

Marker Assisted Selection: Computational Issues

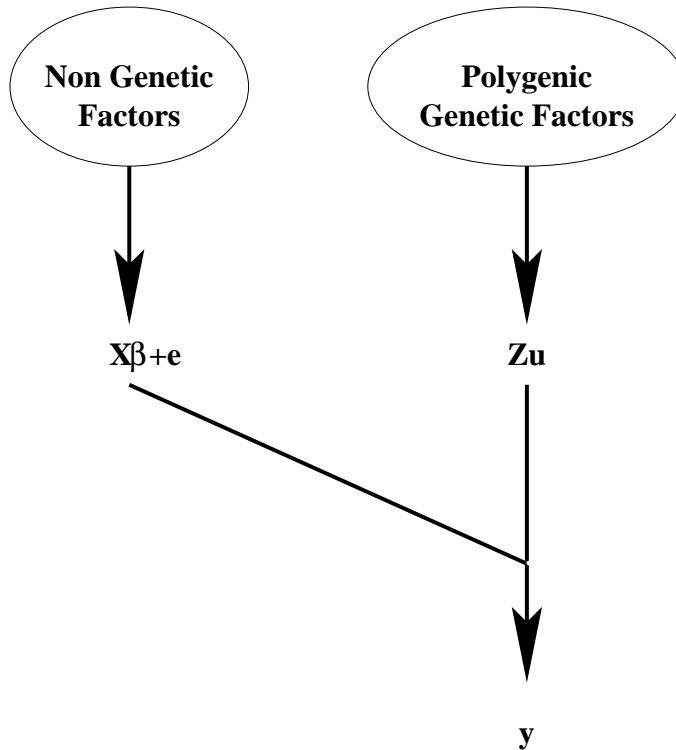
Stephen D. Kachman

January 16, 2001

Introduction

- Genetic improvement of quantitative traits through selection
- Polygenic model
 - Genes at many loci each with a small effect
 - Measure the cumulative effect
 - Analysis with mixed models
- Quantitative Trait Loci
 - Genes at a few loci with a large effect
 - Location is important
- Markers
 - Provide information on the flow of QTL alleles through the population
- Marker Assisted Selection
 - Incorporate marker information about QTL alleles in the prediction

Polygenic Model



$$y = X\beta + Zu + e$$

$$u \sim N(0, G)$$

$$e \sim N(0, R)$$

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

Covariance Matrices

- $\mathbf{R} = \mathbf{I}\sigma_e^2$
 - Environmental variance σ_e^2
 - Environmental covariances are zero
 - Number of observations
 - $\mathbf{R}^{-1} = \mathbf{I}\frac{1}{\sigma_e^2}$

- $\mathbf{G} = \mathbf{A}\sigma_u^2$
- Single loci A
 - additive genetic merit for animal i

$$u_i = u_i^1 + u_i^2$$

where u_i^j is the additive genotypic value of the j allele in animal i

- $\sigma_u^2 = \text{var}(u_i^1) + \text{var}(u_i^2) = 2 \text{var}(u_i^j)$

- G variances and covariances are generated by

Sire	Dam	
	A_{D1}	A_{D2}
A_{S1}	$A_{S1}A_{D1}$	$A_{S1}A_{D2}$
A_{S2}	$A_{S2}A_{D1}$	$A_{S2}A_{D2}$

$$\text{var}(u_i) = \frac{\text{var}(u_S^1 + u_D^1) + \text{var}(u_S^1 + u_D^2) + \dots}{4}$$

$$= \sigma_u^2$$

$$\text{cov}(u_i, u_s) = \frac{\text{cov}(u_S^1 + u_D^1, u_S^1 + u_S^2) + \dots}{4}$$

