

# QTL-by-Environment Interaction

## 1. The problem

Differential expression of a phenotypic trait by genotypes across environments, or Genotype x Environment (GxE) interaction is an old problem of primary importance for quantitative genetics and its applications in breeding, conservation biology, theory of evolution, and human genetics (Eberhard and Russel 1966; Falconer 1981; Via and Lande 1987; Turet et al. 1993). Recent successes in QTL mapping have shifted the focus of GxE analysis from the genotype to gene level (e.g. Paterson et al. 1991; Hayes *et al.* 1993; Sari-Gorla et al. 1997). For breeding purposes, the primary concern is possible environmental instability in manifestation of mapped QTLs that might become candidates for marker-assisted selection. To evaluate stability of QTL effects in crop species dozens of immortal mapping populations have been developed for trait scoring under various environmental conditions (Hayes 1994). Simultaneously, such an approach may provide a significant increase in the resolution power of the QTL analysis (Soller and Beckmann 1990).

However, a serious gap still exists between the demands invoked by real QTL mapping experiments and the power of the available tools. This is also true for QTLxE analysis. An attempt to build into a standard QTL mapping model an additional flexibility allowing for varying across environments QTL effects, is accompanied by a tremendous number of parameters involved in the model, that increase as a product of the identified QTLs and the number of environments where the traits were measured. We will review here some suggestions proposed recently to cope with this problem and a new approximated method of QTLxE analysis (Korol *et al.* 1997; Ronin *et al.* 1997). The last method uses the ideas of classical GxE interaction analysis (Eberhard and Russel 1966) and allows for QTLxE interaction across a large (virtually unlimited) number of environments, without necessarily increasing the number of parameters of the mapping model.

## 2. The state of art

In the approach of multiple-environment QTL analysis, the data from each environment are treated separately, and the final conclusion is derived from the analysis of the estimated QTL effects across environments (Paterson *et al.* 1991; Stuber *et al.* 1992). The usual way of testing for QTLxE interactions is based on application of ANOVA to the resulting individual QTL estimates (e.g. Utz and Melchinger 1996). More sophisticated methods are needed to combine all the data across environments allowing for QTLxE interaction effects. However, a direct application of this idea is quite difficult because the number of parameters of the mapping model increases as a product of the identified QTLs and the number of environments where the traits were measured. Therefore, some

indirect or approximated methods oriented on less general situations have been proposed (Hayes *et al.* 1993; Tinker and Mather 1995; Romagosa *et al.* 1996; Beavis and Keim 1996).

Jiang and Zeng (1995) suggested to employ multiple trait analysis for QTLxE analysis. Namely, in accordance with the classical idea of quantitative genetics (Falconer 1981) they considered trait measurements of the same genotype in different environments as a set of correlated traits. This idea was also discussed by Korol *et al.* (1994) and Ronin *et al.* (1995). However, the multiple trait analysis limits the number of environments because it is associated with an increased number of parameters.

An efficient solution to the foregoing problem was developed by Jensen and co-authors (1995). Their QTL mapping model includes in an obvious way the terms describing the effects of the target QTL and regression cofactors of co-segregation QTLs, the effects of multiple environments, and the terms of QTLxE interactions. Such a description significantly reduces the dimensionality of the problem. However, this formalization is limited by situations where the environments can be obviously characterized by some parameters, or 'environmental factors', like day length, or irrigation-fertilization treatments, altitude, etc. Then, each factor can be represented as an independent variable in the linear model. In most practical cases we are not able to define a small number of such factors. Another approach to analyze QTLxE interaction without direct specification of the 'physical' characteristics of the environments was recently proposed by Romagosa and co-authors (1996). Their algorithm is based on clustering the environments using a few (say two) detected QTLs with most variable effects across environments.

The main distinction of the approximated model proposed by Korol *et al.* (1997) and Ronin *et al.* (1997) is in the chosen form of representation of the dependence of putative QTL effects on environmental states. In reality, each environment is a complex of abiotic (temperatures, humidity, ion concentrations, etc.), biotic (pests and pathogens, competitors, etc.) and agrotechnical features. These could strongly affect the manifestation of quantitative traits and the effects of QTLs, but are difficult to characterize quantitatively. It was suggested that the measured trait values of the mapping population (e.g. trait means) may serve as objective integral characteristic of the environmental state. Accordingly, a larger number of traits should provide a better "bioindication". In the simplest form, one can approximate the environmental dependence of the effect of allele substitution at a QTL by a polynomial over the mean values of the same trait across the environments. For some putative QTLs, the dependence on mean value of the respective traits explains a large part of the environmental variation of the QTL effect. However, the suggested approach does not exclude the possibility to take into account any additional information, like temperature, day length, water regime, etc., that might characterize the environments (e.g. Jensen *et al.* 1995). These ('physical') characteristics can be introduced into the model parallel to the bioindicatory terms (e.g. polynomial over the mean values)

