Genomic Selection in Livestock

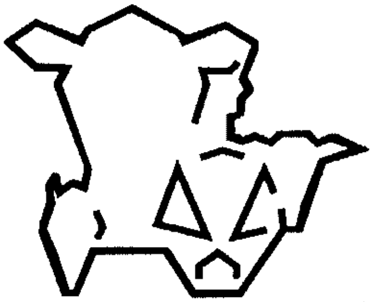
Dorian Garrick
Rohan Fernando
Jack Dekkers

ANIMAL SCIENCE

Animal Breeding & Genetics

Animal Breeding and Genetics
Department of Animal Science
Iowa State University

June 14-18 2010



OUTLINE / TOPICS

- Introduction and motivation (Jack)

- Models to predict single SNP effects (Dorian)
    - Fixed effect models
    - Fitting SNPs as random effects

- Bayesian methods (Rohan)
    - Bayes theorem
    - Gibbs sampler
    - Metropolis Hastings

- Genomic prediction (Dorian)
    - An equivalent (animal) model for genomic prediction
    - Some alternative computing strategies that are not equivalent models
    - Two practical problems of genomic prediction

- Bayesian methods applied to genomic prediction (Rohan)
    - Bayes A
    - Bayes B
    - G-BLUP

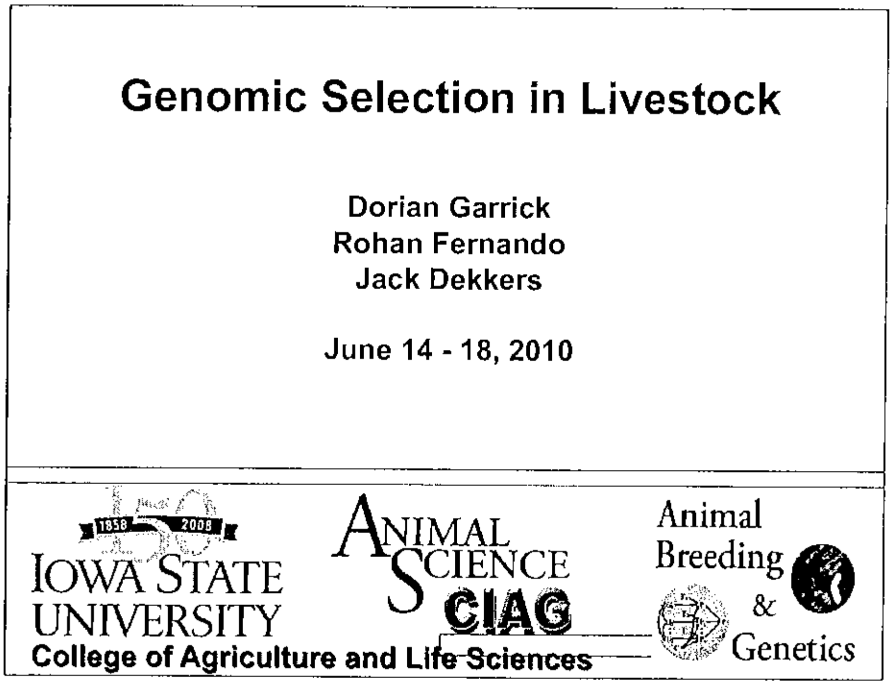
- Interpretation of SNP effect estimates (Jack)
    - Linkage and Linkage Disequilibrium
    - Spurious associations from relationships and breed mixtures

- Application of genomic prediction models to real data (Dorian)
    - Training and validation
    - Problems with validation
    - Improved Validation – simulated real beef cattle applications
    - Validation Statistics

- Other genomic prediction methods (Rohan)
    - Bayes Cπ and estimation of π
    - Estimating the scale factor
    - Alternative distributions – Heavy-tailed vs. Normal distributions

ADDITIONAL TOPICS

- BIGS Genomic Selection Analysis software (Dorian)
- Genomic Prediction across breeds and in crossbreds (Dorian, Jack)
- Low density panels for Genomic Selection (Jack)
- Degression of EBV and weighting information (Dorian)
- Pooling genomic EBV and pedigree or own information (Dorian, Jack)



Genomic Selection in Livestock

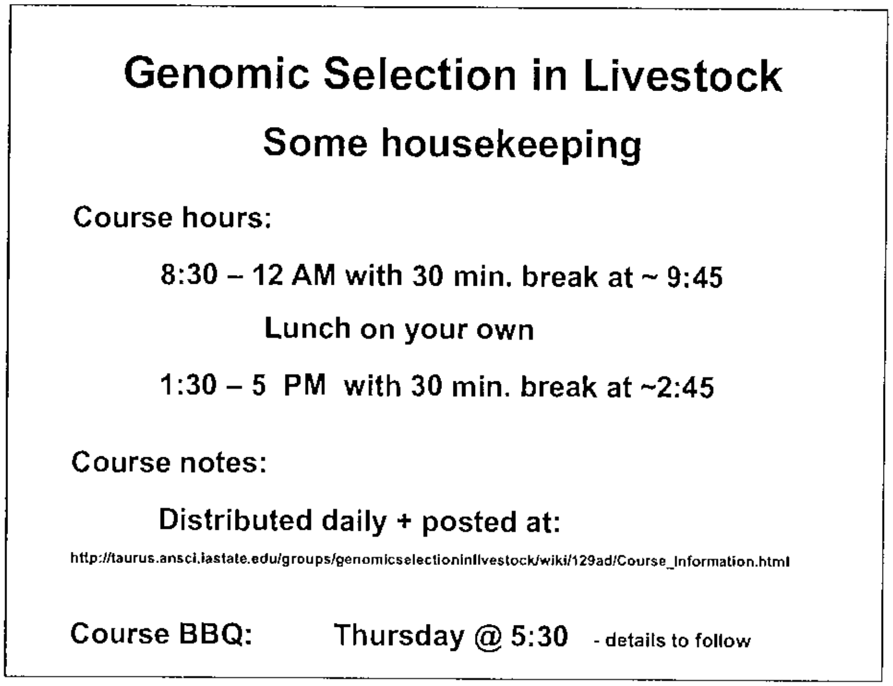
Dorian Garrick
Rohan Fernando
Jack Dekkers

June 14 - 18, 2010

IOWA STATE UNIVERSITY
College of Agriculture and Life Sciences

ANIMAL SCIENCE
CIAG

Animal Breeding & Genetics



Genomic Selection in Livestock

Some housekeeping

Course hours:

8:30 – 12 AM with 30 min. break at ~ 9:45

Lunch on your own

1:30 – 5 PM with 30 min. break at ~2:45

Course notes:

Distributed daily + posted at:

http://taurus.ansci.iastate.edu/groups/genomicselectioninlivestock/wiki/129ad/Course_Information.html

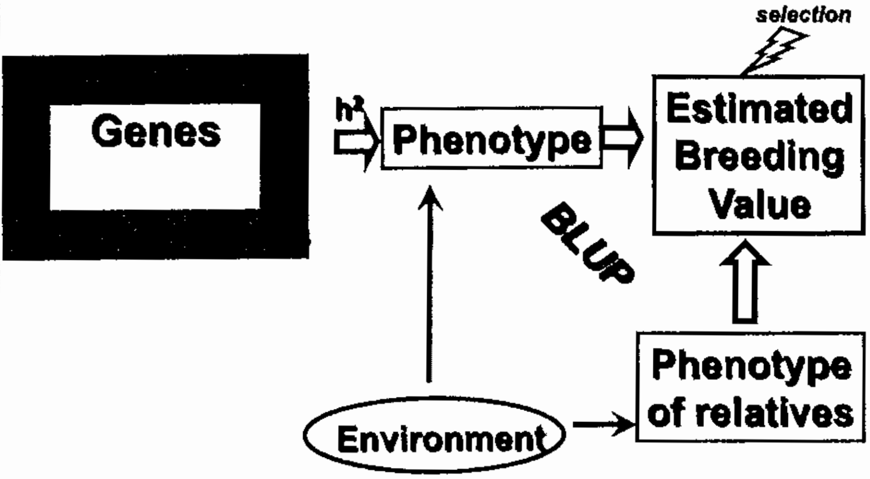
Course BBQ: Thursday @ 5:30 - details to follow

Genomic Selection in Livestock

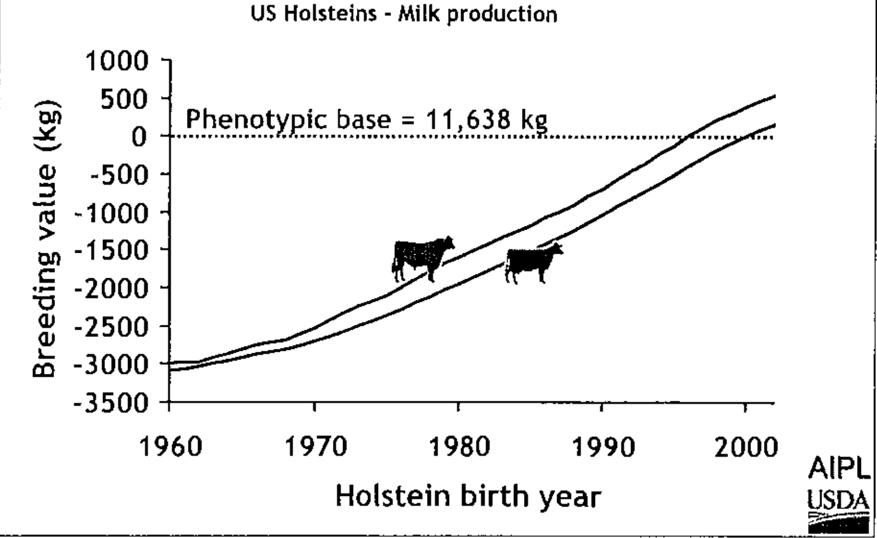
Introduction and Motivation

Jack Dekkers

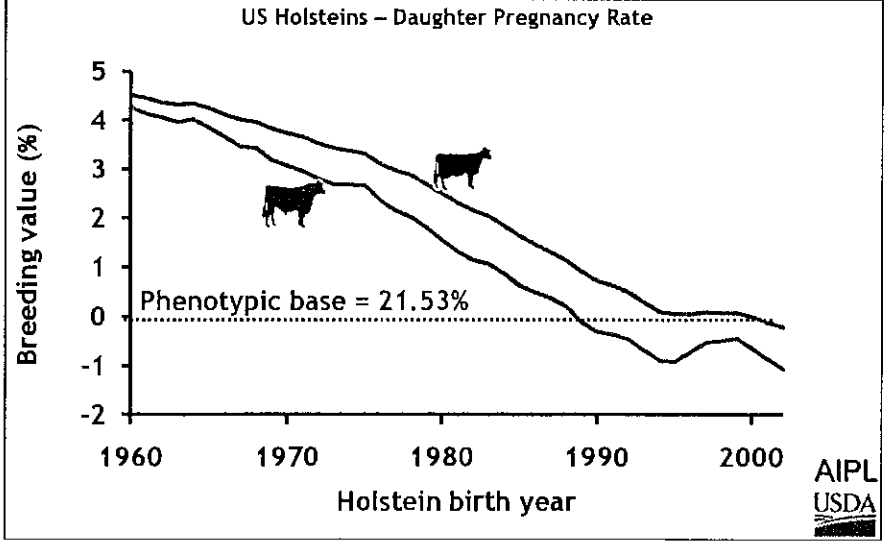
Past and Current Selection Strategies

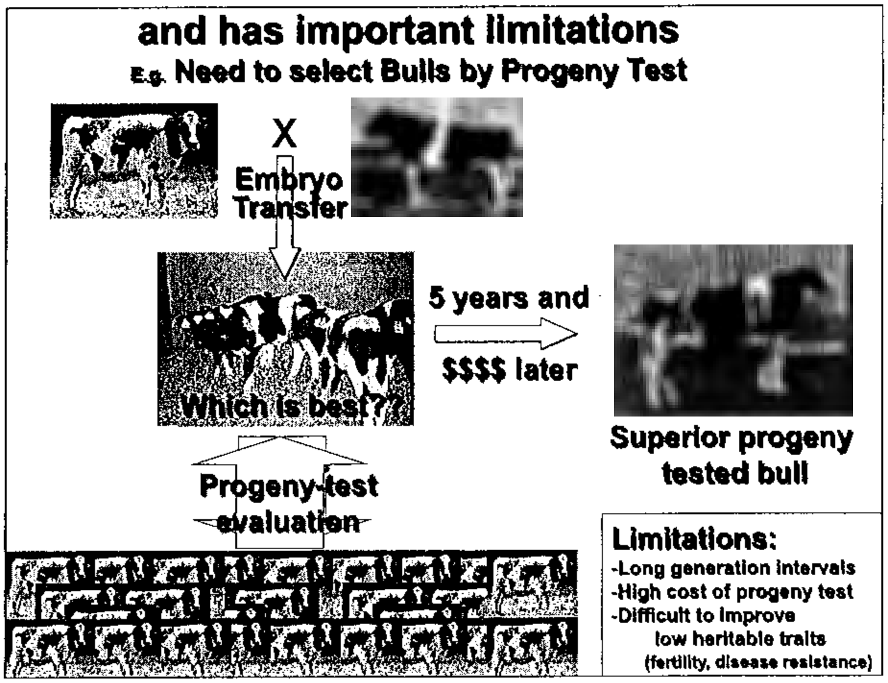


This approach has been very successful for many traits



. . . . . not for others . . . .





and has important limitations

E.g. Need to select Bulls by Progeny Test

X

Embryo Transfer

Which is best??

5 years and

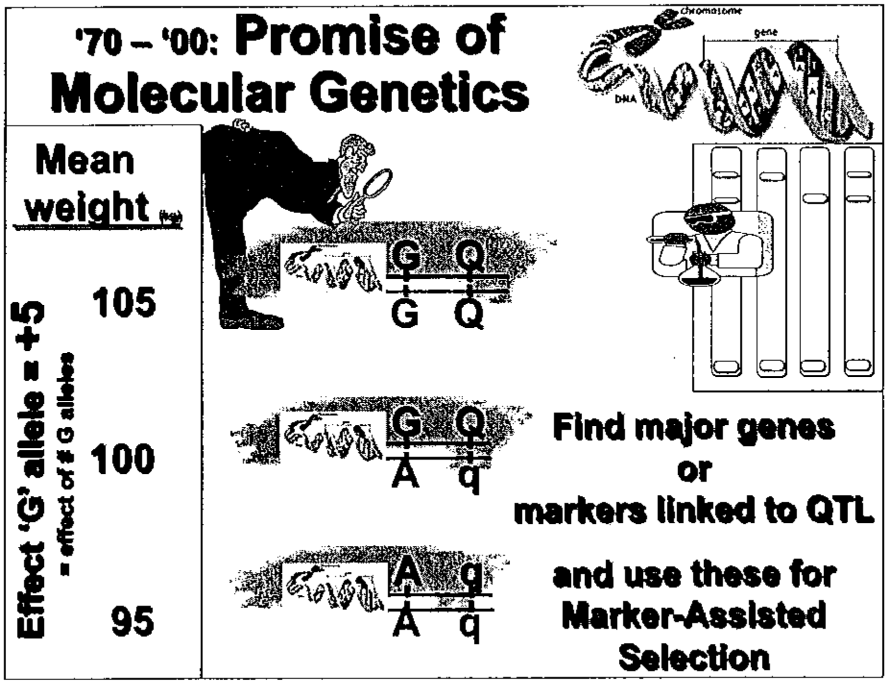
$$$$ later

Superior progeny tested bull

Progeny-test evaluation

Limitations:

- -Long generation intervals
- -High cost of progeny test
- -Difficult to improve low heritable traits (fertility, disease resistance)



'70 – '00: Promise of Molecular Genetics

chromosome
gene
DNA

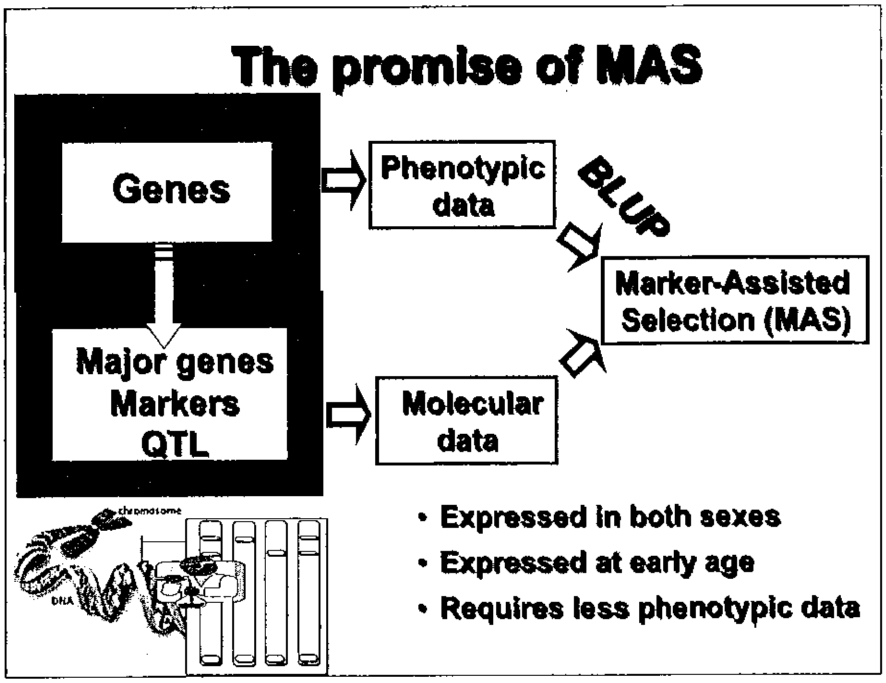
Mean weight (kg)	
105	G Q G Q
100	G Q A q
95	A q A q

Effect 'G' allele = +5

= effect of # G alleles

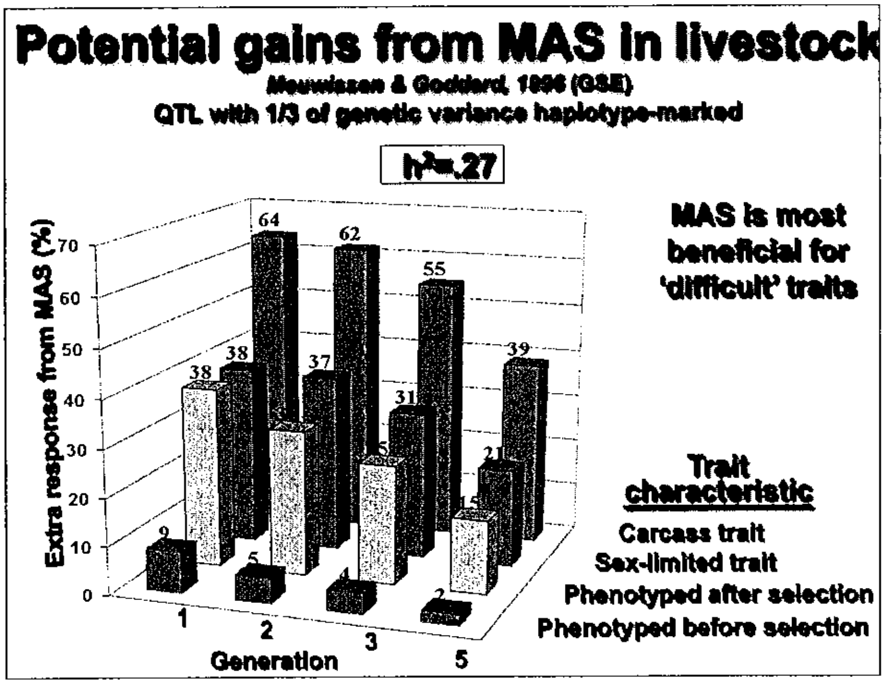
Find major genes
or
markers linked to QTL

and use these for
Marker-Assisted
Selection



The promise of MAS

- Expressed in both sexes
- Expressed at early age
- Requires less phenotypic data



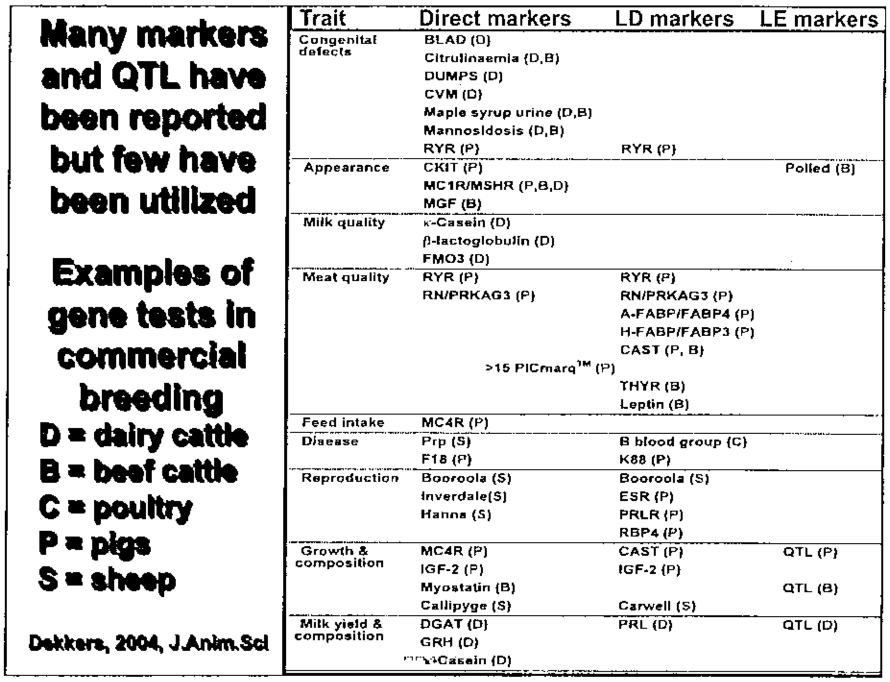
Potential gains from MAS in livestock

Meuwissen & Goddard, 1996 (GSE)

QTL with 1/3 of genetic variance haplotype-marked

$h^2=.27$

MAS is most beneficial for 'difficult' traits



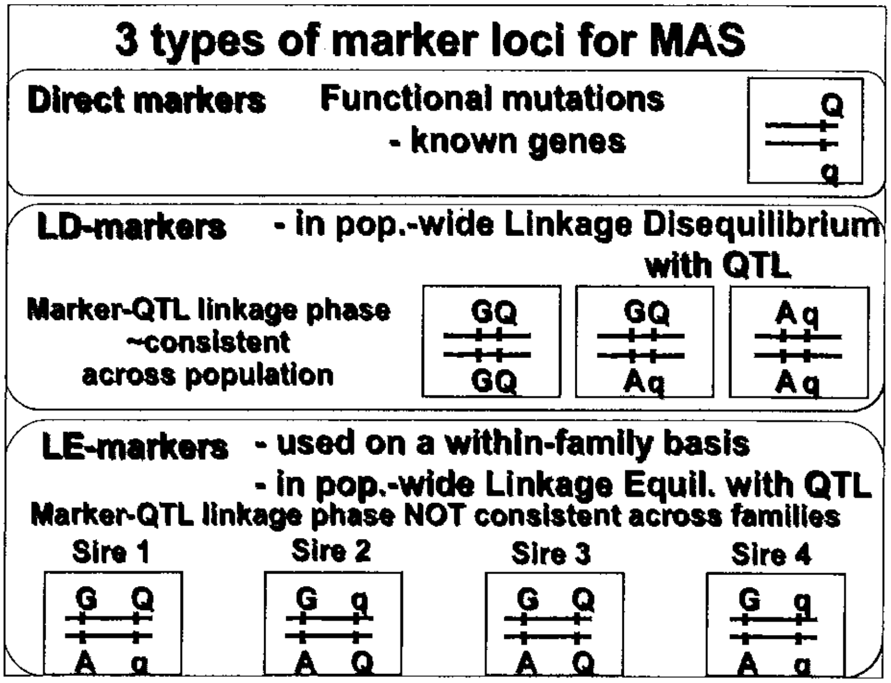
Many markers and QTL have been reported but few have been utilized

Examples of gene tests in commercial breeding

D = dairy cattle
B = beef cattle
C = poultry
P = pigs
S = sheep

Dekkers, 2004, J.Anim.Sci

Trait	Direct markers	LD markers	LE markers
Congenital defects	BLAD (D)		
	Citrulinaemia (D,B)		
	DUMPS (D)		
	CVM (D)		
	Maple syrup urine (D,B)		
	Mannosidosis (D,B)		
	RYR (P)	RYR (P)	
Appearance	CKIT (P)		Polled (B)
	MC1R/MSHR (P,B,D)		
	MGF (B)		
Milk quality	κ-Casein (D)		
	β-lactoglobulin (D)		
	FMO3 (D)		
Meat quality	RYR (P)	RYR (P)	
	RN/PRKAG3 (P)	RN/PRKAG3 (P)	
		A-FABP/FABP4 (P)	
		H-FABP/FABP3 (P)	
		CAST (P, B)	
	>15 PICmarq™ (P)		
		THYR (B)	
		Leptin (B)	
Feed intake	MC4R (P)		
Disease	Prp (S)	B blood group (C)	
	F18 (P)	K88 (P)	
Reproduction	Booroola (S)	Booroola (S)	
	Inverdale(S)	ESR (P)	
	Hanna (S)	PRLR (P)	
		RBP4 (P)	
Growth & composition	MC4R (P)	CAST (P)	QTL (P)
	IGF-2 (P)	IGF-2 (P)	
	Myostatin (B)		QTL (B)
	Callipyge (S)	Carwell (S)	
Milk yield & composition	DGAT (D)	PRL (D)	QTL (D)
	GRH (D)		
	[illegible]Casein (D)		



3 types of marker loci for MAS

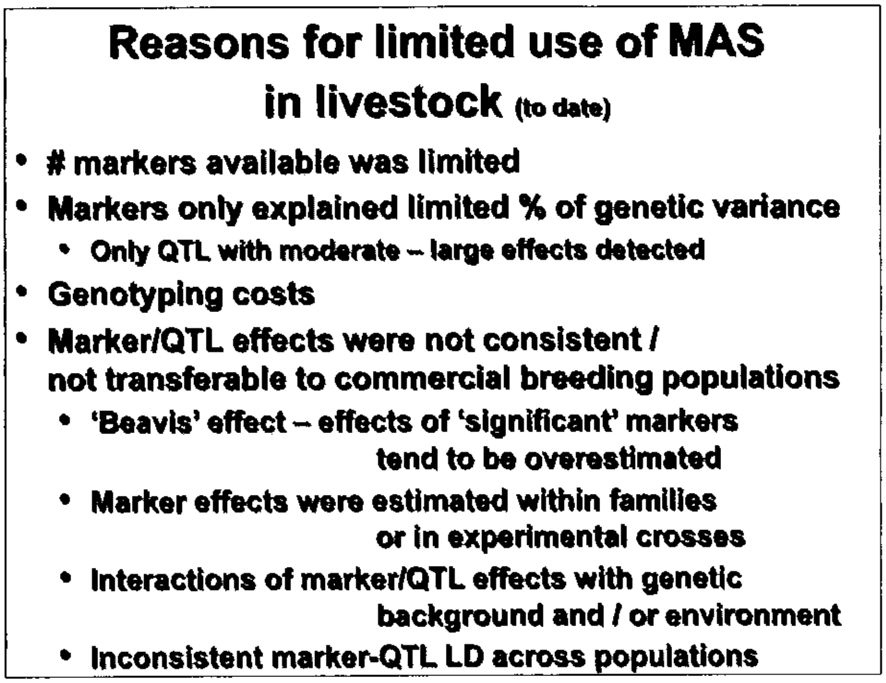
Direct markers **Functional mutations**
- known genes

LD-markers **- in pop.-wide Linkage Disequilibrium with QTL**

Marker-QTL linkage phase ~consistent across population

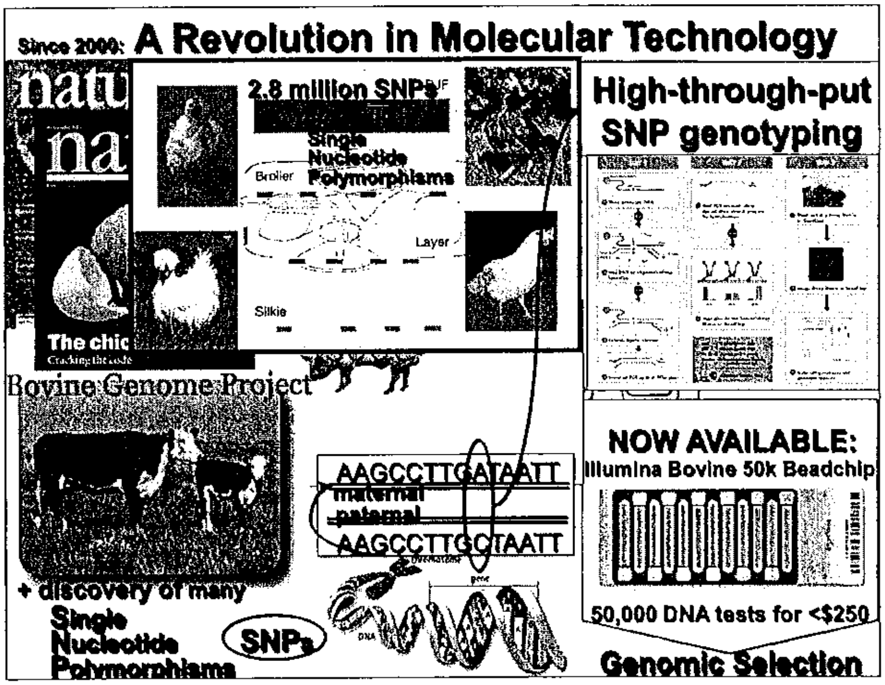
LE-markers **- used on a within-family basis**
- in pop.-wide Linkage Equil. with QTL

