

Marker Based Infinitesimal Model for Quantitative Trait Analysis

Shizhong Xu
Department of Botany and Plant Sciences
University of California
Riverside, CA 92521

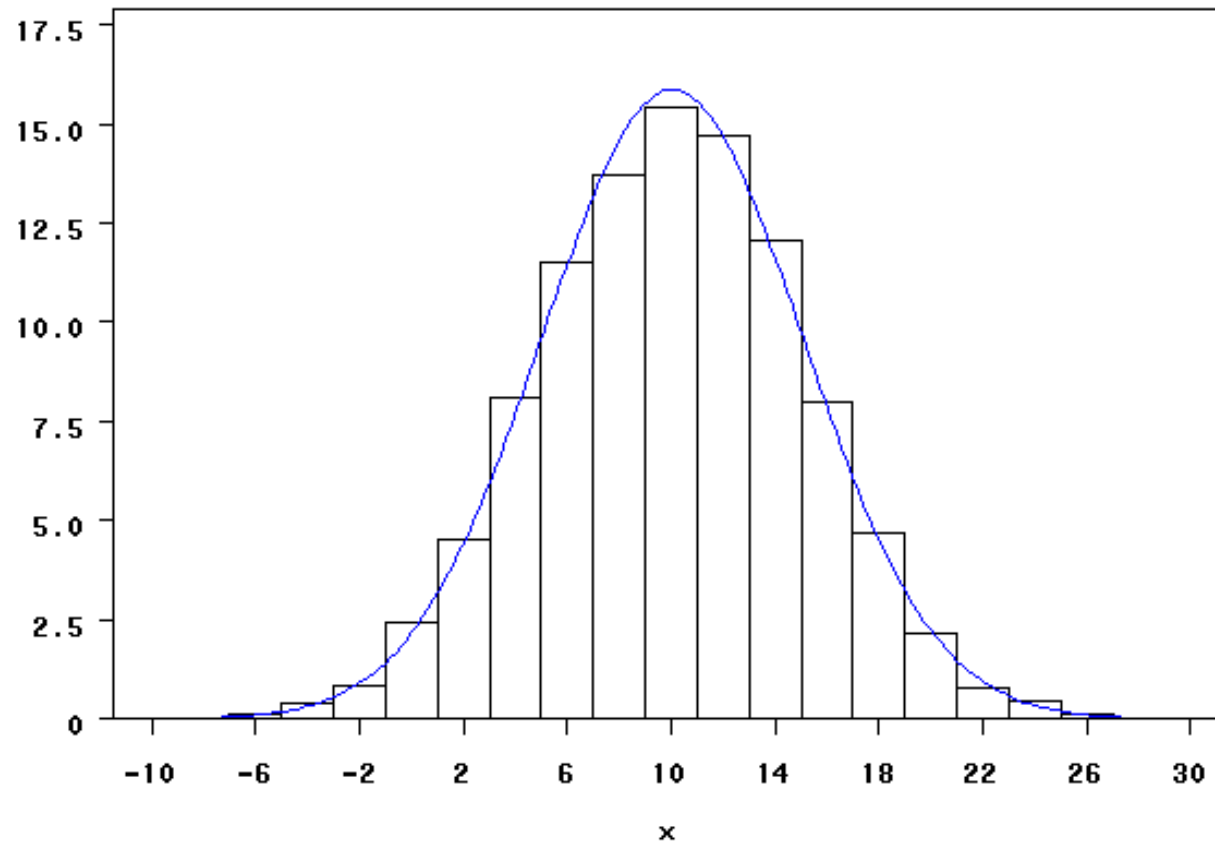
Outline

- Quantitative trait and the infinitesimal model
- Infinitesimal model using marker information
- Adaptive infinitesimal model
- Simulation studies
- Rice and beef cattle data analyses

Outline

- **Quantitative trait and the infinitesimal model**
- Infinitesimal model using marker information
- Adaptive infinitesimal model
- Simulation studies
- Rice and beef cattle data analyses

Quantitative Trait



Quantitative Genetics Model

Phenotype = Genotype + Environment

Infinitesimal Model

- Infinite number of genes
- Infinitely small effect of each gene
- Effect of an individual gene is not recognizable
- Collective effect of all genes are studied using pedigree information (genetic relationship)
- Best linear unbiased prediction (BLUP)

Outline

- Quantitative trait and the infinitesimal model
- **Infinitesimal model using marker information**
- Adaptive infinitesimal model
- Simulation studies
- Rice and beef cattle data analyses

Marker Based Infinitesimal Model

$$y_j = \beta + \sum_{k=1}^p Z_{jk} \gamma_k + \varepsilon_j$$

$$y_j = \beta + \sum_{k=1}^{\infty} Z_{jk} \gamma_k + \varepsilon_j$$

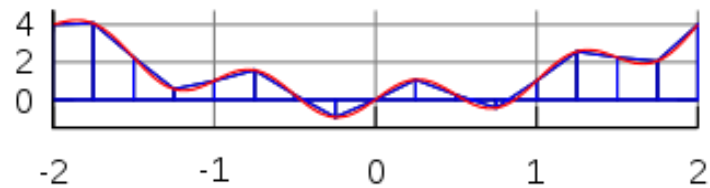
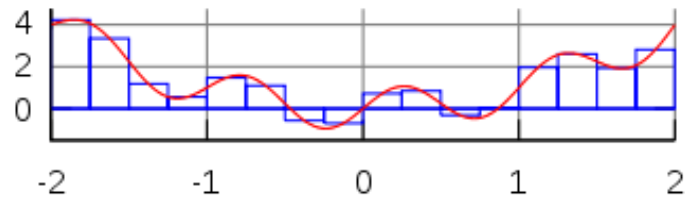
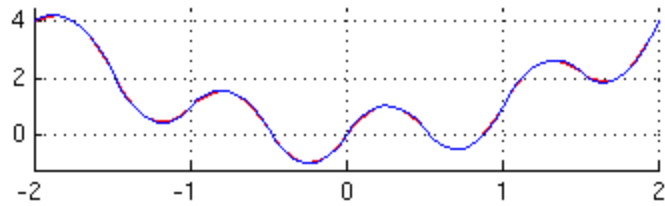
$$y_j = \beta + \int_0^L Z_j(\lambda) \gamma(\lambda) d\lambda + \varepsilon_j$$

Different from Longitudinal Data Analysis

$$y_j = \beta + \int_0^L Z_j(\lambda) \gamma(\lambda) d\lambda + \varepsilon_j$$

$$y_j(t) = \beta + \phi(t) + \varepsilon_j; t \in \Omega$$

Numerical Integration

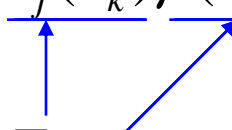


Bin Effect Model

$$y_j = \beta + \sum_{k=1}^{\infty} Z_{jk} \gamma_k + \varepsilon_j$$

$$y_j = \beta + \int_0^L Z_j(\lambda) \gamma(\lambda) d\lambda + \varepsilon_j$$

$$y_j = \beta + \sum_{k=1}^m \bar{Z}_j(\lambda_k) \bar{\gamma}(\lambda_k) \Delta_k + \varepsilon_j$$

$$y_j = \beta + \sum_{k=1}^m Z_{jk} \gamma_k + \varepsilon_j$$


Bin Effects

$$Z_{jk} = \frac{1}{p_k} \sum_{h=1}^{p_k} Z_j(h)$$

Dense markers



Bin



Bin



Recombination Breakpoint Data

$$\text{Marker: } Z_{jk} = \frac{1}{p_k} \sum_{h=1}^{p_k} Z_j(h)$$

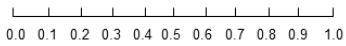
$$\text{Breakpoint: } Z_{jk} = \frac{1}{\Delta_k} \int_0^{\Delta_k} Z_j(\lambda) d\lambda$$

$$Z_{jk} = \frac{1}{\Delta_k} \int_0^{\Delta_k} Z_j(\lambda) d\lambda$$



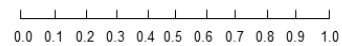
$$Z_{jk} = 1 \times \frac{8}{10} + 0 \times \frac{2}{10} = 0.8$$

Line	Breaking point pattern	Line	Start	Breaking point
1		1	1	0.385
2		2	0	0.795
3		3	1	0.590
4		4	0	
5		5	0	0.730
6		6	1	
7		7	1	0.260
8		8	0	0.320, 0.865, 0.935
9		9	1	0.065, 0.525
10		10	1	0.665
11		11	0	
12		12	1	0.130, 0.460
13		13	0	
14		14	1	
15		15	1	0.190



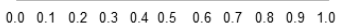
(a)