together with terms characterizing the dependence of the putative QTL effect on interaction between the 'physical' and 'bioindicator' factors. The proposed approach to analyze QTLx $E$  interaction is in a sense similar to that of Romagosa and co-authors (1996) because it also does not specify directly the 'physical' characteristics of the environments. Actually, this is a different version of the same general idea of "bioindicators" as a tool for characterizing "anonymous" environments. The results one could obtain by means of the proposed method of QTLx $E$  analysis will be approximate, allowing, at best, to consider the major part of QTLx $E$  interaction. However, the possibility to work with a virtually unlimited number of environments without increasing the number of parameters needed may significantly offset this drawback resulting in increased power of detection of QTLx $E$  interaction and in higher accuracy of QTL chromosomal location.

### 3. The model

#### 3.1. Mixture model of interval QTL analysis

Consider a simplified situation when the trait of interest ( $x$ ) depends on a single QTL,  $Q/q$ . For the sake of simplicity we confine the analysis to dihaploid mapping populations (which applies also to backcrosses and recombinant inbreds). For an arbitrary genotype of the mapping population, the trait measurement in the  $i$ th environment can be presented as

$$x_i = m_i + 0.5ga_i + e_i, \quad (1)$$

where  $m_i$  is the mean trait value in the  $i$ th environment,  $g$  is either +1 (for  $QQ$  genotypes) or -1 (for  $qq$  genotypes),  $a_i$  is the effect of allele substitution at putative QTL on trait in environment  $i$ , and  $e_i$  is a random variable with zero mean and variance  $\hat{\sigma}_i$ . If we find  $a_i \approx a$  for any  $i$ , then no Gx $E$  interaction is manifested by  $Q/q$ .

Assume that  $Q/q$  resides in some interval  $(k, k+1)$  of a chromosome, with recombination rates  $r_1$  and  $r_2$  in subintervals  $M_k/m_k - Q/q$  and  $Q/q - M_{k+1}/m_{k+1}$ , respectively. The expected densities of the trin marker groups  $U_{mkmk+1}(x) = U_1(x)$ ,  $U_{Mkmk+1}(x) = U_2(x)$ ,  $U_{mkMk+1}(x) = U_3(x)$  and  $U_{MkMk+1}(x) = U_4(x)$  can be written as:

$$U_i(x) = pif_{qq}(x) + (1-p_i)f_{QQ}(x), \quad i=1, \dots, 4 \quad (2)$$

where  $f_{qq}(x)$  and  $f_{QQ}(x)$  are the trait density distributions in the groups  $qq$  and  $QQ$ , respectively.

In a single-environment formulation one could test whether or not the observed variation of  $x$  is associated with segregation in interval  $M_k/m_k - M_{k+1}/m_{k+1}$  and identify the corresponding locus  $Q/q$ . Provided recombination rate between markers is known, the vector of  $n_1$  parameters specifying the putative QTL can be presented as  $\theta_{n_1} = \{r_1, m, a, s^2\}$ . The assumption of no association between  $x$  and segregation in  $M_k/m_k - M_{k+1}/m_{k+1}$  interval can formally be presented by another set of parameters,

$\theta = \theta_{n0} = \{m, s^2\}$ . The null hypothesis  $\{\mathbf{H}_0: \theta = \theta_{n0}\}$  as contrasted with the alternative  $\{\mathbf{H}_1: \theta = \theta_{n1}\}$  can be investigated with the likelihood ratio test (Wilks 1962):

$$\chi^2 (\mathbf{H}_1 \text{ vs. } \mathbf{H}_0) = 2 \ln \left[ \frac{\max_{\theta_{n1} \in S_1} L(\theta_{n1})}{\max_{\theta_{n0} \in S_0} L(\theta_{n0})} \right] \quad (3)$$

where  $S_0$  and  $S_1$  are the parameter spaces corresponding to  $\mathbf{H}_0$  and  $\mathbf{H}_1$ , respectively.

