Bayesian methods for genomic prediction

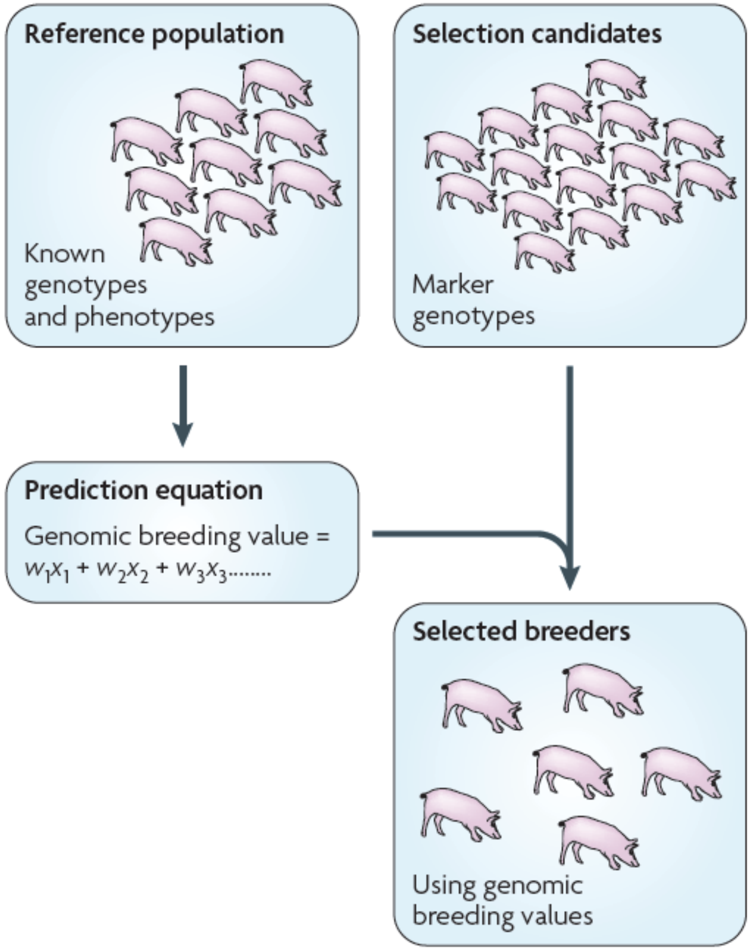
Ben Hayes

Queensland Alliance for Agriculture and Food Innovation, University of Queensland

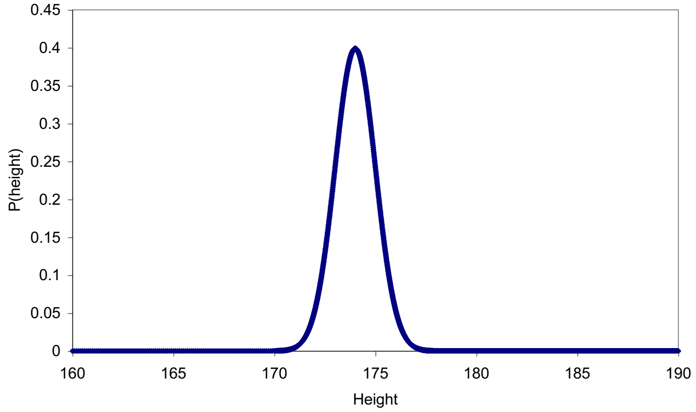
Bayesian methods for Genomic Prediction

- Alternative assumptions regarding the distribution of SNP effects
- Introduction to Bayesian methods
- Genomic prediction with Bayesian methods
- Comparison of accuracy of methods

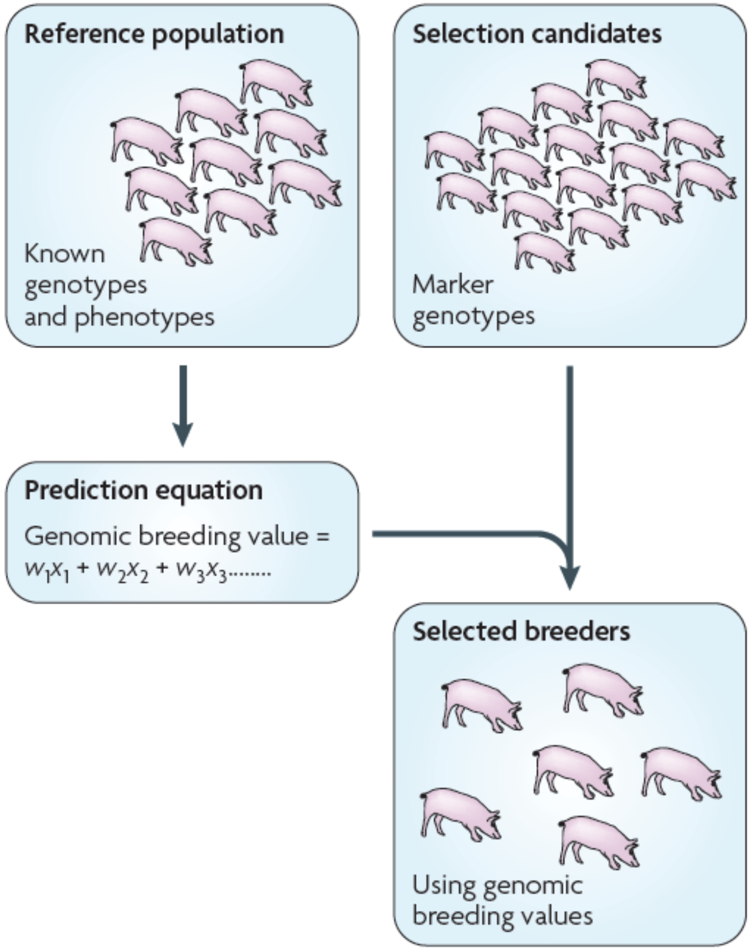
Assumptions regarding distribution of SNP effects



- Best Linear Unbiased Prediction
  - GBLUP, SNPBLUP
- GREML
- *Assumes SNP effects are:*
  - *all non-zero*
  - *very small*
  - *normally distributed*



Assumptions regarding distribution of SNP effects

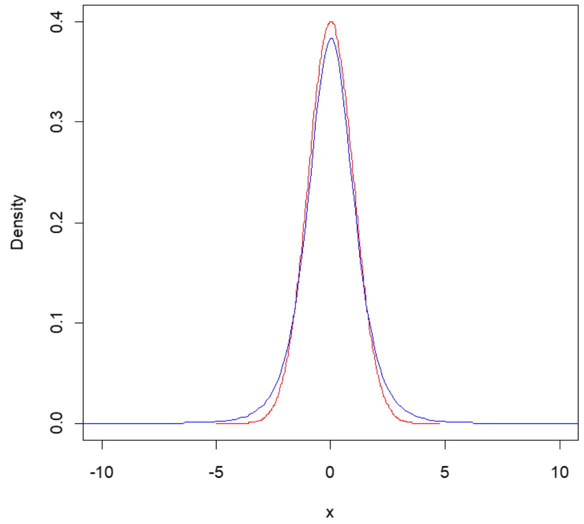


- Alternative distributions?

