Int. J. Epidemiol. Advance Access published October 5, 2007

Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2007;1–2
doi:10.1093/ije/dym205

Commentary: Genetic association studies see light at the end of the tunnel

Timothy M Frayling

Accepted 6 September 2007

In this month's issue Ioannidis *et al*.[1] provides a welcome guide to interpreting data from genetic association studies.

The authors' efforts are important for two main reasons. First, genetic associations have been fraught with difficulty over the past 10 years in their attempts to uncover DNA polymorphisms that alter disease risk. The vast majority of reported associations, typically between a single nucleotide polymorphism (SNP) and a disease, were not replicated. The reasons for this are now well understood and have been discussed before. The main problem is that geneticists have several 100 000 risk factors to study in the form of common polymorphisms but only a few are likely to be involved in any one disease. This makes the prior odds that any one variant is associated very low and therefore very stringent *P*-values are needed to provide any confidence in the statistical evidence.[2] Second, genome wide association (GWA) studies, that test several 100 000 DNA variants in a single experiment, have arrived in abundance in early 2007 making it potentially even harder for epidemiologists to pick their way through the data to decide what is real.

Why bother—are not most genetic effects very small?

With a few exceptions such as the associations between certain human leukocyte antigen (HLA) haplotypes and auto-immune diseases and variants in the complement factor H gene and age-related macular degeneration, most associations between DNA variants and disease are small. The latest results indicate that an odds ratio of 1.3 is a big effect for a common variant with a common disease. However, the authors point out an important aspect of genetic association studies; provided the study is well designed in terms of genotyping quality control and ethnic matching of cases and controls, an association between a polymorphism and disease is much more likely to represent real biology than epidemiological associations studying non-genetic factors. The reason for this is that very few confounding factors can influence the Mendelian process of random assignment of parental alleles during meiosis.[3] Bias, most notably reporting bias, can affect the interpretation of genetic association studies, but other sources of biases are unlikely to influence genetic studies, provided a few simple points are ticked off from a quality control 'checklist'. This means, even with an odds ratio of 1.1, that statistically robust associations are likely to uncover new diseases mechanisms. Ioannidis *et al*. very helpfully guides us through this checklist.

Assessing the evidence

The authors provide some simple guidelines on how confident we can be about genetic associations. They put most weight on the statistical evidence. This is in agreement with other discussions on the subject.[4,5] Given the low *a priori* odds of an association it is emerging that *P*-values in the range of $\sim 1 \times 10^{-7}$ are needed to provide something close to a traditional 1 in 20 chance that the finding is a false positive. The initial results coming from GWA studies indicate that these criteria are being applied and that it is about right: associations with this level of statistical confidence are very likely to be replicated and include variants in or near the *CDKAL1*, *TCF7L2* and *FTO* genes associated with type 2 diabetes,[6–10] variants near the *CDKN2A/2B* genes with coronary artery disease,[6,11–13] several type 1 diabetes variants,[6,14] and an association in the *BTBD9* gene with restless legs syndrome.[15,16] Reporting bias does not affect the interpretation of these results because several large positive studies have been published at a similar time and in each case it would take a negative study of many tens of thousands of individuals to reduce the statistical confidence to 'uncertain' levels.

Once we have assessed the weight of statistical evidence and decided that we can be confident the finding is unlikely to be down to statistical chance, what else could mean the result does not reflect a causal association? The authors very helpfully point out several other issues that readers need to look out for. Fortunately, the good news is that most of these are now being addressed in the first reports of genetic associations making readers' lives a lot easier. The authors discuss sample size in their table 2. They suggest that studies need sample sizes of at least 1000 for the smallest group out of the cases and controls. This is consistent with the first wave of GWA publications, where most studies have used more than 1000 cases and more than 1000 controls.

In their table 3, the authors discuss potential sources of bias and include a list of 'possible/high' risks of bias for certain aspects of genetic association studies. First, these include bias

Genetics of Complex Traits, Peninsula Medical School, University of Exeter, EX1 2LU, UK.

