Summary of Methods

Various Methods

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum \mathbf{M}_i\mathbf{a}_i + \mathbf{e}$$

$estimate\ \sigma^2_{ai}\ and\ \sigma^2_e$

BayesA

Various Methods

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$estimate\ \sigma_{ai}^2\ and\ \sigma_e^2$$

BayesA

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i \delta_i + \mathbf{e}$$

$$estimate\ \delta_i,\ \sigma_{ai}^2\ and\ \sigma_e^2$$

BayesB

Various Methods

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$estimate\ \sigma^2_{ai}\ and\ \sigma^2_e$$

BayesA

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i \delta_i + \mathbf{e}$$

$$estimate\ \delta_i,\ \sigma^2_{ai}\ and\ \sigma^2_e$$

BayesB

$$estimate\ \delta_i,\ \sigma^2_{a}\ and\ \sigma^2_e$$

BayesC

Various Methods

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$estimate\ \sigma_{ai}^2\ and\ \sigma_e^2$$

BayesA

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i \delta_i + \mathbf{e}$$

$$estimate\ \delta_i,\ \sigma_{ai}^2\ and\ \sigma_e^2$$

BayesB

$$estimate\ \delta_i,\ \sigma_a^2\ and\ \sigma_e^2$$

BayesC

$$estimate\ \pi,\ \ \delta_i,\ \ \sigma_a^2\ and\ \sigma_e^2$$

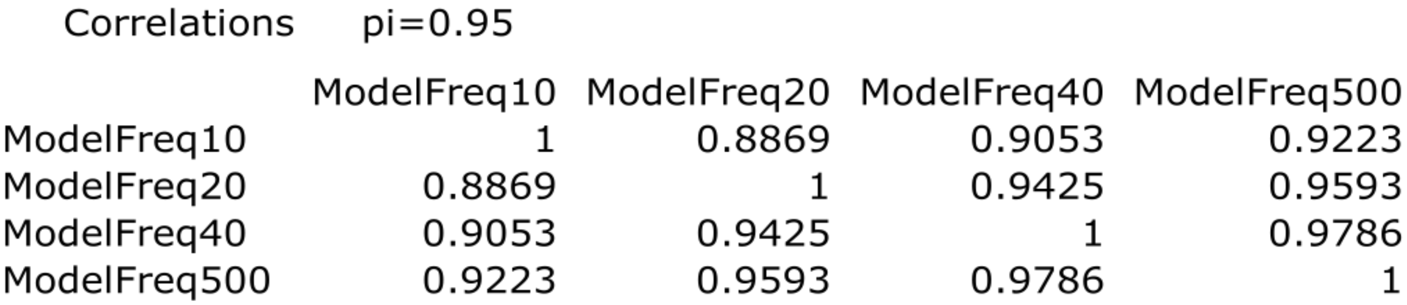
BayesCPi

Various Methods

	Markers in Model	
Marker Effects	All (π=0)	Fraction (1-π)
Random - Individual Variance (Normal)	“Bayes A” (B0)	“Bayes B”
Random - Constant Var (when in model)	Bayes C (C0)=“BLUP”	Bayes C
Random – Constant Var (when in model)		Fraction (1-π) estimated from data=Bayes CPi
Categorical Variants (threshold models)		
Other Variants (estimate scale, heavy tails)		

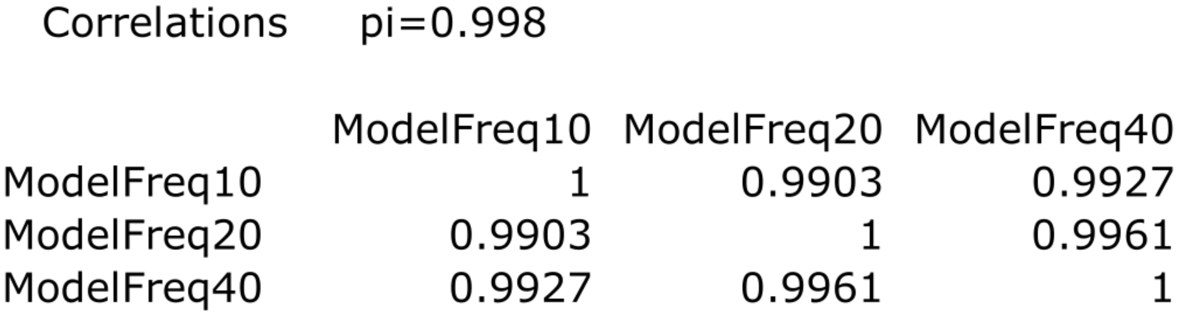
Practical experience and results with various methods using real and simulated data

Pi influences convergence



Correlations pi=0.95

	ModelFreq10	ModelFreq20	ModelFreq40	ModelFreq500
ModelFreq10	1	0.8869	0.9053	0.9223
ModelFreq20	0.8869	1	0.9425	0.9593
ModelFreq40	0.9053	0.9425	1	0.9786
ModelFreq500	0.9223	0.9593	0.9786	1



Correlations pi=0.998

	ModelFreq10	ModelFreq20	ModelFreq40
ModelFreq10	1	0.9903	0.9927
ModelFreq20	0.9903	1	0.9961
ModelFreq40	0.9927	0.9961	1

Genomic Selection

Shrinkage of marker effects

Dorian Garrick

dorian@iastate.edu

ANIMAL SCIENCE

1858 150 2008
IOWA STATE UNIVERSITY

Animal Breeding & Genetics

Simplest Approach

No selection of loci

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$constant\ \sigma_a^2\ and\ \sigma_e^2$$

$$"BLUP"$$

Aassume
normally distributed
- allelic effects
- residual effects

Mixed Model Equations

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Ma} + \mathbf{e}$$

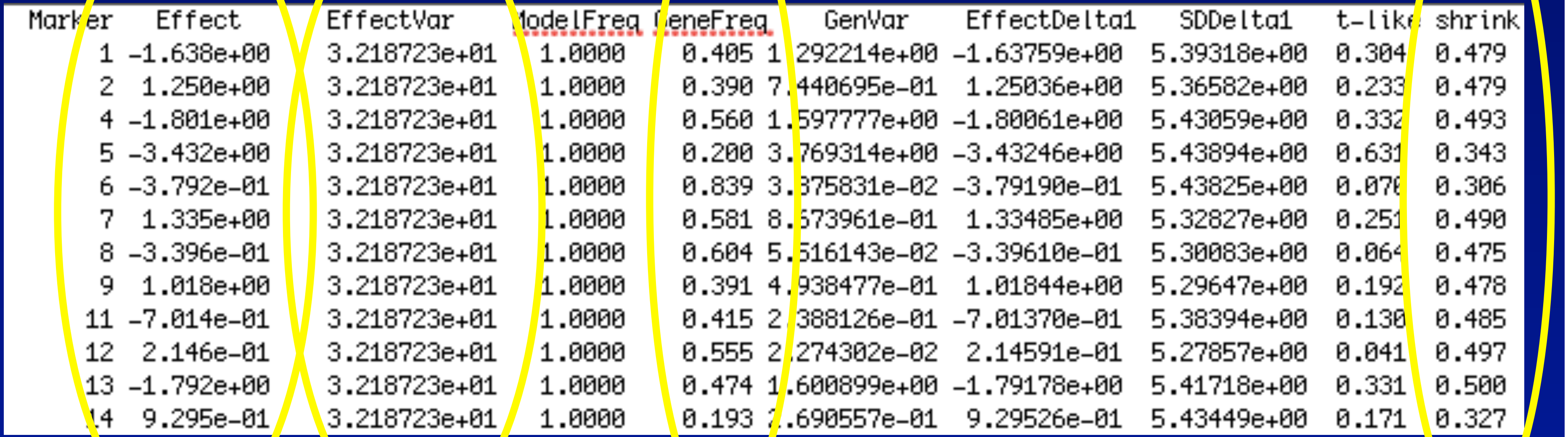
$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'M} \\ \mathbf{M'X} & \mathbf{M'M} + \lambda\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{M'y} \end{bmatrix}$$

$\lambda = \dfrac{\sigma_e^2}{\sigma_a^2}$ *is an unknown that can be estimated eg REML*

These equations have order = number of SNP+1 and are dense

Like Ridge Regression

Estimated Effects



```
Marker    Effect      EffectVar    ModelFreq GeneFreq     GenVar      EffectDelta1    SDDelta1     t-like shrink
     1 -1.638e+00  3.218723e+01  1.0000     0.405 1.292214e+00 -1.63759e+00  5.39318e+00  0.304  0.479
     2  1.250e+00  3.218723e+01  1.0000     0.390 7.440695e-01  1.25036e+00  5.36582e+00  0.233  0.479
     4 -1.801e+00  3.218723e+01  1.0000     0.560 1.597777e+00 -1.80061e+00  5.43059e+00  0.332  0.493
     5 -3.432e+00  3.218723e+01  1.0000     0.200 3.769314e+00 -3.43246e+00  5.43894e+00  0.631  0.343
     6 -3.792e-01  3.218723e+01  1.0000     0.839 3.875831e-02 -3.79190e-01  5.43825e+00  0.070  0.306
     7  1.335e+00  3.218723e+01  1.0000     0.581 8.673961e-01  1.33485e+00  5.32827e+00  0.251  0.490
     8 -3.396e-01  3.218723e+01  1.0000     0.604 5.516143e-02 -3.39610e-01  5.30083e+00  0.064  0.475
     9  1.018e+00  3.218723e+01  1.0000     0.391 4.938477e-01  1.01844e+00  5.29647e+00  0.192  0.478
    11 -7.014e-01  3.218723e+01  1.0000     0.415 2.388126e-01 -7.01370e-01  5.38394e+00  0.130  0.485
    12  2.146e-01  3.218723e+01  1.0000     0.555 2.274302e-02  2.14591e-01  5.27857e+00  0.041  0.497
    13 -1.792e+00  3.218723e+01  1.0000     0.474 1.600899e+00 -1.79178e+00  5.41718e+00  0.331  0.500
    14  9.295e-01  3.218723e+01  1.0000     0.193 2.690557e-01  9.29526e-01  5.43449e+00  0.171  0.327
```

$\hat{\mathbf{a}}$

σ_a^2

$2pq$

$$Shrinkage = \frac{BLUP\ estimate}{OLS\ estimate}$$

Equivalent Model (All SNPs)

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$\mathbf{y} = \mathbf{Xb} + [\mathbf{I}]\left[\sum \mathbf{M}_i \mathbf{a}_i\right] + \mathbf{e}, \quad \mathbf{u} = \sum \mathbf{M}_i \mathbf{a}_i$$

$$\mathrm{var}(\sum M_i a_i) = \sum M_i \,\mathrm{var}(a_i) M_i' = \sigma_a^2 \sum M_i M_i'$$

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'} \\ \mathbf{X} & \mathbf{I} + \lambda \mathbf{G}^{\text{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{y} \end{bmatrix}$$

Current method using genomic G instead of pedigree A

$$\mathbf{G} = \sum M_i M_i'$$

Analytical Methods

No selection of loci

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$constant\ \sigma_a^2\ and\ \sigma_e^2$$

$$"BLUP"$$

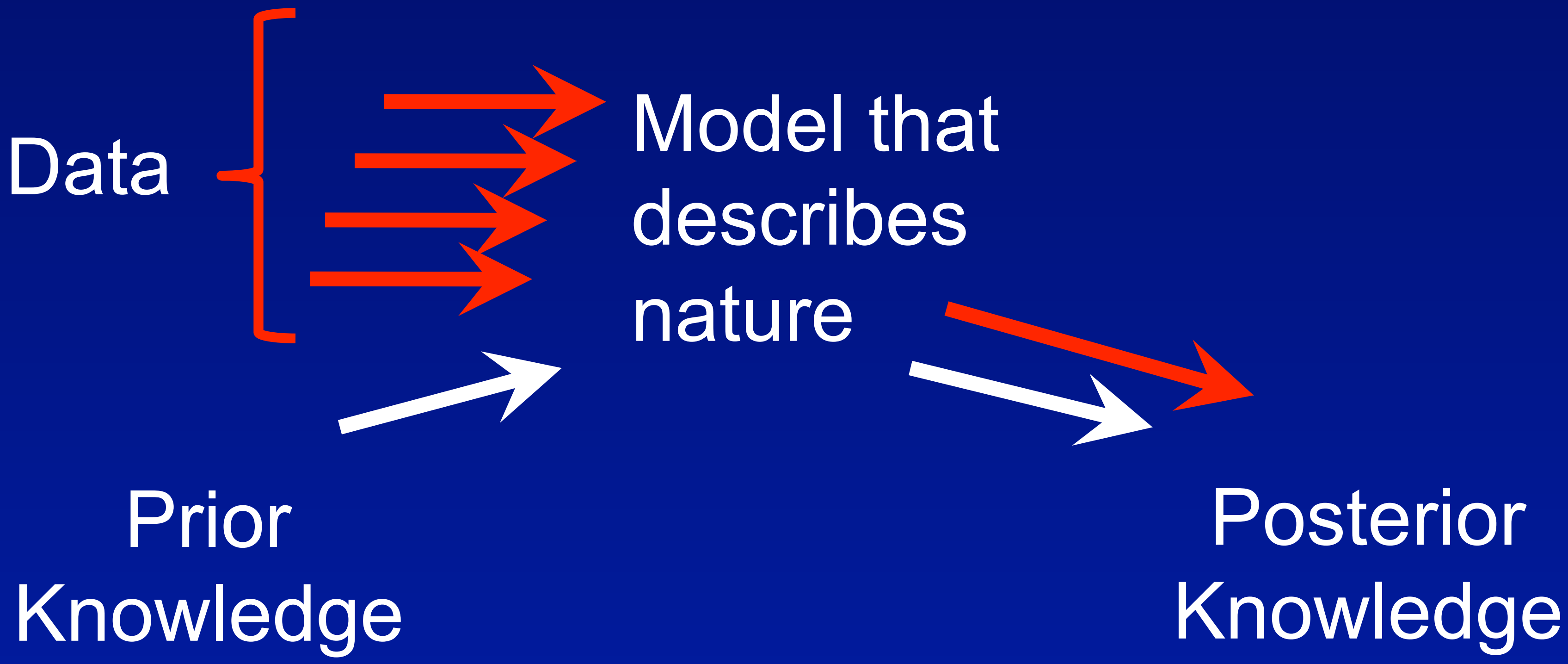
$$SNP - specific\ \sigma_{ai}^2\ and\ \sigma_e^2$$

$$BayesA$$

Need to estimate a variance component for every locus
Markov Chain Monte Carlo
is an efficient method to explore the likelihood surface

Meuwissen, Hayes & Goddard (2001)

Bayesian Methods



Data

Model that describes nature

Prior Knowledge

Posterior Knowledge

Markov Chain Monte Carlo

- Sample unknown parameters based on knowledge of the prior
- Quantify the fit (given the data)
- Sample unknown parameters based on joint knowledge of the prior and the previous fit of each parameter
- Repeat this process until convergence

Bayes A

Prior $\left(a_i/\sigma_i^2\right) \sim N\left(0,\sigma_i^2\right)$

$$\sigma_i^2 \sim v_a S_{v_a}^2 \chi_{v_a}^{-2}$$

Meuwissen, Hayes & Goddard (2001)

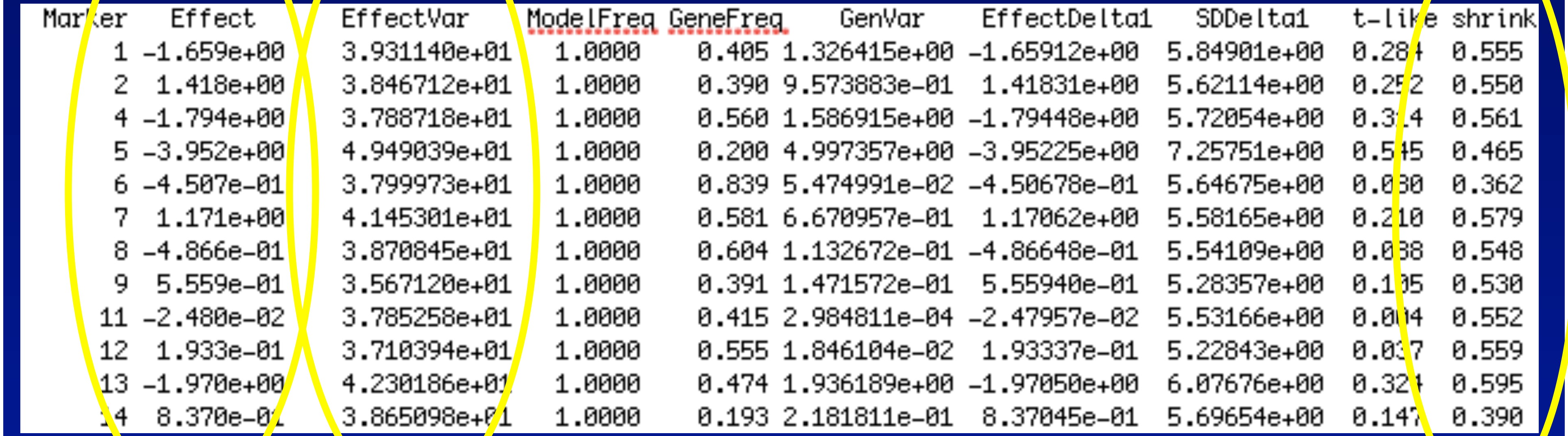
$$so\ that\ \ a_i \sim (iid)t\left(0,S_{v_a}^2,v_a\right)$$

Sorensen & Gianola, 2002

$$Assume\quad \sigma_i^2 = \frac{V_a}{\sum_i 2p_i(1-p_i)} = \frac{V_a}{k2\overline{p}(1-\overline{p})}$$

$$so\quad S_{v_a}^2 = \frac{\left(v_a-2\right)V_a}{v_a k 2\overline{p}(1-\overline{p})}\ for\ k\ SNP$$

8,300 Holstein Bulls w/50k



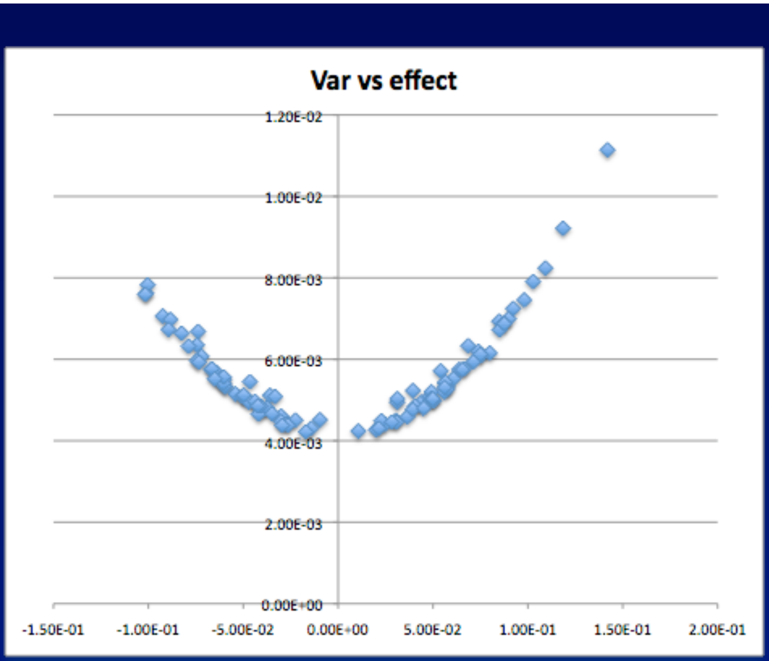
```
Marker    Effect       EffectVar  ModelFreq GeneFreq     GenVar    EffectDelta1    SDDelta1   t-like shrink
     1 -1.659e+00   3.931140e+01   1.0000     0.405 1.326415e+00 -1.65912e+00 5.84901e+00 0.284  0.555
     2  1.418e+00   3.846712e+01   1.0000     0.390 9.573883e-01  1.41831e+00 5.62114e+00 0.252  0.550
     4 -1.794e+00   3.788718e+01   1.0000     0.560 1.586915e+00 -1.79448e+00 5.72054e+00 0.314  0.561
     5 -3.952e+00   4.949039e+01   1.0000     0.200 4.997357e+00 -3.95225e+00 7.25751e+00 0.545  0.465
     6 -4.507e-01   3.799973e+01   1.0000     0.839 5.474991e-02 -4.50678e-01 5.64675e+00 0.080  0.362
     7  1.171e+00   4.145301e+01   1.0000     0.581 6.670957e-01  1.17062e+00 5.58165e+00 0.210  0.579
     8 -4.866e-01   3.870845e+01   1.0000     0.604 1.132672e-01 -4.86648e-01 5.54109e+00 0.088  0.548
     9  5.559e-01   3.567120e+01   1.0000     0.391 1.471572e-01  5.55940e-01 5.28357e+00 0.105  0.530
    11 -2.480e-02   3.785258e+01   1.0000     0.415 2.984811e-04 -2.47957e-02 5.53166e+00 0.004  0.552
    12  1.933e-01   3.710394e+01   1.0000     0.555 1.846104e-02  1.93337e-01 5.22843e+00 0.037  0.559
    13 -1.970e+00   4.230186e+01   1.0000     0.474 1.936189e+00 -1.97050e+00 6.07676e+00 0.324  0.595
    14  8.370e-01   3.865098e+01   1.0000     0.193 2.181811e-01  8.37045e-01 5.69654e+00 0.147  0.390
```

σ_a^2

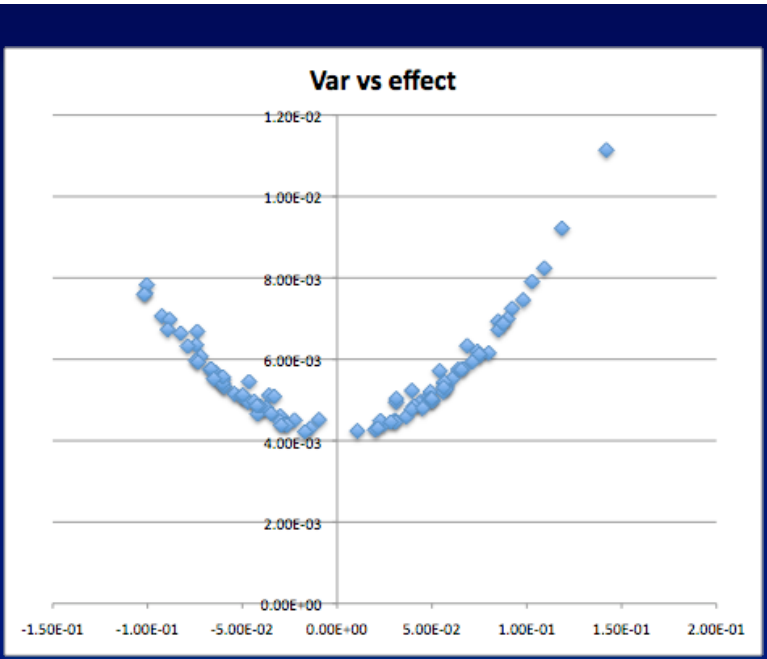
$$Shrinkage = \frac{BLUP\ estimate}{OLS\ estimate}$$

Bayes A

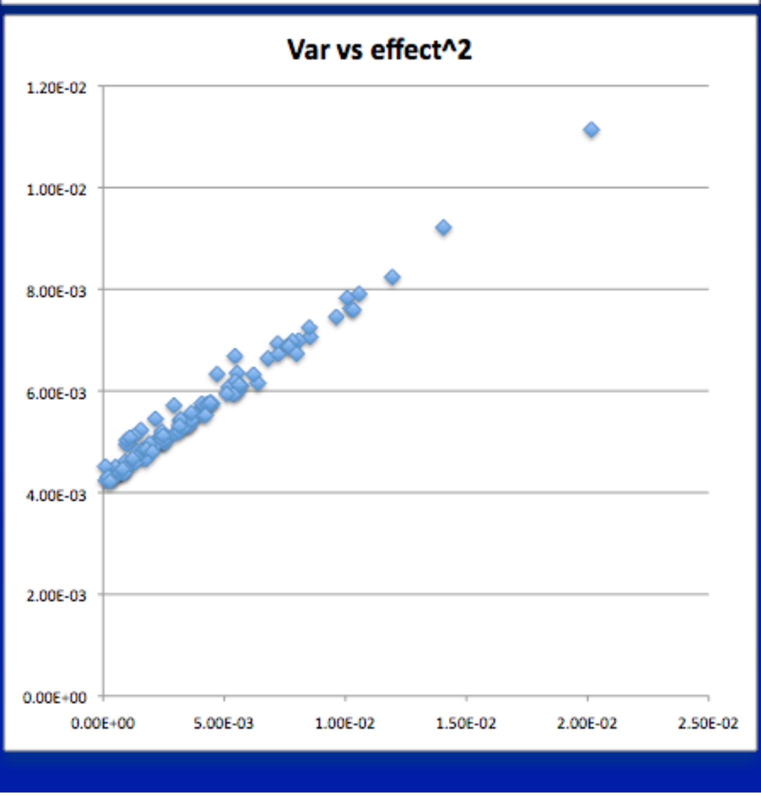
Bayes A df=4


Var vs effect
1.20E-02
1.00E-02
8.00E-03
6.00E-03
4.00E-03
2.00E-03
0.00E+00
-1.50E-01
-1.00E-01
-5.00E-02
0.00E+00
5.00E-02
1.00E-01
1.50E-01
2.00E-01

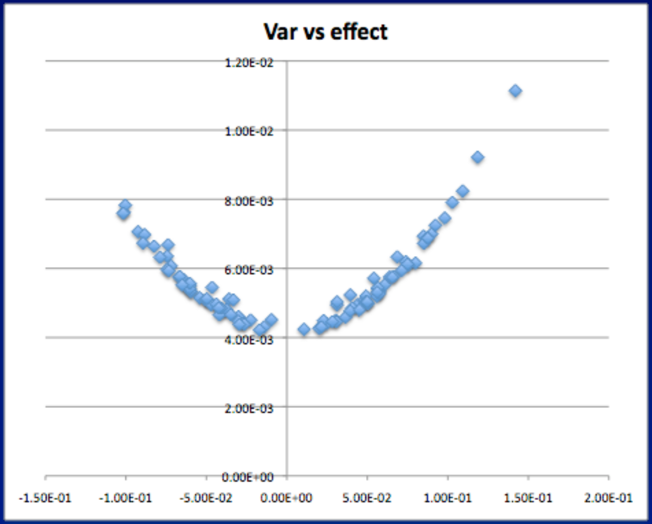
Bayes A df=4

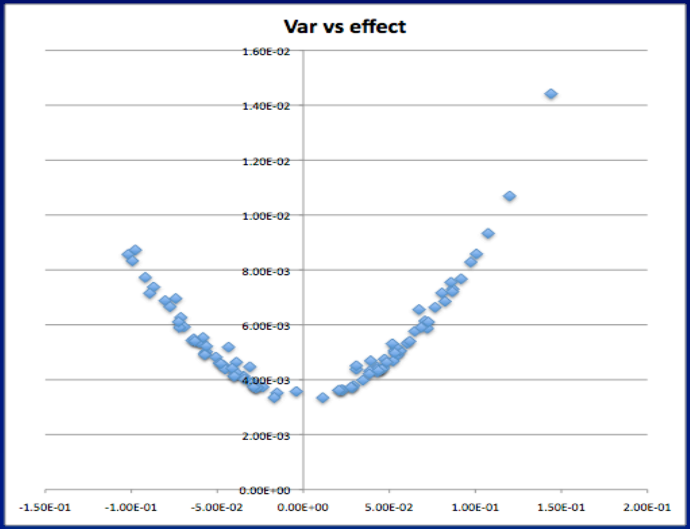

Var vs effect

Bayes A df=4

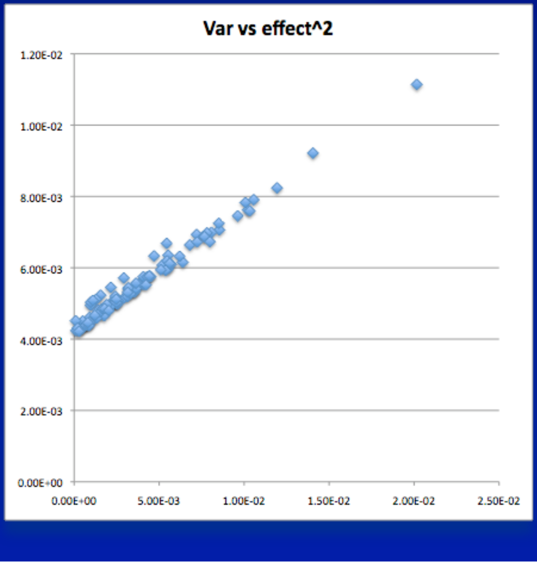

Var vs effect^2

Bayes A Effect vs Var(effect)

