

Multiple testing issues in QTL mapping

Małgorzata Bogdan

Institute of Mathematics and Computer Science, Wrocław University of
Technology, Poland

in cooperation with

J.K.Ghosh, R.W.Doerge, R. Cheng – Purdue University

A. Baierl, F. Frommlet, A. Futschik – Vienna University

A. Chakrabarti - Indian Statistical Institute

P. Biecek, A. Ochman, M. Żak – Wrocław University of Technology

Wrocław, 20.11.2008

Backcross population and recombinant inbred lines

Only two genotypes possible at a given locus

Backcross population and recombinant inbred lines

Only two genotypes possible at a given locus

X_{ij} - dummy variable encoding the genotype of i -th individual at locus j

Backcross population and recombinant inbred lines

Only two genotypes possible at a given locus

X_{ij} - dummy variable encoding the genotype of i -th individual at locus j

$$X_{ij} \in \{-1/2, 1/2\}$$

Backcross population and recombinant inbred lines

Only two genotypes possible at a given locus

X_{ij} - dummy variable encoding the genotype of i -th individual at locus j

$$X_{ij} \in \{-1/2, 1/2\}$$

Multiple regression model:

$$Y_i = \mu + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv} + \varepsilon_i,$$

I - a certain subset of the set $N = \{1, \dots, m\}$, U - is a subset of $N \times N$

Backcross population and recombinant inbred lines

Only two genotypes possible at a given locus

X_{ij} - dummy variable encoding the genotype of i -th individual at locus j

$$X_{ij} \in \{-1/2, 1/2\}$$

Multiple regression model:

$$Y_i = \mu + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv} + \varepsilon_i,$$

I - a certain subset of the set $N = \{1, \dots, m\}$, U - is a subset of $N \times N$

Problem : estimation of the number of influential genes

Bayesian Information Criterion (1)

M_j - j -th linear model with $k_j + l_j < n$ regressors

Bayesian Information Criterion (1)

M_j - j -th linear model with $k_j + l_j < n$ regressors

$\theta_j = (\beta_0, \beta_1, \dots, \beta_{k_j}, \gamma_1, \dots, \gamma_{l_j}, \sigma)$ - vector of model parameters

Bayesian Information Criterion (1)

M_j - j -th linear model with $k_j + l_j < n$ regressors

$\theta_j = (\beta_0, \beta_1, \dots, \beta_{k_j}, \gamma_1, \dots, \gamma_{l_j}, \sigma)$ - vector of model parameters

Bayesian Information Criterion (Schwarz, 1978) –

maximize $BIC = \log L(Y|M_j, \hat{\theta}_j) - \frac{1}{2}(k_j + l_j) \log n$

Bayesian Information Criterion (1)

M_j - j -th linear model with $k_j + l_j < n$ regressors

$\theta_j = (\beta_0, \beta_1, \dots, \beta_{k_j}, \gamma_1, \dots, \gamma_{l_j}, \sigma)$ - vector of model parameters

Bayesian Information Criterion (Schwarz, 1978) –

maximize $BIC = \log L(Y|M_j, \hat{\theta}_j) - \frac{1}{2}(k_j + l_j) \log n$

If m is fixed, $n \rightarrow \infty$ and $X'X/n \rightarrow Q$, where Q is a positive definite matrix, then BIC is consistent - the probability of choosing the proper model converges to 1.

Bayesian Information Criterion (1)

M_j - j -th linear model with $k_j + l_j < n$ regressors

$\theta_j = (\beta_0, \beta_1, \dots, \beta_{k_j}, \gamma_1, \dots, \gamma_{l_j}, \sigma)$ - vector of model parameters

Bayesian Information Criterion (Schwarz, 1978) –

maximize $BIC = \log L(Y|M_j, \hat{\theta}_j) - \frac{1}{2}(k_j + l_j) \log n$

If m is fixed, $n \rightarrow \infty$ and $X'X/n \rightarrow Q$, where Q is a positive definite matrix, then BIC is consistent - the probability of choosing the proper model converges to 1.

When $n \geq 8$ BIC never chooses more regressors than AIC and is usually considered as one of the most restrictive model selection criteria.

Bayesian Information Criterion (1)

M_j - j -th linear model with $k_j + l_j < n$ regressors

$\theta_j = (\beta_0, \beta_1, \dots, \beta_{k_j}, \gamma_1, \dots, \gamma_{l_j}, \sigma)$ - vector of model parameters

Bayesian Information Criterion (Schwarz, 1978) –

maximize $BIC = \log L(Y|M_j, \hat{\theta}_j) - \frac{1}{2}(k_j + l_j) \log n$

If m is fixed, $n \rightarrow \infty$ and $X'X/n \rightarrow Q$, where Q is a positive definite matrix, then BIC is consistent - the probability of choosing the proper model converges to 1.

When $n \geq 8$ BIC never chooses more regressors than AIC and is usually considered as one of the most restrictive model selection criteria.

Surprise ? : - Broman and Speed (JRSS, 2002) report that BIC overestimates the number of regressors when applied to QTL mapping.