Marker-QTL linkage phase NOT consistent across families



Reasons for limited use of MAS in livestock (to date)

- # markers available was limited
- Markers only explained limited % of genetic variance
  - Only QTL with moderate – large effects detected
- Genotyping costs
- Marker/QTL effects were not consistent / not transferable to commercial breeding populations
  - 'Beavis' effect – effects of 'significant' markers tend to be overestimated
  - Marker effects were estimated within families or in experimental crosses
  - Interactions of marker/QTL effects with genetic background and / or environment
  - Inconsistent marker-QTL LD across populations



Since 2000: A Revolution in Molecular Technology

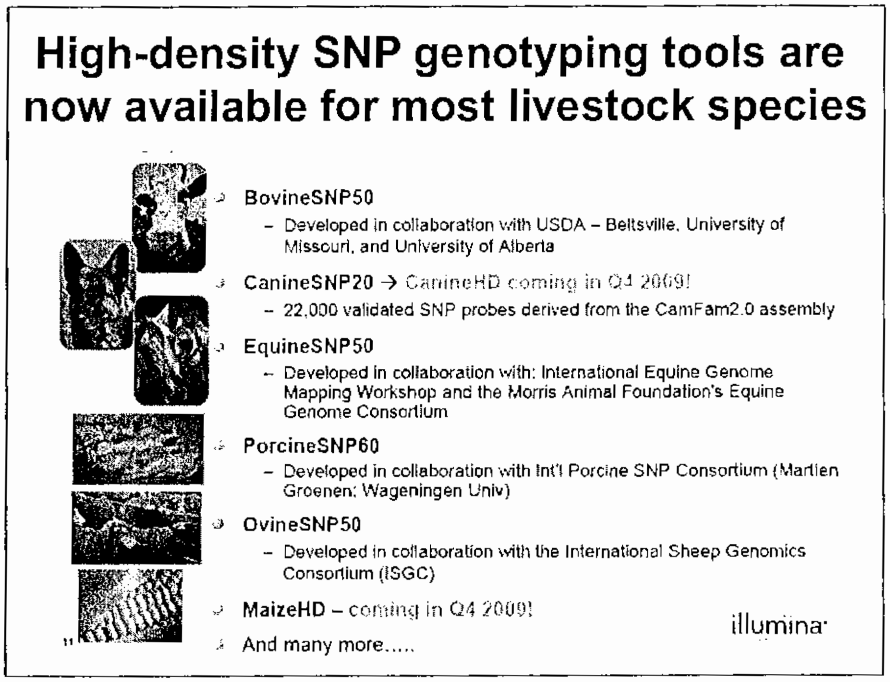
High-through-put SNP genotyping

NOW AVAILABLE: Illumina Bovine 50k Beadchip

50,000 DNA tests for <$250

Genomic Selection

+ discovery of many Single Nucleotide Polymorphisms SNPs

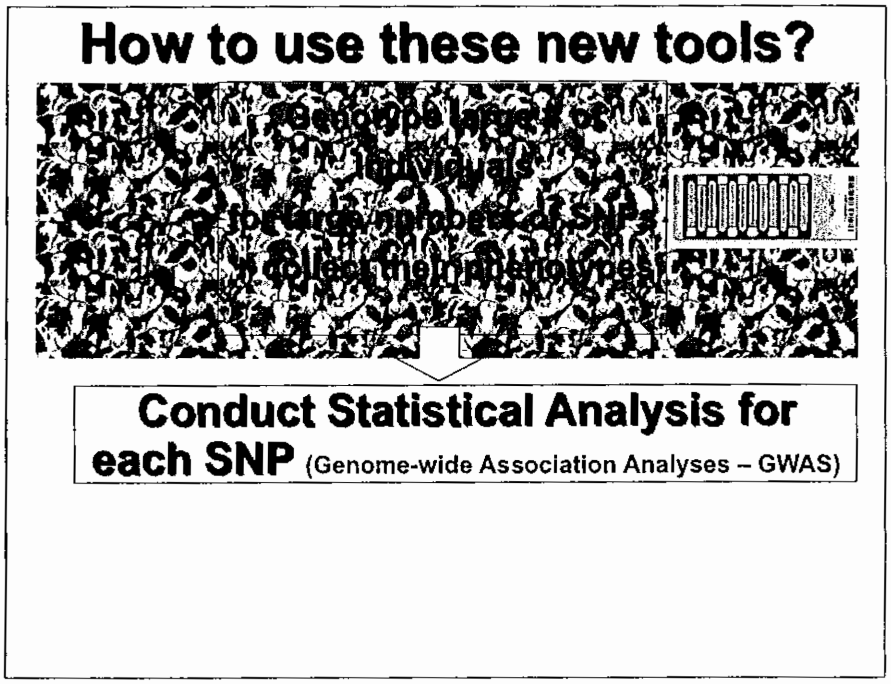


High-density SNP genotyping tools are now available for most livestock species

- BovineSNP50
  - Developed in collaboration with USDA – Beltsville, University of Missouri, and University of Alberta
- CanineSNP20 → CanineHD coming in Q4 2009!
  - 22,000 validated SNP probes derived from the CamFam2.0 assembly
- EquineSNP50
  - Developed in collaboration with: International Equine Genome Mapping Workshop and the Morris Animal Foundation's Equine Genome Consortium
- PorcineSNP60
  - Developed in collaboration with Int'l Porcine SNP Consortium (Martien Groenen: Wageningen Univ)
- OvineSNP50
  - Developed in collaboration with the International Sheep Genomics Consortium (ISGC)
- MaizeHD – coming in Q4 2009!
- And many more.....

illumina

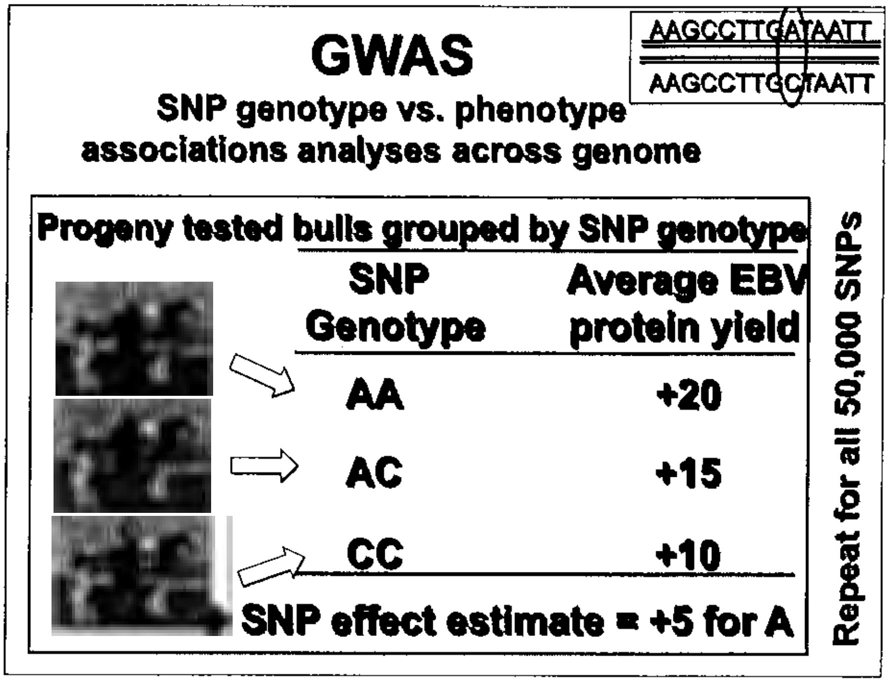
11



How to use these new tools?

Genotype large # of individuals for large numbers of SNPs & collect their phenotypes

Conduct Statistical Analysis for each SNP (Genome-wide Association Analyses – GWAS)



GWAS

AAGCCTTGATAATT

AAGCCTTGCTAATT

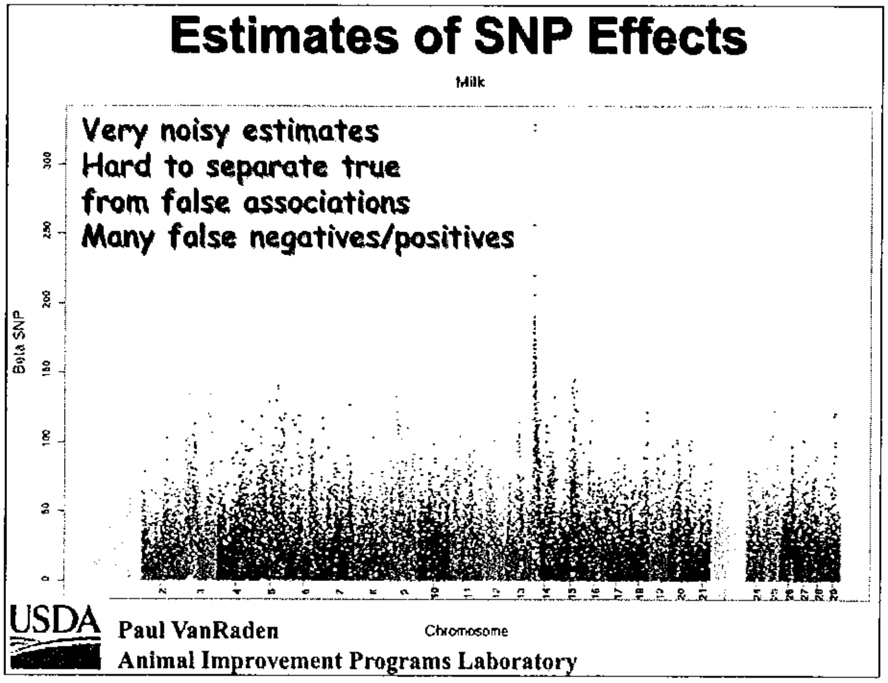
SNP genotype vs. phenotype associations analyses across genome

Progeny tested bulls grouped by SNP genotype

SNP Genotype	Average EBV protein yield
AA	+20
AC	+15
CC	+10

SNP effect estimate = +5 for A

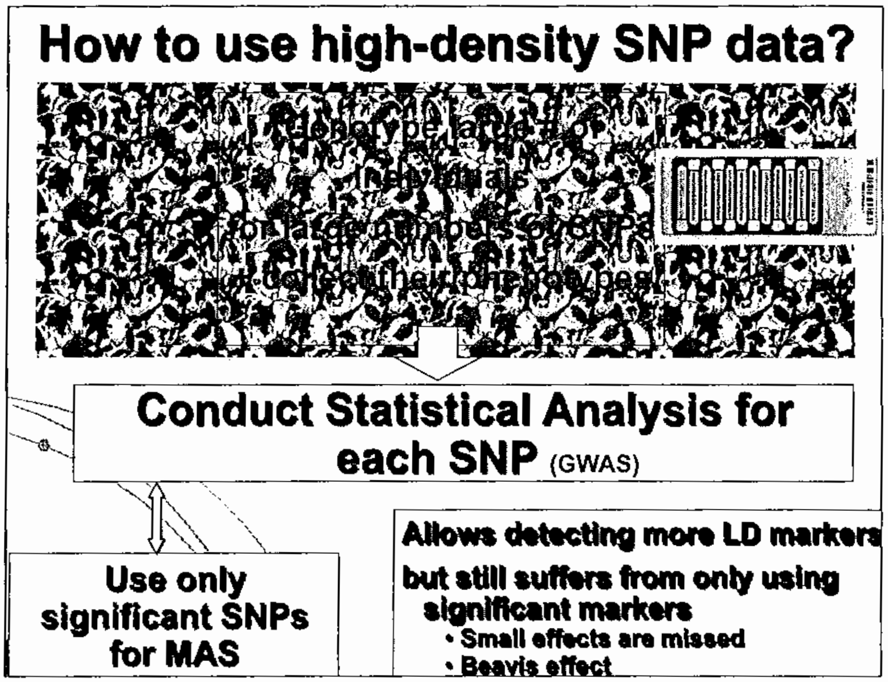
Repeat for all 50,000 SNPs



Estimates of SNP Effects

Very noisy estimates
Hard to separate true
from false associations
Many false negatives/positives

USDA Paul VanRaden
Animal Improvement Programs Laboratory



How to use high-density SNP data?

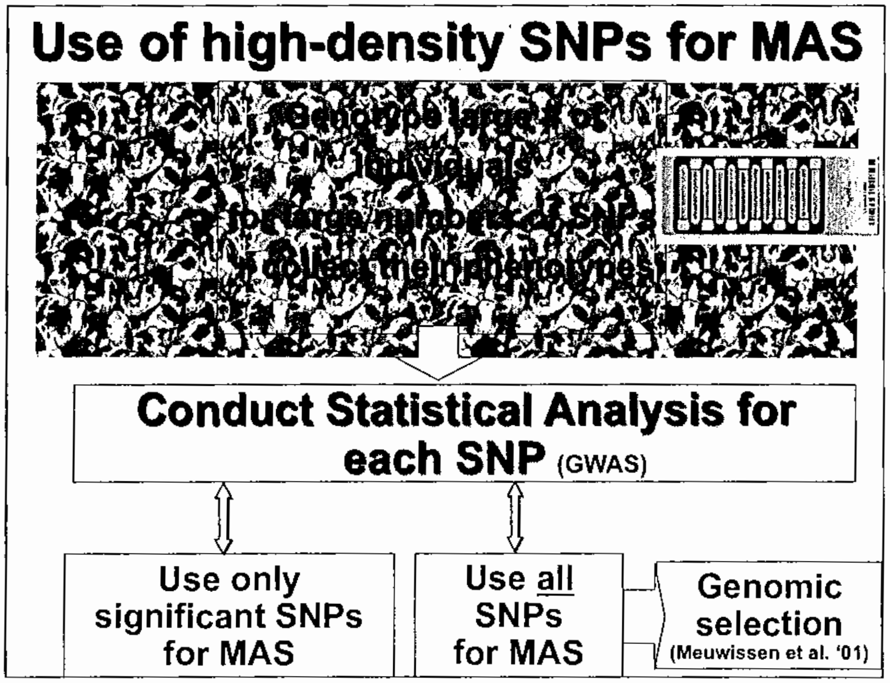
Genotype large # of individuals
for large numbers of SNPs
& collect their phenotypes

Conduct Statistical Analysis for each SNP (GWAS)

Use only significant SNPs for MAS

Allows detecting more LD markers but still suffers from only using significant markers

- Small effects are missed
- Beavis effect



Use of high-density SNPs for MAS

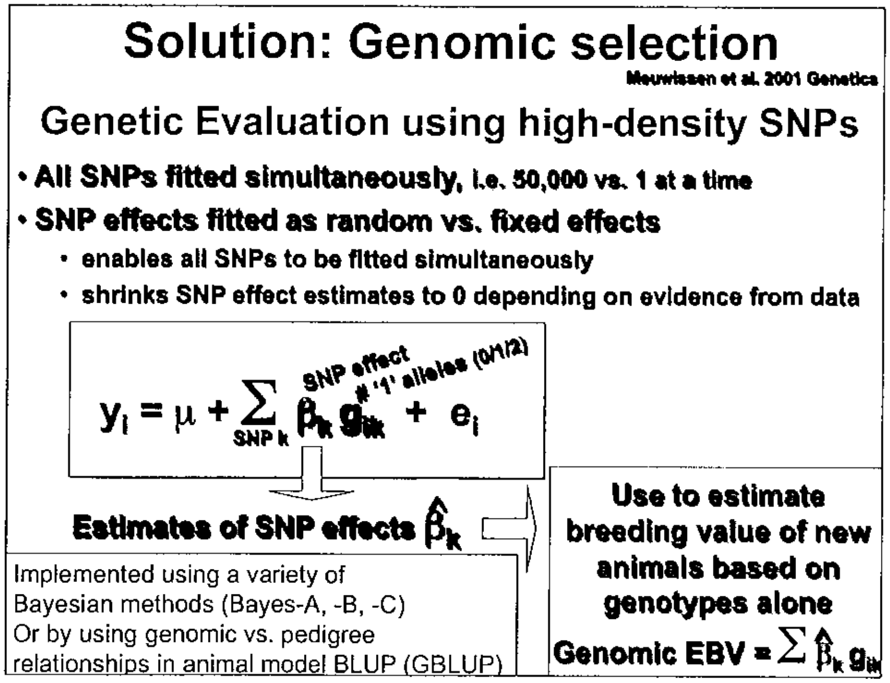
Genotype large # of individuals
for large numbers of SNPs
& collect their phenotypes

Conduct Statistical Analysis for each SNP (GWAS)

Use only significant SNPs for MAS

Use all SNPs for MAS

Genomic selection (Meuwissen et al. '01)



Solution: Genomic selection

Meuwissen et al. 2001 Genetics

Genetic Evaluation using high-density SNPs

- **All SNPs fitted simultaneously,** i.e. 50,000 vs. 1 at a time
- **SNP effects fitted as random vs. fixed effects**
  - enables all SNPs to be fitted simultaneously
  - shrinks SNP effect estimates to 0 depending on evidence from data

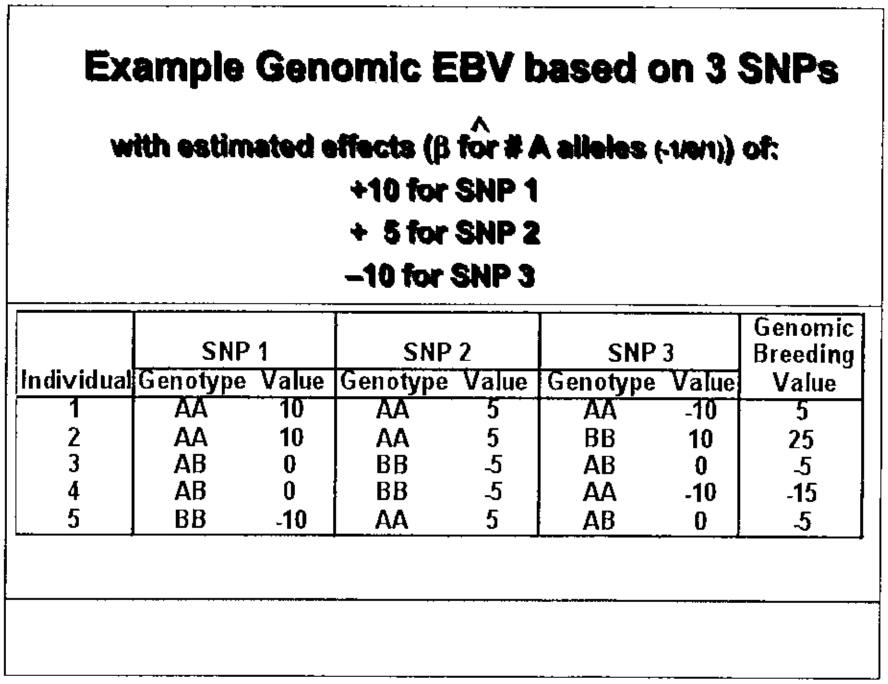
$$y_i = \mu + \sum_{\text{SNP } k} \overset{\text{SNP effect}}{\beta_k} \overset{\text{\# '1' alleles (0/1/2)}}{g_{ik}} + e_i$$

Estimates of SNP effects $\hat{\beta}_k$

Implemented using a variety of Bayesian methods (Bayes-A, -B, -C)
Or by using genomic vs. pedigree relationships in animal model BLUP (GBLUP)

Use to estimate breeding value of new animals based on genotypes alone

Genomic EBV = $\sum \hat{\beta}_k g_{ik}$



Example Genomic EBV based on 3 SNPs

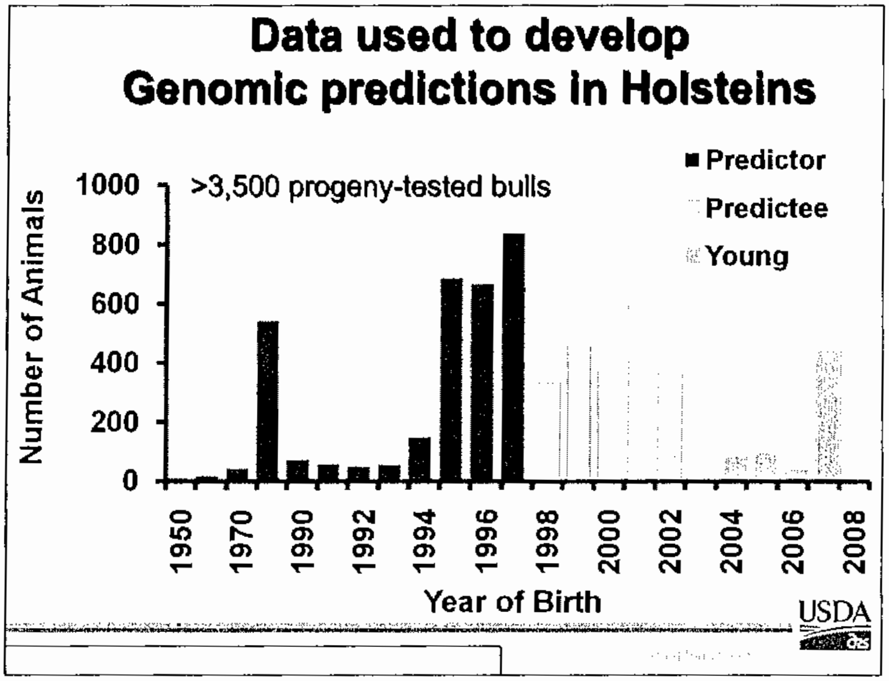
with estimated effects ($\hat{\beta}$ for # A alleles (-1/0/1)) of:

+10 for SNP 1

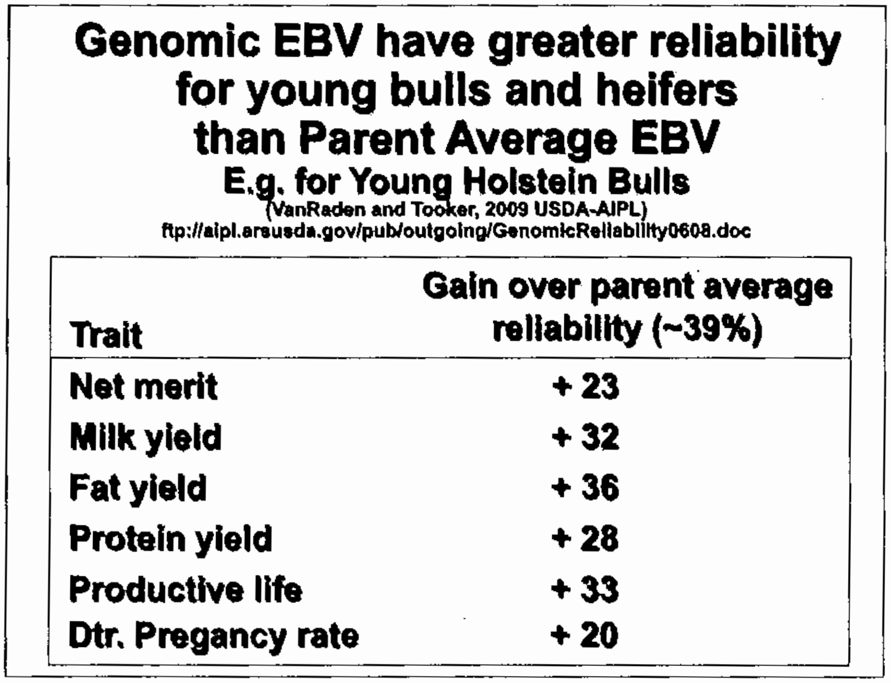
+ 5 for SNP 2

–10 for SNP 3

	SNP 1		SNP 2		SNP 3		Genomic Breeding Value
Individual	Genotype	Value	Genotype	Value	Genotype	Value	
1	AA	10	AA	5	AA	-10	5
2	AA	10	AA	5	BB	10	25
3	AB	0	BB	-5	AB	0	-5
4	AB	0	BB	-5	AA	-10	-15
5	BB	-10	AA	5	AB	0	-5



Data used to develop Genomic predictions in Holsteins



Genomic EBV have greater reliability for young bulls and heifers than Parent Average EBV

E.g. for Young Holstein Bulls

(VanRaden and Tooker, 2009 USDA-AIPL)

ftp://aipl.arsusda.gov/pub/outgoing/GenomicReliability0608.doc

Trait	Gain over parent average reliability (~39%)
Net merit	+ 23
Milk yield	+ 32
Fat yield	+ 36
Protein yield	+ 28
Productive life	+ 33
Dtr. Pregancy rate	+ 20

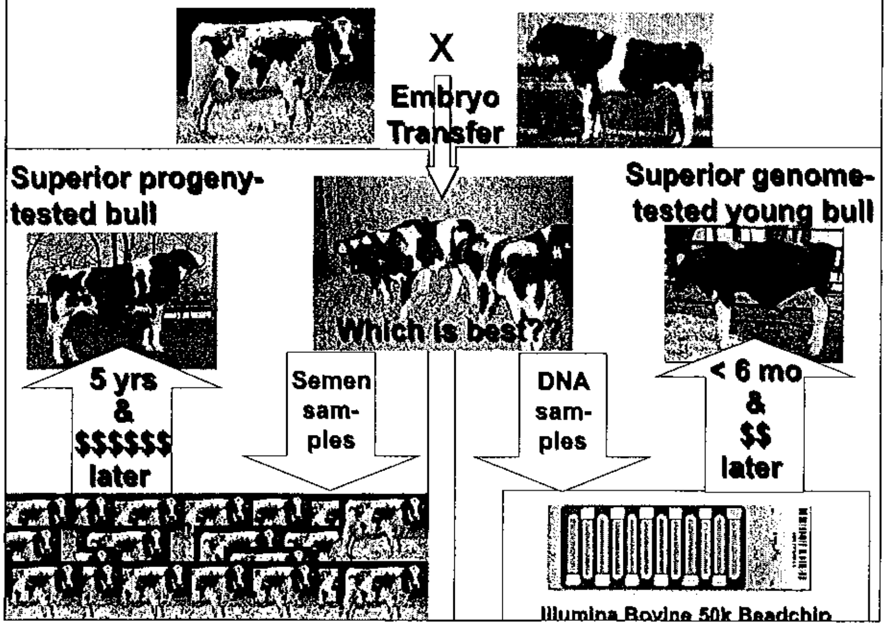
The Promise of Genomic Selection

(based on simulation and some empirical results)

- Increase accuracy of EBV at a young age
- Reduce generation intervals
- Reduce rates of inbreeding
- Reduce need to obtain phenotypes on selection candidates and/or close relatives

This has the potential to revolutionize the design and implementation of breeding programs for livestock (and plants)

Potential impact of GS on dairy cattle breeding



X

Embryo Transfer

Superior progeny-tested bull

Which is best??