In multiple-environments we could employ the foregoing to trait measurements obtained under  $p$  ( $p > 1$ ) environments. When comparing the foregoing alternatives  $\mathbf{H}_0$  and  $\mathbf{H}_1$ , QTLx $E$  interaction effects could be included in the model and tested against the alternative of 'no QTLx $E$  interaction'. Therefore, an additional group of hypotheses  $\{\mathbf{H}_2: \theta = \theta_{n2}\}$  could be considered that assume a dependence of the target QTL effect and, possibly, of the residual variance, on environment. Vector  $\theta_{n2}$  of the full model, corresponding to  $\mathbf{H}_2$  with environment-specific parameters  $a_i$ ,  $s^2_i$  and  $m_i$ , will then contain  $3p+1$  components. Consequently,  $\theta_{n1}$  specifying  $\mathbf{H}_1$  (constant effect  $a_i = a$  of the QTL across environments, though allowing for variable  $s^2_i$  and  $m_i$ ) contains  $2p+2$ , while  $\theta_{n0}$  (no QTL on the tested chromosome)  $2p$  parameters.

In the simplified case of only one QTL segregating in the mapping population, no correlation between trait measurements across environments are expected. With this assumption, instead of the test statistics (3) one can build its multi-environmental equivalent  $\chi^2 (\mathbf{H}_1 \text{ vs. } \mathbf{H}_0)$  with  $df = 2p+2 - 2p = 2$ . If  $\mathbf{H}_0$  is rejected (i.e.  $a_i \neq 0$ ), then the obvious benefit of the multi-environmental model is the increase in the number of measurements, hence a higher precision of the parameter estimates (e.g. Jensen *et al.* 1995). No less important is the possibility to conduct the following two tests:

$$\chi^2 (\mathbf{H}_2 \text{ vs. } \mathbf{H}_0) = 2 \ln \left[ \frac{\max_{\theta_{n2} \in S_2} L(\theta_{n2})}{\max_{\theta_{n0} \in S_0} L(\theta_{n0})} \right] \quad (3a)$$

with  $df = 3p+1 - 2p = p+1$ , and

$$\chi^2 (\mathbf{H}_2 \text{ vs. } \mathbf{H}_1) = 2 \ln \left[ \frac{\max_{\theta_{n2} \in S_2} L(\theta_{n2})}{\max_{\theta_{n1} \in S_1} L(\theta_{n1})} \right] \quad (3b)$$

with  $df = 3p+1 - (2p+2) = p-1$ .

### 3.2. Regression specification of QTLx $E$ interaction

According to the proposed approach, the unknown effects  $a_i$  and (if desirable) the residual variances  $s^2_i$  are represented by low degree polynomials. For instance, with a cubic approximation for  $a_i$  and a quadratic for  $s^2_i$ , we will need only 7 parameters instead of 20 ! Clearly, the main question remains: to what extent the "bioindicating" trait, or a (linear) combination of traits, will indeed be informative with

respect to the dependence of the target QTL effect on environmental states. In the log-likelihood functions  $\ln L(\theta_{nk})$  ( $k=0,1,2$ ) from 3 (3a-3b), we should replace the corresponding coordinates of the parameter vectors  $\theta_{ni}$  by polynomials:

$$\begin{aligned} a_i &= \alpha_0 + \alpha_1 m_i + \alpha_2 m_i^2 + \dots + \alpha_s m_i^s, \\ s_i^2 &= \beta_0 + \beta_1 m_i + \beta_2 m_i^2 + \dots + \beta_t m_i^t, \quad i = 1, \dots, p \end{aligned} \quad (4)$$

The procedure will provide ML-estimates of  $r_1$  and regression coefficients  $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_s$ ,  $\beta_0, \beta_1, \beta_2, \dots, \beta_t$ ; statistically significant deviation of  $\alpha_i$  ( $i > 0$ ) from zero indicates the existence of QTLx $E$  interaction. Clearly, the analysis should include model adjustment with a series of polynomials  $a(m) = P_{as}(m)$  and  $s^2(m) = P_{st}(m)$ .

### 3.3. Results: single QTL analysis

In this situation no 'between-environment' correlation is expected for the residual (within QTL groups) variation. Thus, the log-likelihood functions 3a and 3b for the mixture model 3a -3b could be calculated by summing up over all environments and employing the polynomials of Eqs.4. This assumes implicitly that after removing the effects of the QTL under consideration, the residuals are independent across environments. Clearly, such an idealization is correct if the whole remainder genetic variation of the quantitative trait is taken into account by markers of other genomic regions as co-factors (Jensen and Stam 1994; Zeng 1994). But this may not be the case, and corresponding complications will be considered later.