BIC neglects the prior distribution on the class of possible models

Explanation - Bayesian roots of BIC

BIC neglects the prior distribution on the class of possible models
 \equiv assigning probability $p = \frac{1}{2}$ for inclusion of any given regressor

Explanation - Bayesian roots of BIC

BIC neglects the prior distribution on the class of possible models
 \equiv assigning probability $p = \frac{1}{2}$ for inclusion of any given regressor
 \equiv the prior distribution on the number of additive and interaction effects, $k + l$, is Binomial $B(N_m + N_e, \frac{1}{2})$, where N_m is the number of markers and $N_e = \binom{N_m}{2}$ is the number of possible two-way interaction terms

Explanation - Bayesian roots of BIC

BIC neglects the prior distribution on the class of possible models
 \equiv assigning probability $p = \frac{1}{2}$ for inclusion of any given regressor
 \equiv the prior distribution on the number of additive and interaction effects, $k + l$, is Binomial $B(N_m + N_e, \frac{1}{2})$, where N_m is the number of markers and $N_e = \binom{N_m}{2}$ is the number of possible two-way interaction terms

Example: for $N_m = 200$, $E(k + l) = 10050$

Modified version of BIC, mBIC (1)

M. Bogdan, J.K. Ghosh, R.W. Doerge, *Genetics* (2004)

Proposed solution - supplementing BIC with an informative prior distribution on the set of possible models, proposed in George and McCulloch (1993)

M. Bogdan, J.K. Ghosh, R.W. Doerge, *Genetics* (2004)

Proposed solution - supplementing BIC with an informative prior distribution on the set of possible models, proposed in George and McCulloch (1993)

ν_1 - prior probability that a randomly chosen marker is associated with QTL

ν_2 - prior probability that a randomly chosen pair of markers has an interaction effect

M. Bogdan, J.K. Ghosh, R.W. Doerge, *Genetics* (2004)

Proposed solution - supplementing BIC with an informative prior distribution on the set of possible models, proposed in George and McCulloch (1993)

ν_1 - prior probability that a randomly chosen marker is associated with QTL

ν_2 - prior probability that a randomly chosen pair of markers has an interaction effect

$$\pi(M) = \nu_1^p \nu_2^q (1 - \nu_1)^{N_m - p} (1 - \nu_2)^{N_e - q}$$

Modified version of BIC recommends choosing the model maximizing

$$\begin{aligned} mBIC &= BIC - 2 \log \pi(M) && (0.1) \\ &= n \log RSS + (p + q) \log n + 2p \log \left(\frac{N_m}{E_m} - 1 \right) \\ &+ 2q \log \left(\frac{N_e}{E_e} - 1 \right). \end{aligned}$$

Modified version of BIC recommends choosing the model maximizing

$$\begin{aligned} mBIC &= BIC - 2 \log \pi(M) && (0.1) \\ &= n \log RSS + (p + q) \log n + 2p \log \left(\frac{N_m}{E_m} - 1 \right) \\ &+ 2q \log \left(\frac{N_e}{E_e} - 1 \right). \end{aligned}$$

Standard version $E_m = E_e = 2.2$. For $n \geq 200$ controls FWER at the level below 10%

1. Extending to the intercross + iterative version of mBIC : Baierl, Bogdan, Frommlet, Futschik *Genetics*, 2006
2. Robust versions based on M-estimates: Baierl, Futschik, Bogdan, Biecek *CSDA*, 2007
3. Nonparametric rank version, rBIC: Źak, Baierl, Bogdan, Futschik *Genetics*, 2007
4. Dense maps and interval mapping: Bogdan, Frommlet, Biecek, Cheng, Ghosh, Doerge, *Biometrics*, 2008

Relationship to multiple testing (1)

Bogdan, Ghosh, Żak-Szatkowska, 2008

Relationship to multiple testing (1)

Bogdan, Ghosh, Żak-Szatkowska, 2008

Orthogonal design: $X^T X = nI_{(m+1) \times (m+1)}$, (1)

Relationship to multiple testing (1)

Bogdan, Ghosh, Żak-Szatkowska, 2008

Orthogonal design: $X^T X = nI_{(m+1) \times (m+1)}$, (1)

BIC chooses those X_j 's for which

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \log n$$

Relationship to multiple testing (1)

Bogdan, Ghosh, Żak-Szatkowska, 2008

Orthogonal design: $X^T X = nI_{(m+1) \times (m+1)}$, (1)

BIC chooses those X_j 's for which

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \log n$$

It holds that for large values of n

$$\alpha_n = 2P(Z_j > \sqrt{\log n}) \approx \sqrt{\frac{2}{\pi n \log n}}.$$

Relationship to multiple testing (1)

Bogdan, Ghosh, Żak-Szatkowska, 2008

Orthogonal design: $X^T X = nI_{(m+1) \times (m+1)}$, (1)

BIC chooses those X_j 's for which

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \log n$$

It holds that for large values of n

$$\alpha_n = 2P(Z_j > \sqrt{\log n}) \approx \sqrt{\frac{2}{\pi n \log n}}.$$

When n and m go to infinity and the number of true signals remains fixed, the expected number of “false discoveries” is of the rate $\frac{m}{\sqrt{n \log n}}$.

Relationship to multiple testing (1)

Bogdan, Ghosh, Żak-Szatkowska, 2008

Orthogonal design: $X^T X = nI_{(m+1) \times (m+1)}$, (1)

BIC chooses those X_j 's for which

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \log n$$

It holds that for large values of n

$$\alpha_n = 2P(Z_j > \sqrt{\log n}) \approx \sqrt{\frac{2}{\pi n \log n}}.$$

When n and m go to infinity and the number of true signals remains fixed, the expected number of “false discoveries” is of the rate $\frac{m}{\sqrt{n \log n}}$.

