Published by Oxford University Press on behalf of the International Epidemiological Association 2006 *International Journal of Epidemiology* 2007;**36**:439–445
Advance Access publication 20 December 2006 doi:10.1093/ije/dyl253

GENETICS

On the synthesis and interpretation of consistent but weak gene–disease associations in the era of genome-wide association studies

Muin J Khoury,[1]* Julian Little,[2] Marta Gwinn[1] and John PA Ioannidis[3]

Accepted 21 October 2006

Emerging technologies are allowing researchers to study hundreds of thousands of genetic variants simultaneously as risk factors for common complex diseases. Both theoretical considerations and empirical evidence suggest that specific genetic variants causally associated with common diseases will have small effects (risk ratios mostly <2.0). However, the combination of even a few small effects (e.g. effects of fewer than 20 common genetic variants) could account for a sizeable population attributable fraction of common diseases and shed important light on disease pathogenesis and environmental determinants. Nevertheless, the inauguration of genome-wide association studies only magnifies the challenge of differentiating between the expected, true weak associations from the numerous spurious effects caused by misclassification, confounding and significance-chasing biases. Standards are urgently needed for presenting and interpreting cumulative evidence on gene–disease associations, especially for consistent but weak associations. Criteria for synthesis of the evidence should include sound methods for study conduct and analysis, biological plausibility, experimental evidence and adequate replication in large-scale, collaborative studies. Efforts by the Human Genome Epidemiology Network (HuGENet) are currently ongoing to streamline and operationalize these criteria for data on genetic associations with common diseases.

Keywords Epidemiological methods, genomics, risk ratios, genome-wide analysis

Completion of the Human Genome Project[1] launched a wave of intense investigations of human genetic variation in relation to common complex human diseases.[2] Because most common diseases, such as cancer, diabetes and heart disease, are caused by many genes and environmental factors and their interactions, progress in unravelling genetic risk factors has been slow and tedious. However, common patterns of genetic variation revealed by the International Haplotype Map (HapMap) project are now the basis for lower cost and more efficient genomics technologies.[3,4] Comprehensive analyses of common variation in the human genome in association with specific diseases in case-control or cohort studies are commonly called genome-wide association (GWA) studies. Such studies typically measure sets of special DNA 'tagging' single nucleotide polymorphisms (SNPs) identified in the HapMap project, enriched with non-synonymous and 'quasirandom' SNPs, as well as SNPs in

[1] National Office of Public Health Genomics, Coordinating Center for Health Promotion, Centers for Disease Control and Prevention.

[2] Canada Research Chair in Human Genome Epidemiology, Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Canada.

[3] Department of Hygiene and Epidemiology University of Ioannina School of Medicine, Ioannina, Greece and Department of Medicine, Tufts University School of Medicine, Boston, USA.

* Corresponding author. National Office of Public Health Genomics, Centers for Disease Control and Prevention, 4770 Buford Hwy, Atlanta GA 30341 USA (mailstop K89). E-mail: mkhoury@cdc.gov

evolutionary conserved regions of the genome. Available analytical platforms allow testing for several hundred thousand SNPs (e.g. from 100K to 675K). Although the tested SNPs represent less than one-tenth of all known SNPs, they are in strong linkage disequilibrium with 80–90% of all known SNPs, and therefore achieve high coverage of common variants of the human genome.

It is more debatable whether rare genetic variants are properly covered.[5] As the cost of high-throughput genotyping decreases exponentially, studies performing genome-wide analysis of thousands of individuals are becoming increasingly feasible. These studies will expand the search for genetic risk factors for complex human diseases on an unprecedented scale.[6,7]

Certainly, there is increasing excitement and expectations that GWA approaches will produce disease susceptibility findings such as we have seen recently with age-related macular degeneration,[8] diabetes type II[9] and prostate cancer.[10] In this commentary, we examine the implications of GWA studies for the synthesis and interpretation of associations between single genetic variants and various complex diseases. In particular, we focus on consistently replicated but weak associations (i.e. risk ratios (RR) typically <1.5). Because we expect many gene–disease associations to have small effect sizes and fewer with larger effect sizes,[11,12] even if they are biologically meaningful, weak true effects are difficult to distinguish from spurious effects caused by methodological biases. Thus, standards for the synthesis and interpretation of consistent but weak associations are essential tools for the imminent, large-scale epidemiological 'fishing expedition' in the human genome.