Assumptions regarding distribution of SNP effects

- Alternative distributions?

Assumption	Distribution of SNP effects	Method
Small number of moderate to large effects, many small effects	Students t	BayesA



Assumptions regarding distribution of SNP effects

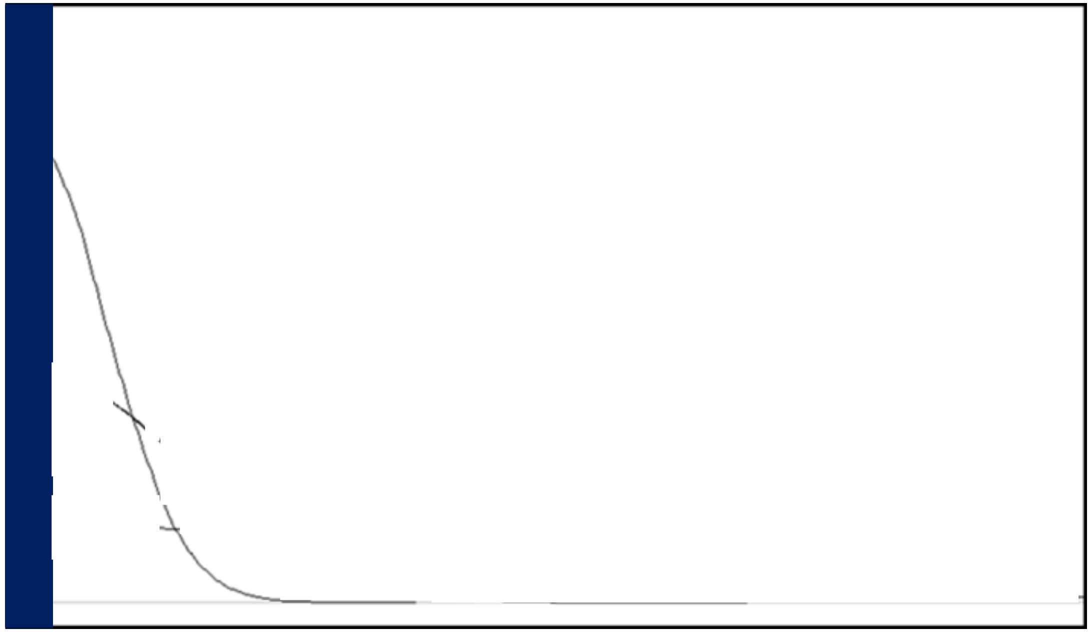
- Alternative distributions?

Assumption	Distribution of SNP effects	Method
Small number of moderate to large effects, many small effects	Students t	BayesA
Small number of moderate to large effects, many zero effects	Mixture, spike at zero, Students t	BayesB
Small number of small effects, many zero effects	Mixture, spike at zero, normal distribution	BayesC
Many zero effects, proportion of small effects, some moderate to large effects	Multi-variate normal	BayesR

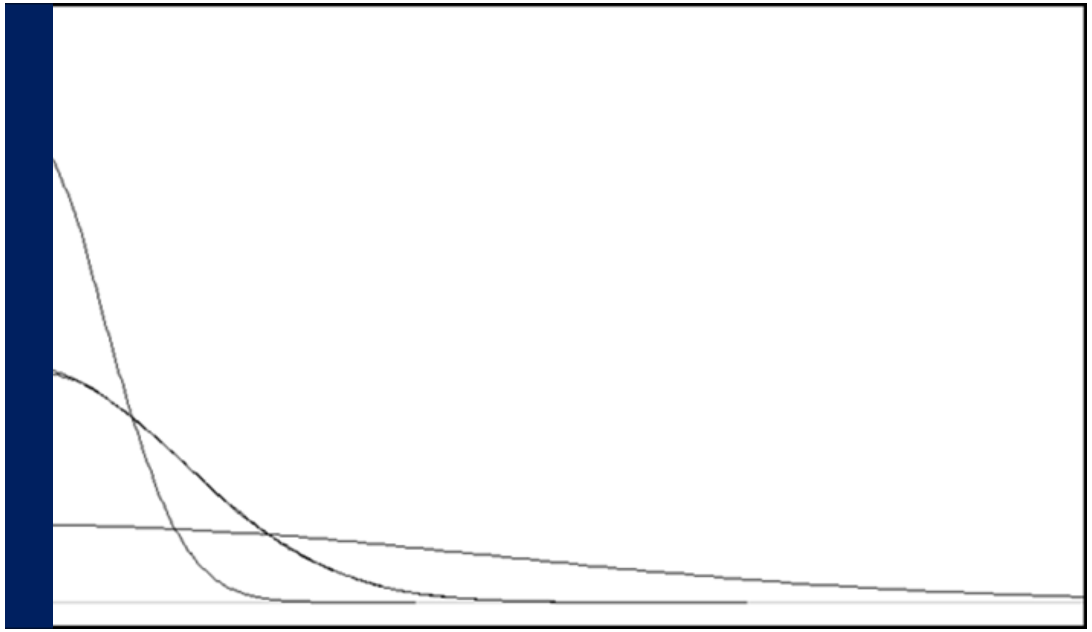
Assumptions regarding distribution of SNP effects

- Alternative distributions?