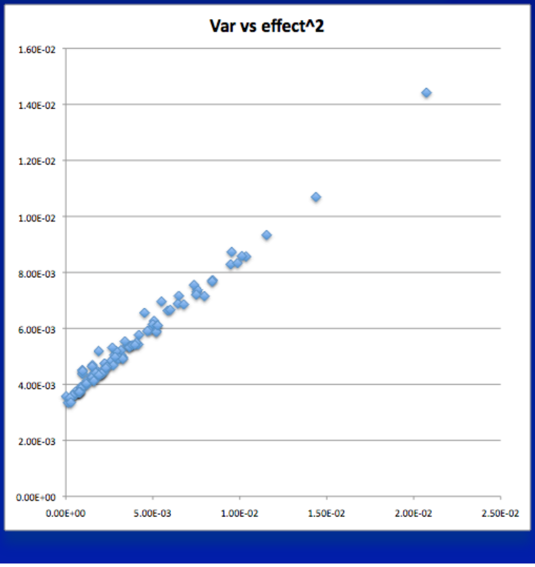




df=4



df=3



Analytical Methods

- Two major classes of mixed models

No selection of loci

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$constant\ \sigma_a^2\ and\ \sigma_e^2$$

$$"BLUP"$$

$$estimate\ \sigma_{ai}^2\ and\ \sigma_e^2$$

$$BayesA$$

Mixture Models (model selection)

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i \delta_i + \mathbf{e}$$

$$estimate\ \delta_i,\ \sigma_{ai}^2\ and\ \sigma_e^2$$

$$BayesB\ (known\ \pi)$$

$$\pi = fraction\ loci\ with\ no\ effect$$

Meuwissen, Hayes & Goddard (2001)

Mixture Models

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i \delta_i + \mathbf{e}$$

nchains

kSNPs

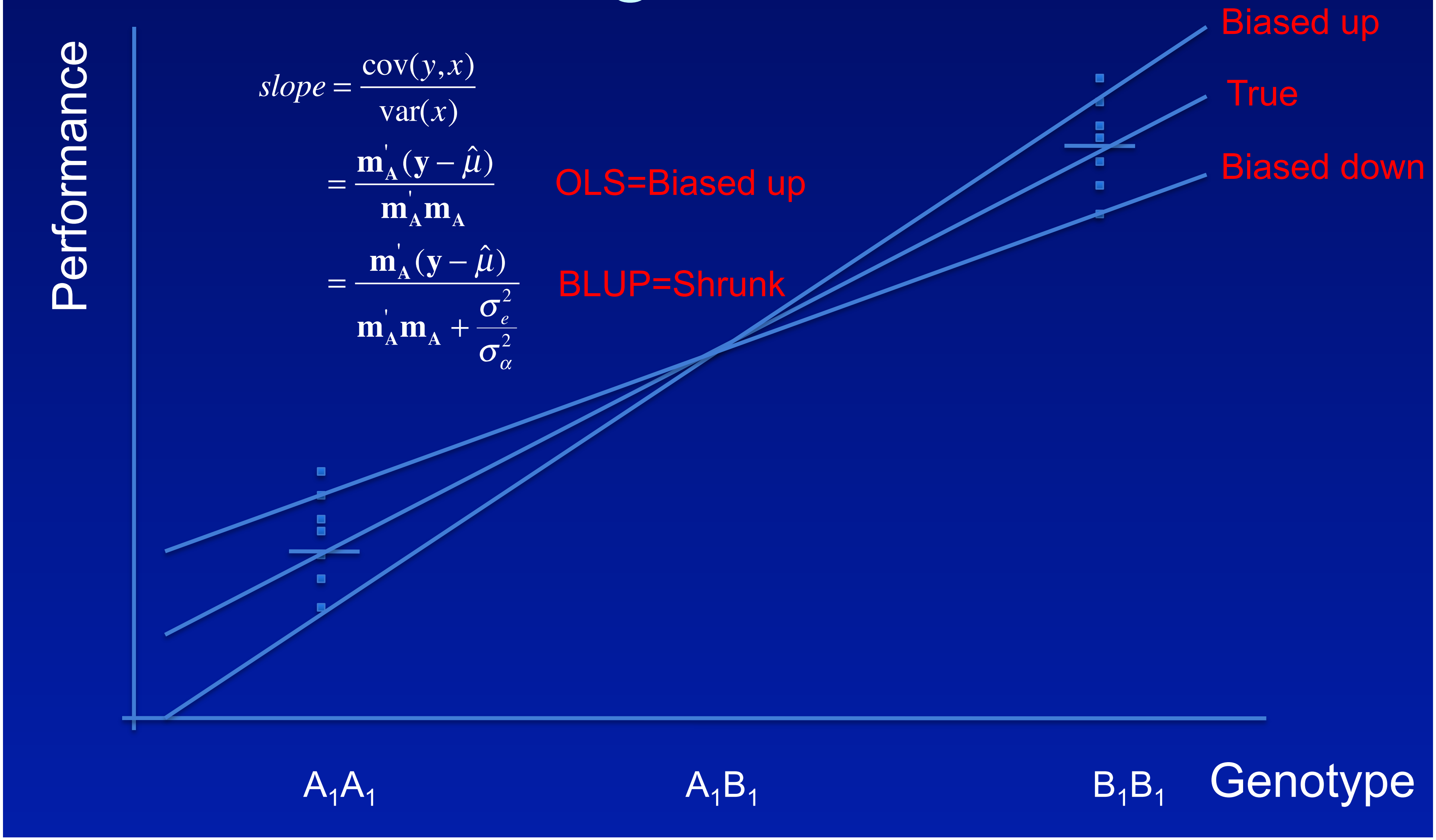
$$\delta_i = 1 \quad L_1 = L(\mathbf{Xb} + \mathbf{M}_i \mathbf{a}_i + \mathbf{e}) \;\; given \;\; (1 - \pi)$$

$$\delta_i = 0 \quad L_0 = L(\mathbf{Xb} + \mathbf{e}) \;\; given \;\; \pi$$

$$Compute \;\; p = \frac{L_1}{L_1 + L_0} \qquad Draw \;\; u = uniform[0,1]$$

$$u < p \;\; then \;\; locus \;\; i \;\; is \;\; in \;\; the \;\; model \;\; this \;\; chain$$

Shrinkage Estimation



$$slope = \frac{\operatorname{cov}(y,x)}{\operatorname{var}(x)} = \frac{\mathbf{m}_{\mathrm{A}}'(\mathbf{y}-\hat{\mu})}{\mathbf{m}_{\mathrm{A}}'\mathbf{m}_{\mathrm{A}}} = \frac{\mathbf{m}_{\mathrm{A}}'(\mathbf{y}-\hat{\mu})}{\mathbf{m}_{\mathrm{A}}'\mathbf{m}_{\mathrm{A}} + \frac{\sigma_e^2}{\sigma_\alpha^2}}$$

OLS=Biased up

BLUP=Shrunk

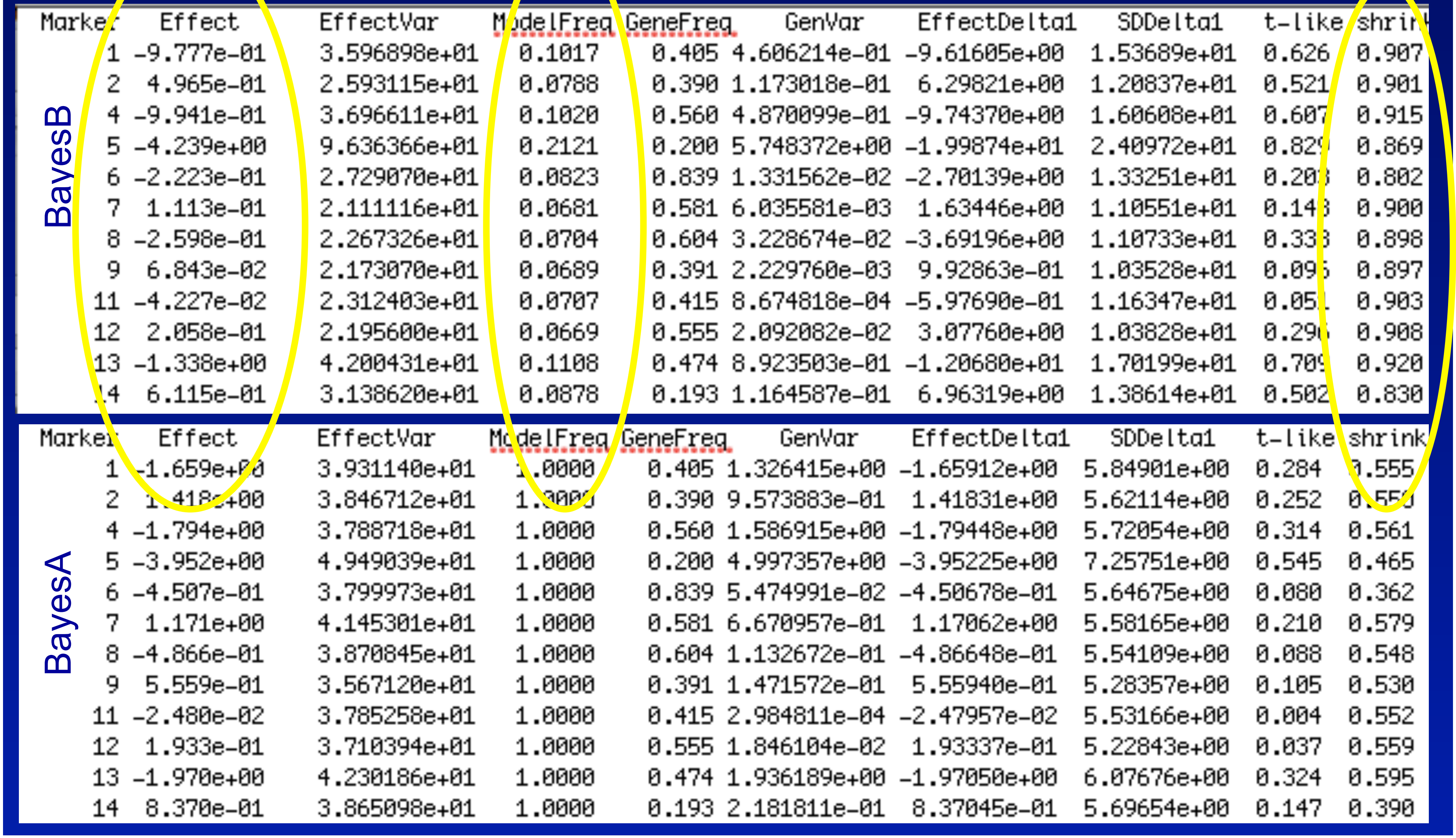
Bayesian Estimation

- Extent of shrinkage that results by treating effects as random (due to uncertainty) depends upon the relative magnitude of $\mathbf{m}_{\mathbf{A}}^{'}\mathbf{m}_{\mathbf{A}}$ *and* $\sigma_e^2 / \sigma_\alpha^2$
  - Less shrinkage than animal models
- Additional shrinkage in mixture models due to model frequency

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i \delta_i + \mathbf{e}$$

$$posterior\ mean\ \ slope = mean(fitted\ slope) \times \Pr(\delta_i = 1)$$

Bayes A vs B marker effects



BayesB

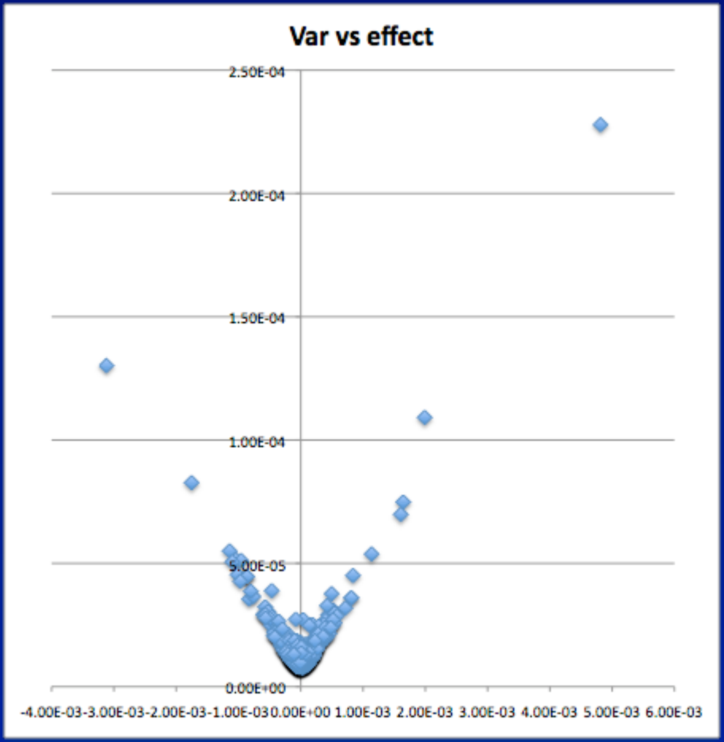
```
Marker  Effect       EffectVar    ModelFreq GeneFreq   GenVar      EffectDelta1   SDDelta1   t-like shrink
    1 -9.777e-01  3.596898e+01  0.1017    0.405 4.606214e-01 -9.61605e+00 1.53689e+01 0.626  0.907
    2  4.965e-01  2.593115e+01  0.0788    0.390 1.173018e-01  6.29821e+00 1.20837e+01 0.521  0.901
    4 -9.941e-01  3.696611e+01  0.1020    0.560 4.870099e-01 -9.74370e+00 1.60608e+01 0.607  0.915
    5 -4.239e+00  9.636366e+01  0.2121    0.200 5.748372e+00 -1.99874e+01 2.40972e+01 0.829  0.869
    6 -2.223e-01  2.729070e+01  0.0823    0.839 1.331562e-02 -2.70139e+00 1.33251e+01 0.203  0.802
    7  1.113e-01  2.111116e+01  0.0681    0.581 6.035581e-03  1.63446e+00 1.10551e+01 0.148  0.900
    8 -2.598e-01  2.267326e+01  0.0704    0.604 3.228674e-02 -3.69196e+00 1.10733e+01 0.333  0.898
    9  6.843e-02  2.173070e+01  0.0689    0.391 2.229760e-03  9.92863e-01 1.03528e+01 0.096  0.897
   11 -4.227e-02  2.312403e+01  0.0707    0.415 8.674818e-04 -5.97690e-01 1.16347e+01 0.051  0.903
   12  2.058e-01  2.195600e+01  0.0669    0.555 2.092082e-02  3.07760e+00 1.03828e+01 0.296  0.908
   13 -1.338e+00  4.200431e+01  0.1108    0.474 8.923503e-01 -1.20680e+01 1.70199e+01 0.709  0.920
   14  6.115e-01  3.138620e+01  0.0878    0.193 1.164587e-01  6.96319e+00 1.38614e+01 0.502  0.830
```

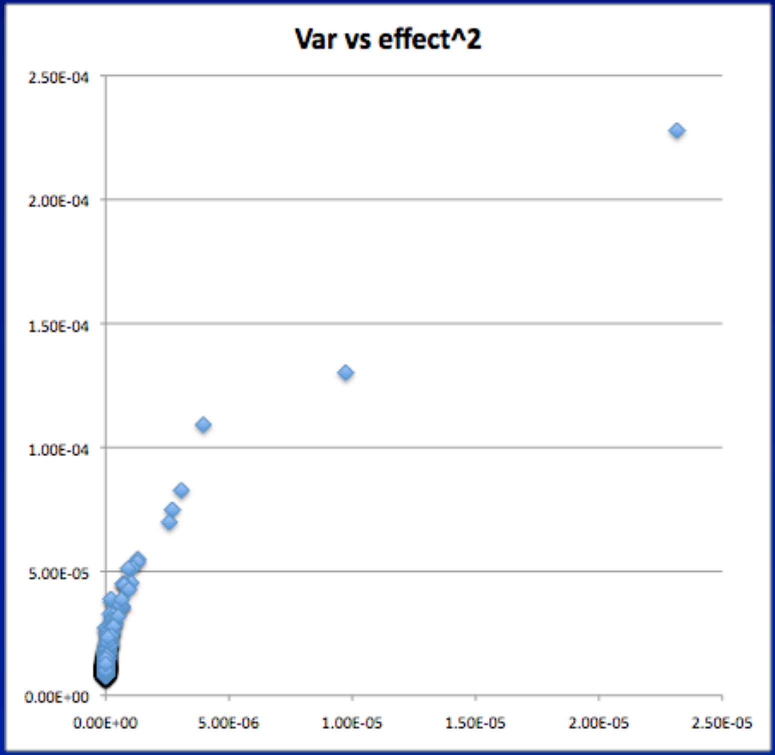
BayesA

```
Marker  Effect       EffectVar    ModelFreq GeneFreq   GenVar      EffectDelta1   SDDelta1   t-like shrink
    1 -1.659e+00  3.931140e+01  1.0000    0.405 1.326415e+00 -1.65912e+00 5.84901e+00 0.284  0.555
    2  1.418e+00  3.846712e+01  1.0000    0.390 9.573883e-01  1.41831e+00 5.62114e+00 0.252  0.550
    4 -1.794e+00  3.788718e+01  1.0000    0.560 1.586915e+00 -1.79448e+00 5.72054e+00 0.314  0.561
    5 -3.952e+00  4.949039e+01  1.0000    0.200 4.997357e+00 -3.95225e+00 7.25751e+00 0.545  0.465
    6 -4.507e-01  3.799973e+01  1.0000    0.839 5.474991e-02 -4.50678e-01 5.64675e+00 0.080  0.362
    7  1.171e+00  4.145301e+01  1.0000    0.581 6.670957e-01  1.17062e+00 5.58165e+00 0.210  0.579
    8 -4.866e-01  3.870845e+01  1.0000    0.604 1.132672e-01 -4.86648e-01 5.54109e+00 0.088  0.548
    9  5.559e-01  3.567120e+01  1.0000    0.391 1.471572e-01  5.55940e-01 5.28357e+00 0.105  0.530
   11 -2.480e-02  3.785258e+01  1.0000    0.415 2.984811e-04 -2.47957e-02 5.53166e+00 0.004  0.552
   12  1.933e-01  3.710394e+01  1.0000    0.555 1.846104e-02  1.93337e-01 5.22843e+00 0.037  0.559
   13 -1.970e+00  4.230186e+01  1.0000    0.474 1.936189e+00 -1.97050e+00 6.07676e+00 0.324  0.595
   14  8.370e-01  3.865098e+01  1.0000    0.193 2.181811e-01  8.37045e-01 5.69654e+00 0.147  0.390
```

Bayes B Effect vs Var(Effect)

$$df = 4 \quad \pi = 0.99$$





Analytical Methods

No selection of loci

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

constant σ_a^2 *and* σ_e^2

"*BLUP*"

estimate σ_{ai}^2 *and* σ_e^2

BayesA

Mixture Models (model selection)

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i \delta_i + \mathbf{e}$$

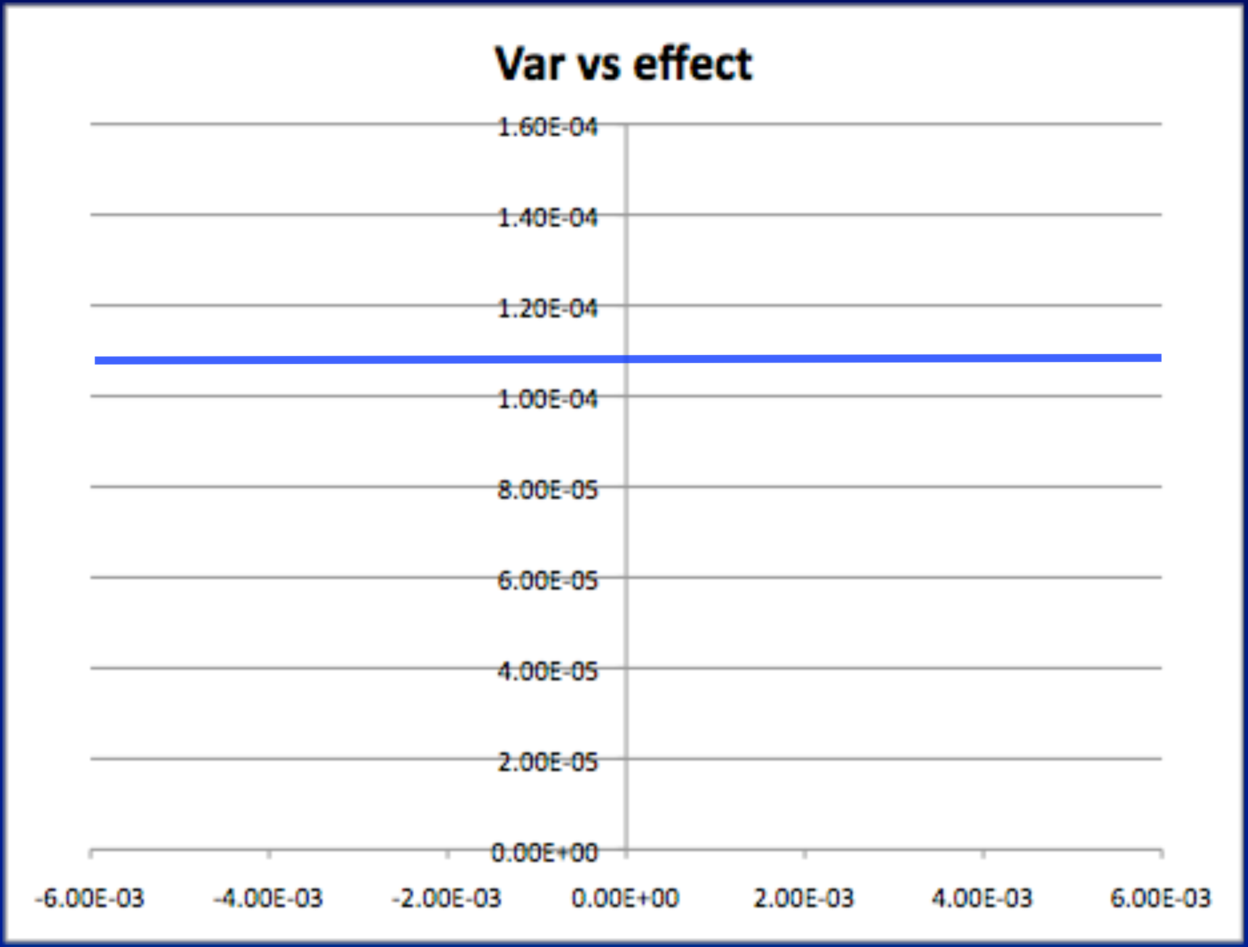
estimate δ_i, σ_{ai}^2 *and* σ_e^2

BayesB (*known* π)

estimate δ_i, σ_a^2 *and* σ_e^2

BayesC (*known* π) "*BLUP*" $= C(0)$

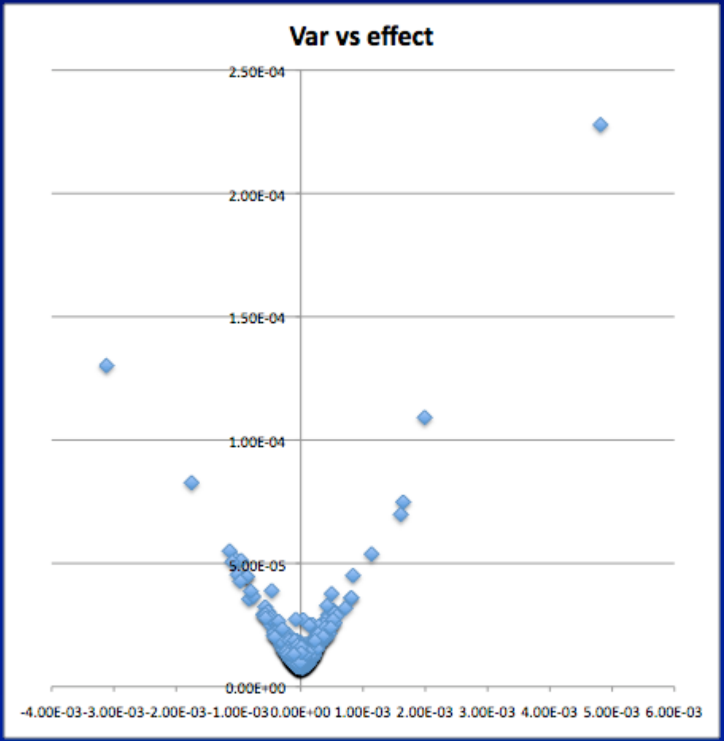
$\pi =$ *fraction loci with no effect*

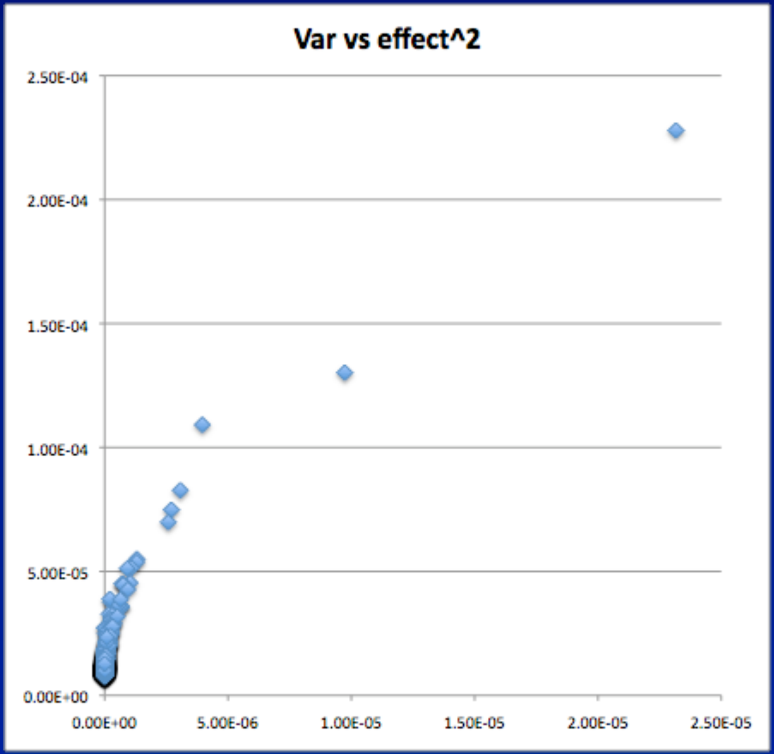

Var vs effect
1.60E-04
1.40E-04
1.20E-04
1.00E-04
8.00E-05
6.00E-05
4.00E-05
2.00E-05
0.00E+00
-6.00E-03
-4.00E-03
-2.00E-03
0.00E+00
2.00E-03
4.00E-03
6.00E-03

Bayes C0

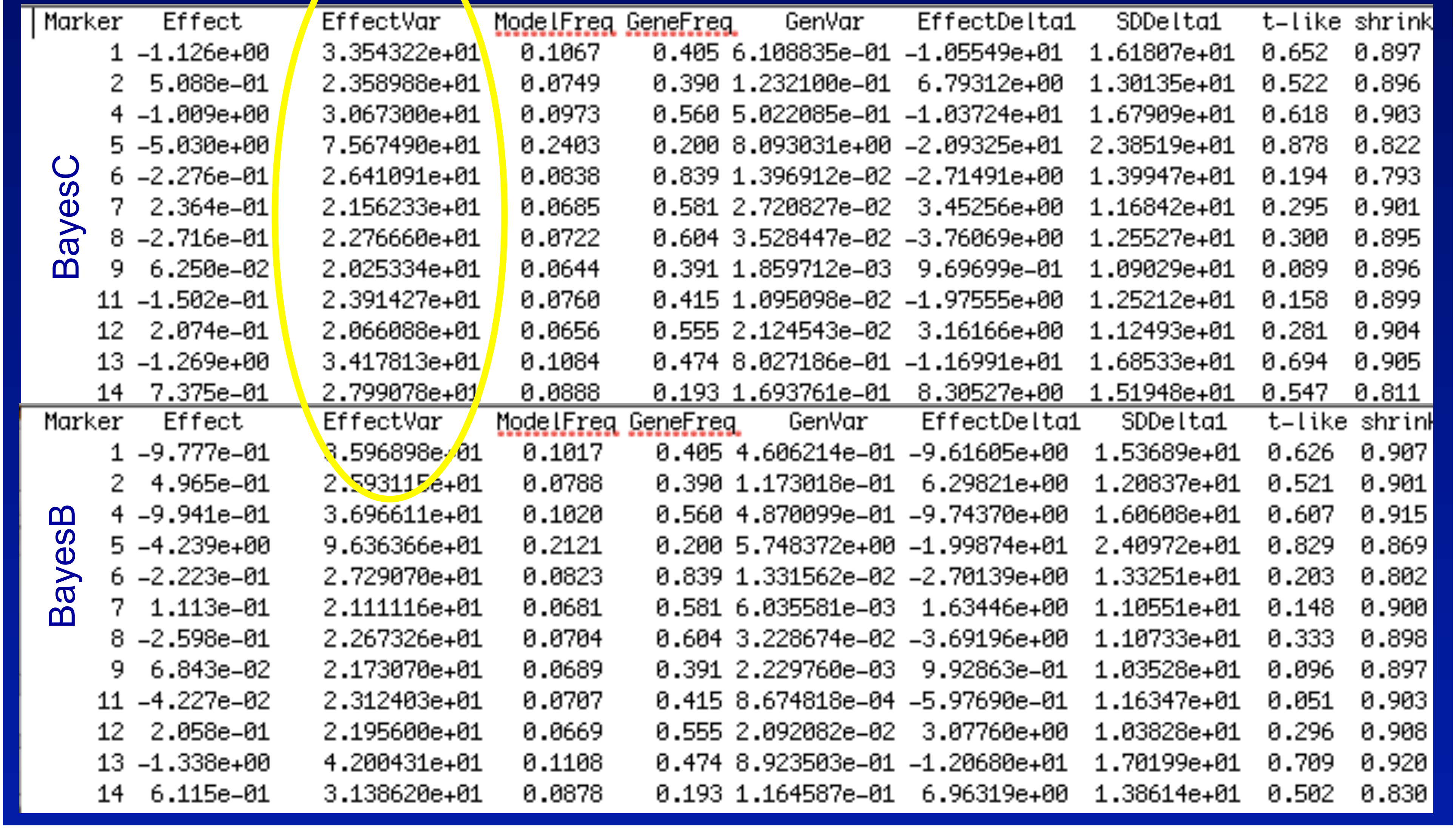
Bayes C (pi>0) or Bayes CPi

Like the following





Bayes C Var(Effect)