Most human genome epidemiology studies still report just one or a few gene–disease associations

Although large GWA studies are now gearing up for various diseases,[13–15] most published gene–disease associations are based on case-control studies assessing just one or a few candidate genes with a postulated role in pathogenesis. For example, as of November 21, 2006, the online CDC database[16] of published epidemiological studies of human genes contained more than 25 000 publications; of these, 84% reported analyses of one or a few (not more than five) specific genetic variants.[17] As we discuss subsequently, many methodological issues remain in the analysis and reporting of single gene–disease associations. Even in studies using GWA methods, typically only a few associations survive the process of replication.

Effects of individual genetic variants are expected to be small, even if they are biologically meaningful

The aetiology and pathogenesis of common chronic diseases such as cancer, diabetes and heart disease reflect the joint effects of numerous genetic and environmental risk factors and their interactions. For any given disease, 'major' gene effects (which manifest as Mendelian or single-gene disorders) account for only a small fraction of cases. For example, familial hypercholesterolaemia,[18] α-1-antitrypsin deficiency,[19] hereditary non-polyposis colorectal cancer[20] and *BRCA* mutations[21] account for <5% of the cases of heart disease, emphysema, colon cancer and breast cancer, respectively. Such Mendelian forms of common diseases have been associated with specific clinical or pathological subtypes, such as early age at onset of heart disease, colon cancer and breast cancer and panacinar histology in emphysema (although ascertainment bias may contribute to their recognition). A complex interplay of genetic and environmental factors likely accounts for the largest attributable fraction of these and other common diseases. Besides Mendelian disorders, larger genetic effects may be observed in some population subgroups defined by genetic background, environmental exposure or disease subtype. Nevertheless, strong genetic effects in subgroup analyses are usually swamped by the large number of spurious findings that can result from data dredging. Schmidt *et al*.[22] have recently proposed an ordered subset analysis (OSA) method to allow for the incorporation of covariates into linkage analysis of disease phenotypes in order to reduce genetic heterogeneity. When multiple genetic and environmental risk factors interact in the pathogenesis of common disease, the effect of any individual factor depends on the relative prevalence of other risk factors, genetic or environmental, that are part of the same sufficient cause.[23] In a simple scenario, all risk factors may be considered to act independently in a multiplicative fashion. Yang *et al*.[24] have recently shown that when the predisposing genetic variants are very common in the population (each with prevalence $\geq$25%), a modest number ($\leq$20) could explain 50% of the burden of a disease in the population, even if the individual genotype associations are relatively small [e.g. relative risk (RR) = 1.2–1.5]. Nevertheless, at this time, it is not possible to assess how likely this scenario is to occur for the majority of common diseases and whether or not it will be consistent with observed patterns of familial aggregation and heritability estimates from these studies. In more complex scenarios, interactions beyond the multiplicative model may be more important and marginal effects may well be tiny. A very strong association that is present in only a small subgroup of the population, defined by the presence of other genetic and environmental factors, can be severely diluted towards the null if these other factors were not specifically measured in the study.[25,26] As the combined prevalence of these interacting risk factors becomes less frequent in the population, the 'average' or marginal effect—obtained from crude analysis—for each genetic variant approaches the null. For example, even a gene–disease association with a RR of 100 will have an observed RR of only 1.1, if the combination of factors required to complete a sufficient cause occurs in just 1/1000 people.[25]

We recently analysed results of 50 meta-analyses (based on a total of 752 studies) that reported statistically significant summary gene–disease associations. In these meta-analyses, the median odds ratio (OR) was 1.43, with an inter-quartile range (IQR) of 1.28–1.65.[27] The true 'average' effects could be even smaller if bias exaggerated some of the results. One of the most consistently replicated but weak associations in genetic

epidemiology, is that of bladder cancer with the *NAT2* genotypes associated with the slow acetylator phenotype. Garcia-Closas *et al*.[28] reported on a meta-analysis of 31 studies of *NAT2* and bladder cancer, including a total of 5091 cases and 6501 controls. The summary RR for *NAT2* slow acetylators compared with rapid/intermediate acetylators was 1.4 (1.2–1.6; $P < 0{\cdot}0001$). The consistency of this association across numerous studies suggests that it reflects more than mere bias; however, its small size suggests that other unmeasured, underlying interactions among genes and environmental exposures also have a role.