Line	Breaking point pattern	Line	Breaking point
1		1	G,0,0,0.385,R,0.385,1,0
2		2	R,0,0,0.795,G,0.795,1,0
3		3	G,0,0,0.59,R,0.59,1,0
4		4	R,0,0,1,0
5		5	R,0,0,0.73,G,0.73,1,0
6		6	G,0,0,1,0
7		7	G,0,0,0.26,R,0.26,1,0
8		8	R,0,0,0.32,G,0.32,0.865,R,0.865,0.935,G,0.935,1,0
9		9	G,0,0,0.065,R,0.065,0.525,G,0.525,1,0
10		10	G,0,0,0.665,R,0.665,1,0
11		11	R,0,0,1,0
12		12	G,0,0,0.13,R,0.13,0.46,G,0.46,1,0
13		13	R,0,0,1,0
14		14	G,0,0,1,0
15		15	G,0,0,0.19,R,0.19,1,0



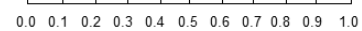
(b)

Line	Bin pattern	Line	Bin data
1		1	1 1 1 1 1 1 0 0 0 0 0 0 0 0 0
2		2	0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
3		3	1 1 1 1 1 1 1 1 1 0 0 0 0 0 0
4		4	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5		5	0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
6		6	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
7		7	1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
8		8	0 0 0 0 0 1 1 1 1 1 1 1 1 0 1
9		9	1 0 0 0 0 0 0 0 1 1 1 1 1 1 1
10		10	1 1 1 1 1 1 1 1 1 0 0 0 0 0 0
11		11	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
12		12	1 1 0 0 0 0 0 0 1 1 1 1 1 1 1
13		13	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14		14	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
15		15	1 1 1 0 0 0 0 0 0 0 0 0 0 0 0



(c)

Line	Bin pattern	Bin 1	Bin 2	Bin 3	Bin 4
1		1	0.54	0	0
2		0	0	0	0.82
3		1	1	0.36	0
4		0	0	0	0
5		0	0	0.08	1
6		1	1	1	1
7		1	0.04	0	0
8		0	0.72	1	0.72
9		0.16	0	0.9	1
10		1	1	0.66	0
11		0	0	0	0
12		0.52	0.16	1	1
13		0	0	0	0
14		1	1	1	1
15		0.76	0	0	0



(d)

What Does a Bin Effect Represent?

$$y_j = \beta + \sum_{k=1}^m Z_{jk} \gamma_k + \varepsilon_j$$

$$Z_{jk} = \frac{1}{\Delta_k} \int_0^{\Delta_k} Z_j(\lambda) d\lambda$$

$$\gamma_k = \Delta_k \times \frac{1}{\Delta_k} \int_0^{\Delta_k} \gamma(\lambda) d\lambda = \int_0^{\Delta_k} \gamma(\lambda) d\lambda$$

Δ_k = size of bin k

λ = uniform variable

Assumptions of the Infinitesimal Model

- High linkage disequilibrium within a bin
- Homogeneous genetic effect within a bin

High Linkage Disequilibrium

$$Z_{jk} = \frac{1}{\Delta_k} \int_0^{\Delta_k} Z_j(\lambda) d\lambda$$

Δ_k = number of crossovers, inversely
related to linkage disequilibrium

$\lim_{\Delta_k \rightarrow 0} \text{var}(Z_{jk}) = \frac{1}{2}$, high linkage disequilibrium (F_2)

$\lim_{\Delta_k \rightarrow \infty} \text{var}(Z_{jk}) = 0$, low linkage disequilibrium

Larger $\text{var}(Z_{jk})$ means higher power

Range of Var(Z)

$$\lim_{\Delta_k \rightarrow 0} \text{var}(Z_{jk}) = \lim_{\Delta_k \rightarrow 0} \frac{2\Delta_k + e^{-2\Delta_k} - 1}{4\Delta_k^2} = \lim_{\Delta_k \rightarrow 0} \frac{1}{2} e^{-2\Delta_k} = \frac{1}{2}$$

$$\lim_{\Delta_k \rightarrow \infty} \text{var}(Z_{jk}) = \lim_{\Delta_k \rightarrow \infty} \frac{2\Delta_k + e^{-2\Delta_k} - 1}{4\Delta_k^2} = \lim_{\Delta_k \rightarrow \infty} \frac{1}{2} e^{-2\Delta_k} = 0$$

$$0 \leq \text{var}(Z_{jk}) \leq 0.5$$

$$\infty \geq \Delta_k \geq 0$$

choose $\text{var}(Z_{jk})$ as close to 0.5 as possible

but with the number of bins small enough to be handled by a program for a given sample size

Outline

- Quantitative trait and the infinitesimal model
- Infinitesimal model using marker information
- **Adaptive infinitesimal model**
- Simulation studies
- Rice and beef cattle data analyses

Adaptive Model Relaxes the Two Assumptions

- High linkage disequilibrium within a bin
 - prevent $\text{var}(Z)$ from being zero
- Homogeneous genetic effect within a bin
 - make all effects positive

Redefine the Bin Size by the Number of Markers Within a Bin

$$y_j = \beta + \sum_{k=1}^m Z_{jk} \gamma_k + \varepsilon_j$$

$$Z_{jk} = \frac{1}{p_k} \sum_{h=1}^{p_k} Z_j(h)$$

$$\gamma_k = \bar{\gamma}_k p_k = \sum_{h=1}^{p_k} \gamma(h)$$

p_k = number of markers in bin k

Weighted Average Effect of a Bin

$$\text{Unweighted: } Z_{jk} = \frac{1}{p_k} \sum_{h=1}^{p_k} Z_j(h); \quad \gamma_k = \bar{\gamma}_k p_k = \sum_{h=1}^{p_k} \gamma(h)$$

$$\text{Weighted: } Z_{jk}^* = \frac{1}{p_k} \sum_{h=1}^{p_k} w(h) Z_j(h); \quad \gamma_k^* = \sum_{h=1}^{p_k} w^{-1}(h) \gamma(h)$$

$$y_j = \beta + \sum_{k=1}^m Z_{jk}^* \gamma_k^* + \varepsilon_j$$

Weight System

Define $c_k = \frac{1}{p_k} \sum_{h=1}^{p_k} |\hat{b}_h| = \text{mean}(|\hat{b}|)$

where \hat{b}_h is the least squares estimate of marker h within bin k

The weight for marker h is defined as

$$w_h = c_k^{-1} \hat{b}_h = \frac{p_k \hat{b}_h}{\sum_{h=1}^{p_k} |\hat{b}_h|} = \frac{\hat{b}_h}{\text{mean}(|\hat{b}|)}$$

Weighted Var(Z^*) > 0

$$\begin{aligned}\text{var}(Z_{jk}^*) &= \frac{1}{p_k^2} \left\{ \sum_{h=1}^{p_k} \text{var}[Z_j^*(h)] + 2 \sum_{l>h}^{p_k} \text{cov}[Z_j^*(h), Z_j^*(l)] \right\} \\ &= \frac{1}{p_k^2} \left\{ \frac{1}{2} \sum_{h=1}^{p_k} w_h^2 + 2 \times \frac{1}{2} \sum_{l>h}^{p_k} w_h w_l (1 - 2\delta_{hl}) \right\} \\ &= \frac{1}{p_k^2} \left\{ \frac{1}{2} \sum_{h=1}^{p_k} w_h^2 \right\}, \text{ when no linkage disequilibrium } (1 - 2\delta_{hl}) = 0 \\ &> 0\end{aligned}$$

Homogenization of Marker Effects Within Bin

$$\gamma_k^* = \sum_{h=1}^{p_k} w_h^{-1} \gamma(h) = c_k \sum_{h=1}^{p_k} \frac{\gamma(h)}{\hat{b}_h} = \rho c_k p_k = \rho \sum_{h=1}^{p_k} |\hat{b}_h|$$

where $\frac{\gamma(h)}{\hat{b}_h} \approx \rho$ (a constant)

$$\gamma_k^* = \rho \sum_{h=1}^{p_k} |\hat{b}_h| \neq 0 \text{ as long as one } \hat{b}_h \neq 0$$

Outline

- Quantitative trait and the infinitesimal model
- Infinitesimal model using marker information
- Adaptive infinitesimal model
- **Simulation studies**
- Rice and beef cattle data analyses

Measurement of Prediction (Cross Validation)

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2, \text{ Mean Squared Error } \downarrow$$