Superior genome-tested young bull

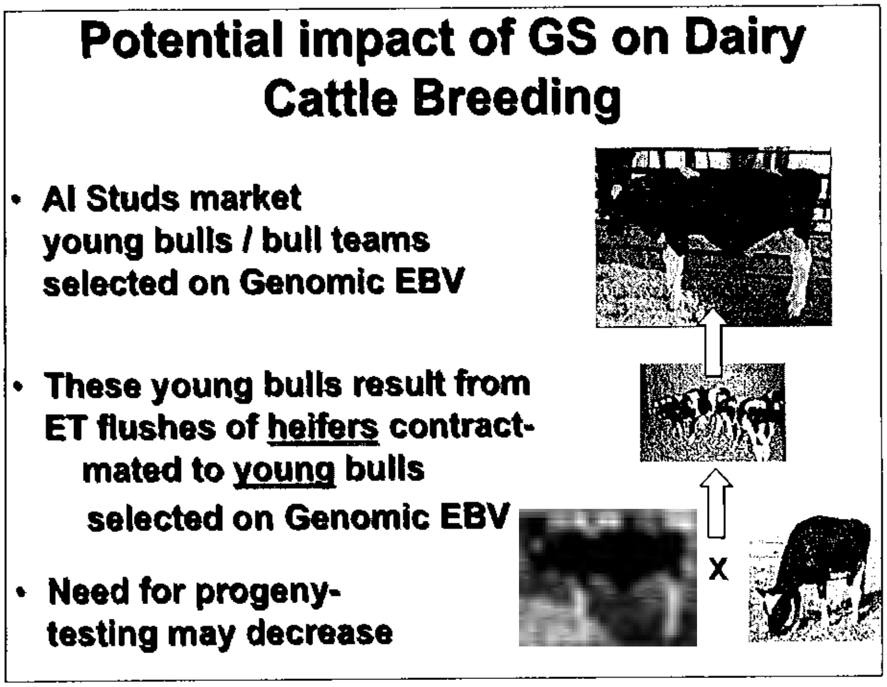
5 yrs & $$$$$$ later

Semen sam-ples

DNA sam-ples

< 6 mo & $$ later

Illumina Bovine 50k Beadchip



Potential impact of GS on Dairy Cattle Breeding

- AI Studs market young bulls / bull teams selected on Genomic EBV
- These young bulls result from ET flushes of <u>heifers</u> contract-mated to <u>young</u> bulls selected on Genomic EBV
- Need for progeny-testing may decrease

Genomic Selection in Livestock

Short course - focus

•Statistical, quantitative genetic, and computational aspects of genomic selection

•Software for genomic selection analyses

•Strategies for implementation of genomic selection in livestock breeding programs

1858 2008

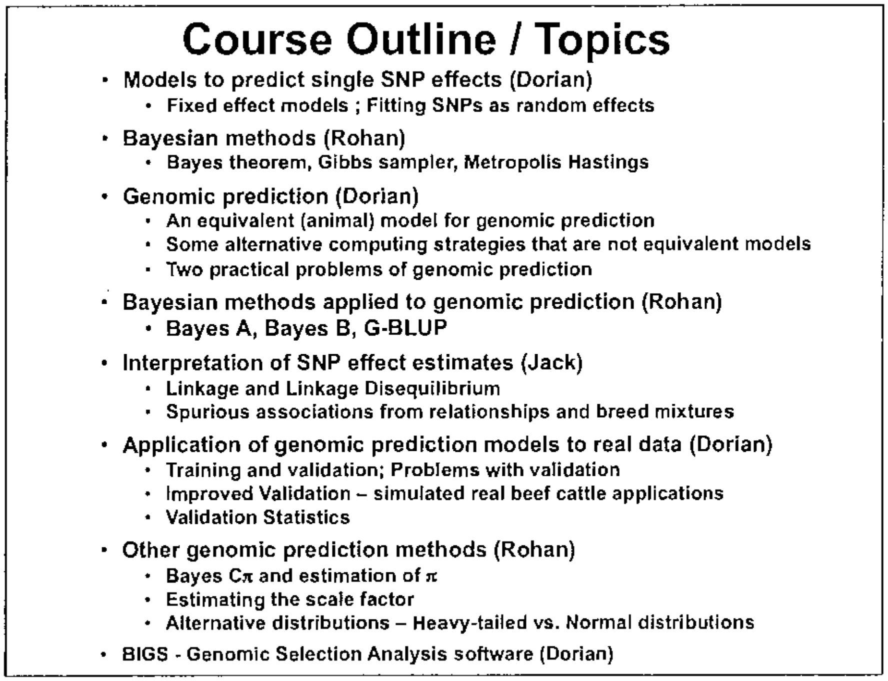
IOWA STATE UNIVERSITY

College of Agriculture and Life Sciences

ANIMAL SCIENCE

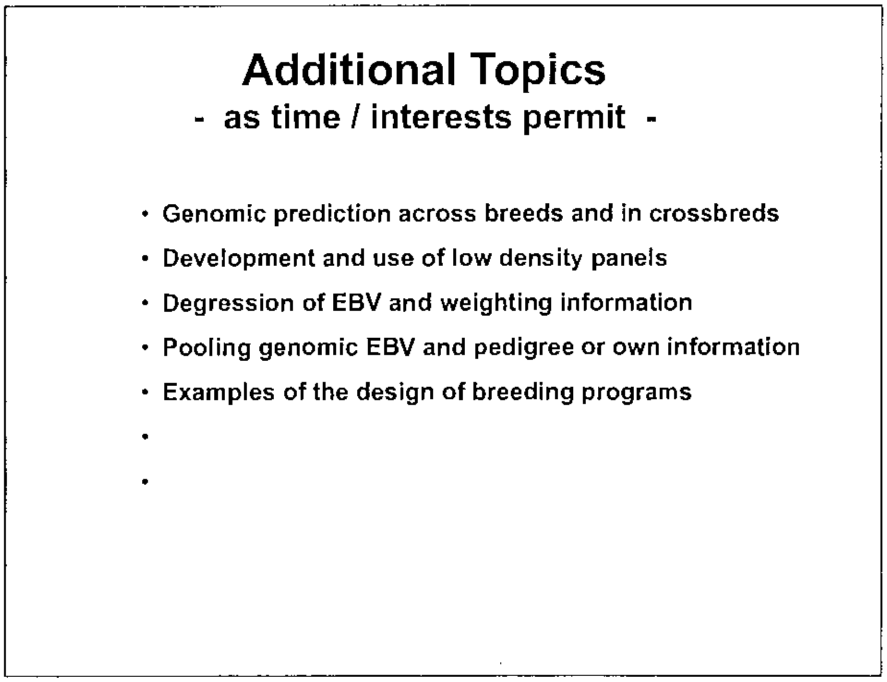
CIAG

Animal Breeding & Genetics



Course Outline / Topics

- Models to predict single SNP effects (Dorian)
  - Fixed effect models ; Fitting SNPs as random effects
- Bayesian methods (Rohan)
  - Bayes theorem, Gibbs sampler, Metropolis Hastings
- Genomic prediction (Dorian)
  - An equivalent (animal) model for genomic prediction
  - Some alternative computing strategies that are not equivalent models
  - Two practical problems of genomic prediction
- Bayesian methods applied to genomic prediction (Rohan)
  - Bayes A, Bayes B, G-BLUP
- Interpretation of SNP effect estimates (Jack)
  - Linkage and Linkage Disequilibrium
  - Spurious associations from relationships and breed mixtures
- Application of genomic prediction models to real data (Dorian)
  - Training and validation; Problems with validation
  - Improved Validation – simulated real beef cattle applications
  - Validation Statistics
- Other genomic prediction methods (Rohan)
  - Bayes Cπ and estimation of π
  - Estimating the scale factor
  - Alternative distributions – Heavy-tailed vs. Normal distributions
- BIGS - Genomic Selection Analysis software (Dorian)



Additional Topics
- as time / interests permit -

- Genomic prediction across breeds and in crossbreds
- Development and use of low density panels
- Degression of EBV and weighting information
- Pooling genomic EBV and pedigree or own information
- Examples of the design of breeding programs
-
-

Genomic Selection
Why does it (not?) work?

Jack Dekkers

AB&G short course 2010

1858 150 2008
IOWA STATE UNIVERSITY
College of Agriculture and Life Sciences

ANIMAL SCIENCE
CIAG

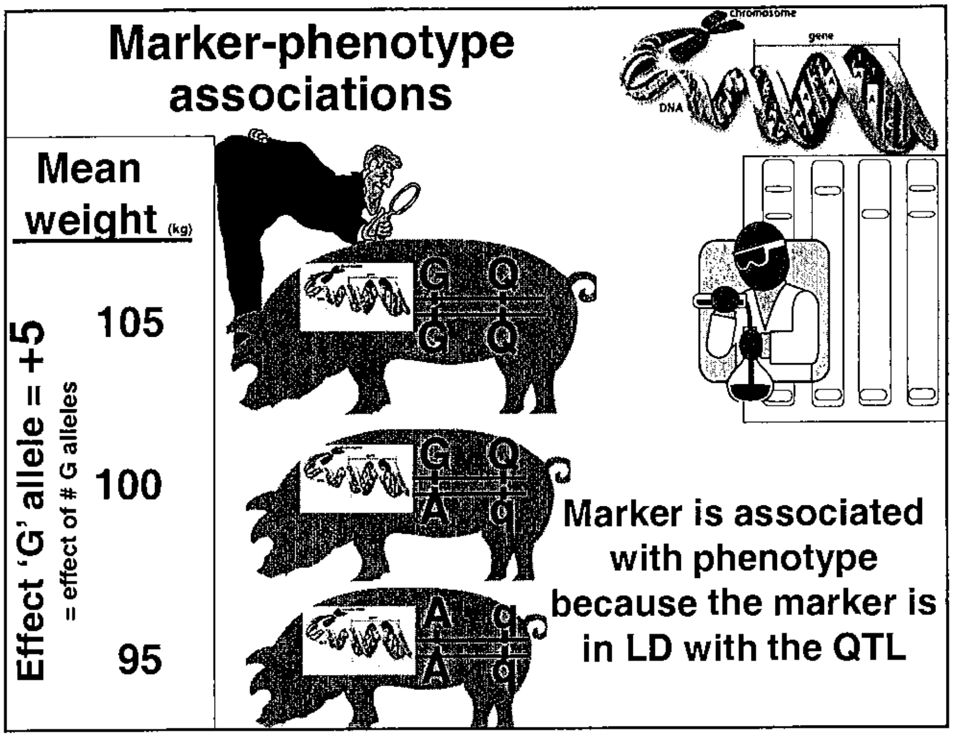
Animal Breeding & Genetics

Original Premise of Genomic Selection

(Meuwissen et al. 2001)

Although SNP panels contain few (if any) genotypes for the actual QTL, they are predictive because causative SNPs capture the effects of closely linked QTL through Linkage Disequilibrium between SNPs and QTL

i.e. associations between SNPs and phenotype result from the QTL being in LD with one or more SNPs



Marker-phenotype associations

Mean weight (kg)

105

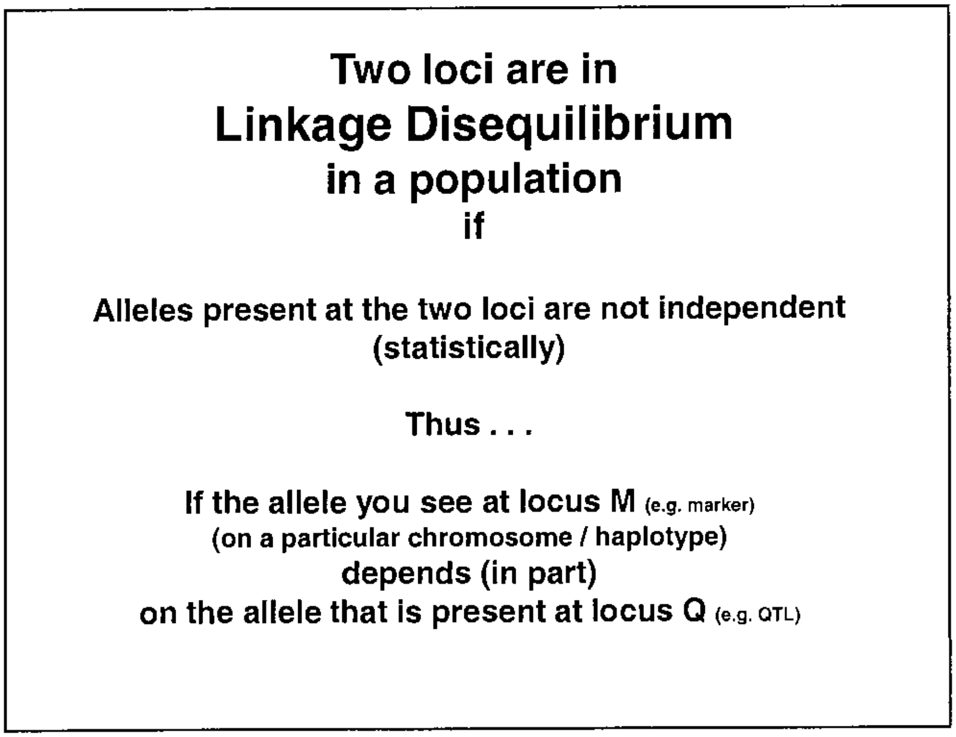
100

95

Effect 'G' allele = +5

= effect of # G alleles

Marker is associated with phenotype because the marker is in LD with the QTL



Two loci are in Linkage Disequilibrium in a population

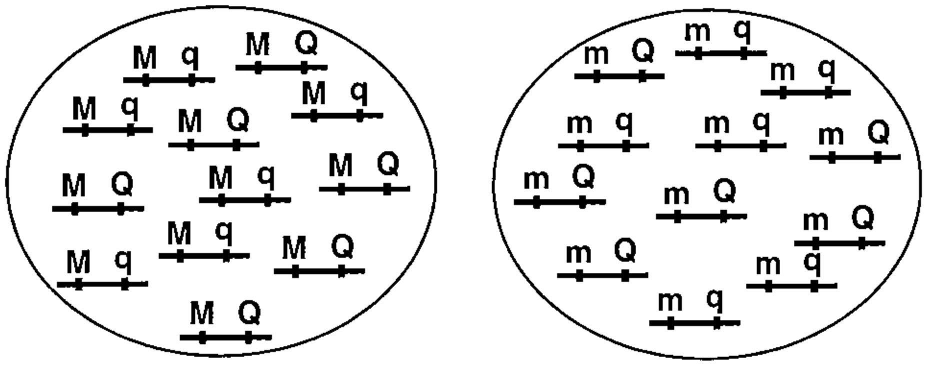
if

Alleles present at the two loci are not independent (statistically)

Thus . . .

If the allele you see at locus M (e.g. marker)
(on a particular chromosome / haplotype)
depends (in part)
on the allele that is present at locus Q (e.g. QTL)

Linkage Equilibrium (LE)

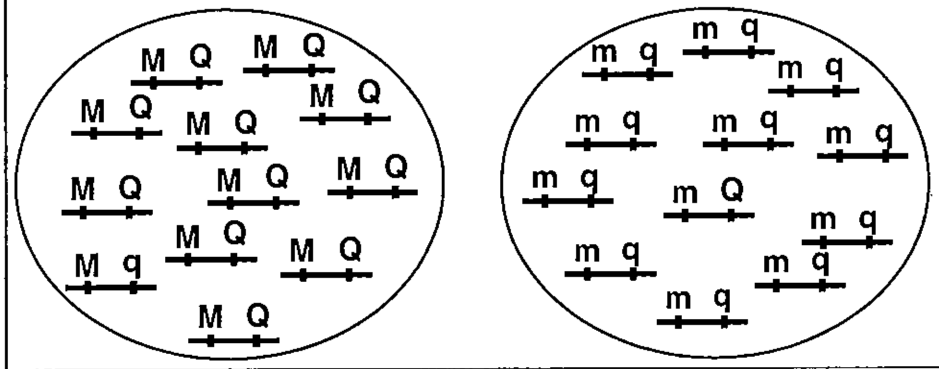


M is as often associated with Q
as m is associated with Q

$D = P_{MQ} - P_M P_Q = 0$

Marker genotype NOT related to phenotype

Linkage Disequilibrium (LD)



M is more often associated with Q
than m is associated with Q

$D = P_{MQ} - P_M P_Q \neq 0$

➔ Marker genotype IS related to phenotype
(if Q/q has effect on phenotype)

A useful measure of LD between two loci

r^2 = squared correlation between allele/genotype present at locus *M* and the allele/genotype present at locus *Q*

Indi-	Parental	Ordered genotypes		# 1 alleles	
vidual	origin	Locus M	Locus Q	Locus M	Locus Q
1	Paternal	0	0	0	0
	Maternal	0	0		
2	Paternal	1	0	2	1
	Maternal	1	1		
3	Paternal	1	1	1	1
	Maternal	0	0		
4	Paternal	0	0	1	0
	Maternal	1	0		
5	Paternal	1	1	2	1
	Maternal	1	0		
		Correl =0.53 r^2 =0.29		Correl =0.76 r^2 =0.58	

r^2 based on alleles and r^2 based on genotypes are expected to be equal if males and females are mated at random

r^2 based on genotypes is much easier to compute (doesn't require haplotyping)

Consider 1 SNP and a nearby single QTL

The SNP will have an association with phenotype iff the SNP is in LD with the QTL

The SNP effect depends on LD between the SNP and QTL

Phenotype = $y = \mu + g_{QTL} + e$ $\quad g_{QTL}$ = additive QTL effect

SNP association analysis: $y = \mu + \beta g_{SNP} + e$ $\quad g_{SNP}$ = 0/1/2 or -1/0/1

$$\beta = \mathrm{Cov}(y, g_{SNP})/\mathrm{Var}(g_{SNP}) = \mathrm{Cov}(g_{QTL}, g_{SNP})/\mathrm{Var}(g_{SNP})$$

$$= r\sqrt{\mathrm{Var}(g_{QTL})/\mathrm{Var}(g_{SNP})} = r\sqrt{\mathrm{Var}(g_{QTL})/2pq}$$

r = correlation between SNP and QTL = $\sqrt{LD}$

Amount of variance explained by the SNP:

$$\mathrm{Var}(\beta g_{SNP}) = \beta^2\mathrm{Var}(g_{SNP}) = [r^2\mathrm{Var}(g_{QTL})/\mathrm{Var}(g_{SNP})]\ \mathrm{Var}(g_{SNP})$$
$$= r^2\ \mathrm{Var}(g_{QTL})$$

→ The proportion of variance at the QTL that is explained (captured) by the SNP = r^2 = LD between SNP and QTL

Mechanisms that Generate and Erode LD

A variety of mechanisms generate linkage disequilibrium, and several of these can operate simultaneously. They can be separated into:

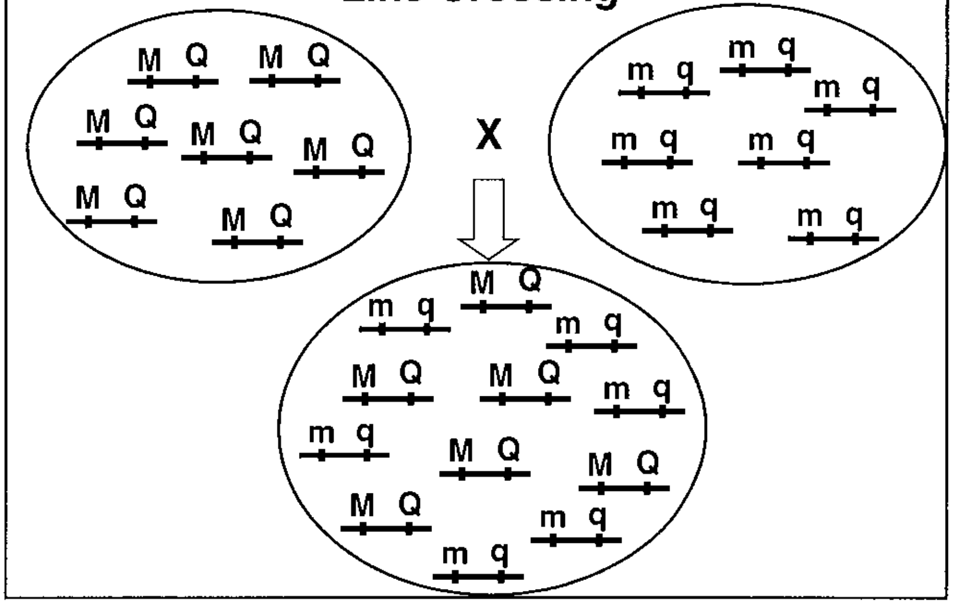
1. **Recurrent factors** – operate to create LD each generation
   a. ***Drift*** (inbreeding) in small populations – by chance or sampling, haplotypes passed on to the next generation are not in LE frequencies
   b. ***Recurrent migration*** – continuous mixing of populations in which haplotypes occur in different frequencies (e.g. Pr(A_1B_1)=1 for pop. 1 and =0 for pop. 2)
   c. ***Selection*** – certain haplotypes may be selected upon and increase in frequency
      – selection also creates LD between loci that are selected upon (= Bulmer effect – see later)
      – selection with epistasis (certain combinations of alleles are favorable) also creates LD between loci involved.
2. **Punctual factors** – operate only sporadically over time to create LD
   a. ***Mutation*** – occurs in a specific haplotype, which is then the only haplotype that contains that mutation, resulting it to be in LD with the mutation.
   b. ***One-time admixture/migration/crossing*** (e.g. producing F_1/F_2) – results in mixing populations with different haplotype frequencies
   c. ***Population bottleneck / founder effects*** – severe drift from 1-time sampling effects

Processes that create LD

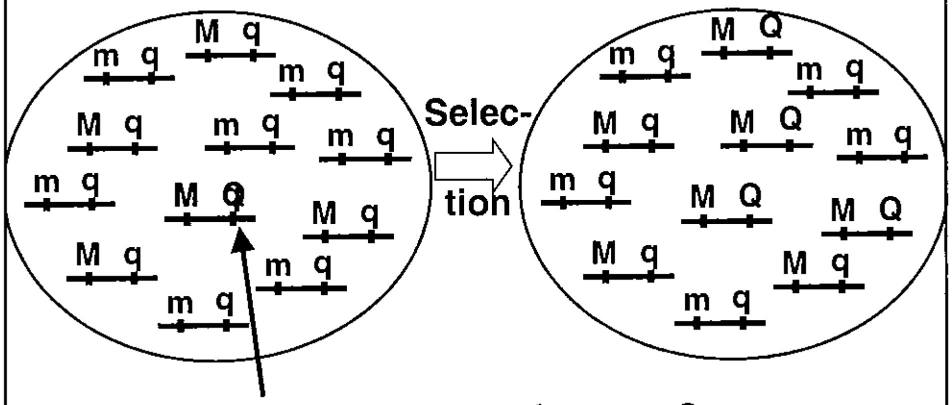
Inbred line I

Line Crossing

Inbred line II



Processes that create LD
Mutation and Selection



Allele on M chromosome mutates from q to Q
and then increases in frequency because of
- random drift
- or selection on Q → selective sweep = LD block around Q

Selective Sweep

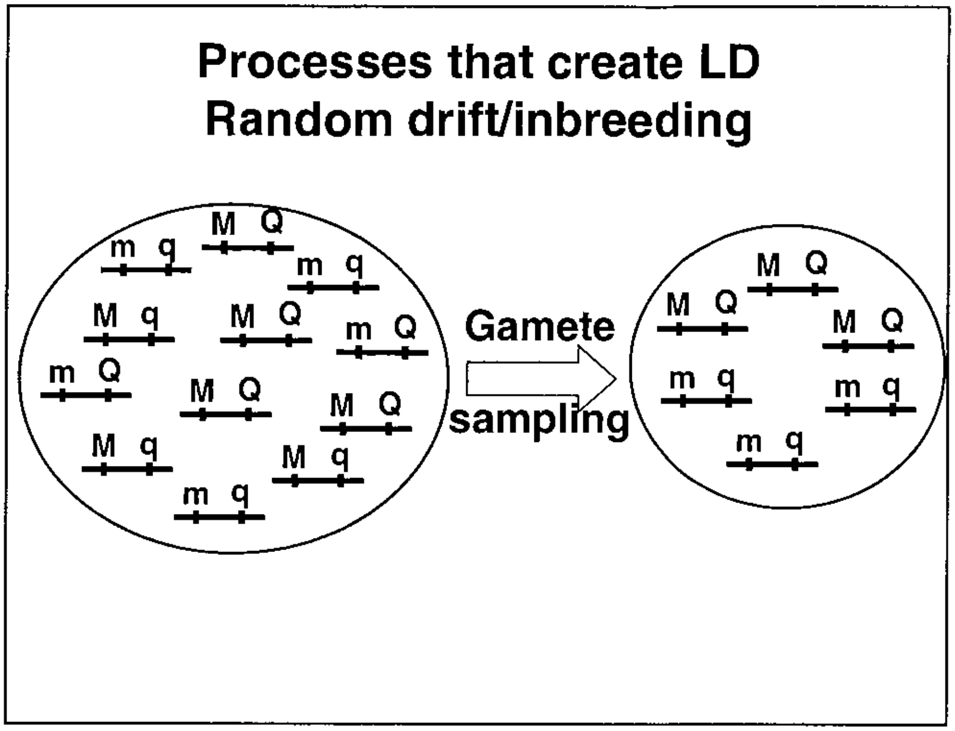
Original mutation (q → Q) occurred in marker haplotype:

001110010Q01001110110

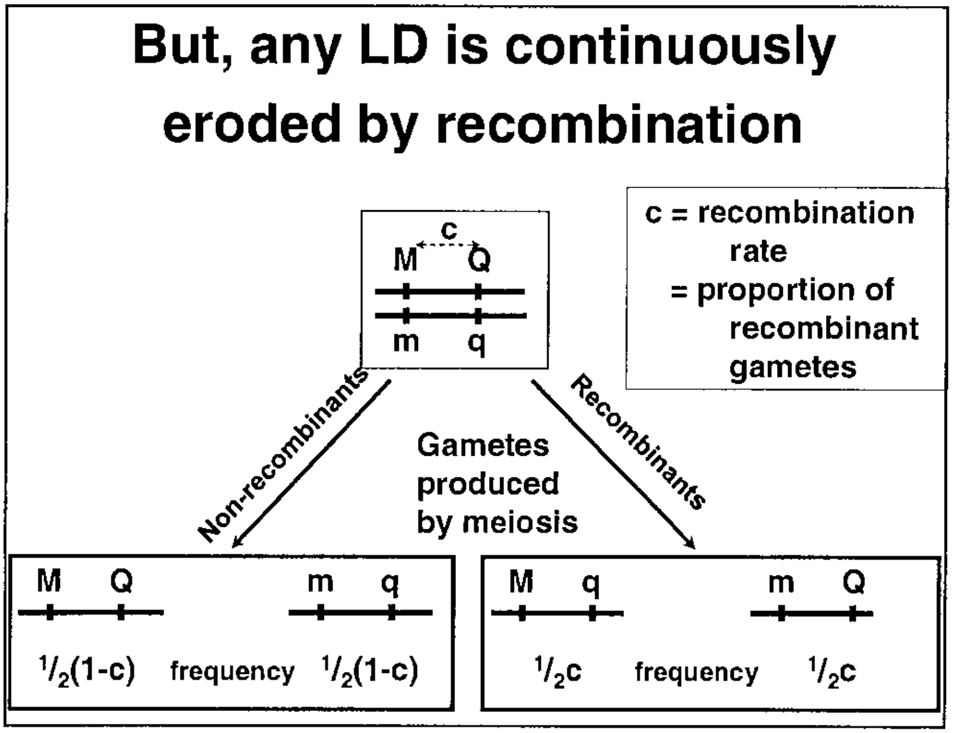
Many generations ⇩ of recombination

1001100 10Q01 100110100
0111100 10Q01 001011010
0010011 10Q01 000010111
0011101 10Q01 101101110
0110100 10Q01 001100010
0001100 10Q01 001000111
1110100 10Q01 011101111
0101100 10Q01 001101010

Unique haplotype associated with Q ⇦



Processes that create LD Random drift/inbreeding



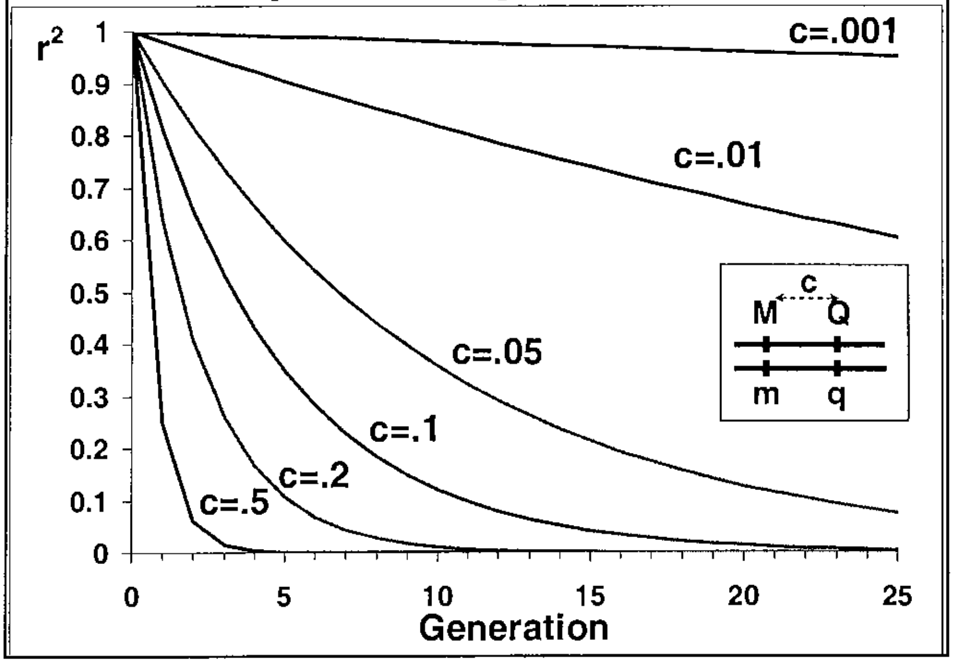
But, any LD is continuously eroded by recombination

c = recombination rate
= proportion of recombinant gametes

Gametes produced by meiosis

c
M Q
m q
Non-recombinants
Recombinants
M Q
m q
$^1/_2(1-c)$ frequency $^1/_2(1-c)$
M q
m Q
$^1/_2c$ frequency $^1/_2c$