BayesC

```
 Marker    Effect          EffectVar    ModelFreq GeneFreq      GenVar     EffectDelta1     SDDelta1    t-like shrink
      1 -1.126e+00    3.354322e+01    0.1067       0.405 6.108835e-01 -1.05549e+01  1.61807e+01  0.652  0.897
      2  5.088e-01    2.358988e+01    0.0749       0.390 1.232100e-01  6.79312e+00  1.30135e+01  0.522  0.896
      4 -1.009e+00    3.067300e+01    0.0973       0.560 5.022085e-01 -1.03724e+01  1.67909e+01  0.618  0.903
      5 -5.030e+00    7.567490e+01    0.2403       0.200 8.093031e+00 -2.09325e+01  2.38519e+01  0.878  0.822
      6 -2.276e-01    2.641091e+01    0.0838       0.839 1.396912e-02 -2.71491e+00  1.39947e+01  0.194  0.793
      7  2.364e-01    2.156233e+01    0.0685       0.581 2.720827e-02  3.45256e+00  1.16842e+01  0.295  0.901
      8 -2.716e-01    2.276660e+01    0.0722       0.604 3.528447e-02 -3.76069e+00  1.25527e+01  0.300  0.895
      9  6.250e-02    2.025334e+01    0.0644       0.391 1.859712e-03  9.69699e-01  1.09029e+01  0.089  0.896
     11 -1.502e-01    2.391427e+01    0.0760       0.415 1.095098e-02 -1.97555e+00  1.25212e+01  0.158  0.899
     12  2.074e-01    2.066088e+01    0.0656       0.555 2.124543e-02  3.16166e+00  1.12493e+01  0.281  0.904
     13 -1.269e+00    3.417813e+01    0.1084       0.474 8.027186e-01 -1.16991e+01  1.68533e+01  0.694  0.905
     14  7.375e-01    2.799078e+01    0.0888       0.193 1.693761e-01  8.30527e+00  1.51948e+01  0.547  0.811
```

BayesB

```
 Marker    Effect          EffectVar    ModelFreq GeneFreq      GenVar     EffectDelta1     SDDelta1    t-like shrink
      1 -9.777e-01    3.596898e+01    0.1017       0.405 4.606214e-01 -9.61605e+00  1.53689e+01  0.626  0.907
      2  4.965e-01    2.593115e+01    0.0788       0.390 1.173018e-01  6.29821e+00  1.20837e+01  0.521  0.901
      4 -9.941e-01    3.696611e+01    0.1020       0.560 4.870099e-01 -9.74370e+00  1.60608e+01  0.607  0.915
      5 -4.239e+00    9.636366e+01    0.2121       0.200 5.748372e+00 -1.99874e+01  2.40972e+01  0.829  0.869
      6 -2.223e-01    2.729070e+01    0.0823       0.839 1.331562e-02 -2.70139e+00  1.33251e+01  0.203  0.802
      7  1.113e-01    2.111116e+01    0.0681       0.581 6.035581e-03  1.63446e+00  1.10551e+01  0.148  0.900
      8 -2.598e-01    2.267326e+01    0.0704       0.604 3.228674e-02 -3.69196e+00  1.10733e+01  0.333  0.898
      9  6.843e-02    2.173070e+01    0.0689       0.391 2.229760e-03  9.92863e-01  1.03528e+01  0.096  0.897
     11 -4.227e-02    2.312403e+01    0.0707       0.415 8.674818e-04 -5.97690e-01  1.16347e+01  0.051  0.903
     12  2.058e-01    2.195600e+01    0.0669       0.555 2.092082e-02  3.07760e+00  1.03828e+01  0.296  0.908
     13 -1.338e+00    4.200431e+01    0.1108       0.474 8.923503e-01 -1.20680e+01  1.70199e+01  0.709  0.920
     14  6.115e-01    3.138620e+01    0.0878       0.193 1.164587e-01  6.96319e+00  1.38614e+01  0.502  0.830
```

Summary

- Genomic Selection methods rely on shrinkage of marker effects to get reliable estimation
- There are several alternatives for shrinking marker effects
  - Treating marker effects as random
  - Fitting mixture models
  - (Using densities less extreme than normal)

Summary

- Fitting Mixture distributions provides a much more powerful method for shrinking marker effects than simply treating marker effects as random

Web-based system

Bioinformatics Infrastructure

- Identify informative regions for fine-mapping and gene discovery
- Provide a platform for collaborating (beef) researchers to undertake genomic training
    - eg US Meat Animal Research Center
    - Federally-funded beef projects
- Provide a platform for delivering genomic predictions to (the beef) industry

Site access

- Follow links from bigs.ansci.iastate.edu
  - BIGS – bioinformatics to implement genomic selection
- Federally-funded project (2010-2012) for US beef cattle researchers
  - Available for limited access to other parties conditional on demand for processors (64 CPUs)
  - Useful for benchmarking

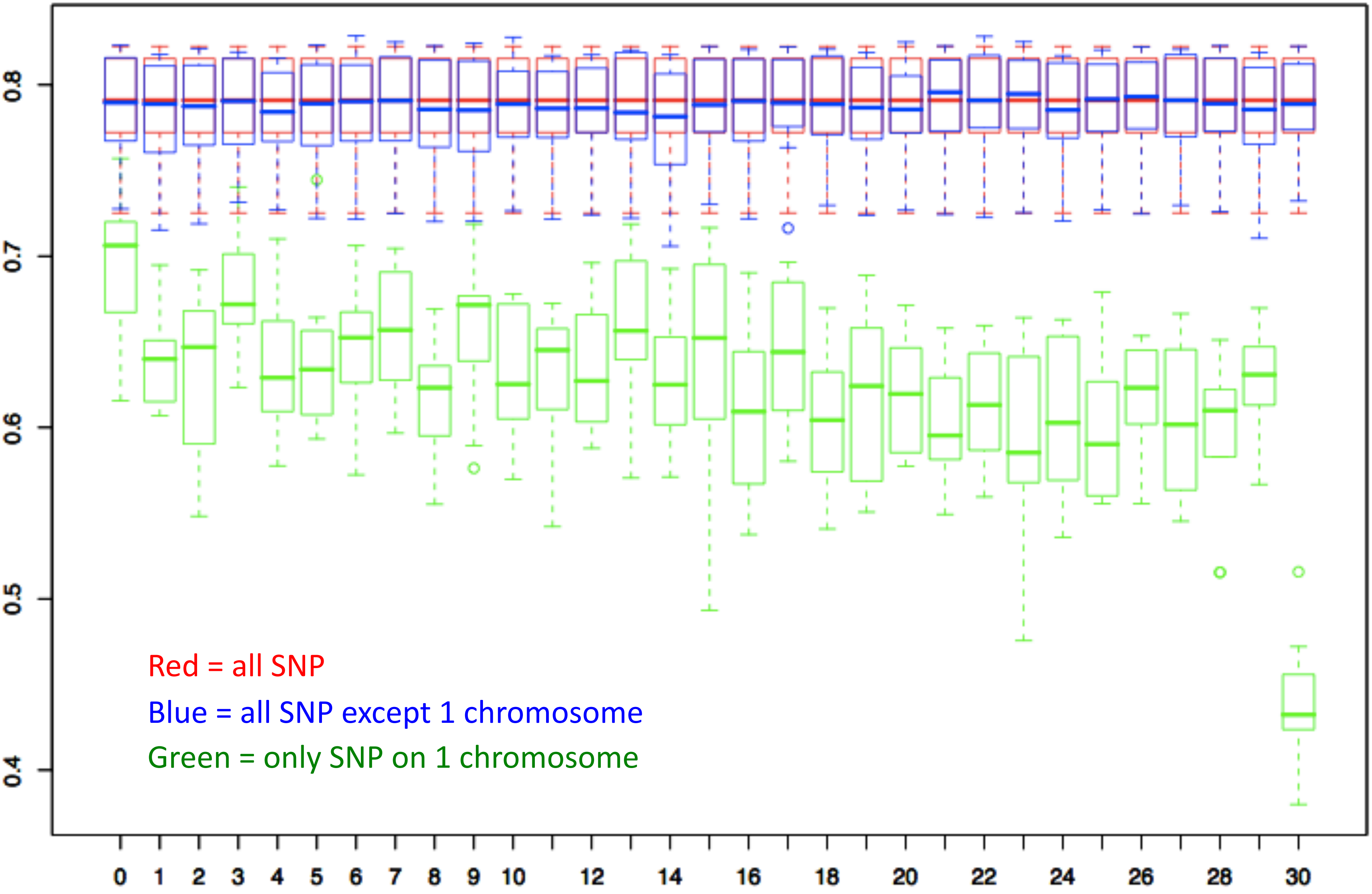
Required Information

- Research from analysis of high-density genotypes to predict merit has several objectives
  - Determine predictive ability of
    - same-density panels in validation/target populations closely related to the training population
    - same-density panels in validation/target populations less related or unrelated to the training population
    - low-density panels in populations closely related to the training population
  - Motivate other genomic selection research

Predictive ability of Individual Chromosomes

Milkfat

Data kindly shared by Vlad, LIC



Red = all SNP

Blue = all SNP except 1 chromosome

Green = only SNP on 1 chromosome

Problems with Validation

BayesB then BayesA (100 markers)

“Heritability” for 100 markers chosen for trait in row, applied to trait in column

0.64	0.50	0.23	0.33	0.29	0.22	0.45	0.30	0.24
0.53	**0.61**	0.24	0.33	0.29	0.23	0.45	0.30	0.26
0.27	0.29	**0.57**	0.33	0.29	0.22	0.36	0.30	0.25
0.27	0.27	0.23	**0.67**	0.29	0.26	0.42	0.30	0.29
0.28	0.24	0.23	0.33	**0.57**	0.25	0.40	0.35	0.27
0.27	0.29	0.26	0.33	0.29	**0.53**	0.42	0.30	0.25
0.29	0.29	0.23	0.33	0.29	0.25	**0.70**	0.26	0.25
0.29	0.27	0.24	0.33	0.29	0.22	0.36	**0.63**	0.24
0.32	0.27	0.26	0.33	0.29	0.25	0.42	0.30	**0.65**

Bayes B then Bayes A (100 markers)

Correlation in training data
chosen for trait in row applied to trait in column

0.79	0.68	0.37	0.41	0.42	0.33	0.56	0.46	0.39
0.69	**0.76**	0.38	0.4	0.44	0.34	0.54	0.42	0.41
0.39	0.41	**0.77**	0.4	0.39	0.35	0.5	0.4	0.39
0.36	0.36	0.35	**0.78**	0.41	0.41	0.53	0.45	0.43
0.41	0.4	0.38	0.36	**0.79**	0.39	0.51	0.51	0.41
0.39	0.4	0.39	0.45	0.41	**0.72**	0.55	0.41	0.38
0.41	0.4	0.35	0.45	0.4	0.41	**0.87**	0.4	0.41
0.43	0.41	0.37	0.4	0.48	0.37	0.5	**0.79**	0.37
0.44	0.4	0.39	0.44	0.38	0.37	0.5	0.45	**0.78**

1st attempt Cross Validation

- Dataset 1 comprising 8 breeds
- Select best 100 markers in all data using BayesB

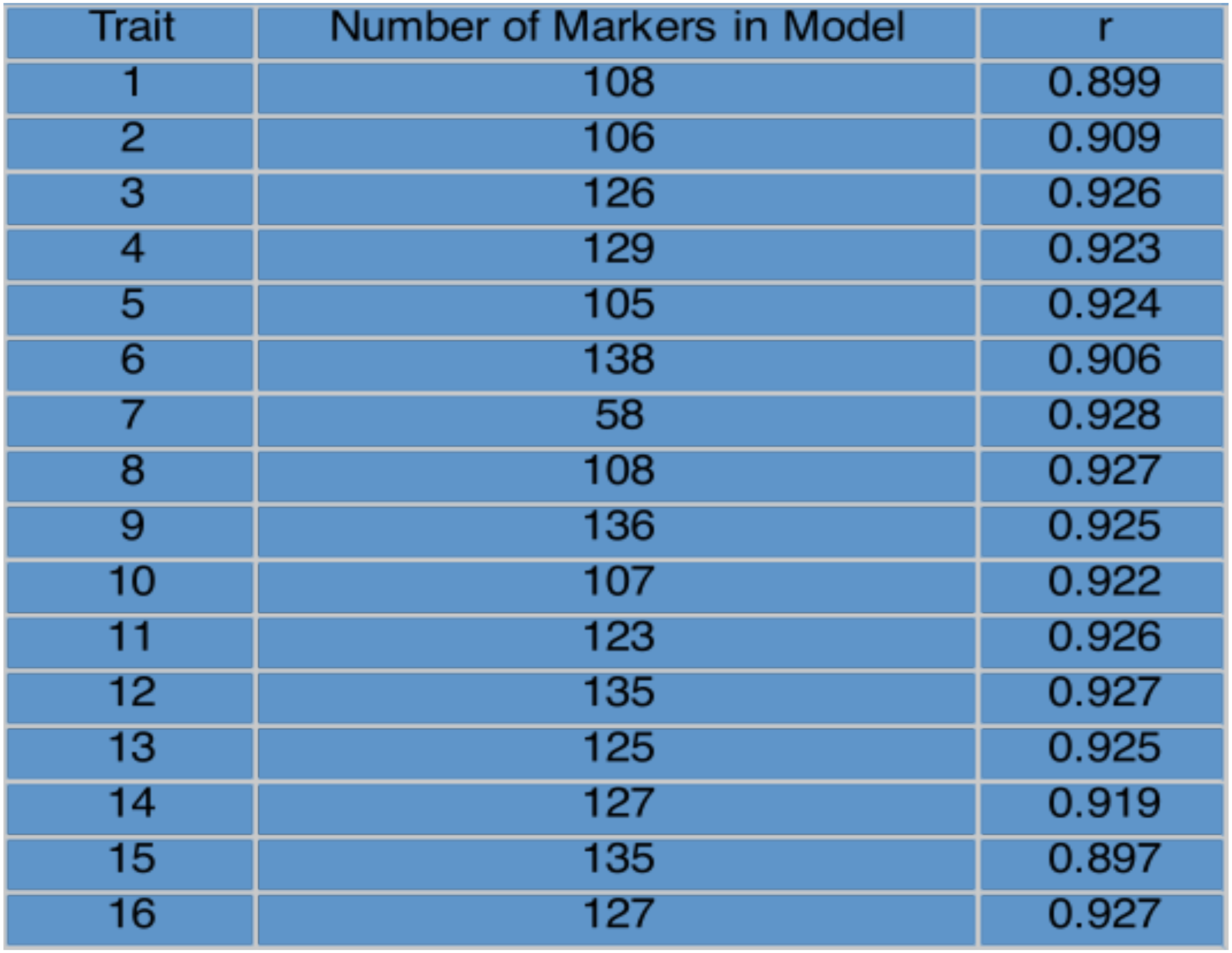
Training	B1		✓	✓	✓	✓	✓	✓	✓
	B2	✓		✓	✓	✓	✓	✓	✓
	B3	✓	✓		✓	✓	✓	✓	✓
	B4	✓	✓	✓		✓	✓	✓	✓
	B5	✓	✓	✓	✓		✓	✓	✓
	B6	✓	✓	✓	✓	✓		✓	✓
	B7	✓	✓	✓	✓	✓	✓		✓
	B8	✓	✓	✓	✓	✓	✓	✓	
Validation		B1	B2	B3	B4	B5	B6	B7	B8

Bayes B then Bayes A (100 markers)

markers in row chosen from Bayes B on all data, Bayes A trained in cross-validation for trait in column, predicting merit in omitted data

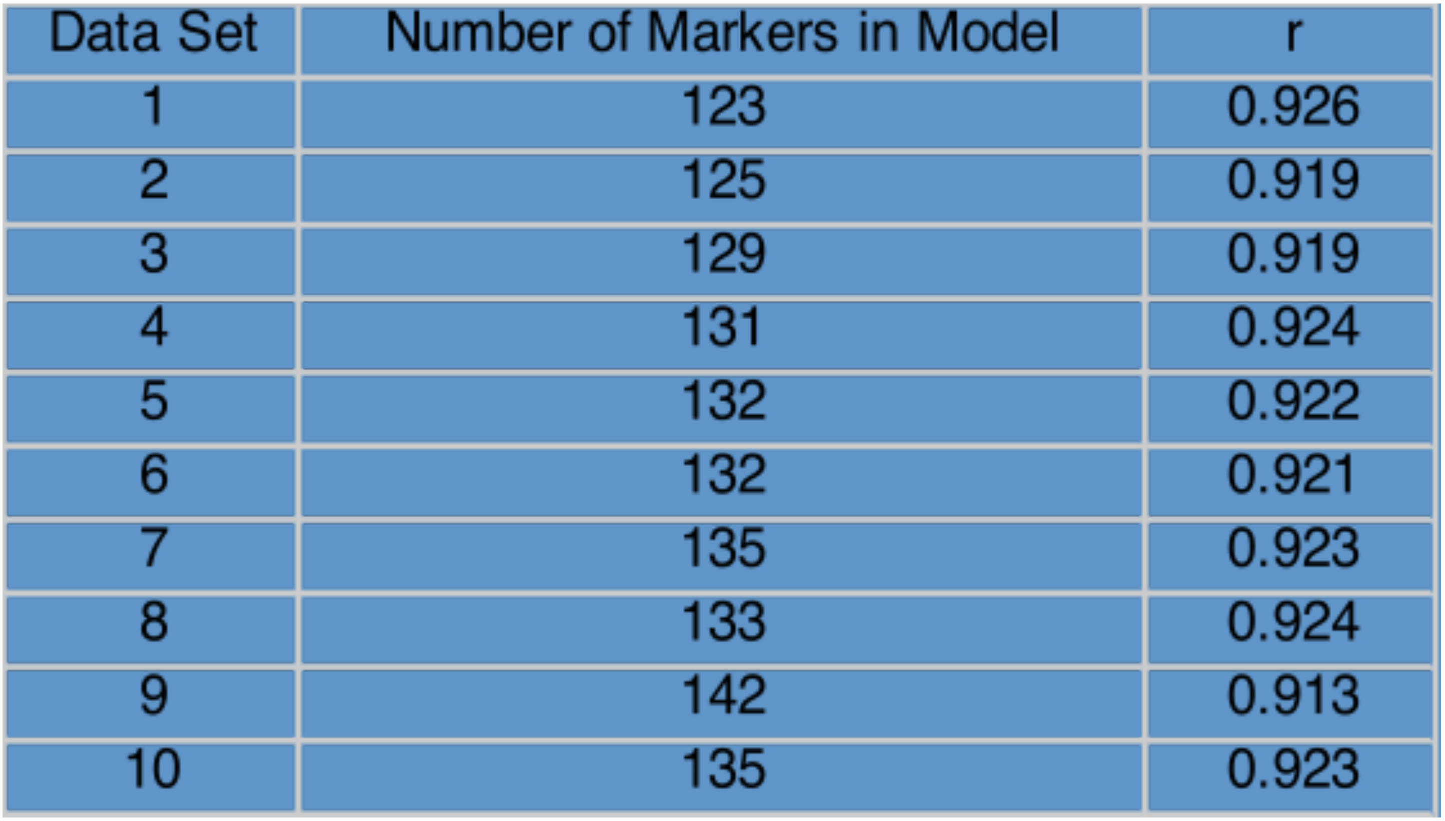
0.66	0.53	-0.02	0.09	0.02	-0.06	0.07	0.08	-0.03
0.53	**0.65**	0.01	0.03	0.1	-0.02	0.06	-0.02	0.06
0.01	0.03	**0.68**	0.02	-0.03	-0.02	-0.04	-0.01	-0.05
-0.05	-0.06	0.01	**0.68**	0.02	0.04	0.02	0.08	0.11
0.09	0.07	-0.02	0	**0.68**	0.04	0	0.2	0.04
-0.02	0.01	0.06	0.14	0.08	**0.58**	0.11	0.03	-0.03
-0.01	0.01	-0.04	0.14	0	0.1	**0.74**	-0.07	0.04
0.06	0.05	0.01	0.05	0.22	0.07	0.06	**0.69**	-0.05
0.08	-0.02	0.02	0.15	-0.08	-0.01	0.01	0.14	**0.7**

StepWise then BayesA



Trait	Number of Markers in Model	r
1	108	0.899
2	106	0.909
3	126	0.926
4	129	0.923
5	105	0.924
6	138	0.906
7	58	0.928
8	108	0.927
9	136	0.925
10	107	0.922
11	123	0.926
12	135	0.927
13	125	0.925
14	127	0.919
15	135	0.897
16	127	0.927

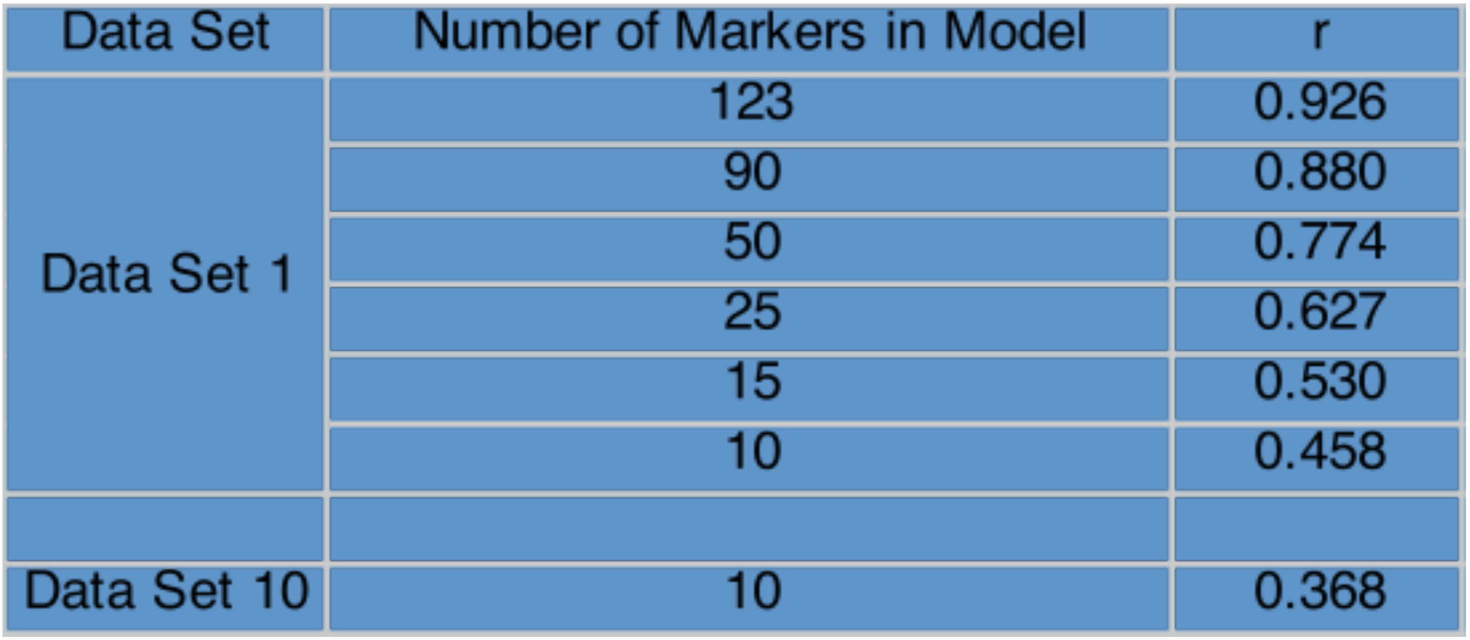
StepWise then BayesA



Data Set	Number of Markers in Model	r
1	123	0.926
2	125	0.919
3	129	0.919
4	131	0.924
5	132	0.922
6	132	0.921
7	135	0.923
8	133	0.924
9	142	0.913
10	135	0.923

Successive datasets have previously best markers removed

StepWise and BayesA



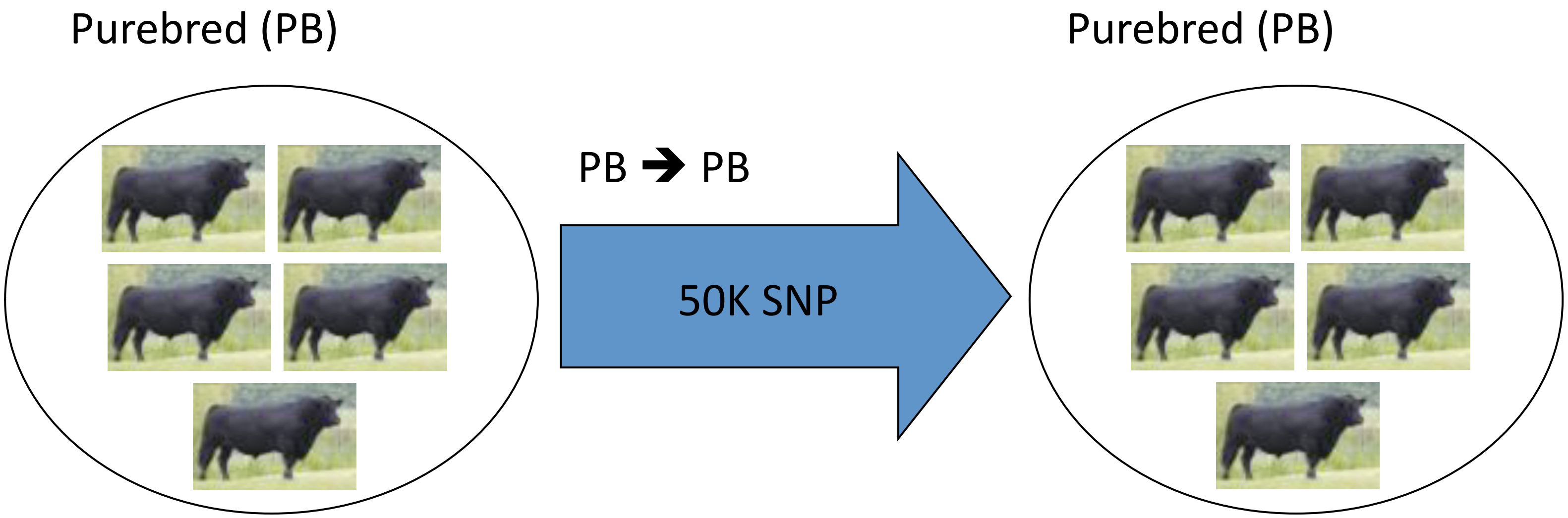
Data Set	Number of Markers in Model	r
Data Set 1	123	0.926
	90	0.880
	50	0.774
	25	0.627
	15	0.530
	10	0.458
Data Set 10	10	0.368

Improved Validation

Proper cross-validation

- Marker subset selection and marker estimation are undertaken on each training data subset and used to predict "virgin" data
- Correlation dropped to 0.18 (at best) when properly (100 marker subset chosen in training data) cross-validated

Training and Validation



Purebred (PB)