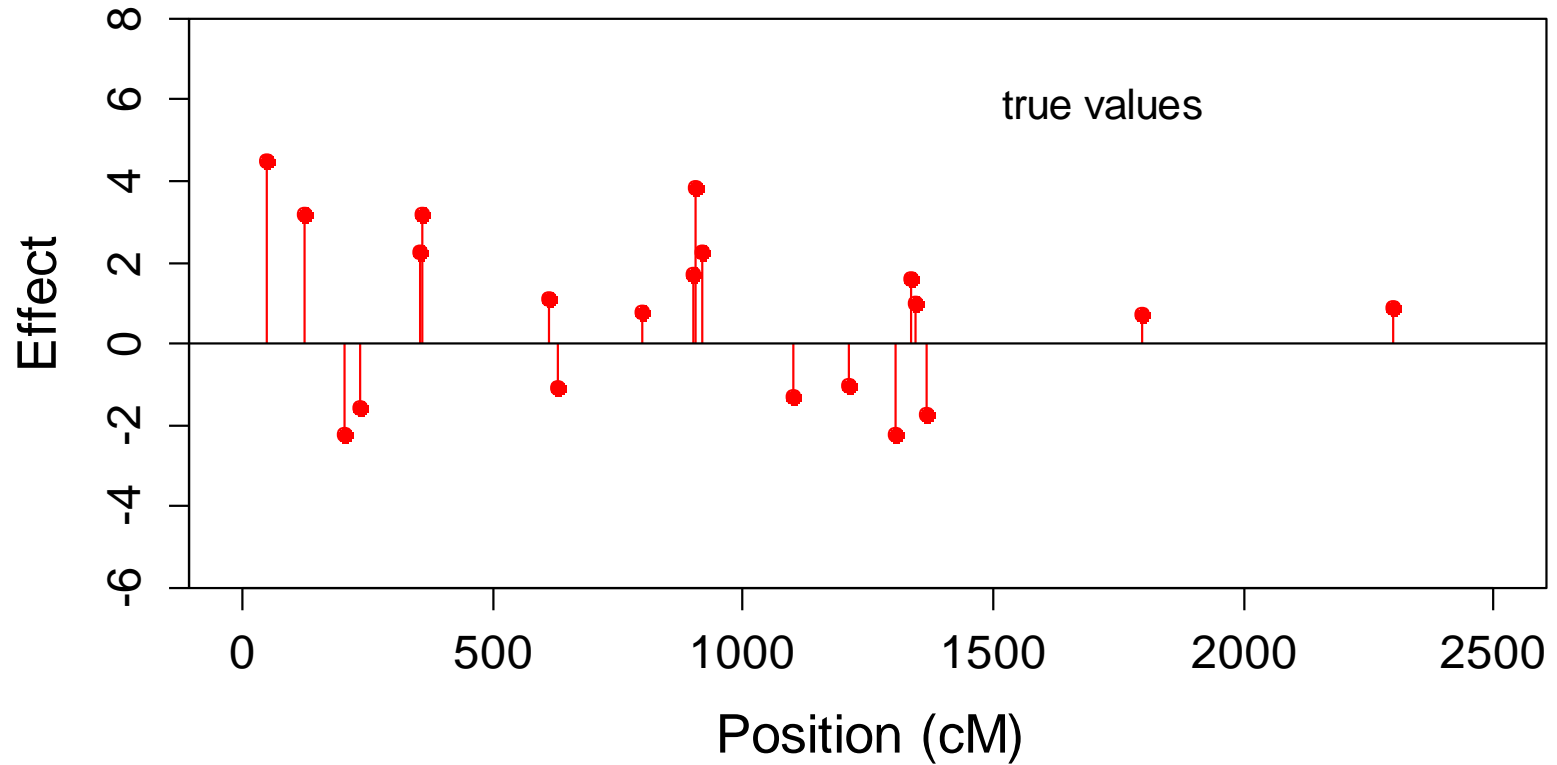
$$MSY = \frac{1}{n} \sum_{j=1}^n (y_j - \bar{y}_j)^2, \text{ Phenotypic Variance}$$

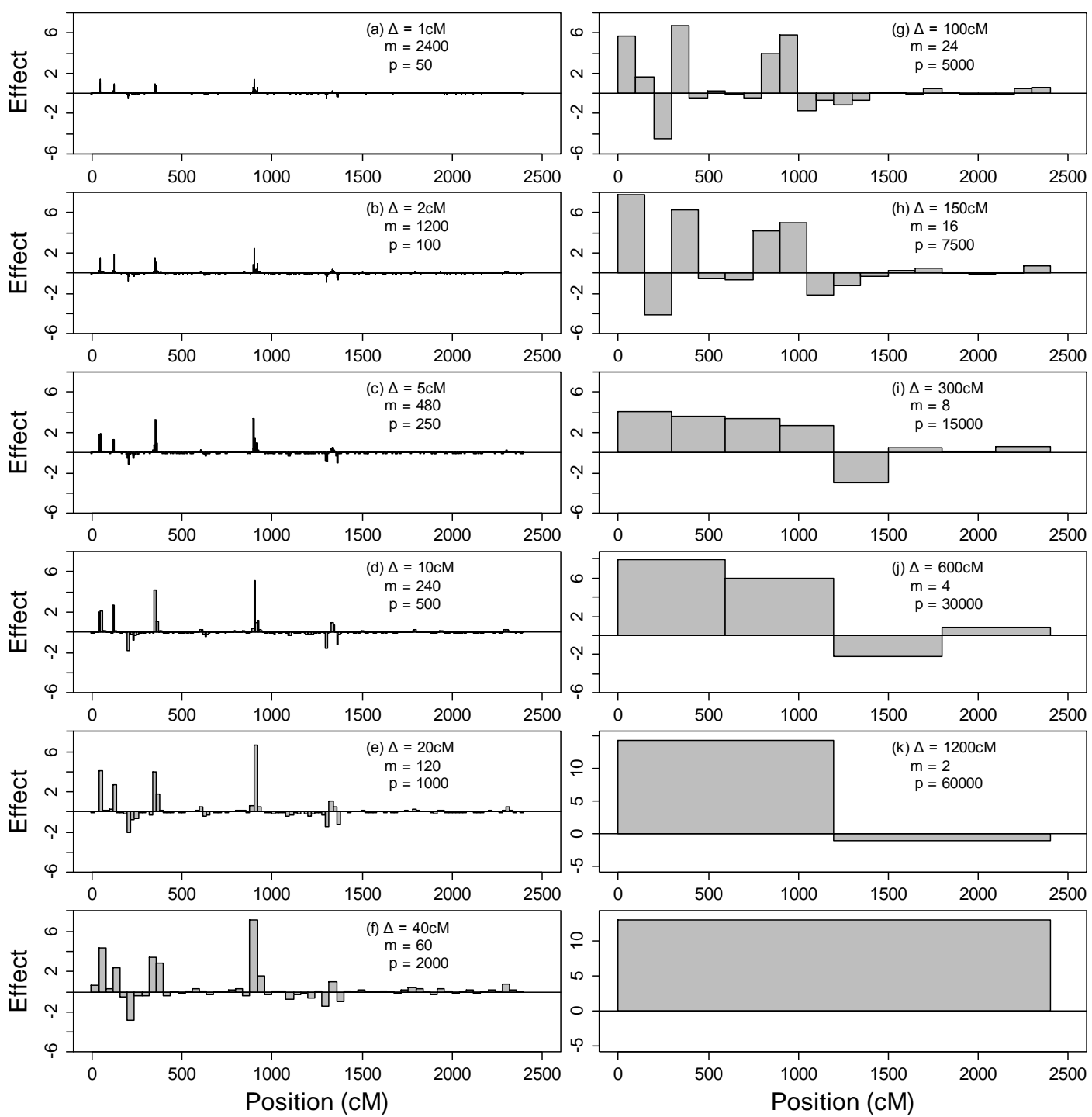
$$R^2 = \frac{MSY - MSE}{MSY}, \text{ Squared Correlation } \uparrow$$