The simulated experiment (Korol *et al.* 1997) included ten environments with mean value of the trait ( $m$ ) linearly increasing from  $m_1=0$  up to  $m_{10}=3.6$ . The target QTL was positioned in the middle of the third interval of six of the marked chromosome, each interval was 20 cM. The assumed size of the mapping population was  $N=200$ . The intention was to compare the general model (**MG**) specifying all effects  $a_i$  and residual variances  $s_i^2$  with the approximated model (**MA**), where the QTL effect and residual variance are represented as functions of an environmental "bioindicator" (population mean values of the same trait were employed). The power of detection of QTL effect and QTLx $E$  interaction, and the precision of chromosomal location of the detected QTL were used for comparison of the models. Table 1 demonstrates that adequate approximation of  $a(m)$  results in an appreciable increase in power of both tests: **H**<sub>1</sub> vs. **H**<sub>0</sub> (presence of a QTL, allowing for  $a_i = \text{const}$  and  $s_i^2 \neq \text{const}$ ), and **H**<sub>2</sub> vs. **H**<sub>1</sub> (presence of QTLx $E$  interaction, allowing for  $s_i^2 \neq \text{const}$ ). Polynomial approximation has resulted in an improved accuracy in the estimated QTL position. Note, that the increased resolution of the **MA** model as compared to the **MG** model was obtained using only 8

parameters (less than half of that in **MG**). Likewise, the **MA** model provided about 1.5-to-2-fold reduction in the confidence interval for the  $a_i$  estimates across environments (not shown).

**Table 1.** Estimated location and power of detection of QTL effect ( $\beta_f$ ) and QTL×E interaction ( $\beta_e$ ) employing the general (**MG**) and approximated (**MA**) models in single-QTL situations.

$h^2\%$		$L$ (cM)	$\beta_{et}$ (%)			$\beta_{ft}$ (%)			$n_p$	$df_e$	$df_f$	
			$\alpha\% \rightarrow 5$	1	0.1	5	1	0.1				
$S_1$	<b>MA</b>	66.0±2.03	67	41	16	67	45	22	8	3	5	0.52
	<b>MG</b>	65.4±2.56	57	29	6	64	34	8	21	9	11	
$S_2$	<b>MA</b>	61.7±0.89	97	87	70	94	87	67	8	3	5	1.26
	<b>MG</b>	62.5±1.45	88	71	47	91	78	49	21	9	11	
$S_3$	<b>MA</b>	59.8±0.49	100	99	94	100	99	97	8	3	5	2.27
	<b>MG</b>	61.1±0.68	100	95	81	100	98	90	21	9	11	

The results of 200 Monte-Carlo runs are presented for single-QTL situations.  $L$  is the estimated QTL location (the simulated value of  $L$  is 60 cM);  $\alpha$  is significance level,  $n_p$  is the number of parameters specifying the model;  $df_f$  and  $df_e$  - degree of freedom for the tests of QTL presence and QTL×E interaction, respectively

## 4. Complications

### 4.1 How to choose a good approximation

With simulated data, it is easy to compare "the adequate" and "non-adequate" approximations simply because we know the employed model. However, the situation will be quite different when real data will be analysed, i.e. no *a priori* information exists on the form of  $a(m)$ . Thus, the decision about the "adequacy" should be justified statistically. Table 2 illustrates the possibility to deduce the adequate approximation of the QTL×E interaction from the data. The columns  $\beta_{et}=\beta(\alpha)$  show the power of detection of QTL×E interaction for each of the presented models. The critical values ( $\alpha$ ) of the test

statistics [see Eq. 3b] were determined by using: (a) the asymptotic  $\chi^2$  distribution, and (b) Monte-Carlo simulations with 5000 runs for each of the models (data in brackets). These results demonstrate a remarkable proximity of these two estimates of the power for all of the models. The highest power of detection QTLx E interaction and the highest accuracy of the QTL location were obtained with the adequate model (**MA<sub>3</sub>**). It is not surprising that **MA<sub>3</sub>** is superior over **MG**. But less expected is the fact that the non-adequate approximations **MA<sub>2</sub>** and **MA<sub>4</sub>** were also superior over **MG** (but with much less needed parameters). Thus, it is not mandatory to have the 'adequate' approximation in order to take advantage of the proposed method: it will be sufficient to provide 'a good one'. Nevertheless, how can we decide about the adequate model *a posteriori*, provided the class of the approximation functions is chosen correctly?