Although interaction effects abound in the literature, most are derived from small studies that are underpowered even for detecting main effects, suggesting that they may be the result of *post hoc* analysis with little chance of replication.[29] Among *NAT2* slow acetylators, the consistent finding of higher bladder cancer risk in cigarette smokers than in never-smokers bolsters the case for true interaction. Case-only meta-analyses provided support for an interaction between *NAT2* and smoking (P for interaction 0.009).[28]

But wait! How will we find such needles in the haystack?

It will be exceedingly difficult from individual studies to distinguish small effects that are valid and biologically meaningful from those arising spuriously from the methodological problems that are known to plague the field of genetic association studies. Many of these problems have been described and quantified for individual studies,[30,31] and for meta-analyses;[31] they include frank biases as well as the inherent complexity of the task at hand.[32–43] Biases include significance-chasing (including publication bias, selective analysis and reporting biases), confounding by population stratification, faulty selection of subjects for comparisons, differences in storage and genetic analysis of samples collected from cases and controls, genotyping errors, deviations from Hardy–Weinberg equilibrium, linkage disequilibrium issues and misclassification of exposures and outcomes. Other inherent problems include presence of undetected gene–gene and gene–environment interactions, limited sample size and statistical power and type I errors (false positive associations).

All of these issues have been reviewed previously and will not be discussed further, except to point out that type I errors are particularly relevant to the conduct of GWA studies. A large search among hundreds of thousands of genetic variants can be expected by chance alone to find thousands of false positive signals (RRs significantly different from 1.0). Many approaches to this problem have been proposed, including Bonferroni adjustment for multiple comparisons, Bayesian inference based on prior probabilities, and more exploratory methods such as data visualization, multi-dimensionality reduction and neural network analyses.[44–46] While these methods may facilitate the detection of true associations, they are likely to greatly increase the number of false positive associations as well. Perhaps a good summary of the prevailing notion in the genetic association field is the recent editorial by Duncan Thomas. He states: 'Clearly, the next few years should be an exciting time as GWA studies get under way . . . The interpretation of the mass of data that will result can be expected to keep investigators and pundits entertained long into the future'.[6] Perhaps the process of careful replication will help 'average' out the biases, but a prolonged cycle of initial positive, reported findings and subsequent refutation could consume a great deal of energy and resources. Therefore, we believe that more systematic attempts at epidemiological data integration and interpretation in the era of GWA studies need to be developed.

Synthesis and interpretation of the evidence on consistent but weak associations

Recognizing methodological challenges in the field of genetic associations, the Human Genome Epidemiology Network (HuGENet) was founded in 1998 as an open, global collaboration of individuals and organizations committed to creating the knowledge base on genetic variation and human diseases. Recently, the collaboration has undertaken a number of steps to move this rather difficult field forward.[47,48] A 'network of networks' of investigators has been created to share best practices, tools and methods for analysis of associations between genetic variation and common diseases.[49] A 'road map' was published in January 2006 to define near-term plans for HuGENet and the networks. These include developing consensus guidelines for reporting results of genetic association studies; augmenting published associations with unpublished data by promoting publication of 'negative' studies and collaboration on comprehensive analyses within investigator networks; expanding systematic HuGE reviews with meta-analyses of individual-level data and prospective meta-analyses; and developing field synopses that will offer regularly updated overviews online and in selected journals.[50,51] Detailed guidance for conducting systematic reviews of gene–disease associations with quantitative synthesis of the results was published online in March 2006.[52] As shown in Table 1, 366 meta-analyses of gene–disease associations have been published (of which 49 are formal HuGE reviews) and their yearly numbers have more than tripled between 2001 and 2005.