Simulation Experiment

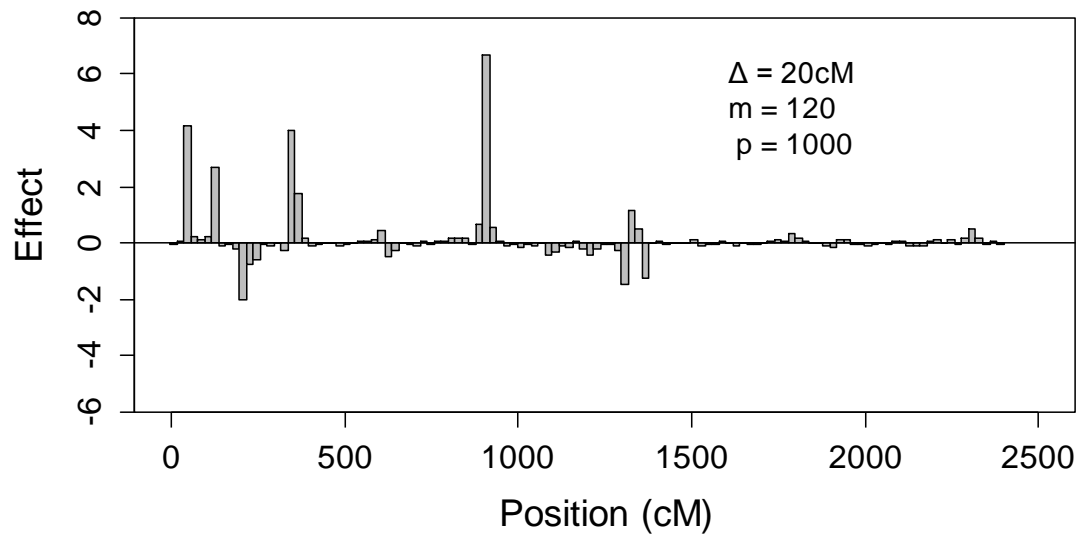
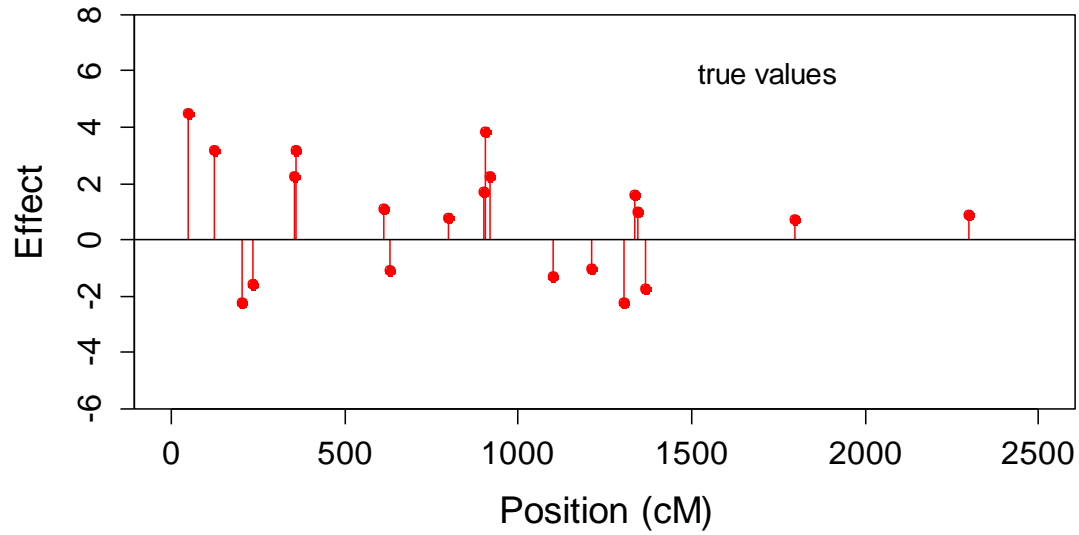
- Genome size = 2,500 cM
- Number of markers = 120,000
- Marker interval = 0.02 cM
- Cross validation (MSE)
- Design I = 20 QTL
- Design II = Clustered polygenic model
- Design III = Polygenic model
- Design IV = Design I with 2,500 x100 cM

True QTL Effect





True and Estimated QTL Effect



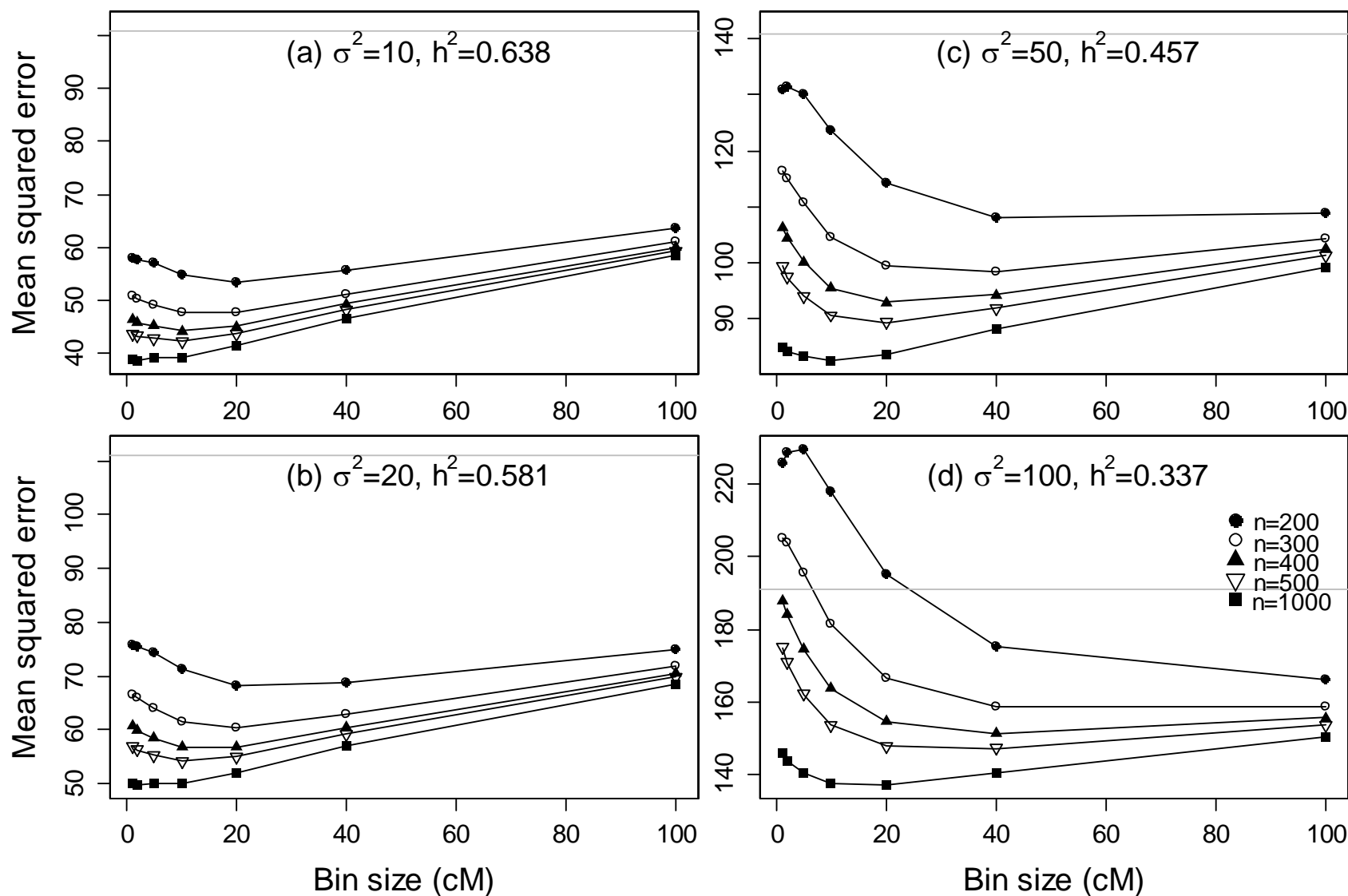


Figure 1. Mean squared error expressed as a function of bin size for Design I. The mean squared errors were obtained from 100 replicated simulations. The overall proportion of the phenotypic variance contributed by the 20 simulated QTL was calculated using $h^2 = 64.41 / (64.41 + 26.53 + \sigma^2)$. Each panel contains the result of five different sample sizes (n). The phenotypic variance of the simulated trait is indicated by the light horizontal line in each panel (each panel represents one of the four different scenarios).

