Lecture 7
Genomic Selection

Guilherme J. M. Rosa

University of Wisconsin-Madison

Mixed Models in Quantitative Genetics
SISG, Seattle
22 - 24 July 2020

Marker Assisted Selection

MAS: Use of genetic markers to improve the efficiency of genetic selection

Basic idea behind of MAS:

- Most traits of economic importance are controlled by a fairly large number of genes
- Some of these genes, however, with larger effect
- Following the pattern of inheritance of such genes might assist in selection

MAS Could Help Improve

Low heritability traits

Phenotypes that can be measured on one sex only

Characteristics that are not measurable before sexual maturity

Traits that are difficult to measured or require sacrifice

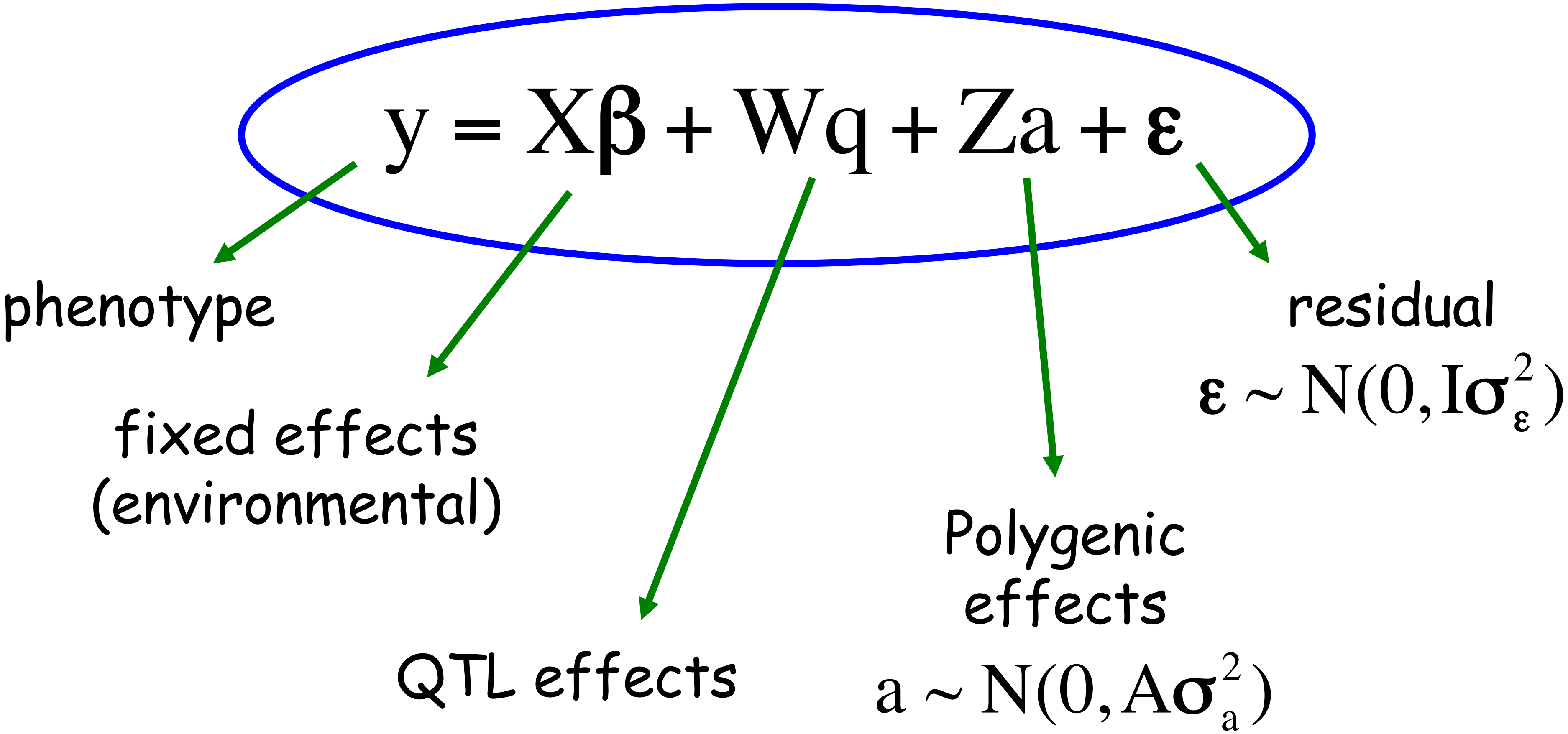
Efficiency of MAS

Size (effect) of QTL

Frequency of favorable allele

Recombination rate between marker(s) and QTL

Modeling Effects at The QTL Genotype



$$y = X\beta + Wq + Za + \varepsilon$$

phenotype

fixed effects
(environmental)

QTL effects

Polygenic
effects
$a \sim N(0, A\sigma_a^2)$

residual
$\varepsilon \sim N(0, I\sigma_\varepsilon^2)$

Modeling Effects at the QTL Genotype

QTL-genotype as a fixed effect: Regression of phenotypes using QTL genotype probabilities from segregation analysis (Kinghorn et al. 1993, Meuwissen and Goddard 1997)

QTL-genotype as a random effect: QTL effect is modeled as the sum of the two gametic effects (Fernando and Grossman 1989)

$$y = X\beta + Wv + Za + \varepsilon \ , \qquad \mathrm{Var}\begin{pmatrix} v \\ a \\ \varepsilon \end{pmatrix} = \begin{pmatrix} G_v\sigma_v^2 & 0 & 0 \\ 0 & A\sigma_a^2 & 0 \\ 0 & 0 & I\sigma_\varepsilon^2 \end{pmatrix}$$

Gametic relationship matrix

PAUSE

- Marker Assisted Selection
- Different modelling approaches, but always with just a few markers included

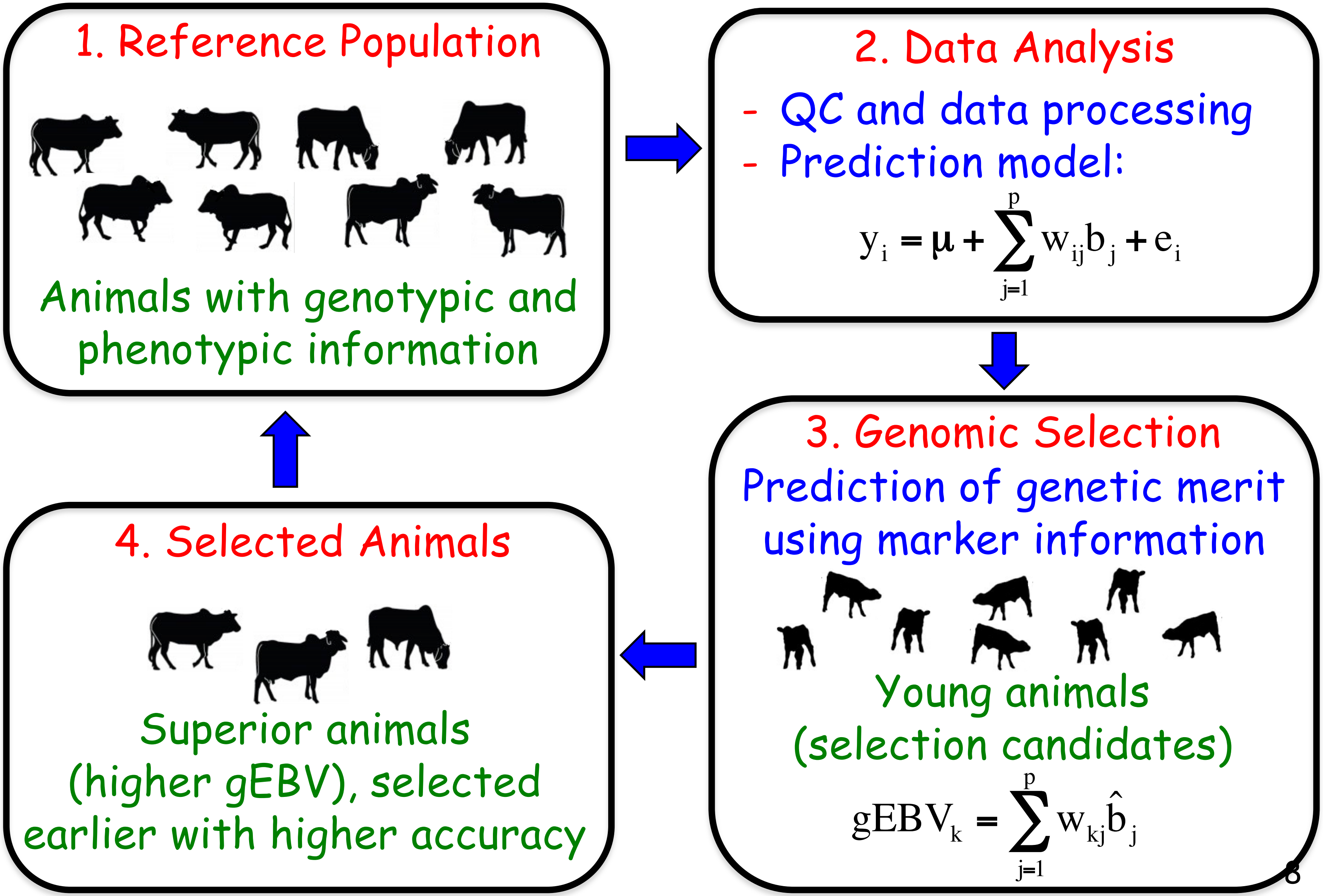
Next PAUSE, slide 21

Genomic Selection
(Genome-wide Marker Assisted Selection)