Beyond the mechanics of meta-analysis and systematic synthesis of gene–disease associations, what evidence do we

Table 1 Reported gene-disease associations in the published literature[a] by type of publication and year, 2001–05

Publication year	Gene–disease association	Gene–gene gene–environment joint effects	Meta-analysis or HuGE review
2001	2142	435	33
2002	2796	568	36
2003	3020	600	61
2004	3761	663	93
2005	4611	924	143

[a] HuGE published literature database[16,17] searched online April 10, 2006 at http://apps.nccd.cdc.gov/genomics/GDPQueryTool/frmQueryAdvPage.asp

need to conclude that such associations are 'causal'? As discussed by Weed,[53] meta-analysis may allow us to 'summarize evidence from biological, clinical and social levels of knowledge . . . [but] combining evidence across levels is beyond its current capacity. Meta-analysis has a real but limited role in causal inference, adding to an understanding of some causal criteria.' Hill's[54] nine criteria for causal inference have been used to interpret observed associations between environmental exposures and disease outcomes in various fields, such as nutritional epidemiology,[55] but only in a limited way in the field of genetic association.[55] Although none of the nine criteria (including strength of association) are absolute, the question of causal interpretation of genetic associations is of timely interest.

By creating the opportunity for millions of comparisons, GWA might be expected to generate an outpouring of false positive results. However, by its very nature, GWA may also supply a definitive solution to the problem of selective reporting, which is not limited to genetic epidemiology. Typically, epidemiological studies target only a few risk factors at a time and only selected findings are published. In theory, GWA studies could collect information simultaneously on a very large number of genetic variants and make the entire database transparent and available online.[7]

One important near-term activity for HuGENet is developing a systematic approach to assessing cumulative evidence and inferring causality for genetic associations. In a recent commentary, one of us proposed a schema (Table 2) for qualitative scoring on five 'axes,' including effect size, replication, protection from bias, biological plausibility and relevance to medicine or public health.[57] This schema has much in common with the criteria, guidelines or viewpoints discussed earlier but will need to be further modified based on accumulated experience from ongoing GWA efforts (see Table 2 for comments). For example, in this schema, 'weak' associations (RRs <2) are viewed as least credible, yet we can expect most true associations to fall below this criterion. Indeed, many may have relative risks <1.2, in the range where very large sample sizes are needed (tens of thousands of cases). Under this scenario, the analytical ability of epidemiological methods will break down, even with limited bias. For example, a recent GWA-based discovery of a genetic variant that increases the risk of obesity 1.22-fold has not been replicated consistently.[15] How credible can this association be and how large a sample size do we need to validate such an association, even if it is true?

Replication of evidence, while an absolute necessity, could become more problematic as researchers debate how much replication is enough, especially in the case of small effect sizes. Perhaps in genetic epidemiology, replication may be a continuous process without end. In research on medical interventions, too much replication is unacceptable because it exposes people to documented risks from harmful interventions or withholds benefits from effective interventions. In genetic epidemiology, the downside of excess replication lies in the opportunity costs—research funds and investigator efforts that could be better applied to other endeavours. Even accumulated evidence from a large number of studies may have modest credibility, and better and larger studies may still be needed. The cost of replicating associations with individual genetic variants emerging as candidates from a GWA study will be considerably less costly than the GWA study itself. Certainly, an open model for sharing of individual level data in GWA studies—as we are beginning to see from the NIH-sponsored Genetic Association Information Network (GAIN) initiative[58] and the Wellcome Trust-sponsored Wellcome Trust Case-Control Consortium (WTCCC) consortium[59]—may help the validation/replication process by enhancing transparency and minimizing selective reporting biases.