BayesC



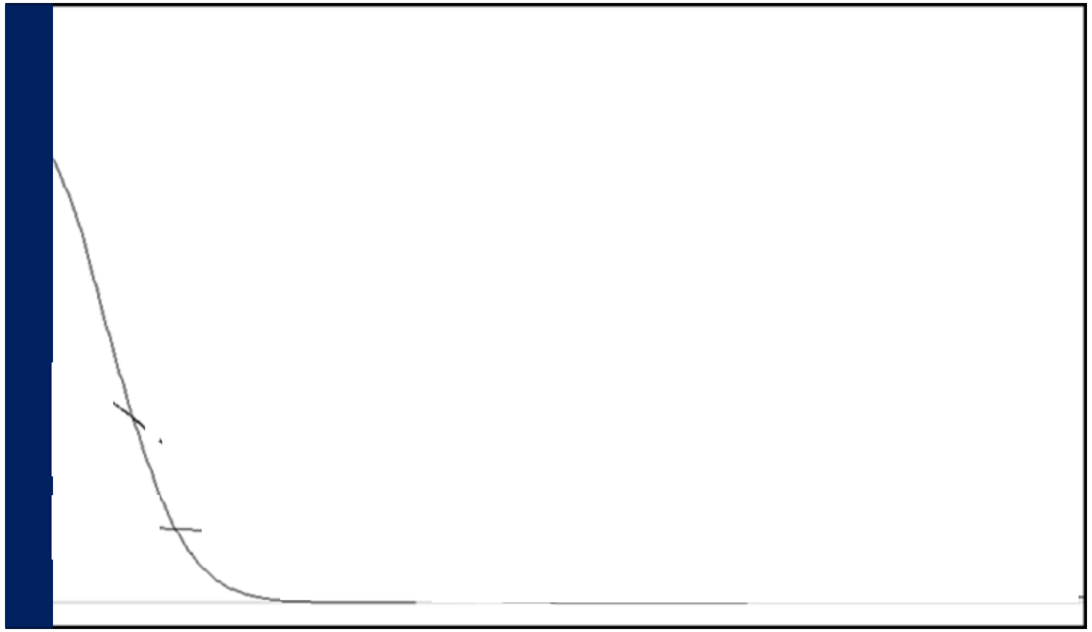
BayesR



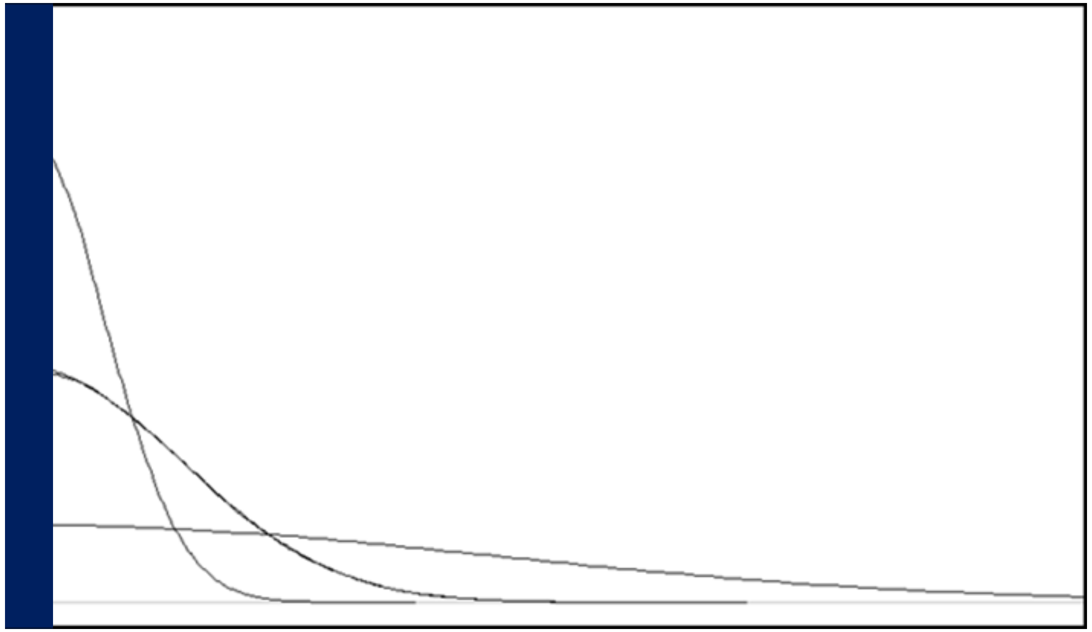
Assumptions regarding distribution of SNP effects

- Alternative distributions?

BayesC



BayesR



Bayesian approach allows us to incorporate this prior knowledge in the prediction of SNP effects

Bayesian methods for Genomic Prediction

- Alternative assumptions regarding the distribution of SNP effects
- Introduction to Bayesian methods
- Genomic prediction with Bayesian methods
- Comparison of accuracy of methods

Bayesian methods

- Bayes theorem

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

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

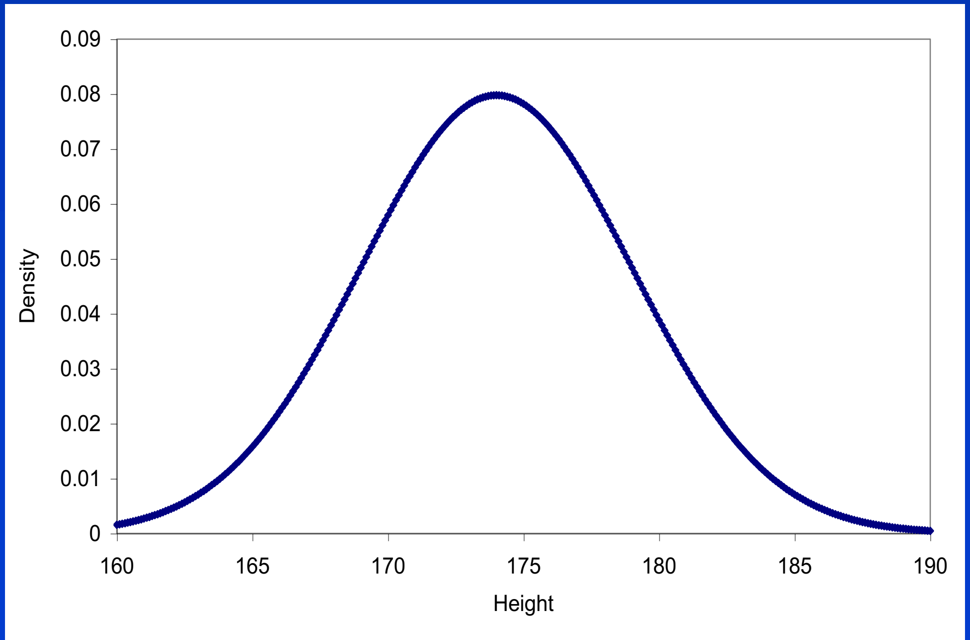
$$y = \text{average height} + e$$

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

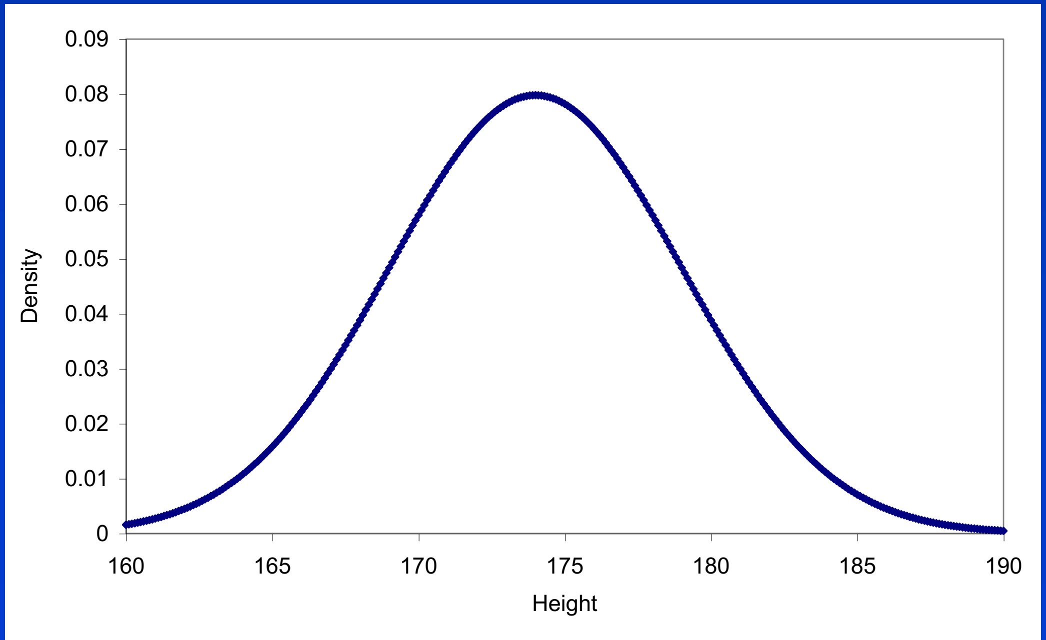
$$P(x \mid y) \propto P(y \mid x)P(x)$$

From the data……

$$\bar{x} = 178$$

$$s.e = 5$$

Prior probability of x (average height)



