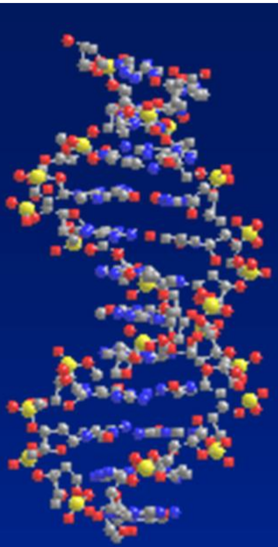
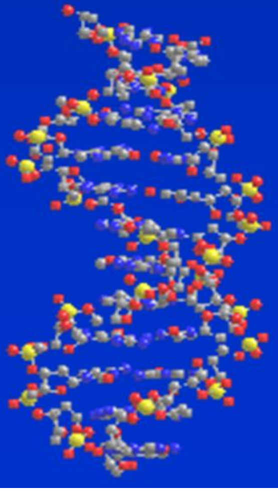
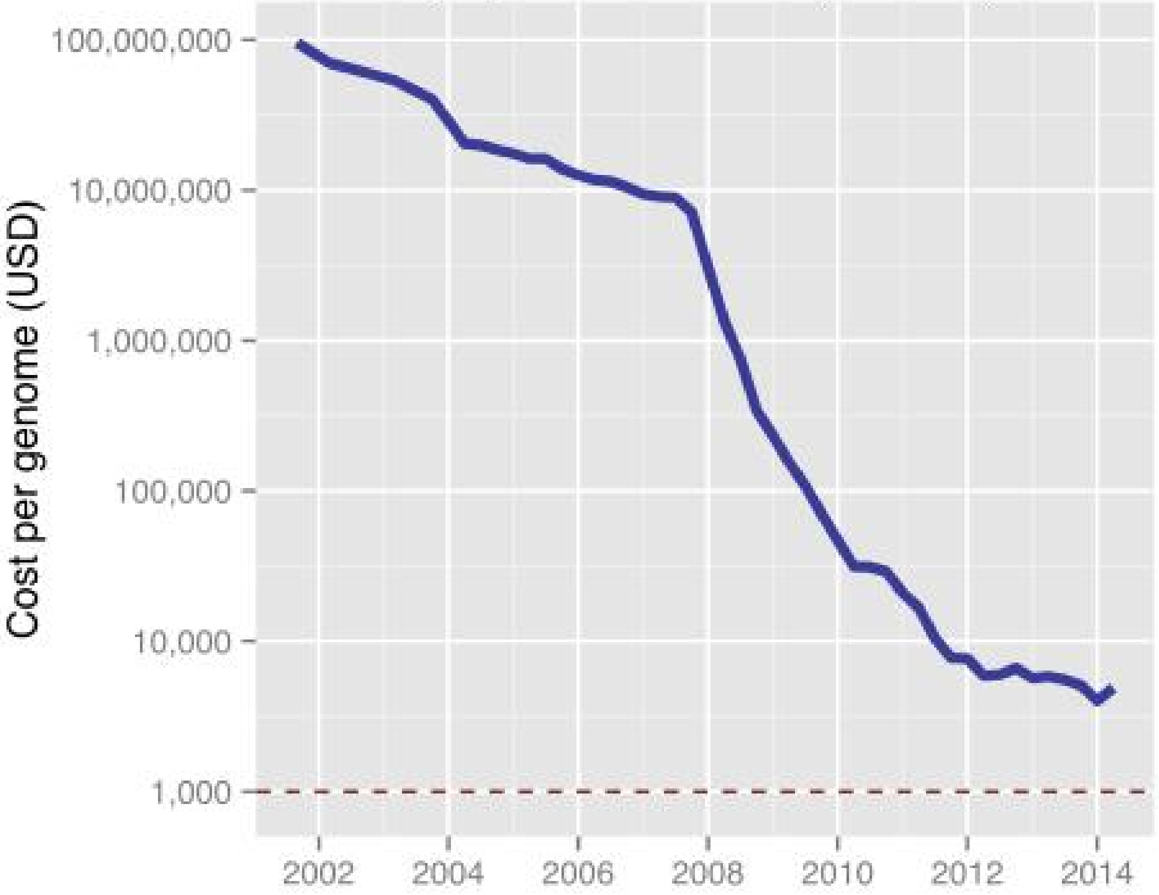


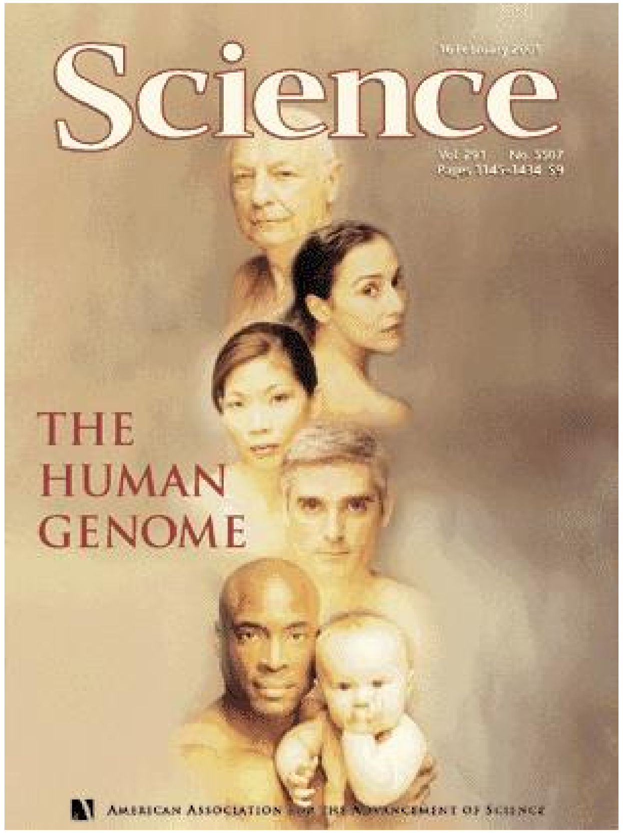
Genomic Prediction and Selection



INFUSE
DEFIER

Genome sequencing cost as estimated by NHGRI
(September 2001 to April 2014)





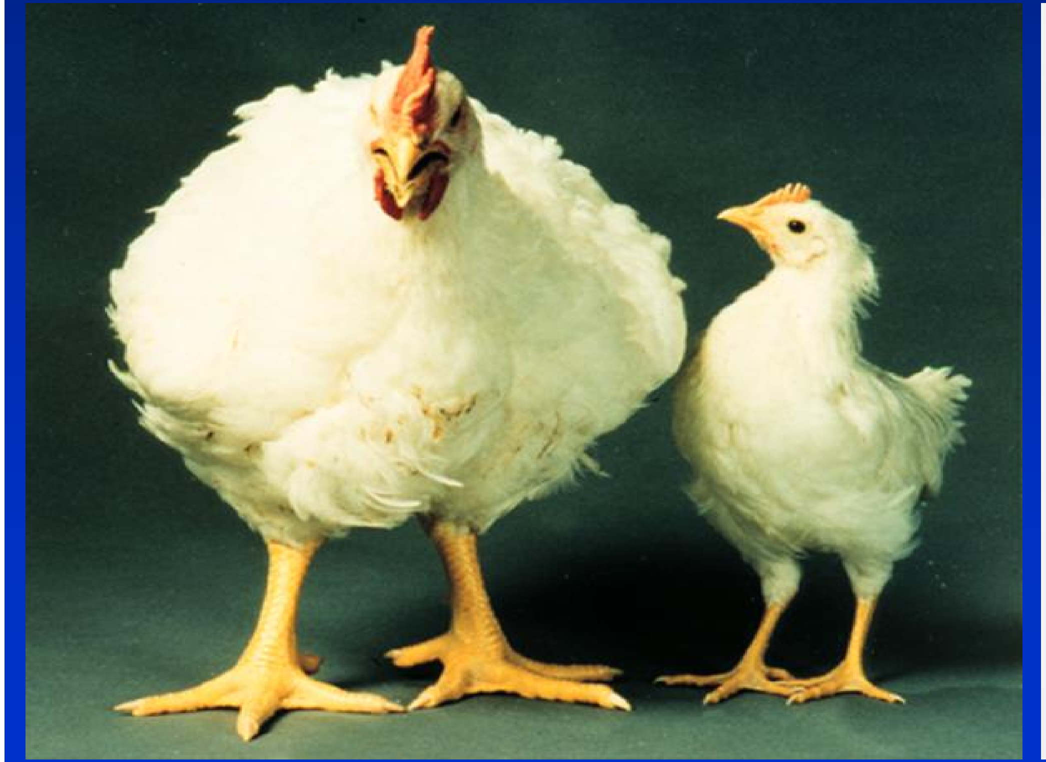
Course overview

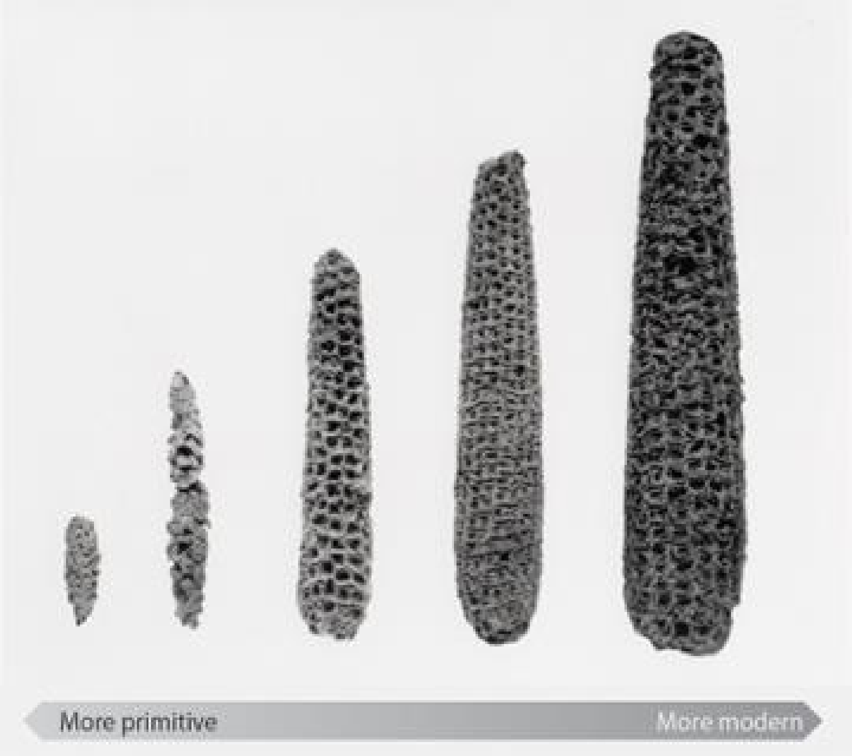
- Day 1
  - Genomic prediction with SNP array genotypes (BLUP, GBLUP and Bayesian methods)
  - What drives the accuracy of genomic prediction – theory and in practise

- Day 2
  - Brief introduction to sequence data and implications for genomic prediction
  - Genomic prediction in practise 1. Combining genomic prediction with other information – selection index and multiple trait approaches
  - Genomic prediction in practise 2. Genomic selection in breeding programs

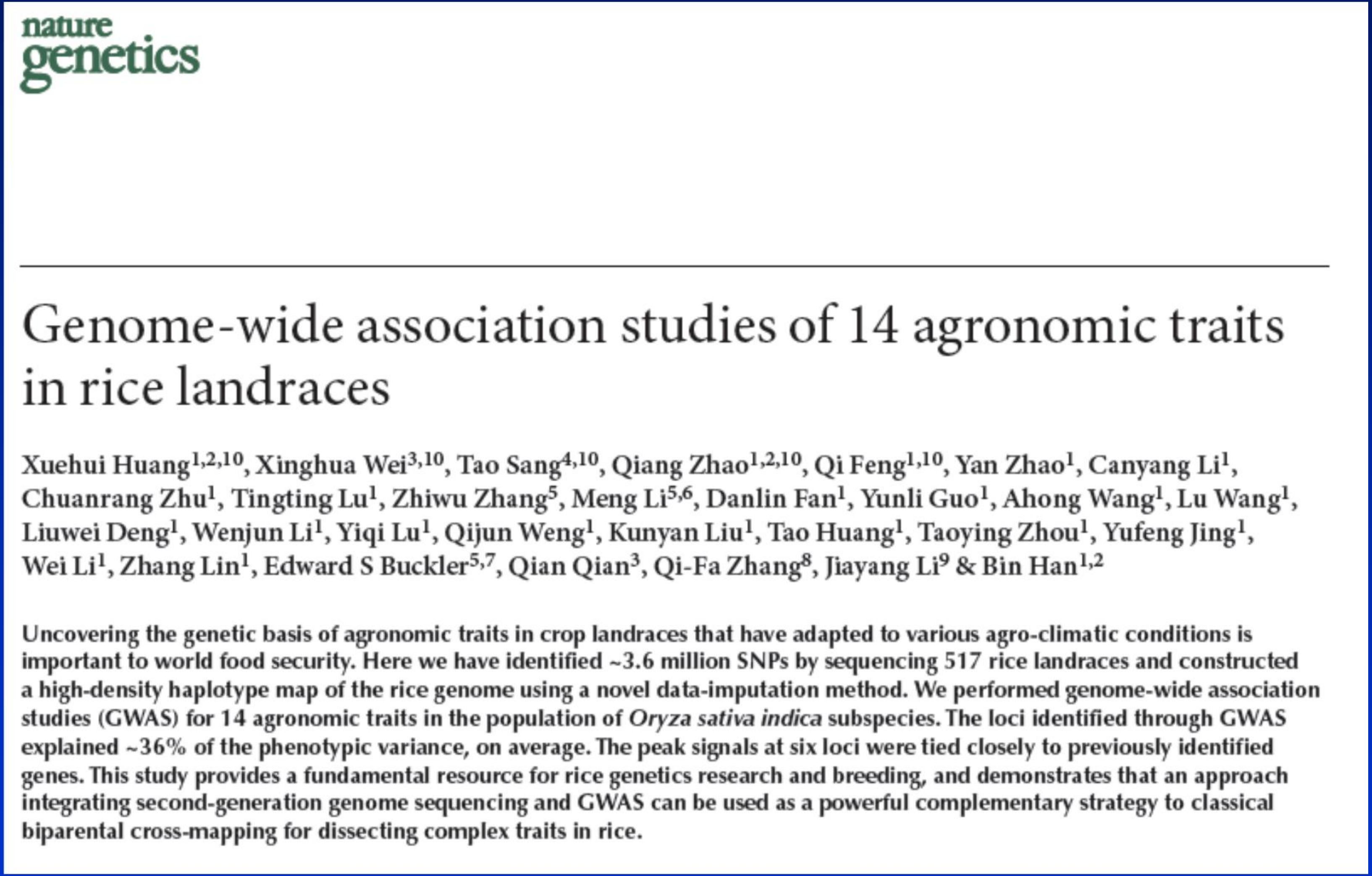
Quantitative traits

- Genetic variation observed for many (all?) traits of economic importance in livestock and plant species
- One gene or many?





Yield in Rice



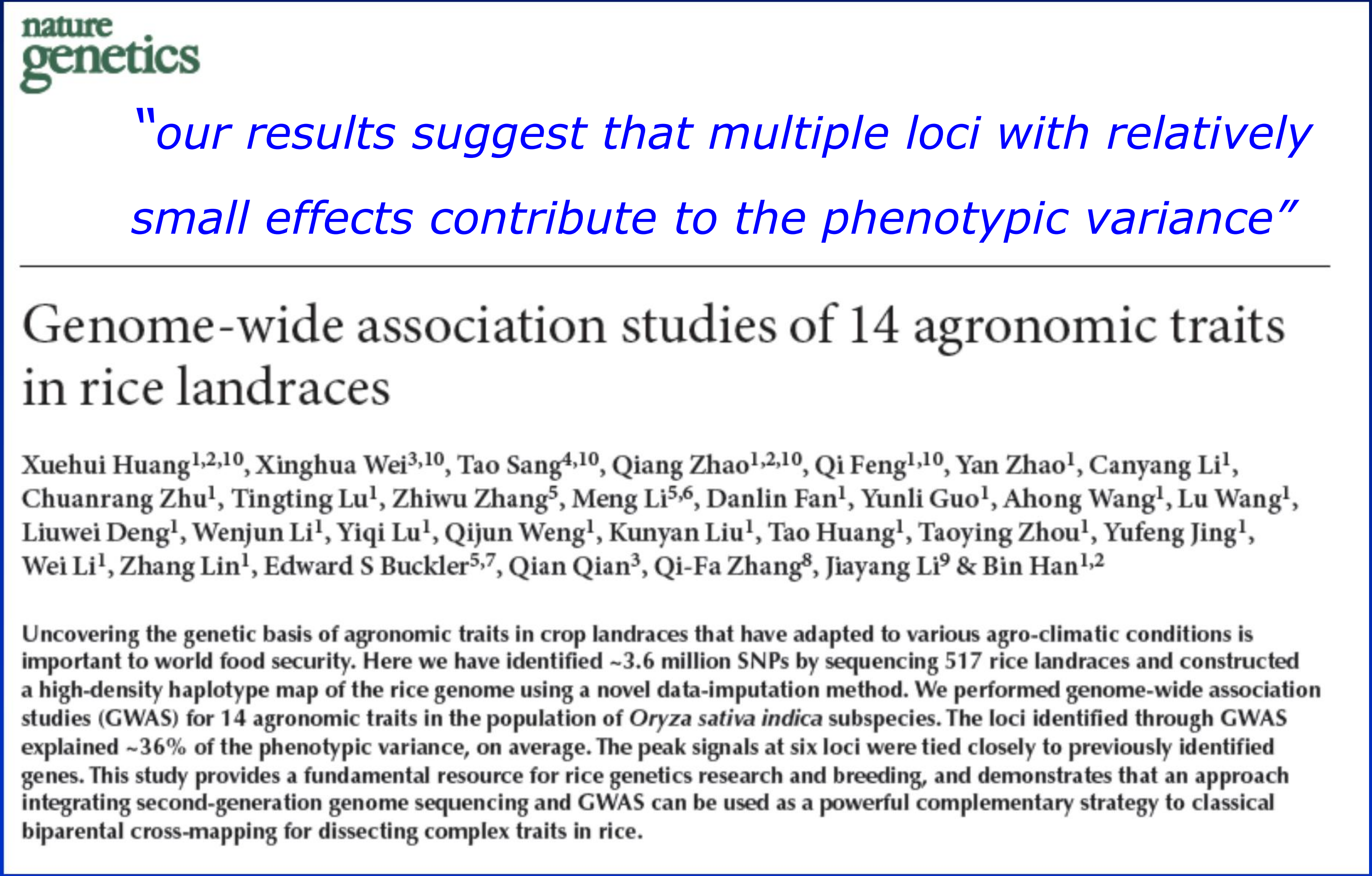
nature genetics

Genome-wide association studies of 14 agronomic traits in rice landraces

Xuehui Huang[1,2,10], Xinghua Wei[3,10], Tao Sang[4,10], Qiang Zhao[1,2,10], Qi Feng[1,10], Yan Zhao[1], Canyang Li[1], Chuanrang Zhu[1], Tingting Lu[1], Zhiwu Zhang[5], Meng Li[5,6], Danlin Fan[1], Yunli Guo[1], Ahong Wang[1], Lu Wang[1], Liuwei Deng[1], Wenjun Li[1], Yiqi Lu[1], Qijun Weng[1], Kunyan Liu[1], Tao Huang[1], Taoying Zhou[1], Yufeng Jing[1], Wei Li[1], Zhang Lin[1], Edward S Buckler[5,7], Qian Qian[3], Qi-Fa Zhang[8], Jiayang Li[9] & Bin Han[1,2]

Uncovering the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using a novel data-imputation method. We performed genome-wide association studies (GWAS) for 14 agronomic traits in the population of *Oryza sativa indica* subspecies. The loci identified through GWAS explained ~36% of the phenotypic variance, on average. The peak signals at six loci were tied closely to previously identified genes. This study provides a fundamental resource for rice genetics research and breeding, and demonstrates that an approach integrating second-generation genome sequencing and GWAS can be used as a powerful complementary strategy to classical biparental cross-mapping for dissecting complex traits in rice.

