Limitations of QTL Mapping

Reading:


Supplementary Reading:


We have already touched on some of the problem areas of QTL mapping, including how does one decide whether a genome region carries a QTL or not (setting a testing threshold), problems of sample size, and issues in model building. These questions have been most reliably explored in simulation studies and in a few experimental studies using very large population sizes.

Beavis’ (1998) study on “power, precision, and accuracy” of QTL mapping is a classic. His work (first reported at a symposium in 1994) was the first demonstration of the effects of small sample sizes on the reliability of QTL mapping. This work demonstrated that when sample sizes (n) were small, power to detect true QTL was low, and the bias in the estimated proportion of the genetic variance explained by mapped QTLs was huge:

<table>
<thead>
<tr>
<th># of true QTL</th>
<th>h²</th>
<th>n</th>
<th>Power</th>
<th>Bias of $\hat{\sigma}_G^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.30</td>
<td>100</td>
<td>9%</td>
<td>+559%</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>500</td>
<td>57%</td>
<td>+144%</td>
</tr>
<tr>
<td>10</td>
<td>0.95</td>
<td>100</td>
<td>39%</td>
<td>+197%</td>
</tr>
<tr>
<td>40</td>
<td>0.95</td>
<td>500</td>
<td>46%</td>
<td>+165%</td>
</tr>
</tbody>
</table>

The bias in estimated genetic variance occurs mainly due to sampling of small populations; the true QTLs that are not detected (most of them in small sample sizes) have effects that are detected at the regions that are detected as having QTLs. So the declared QTL regions have effects that appear much larger than they really are. This phenomenon is known as “The Beavis Effect”.
Note carefully that this study was based on a simulation of F2 populations. The test statistics (such as the t-test in a QTL regression) for additive effects are expected to be only equal between F2 populations of size \( n \) and homozygous (RIL or doubled haploid) populations of size \( 2n \). In other words, statistically you should be substantially better off using RILs than F2s if population sizes are identical.

The implications for Beavis’ study for setting threshold levels for declaring QTLs are not entirely clear. Since power of detection suffers badly with smaller population sizes, perhaps Type I error control should not be a major concern. Rather, the major concern might actually be Type II error and the bias on estimating QTL effects.

An alternative to trying to control Type I error rates using genome-wise error control (such as permutation tests) is the false discovery rate. The false discovery rate (FDR) is the proportion of false positives among all tests that are declared significant. The FDR is not as stringent as genome-wise error control because you are accepting that some proportion of the QTLs you find will be false positives (rather than trying to keep the number of false positives at zero). Benjamini and Yekutieli (2005) present strong theoretical arguments for the use of the FDR in QTL mapping and demonstrate several ways in which it can be calculated, given a real data set.

Bernardo (2004) conducted a simulation study to determine the relationship between the single-marker threshold alpha value and the FDR using a simulation study. He simulated different numbers of QTLs and different heritabilities in an F2 population of 150 individuals. Then he conducted QTL tests and asked the question: what proportion of the QTLs that I found at different alpha values were false positives? His answer was that, if there were few true QTL, the FDR tended to be quite high (e.g., FDR of about 80% if alpha = 0.05 is used and heritability is 80% and only one QTL exists). FDR tended to decrease with more true QTL and with higher heritabilities. In addition, the alpha value had a big impact on FDR (as one might expect). Bernardo concluded that alpha values of 0.0001 be used in QTL mapping to ensure that FDR levels would most likely be 0.01 or lower.

I have some uncertainties about this study. First, the simulation again was based on F2 populations, which are equivalent in some ways to RIL populations of half the size (e.g., an F2 population of 150 will provide an equivalent t-value as an RIL population of 75 for a given true QTL). Second, he simulated a multiple regression procedure, but it is not clear at what stage the alpha level he talks about is used. It is clearly used in the first stage, each chromosome is tested one-at-a-time. For a given chromosome, all markers are tested together, then backwards elimination is used to drop the number of markers down to a set that is significant at the given alpha level. Then a multiple regression model is built that includes all of the markers that remained significant in the individual chromosome analyses. It is not clear if the markers are then re-tested and dropped if they are not at the declared alpha level. It seems like this was not done, in which case, one would expect higher FDR. (Consider the result stated above where there is only one true QTL; one would expect the marker nearest that QTL would fit in a final model but knock out most of the markers on other chromosomes in the final model fit, thus reducing the
FDR dramatically). Finally, the desire for an FDR of no more than 1% seems stringent to me. The appropriate FDR level should be up to the experimenter and readers can decide themselves if they are willing to believe in a QTL declared at a stated FDR level. Based on his results, the heritability of the trait has a big impact on FDR, and this is an estimable quantity. So, researchers should estimate the heritability in their experiment, then decide based on that what an appropriate FDR is, given some guesses about the number of true QTL for the trait. For example, if heritability is 80% and there are ten true QTL an alpha value of 0.005 would give an FDR of about 10%. This seems reasonable, and this FDR level is recommended by Bejamini and Yekutieli (2005).

The studies above were based on simulations: what happens in real experiments? Schon et al. (2004) addressed this question by analyzing a population of 990 F5 lines from a cross of two maize lines in 19 environments. They used cross-validation to estimate the bias in the estimates of the genetic variance associated with markers. With this huge data set, they detected 18 QTL for grain yield, but only about 42% of the genetic variation could be explained by the mapped QTL. The cross-validation results suggest that even this estimate is biased upward…on average, only associate about 30% of the genotypic variation in validation subsets of the population was explained by QTLs mapped in independent sets of the data. But things got markedly worse with smaller data sets. These results suggest that QTL mapping of very complex traits controlled by many genes with small effects will not typically produce reliable results.

Laurie et al. (2005) reported similar results from the next largest plant QTL study I am aware of. They mapped QTLs for oil content in 500 S1-derived lines from a maize population that was random-mated for 10 generations after crossing extremely high oil and low oil lines. They reported about 50 QTLs for oil content, but again, these mapped QTLs were associated with only about 50% of the genetic variance for oil content.