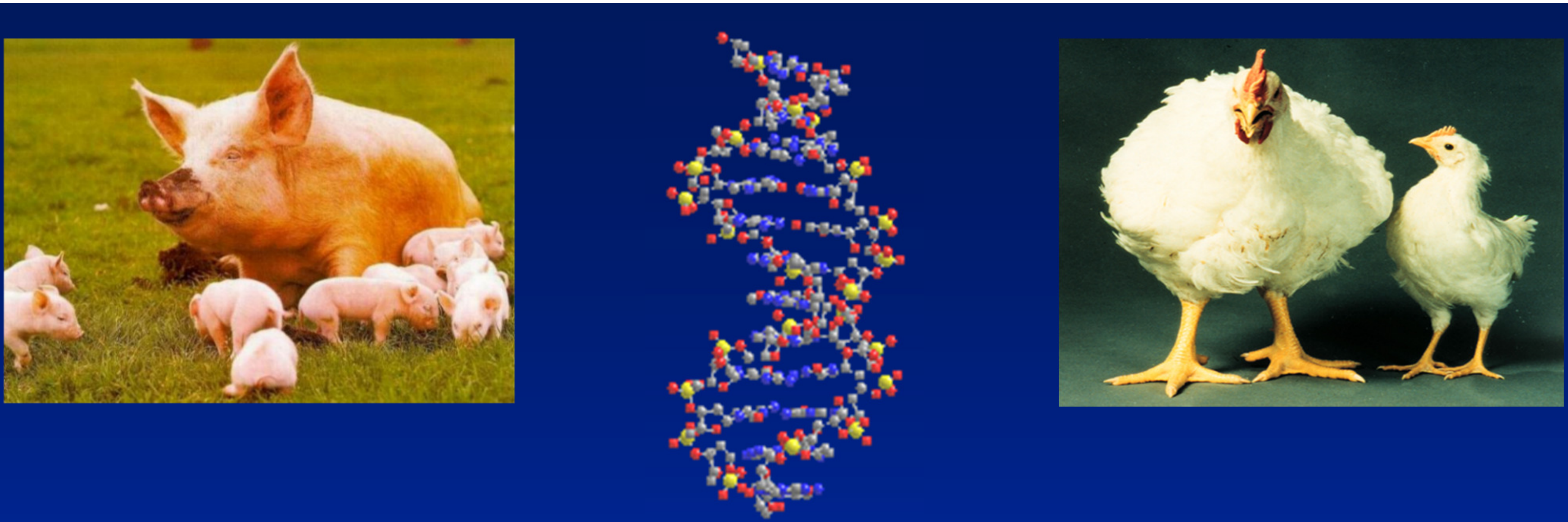
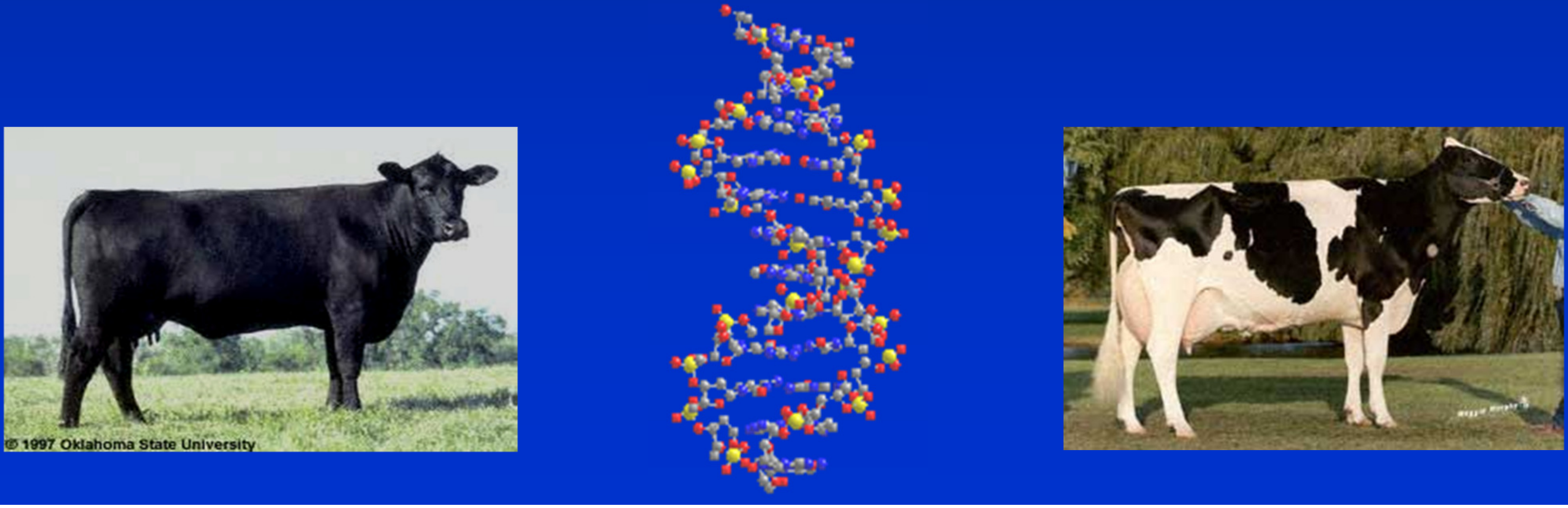


Genomic Selection in the era of Genome sequencing



Course overview

- Day 1
  - Linkage disequilibrium in animal and plant genomes
- Day 2
  - Genome wide association studies
- Day 3
  - Genomic selection
- Day 4
  - Genomic selection
- Day 5
  - Imputation and whole genome sequencing for genomic selection

Genomic selection

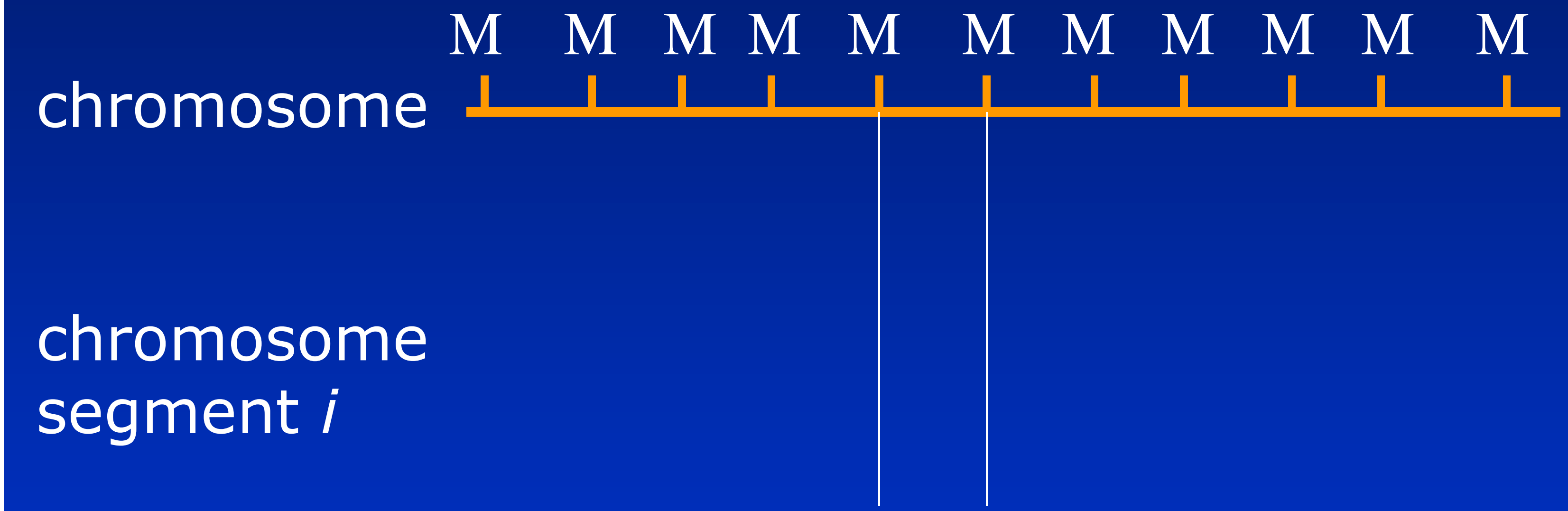
- Introduction
- Genomic selection with Least Squares and BLUP
- Introduction to Bayesian methods
- Genomic selection with Bayesian methods
- Comparison of accuracy of methods

Genomic selection

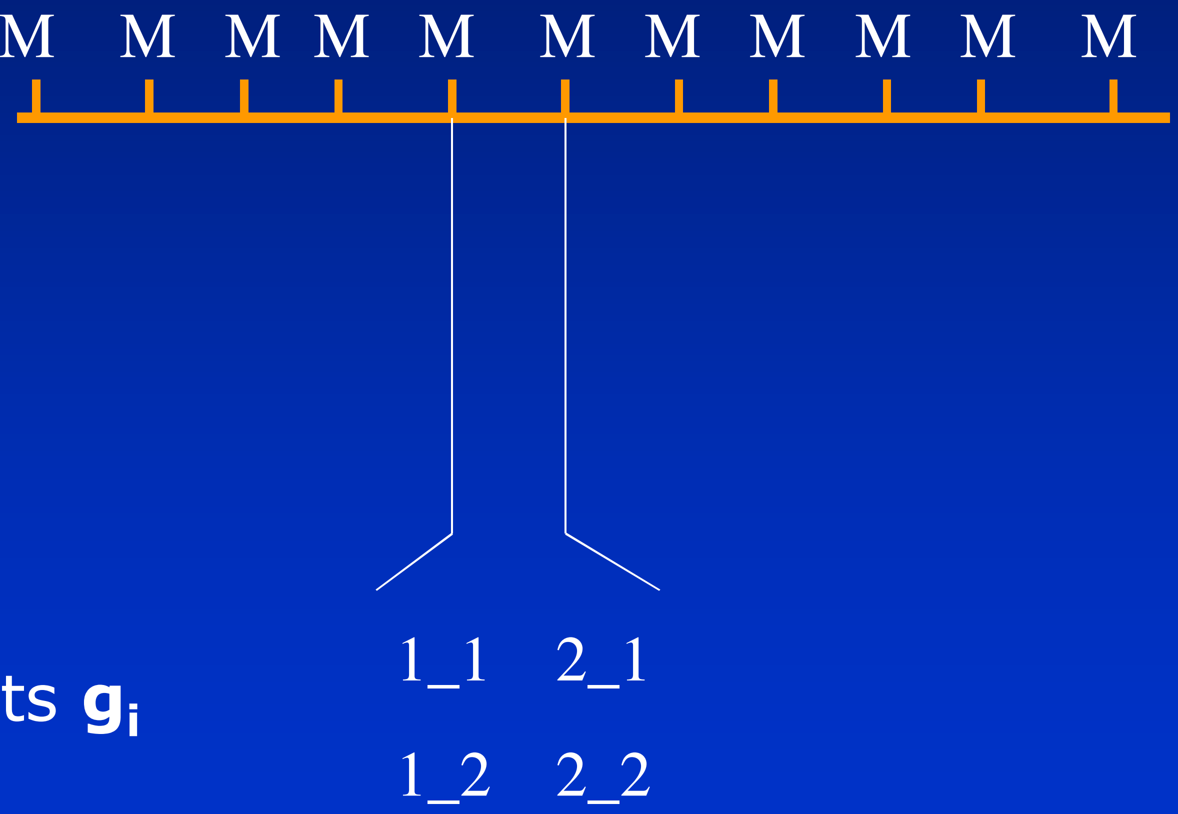
- Problem marker assisted selection is only a proportion of genetic variance is tracked with markers
  - Eg. 10 QTL << 5% of the genetic variance
- Alternative is to trace all segments of the genome with markers
  - Divide genome into chromosome segments based on marker intervals?
  - Capture all QTL = all genetic variance

Genomic selection

Genomic selection



Genomic selection



chromosome

chromosome
segment *i*

chromosome
segment effects $\mathbf{g}_\mathbf{i}$

Genomic selection

- Predict genomic breeding values as sum of effects over *all* segments

$$\boxed{\mathbf{GEBV} = \sum_{i}^{p} \mathbf{X}_i \hat{\mathbf{g}}_i}$$

Genomic selection

- Predict genomic breeding values as sum of effects over *all* segments

$$\mathbf{GEBV} = \sum_{i}^{p} \mathbf{X}_i \hat{\mathbf{g}}_i$$

Number of chromosome segments

Genomic selection

- Genomic selection can be implemented
  - with marker haplotypes within chromosome segments

$$\mathbf{GEBV} = \sum_{i}^{p} \mathbf{X}_i \hat{\mathbf{g}}_i$$

1_1	0.3
1_2	0.0
2_1	-0.2
2_2	-0.1

Genomic selection

- Genomic selection can be implemented
  - with marker haplotypes within chromosome segments
  - with single markers

$$\mathbf{GEBV} = \sum_{i}^{p} \mathbf{X}_i \, \hat{\mathbf{g}}_i$$

$$\mathbf{GEBV} = \sum_{i}^{p} \mathbf{X}_i \, \hat{g}_i$$

2 -0.5

Genomic selection

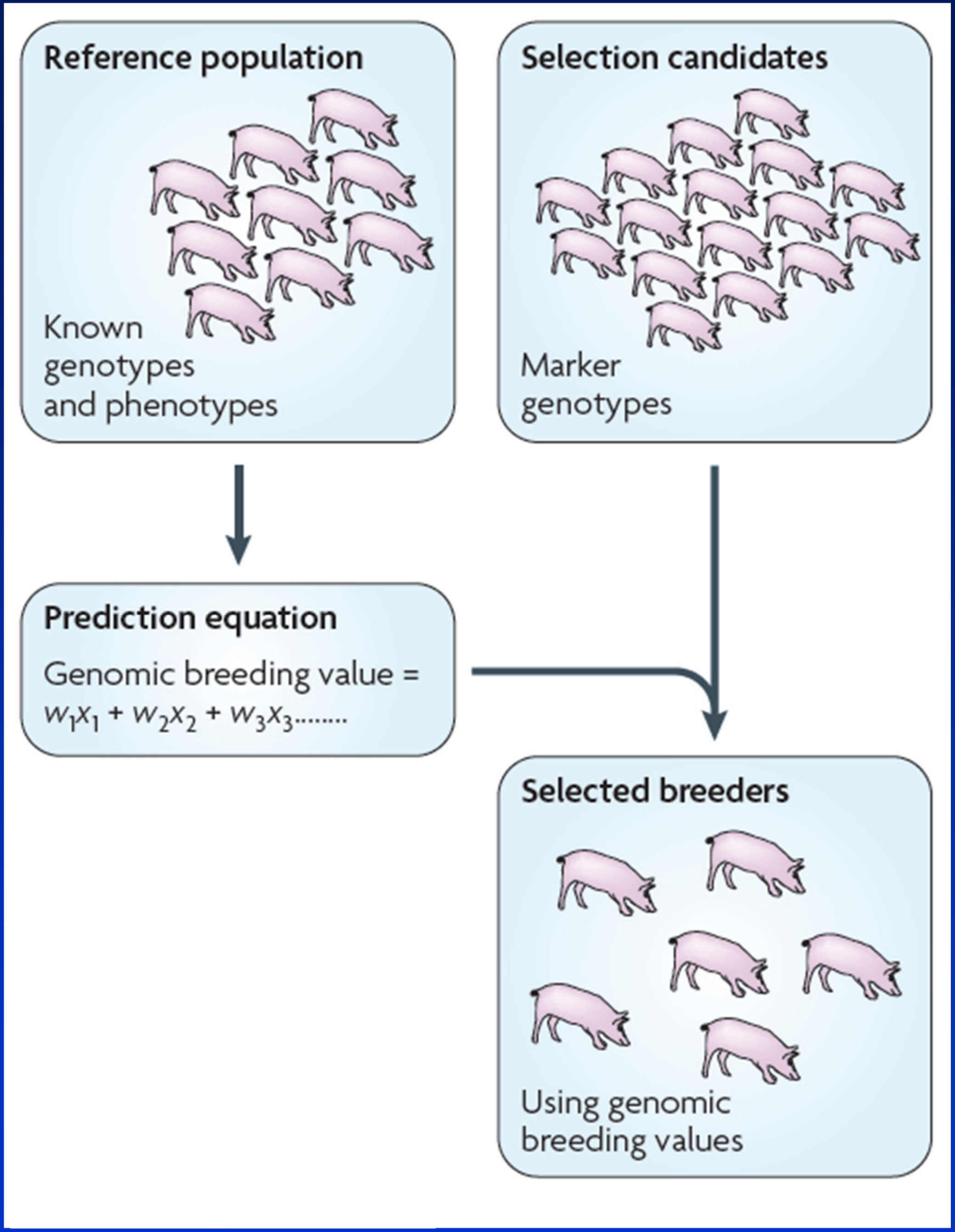
- Genomic selection exploits linkage disequilibrium
  - Assumption is that effect of haplotypes or markers within chromosome segments picking up QTL and will have same effect across the whole population
- Possible within dense marker maps now available

1_1	0.3
1_2	0.0
2_1	-0.2
2_2	-0.1

Genomic selection

- Genomic selection avoids bias in estimation of effects due to multiple testing, as all effects fitted simultaneously

Genomic selection



Genomic selection

- First step is to predict the chromosome segment effects in a reference population
- Number of effects >>> than number of records
- Eg. 10 000 intervals * 4 haplotypes = 40 000 haplotype effects
- From ~ 2000 records?
- Need methods that can deal with this

Genomic selection

- Introduction
- Genomic selection with Least Squares and BLUP
- Introduction to Bayesian methods
- Genomic selection with Bayesian methods
- Comparison of accuracy of methods

Least squares Genomic selection

- Two step procedure
  - Test each chromosome segment for presence of QTL (fitting haplotypes within segment), take significant effects
  - Fit the significant effects simultaneously in multiple regression
  - Predict GEBVs
- Identical to Marker assisted selection with multiple markers
- Problems remain
  - Do not capture all QTL
  - Over-estimation of haplotype effects due to setting of significance threshold

Genomic selection with BLUP

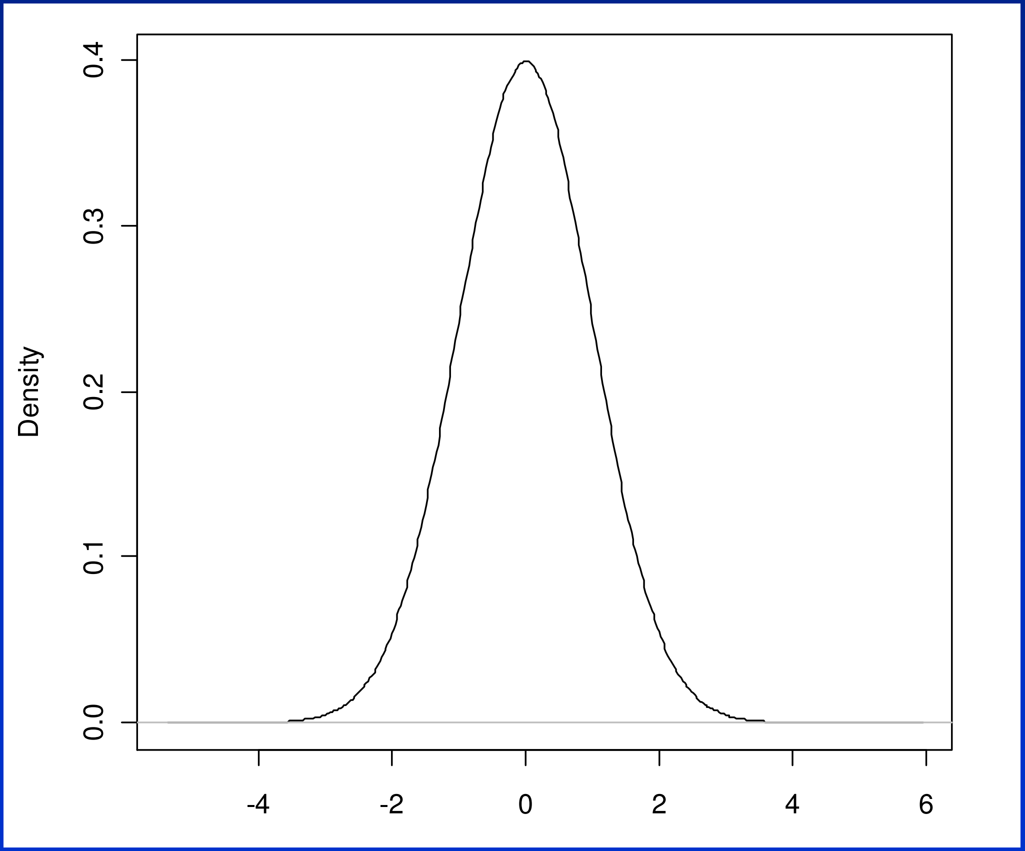
- BLUP = best linear unbiased prediction
- Model:

$$\mathbf{y} = \mu \mathbf{1}_{\mathbf{n}} + \sum_{i=1}^{p} \mathbf{X}_{\mathbf{i}} \mathbf{g}_{\mathbf{i}} + \mathbf{e}$$

- In BLUP we assume variance of haplotype effects across all segments is equal, eg E(**g**) ~ $N(0,\sigma_g^2)$, where g = $[g_1 g_2 g_3 .. g_p]$

Genomic selection with BLUP

- BLUP assumes normal distribution of SNP/haplotype effects



Genomic selection with BLUP

- **BLUP** = best linear unbiased prediction
- Then we can estimate segment effects as:

$$\boxed{\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X+I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}}$$

- $\lambda = \sigma_e^2 / \sigma_g^2$

Genomic selection with BLUP

- Example
- A “simulated” data set
- Single chromosome, with 10 markers
- Phenotypes “simulated”
  - overall mean of 1
  - an effect for SNP 1 of 2 allele of 1
  - normally distributed error term with mean 0 and variance 1.

Genomic selection with BLUP

- Example

		X										
Animal	Y		1	2	3	4	5	6	7	8	9	10
1	0.19		0	0	0	0	0	0	1	2	0	2
2	1.23		1	0	0	1	1	1	2	1	0	1
3	0.86		1	0	0	1	0	0	1	1	1	1
4	1.23		1	1	1	1	0	1	2	1	1	1
5	0.45		0	1	1	1	1	1	2	1	0	1

- 10 SNPs
- Only 5 phenotypic records.