Yield in Rice



nature genetics

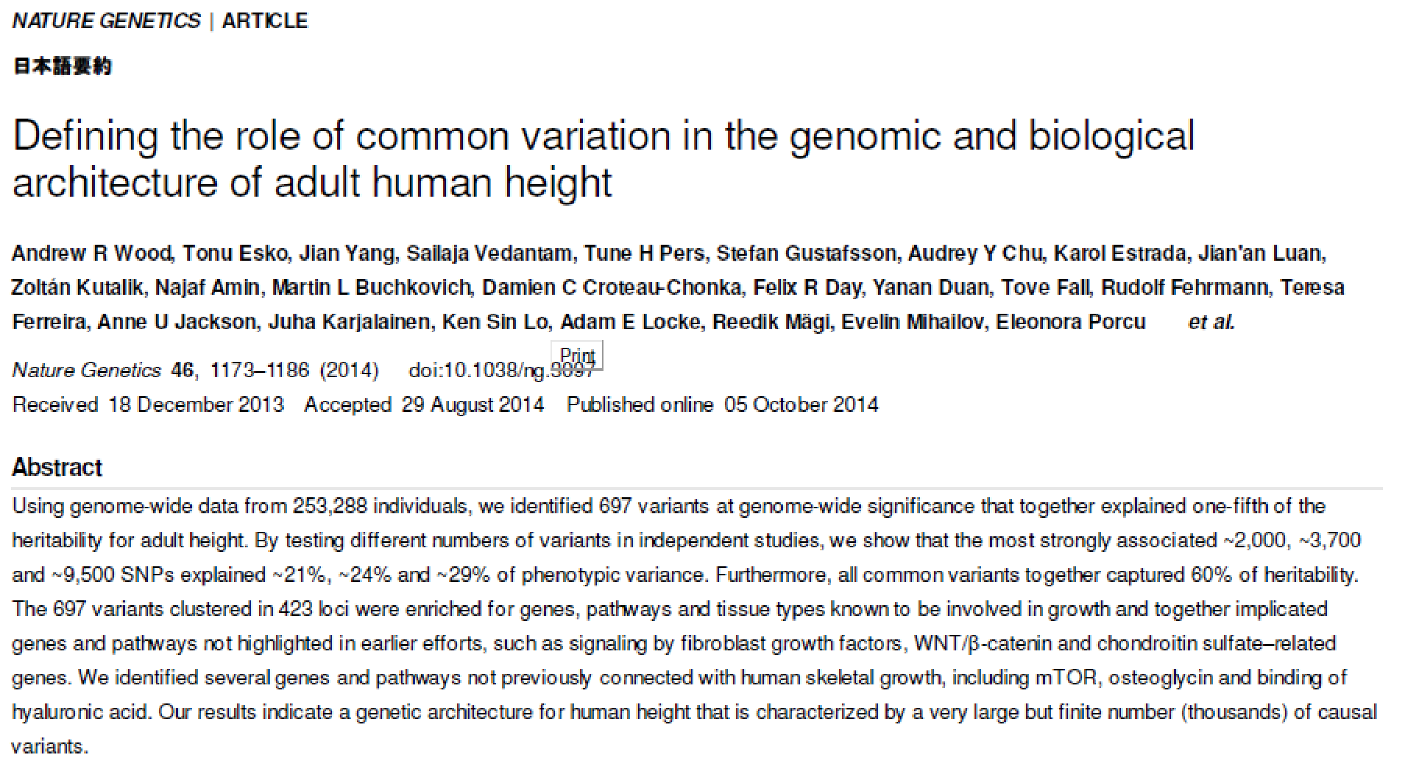
"our results suggest that multiple loci with relatively small effects contribute to the phenotypic variance"

Genome-wide association studies of 14 agronomic traits in rice landraces

Xuehui Huang[1,2,10], Xinghua Wei[3,10], Tao Sang[4,10], Qiang Zhao[1,2,10], Qi Feng[1,10], Yan Zhao[1], Canyang Li[1], Chuanrang Zhu[1], Tingting Lu[1], Zhiwu Zhang[5], Meng Li[5,6], Danlin Fan[1], Yunli Guo[1], Ahong Wang[1], Lu Wang[1], Liuwei Deng[1], Wenjun Li[1], Yiqi Lu[1], Qijun Weng[1], Kunyan Liu[1], Tao Huang[1], Taoying Zhou[1], Yufeng Jing[1], Wei Li[1], Zhang Lin[1], Edward S Buckler[5,7], Qian Qian[3], Qi-Fa Zhang[8], Jiayang Li[9] & Bin Han[1,2]

Uncovering the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using a novel data-imputation method. We performed genome-wide association studies (GWAS) for 14 agronomic traits in the population of *Oryza sativa indica* subspecies. The loci identified through GWAS explained ~36% of the phenotypic variance, on average. The peak signals at six loci were tied closely to previously identified genes. This study provides a fundamental resource for rice genetics research and breeding, and demonstrates that an approach integrating second-generation genome sequencing and GWAS can be used as a powerful complementary strategy to classical biparental cross-mapping for dissecting complex traits in rice.

Human height



NATURE GENETICS | ARTICLE

日本語要約

Defining the role of common variation in the genomic and biological architecture of adult human height

Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, Najaf Amin, Martin L Buchkovich, Damien C Croteau-Chonka, Felix R Day, Yanan Duan, Tove Fall, Rudolf Fehrmann, Teresa Ferreira, Anne U Jackson, Juha Karjalainen, Ken Sin Lo, Adam E Locke, Reedik Mägi, Evelin Mihailov, Eleonora Porcu ***et al.***

Nature Genetics **46**, 1173–1186 (2014) doi:10.1038/ng.3097
Received 18 December 2013 Accepted 29 August 2014 Published online 05 October 2014

Abstract

Using genome-wide data from 253,288 individuals, we identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated ~2,000, ~3,700 and ~9,500 SNPs explained ~21%, ~24% and ~29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The 697 variants clustered in 423 loci were enriched for genes, pathways and tissue types known to be involved in growth and together implicated genes and pathways not highlighted in earlier efforts, such as signaling by fibroblast growth factors, WNT/β-catenin and chondroitin sulfate–related genes. We identified several genes and pathways not previously connected with human skeletal growth, including mTOR, osteoglycin and binding of hyaluronic acid. Our results indicate a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants.

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

USING YOUR
GENOME

If you want to predict how tall your children might one day be, a good bet would be to look in the mirror, and at your mate. Studies going back almost a century have

Even though these genome-wide association studies (GWAS) turned up dozens of variants, they did "very little of the prediction that you would do just by asking people how tall their parents are", says Joel Hirschhorn at the Broad Institute in Cambridge, Massachusetts, who led one of the studies[3].

contribute to a variety of traits and common diseases. But even when dozens of genes have been linked to a trait, both the individual and cumulative effects are disappointingly small and nowhere near enough to explain earlier estimates of heritability. "It is the big topic in the genetics of common disease right

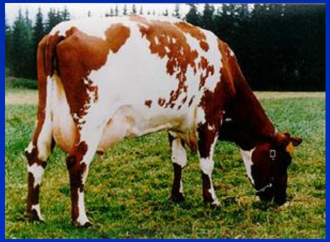
ILLUSTRATIONS BY D. PARKINS

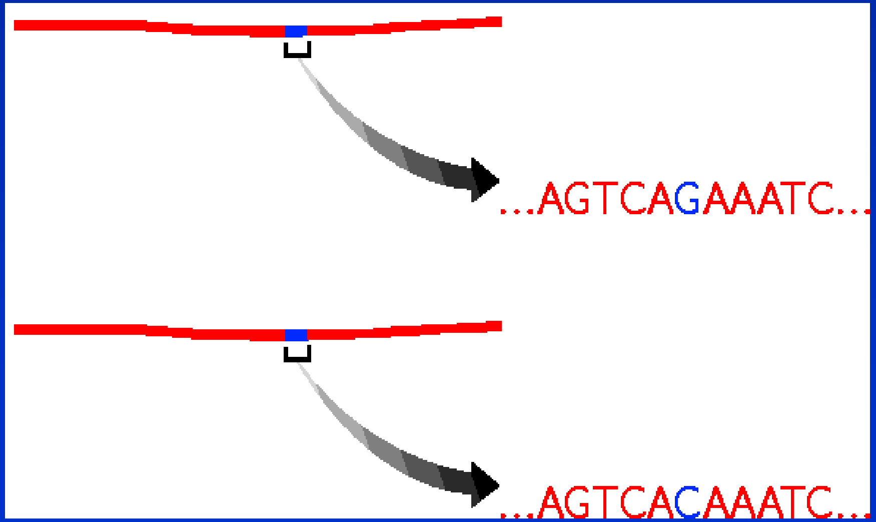
Complex traits

- Large number of causative mutations (quantitative trait loci, QTL) for most complex traits
- Variance explained by individual markers will be small
- Use large numbers of DNA markers to simultaneously track all QTL
- Sequence data -> includes causative mutations

The Revolution

- As a result of sequencing animal and plant genomes, have a huge amount of information on variation in the genome
  - at the DNA level
- Most abundant form of variation are Single Nucleotide Polymorphisms (SNPs)

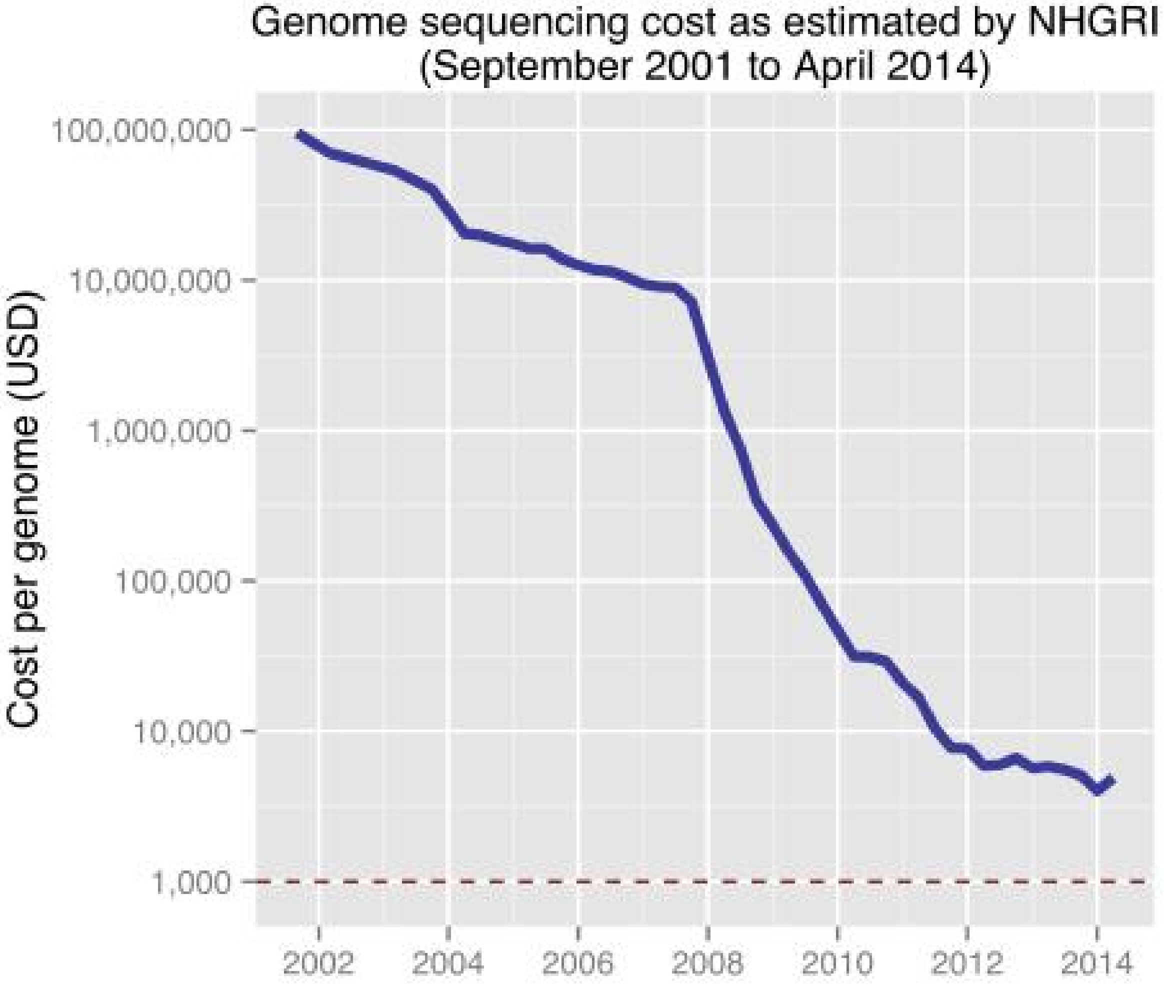




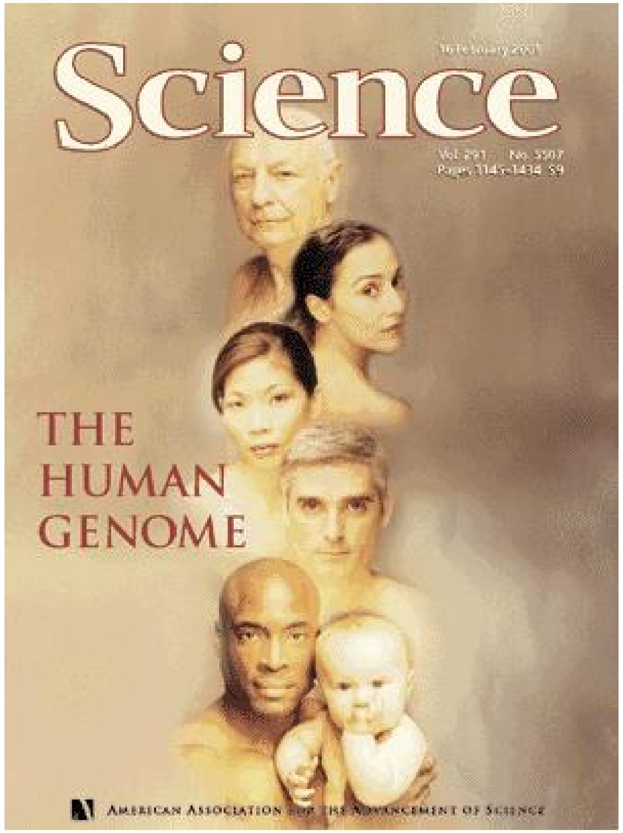
- 1000 Genomes project (Pilot)
- ~15 mill SNPs
- ~7 mill SNPs with minor allele >5%
- ~100,000-300,000 cSNPs
- ~50,000 nonsynonymous cSNPs -> change protein structure
- Every individual carries 250-300 loss of function mutations!

The Revolution

- SNP chips available for
  - Sheep, Cattle (50K, 800K), Pigs,
  - Chickens
  - Salmon
  - Horse, Dog
- Plants
  - Maize, Wheat
  - Cotton, Soybean under development
- Cost?
  - ~ $100-200 USD for 60K SNPs
- Genotyping by re-sequencing?
  - 40 million SNPs in cattle
  - Insertion deletions
  - Copy number variants?



Genome sequencing cost as estimated by NHGRI
(September 2001 to April 2014)



Sequence data vs SNP arrays

- Genomic selection (all hypotheses!)
  - No longer have to rely on LD, causative mutation actually in data set
    - Higher accuracy of prediction?
    - Better prediction across breeds/populations?
    - Better persistence of accuracy across generations

- **But** have sequencing errors, genotype errors, expense…….

Aim

- Provide you with genome wide association and genomic prediction methodologies to exploit high density genotypes, up to whole genome sequence data, in livestock and plant improvement

Course overview

- Day 1
  - Genomic prediction with SNP array genotypes (BLUP, GBLUP and Bayesian methods)
  - What drives the accuracy of genomic prediction – theory and in practise

- Day 2
  - Brief introduction to sequence data and implications for genomic prediction
  - Genomic prediction in practise 1. Combining genomic prediction with other information – selection index and multiple trait approaches
  - Genomic prediction in practise 2. Genomic selection in breeding programs

Genomic prediction

- Introduction to genomic prediction

- Genomic prediction with BLUP/GBLUP

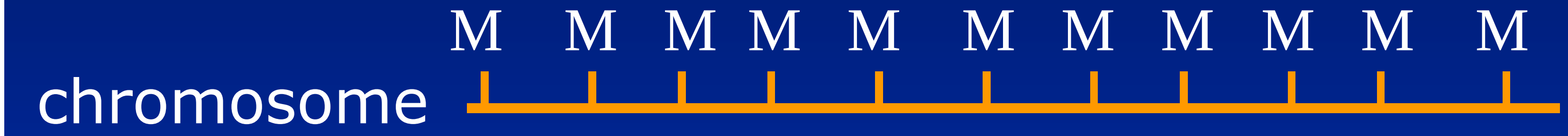
- Genomic prediction with Bayesian methods