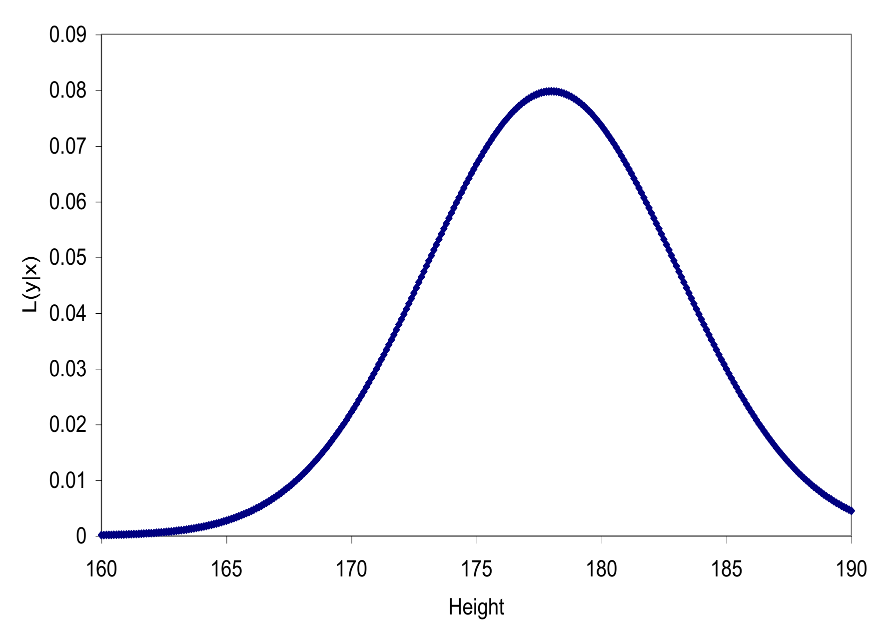
Bayesian methods

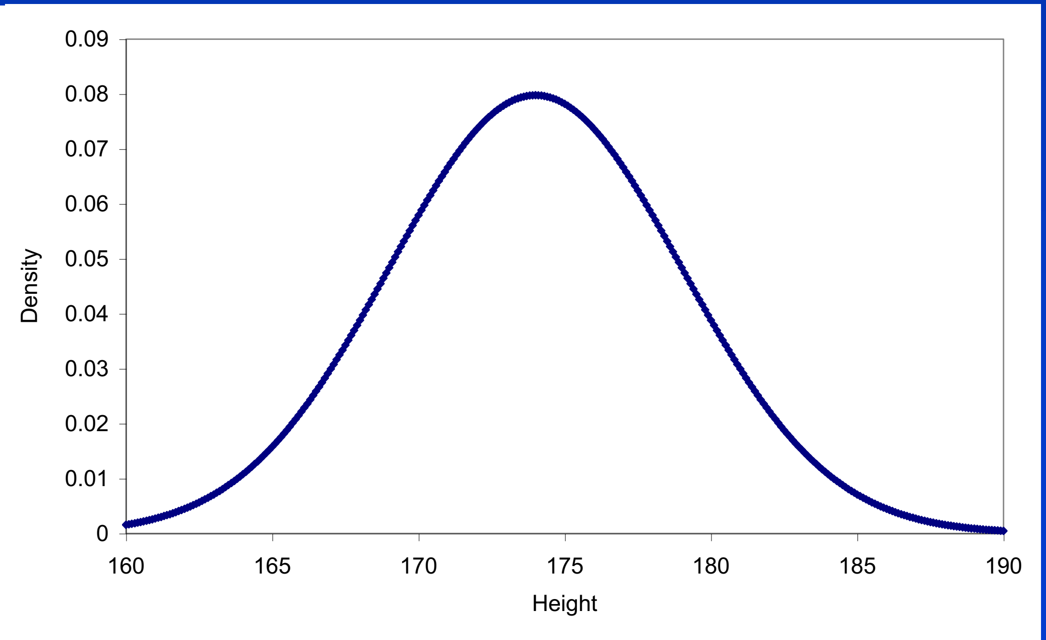
- Bayes theorem

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

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

Prior probability of x (average height)





Bayesian methods

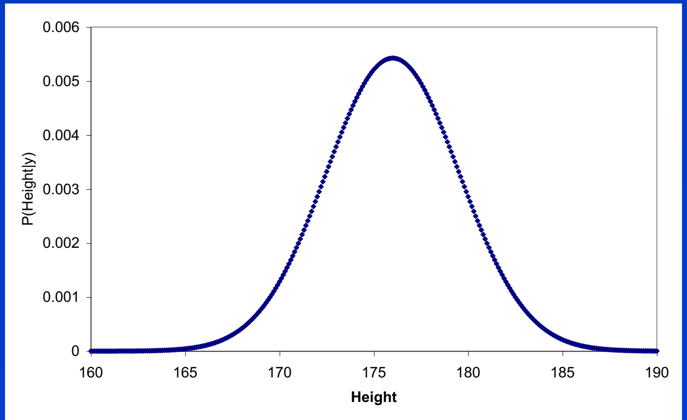
- Bayes theorem

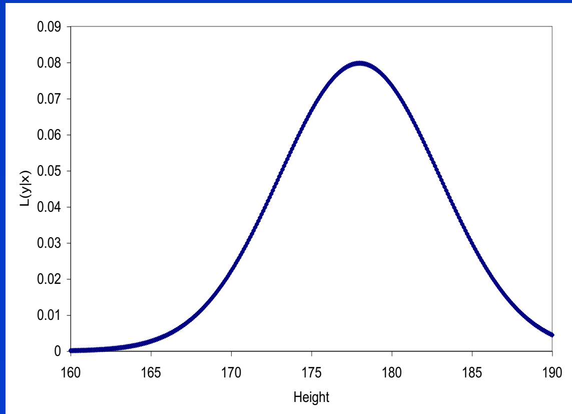
$$P(x \mid y) \propto P(y \mid x)P(x)$$

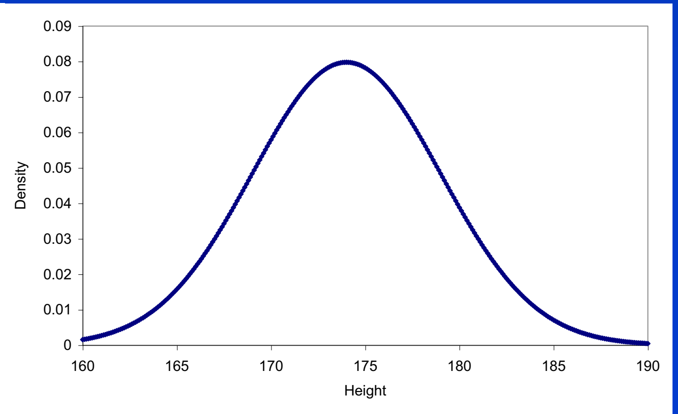
P(x|y) mean = 176cm

L(y|x)

