Is this a fake QTL? A short note on analysis of multiple QTL peaks on the same chromosome

Mathias Lorieux

RICE team, DIADE Research Unit, Institut de Recherche pour le Développement (IRD), Montpellier, France, and Rice Genetics and Genomics Laboratory, International Center for Tropical Agriculture (CIAT), Cali, Colombia. Email: mathias.lorieux@ird.fr, m.lorieux@cgiar.org

Abstract

A simple method is proposed to detect fake QTL provoked by spurious linkage with a true QTL. The method is based on the calculation of the expected value of the test-statistic (F-test and LOD score are considered) at the putatively fake QTL position.

The problem

In quantitative trait locus (QTL) mapping, it is frequent to observe several peaks in the plot of the test statistic (LOD score, F-test, etc.) against the linkage group position – the statistic decays sufficiently between local peaks to allow declaring each peak as a different QTL. In this case, one can legitimately ask if some of the peaks are actually spurious QTLs generated by, for example, a departure of the data from its expected structure model. Distant genomic regions on the same chromosome that show unexpected linkage strength are an example of such departures. Methods like Multiple Interval Mapping (MIM) (Jansen 1993; Kao *et al.* 1999), Composite Interval Mapping (CIM) (Zeng 1994) or Inclusive Composite Interval Mapping (ICIM) (Li *et al.* 2007) are designed to co-analyze several QTL peaks, however they might not be adequate for such situations (I'd like to leave this point open for discussion).

Method

For simplicity, let's consider the case of two QTL peaks found on a linkage group for the same trait. Suppose an F-test statistic has been obtained from the regression of phenotypes onto genotypes in a segregating population, and then plotted along a chromosome at every marker position. Suppose also that the plot shows a main peak that reveals the presence of a QTL, but also seems to reveal a second, lower QTL peak, at some distance from the main peak. One needs to decide if the secondary QTL peak is actually due to a real QTL effect, or if it is simply the result of linkage with the main QTL. Normally, if there was only one *real* QTL on a chromosome, one would expect to observe a propagation effect from the main QTL: the F-test value should decrease continuously with the genetic distance (i.e., the recombination fraction) from the true QTL position. Moreover, no other QTL local peak should show up, apart from small stochastic variations due, for example, to missing or erroneous data at some marker positions. But this is only true under the assumption that the recombination odds are following their expected values.

Sometimes, abnormally tight linkage between markers located in distant chromosomal regions can be observed. This can be due to different causes – for example, epistatic interaction of sterility genes. One can consider this situation as a special case of inadequacy between the data structure and the data model used in the QTL analysis, where the data model says that the pairwise recombination fraction between any pair of markers, and the sum of adjacent recombination fractions (SARF) of all the intervals that separate the two markers are equivalent, while in the real data they are *not*. In such cases, one could question if the secondary QTL peak is primarily explained by its spurious linkage with the main QTL peak-region. This hypothesis could be tested by a careful inspection of the *expected* value of the F-test at a given marker position, $E(F_M)$.

A way to calculate $E(F_M)$ is as follows. Consider a QTL Q and a marker M linked to Q . Let F_Q be the observed value for the F-test at the Q position, and R_Q^2 be the fraction of the trait variation explained by Q . Let also r be the recombination fraction between Q and M , n be the population size and p be the number of genotypic classes observed in the marker. Then, we know that the expected value of R^2 at any marker M is

$$E(R_M^2) = R_Q^2(1 - 2r)^2 \quad (\text{A.1})$$

Since we know from linear regression theory that

$$F = \left(\frac{df2}{df1} \right) \frac{R^2}{1-R^2} \quad (\text{A.2})$$

which leads to

$$R^2 = \frac{F}{F + \left(\frac{df2}{df1} \right)} \quad (\text{A.2b})$$

where $df2 = n - p$ and $df1 = p - 1$ are the degrees of freedom used to calculate F , we can plug (A.2b) in (A.1) to derive the expected value of the F-test at marker M

$$E(F_M) = \left(\frac{df2}{df1} \right) \left[\frac{F_Q(1-2r)^2}{F_Q + \left(\frac{df2}{df1} \right)} \right] \left/ \left[1 - \frac{F_Q(1-2r)^2}{F_Q + \left(\frac{df2}{df1} \right)} \right] \right. \quad (\text{A.3})$$

which slightly simplifies to

$$E(F_M) = \left(\frac{df2}{df1} \right) \left(\frac{F_Q(1-2r)^2}{\left(\frac{df2}{df1} \right) + F_Q - F_Q(1-2r)^2} \right) \quad (\text{A.4})$$

In backcross or double haploid populations, there are only two genotypic classes, so $df2 = n - 2$ and $df1 = 1$, and

$$E(F_M) = \frac{(n-2)F_Q(1-2r)^2}{n-2+F_Q-F_Q(1-2r)^2} \quad (\text{A.5})$$

In a recombinant inbred line (RIL) population derived by single-seed descent, r is replaced by $2r/(1 + 2r)$ in (A.5).