PB → PB

50K SNP

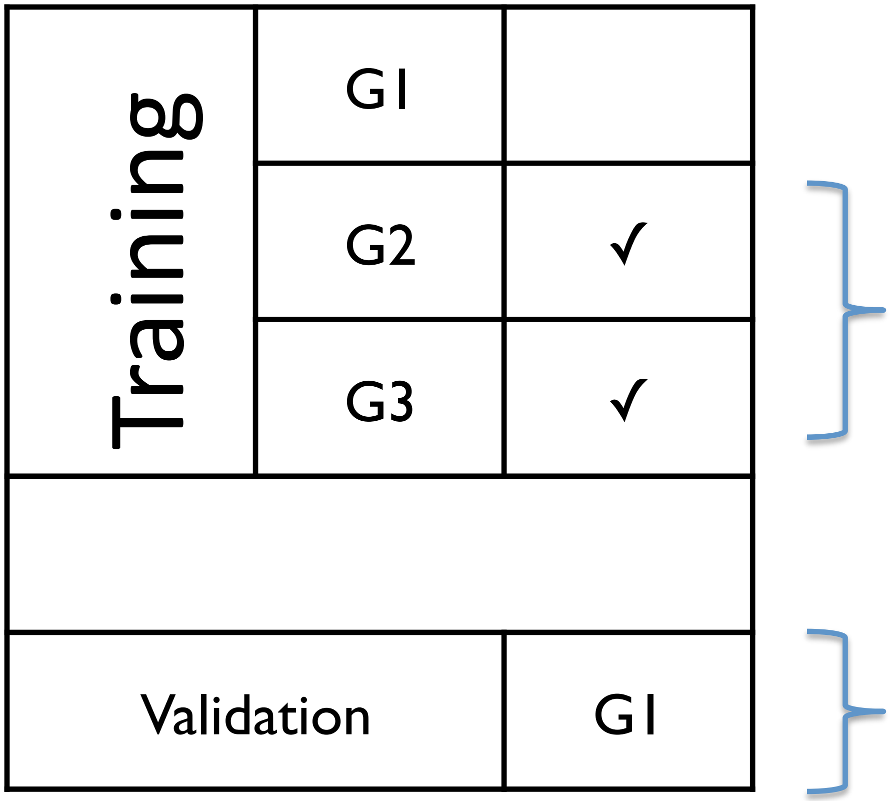
Purebred (PB)

Validation

- Almost always SNP that spuriously fit the data well
  - Having a model that fits the training data well provides relatively little information about how good the prediction will be in new data
    - Many world-changing research discoveries are announced in news releases and then never-to-be-heard-of-again
- Training & Validation can be done together to quantify the likely confidence in predictions

Cross Validation

- Partition the dataset (by sire) into say three groups



Training	G1	
	G2	✓
	G3	✓
Validation		G1

Derive g-EPD

Compute the correlation between predicted genetic merit from g-EPD and observed performance

Cross Validation

- Every animal is in exactly one validation set

Training	G1		✓	✓
	G2	✓		✓
	G3	✓	✓	
Validation		G1	G2	G3

Cross-Validation

- 1800 bulls with EPDs - split into 3
  - – At random
  - – By sire ID - sire of bulls nested in subset
  - – By sire ID - sires also fitted as fixed effects
  - – By time - oldest, middle-aged, youngest

Results

41028m	Random	Sire	Sire+cg	Time
Bayes A (B0)	0.745	0.726	0.646	0.732
Bayes B (.99)	0.722	0.700	0.618	0.712
Bayes C0	0.746	0.728	0.648	0.730
Bayes C(.50)	0.746	0.728	0.647	0.730
Bayes C(.99)	0.728	0.708	0.625	0.717
100m				
C.99/C100m	0.553	0.567	0.389	0.583
StepWise	0.547	0.558	0.393	0.542
PRESS	0.523	0.539	0.365	0.574

Simulated SNP Results - 1184 QTL

52566 markers	Number of training animals			
π=0.977	1000	2000	3000	4000
B(true)	0.65	0.76	0.82	0.84
C(true)	0.62	0.74	0.80	0.83
B(inflated)	0.63	0.75	0.80	0.83
C(inflated)	0.60	0.71	0.77	0.80
B(0.50)	0.62	0.74	0.79	0.82
C(0.50)	0.60	0.70	0.75	0.78
B(0)	0.64	0.74	0.79	0.81
C(0)	0.59	0.70	0.75	0.78

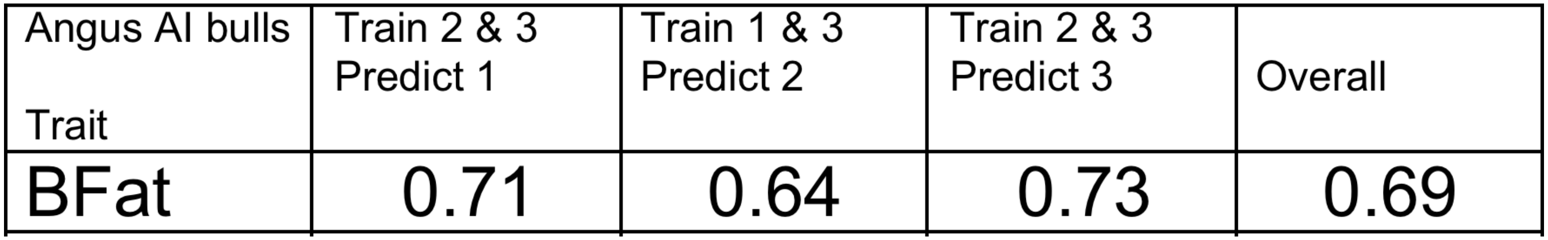
True=#QTL/#markers; inflated=0.9 true; heritability=0.5
(Christian Stricker for Swiss Cattle Breeders)

Simulated Results

2000 animals	Number of QTL		
	171	493	1184
B(true)	0.88	0.82	0.76
C(true)	0.88	0.81	0.74
B(inflated)	0.84	0.79	0.75
C(inflated)	0.70	0.74	0.71
B(0.50)	0.81	0.78	0.74
C(0.50)	0.65	0.72	0.70
B(0)	0.82	0.77	0.74
C(0)	0.64	0.72	0.70

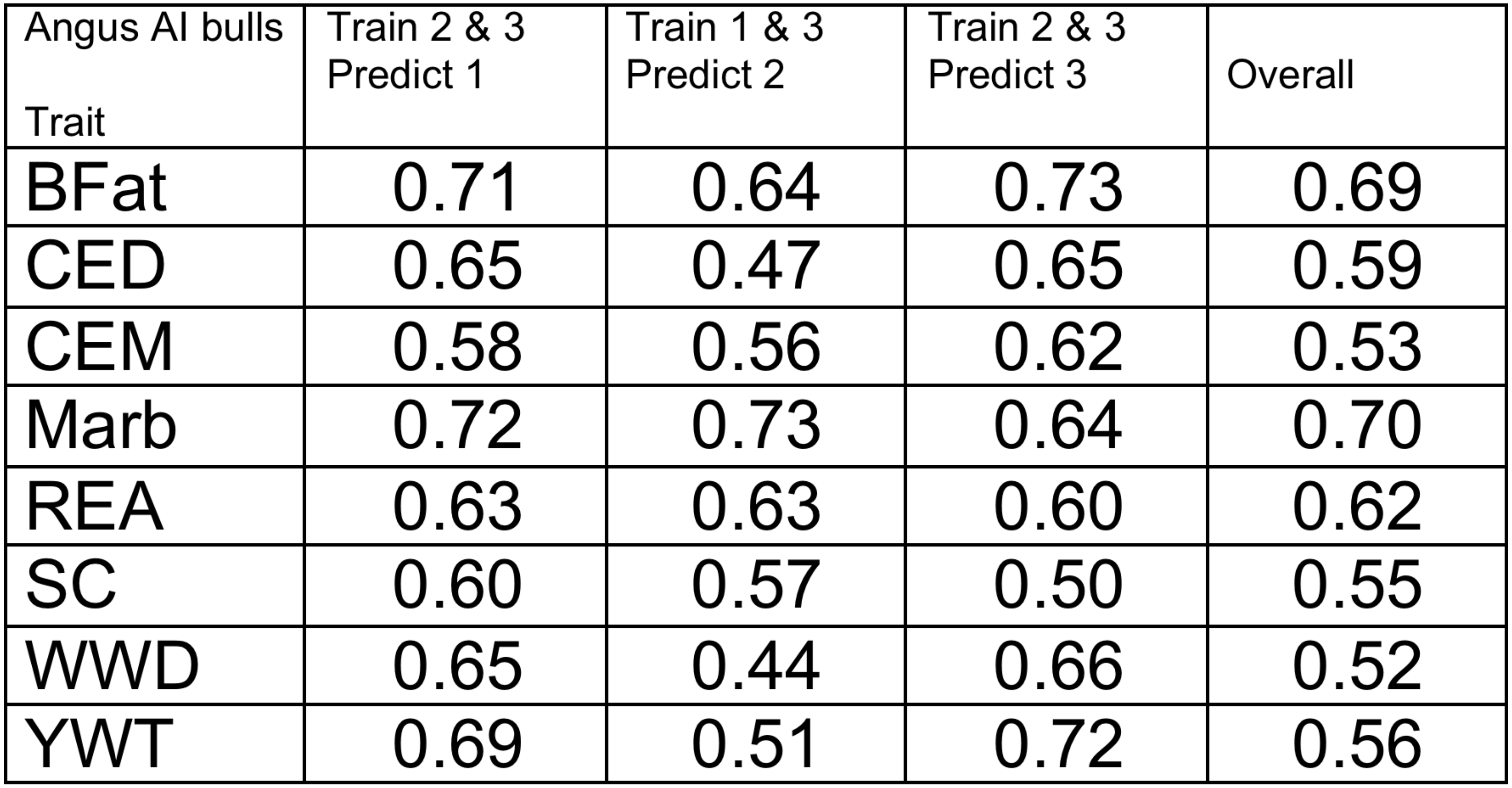
True=#QTL/#markers; inflated=0.9 true; heritability=0.5
(Christian Stricker for Swiss Cattle Breeders)

50k within-breed predictions



Angus AI bulls Trait	Train 2 & 3 Predict 1	Train 1 & 3 Predict 2	Train 2 & 3 Predict 3	Overall
BFat	0.71	0.64	0.73	0.69

50k within-breed predictions



Angus AI bulls Trait	Train 2 & 3 Predict 1	Train 1 & 3 Predict 2	Train 2 & 3 Predict 3	Overall
BFat	0.71	0.64	0.73	0.69
CED	0.65	0.47	0.65	0.59
CEM	0.58	0.56	0.62	0.53
Marb	0.72	0.73	0.64	0.70
REA	0.63	0.63	0.60	0.62
SC	0.60	0.57	0.50	0.55
WWD	0.65	0.44	0.66	0.52
YWT	0.69	0.51	0.72	0.56

50k within-breed predictions

- These predictions are characterized by correlations between genomic merit and realized performance from 0.5 to 0.7
  - They will account for 25 (0.5^2) to 50% (0.7^2) genetic variation
  - Compared to a trait with heritability of 25%, the genomic predictions would be equivalent to observing 6 to 15 offspring in a progeny test
- Correlations of 0.7 are similar to the performance of genomic predictions in dairy cattle

50k within-breed predictions

- These predictions are not as highly accurate as can be achieved in a well designed and managed progeny test, say with 100 or more offspring
- However, for many traits they are much more reliable for animals of a young age (eg prior to first selection) than is currently achievable from individual performance

Across-breed prediction

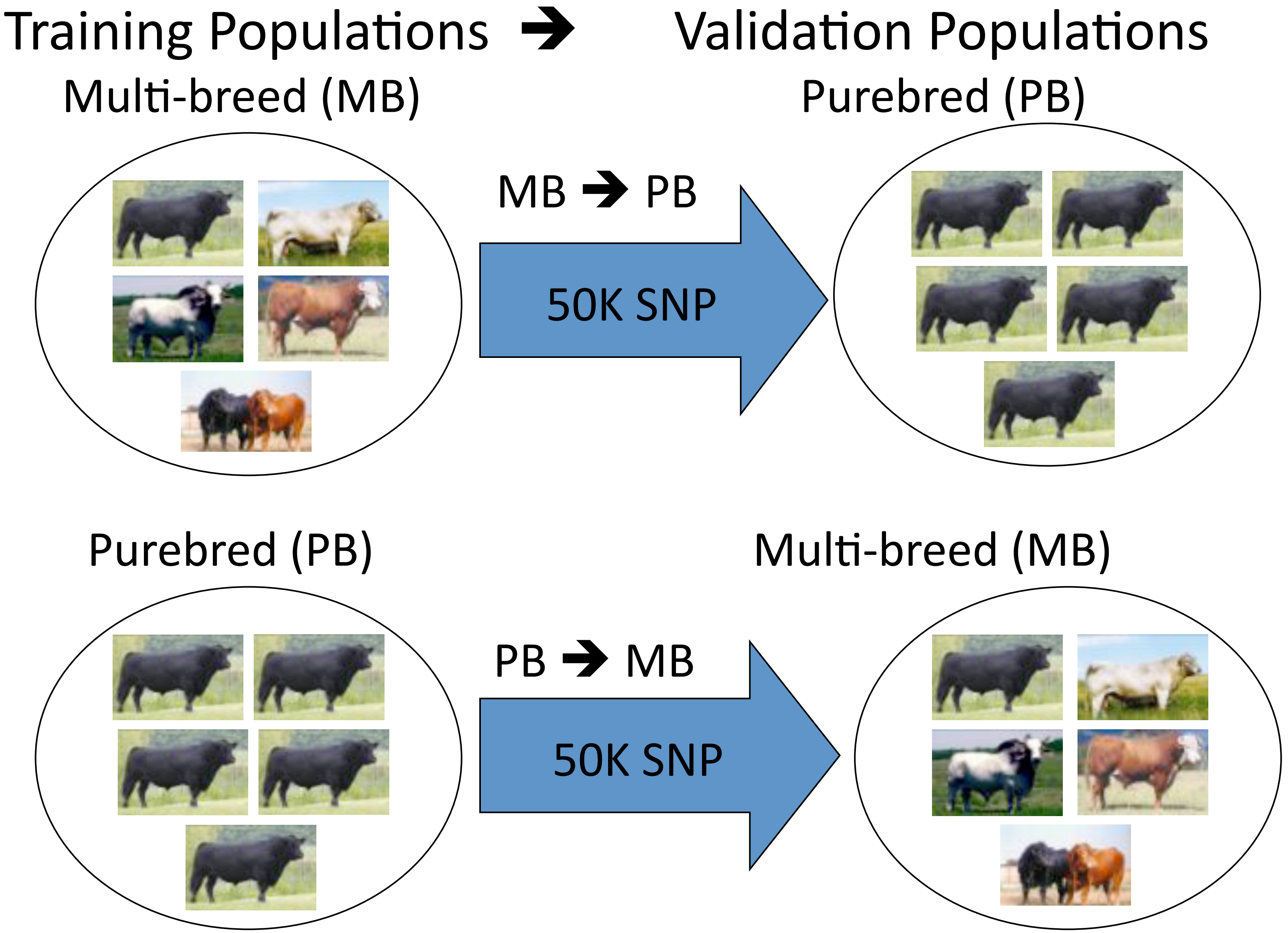
- Refers to the process of predicting performance for a breed or cross that was not in the training dataset
- Critical interest to those selecting breeds that are not well represented in the training populations
- May not be as reliable as within-breed predictions due to complexities associated with non-additive genetic effects (dominance and epistasis)
- Potential can be assessed by simulating the effects of major genes using real SNP genotypes on various populations

Introduction

- Toosi et al.,(2008) simulated genotypic and phenotypic data
  - Training in crossbred and MB populations
  - Successful selection of PB for MB performance

- Linkage Disequilibrium (LD)
  - Simulated LD in pure and MB populations may not accurately reflect real LD in beef cattle populations

Objective



Training Populations → Validation Populations

Multi-breed (MB)

Purebred (PB)

MB → PB

50K SNP

Purebred (PB)

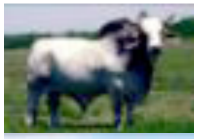
Multi-breed (MB)

PB → MB

50K SNP

50K SNP Datasets

MB Population (N=924)

	Angus	239
	Brahman	10
	Charolais	183
	Hereford	78
	Limousin	45
	Maine-Anjou	137
	Shorthorn	97
	South Devon	135

PB Population (N=1086)

	Angus	1086

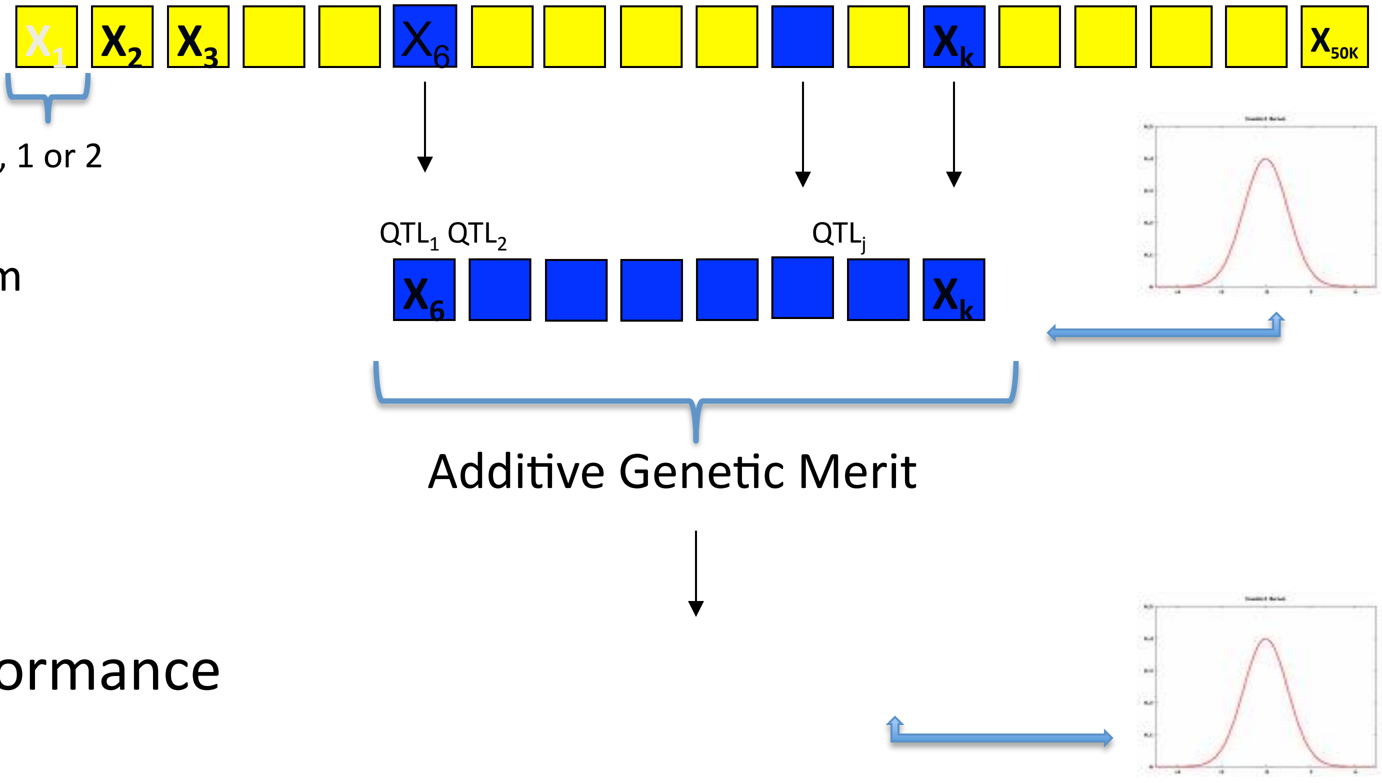
Simulation of Additive Genetic Merit and Phenotypic Performance

50K SNP

SNP chosen at random

$\text{QTL}_{50,\ 100,\ 250,\ 500}$

Phenotypic performance



Additive Genetic Merit

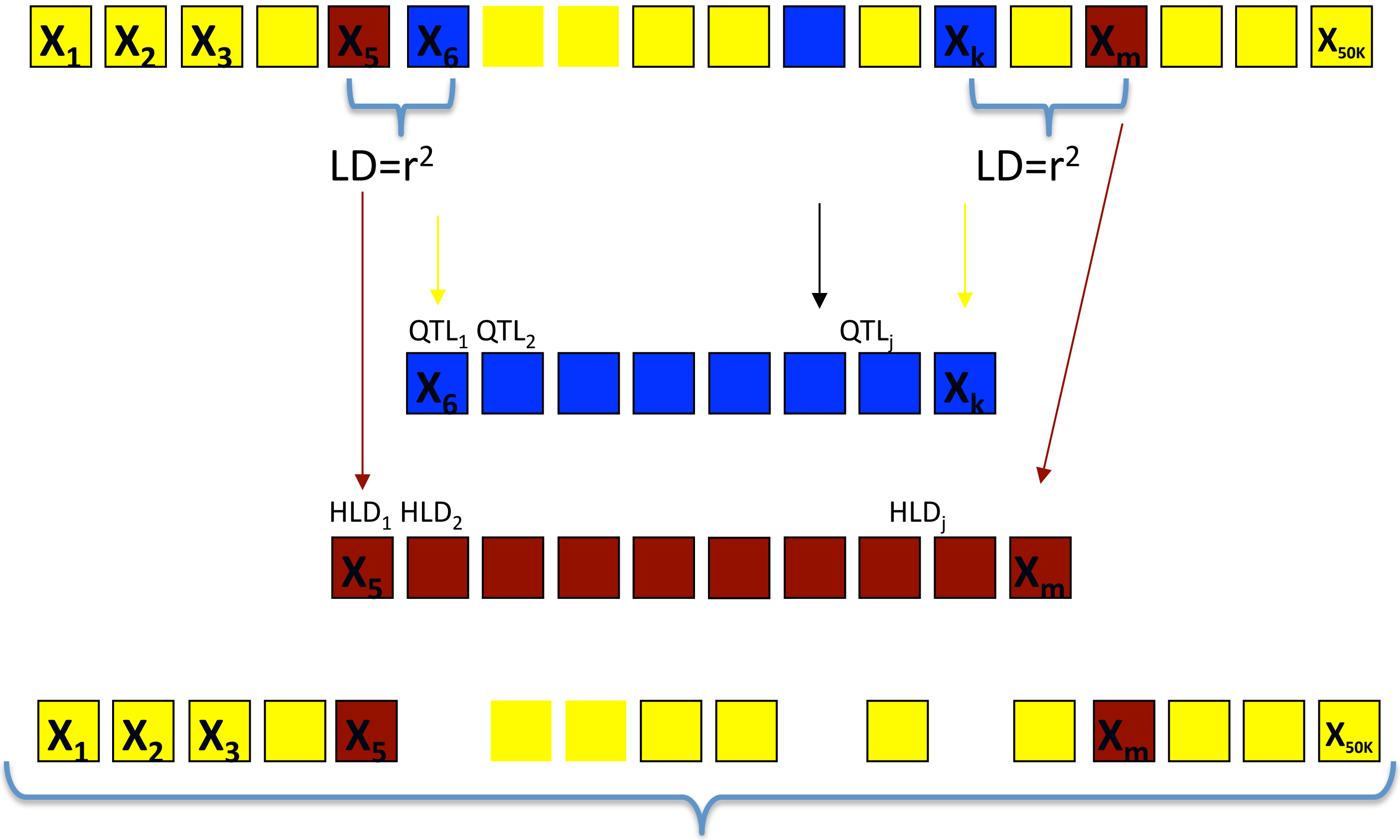
Marker Panels

50K SNP

QTL $_{50,\ 100,\ 250,\ 500}$

HLD$_{50,\ 100,\ 250,\ 500}$

50K w/o QTL



X_1 X_2 X_3 X_5 X_6 X_k X_m X_{50K}
LD=r^2
LD=r^2
QTL_1 QTL_2 QTL_j
X_6 X_k
HLD_1 HLD_2 HLD_j
X_5 X_m
X_1 X_2 X_3 X_5 X_m X_{50K}

Bayesian Analysis

Simulated Phenotypes/real 50k Data

- Effect of number of available markers

50 QTL	Train in Multibreed Validate in Purebreed	Train in Purebreed Validate in Multibreed
Just QTL	0.953	0.962
QTL + Best markers	0.931	0.938
QTL + 50k	0.766	0.842

Simulated Phenotypes/real 50k Data

- Effect of number of available markers

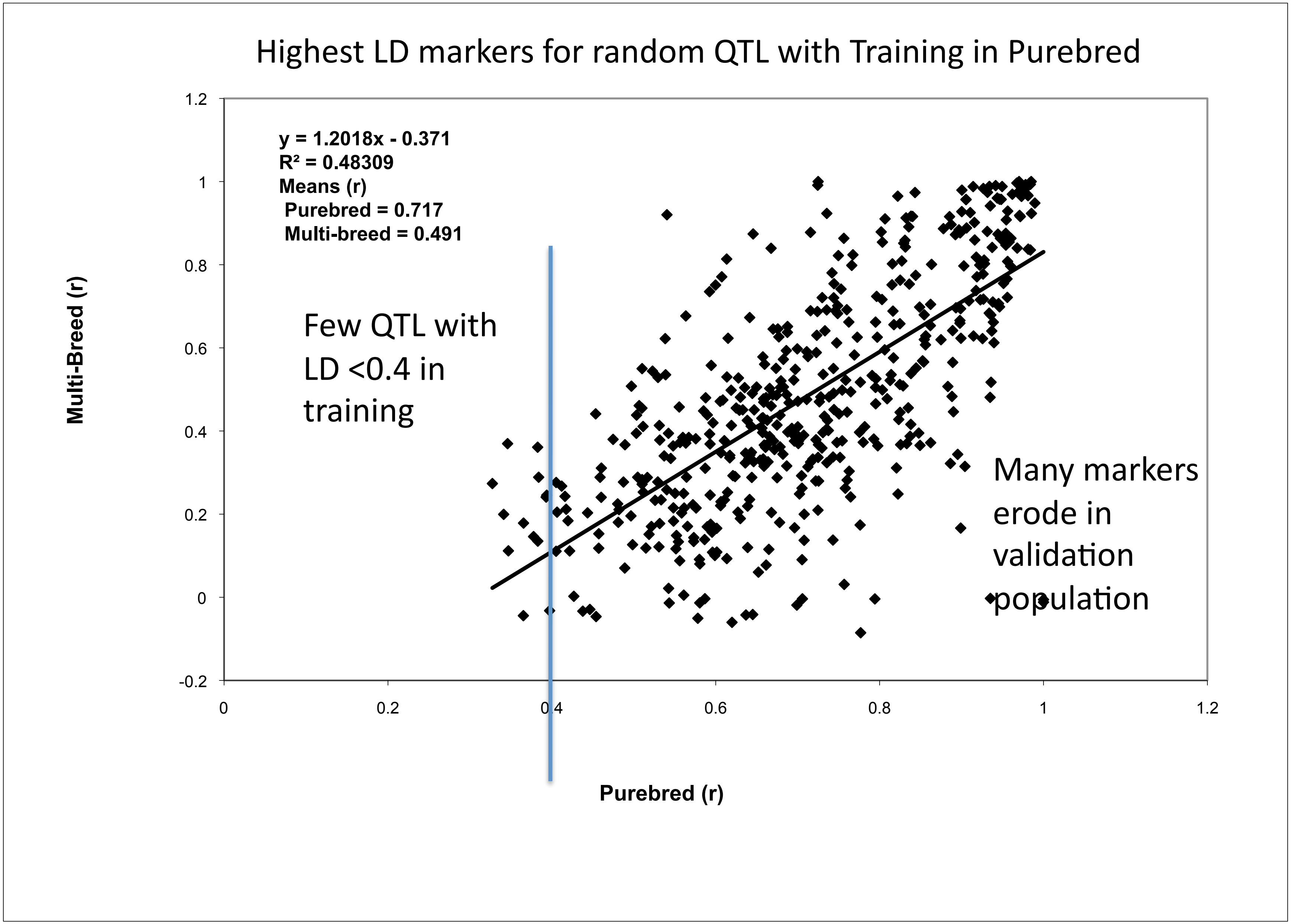
50 QTL	Train in Multibreed Validate in Purebreed	Train in Purebreed Validate in Multibreed
Just QTL	0.953	0.962
QTL + Best markers	0.931	0.938
QTL + 50k	0.766	0.842
Just Best markers	0.570	0.489
50k w/o QTL (real life)	0.388	0.422

Kizilkaya et al, ASAS, 2009

Effect of number of available markers

- Redundant markers reduce accuracy
  - Increased type I errors
- Accuracy suffers greatly when QTL not on panel
  - Not enough markers of sufficiently high LD to act as good proxies on a one-for-one basis
- Multibreed population generally inferior to purebred