P(x)







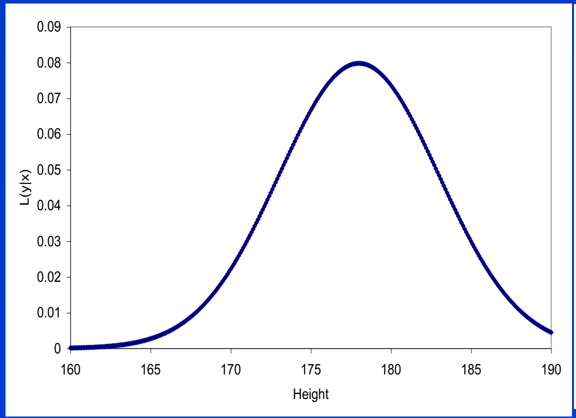
Bayesian methods

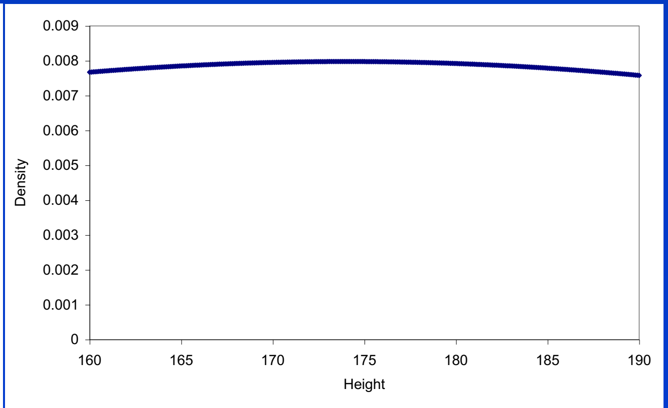
- Bayes theorem
- Less certainty about prior information? Use *less* informative (flat) prior

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

L(y|x)

P(x)





Bayesian methods

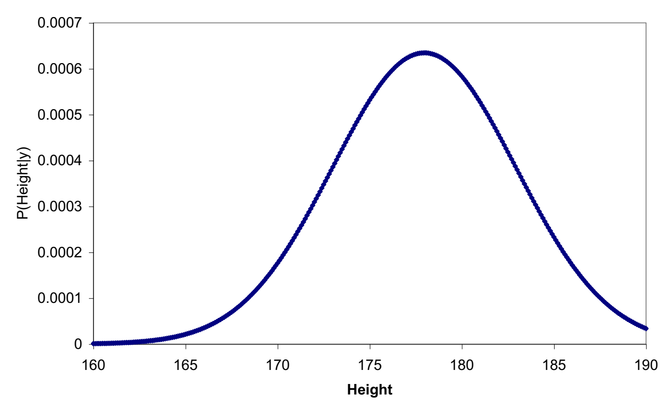
- Bayes theorem
- Less certainty about prior information? Use *less* informative (flat) prior

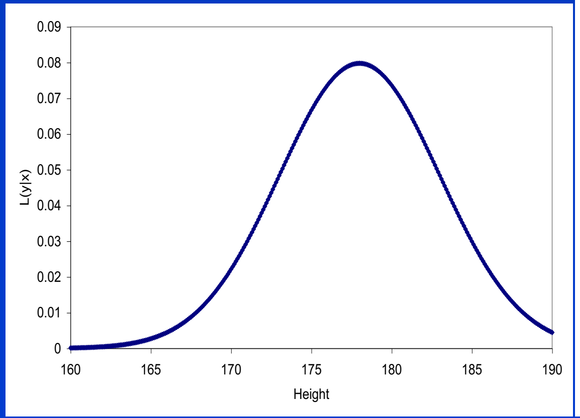
$$P(x \mid y) \propto P(y \mid x) P(x)$$

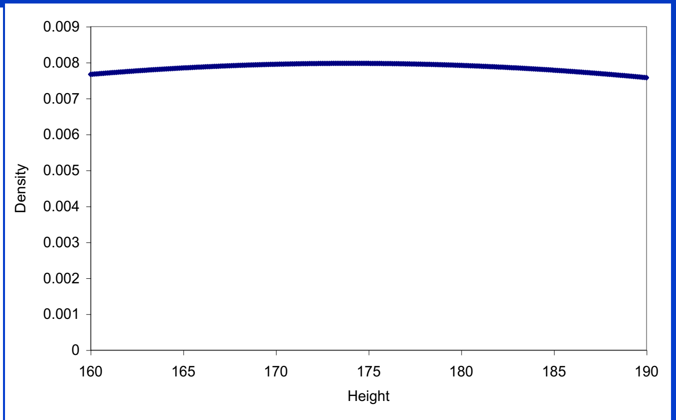
P(x|y) mean = 178cm

L(y|x)

P(x)





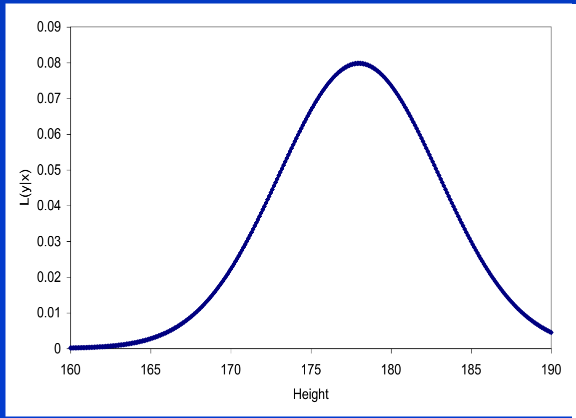


Bayesian methods

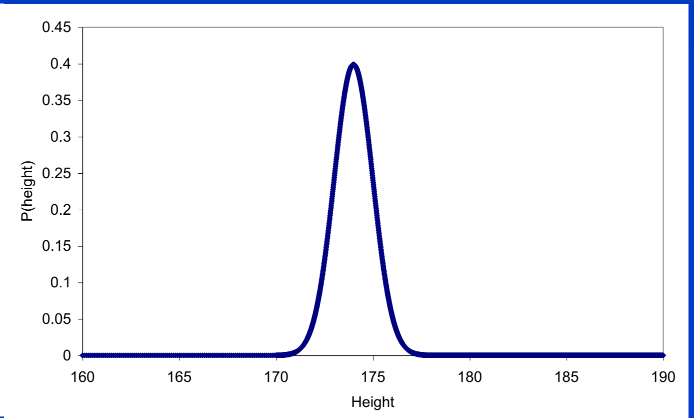
- Bayes theorem
- More certainty about prior information? Use *more* informative prior