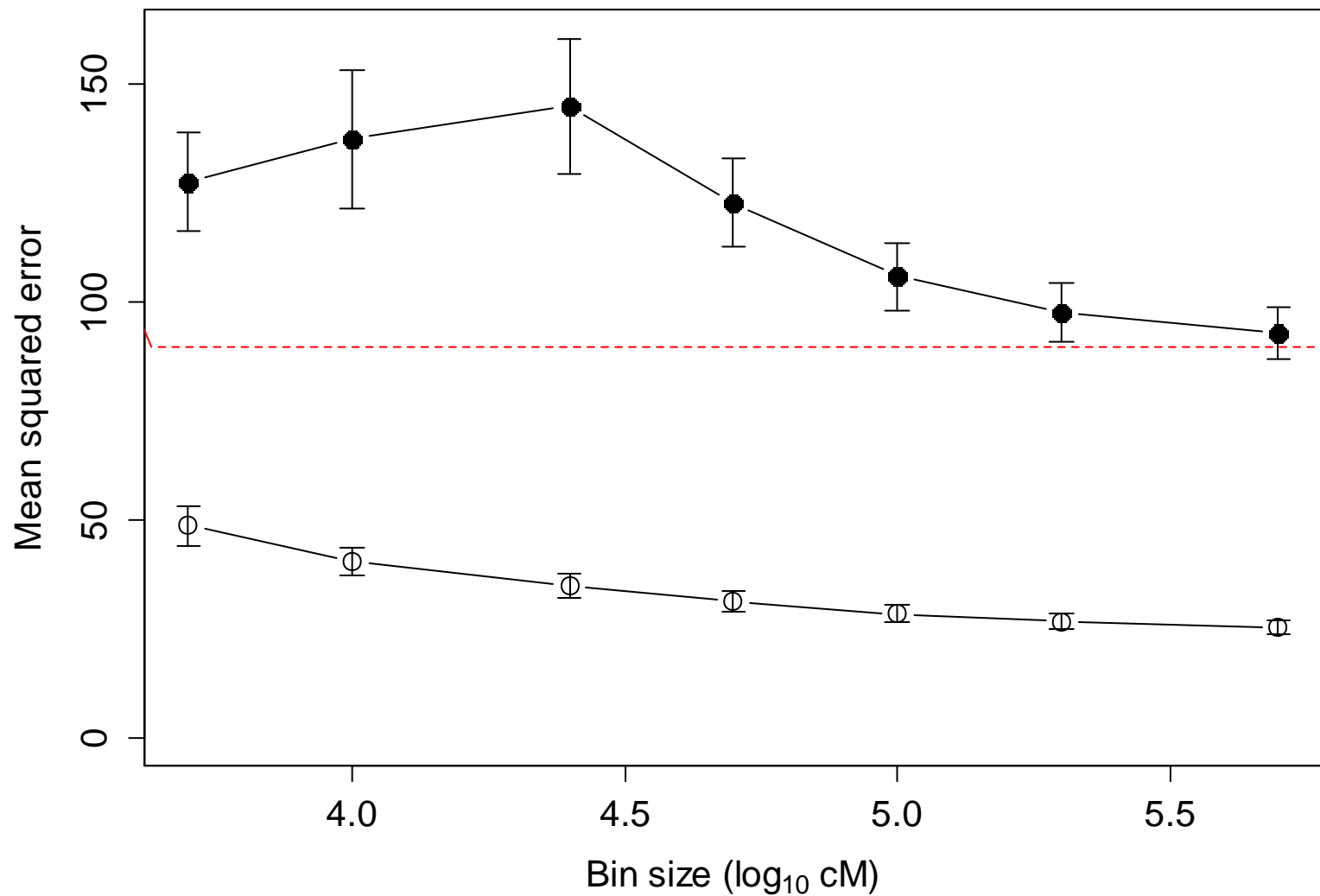


Figure 6. Mean squared error for the simulated data under design IV (low linkage disequilibrium plotted against the bin size. The sample size of the simulated population was $n = 500$. The residual error variance was $\sigma^2 = 20$, corresponding to $h^2 = 0.777$. The filled circles indicate the MSE under the infinitesimal model while the open circles indicate the MSE under the adaptive infinitesimal model. The dashed horizontal line represents the phenotypic variance of the simulated trait (89.71).

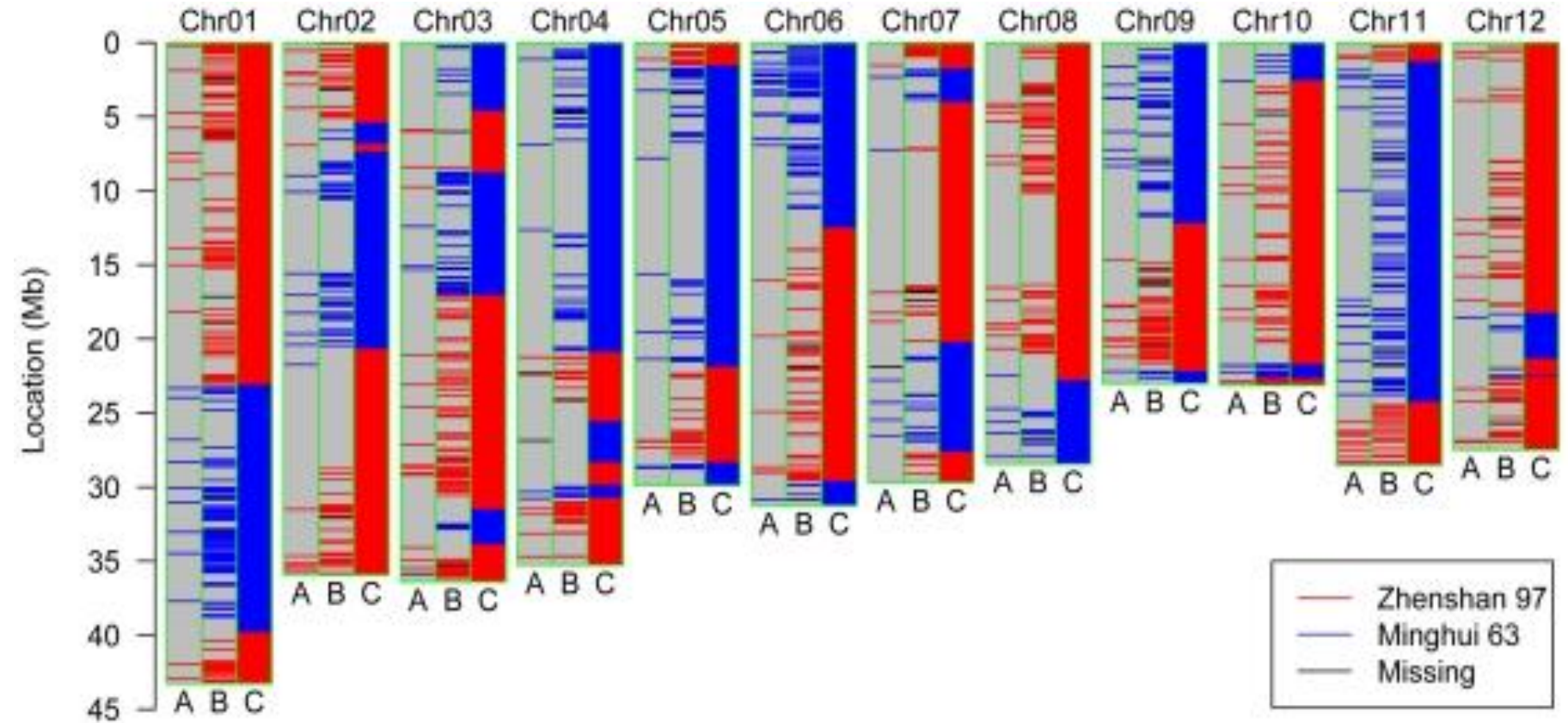
Outline

- Quantitative trait and the infinitesimal model
- Infinitesimal model using marker information
- Adaptive infinitesimal model
- Simulation studies
- **Rice and beef cattle data analyses**

Rice Tiller Number (Yu et al. 2011)

- Number of recombinant inbred lines: 210
- Number of SNP: 270,820
- Number of natural bins: 1619
- Number of artificial bins: vary from small to large
- Method: Empirical Bayes (eBayes)
- Cross validation: MSE and R-square