Genomic selection with BLUP

- Example

		X									
Animal	Y	1	2	3	4	5	6	7	8	9	10
1	0.19	0	0	0	0	0	0	1	2	0	2
2	1.23	1	0	0	1	1	1	2	1	0	1
3	0.86	1	0	0	1	0	0	1	1	1	1
4	1.23	1	1	1	1	0	1	2	1	1	1
5	0.45	0	1	1	1	1	1	2	1	0	1

- Assume value of 1 for λ
- $1_n = [1\ 1\ 1\ 1\ 1]$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n}'\mathbf{X} \\ \mathbf{X}'\mathbf{1_n} & \mathbf{X}'\mathbf{X}+\mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}$$

Genomic selection with BLUP

- Example

Mean	0.47
SNP1	0.29
SNP2	-0.05
SNP3	-0.05
SNP4	0.08
SNP5	-0.02
SNP6	0.13
SNP7	0.13
SNP8	-0.08
SNP9	0.11
SNP10	-0.08

Genomic selection with BLUP

- Now we want to predict GEBV for a group of young animals without phenotypes.

$$\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

- We have the g_hat, and we can get **X** from their haplotypes (after genotyping)…………

Progeny	X								
1	1	1	1	1	1	1	2	1	0	1
2	1	0	0	1	1	1	2	1	0	1
3	1	0	0	1	1	1	2	1	0	1
4	1	0	0	1	1	1	2	1	0	1
5	0	0	0	0	0	0	1	2	0	2

Genomic selection with BLUP

- GEBV

$$\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

X										$\hat{\mathbf{g}}$		GEBV
1	1	1	1	1	1	2	1	0	1	0.29		0.47
1	0	0	1	1	1	2	1	0	1	-0.05		0.58
1	0	0	1	1	1	2	1	0	1	-0.05		0.58
1	0	0	1	1	1	2	1	0	1	0.08		0.58
0	0	0	0	0	0	1	2	0	2	-0.02		-0.20
										0.13		
										0.13		
										-0.08		
										0.11		
										-0.08		

Genomic selection with BLUP

- Where do we get σ_g^2 from?
- Can estimate total additive genetic variance and divide by number of segments, eg $\sigma_g^2 = \sigma_a^2 / p$
- If using single markers take account of heterozygosity

$$\sigma_g^2 = \sigma_a^2 / 2\sum_{i=1}^{p} q_i(1-q_i)$$

- Ridge regression (Bayesian approach)
- Cross validation

Genomic selection with BLUP

- An equivalent model
- If there are many QTLs whose effects are normally distributed with constant variance,
- Then genomic selection equivalent to replacing the expected relationship matrix with the realised or genomic relationship matrix (**G**) estimated from DNA markers in normal BLUP equations.
  - G_{ij} = proportion of genome that is IBD between animals i and j

Genomic selection with BLUP

- An equivalent model
- Rescale X to account for allele frequencies
  - $w_{ij} = x_{ij} - 2p_j$
- Then breeding values are
  - $\mathbf{v} = \mathbf{W}\mathbf{g}$ ($\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$)
- And

$$\mathbf{G} = \mathbf{W}\mathbf{W}'/2\sum_{j=1}^{p} p_j(1-p_j)$$

- Then

$$V(\mathbf{v}) = \mathbf{G}\sigma_a^2$$

Genomic selection with BLUP

- An equivalent model

$$\mathbf{y} = \mu\mathbf{1}_{\mathrm{n}} + \mathbf{Zv} + \mathbf{e}$$

$$\begin{bmatrix} \overset{\wedge}{\mu} \\ \overset{\wedge}{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{\mathrm{n}}'\mathbf{1}_{\mathrm{n}} & \mathbf{1}_{\mathrm{v}}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1}_{\mathrm{n}} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_{\mathrm{n}}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Genomic selection with BLUP

- An equivalent model
  - Model 1.

$$\mathbf{y} = \mu \mathbf{1}_\mathrm{n} + \sum_{i=1}^{p} \mathbf{X}_\mathrm{i} \mathbf{g}_\mathrm{i} + \mathbf{e}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_\mathrm{n}'\mathbf{1}_\mathrm{n} & \mathbf{1}_\mathrm{n}'\mathbf{X} \\ \mathbf{X}'\mathbf{1}_\mathrm{n} & \mathbf{X}'\mathbf{X} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_\mathrm{n}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}$$

$$\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

  - Model 2.

Genomic selection with BLUP

- An equivalent model
  - Model 1.

$$\mathbf{y} = \mu \mathbf{1}_{\mathrm{n}} + \sum_{i=1}^{p} \mathbf{X}_{\mathrm{i}} \mathbf{g}_{\mathrm{i}} + \mathbf{e}$$

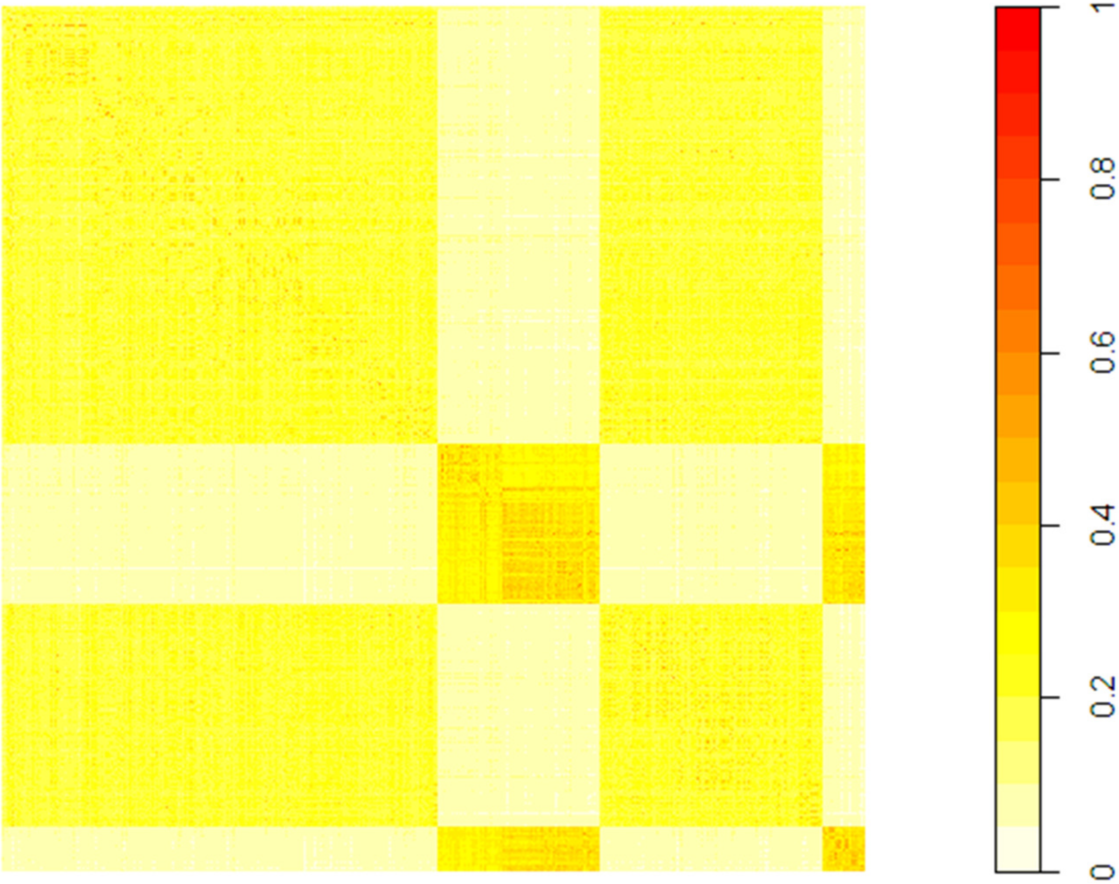
$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{\mathrm{n}}'\mathbf{1}_{\mathrm{n}} & \mathbf{1}_{\mathrm{n}}'\mathbf{X} \\ \mathbf{X}'\mathbf{1}_{\mathrm{n}} & \mathbf{X}'\mathbf{X} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_{\mathrm{n}}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}$$

$$\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

  - Model 2.

$$\mathbf{y} = \mu \mathbf{1}_{\mathrm{n}} + \mathbf{Z}\mathbf{v} + \mathbf{e}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{\mathrm{n}}'\mathbf{1}_{\mathrm{n}} & \mathbf{1}_{\mathrm{v}}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1}_{\mathrm{n}} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\dfrac{\sigma_e^2}{\sigma_v^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_{\mathrm{n}}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$



Holstein reference *n = 781*

Jersey reference *n = 287*

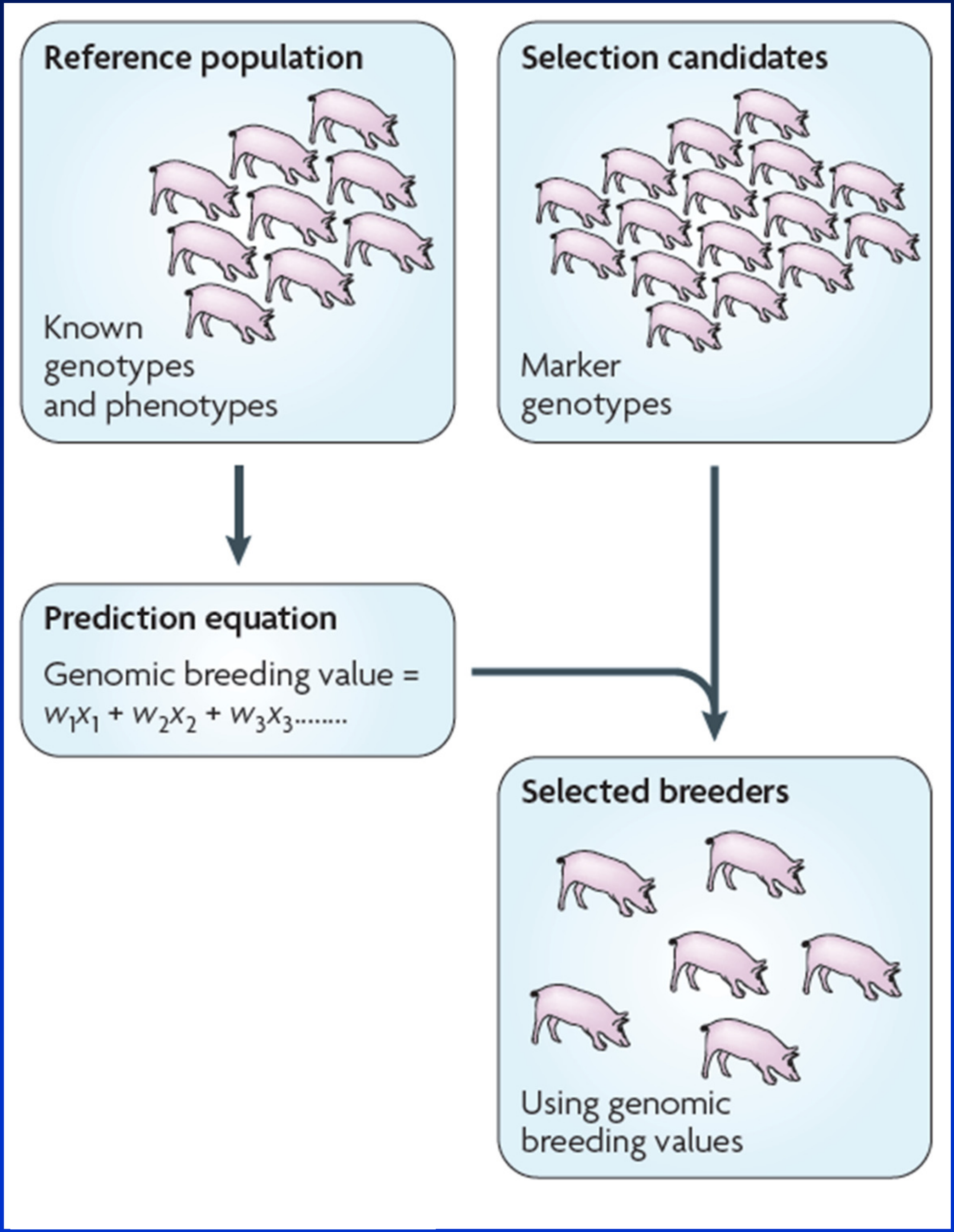
Holstein validation *n = 400*

Jersey validation *n = 77*

Genomic selection with BLUP

- An equivalent model
- Why use model 2.
  - If number of markers >>> large than number of animals, more computationally efficient
  - Can be integrated into national evaluations more readily?
  - Calculate accuracy of GEBV from inverse coefficient matrix

Genomic selection



Reference population

Known genotypes and phenotypes

Selection candidates

Marker genotypes

Prediction equation

Genomic breeding value = $w_1x_1 + w_2x_2 + w_3x_3$........

Selected breeders

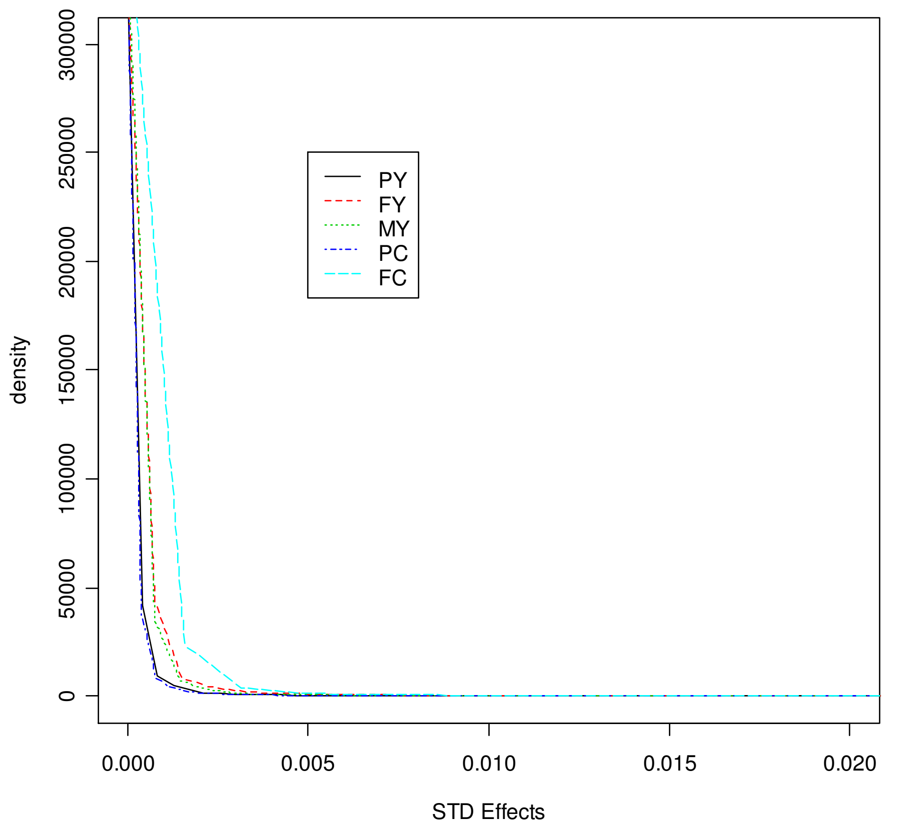
Using genomic breeding values

Genomic selection

- Introduction
- Genomic selection with Least Squares and BLUP
- Introduction to Bayesian methods
- Genomic selection with Bayesian methods
- Comparison of accuracy of methods

Bayesian methods

- BLUP assumes normally distributed QTL effects
- Does not match prior knowledge of distributions of QTL effects for some traits
- Use Bayesian approaches to incorporate prior knowledge



Bayesian methods

- Bayes theorem

$$\boxed{P(x \mid y) \propto P(y \mid x)P(x)}$$

Bayesian methods

- Bayes theorem

$$P(x \mid y) \propto P(y \mid x) P(x)$$

Probability of
parameters x given
the data y (posterior)

Bayesian methods

- Bayes theorem

$$P(x \mid y) \propto P(y \mid x)P(x)$$

Probability of parameters x given the data y (posterior)

Is proportional to

Bayesian methods

- Bayes theorem

$$P(x \mid y) \propto P(y \mid x)P(x)$$

Probability of parameters x given the data y (posterior)

Is proportional to

Probability of data y given the x (likelihood of data)

Bayesian methods

- Bayes theorem

$$P(x \mid y) \propto P(y \mid x) P(x)$$

Probability of parameters x given the data y (posterior)

Is proportional to

Probability of data y given the x (likelihood of data)

Prior probability of x

Bayesian methods

- Consider an experiment where we measure height of 10 people to estimate average height
- We want to use prior knowledge from many previous studies that average height is 174cm with standard error 5cm

y=average height + e

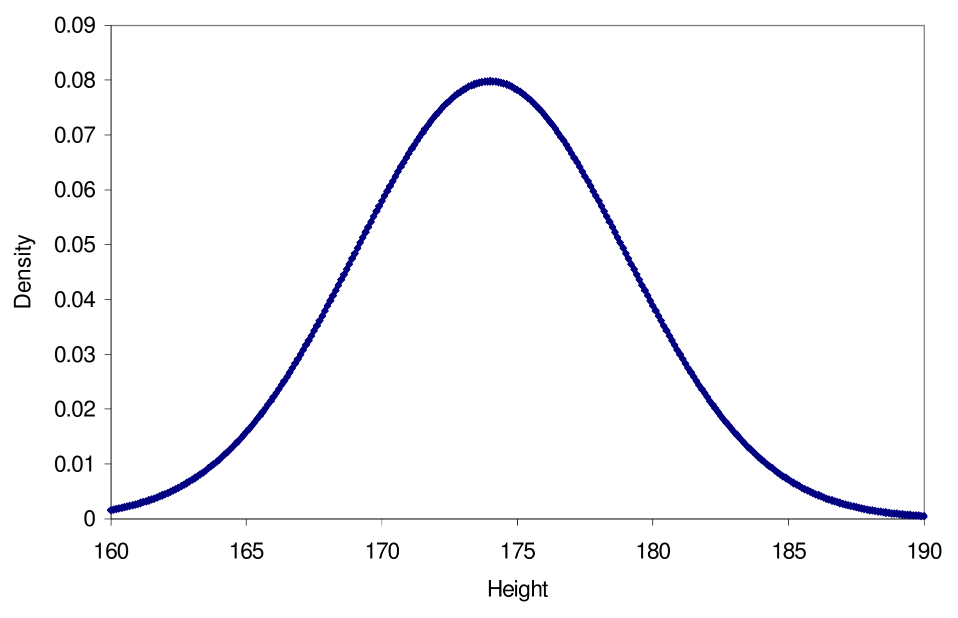
Bayesian methods

- Bayes theorem

$$P(x \mid y) \propto P(y \mid x)P(x)$$

↑

Prior probability of x (average height)



Bayesian methods

- Bayes theorem

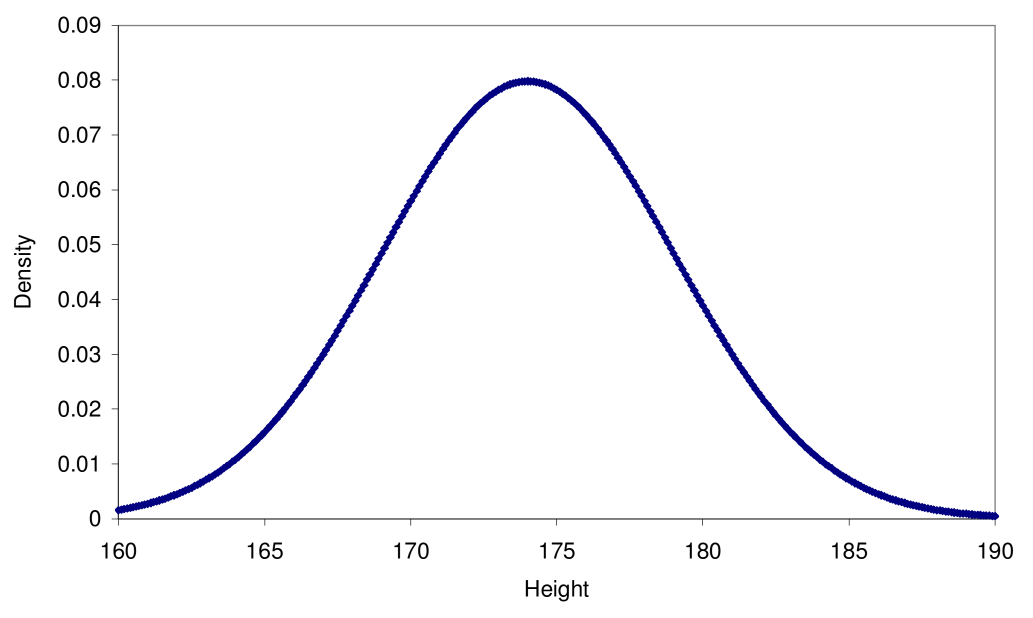
$$P(x \mid y) \propto P(y \mid x)P(x)$$

From the data……

$$\bar{x} = 178$$

$$s.e = 5$$

Prior probability of x (average height)