As most quantitative traits are influenced by many genes, tracking a small number of them using molecular markers will explain only a small fraction of the total genetic variance

GWMAS, on the other hand, makes use of a very dense set of markers covering the entire genome, which potentially explain all genetic variance

Genomic Selection



1. Reference Population

Animals with genotypic and phenotypic information

2. Data Analysis

- QC and data processing
- Prediction model:

$$y_i = \mu + \sum_{j=1}^{p} w_{ij} b_j + e_i$$

3. Genomic Selection

Prediction of genetic merit using marker information

Young animals (selection candidates)

$$gEBV_k = \sum_{j=1}^{p} w_{kj} \hat{b}_j$$

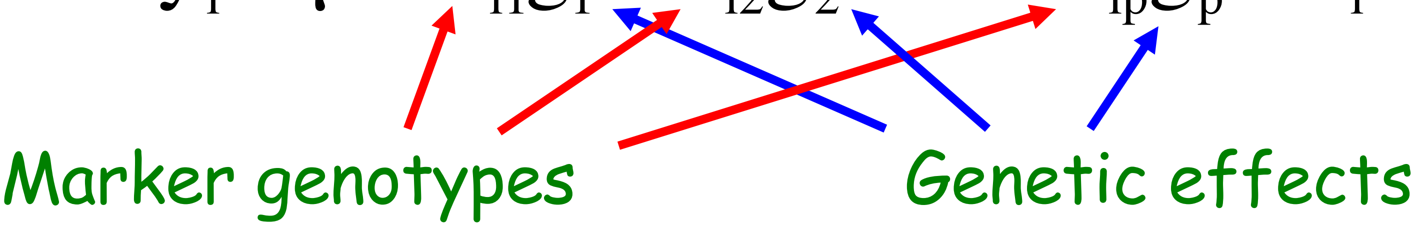
4. Selected Animals

Superior animals (higher gEBV), selected earlier with higher accuracy

Genomic Selection

(Meuwissen et al., 2001)

$$y_i = \mu + x_{i1}g_1 + x_{i2}g_2 + \ldots + x_{ip}g_p + e_i$$



Marker genotypes

Genetic effects

Genomic EBV: $\mathrm{GEBV} = x_{i1}\hat{g}_1 + x_{i2}\hat{g}_2 + \ldots + x_{ip}\hat{g}_p = \sum_{j=1}^{p} x_{ij}\hat{g}_j$

- ‘big p small n paradigm’
- Dimension reduction techniques (e.g. SVD and PLS), and stepwise strategies
- Alternatively, ridge regression, random effects models, and hierarchical modeling

Least Squares

Two-step Procedure:

- Test each marker (chromosome segment) for presence of QTL and select those with significant effects
- Fit selected markers simultaneously using multiple regression
- Predict breeding values using fitted regression (similar to LD- MAS approach with multiple markers)

Problems:

- Over estimation of markers effects due to first-step (selection)
- Do not capture all QTL

BLUP

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j g_j + \mathbf{e} \qquad \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{X} \\ \mathbf{X}'\mathbf{1} & \mathbf{X}'\mathbf{X} + \mathbf{I}\gamma \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}$$

$$g_j \sim N(0, \sigma_0^2) \qquad \gamma = \sigma_e^2 / \sigma_0^2$$

How to choose σ_0^2 ?

- Arbitrary; but σ_0^2 controls amount of shrinkage
- Alternative: set $\sigma_0^2 = \sigma_u^2 / p$, where σ_u^2 is an estimate (prior) of total additive genetic variance

Bayes A

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j + \mathbf{e} \quad \longrightarrow \quad \mathbf{y} \mid \mu, \mathbf{g}_j, \sigma_e^2 \sim N(\mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j, \mathbf{I}\sigma_e^2)$$

Prior distributions:

$$\begin{cases} g_j \mid \sigma_j^2 \sim N(0, \sigma_j^2) \\ \sigma_j^2 \sim \chi^{-2}(\nu, S) \\ \quad \text{(scaled inverted chi-square distribution with} \\ \quad \text{scale parameter S and } \nu \text{ degrees of freedom)} \\ \sigma_e^2 \sim \chi^{-2}(-2, 0) \end{cases}$$

Bayes B

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j + \mathbf{e} \quad \longrightarrow \quad \mathbf{y} \mid \mu, \mathbf{g}_j, \sigma_e^2 \sim N(\mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j, \mathbf{I}\sigma_e^2)$$

Prior distributions:

$$\begin{cases} \left[\begin{array}{ll} g_j = 0 & \text{with probability } \pi \\ g_j \mid \sigma_j^2 \sim N(0, \sigma_j^2) & \text{with probability } (1 - \pi) \end{array} \right. \\ \sigma_j^2 \sim \chi^{-2}(\nu, S) \\ \sigma_e^2 \sim \chi^{-2}(-2, 0) \end{cases}$$

Simulation Study

Genome: 1000 cM with markers every 1 cM

Markers surrounding each 1 cM region combined into haplotypes

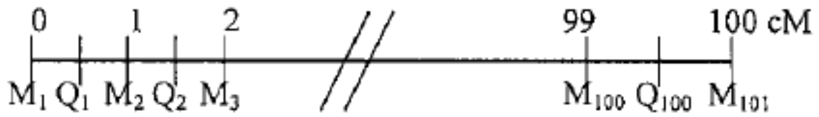
LD between marker and QTLs due to finite population size (N_e = 100)

Training sample: single generation with 2,000 animals

Test sample: prediction of breeding values of their progeny based on marker genotypes

Simulation Study

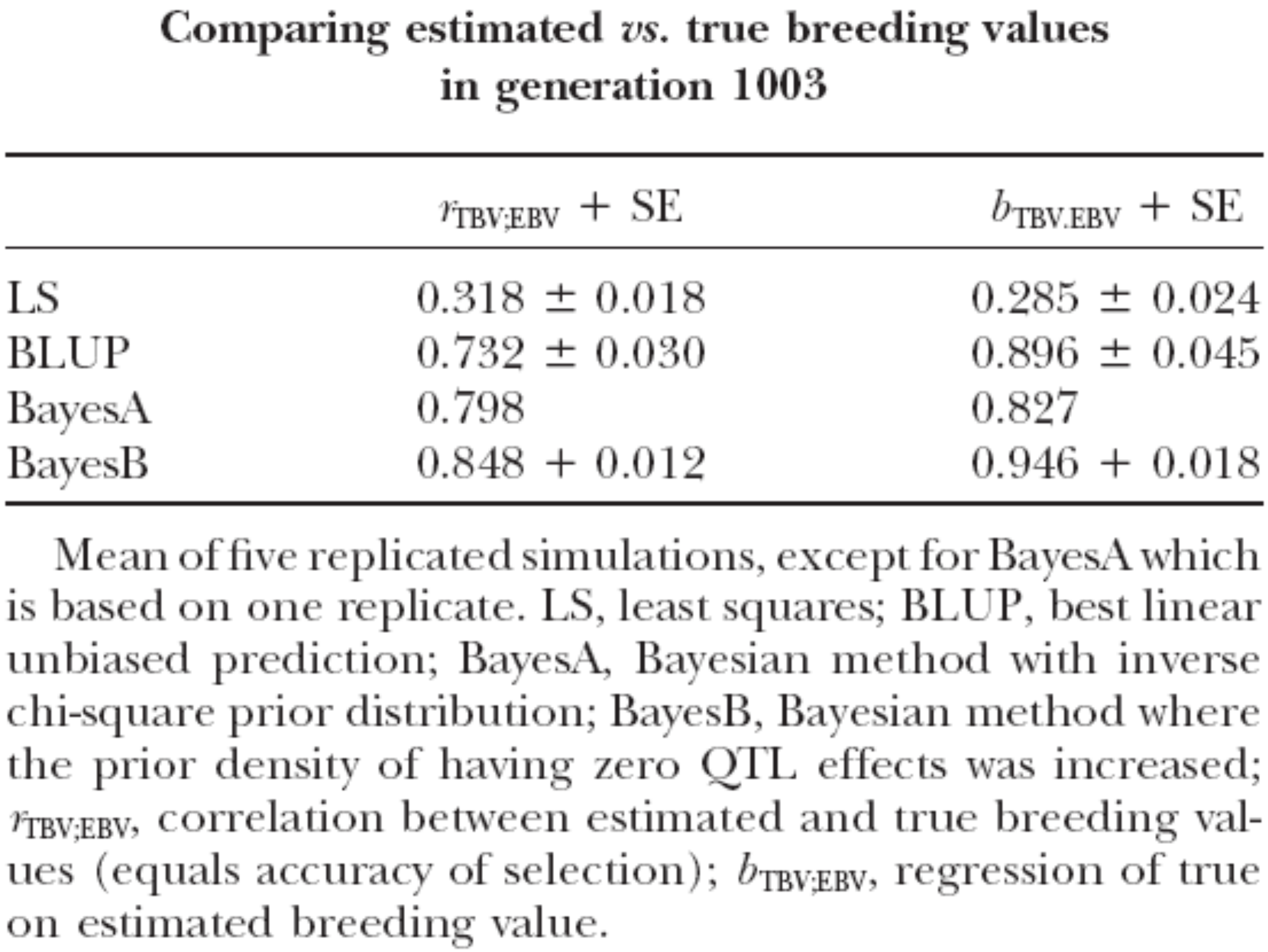
The parameters of the simulated genetic model