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

L(y|x)



P(x)



Bayesian methods

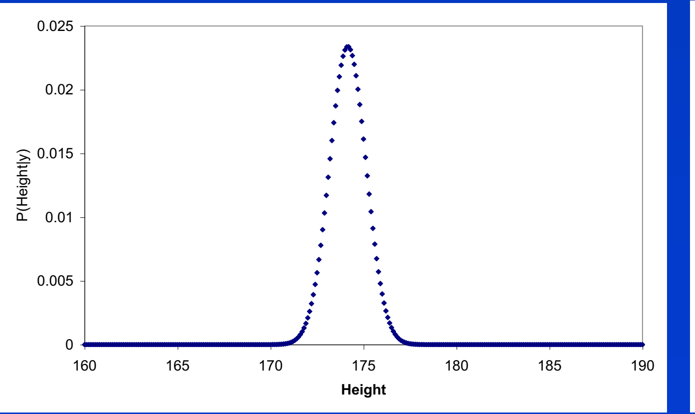
- Bayes theorem
- More certainty about prior information? Use *more* informative prior

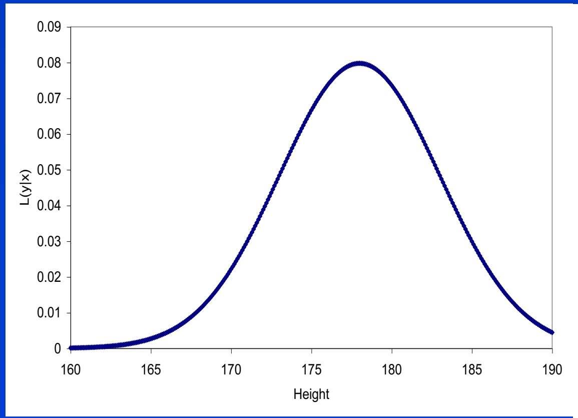
$$P(x \mid y) \propto P(y \mid x)P(x)$$

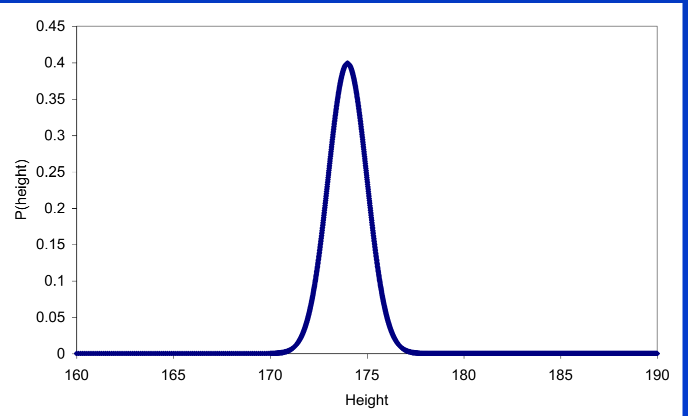
P(x|y) mean = 174.5cm

L(y|x)

P(x)







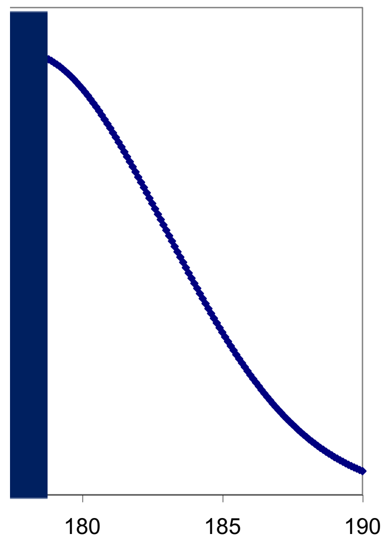
Bayesian methods for Genomic Prediction

- Alternative assumptions regarding the distribution of SNP effects
- Introduction to Bayesian methods
- Genomic prediction with Bayesian methods
- Comparison of accuracy of methods

BayesC

- $y = 1_n\mu + X\beta + e$

$$\beta_j \begin{cases} \sim N(0, \sigma_\beta^2) & \text{with probability } \pi \\ = 0 & \text{with probability } 1 - \pi \end{cases}$$



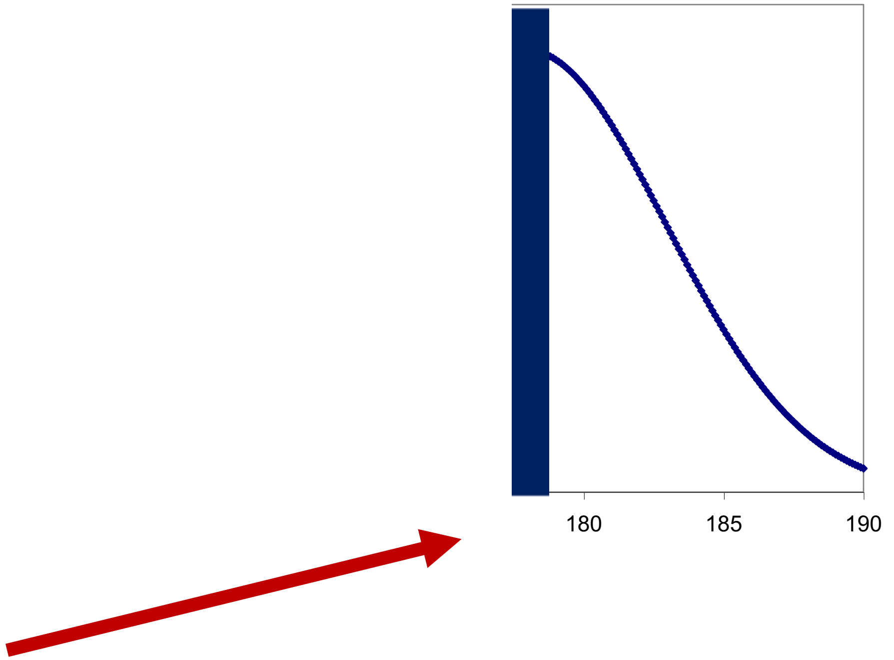
$$P(\boldsymbol{\beta}, \mu | \boldsymbol{y}) \propto P(\boldsymbol{y} | \boldsymbol{\beta}, \mu) P(\boldsymbol{\beta}, \mu)$$