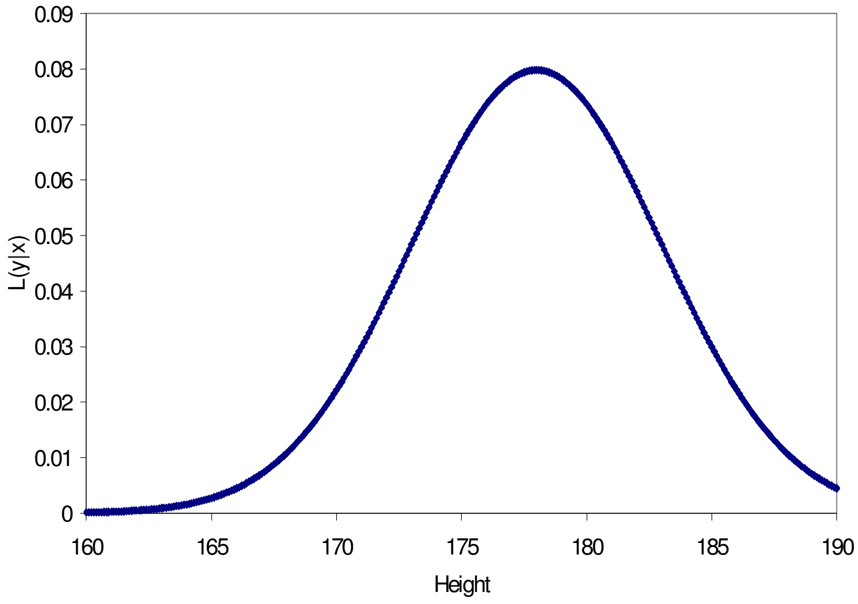
Bayesian methods

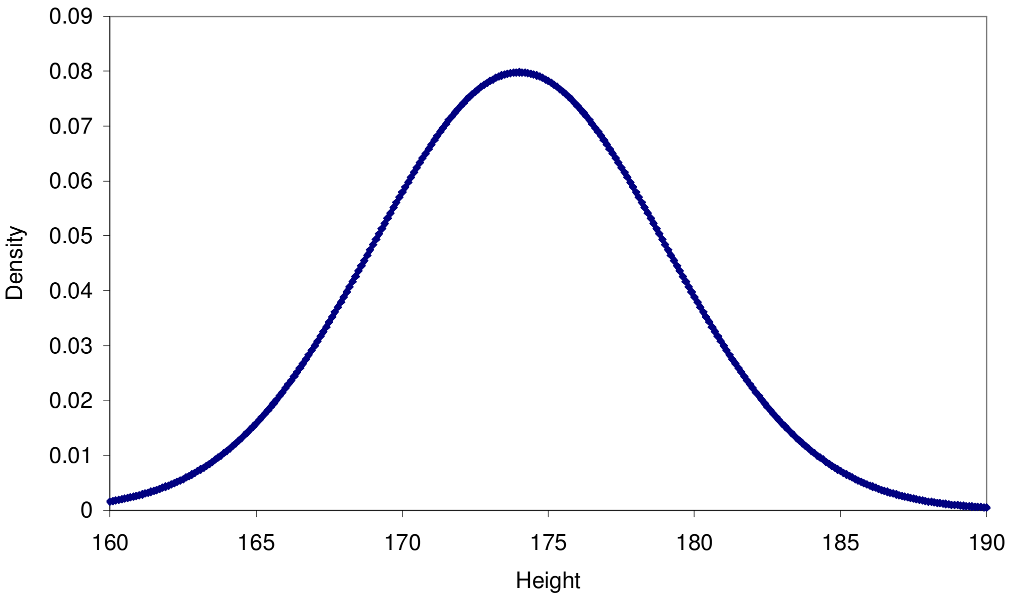
- Bayes theorem

$$P(x \mid y) \propto P(y \mid x)P(x)$$

Likelihood of data (y) given height x, most likely x = 178cm

Prior probability of x (average height)





Bayesian methods

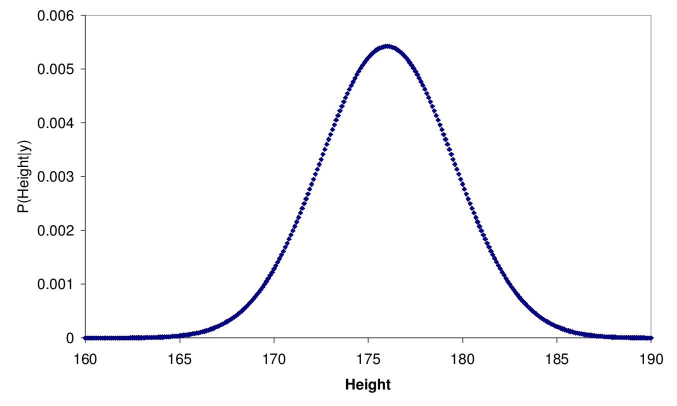
- Bayes theorem

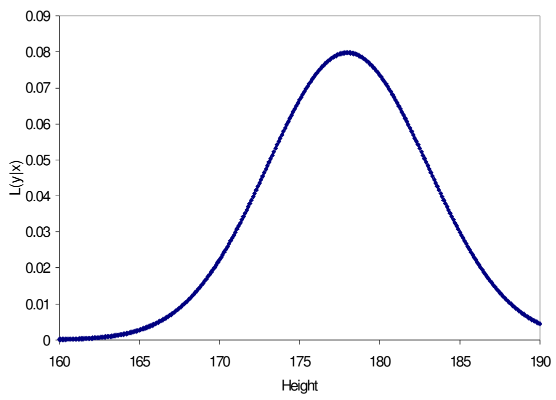
$$P(x \mid y) \propto P(y \mid x)P(x)$$

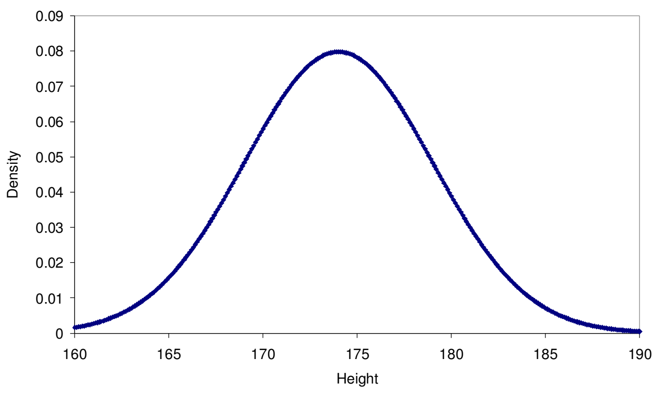
P(x|y) mean = 176cm

L(y|x)

P(x)





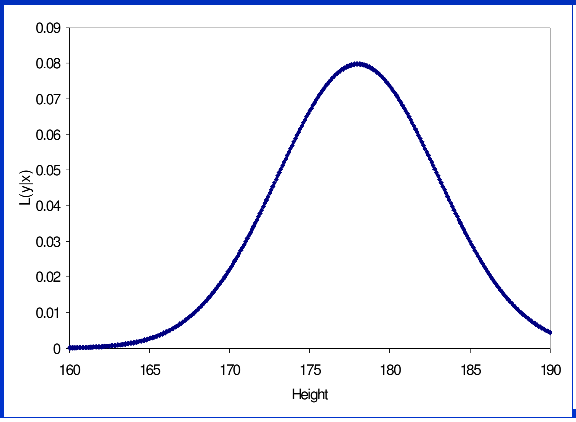


Bayesian methods

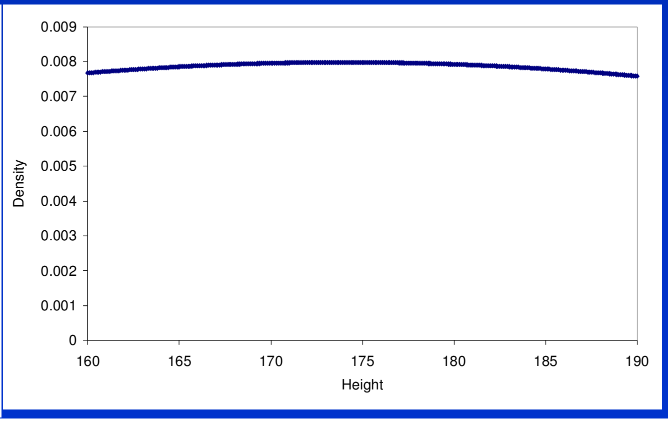
- Bayes theorem
- Less certainty about prior information? Use *less* informative (flat) prior

$$P(x \mid y) \propto P(y \mid x) P(x)$$

L(y|x)



P(x)



Bayesian methods

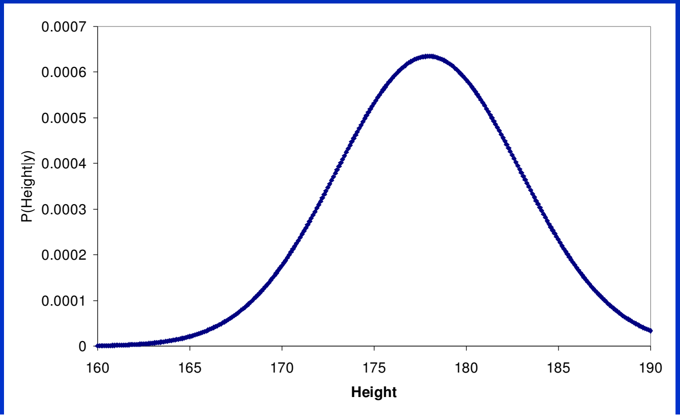
- Bayes theorem
- Less certainty about prior information? Use *less* informative (flat) prior

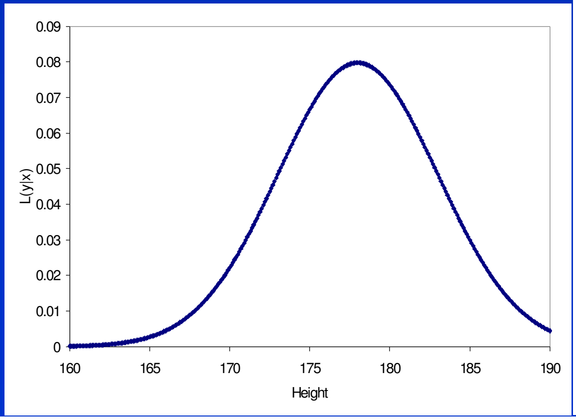
$$P(x \mid y) \propto P(y \mid x) P(x)$$

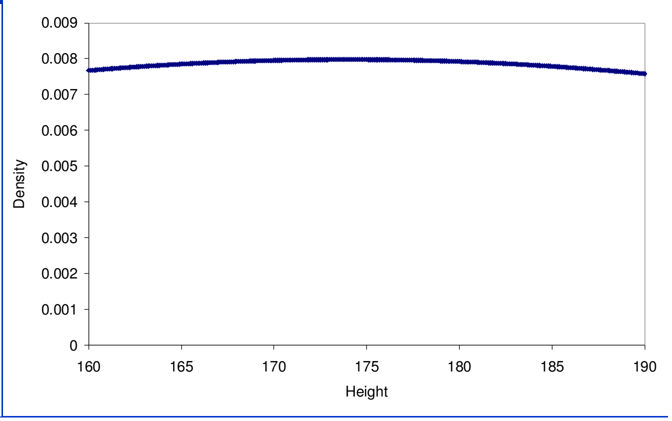
P(x|y) mean = 178cm

L(y|x)

P(x)







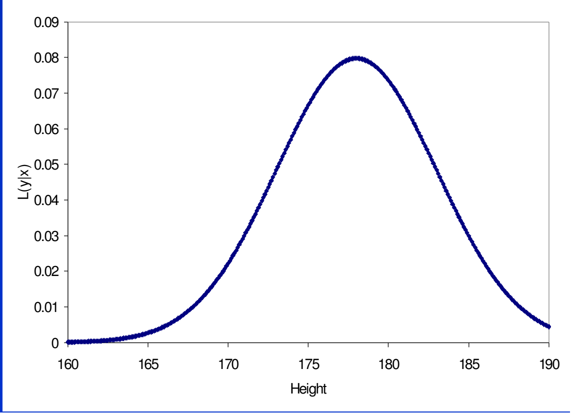
Bayesian methods

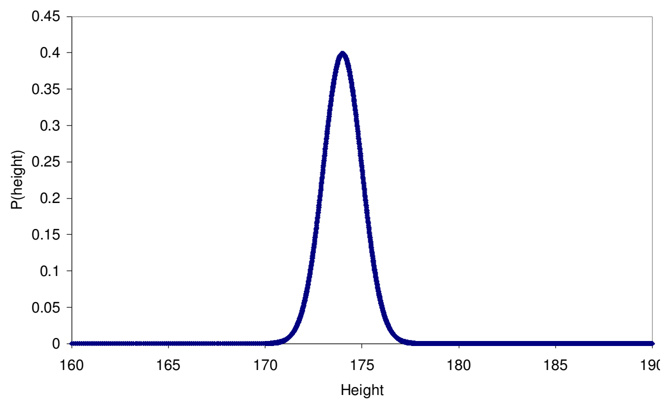
- Bayes theorem
- More certainty about prior information? Use *more* informative prior

$$P(x \mid y) \propto P(y \mid x) P(x)$$

L(y|x)

P(x)





Bayesian methods

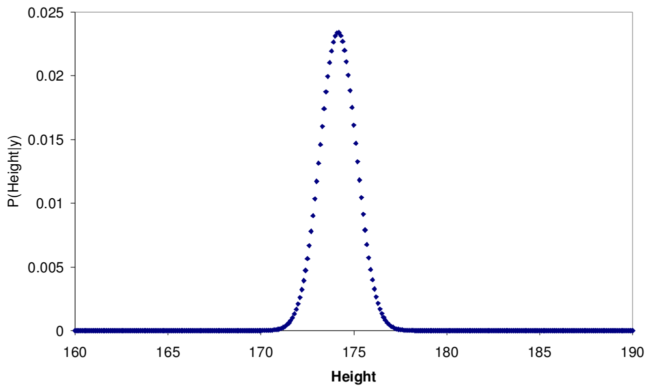
- Bayes theorem
- More certainty about prior information? Use *more* informative prior

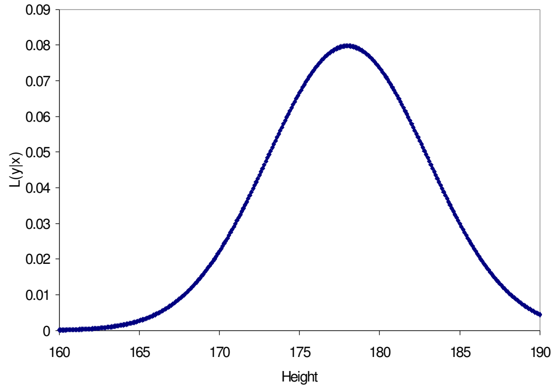
$$P(x \mid y) \propto P(y \mid x)P(x)$$

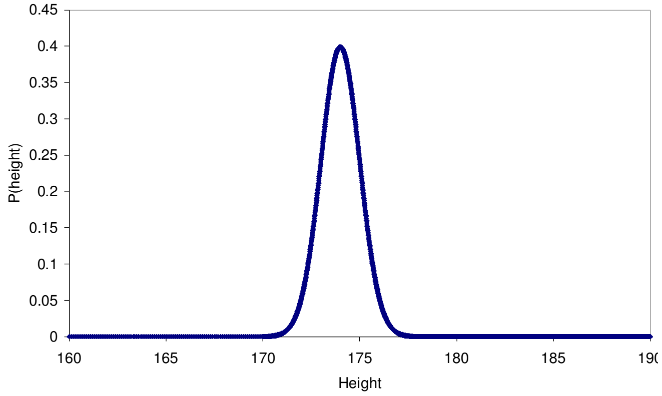
P(x|y) mean = 174.5cm

L(y|x)

P(x)





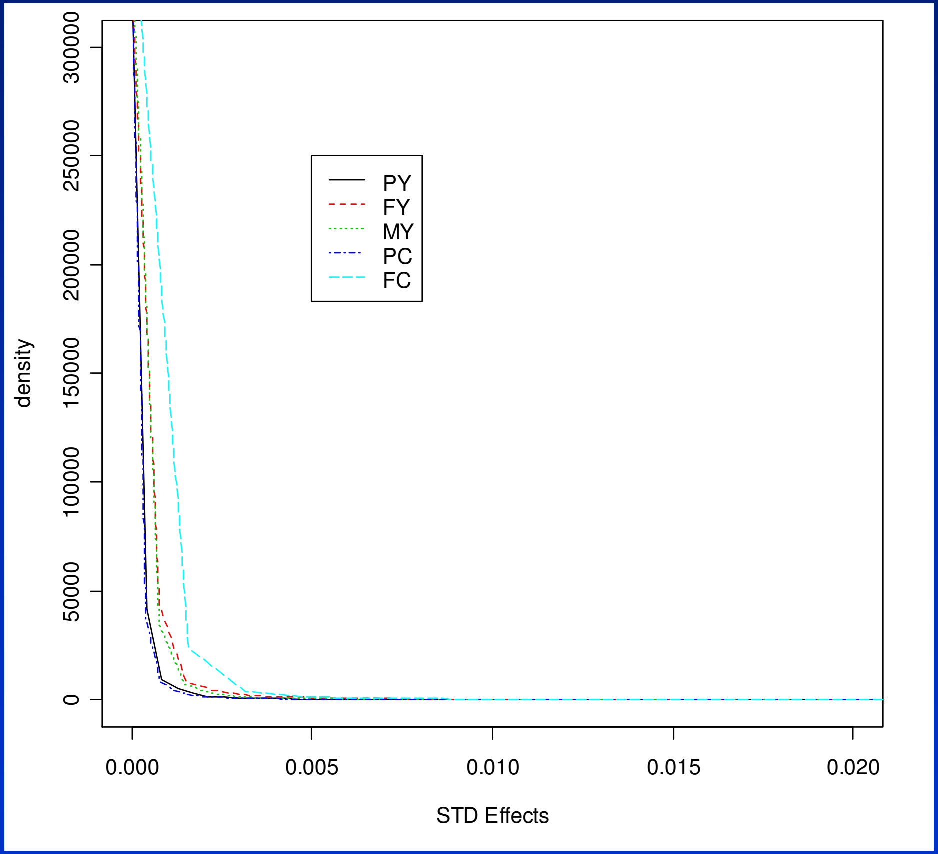


Genomic selection

- Introduction
- Genomic selection with Least Squares and BLUP
- Introduction to Bayesian methods
- Genomic selection with Bayesian methods
- Comparison of accuracy of methods

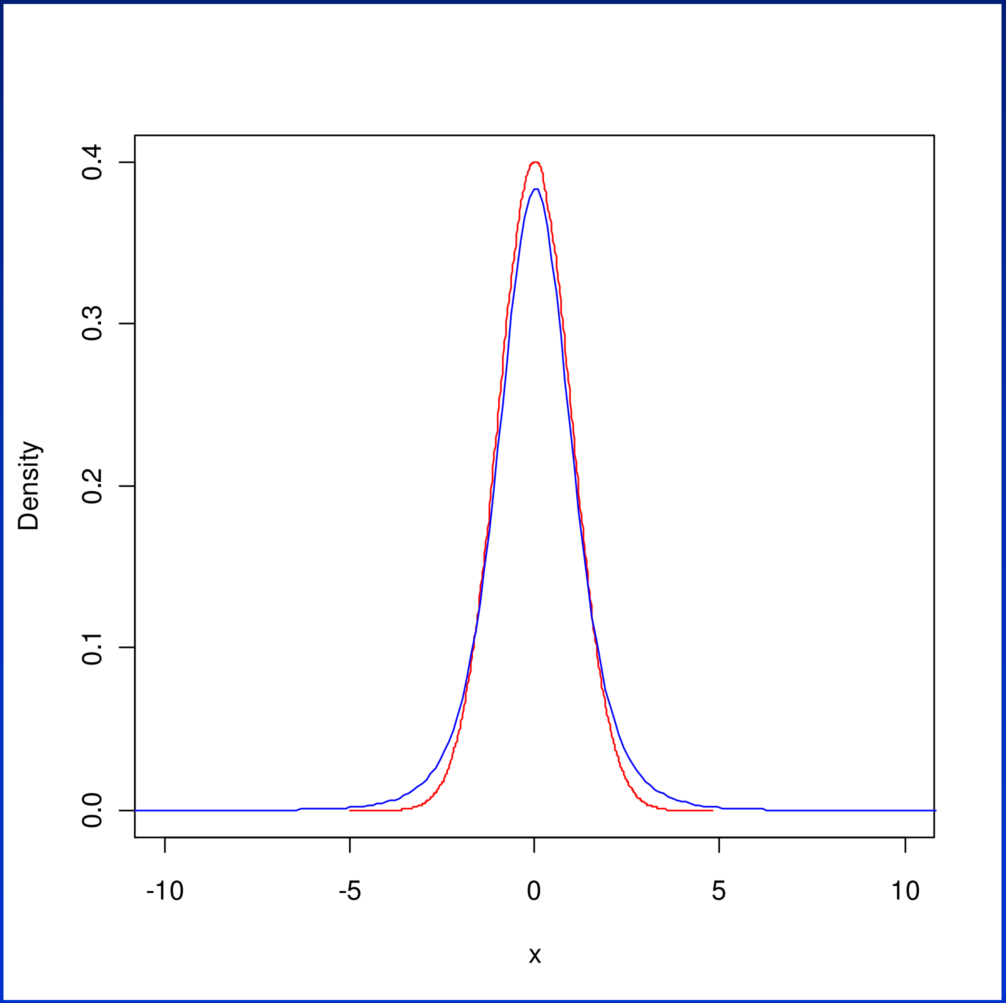
Genomic selection

- For some traits prior knowledge suggests t-distribution of effects
- How to incorporate this into our predictions?



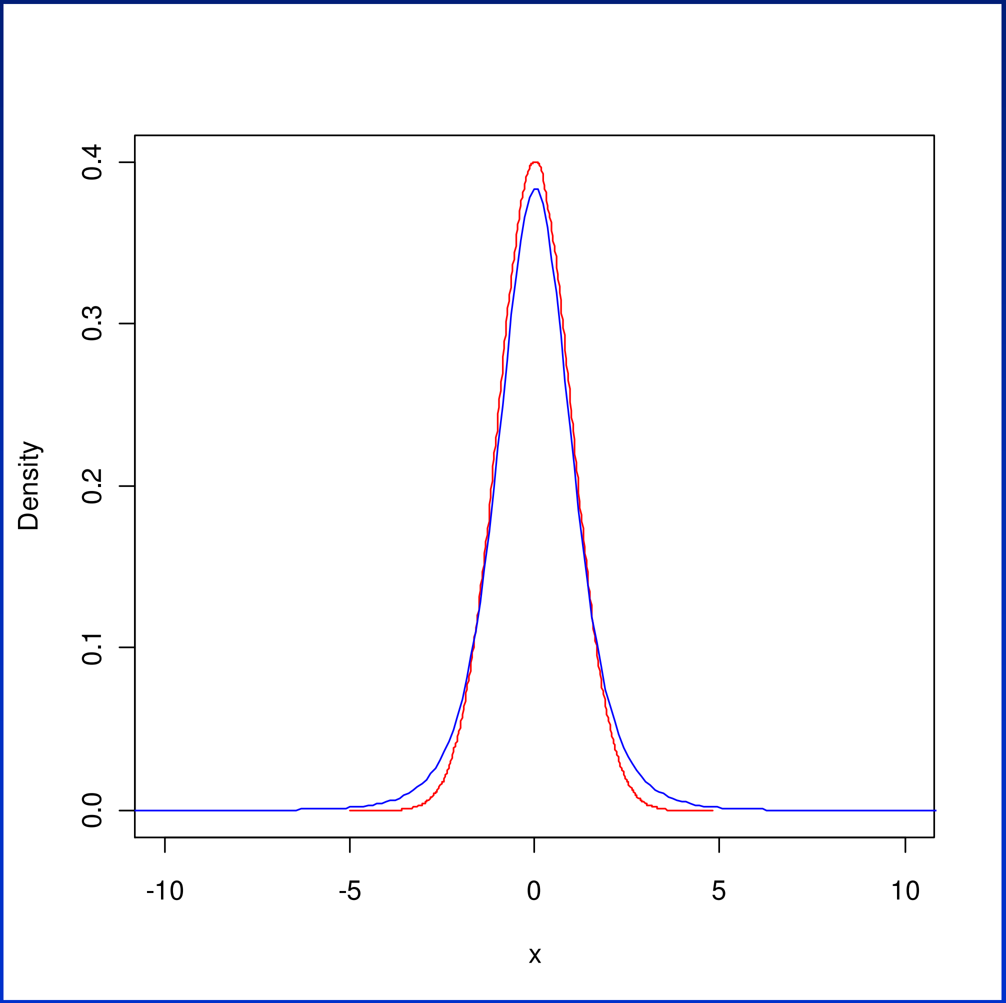
Genomic selection

- For some traits prior knowledge suggests t-distribution of effects
- How to incorporate this into our predictions?