- Examples in real data

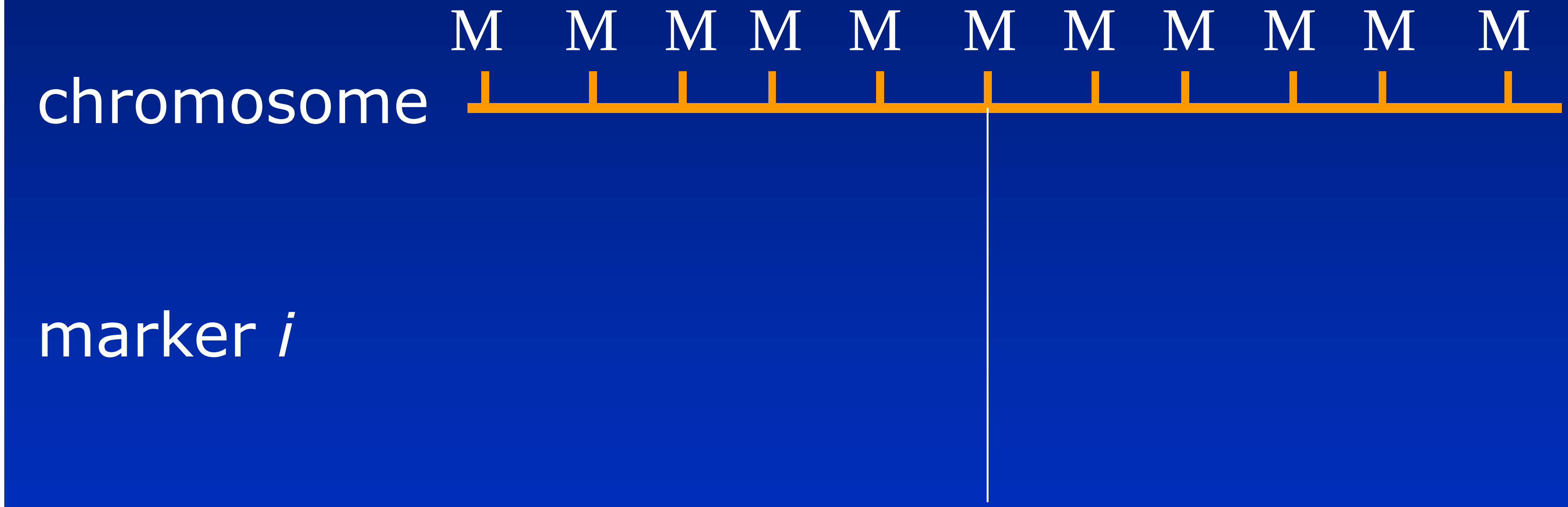
Genomic prediction

- 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

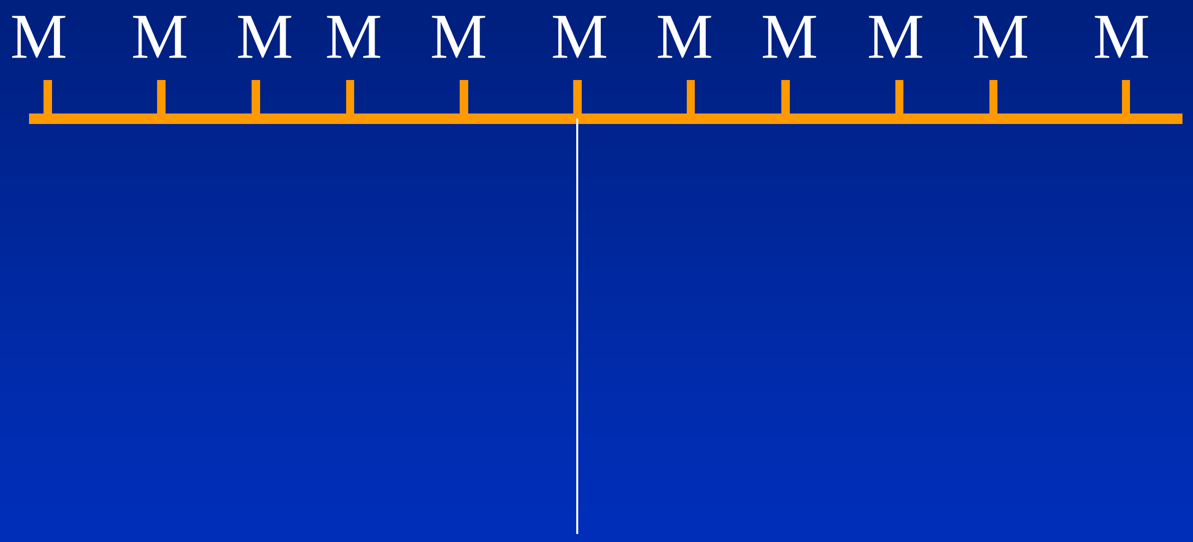


Genomic prediction



Genomic prediction

chromosome

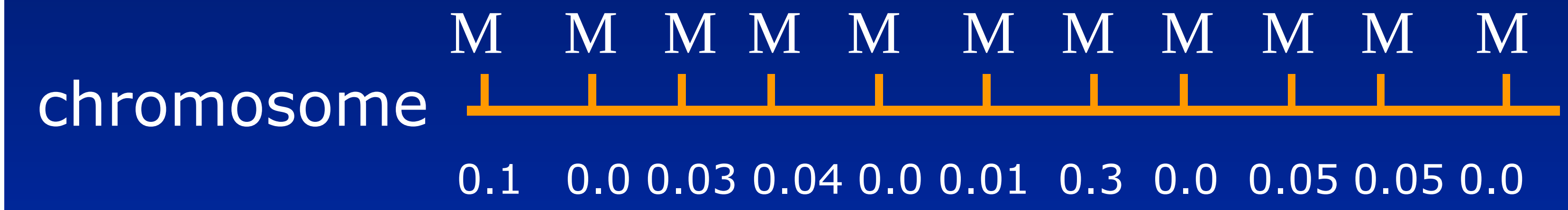


marker *i*

Effect of “2” allele

+0.3 L milk

Genomic prediction



M M M M M M M M M M M

chromosome

0.1 0.0 0.03 0.04 0.0 0.01 0.3 0.0 0.05 0.05 0.0

Genomic prediction

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

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

Genomic prediction

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

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

Number of SNP

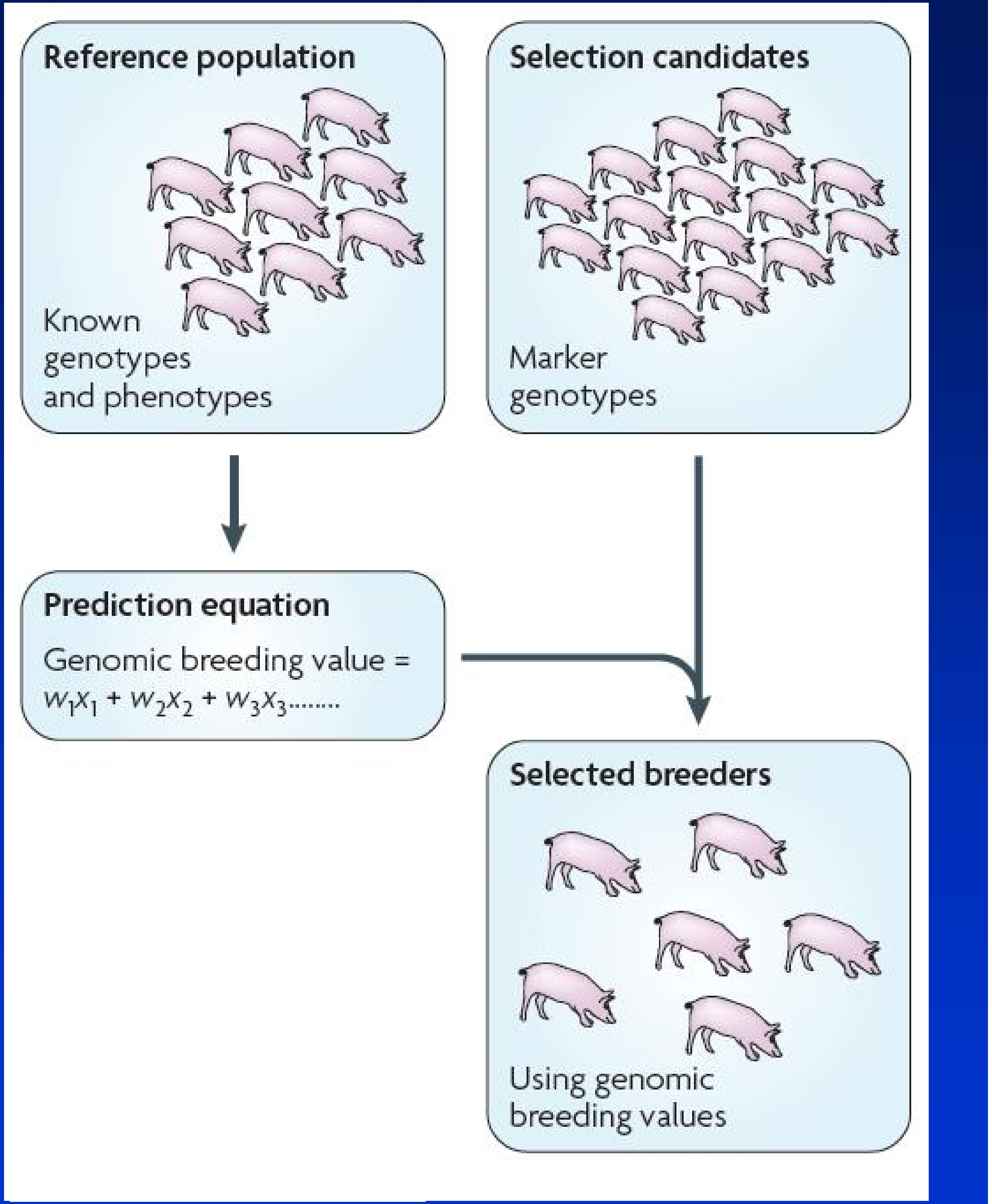
Genomic prediction

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

Genomic prediction

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

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 prediction

- First step is to predict the chromosome segment effects in a reference population
- Number of effects >>> than number of records
- Eg. 50,000 SNPs
- From ~ 2000 records?
- Need methods that can deal with this

Genomic prediction

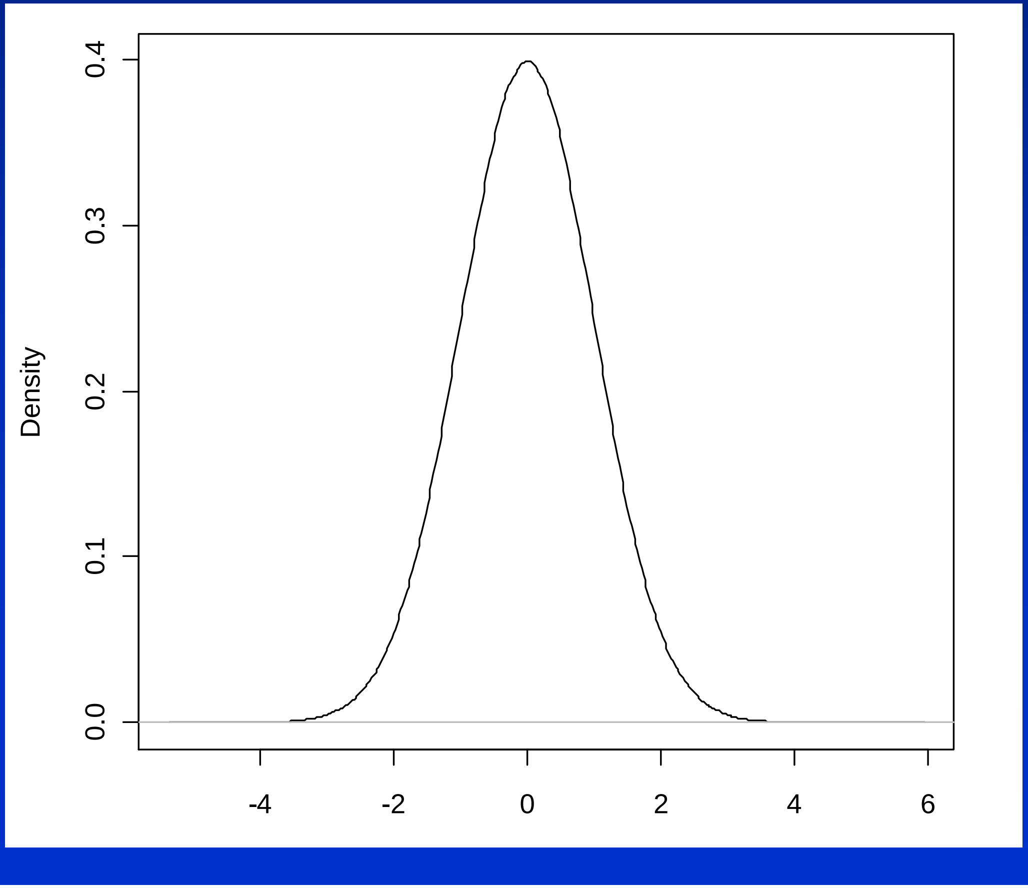
- BLUP = best linear unbiased prediction
- Model:

$$\mathbf{y} = \mathbf{1_n}\mu + \sum_{i=1}^{p} \mathbf{X_i g_i} + \mathbf{e}$$

- In BLUP we assume SNP effects come from normal distribution with same variance
  E(**g**) ~ N(0,σ_g^2)

Genomic prediction with BLUP

- BLUP assumes normal distribution of SNP effects



Genomic prediction with BLUP

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

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

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

Genomic prediction 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 prediction with BLUP

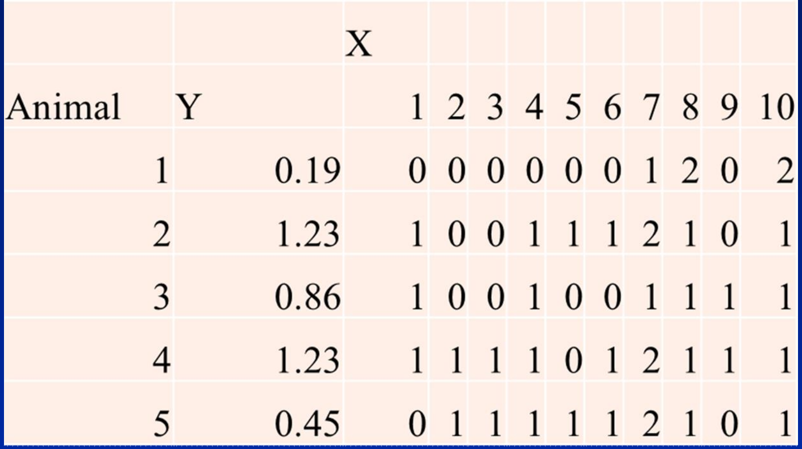
- Example

		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 prediction 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]

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

Genomic prediction 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 prediction with BLUP

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

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

- 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 prediction 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 prediction 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 prediction 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 prediction with BLUP

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

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

- Then

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

Genomic prediction with BLUP

- An equivalent model

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Zv} + \mathbf{e}$$

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

Genomic prediction with BLUP

- An equivalent model
  - Model 1.

$$\mathbf{y} = \mathbf{1_n}\mu + \sum_{i=1}^{p} \mathbf{X_i g_i} + \mathbf{e}$$

$$\begin{bmatrix} \overset{\wedge}{\mu} \\ \overset{\wedge}{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

$$\mathbf{GEBV} = \mathbf{X}\overset{\wedge}{\mathbf{g}}$$

  - Model 2.

Genomic prediction with BLUP

- An equivalent model
  - Model 1.

$$\mathbf{y} = \mathbf{1_n}\mu + \sum_{i=1}^{p}\mathbf{X_i g_i} + \mathbf{e}$$

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

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

  - Model 2.

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Zv} + \mathbf{e}$$

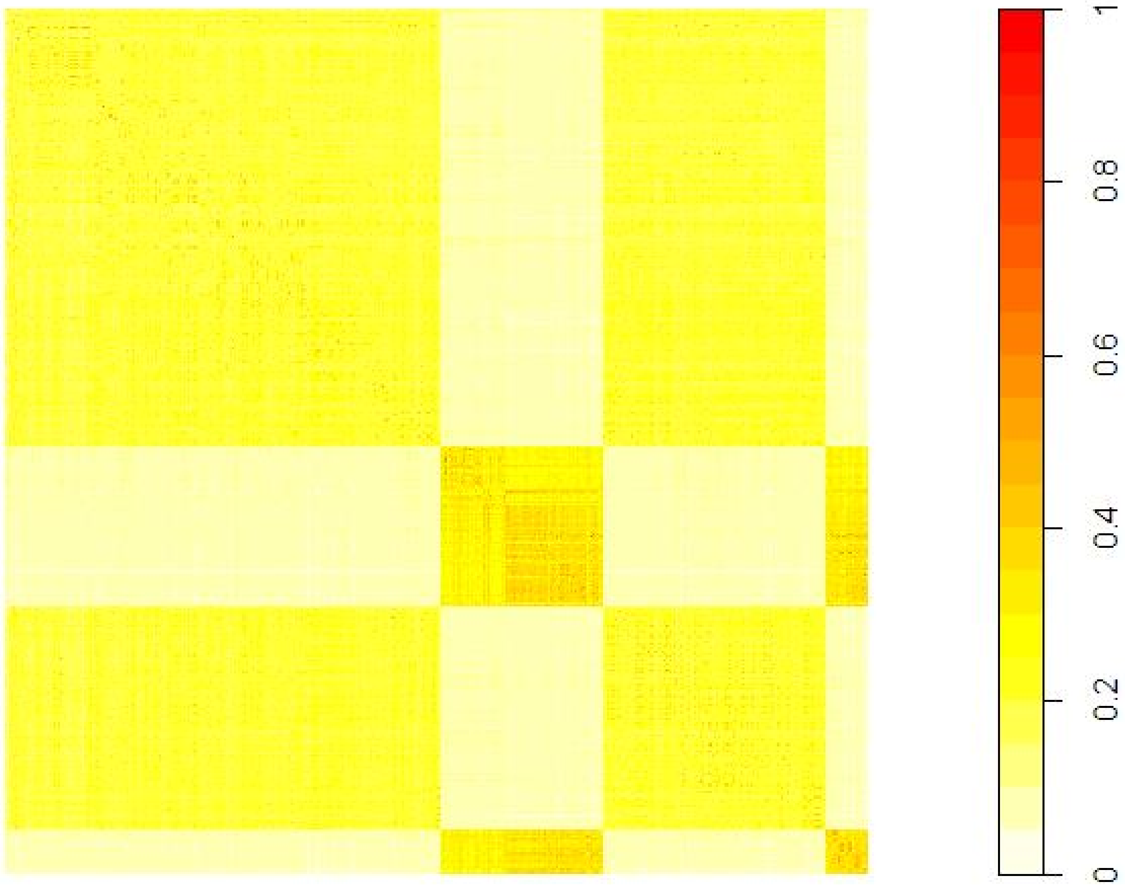
$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_v'Z} \\ \mathbf{Z'1_n} & \mathbf{Z'Z} + \mathbf{G}^{-1}\dfrac{\sigma_e^2}{\sigma_v^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{Z'y} \end{bmatrix}$$