Purebred or Crossbred



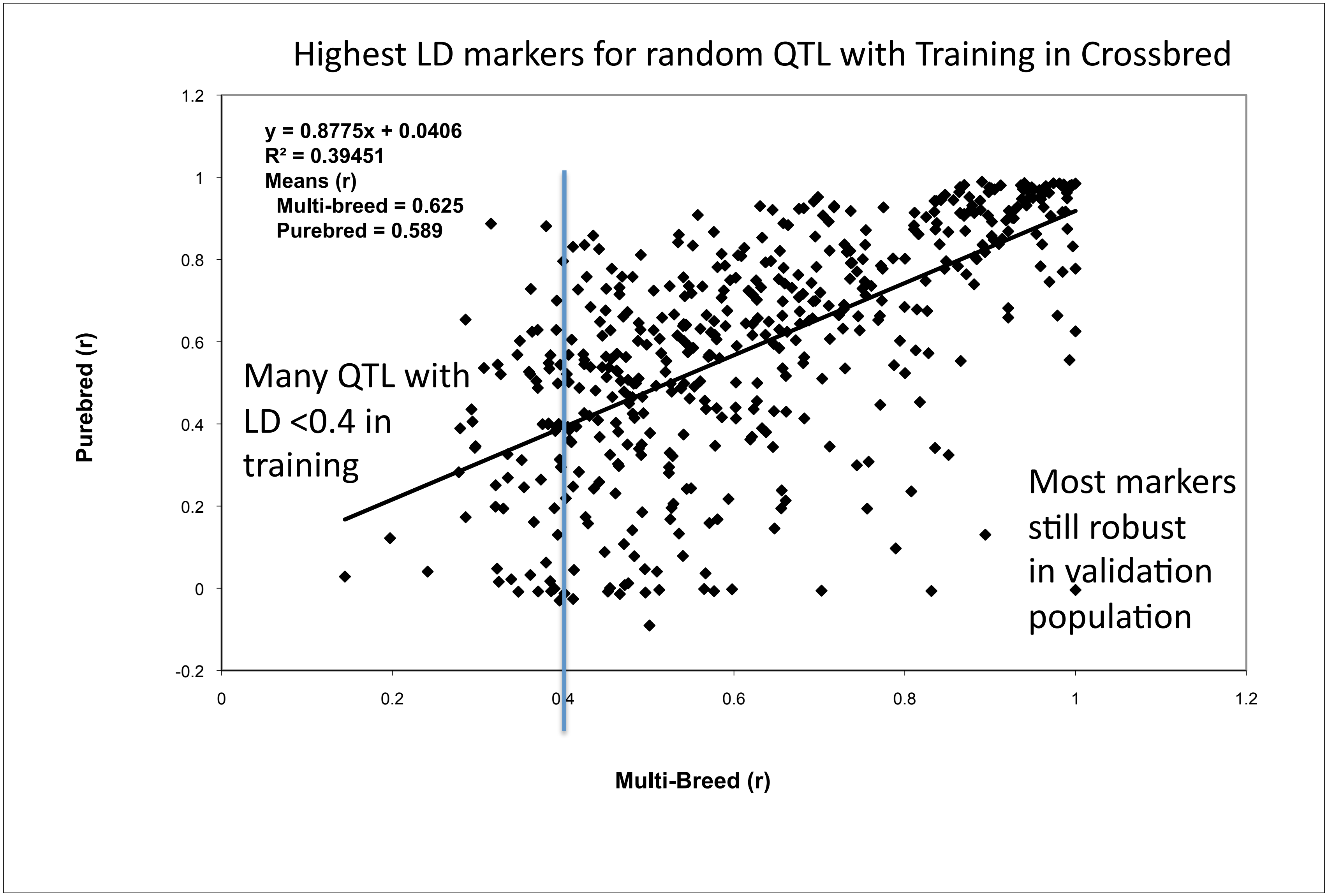
Highest LD markers for random QTL with Training in Purebred

y = 1.2018x - 0.371
$R^2 = 0.48309$
Means (r)
Purebred = 0.717
Multi-breed = 0.491

Few QTL with LD <0.4 in training

Many markers erode in validation population

Purebred or Crossbred



Many QTL with LD <0.4 in training

Most markers still robust in validation population

Effect of number of available markers

- Easier to find high LD markers in purebreds than multibreed populations because average LD is higher
  - Favors the use of purebred populations
  - Necessitates higher density SNP panels in multibreeds
- Markers chosen in purebreds may be less informative in multibreed populations as they will have less LD
- Markers that work well in multibreed populations seem to work just as well in purebred populations
- Nice to have larger multibreed populations & denser panels

Correlations between true and predicted genetic merits in validation population

Panel: QTL

QTL	MB→PB	PB→MB
50	0.953	0.962
100	0.938	0.941
250	0.840	0.853
500	0.720	0.786

Simulated Phenotypes/real 50k Data

- Effect of number of QTL

50k w/o QTL	Train in Multibreed Validate in Purebreed	Train in Purebreed Validate in Multibreed
50 QTL	0.388	0.422
100 QTL	0.289	0.308
250 QTL	0.247	0.276
500 QTL	0.200	0.299

- Identical trends when panel comprises QTL only
- These correlations a/c for < 20% variation at best

Correlations between true and predicted genetic merits in validation population

Panel: HLD

QTL	MB→PB	PB→MB
50	0.570	0.486
100	0.513	0.480
250	0.510	0.429
500	0.372	0.391

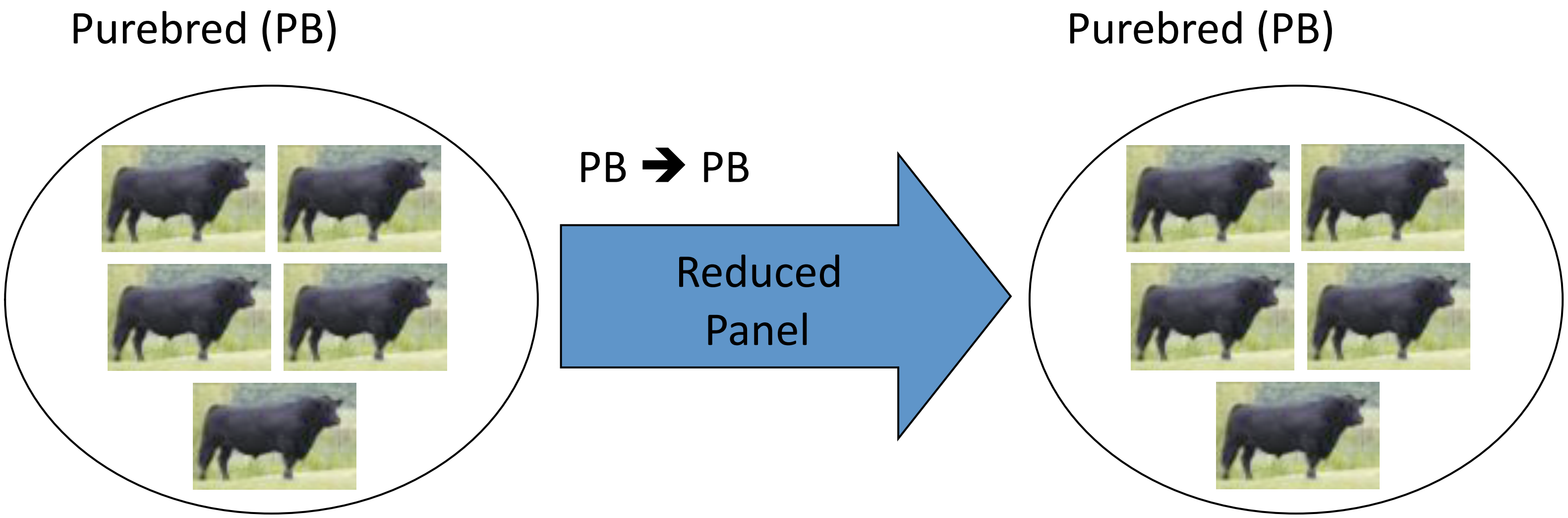
Average LD between QTL and HLD marker in PB or MB populations

HLD to QTL chosen from	HLD-QTL LD assessed in	
	PB	**MB**
PB	0.549	0.322
MB	0.412	0.408

Conclusions

- MB population
  - A good choice to carry out genomic selection
  - Reasonably accurate estimate of genetic merits of selection candidates in a PB population
- Accuracy of genetic merit in genomic selection
  - Higher with fewer QTL
  - Erodes when more uninformative SNPs added
- The extent of LD hence r^2 are highly variable
  - Lower average r^2 in MB than PB populations
  - No complete LD for all QTL with SNPs
  - Denser markers are needed

Training and Validation



Purebred (PB)

PB → PB

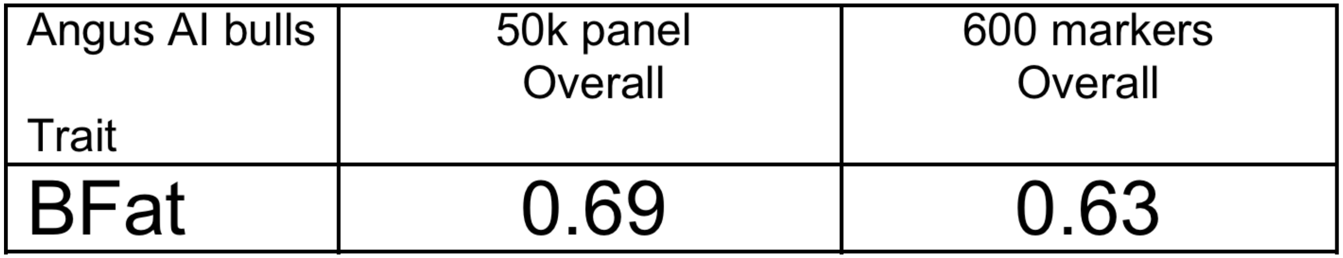
Reduced Panel

Purebred (PB)

Reduced panel within-breed selection

- Two-stage Bayesian analysis
  - Run all 50k markers
    - in each of the three training sets (2&3, 1&3, 1&2)
  - Select the best 600 markers on model frequency and genomic coverage
  - Rerun the training and validation analyses using only the markers on the 600 marker panel

50k versus 600 markers



Angus AI bulls Trait	50k panel Overall	600 markers Overall
BFat	0.69	0.63

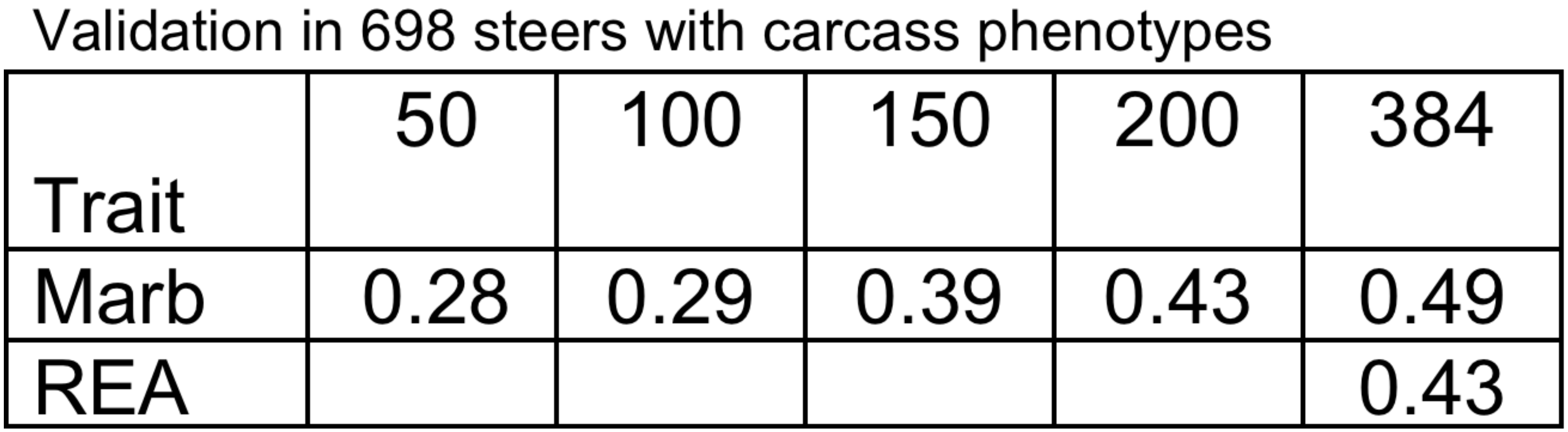
50k versus 600 markers

Angus AI bulls Trait	50k panel Overall	600 markers Overall
BFat	0.69	0.63
CED	0.59	0.61
CEM	0.53	0.55
Marb	0.70	0.67
REA	0.62	0.56
SC	0.55	0.51
WWD	0.52	0.49
YWT	0.56	0.55

384 SNP Panels

- Panels of 600 markers per trait for 8 traits would require a single panel of 4,800 markers
- Technology is moving such that larger panels are costing the same as smaller panels used to, rather than reducing the cost of smaller panels
- Significantly cheaper panels are currently limited to 384 (or less) SNP
  - Allow 100 or so of the best SNP for 3-4 key traits

Even Smaller Panels



Validation in 698 steers with carcass phenotypes

Trait	50	100	150	200	384
Marb	0.28	0.29	0.39	0.43	0.49
REA					0.43

Validation in New AI Bulls

Trait	50k	600	384
Validation	3-way		275
BFat	0.69	0.63	0.32
Marb	0.70	0.67	0.59
REA	0.62	0.56	0.58
YWT	0.56	0.55	0.35
CCWT			0.44
HP			0.39

Summary – beef cattle in US

- 50k within breed (like 5-15 progeny)
- 50k across breed
  (like 1 individual record or 5 progeny)
- Reduced panel within breed
  (varies up to 50k accuracy)

Validation Statistics

Validation Statistics

- Proportion of additive variation accounted for by the genomic prediction
  - Molecular BV used as an observation

1/ Multivariate model using the MBV as a trait to estimate (eg ASREML) the genetic correlation

2/ Reduction in estimated sire variance when the MBV is included as a fixed effect in the model

3/ Regression of phenotype on MBV

Thallman et al, 2009 BIF

Thallman et al, 2009 BIF

Data on 1,000 animals representing 100 sires

		Proportion of additive variance explained by MBV			
		BVN res cov estd	BVN res cov=0	Reduction	Regression
heritability	rg				
Data Simulated from Additive Model Only					
0.1	0.04	0.11	0.08	0.02	~~0.05~~
0.1	0.16	0.21	0.23	0.17	~~0.21~~
0.1	0.36	0.38	0.44	1.40	~~6.62~~
0.1	0.64	0.54	0.64	0.29	~~-0.23~~
0.3	0.04	0.06	0.05	0.04	0.05
0.3	0.16	0.17	0.19	0.15	0.20
0.3	0.36	0.35	0.40	0.35	0.42
0.3	0.64	0.64	0.68	0.66	0.83
0.5	0.04	0.05	0.05	0.04	0.05
0.5	0.16	0.16	0.18	0.16	0.18
0.5	0.36	0.35	0.39	0.36	0.39
0.5	0.64	0.63	0.66	0.63	0.72

http://www.bifconference.com/bif2009/proceedings/C4_5_pro_Quass.pdf

Some observations on across-breed prediction in dairy cattle

Comparison of the 5-SNP window variance in unrelated animals

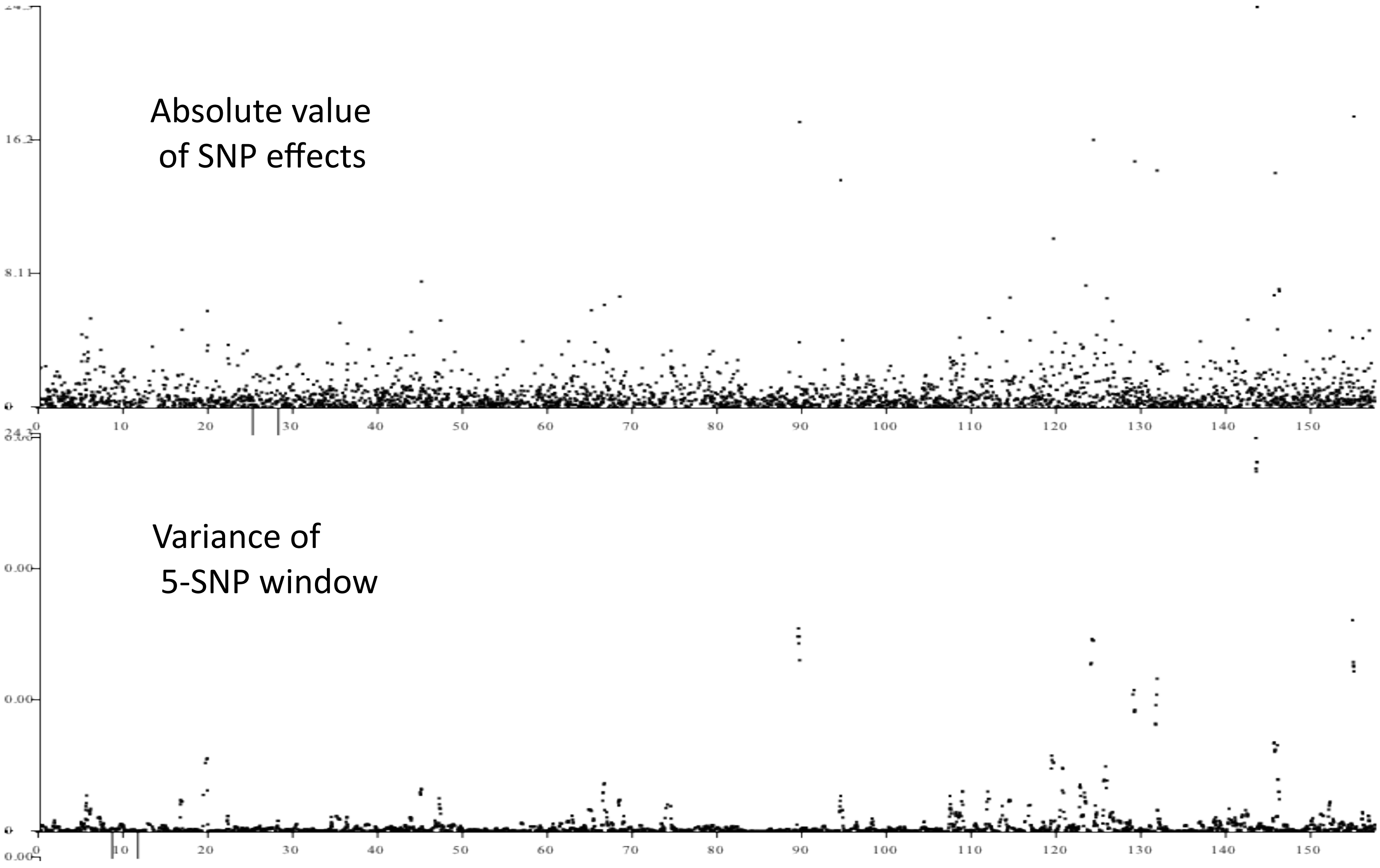
Holstein (HO) using 8512 bulls
Jersey (JE) using 1915 bulls
Brown Swiss (BS) using 742 bulls

Milk Production

Correlations Genomic & ProgenyTest

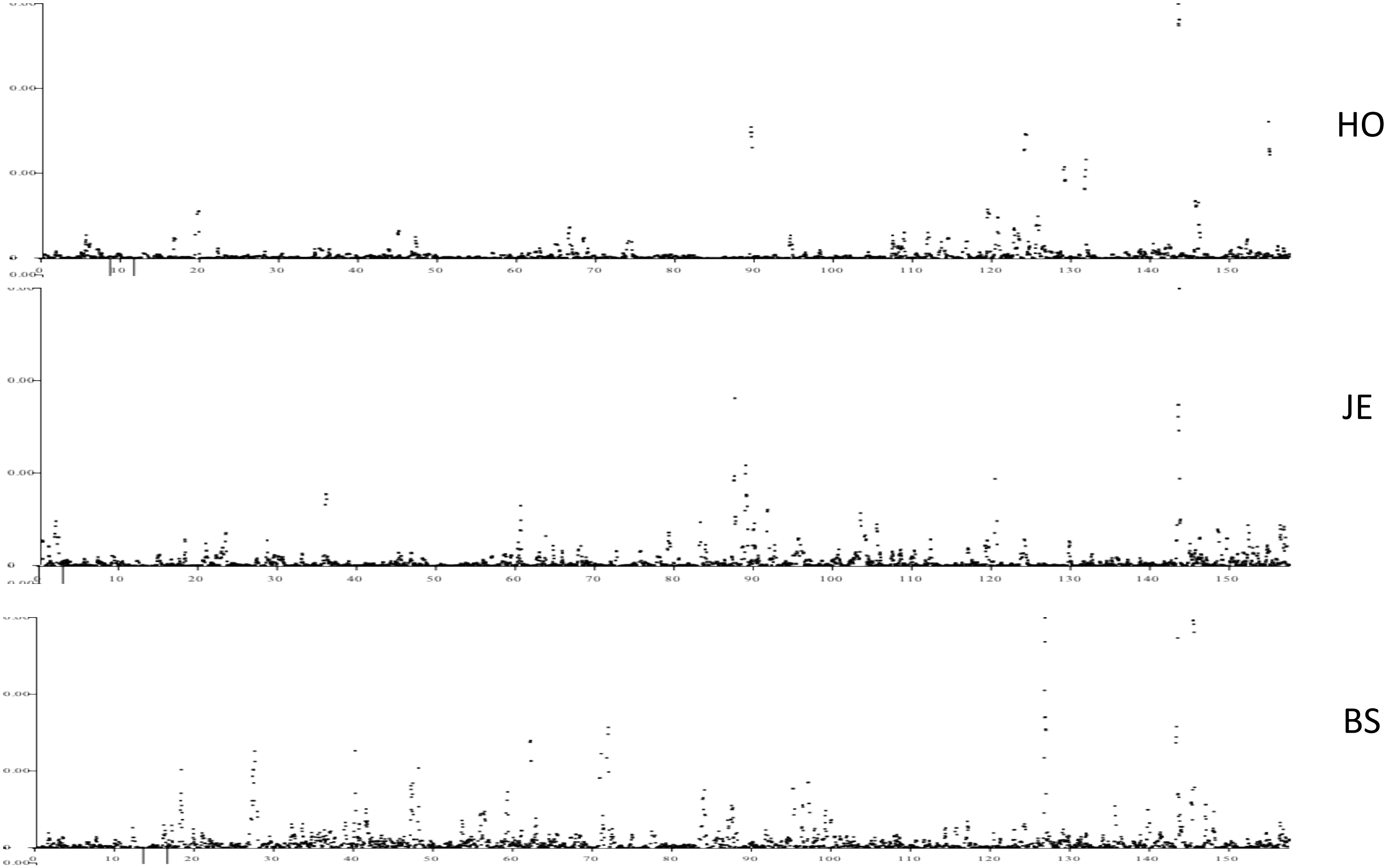
Method	Brown Swiss	Jersey	Holstein
Bayes A	0.194	0.198	
	0.191	0.201	
Bayes B (π=0.9)	0.141	0.244	
+FindScale	0.143	0.247	
Bayes C (π=0.9)	0.141	0.180	
+FindScale	0.145	0.183	
+FindScale	0.077 (JE & HO)	0.197 (BS & HO)	0.253 (BS & JE)
Bayes C0	0.180	0.084	
+FindScale	0.184	0.082	
Bayes CPi	0.146	0.172	
+FindScale	0.152	0.169	

Holstein BTA1 Milk

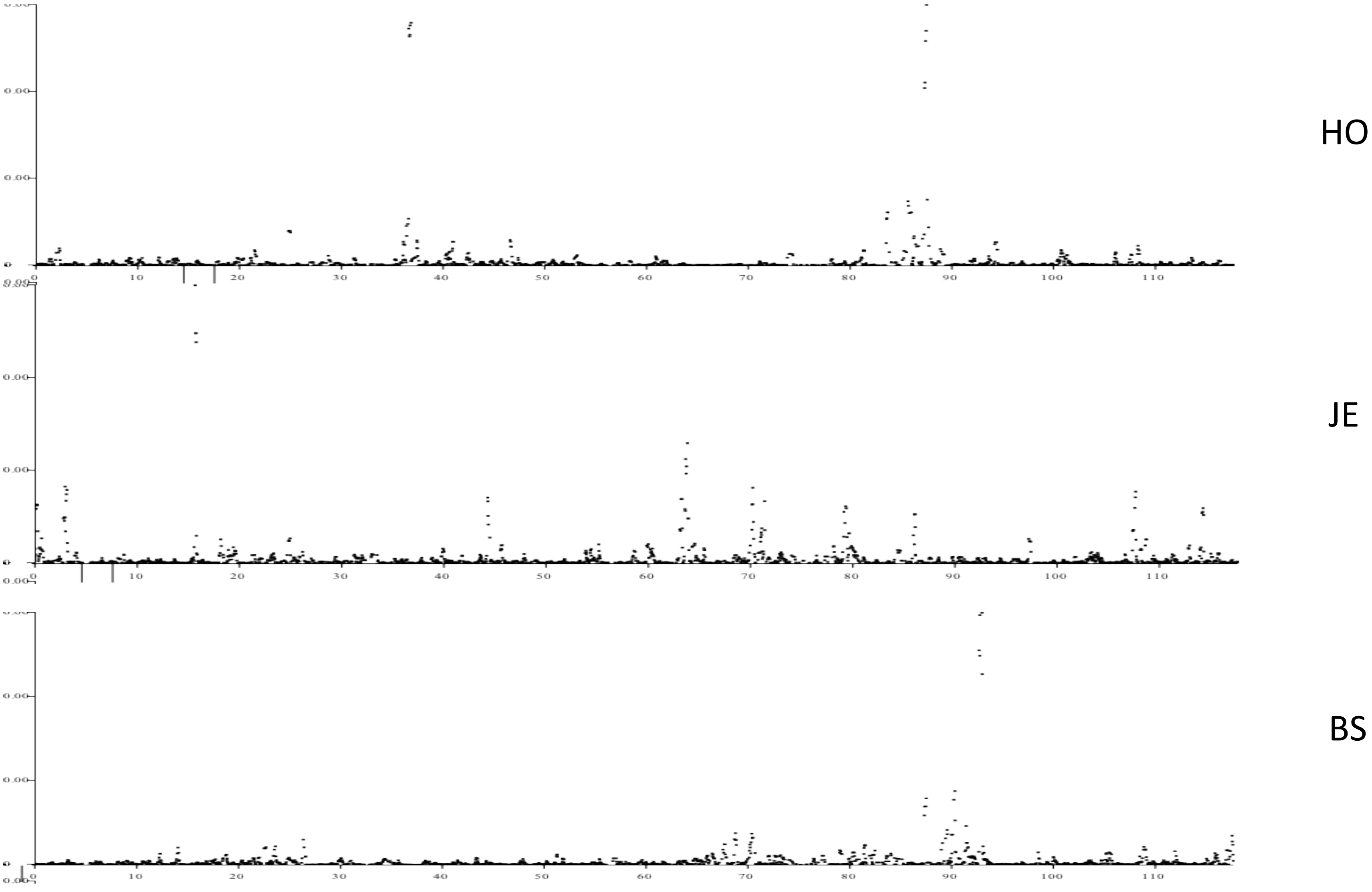

Absolute value of SNP effects

Variance of 5-SNP window

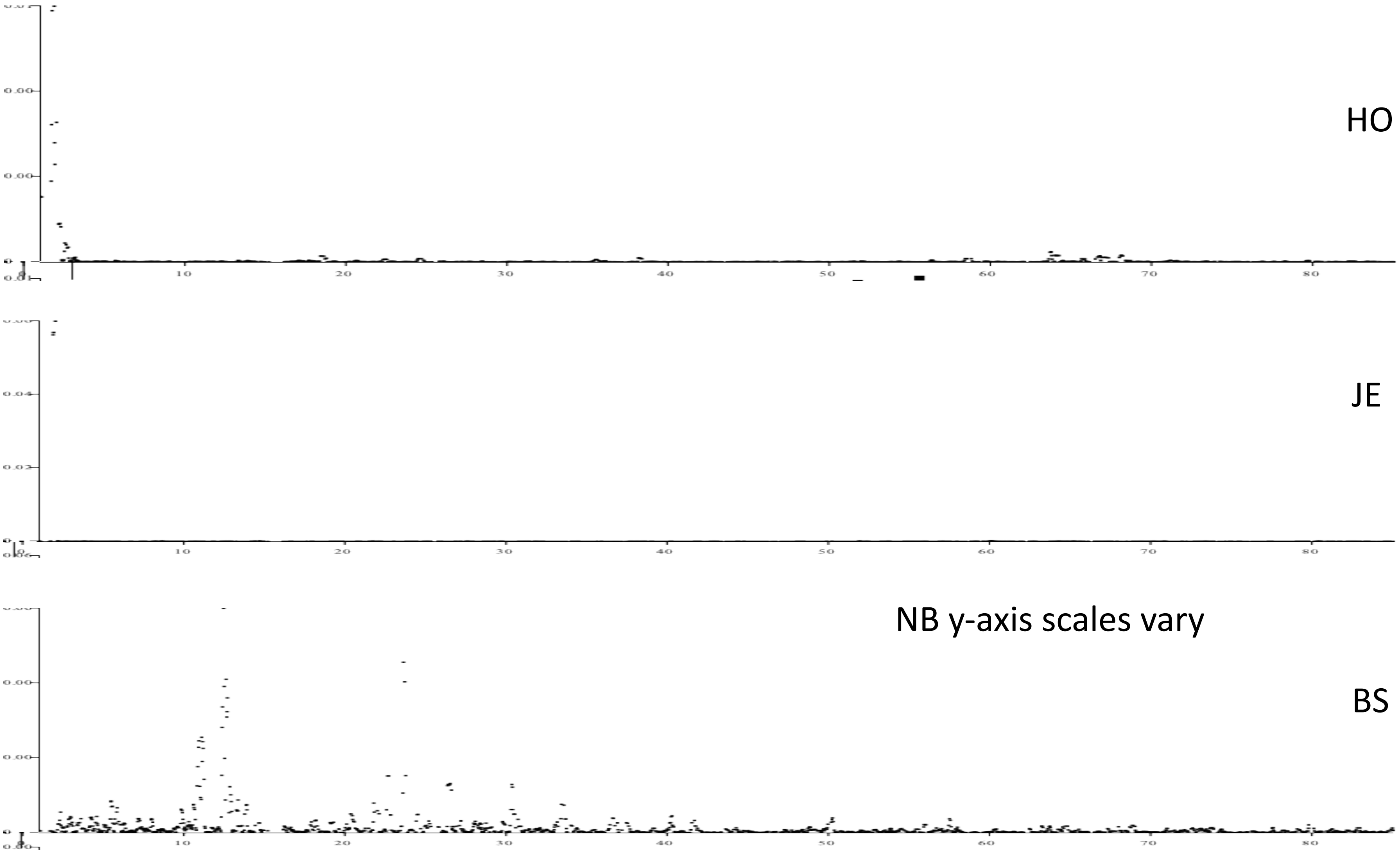
BTA1 - Milk


0.00
0.00
0.00
0
0.00
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
HO
0.00
0.00
0
0.00
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
JE
0.00
0.00
0
0.00
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
BS