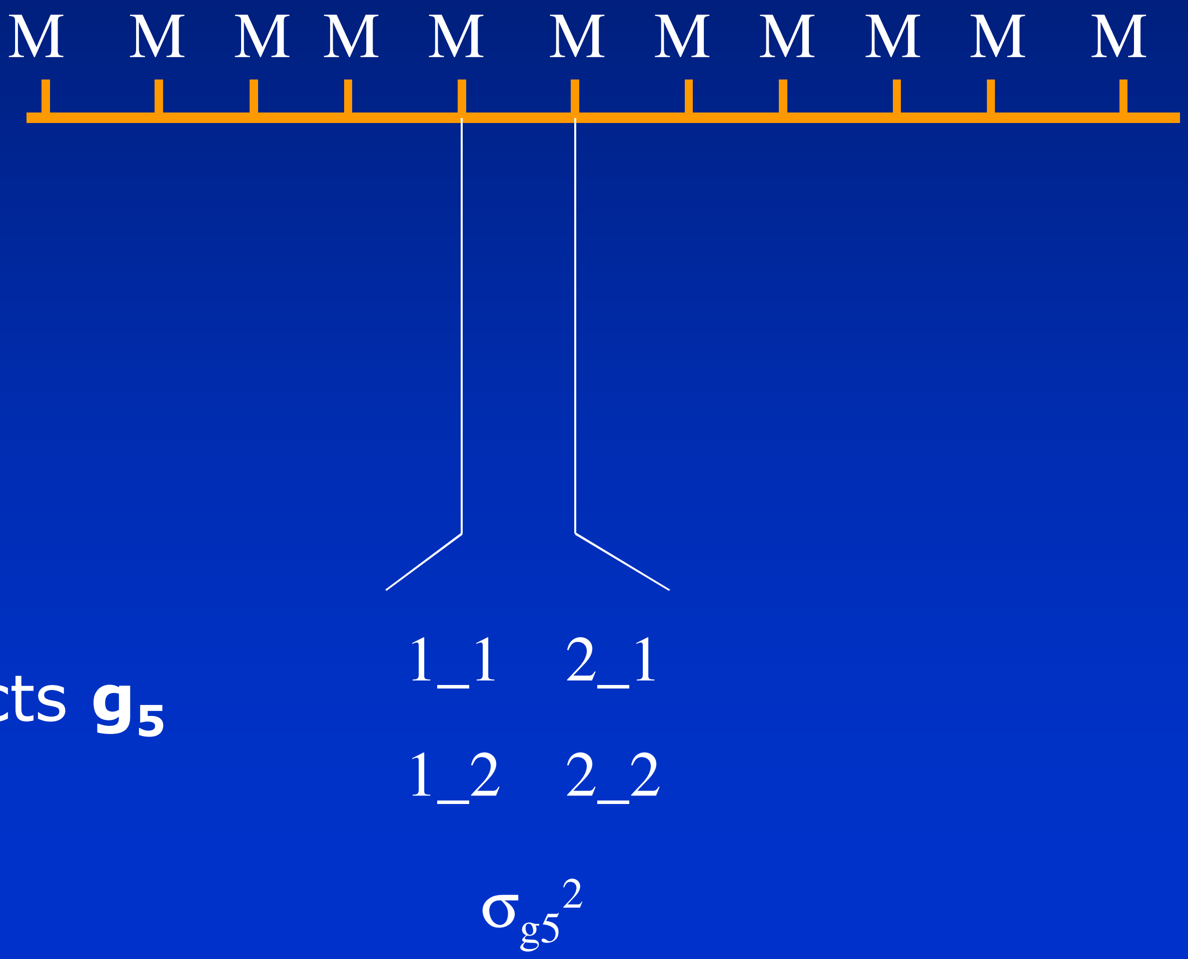


Genomic selection

- The **t distribution** can be presented as a two level hierarchical model
- Allow different variances between chromosome segments
- Assume a distribution of these variances
- Computationally easier to deal with than original form



Genomic selection

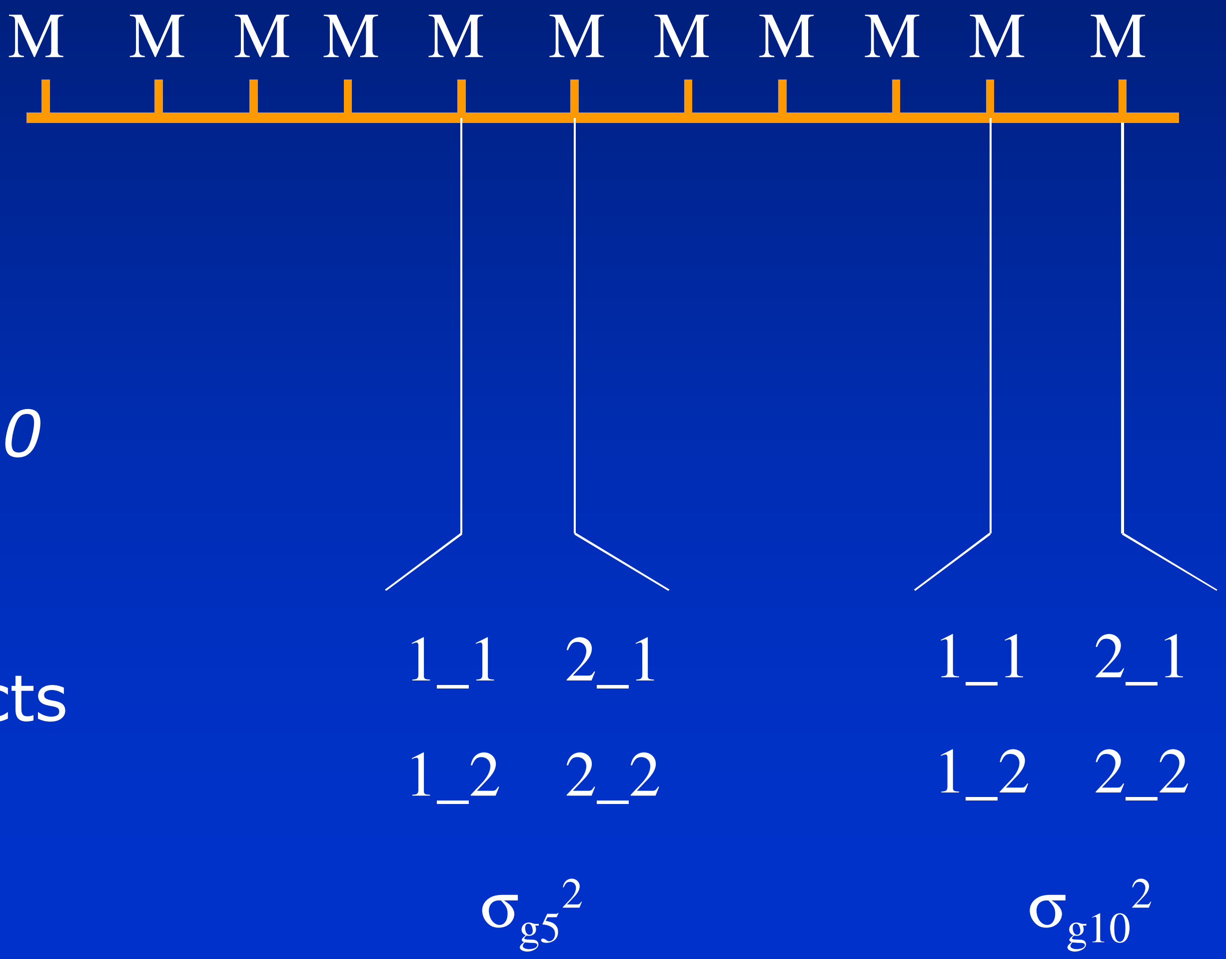


chromosome

chromosome
segment *5*

chromosome
segment effects $\mathbf{g_5}$

Genomic selection

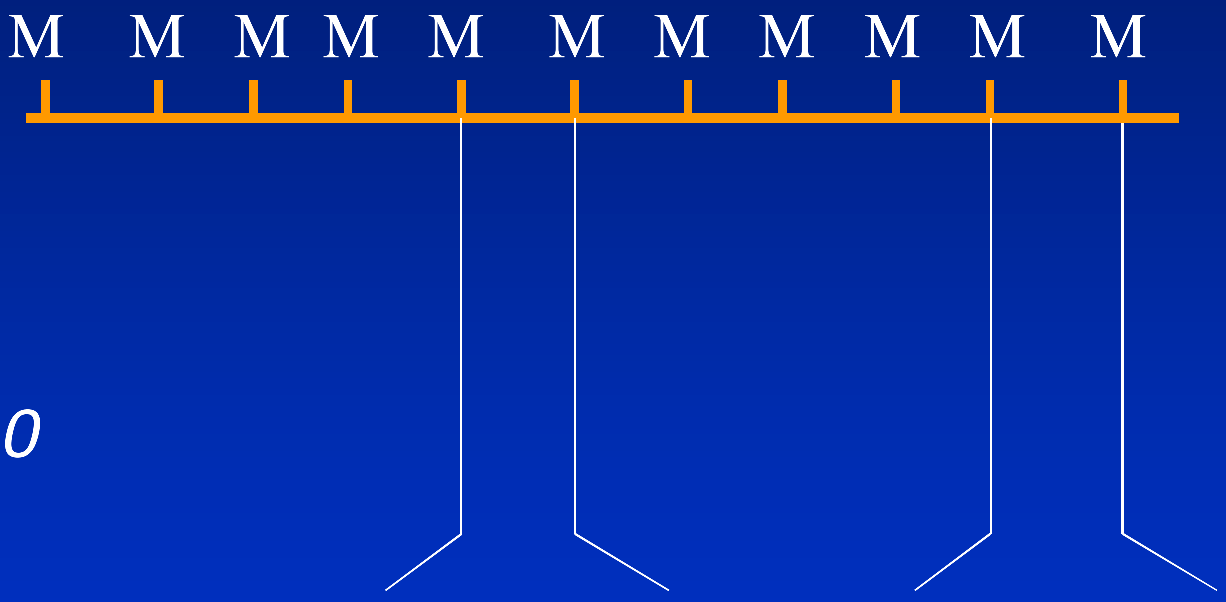


chromosome

chromosome
segment *5, 10*

chromosome
segment effects
g_5, g_{10}

Genomic selection



chromosome

chromosome
segment *5, 10*

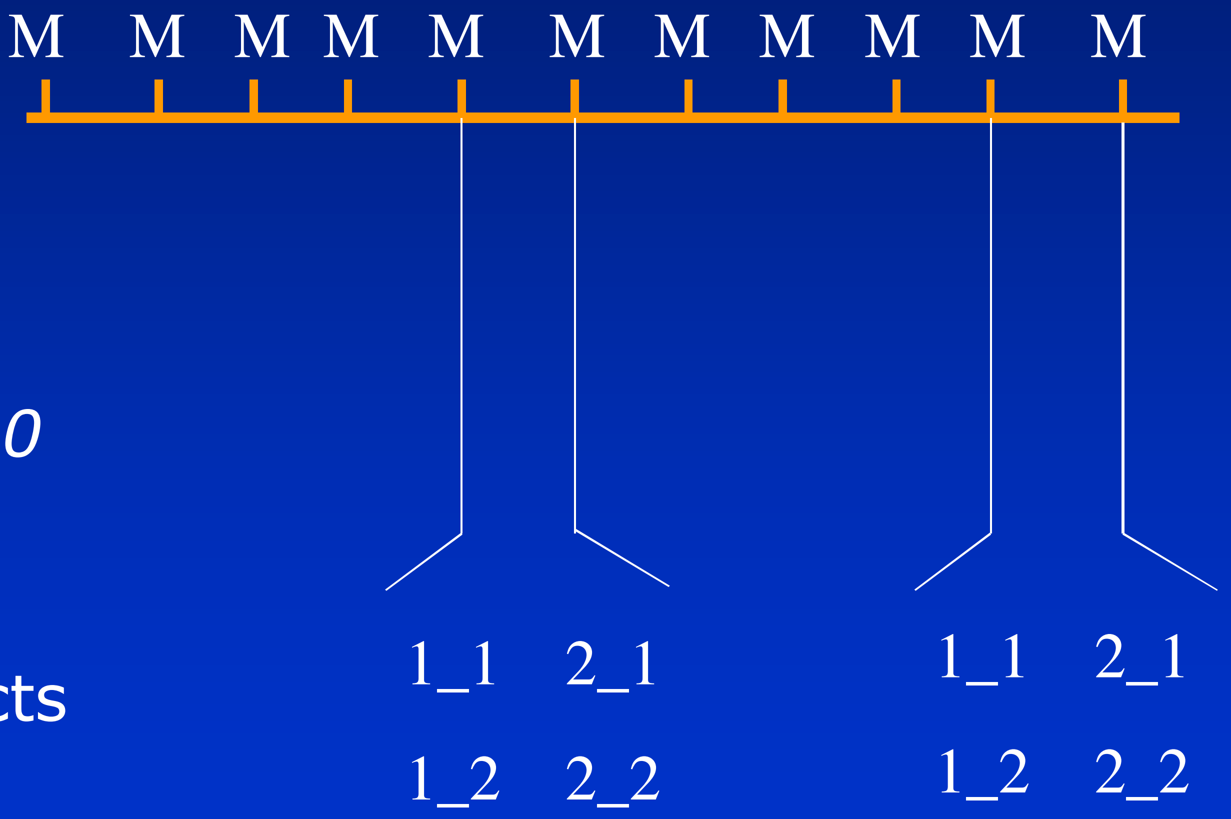
chromosome
segment effects
$\mathbf{g}_{5}, \mathbf{g}_{10}$

1_1 2_1 1_1 2_1

1_2 2_2 1_2 2_2

$$\sigma_{g5}^2 = \sigma_{g10}^2$$

Genomic selection



chromosome

chromosome segment *5, 10*

chromosome segment effects **g_5, g_{10}**

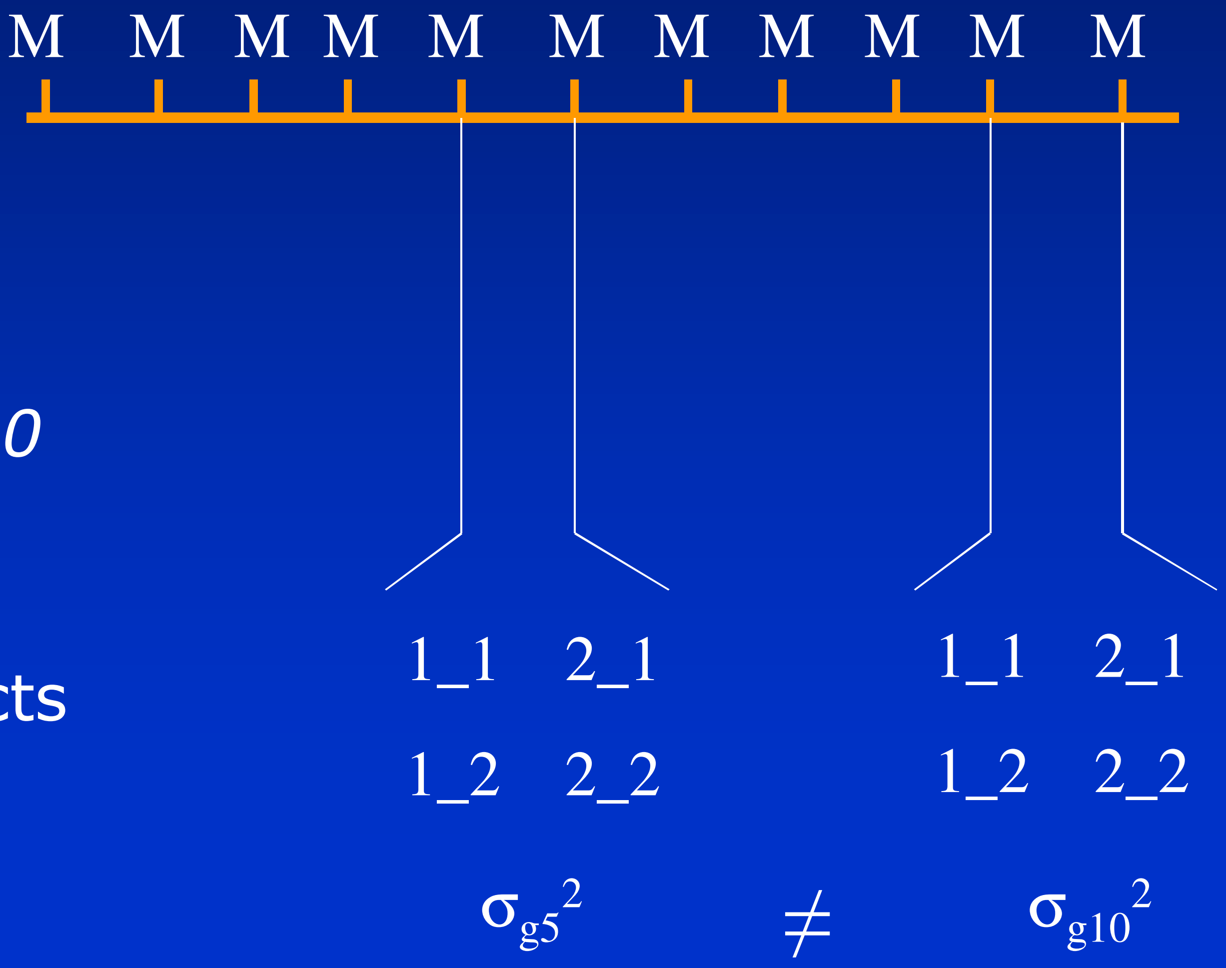
$g_1, g_2, g3, g4, g5 \ldots\ldots \sim N(0, \sigma_g^2)$ $\sigma_{g5}^2 = \sigma_{g10}^2$