**Table 2.** Comparison of the general model with the polynomial approximations for the detection power of QTLx E interaction ( $\beta_e$ ), and accuracy of QTL location ( $L$ ). Situation  $S_2$  (see Table 1) is considered. The power of the test was obtained using Monte-Carlo simulations (see text); corresponding results based on  $\chi^2$  asymptotic distribution of the test statistic are given in brackets). The distribution of frequencies of the chosen approximations is presented in the last three columns ( $f_M$ ) (resulted from competition between **MA<sub>1</sub>**-**MA<sub>4</sub>** and **MG**, see text).

Model	$L$ (cM)		$\beta_e$ (%)		$df$	$n_p$	$f_M$		
	$\alpha\% \rightarrow 5$		1	0.1			$\alpha\% \rightarrow 5$	1	0.1
<b>MA<sub>1</sub></b>	62.0±1.46	88(88)	68(69)	48(46)	1	6	0.18	0.16	0.13
<b>MA<sub>2</sub></b>	61.1±1.31	92(92)	78(77)	56(59)	2	7	0.18	0.17	0.15
<b>MA<sub>3</sub></b>	61.7±0.89	96(97)	86(87)	66(70)	3	8	0.54	0.49	0.45
<b>MA<sub>4</sub></b>	61.6±0.92	95(95)	80(82)	61(61)	4	9	0.05	0.05	0.05
<b>MG</b>	62.5±1.45	88	71	47	9	21	0.03	0.03	0.03
Total frequency of cases where QTLx E interaction was detected							0.98	0.90	0.78

To address the last question, the following procedure was employed (Korol *et al.* 1997). For each run, the data were analysed using all of the models (**MA<sub>1</sub>**-**MA<sub>4</sub>** and **MG**) and the models that have detected QTLx E interaction at the level of significance  $\alpha$  were chosen. Then, the model that (i) exceeded significantly (at some level  $\alpha$ ) all of the more simple models, and (ii) did not differ significantly (at  $\alpha$ ) from more complex models, was selected as an 'adequate'. The general model also participated in this competition, as the most complex one (because of the number of parameters needed). The resulting distribution of the choices of the 'adequate' model is presented in the last three columns of Table 2. The conclusions are that: (1) the adequate model **MA<sub>3</sub>** is the absolute winner (chosen in more than half of the runs where the QTLx E interaction was detected, and with a frequency that is three-fold higher than the next best choice); (2) the models of the polynomial class

were chosen by a factor of 25-30 than the exact general model **MG**. Moreover, even the simplest approximation, **MA**<sub>1</sub>, appears to be chosen 4-6 fold more frequently than **MG**.

#### **4.2. Robustness to the effect of correlations caused by unaccounted for QTLs**

When several QTLs segregate simultaneously in the mapping population, their effects will generate correlations between trait measurements across environments which should be taken into account. One of the possible ways to account for this correlation is through simultaneous analysis of multiple traits, taking the trait values in different environments as different quantitative traits (Korol *et al.* 1994, 1995; Jiang and Zeng 1995; Ronin *et al.* 1995). However, the multiple trait analysis is associated with an increased number of parameters. The approach proposed by Korol *et al.* (1997) and Ronin *et al.* (1997) is free of this problem, but introduces other sources of distortions: (1) correlations caused by unaccounted QTLs, and (2) approximated description of QTL dependence on environment.

Consider the first problem. We should evaluate to what extent correlations between environments caused by unaccounted QTLs may affect the efficiency of the proposed approach. Clearly, the approximated model of the environmental dependence of QTL effect as a function of the mean value of the trait in a given environment can be applied not only to the "target" QTL, but also to the cosegregating QTLs. This approach may be especially attractive when there are many cosegregating QTLs with environmental dependent effects (e.g., as regression cofactors on respective marker loci). This will result in far fewer parameters. However, it is important to know also the robustness of the method to violation of model assumptions when some QTLs remain undetected. The simulation results (Korol *et al.* 1997) show that a strong QTL, if not accounted by the model, may cause correlations between environments resulting in reduced accuracy of estimated parameters. Nevertheless, the distortion caused by a QTL comparable with the target one (e.g. exceeding the target effect not more than 2 times) is not dramatic for the application.

