutionize medicine? *N Engl J Med* 2000;343:141–4.
29. Yang Q, Khoury MJ, Botto LD, et al. Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *Am J Hum Genet* 2003;72:636–49.
30. Rebbeck TR, Spitz M, Wu X. Assessing the function of genetic variants in candidate gene association studies. *Nat Rev Genet* 2004;5:589–97.
31. Ioannidis JP, Bernstein J, Boffetta P, et al. A network of investigator networks in human genome epidemiology. *Am J Epidemiol* 2005;162:302–4.
32. Ioannidis JPA, Gwinn ML, Little J, et al. A roadmap for efficient and reliable human genome epidemiology. *Nat Genet* 2006;38:3–5.