$$= \frac{\text{var}(u_S^1) + \dots}{4}$$

$$= \frac{\sigma_u^2}{2}$$

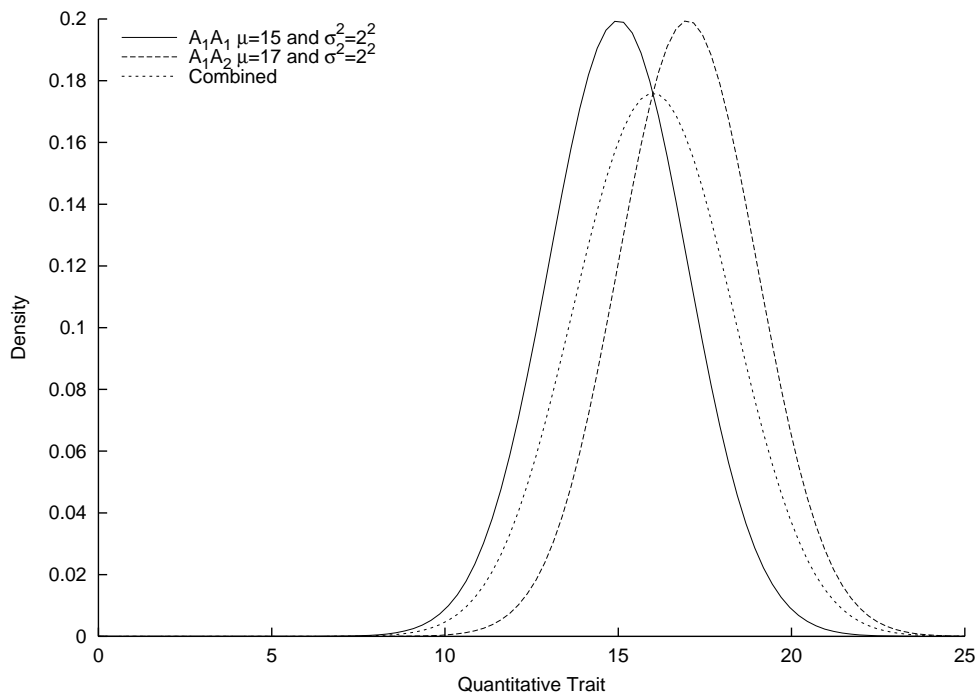
- Number of animals being evaluated

Computational Issues

- Consider 100,000 animals
 - One observation per animal then
 - * R has $100,000^2$ elements
 - * Storage 74 gigabytes
 - * Only 100,000 non-zero elements in R^{-1}
 - * Storage 781 kilobytes
 - * A also has $100,000^2$ elements
 - * A^{-1} is also mostly zeros
- Holstein uses records on approximately a million cows
 - Full storage would be in the terabytes
 - To solve a million equations by brute force in one day would involve a computer with teraflop computing capability
- Computational feasibility depends heavily on the sparseness of the matrices involved.

