The nature of confounding in genome-wide association studies

Bjarni J. Vilhjálmsson[1,2] *and Magnus Nordborg*[3,4]

The authors argue that population structure per se is not a problem in genome-wide association studies — the true sources are the environment and the genetic background, and the latter is greatly underappreciated. They conclude that mixed models effectively address this issue.

> “population structure is not the fundamental source of the problem, and removing it is not the solution”

[1]*Harvard School of Public Health, Boston, Massachusetts 02115, USA.*
[2]*The Broad Institute, Cambridge, Massachusetts 02143, USA.*
[3]*Gregor Mendel Institute, Austrian Academy of Sciences, 1030 Vienna, Austria.*
[4]*Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089, USA.*
e-mails: bvilhjal@hsph.harvard.edu; magnus.nordborg@gmi.oeaw.ac.at
doi:10.1038/nrg3382
Published online 20 November 2012

Thanks to dramatically decreasing genotyping and sequencing costs, genome-wide association studies (GWASs) are becoming the default method for studying the genetics of natural variation. The increasing number and diversity of GWASs will require appropriate statistical analysis methods. The most basic problem is assessing the significance of an association in the light of confounding effects that may cause spurious associations.

The aspect of this problem that has received the most attention is the danger of false positives in structured populations. If the study population is a mixture of populations that differ with respect to allele frequencies as well as the trait of interest, spurious correlations will arise. To take a classic example, a GWAS for skill with chopsticks carried out in San Francisco might identify human leukocyte antigen A1 (*HLA-A1*) as an allele associated with chopstick skill simply because this allele is more common in people of East Asian origin[1].

Thus stated, the problem is straightforward. Population structure acts as a confounding factor that must either be eliminated through better study design or controlled in the analysis[2–4]. However, although compelling, the chopstick example is actually misleading: population structure is not the fundamental source of the problem, and removing it is not the solution.

The source of confounding in the chopstick example is better thought of as the environment. The problem arises because different subgroups have different levels of exposure to chopsticks. This type of confounding is extremely familiar to genetic epidemiologists, but it is unimportant in settings where the environment can be experimentally controlled or randomized (as is routinely done in plant breeding, for example).

There is another source of confounding, however, and that is the genetic background. The estimate of the effect of a particular locus can be confounded by the other causal loci in the genome. This genetic background effect will always be present to some extent, even in ‘unrelated’ individuals. Variation in relatedness is a basic property of natural populations, as is correlation between causative loci. This issue is familiar to quantitative geneticists[5] but has not been widely appreciated in other fields. It is important for GWASs and will become crucial as sample sizes increase.

To demonstrate this, let us return to the chopstick example but fast-forward to the era of millions of SNPs. Genetic differentiation between East Asians and other populations means that vast numbers of markers in addition to *HLA-A1* would be associated with chopstick skill. These markers would also be correlated with *HLA-A1*, with each other and with any trait (genetic or not) that differed systematically between East Asians and other populations. A naive GWAS of any such trait would identify large numbers of false positives in addition to the true positives. For some traits, confounding would be due both to differences in environmental factors across groups and the genetic background, whereas for other traits (for example, eye colour), the environmental effect would be negligible. Only for traits that really have no genetic basis (such as chopstick skills) could we ignore the genetic background.

But why distinguish between genetic and environmental confounding if both can be controlled using population structure as a proxy? The answer is that this approach will work well only in very simple cases. An obvious problem is that San Francisco is not just a simple mixture of two (or more) homogeneous populations. More generally, there really is no such a thing as a homogeneous population. As discussed above, any sample will contain various levels of relatedness, and whether this matters depends not only on the magnitude of this variation but also on the genetics of the trait. A genetic background effect that is trivial compared to the marginal effect of a major locus may be enormous compared to the effect of a minor locus and may even be strong enough to cause false positives at loci with no true effect.

> the underlying sources of confounding in GWASs are environmental and genetic

What, then, is the solution? In a groundbreaking 1918 paper, Fisher[6] showed how the phenotypic covariance between relatives depends on their genetic relatedness, assuming that phenotypic variation was due to the additive effects of a large number of loci of small effect. The basic idea is very simple: the more alleles that individuals share, the more similar they will be. In human genetics, the same logic underlies the Haseman–Elston regression[7], and in animal breeding the classic Henderson mixed model[8] has been used to reduce the confounding effects of genetic background when mapping in pedigrees[5].

By estimating relatedness from SNPs instead of pedigrees, Henderson's model can also be used in GWASs[9]. There have been two distinct applications. In the first, a mixed model is used to assess the effect of a specific locus while controlling the effect of the genetic background by estimating genome-wide relatedness[9]. This approach has been shown to outperform methods that try to estimate population structure directly and to include it as a fixed effect[9–12]. The second application focuses on estimating the heritability (again via genome-wide relatedness) and does not try to map any genes. This approach has been used to demonstrate that SNP variation accounts for a substantial fraction of human height, despite the fact that the variants identified in GWASs jointly explain very little (the so-called 'missing heritability')[13].