Genomic selection

chromosome

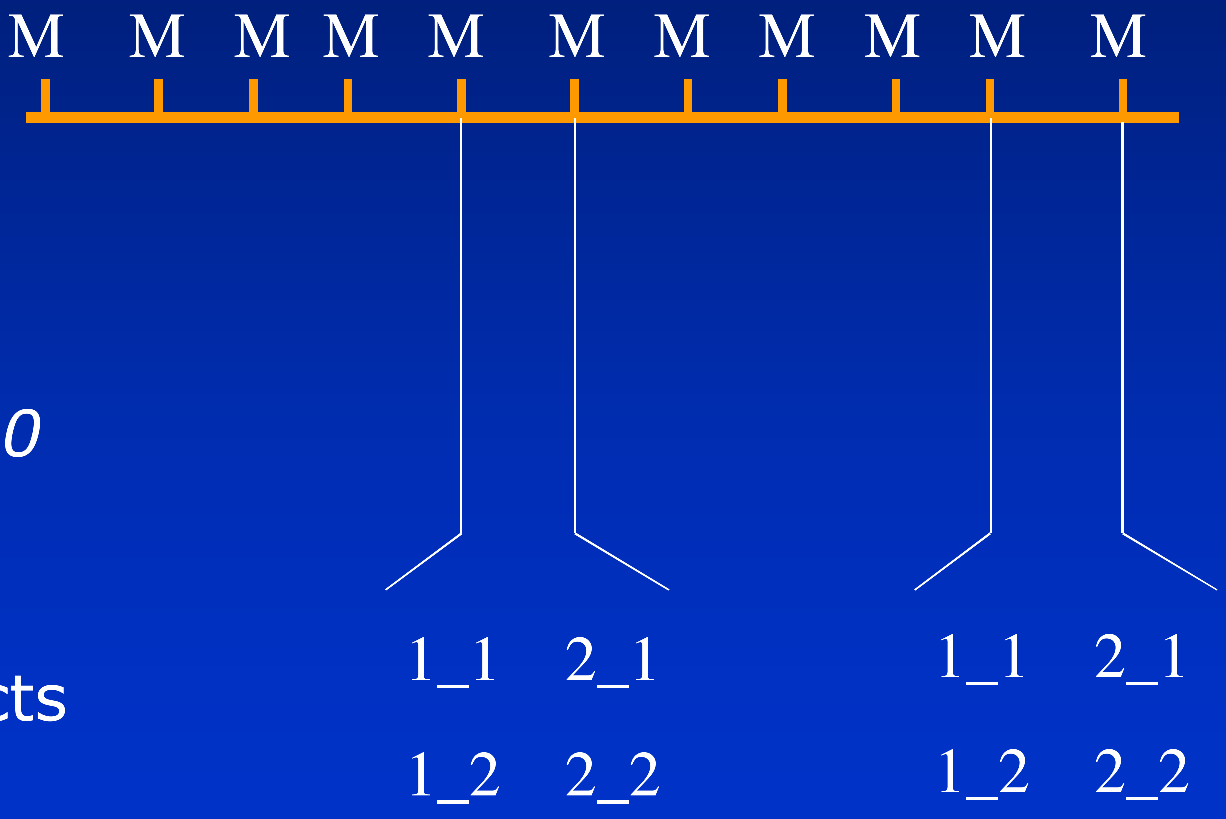
chromosome segment *5, 10*

chromosome segment effects $\mathbf{g}_5, \mathbf{g}_{10}$



$\sigma_{g5}^2 \neq \sigma_{g10}^2$

Genomic selection



M M M M M M M M M M M

chromosome

chromosome
segment *5, 10*

chromosome
segment effects
g_5, g_{10}

1_1 2_1 1_1 2_1

1_2 2_2 1_2 2_2

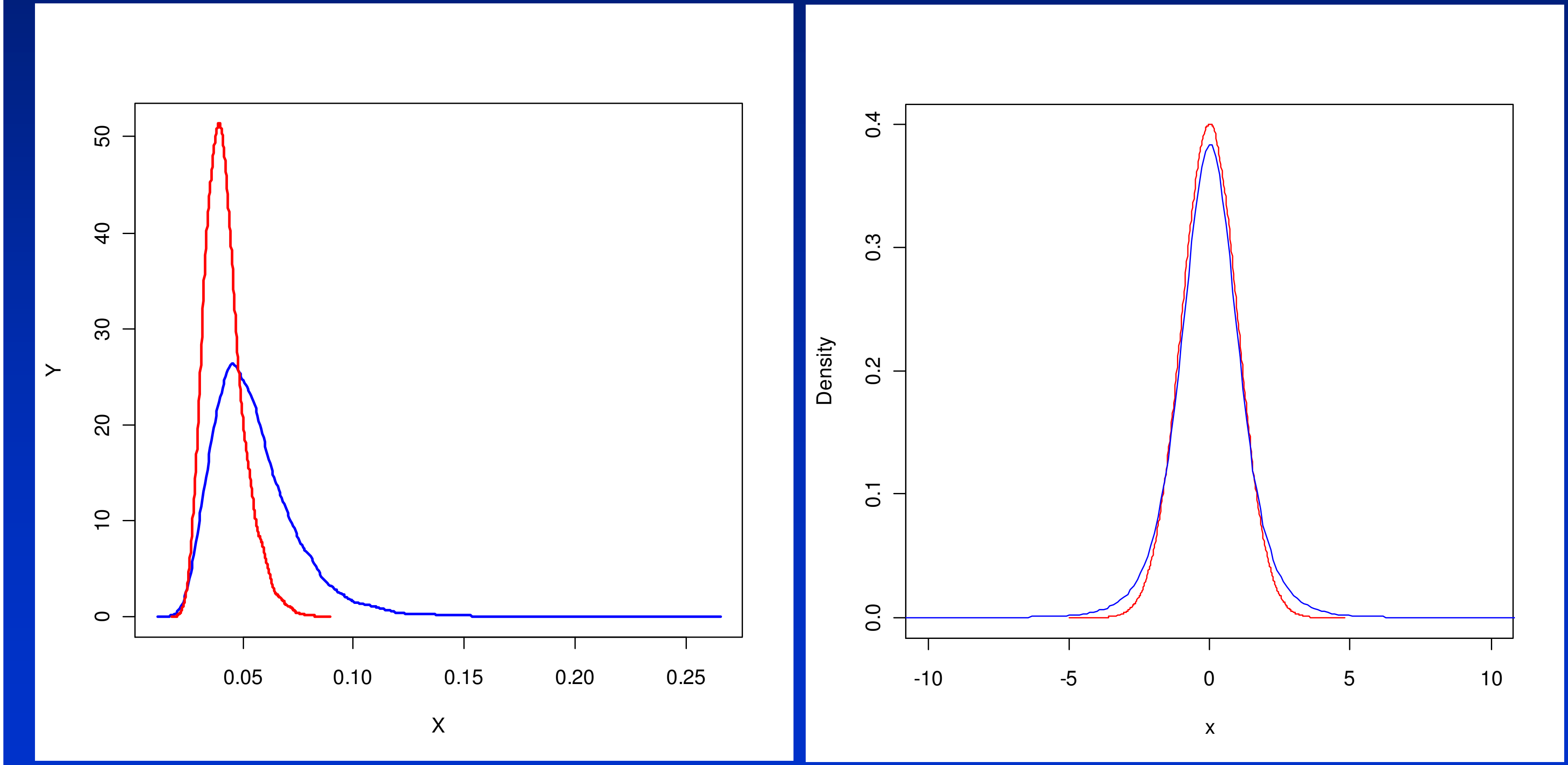
$g_5 \sim N(0, \sigma_{g5}^2)$, $g_{10} \sim N(0, \sigma_{g10}^2)$ $\sigma_{g5}^2 \neq \sigma_{g10}^2$

Bayesian methods

- Now lets allow different variances of chromosome segment effects

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}}_1 \\ \cdot \\ \hat{\mathbf{g}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n}'\mathbf{X_1} & \cdot & \mathbf{1_n}'\mathbf{X_p} \\ \mathbf{X_1}'\mathbf{1_n} & \mathbf{X_1}'\mathbf{X_1} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_{g1}^2} & \cdot & \mathbf{X_1}'\mathbf{X_p} \\ \cdot & \cdot & \cdot & \cdot \\ \mathbf{X_p}'\mathbf{1_n} & \mathbf{X_p}'\mathbf{X_1} & \cdot & \mathbf{X_p}'\mathbf{X_p} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_{gp}^2} \end{bmatrix}^{-1} \begin{bmatrix} 1_n' y \\ X_1' y \\ \cdot \\ X_p' y \end{bmatrix}$$

Distribution of σ_{gj}^2 --> Distribution of g_j



Bayesian methods

- Now lets allow different variances of chromosome segment effects
- Need two levels of models
  - Data

$$\boxed{P(\mathbf{g},\mu \mid y) \propto P(y \mid \mathbf{g},\mu)P(\mathbf{g},\mu)}$$

  - Variances of chromosome segment effects

$$\boxed{P(\sigma_{gi}^2 \mid g_i) \propto P(g_i \mid \sigma_{gi}^2)P(\sigma_{gi}^2)}$$

Bayesian methods

- Now lets allow different variances of chromosome segment effects
- Data

$$P(\mathbf{g},\mu \mid y) \propto P(y \mid \mathbf{g},\mu)P(\mathbf{g},\mu)$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}}_1 \\ . \\ \hat{\mathbf{g}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X_1} & . & \mathbf{1_n'X_p} \\ \mathbf{X_1'1_n} & \mathbf{X_1'X_1} + \mathbf{I}\frac{\sigma_e^2}{\sigma_{g1}^2} & . & \mathbf{X_1'X_p} \\ . & . & . & . \\ \mathbf{X_p'1_n} & \mathbf{X_p'X_1} & . & \mathbf{X_p'X_p} + \mathbf{I}\frac{\sigma_e^2}{\sigma_{gp}^2} \end{bmatrix}^{-1} \begin{bmatrix} 1_n'y \\ X_1'y \\ . \\ X_p'y \end{bmatrix}$$

Bayesian methods

- Variances of chromosome segments

$$P(\sigma_{gi}^2 \mid g_i) \propto P(g_i \mid \sigma_{gi}^2)P(\sigma_{gi}^2)$$

- Note that these variance components are not the parameters of interest
- However they are useful intermediates to arrive at better inferences for the g_i
- Amount of shrinkage of effects varies between segments

Bayesian methods

- Variances of chromosome segments

$$P(\sigma_{gi}^2 \mid g_i) \propto P(g_i \mid \sigma_{gi}^2) P(\sigma_{gi}^2)$$

- Prior?
  - – Inverted chi square convenient for variances

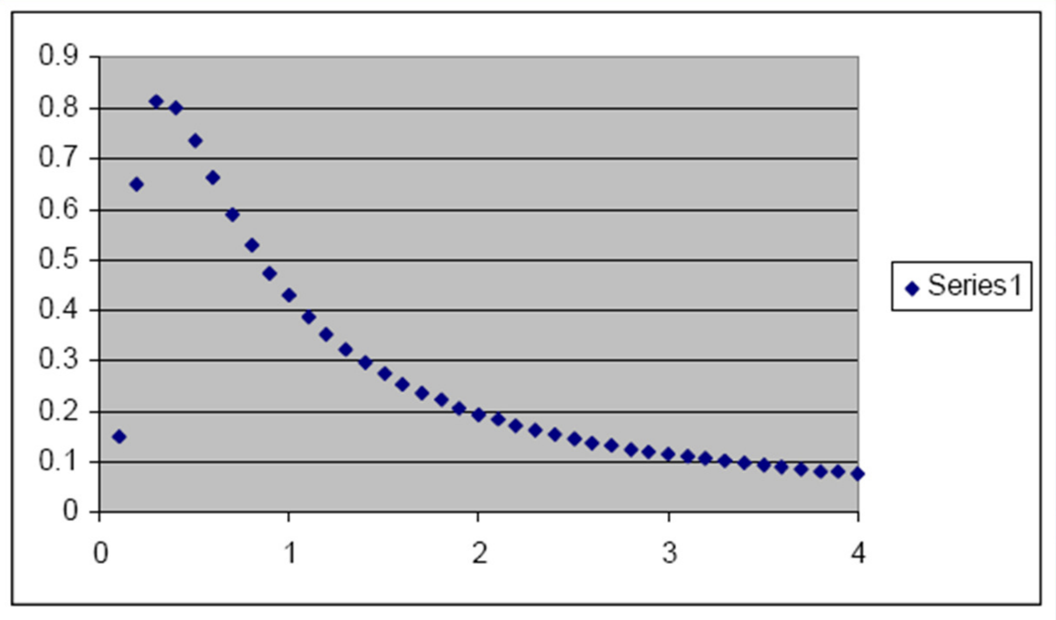
Bayesian methods

- Prior?
  - Inverted chi square convenient for variances
  - An inverted chi square with v degrees of freedom and scaled by S^2, eg.

$$S^2 / \chi_v^2$$

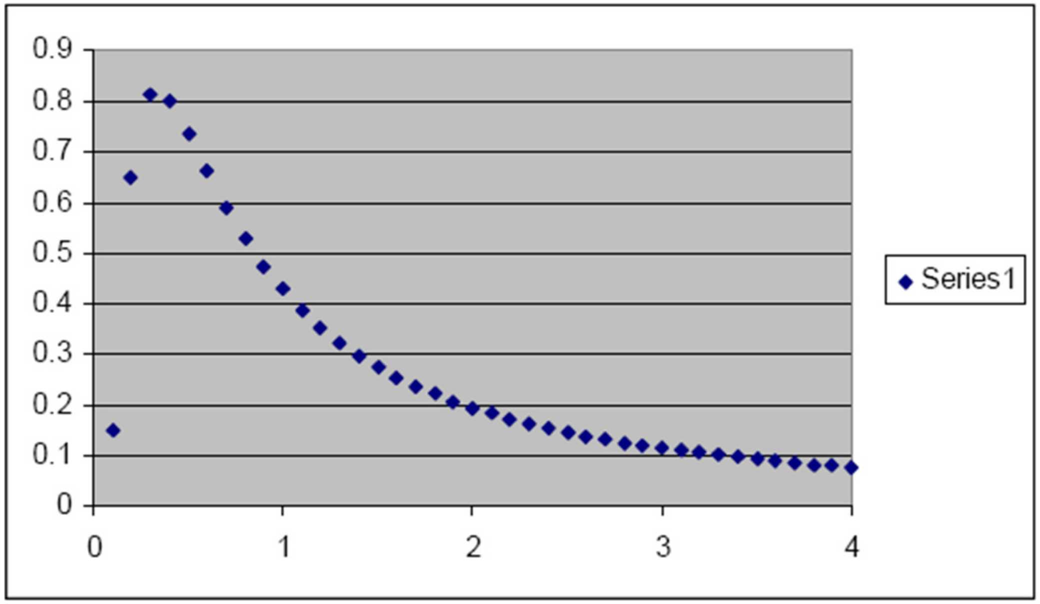
  - Describes a distribution with
    - mean $vS^2/(v-2)$
    - variance $\dfrac{2v^2S^4}{(v-2)^2(v-4)}$
  - Larger v, more informative prior = more belief about variance

Bayesian methods

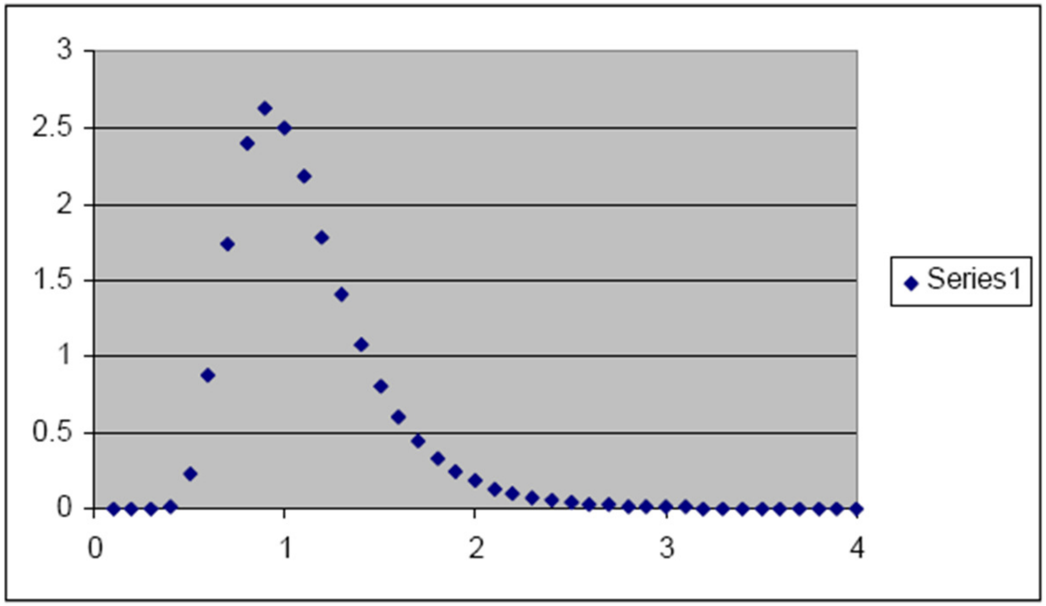


$\nu=2$

Bayesian methods



v=2



v=20

Bayesian methods

- Variances of chromosome segments

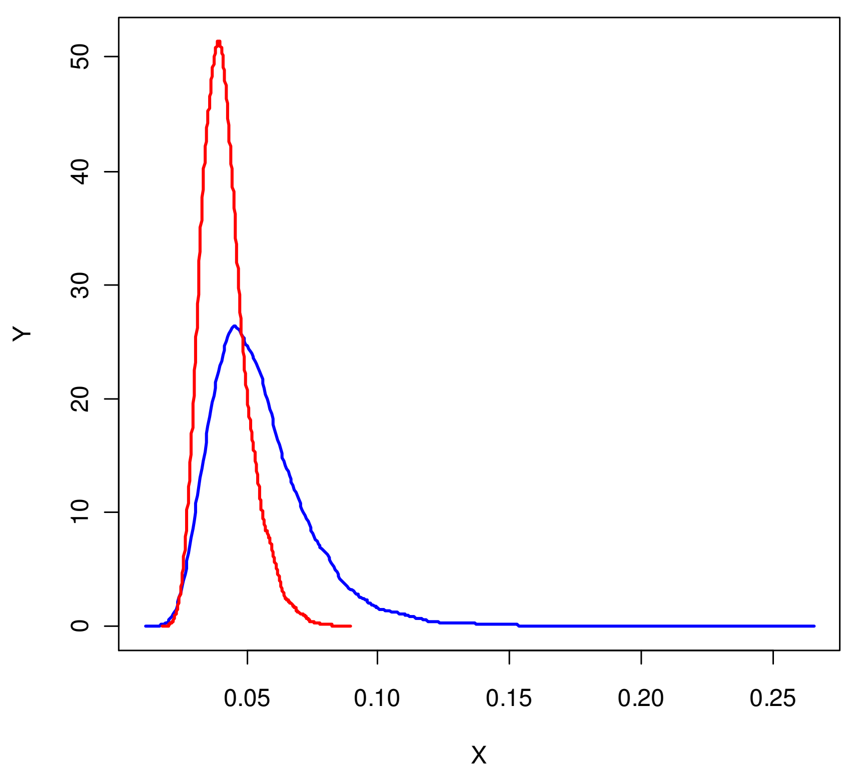
$$P(\sigma_{gi}^2 \mid g_i) \propto P(g_i \mid \sigma_{gi}^2) P(\sigma_{gi}^2)$$

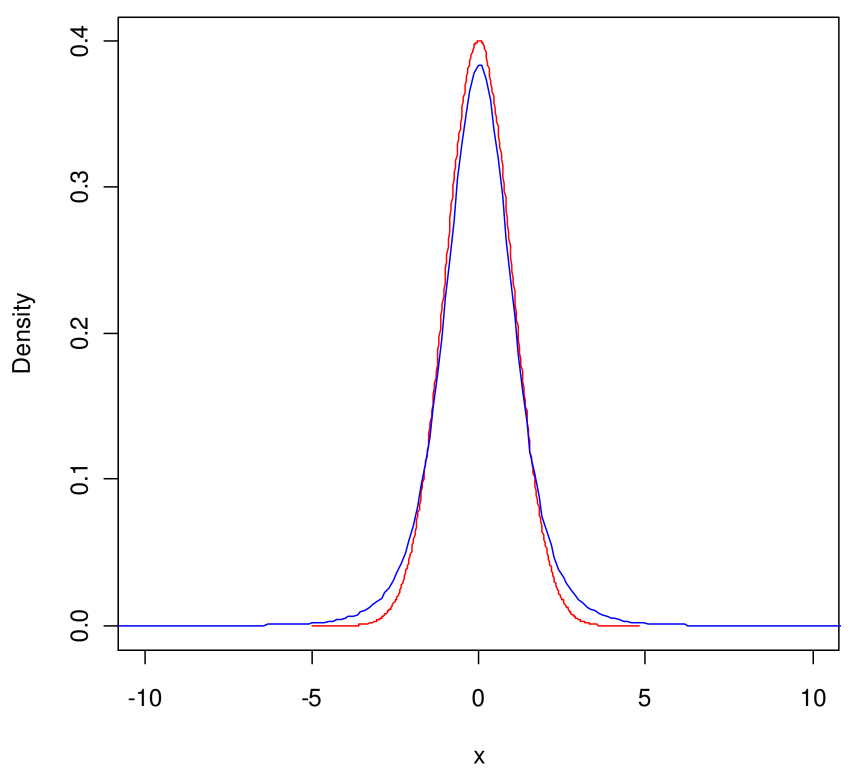
- Prior?

$$S^2 / \chi_v^2$$

- We can choose *v* and S^2 so that the prior reflects our knowledge that there are many QTL of small effect and few of large effect

Distribution of σ_{gj}^2 --> Distribution of g_j



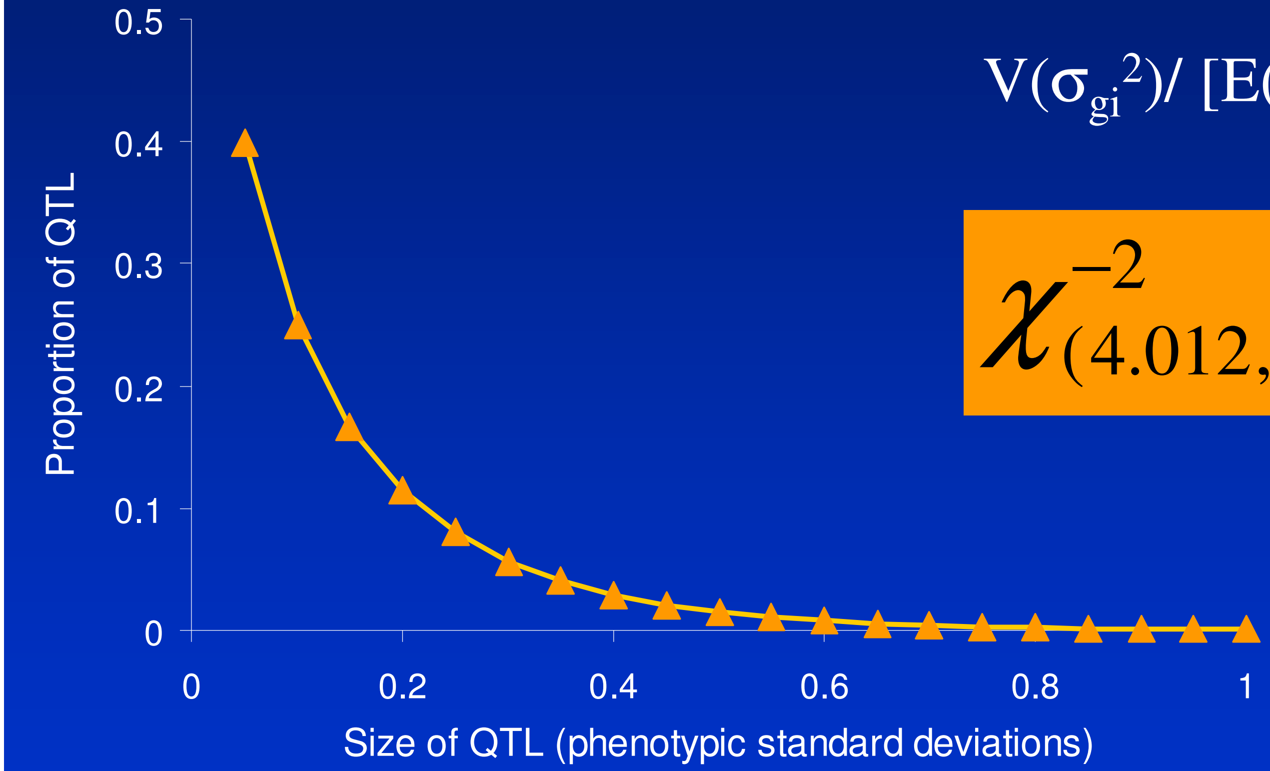


Bayesian methods

$E(\sigma_{gi}^{2})=S/(v\text{-}2)$

$V(\sigma_{gi}^{2})/\ [E(\sigma gi2)]^{2}=2/(v\text{-}4)$

$$\chi^{-2}_{(4.012,0.002)}$$



Bayesian methods

- Variances of chromosome segments

$$P(\sigma_{gi}^{2} \mid \mathbf{g_i}) \propto P(\mathbf{g_i} \mid \sigma_{gi}^{2}) P(\sigma_{gi}^{2})$$

- Posterior?
  - An advantage of choosing the inverse chi-square distribution for the prior is that the posterior will also be an inverse chi-square distribution
    - Degrees of freedom = prior + data
    - Scaling factor = sums of squares prior (S^2) + sums of squares from data

Bayesian methods

- Variances of chromosome segments

$$P(\sigma^2_{gi} \mid \mathbf{g_i}) \propto P(\mathbf{g_i} \mid \sigma^2_{gi}) P(\sigma^2_{gi})$$

- Posterior?
  - n_i = number of haplotype effects

$$\chi^{-2}_{(v+n_i, S^2+\mathbf{g_i}'\mathbf{g_i})}$$

Bayesian methods

- Variances of chromosome segments

$$P(\sigma^2_{gi} \mid \mathbf{g_i}) \propto P(\mathbf{g_i} \mid \sigma^2_{gi}) P(\sigma^2_{gi})$$

- Posterior?

$$\chi^{-2}_{(4.012+n_i,\, 0.002+\mathbf{g_i}'\mathbf{g_i})}$$

- But posterior cannot be estimated directly, dependent on $\mathbf{g_i}$!!