Break-up of LD by recombination

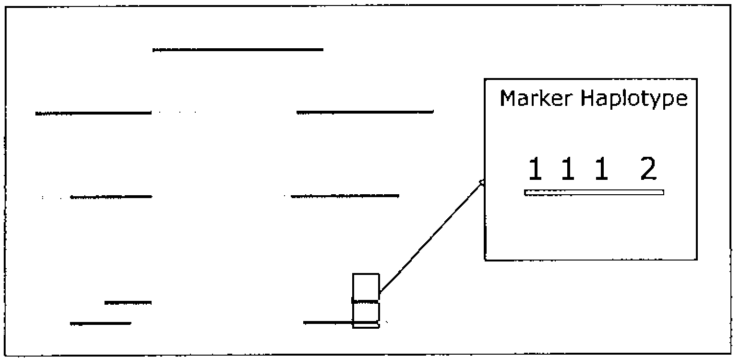


$(1-c)$ [illegible]

[illegible]

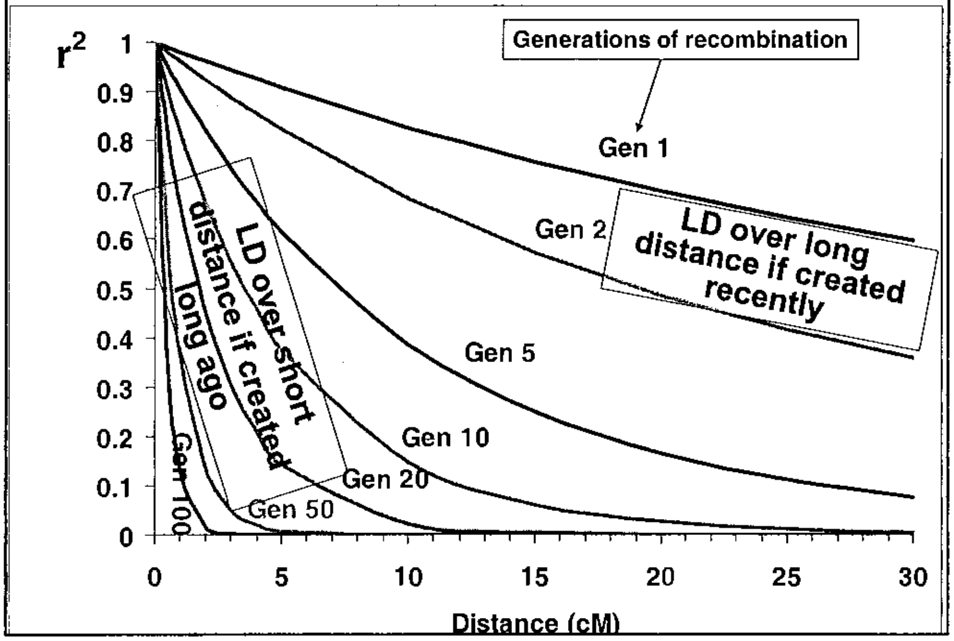
Another way of looking at LD

Conservation of chunks of ancestral chromosomes



Size of conserved chunks depends on how long ago LD was created – longer if N_e larger

Historic LD expected only over short distances

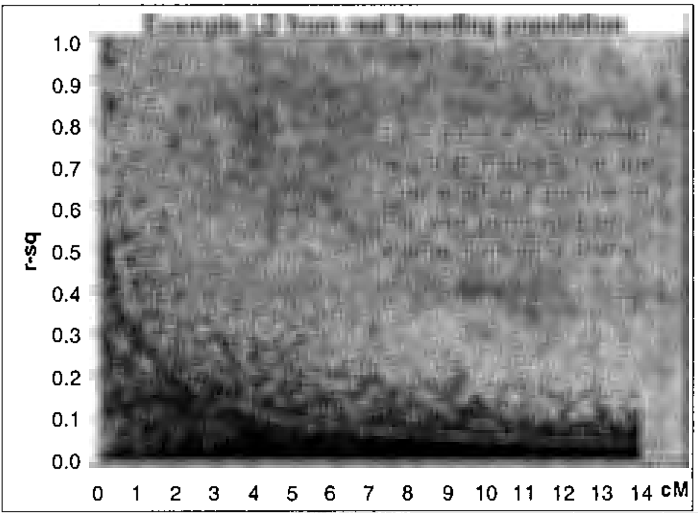


Balance between Drift and Recombination

In *small(er) closed populations*

- LD is continuously created by drift – more with smaller effective pop. size, N_e
- LD is continuously eroded by recombination – faster at longer distances

This results in a balance/equilibrium of average LD at a given distance: $E(r^2_{\infty,c}) = \frac{1}{1+4N_e c}$ (Sved 1971)



LD is very variable

LD at short distances is often lower than expected based on a given effective population size (= yellow line)

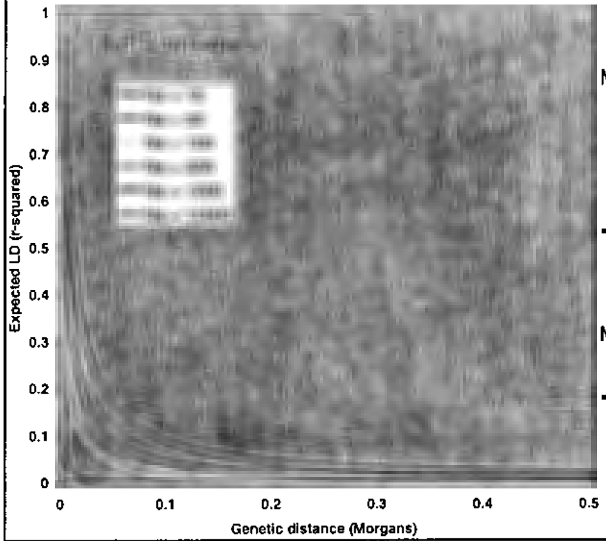
Because LD reflects historical effective population size and this has not been constant

Balance between Drift and Recombination

In *small(er) closed populations*

- LD is continuously created by drift – more with smaller effective pop. size, N_e
- LD is continuously eroded by recombination – faster at longer distances

This results in a balance/equilibrium of average LD at a given distance: $E(r^2_{\infty,c}) = \frac{1}{1+4N_e c}$ (Sved 1971)



Most outbred domesticated plant and animal populations have small(er) (historical) effective population size and drift-recombination balance is expected to be the main contributor to LD

→ LD is expected to be sizeable at short distances, but small at longer distances.

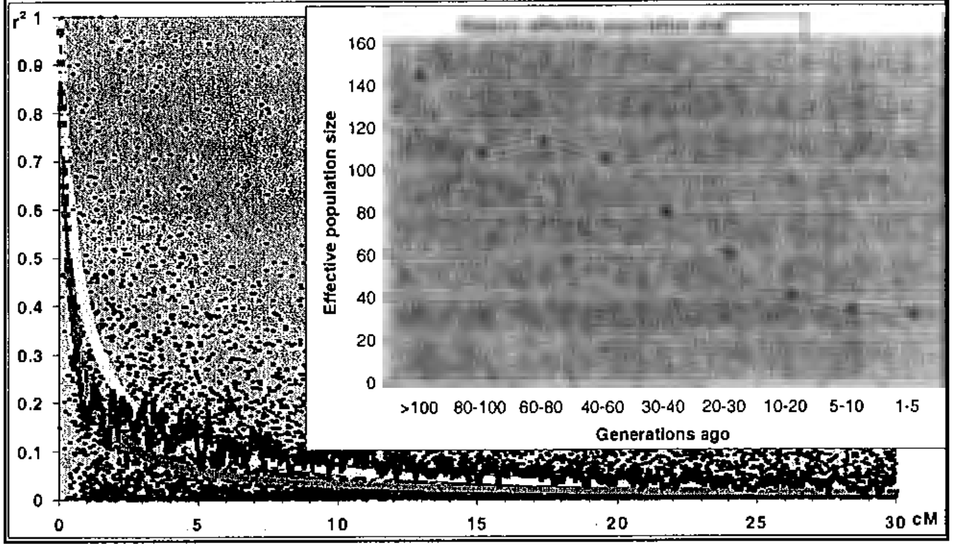
Most human populations have large (historical) effective size

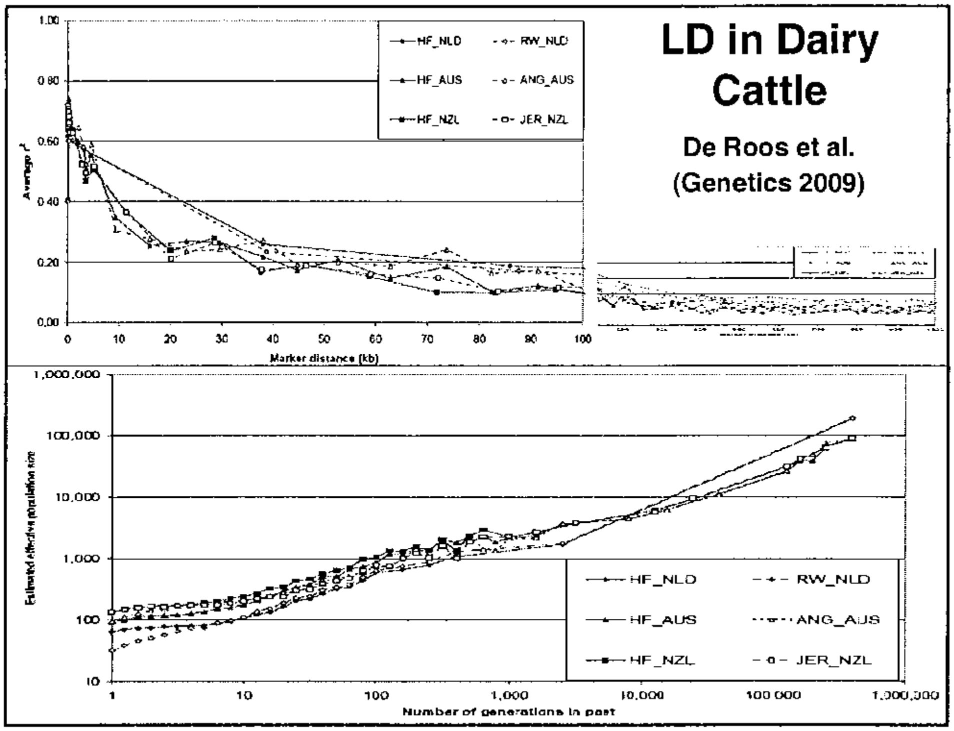
→ $E(r^2) = \frac{1}{1+4N_e c}$

→ LD is smaller at given distance.

Estimating historical N_e from average LD at a given distance

$$E(r_c^2) = \frac{1}{1+4N_{e,t}\,c} \quad ==> \quad \hat{N}_{e,t} = \frac{1}{4c}\left(\frac{1}{\hat{r}_c^2} - 1\right)$$

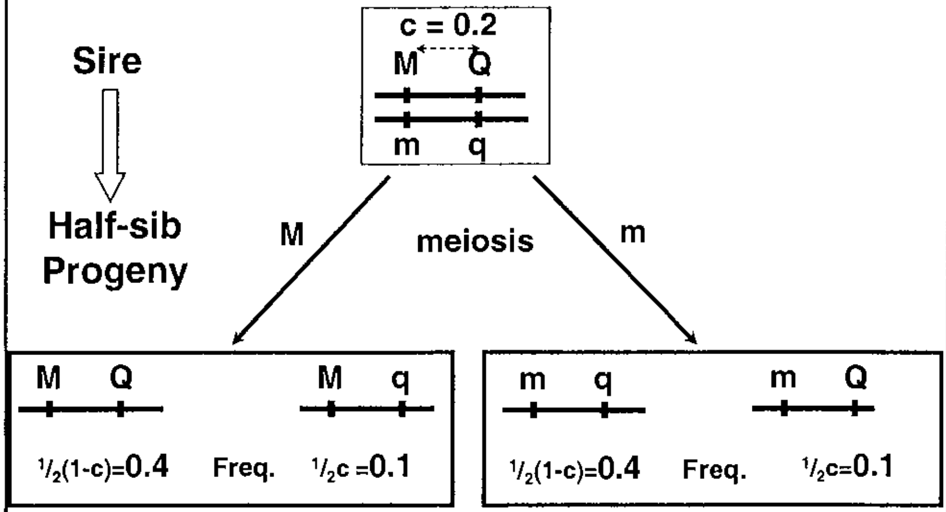




LD in Dairy Cattle

De Roos et al.
(Genetics 2009)

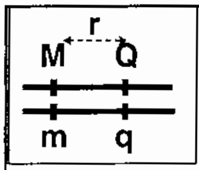
But: LD always exists WITHIN families



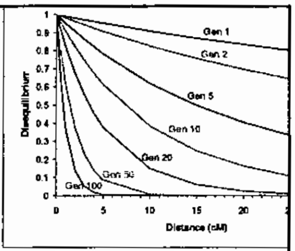
$^1/_2(1\text{-}c)$=0.4 Freq. $^1/_2c$ =0.1

$^1/_2(1\text{-}c)$=0.4 Freq. $^1/_2c$=0.1

And this LD extends over long distances - only 1 round of recombination



Implications for QTL detection and MAS



- In closed breeding populations
  - Population-wide LD only over short distances (< 2 cM)
    - ➢ need many markers or carefully placed markers (candidate genes) to detect QTL
    - ➢ Positive markers expected to be close to QTL

- Recent crosses
- Within-families

} - LD over long distances (20 cM)

- ➢ Fewer markers needed to detect QTL
- ➢ Positive markers may be far from QTL

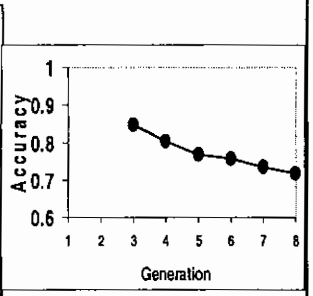
Accuracy of EBV from Genomic selection
Does it result from historic SNP-QTL LD?

Meuwissen et al. (2001 Genetics: 1819)

$N_e = 100$

Estimates from 2200 individuals

GEBV accuracy	Marker distance
0.85	1 cM
0.81	2 cM
0.75	4 cM



And does the decline of accuracy over generations result from erosion of LD by recombination?

Impact of historic LD on accuracy of GEBV A Simulation study – Habier, Dekkers, Fernando (unpublished)

- 8 chromosomes
- 200 QTL/chromosome
- Heritability 0.5 for female phenotypes, 0.8 for male phenotypes
- No historic LD, only LD from the pedigree

D. Habier

Simulations – without historic LD

Generation 0	Population in equilibrium N=500
Real pedigree (13 generations)	1500 males + 1500 females (Matings: 50 sires + 500 dams)

4 training generations start

Gen 1 Training

[illegible]

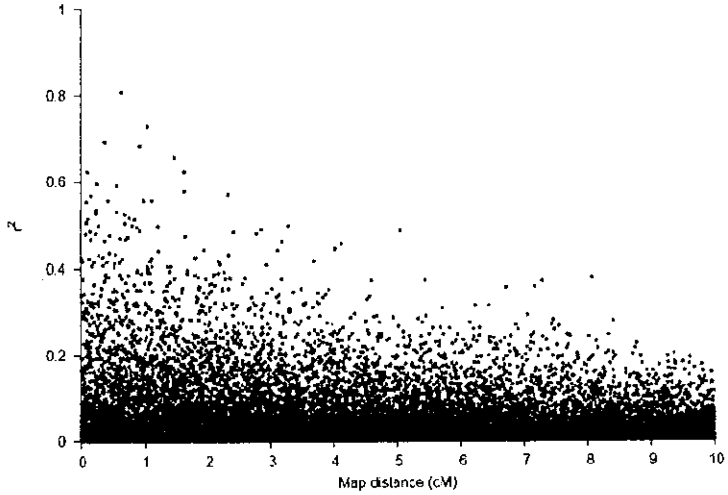
50 sires 500 dams

Gen 2-4 [illegible]

D. Habier

Linkage disequilibrium

No Historic LD – Real pedigree



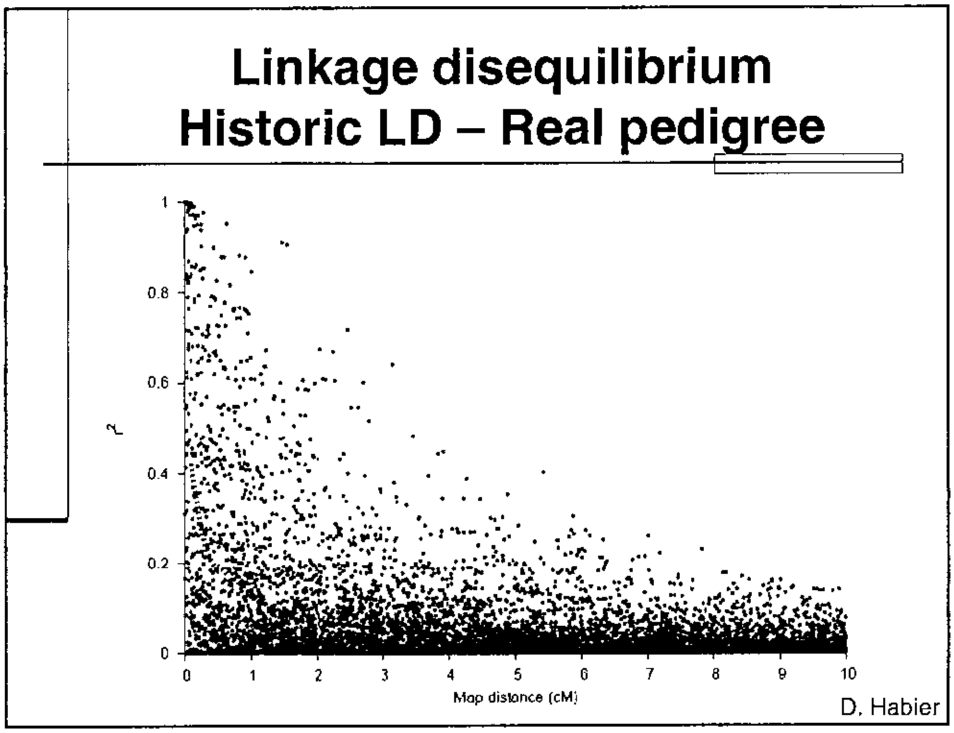
D. Habier

Simulations – WITH Historic LD

Generation -1050	Random mating (N=500)
Generation - 50	Random mating (N=100)
Real pedigree (13 generations)	1500 males + 1500 females (50 sires + 500 dams)

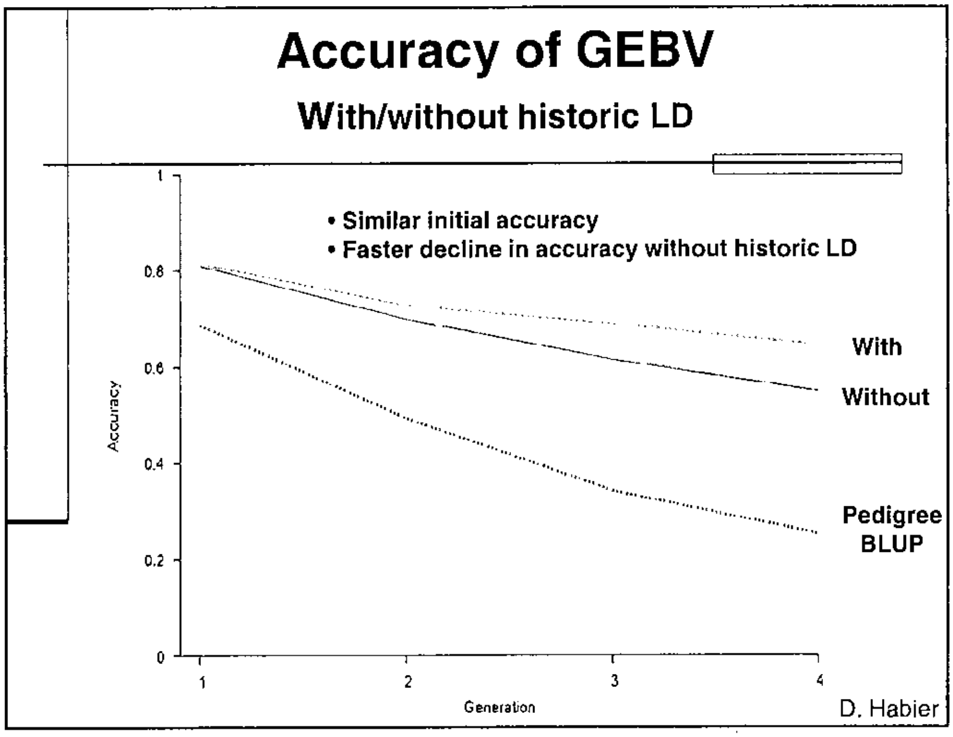
4 training generations start

D. Habier



Linkage disequilibrium Historic LD – Real pedigree

D. Habier



Accuracy of GEBV

With/without historic LD

- Similar initial accuracy
- Faster decline in accuracy without historic LD

D. Habier

Factors that contribute to accuracy of Genomic Selection EBV

- Historic marker-QTL LD – the original premise of GS
- Pedigree relationships captured by markers
    - Does not require marker-QTL LD or linkage

  Habier, D. et al. Genetics 2007;177:2389-2397
- Deviations from pedigree relationships

  (genomic vs. pedigree relationships)
    - Requires marker-QTL LD or linkage to be useful
- Population structure / stratification

  (e.g. mixed populations, unbalanced family structure)
- Within-family linkage / cosegregation information
    - Cosegregation between markers and linked QTL from parents to progeny

More important for accuracy in short term
(generations immediately following training)

Pedigree vs. Genomic Relationships

Sire, dam, and 4 full sibs (from Dorian Part I)

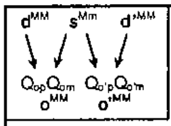
A matrix

$$\begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .5 \\ 0 & 1 & .5 & .5 & .5 & .5 \\ .5 & .5 & 1 & .5 & .5 & .5 \\ .5 & .5 & .5 & 1 & .5 & .5 \\ .5 & .5 & .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & .5 & .5 & 1 \end{bmatrix}$$

G matrix

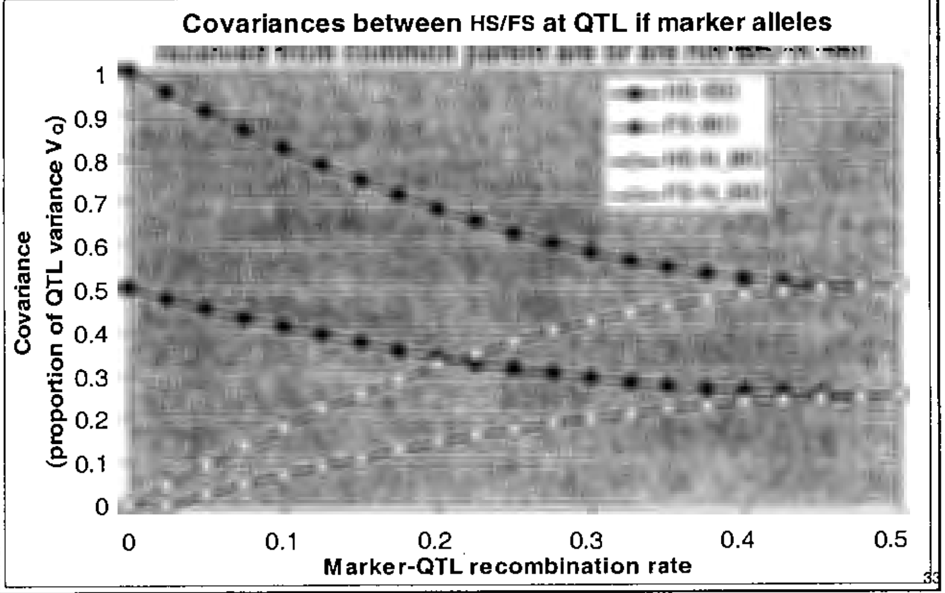
$$\begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .5 \\ 0 & 1 & .5 & .5 & .5 & .5 \\ .5 & .5 & 1 & .6 & .4 & .4 \\ .5 & .5 & .6 & 1 & .4 & .4 \\ .5 & .5 & .4 & .4 & 1 & .6 \\ .5 & .5 & .4 & .4 & .6 & 1 \end{bmatrix}$$

Deviations of genomic from pedigree relationships are predictive only if SNPs are in LD or linked to QTL



COVARIANCE BETWEEN RELATIVES BASED ON A LINKED MARKER

Example for paternal half-sibs and for full-sibs based on genotypes at a marker linked to a QTL with recomb. rate *c*.



