Missing heritability of complex traits and G-E interactions

André Scherag[1]

1 Institute for Medical Informatics, Biometry and Epidemiology, University of Duisburg-Essen, Essen, Germany

Missing heritability of complex traits and G-E interactions

The development of high-density single nucleotide polymorphism (SNP) arrays has resulted in a plethora of new molecular SNP markers robustly associated with complex traits [1], [2]. The success of these genome-wide association studies (GWAS) is largely based on stringent significance levels ($\alpha = 5 \times 10^{-8}$) and high statistical power due to meta-analytic summaries of individual cohort results in large-scale consortia. Moreover, consistent replication has become the gold standard for GWAS publications. A catalogue of GWAS results is available at http://www.genome.gov/gwastudies/ [3].

Heritability and genome-wide association studies

Despite the success of GWAS to detect many new and robust SNP associations for complex traits, GWAS have also been frequently criticized [4]. One example is the effect size of the molecular markers which is often quite small – e.g., odds ratios per effect allele <1.2 in case-control GWAS – although considerably larger effects have been reported for phenotypes that are "closer to the biology" such as metabolic outcomes [5]. For the purpose of this report but without loss of generality, let us focus on the body-mass-index (BMI measured in kg/m^2 units) as an example of a complex quantitative trait. Using a simple linear model for the quantitative trait BMI as outcome Y and SNP genotype as predictor X coded as 0, 1, and 2 to quantify the presence of 0, 1 or 2 effect (e.g., BMI increasing) alleles in one individual i one may write:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

where β_0 is the intercept while β_1 refers to the "dosage" effect of one allele of a single SNP assuming an additive genetic model. In this simple model, environmental effects are assumed to be part of ε_i (a normally distributed error term). Alternatively this model may be extended to include (environmental) covariate information. The narrow sense heritability (the heritability under an additive genetic model) for a single SNP can be estimated by the proportion of variance of the BMI that can be "explained" by the SNP alleles as compared to the total variability of the BMI. Model (1) can be generalized to an oligogenic model which includes more than a single SNP. For J SNPs one may write:

$$Y_i = \beta_0 + \sum_{j=1}^{J} \beta_j X_{ij} + \varepsilon_i \quad (2)$$

where each allele effect size β_j is allowed to be different; in a modified version of model (2) effects sizes are sometimes forced to be the same. Given that the effects are often of similar (small) size, fewer parameters need to be estimated.

In a GWAS meta-analysis by the Genetic Investigation of Anthropometric Traits (GIANT) consortium including 249,796 individuals, 32 SNP alleles were found to be robustly associated with BMI variability [6]. The frequency of effect alleles ranged between 0.04 and 0.83 while the effect sizes (change of BMI per effect allele) ranged between 0.06 and 0.39 kg/m^2. Furthermore, the narrow sense heritably estimates for each SNP ranged between <0.01% and 0.34%. When estimating the narrow sense heritability across all 32 SNPs in a model similar to model (2) this changed to ~1.5%. This number is in striking contrast to the heritability estimates derived from formal genetic studies of BMI such as twin, family and adoption studies (reviewed in [7]). In these formal genetic studies in which no molecular data had been utilized, heritability estimates ranged between 40% and 70% [6], [7]. The gap between GWAS-based and formal genetics based narrow sense heritability estimates has been observed for many other complex traits and has been referred to as "missing heritability" [8], [9], [10].

G-E interactions as an explanation for missing heritability

Many explanations have been provided for the "missing heritability" [8], [9], [10]. The usual explanations deal with the limited perspective on genetic variation when focussing on SNPs, the choice of the analysed phenotypes, more complex inheritance patterns including epigenetic processes or the choice of the statistical model [11]. In addition, interactions have also been identified

as a possible culprit. As both "heritability" and gene-environment (G-E) issues have been extensively discussed in the past, it is beyond the scope of this short article to provide a comprehensive overview (for a review see [12]). However, referring to the landmark paper by Lewontin in 1974 [13] it is obvious that all modelling assumptions that we should usually check by model diagnostics also provide the limits for G-E assessments. Quoting Lewontin "...The simple analysis of variance is useless for these purposes [the analysis into genetic and environmental components of variation] and indeed it has no use at all. In view of the terrible mischief that has been done by confusing the spatiotemporally local analysis of variance with the global analysis of causes, I suggest that we stop the endless search for better methods of estimating useless quantities. There are plenty of real problems." Despite the awareness of this general problem, new methods have been proposed to screen for biologically plausible interactions using statistical methodology (for a review see [14]). For our BMI example the first robust G-E findings have now been published [15] showing that the effect of the variant with the strongest effect in GWAS is attenuated in the physically active individuals. Genome-wide G-E assessments e.g., focussing on interactions with physical activity or smoking will be the next step.

From missing to hidden and phantom heritability?