'Protection from bias' is difficult to assess. Most known biases in epidemiology cause spurious associations that can easily mimic a true small effect size. Biases introduced by genotyping

Table 2 Grading the credibility of the evidence for individual gene–disease associations: some proposed grading criteria and their limitations in interpreting recurring weak associations

Axis	Proposed grading[a]	Comments
Effect size	Small effect size (RR<2) has lowest grade while large effect size considered best (RR>5)	Most biologically causal factors are expected to have RR < 2. Many may be beyond the limit of analytical ability
Amount of evidence/replication	Single or few scattered studies have lowest grade while large-scale inclusive analyses are best	The more information the better the inference, although it may be difficult to set hard rules for the amount of replication for weak associations. There is a risk for endless replication
Protection from bias	Clear presence of bias gets poor grade while clear strong protection gets high grade	Most studies will be in between. Absolute protection from bias is hard to achieve. More empirical evidence and consensus is needed on which biases are more serious than others.
Biological plausibility	No functional data scores lowest while convincing biological data scores highest	Need consensus and empirical evidence for the importance of specific items of biological plausibility
Relevance	Graded according to clinical or public health application	Individual weak associations will have little relevance to use for genetic testing because of their poor predictive ability especially for rare conditions

[a] Grading proposed by Ioannidis.[57]

errors and population stratification can at least be measured. Probably much more important are common epidemiological biases related to participant selection, outcome ascertainment and measurement of exposures and interactions. It may be possible to reach consensus on the relative importance and control of biases in the design, conduct and presentation of single studies and the assembly and presentation of large sets of evidence; however, additional, empirical studies are needed to elucidate how such biases operate alone and in combination.

Biological plausibility also remains an unsettled topic as researchers begin to integrate results of genetic association studies with other lines of evidence, e.g. from gene expression studies, online bioinformatics databases and experimental studies in animal models.[60–62]

The bottom line in grading evidence is its relevance to clinical and public health practice. Even 'true' genetic associations may not explain clinically meaningful or preventable outcomes. The strongest genetic associations may not be with clinical endpoints but with intermediate phenotypes or other biological markers far upstream from the outcome of interest. For example, a genetic variant might have an imperceptible association with myocardial infarction, a somewhat stronger association with serum pro-thrombotic profile, and a strong relationship to a gene expression profile, reflecting complex biological processes that may or may not lead to simple interventions.

Weak risk factors, including genetic variants, have little validity as predictive or diagnostic tests and are thus clearly inadequate tools for screening populations and testing individual patients.[63,64] Information about individual genetic variants with RRs of $\leq$2.0 will probably not find direct application in medicine or public health;[65] however, they may offer clues to disease pathogenesis, natural history and environmental risk factors through 'Mendelian randomization'.[66–68] Combining several genetic variants in a single test could improve positive predictive value but only at the cost of reduced sensitivity.[69] More sophisticated approaches are needed to integrate information on multiple genetic variants with other biomarkers, as well as physiological and clinical data, for use in population medicine. A useful framework to address this challenge is the emergence of collaborative population-based biobanks with adequate consent procedures, storage and sharing of biological samples for studying the joint role of genetic and environmental factors on the occurrence of common diseases.[70]

Conclusions

In evaluating associations between genetic variants and common complex diseases, we should fully expect biologically meaningful associations with small effects. The usual criteria for grading the evidence and for causal inference need to be adapted and modified. As part of the HuGENet 'road map', an ongoing effort has been made to streamline and operationalize criteria for genetic associations with various common diseases. Because of the lack of clinical or public health utility of these weak associations for genetic testing, many may dismiss such findings as hype, focusing on the obvious methodological issues that plague genetic association studies. We do not think this should be discouraging. Weak associations will be the norm rather than the exception and in the current era of genome-wide association studies, we have the opportunity to develop a validated and continuously updated 'knowledge base' on the relationship between genetic factors and human diseases. Studying bias and false positive findings will be very informative and useful. The next few years will provide a crucial window of opportunity to develop methods and standards for measuring, validating and interpreting genetic associations. The simple answer to 'are we there yet?' may be 'no' for years to come. Ultimately, the promise of the Human Genome Project rests on our ability to accurately characterize the relationship between genetic variation and human disease and use this information for the benefit of population health.

Conflict of interest: None declared.

KEY MESSAGES

- Most human genome epidemiology studies still report just one or a few gene–disease associations.
- Effects of individual genetic variants are expected to be small, even if they are biologically meaningful.