Holstein reference *n = 781*

Jersey reference *n = 287*

Holstein validation *n = 400*

Jersey validation *n = 77*



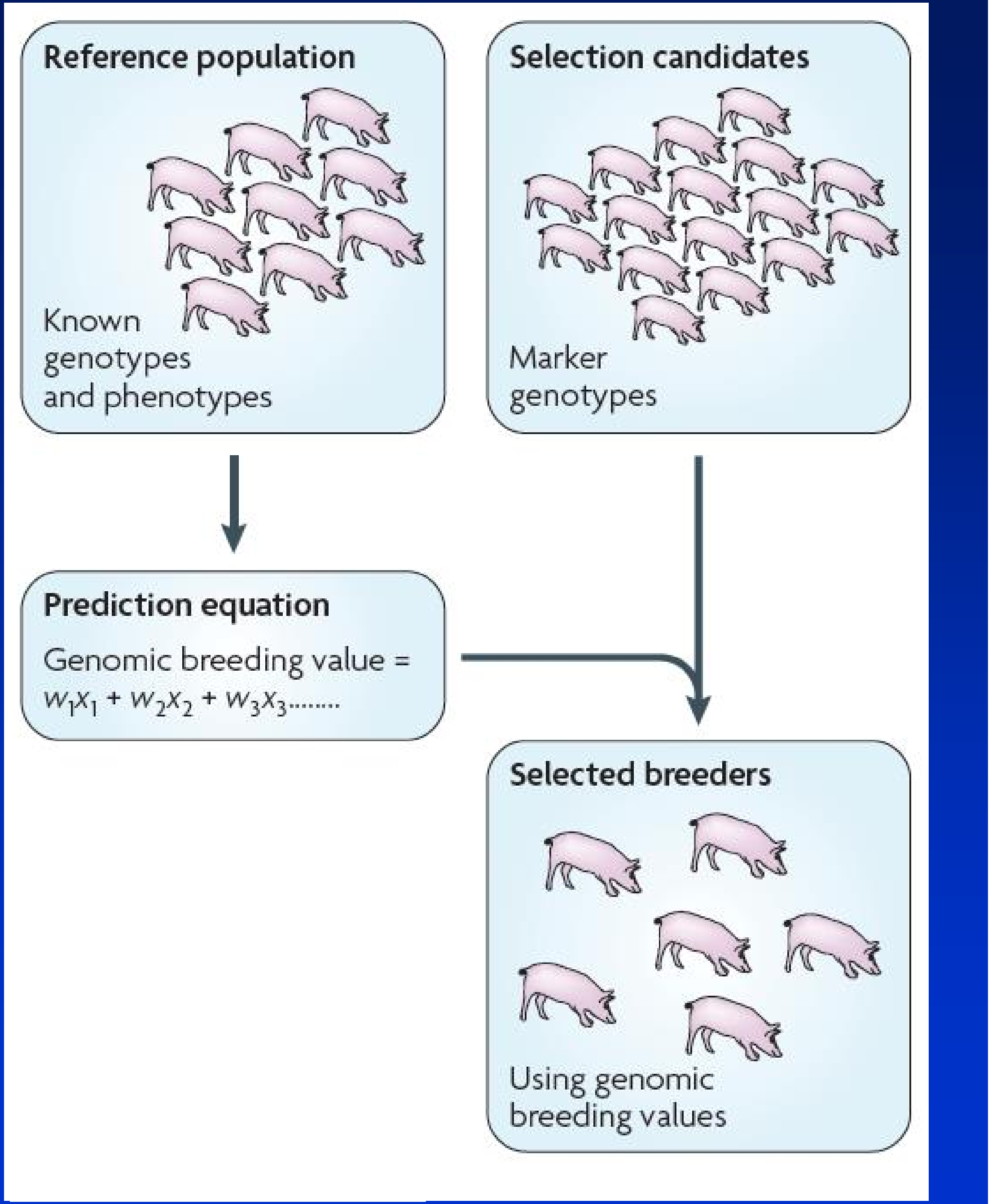
Genomic prediction with BLUP

- An equivalent model
- Why use model 2 (GBLUP).
  - 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 prediction

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

Genomic selection



Reference population

Known genotypes and phenotypes

Selection candidates

Marker genotypes

Prediction equation

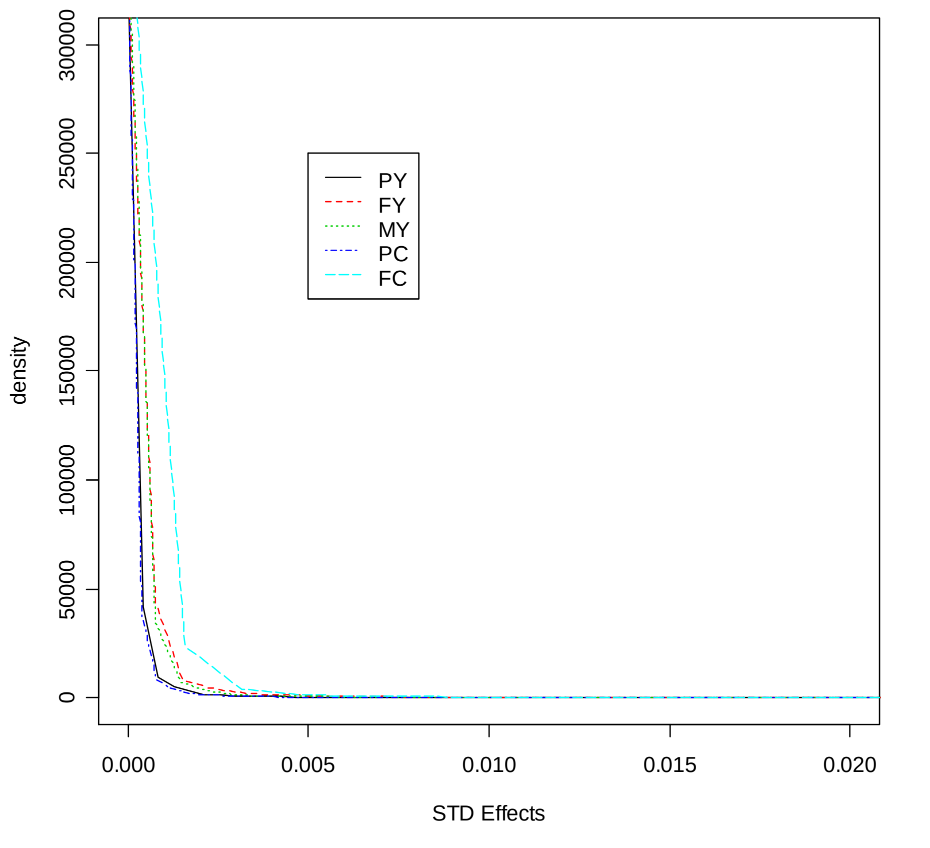
Genomic breeding value = $w_1x_1 + w_2x_2 + w_3x_3$........

Selected breeders

Using genomic breeding values

Alternative prior assumptions for SNP effects

- BLUP assumes normally distributed QTL effects
- Does not match prior knowledge of distributions of QTL effects for some traits



Alternative prior assumptions for SNP effects

- Students t distribution?
  - BayesA
- Many zero effects and proportion Students t distribution?
  - BayesB
- Many zero effect and rest normal distribution
  - BayesCpi
- Double exponential effects
  - BayesianLASSO
- Multiple normal distributions
  - BayesMulti, BayesR

Bayesian methods

- Bayes theorem

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

Bayesian methods

- Bayes theorem

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

Probability of parameters x given the data y (posterior)

Bayesian methods

- Bayes theorem

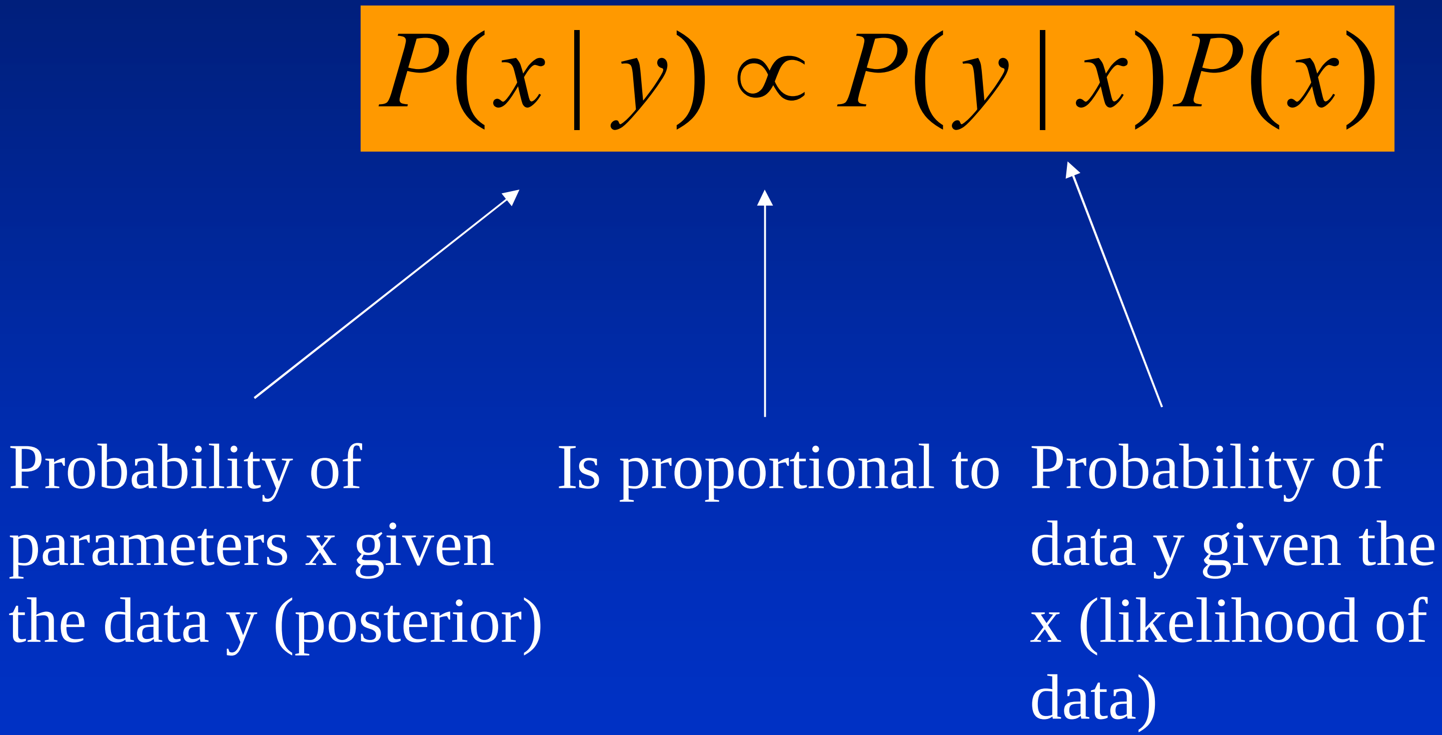
$$\boxed{P(x \mid y) \propto P(y \mid x) P(x)}$$

Probability of parameters x given the data y (posterior)

Is proportional to

Bayesian methods

- Bayes theorem



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

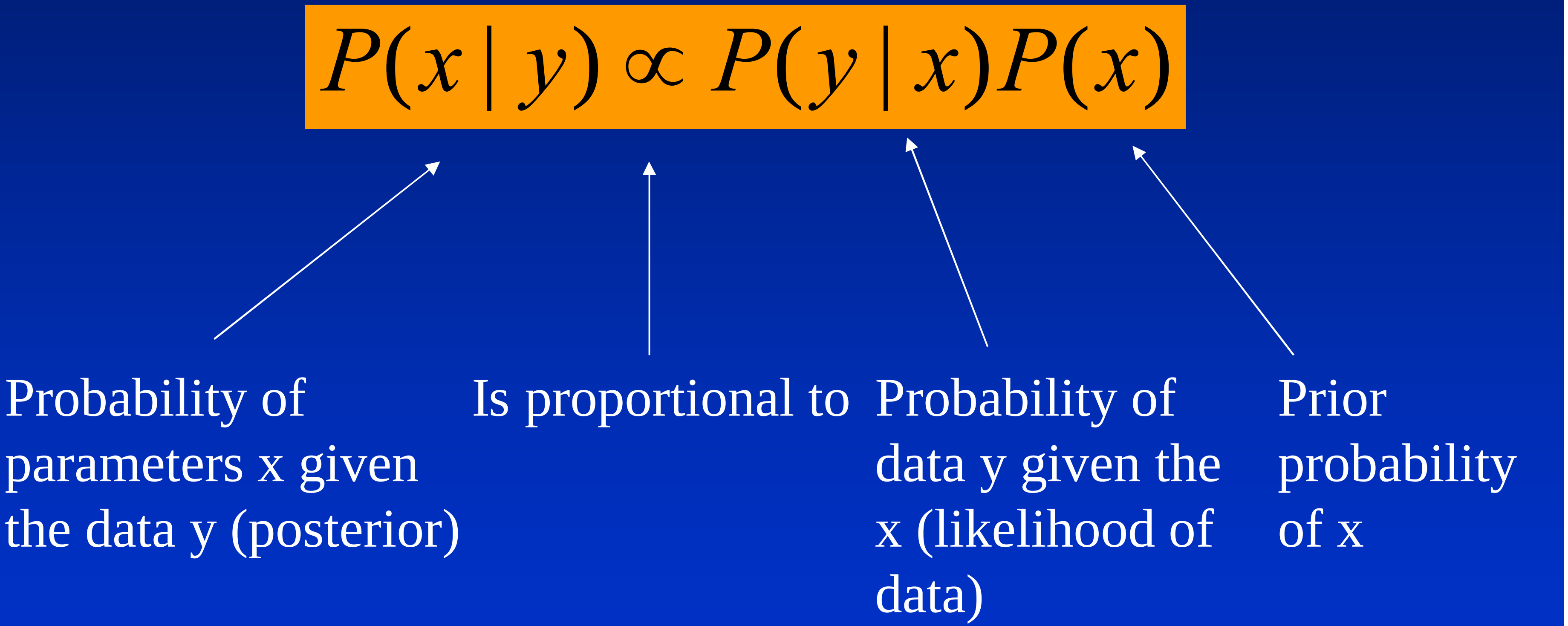
Probability of parameters x given the data y (posterior)

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

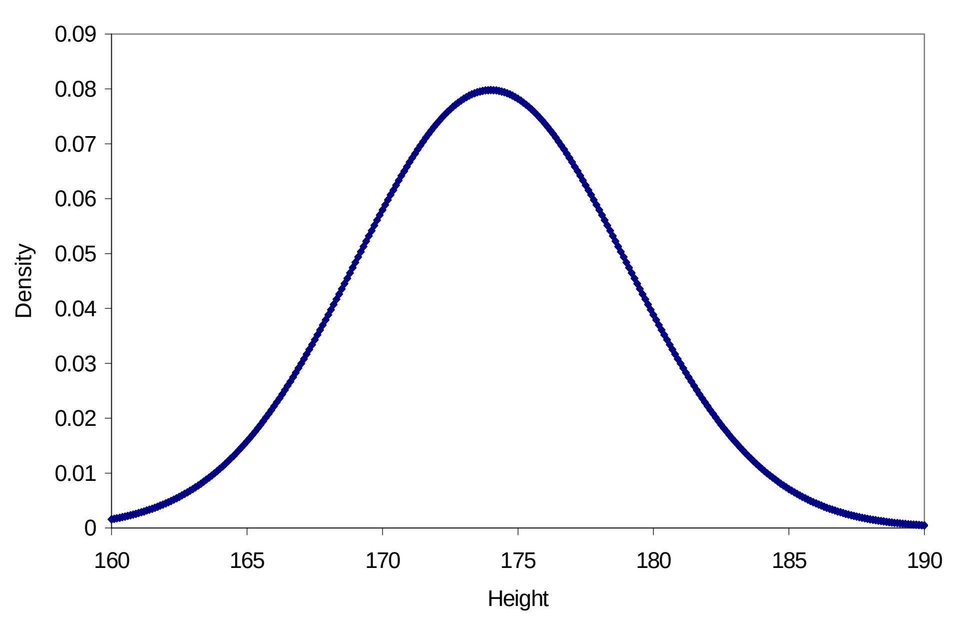
$$y = \text{average height} + e$$

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

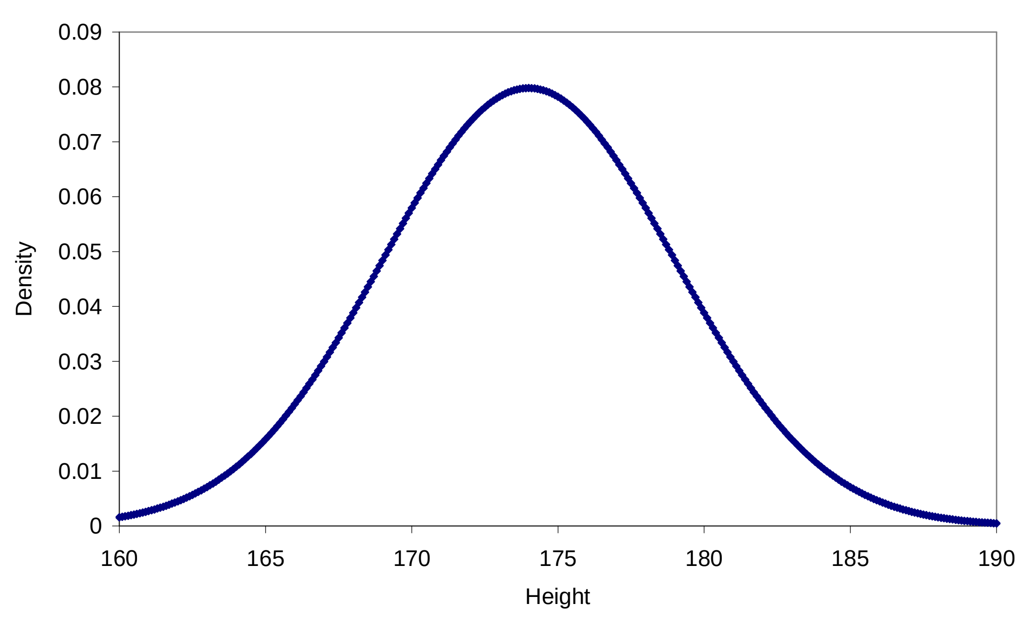
$$P(x\mid y)\propto P(y\mid x)P(x)$$

Prior probability of x (average height)

