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

\[ y_j = \beta + \int_0^L Z_j(\lambda) \gamma(\lambda) d\lambda + \epsilon_j \]

\[ y_j(t) = \beta + \phi(t) + \epsilon_j; \quad t \in \Omega \]
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

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

Breakpoint: \( Z_{jk} = \frac{1}{\Delta_k} \int_{0}^{\Delta_k} Z_j(\lambda) d\lambda \)
\[ Z_{jk} = \frac{1}{\Delta_k} \int_0^{\Delta_k} Z_j(\lambda) d\lambda \]

\[ Z_{jk} = 1 \times \frac{8}{10} + 0 \times \frac{2}{10} = 0.8 \]
What Does a Bin Effect Represent?

\[ y_j = \beta + \sum_{k=1}^{m} Z_{jk} \gamma_k + \epsilon_j \]

\[ Z_{jk} = \frac{1}{\Delta_k} \int_{0}^{\Delta_k} Z_j(\lambda) d\lambda \]

\[ \gamma_k = \Delta_k \times \frac{1}{\Delta_k} \int_{0}^{\Delta_k} \gamma(\lambda) d\lambda = \int_{0}^{\Delta_k} \gamma(\lambda) d\lambda \]

\[ \Delta_k = \text{size of bin } k \]

\[ \lambda = \text{uniform variable} \]
Assumptions of the Infinitesimal Model

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

\[ Z_{jk} = \frac{1}{\Delta_k} \int_{0}^{\Delta_k} Z_j(\lambda)d\lambda \]

\( \Delta_k \) = number of crossovers, inversely related to linkage disequilibrium

\[ \lim_{\Delta_k \to 0} \text{var}(Z_{jk}) = \frac{1}{2}, \text{ high linkage disequilibrium (F}_2\text{) } \]

\[ \lim_{\Delta_k \to \infty} \text{var}(Z_{jk}) = 0, \text{ low linkage disequilibrium } \]

Larger \( \text{var}(Z_{jk}) \) means higher power
Range of Var(Z)

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

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

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

\[\infty \geq \Delta_k \geq 0\]

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

\[ y_j = \beta + \sum_{k=1}^{m} Z_{jk} \gamma_k + \epsilon_j \]

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

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

\[ p_k = \text{number of markers in bin } k \]
Weighted Average Effect of a Bin

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

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

\( \gamma_j = \beta + \sum_{k=1}^{m} Z_{jk}^* \gamma_k^* + \epsilon_j \)
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 $\text{Var}(Z^*) > 0$

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

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

\[
MSY = \frac{1}{n} \sum_{j=1}^{n} (y_j - \overline{y}_j)^2 \quad \text{, Phenotypic Variance}
\]

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

Effect vs. Position (cM) plot showing true values.
Estimated Bin Effects

(a) $\Delta = 1\text{cM}$
   $m = 2400$
   $p = 50$

(b) $\Delta = 2\text{cM}$
   $m = 1200$
   $p = 100$

(c) $\Delta = 5\text{cM}$
   $m = 480$
   $p = 250$

(d) $\Delta = 10\text{cM}$
   $m = 240$
   $p = 500$

(e) $\Delta = 20\text{cM}$
   $m = 120$
   $p = 1000$

(f) $\Delta = 40\text{cM}$
   $m = 60$
   $p = 2000$

(g) $\Delta = 100\text{cM}$
   $m = 24$
   $p = 5000$

(h) $\Delta = 150\text{cM}$
   $m = 16$
   $p = 7500$

(i) $\Delta = 300\text{cM}$
   $m = 8$
   $p = 15000$

(j) $\Delta = 600\text{cM}$
   $m = 4$
   $p = 30000$

(k) $\Delta = 1200\text{cM}$
   $m = 2$
   $p = 60000$

(l) $\Delta = 2400\text{cM}$
   $m = 1$
   $p = 120000$
True and Estimated QTL Effect

0 500 1000 1500 2000 2500
-6 -4 -2 0 2 4 6 8
Parameter setting
ture values
Effect
Position (cM)

Estimate
Δ = 20cM
m = 120
p = 1000

0 500 1000 1500 2000 2500
-6 -4 -2 0 2 4 6 8
Effect
Position (cM)
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

\[ h^2 = \frac{64.41}{64.41 + 26.53 + \sigma^2} \].

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 $8.5 \times 10^5$ 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 $\log_{10} \text{bp}$. For example, the largest bin size $\log_{10} \text{bp} = 8.5$ means that the bin size contains $8.5 \times 10^5$ 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE(^2)</th>
<th>R-square</th>
</tr>
</thead>
<tbody>
<tr>
<td>eBayes</td>
<td>648.11</td>
<td>0.0332</td>
</tr>
<tr>
<td>G-Blup</td>
<td>632.46</td>
<td>0.0565</td>
</tr>
<tr>
<td>BayesB-1</td>
<td>655.59</td>
<td>0.0220</td>
</tr>
<tr>
<td>BayesB-2(^1)</td>
<td>658.19</td>
<td>0.0182</td>
</tr>
<tr>
<td>Lasso</td>
<td>603.75</td>
<td>0.0994</td>
</tr>
<tr>
<td>Bin model</td>
<td>447.10</td>
<td>0.3330</td>
</tr>
</tbody>
</table>

\(^1\)The Pi value for BayesB-2 is set at 0.95.

\(^2\)The 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!