Yu et al. 2007, PLoS One 6(3) e17595



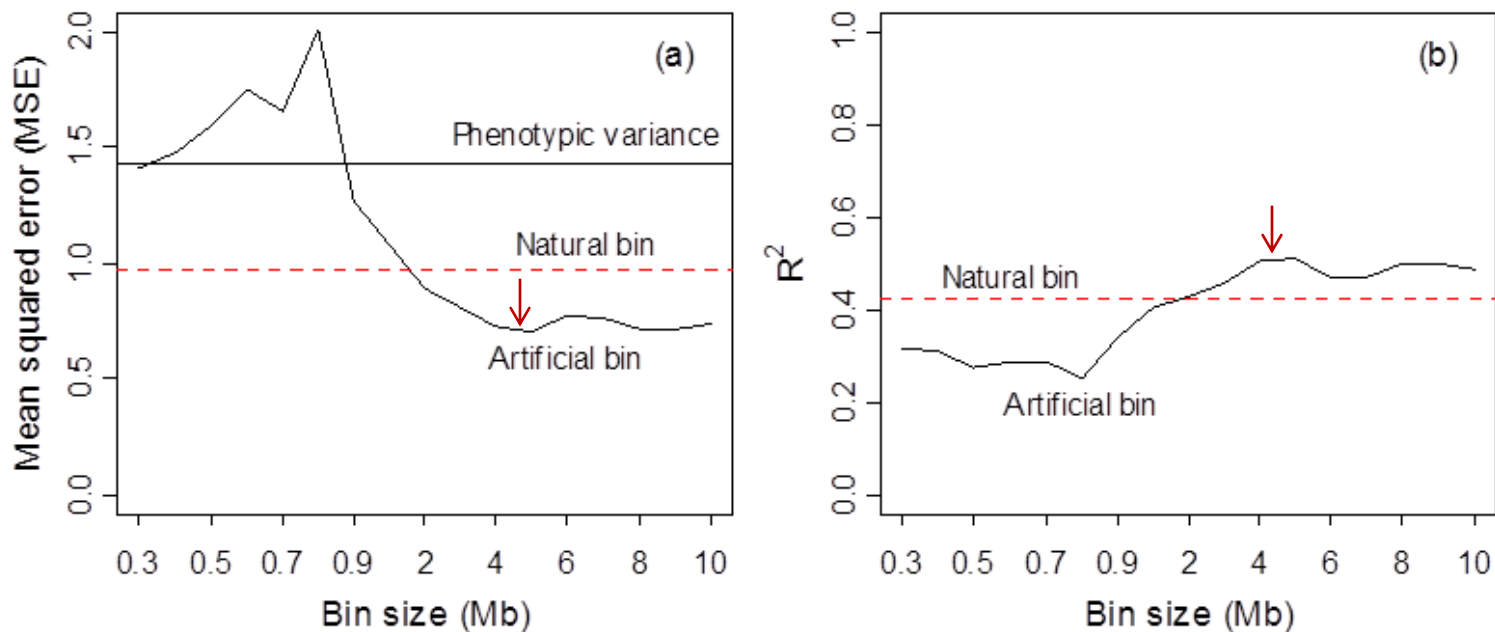


Figure 5. The MSE (curve in the left panel) and the R-square (curve in the right panel) of the rice tiller number trait analysis, expressed as a function of bin size (artificial bins). The black dashed horizontal line in the left panel is the phenotypic variance. The red dashed horizontal line in the left panel is the MSE of the natural bin (without breakpoints with bin) analysis. The red dashed horizontal line in the right panel is the R-square of the natural bin analysis. **R-square increased from 0.42 to 0.55.**

Beef Cattle Data Analysis

- Trait = carcass weight
- Number of beef = 922
- Number of SNP markers = 40809
- Number of chromosomes = 29
- Methods = unweighted and weighted