Map per chromosome[a]	 0 1 2 99 100 cM M_1 Q_1 M_2 Q_2 M_3 M_{100} Q_{100} M_{101}
Number of chromosomes is the total number of morgans	10
Mutation rate of QTL	2.5×10^{-5}
Distribution of additive mutational effects	Gamma(1.66; 0.4)
Dominance of QTL effects	0
Mutation rate of marker loci	2.5×10^{-3}
Population structure	
Generations 1–1000	Ideal[b], $N = 100$
Generation 1001	Ideal[b], $N = 200$
Generation 1002	20 half-sib families, $N = 2000$
Generation 1003 and later	Ideal[b], $N = 2000$
Marker genotyping	Generations 1001 and later
Phenotypic recording	Generations 1001 and 1002

[a] M, marker position; Q, QTL position.

[b] Ideal denotes a population structure where the effective size equals the actual population size. This structure is simulated by giving every male (female) in generation $t - 1$ an equal probability of becoming the sire (dam) of animal i in generation t, which implies no selection and random mating of males and females.

Simulation Study

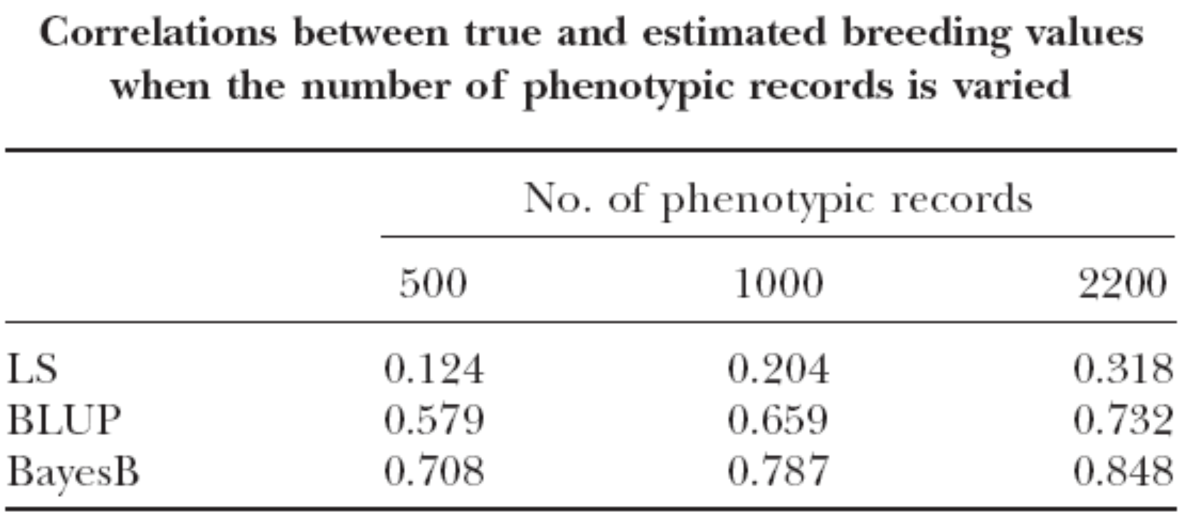


Comparing estimated *vs.* true breeding values in generation 1003

	$r_{\mathrm{TBV;EBV}}$ + 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

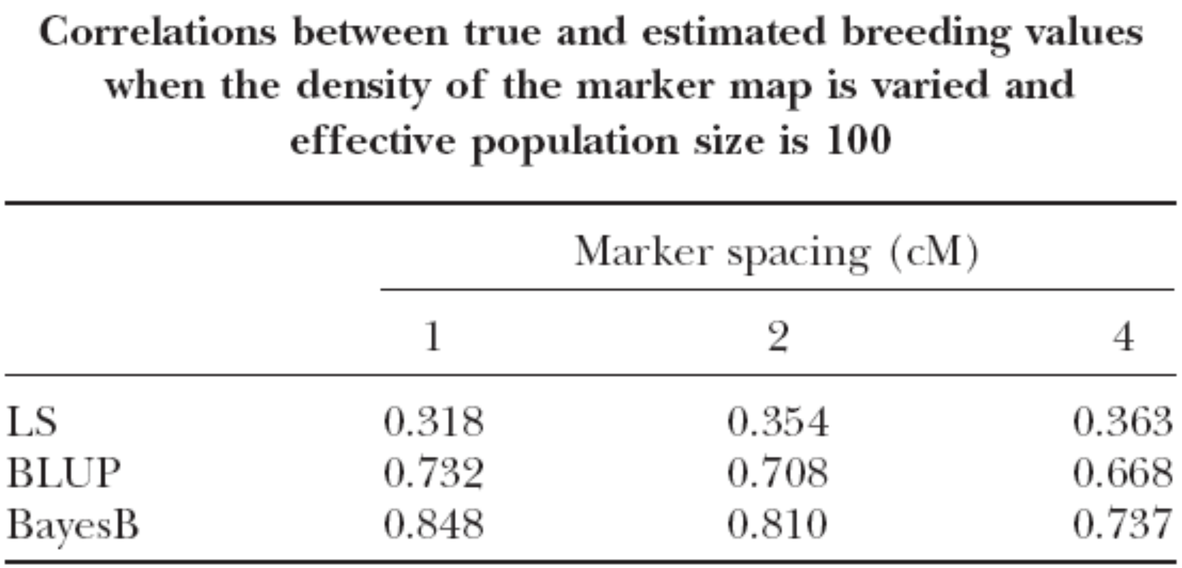
Mean of five replicated simulations, except for BayesA which is based on one replicate. LS, least squares; BLUP, best linear unbiased prediction; BayesA, Bayesian method with inverse chi-square prior distribution; BayesB, Bayesian method where the prior density of having zero QTL effects was increased; $r_{\mathrm{TBV;EBV}}$, correlation between estimated and true breeding values (equals accuracy of selection); $b_{\mathrm{TBV;EBV}}$, regression of true on estimated breeding value.

Simulation Study



Correlations between true and estimated breeding values when the number of phenotypic records is varied

