
Article:
Slate, J. orcid.org/0000-0003-3356-5123 (2017) Robust inference of genetic architecture in mapping studies. Molecular Ecology, 26 (6). pp. 1453-1455. ISSN 0962-1083

https://doi.org/10.1111/mec.14052

Robust inference of genetic architecture in mapping studies

Jon Slate

Department of Animal & Plant Sciences, University of Sheffield, Sheffield, S10 2TN, UK

j.slate@sheffield.ac.uk

Due to the development of next-generation sequencing and related tools, the feasibility of gene mapping studies in wild populations has improved dramatically. However, phenotypic data collection remains challenging and sample sizes are typically orders of magnitude smaller than are seen in genome wide association studies (GWAS) of human populations, where hundreds of thousands of people may be screened (Wood *et al.* 2014). Consequently, the power to detect quantitative trait loci (QTL) remains modest, unless the focal trait is segregating for genes of major effect. Worse still, small sample sizes result in crude estimates of effect sizes; those that are mostly severely overestimated, will be the ones most likely to reach statistical significance - the well-known 'Beavis Effect' (Beavis 1994). This makes inference from mapping studies very problematic. Does a significant peak with a large effect on phenotypic variation represent a true hit, with the trait being determined by relatively few genes of large effect (an 'oligogenic' architecture)? Or does the peak simply represent an upwardly biased estimate and a false positive QTL? Without replication, it is very hard to know which scenario is true. Similarly, interpretation of a null result (no significant QTL) can be problematic. Does this mean that the study was underpowered to pick up any medium-large effect loci that are present? Or does the trait have a genuinely polygenic architecture, caused by many loci of small effect, each of which is undetectable in that particular experiment? It's not hard to see why there remains scepticism about the value of gene mapping studies to evolutionary research (Rockman 2012; Travisano & Shaw 2013). In this issue of Molecular Ecology, Li *et al.* (2017) describe a gene mapping experiment that goes some way to addressing the issues of low power and inflated effect size that have plagued previous studies. The authors have conducted a mapping study of brain traits in nine-spined sticklebacks *Pungitius pungitius* to tackle two alternative hypotheses about the genetic architecture of brain morphology. At face value, the study could be seen as a relatively standard GWAS, albeit with an impressive number of markers for what is not a classical model organism. However, scratch beneath the surface a little, and it becomes clear that some sophisticated analytical approaches have been used to try to understand trait architecture in a more rigorous way than is typical.

In the study Li and colleagues used an F2 mapping population, derived by crossing a marine female from the Baltic Sea with a freshwater male from a pond in Northern Finland. There were 239 phenotyped and genotyped F2 individuals, and a little over 15,000 SNPs obtained from genotyping-by-sequencing, which were mapped to 21 linkage groups (the known number of chromosomes in this species). The authors measured the volume of five different parts of the brain, and were interested in comparing two alternative hypotheses about the genetic architecture of brain traits. Under the *mosaic model* each brain component has a distinct genetic architecture, and it is free to evolve without genetic constraint from other brain components. The alternative idea, the *concerted model*, posits that brain component evolution is constrained, perhaps due to a common genetic architecture influencing the different parts. Recognising that the experimental design was exactly the kind where QTL of major effect could be identified spuriously in a standard linkage mapping (or GWAS) experiment, the authors utilised an approach known as *de-biased Least Absolute Shrinkage and Selection Operator (LASSO) mapping* (Van de Geer et al. 2014; Zhang & Zhang 2014). De-biased LASSO

has not been widely employed in gene-mapping studies but it has several advantageous properties. Perhaps the most obviously different feature is that multiple markers are modelled simultaneously. This reduces the risk of effect size overestimation, and can also lead to an increase in power. It also facilitates the estimation of heritability by summing the effect of SNPs fitted in the model. Thus, Li *et al.* (2017) had two main aims. The empirical goal was to understand the genetic architecture of brain traits and evaluate whether the mosaic or concerted model was more plausible. The methodological goal was to compare different mapping approaches using both real and simulated data, to establish whether de-biased LASSO gave more reliable parameter estimates than approaches that fit single markers consecutively.