Proportion of alleles shared by fullsib pair	Pro-ba-bility
1 locus	
0/2 = 0	¼
1/2 = 0.25	½
2/2 = 1	¼
Average	0.5
St.Dev.	0.35
2 loci	
0/4 = 0	$^1/_8$
1/4 = 0.25	¼
2/4 = 0.50	$^3/_8$
3/4 = 0.75	¼
4/4 = 1	$^1/_8$
Average	0.5
St.Dev.	0.25

Distribution of the proportion of alleles shared by Sibs

Based on Van Raden 2007, Interbull

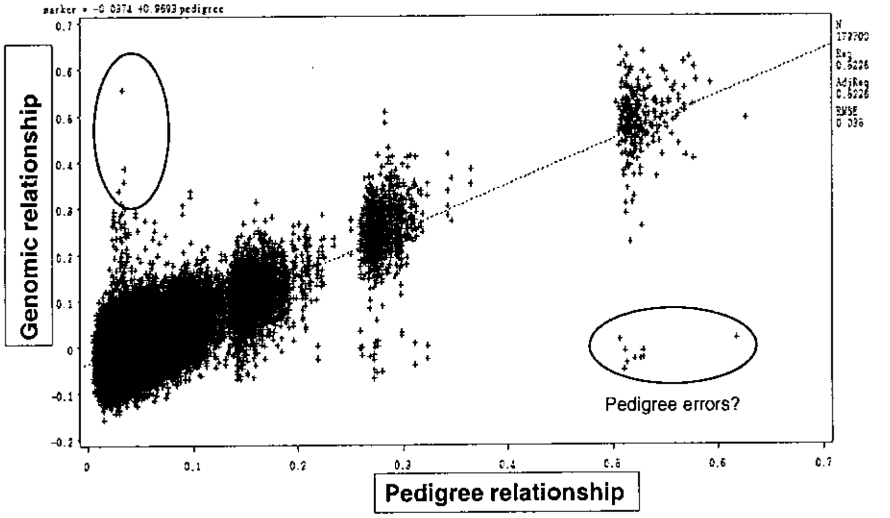
	# Loci	Percentage of alleles shared			
		Full sibs		Half sibs	
		Mean	SD	Mean	SD
Unlinked loci	1	50	35.4	25	17.7
	5	50	15.8	25	7.9
	10	50	11.2	25	5.6
	50	50	5.0	25	2.5
	100	50	3.5	25	1.8
	Infinite	50	0.0	25	0.0
Linked loci		50	≥3.5	25	≥1.8

Genomic relationships capture some of the Mendelian sampling terms if the SNPs are linked to QTL

Note that a parent and offspring always share exactly 50% of their alleles

Genomic vs. pedigree relationships in real data

Wolc and Dekkers, unpublished



The impact of genetic relationships on genome-assisted breeding values in German Holstein cattle

David Habier, J. Tetens, F. Seefried, P. Lichtner, G. Thaller

Institute of Animal Breeding and Husbandry, Christian-Albrechts University of Kiel

GSE 2010 42:5

- 3,863 progeny-tested German Holstein bulls
- Genotyped for 54,001 SNPs
- Traits: Milk, fat and protein yield, somatic cell score
- Family structure: Half- and full sib families, fathers and sons

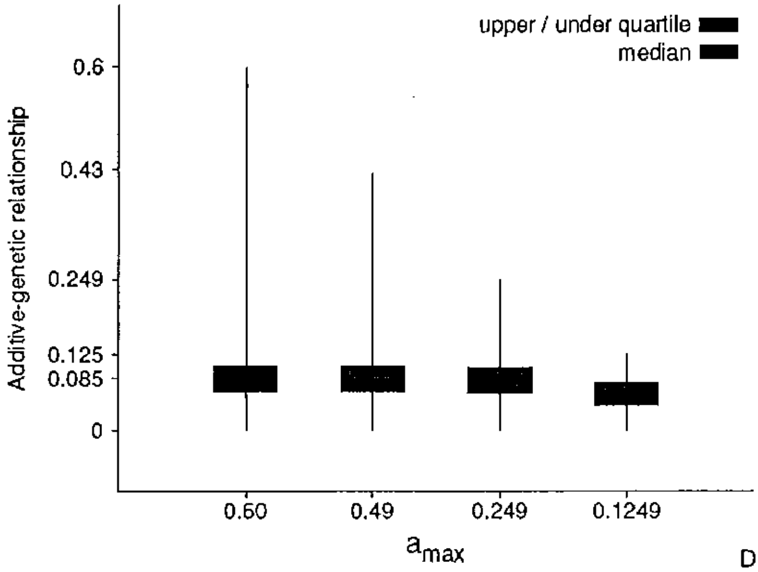
- Sampling of bulls into training and validation
- Excluding bulls that cause to exceed a_{max}
- Training size: 2,084 and 1,042 bulls
- Validation size: 490 bulls

Controled the maximum additive-genetic relationship (a_{max}) between bulls in training and validation

a_{max}	Close relatives in training
0.6	Fathers, full sibs, half sibs
0.49	Half sibs
0.249	-
0.1249	-

D. Habier

Additive-genetic relationships between bulls in training and validation



D. Habier

Estimation of GEBV

BayesB (Meuwissen et al., 2001)

- 1% of available SNPs are fitted

G-BLUP

- Genomic relationship matrix (uses all SNPs)

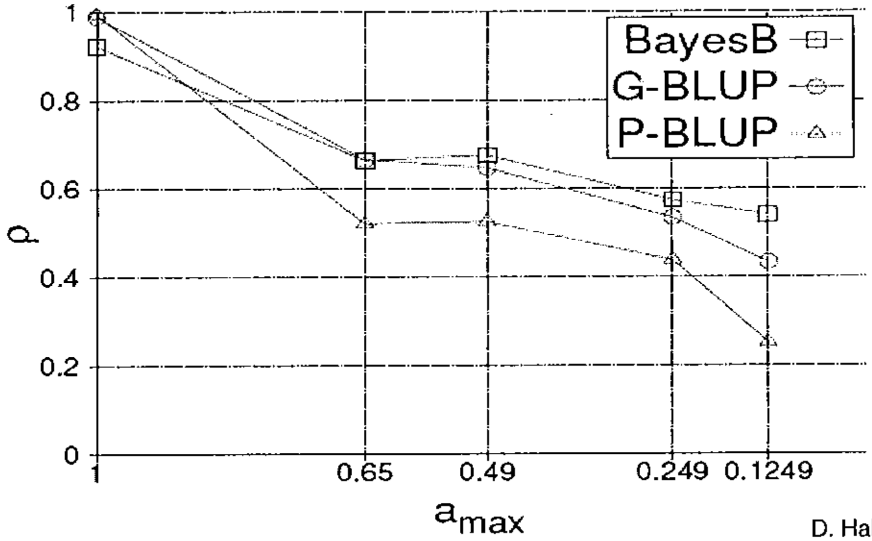
P-BLUP

- Numerator-relationship matrix

D. Habier

Accuracy of GEBV against a_{max} between training and validation populations

Milk yield



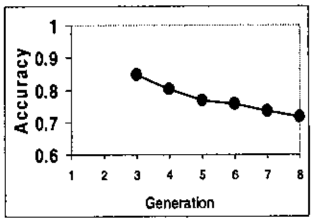
D. Habier

Conclusions

- Genetic relationships (between individuals used for training and validation / prediction) can contribute substantially to the accuracy of GEBV

Implications

- Accuracy of GEBV will be lower for individuals that are not well connected to the training data.
- Part of the decline in the accuracy of GEBV over generations results from declining genetic relationships with the training data.



- ➔Ongoing phenotyping and re-training will be needed to maintain accuracies of GEBV
- How accurate are GEBV when used across breeds? See later

Motivation
The problem of predicting genetic merit

What's wrong with what we do now?

The Prediction Problem

[illegible]

Model Equation

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

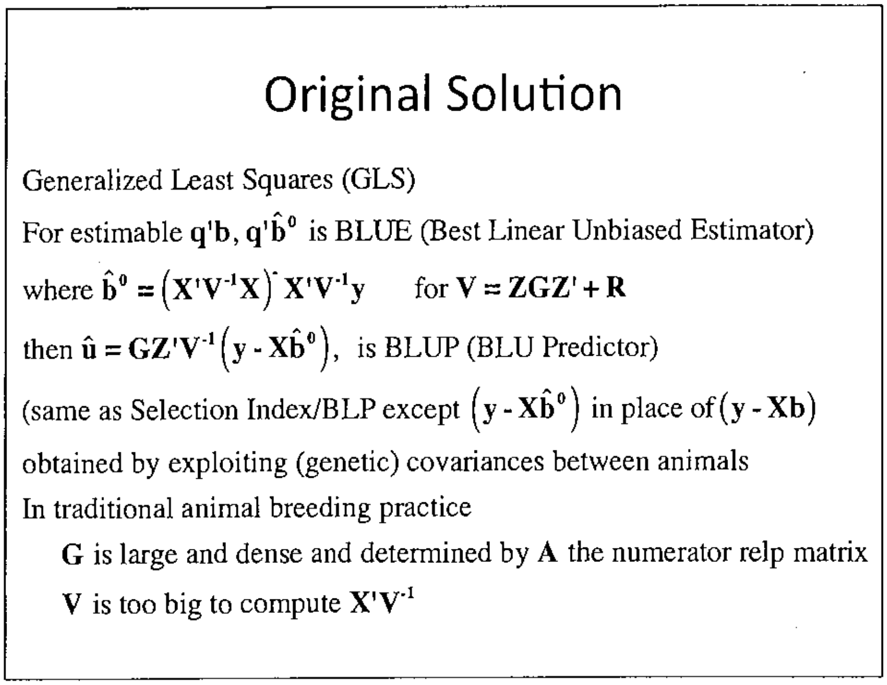
Other aspects of the model

First moments $\mathbf{E[u] = 0, E[e] = 0}$, therefore $\mathbf{E[y] = Xb}$

Second moments $\mathbf{var[u] = G, var[e] = R, cov[u,e'] = 0}$

Distributional Assumptions e.g. $\mathbf{u, e} \sim \text{MVN}$

Want to predict $\mathbf{u}$ or linear functions like $\mathbf{k'u}$



Original Solution

Generalized Least Squares (GLS)

For estimable $\mathbf{q'b}$, $\mathbf{q'\hat{b}^0}$ is BLUE (Best Linear Unbiased Estimator)

where $\hat{\mathbf{b}}^0 = \left(\mathbf{X'V^{-1}X}\right)^{-}\mathbf{X'V^{-1}y}$ for $\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$

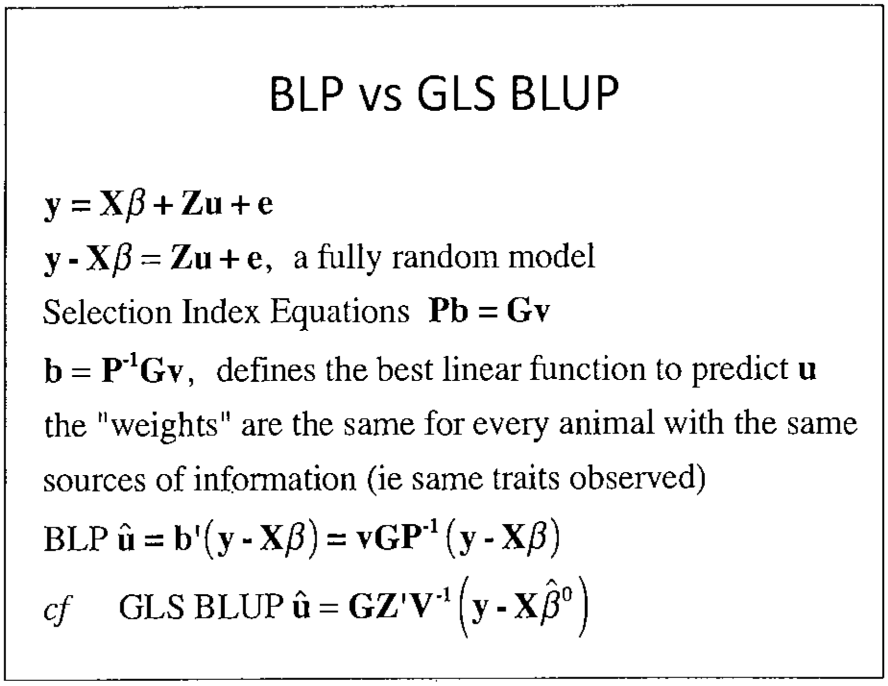
then $\hat{\mathbf{u}} = \mathbf{GZ'V^{-1}}\left(\mathbf{y} - \mathbf{X\hat{b}^0}\right)$, is BLUP (BLU Predictor)

(same as Selection Index/BLP except $\left(\mathbf{y} - \mathbf{X\hat{b}^0}\right)$ in place of $\left(\mathbf{y} - \mathbf{Xb}\right)$

obtained by exploiting (genetic) covariances between animals

In traditional animal breeding practice

- $\mathbf{G}$ is large and dense and determined by $\mathbf{A}$ the numerator relp matrix
- $\mathbf{V}$ is too big to compute $\mathbf{X'V^{-1}}$



BLP vs GLS BLUP

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e}$

$\mathbf{y} - \mathbf{X}\beta = \mathbf{Zu} + \mathbf{e}$, a fully random model

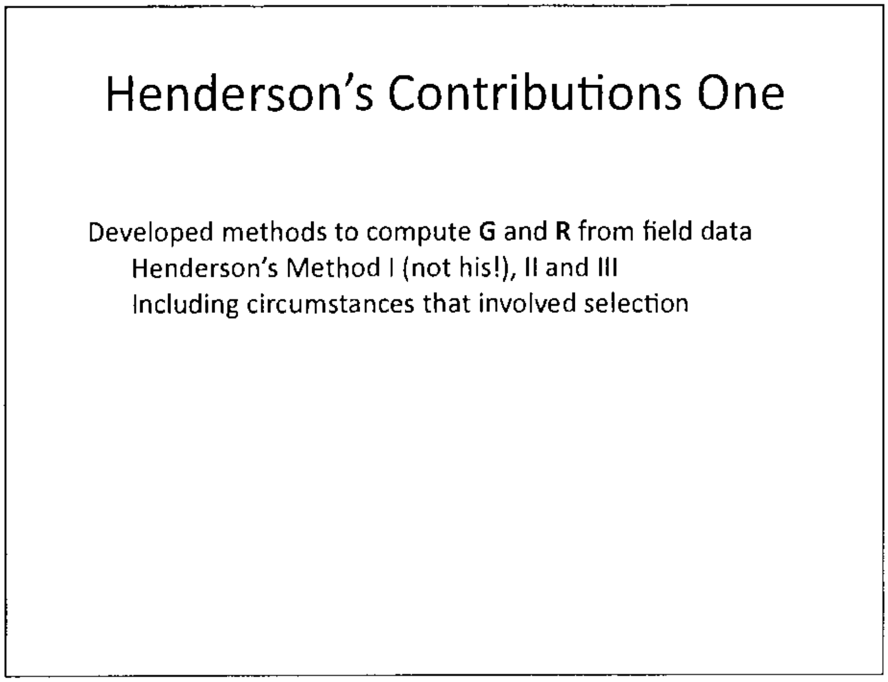
Selection Index Equations $\mathbf{Pb} = \mathbf{Gv}$

$\mathbf{b} = \mathbf{P^{-1}Gv}$, defines the best linear function to predict $\mathbf{u}$

the "weights" are the same for every animal with the same sources of information (ie same traits observed)

BLP $\hat{\mathbf{u}} = \mathbf{b}'\left(\mathbf{y} - \mathbf{X}\beta\right) = \mathbf{vGP^{-1}}\left(\mathbf{y} - \mathbf{X}\beta\right)$

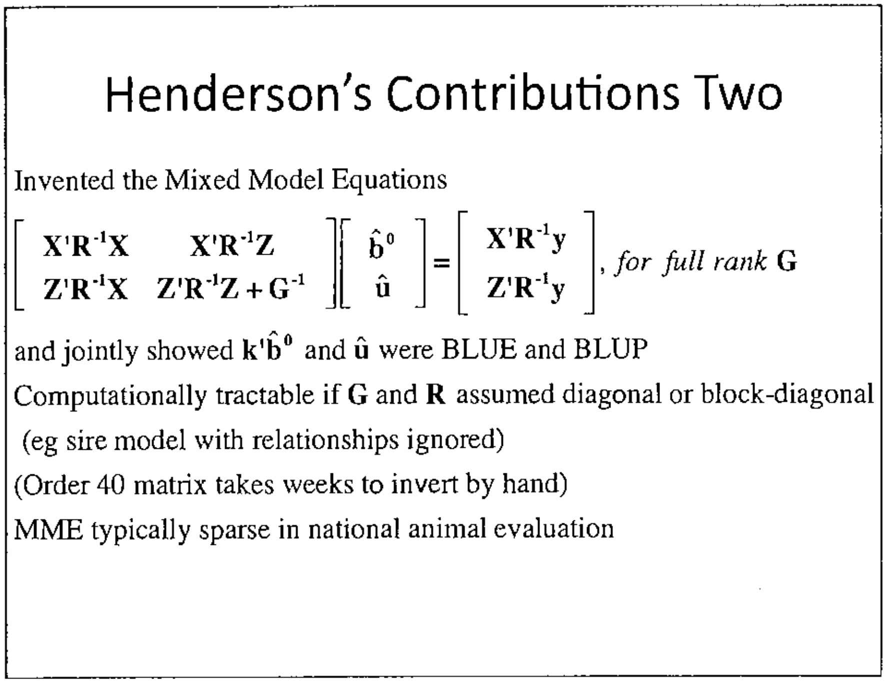
cf GLS BLUP $\hat{\mathbf{u}} = \mathbf{GZ'V^{-1}}\left(\mathbf{y} - \mathbf{X}\hat{\beta}^0\right)$



Henderson's Contributions One

Developed methods to compute **G** and **R** from field data

- Henderson's Method I (not his!), II and III
- Including circumstances that involved selection



Henderson's Contributions Two

Invented the Mixed Model Equations

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}^0 \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}, \textit{for full rank } \mathbf{G}$$

and jointly showed $\mathbf{k'}\hat{\mathbf{b}}^{\mathbf{0}}$ and $\hat{\mathbf{u}}$ were BLUE and BLUP

Computationally tractable if **G** and **R** assumed diagonal or block-diagonal

(eg sire model with relationships ignored)

(Order 40 matrix takes weeks to invert by hand)

MME typically sparse in national animal evaluation

Henderson's Contributions Three

Invented an algorithm to directly form $\mathbf{A}^{-1}$ from a pedigree list

Then $\mathbf{G}^{-1}$ can be formed as a scalar product or kronecker product

define d to be "mendelian" sampling variance

d=(1, 3/4, 1/2) for 0, 1 or 2 parents known

define s'=(-1/2, -1/2, 1) to represent sire (if known), dam (if known) and individual equations

accumulate $\mathbf{s}d^{-1}\mathbf{s}'$ in the sire, dam and individual rows/columns for every trio of animals in the pedigree list

Consequence of $\mathbf{A}^{-1}$ structure

Accumulate for each animal
$$\begin{array}{cc} & \begin{array}{ccc} \textit{sire} & \textit{dam} & i \end{array} \\ \begin{array}{c} \textit{sire} \\ \textit{dam} \\ i \end{array} & \left[\begin{array}{ccc} 0.25 & 0.25 & -0.5 \\ 0.25 & 0.25 & -0.5 \\ -0.5 & -0.5 & 1 \end{array}\right] d^{-1} \end{array}$$

When both parents are known

Nonparents (ie terminal offspring)

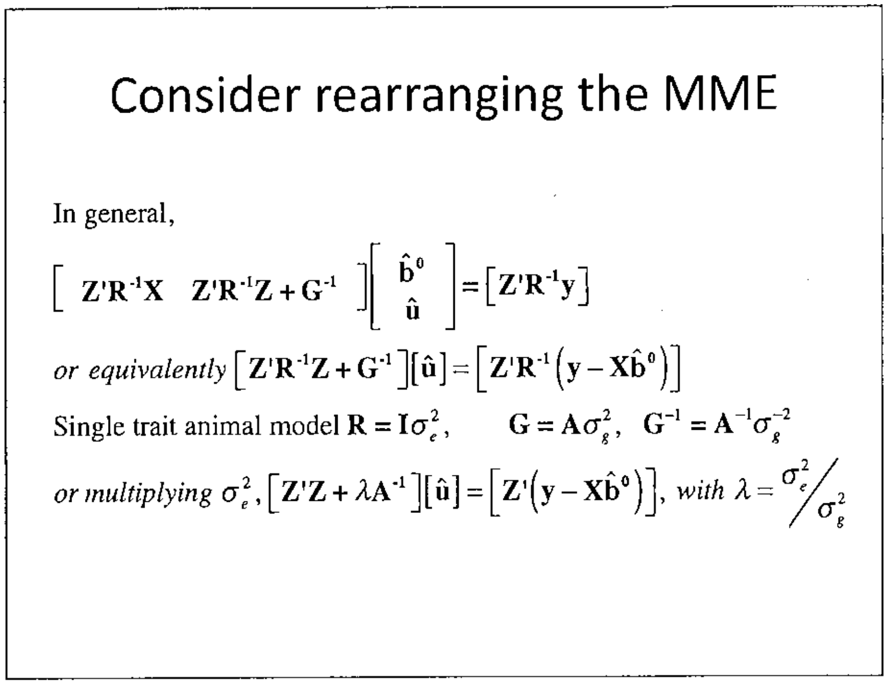
Own equation (ie row) has 2 on diagonal, -1 in sire column -1 in dam column

Parent with one offspring

Own equation has 2+1/2 on diagonal, -1 in sire and dam columns in addition to -1/2 in the column of its mate, -1 in column of offspring

Parent with many offspring to different mates

accumulates a large diagonal element, many small negative offdiagonals



Consider rearranging the MME

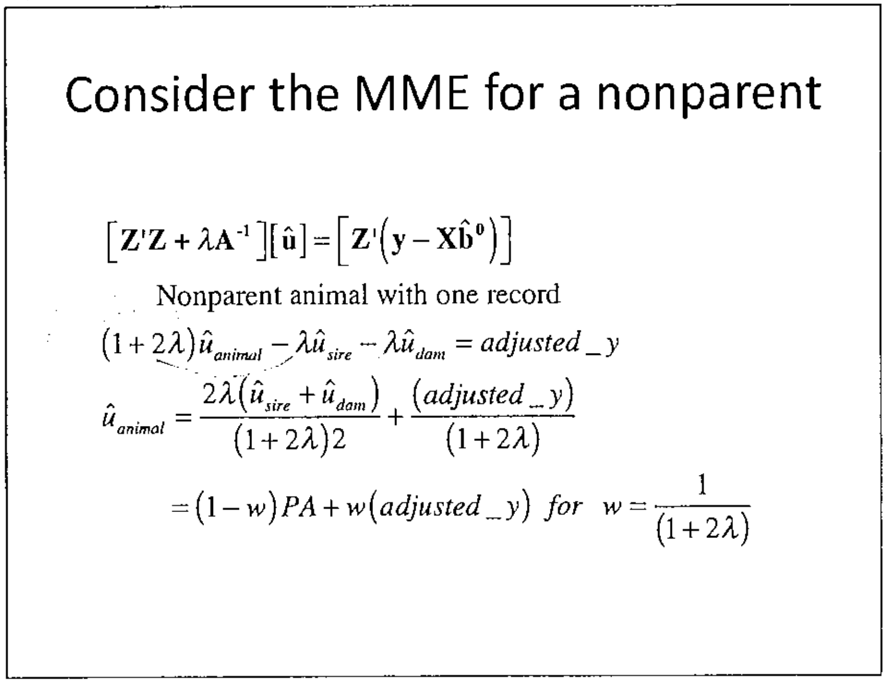
In general,

$$\begin{bmatrix} \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix}\begin{bmatrix} \mathbf{\hat{b}^0} \\ \mathbf{\hat{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z'R^{-1}y} \end{bmatrix}$$

or equivalently $\left[\mathbf{Z'R^{-1}Z+G^{-1}}\right]\left[\mathbf{\hat{u}}\right] = \left[\mathbf{Z'R^{-1}}\left(\mathbf{y}-\mathbf{X\hat{b}^0}\right)\right]$

Single trait animal model $\mathbf{R} = \mathbf{I}\sigma_e^2$, $\mathbf{G} = \mathbf{A}\sigma_g^2$, $\mathbf{G}^{-1} = \mathbf{A}^{-1}\sigma_g^{-2}$

or multiplying σ_e^2, $\left[\mathbf{Z'Z}+\lambda\mathbf{A}^{-1}\right]\left[\mathbf{\hat{u}}\right] = \left[\mathbf{Z'}\left(\mathbf{y}-\mathbf{X\hat{b}^0}\right)\right]$, *with* $\lambda = \sigma_e^2 / \sigma_g^2$



Consider the MME for a nonparent

$$\left[\mathbf{Z'Z}+\lambda\mathbf{A}^{-1}\right]\left[\mathbf{\hat{u}}\right] = \left[\mathbf{Z'}\left(\mathbf{y}-\mathbf{X\hat{b}^0}\right)\right]$$

Nonparent animal with one record

$$(1+2\lambda)\hat{u}_{animal} - \lambda\hat{u}_{sire} - \lambda\hat{u}_{dam} = adjusted\_y$$

$$\hat{u}_{animal} = \frac{2\lambda\left(\hat{u}_{sire}+\hat{u}_{dam}\right)}{(1+2\lambda)2} + \frac{(adjusted\_y)}{(1+2\lambda)}$$

$$= (1-w)PA + w(adjusted\_y) \;\; for \;\; w = \frac{1}{(1+2\lambda)}$$

$2\lambda^{-1}$

Consider the MME for a nonparent

$\hat{u}_{animal} = (1-w)PA + w(adjusted\_y) \;\; for \;\; w = \frac{1}{(1+2\lambda)}$

$\lambda = \frac{1-h^2}{h^2} \; so \; for \; h^2 = 1, \; \lambda = 0, w = 1, \; (no \;\; shrinkage)$

$for \, h^2 = low, \;\; \lambda = big, \;\; w = small, \; (shrink \; the \; deviation)$

Two sources of BV information are pooled

The parent average PA

The individual prediction (shrunk deviation)

with heritability influencing shrinkage

Consider the MME for a nonparent

$$\left[\mathbf{Z'Z} + \lambda\mathbf{A}^{-1}\right]\left[\hat{\mathbf{u}}\right] = \left[\mathbf{Z'}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^{0}\right)\right]$$

Nonparent animal with one record

$$\hat{u}_{animal} \;\; = (1-w)PA + w(adjusted\_y)$$

Nonparent animal with no record

$$2\lambda\hat{u}_{animal} - \lambda\hat{u}_{sire} - \lambda\hat{u}_{dam} = 0$$

$$\hat{u}_{animal} = \frac{\lambda\left(\hat{u}_{sire} + \hat{u}_{dam}\right)}{\lambda 2} = \frac{\left(\hat{u}_{sire} + \hat{u}_{dam}\right)}{2} = PA$$

Reliability of nonparents

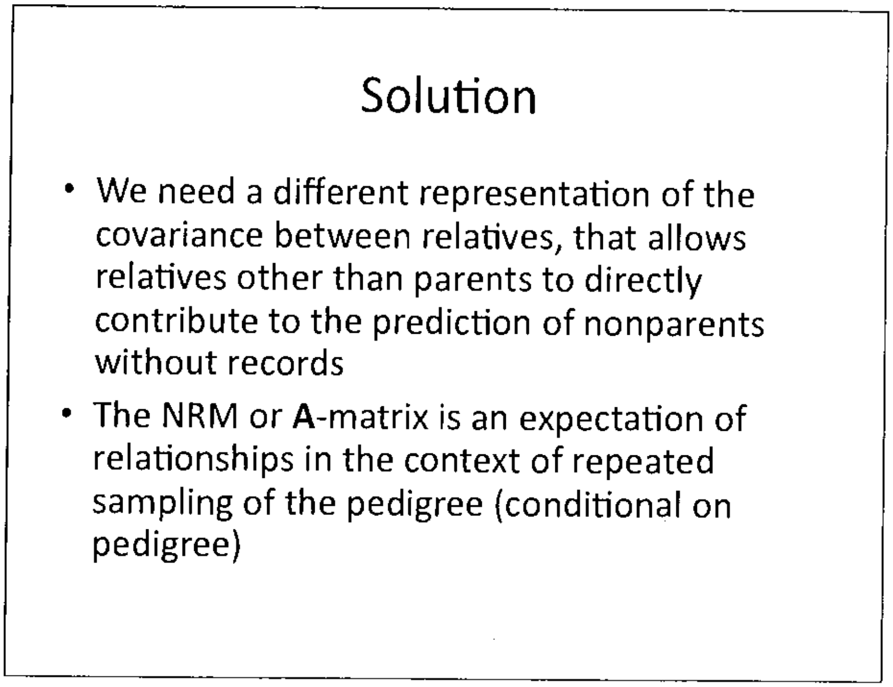
Property of BLP/BLUP is $\mathrm{cov}(u,\hat{u}) = \mathrm{var}(\hat{u})$ *so* $\mathrm{r}^2 = \dfrac{\mathrm{var}(\hat{u})}{\mathrm{var}(u)}$

but $\hat{u}_{nonparent} = \dfrac{\hat{u}_{sire}}{2} + \dfrac{\hat{u}_{dam}}{2}$, *for nonparent without a record*

so $\mathrm{r}^2_{nonparent} = \dfrac{r^2_{sire}}{4} + \dfrac{r^2_{dam}}{4} \leq \dfrac{1}{2}$

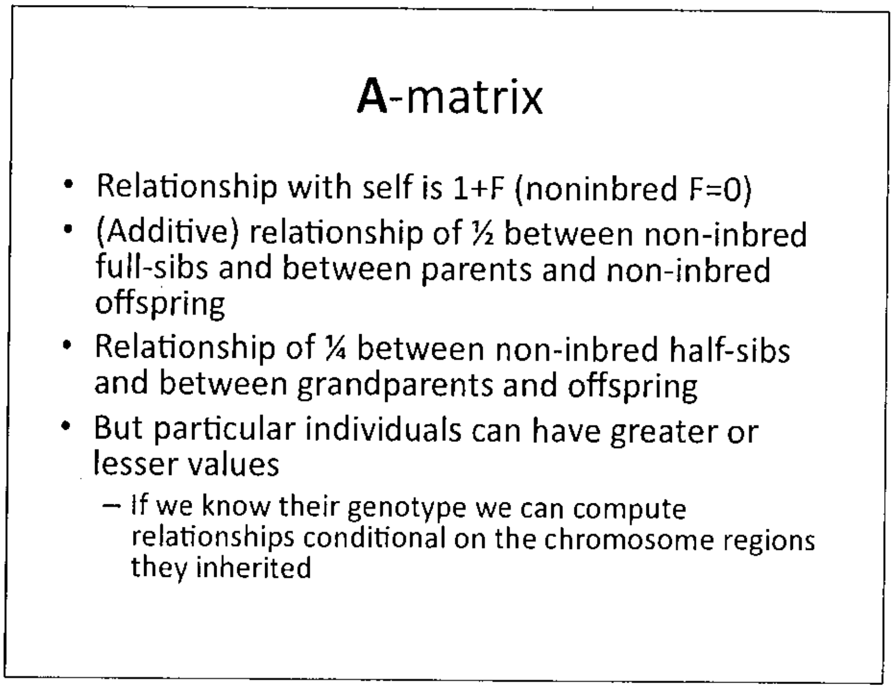
Finally $\Delta G = \dfrac{i r_{nonparent}\sigma_g}{L}$, limiting selection response when candidates at puberty lack phenotypic information