Including the effects of cosegregating QTLs into the model solves the last problem. This can be done by combining the proposed approach with regression cofactors (Jansens and Stam 1994). However, an appreciable proportion of genetic variation for the analyzed trait may still remain in the residuals because of combined effect of many small QTLs. This residual genetic variation may be several-fold larger than the effect of the target QTL. Would the resulting correlation between environments preclude the application of the method?. To address this question Korol *et al.* (1997) simulated a situation where the genetic variation of the trait depended on the target QTL ( $Q_1/q_1$ ) (with an average  $h^2 \sim 2.5\%$ ) and 10 additional unlinked QTLs ( $Q_2/q_2-Q_{11}/q_{11}$ ). The average (across

environments) effect of  $Q_2/q_2$  was  $h^2 \sim 10\%$ , whereas the combined average effect of  $Q_3/q_3-Q_{11}/q_{11}$  was 15%. Thus, the total effect of  $Q_2/q_2-Q_{11}/q_{11}$  was 10-fold as compared to that of  $Q_1/q_1$ , whereas the effect of  $Q_3/q_3-Q_{11}/q_{11}$  was 6-fold as compared to  $Q_1/q_1$ . One may expect that the power of detection of  $Q_1/q_1 \times E$  interaction will be very low if the segregation of  $Q_2/q_2-Q_{11}/q_{11}$  is not accounted for by the model hence causing correlation between the environments. That was indeed the case (see Table 3, first row). Note, that in this case employment of the asymptotic distribution for the critical values gave seriously biased upwards estimates  $\beta$  of the power of detection of  $Q_1/q_1 \times E$  interaction, compared to the estimates obtained using Monte-Carlo simulations with 5000 runs.

**Table 3.** The effect of co-factors on the power of detection of QTL $\times$ E interaction ( $\beta_e$ ), and accuracy of QTL location ( $L$ ). The power of the test was obtained using Monte-Carlo simulations (see text); corresponding results based on  $\chi^2$  distribution of the test statistic are given in brackets). Three models of the analysis of the residual variation were employed: (1) the co-factors are totally ignored; (2) the effect of the strongest co-factor is eliminated; (3) the genetic component of the residual variation is replaced by equivalent non-genetic variation.  $D_{gei}$  is the variance of QTL $\times$ E interaction for the target QTL,  $h^2\%$  is the averaged heritability over environments attributed to the target QTL

Model	$h^2\%$	$D_{gei}$	$L(\text{cM})$	$\beta_e(\%)$		
				$\alpha\% \rightarrow 5$	1	0.1
1	2.4	0.101	62.9 $\pm$ 1.66	45(70)	27(58)	15(42)
2			59.9 $\pm$ 0.74	96(96)	91(90)	76(76)
3			59.2 $\pm$ 0.49	99(99)	99(99)	96(97)

Including  $Q_2/q_2$  as a co-factor into the model substantially improved the situation, increasing the detection power of  $Q_1/q_1 \times E$  interaction from two- to five-fold (fa ranging from 0.05 to 0.001) and the accuracy of  $Q_1/q_1$  location more than two-fold (second row of Table 3). Note, that in this case  $\chi^2$  distribution appeared to be a good approximation for the test statistic, in spite of the noise caused by  $Q_3/q_3-Q_{11}/q_{11}$ . For comparison, the third row of the Table shows the results for the case where the whole residual genetic variation caused by  $Q_3/q_3-Q_{11}/q_{11}$  is replaced by non-genetic variation. We can conclude that distortion of the basic model assumption of 'no correlation between environments' caused by the presence of  $Q_3/q_3-Q_{11}/q_{11}$ , which collectively exceed by a factor six the effect of the target QTL  $Q_1/q_1$ , is incomparably smaller than that caused by a single QTL,  $Q_2/q_2$ , which exceeds the target QTL only by a factor of four.

### 4.3. Missing data

One can hardly expect that all genotypes will be perfectly represented in all of the environments where the experiment was conducted. Some data will be missed, hence it is of primary interest to get some idea how it could affect the power of QTLx $E$  detection. The approximate model allows to treat this problem easily. It appeared that with a large number of environments, even if a large proportion of genotypes is not represented in each environment, the resulting power of the test of QTLx $E$  interaction and location accuracy of the target QTL are quite high. Monte-Carlo simulations presented in Table 4 illustrate this point. It is noteworthy, that if only 20-50% of the data are available in each of the 50-100 environments, still the approximated model give very satisfactory results even when not the best (optimal) approximation was used (compare the results for  $MA_1$ ,  $MA_2$  and  $MA_3$  for the two examples with the situation  $S_4$  with  $h^2 \approx 2.3\%$  for the target QTL  $Q_1/q_1$  and  $h^2 \approx 6\%$  for a co-segregating QTL  $Q_2/q_2$ ). Clearly, an attempt to apply the general model would mean an unrealistic task of estimation of about 100-200 parameters, in contrast to our model which needs only 8 parameters.