Corollary: BIC is not consistent when $\frac{m}{\sqrt{n \log n}} \rightarrow \infty$

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

Probability of detecting at least one “false positive”: $\text{FWER} \leq \alpha_n$

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

Probability of detecting at least one "false positive": $\text{FWER} \leq \alpha_n$

$$2(1 - \Phi(\sqrt{c_{Bon}})) = \frac{\alpha_n}{m}$$

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

Probability of detecting at least one “false positive”: $\text{FWER} \leq \alpha_n$

$$2(1 - \Phi(\sqrt{c_{Bon}})) = \frac{\alpha_n}{m}$$

$$c_{Bon} = 2 \log \left(\frac{m}{\alpha_n} \right) (1 + o_{n,m}) = (\log n + 2 \log m)(1 + o_{n,m})$$

where $o_{n,m}$ converges to zero when n or m tends to infinity.

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

Probability of detecting at least one “false positive”: $\text{FWER} \leq \alpha_n$

$$2(1 - \Phi(\sqrt{c_{Bon}})) = \frac{\alpha_n}{m}$$

$$c_{Bon} = 2 \log \left(\frac{m}{\alpha_n} \right) (1 + o_{n,m}) = (\log n + 2 \log m)(1 + o_{n,m})$$

where $o_{n,m}$ converges to zero when n or m tends to infinity.

$$c_{mBIC} = \log n + 2 \log \left(\frac{m}{c} - 1 \right) \approx \log n + 2 \log m - 2 \log c$$

$$1. FWER \approx \sqrt{\frac{2}{\pi}} \frac{c}{\sqrt{n(\log n + 2 \log m - 2 \log c)}}$$

1. $FWER \approx \sqrt{\frac{2}{\pi}} \frac{c}{\sqrt{n(\log n + 2 \log m - 2 \log c)}}$

2. The power of detecting the explanatory variable with $\beta_j \neq 0$ quickly converges to 1;

$$1 - P \left(-\sqrt{c_{mBIC}} - \frac{\sqrt{n}\beta_j}{\sigma} < \frac{\sqrt{n}(\hat{\beta}_j - \beta_j)}{\sigma} < \sqrt{c_{mBIC}} - \frac{\sqrt{n}\beta_j}{\sigma} \right) \\ > 1 - \Phi \left(\sqrt{c_{mBIC}} - \left| \frac{\sqrt{n}\beta_j}{\sigma} \right| \right) \rightarrow 1 \quad ,$$

1. $FWER \approx \sqrt{\frac{2}{\pi}} \frac{c}{\sqrt{n(\log n + 2 \log m - 2 \log c)}}$

2. The power of detecting the explanatory variable with $\beta_j \neq 0$ quickly converges to 1;

$$1 - P \left(-\sqrt{c_{mBIC}} - \frac{\sqrt{n}\beta_j}{\sigma} < \frac{\sqrt{n}(\hat{\beta}_j - \beta_j)}{\sigma} < \sqrt{c_{mBIC}} - \frac{\sqrt{n}\beta_j}{\sigma} \right) \\ > 1 - \Phi \left(\sqrt{c_{mBIC}} - \left| \frac{\sqrt{n}\beta_j}{\sigma} \right| \right) \rightarrow 1 \quad ,$$

Corollary: Independently on the choice of c mBIC is consistent

Asymptotic optimality of mBIC (1)

γ_0 - cost of the false discovery, γ_A - cost of missing the true signals

Asymptotic optimality of mBIC (1)

γ_0 - cost of the false discovery, γ_A - cost of missing the true signals

$$\beta_j \sim (1 - p)\delta_0 + pN(0, \tau^2)$$

Asymptotic optimality of mBIC (1)

γ_0 - cost of the false discovery, γ_A - cost of missing the true signals

$$\beta_j \sim (1 - p)\delta_0 + pN(0, \tau^2)$$

Expected value of the experiment cost:

$$R = m(\gamma_0 t_1 (1 - p) + \gamma_A t_2 p),$$

where t_1 and t_2 are type I and type II errors

Asymptotic optimality of mBIC (1)

γ_0 - cost of the false discovery, γ_A - cost of missing the true signals

$$\beta_j \sim (1 - p)\delta_0 + pN(0, \tau^2)$$

Expected value of the experiment cost:

$$R = m(\gamma_0 t_1 (1 - p) + \gamma_A t_2 p),$$

where t_1 and t_2 are type I and type II errors

Optimal rule: Bayes oracle

$$\frac{f_A(\hat{\beta}_j)}{f_0(\hat{\beta}_j)} > \frac{(1 - p)\gamma_0}{p\gamma_A},$$

where $f_A(\hat{\beta}_j) \sim N(0, \tau^2 + \frac{\sigma^2}{n})$ and $f_0(\hat{\beta}_j) \sim N(0, \frac{\sigma^2}{n})$

Asymptotic optimality of mBIC (2)

Bayes oracle

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \frac{\sigma^2 + n\tau^2}{n\tau^2} \left[\log \left(\frac{n\tau^2 + \sigma^2}{\sigma^2} \right) + 2 \log \left(\frac{1-p}{p} \right) + 2 \log \left(\frac{\gamma_0}{\gamma_A} \right) \right]$$

Asymptotic optimality of mBIC (2)

Bayes oracle

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \frac{\sigma^2 + n\tau^2}{n\tau^2} \left[\log \left(\frac{n\tau^2 + \sigma^2}{\sigma^2} \right) + 2 \log \left(\frac{1-p}{p} \right) + 2 \log \left(\frac{\gamma_0}{\gamma_A} \right) \right]$$