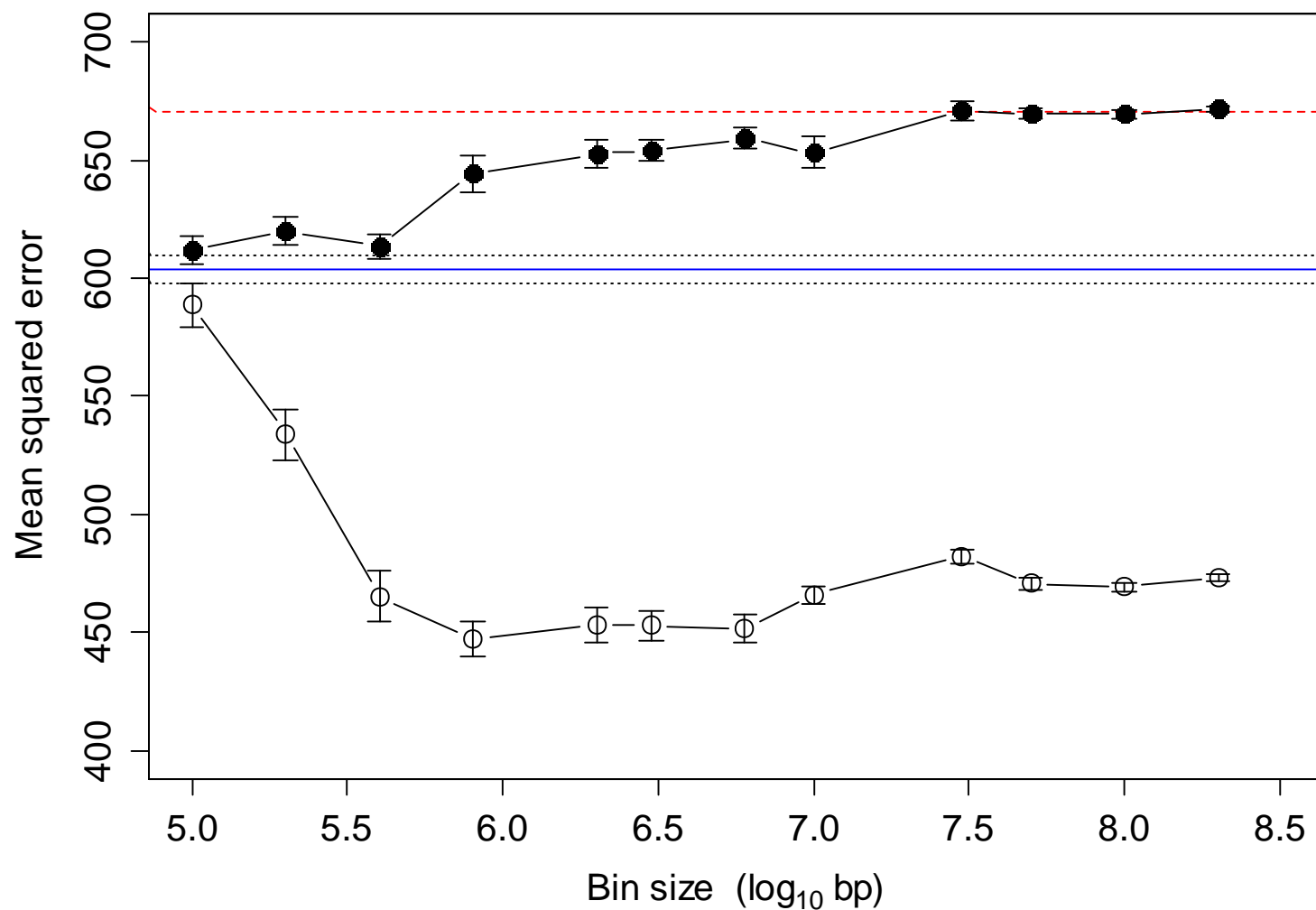
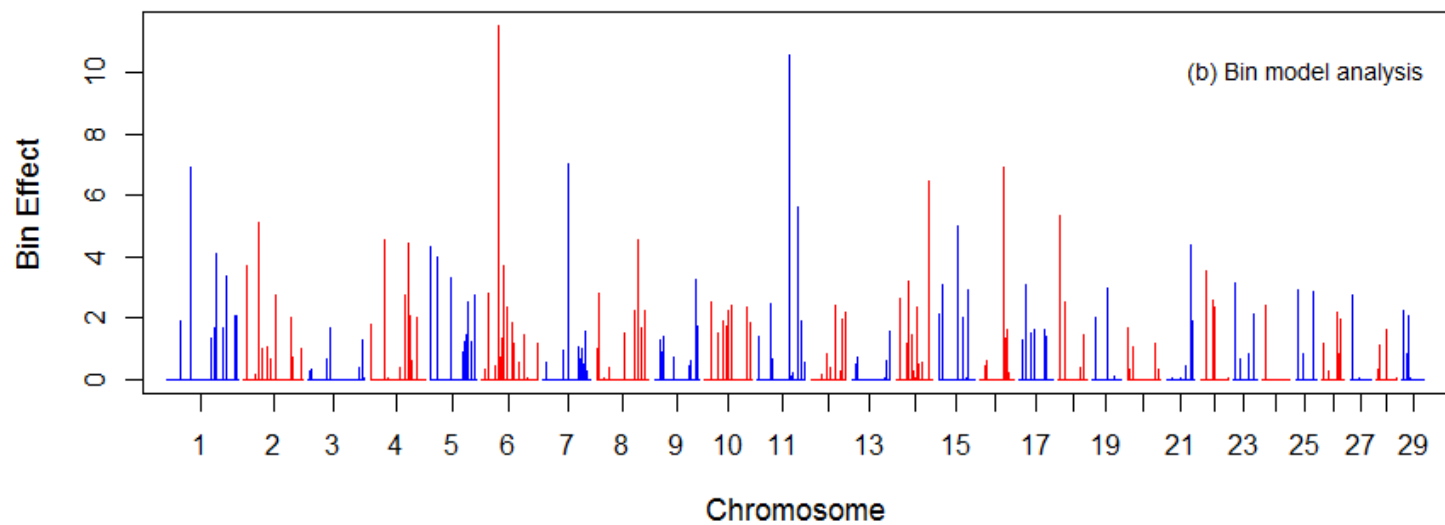
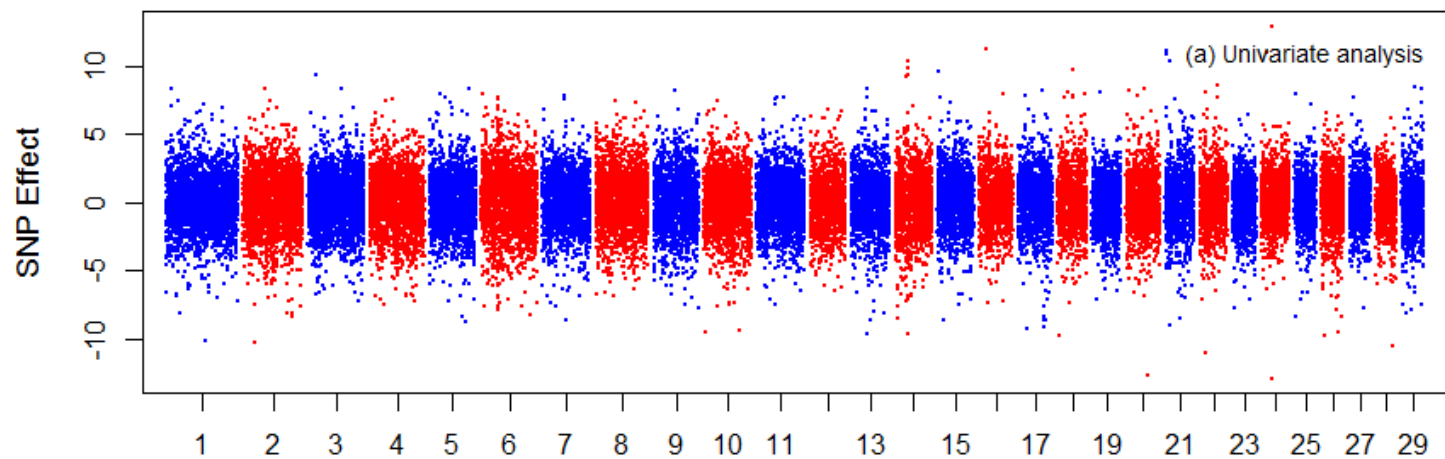
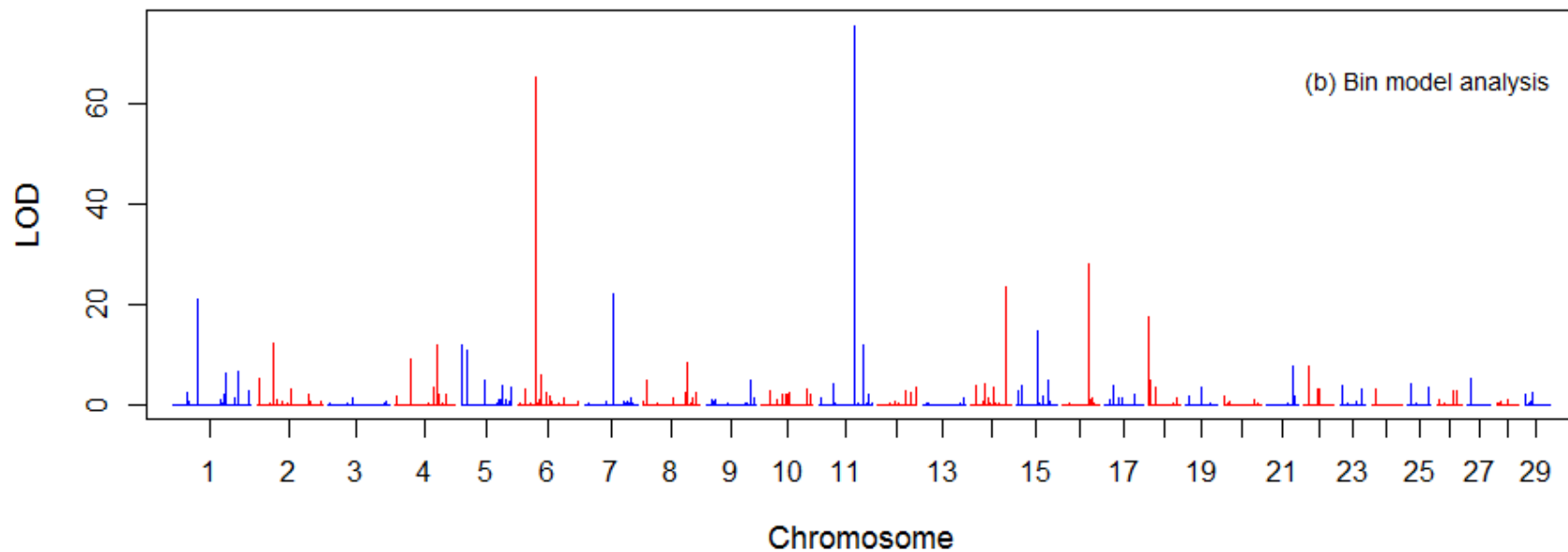
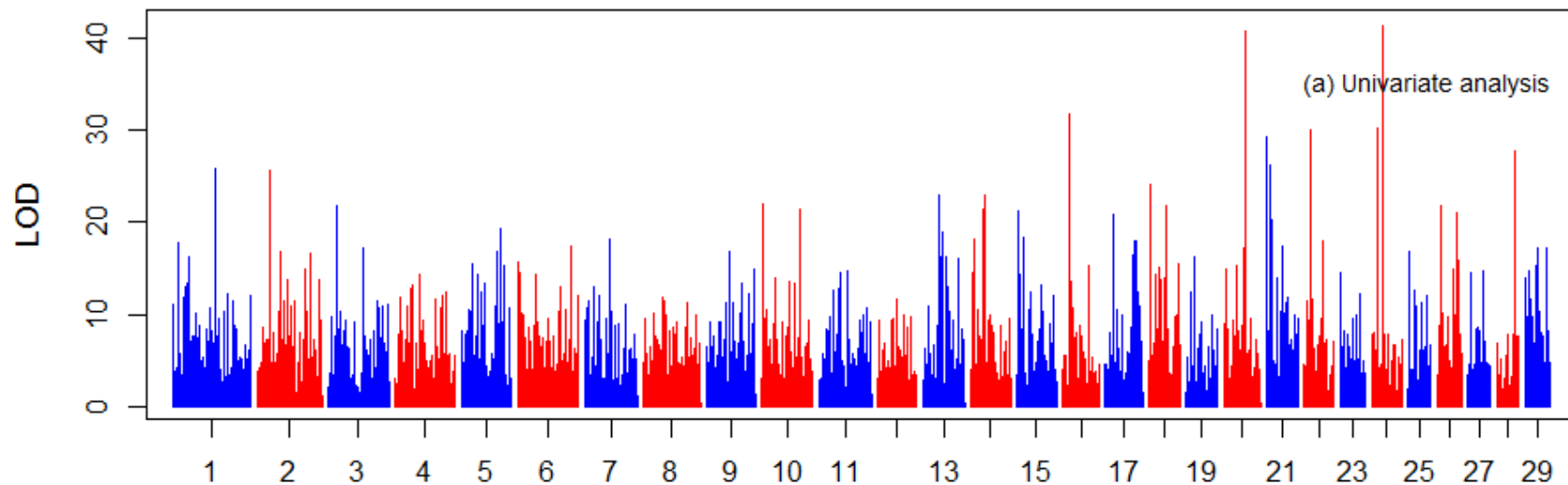


Figure 7. Mean squared error for the carcass trait of beef cattle plotted against the bin size. The filled circles indicate the MSE under the infinitesimal model while the open circles indicate the MSE under the adaptive infinitesimal model. The dashed horizontal line represents the phenotypic variance of the simulated trait (670.36). The blue horizontal line along with the two dotted lines represents the MSE and the standard deviation of the MSE in the situation where the bin size was one (one marker per bin). The sample size was $n = 921$ and the number of SNP markers was $p = 40809$. The bin size was defined as \log_{10} bp. For example, the largest bin size \log_{10} bp = 8.5 means that the bin size contains 8.5×10^5 base pairs.