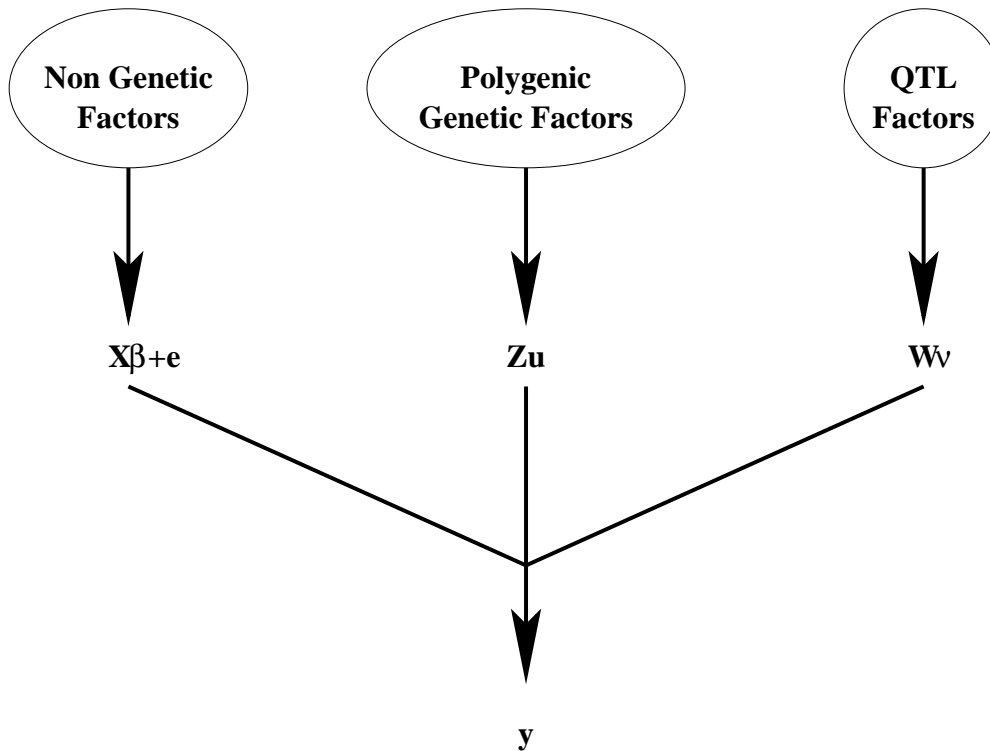
Quantitative Trait Loci

- A locus which deserves special attention
 - Difference in an allele at a QTL will have a “significant” impact on progeny performance



- Can still have many loci where the effect is small but the total effect is significant
- Many alleles at the QTL

Additive QTL effects



- Sample of alleles at a QTL
- ν_i be the additive effect of QTL allele i
- $\text{var}(\nu_i) = \sigma_\nu^2$ and $\text{var}(\boldsymbol{\nu}) = \boldsymbol{\Lambda}\sigma_\nu^2$

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i + \nu_i^p + \nu_i^m + e_i$$

$$a_i = u_i + \nu_i^p + \nu_i^m$$

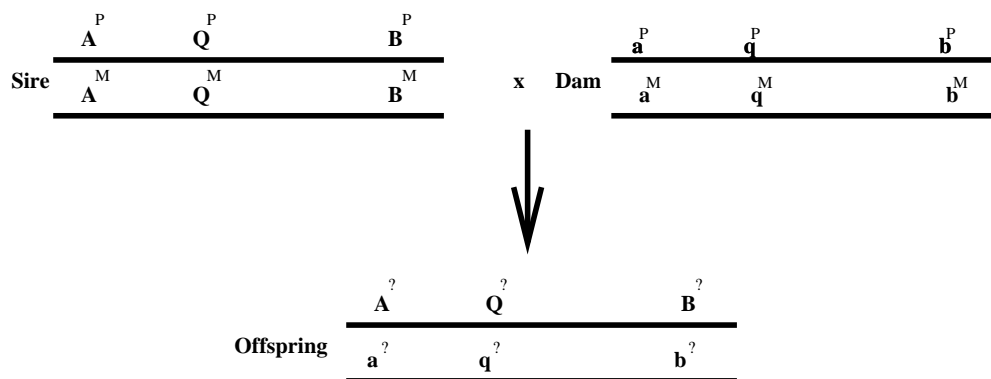
- Which is a mixed model

Estimation

- If we knew which QTL allele is passed on
 - There would 2×number of foundation animals QTL effects to be estimated

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z & X'R^{-1}W \\ Z'R^{-1}X & Z'R^{-1}Z + A^{-1}\sigma_u^{-2} & Z'R^{-1}W \\ X'R^{-1}X & W'R^{-1}Z & W'R^{-1}W + I\sigma_v^{-2} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} X'R^{-1}y \\ Z'R^{-1}y \\ W'R^{-1}y \end{pmatrix}$$

- However, we don't know which one is passed on
- Markers provide partial information



Variance of QTL effects given marker information

- Consider full-sibs 101 and 102

$$\frac{A_{101}^? Q_{101}^?}{a_{101}^? q_{101}^?} \text{ and } \frac{A_{102}^? Q_{102}^?}{a_{102}^? q_{102}^?}$$

- The Covariance between the paternal QTL effects ν_{101}^P and ν_{102}^P is

$$\begin{aligned} \text{cov}(\nu_{101}^P, \nu_{102}^P) &= \Pr(Q_{101}^P \text{ and } Q_{102}^P) \sigma_\nu^2 \\ &\quad + \Pr(Q_{101}^M \text{ and } Q_{102}^M) \sigma_\nu^2 \\ &= \frac{1}{2} \sigma_\nu^2 \end{aligned}$$

- Which is proportional to $\text{cov}(a_{101}, a_{102})$.