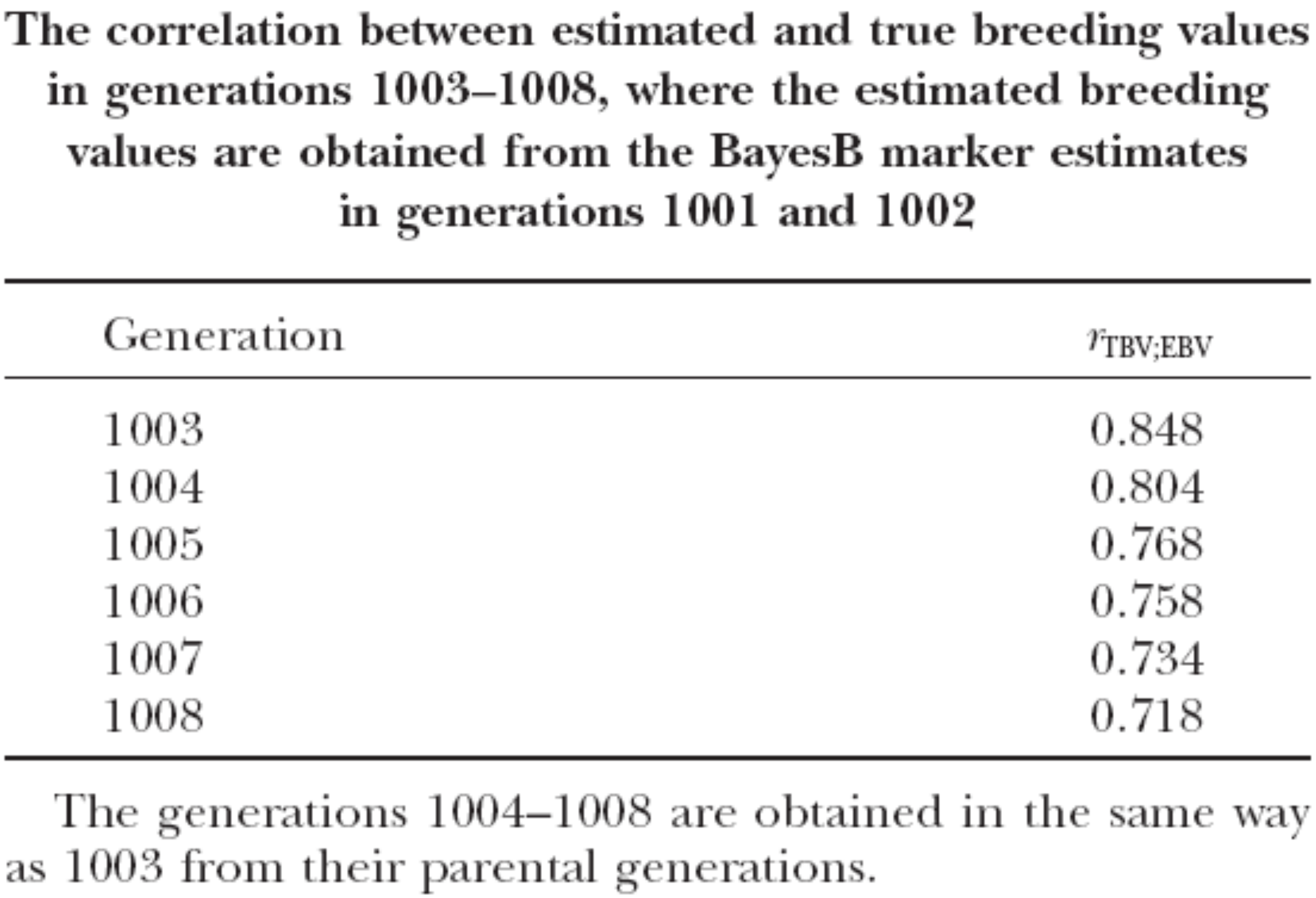
	No. of phenotypic records		
	500	1000	2200
LS	0.124	0.204	0.318
BLUP	0.579	0.659	0.732
BayesB	0.708	0.787	0.848



Correlations between true and estimated breeding values when the density of the marker map is varied and effective population size is 100

	Marker spacing (cM)		
	1	2	4
LS	0.318	0.354	0.363
BLUP	0.732	0.708	0.668
BayesB	0.848	0.810	0.737

Simulation Study

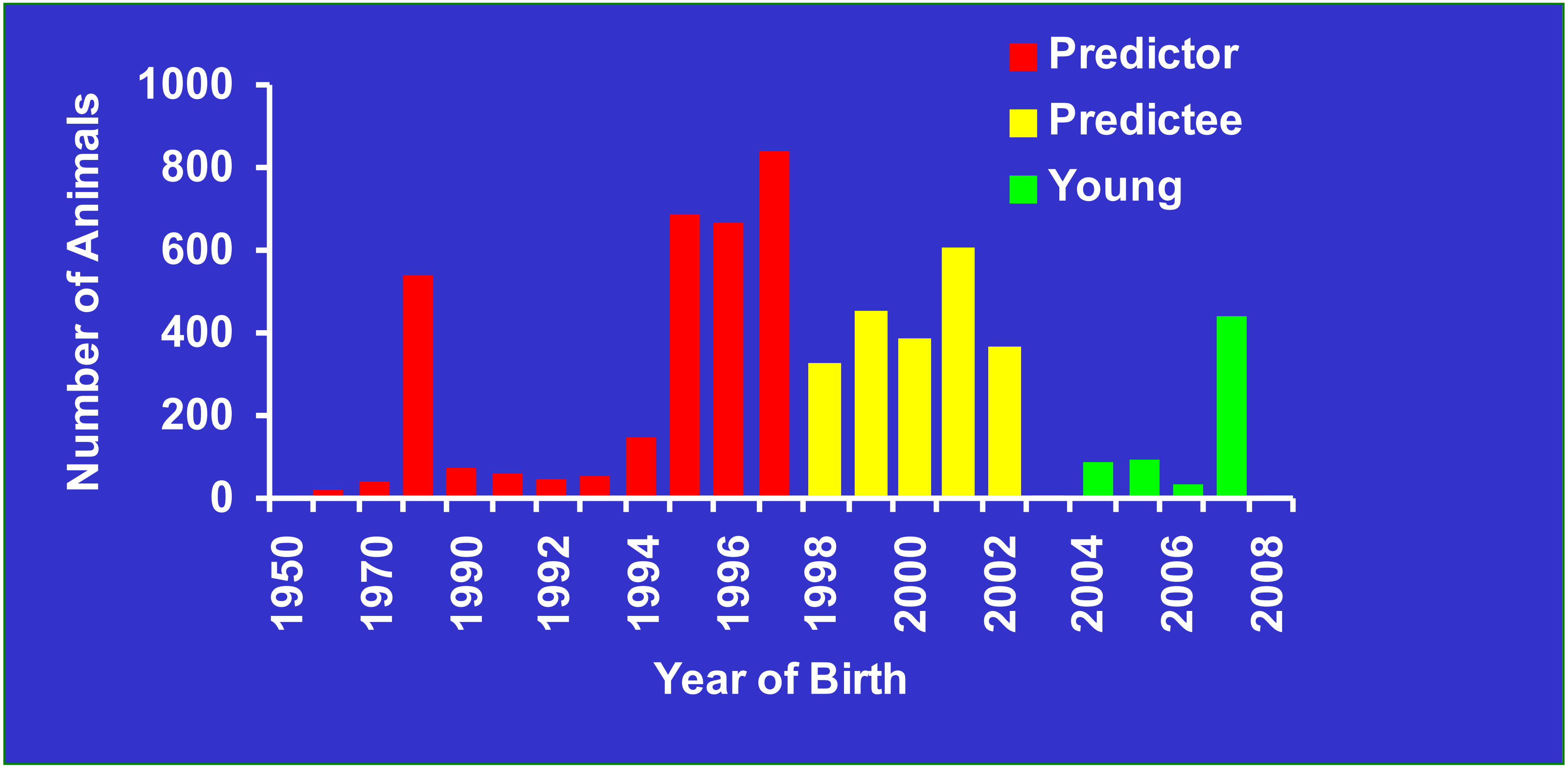


The correlation between estimated and true breeding values in generations 1003–1008, where the estimated breeding values are obtained from the BayesB marker estimates in generations 1001 and 1002

Generation	$r_{\mathrm{TBV;EBV}}$
1003	0.848
1004	0.804
1005	0.768
1006	0.758
1007	0.734
1008	0.718

The generations 1004–1008 are obtained in the same way as 1003 from their parental generations.

Application with Real Data



(VanRaden et al., 2008)