Asymptotic Optimality: the model selection rule V is asymptotically optimal if

$$\lim_{n \rightarrow \infty, m \rightarrow \infty} \frac{R_V}{R_{BO}} = 1 .$$

Asymptotic optimality of mBIC (2)

Bayes oracle

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \frac{\sigma^2 + n\tau^2}{n\tau^2} \left[\log \left(\frac{n\tau^2 + \sigma^2}{\sigma^2} \right) + 2 \log \left(\frac{1-p}{p} \right) + 2 \log \left(\frac{\gamma_0}{\gamma_A} \right) \right]$$

Asymptotic Optimality: the model selection rule V is asymptotically optimal if

$$\lim_{n \rightarrow \infty, m \rightarrow \infty} \frac{R_V}{R_{BO}} = 1 .$$

Theorem 1 (Bogdan, Chakrabarti, Ghosh, 2008). Under orthogonal design (1) mBIC is asymptotically optimal when $\lim_{m \rightarrow \infty} mp = s$, where $s \in \mathbf{R}$.

Asymptotic optimality of mBIC (2)

Bayes oracle

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \frac{\sigma^2 + n\tau^2}{n\tau^2} \left[\log \left(\frac{n\tau^2 + \sigma^2}{\sigma^2} \right) + 2 \log \left(\frac{1-p}{p} \right) + 2 \log \left(\frac{\gamma_0}{\gamma_A} \right) \right]$$

Asymptotic Optimality: the model selection rule V is asymptotically optimal if

$$\lim_{n \rightarrow \infty, m \rightarrow \infty} \frac{R_V}{R_{BO}} = 1 .$$

Theorem 1 (Bogdan, Chakrabarti, Ghosh, 2008). Under orthogonal design (1) mBIC is asymptotically optimal when $\lim_{m \rightarrow \infty} mp = s$, where $s \in \mathbf{R}$.

Conjecture (Frommlet, Bogdan, 2008). Theorem 1 holds also when

$$\beta_j \sim (1-p)\delta_0 + pF_A,$$

where F_A has a positive density at 0.

Computer simulations(1)

Setting : $n = 200$, $m = 300$, entries of $X \sim N(0, \sigma = 0.5)$,

$k \sim \text{Binomial}(m, p)$, with $p = \frac{1}{30}$ ($mp = 10$), $\beta_i \sim N(0, \sigma = 1.5)$,

$\varepsilon \sim N(0, 1)$ and Tukey's gross error model:

$\varepsilon \sim \text{Tukey}(0.95, 100, 1) = 0.95 * N(0, 1) + 0.05 * N(0, 10)$.

Computer simulations(1)

Setting : $n = 200$, $m = 300$, entries of $X \sim N(0, \sigma = 0.5)$,

$k \sim \text{Binomial}(m, p)$, with $p = \frac{1}{30}$ ($mp = 10$), $\beta_i \sim N(0, \sigma = 1.5)$,

$\varepsilon \sim N(0, 1)$ and Tukey's gross error model:

$\varepsilon \sim \text{Tukey}(0.95, 100, 1) = 0.95 * N(0, 1) + 0.05 * N(0, 10)$.

Characteristics : Power, $FDR = \frac{FP}{AP}$, $MR = FP + FN$,

$$l_2 = \sum_{j=1}^m (\beta_j - \hat{\beta}_j)^2$$

mean value of the absolute prediction error based on 50 additional observations, d

Table: Results for 1000 replications.

noise criterion	N(0,1)			Tukey(0.95, 100, 1)		
	BIC	mBIC	rBIC	BIC	mBIC	rBIC
FP	13.3	0.073	0.08	12.5	0.08	0.1
FN	1.84	2.97	3.45	3.95	6.11	4.29
Power	0.8155	0.7030	0.6586	0.6087	0.3923	0.5806
FDR	0.5889	0.0107	0.0116	0.6487	0.0210	0.0162
MR	15.1480	3.0410	3.5310	16.4440	6.1910	4.3910
l_2	2.3610	0.6025	0.8500	13.51	4.732	1.597
d	0.9460	0.8505	0.8687	1.714	1.503	1.298

$$E|\varepsilon_1| \approx 0.8 \quad , \quad E|\varepsilon_2| \approx 1.16$$

Ckkrabarti and Ghosh (2006)

Loss - $\|X\beta_0 - X\hat{\beta}\|^2$.

mAIC performs better than AIC when the number of true regressors is small or when the number of true regressors is large but they have very small effects

Similar conclusions can drawn from Abramovich, Benjamini, Donoho i Johnstone (2006)

Correction for correlation

M. Bogdan, F. Frommlet, P. Biecek, R. Cheng, J.K. Ghosh, R.W. Doerge - *Biometrics*, 2008.

Correction for correlation

M. Bogdan, F. Frommlet, P. Biecek, R. Cheng, J.K. Ghosh, R.W. Doerge - *Biometrics*, 2008.

When the distance between markers converges to 0 then $\rho(X_{i+1}, X_i) \rightarrow 1$ and the penalty in mBIC goes to infinity.

Correction for correlation

M. Bogdan, F. Frommlet, P. Biecek, R. Cheng, J.K. Ghosh, R.W. Doerge - *Biometrics*, 2008.

When the distance between markers converges to 0 then $\rho(X_{i+1}, X_i) \rightarrow 1$ and the penalty in mBIC goes to infinity.

Consider m markers.

$$P_{H_0} \left(\max_{i \in \{1, \dots, m\}} LRT(i) > c \right) = \alpha$$

Correction for correlation

M. Bogdan, F. Frommlet, P. Biecek, R. Cheng, J.K. Ghosh, R.W. Doerge - *Biometrics*, 2008.

When the distance between markers converges to 0 then $\rho(X_{i+1}, X_i) \rightarrow 1$ and the penalty in mBIC goes to infinity.