BTA6 - Milk

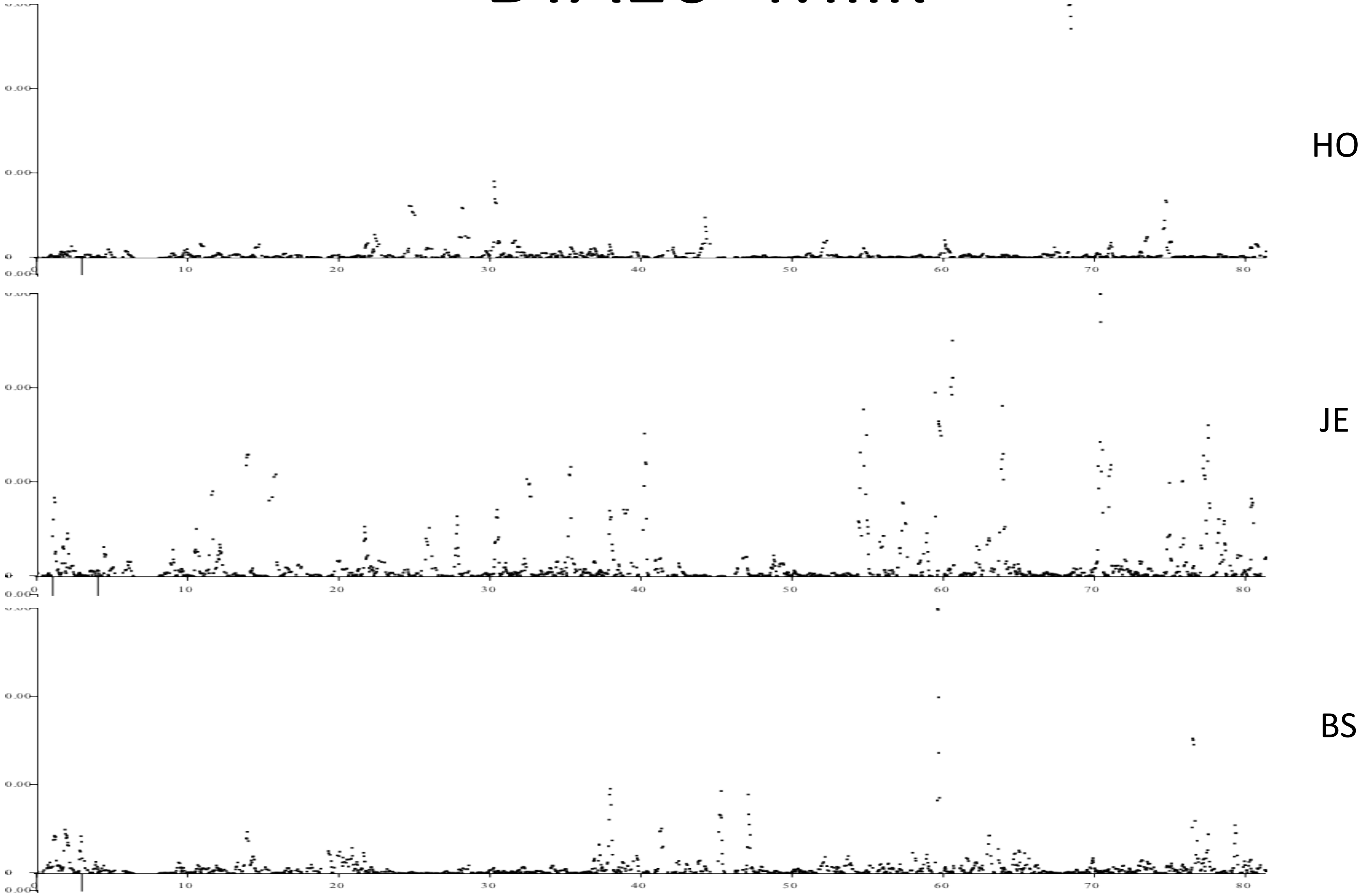


BTA- 14 (location of DGAT1)



NB y-axis scales vary

BTA16 -Milk



Genomic Selection
Estimation of the mixture fraction

Dorian Garrick

dorian@iastate.edu

Animal Science

1858 150 2008
Iowa State University

Animal Breeding & Genetics

Analytical Methods

	“BLUP”	BayesA	BayesB	BayesC	BayesCPi
Number SNP	All	All			
			1-pi	1-pi	1-pi
SNP Variance	constant			constant	constant
		variable	variable		
pi	NA	NA			
			known	known	
					unknown

Simulated Results

2000 animals	Number of QTL		
52,566 SNP markers	171	493	1184
BayesB(true pi)	0.88	0.82	0.76
BayesB(inflated pi)	0.84	0.79	0.75
BayesB(0.50)	0.81	0.78	0.74
Bayes A=B(0)	0.82	0.77	0.74
“BLUP”=C(0)	0.64	0.72	0.70

True=#QTL/#markers; inflated=0.9 true; heritability=0.5
(Christian Stricker for Swiss Cattle Breeders)

pi matters!

How do you know pi?

Mixture Models (model selection)

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i \delta_i + \mathbf{e}$$

$estimate\ \delta_i,\ \sigma_a^2\ and\ \sigma_e^2$

$BayesC\ (known\ \pi)\ "BLUP" = C(0)$

$\pi =\ fraction\ loci\ with\ no\ effect$

$estimate\ \pi\ \ prior\ U[0,1],\ \delta_i,\ \sigma_a^2\ and\ \sigma_e^2$

$BayesC\pi$

Fernando et al 2009
(in preparation)

Simulated Results

- 2000 unlinked loci, Q QTL, N training animals, 1000 validation animals, heritability =0.5

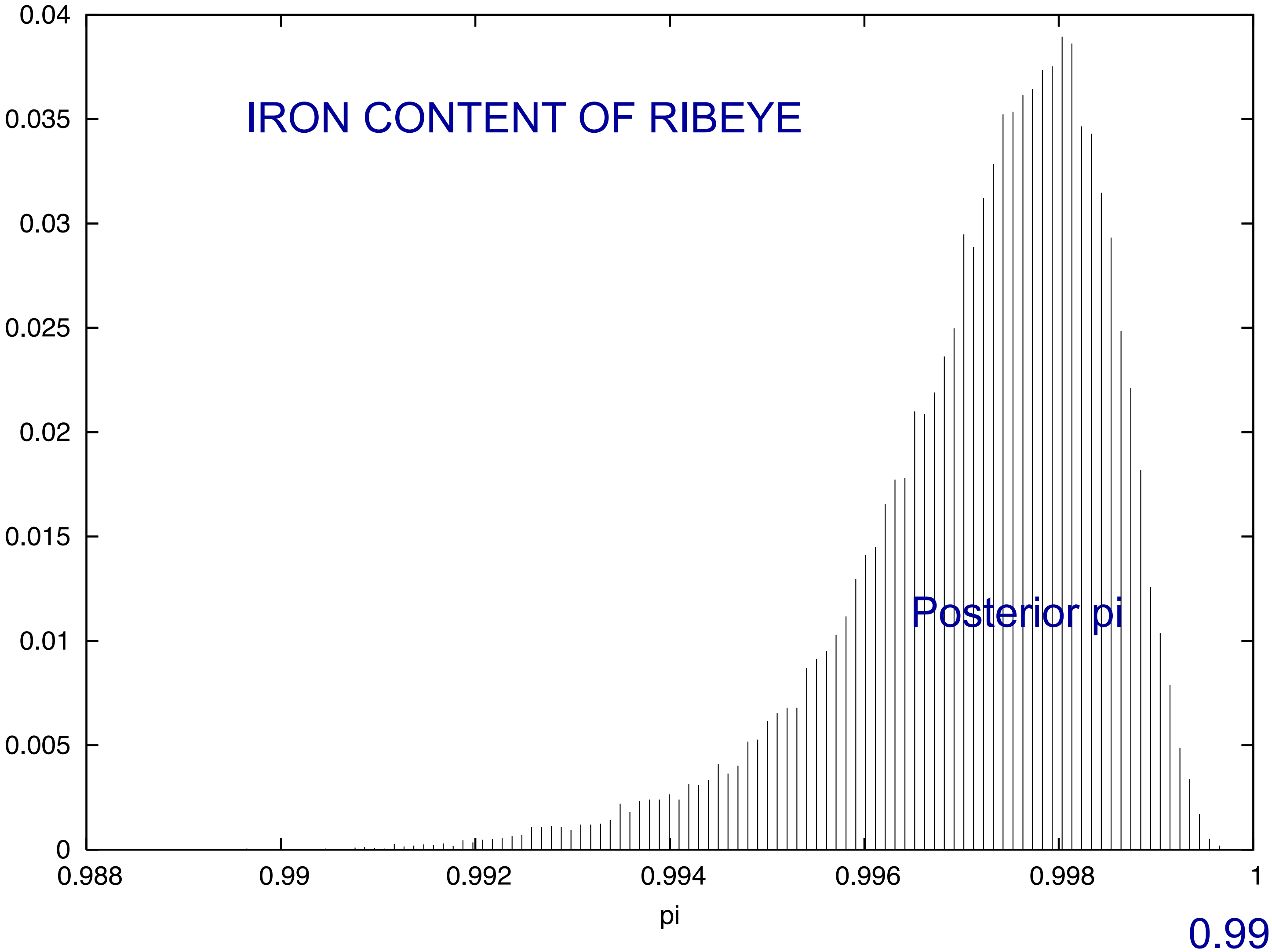
			BayesB (.5) **(pi known)**	**Bayes Cpi** **(pi unknown)**	
N	Q	pi	**Correlation**	pi-hat	**Correlation**
2000	10	0.995	**0.937**	0.994	**0.995**
2000	200	0.90	**0.834**	0.899	**0.866**
2000	1900	0.05	**0.571**	0.202	**0.613**
4000	1900	0.05	**0.722**	0.096	**0.763**

Simulated Results - Real 50k

- Train 1086 purebred animals
- Validate 984 multibreed animals
- Random 50 SNP = QTL (pi=0.999)
- Heritability=0.25

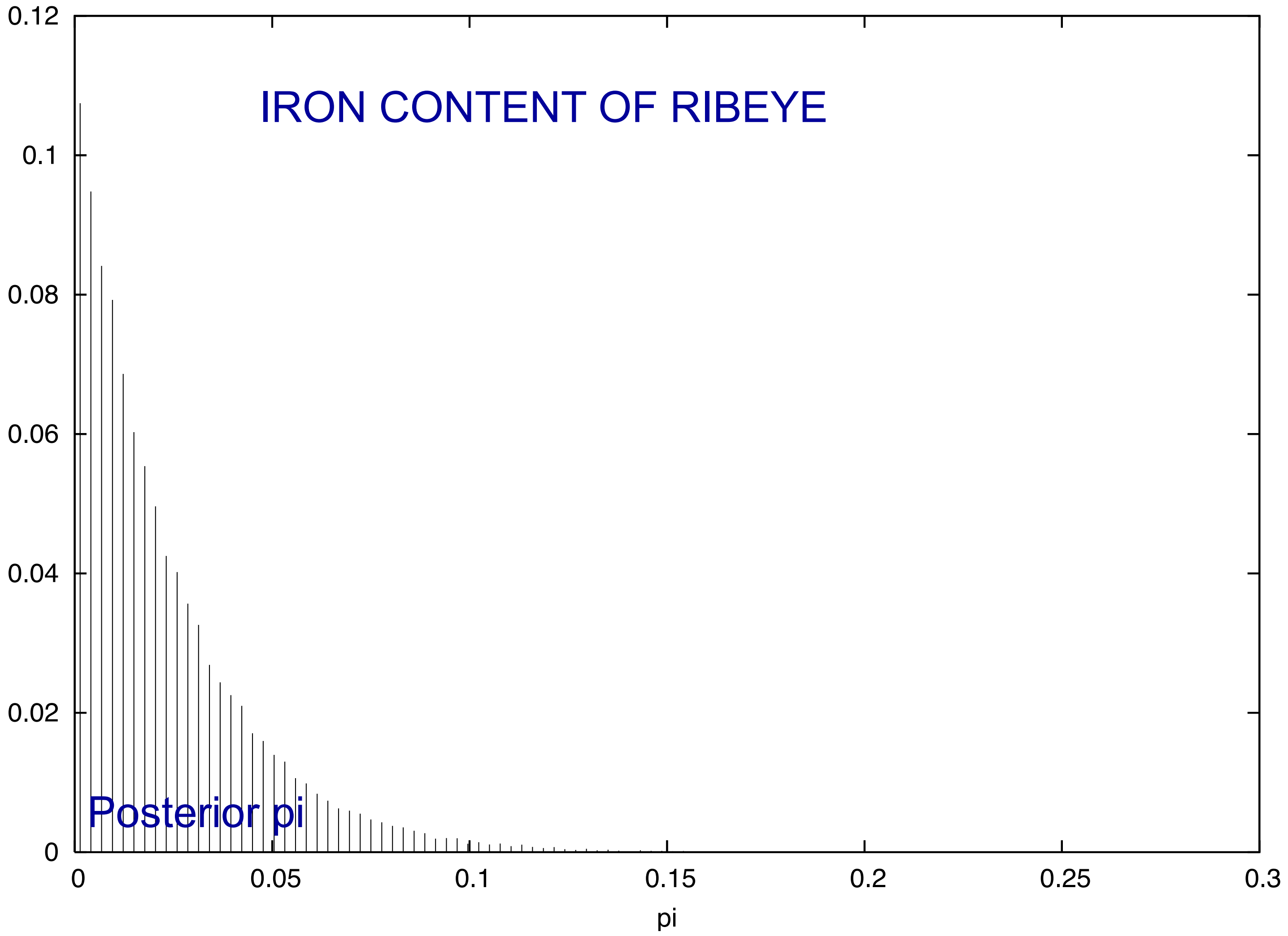
	Correlation True and Predicted Merit		
Assumed pi	Bayes B (pi known)	Bayes C (pi known)	Bayes Cpi (pi unknown)
0.999	0.86	0.86	
0.25	0.70	0.26	
N/A			0.86

50,000 markers (bovine)

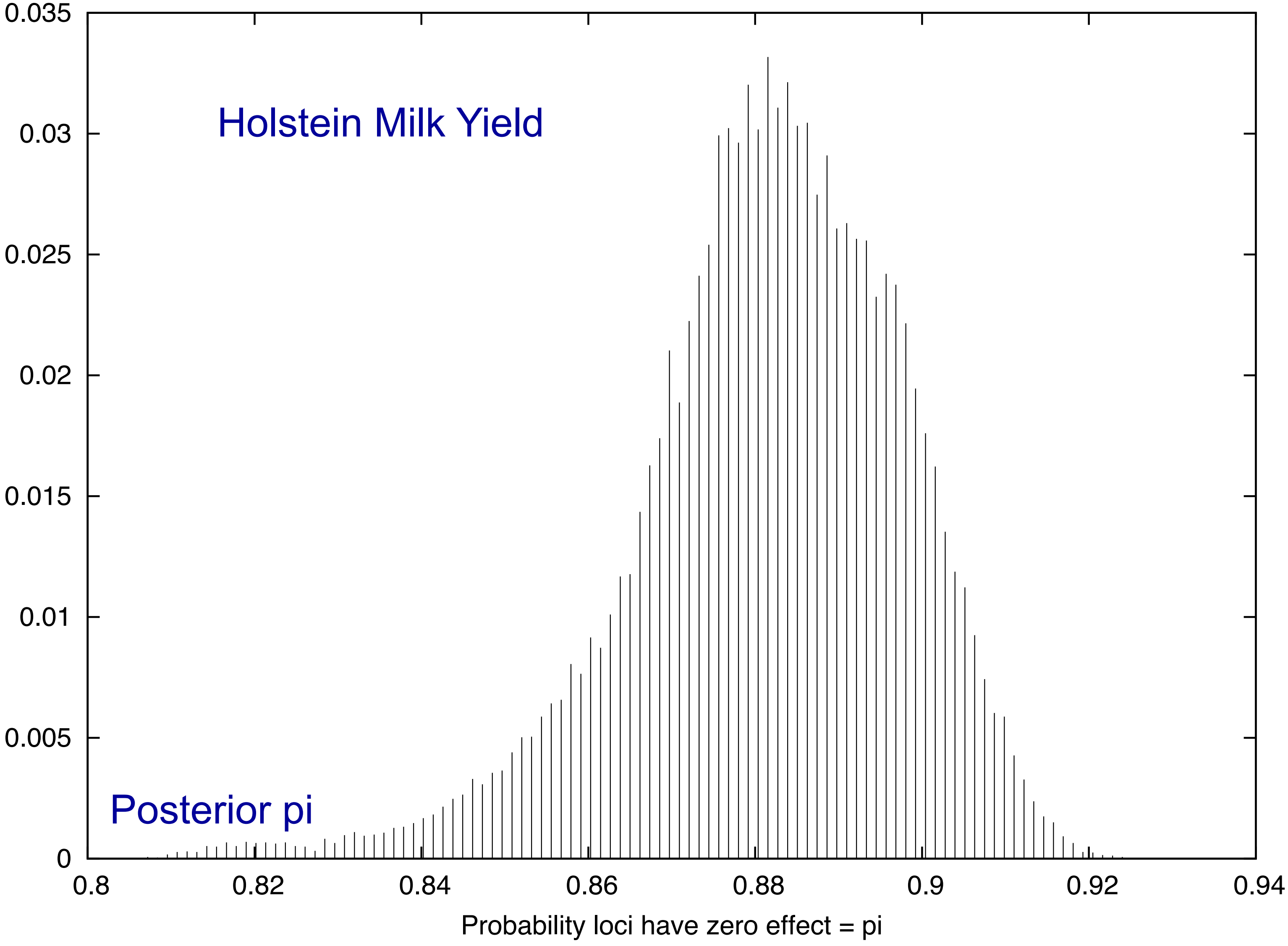


0.998=100 loci

“Best” 100 markers



Bayes C pi on 8,300 bulls



Summary

- The mixture fraction (pi) is an important parameter in determining the relative performance of alternative methods for genomic selection
- The mixture fraction can be concurrently estimated from the data, more easily in Bayes C than in Bayes A

Genomic Selection

Scale Factor Estimation

Dorian Garrick

dorian@iastate.edu

ANIMAL SCIENCE

150 1858 2008 IOWA STATE UNIVERSITY

Animal Breeding & Genetics

Bayes A

Prior $\left(a_i / \sigma_i^2 \right) \sim N\left(0, \sigma_i^2\right)$

$$\sigma_i^2 \sim v_a S_{v_a}^2 \chi_{v_a}^{-2}$$

Meuwissen, Hayes & Goddard (2001)

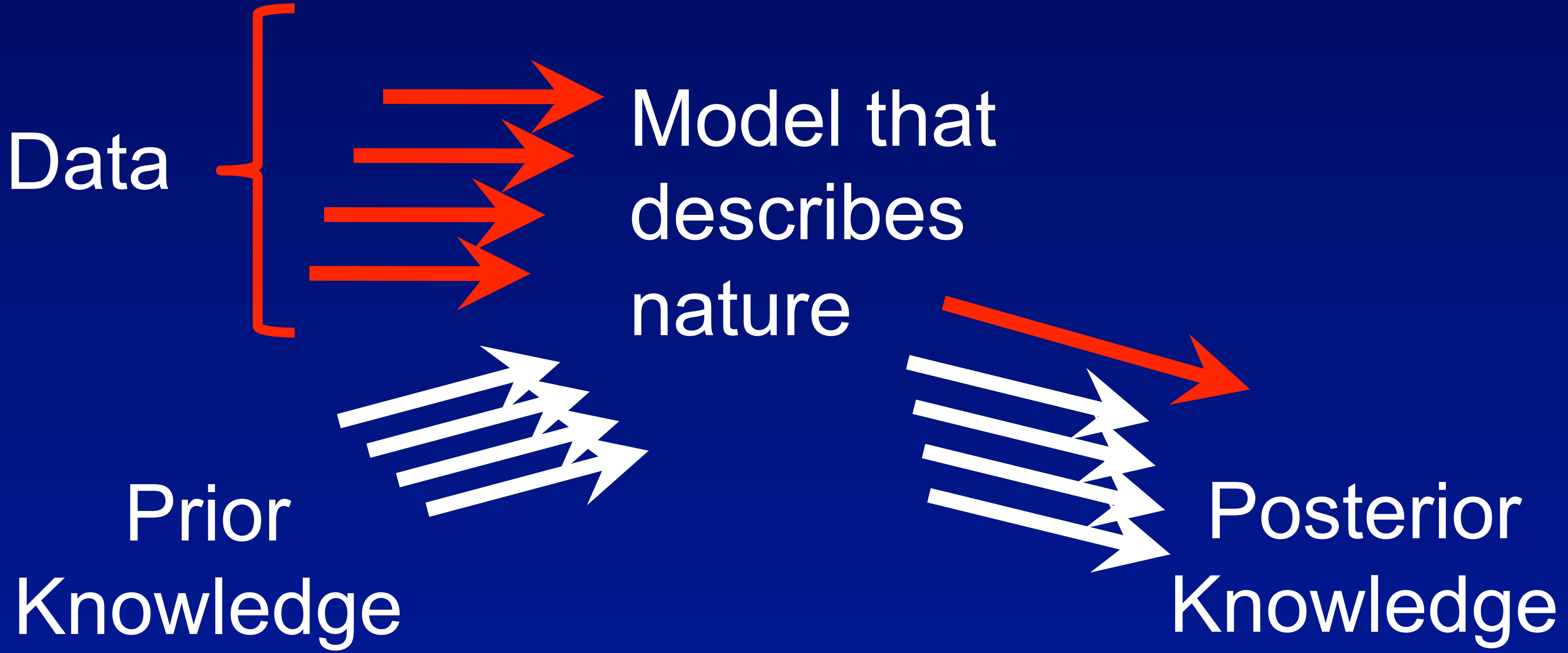
$$so\ that\ \ a_i \sim (iid)t\left(0, S_{v_a}^2, v_a\right)$$

Sorensen & Gianola, 2002

$$Assume \quad \sigma_i^2 = \frac{V_a}{\sum_i 2p_i(1-p_i)} = \frac{V_a}{k2\bar{p}(1-\bar{p})}$$

$$so \quad S_{v_a}^2 = \frac{\left(v_a - 2\right)V_a}{v_a k 2\bar{p}(1-\bar{p})} \ for\ k\ SNP$$

BayesA/B not Bayesian Methods



Data

Model that describes nature

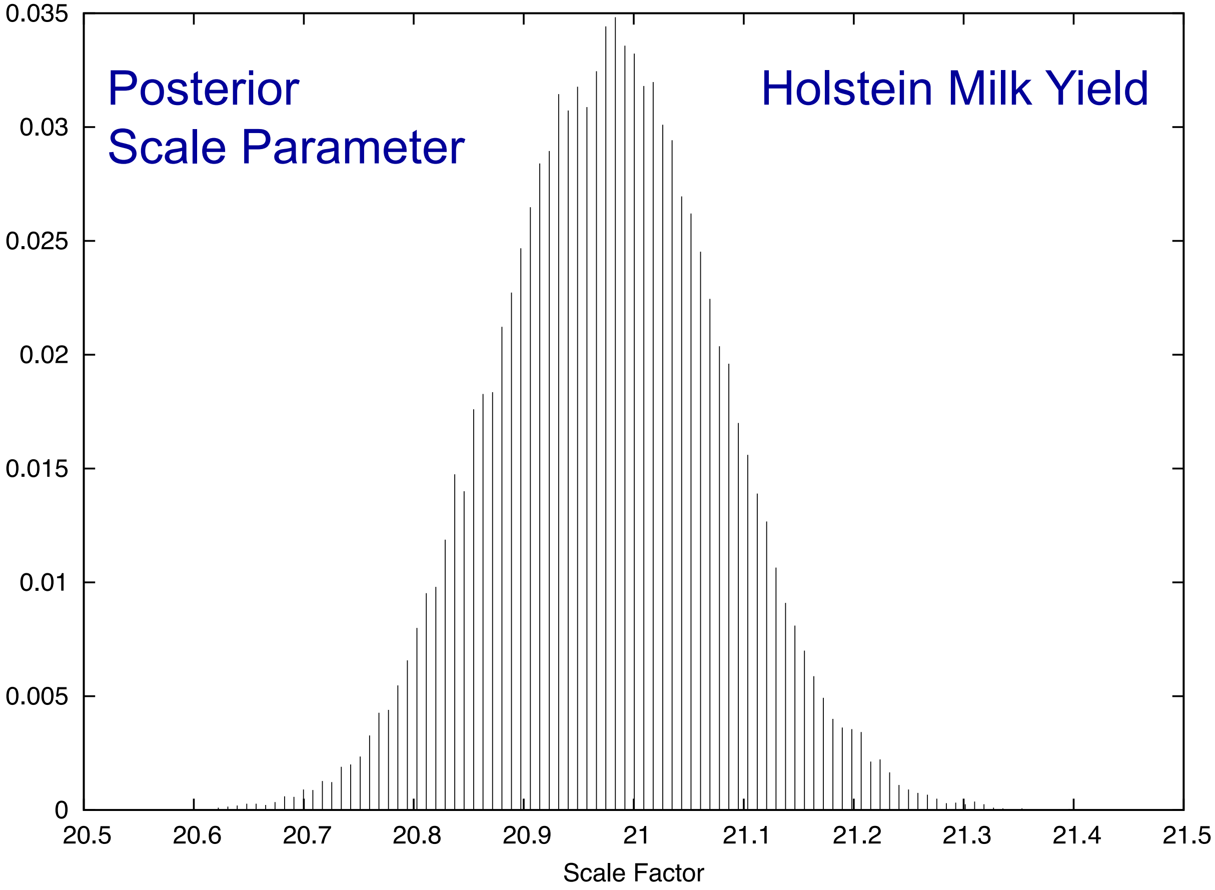
Prior Knowledge

Posterior Knowledge

Gianola et al “Bayesian Alphabet” 2009

But they work very well in practice!