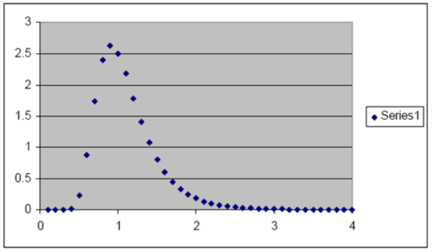
Bayesian methods

- Solution is to use Gibbs sampling
  - Draw samples from the posterior distributions of parameters conditional on all other effects
  - The average of these samples can be used as the estimates of the parameters

Bayesian methods

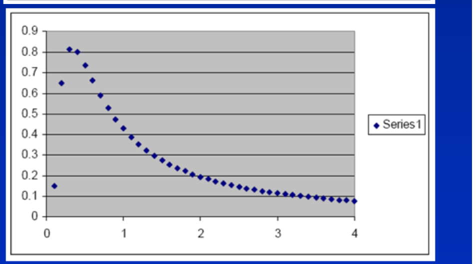
- Gibbs sampling scheme
  - Parameters to estimate and their posteriors
  - P(σ_{gi}^2|g_i)

$$\chi^{-2}_{(4.012+n_i,\,0.002+\mathbf{g_i}'\mathbf{g_i})}$$



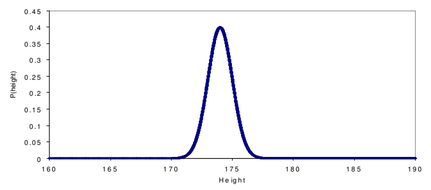
  - P(σ_e^2|**e**)

$$\chi^{-2}_{(n-2,\,\mathbf{e}'\mathbf{e})}$$



  - P(μ|**y**,**e**,**g**, σ_e

$$N\left(\frac{1}{n}\left(\mathbf{1}_{\mathbf{n}}'\mathbf{y}-\mathbf{1}_{\mathbf{n}}'\mathbf{X}\mathbf{g}\right),\sigma_e^2/n\right)$$



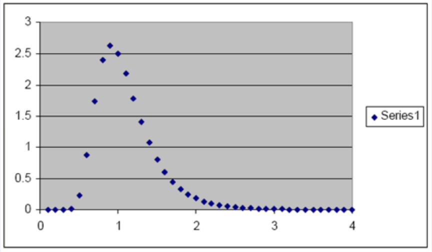
  - P(g_{ij}|**y**,μ,**g**≠ij,σ_{gi}^2,σ_e^2

$$N\left(\frac{\mathbf{X}_{\mathbf{ij}}'\mathbf{y}-\mathbf{X}_{\mathbf{ij}}'\mathbf{X}\mathbf{g}_{(\mathbf{ij=0})}-\mathbf{X}_{\mathbf{ij}}'\mathbf{1}_{\mathbf{n}}\mu}{\mathbf{X}_{\mathbf{ij}}'\mathbf{X}_{\mathbf{ij}}+\sigma_e^2/\sigma_{gi}^2},\sigma_e^2/\left(\mathbf{X}_{\mathbf{ij}}'\mathbf{X}_{ij}+\sigma_e^2/\sigma_{gi}^2\right)\right)$$

Bayesian methods

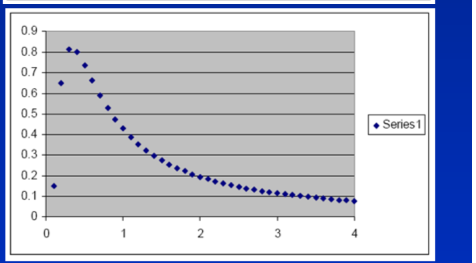
- Gibbs sampling scheme
  - Parameters to estimate and their posteriors
  - P(σ_{gi}^2|g_i)

$$\chi^{-2}_{(4.012+n_i\,,0.002+\mathbf{g_i}\,'\mathbf{g_i}\,)}$$



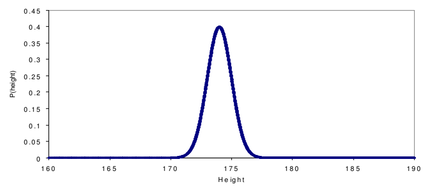
  - P(σ_e^2|**e**)

$$\chi^{-2}_{(n-2,\mathbf{e'e})}$$



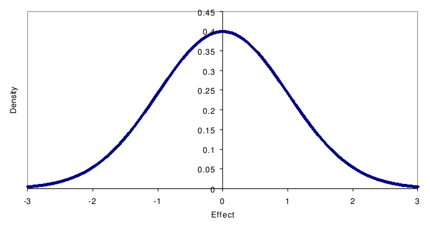
  - P(μ|**y**,**e**,**g**, σ_e

$$N\left(\frac{1}{n}\left(\mathbf{1}_{\mathbf{n}}^{'}\mathbf{y}-\mathbf{1}_{\mathbf{n}}^{'}\mathbf{Xg}\right),\sigma_e^2\,/\,n\right)$$



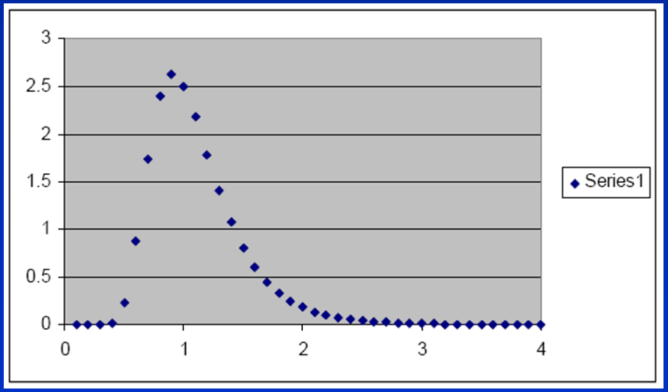
  - P(g_{ij}|**y**,μ,**g**≠ij,σ_{gi}^2,σ_e^2

$$N\left(\frac{\mathbf{X}_{\mathbf{ij}}^{'}\mathbf{y}-\mathbf{X}_{\mathbf{ij}}^{'}\mathbf{Xg}_{\mathbf{(ij=0)}}-\mathbf{X}_{\mathbf{ij}}^{'}\mathbf{1}_{\mathbf{n}}\mu}{\mathbf{X}_{\mathbf{ij}}^{'}\mathbf{X}_{\mathbf{ij}}+\sigma_e^2/\sigma_{gi}^2},\sigma_e^2/\left(\mathbf{X}_{\mathbf{ij}}\right.\right.$$



Bayesian methods

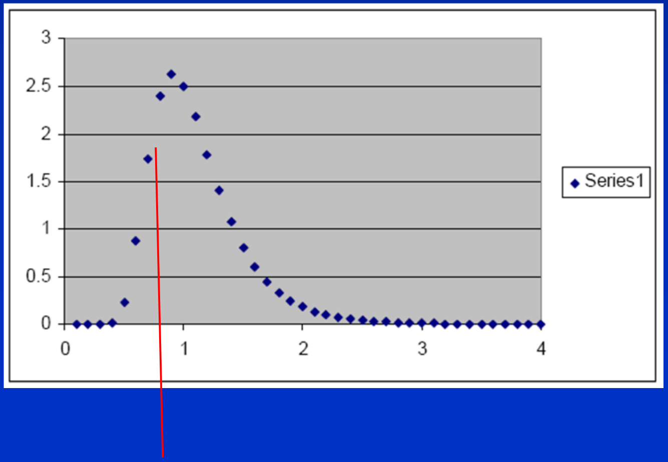
- The Gibbs chain
  - Step 1. Initialise value of **g**, eg. **g**=0.01 and μ, eg $\mu=0.01$
  - Step 2. For each *i*, draw from $P(\sigma_{gi}^2|g_i)$



$$\chi^{-2}_{(4.012+n_i,0.002+\mathbf{g_i}'\mathbf{g_i})}$$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, eg. **g**=0.01 and μ, eg μ=0.01
  - Step 2. For each *i*, draw from $P(\sigma_{gi}^2|g_i)$



$$\chi^{-2}_{(4.012+n_i,\,0.002+\mathbf{g_i}'\mathbf{g_i})}$$

- $\sigma_{g1}^2=0.95$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, eg. **g**=0.01 and μ, eg μ=0.01
  - Step 2. For each *i*, draw from $P(\sigma_{gi}^2|g_i)$
  - Step 3. Draw a sample from $P(\sigma_e^2|e)$ First calculate the **e** as

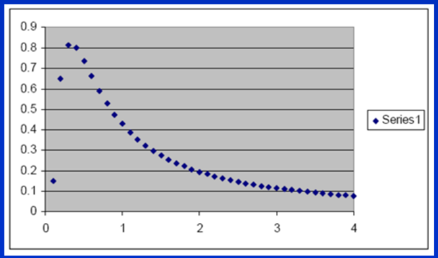
$$\boxed{\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{g} - 1_n^{'}\mu}$$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, eg. **g**=0.01 and μ, eg μ=0.01
  - Step 2. For each *i*, draw from $P(\sigma_{gi}^2 | g_i)$
  - Step 3. Draw a sample from $P(\sigma_e^2 | e)$ First calculate the **e** as

$$\mathbf{e} = \mathbf{y} - \mathbf{Xg} - 1_n^{'}\mu$$

  - Then sample…



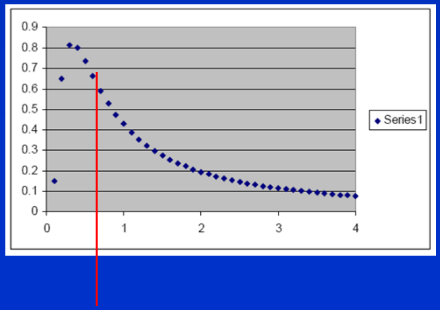
$$\chi^{-2}_{(n-2,\mathbf{e'e})}$$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, eg. **g**=0.01 and μ, eg μ=0.01
  - Step 2. For each *i*, draw from $P(\sigma_{gi}^2|g_i)$
  - Step 3. Draw a sample from $P(\sigma_e^2|e)$ First calculate the **e** as

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{g} - 1_n\mu$$

  - Then sample…



$$\chi^{-2}_{(n-2,\mathbf{e'e})}$$

– $\sigma_e^2 = 0.5$

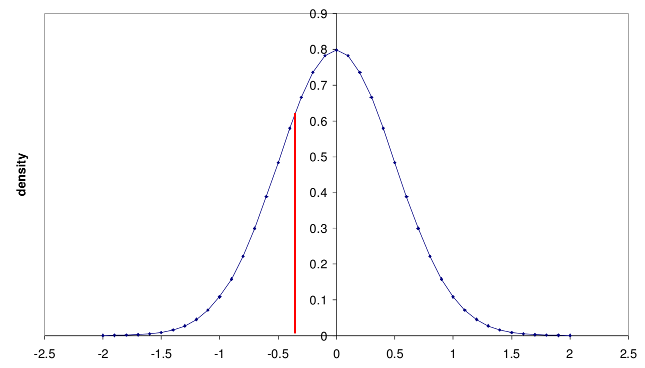
Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, eg. **g**=0.01 and μ, eg $\mu=0.01$
  - Step 2. For each *i*, draw from $P(\sigma_{gi}^2|g_i)$
  - Step 3. Draw a sample from $P(\sigma_e^2|e)$
  - Step 4. Draw a sample from $P(\mu|y,g,\sigma_e^2)$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, eg. **g**=0.01 and μ, eg μ=0.01
  - Step 2. For each *i*, draw from $P(\sigma_{gi}^2|g_i)$
  - Step 3. Draw a sample from $P(\sigma_e^2|e)$
  - Step 4. Draw a sample from $P(\mu|y,g,\sigma_e^2)$

    - μ=-0.1

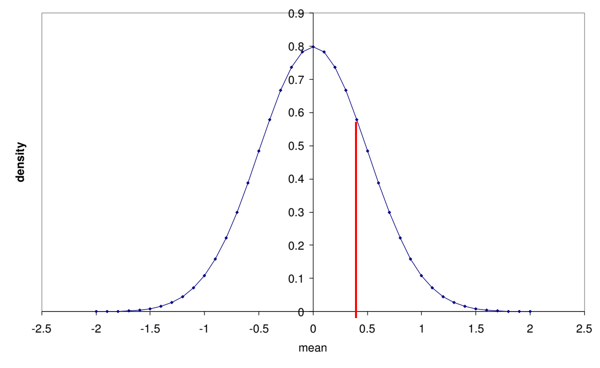


$$N\left(\frac{1}{n}\left(\mathbf{1}_{\mathbf{n}}'\mathbf{y}-\mathbf{1}_{\mathbf{n}}'\mathbf{Xg}\right),\sigma_e^2/n\right)$$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, eg. **g**=0.01 and μ, eg $\mu=0.01$
  - Step 2. For each *i*, draw from $P(\sigma_{gi}^2|g_i)$
  - Step 3. Draw a sample from $P(\sigma_e^2|e)$
  - Step 4. Draw a sample from $P(\mu|y,g,\sigma_e^2)$
  - Step 5. For each g_{ij}, draw from $P(g_{ij}|y,\mu,g,\sigma_{gi}^2,\sigma_e^2)$

    - $g_{11} = 0.5$



Bayesian methods

- The Gibbs chain
  - Repeat steps 2-5 many times to build up samples from posterior distributions of the parameters

Bayesian methods

- The Gibbs chain
  - Repeat steps 2-5 many times to build up samples from posterior distributions of the parameters
  - Finally, take estimates of parameters as average over many cycles
  - Discard first ~ 100 cycles as dependent on starting values

Bayesian methods

- Example
  - Consider a data set with three markers. The data set was simulated as:
  - the effect of a 2 allele at the first marker is 3, the effect of a 2 allele at the second marker is 0, and the effect of a 2 allele at the third marker was -2.
  - the μ was 3
  - σ_e^2 was 0.23. The data set was:

Bayesian methods

- Example

Animal	Phenotype	Marker1 allele 1	Marker1 allele 2	Marker2 allele 1	Marker 2 allele 2	Marker3 allele 1	Marker 3 allele 2
1	9.68	2	2	2	1	1	1
2	5.69	2	2	2	2	2	2
3	2.29	1	2	2	2	2	2
4	3.42	1	1	2	1	1	1
5	5.92	2	1	1	1	1	1
6	2.82	2	1	2	1	2	2
7	5.07	2	2	2	1	2	2
8	8.92	2	2	2	2	1	1
9	2.4	1	1	2	2	1	2
10	9.01	2	2	2	2	1	1
11	4.24	1	2	1	2	2	1
12	6.35	2	2	1	1	1	2
13	8.92	2	2	1	2	1	1
14	-0.64	1	1	2	2	2	2
15	5.95	2	1	1	1	1	1
16	6.13	1	2	2	1	1	1
17	6.72	2	1	2	1	1	1
18	4.86	1	2	2	1	1	2
19	6.36	2	2	2	2	2	2
20	0.81	1	1	2	1	1	2
21	9.67	2	2	1	2	1	1
22	7.74	2	2	2	1	1	2
23	1.45	1	1	2	2	2	1
24	1.22	1	1	2	1	2	1
25	-0.52	1	1	2	2	2	2

Bayesian methods

- Example
  - The Bayesian approach was applied, fitting single marker effects
  - X matrix
    - Number of copies of two allele for each animal, eg. 2 1 0 for animal 1.

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, μ
    - g1=0.01, g2=0.01,g3=0.01, μ=0.1

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, μ
    - g1=0.01, g2=0.01,g3=0.01, μ=0.1
  - Step 2. For *i=1,2,3*, draw from $P(\sigma_{gi}^2|g_i)$

$$\chi^{-2}_{(4.012+n_i,0.002+\mathbf{g_i}'\mathbf{g_i})}$$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, μ
    - g1=0.01, g2=0.01,g3=0.01, μ=0.1
  - Step 2. For *i=1,2,3*, draw from P(σ_{gi}^2|g_i)

$$\boxed{\chi^{-2}_{(4.012+1,0.002+0.001)}}$$

- σ_{g1}^2=0.002, σ_{g2}^2=0.06, σ_{g3}^2=0.009

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, μ
    - g1=0.01, g2=0.01,g3=0.01, μ=0.1
  - Step 2. For *i=1,2,3*, draw from P(σ_{gi}^2|g_i)
    - σ_{g1}^2=0.002, σ_{g2}^2=0.06, σ_{g3}^2=0.009
  - Step 3. Draw a sample from P(σ_e^2|e)

$$\chi^{-2}_{(n-2,\mathbf{e}'\mathbf{e})}$$

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{g} - 1_n \mu$$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, μ
    - g1=0.01, g2=0.01,g3=0.01, $\mu=0.1$
  - Step 2. For *i=1,2,3*, draw from $P(\sigma_{gi}^2|g_i)$
    - $\sigma_{g1}^2=0.002, \sigma_{g2}^2=0.06, \sigma_{g3}^2=0.009$
  - Step 3. Draw a sample from $P(\sigma_e^2|e)$

$$\boxed{\chi^{-2}_{(23,812.031)}}$$

- $\sigma_e^2= 53.38$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, μ
    - g1=0.01, g2=0.01,g3=0.01, μ=0.1
  - Step 2. For *i=1,2,3*, draw from $P(\sigma_{gi}^2 | g_i)$
    - $\sigma_{g1}^2=0.002$, $\sigma_{g2}^2=0.06$, $\sigma_{g3}^2=0.009$
  - Step 3. Draw a sample from $P(\sigma_e^2 | e)$
    - $\sigma_e^2= 53.38$
  - Step 4. Draw a sample from $P(\mu | y,g,\sigma_e^2)$

$$\boxed{N\left(\frac{1}{n}\left(\mathbf{1}_{\mathbf{n}}^{'}\mathbf{y}-\mathbf{1}_{\mathbf{n}}^{'}\mathbf{Xg}\right),\sigma_e^2/n\right)}$$

    - $\mu=3.25$

Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, μ
    - g1=0.01, g2=0.01,g3=0.01, μ=0.1
  - Step 2. For *i=1,2,3*, draw from $P(\sigma_{gi}^2|g_i)$
    - $\sigma_{g1}^2=0.002, \sigma_{g2}^2=0.06, \sigma_{g3}^2=0.009$
  - Step 3. Draw a sample from $P(\sigma_e^2|e)$
    - $\sigma_e^2= 53.38$
  - Step 4. Draw a sample from $P(\mu|y,g,\sigma_e^2)$
    - $\mu=3.25$
  - Step 5. Draw a sample from $P(g_{ij}|y,\mu,\mathbf{g}\neq ij,\sigma_{gi}^2,\sigma_e^2)$

$$N\left(\frac{\mathbf{X}_{\mathbf{ij}}'\mathbf{y}-\mathbf{X}_{\mathbf{ij}}'\mathbf{X}\mathbf{g}_{(\mathbf{ij}=\mathbf{0})}-\mathbf{X}_{\mathbf{ij}}'\mathbf{1}_{\mathbf{n}}\mu}{\mathbf{X}_{\mathbf{ij}}'\mathbf{X}_{\mathbf{ij}}+\sigma_e^2/\sigma_{gi}^2},\sigma_e^2/\left(\mathbf{X}_{\mathbf{ij}}'\mathbf{X}_{ij}+\sigma_e^2/\sigma_{gi}^2\right)\right)$$

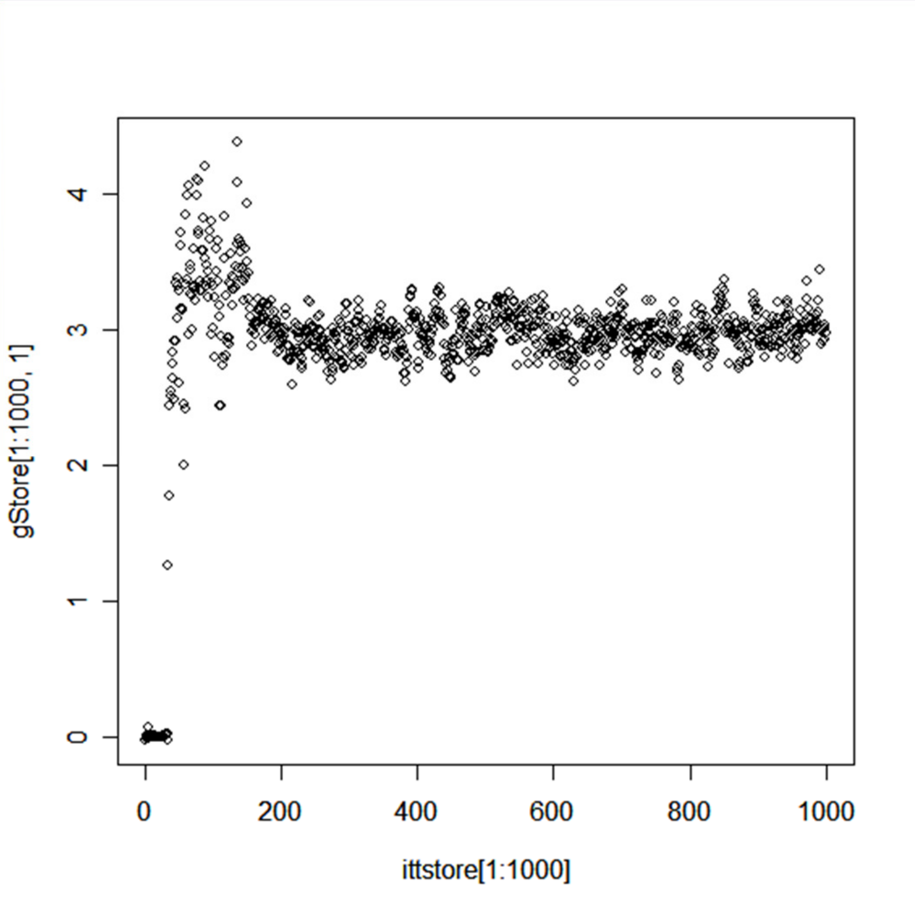
Bayesian methods

- The Gibbs chain
  - Step 1. Initialise value of **g**, μ
    - g1=0.01, g2=0.01,g3=0.01, μ=0.1
  - Step 2. For *i=1,2,3*, draw from $P(\sigma_{gi}^2|g_i)$
    - $\sigma_{g1}^2=0.002$, $\sigma_{g2}^2=0.06$, $\sigma_{g3}^2=0.009$
  - Step 3. Draw a sample from $P(\sigma_e^2|e)$
    - $\sigma_e^2= 53.38$
  - Step 4. Draw a sample from $P(\mu|y,g, \sigma_e^2,e)$
    - $\mu=3.25$
  - Step 5. Draw a sample from $P(g_{ij}|y,\mu,\mathbf{g}\neq ij,\sigma_{gi}^2,\sigma_e^2)$
    - g1=-0.02, g2=-0.81,g3=-0.005