Consider m markers.

$$P_{H_0} \left(\max_{i \in \{1, \dots, m\}} LRT(i) > c \right) = \alpha$$

Consider N independent tests.

$$P_{H_0} \left(\max_{i \in \{1, \dots, N\}} LRT(i) > c \right) \approx 1 - (1 - 2(1 - \Phi(\sqrt{c})))^N$$

Correction for correlation

M. Bogdan, F. Frommlet, P. Biecek, R. Cheng, J.K. Ghosh, R.W. Doerge - *Biometrics*, 2008.

When the distance between markers converges to 0 then $\rho(X_{i+1}, X_i) \rightarrow 1$ and the penalty in mBIC goes to infinity.

Consider m markers.

$$P_{H_0} \left(\max_{i \in \{1, \dots, m\}} LRT(i) > c \right) = \alpha$$

Consider N independent tests.

$$P_{H_0} \left(\max_{i \in \{1, \dots, N\}} LRT(i) > c \right) \approx 1 - (1 - 2(1 - \Phi(\sqrt{c})))^N$$

$$N = \frac{\log(1 - \alpha)}{\log(2\Phi(\sqrt{c}) - 1)} \quad \text{weight; } w = \frac{N}{m}$$

How to find c ? (1)

Siegmund and Dupuis (Genetics, 1999): Approximation of the distribution of the sequence of LRT statistics by the square of the Orenstein-Uhlenbeck process

How to find c ? (1)

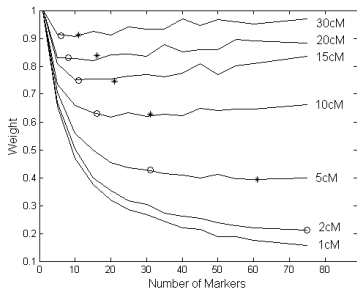
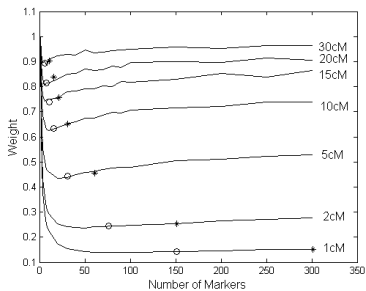
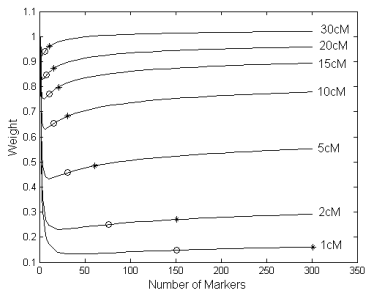
Siegmund and Dupuis (Genetics, 1999): Approximation of the distribution of the sequence of LRT statistics by the square of the Orenstein-Uhlenbeck process

δ - distance between markers, L - the chromosome length

$$P\left(\max_{i \in \{1, \dots, N\}} LRT(i) > c\right) \approx 1 - \exp\left(-2\left[1 - \Phi(\sqrt{c})\right] - 0.04L\sqrt{c}\phi(\sqrt{c})\nu\left(\sqrt{0.04c\delta}\right)\right)$$

$$\nu(t) = 2t^{-2} \exp\left[-2 \sum_{n=1}^{\infty} n^{-1} \Phi\left(-\frac{1}{2}|t|n^{1/2}\right)\right].$$

Weights (1)



Interval mapping (1)

Performing tests at a dense set of locations within markers, based on the information in flanking markers. Two most popular methods - EM algorithm (Lander and Botstein , 1989), Regression Mapping (Haley and Knott, 1992)

Interval mapping (1)

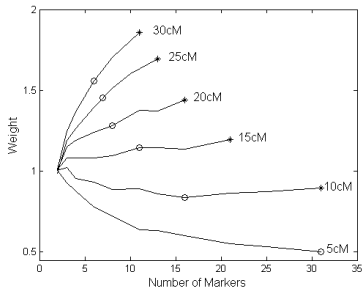
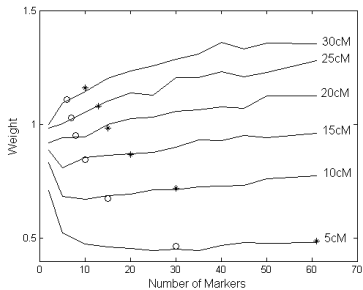
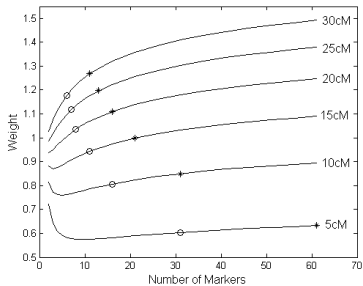
Performing tests at a dense set of locations within markers, based on the information in flanking markers. Two most popular methods - EM algorithm (Lander and Botstein , 1989), Regression Mapping (Haley and Knott, 1992)

Rebaï et al. (1993) (location - additional parameter):

$$Pr \left(\sup_{0 \leq x \leq \sum_{i=1}^k \delta_i} LTR(x) > c^2 \right) \approx 2\Phi(-c) + \frac{2}{\pi} \exp \left(-\frac{1}{2}c^2 \right) \sum_{i=1}^k \arctan \left(\sqrt{\frac{r_i}{1-r_i}} \right)$$

k - the number of intervals, δ_i - the length of i th interval r_i - the probability of recombination on i th interval.