BayesC

- $y = 1_n\mu + X\beta + e$

$$\beta_j \begin{cases} \sim N(0, \sigma_\beta^2) & \text{with probability } \pi \\ = 0 & \text{with probability } 1 - \pi \end{cases}$$

$$P(\boldsymbol{\beta}, \mu | \boldsymbol{y}) \propto P(\boldsymbol{y} | \boldsymbol{\beta}, \mu) P(\boldsymbol{\beta}, \mu)$$



BayesC -> Gibbs Sampling

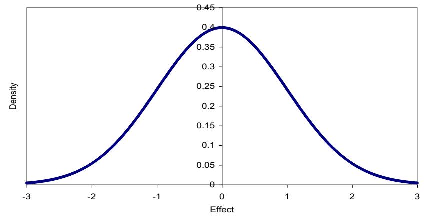
- Cannot solve directly, as estimates of parameters depend on other parameters -> no closed form solution
- For example, estimate of a SNP effect depends on whether or not the SNP is in the zero variance part of distribution or non-zero variance part of the distribution
- Use Gibbs sampling!
- Sample from posterior distribution of parameter **conditional** on all other parameters

BayesC -> Gibbs Sampling

- Sample from posterior distribution of parameter **conditional** on all other parameters

- For example, for SNP effect β_i
  - First sample if in zero effect or non zero effect part of distribution (δ_i)
  - Then if in non-zero part of the distribution, sample from

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

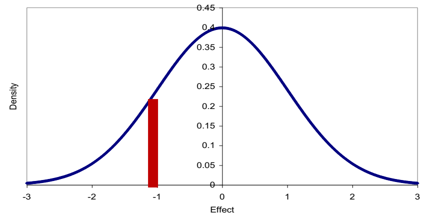


BayesC -> Gibbs Sampling

- Sample from posterior distribution of parameter **conditional** on all other parameters

- For example, for SNP effect β_i
  - First sample if in zero effect or non zero effect part of distribution (δ_i)
  - Then if in non-zero part of the distribution, sample from

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

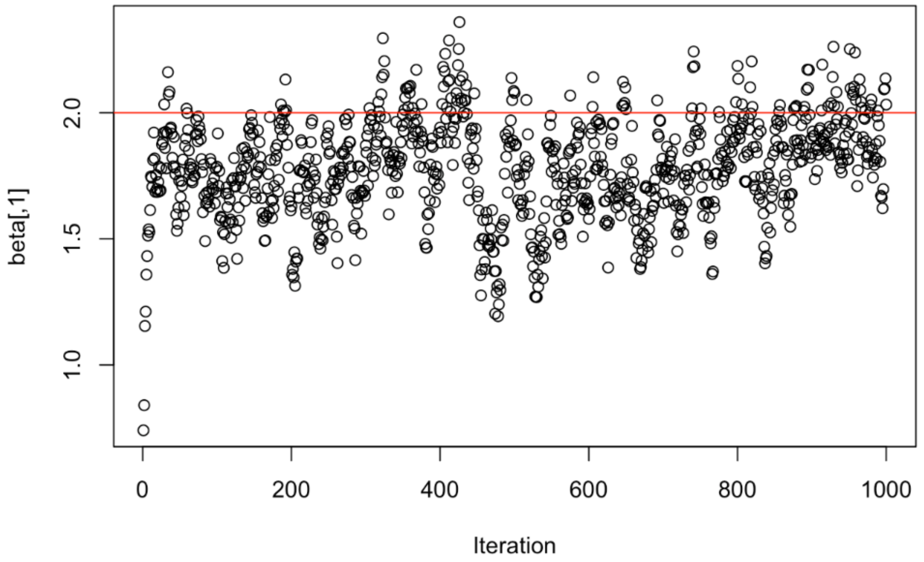


BayesC -> Gibbs Chain

- Set starting values for σ_e^2, μ, δ)
- Then (for many iterations)
  - For each SNP, sample δ_i, β_i conditional on other parameters
  - Sample σ_e^2, μ with updated δ_i, β_i
- Samples reconstruct posterior distributions of parameters