References

[1] Collins FS, Guttmacher AE. Welcome to the genomic era. *N Engl J Med* 2003;**349**:996–98.

[2] Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* 2005;**366**:1223–34.

[3] International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;**437**:1299–320.

[4] Dove A. The SNPs are down: genotyping for the rest of us. *Nat Methods* 2005;**2**:989–94.

[5] Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. *Nat Genet* 2006;**38**:659–62.

[6] Thomas D. Are we ready for genome-wide association studies. *CEBP* 2006;**15**:595–98.

[7] Lawrence RW, Evans DM, Cardon LR. Prospects and pitfalls in whole genome association studies. *Philo Trans Royal Soc B: Biol Sci* 2005;**360**:1589–95.

[8] Klein RJ, Zeiss C, Chew EY *et al*. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005;**308**:385–89.

[9] Grant SF, Thorleifsson G, Reynisdottir I *et al*. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 2006;**38**:320–23.

[10] Amundadottir LT, Sulem P, Gudmundsson J *et al*. A common variant associated with prostate cancer in European and African populations. *Nat Genet* 2006;**38**:652–58.

[11] Wright A, Charlesworth B, Rudan I *et al*. A polygenic basis for late-onset disease. *Trends Genet* 2003;**19**:97–106.

[12] Barton NH, Keightley PD. Understanding quantitative genetic variation. *Nat Rev Genet* 2002;**3**:11–21.

[13] Maraganore DM, de Andrade M, Lesnick TJ *et al*. Whole genome association study for Parkinson disease. *Am J Hum Genet* 2005;**77**:684–93.

[14] Arking DE, Pfeufer A, Post W *et al*. A common genetic variant in the NOS1 regulator *NOS1AP* modulates cardiac repolarization. *Nat Genet* 2006; online publication April 30, 2006; DOI:10.1038/ng1790.

[15] Herbert A, Gerry NP, McQueen MB *et al*. A common genetic variant is associated with adult and childhood obesity. *Science* 2006;**312**:279–83.

[16] Lin B, Clyne M, Walsh M *et al*. Tracking the epidemiology of human genes in the literature: the HuGE published literature database. *Am J Epidemiol* 2006 [Epub ahead of print] PMID: 16641305.

[17] Centers for Disease Control and Prevention. Genomics and Disease Prevention Information System (GDP Info). Accessed online on May 1, 2006 at: http://www.cdc.gov/genomics/search/aboutHPLD.htm.

[18] Austin MA, Zimmern RL, Humphries SE. High 'population attributable fraction' for coronary heart disease mortality among relatives in monogenic familial hypercholesterolemia. *Genet Med* 2002;**4**:275–78.

[19] de Serres FJ. Worldwide Racial and Ethnic Distribution of Alpha1-Antitrypsin Deficiency: Summary of an Analysis of Published Genetic Epidemiologic Surveys. *Chest* 2002;**122**:1818–29.

[20] National Cancer Institute. Genetics of colorectal cancer. (PDQ). Accessed online May 1, 2006 at: http://www.cancer.gov/cancertopics/pdq/genetics/colorectal/healthprofessional.

[21] National Cancer Institute. Genetics of breast and ovarian colorectal cancer. (PDQ). Accessed online May 1, 2006 at: http://www.cancer.gov/cancertopics/pdq/genetics/breast-and-ovarian/healthprofessional.

[22] Schmidt S, Schnmidt MA, Qin X *et al*. Linkage analysis with gene-environment interaction: model illustration and performance of ordered subset analysis. *Genet Epidemiol* 2006;**30**:409–22.

[23] Rothman KJ, Greenland D. Causation and causal inference in epidemiology. *Am J Publ Health* 2005;**95(Suppl. 1)**:S144–50.

[24] Yang Q, Khoury MJ, Friedman J, Little J, Flanders WD. How many genes underlie the occurrence of common complex diseases in the population? *Int J Epidemiol* 2005;**34**:1129–37.

[25] Khoury MJ, Adams MJ, Flanders WD. An epidemiologic approach to ecogenetics. *Am J Hum Genet* 1988;**42**:89–95.