Bayesian methods

- Gibbs chain for 1000 cycles

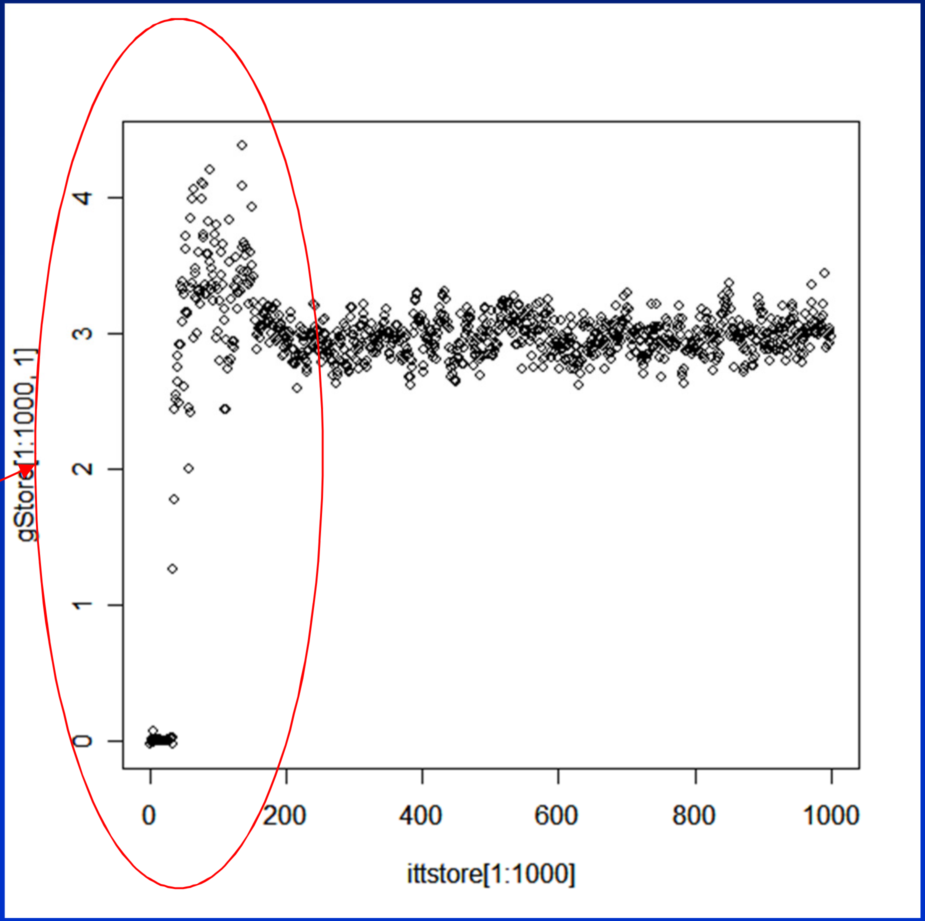
  - $P(g_1|y,\mu,\mathbf{g}\neq 1,\sigma_{g1}^2,\sigma_e^2)$



Bayesian methods

- Gibbs chain for 1000 cycles
  - $P(g_1 | y, \mu, \mathbf{g} \neq 1, \sigma_{g1}^2, \sigma_e^2)$

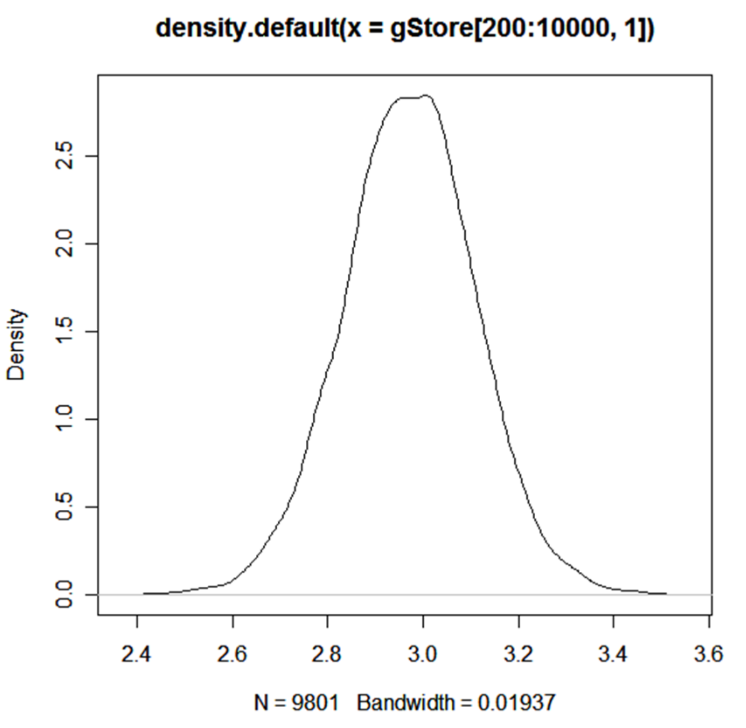
"Burn in"



Bayesian methods

- Gibbs chain for 1000 cycles
  - $P(g_1 | y,\mu,\mathbf{g}\neq 1,\sigma_{g1}^2,\sigma_e^2)$

$$\hat{g}_1 = 2.97$$



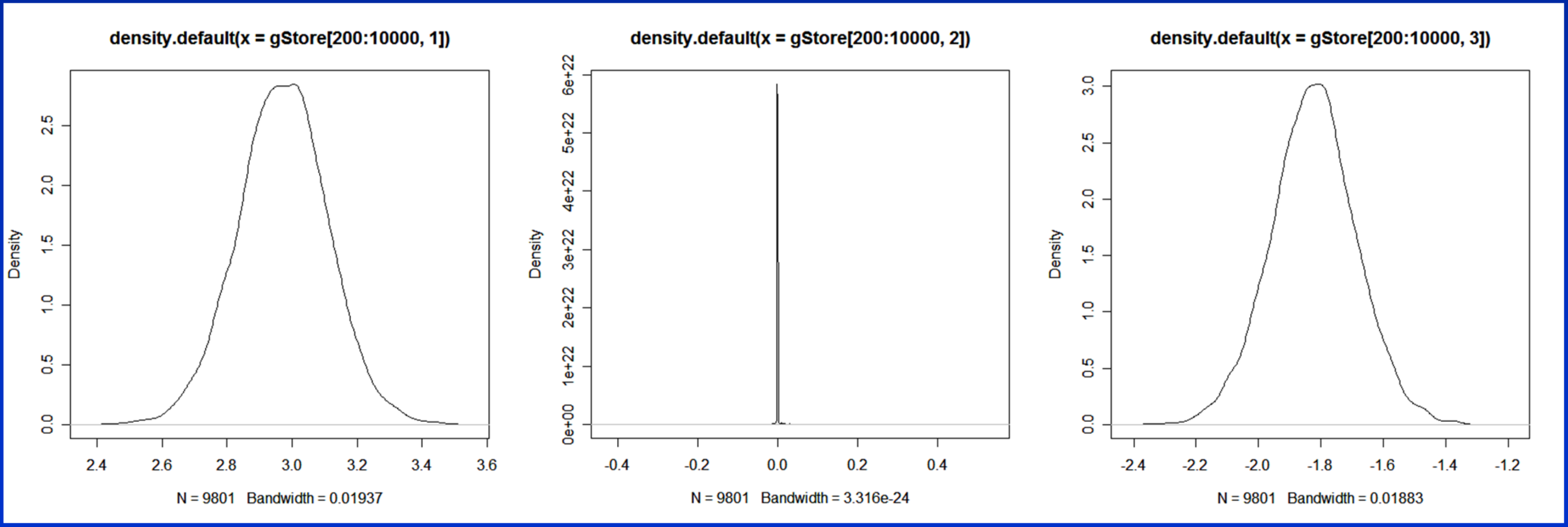
Bayesian methods

- Gibbs chain for 1000 cycles

$\hat{g}_1 = 2.97$

$\hat{g}_2 = 0.002$

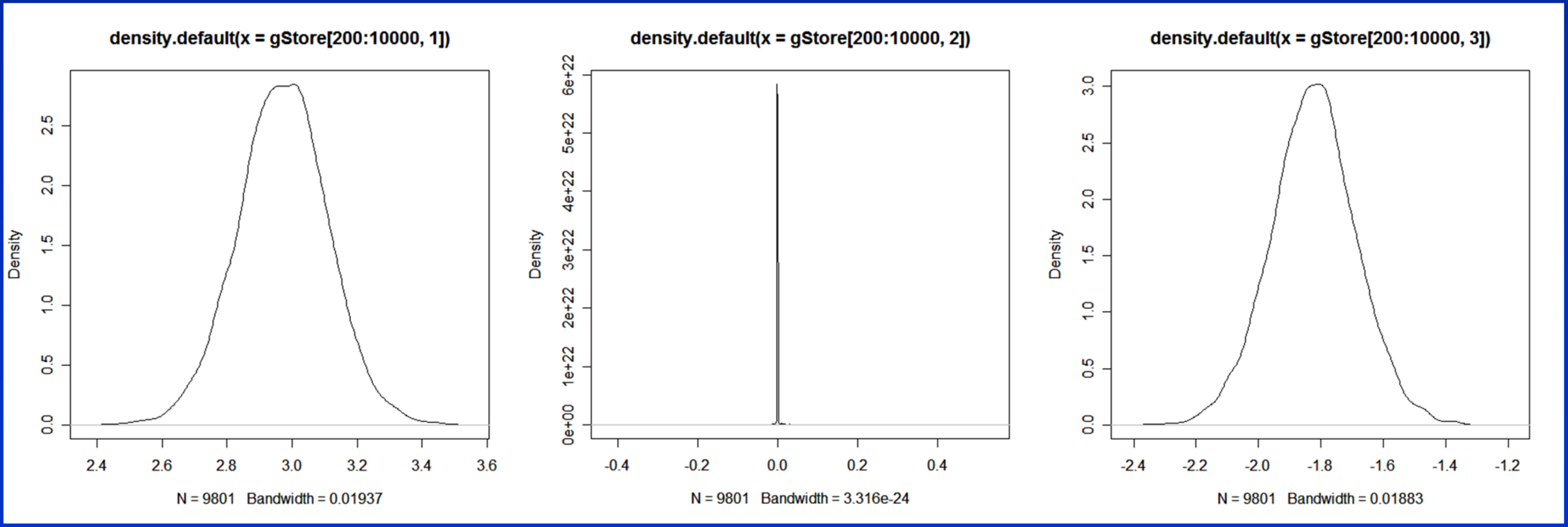
$\hat{g}_1 = -1.81$



Bayesian methods

Vector of SNP effects for calculating GEBV

$\hat{g}_1 = 2.97$ $\hat{g}_2 = 0.002$ $\hat{g}_1 = -1.81$



Bayesian methods

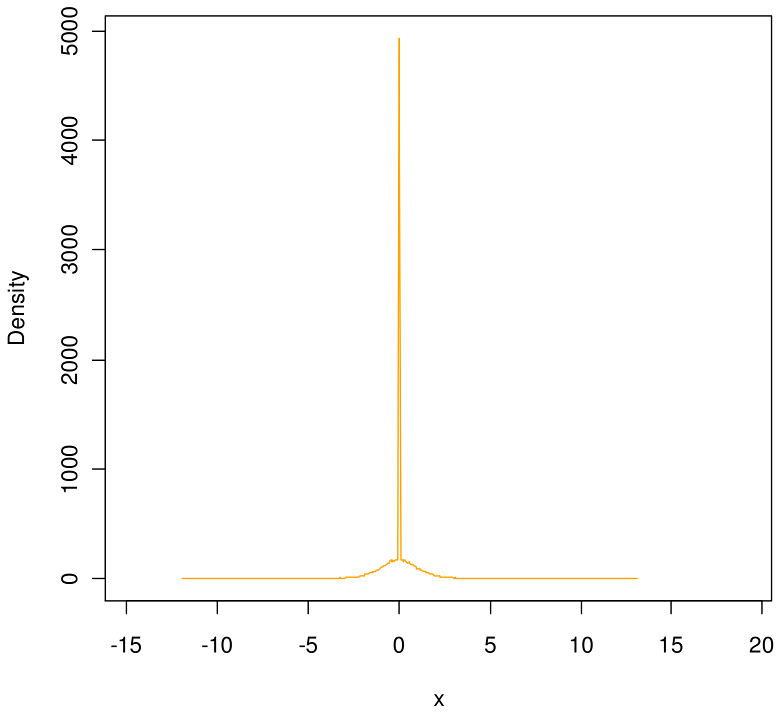
- Alternative priors for variance of segment haplotype/snp effects
  - Meuwissen BayesA $\chi^{-2}_{(4.012,0.002)}$ $\chi^{-2}_{(4.012+n_i,0.002+\mathbf{g_i}'\mathbf{g_i})}$
  - Xu (2003) $\chi^{-2}_{(0,0)}$ $\chi^{-2}_{(1,\mathbf{g}'\mathbf{g})}$
    - Uninformative
  - Te Braak (2006) $p(\sigma^2_{gi}) \propto (\sigma^2_{gi})^{-1+\alpha}$ $g_i'g_i/\chi^{-2}_{1-2a}$
  - Meuwissen BayesB

$$\begin{aligned}\sigma^2_{gi} &= 0 \text{ with probability } \pi,\\ \sigma^2_{gi} &\sim \chi^{-2}(\nu,\ S) \text{ with probability } (1-\pi),\end{aligned}$$

Bayesian methods

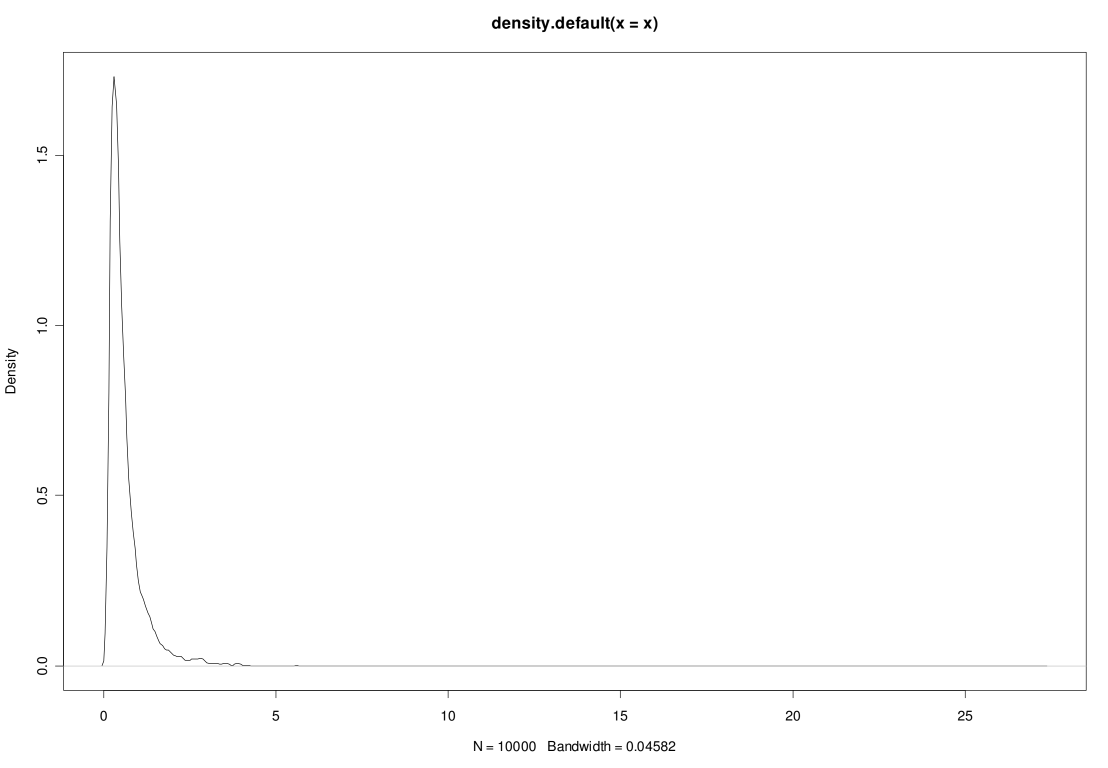
- Meuwissen BayesB
  - BayesA prior information is many QTL with small effects and few with moderate effects
  - But we have more prior knowledge than this – some chromosome segments will have no effect at all (*contain no QTL*)
    - $\sigma_{gi}^2=0, g_i=0$
  - How to sample from the posterior?

$$\sigma_{gi}^2 = 0 \text{ with probability } \pi,$$
$$\sigma_{gi}^2 \sim \chi^{-2}(\nu,\ S) \text{ with probability } (1-\pi),$$



Bayesian methods

- Meuwissen BayesB
  - If we sample σ_{gi}^2 from $\chi^{-2}_{(4.012+n_i,0.002+\mathbf{g_i}'\mathbf{g_i})}$
  - We will never sample 0, as the distribution has no mass at zero.

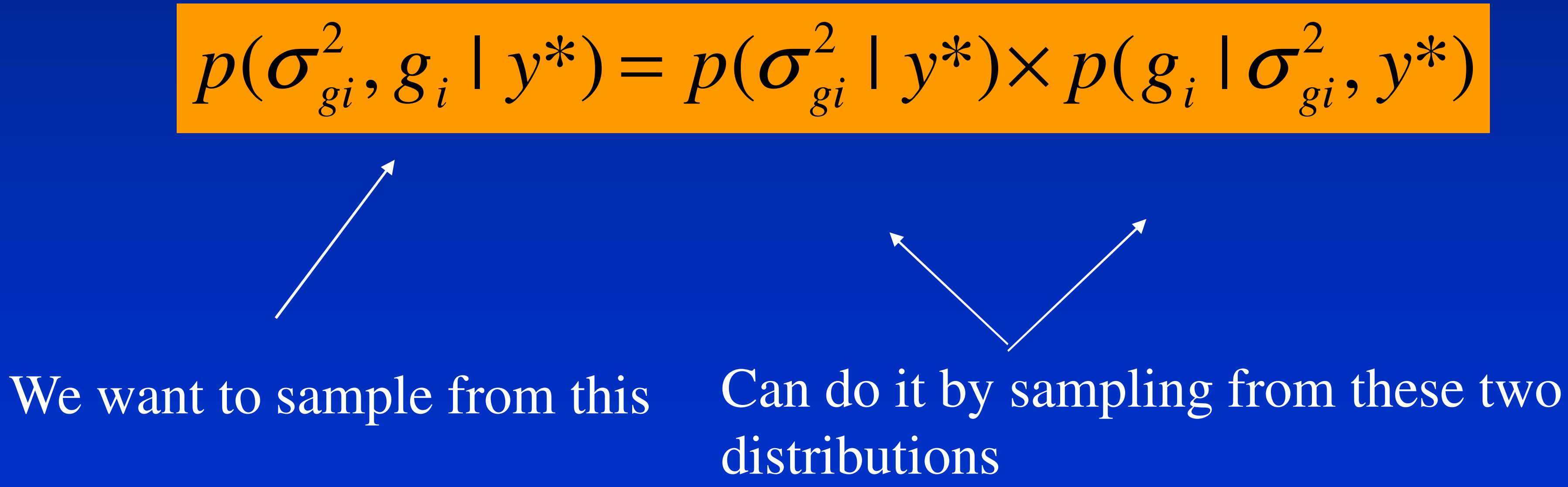


Bayesian methods

- Meuwissen BayesB
  - If we sample σ_{gi}^2 from $\chi^{-2}_{(4.012+n_i,0.002+\mathbf{g_i}'\mathbf{g_i})}$
  - We will never sample 0 if $g_i'g_i>0$, as the distribution has no mass at zero.
  - But if $\sigma_{gi}^2>0$, then sampling $g_i = 0$ has infinitesimal (basically zero) probability

Bayesian methods

- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:



$$p(\sigma_{gi}^2, g_i \mid y^*) = p(\sigma_{gi}^2 \mid y^*) \times p(g_i \mid \sigma_{gi}^2, y^*)$$

We want to sample from this

Can do it by sampling from these two distributions

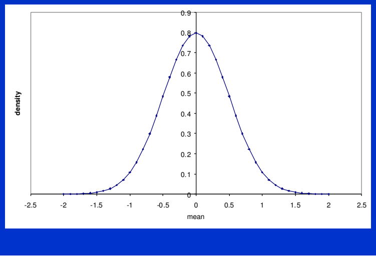
Bayesian methods

- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:

$$p(\sigma^2_{gi}, g_i \mid y^*) = \boxed{p(\sigma^2_{gi} \mid y^*)} \times p(g_i \mid \sigma^2_{gi}, y^*)$$

We want to sample from this

$P(g_i|y,\mu,g,\sigma_{gi}^2,\sigma_e^2)$



Bayesian methods

- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:

$$p(\sigma_{gi}^2, g_i \mid y^*) = \boxed{p(\sigma_{gi}^2 \mid y^*)} \times p(g_i \mid \sigma_{gi}^2, y^*)$$

??