From the dataé é

$$\bar{x}=178$$

$$s.e=5$$



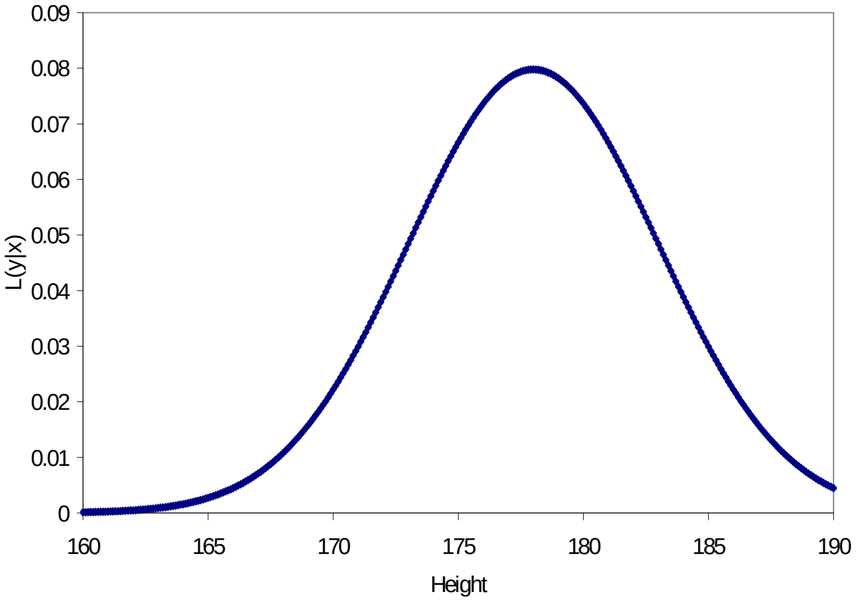
Bayesian methods

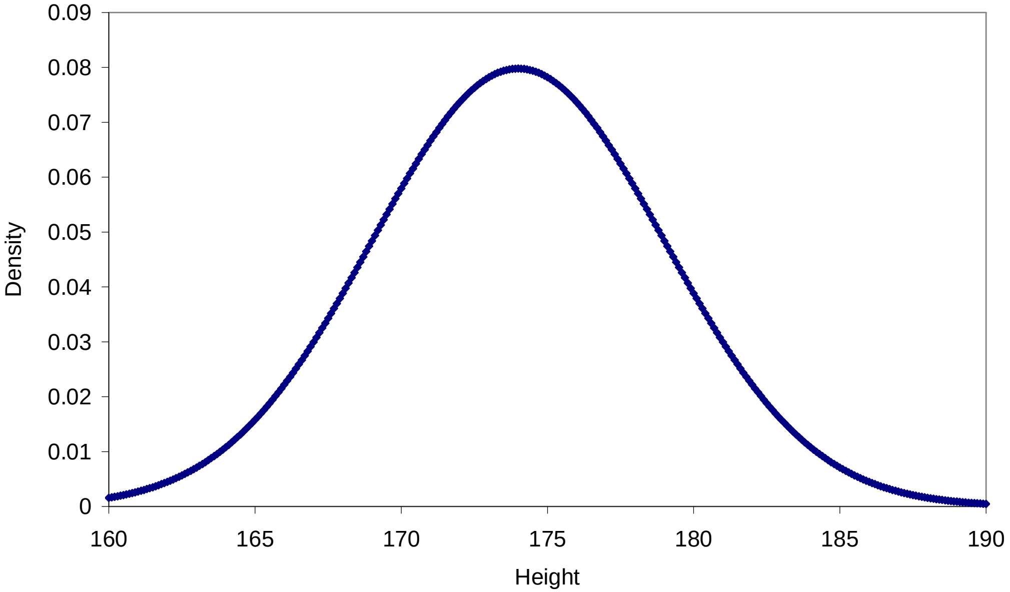
- Bayes theorem

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

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

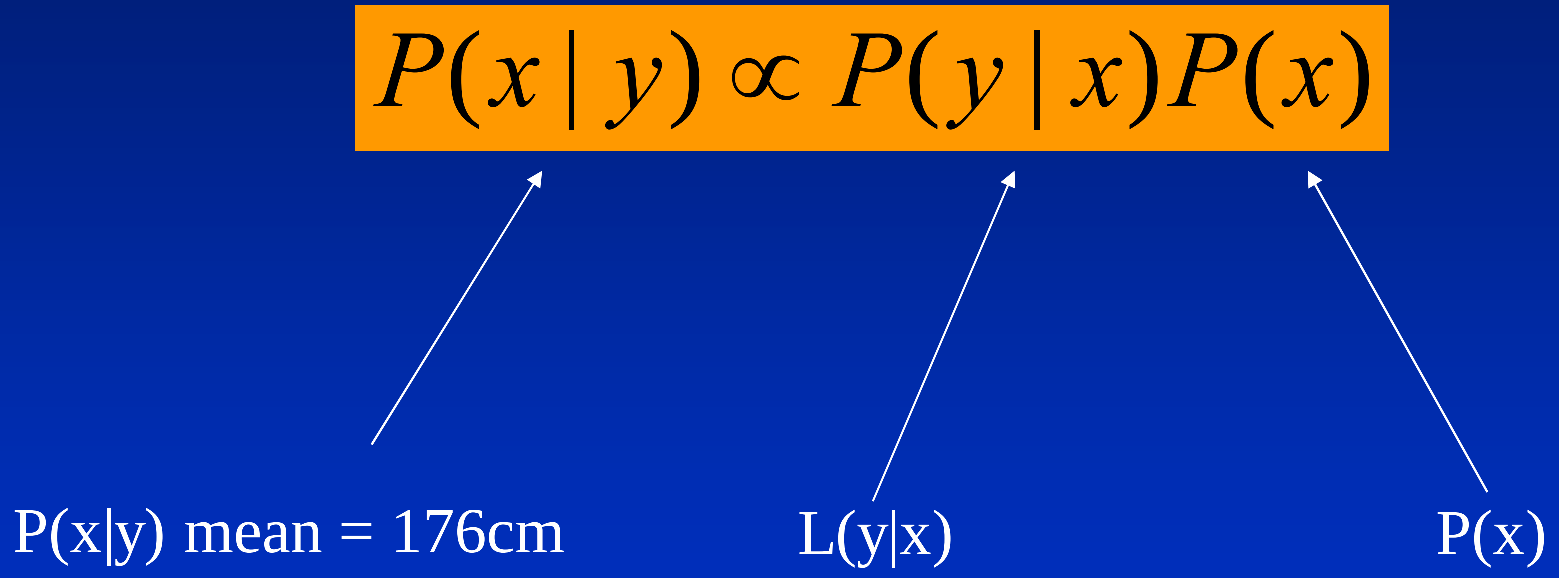
Prior probability of x (average height)





Bayesian methods

- Bayes theorem

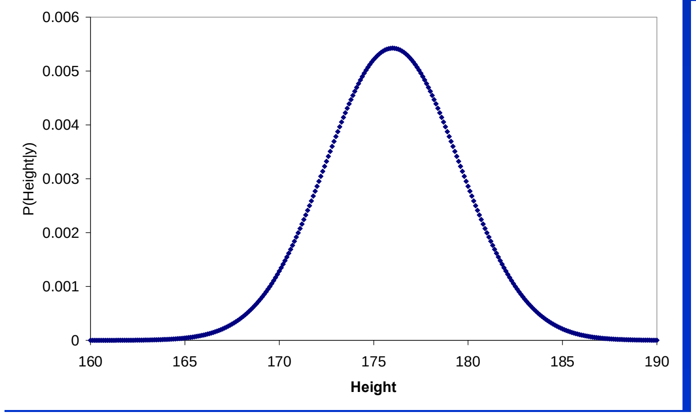


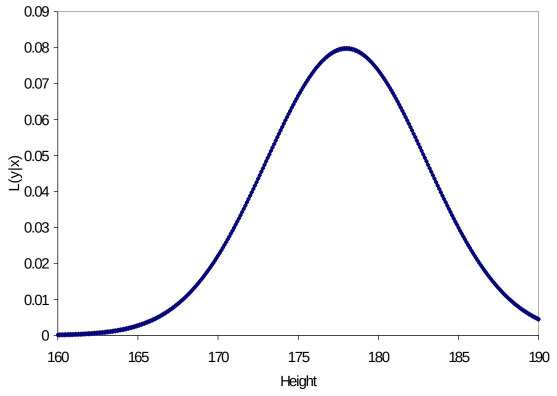
$$P(x \mid y) \propto P(y \mid x)P(x)$$

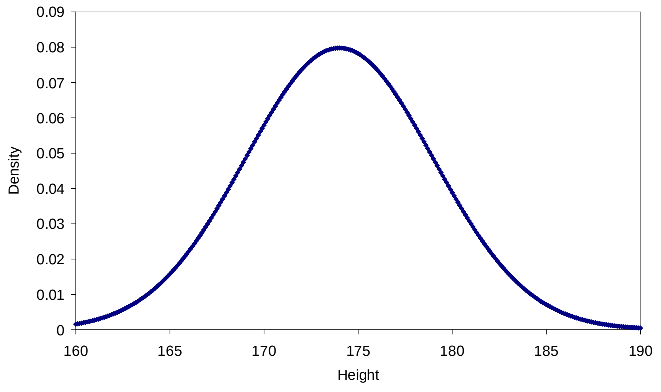
P(x|y) mean = 176cm

L(y|x)

P(x)





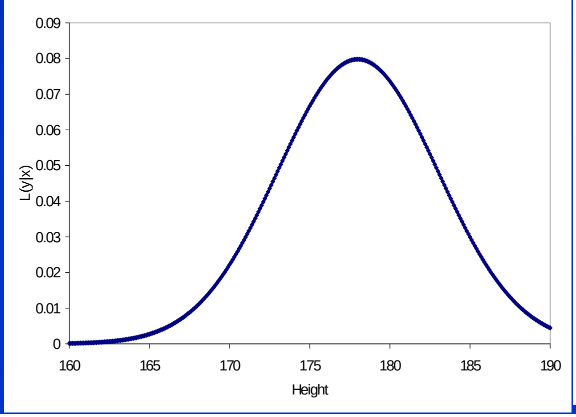


Bayesian methods

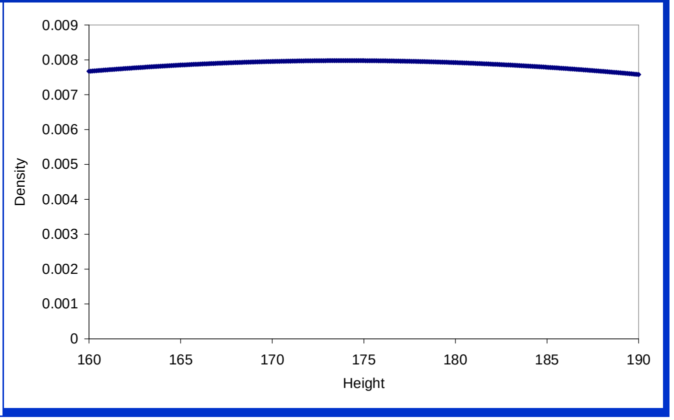
- Bayes theorem
- Less certainty about prior information? Use *less* informative (flat) prior

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

L(y|x)



P(x)



Bayesian methods

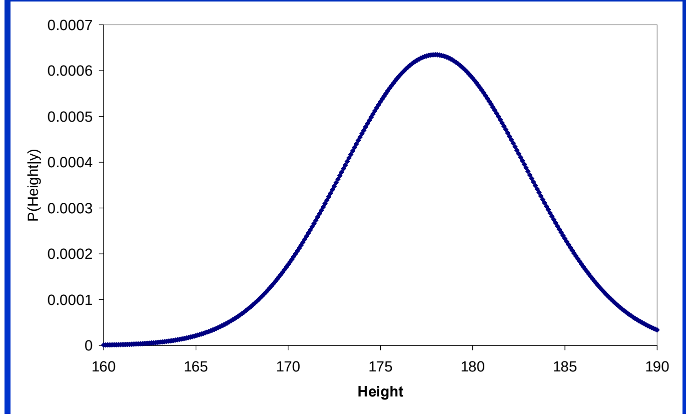
- Bayes theorem
- Less certainty about prior information? Use *less* informative (flat) prior

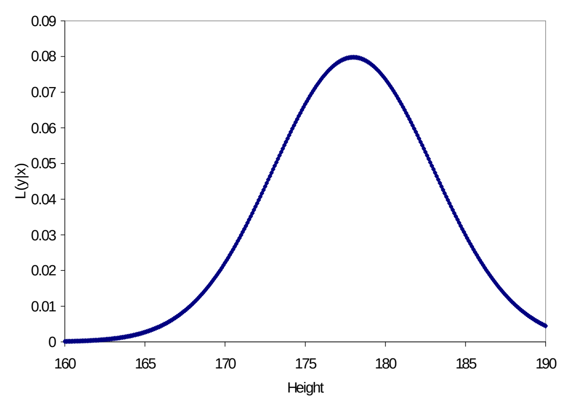
$$P(x \mid y) \propto P(y \mid x)P(x)$$

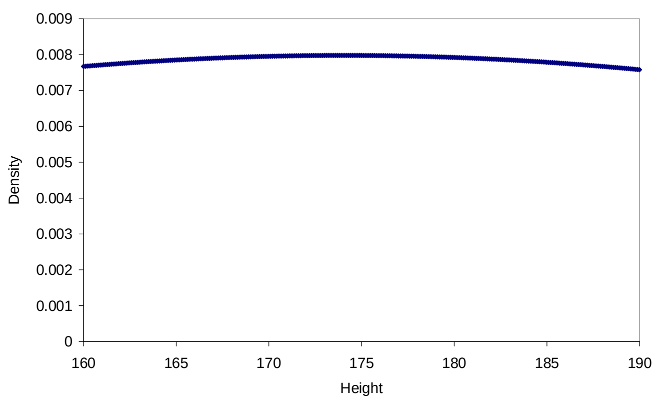
P(x|y) mean = 178cm

L(y|x)

P(x)







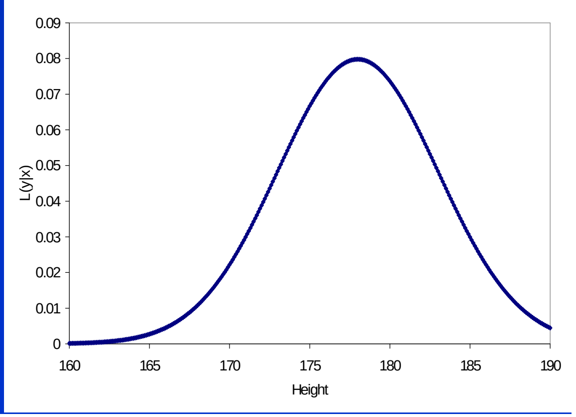
Bayesian methods

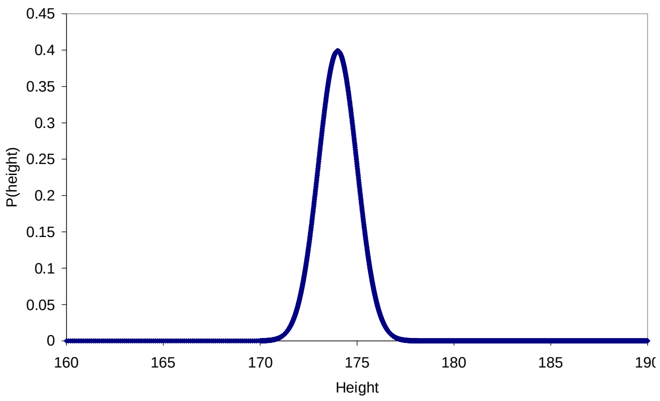
- Bayes theorem
- More certainty about prior information? Use *more* informative prior

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

L(y|x)

P(x)





Bayesian methods

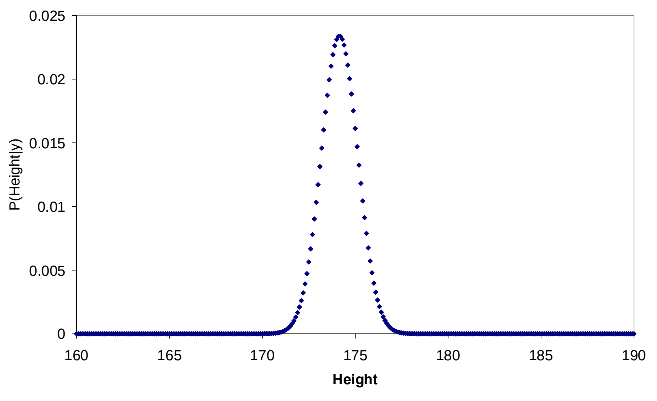
- Bayes theorem
- More certainty about prior information? Use *more* informative prior

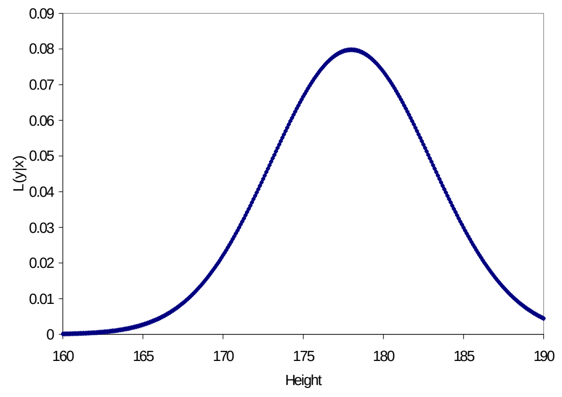
$$P(x \mid y) \propto P(y \mid x)P(x)$$

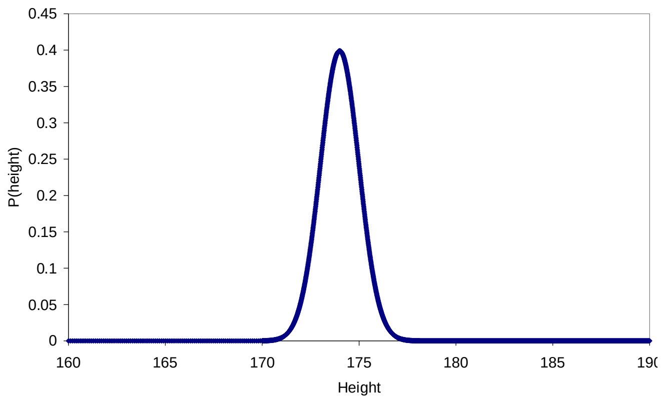
P(x|y) mean = 174.5cm

L(y|x)

P(x)





