Eur J Epidemiol (2009) 24:585–587
DOI 10.1007/s10654-009-9387-y

COMMENTARY

Genetic Scoring Analysis: a way forward in Genome Wide Association Studies?

Najaf Amin · Cornelia M. van Duijn · A. Cecile J. W. Janssens

Received: 6 August 2009 / Accepted: 20 August 2009 / Published online: 2 September 2009

For the past 5 years genome-wide association studies (GWAS) have dominated the search for new genes for complex diseases overtaking other approaches of gene finding such as candidate gene and linkage analyses. Facilitated by technological developments in molecular biology, genetic epidemiologists have so far discovered many variants associated with several common diseases and traits such as Type 2 Diabetes, age-related macular degeneration and Crohn's disease [1]. There currently are 26 established susceptibility genes published for type 2 diabetes [2], 54 for human height and 22 for lipid levels [3, 4]. These variants still explain only a small part of the genetic variance or heritability, for human height and lipids up to 4–6% [5, 6], and subsequently the search for novel variants continues to unravel 'missing heritability'.

This missing heritability is explained by additional rare variants with strong effects and/or common variants with weak effects, acting additively and/or interacting with other genetic and environmental variants. To discover these additional genetic factors, GWAS need to enlarge, and this has led to further expansion of existing consortia and the establishment of new ones. Since the first publication in 2005 [7], GWAS have undergone enormous evolution: from 10,000 single nucleotide polymorphisms (SNPs) in 100 individuals of a single sample [7] to 1 million genotyped and ~2.5 million imputed SNPs in more than 80,000 individuals of multiple samples [8]. The decreasing costs of genotyping, new statistical methodologies, and increasing willingness of the scientists to share and pool data sets have facilitated these rapid developments and made this approach very successful also in the setting of epidemiology. For instance, the Cohort for Health and Aging Research (CHARGE) is studying multiple common traits in 50,000–70,000 individuals from US and European follow-up studies [9–11], and the Dutch three-generation study LifeLines is going to include 165,000 participants [12].

While increasing size will help in finding new variants with smaller effects, there will also be true positives that remain undetected in the larger consortia because of the stringent threshold levels of statistical significance imposed in GWAS ($P < 5 \times 10^{-8}$) to adjust for multiple testing. The chances of success of consortia are further reduced if confounding due to population heterogeneity, also refer to as population admixture, is to be adjusted for, which is the case when populations are of different genetic origins. Therefore new approaches are needed to identify genetic variants explaining the missing heritability and one such new approach was used successfully in a recent GWAS in schizophrenia that was published in *Nature*, online on July 1 [13]. The classical GWAS analysis produced only one genome-wide significant polymorphism, but the authors used a new 'genetic scoring' method through which they demonstrated that there indeed existed undetected variants below the threshold. How to detect variants that are not detected? Basically, the method tests the association of a score variable that manifests a combined effect of many SNPs. The polymorphisms in the score are selected on the basis of their nominal *P* value in the predefined discovery sample. Scores can be generated for any arbitrarily chosen threshold of nominal statistical significance, for instance selecting all SNPs with *P* values lower than e.g. 0.01, 0.1 or 0.5. The significance of the score is then tested by using it as a predictor in a simple regression model in an independent 'target sample'. In this target sample, a one-parameter test for all SNPs can be used, thus relaxing the

N. Amin · C. M. van Duijn · A. C. J. W. Janssens (✉)
Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands
e-mail: a.janssens@erasmusmc.nl

conservative P value of 5×10^{-8} needed for testing all SNPs in GWAS to classical significance level of 0.05. Using data from the International Schizophrenia Consortium with men defined as the discovery sample and women as the target sample, the authors showed that a score based on all SNPs with $P < 0.5$ was most strongly and significantly correlated with schizophrenia in the target sample compared to the scores based on other thresholds. The fact that the set of SNPs with $P < 0.5$, including both many falsely and an unknown number of truly associated SNPs, predicted better than the score with $P < 5 \times 10^{-8}$ suggests that both the number of undetected relevant variants as well as their joint effect on the outcome is substantial [13]. The authors further showed that the score correlated significantly with related diseases as bipolar disorder, but not with unrelated outcomes such as Crohn's disease, coronary artery disease, hypertension, rheumatoid arthritis or type 1 and type 2 diabetes. This suggests that schizophrenia and bipolar disorder have a shared genetic component and also that the selected alleles were specific to schizophrenia and related disorders [13].

The genetic scoring method is logical and simple as among the SNPs that fail to reach the significance threshold in the GWAS there ought to be true associations, which just do not reach the threshold because the study does not have enough power [14]. There may, however, be several caveats. First, the informative value of the approach depends on the size of the discovery sample. If the discovery sample is small, more falsely associated SNPs will be selected at each threshold, and consequently scores do not explain much of the phenotypic variance in the target sample. The second caveat is that also a score based on 38,000 SNPs with P value lower than 0.5, derived in a discovery sample of 3,800 individuals, explained only 3% of the heritable variance in the target population of 3,100 persons. It can be expected that a larger discovery set will select more true positives among those with a P value lower than 0.5 and therefore explain a higher percentage of the variance in the target sample. However, simulations showed that the variance explained by the scores can increase from 3 to 20% if the size of the discovery sample is increased to 20,000 individuals [13]. Thus, also for this new method the size of the discovery sample is an important determinant of success. Third, one of the major conclusions on the basis of this method is that there are undetected common genetic contributions. Of course one may argue that this observation could already be inferred from the fact that there is 'missing heritability'. But perhaps an even more important limitation of the genetic scoring method is that it does not tell which one(s) of the variants included is responsible for the statistical significance.