Weights (2)



Approximate formulas for weights

$$\hat{w}_{SM}^{add}(\delta) = 1 - 0.9e^{(-10\delta/100+10(\delta/100)^2)} ,$$

$$\hat{w}_{SM}^{epi}(\delta) = 1 - e^{(-10.7\delta/100+8.7(\delta/100)^2)} ,$$

$$\hat{w}_{IM}^{add}(\delta) = -0.15 + 3.1\sqrt{\delta/100} - 1.3\delta/100 ,$$

$$\hat{w}_{IM}^{epi}(\delta) = -0.53 + 5.4\sqrt{\delta/100} - 2.7\delta/100 ,$$

When markers are differently spaced we use the weights w^{add} and w^{epi} corresponding to the average intermarker distance.

When markers are differently spaced we use the weights w^{add} and w^{epi} corresponding to the average intermarker distance.

Standard version of mBIC:

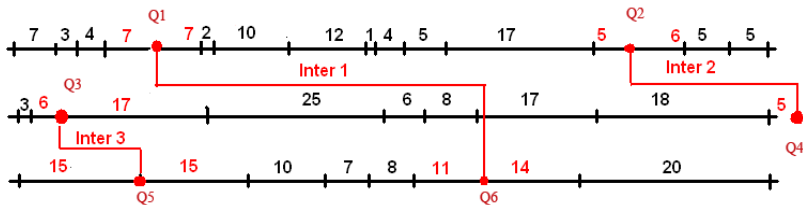
$$n \log RSS + (p + q) \log n + 2p \log \left(\frac{w^{add} N_m}{2.2} \right) + 2r \log \left(\frac{w^{epi} N_e}{2.2} \right)$$

Dense marker spacings - 2, 5, 10 cM; Interval Mapping - markers every 5, 10, and 25cM

Computer simulations

Dense marker spacings - 2, 5, 10 cM; Interval Mapping - markers every 5, 10, and 25cM

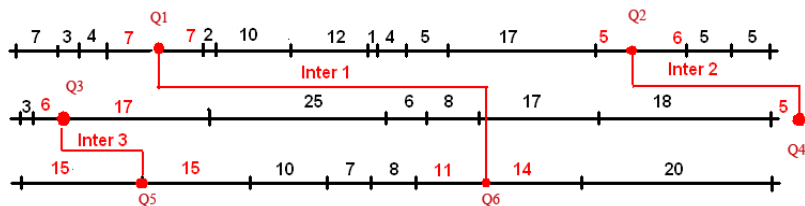
More realistic scenario:



Computer simulations

Dense marker spacings - 2, 5, 10 cM; Interval Mapping - markers every 5, 10, and 25cM

More realistic scenario:



For power simulations : $\beta_{Q1} = 0.6$, $\beta_{Q4} = 0.7$, $\beta_{Q6} = 0.5$,
 $\gamma_{Q1Q6} = 1.2$, $\gamma_{Q2Q4} = 1.4$ and $\gamma_{Q3Q5} = 1$, $\varepsilon = 1$. The overall
broad sense trait heritability $H^2 = 0.355$: $h^2_{Q1} = h^2_{Q1Q6} = 0.058$,
 $h^2_{Q4} = h^2_{Q2Q4} = 0.08$, $h^2_{Q6} = h^2_{Q3Q5} = 0.04$.

Type I error

n	method	δ	e_m	e_e	e_t
200	SM	2cM	4.4	2.4	6.9
	SM	5cM	4.8	2.9	7.8
	SM	10cM	3.7	3.6	7.3
	SM	Fig.	4.8	3.0	7.8
200	IM	5 cM	4.9	2.9	7.8
	IM	10 cM	4.1	4.0	8.1
	IM	25 cM	3.9	4.3	8.2
	IM	Fig. 3	4.8	3.1	7.9
500	SM	2cM	1.7	1.5	3.2
	SM	5cM	2.9	2.4	5.3
	SM	10cM	2.8	2	4.8
	SM	Fig. 3	3.3	2.4	5.7
500	IM	5 cM	3	2.6	5.6
	IM	10 cM	2.4	2.6	5.0
	IM	25 cM	3.1	2.2	5.3
	IM	Fig. 3	3.2	2.1	5.3

Power(1)

n	scenario	Q1 <i>pow</i> <i>std</i>	Q4 <i>pow</i> <i>std</i>	Q6 <i>pow</i> <i>std</i>	Int 1 <i>pow</i> <i>std1</i> <i>std6</i>	Int 2 <i>pow</i> <i>std2</i> <i>std4</i>	Int 3 <i>pow</i> <i>std3</i> <i>std5</i>	<i>pdf</i>
200	SM	0.67 10.2	0.76 6.9	0.35 14.8	0.24 10.9 14.5	0.47 9.2 7.8	0.07 10.1 16.6	0.06
200	IM	0.69 9.8	0.76 8.1	0.39 13.2	0.28 9.7 12.1	0.47 8.0 8.6	0.10 9.0 11.4	0.07
200	2cM	0.71 7.8	0.75 9.6	0.53 9.8	0.40 8.3 8.2	0.45 7.3 9.7	0.19 7.1 8.2	0.06

Power (2)

n	scenario	Q1 <i>pow</i> <i>std</i>	Q4 <i>pow</i> <i>std</i>	Q6 <i>pow</i> <i>std</i>	Int 1 <i>pow</i> <i>std1</i> <i>std6</i>	Int 2 <i>pow</i> <i>std2</i> <i>std4</i>	Int 3 <i>pow</i> <i>std3</i> <i>std5</i>	<i>pfp</i>
500	SM	0.96 8.0	0.98 5.0	0.86 12.9	0.79 8.2 13.0	0.94 5.8 5.0	0.40 9.2 15.0	0.07
500	IM	0.97 5.6	0.997 7.0	0.83 6.3	0.80 6.2 6.1	0.97 4.7 6.9	0.48 5.7 6.8	0.07
500	2cM	0.99 3.6	0.985 6.3	0.93 4.3	0.955 3.8 3.4	0.97 3.2 6.2	0.84 3.8 4.4	0.04

The proposed modifications of mBIC allow to keep type I error at the desired level. Both type I and type II errors quickly decrease when the sample size increases.