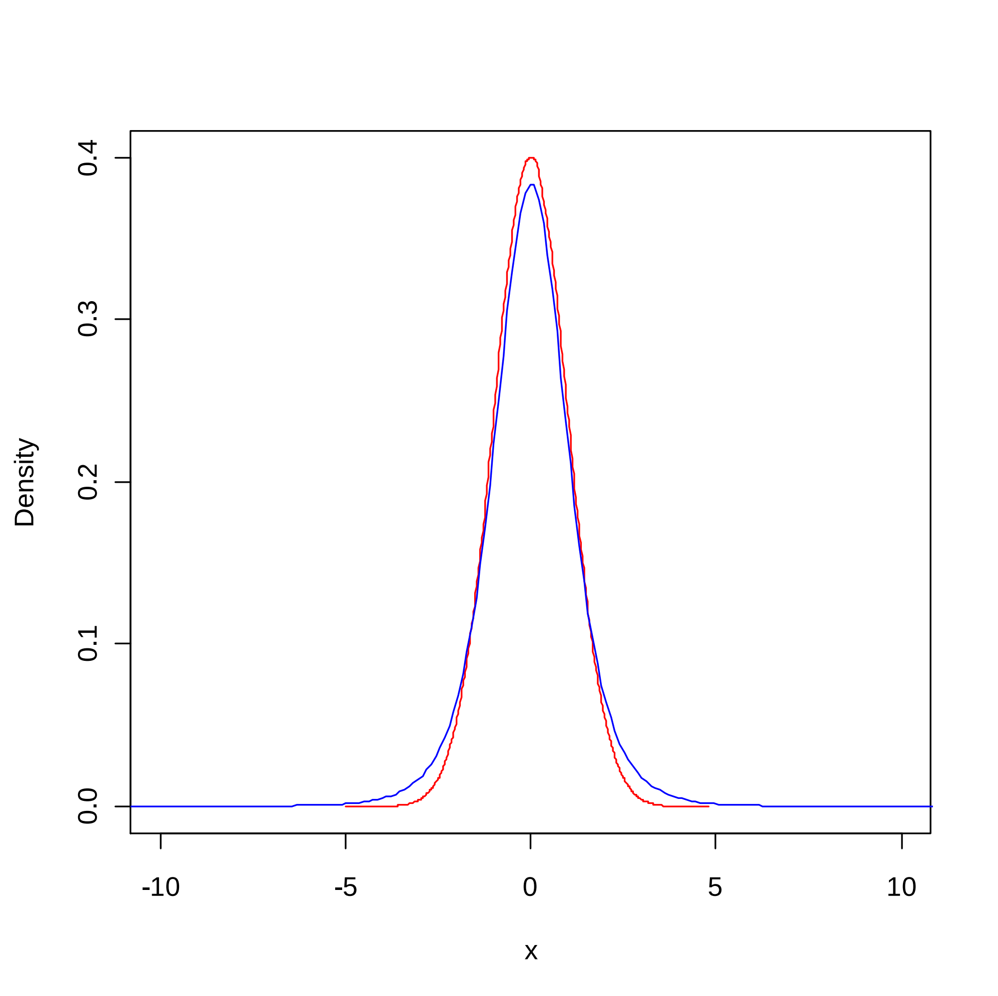


Genomic prediction

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

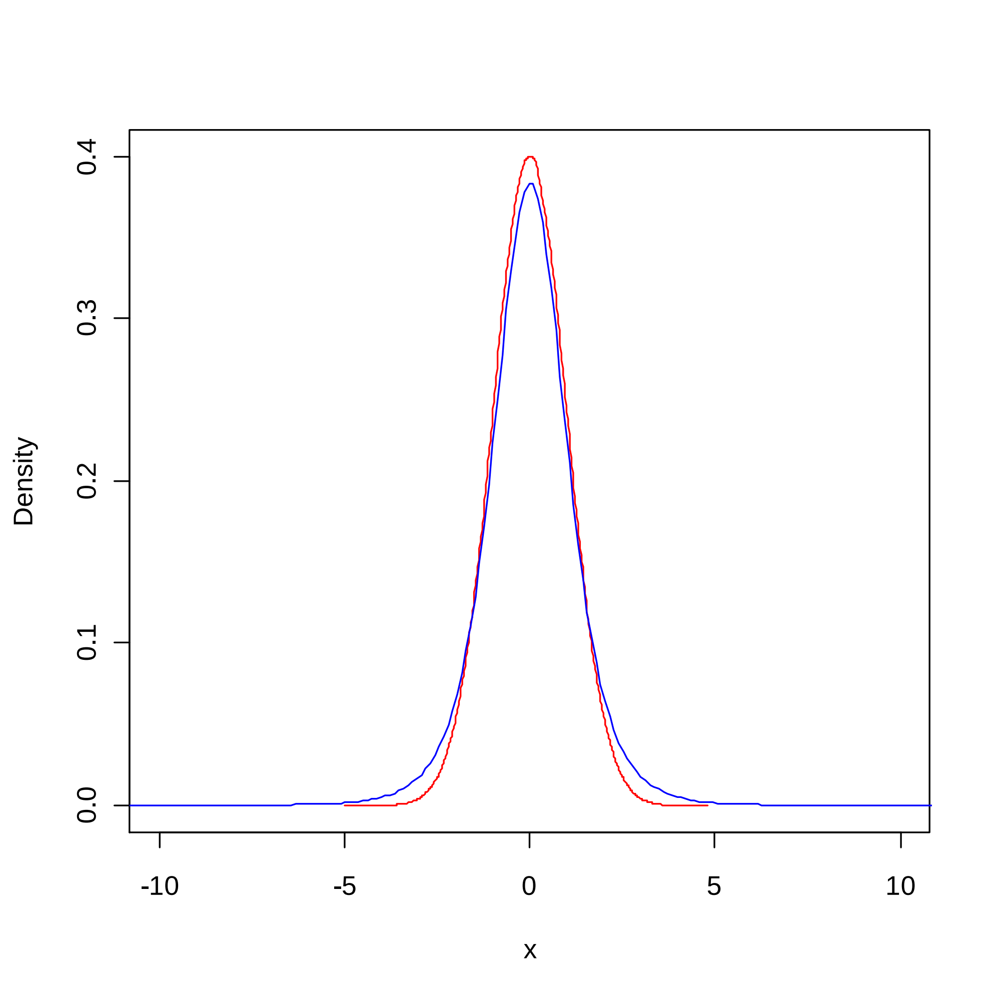
Genomic selection

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



Genomic selection

- The **t distribution** can be presented as a two level hierarchical model
- Allow different variances between SNPs (SNP specific shrinkage)
- Assume a distribution of these variances
- Computationally easier to deal with than original form

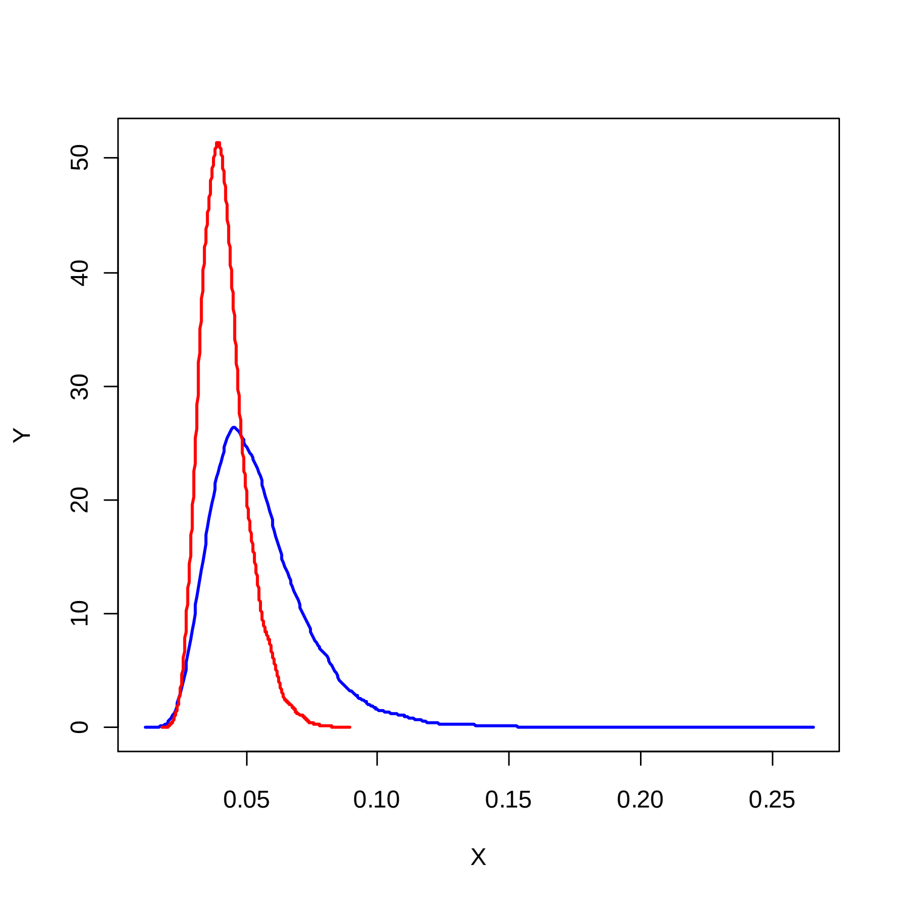


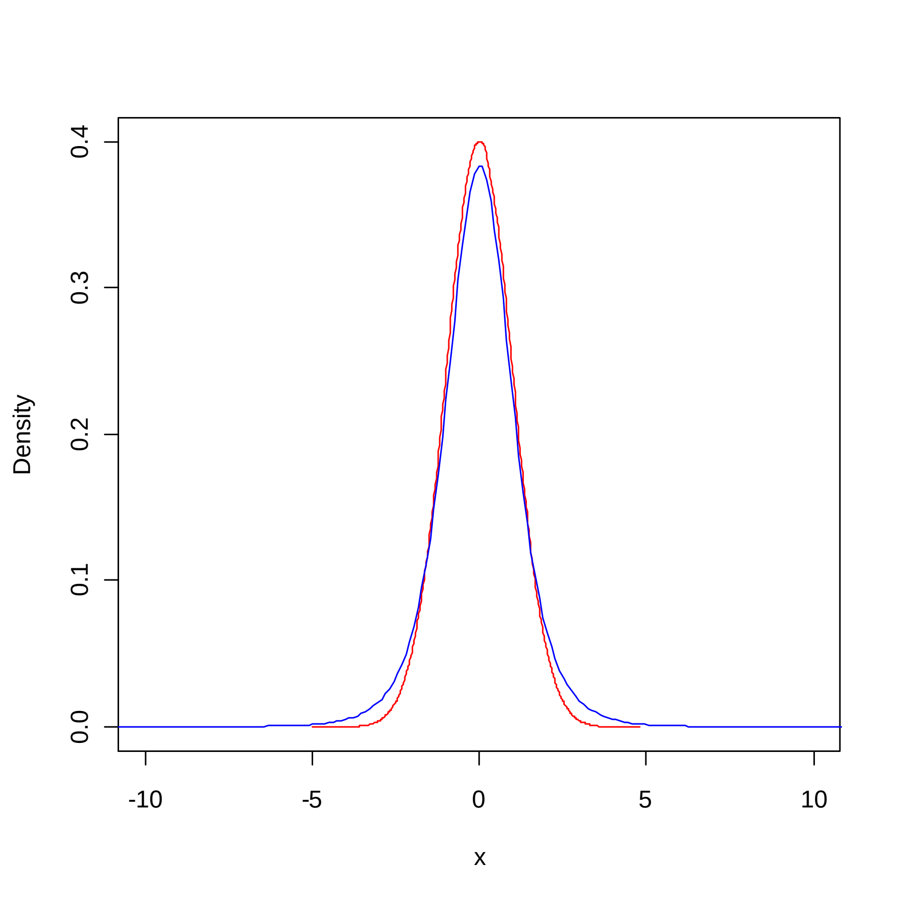
Bayesian methods

- Now lets allow different variances of SNP effects

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

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





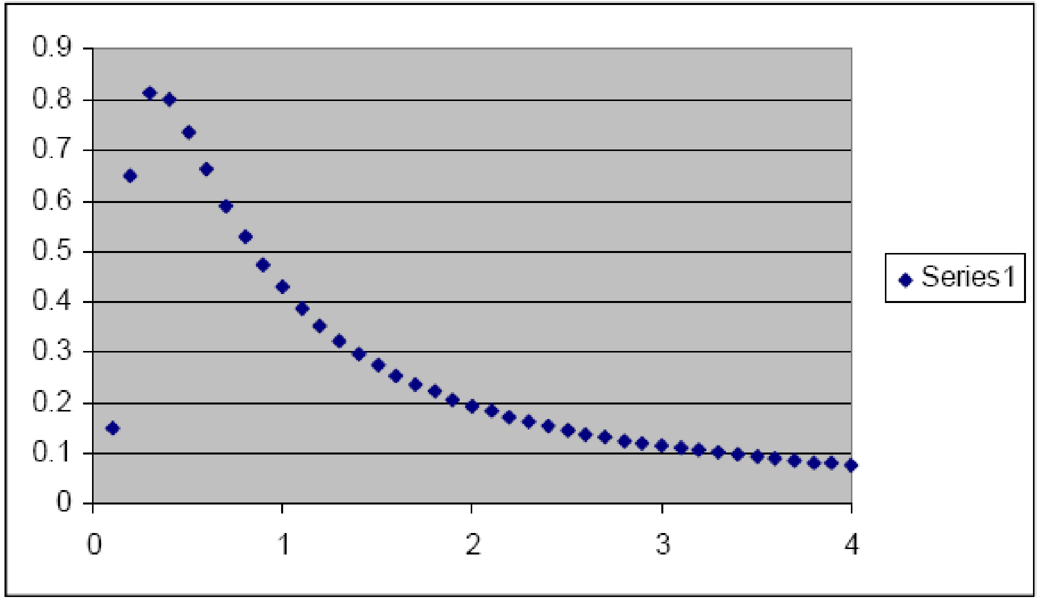
Bayesian methods

- Prior?
  - Inverted chi square convenient for variances
  - An inverted chi square with v degrees of freedom and scaled by S², eg.

$$\boxed{S^2 / \chi_v^2}$$

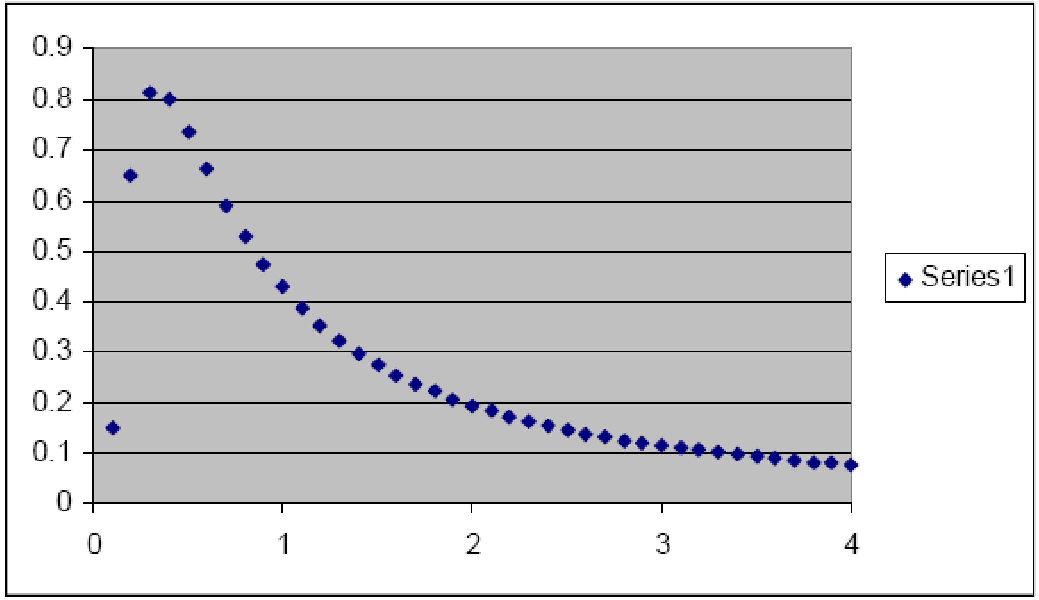
  - Describes a distribution with
    - mean $\boxed{vS^2 /(v-2)}$
    - variance $\boxed{\dfrac{2v^2 S^4}{(v-2)^2 (v-4)}}$
  - Larger v, more informative prior = more belief about variance