In F_2 populations, three genotypic classes are observed, so $df2 = n - 3$ and $df1 = 2$, and

$$E(F_M) = \frac{(n-3)F_Q(1-2r)^2}{n-3+2F_Q-2F_Q(1-2r)^2} \quad (\text{A.6})$$

To test if a second, smaller QTL is actually a fake one, we compare two different values for $E(F_M)$, $E_p(F_M)$ and $E_d(F_M)$, with the observed F-test values (by plotting all three values against the marker positions on the chromosome, expressed in centimorgans or cM). $E_p(F_M)$ and $E_d(F_M)$ are calculated using two different estimates for r , respectively: the first one, $\hat{r}_p$, is the observed *pairwise* recombination fraction between

marker M and the marker that has the highest F-test, M_Q , and the second one, $\hat{r}_d$, is derived from the sum d of the sizes in cM of the adjacent intervals defined by the markers that lay between M and M_Q . $\hat{r}_d$ is then obtained from d by inverting the mapping function – e.g., Haldane's or Kosambi's function – that was used to calculate the genetic map. In case of spurious linkage, $\hat{r}_p$ can be significantly lower than $\hat{r}_d$, which is likely to be closer to the true recombination fraction that *would* be observed in a population that follows the expected values. Note that the markers are assumed to be densely distributed within the $M - M_Q$ interval, so we avoid the "ghost QTL" effect (see Haley & Knott 1992; Martínez & Curnow 1992); this is important because we need the main QTL to be precisely located.

The difference between the $E_p(F_M)$ and $E_d(F_M)$ values reflects the effect of the discrepancy between $\hat{r}_p$ and $\hat{r}_d$ on $E(F_M)$. Consequently, if the second, smaller QTL is due to spurious linkage with the main QTL, we expect the curves for the observed F-test and its expectation $E_p(F_M)$ to have similar shapes and close values, and $E_d(F_M)$ to show a regular decrease from the QTL peak and an obvious departure from $E_p(F_M)$. This departure could be tested by calculating the probability log-ratio

$$\log \left(\frac{p(E_d)}{p(E_p)} \right) \quad (\text{A.7})$$

The expected LOD score at QTL position M , $E(Z_M)$, can be derived using a similar approach. We can use the generalized R^2 expression, proposed by Nagelkerke (1991), to derive the relationship between R^2 and Z

$$R^2 = 1 - 10^{\left(\frac{-2Z}{n}\right)} \quad (\text{A.11})$$

which leads to the following expression for $E(Z_M)$

$$E(Z_M) = -\frac{n}{2} \log_{10} \left[1 - \left[1 - 10^{\left(\frac{-2Z_Q}{n}\right)} \right] (1 - 2r)^2 \right] \quad (\text{A.12})$$

where Z_Q is the LOD score statistic at QTL position Q .

Then we can calculate $E(Z_M)$ for $\hat{r}_p$ and $\hat{r}_d$ just as we did for $E_p(F_M)$ and $E_d(F_M)$, and compare the curves to the observed LOD score in order to conclude on the presence of a fake QTL. Note, (A.12) is valid for any population type.

Figure 1 illustrates an application of the method to a QTL study of resistance to the parasitic plant *Striga hermonthica* in rice.

Figure 1. Observed and expected LOD score, for resistance to *Striga hermonthica* along a rice chromosome, in a RIL population. A secondary peak is observed at ~10 cM. Observe the resemblance of shape between the observed LOD score and its expected value $E_p(Z_M)$ calculated using pairwise recombination fractions between the QTL peak and every marker position. Also observe the difference between $E_p(Z_M)$ and $E_d(Z_M)$ at ~10 cM. We thus declare the secondary peak as a fake QTL.

Conclusion

The proposed method allows detecting a fake QTL induced by spurious linkage with a true QTL. It can be used as a complement to other methods like CIM, ICIM or MIM as an additional test for QTL presence. We haven't considered the case of a fake QTL provoked by spurious linkage with a true QTL located on a different chromosome, however we don't see a reason why the method couldn't be applicable to this case.

Additional note. The sometimes used, simpler formula $E_b(F_M) = F_Q(1 - 2r)^2$, although biased, remains a fair approximation of $E(F_M)$ for large populations since the ratio between the two estimates is, for populations with two genotypic classes

$$E(F_M)/E_b(F_M) = \frac{n-2}{n-2+F_Q-F_Q(1-2r)^2} \quad (\text{A.8})$$

and n is usually much larger than F_Q . Nevertheless, the quality of the approximation decreases as the contribution of the QTL to the trait increases.

However the bias of $E_b(F_M)$ can be quite serious for smaller n values and highly significant QTL, especially when r is large. This bias is, after simplification

$$E_b(F_M) - E(F_M) = \frac{4F_Q^2(1-2r)^2r(1-r)}{-4F_Qr^2+4F_Qr+n-2} \quad (\text{A.9})$$

And the relative bias is

$$\frac{E_b(F_M)-E(F_M)}{E(F_M)} \quad (\text{A.10})$$

For example, in a backcross population of $n = 200$ individuals, if $r = 0.45$ and $F_Q = 40$, the relative bias is 20%, and it increases to an unacceptable $\sim 40\%$ if $n = 100$.

Acknowledgements

Many thanks to Julie Scholes, James Bradley and Jon Slate from the University of Sheffield (UK) for their valuable comments and edits on the manuscript.

References

- Haley, C., & Knott, S. (1992). A simple regression method for mapping QTL in line crosses using flanking markers. *Heredity*, 69, 315–324.
- Jansen, R. (1993). Interval mapping of multiple quantitative trait loci. *Genetics*, 135, 205–211.

Kao, C. H., Zeng, Z.-B., & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics*, 152(3), 1203–1216.

Li, H., Ye, G., & Wang, J. (2007). A modified algorithm for the improvement of composite interval mapping. *Genetics*, 175, 361–374

Martinez, O., & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *TAG Theoretical and Applied Genetics*, 85(4), 1–10.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691-692.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics*, 136, 1457–1468.