An option to do better



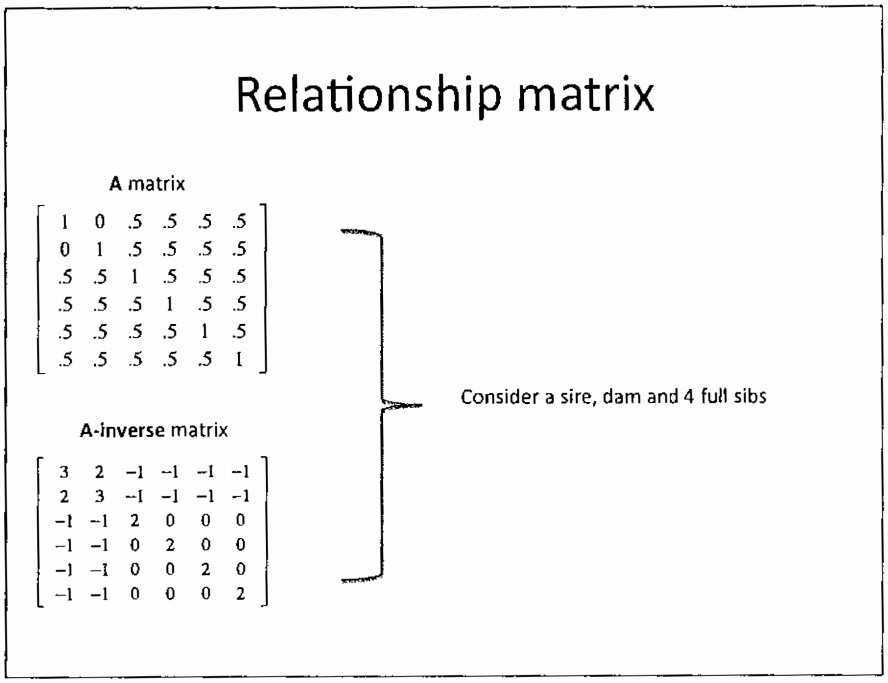
Solution

- We need a different representation of the covariance between relatives, that allows relatives other than parents to directly contribute to the prediction of nonparents without records
- The NRM or **A**-matrix is an expectation of relationships in the context of repeated sampling of the pedigree (conditional on pedigree)



A-matrix

- Relationship with self is 1+F (noninbred F=0)
- (Additive) relationship of ½ between non-inbred full-sibs and between parents and non-inbred offspring
- Relationship of ¼ between non-inbred half-sibs and between grandparents and offspring
- But particular individuals can have greater or lesser values
  - If we know their genotype we can compute relationships conditional on the chromosome regions they inherited



Relationship matrix

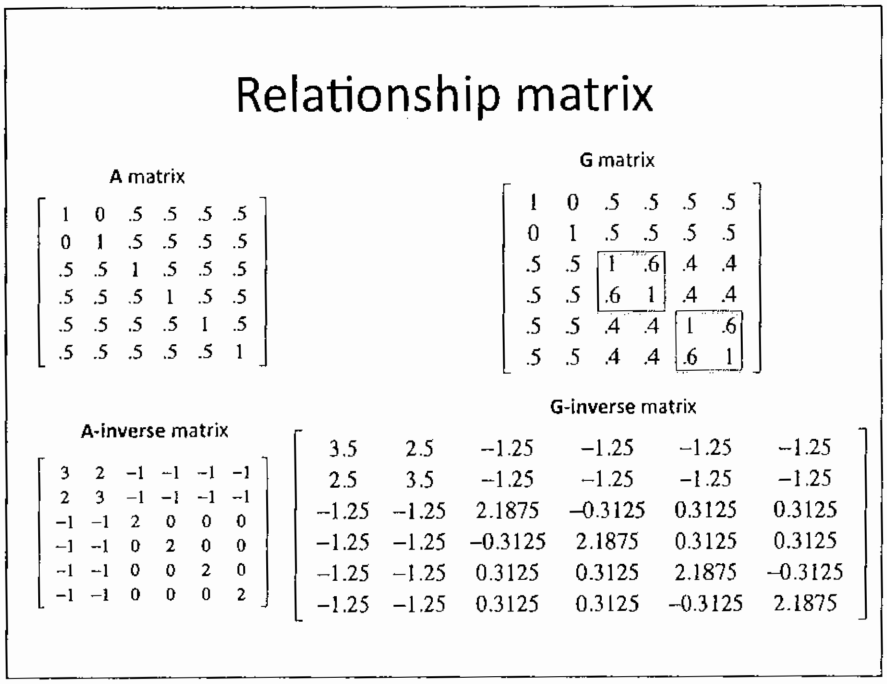
A matrix

$$\begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .5 \\ 0 & 1 & .5 & .5 & .5 & .5 \\ .5 & .5 & 1 & .5 & .5 & .5 \\ .5 & .5 & .5 & 1 & .5 & .5 \\ .5 & .5 & .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & .5 & .5 & 1 \end{bmatrix}$$

A-inverse matrix

$$\begin{bmatrix} 3 & 2 & -1 & -1 & -1 & -1 \\ 2 & 3 & -1 & -1 & -1 & -1 \\ -1 & -1 & 2 & 0 & 0 & 0 \\ -1 & -1 & 0 & 2 & 0 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 \\ -1 & -1 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Consider a sire, dam and 4 full sibs



Relationship matrix

A matrix

$$\begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .5 \\ 0 & 1 & .5 & .5 & .5 & .5 \\ .5 & .5 & 1 & .5 & .5 & .5 \\ .5 & .5 & .5 & 1 & .5 & .5 \\ .5 & .5 & .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & .5 & .5 & 1 \end{bmatrix}$$

G matrix

$$\begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .5 \\ 0 & 1 & .5 & .5 & .5 & .5 \\ .5 & .5 & 1 & .6 & .4 & .4 \\ .5 & .5 & .6 & 1 & .4 & .4 \\ .5 & .5 & .4 & .4 & 1 & .6 \\ .5 & .5 & .4 & .4 & .6 & 1 \end{bmatrix}$$

A-inverse matrix

$$\begin{bmatrix} 3 & 2 & -1 & -1 & -1 & -1 \\ 2 & 3 & -1 & -1 & -1 & -1 \\ -1 & -1 & 2 & 0 & 0 & 0 \\ -1 & -1 & 0 & 2 & 0 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 \\ -1 & -1 & 0 & 0 & 0 & 2 \end{bmatrix}$$

G-inverse matrix

$$\begin{bmatrix} 3.5 & 2.5 & -1.25 & -1.25 & -1.25 & -1.25 \\ 2.5 & 3.5 & -1.25 & -1.25 & -1.25 & -1.25 \\ -1.25 & -1.25 & 2.1875 & -0.3125 & 0.3125 & 0.3125 \\ -1.25 & -1.25 & -0.3125 & 2.1875 & 0.3125 & 0.3125 \\ -1.25 & -1.25 & 0.3125 & 0.3125 & 2.1875 & -0.3125 \\ -1.25 & -1.25 & 0.3125 & 0.3125 & -0.3125 & 2.1875 \end{bmatrix}$$

Predict the last animal with no data

$$\begin{bmatrix} -1.25\hat{u}_{sire} & -1.25\hat{u}_{dam} & .3125\hat{u}_{sib1} & .3125\hat{u}_{sib2} & -.3125\hat{u}_{sib3} & 2.1875\hat{u}_{candidate} \end{bmatrix} = [0]$$

$$\hat{u}_{candidate} = \frac{1.25(\hat{u}_{sire} + \hat{u}_{dam}) - 0.3125(\hat{u}_{sib1} + \hat{u}_{sib2}) + 0.3125\hat{u}_{sib3}}{2.1875}$$

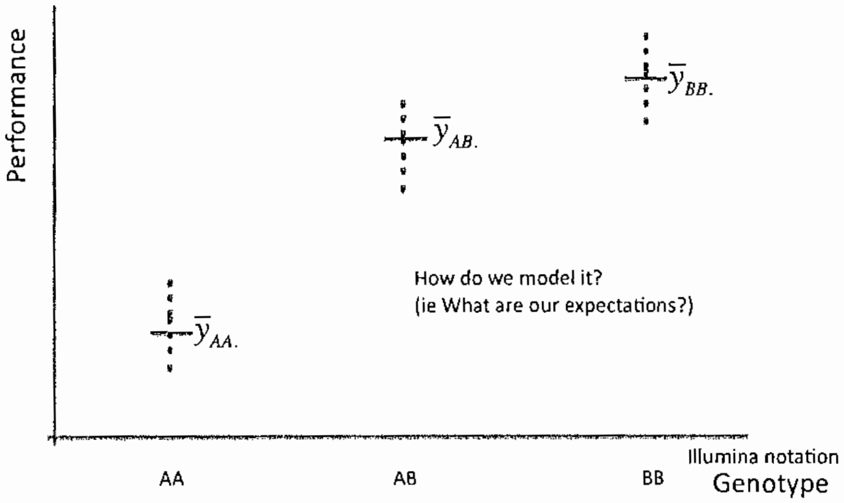
But to form **G**, we needed to know which loci/QTL contribute to variation in performance

Fixed effects models to predict SNP effects

Genomic Prediction

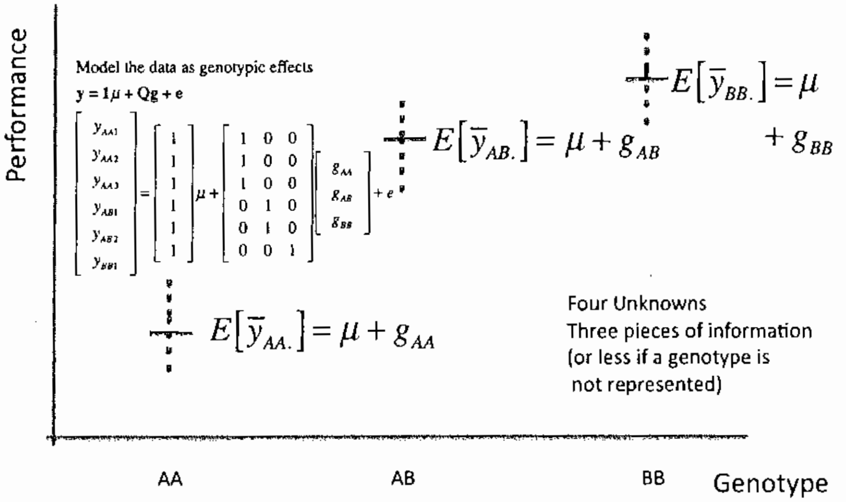
- Two-step process
  - Training population
    - Predict the breeding value of (every) (small) genomic region (to find the informative regions ie QTL)
  - Target population
    - Predict the breeding value of the selection candidates by summing up the breeding values of all the genomic regions they inherited

Data on some locus



How do we model it?
(ie What are our expectations?)

Data on some locus



Model the data as genotypic effects

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

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} g_{AA} \\ g_{AB} \\ g_{BB} \end{bmatrix} + e$$

$E[\bar{y}_{AA.}] = \mu + g_{AA}$

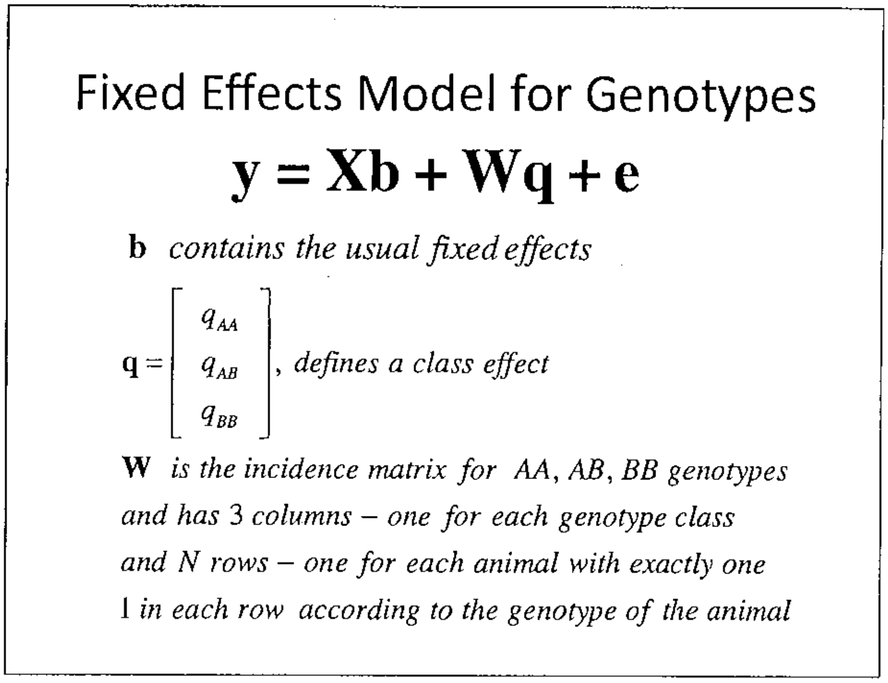
$E[\bar{y}_{AB.}] = \mu + g_{AB}$

$E[\bar{y}_{BB.}] = \mu + g_{BB}$

Four Unknowns
Three pieces of information
(or less if a genotype is
not represented)

Parameters and Information Content

- The information content (in fixed effects model) is partly reflected in the degrees of freedom
  - Some degrees of freedom are available to estimate functions of fitted parameters
  - The remainder, if any, contribute to the error sum of squares
- Overparameterized models have more parameters than estimable functions



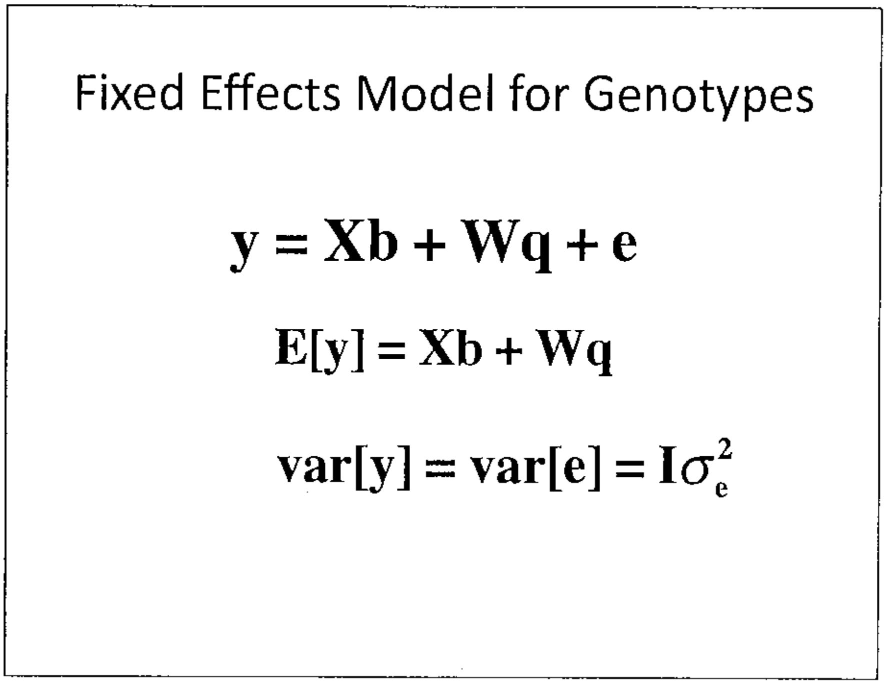
Fixed Effects Model for Genotypes

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

$\mathbf{b}$ *contains the usual fixed effects*

$\mathbf{q} = \begin{bmatrix} q_{AA} \\ q_{AB} \\ q_{BB} \end{bmatrix}$, *defines a class effect*

$\mathbf{W}$ *is the incidence matrix for AA, AB, BB genotypes and has 3 columns – one for each genotype class and N rows – one for each animal with exactly one 1 in each row according to the genotype of the animal*



Fixed Effects Model for Genotypes

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

$$\mathrm{E}[\mathbf{y}] = \mathbf{Xb} + \mathbf{Wq}$$

$$\mathrm{var}[\mathbf{y}] = \mathrm{var}[\mathbf{e}] = \mathbf{I}\sigma_e^2$$

Least Squares Equations

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{q}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

$$For \; [\mathbf{b}] = [\mu], \mathbf{X} = \mathbf{1}$$

$$LHS = \begin{bmatrix} N & n_{AA} & n_{AB} & n_{BB} \\ n_{AA} & n_{AA} & 0 & 0 \\ n_{AB} & 0 & n_{AB} & 0 \\ n_{BB} & 0 & 0 & n_{BB} \end{bmatrix} RHS = \begin{bmatrix} y_{..} \\ y_{AA}. \\ y_{AB}. \\ y_{BB}. \end{bmatrix}$$

Equations have order equal to number of fixed effects plus genotypes

No unique solution

$$LHS = \begin{bmatrix} \cancel{N} & \cancel{n_{AA}} & \cancel{n_{AB}} & \cancel{n_{BB}} \\ \cancel{n_{AA}} & n_{AA} & 0 & 0 \\ \cancel{n_{AB}} & 0 & n_{AB} & 0 \\ \cancel{n_{BB}} & 0 & 0 & n_{BB} \end{bmatrix} RHS = \begin{bmatrix} \cancel{y_{..}} \\ y_{AA}. \\ y_{AB}. \\ y_{BB}. \end{bmatrix}$$

$$\hat{\mathbf{b}} = \begin{bmatrix} 0 \\ \widehat{\mu + q_{AA}} \\ \widehat{\mu + q_{AB}} \\ \widehat{\mu + q_{BB}} \end{bmatrix}, \; \textit{is one possible solution}$$

No unique solution

$$\hat{\mathbf{b}} = \begin{bmatrix} \widehat{\mu + q_{BB}} \\ \widehat{q_{AA} - q_{BB}} \\ \widehat{q_{AB} - q_{BB}} \\ 0 \end{bmatrix}, \textit{is another possible solution}$$

$$LHS = \begin{bmatrix} N & n_{AA} & n_{AB} & \cancel{n_{BB}} \\ n_{AA} & n_{AA} & 0 & \cancel{0} \\ n_{AB} & 0 & n_{AB} & \cancel{0} \\ \cancel{n_{BB}} & \cancel{0} & \cancel{0} & \cancel{n_{BB}} \end{bmatrix} \quad RHS = \begin{bmatrix} y_{..} \\ y_{AA\cdot} \\ y_{AB\cdot} \\ \cancel{y_{BB\cdot}} \end{bmatrix}$$

Different Solutions have same Estimable Functions

$$\hat{\mathbf{b}}_1 = \begin{bmatrix} \widehat{\mu + q_{BB}} \\ \widehat{q_{AA} - q_{BB}} \\ \widehat{q_{AB} - q_{BB}} \\ 0 \end{bmatrix} \qquad \hat{\mathbf{b}}_2 = \begin{bmatrix} 0 \\ \widehat{\mu + q_{AA}} \\ \widehat{\mu + q_{AB}} \\ \widehat{\mu + q_{BB}} \end{bmatrix}$$

Interesting contrasts

$$\mathbf{k}' = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix} \textit{ then } \mathbf{k}'\hat{\mathbf{b}}_1 = \mathbf{k}'\hat{\mathbf{b}}_2 = \widehat{\mu + q_{AA}}$$

$$\mathbf{k}' = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix} \textit{ then } \mathbf{k}'\hat{\mathbf{b}}_1 = \mathbf{k}'\hat{\mathbf{b}}_2 = \widehat{q_{AA} - q_{AB}}$$

Estimable Functions

- In fixed effects models, many model parameters or functions of model parameters are not estimable, even though a numeric value can be obtained by solving the least squares equations (eg by generalized inverse)

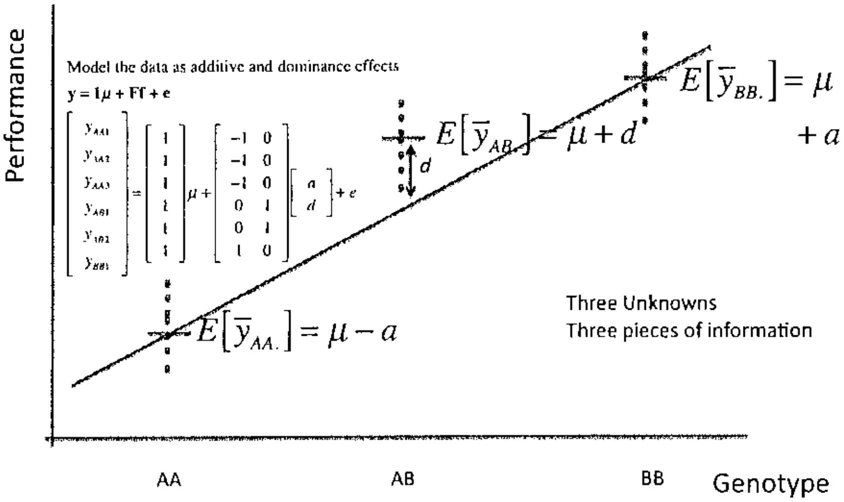
$[\mathbf{X'X}]^-$ is any generalized inverse of $\mathbf{X'X}$ if $(\mathbf{X'X})[\mathbf{X'X}]^-(\mathbf{X'X}) = \mathbf{X'X}$

Define $\mathbf{H} = [\mathbf{X'X}]^-(\mathbf{X'X})$

A linear function $\mathbf{k'b}^0$ is estimable if $\mathbf{k'H} = \mathbf{k'}$

$\text{var}(\mathbf{k'b}^0) = \mathbf{k'}[\mathbf{X'X}]^-\mathbf{k}$ $\{or\, \mathbf{k'}[\mathbf{X'X}]^-\mathbf{k}\,\sigma^2$ (if $\mathbf{R}$ was not explicitly fitted)$\}$

Data on some locus



Model the data as additive and dominance effects

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ff} + \mathbf{e}$$

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}\mu + \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} a \\ d \end{bmatrix} + e$$

$E[\overline{y}_{AA.}] = \mu - a$

$E[\overline{y}_{AB.}] = \mu + d$

$E[\overline{y}_{BB.}] = \mu + a$

Three Unknowns
Three pieces of information

Genotypic vs genetic effects

$$\mathbf{g} = \begin{bmatrix} g_{AA} \\ g_{AB} \\ g_{BB} \end{bmatrix}, \textit{ genotypic class effects} \quad \mathbf{a} = \begin{bmatrix} -a \\ d \\ a \end{bmatrix}, \textit{ additive and dominance effects}$$

$$a = \frac{g_{BB} - g_{AA}}{2}, \textit{and } d = g_{AB} - \frac{g_{AA} + g_{BB}}{2}$$

$$\mathbf{K}' = \begin{bmatrix} \mathbf{k}_1' \\ \mathbf{k}_2' \end{bmatrix} = \begin{bmatrix} \frac{-1}{2} & 0 & \frac{1}{2} \\ \frac{-1}{2} & 1 & \frac{-1}{2} \end{bmatrix}, \mathbf{K}'\mathbf{q} = \mathbf{a}, \textit{ columns of } \mathbf{K} \textit{ are othogonal } \mathbf{k}_1'\mathbf{k}_2 = 0$$

but note $\mathbf{g}$ *itself is not estimable, but functions like* $g_{BB} - g_{AA}$ *are*

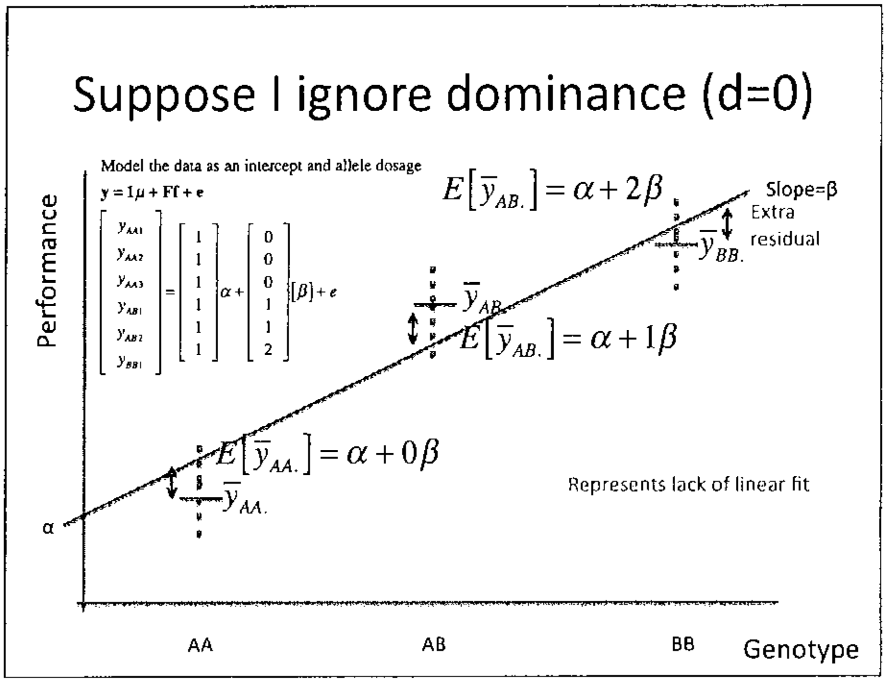
Equivalent Models

	Genotypic	E[ ]	Falconer	E[ ]
AA	$\mu+g_{AA}$	10	μ-a	10=13-3
AB	$\mu+g_{AB}$	14	μ+d	14=13+1
BB	$\mu+g_{BB}$	16	μ+a	16=13+3

μ=0	μ=10	μ=16	μ=13
g_{AA}= 10	g_{AA}= 0	g_{AA}= -6	a= 3
g_{AB}= 14	g_{AB}= 4	g_{AB}= -2	d= 1
g_{BB}= 16	g_{BB}= 6	g_{BB}= 0	

Both models have the same expectation
Both models have the same variance

Therefore the models are equivalent
(I can fit either model and migrate from one to the other)

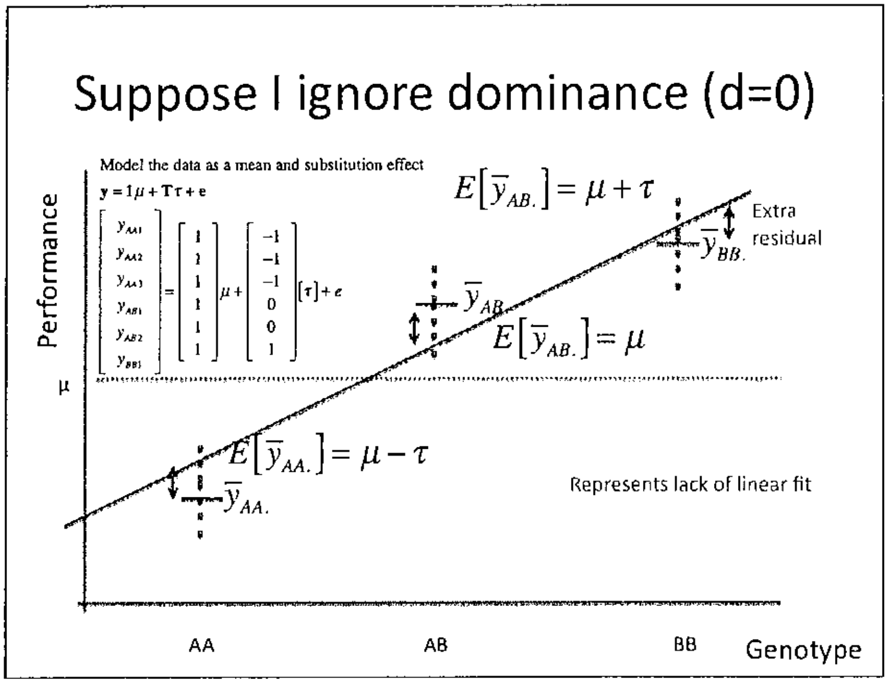


Suppose I ignore dominance (d=0)

Model the data as an intercept and allele dosage

$\mathbf{y} = \mathbf{1}\mu + \mathbf{F}\mathbf{f} + \mathbf{e}$

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \alpha + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 2 \end{bmatrix} [\beta] + e$$

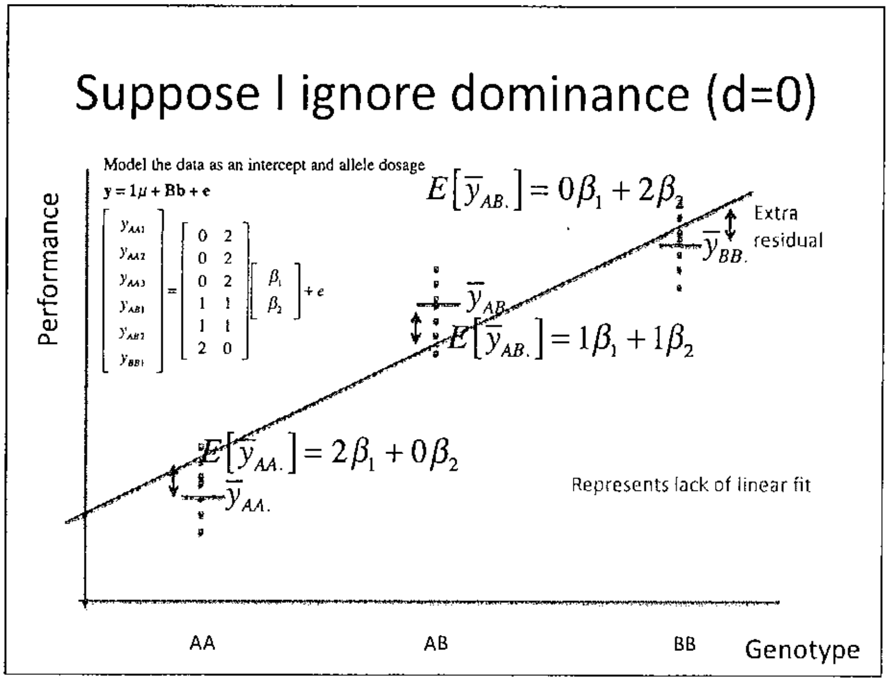


Suppose I ignore dominance (d=0)