Then what can we do with this information? First, the method may be used to improve our understanding the genetic architecture of the disease or trait. Scores can be calculated and tested for multiple different significance thresholds levels of statistical significance. By comparing the proportions of explained variance across these thresholds, a pattern may be observed. When going up from a very low threshold, e.g., $P < 10^{-7}$ to $P < 0.5$, we may see that scores may rise to a certain point and then either decline or become stable, a pattern which suggests that a few genes with stronger effects may be involved. When the proportion of explained variance monotonically increases until all SNPs are included in the scores, there are likely to be a large number of common variants with small effects. So the scores calculated over several different cut offs can give an indication on how complex the trait is, on the likelihood that the trait has a polygenic basis. For example, for schizophrenia the score goes up from 0.004 to 0.025 by moving up from a threshold of $P < 0.01$ to $P < 0.5$ [13], which is an indication that many more common low risk variants are likely involved in schizophrenia.

Second, this method could be considered as an intermediate step in the gene discovery process. When scores are statistically significant, one may consider to only analyzing the included SNPs in the independent samples. For replication purposes this leads to a less stringent level of statistical significance, and potentially to a higher likelihood of finding susceptibility variants. Because the success of this approach will depend on the size of the discovery sample—the larger the discovery sample the more likely true susceptibility genes will be selected in the scores—its added value of selecting SNPs in much smaller independent populations may not be efficient. More promising is to use the score approach to select SNPs for use in complex modeling of the trait for instance to study gene by gene interactions which otherwise seems impossible with 2.5 million SNPs.

Third, the method could be used to predict disease for preventive and clinical purposes. Evans and colleagues applied the score approach and assessed the discriminative ability for several threshold levels of statistical significance in several complex diseases [15]. When significance thresholds were varied from 10^{-5} to 0.8, discriminative ability improved for bipolar depression, coronary heart disease, hypertension and type 2 diabetes, but decreased for rheumatoid arthritis and type 1 diabetes prediction. For all diseases, the discriminative ability was lower than what would be obtained when testing known susceptibility genes, except for hypertension where no susceptibility variants were known at the time and for bipolar disorder for which the score performed better than the known variants, but only for the liberal and not for the stringent significance thresholds. It is also suggested that the shared genetic liability between schizophrenia and bipolar disorder would make the genetic based refinement of the diagnosis of these

diseases possible, which may also be tried for other diseases with overlapping symptoms [13]. Of all potential applications of the genetic scoring method, this is the least substantiated, and it may still be too premature [16]. However, when the proportion of variance explained by the scores can increase from 3 to 20%, as suggested by simulation analyses by improving the power of the discovery set [13], the discriminative accuracy could be in the range of what we commonly see for non-genetic risk prediction models in e.g., cardiovascular diseases, diabetes and mortality [17–19].

GWAS have been very successful in finding multiple variants for many traits, but we are reaching the limits of what can be found through this approach sooner or later. Whether the genetic scoring method will be successful in finding more risk variants for complex traits and in unraveling 'missing heritability' remains to be determined. The new genetic score method is one approach, approaches aiming at testing of more complex models with gene by gene and gene by environment interactions may be another avenue. Last but not least technological developments may come to rescue with new development in high throughput sequencing.

References

1. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest. 2008;118:1590–605.
2. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet. 2008;40:638–45.
3. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet. 2009;41:47–55.
4. Aulchenko YS, Struchalin MV, Belonogova NM, Axenovich TI, Weedon MN, Hofman A et al. Predicting human height by Victorian and genomic methods. Eur J Hum Genet. 2009;17:1070–5.
5. Isaacs A, Sayed-Tabatabaei FA, Aulchenko YS, Zillikens MC, Sijbrands EJ, Schut AF, et al. Heritabilities, apolipoprotein E, and effects of inbreeding on plasma lipids in a genetically isolated population: the Erasmus Rucphen Family Study. Eur J Epidemiol. 2007;22:99–105.
6. Weedon MN, Frayling TM. Reaching new heights: insights into the genetics of human stature. Trends Genet. 2008;24:595–603.
7. Hu N, Wang C, Hu Y, Yang HH, Giffen C, Tang ZZ, et al. Genome-wide association study in esophageal cancer using GeneChip mapping 10 K array. Cancer Res. 2005;65:2542–6.
8. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, Helgadottir A, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet. 2009;41:18–24.
9. Dawber TR, Meadors GF, Moore FE Jr. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health. 1951;41:279–81.
10. Hofman A, Breteler MM, van Duijn CM, Krestin GP, Pols HA, Stricker BH, et al. The Rotterdam Study: objectives and design update. Eur J Epidemiol. 2007;22:819–29.
11. Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet. 2009;2:73–80.
12. Stolk RP, Rosmalen JG, Postma DS, de Boer RA, Navis G, Slaets JP, et al. Universal risk factors for multifactorial diseases: LifeLines: a three-generation population-based study. Eur J Epidemiol. 2008;23:67–74.
13. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460:748–52.
14. Miettinen OS. Up from 'false positives' in genetic-and other-epidemiology. Eur J Epidemiol. 2009;24:1–5.
15. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet. 2009.
16. Janssens AC. Is the time right for translation research in genomics? Eur J Epidemiol. 2008;23:707–10.
17. Drame M, Novella JL, Lang PO, Somme D, Jovenin N, Laniece I, et al. Derivation and validation of a mortality-risk index from a cohort of frail elderly patients hospitalised in medical wards via emergencies: the SAFES study. Eur J Epidemiol. 2008;23:783–91.
18. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med. 2006;8:395–400.
19. van Hoek M, Dehghan A, Witteman JC, van Duijn CM, Uitterlinden AG, Oostra BA, et al. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes. 2008;57:3122–8.