Diabetologia (2005) 48: 1241–1243
DOI 10.1007/s00125-005-1807-2

COMMENTARY

M. I. McCarthy · P.-H. Groop · T. Hansen

Making the right associations

Published online: 26 May 2005

Abbreviations SNP: single nucleotide polymorphism

Increasingly, genetic studies contribute to our understanding of the pathogenesis of diabetes and its complications at the molecular level [1, 2]. With the advent of powerful new high-throughput analytical methods [3], there is every reason to expect further insights. Virtually all such genetic studies include some test of the association between genome sequence variation and the phenotype of interest, be that diabetes itself, the presence of a given complication, or measures of some diabetes-related intermediate trait. Association studies of this type have scored several recent successes in relation to both type 1 (*CTLA4, INS* and *PTPN22* genes) [2, 4, 5] and type 2 diabetes (*PPARG, KCNJ11*) [6–8], identifying variants that are unequivocally associated with disease. Two further association studies, examining variation in the genes encoding the Krüppel-like transcription factor family and the receptors for adiponectin, are published in this issue of *Diabetologia* [9, 10].

Despite these advances, there are major concerns regarding the overall performance and robustness of genetic association studies. All too often, initial positive (or negative) association findings fail the test of replication, leaving the literature littered with the detritus of uncertain and poorly reproducible associations. As set out in a number of excellent review articles [11–15], the origins of such inconsistency lie in a series of common methodological failings. These include (but are not limited to) the use of sample sizes that are inadequately powered for the task in hand, incomplete assessments of sequence variation within the locus of interest, technical errors in genotyping, and the inappropriate interpretation of data when large number of statistical tests have been performed. Furthermore, discrepancies introduced by these failings are compounded by the rather low prior odds that any given variant contributes to susceptibility to a given trait and by the understandable bias of reviewers and journal editors towards the publication of novel, superficially interesting, positive associations (whilst otherwise well-performed association studies reporting no association are often dismissed as 'negative'). In fact, where power is low and liberal thresholds for declaring significance are used, the vast majority of such 'positive' associations will be erroneous [16].

Of course, it is not all bad news. Susceptibility variants for diabetes and its complications do exist. However, with the exception of HLA in type 1 diabetes, the effect sizes are usually modest. Such effects can be reliably and reproducibly detected provided the studies are adequately powered and feature appropriately constructed sample sets, careful genotyping and appropriate analysis [2, 4–8]. For example, the variants P12A in *PPARG* and E23K in *KCNJ11* have been detected repeatedly within large type 2 diabetes case-control samples, such that the overall evidence for association now exceeds any reasonable correction for genome-wide significance [17]. Other type 2 diabetes-susceptibility effects—at *CAPN10* for example—look in-

M. I. McCarthy (✉)
Oxford Centre for Diabetes, Endocrinology and Metabolism,
University of Oxford,
Headington, Oxford, OX3 7LJ, UK
e-mail: mark.mccarthy@drl.ox.ac.uk
Tel.: +44-1865-857298
Fax: +44-1865-857299

M. I. McCarthy
Wellcome Trust Centre for Human Genetics,
University of Oxford,
Oxford, UK

P.-H. Groop
Folkhälsan Research Centre, Biomedicum Helsinki,
University of Helsinki,
Helsinki, Finland

P.-H. Groop
Division of Nephrology, Department of Medicine,
Helsinki University, Central Hospital,
Helsinki, Finland

T. Hansen
Steno Diabetes Center and Hagedorn Research Institute,
Gentofte, Denmark

creasingly convincing as larger sample sizes have been assembled, though the combined evidence is not, as yet, incontrovertible [18].

As these successes show, and as connoisseurs of the association game will know, single studies rarely, if ever, provide conclusive evidence that a given variant is (or is not) associated with disease. Instead, evidence typically accumulates over multiple studies. Crucially, replication not only provides access to increasingly large cumulative sample sizes but also insures against the inevitable biases, confounders and technical problems (latent population stratification and genotyping errors, for example) that can unavoidably afflict any individual study [19]. Only when extensive replication studies have been completed does a meaningful estimate of the true effect size emerge.