Model the data as a mean and substitution effect

$\mathbf{y} = \mathbf{1}\mu + \mathbf{T}\tau + \mathbf{e}$

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} -1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix} [\tau] + e$$



Suppose I ignore dominance (d=0)

Model the data as an intercept and allele dosage

$\mathbf{y} = \mathbf{1}\mu + \mathbf{B}\mathbf{b} + \mathbf{e}$

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 0 & 2 \\ 0 & 2 \\ 1 & 1 \\ 1 & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + e$$

$$E[\bar{y}_{AB.}] = 0\beta_1 + 2\beta_2$$

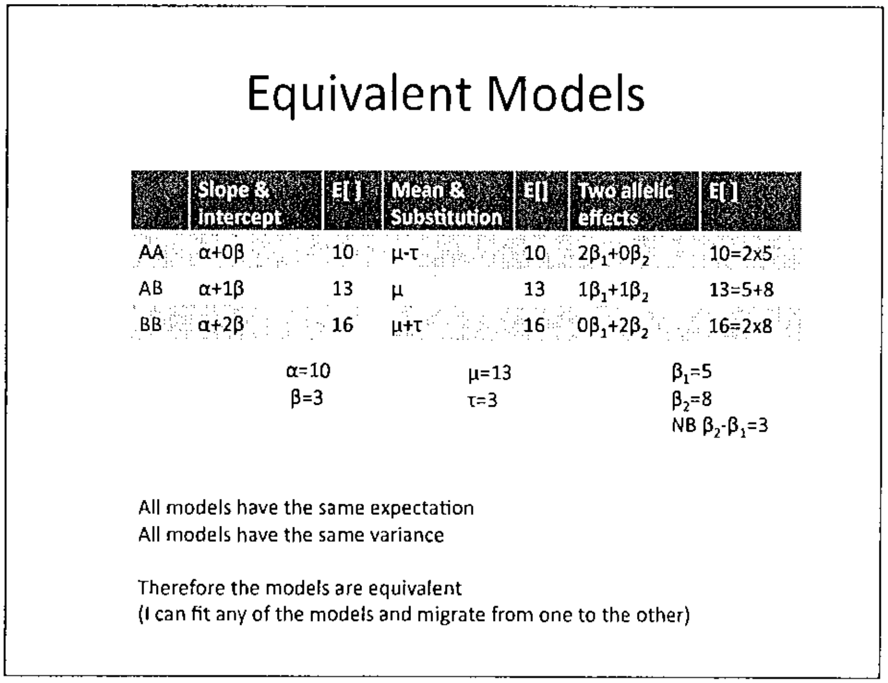
$$E[\bar{y}_{AB.}] = 1\beta_1 + 1\beta_2$$

$$E[\bar{y}_{AA.}] = 2\beta_1 + 0\beta_2$$

Extra residual

Represents lack of linear fit

Performance
$\bar{y}_{BB.}$
$\bar{y}_{AB.}$
$\bar{y}_{AA.}$
AA
AB
BB
Genotype

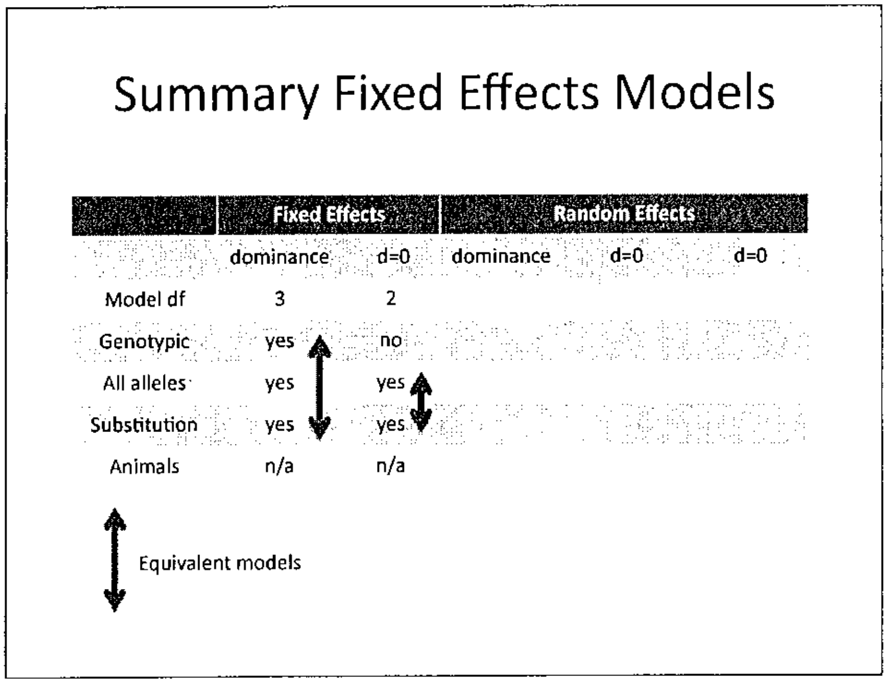


Equivalent Models

	Slope & Intercept	E[ ]	Mean & Substitution	E[]	Two allelic effects	E[ ]
AA	α+0β	10	μ-τ	10	$2\beta_1+0\beta_2$	10=2x5
AB	α+1β	13	μ	13	$1\beta_1+1\beta_2$	13=5+8
BB	α+2β	16	μ+τ	16	$0\beta_1+2\beta_2$	16=2x8
	α=10 β=3		μ=13 τ=3		β_1=5 β_2=8 NB β_2-β_1=3	

All models have the same expectation
All models have the same variance

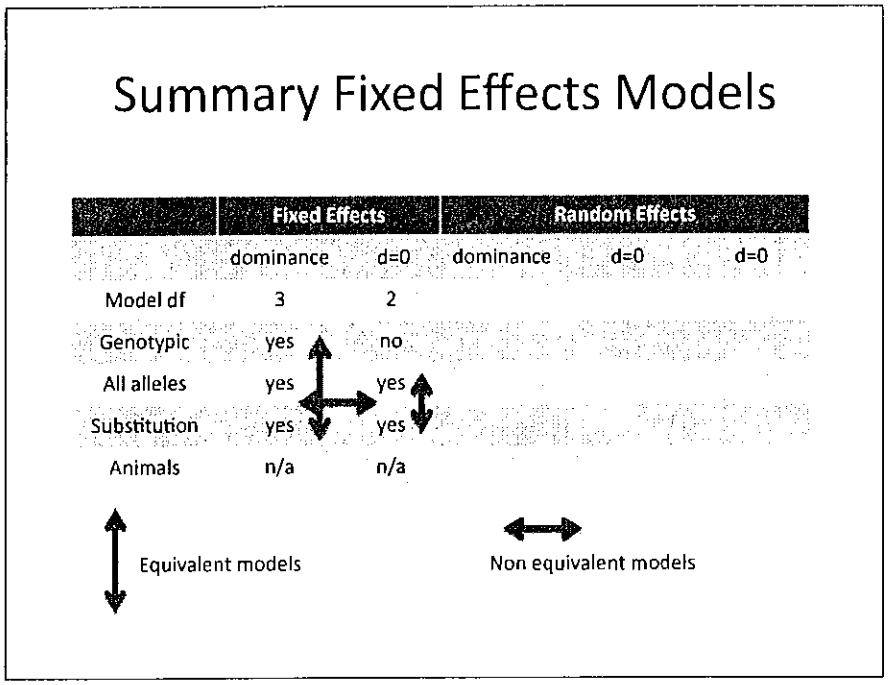
Therefore the models are equivalent
(I can fit any of the models and migrate from one to the other)



Summary Fixed Effects Models

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2			
Genotypic	yes	no			
All alleles	yes	yes			
Substitution	yes	yes			
Animals	n/a	n/a			

↕ Equivalent models



Summary Fixed Effects Models

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2			
Genotypic	yes	no			
All alleles	yes	yes			
Substitution	yes	yes			
Animals	n/a	n/a			

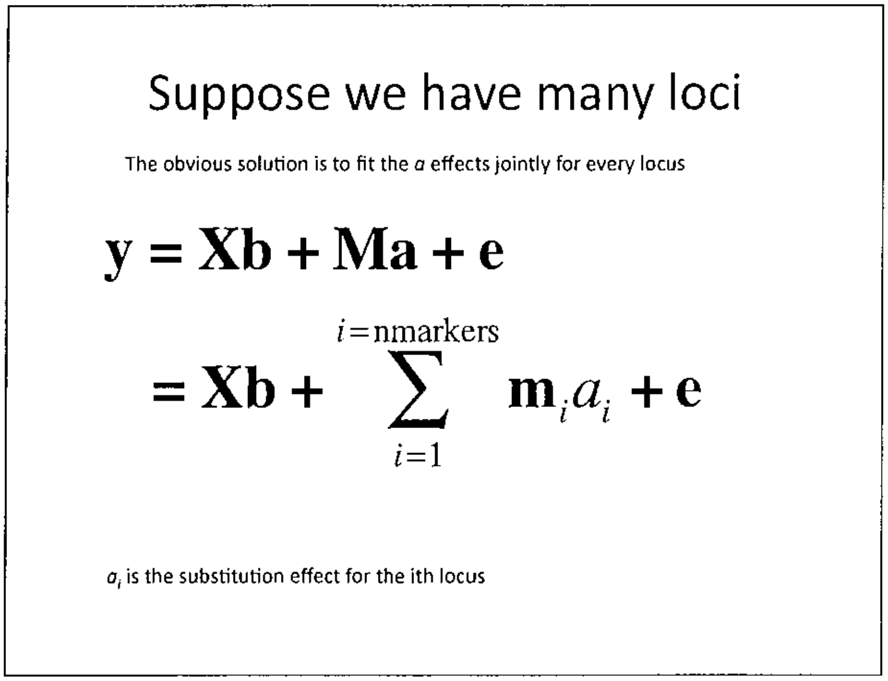
↕ Equivalent models

↔ Non equivalent models

Fitting SNPs as random effects

Fixed or Random

- Reasonable to consider animal effects as random in the usual context
    - Variation in alleles (ie genotype) between animals that contributes to the genetic variance
        - Not variation in allelic value at a particular locus
- Not so clear that an individual locus (or every loci) should be treated as random
    - Especially when the genotypes are observed and treated as known in the incidence matrix



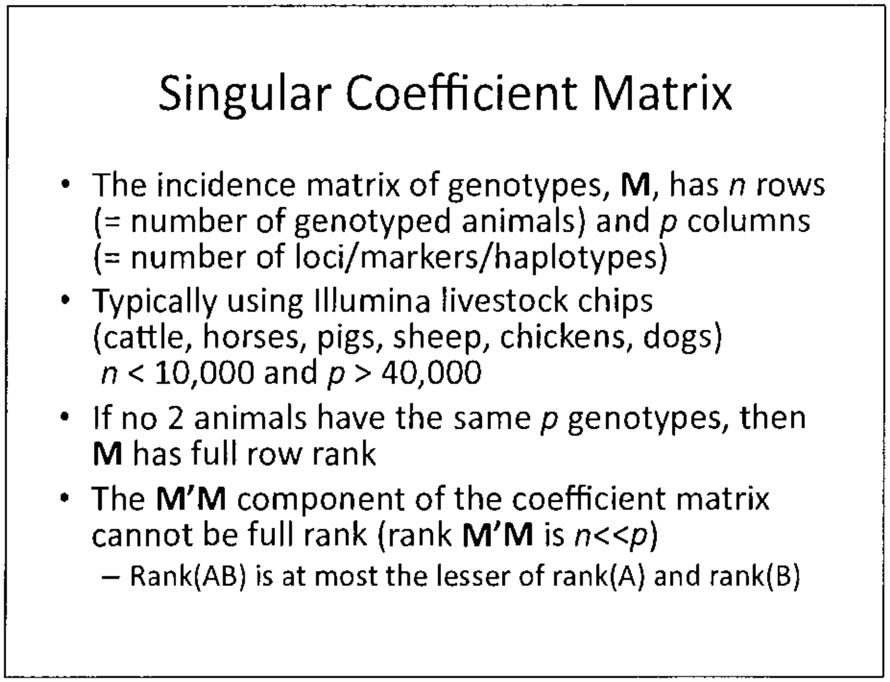
Suppose we have many loci

The obvious solution is to fit the *a* effects jointly for every locus

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

$$= \mathbf{Xb} + \sum_{i=1}^{i=\text{nmarkers}} \mathbf{m}_i a_i + \mathbf{e}$$

a_i is the substitution effect for the ith locus



Singular Coefficient Matrix

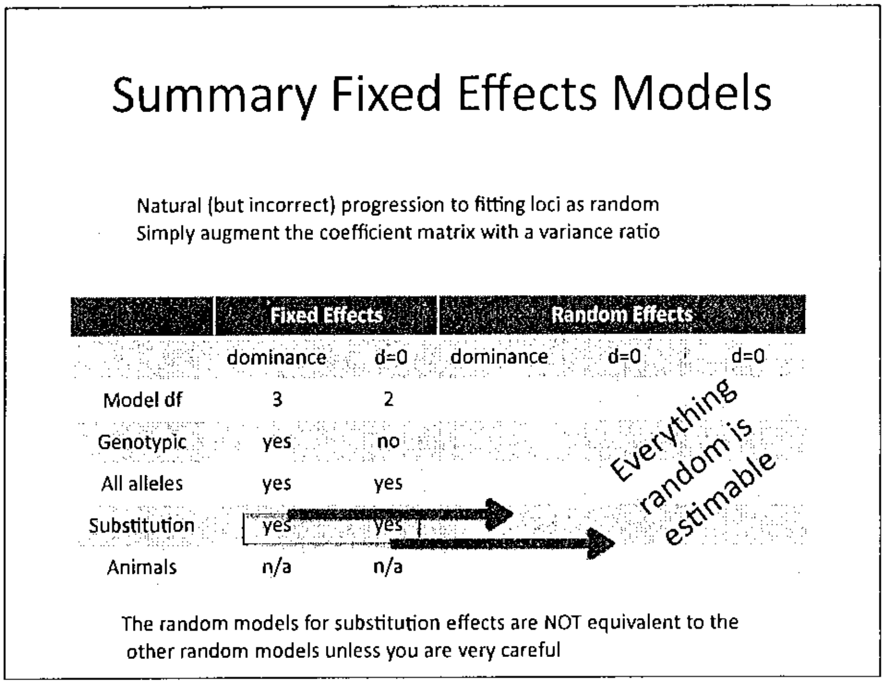
- The incidence matrix of genotypes, **M**, has *n* rows (= number of genotyped animals) and *p* columns (= number of loci/markers/haplotypes)
- Typically using Illumina livestock chips (cattle, horses, pigs, sheep, chickens, dogs) *n* < 10,000 and *p* > 40,000
- If no 2 animals have the same *p* genotypes, then **M** has full row rank
- The **M'M** component of the coefficient matrix cannot be full rank (rank **M'M** is *n*<<*p*)
  - Rank(AB) is at most the lesser of rank(A) and rank(B)

Practical Consequence

- It is not possible using ordinary least squares to simultaneously estimate more than *n* effects of loci plus other fixed effects
  - Can use stepwise approaches to successively add loci and determine a subset of markers that are informative in the training data
    - But least squares tend to produce upwards biased estimates of effects (especially when power is limiting)
  - Cannot use all markers to predict genomic merit

Alternative Approaches

- Modifications to Least Squares
  - Ridge Regression, Partial Least Squares etc
- Treat *a* effects as random rather than fixed
  - We routinely fit single and multi-trait animal models with many more effects than observations
  - Provides opportunities for many mixed model procedures, such as BLUP, REML, Bayesian analyses
  - These methods will also "shrink" estimates



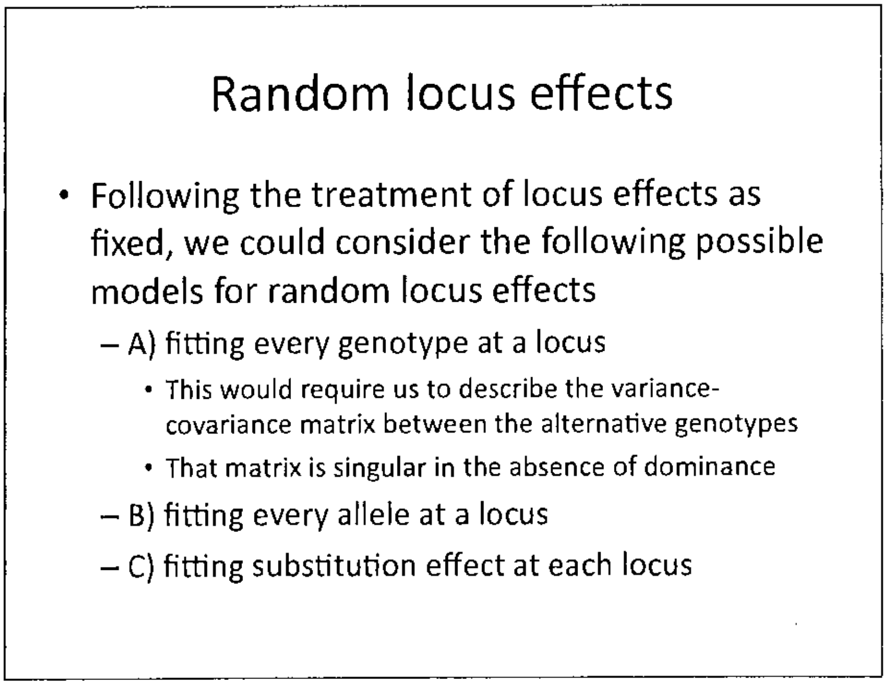
Summary Fixed Effects Models

Natural (but incorrect) progression to fitting loci as random
Simply augment the coefficient matrix with a variance ratio

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2			
Genotypic	yes	no			
All alleles	yes	yes			
Substitution	yes	yes			
Animals	n/a	n/a			

Everything random is estimable

The random models for substitution effects are NOT equivalent to the other random models unless you are very careful



Random locus effects

- Following the treatment of locus effects as fixed, we could consider the following possible models for random locus effects
  - A) fitting every genotype at a locus
    - This would require us to describe the variance-covariance matrix between the alternative genotypes
    - That matrix is singular in the absence of dominance
  - B) fitting every allele at a locus
  - C) fitting substitution effect at each locus

Mixed Model Theory

- Prediction and estimation follow logically once we define relevant variance-covariance matrices
  - All effects are estimable (unlike least squares)

$$\mathrm{var}(\mathbf{g}) = \mathbf{G} \quad \mathrm{var}(\hat{\mathbf{g}}) = \mathbf{G} - \mathbf{C}^{22} \quad \mathrm{var}(\hat{\mathbf{g}} - \mathbf{g}) = \mathbf{C}^{22} \quad r^2_{g\hat{g}} = {}^{\mathrm{var}(\hat{g})}\!\Big/_{\mathrm{var}(g)}$$

$$\mathrm{var}(\mathbf{k'g}) = \mathbf{k'Gk} \quad \mathrm{var}(\mathbf{k'\hat{g}}) = \mathbf{k'}\left(\mathbf{G} - \mathbf{C}^{22}\right)\mathbf{k}$$

- The analogous terms in routinely applied animal models are the numerator relationship matrix, genetic and residual variances
  - Random effects might be interpreted in the context of resampling in repeat experiments

Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes		This model follows	
Substitution	yes	yes		R≠D	R=D
Animals	n/a	n/a			

Fit all allelic effects as random

- Assuming no dominance we could fit effects of two (or more) individual alleles

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

- ***M*** is a matrix of covariates, one column for each allele (or haplotype), that counts the number of copies – each row sums to two

$$\textit{rows of } \mathbf{M} \textit{ are one of} \begin{bmatrix} 0 & 2 \\ 1 & 1 \\ 2 & 0 \end{bmatrix}, \mathrm{a} = \begin{bmatrix} a_A \\ a_B \end{bmatrix}, \textit{ for} \begin{bmatrix} y_{AA} \\ y_{AB} \\ y_{BB} \end{bmatrix}$$

Estimable Functions in Fixed Models

- Class variables of fixed effects are not estimable
  - Differences between levels in the same class are estimable
  - The sum of any one level and the mean are estimable (in a 1-way model)
  - Fitting a fixed class variable is typically done by
    - deleting the row and column of the coefficient matrix for any one level of the class
    - Introducing a lagrange multiplier to fit a sigma constraint

Sum to Zero in Random Models

- Class variables of random effects (e.g. sire or animal) are all estimable
  - Typically all levels are fitted, even though interest may be focused on differences between levels (eg one sire compared to another)
- A feature of BLUP(u) is that certain sums of the elements are zero
  - A biallelic factor fitting say a_1 and a_2 will have solutions that sum to zero (ie a-hat$_1$ = - a-hat$_2$)
  - In a model fitting many biallelic loci as random effects, the number of equations can be halved

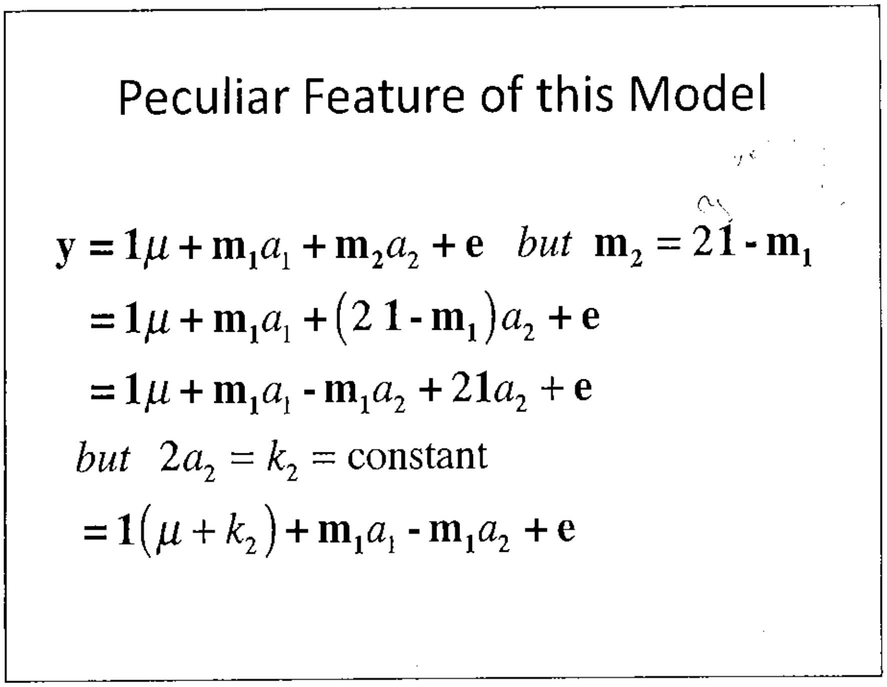
Var(**a**) (ie allelic effects)

$$\mathrm{var}(\mathbf{a}) = \mathbf{A} = \mathrm{var}\begin{bmatrix} a_A \\ a_B \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\sigma_A^2 = \mathbf{I}\sigma_A^2$$

For the 3 possible biallelic genotypes

$$\mathrm{var}(\mathbf{MA}) = \mathbf{MAM'} = \begin{bmatrix} 0 & 2 \\ 1 & 1 \\ 2 & 0 \end{bmatrix}\mathbf{A}\begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 2 & 4 \end{bmatrix}\sigma_A^2$$

Note this **A** is the variance-covariance matrix of allelic effects, not the NRM



Peculiar Feature of this Model

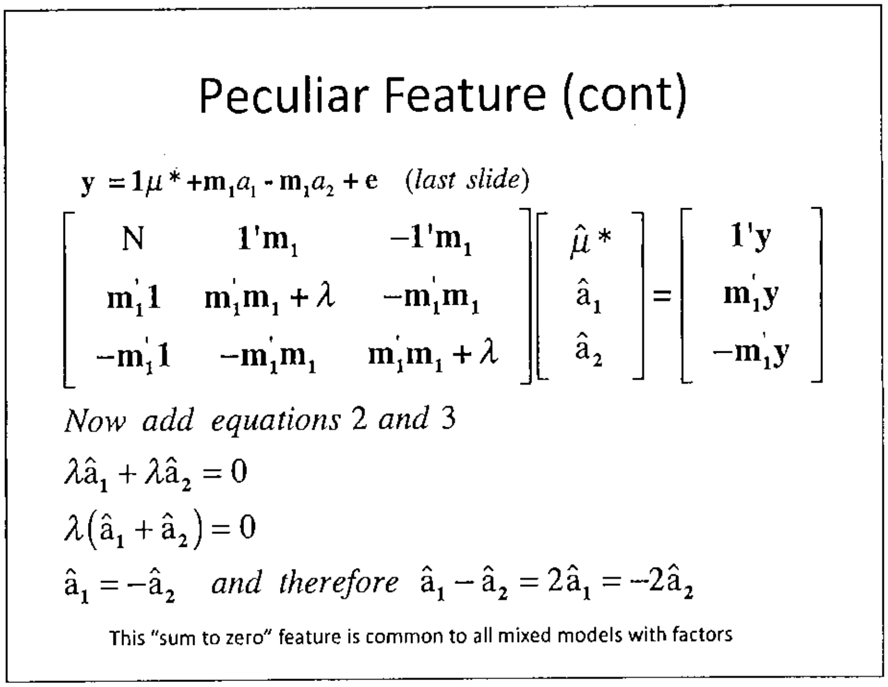
$$\mathbf{y} = \mathbf{1}\mu + \mathbf{m}_1 a_1 + \mathbf{m}_2 a_2 + \mathbf{e} \quad but \quad \mathbf{m}_2 = 2\mathbf{1} - \mathbf{m}_1$$

$$= \mathbf{1}\mu + \mathbf{m}_1 a_1 + (2\,\mathbf{1} - \mathbf{m}_1) a_2 + \mathbf{e}$$

$$= \mathbf{1}\mu + \mathbf{m}_1 a_1 - \mathbf{m}_1 a_2 + 2\mathbf{1} a_2 + \mathbf{e}$$

$$but \quad 2a_2 = k_2 = \text{constant}$$

$$= \mathbf{1}(\mu + k_2) + \mathbf{m}_1 a_1 - \mathbf{m}_1 a_2 + \mathbf{e}$$



Peculiar Feature (cont)

$$\mathbf{y} = \mathbf{1}\mu^* + \mathbf{m}_1 a_1 - \mathbf{m}_1 a_2 + \mathbf{e} \quad \textit{(last slide)}$$

$$\begin{bmatrix} \mathrm{N} & \mathbf{1'm}_1 & -\mathbf{1'm}_1 \\ \mathbf{m}_1'\mathbf{1} & \mathbf{m}_1'\mathbf{m}_1 + \lambda & -\mathbf{m}_1'\mathbf{m}_1 \\ -\mathbf{m}_1'\mathbf{1} & -\mathbf{m}_1'\mathbf{m}_1 & \mathbf{m}_1'\mathbf{m}_1 + \lambda \end{bmatrix} \begin{bmatrix} \hat{\mu}^* \\ \hat{\mathrm{a}}_1 \\ \hat{\mathrm{a}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{m}_1'\mathbf{y} \\ -\mathbf{m}_1'\mathbf{y} \end{bmatrix}$$

Now add equations 2 *and* 3

$$\lambda \hat{\mathrm{a}}_1 + \lambda \hat{\mathrm{a}}_2 = 0$$

$$\lambda(\hat{\mathrm{a}}_1 + \hat{\mathrm{a}}_2) = 0$$

$$\hat{\mathrm{a}}_1 = -\hat{\mathrm{a}}_2 \quad \textit{and therefore} \quad \hat{\mathrm{a}}_1 - \hat{\mathrm{a}}_2 = 2\hat{\mathrm{a}}_1 = -2\hat{\mathrm{a}}_2$$

This "sum to zero" feature is common to all mixed models with factors

Extension to multiple loci

Allelic effects

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e} \quad (1\ locus)$$

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{i=ploci} \mathbf{M}_i\mathbf{a}_i + \mathbf{e} \quad (p\,loci)$$

MME for two uncorrelated loci (order is 1+ 2 x 2 = 4 allelic effects)