E-mail: Tim.frayling@pms.ac.uk

in phenotype definition. A real example of this was recently provided by the type 2 diabetes genome wide scans. Some had used cases and controls of very similar, or even matched, body mass index (BMI),[7,17] the most important risk factor for diabetes, whilst other studies had not selected cases and controls on the basis of BMI. This meant some studies identified common variation in the *FTO* gene, which was shown to alter BMI in the general population, as a type 2 diabetes gene whilst others did not. Importantly, this kind of bias is unlikely to result in false-positive results, but did result in false-negative results in the studies that effectively corrected for something on the causal pathway. Second, the authors point out the importance of genotyping quality control ('bias in genotyping'). Readers need to be confident that studies have adhered to all the appropriate quality control checks. Often this is easy to see if the authors report that a second variant, highly correlated with the main one reported, is also associated—the chances that two variants have produced spurious results is greatly reduced.

Third, the authors discuss population stratification. This can be a genuine confounding factor in genetic association studies if not properly controlled for. If there are two background populations and disease frequency and allele frequencies are different between these background populations then false-positive and false-negative results can occur. Fortunately, again this problem can be overcome if studies include and correct for ancestry informative markers—DNA polymorphisms that vary a lot in allele frequency across the geographical region of study. Genome wide studies offer even greater protection from potential population stratification because data from several 100 000 variants can be used to very accurately correct for or exclude individuals from different ethnic backgrounds.[18]

Reporting bias has been a major concern for genetic association studies. Many reports, even those with more than one study may be <100% honest about including all data from all studies they have access to. Here, again things are improving rapidly. The authors point out that one effective way round this is to set up consortia and perform meta-analyses. This is happening. The first wave of genome wide analyses usually report several studies in the first publication. The signs are that there is little in the way of selective reporting: for several diseases new associations have been reported in more than one study and results are replicating.

In conclusion, the 'interim' guidelines from Ioannidis *et al*. provide a useful framework for epidemiologists to assess the robustness of genetic associations. Fortunately, and perhaps paradoxically, given the wealth of data they are producing, GWA studies, together with a greatly improved understanding of statistical and quality control issues, are making life a lot easier. If genetic association studies meet what are now routine quality criteria, they offer an unprecedented increase to our understanding of common diseases and conditions.

Conflict of interest: None declared.

References

[1] Ioannidis JPA, Boffetta P, Little J *et al*. Assessment of cumulative evidence on genetic associations: interim guidelines. *Int J Epidemiol* (In press).

[2] Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000;**405:**847–56.

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

[4] Todd JA. Statistical false positive or true disease pathway? *Nat Genet* 2006;**38:**731–33.

[5] Chanock SJ, Manolio T, Boehnke M *et al*. Replicating genotype-phenotype associations. *Nature* 2007;**447:**655–60.

[6] The Wellcome Trust Case Control consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–78.

[7] Saxena R, Voight BF, Lyssenko V *et al*. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;**316:**1331–36.

[8] Scott LJ, Mohlke KL, Bonnycastle LL *et al*. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;**316:**1341–45.

[9] Steinthorsdottir V, Thorleifsson G, Reynisdottir I *et al*. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* 2007;**39:**770–75.

[10] Zeggini E, Weedon MN, Lindgren CM *et al*. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007;**316:**1336–41.

[11] Helgadottir A, Thorleifsson G, Manolescu A *et al*. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* 2007;**316:**1491–93.

[12] McPherson R, Pertsemlidis A, Kavaslar N *et al*. A common allele on chromosome 9 associated with coronary heart disease. *Science* 2007;**316:**1488–91.

[13] Samani NJ, Erdmann J, Hall AS *et al*. Genomewide association analysis of coronary artery disease. *N Engl J Med* 2007;**357:**443–53.

[14] Todd JA, Walker NM, Cooper JD *et al*. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 2007;**39:**857–64.

[15] Stefansson H, Rye DB, Hicks A *et al*. A genetic risk factor for periodic limb movements in sleep. *N Engl J Med* 2007;**357:**639–47.

[16] Winkelmann J, Schormair B, Lichtner P *et al*. Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. *Nat Genet* 2007;**39:**1000–6.

[17] Sladek R, Rocheleau G, Rung J *et al*. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007;**445:**881–85.

[18] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;**38:**904–9.