The proposed modifications of mBIC allow to keep type I error at the desired level. Both type I and type II errors quickly decrease when the sample size increases.

Using dense genetic maps can substantially increase the power of detecting QTL and the precision of their localization

The proposed modifications of mBIC allow to keep type I error at the desired level. Both type I and type II errors quickly decrease when the sample size increases.

Using dense genetic maps can substantially increase the power of detecting QTL and the precision of their localization

Using multiple interval mapping has a relatively small influence on the power of QTL detection but it substantially increases the precision of QTL localization.

Real Data Analysis (1)

Zeng et al. (2000) data on the morphological differences between two species of *Drosophila*, *Drosophila simulans* and *Drosophila mauritana*

Real Data Analysis (1)

Zeng et al. (2000) data on the morphological differences between two species of *Drosophila*, *Drosophila simulans* and *Drosophila mauritana*

Trait - the size and the shape of the posterior lobe of the male genital arch, quantified by a morphometric descriptor.

Real Data Analysis (1)

Zeng et al. (2000) data on the morphological differences between two species of *Drosophila*, *Drosophila simulans* and *Drosophila mauritana*

Trait - the size and the shape of the posterior lobe of the male genital arch, quantified by a morphometric descriptor.

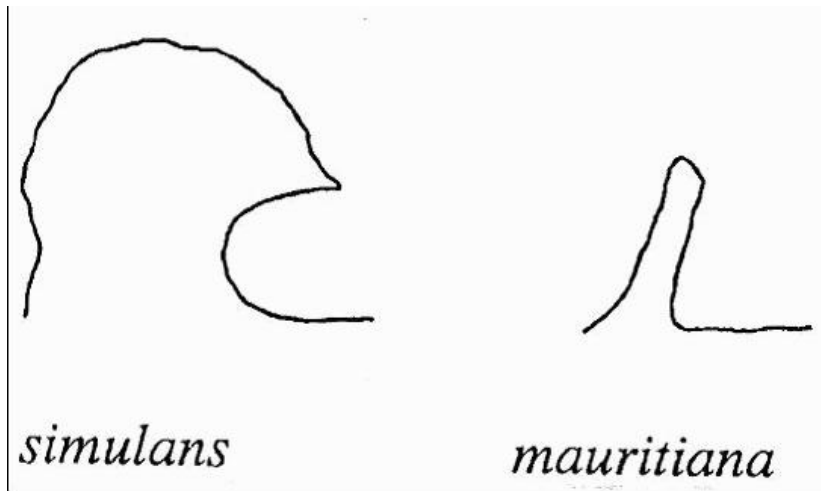
Fourier analysis:

$$x(s) = a_0 + \sum_{i=1}^{25} (a_i \cos(is) + b_i \sin(is))$$

$$y(s) = c_0 + \sum_{i=1}^{25} (c_i \cos(is) + d_i \sin(is))$$

Analyzed trait - first principal component of $[a_1, \dots, a_{25}, b_1, \dots, d_{25}]$

Real Data Analysis (2)



F1 - the cross resulting from hybridizing simulans females with mauritana males

F1 females are fertile and F1 males have intermediate posterior lobe morphology

F1 females are backcrossed to parental males - continuous series of morphologies, suggesting polygenic inheritance.

Data on both backcrosses - BS and BM.

Each backcross - two independent samples, collected at different times.

Marker data - three chromosomes: chromosome X and two autosomal chromosomes, 6 markers on chromosome X, 16 markers on chromosome 2, and 23 markers on chromosome 3.

Zeng et al. (2000) analyze all data together using the model

$$y_{ijk} = \mu_{ij} + \sum_{r=1}^m \beta_{ir} x_{ijk r} + \sum_{r \neq s \subset (1, \dots, m)} \gamma_{irs} x_{ijk r} x_{ijk s} + \varepsilon_{ijk} .$$

sample size, $n=962$

Zeng et. al. (2000) use MIM based on EM algorithm.

Search procedure starts from the initial set of locations identified with Composite Interval Mapping. The threshold value for including, deleting effects is computed using the residual permutation test (Doerge and Churchill, 1996) at the genomewide significance level 0.05. The search for interactions is limited to the pairs of identified QTL.

Using the aggregated sample, Zeng et al. (2000) identify 17 QTL on the two autosomal chromosomes and six interactions (only in BM). $R^2 = 92\%$, all main effects have the same sign - the trait is subject to strong selection.

Real Data Analysis (6)

We analyze data separately for each backcross. Sample sizes - 471 for BS and 491 for BM.

We use interval mapping version of mBIC. The average length of the intermarker interval on the three chromosomes is 9.04cM. The corresponding weights for main and epistatic effects are equal to $\hat{w}_{IM}^{add} = 0.66$ and $\hat{w}_{IM}^{epi} = 0.85$.