Importantly, while both approaches model the genetic background via estimates of pairwise relatedness, relatedness is really only a proxy for allele sharing at causative loci. Thus, although any sample from a natural population will contain different levels of relatedness, the mixed model approach would work even in an idealized population without any such differences. If we simulate a sample by independently drawing each individual from the population allele frequencies, there will still be stochastic differences in allele sharing between individuals, and these can be captured using Henderson's model, at least as long as the basic assumptions hold[6]. Simulations suggest that this will be the case under a range of genetic architectures, at least as long as interactions are additive[14,15].

Deviations from additivity may cause biases[16], and selection can clearly have a dramatic effect, because it will, by definition, make the causal polymorphisms have a different distribution than the non-causal ones, and there is thus no reason to believe that the latter would predict the phenotypic covariance well. Under these conditions at least, other methods for reducing confounding may be required. Examples include flowering time in maize and *Arabidopsis thaliana*[9,10], in which mixed models that included both relatedness and direct estimates of population structure were found to outperform models that included relatedness only. The probable reason is that variation for these traits is partly due to genes that have been under very strong selection and hence have a different distribution from the rest of the genome[9]. Indeed, given that the background genes are the true confounding factors, they should be included if possible[17,18]. Linkage disequilibrium between closely linked causative sites (as is often observed in cases of allelic heterogeneity), especially in combination with background confounding, can make fine-mapping extremely challenging[17,19,20]. The problem is, of course, that we usually do not know what the causal loci are, and methods that try to identify them are prone to over-fitting. Nonetheless, including known causal polymorphisms, either as fixed effects[14] or as estimators of the phenotypic covariance matrix (an approach that is closely related to Bayesian linear regression)[21,22], may greatly increase power. Intuitively, the more we know about the genetics of a trait, the greater our power is to detect the rest of the genetic contribution.

To conclude, the underlying sources of confounding in GWASs are environmental and genetic. Population structure per se is not the problem, nor is relatedness: estimates of either can help us to reduce confounding, but to do this well, it is helpful to understand its true source. Mixed models that attempt to describe phenotypic covariance are a natural way to model this confounding. They have a solid mechanistic basis, and the variance components estimated are easily interpreted, allowing us to distinguish genetic from environmental components[23,24].

1. Lander, E. S. & Schork, N. J. *Science* **265**, 2037–2048 (1994).
2. Devlin, B. & Roeder, K. *Biometrics* **55**, 997–1004 (1999).
3. Pritchard, J. K. *et al. Am. J. Hum. Genet.* **67**, 170–181 (2000).
4. Price, A. L. *et al. Nature Genet.* **38**, 904–909 (2006).
5. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, 1998).
6. Fisher, R. A. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
7. Haseman, J. K. & Elston, R. C. *Behav. Genet.* **2**, 3–19 (1972).
8. Henderson, C. R. *Applications of Linear Models in Animal Breeding* (Univ. Guelph Press, 1984).
9. Yu, J. *et al. Nature Genet.* **38**, 203–208 (2006).
10. Zhao, K. *et al. PLoS Genet.* **3**, e4 (2007).
11. Kang, H. M. *et al. Nature Genet.* **42**, 348–354 (2010).
12. Price, A. L. *et al. Nature Rev. Genet.* **11**, 459–463 (2010).
13. Yang, J. *et al. Nature Genet.* **42**, 565–569 (2010).
14. Segura, V. *et al. Nature Genet.* **44**, 825–830 (2012).
15. Zaitlen, N. & Kraft, P. *Hum. Genet.* **131**, 1655–1664 (2012).
16. Zuk, O. *et al. Proc. Natl Acad. Sci. USA* **109**, 1193–1198 (2012).
17. Atwell, S. *et al. Nature* **465**, 627–631 (2010).
18. Platt, A., Vilhjálmsson, B. J. & Nordborg, M. *Genetics* **186**, 1045–1052 (2010).
19. Dickson, S. P. *et al. PLoS Biol.* **8**, e1000294 (2010).
20. Huang, X. *et al. Nature Genet.* **42**, 961–967 (2010).
21. Listgarten, J. *et al. Nature Methods* **9**, 525–526 (2012).
22. Zhou, X., Carbonetto, P. & Stephens, M. Preprint at arXiv [online], http://arxiv.org/pdf/1209.1341v1.pdf (2012).
23. Deary, I. J. *et al. Nature* **482**, 212–215 (2012).
24. Korte, A. *et al. Nature Genet.* **44**, 1066–1071 (2012).

Acknowledgements

We thank E. Buckler, D. Balding, P. Donnelly, A. Hancock, I. Hellmann, M. Horton, A. Korte, Q. Long, D. Meng, N. Patterson, A. Platt, A. Price, V. Segura, O. Stegle and Q. Zhang for discussions and/or comments on the manuscript. We are especially grateful to A. Price for sharing the observation that mixed models can explain a substantial fraction of the phenotypic covariance in randomly generated individuals. Remaining errors or omissions are our responsibility. This work was supported by US National Institutes of Health grant HG002790, European Research Council grant AdG-268962 and the Gregor Mendel Institute. We apologize to authors whose work could not be cited owing to space constraints.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Bjarni J. Vilhjálmsson's homepage: http://sites.google.com/site/bjarnijvilhjalmsson

Magnus Nordborg's homepage: http://www.gmi.oeaw.ac.at/research-groups/magnus-nordborg

ALL LINKS ARE ACTIVE IN THE ONLINE PDF