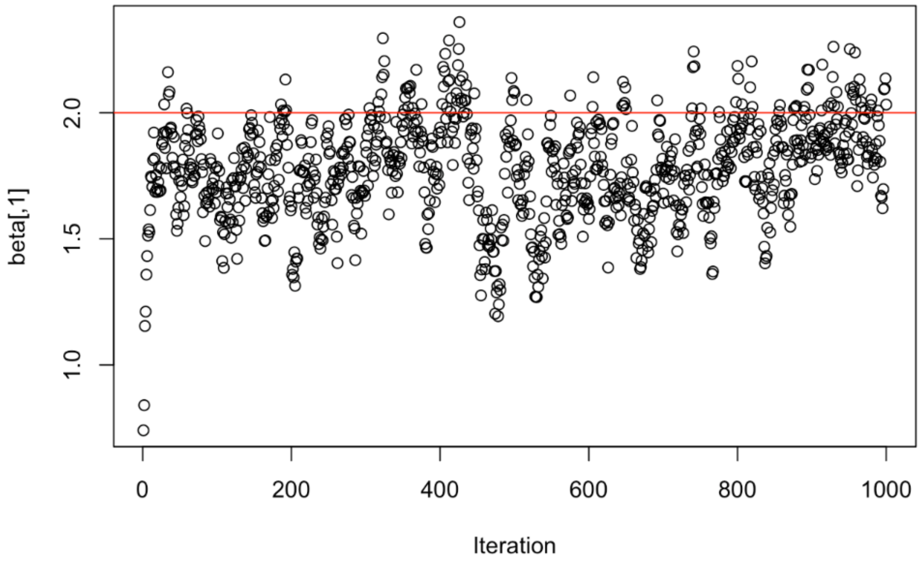
BayesC -> Gibbs Chain

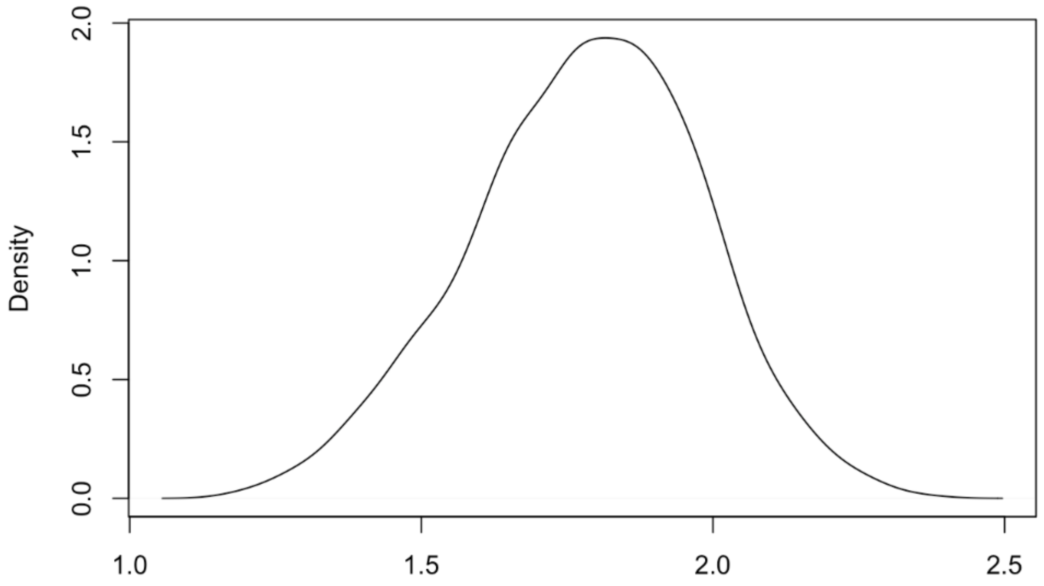
- Set starting values for σ_e^2, μ, δ)
- Then (for many iterations)
  - For each SNP, sample δ_i, β_i conditional on other parameters
  - Sample σ_e^2, μ with updated δ_i, β_i
- Samples reconstruct posterior distributions of parameters



BayesC -> Gibbs Chain

- Set starting values for σ_e^2, μ, δ)
- Then (for many iterations)
  - For each SNP, sample δ_i, β_i conditional on other parameters
  - Sample σ_e^2, μ with updated δ_i, β_i
- Samples reconstruct posterior distributions of parameters





Bayesian methods for Genomic Prediction

- Alternative assumptions regarding the distribution of SNP effects
- Introduction to Bayesian methods
- Genomic prediction with Bayesian methods
- Comparison of accuracy of methods

Real Data, 800K

- Reference
  - Holstein = 3049 bulls, 8478 cows
  - Jersey = 770 bulls, 3917 cows

- Validation
  - Holstein = 262 bulls
  - Jersey = 105 bulls
  - *Australian Reds = 114 bulls*

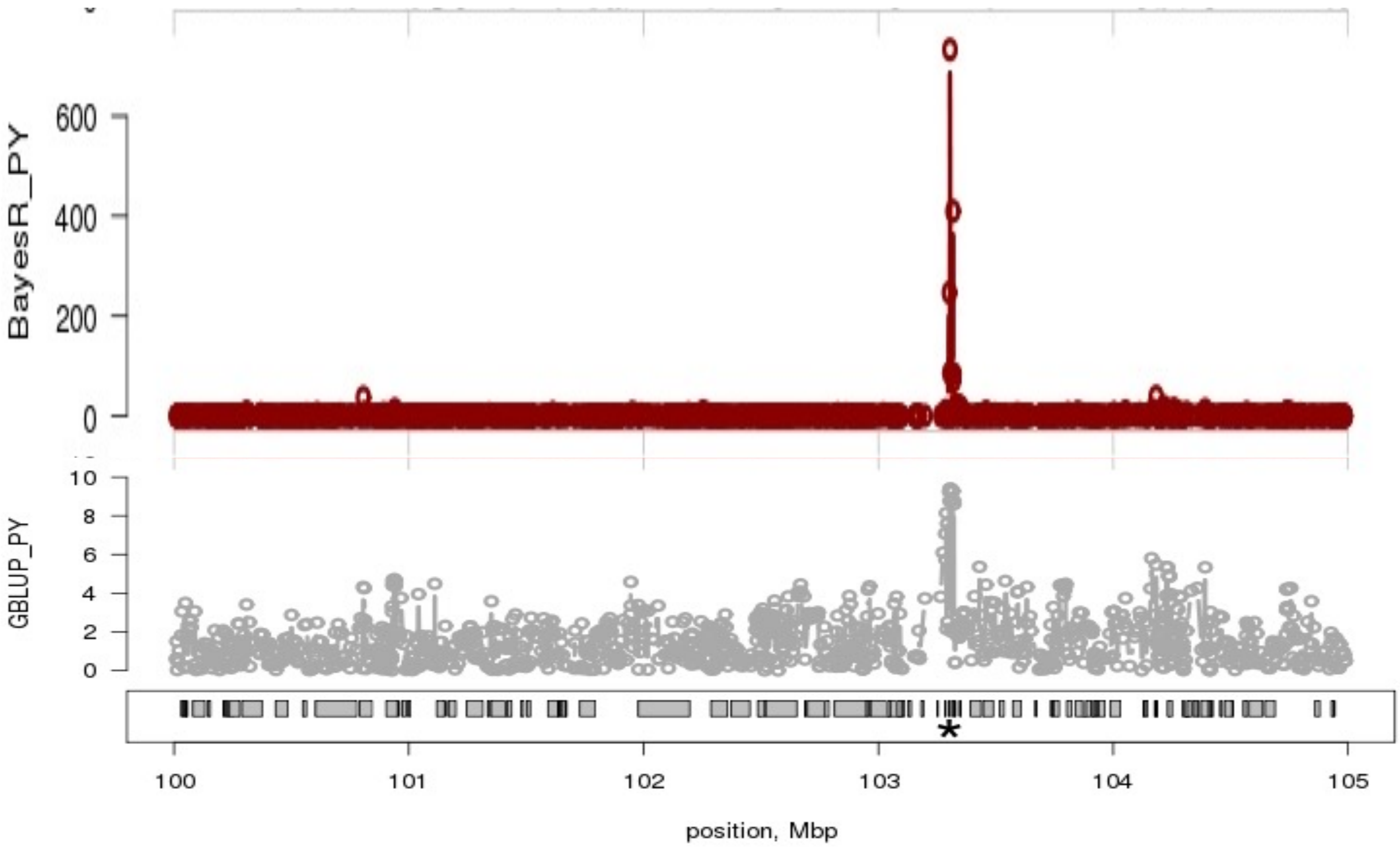
- GEBV with GBLUP, BayesR
- (Kemper et al GSE, 2014)

Real Data, 800K

- r(GEBV,DTD)

	Fat	Milk	Protein	Fat%	Protein%	**Average**
Holstein						
GBLUP	0.60	0.59	0.58	0.72	0.83	**0.66**
BAYESR	0.64	0.62	0.57	0.81	0.84	**0.69**
Jersey						
GBLUP	0.56	0.62	0.67	0.64	0.76	**0.65**
BAYESR	0.56	0.69	0.71	0.76	0.79	**0.70**
Australian Reds						
GBLUP	0.20	0.16	0.11	0.32	0.34	**0.22**
BAYES	0.26	0.21	0.13	0.44	0.36	**0.28**

BayesR



code 00025B

Bayesian methods for Genomic Prediction

Bayesian approach allows us to incorporate prior knowledge in prediction of SNP effects

Bayesian methods can have an advantage when:

QTL of moderate to large effect on the trait (eg Fat%, DGAT1)

Very large numbers of SNP (800K, sequence) -> set some SNP effects to zero

Multi-breed, across population genomic predictions