Bayesian methods

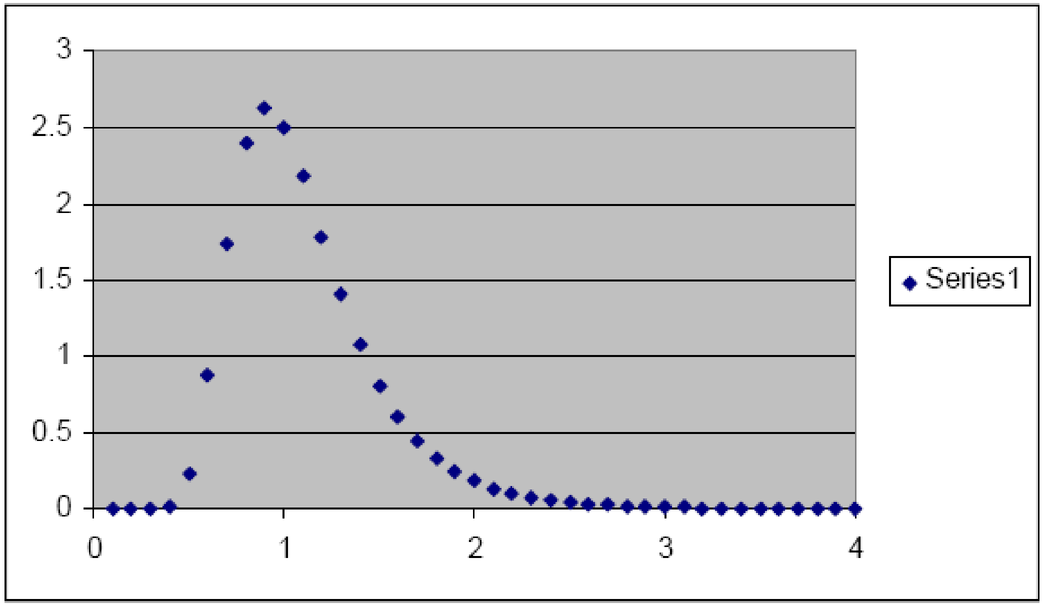


ν=2

Bayesian methods

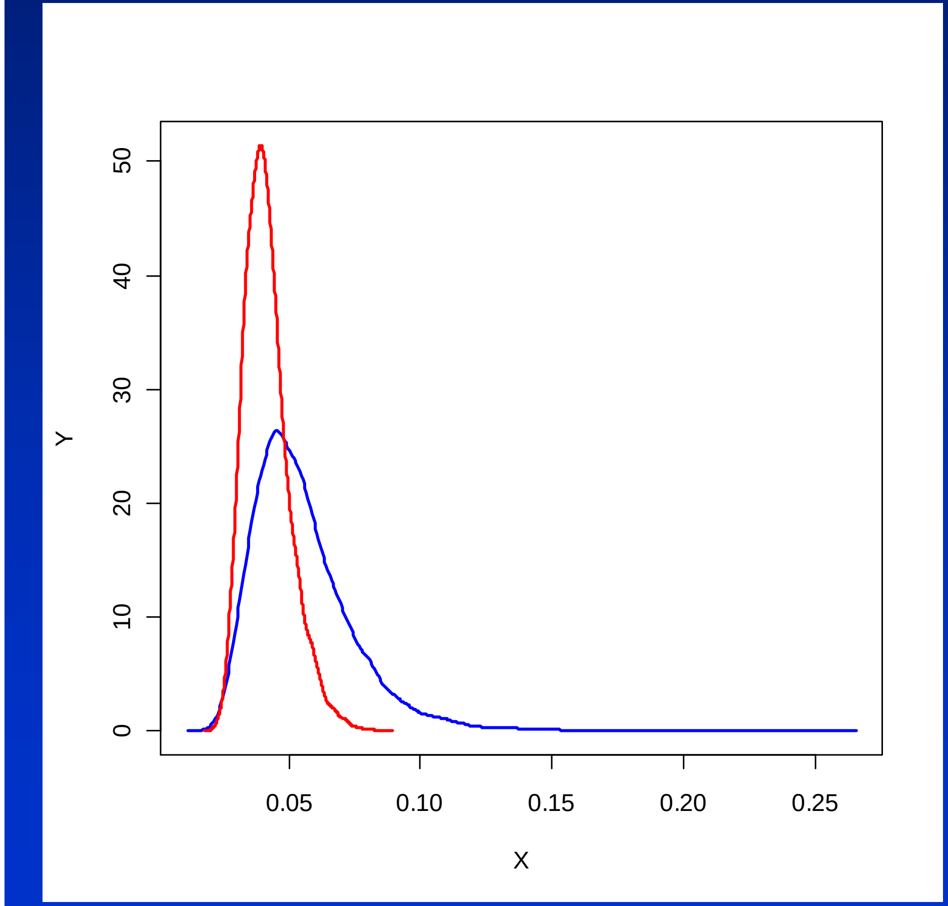


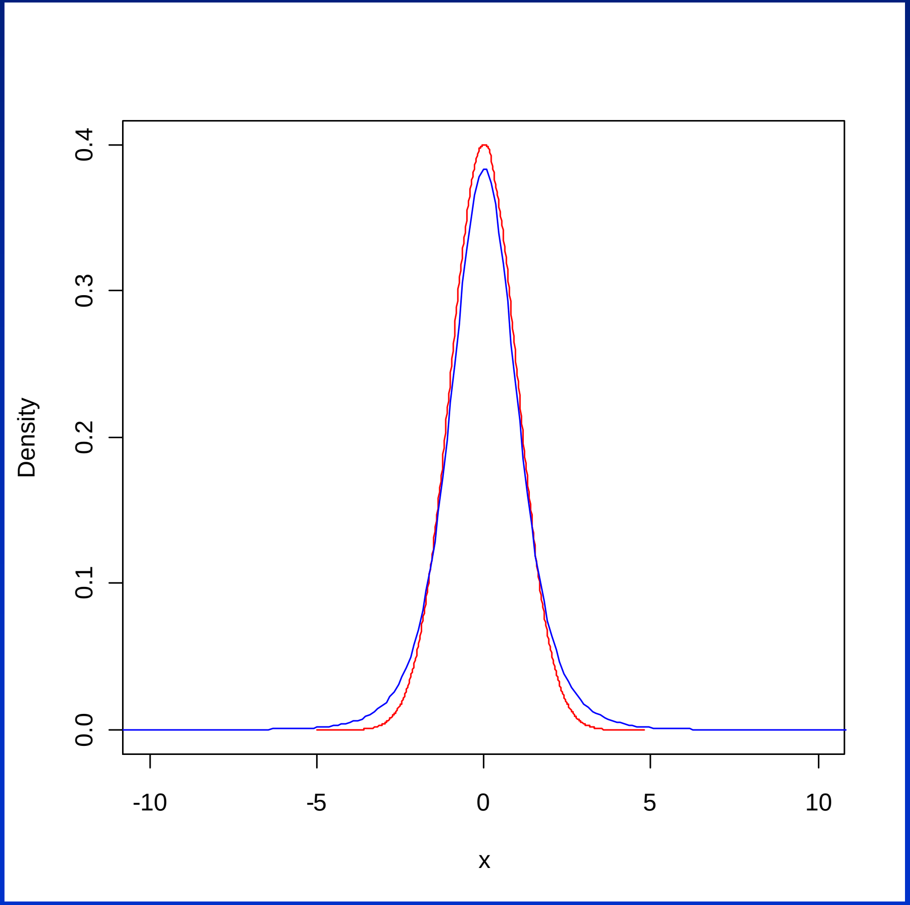
$\nu=2$



$\nu=20$

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

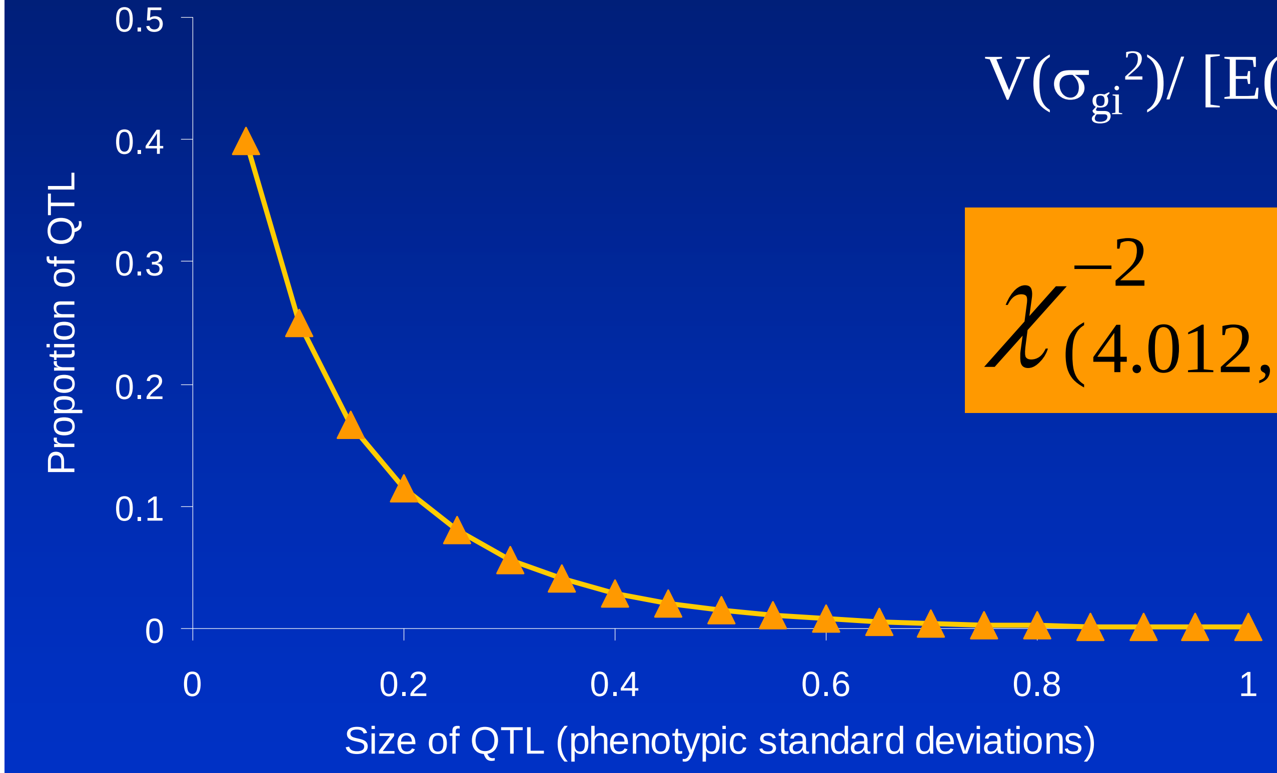




Bayesian methods

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

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



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

Bayesian methods

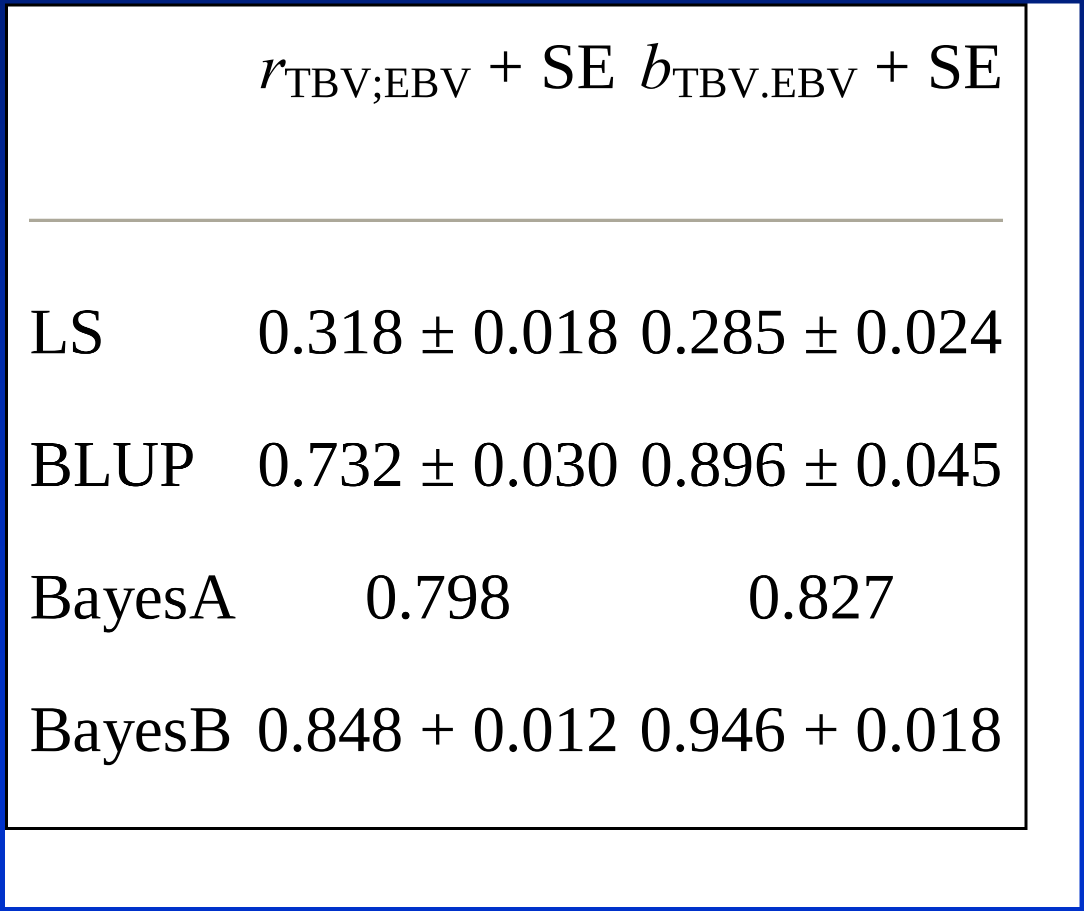
- No closed form solution
- Markov Chain Monte Carlo sampling (Gibbs, Metropolis Hastings)
- Expectation Maximisation algorithms
- Practical

Genomic prediction

- 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 prediction

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



	$r_{\mathrm{TBV;EBV}}$ + SE	$b_{\mathrm{TBV.EBV}}$ + 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 prediction

- 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 prediction

- 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 prediction

- 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 prediction

- 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

Genomic prediction

- Introduction to genomic selection
- Genomic prediction with BLUP
- Genomic prediction with Bayesian methods
- Examples in real data

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

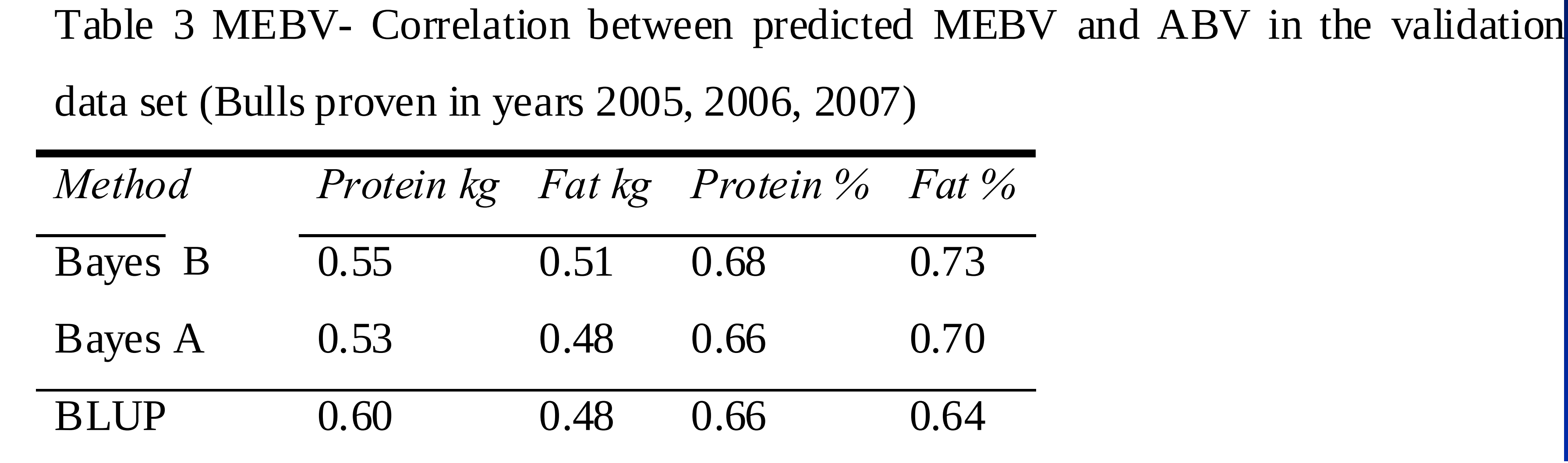
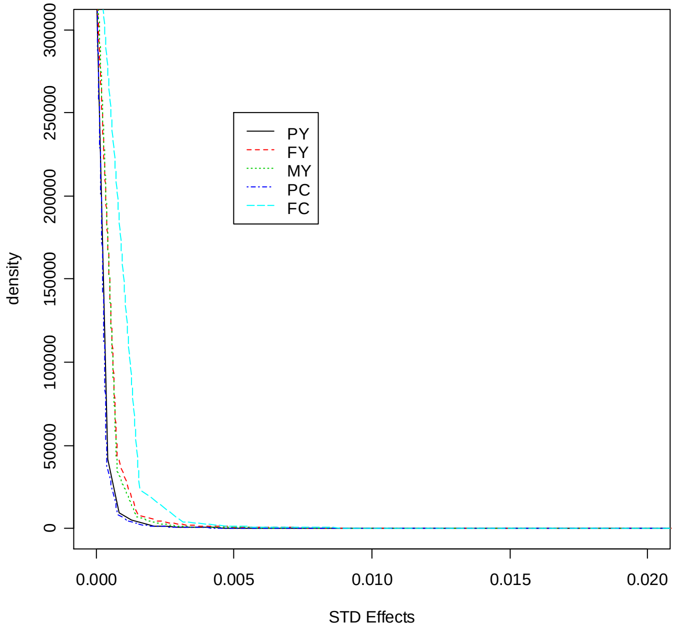
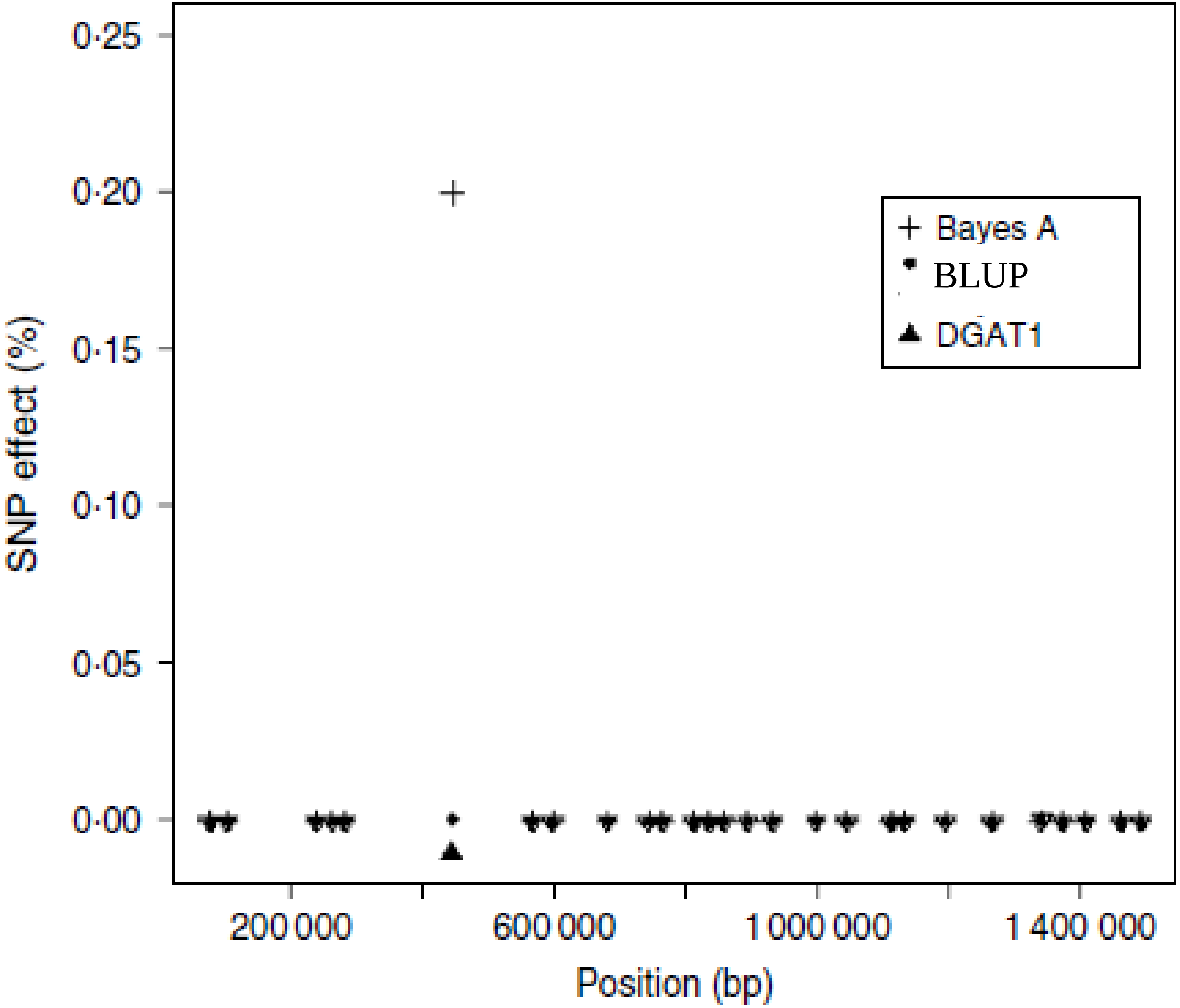


Table 3 MEBV- Correlation between predicted MEBV and ABV in the validation data set (Bulls proven in years 2005, 2006, 2007)

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 prediction

- Bayesian Cπ (Habier et al 2011)

- Two criticisms of BayesA/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
  - π is treated as known, not estimated from the data

Genomic prediction

- Bayesian CΠ (Habier et al 2011)

- Use a common σ_{gi}^{2} across all SNP in model
  - Many degrees of freedom from data
  - A “BLUP” for SNP in model

- Estimate Π from data

Bayesian methods

- Now lets allow different variances of chromosome segment effects

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

Genomic prediction

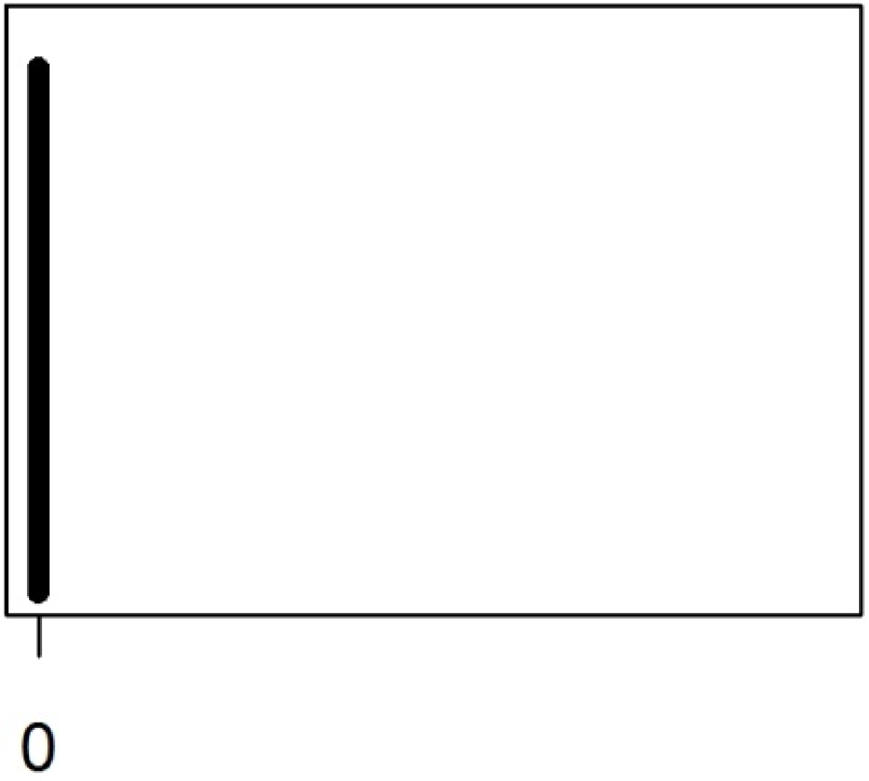
- Bayesian Cπ (Habier et al 2011)
  - Accuracy in German Holstein Friesian data set

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

- Can draw inferences about trait architecture?