All of this leaves journals such as *Diabetologia* in something of a quandary. On the one hand, no journal wants to publish papers that have a depressingly high chance of being 'wrong' [16]. On the other hand, to insist that each individual study should reach some standard of unequivocal proof would paralyse the field. It is hard to think of any genetics paper published in *Diabetologia* that could be considered to have cleared the latter hurdle.

Journal editors with limited page numbers also struggle with the question of bias referred to above. Manuscripts reporting positive associations are undoubtedly more 'newsworthy' than those that find no association (not least because the large number of variants within the genome means that only a small percentage of variants will ever be truly associated with a given trait). Yet, in the interests of establishing a true, unbiased picture of the overall strength of the evidence for (or against) association at a given variant, it is important that all well-performed studies stand an equitable chance of publication, whatever their outcome. To facilitate more complete capture of association data, there have been calls for electronic registration [14, 20], and some provisional efforts have been made to establish such registries (e.g. http://geneticassociationdb.nih.gov, last accessed in April 2005). However, for the time being, scientific probity requires that well-performed studies that fail to detect associations (especially those that convincingly fail to replicate published positive results) should be regarded as competitive for publication.

Two papers in this issue of *Diabetologia* illustrate the issues discussed above rather well. Kanazawa et al. [9] describe association studies in a series of Japanese case-control samples, which were designed to evaluate a family of biological candidates (the Krüppel-like transcription factors). One of the single nucleotide polymorphisms (SNPs) typed (in *KLF7*) shows an apparently impressive association (p=0.000057, odds ratio=1.59) with type 2 diabetes. There is evidence of replication across the two sample sets and some functional data to support the candidacy of the gene concerned. Although these results are interesting, the authors are right to be circumspect in their interpretation of the data given the chequered history of association findings. At the very least, this study is likely to have overestimated the effect size at the *KLF7* SNP (the so-called 'winner's curse' [20]). At worst, future replication studies (in Japanese and other populations) may fail to confirm the association at all, and the candidacy of *KLF7* will founder on the rocks that have claimed so many other promising associations. These results are therefore far from conclusive. However, publication of this well-performed study enables other groups to test the robustness of this association in additional data sets.

In the second paper, Hara et al. [10] have undertaken a detailed assessment of the relationship between the variation in two excellent candidate genes (encoding the receptors for the adipocytokine, adiponectin) and type 2 diabetes, considering both the discrete disease phenotype and diabetes-related intermediate traits. A dense inventory of common variants in these genes was examined and, to the extent allowed by the sample sizes deployed, no significant association effects were detected. Once again, the authors have been appropriately cautious in their interpretation, pointing out that the variants typed do not represent a complete inventory of sequence variation in and around the genes, and that larger sample sizes would have provided greater power to detect (and/or exclude) more modest effects. However, given the biological credentials of the genes concerned as candidates for susceptibility to type 2 diabetes, these remain interesting data which, despite these limitations, add significantly to our knowledge.

Editorial judgements about which papers should be published in a journal such as *Diabetologia* are inevitably somewhat subjective. The vagaries of the peer-review process are well known to all who participate. In an attempt to foster consistency in the treatment of such manuscripts and to encourage transparency in the decisions reached, *Diabetologia* has developed a set of guidelines for genetic association studies. These are posted on the journal's website (http://www.diabetologia-journal.org/genetics%20guidelines.htm).

The term 'guidelines' is used advisedly. These are very definitely not meant to be thresholds that must be cleared if a paper is to be accepted for publication in *Diabetologia*. Very few association studies are entirely blameless, and the authors of those that are might reasonably be seeking publication in the very highest impact journals. What these guidelines do reflect are the criteria that we expect those participating in the peer-review process for *Diabetologia* to use when evaluating association studies. As well as encouraging consistency and transparency, we also hope that these guidelines will lead to a greater awareness of these important methodological issues by those undertaking association studies. In addition, by asking authors to ensure that submissions include information essential for their evaluation (e.g. rs numbers of variants, information on genotyping accuracy), the guidelines may even accelerate the review process.