For many of the traits that Li and colleagues studied, they found genomic regions that explained significant genetic variation, even at a stringent genomewide significance threshold. There was very little between-trait overlap in QTL locations, so the data were consistent with the mosaic model (different genomic regions affecting different brain components). When running single SNP (i.e. conventional) analyses, there were frequently numerous, tightly-linked significant SNPs, and their estimated effect sizes were frequently 5-10% of the overall trait variation. These would be regarded as genes of reasonably large effect. The multi-marker (i.e. de-biased LASSO) approach usually identified the same genomic regions. However, the effect sizes were typically much smaller - 1% or less of the phenotypic variation. The multimarker analyses could also measure the effect of each linkage group on trait variation. Summing these effects across linkage groups provided estimates of trait heritability; depending on the trait these ranged from ~0.10 - ~0.45. Some linkage groups contributed disproportionately to additive genetic variation, but not in a way that supported the concerted model (which would have predicted that the same linkage group would contribute to different traits). Overall, the data suggest that brain traits are moderately heritable in this cross, and that the trait architecture is consistent with the mosaic model of brain evolution. One slight caveat is that the experimental cross was derived from just a single pair of fish, and we are largely ignorant of how much genetic variation is segregating within *versus* between marine and freshwater populations. Other experimental designs may have yielded quite different conclusions.

How well does the multi-marker de-biased LASSO method perform? The simulations, which are presented in the supplementary material, explored both an oligogenic and a polygenic scenario. Unsurprisingly, the single-locus approach had a high false positive rate, and a lower power to detect true positives than de-biased LASSO (although neither approach had high power when the simulated QTL were small). Effect size estimation of individual QTL was actually greatly downward-biased with de-biased LASSO and tended to be upwardly biased with single-locus estimates. Encouragingly though, de-biased LASSO provided accurate estimates of overall trait heritability, regardless of trait architecture. Single-locus approaches fail in this regard, especially when the true architecture is polygenic. It may well be the case that the downward bias of de-biased LASSO effect size estimates can be rectified by summing the effects of linked SNPs in the region, or by first pruning the marker data, so that retained SNPs are not in high linkage disequilibrium. The F2 design used by Li and colleagues, probably causes linked SNPs to be in strong linkage disequilibrium. More generally, de-biased LASSO is one of several recently introduced approaches that fit multiple markers simultaneously (Moser *et al.* 2015; Zhou *et al.* 2013). These methods are beginning to be adopted in evolutionary / ecological studies (Comeault *et al.* 2015) and they are attractive for several reasons. First, they facilitate a more holistic approach to studying trait architecture, where instead of paying slavish attention to 'significant peaks', QTL effect sizes, heritabilities, individual breeding values and

whole-chromosome contributions (*sensu* Yang et al. (2011) to genetic variation can be estimated simultaneously. Second, they go a long way towards avoiding incorrect inference of an oligogenic trait architecture that almost inevitably comes about from upwardly-biased single-marker effect size estimates of true or false QTLs. It remains to be seen which multimarker method performs best, although benchmarking studies of some approaches do exist (Moser *et al.* 2015). Perhaps, the most heartening thing is that ecological genetic mapping studies such as the one by Li and colleagues, are beginning to mature in a way that hypothesis-driven questions can be addressed without the reliance on the detection and identification of specific QTL that may or may not true positives.

References

Beavis WD (1994) The power and deceit of QTL experiments: lessons from comparative QTL studies, 250-266.

Comeault AA, Flaxman SM, Riesch R*, et al.* (2015) Selection on a Genetic Polymorphism Counteracts Ecological Speciation in a Stick Insect. *Current Biology* **25**, 1975-1981.

Li Z, Guo B, Yang J*, et al.* (2017) Deciphering the genomic architecture of the stickleback brain with a novel multi-locus gene-mapping approach. *Molecular Ecology* **26**, xxx-xxx.

Moser G, Lee SH, Hayes BJ*, et al.* (2015) Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *Plos Genetics* **11**.

Rockman MV (2012) THE QTN PROGRAM AND THE ALLELES THAT MATTER FOR EVOLUTION: ALL THAT'S GOLD DOES NOT GLITTER. *Evolution* **66**, 1-17.

Travisano M, Shaw RG (2013) LOST IN THE MAP. *Evolution* **67**, 305-314.

Van de Geer S, Buhlmann P, Ritov Y, Dezeure R (2014) On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models. *Annals of Statistics* **42**, 1166-1202.

Wood AR, Esko T, Yang J*, et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-1186.

Yang J, Manolio TA, Pasquale LR*, et al.* (2011) Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43**, 519-525.

Zhang CH, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **76**, 217-242.

Zhou X, Carbonetto P, Stephens M (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *Plos Genetics* **9**.