Using a more parsimonious model ignoring statistical interactions, Peter Visscher and colleagues [16], [17] have estimated the variance explained by all SNPs together using a linear mixed model framework instead of focussing only on those SNPs that meet a stringent significance threshold. For BMI they report a narrow sense heritability of ~16% when using all autosomal SNPs which is closer to the estimates of formal genetic studies. This finding has been used to introduce the term "hidden heritability" which simply means that there are more SNPs truly associated with the complex trait of interest which have not been discovered yet. These polymorphisms are likely either less frequent variants or variants with even smaller genetic effects. Alternatively, Eric S. Lander and colleagues [18] have introduced the term "phantom heritability". They argue that models including gene x gene interactions are also consistent with the available empirical data. Given the presence of such interactions the "missing heritability" gap will become much smaller.

If a parsimonious model of additive genetic effects or a more complex model truly reflects large parts of the underlying biology will be part of future discussions on the genetic architecture of complex traits once whole genome sequencing will become reality in large scale consortia. Most likely the answer will be different for different phenotypes. Apart from this theoretical discussion a recent finding for body height should warn us [19]. Makowsky et al. [19] used all SNPs and derived whole genome prediction models built in a training data set. This prediction model based on all SNPS was sub-

sequently applied to an independent test data set from the same population to predict body height. Based on the samples they used, predictions were dramatically worse (a reduction of about 80% in terms of variance explained) in the test data set as compared to the variance explained in the training set. If replicated in larger samples and for disease phenotypes, such a finding will also show the practical limits of predictive genetic tests using SNPs [20]. The currently discussed genetic risk scores for complex diseases so far only implement a few SNPs. Given the relatively poor performance of most of these scores improved performance is often expected the more SNPs are included in the score. The finding by Makowsky et al. [19] shows that this may not be the case. However, a different picture may arise for rare variants. In this field where standard statistical methodology relying on asymptotic properties frequently fails, rigorous statistical evaluation and detailed reporting following guidelines like GRIPS [20] or REMARK [21] is an urgent request. Most importantly, the "incorporation of the underlying biology into our conceptual models" citing Duncan C. Thomas [22] will become more and more central to the field of genetic epidemiology.

Notes

Competing interests

The author declares that he has no competing interests.

References

1. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010;363(2):166-76. DOI: 10.1056/NEJMra0905980

2. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7-24. DOI: 10.1016/j.ajhg.2011.11.029

3. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009;106(23):9362-7. DOI: 10.1073/pnas.0903103106

4. Goldstein DB. Common genetic variation and human traits. N Engl J Med. 2009;360(17):1696-8. DOI: 10.1056/NEJMp0806284

5. Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, Pistis G, Serbanovic-Canic J, Elling U, Goodall AH, Labrune Y, et al. New gene functions in megakaryopoiesis and platelet formation. Nature. 2011;480(7376):201-8. DOI: 10.1038/nature10659

6. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Mägi R, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010;42(11):937-48. DOI: 10.1038/ng.686

7. Maes HH, Neale MC, Eaves LJ. Genetic and environmental factors in relative body weight and human adiposity. Behav Genet. 1997;27(4):325-51. DOI: 10.1023/A:1025635913927

8. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008;456(7218):18-21. DOI: 10.1038/456018a

9. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53. DOI: 10.1038/nature08494

10. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446-50. DOI: 10.1038/nrg2809

11. Pütter C, Pechlivanis S, Nöthen MM, Jöckel KH, Wichmann HE, Scherag A. Missing heritability in the tails of quantitative traits? A simulation study on the impact of slightly altered true genetic models. Hum Hered. 2011;72(3):173-81. DOI: 10.1159/000332824

12. Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, Schäfer H. Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet. 2008;16(10):1164-72. DOI: 10.1038/ejhg.2008.106

13. Lewontin RC. The analysis of variance and the analysis of causes. 1974. Int J Epidemiol. 2006;35(3):520-5. DOI: 10.1093/ije/dyl062

14. Thomas D. Gene–environment-wide association studies: emerging approaches. Nat Rev Genet. 2010;11(4):259-72. DOI: 10.1038/nrg2764

15. Kilpeläinen TO, Qi L, Brage S, Sharp SJ, Sonestedt E, Demerath E, Ahmad T, Mora S, Kaakinen M, Sandholt CH, et al. Physical activity attenuates the influence of FTO variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. PLoS Med. 2011;8(11):e1001116. DOI: 10.1371/journal.pmed.1001116

16. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011;43(6):519-25. DOI: 10.1038/ng.823

17. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76-82. DOI: 10.1016/j.ajhg.2010.11.011

18. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci USA. 2012;109(4):1193-8. DOI: 10.1073/pnas.1119675109

19. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G. Beyond missing heritability: prediction of complex traits. PLoS Genet. 2011;7(4):e1002051. DOI: 10.1371/journal.pgen.1002051

20. Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ; GRIPS Group. Strengthening the reporting of genetic risk prediction studies: the GRIPS Statement. PLoS Med. 2011;8(3):e1000420. DOI: 10.1371/journal.pmed.1000420

21. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. PLoS Med. 2012;9(5):e1001216. DOI: 10.1371/journal.pmed.1001216

22. Thomas DC. Genetic epidemiology with a capital E: where will we be in another 10 years?. Genet Epidemiol. 2012;36(3):179-82. DOI: 10.1002/gepi.21612

Corresponding author:

Dr. André Scherag
Institute for Medical Informatics, Biometry and Epidemiology, University of Duisburg-Essen, Hufelandstraße 55, Essen, Germany
andre.scherag@uk-essen.de