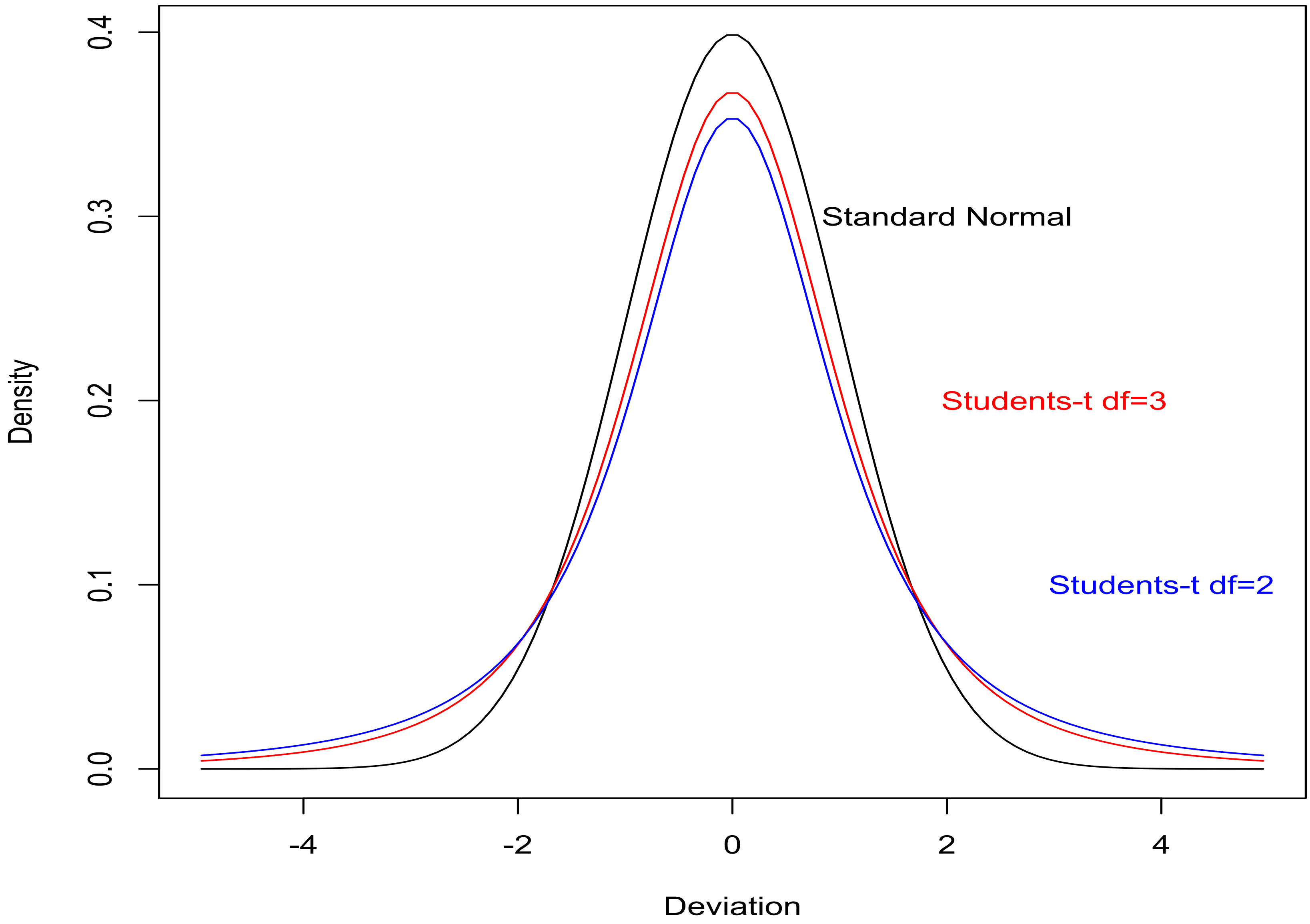
Bayes A on 8,300 bulls



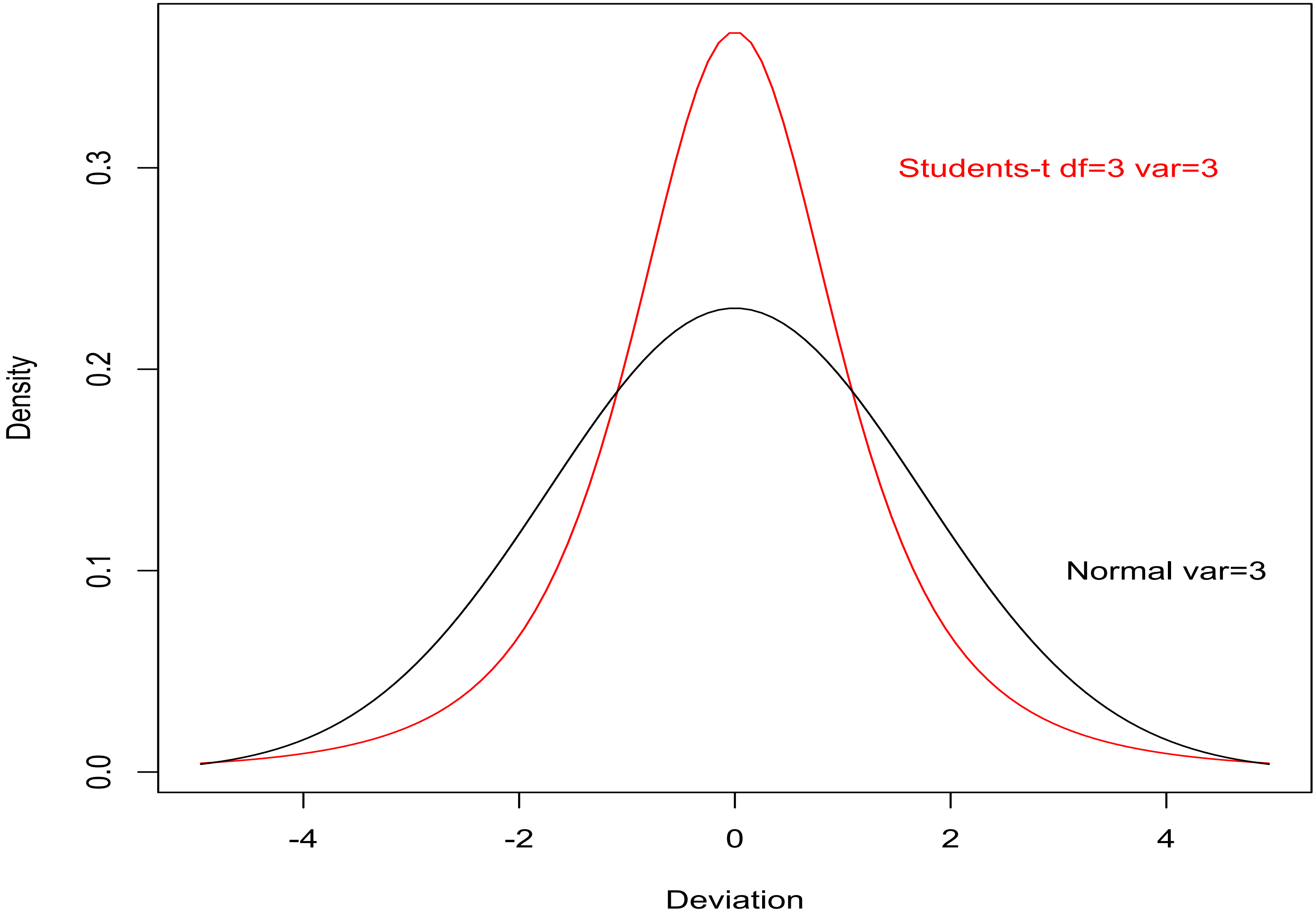
$$S_{v_a}^2 = \frac{(v_a - 2)V_a}{v_a k 2\overline{p}(1-\overline{p})} = \frac{(4-2)\times 646100}{4\times 43043\times 0.36} = 20.85$$

Alternative Distributions (to the normal)

Students-*t* Distributions



At Constant Variance



Real SNPs - Simulated Traits

- Training Data
  - 2,869 Angus and Angus-cross (steers)
- Validation Data
  - 1,086 ISU Angus
  - 972 CMP half-sib groups representing 8 sire breeds (predominantly Angus)
- Random 50 or 500 SNPs were QTL
- Panels were the QTL, 50k+QTL, 50k-QTL

Error Distributions

- The impact of normally distributed vs students-*t* distributed residual effects in the true and/or the fitted model
  - Simulated effects had 3 degrees of freedom
  - Fitted effects estimated degrees of freedom simultaneously with all other relevant parameters

50 QTL

True = Markers Normal Residuals Normal
Fitted = Markers Normal Residuals Normal

50QTL	BayesC	Training-Y	Training-G	ISU	CMP
50SNP=QTL	π=0.	0.725	0.991	0.988	0.991
50k+QTL	π=0.999	0.743	0.975	0.973	0.974
50k-QTL	π=0.999	0.661	0.763	0.649	0.591
50k-QTL	Cpi π=0.996	0.763	0.806	0.657	0.599

Fitted = Markers Normal Residuals *t*

50QTL	BayesC	df	Training-G	ISU	CMP
50SNP=QTL	π=0.	91	0.991	0.988	0.991
50k+QTL	π=0.999	91	0.975	0.973	0.974
50k-QTL	π=0.999	80	0.764	0.650	0.590
50k-QTL	Cpi π=0.996	59	0.807	0.658	0.598

500 QTL

True = Markers Normal Residuals Normal
Fitted = Markers Normal Residuals Normal

500QTL	BayesC	Training-Y	Training-G	ISU	CMP
50SNP=QTL	π=0.	0.776	0.932	0.910	0.910
50k+QTL	π=0.99	0.878	0.821	0.619	0.620
50k-QTL	π=0.99	0.853	0.760	0.370	0.318
50k-QTL	Cpi π=0.701	0.915	0.773	0.358	0.301

Fitted = Markers Normal Residuals *t*

500QTL	BayesC	df	Training-G	ISU	CMP
50SNP=QTL	π=0.	78	0.932	0.910	0.910
50k+QTL	π=0.99	57	0.821	0.619	0.620
50k-QTL	π=0.99	53	0.760	0.370	0.319
50k-QTL	Cpi π=0.701	51	0.771	0.352	0.285

Conclusion (1)

- There is no real harm in fitting a model that assumes residuals follow a students-*t* distribution with unknown df when the true model has normally distributed residuals

50 QTL

True = Markers Normal Residuals *t*
Fitted = Markers Normal Residuals Normal

50QTL	BayesC	Training-Y	Training-G	ISU	CMP
50SNP=QTL	π=0.	0.552	0.977	0.977	0.973
50k+QTL	π=0.999	0.592	0.901	0.893	0.877
50k-QTL	π=0.999	0.551	0.664	0.529	0.472

Fitted = Markers Normal Residuals *t*

50QTL	BayesC	df	Training-G	ISU	CMP
50SNP=QTL	π=0.	3	0.989	0.988	0.987
50k+QTL	π=0.999	3	0.953	0.947	0.942
50k-QTL	π=0.999	3.6	0.724	0.599	0.531

500 QTL

True = Markers Normal Residuals *t*
Fitted = Markers Normal Residuals Normal

500QTL	BayesC	Training-Y	Training-G	ISU	CMP
50SNP=QTL	π=0.	0.613	0.848	0.800	0.800
50k+QTL	π=0.99	0.778	0.652	0.405	0.414
50k-QTL	π=0.99	0.763	0.608	0.270	0.247

Fitted = Markers Normal Residuals *t*

500QTL	BayesC	df	Training-G	ISU	CMP
50SNP=QTL	π=0.	3	0.897	0.869	0.868
50k+QTL	π=0.99	3.1	0.723	0.501	0.480
50k-QTL	π=0.99	3.4	0.669	0.324	0.268

Conclusion (2)

- If residuals follow a students-*t* distribution with few degrees of freedom, there are modest benefits of fitting models that estimates the degrees of freedom from the data

Marker Effects Distributions

- The impact of normally distributed vs students-*t* distributed marker effects in the true and/or the fitted model
  - Simulated effects had 3 degrees of freedom
  - Fitted effects estimated degrees of freedom simultaneously with all other relevant parameters

50 QTL

True = Markers Normal Residuals Normal

Fitted = Markers Normal Residuals Normal

50QTL	50k-QTL	Training-Y	Training-G	ISU	CMP
Bayes B	π=0.999	0.656	0.761	0.648	0.589
Bayes C	π=0.	0.905	0.765	0.345	0.300

Fitted = Markers *t* Residuals Normal

50QTL	50k-QTL	df	Training-G	ISU	CMP
Bayes C	π=0.999	31	0.770	0.646	0.580
Bayes C	π=0.	2	0.822	0.663	0.593

500 QTL

True = Markers Normal Residuals Normal

Fitted = Markers Normal Residuals Normal

500QTL	50k-QTL	Training-Y	Training-G	ISU	CMP
Bayes B	π=0.99	0.836	0.753	0.362	0.314
Bayes C	π=0.	0.916	0.770	0.348	0.281

Fitted = Markers *t* Residuals Normal

500QTL	50k-QTL	df	Training-G	ISU	CMP
Bayes C	π=0.99	48	0.762	0.370	0.319
Bayes C	π=0.	3.3	0.775	0.369	0.320

Conclusion (3)

- Recall the usual approaches (Bayes B or C) suffer from incorrect values of π
  - When π is correct, and effects are really normal, the estimated degrees of freedom are large and no harm is done to prediction accuracy
  - When π is too low, and effects are really normal, the estimated degrees of freedom are small, shrinking the effects of spurious markers and overcoming the erosion of accuracy from fitting too many markers

50 QTL

True = Markers *t* Residuals Normal

Fitted = Markers Normal Residuals Normal

50QTL	50k-QTL	Training-Y	Training-G	ISU	CMP
Bayes B	π=0.999	0.637	0.769	0.647	0.581
Bayes C	π=0.	0.891	0.732	0.319	0.274

Fitted = Markers *t* Residuals Normal

50QTL	50k-QTL	df	Training-G	ISU	CMP
Bayes C	π=0.999	19	0.767	0.646	0.587
Bayes C	π=0.	2.2	0.807	0.640	0.586

500 QTL

True = Markers *t* Residuals Normal

Fitted = Markers Normal Residuals Normal

500QTL	50k-QTL	Training-Y	Training-G	ISU	CMP
Bayes B	π=0.99	0.828	0.765	0.462	0.395
Bayes C	π=0.	0.907	0.754	0.298	0.247

Fitted = Markers *t* Residuals Normal

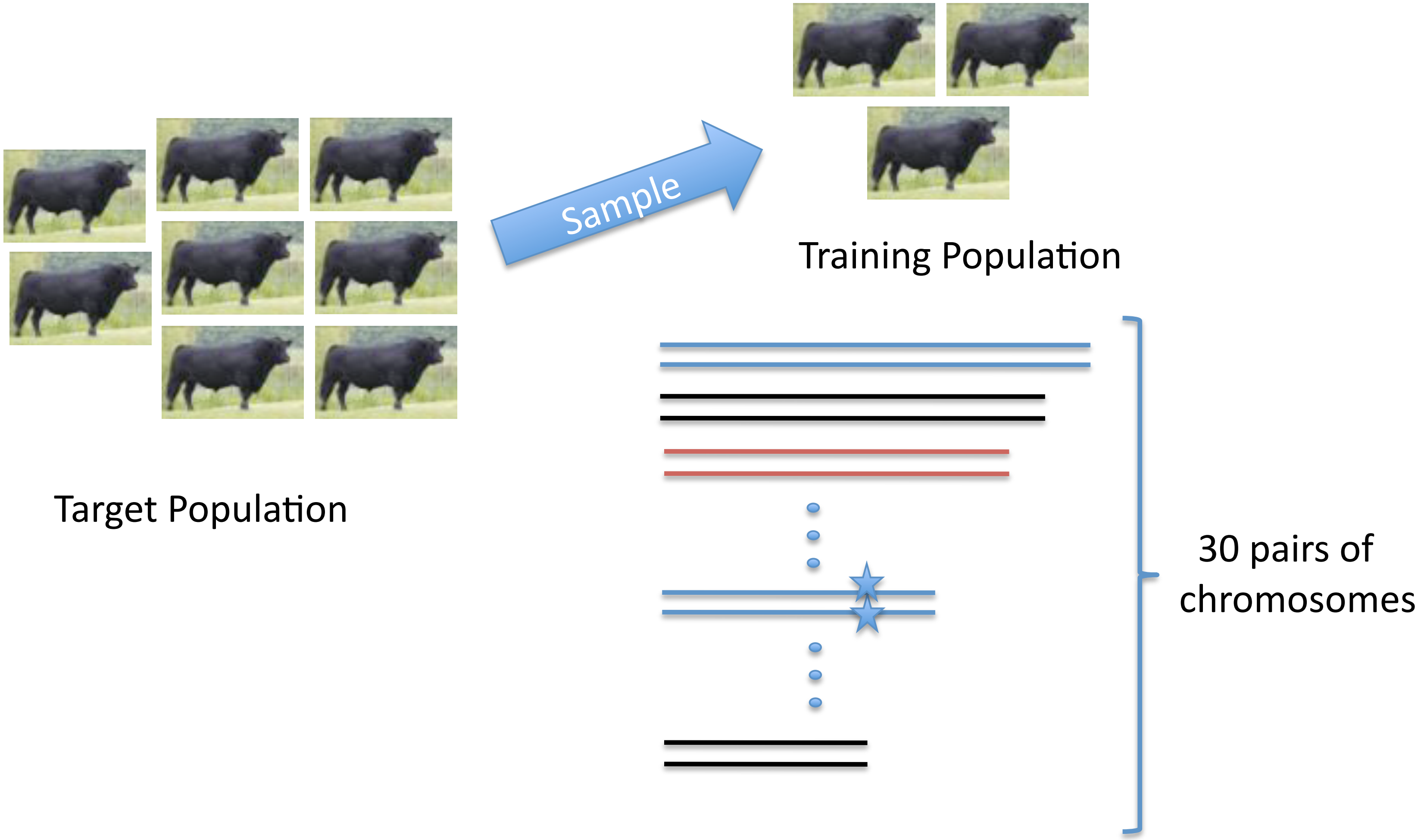
500QTL	50k-QTL	df	Training-G	ISU	CMP
Bayes C	π=0.99	8.7	0.779	0.476	0.404
Bayes C	π=0.	2.9	0.776	0.457	0.395

Conclusions (4)

- When marker effects are distributed as students-*t* with small degrees of freedom
  - there is little accuracy loss if appropriate π is used and effects are fitted as if normally distributed
  - When too many markers are in the model, that is π is too small, this has little impact on prediction if degrees of freedom are estimated from the data

Spurious Markers Effects Can Validate in Relatives

Goal in Marker/Gene Discovery



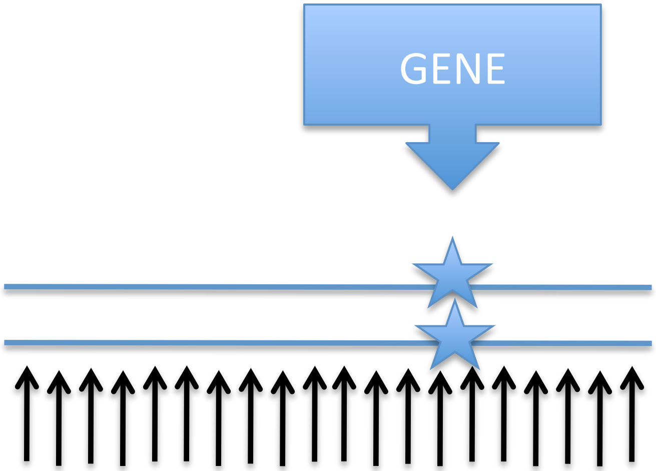
Sample

Training Population

Target Population

30 pairs of chromosomes

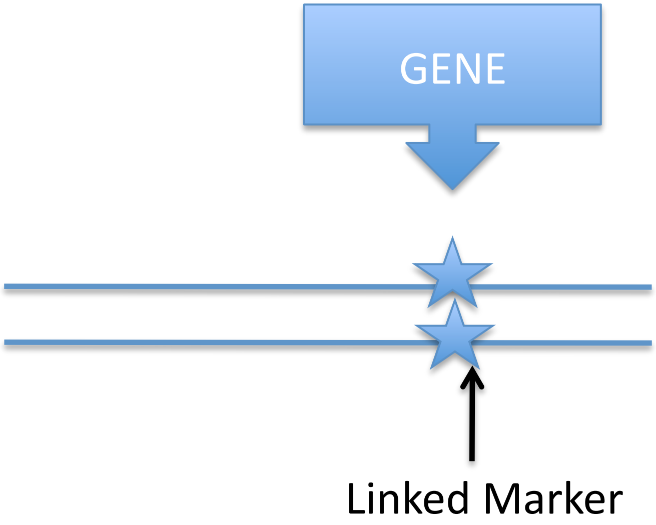
Goal in Marker/Gene Discovery



DNA markers (e.g. SNPs)
>1,000 per chromosome

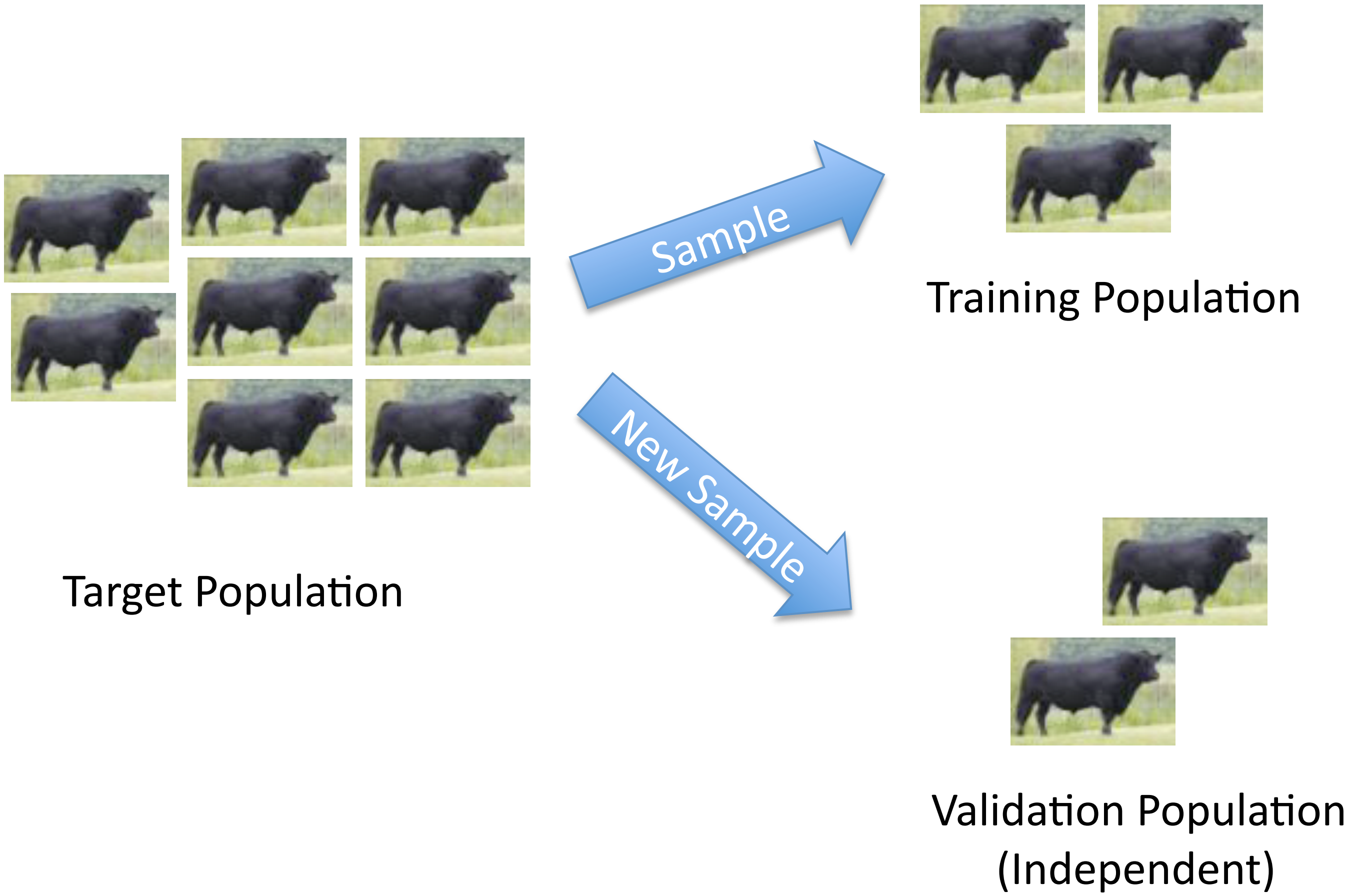
Goal in Marker/Gene Discovery

Research is looking for markers in tight linkage disequilibrium (LD) due to close physical proximity to causal mutations



Inheritance of a marker allele is indicative of inheritance of favorable allele in gene

Ideal Validation of Good Marker



Sample

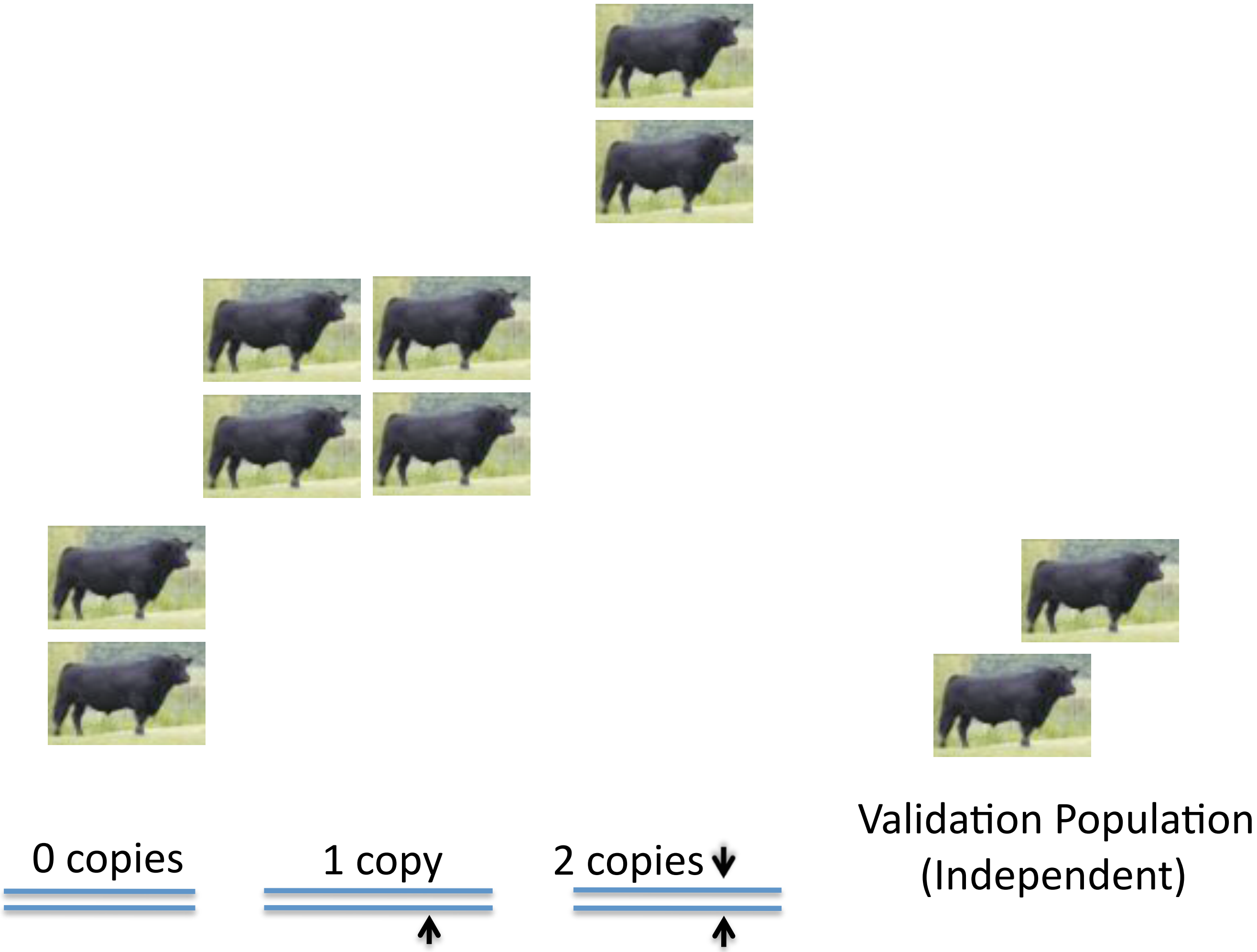
Training Population

New Sample

Target Population

Validation Population (Independent)