BayesR

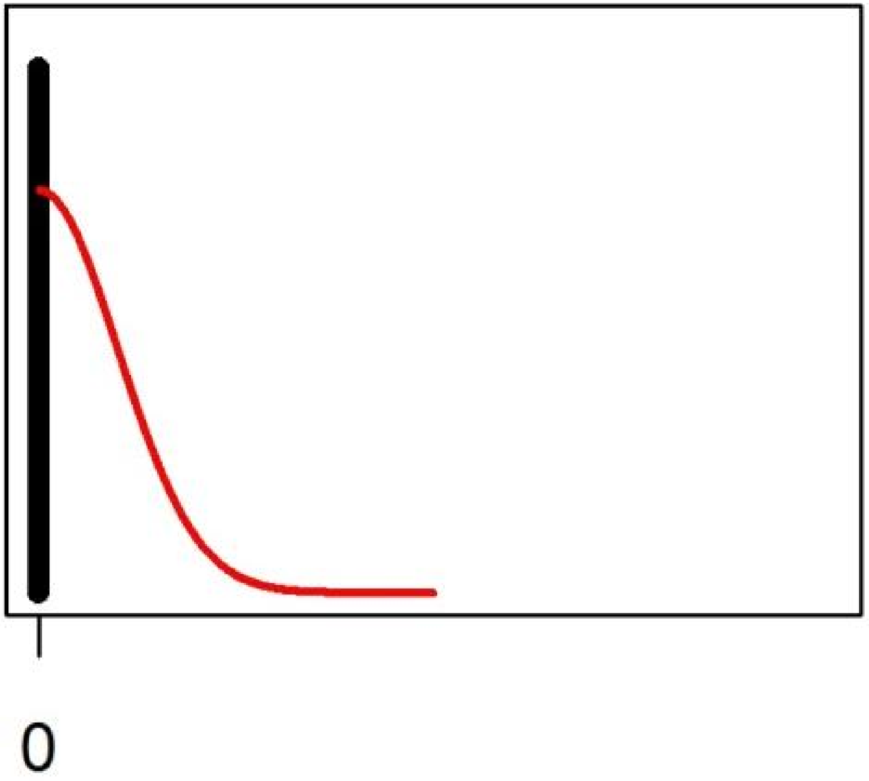
idea: SNP effects from one of four normal distributions which have different variances



Erbe et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012 Jul;95(7):4114-29.

BayesR

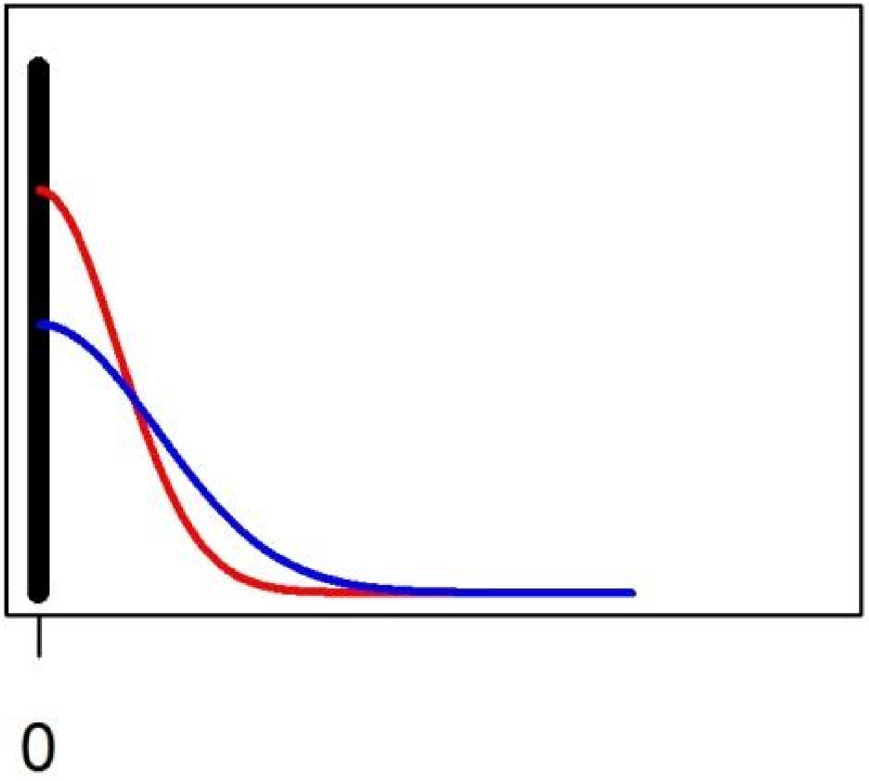
idea: SNP effects from one of four normal distributions which have different variances



Erbe et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012 Jul;95(7):4114-29.

BayesR

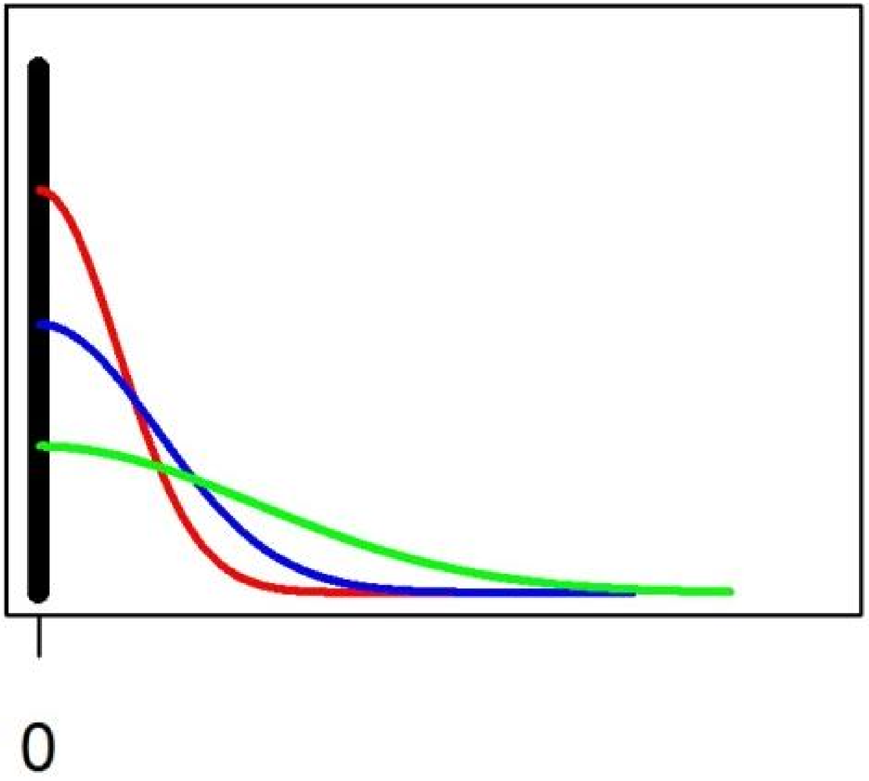
idea: SNP effects from one of four normal distributions which have different variances



Erbe et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012 Jul;95(7):4114-29.

BayesR

idea: SNP effects from one of four normal distributions which have different variances



Erbe et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012 Jul;95(7):4114-29.

Model

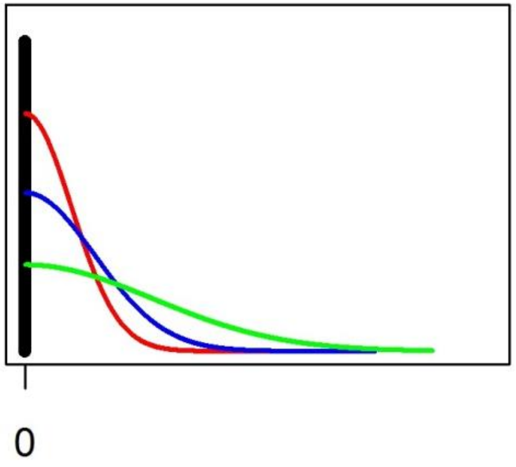
$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{W}\mathbf{g} + \mathbf{e}$$

Å y: vector of phenotypes

Å ε: overall mean

Å g: vector of SNP effects

$$\sigma^2_{g_i} = \begin{cases} 0 \cdot \sigma^2_a \textit{ with probability } p_1 \\ 0.0001 \cdot \sigma^2_a \textit{ with probability } p_2 \\ 0.001 \cdot \sigma^2_a \textit{ with probability } p_3 \\ 0.01 \cdot \sigma^2_a \textit{ with probability } p_4 \end{cases}$$



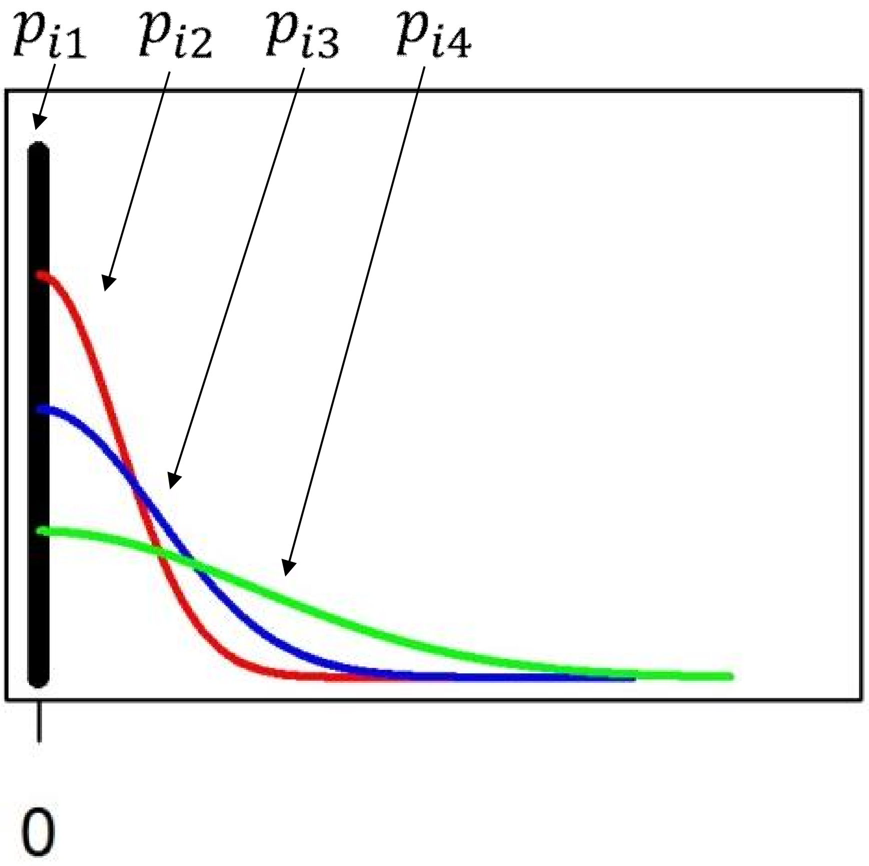
BayesR

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}}_1 \\ . \\ \hat{\mathbf{g}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n}'\mathbf{X_1} & . & \mathbf{1_n}'\mathbf{X_p} \\ \mathbf{X_1}'\mathbf{1_n} & \mathbf{X_1}'\mathbf{X_1} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_{g1}^2} & . & \mathbf{X_1}'\mathbf{X_p} \\ . & . & . & . \\ \mathbf{X_p}'\mathbf{1_n} & \mathbf{X_p}'\mathbf{X_1} & . & \mathbf{X_p}'\mathbf{X_p} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_{g4}^2} \end{bmatrix}^{-1} \begin{bmatrix} 1_n'y \\ X_1'y \\ . \\ X_p'y \end{bmatrix}$$

BayesR

Å Each SNP has a probability of being in each of the four distributions.

BayesR



BayesR

Gibbs Sampling:

Å For each SNP in each iteration, calculate likelihood of data if SNP in distribution 1, 2, 3 or 4 given BLUP estimate of effect, proportion of SNP in that distribution

Å Take SNP effect for distribution with highest likelihood, update p_{ij}, Ŭ (count of SNP in each distribution)

Å Use Dirichlet distribution to sample distribution proportions, P∼Dirichlet(Ŭ+ɓ)

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*

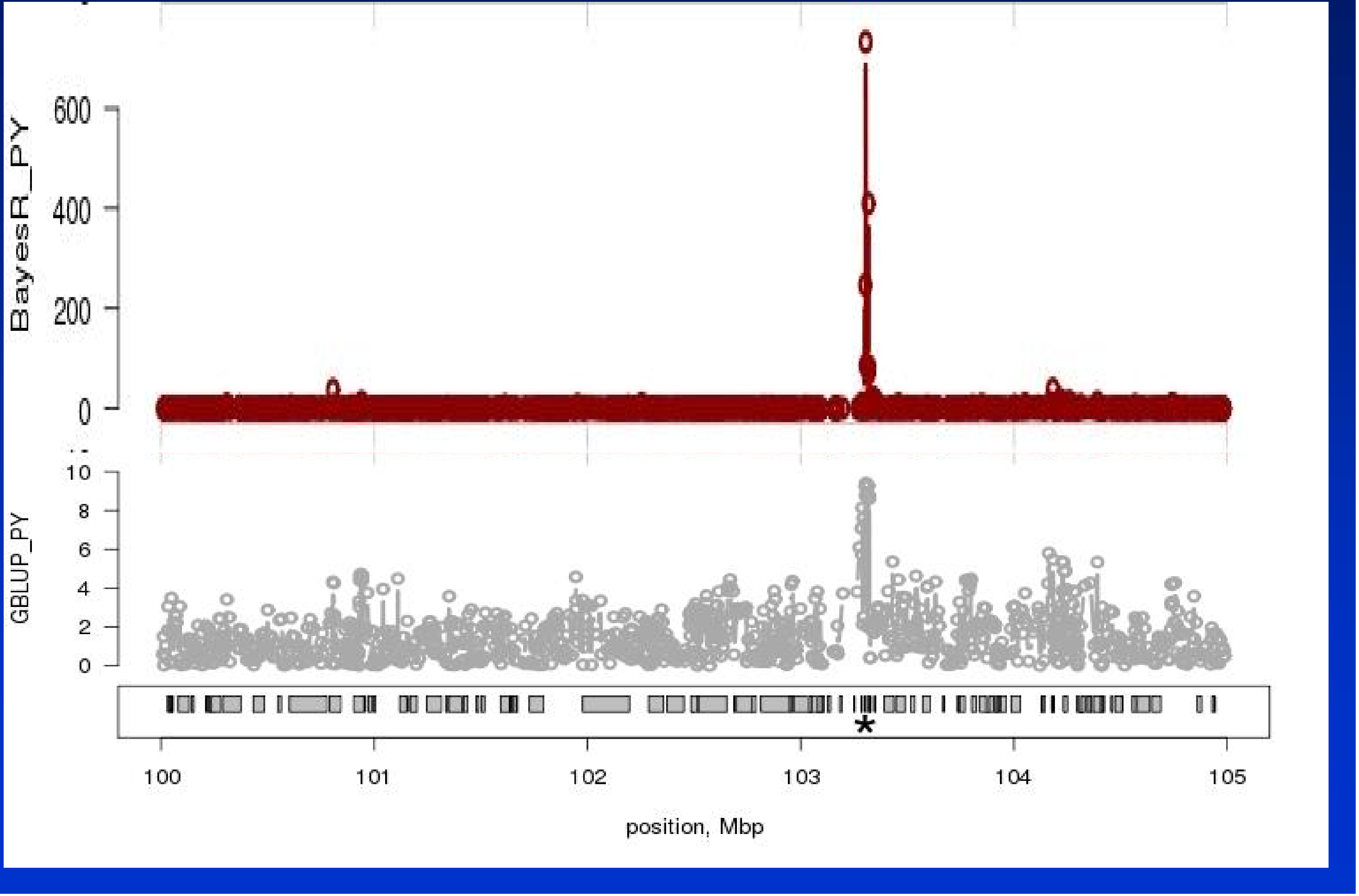
- GEBV with GBLUP, BayesR
- (Kemper et al GSE, 2014)

Real Data, 800K

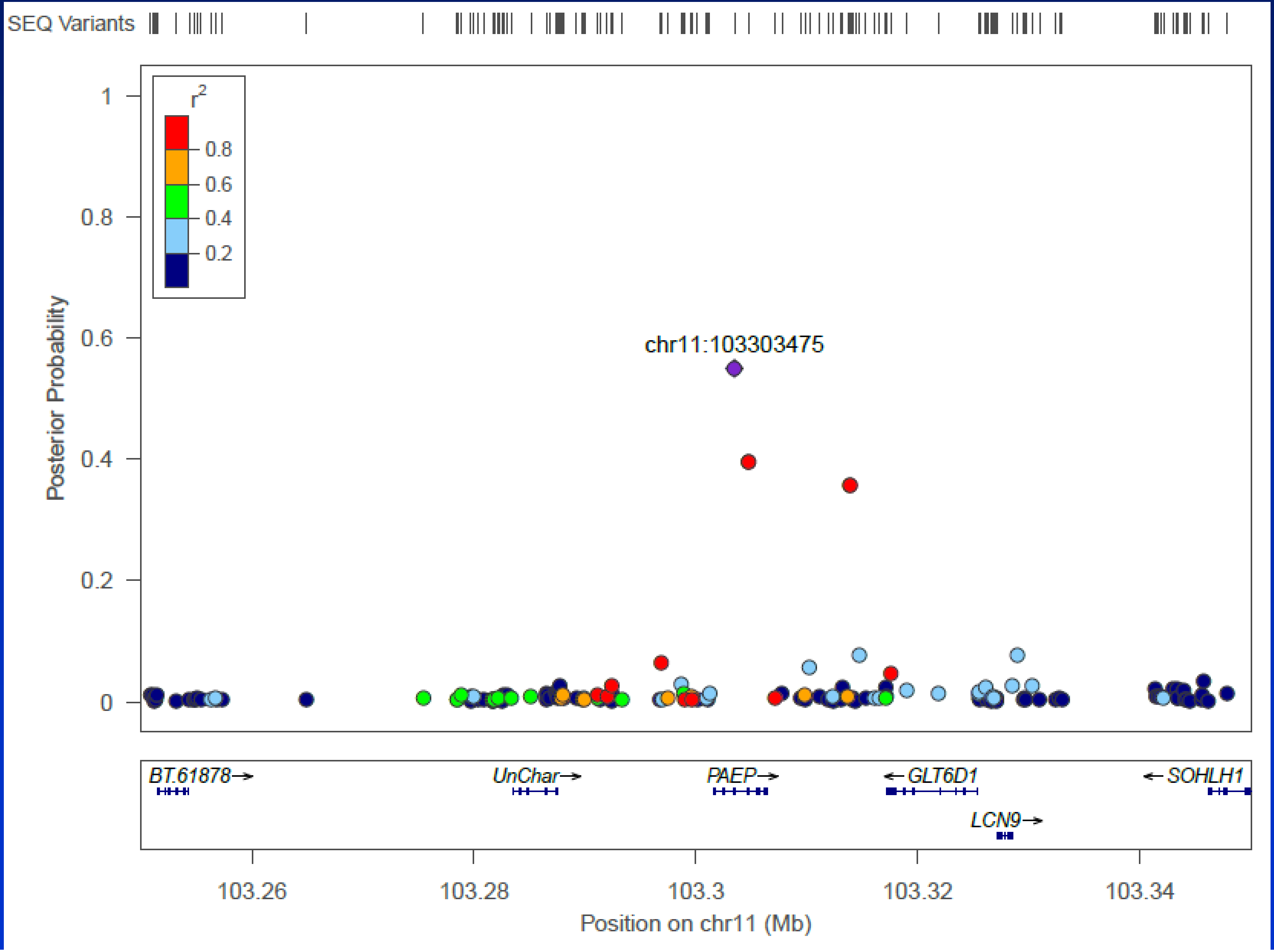
- r(GEBV,DTD)

	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**
BAYESR	0.26	0.21	0.13	0.44	0.36	**0.28**

BayesR


BayesR_PY
600
400
200
0
GBLUP_PY
10
8
6
4
2
0
*
100
101
102
103
104
105
position, Mbp

BayesR -> QTL mapping



BayesR

https://github.com/syntheke/bayesR

Genomic prediction

- 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 ∏ from data, common variance across SNP
  - BayesR = multiple normal distributions

Genomic prediction

- Bayesian methods can have an advantage when:

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

- Very large numbers of SNP (eg 800K) (but need large reference sets) – set some SNP effects to zero

- Multi-breed, across breed genomic predictions