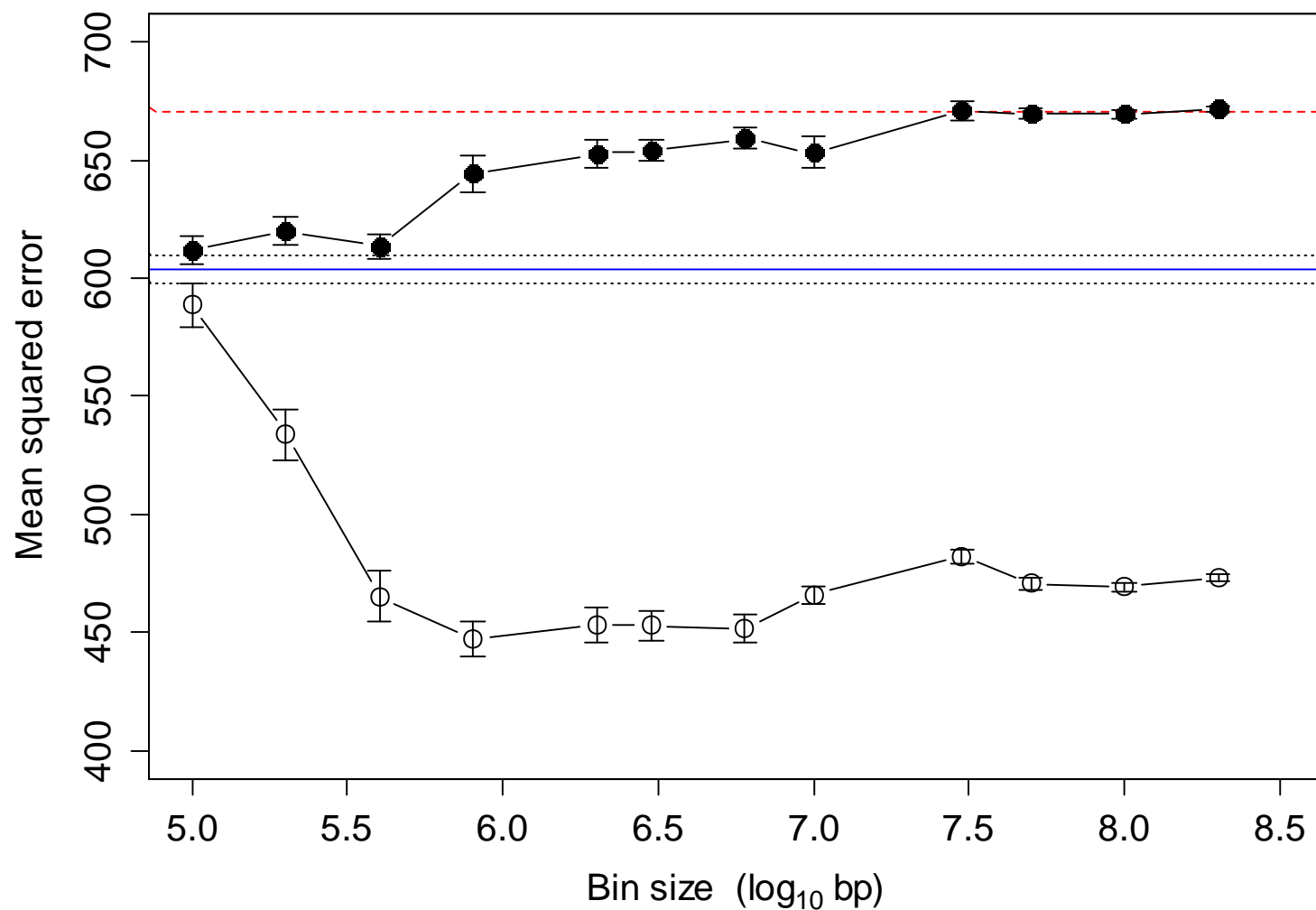
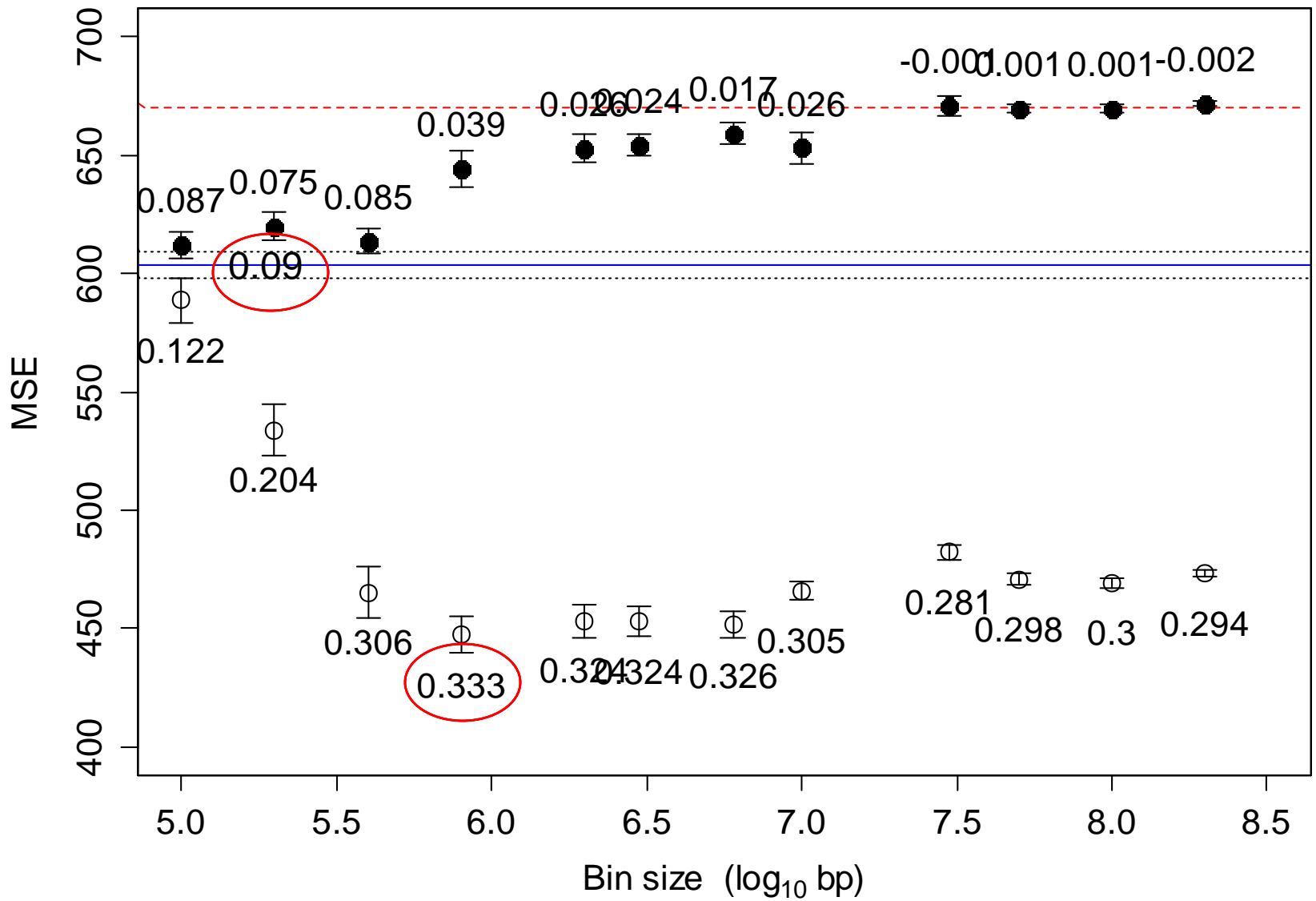


Figure 7. Mean squared error for the carcass trait of beef cattle plotted against the bin size. The filled circles indicate the MSE under the infinitesimal model while the open circles indicate the MSE under the adaptive infinitesimal model. The dashed horizontal line represents the phenotypic variance of the simulated trait (670.36). The blue horizontal line along with the two dotted lines represents the MSE and the standard deviation of the MSE in the situation where the bin size was one (one marker per bin). The sample size was $n = 921$ and the number of SNP markers was $p = 40809$. The bin size was defined as \log_{10} bp. For example, the largest bin size \log_{10} bp = 8.5 means that the bin size contains 8.5×10^5 base pairs.



Marker Analysis $p = 40809$
 MSE = 600
 $R^2 = (670-600)/670 = 0.09$

Table 1. Mean squared error (MSE) and R-square values obtained from the 10-fold cross validation analysis for the beef carcass trait using five competing models and the proposed bin model.

Model	MSE ²	R-square
eBayes	648.11	0.0332
G-Blup	632.46	0.0565
BayesB-1	655.59	0.0220
BayesB-2 ¹	658.19	0.0182
Lasso	603.75	0.0994
Bin model	447.10	0.3330

¹The Pi value for BayesB-2 is set at 0.95.

²The phenotypic variance of the beef carcass trait is 670.36. The magnitude of MSE value smaller than 670.36 indicates the effectiveness of the model predictability.

Outline

- Quantitative trait and the infinitesimal model
- Infinitesimal model using marker information
- Adaptive infinitesimal model
- Simulation studies
- Rice and beef cattle data analyses

Acknowledgements

- Zhiqiu Hu (postdoc)
- Qifa Zhang (rice data)
- Zhiqiun Wang (beef data)
- USDA Grant 2007-02784

Thank You !