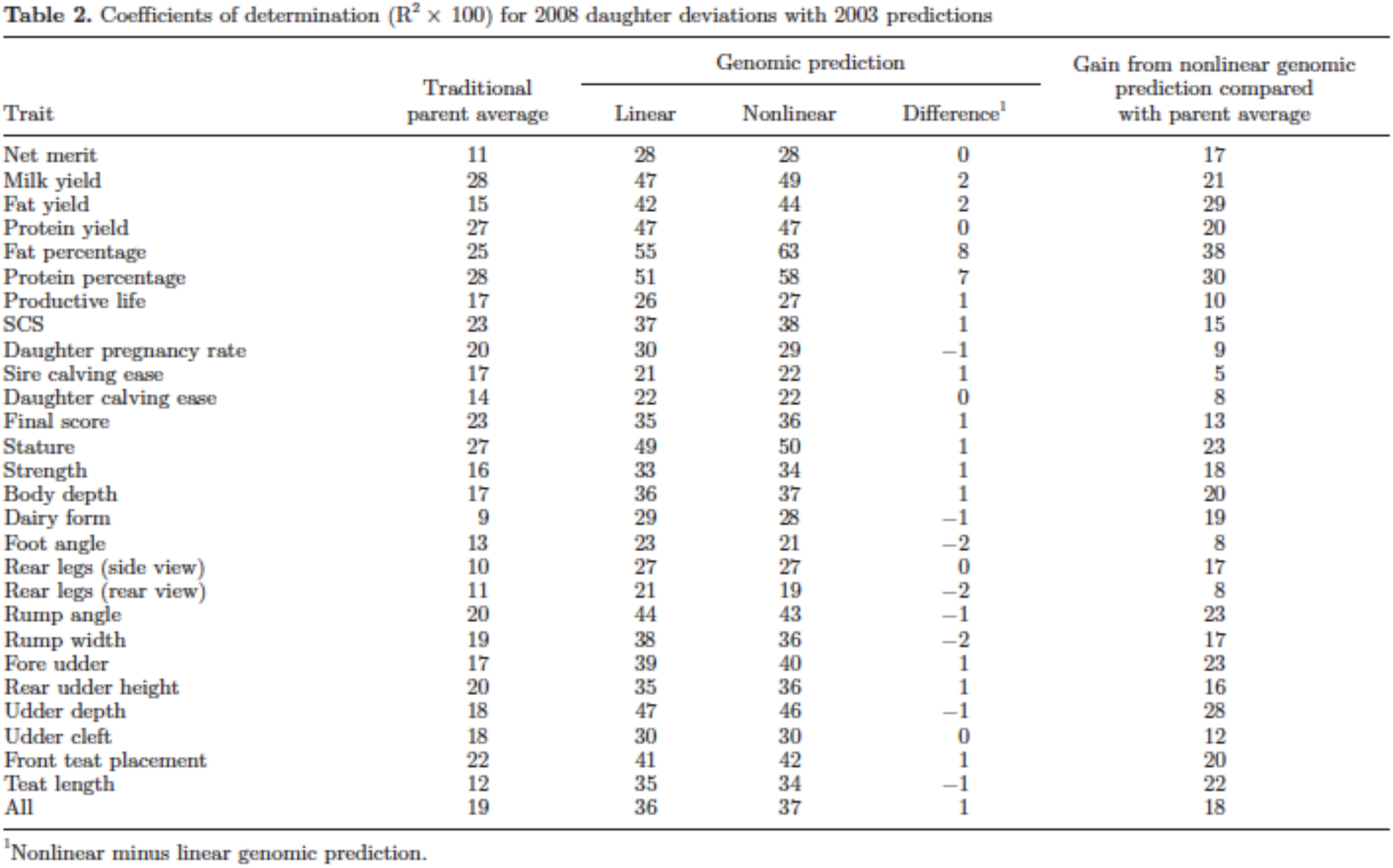


Table 2. Coefficients of determination ($R^2 \times 100$) for 2008 daughter deviations with 2003 predictions

Trait	Traditional parent average	Genomic prediction			Gain from nonlinear genomic prediction compared with parent average
		Linear	Nonlinear	Difference[1]	
Net merit	11	28	28	0	17
Milk yield	28	47	49	2	21
Fat yield	15	42	44	2	29
Protein yield	27	47	47	0	20
Fat percentage	25	55	63	8	38
Protein percentage	28	51	58	7	30
Productive life	17	26	27	1	10
SCS	23	37	38	1	15
Daughter pregnancy rate	20	30	29	−1	9
Sire calving ease	17	21	22	1	5
Daughter calving ease	14	22	22	0	8
Final score	23	35	36	1	13
Stature	27	49	50	1	23
Strength	16	33	34	1	18
Body depth	17	36	37	1	20
Dairy form	9	29	28	−1	19
Foot angle	13	23	21	−2	8
Rear legs (side view)	10	27	27	0	17
Rear legs (rear view)	11	21	19	−2	8
Rump angle	20	44	43	−1	23
Rump width	19	38	36	−2	17
Fore udder	17	39	40	1	23
Rear udder height	20	35	36	1	16
Udder depth	18	47	46	−1	28
Udder cleft	18	30	30	0	12
Front teat placement	22	41	42	1	20
Teat length	12	35	34	−1	22
All	19	36	37	1	18

[1] Nonlinear minus linear genomic prediction.

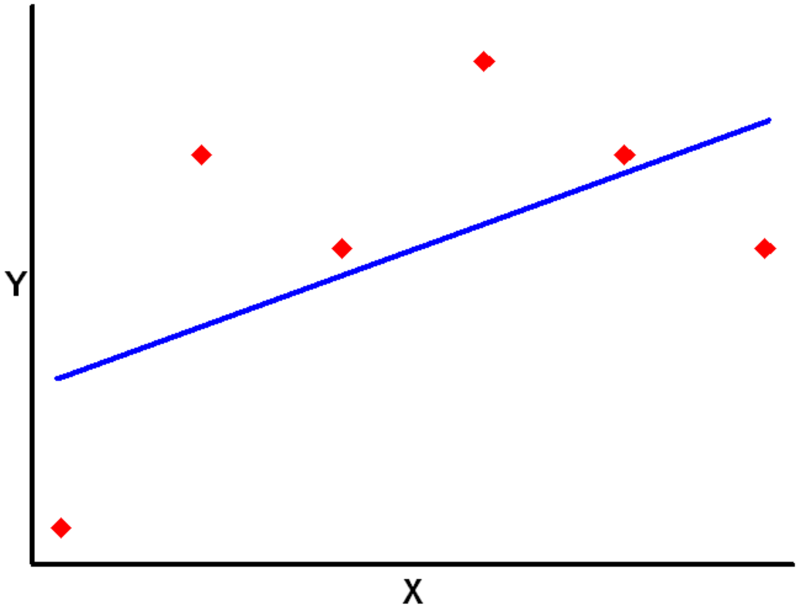
PAUSE

- ⇨ Principles of Genomic Selection
- ⇨ Some additional details on statistical modelling next

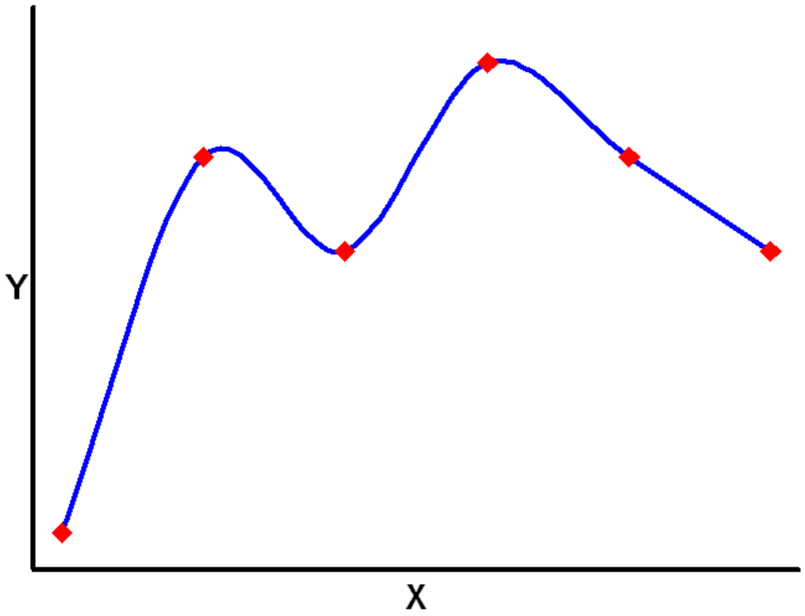
Next PAUSE, slide 30

Model Selection

- ⇨ Goodness-of-fit vs. Model Complexity (Bias-variance tradeoff)



Over-reduction



Over-fit

Model Selection

⇨ Goodness-of-fit

- likelihood ratio approach (LRT; nested models)

$$\mathrm{LRT} = -2\ln\left(\frac{\mathrm{L}_1}{\mathrm{L}_2}\right) \sim \chi^2_{(\mathrm{p}_1 - \mathrm{p}_2)}$$

⇨ Model complexity

- number of free parameters, p (effective number)

Linear (regularized) fitting: $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} \longrightarrow \mathrm{p} = \mathrm{trace}(\mathbf{S})$

Model Selection

⇨ Balancing goodness-of-fit and complexity

- Akaike information criterion (AIC):

$$\mathrm{AIC} = 2\mathrm{p} - \ln(\mathrm{L})$$

- Bayesian information criterion (BIC):
  (or Schwarz Criterion)

$$\mathrm{BIC} = \mathrm{p}\ln(\mathrm{n}) - 2\ln(\mathrm{L})$$

☞ If $\mathrm{e}_\mathrm{i} \overset{\mathrm{iid}}{\sim} \mathrm{N}(0, \sigma_\mathrm{e}^2)$ then:

$$\mathrm{AIC} = 2\mathrm{p} + \mathrm{n}\ln\left(\frac{\mathrm{RSS}}{\mathrm{n}}\right) \quad \text{and} \quad \mathrm{BIC} = \frac{1}{\sigma_\mathrm{e}^2}\mathrm{RSS} + \mathrm{p}\ln(\mathrm{L})$$

Ridge Regression

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

λ ≥ 0 (complexity parameter)

$$\text{or, equivalently:} \quad \hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

$$\text{subject to:} \sum_{j=1}^{p} \beta_j^2 \le s$$

Ridge Regression

$$\begin{cases} \hat{\beta}_0 = \bar{y} = \sum y_i / N \\ \text{after centering } y_i \text{ and } x_i\text{'s (i.e., } y_i - \bar{y} \text{ and } x_i - \bar{x}) \end{cases}$$