Finally, we hope that these guidelines will contribute to a more mature dialogue between authors and the readership of the journal, one marked less by emphasis on the 'headline' p value (which all too often induces authors to construct some post-hoc narrative around the single nominally significant result that emerged from the forest of statistical output) and more by methodological and technical probity,

adequate sample size, and appropriate and considered interpretation of the findings (including a realistic assessment of the limitations of the study).

The perceived unreliability of association studies has been a weight around the neck of genetic researchers over the past decade. In years to come, genome-wide association studies will generate data on a previously unimaginable scale (billions of genotypes). These studies should provide a comprehensive view of the association landscape of type 2 diabetes, but only if they are well performed and correctly interpreted. Greater emphasis on the application of robust methodologies for association studies is therefore particularly timely.

Acknowledgements We would like to thank those who commented on and contributed to generation of the guidelines listed on the *Diabetologia* website: E. Zeggini, S. Wiltshire (Oxford, UK); J. Hirschhorn, D. Altshuler, J. Florez (Cambridge, MA, USA), A. Hattersley, T. Frayling, M. Weedon (Exeter, UK), O. Pedersen (Gentofte, Denmark), and L. Groop (Malmö, Sweden).

References

1. McCarthy MI (2004) Progress in defining the molecular basis of type 2 diabetes through susceptibility gene identification. Hum Mol Genet 13(Suppl 1):R33–R41
2. Ueda H, Howson JMM, Esposito L et al (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature 423:506–511
3. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108
4. Barratt BJ, Payne F, Lowe C et al (2004) Remapping the insulin gene/IDDM2 locus in type 1 diabetes. Diabetes 53:1884–1889
5. Smyth D, Cooper JD, Collins JE et al (2004) Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. Diabetes 53:3020–3023
6. Altshuler D, Hirschhorn JN, Klannemark M et al (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 26:76–80
7. Gloyn AL, Weedon MN, Owen KR et al (2003) Large scale association studies of variants in genes encoding the pancreatic beta-cell K-ATP channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with increased risk of type 2 diabetes. Diabetes 52:568–572
8. Nielsen ED, Hansen L, Carstensen B et al (2003) The E23K variant of Kir6.2 associates with impaired post-OGTT serum insulin response and increased risk of type 2 diabetes. Diabetes 52:573–577
9. Kanazawa A, Kawamura Y, Sekine A et al (2005) Single nucleotide polymorphisms in the gene encoding Krüppel-like factor 7 are associated with type 2 diabetes. Diabetologia DOI 10.1007/s00125-005-1797-0
10. Hara K, Horikoshi M, Kitazato H et al (2005) Absence of an association between the polymorphisms in the genes encoding adiponectin receptors and type 2 diabetes. Diabetologia DOI 10.1007/s00125-005-1806-3
11. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45–61
12. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis JG (2001) Replication validity of genetic association studies. Nat Genet 29:306–309
13. Ioannidis JPA, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis JG (2003) Genetic associations in large versus small studies: an empirical assessment. Lancet 361:567–571
14. Colhoun HM, McKeigue PM, Davey Smith G (2003) Problems of reporting genetic associations with complex outcomes. Lancet 361:865–872
15. Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5:89–100
16. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 96:434–442
17. Parikh H, Groop L (2004) Candidate genes for type 2 diabetes. Rev Endocr Metab Disord 5:151–176
18. Weedon MN, Schwarz PEH, Horikawa Y et al (2003) Meta-analysis confirms a role for calpain-10 variation in type 2 diabetes susceptibility. Am J Hum Genet 73:1208–1212
19. Page GP, George V, Page PZ, Allison DB (2003) "Are we there yet?": deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. Am J Hum Genet 73:711–719
20. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177–182