**Table 4.** The effect of missing data on the power of detection of QTLx $E$  interaction and accuracy of QTL location, when the number of environments is large. Here  $N_{lin}$  is the total number of genotypes (lines) in the mapping population;  $N_{env}$  is the number of environments, and  $n$  is the mean number of genotypes scored per environments.

$S_i$	Model	$N_{lin}$	$n$	$N_{env}$	$L(cM)$	$\beta_{et}(\%)$		
						$\alpha\% \rightarrow 5$	1	0.1
$S_1$	$MA_3$	200	10	200	66.0 $\pm$ 2.03	67	41	16
		200	100	50	59.7 $\pm$ 0.94	94	82	60
		200	100	100	60.0 $\pm$ 0.45	100	99	98
$S_4$	$MA_1$	200	50	40	61.4 $\pm$ 1.10	81	69	46
	$MA_2$	200	50	40	61.0 $\pm$ 1.10	89	77	56
	$MA_3$	200	50	40	60.0 $\pm$ 0.72	98	93	82

### 4.4. How to recognize the situations when the approximated model is not valid ?

The conducted simulations demonstrate the utility of the proposed approximate approach for analyzing QTLx $E$  interaction. Its main benefit is the ability to use data collected from a large number of environments without the necessity of increasing the number of parameters. Expressing the

dependence of a QTL effect on environmental conditions as a function of environmental mean value of the trait can also be applied to multiple QTLs from independent genomic regions. Therefore, the proposed approach could be very helpful in coping, albeit in an approximate form, with a difficult problem of QTL mapping analysis, i.e. rapid increase in the number of parameters with increasing number of effective QTLs and environments. This improves our ability to extract more mapping information when more environments are used to evaluate the quantitative trait. The most difficult problem with this approach, is the recognizing the situations when the applied approximation is not valid. If the opposite is true, that is if the dependence of the QTL effect on environmental conditions can indeed be presented in the form of regression on mean values or any other "bioindicators", then the proposed approximated method proved to give a higher statistical performance compared to the precise general model (**MG**). Thus, one can start the procedure using the approximated method. However, if the approximated analysis revealed no significant QTLx $E$  interaction, does it really mean an independence of the QTL effect from environmental conditions? Or alternatively, the interaction may exist, but it cannot be represented as a regression of the target QTL effect on the mean values of the trait or some other "bioindicators"?

Consider one of the possible ways to cope with this problem. If the general model is applicable (i.e. the number of parameters is not too large), it may be used as a tool to answer the foregoing question. Rejection of the  $H_0$  hypothesis 'no QTLx $E$  interaction' by **MG** will mean that our basic assumption (regression on the "bioindicator") does not fit the data. If the number of environments is too large, the general model can be applied for randomly chosen groups of environments. Then, the significance of the interaction may be evaluated from the obtained distribution of the tests using the Bonferroni correction. For example with  $N=100$  environments, one can produce  $k=20$  samples, each including data of  $m=10$  randomly chosen environments. Let  $\alpha$  be the accepted level of significance for the QTLx $E$  interaction test for the whole set of the samples. Then, assuming independence of these samples, one can reject the  $H_0$  hypothesis if at least for one of the samples the test statistics achieved the significance level of  $\alpha/k$ . Clearly, due to the postulated independence (which is not the case for  $mk > N$ ), this is a conservative test of QTLx $E$  interaction. Nevertheless, it seems to be more preferable than the standard way of multiple-environment data analysis, when the data from each environment are treated separately, and the final conclusion is derived from the analysis of the estimated QTL effects across environments (Paterson et al. 1991; Stuber et al. 1992; Utz and Melchinger 1996).