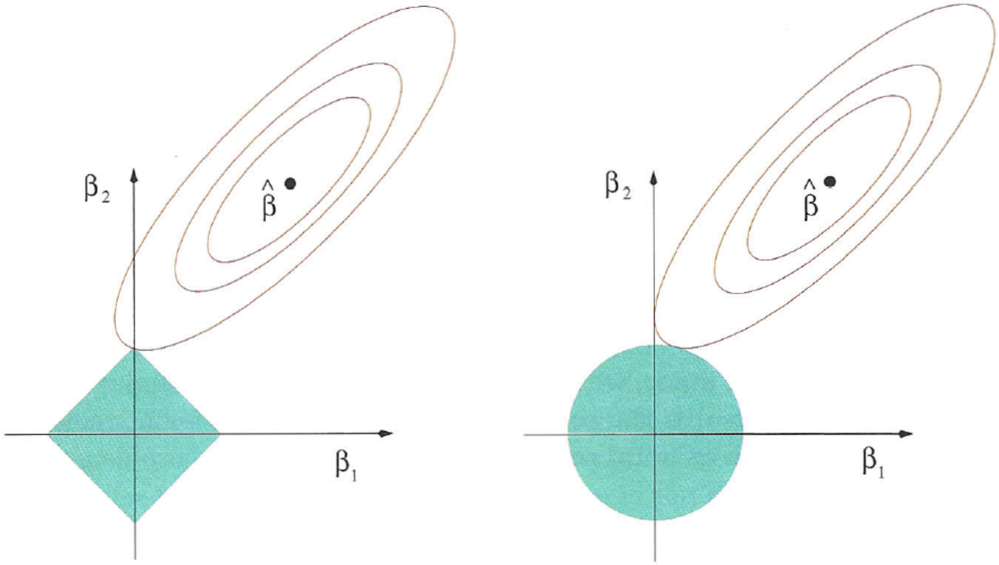
$$\mathrm{RSS}(\boldsymbol{\lambda}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\lambda}\boldsymbol{\beta}'\boldsymbol{\beta}$$

$$\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} = (\mathbf{X}'\mathbf{X} + \boldsymbol{\lambda}\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

LASSO

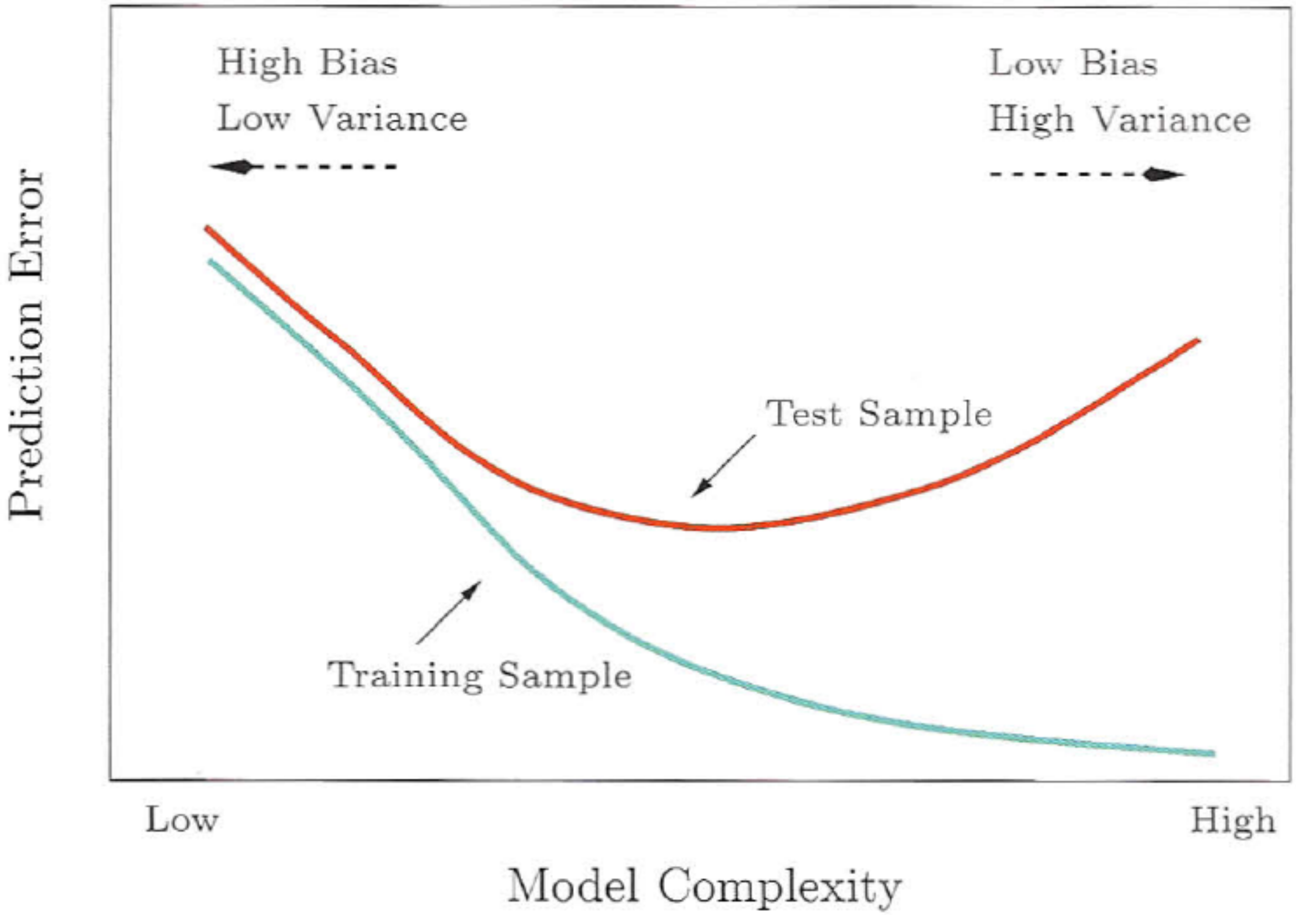
$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \text{ subject to}: \sum_{j=1}^{p} |\beta_j| \le t$$

- Estimation picture for the LASSO (left) and Ridge Regression (right)



The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ (lasso) and $\beta_1^2 + \beta_2^2 \le t^2$ (ridge regression), while the red ellipses are the contours of the least squares error function.

Predictive Ability

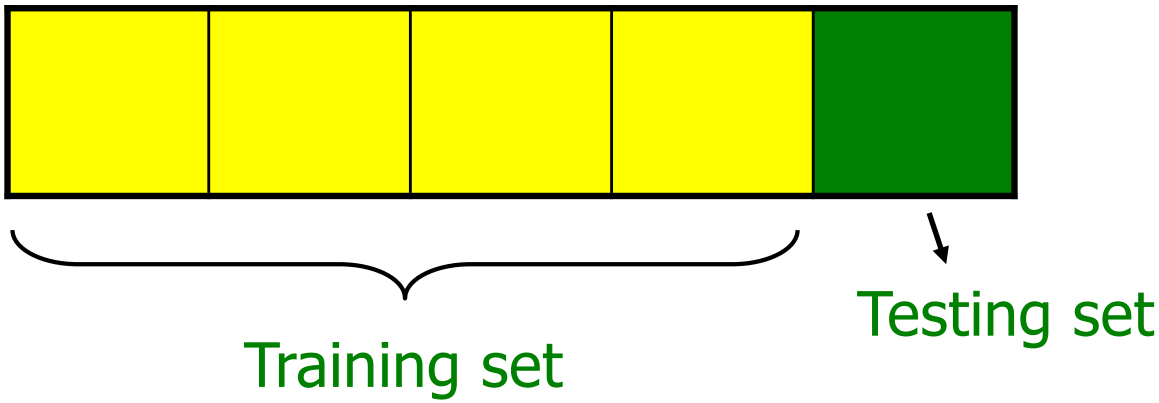


(Hastie et al 2009)

Behavior of test sample and training sample error as the model complexity is varied

Cross-validation

- *K*-FOLD



Training set

Testing set

$$\begin{cases} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\ \hat{\boldsymbol{\beta}} : \text{estimate of } \boldsymbol{\beta} \end{cases} \Rightarrow \begin{cases} \text{PMSE} = \frac{1}{m}\sum_i (y_i - \hat{y}_i)^2 \\ \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \end{cases}$$

- LEAVE-ONE-OUT (“*n*-FOLD”)

PAUSE

- ⇨ Model Comparison, variable selection
- ⇨ Penalized regression
- ⇨ Cross-validation

Next PAUSE, slide 38 (end)