Ideal Validation of Good Marker



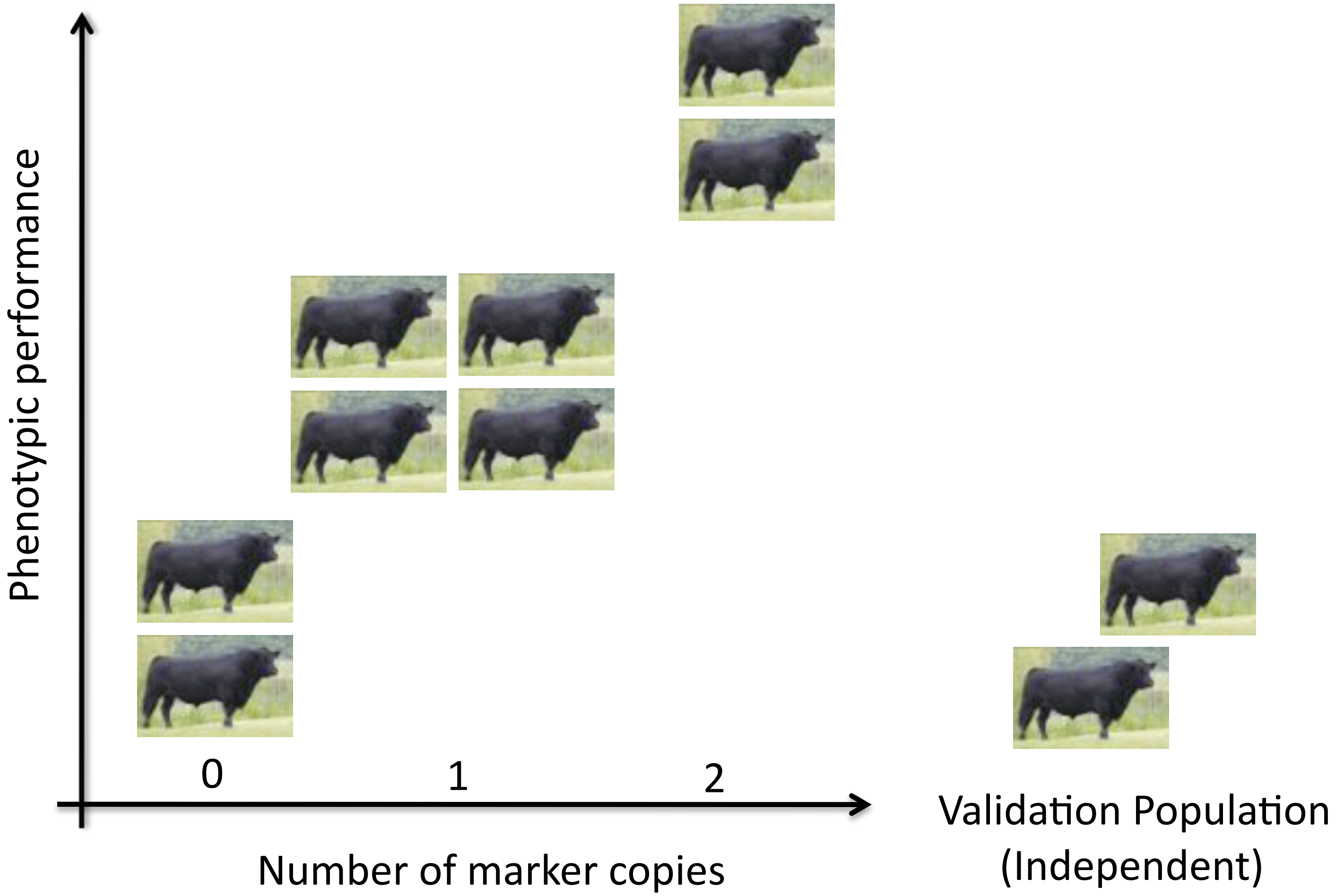
0 copies

1 copy

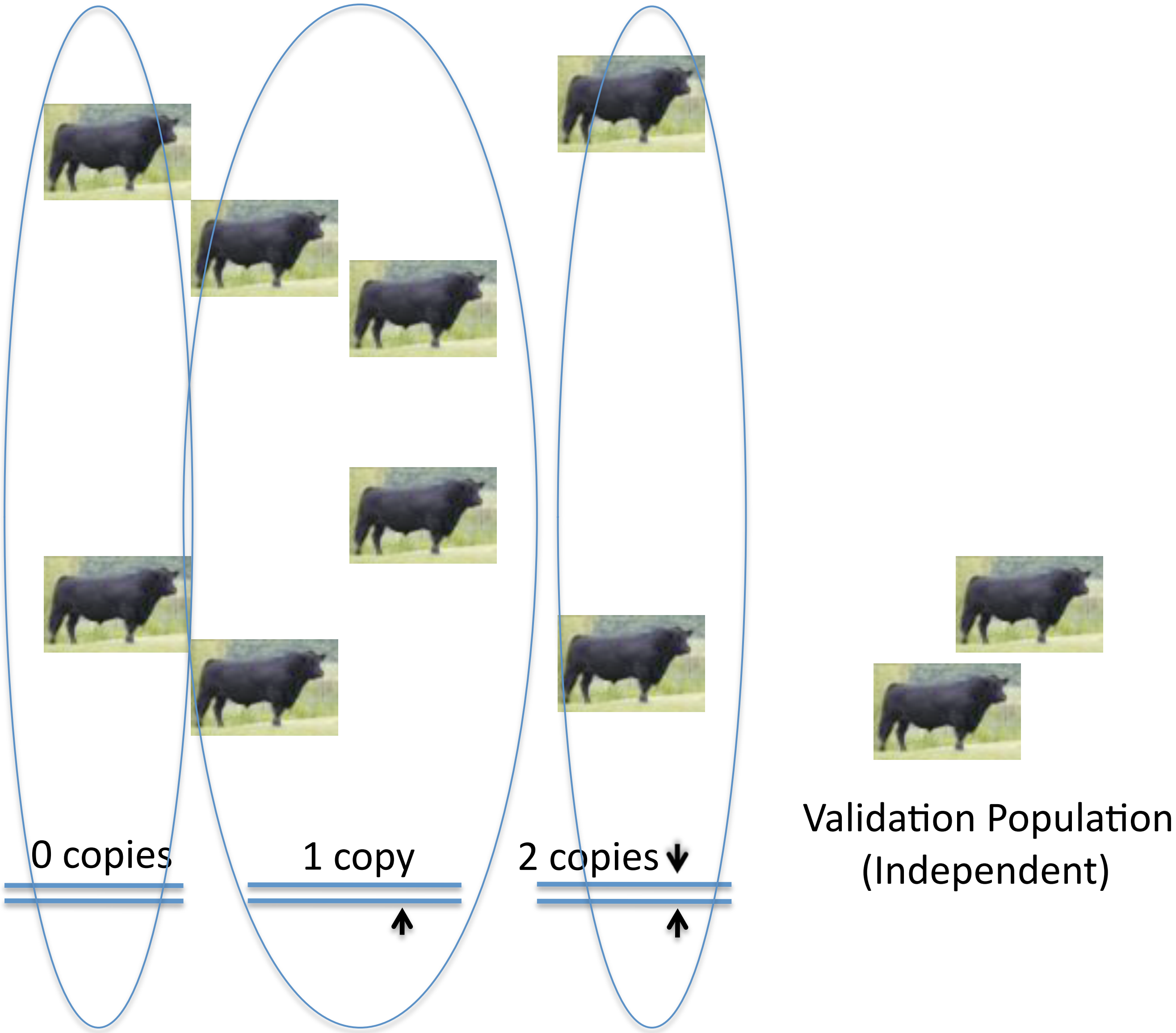
2 copies

Validation Population
(Independent)

Ideal Validation of Good Marker



Ideal Failed Validation of Bad Marker



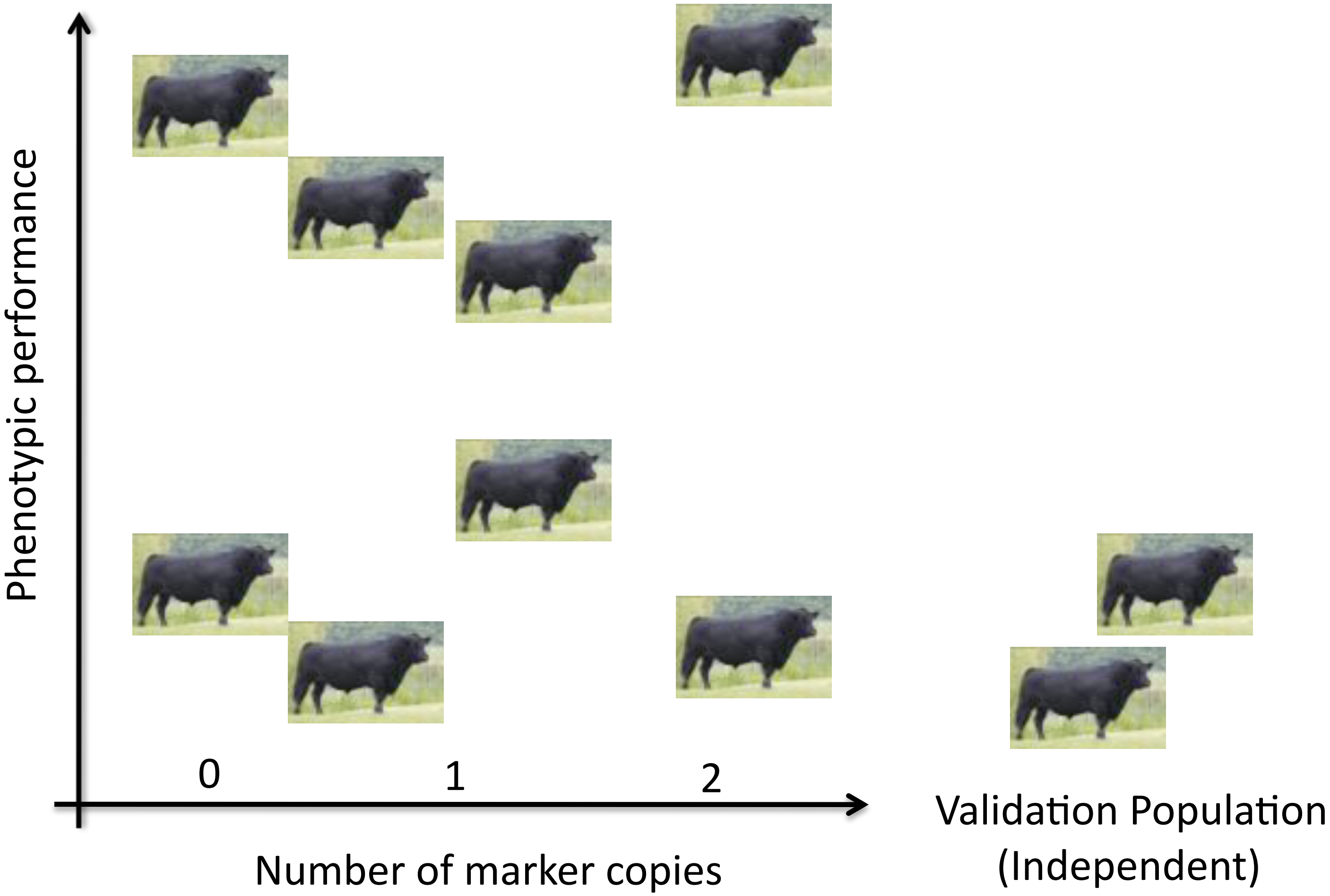
0 copies

1 copy

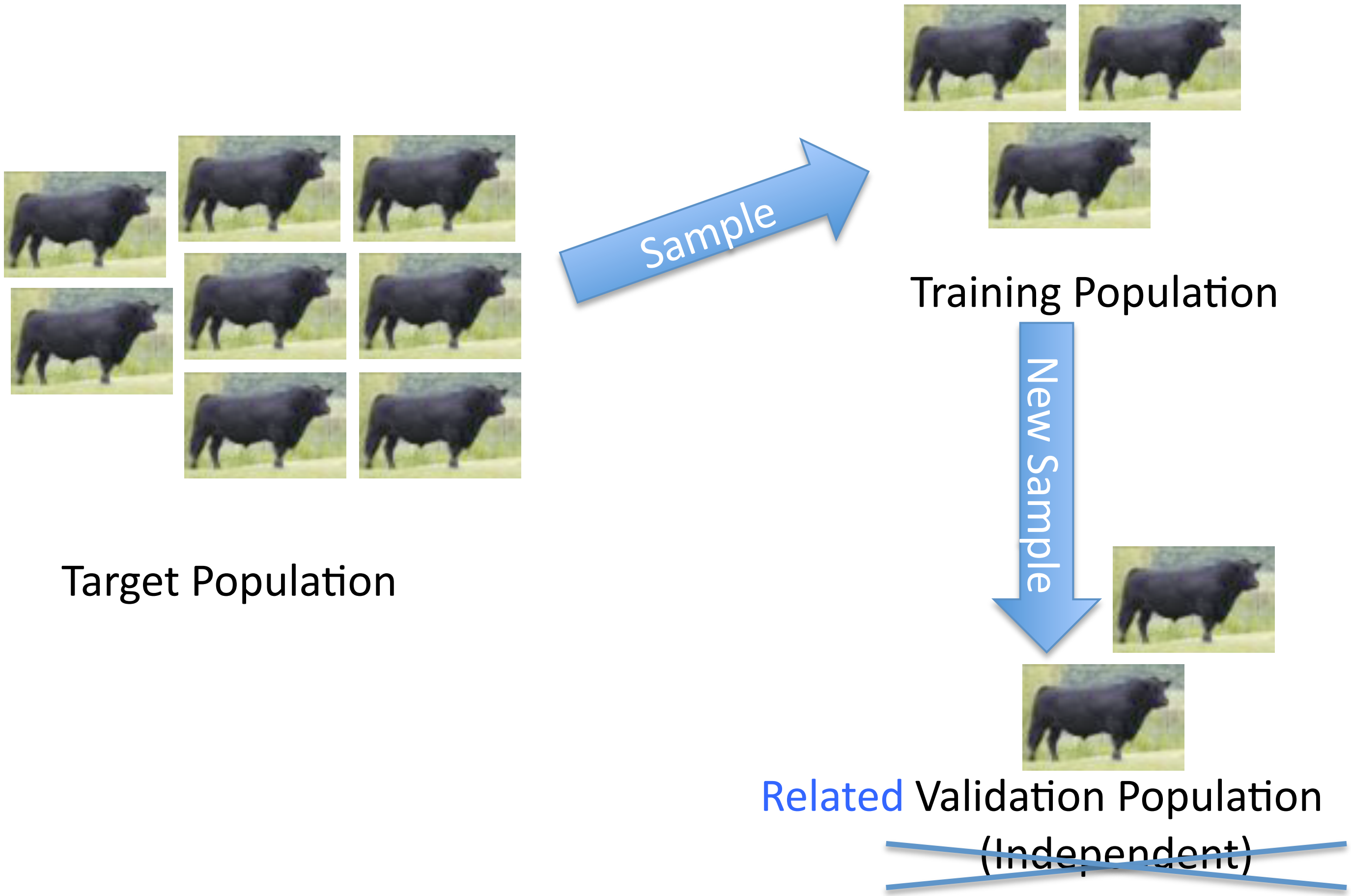
2 copies

Validation Population
(Independent)

Ideal Failed Validation of Bad Marker



Validation in Practice



Target Population

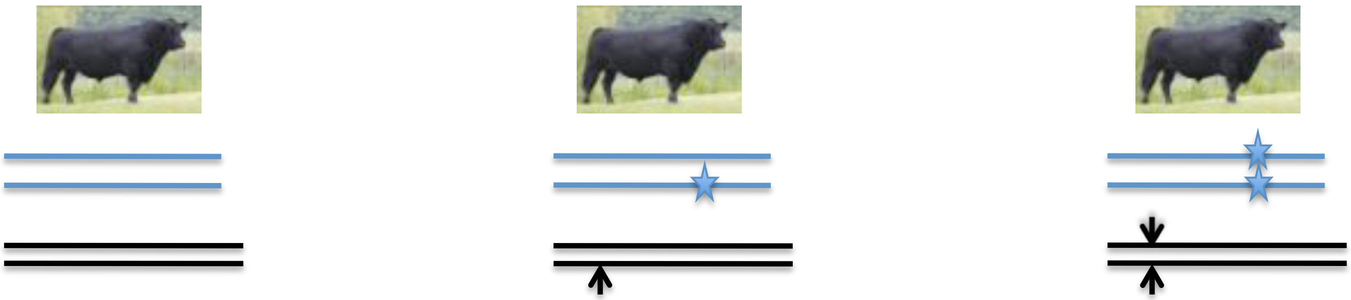
Sample

Training Population

New Sample

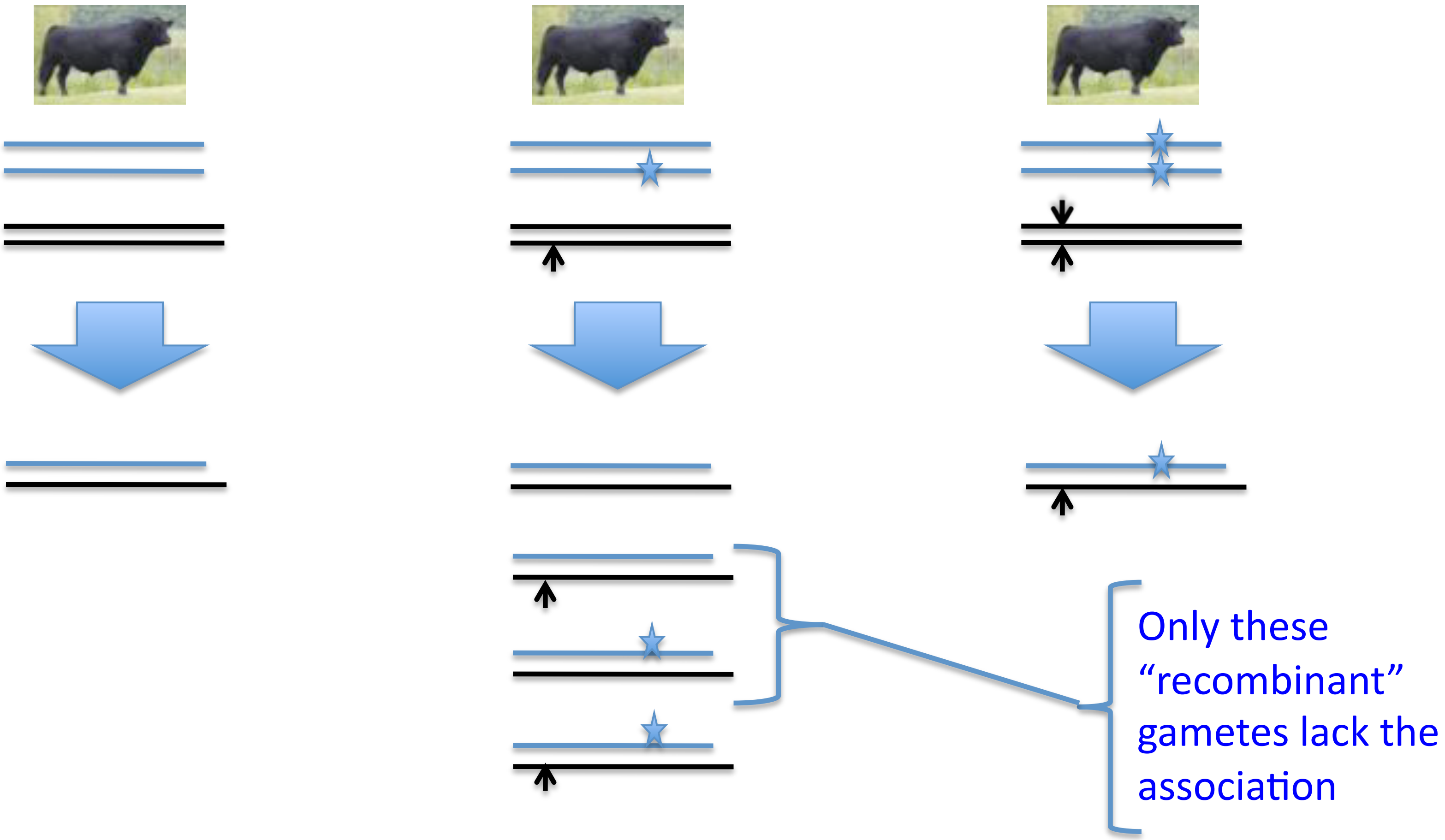
Related Validation Population ~~(Independent)~~

Problems with Related Validation and Discovery Populations



Totally spurious markers can be discovered in the training population especially when there are many more (e.g. 50k) markers to consider then there are training animals

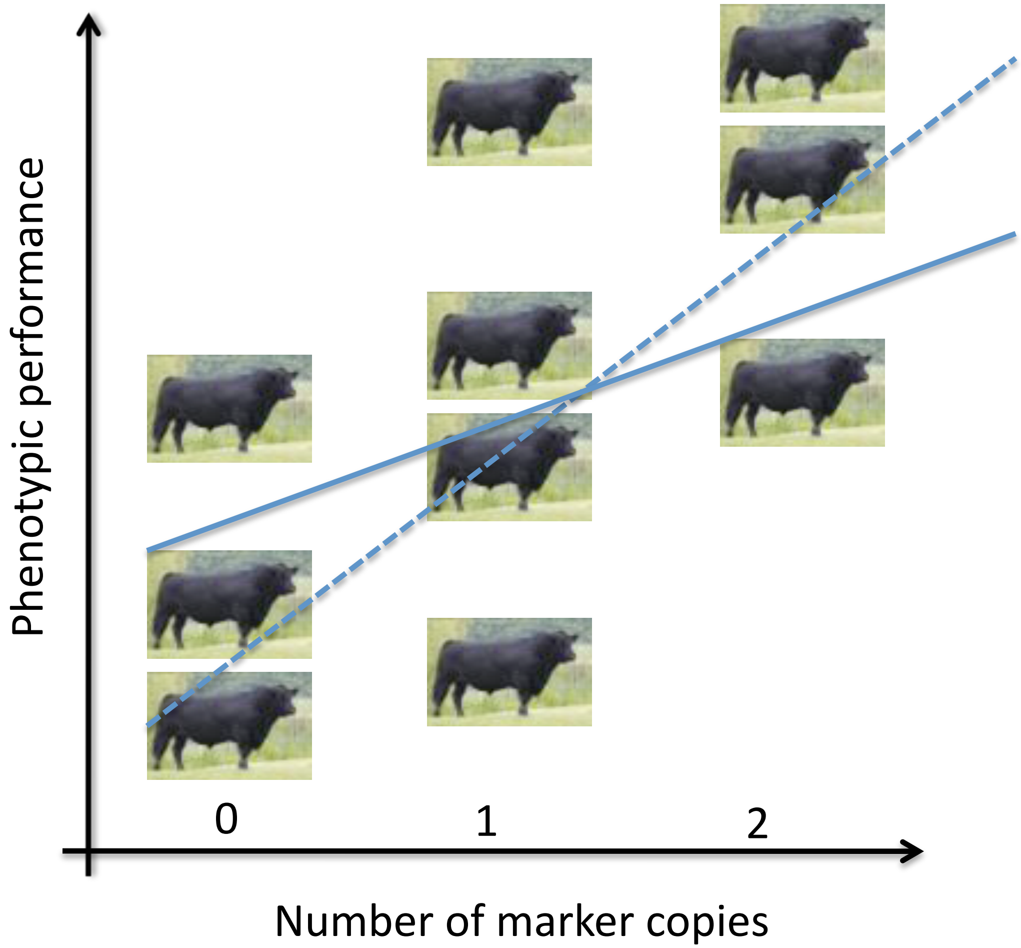
Problems with Related Validation and Discovery Populations



Only these "recombinant" gametes lack the association

Gametes from a parent in the discovery population show a marker effect

Problems Validating in Relatives



Regression in Discovery

Regression in Validation
(when offspring of Discovery)

Spurious markers validate
with HALF their discovery
effect, rather than NO effect

Validating in Relatives

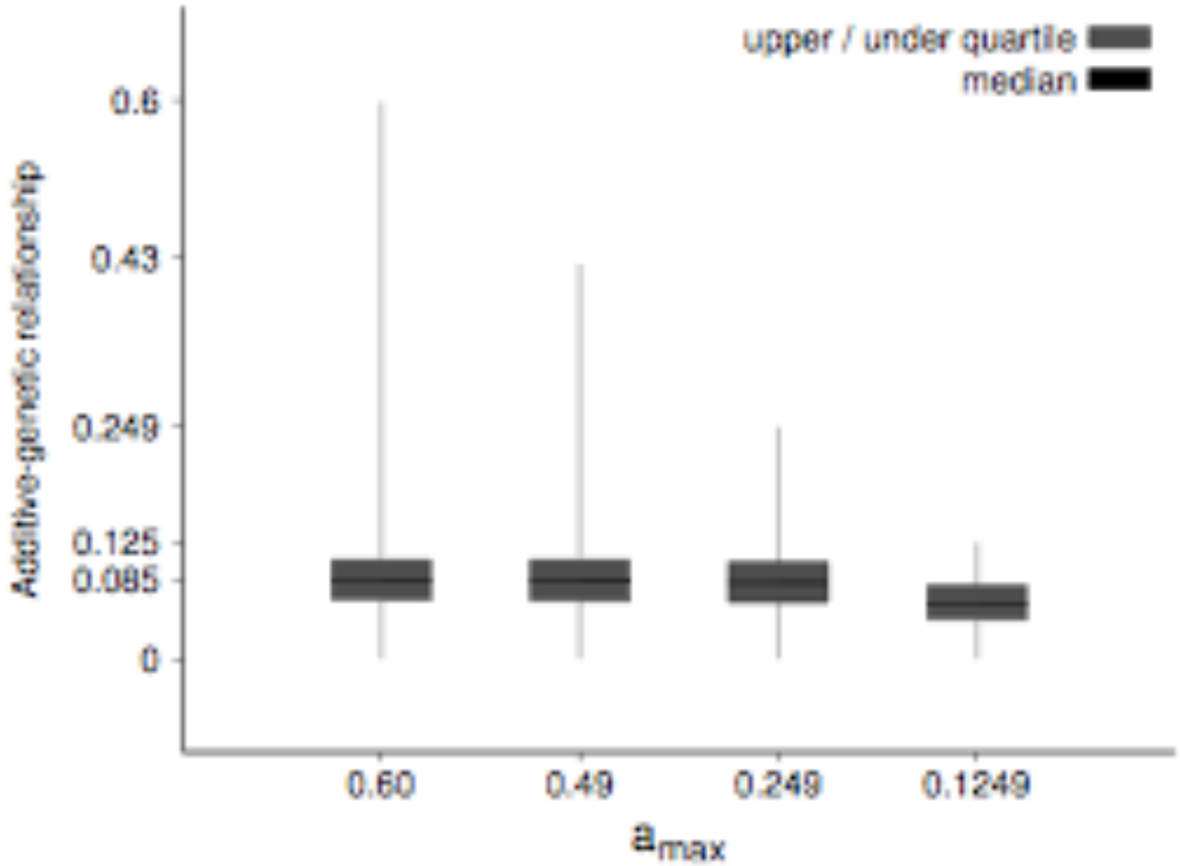
- The marker effect of
  - real associations will be retained
  - spurious associations will halve each generation if the marker and gene are not linked
- In general, the marker effect reduces by $(1\text{-}r_{QM})$ each generation
- Marker panels that comprise a mixture of real and spurious results, validated in relatives, will gradually erode over time
  - Validation will overestimate their real value

Practical Demonstration - Habier et al

amax is the maximum additive relationship between any bull in training and any bull in validation

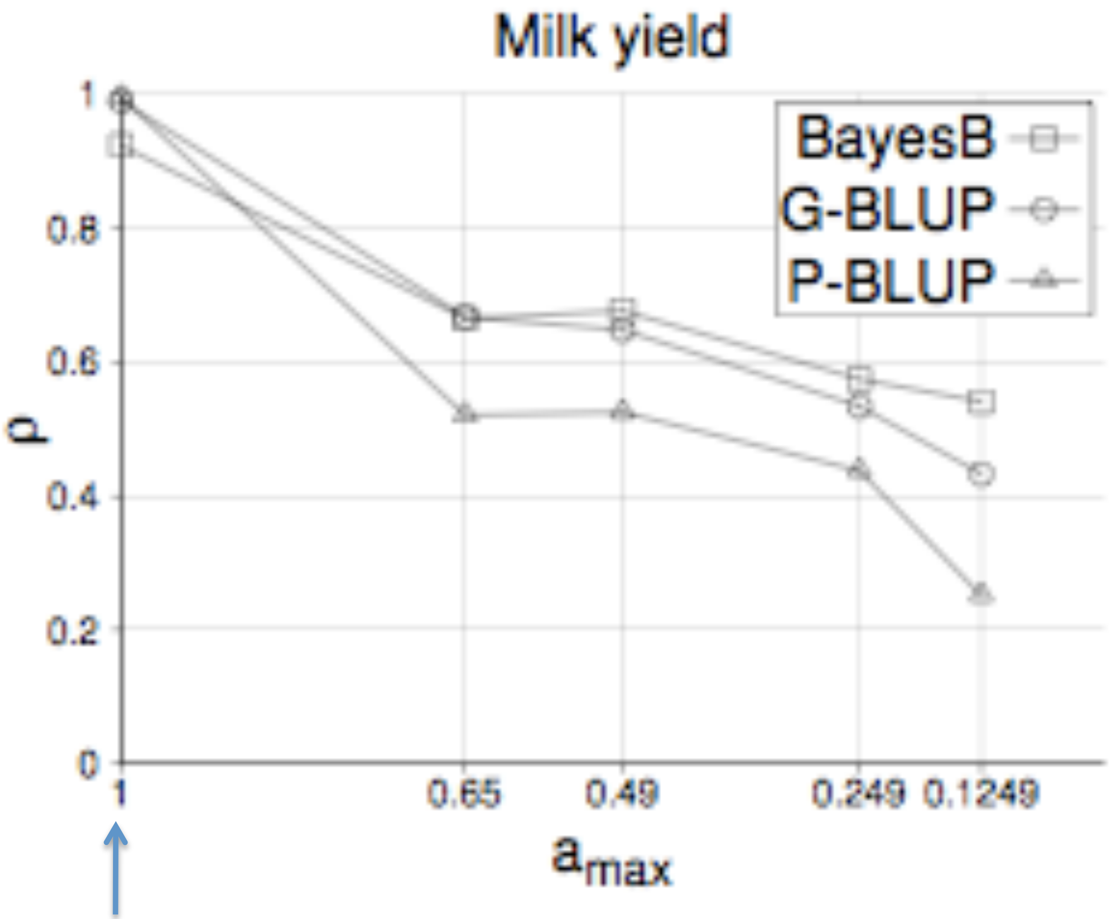
Scenarios:
amax of 0.6, 0.49, 0.249 and 0.1249
0.6: Fathers, full-and half sibs in training
0.49: Half sibs in training
<0.25: No half sibs

Additive genetic relationships between training and validation subsets



These represent four different partitionings of the data into training & validation

Accuracy of genomic EBVs vs amax



r=0.7 50% variance

r=0.5 25% variance

r in training data

2084 training bulls

Conclusions

- Presence of parent-offspring links, or of half-sibs represented in both the training and validation data leads to genomic predictions that appear to account for 2x as much variance compared to using less related animals in validation
- Discovery populations that use all AI bulls in a breed will make it very difficult to form a reliable validation dataset
- Validation results will overstate the real value of genomic tests