- Now if we add that both full sibs received A^P , then

$$\begin{aligned}
\text{cov}(\nu_{101}^P, \nu_{102}^P | A_{101}^? = A_{102}^? = A^P) \\
&= \Pr(Q_{101}^P \text{ and } Q_{102}^P | \dots) \sigma_\nu^2 \\
&\quad + \Pr(Q_{101}^M \text{ and } Q_{102}^M | \dots) \sigma_\nu^2 \\
&= [(1 - \theta_{AQ})^2 + \theta_{AQ}^2] \sigma_\nu^2
\end{aligned}$$

where θ_{AQ} is the recombination fraction between loci A and Q.

- Which is a bit of a simplification because we ignored the possibility the $Q^P \equiv Q^M$.

$$\Pr(Q^P \equiv Q^M | M) = \text{cov}(\nu_S^P, \nu_S^M) \frac{1}{\sigma_\nu^2}$$

which yields

$$\begin{aligned}
\text{cov}(\nu_{101}^P, \nu_{102}^P | A_{101}^? = A_{102}^? = A^P) \\
&= \left[(1 - \theta_{AQ})^2 + \theta_{AQ}^2 \right. \\
&\quad \left. + 2\theta_{AQ}(1 - \theta_{AQ}) \Pr(Q^P \equiv Q^M | M) \right] \sigma_\nu^2 \\
&= \Pr(Q_{101}^? \equiv Q_{102}^? | M) \sigma_\nu^2
\end{aligned}$$

Building $\text{var}(\boldsymbol{\nu})$

- Built iteratively, similar to the \mathbf{A}

$$\text{var}(\boldsymbol{\nu}) = \boldsymbol{\Lambda} \sigma_{\nu}^2$$

$$\Lambda_{ikjl} = \Pr(Q_i^k \equiv Q_j^l | M)$$

- Each individual has two QTL effects
 - $\boldsymbol{\Lambda}$ is $2n \times 2n$
 - Each QTL is equivalent to adding 2 more traits to the analysis
- Start with oldest animals

$$\boldsymbol{\Lambda}_i = \begin{pmatrix} \boldsymbol{\Lambda}_{i-1} & \boldsymbol{\Lambda}_{i-1} \mathbf{q}_i \\ \mathbf{q}_i' \boldsymbol{\Lambda}_{i-1} & \mathbf{C}_i \end{pmatrix}$$

where

$$\mathbf{C}_i = \begin{pmatrix} 1 & \Pr(Q_i^1 \equiv Q_i^2 | M) \\ \Pr(Q_i^2 \equiv Q_i^1 | M) & 1 \end{pmatrix}$$

$$\mathbf{q}_i' = \begin{pmatrix} 0 & \Pr(Q_i^1 \leftarrow Q_s^1 | M) & \Pr(Q_i^1 \leftarrow Q_s^2 | M) & 0 & \dots \\ 0 & \Pr(Q_i^2 \leftarrow Q_s^1 | M) & \Pr(Q_i^2 \leftarrow Q_s^2 | M) & 0 & \dots \end{pmatrix}$$

- \mathbf{q}_i only has a few non-zero elements
- However we need $\boldsymbol{\Lambda}^{-1}$

Building $\text{var}(\nu)^{-1}$

- Built iteratively using the following result for portioned matrices

$$\Lambda_i^{-1} = \begin{pmatrix} \Lambda_{i-1}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{q}_i \mathbf{D}_i^{-1} \mathbf{q}_i' & -\mathbf{q}_i \mathbf{D}_i^{-1} \\ -\mathbf{q}_i' \mathbf{D}_i^{-1} & \mathbf{D}_i^{-1} \end{pmatrix}$$

where

$$\mathbf{D}_i = \mathbf{C}_i - \mathbf{q}_i' \Lambda_{i-1} \mathbf{q}_i$$

- The ease of computing this depends heavily on the completeness of the marker data.
- Without inbreeding,

$$\mathbf{q}_i' \Lambda_{i-1} \mathbf{q}_i = \mathbf{q}_i' \mathbf{q}_i$$

Summary

- Mixed model
 - Challenge is in computing Λ^{-1} .
- Turns a single trait model into 3 trait model
 - Full storage grows by a factor of 9
 - Direct solution grows by a factor 27
- Each additional QTL is equivalent to adding two more traits
- Trade off between the number of animals that can be included in the evaluation and the number of QTL that are included
- Assumes that marker data is 100% correct