

# Experimental Design and Sample Size Requirement for QTL Mapping

Zhao-Bang Zeng  
Bioinformatics Research Center  
Departments of Statistics and Genetics  
North Carolina State University  
zeng@stat.ncsu.edu

## Experimental Designs

Crosses from divergent inbred lines, populations and species

- Backcross cross (BC):
  - Two genotypes at a locus (similar to RI)
  - Simple to analyze
- F2:
  - Three genotypes at a locus, can estimate both additive and dominance effects
  - More complex for data analysis particularly for multiple QTL with epistasis
  - More opportunity and information to examine genetic structure or architecture of QTL
  - Have more power than BC for QTL analysis

- Recombinant inbred lines (RI)
  - More mapping resolution as more recombination occurred in constructing RI
  - Can improve the measurement of mean phenotype of a line with multiple individuals, i.e. can increase heritability. Potentially a very big, big advantage for QTL analysis and a big factor for power calculation and sample size requirement.

- Advanced generation of cross: F3, F4, ...
  - By selfing: lead to RI
  - By random mating: increase recombination, extend the length of linkage map, increase the mapping resolution (estimation of QTL position)
- Doubled haploid: similar to BC and RI in analysis
- Repeated backcross
- Testcross
- NC design III (marker genotype data on F2 or F3 and trait phenotype data on both backcrosses from F2 or F3)

## Other populations used for QTL analysis

- Cross from segregating populations (no inbred available):
  - Similar model and analysis procedure used as inbred cross, but more complex in analysis. Need to estimate the probability of allelic origin for each genomic point from observed markers.
  - Less powerful for QTL analysis (QTL alleles may not be preferentially fixed in the parental populations);
  - More difficult for power calculation (more unknown).

- Half sibs:
  - Analyze the segregation of one parent; similar to back-cross in model and analysis.
  - Less powerful for QTL detection – more uncontrollable variability in the other parents.
  - Analyze allelic effect difference in one parent, not the allelic effect difference between widely differentiated inbred lines, populations and species. Generally the relevant heritability is low for QTL analysis.

- Full sibs:
  - Four genotypes at a locus; can estimate allelic substitution effects for male and female parents and their interaction (dominance).
  - Doubled information for QTL analysis than half-sibs; should be more powerful.
  - Note: However, if we use the double pseudo-backcross approach for mapping analysis, we do NOT utilize full genetic information, (actually use less than half the information available). Not powerful for QTL identification.  
**Power calculation depends on how the data is analyzed.**
- Complex pedigree: go fishing

## Power and sample size calculation

First a simple case (a point for departure): One marker and One QTL for F2

Assume that the QTL genotypic effects are

$$\begin{array}{ccc} AA & Aa & aa \\ a & d & -a \end{array}$$

The test for marker effects

$$t_1 = \frac{\mu_{MM} - \mu_{mm}}{\sqrt{\frac{\sigma_r^2}{n/4} + \frac{\sigma_r^2}{n/4}}} = \frac{(1 - 2r)2a}{\sqrt{8\sigma_r^2/n}} \quad (1)$$

and

$$t_2 = \frac{\mu_{Mm} - \frac{\mu_{MM} + \mu_{mm}}{2}}{\sqrt{\frac{\sigma_r^2}{n/2} + \frac{\sigma_r^2}{n} + \frac{\sigma_r^2}{n}}} = \frac{(1 - 2r)d}{\sqrt{4\sigma_r^2/n}} \quad (2)$$

Note that  $\mu_{Mm}$  does not contribute to the test in (1); adding  $\mu_{Mm}$  in (1) does not increase the efficiency of the test unless  $|d| \geq a/2$  (but see below for the calculation of sample size required with dominance).

When  $n$  is large, the observed difference  $\hat{t}$  is approximately normal distributed, and the power  $1 - \beta$  to detect the difference (for one-tailed test) is

$$1 - \beta = \text{Prob}[\hat{t} > z_\alpha \text{ with } \hat{t} \sim \mathcal{N}(t, 1)] \quad (3)$$

$$= 1 - \Phi(z_\alpha - t) \quad (4)$$

where  $z_\alpha$  is the  $z$  critical value of the test with  $(1 - \alpha)$  confidence under the null hypothesis  $t = 0$  and  $\Phi(x)$  is the standard normal cumulative distribution function.  $\alpha$  is the type I error and  $\beta$  is the type II error.

For given  $\alpha$  and  $\beta$  for the test the sample size  $n$  required is determined by

$$n_1 = 8 \left[ \frac{z_\alpha + z_\beta}{(1 - 2r)2a/\sigma_r} \right]^2 \quad \text{for additive effect} \quad (5)$$

$$n_2 = 4 \left[ \frac{z_\alpha + z_\beta}{(1 - 2r)d/\sigma_r} \right]^2 \quad \text{for dominance effect.} \quad (6)$$

## Several points on determining the required sample size

1. If the test is two-tailed (the usual case),  $z_\alpha$  should be replaced by  $z_{\alpha/2}$ .
2. For interval mapping the required sample size can be reduced by a factor of  $(1 - r^*)$  where  $r^*$  is the recombination frequency between an interval of two marker loci. Example: if  $r^*$  is about 0.23 for a 30 cM interval. Then,  $(1 - 2r)^2$  in (5) and (6) can be replaced by  $(1 - r^*) = 0.77$  to account for the worst case when a QTL is located in the middle of an interval ( $r \simeq r^*/2$ ).