Sample σ_{gi}^2 without conditioning on g_i

Bayesian methods

- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:

$$p(\sigma_{gi}^2, g_i \mid y^*) = \enclose{circle}{p(\sigma_{gi}^2 \mid y^*)} \times p(g_i \mid \sigma_{gi}^2, y^*)$$

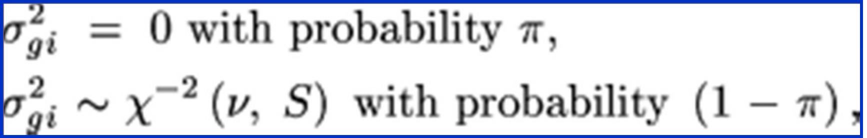
  - Cannot be expressed as a known distribution = cannot use Gibbs for this bit
  - Use a Metropolis Hastings algorithm

Bayesian methods

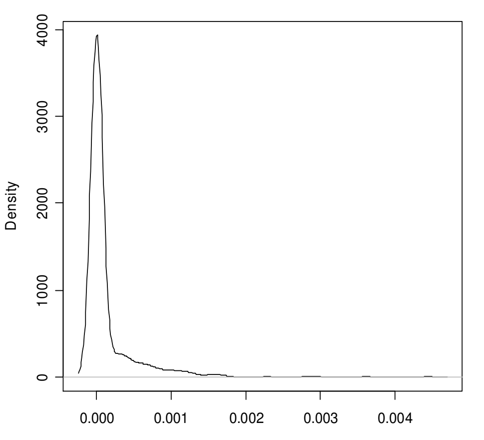
- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:

$$p(\sigma_{gi}^2, g_i \mid y^*) = \boxed{p(\sigma_{gi}^2 \mid y^*)} \times p(g_i \mid \sigma_{gi}^2, y^*)$$

  - Step1 Sample $\sigma_{g\_new}^2$,from prior($\sigma_{g\_new}^2$)



$$\begin{aligned} \sigma_{gi}^2 &= 0 \text{ with probability } \pi, \\ \sigma_{gi}^2 &\sim \chi^{-2}(\nu,\ S) \text{ with probability } (1-\pi), \end{aligned}$$

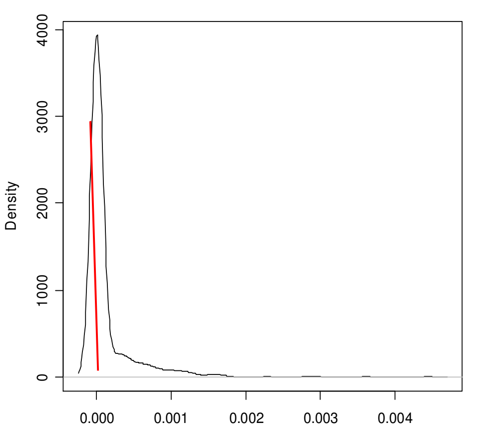


Bayesian methods

- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:

$$p(\sigma_{gi}^2, g_i \mid y^*) = \boxed{p(\sigma_{gi}^2 \mid y^*)} \times p(g_i \mid \sigma_{gi}^2, y^*)$$

  - Step1 Sample $\sigma_{g\_new}^2$,from prior($\sigma_{g\_new}^2$)
    - $\sigma_{g\_new}^2=0$

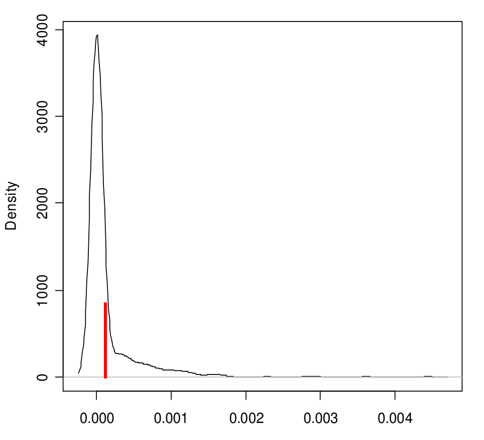


Bayesian methods

- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:

$$p(\sigma^2_{gi}, g_i \mid y^*) = \boxed{p(\sigma^2_{gi} \mid y^*)} \times p(g_i \mid \sigma^2_{gi}, y^*)$$

  - Step1 Sample $\sigma_{g\_new}^2$,from prior($\sigma_{g\_new}^2$)
    - $\sigma_{g\_new}^2$=0.5



Bayesian methods

- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:

$$p(\sigma_{gi}^2, g_i \mid y^*) = \boxed{p(\sigma_{gi}^2 \mid y^*)} \times p(g_i \mid \sigma_{gi}^2, y^*)$$

  - Step 1 Sample $\sigma_{g\_new}^2$,from prior($\sigma_{g\_new}^2$)
  - Step 2 Evaluate p(y*| $\sigma_{g\_new}^2$) (Likelihood)

$$L(\mathbf{y}^* \mid \sigma_{ginew}^2 = \frac{1}{2\pi^{1/2n} \mid \mathbf{V} \mid^{1/2}} e(-0.5 * (\mathbf{y}^{*\prime} \mathbf{V}^{-1} \mathbf{y}^*))$$

$$\mathbf{V} = \mathbf{X}(\mathbf{I}\sigma_{ignew}^2)\mathbf{X}' + \mathbf{I}\boldsymbol{\sigma}_{\mathbf{e}}^2)$$

Bayesian methods

- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:

$$p(\sigma_{gi}^2, g_i \mid y^*) = p(\sigma_{gi}^2 \mid y^*) \times p(g_i \mid \sigma_{gi}^2, y^*)$$

  - Step 1 Sample $\sigma_{g\_new}^2$,from prior($\sigma_{g\_new}^2$)
  - Step 2 Evaluate p(y*| $\sigma_{g\_new}^2$) (Likelihood)
  - Step 3 Replace σ_{gi}^2 with $\sigma_{g\_new}^2$ probability min[p(y*| $\sigma_{g\_new}^2$)/ p(y*| σ_{gi}^2):1]

Bayesian methods

- Meuwissen BayesB
  - Solution: sample σ_{gi}^2,g_i simultaneously from the distribution:

$$p(\sigma_{gi}^2, g_i \mid y^*) = p(\sigma_{gi}^2 \mid y^*) \times p(g_i \mid \sigma_{gi}^2, y^*)$$

  - Step 1 Sample $\sigma_{g\_new}^2$,from prior($\sigma_{g\_new}^2$)
  - Step 2 Evaluate $p(y^*| \sigma_{g\_new}^2)$ (Likelihood)
  - Step 3 Replace σ_{gi}^2 with $\sigma_{g\_new}^2$ probability $\min[p(y^*| \sigma_{g\_new}^2)/ p(y^*| \sigma_{gi}^2):1]$
  - Step 4 Repeat ~ 100 cycles

Genomic selection

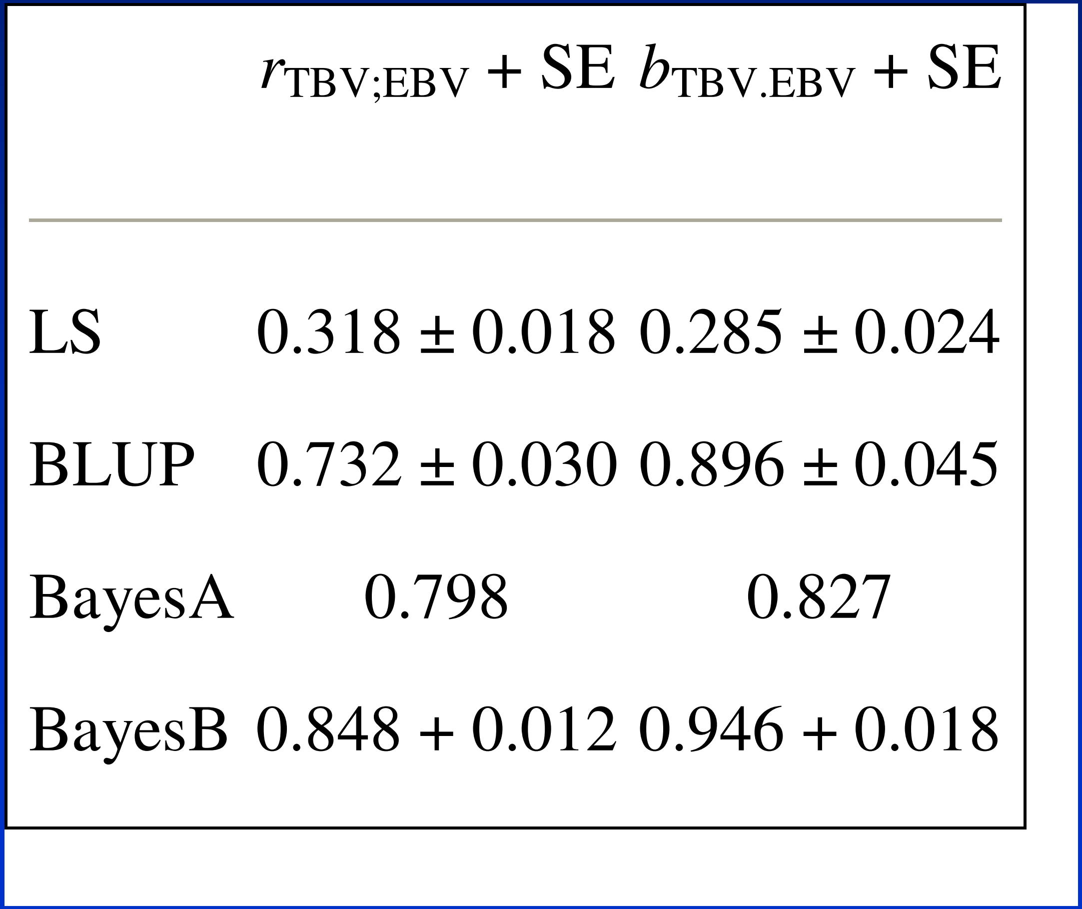
- Introduction
- Genomic selection with Least Squares and BLUP
- Introduction to Bayesian methods
- Genomic selection with Bayesian methods
- Comparison of accuracy of methods

Genomic selection

- Comparison of accuracy of methods (Meuwissen et al. 2001)
    - Genome of 1000 cM simulated, marker spacing of 1 cM.
    - Markers surrounding each 1-cM region combined into haplotypes.
    - Due to finite population size (Ne = 100), marker haplotypes were in linkage disequilibrium with QTL between markers.
    - Effects of haplotypes predicted in one generation of 2000 animals
    - Breeding values for progeny of these animals predicted based on marker genotypes

Genomic selection

- Comparison of accuracy of methods (Meuwissen et al. 2001)



	$r_{TBV;EBV}$ + SE	$b_{TBV.EBV}$ + SE
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
BayesA	0.798	0.827
BayesB	0.848 + 0.012	0.946 + 0.018

Genomic selection

- Comparison of accuracy of methods (Meuwissen et al. 2001)
  - The least squares method does very poorly, primarily because the haplotype effects are over-estimated.

Genomic selection

- Comparison of accuracy of methods (Meuwissen et al. 2001)
  - The least squares method does very poorly, primarily because the haplotype effects are over-estimated.
  - Increased accuracy of the Bayesian approach because method sets many of the effects of the chromosome segments close to zero in BayesA, or zero in BayesB

Genomic selection

- Comparison of accuracy of methods (Meuwissen et al. 2001)
  - The least squares method does very poorly, primarily because the haplotype effects are over-estimated.
  - Increased accuracy of the Bayesian approach because method sets many of the effects of the chromosome segments close to zero in BayesA, or zero in BayesB
  - Also “shrinks” estimates of effects of other chromosome segments based on a prior distribution of QTL effects.

Genomic selection

- Comparison of accuracy of methods (Meuwissen et al. 2001)
  - The least squares method does very poorly, primarily because the haplotype effects are over-estimated.
  - Increased accuracy of the Bayesian approach because method sets many of the effects of the chromosome segments close to zero in BayesA, or zero in BayesB
  - Also “shrinks” estimates of effects of other chromosome segments based on a prior distribution of QTL effects.
  - Accuracies were very high, as high as following progeny testing for example

In real data

- 1500 Australian dairy bulls
- genotyped for 56000 genome wide SNPs
- Phenotypes average of daughters milk production

In real data

- Split data into two sub-populations
  - Reference: Bulls born < 2003
  - Validation: Bulls born >= 2003

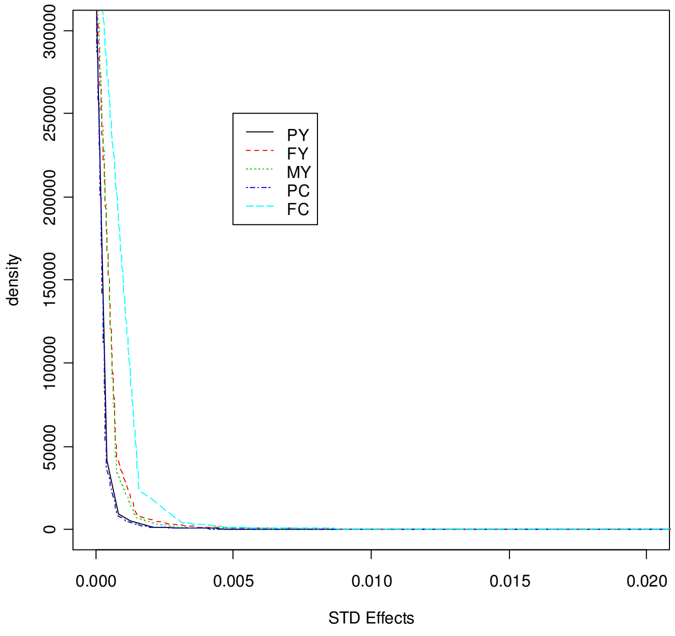
In real data

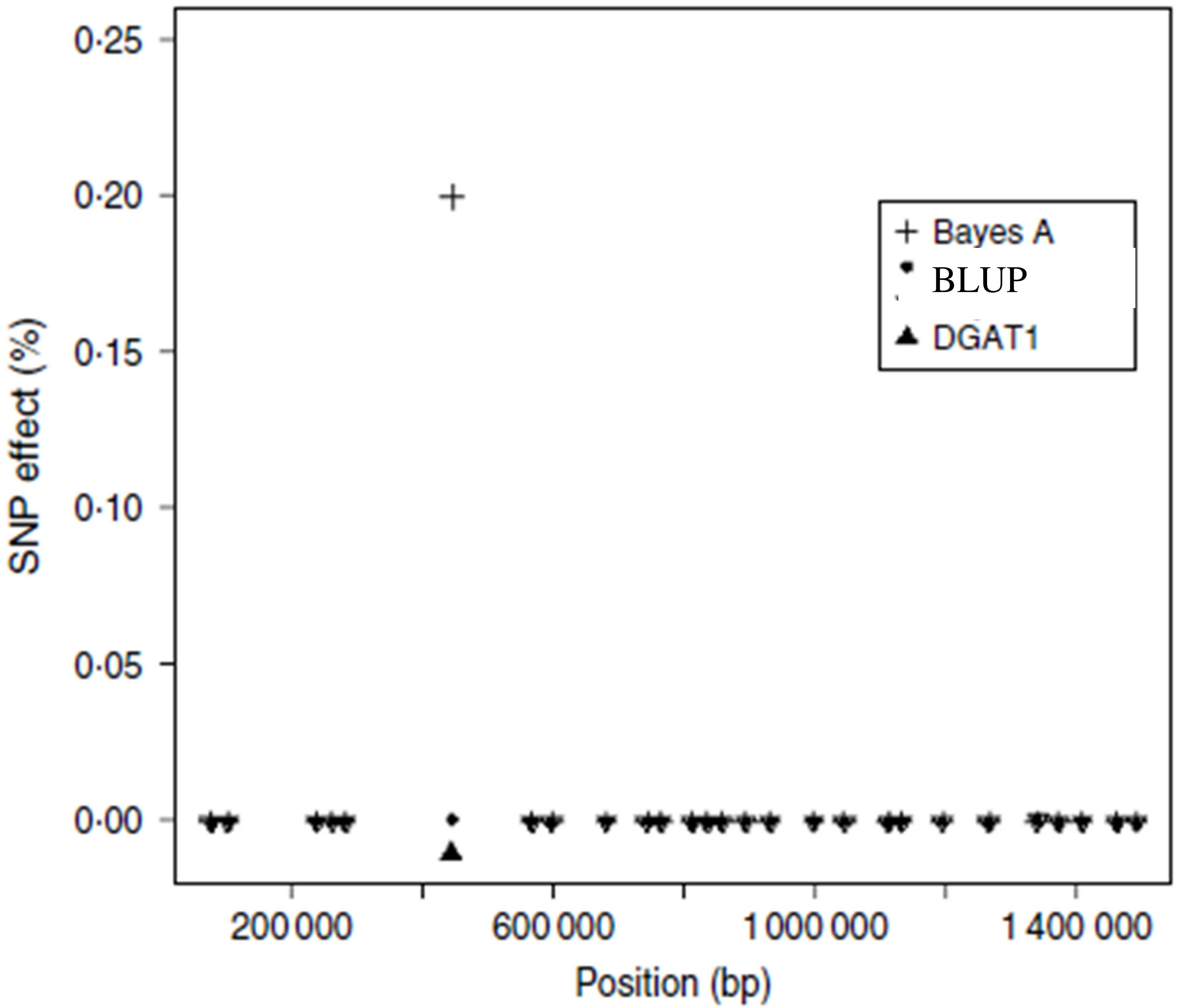
- Split data into two sub-populations
  - Reference: Bulls born < 2003
  - Validation: Bulls born >= 2003
- Accuracy
  - Correlation of genomic breeding values with EBVs (which include daughter information) in validation set

In real data

Table 3 MEBV- Correlation between predicted MEBV and ABV in the validation data set (Bulls proven in years 2005, 2006, 2007)

Method	*Protein kg*	*Fat kg*	*Protein %*	*Fat %*
Bayes B	0.55	0.51	0.68	0.73
Bayes A	0.53	0.48	0.66	0.70
BLUP	0.60	0.48	0.66	0.64




0·25
0·20
0·15
0·10
0·05
0·00
SNP effect (%)
200 000
600 000
1 000 000
1 400 000
Position (bp)
+ Bayes A
● BLUP
▲ DGAT1

Genomic selection

- Yi and Xu 2008 (Genetics)

- Sample from inverse chi square distribution, but then sample shape (v) and scale (S^2) of the distribution
  - Reflect absence of knowledge of distribution of QTL effects?
  - Prior on S^2 is uniform, then posterior is gamma

$$S^2 \mid y, \mu, \beta, \sigma^2, v \sim gamma\left(\frac{pv}{2}, \frac{v}{2}\sum_{j=1}^{p}\frac{1}{\sigma_{gj}^2}\right)$$

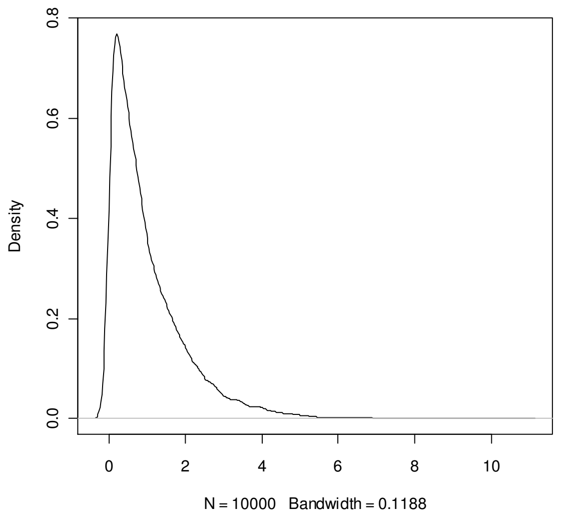
  - Prior on v of 1/v, not a conjugate prior = metropolis hastings

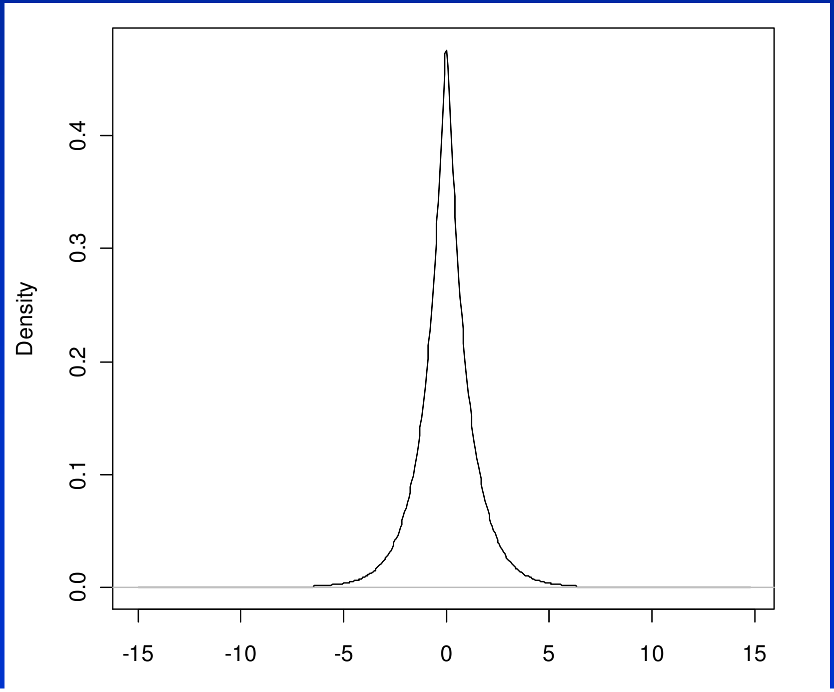
Genomic selection

- Yi and Xu 2008 (Genetics)
- Propose sampling σ_{gi}^2 from an exponential distribution (Bayesian LASSO)

Distribution of σ_{gj}^2 --> Distribution of g_j

$$P(\sigma_{gi}^2) \sim \frac{\lambda}{2} e^{-\lambda \sigma_{gi}^2 / 2}$$





Genomic selection

- Bayesian LASSO
  - $P(\sigma_{gi}^2 | g_i)$

$$\sigma_j^2 \mid y, \mu, g, \sigma_e^2, \lambda^2 \sim InvGauss(\sqrt{\frac{\lambda^2 \sigma_e^2}{g_j^2}}, \lambda^2$$

  - $P(\lambda^2 | y, \mu, g, \sigma_e^2, \sigma_g^2)$

$$\lambda^2 \mid y, \mu, g, \sigma_e^2, \sigma_j^2 \sim gamma(p + a, \sum_{j=1}^{p} \sigma_j^2 / 2 + b)$$

Genomic selection

- Bayesian C∏ (Habier et al 2011)

- Two criticisms of BayesB
  - Posterior of locus-specific variance has only one additional degree of freedom, compared to its prior regardless of the number of genotypes, so
    - Degree of shrinkage of depends strongly on prior
    - Little information coming from data
  - ∏ is treated as known, not estimated from the data

Genomic selection

- Bayesian CΠ (Habier et al 2011)

- Use a common σ_{gi}^2 across all SNP
  - Many degrees of freedom from data
  - A "BLUP" for SNP in model
- Estimate Π from data
  - Sample from
    - Beta(K - m(t) + 1, m(t) + 1).
    - Where K is number of SNP, m(t) is the number of SNP in the model at iteration t (eg. Those not set to zero)

Genomic selection

- Bayesian Cπ (Habier et al 2011)
  - Accuracy in German Holstein Friesian data set

Trait	**GBLUP**	**BayesA**	**BayesB**	**BayesCpi**
Milk Yield	0.48	0.48	0.40	0.43
Fat Yield	0.51	0.56	0.52	0.54
Protein Yield	0.21	0.22	0.17	0.21
Somatic cells	0.17	0.17	0.12	0.14

- Little improvement in accuracy
- But can draw inferences about trait architecture?

Genomic selection

- Methods for deriving prediction equation differ in assumptions about distribution of QTL effects
  - BLUP = normal distribution with known variance
  - Ridge regression = normal distribution with prior assumption about variance
  - BayesA = t-distribution, degree of shrinkage known a-priori, or sampled
  - BayesB = mixture distribution, many effects zero
  - BayesianLASSO, double exponential distribution of effects
  - Bayesian Cπ, estimate Π from data, common variance across SNP