The foregoing test based on the general model may result in the same conclusion as the approximated model, i.e. 'no QTLx $E$  interaction'. By contrast, if the general model allowed to detect QTLx $E$  interaction but the approximated model did not, it will indicate that the proposed "bioindicator(s)" is(are) not informative and other explanatory factors should be found. Further studies are needed to develop a more optimal algorithms of application of the proposed approach

when applied to a large number of environments (when direct utilization of the general model is impossible). However, even in the current form, the drawbacks of the proposed method are by far compensated by the possibility to work with actually unlimited number of environments, under massive missing data, and at a remarkable reduction in the number of parameters needed.

## References

- Beavis, W.D. and P. Keim, 1997. Identification of QTL that are affected by environment, In *Genotype by Environment Interactions*, edited by M. S. Kang and H. Gauch, Jr. CRC Press.
- Eberhard, S.A. and W.A. Russel, 1966 Stability parameters for comparing varieties. *Crop Sci.* **6**: 36-40.
- Falconer, D.S., 1989. *Introduction to Quantitative Genetics*. Longman Scientific & Technical, Essex, England.
- Hayes, P.M., 1994. Genetic stocks available through the North American Barley Genome Mapping Project. *Barley Genetics Newsletter* **24**: 113-116.
- Hayes, P.M., B.H. Liu, S.J. Knapp, F. Chen, B. Jones, T. Blake, J. Franckowiak, D. Rasmusson, M. Sorrels, S.E. Ullrich, D. Wesenberg, and A. Kleinhofs, 1993. Quantitative trait locus effects and environmental interaction in a sample of North American barley germplasm. *Theor. Appl. Genet.* **87**: 392-401.
- Jansen, R.C., and P. Stam, 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**: 1447-1455.
- Jansen, R.C., J.M. Van Ooijen, P. Stam, C. Lister, and C. Dean, 1995. Genotype-by-environment interaction in genetic mapping of multiple quantitative trait loci. *Theor. Appl. Genet.* **91**: 33-37.
- Jiang, C. and Z.-B. Zeng, 1995. Multiple trait analysis and genetic mapping for quantitative trait lo. *Genetics* **140**: 1111-1127.
- Korol, A.B., I.A. Preygel, and S.I. Preygel, 1994. *Recombination Variability and Evolution*. Chapman & Hall, London.
- Korol, A.B., Y.I. Ronin, and V.M. Kirzhner, 1995. Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics* **140**: 1137-1147.
- Paterson, A.H., S. Damon, J.D. Hewitt, D. Zamir, H.D. Rabinowitch, S.E. Lincoln, E.S. Lander, and S.D. Tanksley, 1991. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations and environments. *Genetics* **127**: 181-197.
- Romagosa, I., S.E.Ullrich, F. Han, and P.M.Hayes, 1996. Use of additive main effects and multiplikative interaction model in QTL mapping for adaptation in barley. *Theor. Appl. Genet.* **93**: 30-37.
- Ronin, Y.I., V.M. Kirzhner, and A.B. Korol, 1995. Linkage between loci of quantitative traits

- and marker loci. Multi-trait analysis with a single marker. *Theor. Appl. Genet.* **90**: 776-786.
- Sari-Gorla, M., T. Calinski, Z. Kaczmarek, and P. Krajewski, 1997. Detection of QTL-environment interaction in maize by a least squares interval mapping method. *Heredity* **78**:146-157.
- Stuber, C.W., Linkoln, S.E., Wolff, S.E., Helentjaris, T. and Lander, E.S. 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Genetics* **132**: 823-839.
- Tinker, N.A. and D.E.Mather 1995. Methods for QTL analysis with progeny replicated in multiple environments. *JQTL* **1** (2). <http://probe.nalusda.gov:8000/otherdocs/jqtl/jqtl1995-02/jqtl16r2.html>
- Tiret, L., L. Abel, and R. Rakotovao, 1993. Effect of ignoring genotype-environment interaction on segregation analysis of quantitative traits. *Genetic Epidemiology* **10**: 581-586.
- Utz, H.F. and A.E. Melchinger, 1996. PLABQTL: A program for composite interval mapping of QTL. *JQTL* **2** (1). <http://probe.nalusda.gov:8000/otherdocs/jqtl/jqtl1996-01/utz.html>
- Via, S. and R. Lande, 1987. Evolution of genetic variability in a spatially heterogeneous environment: effects of genotype-environment interaction. *Genet. Res.* **49**: 147-156.
- Wilks, S.S., 1962. *Mathematical Statistics*. Wiley, New York.
- Zeng, Z.-B., 1994. Precision mapping of quantitative trait loci. *Genetics* **136**: 1457-1468.