3. In the test, if we also use many unlinked markers for controlling genetic background, most of genetic variance in the population can be removed from the residual variance (the idea of composite interval mapping), and  $\sigma_r^2$  may be roughly approximated by the environment variance  $\sigma_e^2$ . The overall heritability of the trait matters enormously.
4. For a systematical search for QTL in a genome, the type I error  $\alpha$  for each test should be substantially lower to account for increased false positive probability in an overall search. In most cases, the use of  $\alpha^* = 0.001$  (a very conservative level) for each individual test should be sufficient to ensure an overall false positive rate of less than 5%.

These suggest that the relevant number be calculated as

$$n_1 \simeq \frac{8}{0.77} \left[ \frac{z_{\alpha^*} + z_{\beta}}{2a/\sigma_e} \right]^2 \quad \text{for additive effect} \quad (7)$$

Now it remains to determine the likely magnitudes of  $2a/\sigma_e$ . Suppose that a QTL contributes to a proportion  $f$  of the genetic variance  $\sigma_g^2$  in a  $F_2$  population. Assuming that no other genes are linked to the QTL and ignoring the dominance  $d = 0$  (see below),

$$\frac{(2a)^2}{8\sigma_e^2} = f\sigma_g^2/\sigma_e^2.$$

$\sigma_g^2/\sigma_e^2$  is an unknown quantity.

Example: assuming  $h_{F_2}^2 = \sigma_g^2 / (\sigma_e^2 + \sigma_g^2) = 0.6$  means

$$\frac{\sigma_g^2}{\sigma_e^2} = 1.5 \quad \text{and} \quad \frac{(2a)^2}{\sigma_e^2} = 12f$$

Given that  $\alpha^* = 0.001$  and  $\beta = 0.1$  ( $z_{0.001} + z_{0.1} = 3.09 + 1.28 = 4.37$ ), the required sample sizes for detecting leading QTL for  $f = 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4$  and  $0.5$  are

$f$	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5
$n$	1653	826	330	165	82	55	41	33

## Effects of dominance

Depending on the degree of the dominance effect, the sample size required for detecting dominance effect may need to be substantially increased. Dominance does not, however, affect the calculation of the power detecting QTL. For example, suppose  $d = a$ . In this case we may use

$$t_3 = \frac{\mu_{M_-} - \mu_{mm}}{\sqrt{\frac{\sigma_r^2}{3n/4} + \frac{\sigma_r^2}{n/4}}} = \frac{(1 - 2r)2a}{\sqrt{16\sigma_r^2/3n}}.$$

But because of dominance

$$\frac{3(2a)^2}{16} = f\sigma_g^2.$$

Thus as long as  $f$ , the proportion of the genetic variation attributed to the QTL, is fixed, the required sample size for the test is unchanged.

## Effect of linkage: multiple linked QTL

Two issues

- Detection of QTL on the chromosome: For two linked QTL, if the model is misidentified (two QTL analyzed as one), the power to identify the "one QTL" is based on the joint effect of QTL (a weighted sum).
  - If the two QTL are in coupling linkage, the joint effect is aggregated. Power is increased.
  - If the two QTL are in repulsion linkage, the joint effect is reduced. Power is decreased, and can be very, very low. However, if we can identify the correct model (searching for two QTL or conditional searching), the issue is about separating linked QTL, and the power to identify repulsion-linked QTL is not necessarily very

low.

- Separating linked QTL (identifying both QTL)

The required sample size is increased by a factor (Zeng 1993)

$$\frac{\sigma_i^2}{\sigma_{i \cdot j}^2} = \frac{1/4}{r(1-r)}$$

$r$	0.5	0.4	0.3	0.2	0.15	0.1
$\frac{1}{4r(1-r)}$	1	1.04	1.19	1.56	1.96	2.78

$r$	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
$\frac{1}{4r(1-r)}$	3.05	3.40	3.84	4.43	5.26	6.51	8.59	12.76	25.25

## Comments

- QTL detection and power calculation depend on QTL mapping analysis procedure: Composite interval mapping is more powerful than simple interval mapping; Multiple interval mapping is more powerful than composite interval mapping.
- The power of the test can be increased by combining information from multiple related traits, multiple crosses, multiple environments, ...

The genetic structure becomes more complex, so is the statistical analysis. But, there are definite advantages in the joint multiple trait analysis for QTL identification (Jiang and Zeng 1995), and of course for hypothesis testing (pleiotropy) and parameter estimation.

## How large sample size do I need for my QTL mapping experiment?

- What is heritability for your trait (any knowledge or guess)?
- How large effect of a QTL (as a minimum) do you target to detect? Detect a QTL that explains 5% variation for example.
- Likely complexity of genetic architecture of QTL? How many QTL, distribution of effects, epistasis, ....