First round of forward selection identifies 16 main effects and 1 epistatic effect for BS and 13 main effects and 2 epistatic effects for BM.

Second step with modified coefficients according to Baierl et al. *Genetics* 2006 (2.2 replaced by 16 for BS and by 13 for BM) identifies one additional main effect, both for BS and BM.

Real Data Analysis (6)

We analyze data separately for each backcross. Sample sizes - 471 for BS and 491 for BM.

We use interval mapping version of mBIC. The average length of the intermarker interval on the three chromosomes is 9.04cM. The corresponding weights for main and epistatic effects are equal to $\hat{w}_{IM}^{add} = 0.66$ and $\hat{w}_{IM}^{epi} = 0.85$.

First round of forward selection identifies 16 main effects and 1 epistatic effect for BS and 13 main effects and 2 epistatic effects for BM.

Second step with modified coefficients according to Baierl et al. *Genetics* 2006 (2.2 replaced by 16 for BS and by 13 for BM) identifies one additional main effect, both for BS and BM.

We use all subsets model selection with the relaxed mBIC to choose between our effects and Zeng et al. (2000) effects.

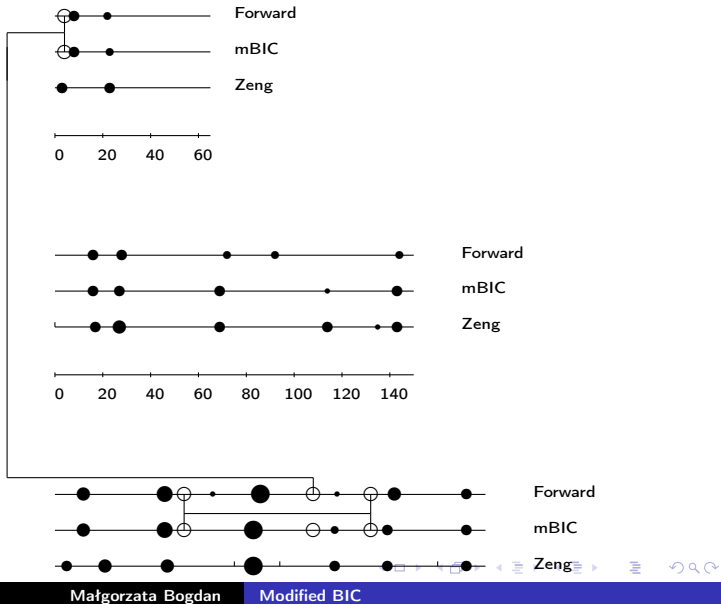
We identify 15 QTL with main effects for BS and 13 QTL with main effects and 2 epistatic effects for BM. All main effects have the same sign.

$R^2 = 90\%$ for BS and $R^2 = 84\%$ for BM.

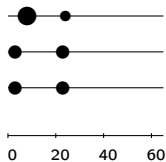
The estimated heritability of the weakest identified QTL $\approx 2\%$.

The estimates of heritabilities should be treated with caution. Since the estimation was preceded with model selection these estimates are biased upwards.

Real Data Analysis, BM



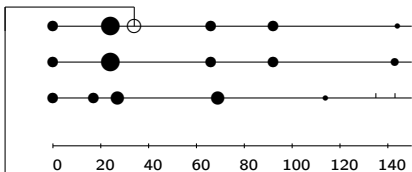
Real Data Analysis (8)



Forward

mBIC

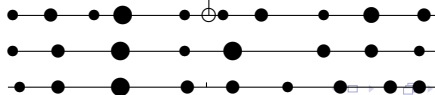
MIM



Forward

mBIC

MIM



Forward

mBIC

MIM

1. Relaxing the penalty so as to control FDR instead of FWER, expected optimality for a wider range of values of p - with F. Frommlet, J. K. Ghosh, A. Chakrabarti and M. Murawska.
2. Application for association mapping - with F. Frommlet and M. Murawska.
3. Application for GLM and Zero Inflated Generalized Poisson Regression, with M. Żak-Szatkowska, C. Czado, V. Earhardt.
4. Application for model selection in logic regression and comparison with Bayesian Regression Trees - with M. Malina, K. Ickstadt, H. Schwender.

1. Baierl, A., Bogdan, M., Frommlet, F., Futschik, A., 2006. On Locating multiple interacting quantitative trait loci in intercross designs. *Genetics* 173, 1693-1703.
2. Baierl, A., Futschik, A., Bogdan, M., Biecek, P., 2007. Locating multiple interacting quantitative trait loci using robust model selection, *Computational Statistics and Data Analysis* 51, 6423-6434.
3. Bogdan, M., Ghosh, J.K., Doerge, R.W., 2004. Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. *Genetics* 167, 989-999.
4. Bogdan, M., Frommlet, F., Biecek, P., Cheng, R., Ghosh, J. K., Doerge R. W. 2008 Extending the Modified Bayesian Information Criterion (mBIC) to dense markers and multiple interval mapping. *Biometrics*, doi: 10.1111/j.1541-0420.2008.00989.x.
5. Broman, K.W., Speed, T.P., 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Stat. Soc. B* 64, 641-656.
6. George, E.I., McCulloch, R.E., 1993. Variable Selection Via Gibbs Sampling. *J. Amer. Statist. Assoc.* 88 : 881-889.
7. Žak, M., Baierl, A., Bogdan, M., Futschik, A., 2007. Locating multiple interacting quantitative trait loci using rank-based model selection. *Genetics* 176, 1845-1854.