[26] Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;**37**:413–17.

[27] Ioannidis JPA, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol* 2006;**164**:609–14.

[28] Garcia-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**:649–59.

[29] Ioannidis JPA. Why Most Published Research Findings Are False. *PLoS Med* 2005;**2**:e124.

[30] Bogardus ST, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research: the need for methodologic standards. *JAMA* 1999;**281**:1919–26.

[31] Attia J, Thakkinstian A, D'Este C. Meta-analyses of molecular association studies: methodologic lessons for genetic epidemiology. *J Clin Epidemiol* 2003;**56**:297–303.

[32] Ioannidis JP. Genetic associations. False or true? *Trends Mol Med* 2003;**9**:135–38.

[33] Thomas DC, Witte JS. Point: Population stratification: a problem for case-control studies of candidate-gene associations? *CEBP* 2002;**11**:505–12.

[34] Hattersley AT, McCarthy MI. What makes a good association study? *Lancet* 2005;**366**:1315–23.

[35] Cordell HJ, Clayton DG. Genetic association studies. *Lancet* 2005;**366**:1121–31.

[36] Lawrence RW, Evans DM, Cardon LR. Prospects and pitfalls in whole genome association studies. *Phil Trans Royal Soc B* 2005;**360**:1589–95.

[37] Kraft P, Hunter D. Integrating epidemiology and genetic association: the challenge of gene-environment interaction, *Phil Trans R Soc B* 2005;**360**:1609–16.

[38] Salanti G, Sanderson S, Higgins JP. Obstacles and opportunities in meta-analysis of genetic association studies. *Genet Med* 2005;**7**:13–20.

[39] 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.

[40] Bracken MB. Genomic epidemiology of complex disease: the need for an electronic evidence-based approach to research synthesis. *Am J Epidemiol* 2005;**162**:297–301.

[41] Tang JL. Selection bias in meta-analyses of gene-disease Associations. *PloS Med* 2005;**2**:e249.

[42] Clayton DG, Walker NM, Smyth DJ *et al*. Population structure, differential bias, and genomic control in a large-scale, case-control association study. *Nat Genet* 2005;**37**:1243–46.

[43] Ioannidis J, Salanti G, Trikalinos T *et al*. Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations. *Am J Epidemiol* 2006;**163**:300–09.

[44] Khoury MJ, Millikan R, Little J, Gwinn M. The emergence of epidemiology in the genomics age. *Int J Epidemiol* 2004;**33**:936–44.

[45] Farrall M, Morris AP. Gearing up for genome-wide gene-association studies. *Hum Mol Genet* 2005;**14**:R157–62.

[46] Wacholder S, Chanock S, Garcia-Closas M *et al*. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *JNCI* 2004;**96**:434–42.

[47] Khoury MJ, Dorman JS. The human genome epidemiology network. *Am J Epidemiol* 1998;**148**:1–3.

[48] Centers for Disease Control and Prevention. The Human Genome Epidemiology Network Homepage. Accessed online May 1, 2005 at: http://www.cdc.gov/genomics/hugenet/default.htm.

[49] Ioannidis JP, Bernstein J, Bofetta P *et al*. A network of investigator networks in human genome epidemiology. *Am J Epidemiol* 2005;**162**:302–04.

[50] Ioannidis JP, Gwinn M, Little J *et al*. A road map for efficient and reliable human genome epidemiology. *Nat Genet* 2006;**38**:3–5.

[51] Editorial. Embracing risk. *Nat Genet* 2006;**38**:1.

[52] Little J, Higgins JP (eds). HuGENet handbook of systematic reviews version 1.0 released March 2006. Accessed online May 1, 2006 at: http://www.genesens.net/_intranet/doc_nouvelles/HuGE%20Review%20Handbook%20v11.pdf.

[53] Weed DL. Hill criteria applications Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. *Int J Epidemiol* 2000;**29**:387–90.

[54] Hill A. The environment and disease: association or causation? *Proc R Soc Med* 1965;**58**:295–300.

[55] Potischman N, Weed DL. Causal criteria in nutritional epidemiology. *Am J Clin Nutr* 1999;**69**:1309S–14S.