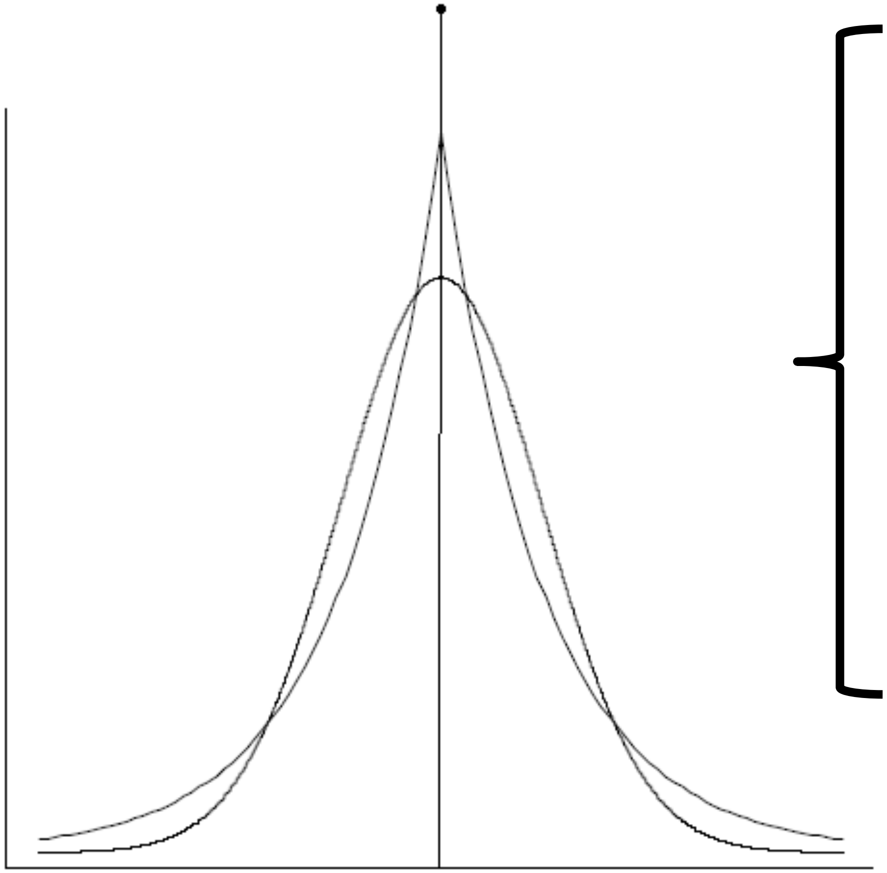
Bayesian Alternative

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j + \mathbf{e} \;\longrightarrow\; \mathbf{y} \mid \boldsymbol{\mu}, \mathbf{g}_j, \boldsymbol{\sigma}_e^2 \sim \mathrm{N}(\mathbf{1}\boldsymbol{\mu} + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j, \mathbf{I}\boldsymbol{\sigma}_e^2)$$

$$\left\{\begin{array}{l}
\text{BRR:}\quad g_j \mid \boldsymbol{\sigma}_0^2 \sim \mathrm{N}(0, \boldsymbol{\sigma}_0^2) \\
\text{Bayes A:}\quad g_j \mid \boldsymbol{\sigma}_j^2 \sim \mathrm{N}(0, \boldsymbol{\sigma}_j^2),\;\; \boldsymbol{\sigma}_j^2 \sim \boldsymbol{\chi}^{-2}(\boldsymbol{\nu}, \mathrm{S}) \\
\text{Bayes B,C:}\quad g_j \mid k, \boldsymbol{\sigma}_j^2 \sim \boldsymbol{\pi} \times \mathrm{N}(0, k\boldsymbol{\sigma}_j^2) + (1 - \boldsymbol{\pi}) \times \mathrm{N}(0, \boldsymbol{\sigma}_j^2) \\
\text{BLasso:}\quad g_j \mid \boldsymbol{\sigma}_j^2 \sim \mathrm{N}(0, \boldsymbol{\sigma}_j^2),\;\; \boldsymbol{\sigma}_j^2 \sim \mathrm{Exponential}(\boldsymbol{\lambda}) \\
\text{BX:}\quad g_j \mid \boldsymbol{\sigma}_j^2 \sim \mathrm{N}(0, \boldsymbol{\sigma}_j^2),\;\; \boldsymbol{\sigma}_j^2 \sim \mathrm{X}
\end{array}\right.$$

Normal/Independent Distributions

$$p(g_j) = \int_{\sigma_j^2} p(g_j \mid \boldsymbol{\sigma}_j^2)\, p(\boldsymbol{\sigma}_j^2)\, d\boldsymbol{\sigma}_j^2$$



BRR: Normal

Bayes A: Student-t

Bayes B,C: Mixtures

BLasso: Double exponential

GBLUP

Regression with genetic effects with normal distribution with common variance

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j + \mathbf{e} \quad \text{, with:} \quad g_j \mid \boldsymbol{\sigma}_g^2 \sim \mathrm{N}(0, \boldsymbol{\sigma}_g^2)$$

Equivalent Model

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{a} + \mathbf{e} \quad \text{, with:} \quad \mathbf{a} \mid \boldsymbol{\sigma}_a^2 \sim \mathrm{N}(\mathbf{0}, \mathbf{G}\boldsymbol{\sigma}_a^2)$$

⇨ **G** is the genomic relationship matrix:

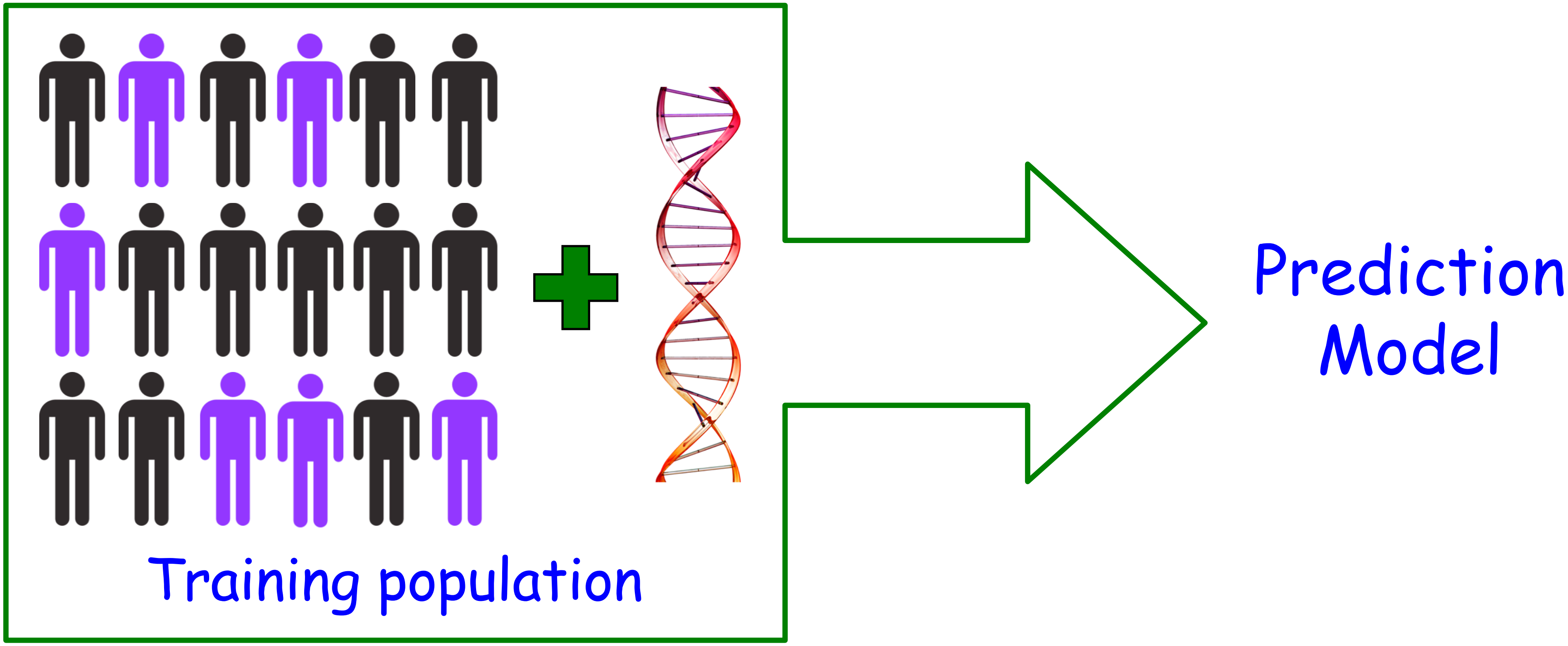
$$\mathbf{G} = \left(2 \sum_{j=1}^{p} p_j (1 - p_j) \right)^{-1} (\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})'$$

ssGBLUP

Single-step GBLUP: Single mixed model with all animals (genotyped and non-genotyped) included, with matrix **A** replaced by **H**