[56] Page GP, George V, Go RC *et al*. ''Are we there yet?'': Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am J Hum Genet* 2003;**73**:711–19.

[57] Ioannidis JP. Commentary: grading the credibility of molecular evidence for complex diseases. *Int J Epidemiol* 2006; [Epub ahead of print] PMID: 16540537.

[58] Genetic Association Information Network (GAIN). Accessed online on 10/4/2006 at: http://www.fnih.org/GAIN/GAIN_home.shtml.

[59] The Wellcome Trust Case Control Consortium. (WTCCC). Accessed online on 10/4/2006 at: http://www.wtccc.org.uk/.

[60] Rebbeck TR, Spitz M, Wu X. Assessing the function of genetic variants in candidate gene association studies. *Nat Rev Genet* 2004;**5**:589–94.

[61] Jais PH. How frequent is altered gene expression among susceptibility genes to human complex disorders? *Genet Med* 2005;**7**:83–96.

[62] Campbell H, Rudan I. Interpretation of genetic association studies in complex disease. *Pharmacogen J* 2002;**2**:349–60.

[63] Khoury MJ, Newill CA, Chase GC. Epidemiologic evaluation of screening for risk factors: application to genetic screening. *Am J Public Health* 1985;**75**:1204–08.

[64] Pepe MS, Janes H, Longton G *et al*. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;**159**:882–90.

[65] Holtzman NA, Marteau T. Will genetics revolutionize medicine? *New Engl J Med* 2000;**343**:141–44.

[66] Davey-Smith G, Ibrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32**:1–22.

[67] Khoury MJ, Davis RL, Gwinn M *et al*. Do we need genomic research for the prevention of common diseases with environmental causes? *Am J Epidemiol* 2005;**161**:799–805.

[68] Nitsch D, Molokhia M, Smeeth L *et al*. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *Am J Epidemiol* 2006;**163**:397–403.

[69] Khoury MJ, Yang Q, Gwinn M *et al*. An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. *Genet Med* 2004;**6**:38–47.

[70] Public Population Project in Genomics (P3G). Accessed online on October 4, 2006 at: http://www.p3gconsortium.org/index.cfm.

Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2007;**36**:445–448
doi:10.1093/ije/dym055

Commentary: Rare alleles, modest genetic effects and the need for collaboration

H Campbell* and T Manolio[4]

Accepted 1 March 2007

The article by Khoury *et al*.[1] presents a useful overview of some of the complex issues facing those trying to identify genetic variants underlying common complex disease. They focus on the common disease—common variant model where effect sizes associated with individual genetic variants are small. Undoubtedly this will be the case for most, but not all, variants. An L-shaped or exponential distribution of mutation effect sizes has wide support [2–4] with many variants with small effects, a smaller number with intermediate effects and relatively few with large effects. It could be argued that the genetic variants related to human disease that have been identified to date primarily reflect the study designs used to identify them. Linkage studies conducted among families with multiple cases of disease were successful in identifying highly penetrant variants with large effects. Association studies conducted in general population samples using common genetic markers typically find low penetrance variants with (very) small effects, as noted by Khoury. This is not unexpected given that these common genetic variants are ancient and will have been subject to some selective pressure over time.[3]

We can predict that re-sequencing studies in the near future which study rarer variants (say 0.05–5%) will identify many variants of intermediate effect associated with common complex disease. This paradigm shift has already begun with the seminal work of Cohen, who compared non-synonymous sequence variations in individuals at the extremes of the population distribution of LDL-cholesterol levels, and determined that a significant fraction of genetic variance is due to multiple alleles with intermediate effects that are present at low frequencies (0.05–5%) in the population, particularly persons of African ancestry.[5] Until many such studies are reported it will be premature to decide on the relative importance of the common variant—common disease model and the alternative rare variant—common disease model which states that disease susceptibility to common diseases is the result of multiple low frequency/rare variants with larger phenotypic effects. As Cohen notes, although individually rare, these variants may be

* Corresponding author. Division of Community Health Sciences, Public Health Sciences, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK. E-mail: harry.campbell@ed.ac.uk