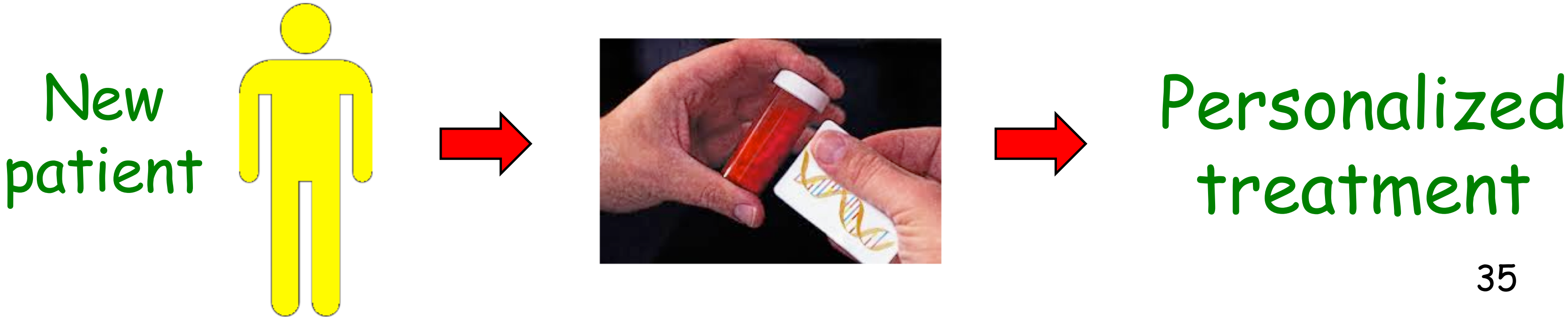
$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

Preventive and Personalized Medicine



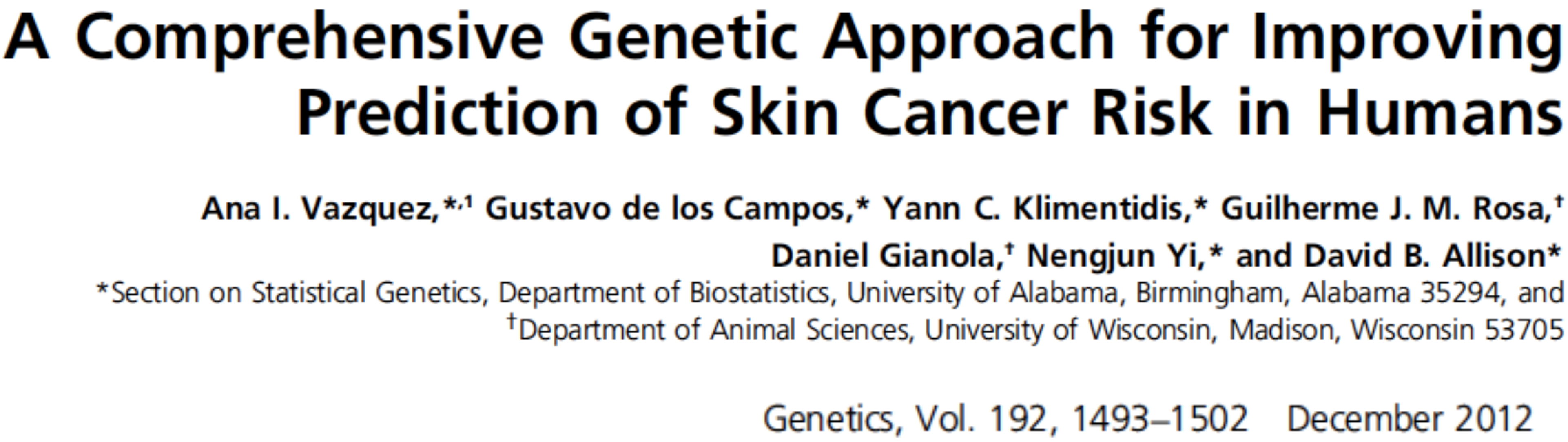
Training population

Prediction Model



New patient

Personalized treatment



A Comprehensive Genetic Approach for Improving Prediction of Skin Cancer Risk in Humans

Ana I. Vazquez,[*,1] Gustavo de los Campos,[*] Yann C. Klimentidis,[*] Guilherme J. M. Rosa,[†] Daniel Gianola,[†] Nengjun Yi,[*] and David B. Allison[*]

[*]Section on Statistical Genetics, Department of Biostatistics, University of Alabama, Birmingham, Alabama 35294, and [†]Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin 53705

Genetics, Vol. 192, 1493–1502 December 2012

- ⇨ 5,132 subjects from Framingham Heart Study
- ⇨ Phenotypes measured from 1948 until death
- ⇨ Genotypes: Affymetrix 500K SNPs

Three generations of participants.

Photo: http://www.framinghamheartstudy.org/

Models

1. No-SNP: standard covariables
2. Covariates + familial relationships
3. Covariates + SNPs (PC or Bayesian LASSO)

Probit B-LASSO $$p(\mathbf{y}|\boldsymbol{\beta},\mathbf{u}) = \prod_{i=1}^{5132}\left\{[\Phi(\eta_i)]^{y_i}[1-\Phi(\eta_i)]^{1-y_i}\right\}$$

$$\eta_i = \beta_0 + \sum_{j=1}^{p_1} x_{1ij}\beta_{1j} + \sum_{j=1}^{p_2} x_{2ij}\beta_{2j}$$

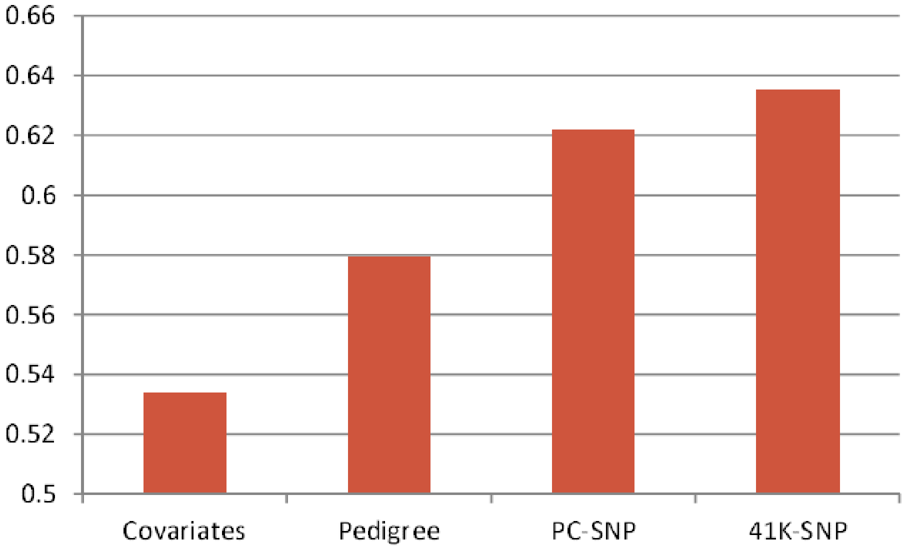
or

$$\eta_i = \beta_0 + \sum_{j=1}^{p_1} x_{1ij}\beta_{1j} + u_i$$

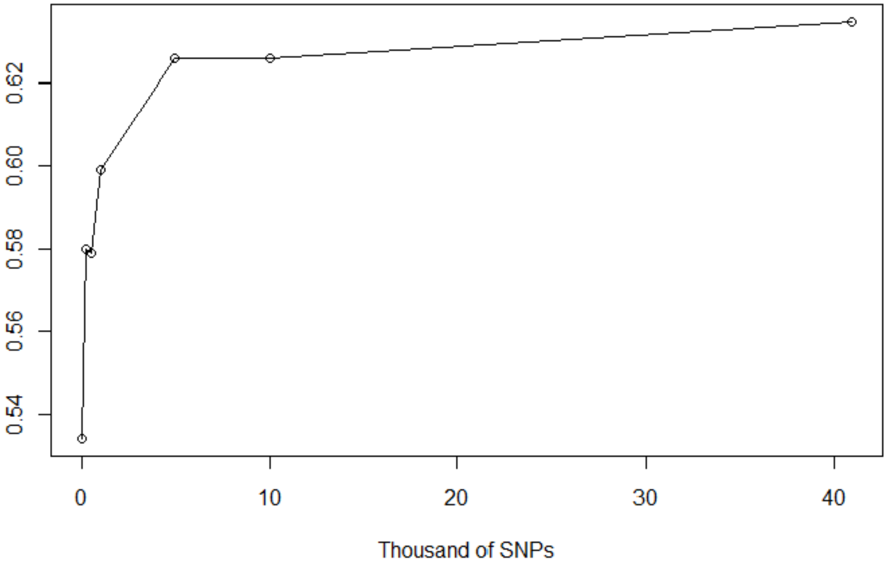
$$\begin{aligned} p(\beta_0,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{u},\boldsymbol{\tau}^2,\lambda) \propto {} & \left[\prod_{j=1}^{p} N\left(\beta_{2,j}\middle|0,\tau_j^2\right)\right] \\ & \times \left[\prod_{j=1}^{p} \mathrm{Exp}\left(\tau_j^2|\lambda^2\right)\right] \times G(\lambda^2|\alpha_1,\alpha_2) \\ & \times N(\mathbf{u}|\mathbf{0},\mathbf{A}\sigma_u^2)\times\chi^{-2}(\sigma_u^2|S,df), \end{aligned}$$

Results (ROC, Area Under the Curve)

Comparison of Models



Models with increasing number of SNPs


Thousand of SNPs