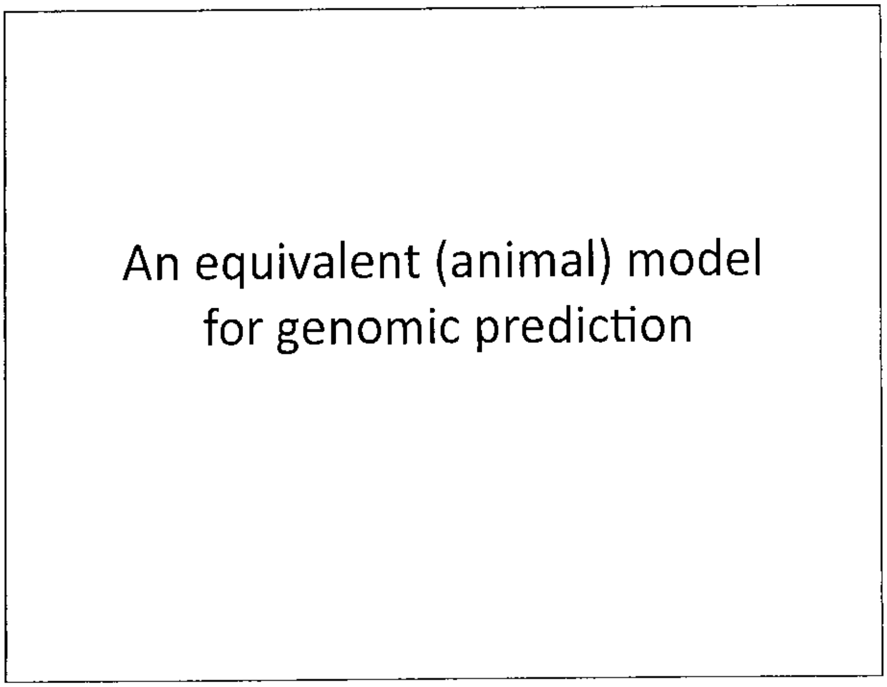
$$\begin{bmatrix} N & \mathbf{1'M}_1 & \mathbf{1'M}_2 \\ \mathbf{M}_1^{'}\mathbf{1} & \mathbf{M}_1^{'}\mathbf{M}_1 + \lambda_1 & \mathbf{M}_1^{'}\mathbf{M}_2 \\ \mathbf{M}_2^{'}\mathbf{1} & \mathbf{M}_2^{'}\mathbf{M}_1 & \mathbf{M}_2^{'}\mathbf{M}_2 + \lambda_2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{a}}_1 \\ \hat{\mathbf{a}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{M}_1^{'}\mathbf{y} \\ \mathbf{M}_2^{'}\mathbf{y} \end{bmatrix}$$

Order of MME is number of fixed effects plus twice number loci (if biallelic)
Consider the implications for 100-1,000 animals with 50,000 loci

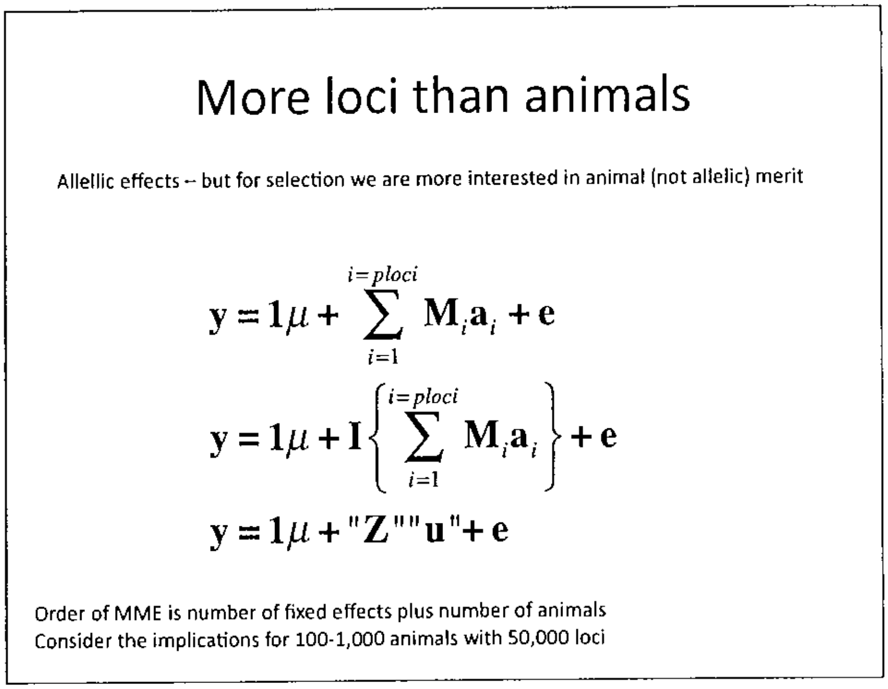
$$\lambda_i = \frac{\sigma_e^2}{\sigma_{ai}^2}$$

Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes			
Substitution	yes	yes		R≠D	R=D
Animals	n/a	n/a		This model follows	



An equivalent (animal) model for genomic prediction



More loci than animals

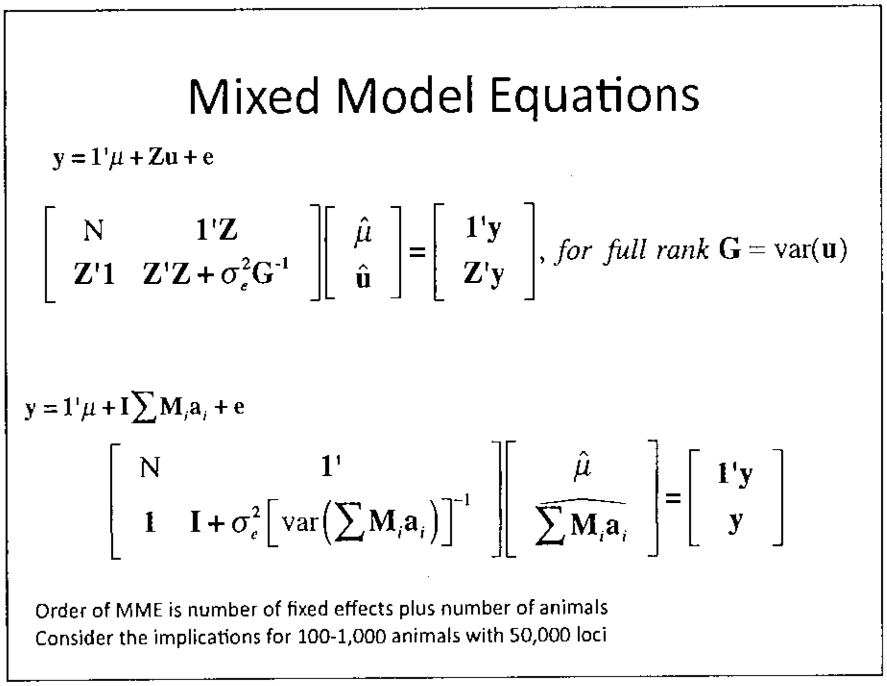
Allellic effects – but for selection we are more interested in animal (not allelic) merit

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{i=ploci} \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{I}\left\{ \sum_{i=1}^{i=ploci} \mathbf{M}_i \mathbf{a}_i \right\} + \mathbf{e}$$

$$\mathbf{y} = \mathbf{1}\mu + "\mathbf{Z}""\mathbf{u}" + \mathbf{e}$$

Order of MME is number of fixed effects plus number of animals
Consider the implications for 100-1,000 animals with 50,000 loci



Mixed Model Equations

$y = 1'\mu + Zu + e$

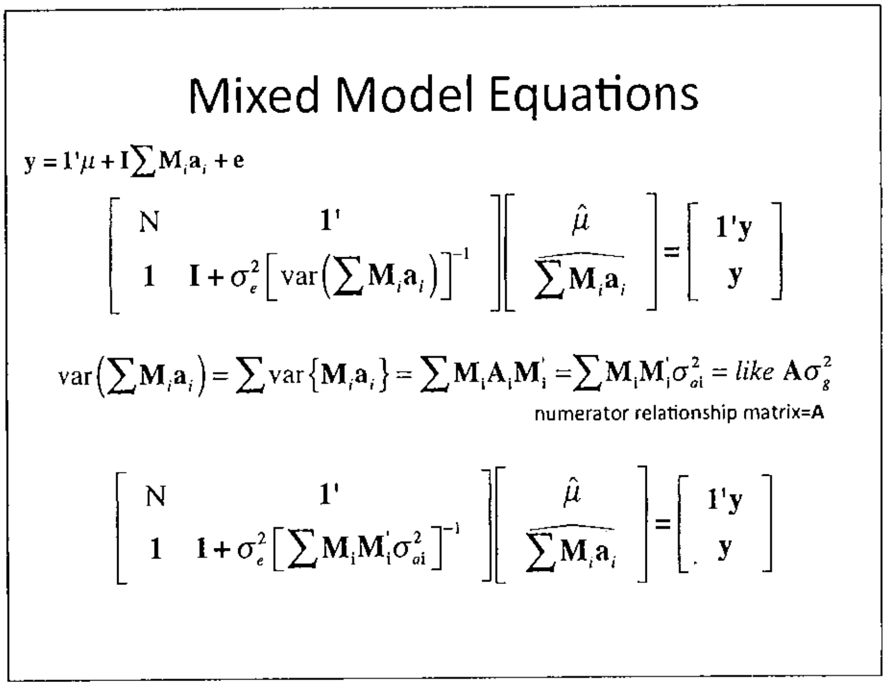
$$\begin{bmatrix} N & 1'Z \\ Z'1 & Z'Z + \sigma_e^2 G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} 1'y \\ Z'y \end{bmatrix}, \textit{for full rank } \mathbf{G} = \text{var}(\mathbf{u})$$

$y = 1'\mu + I\sum M_i a_i + e$

$$\begin{bmatrix} N & 1' \\ 1 & I + \sigma_e^2 \left[\text{var}\left(\sum M_i a_i\right)\right]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\sum M_i a_i} \end{bmatrix} = \begin{bmatrix} 1'y \\ y \end{bmatrix}$$

Order of MME is number of fixed effects plus number of animals

Consider the implications for 100-1,000 animals with 50,000 loci



Mixed Model Equations

$y = 1'\mu + I\sum M_i a_i + e$

$$\begin{bmatrix} N & 1' \\ 1 & I + \sigma_e^2 \left[\text{var}\left(\sum M_i a_i\right)\right]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\sum M_i a_i} \end{bmatrix} = \begin{bmatrix} 1'y \\ y \end{bmatrix}$$

$$\text{var}\left(\sum M_i a_i\right) = \sum \text{var}\left\{M_i a_i\right\} = \sum M_i A_i M_i' = \sum M_i M_i' \sigma_{ai}^2 = \textit{like } A\sigma_g^2$$

numerator relationship matrix=A

$$\begin{bmatrix} N & 1' \\ 1 & 1 + \sigma_e^2 \left[\sum M_i M_i' \sigma_{ai}^2\right]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\sum M_i a_i} \end{bmatrix} = \begin{bmatrix} 1'y \\ y \end{bmatrix}$$

An Equivalent Animal Model

$$\mathbf{M}_i\mathbf{M}_i'\sigma^2_{ai} \textit{ contains elements like} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} 2\sigma^2_{ai}$$

$\mathbf{M}_i\mathbf{M}_i'$ has order equal to number of animals (N)

$\sum \mathbf{M}_i\mathbf{M}_i'$ is summed over p loci

A diagonal element for a totally heterozygous animal is $1 \times 2\sum \sigma^2_{ai}$

Therefore σ^2_u in a typical animal model is (at least) $2\sum \sigma^2_{ai}$

A diagonal element for a totally homozygous animals is (1+F)=$2 \times 2\sum \sigma^2_{ai}$

A typical offdiagonal element is a weighted function of 0, 1 or 2

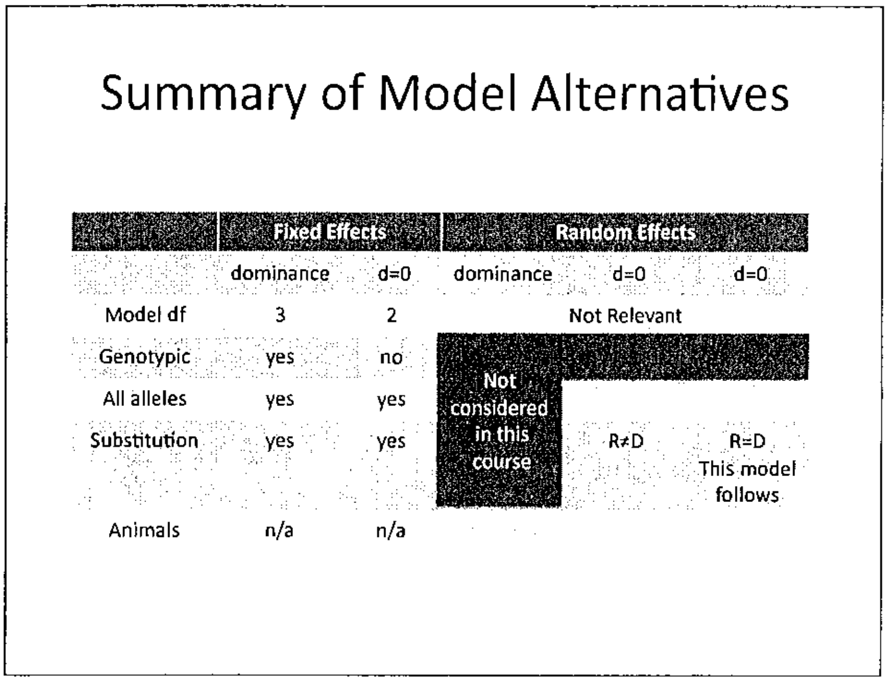
The number of 0's is the number of loci that the 2 animals are alternate homozygotes

The number of 2's is the number of loci that the 2 animals are the same homozygote

The number of 1's is N minus the number of 0's and 2's

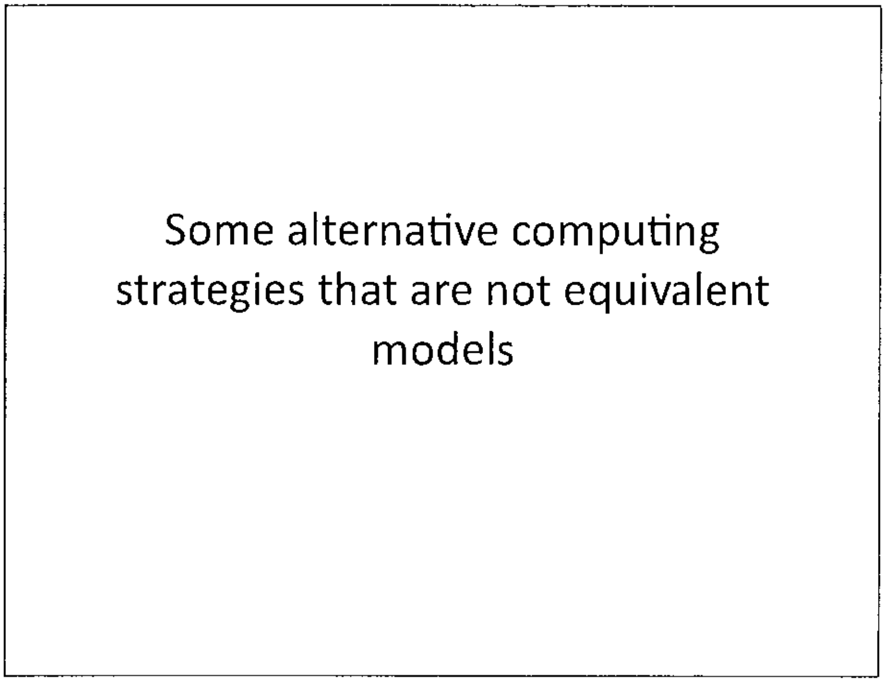
Non-inbred animal

- In the usual context, a non-inbred animal is IBS but not IBD (with a_{ii}=1)
- The fraction of homozygosity across loci is expected to be the sum over all loci of p^2+q^2 in the absence of inbreeding
- Such an animal would have an average diagonal of the genomic matrix >> 1



Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes			
Substitution	yes	yes		R≠D	R=D This model follows
Animals	n/a	n/a			



Some alternative computing strategies that are not equivalent models

Reconsider a single locus

$$\mathbf{y}=\mathbf{1}\mu+\mathbf{Ma}+\mathbf{e} \qquad or \qquad \mathbf{y}=\mathbf{1}\mu+\mathbf{m}_1a_1+\mathbf{m}_2a_2+\mathbf{e}$$

$$\begin{bmatrix} \mathrm{N} & \mathbf{1'm}_1 & \mathbf{1'm}_2 \\ \mathbf{m}_1'\mathbf{1} & \mathbf{m}_1'\mathbf{m}_1+\lambda & \mathbf{m}_1'\mathbf{m}_2 \\ \mathbf{m}_2'\mathbf{1} & \mathbf{m}_2'\mathbf{m}_1 & \mathbf{m}_2'\mathbf{m}_2+\lambda \end{bmatrix}\begin{bmatrix} \hat{\mu} \\ \hat{\mathrm{a}}_1 \\ \hat{\mathrm{a}}_2 \end{bmatrix}=\begin{bmatrix} \mathbf{1'y} \\ \mathbf{m}_1'\mathbf{y} \\ \mathbf{m}_2'\mathbf{y} \end{bmatrix}$$

For $\lambda=\frac{\sigma_e^2}{\sigma_a^2}$, *these MME have the same solution for* $\hat{a}_1-\hat{a}_2$ $(but\ not\ \hat{\mu})$ *as*

$$\begin{bmatrix} N & \mathbf{1'm}_1 \\ \mathbf{m}_1'\mathbf{1} & \mathbf{m}_1'\mathbf{m}_1+\lambda/2 \end{bmatrix}\begin{bmatrix} \hat{\mu}* \\ \widehat{a_1-a_2} \end{bmatrix}=\begin{bmatrix} \mathbf{1'y} \\ \mathbf{m}_1'\mathbf{y} \end{bmatrix}$$

As if we fitted $\mathbf{y}=\mathbf{1}\mu+\mathbf{m}_1a_1+\mathbf{e}$ *with different* λ

Hint of Identical Solutions

$\mathbf{y}=\mathbf{1}\mu+\mathbf{Ma}+\mathbf{e}$ (Model I), *with* $\mathbf{M'1}=2\mathbf{1}$

$E[y]=\mu$, $\mathrm{var}[\mathbf{y}]=\mathbf{MM'}\sigma_a^2+\mathbf{I}\sigma_e^2 \quad \lambda_I={\sigma_e^2}/{\sigma_a^2}$

$$\begin{aligned}\mathbf{y}&=\mathbf{1}\mu+\mathbf{m}_1a_1+\mathbf{m}_2a_2+\mathbf{e} \quad but\ \ \mathbf{m}_2=2\mathbf{1}-\mathbf{m}_1\\ &=\mathbf{1}\mu+\mathbf{m}_1a_1+(2\,\mathbf{1}-\mathbf{m}_1)a_2+\mathbf{e} \quad but\ \ 2a_2=k_2=\text{constant}\\ &=\mathbf{1}(\mu+k_2)+\mathbf{m}_1a_1-\mathbf{m}_1a_2+\mathbf{e}\\ &=\mathbf{1}(\mu+k_2)+\mathbf{m}_1(a_1-a_2)+\mathbf{e} \quad \text{(Model II)}\end{aligned}$$

$E[y]=(\mu+k_2)$, $\mathrm{var}[\mathbf{y}]=\mathbf{m}_1\mathbf{m}_1'2\sigma_a^2+\mathbf{I}\sigma_e^2 \quad \lambda_{II}={\sigma_e^2}/{2\sigma_a^2}={\lambda_I}/{2}$

Clearly the first and second moments are different in models I and II

Proof of Identical Solutions

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e} \\ &= \mathbf{1}\mu + \mathbf{MT}^{-1}\mathbf{Ta} + \mathbf{e} \\ &= \mathbf{1}\mu + \left[\mathbf{MT}^{-1}\right]\left[\mathbf{Ta}\right] + \mathbf{e} \end{aligned}$$

$$\mathbf{Ta} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2 \\ a_2 - a_1 \end{bmatrix}$$

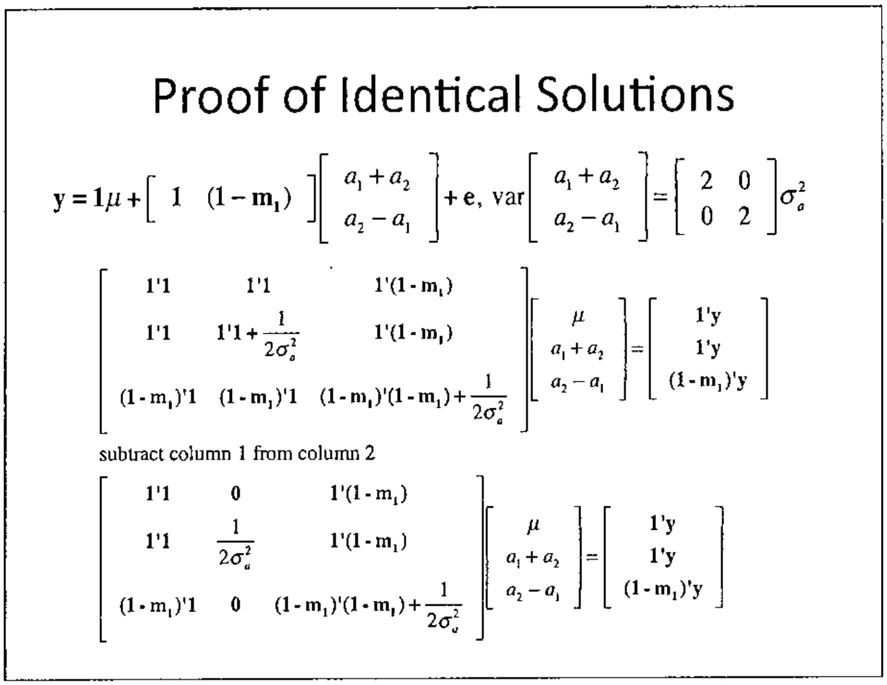
$$\textit{and } \mathbf{MT}^{-1} = \mathbf{M}\frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \mathbf{m}_1 + \mathbf{m}_2 & \mathbf{m}_2 - \mathbf{m}_1 \end{bmatrix}$$

Proof of Identical Solutions

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\mu + \left[\mathbf{MT}^{-1}\right]\left[\mathbf{Ta}\right] + \mathbf{e} \\ &= \mathbf{1}\mu + \frac{1}{2}\begin{bmatrix} \mathbf{m}_1 + \mathbf{m}_2 & \mathbf{m}_2 - \mathbf{m}_1 \end{bmatrix} \begin{bmatrix} a_1 + a_2 \\ a_2 - a_1 \end{bmatrix} + \mathbf{e} \end{aligned}$$

$$\textit{but } \mathbf{m}_1 + \mathbf{m}_2 = 2\mathbf{1} \quad \textit{and} \quad \mathbf{m}_2 - \mathbf{m}_1 = 2(\mathbf{1} - \mathbf{m}_1)$$

$$= \mathbf{1}\mu + \begin{bmatrix} \mathbf{1} & (\mathbf{1} - \mathbf{m}_1) \end{bmatrix} \begin{bmatrix} a_1 + a_2 \\ a_2 - a_1 \end{bmatrix} + \mathbf{e}$$



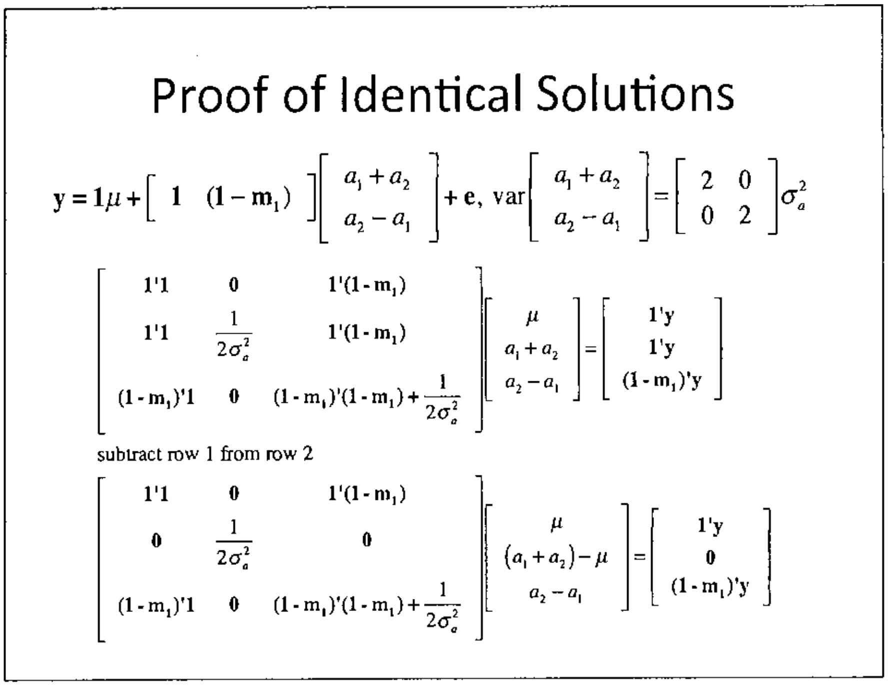
Proof of Identical Solutions

$$\mathbf{y} = \mathbf{1}\mu + \begin{bmatrix} 1 & (1-\mathbf{m}_1) \end{bmatrix}\begin{bmatrix} a_1+a_2 \\ a_2-a_1 \end{bmatrix} + \mathbf{e},\ \text{var}\begin{bmatrix} a_1+a_2 \\ a_2-a_1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\sigma_a^2$$

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{1'1} & \mathbf{1'(1-m_1)} \\ \mathbf{1'1} & \mathbf{1'1}+\frac{1}{2\sigma_a^2} & \mathbf{1'(1-m_1)} \\ \mathbf{(1-m_1)'1} & \mathbf{(1-m_1)'1} & \mathbf{(1-m_1)'(1-m_1)}+\frac{1}{2\sigma_a^2} \end{bmatrix}\begin{bmatrix} \mu \\ a_1+a_2 \\ a_2-a_1 \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{1'y} \\ \mathbf{(1-m_1)'y} \end{bmatrix}$$

subtract column 1 from column 2

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{0} & \mathbf{1'(1-m_1)} \\ \mathbf{1'1} & \frac{1}{2\sigma_a^2} & \mathbf{1'(1-m_1)} \\ \mathbf{(1-m_1)'1} & \mathbf{0} & \mathbf{(1-m_1)'(1-m_1)}+\frac{1}{2\sigma_a^2} \end{bmatrix}\begin{bmatrix} \mu \\ a_1+a_2 \\ a_2-a_1 \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{1'y} \\ \mathbf{(1-m_1)'y} \end{bmatrix}$$



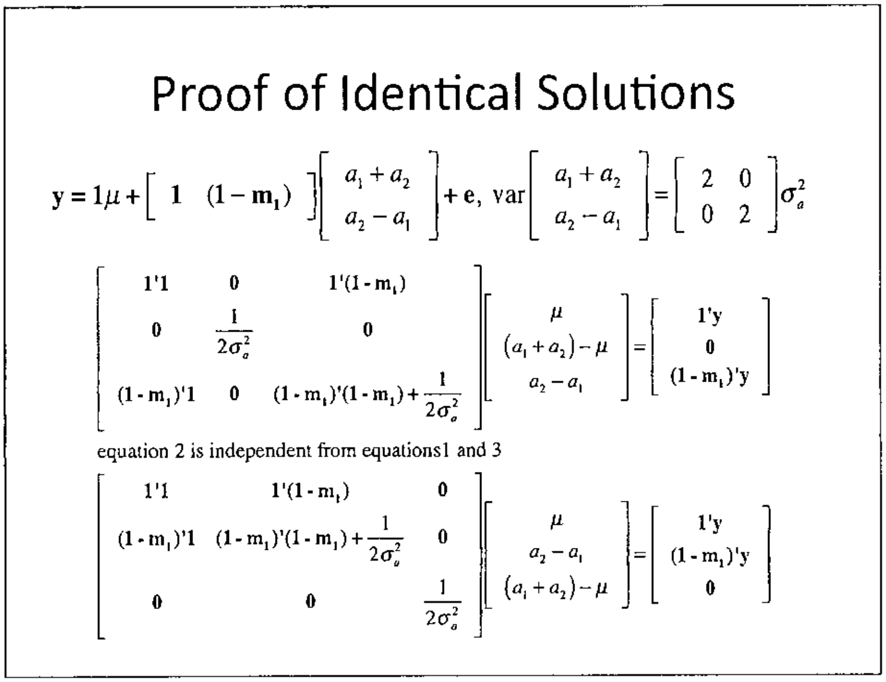
Proof of Identical Solutions

$$\mathbf{y} = \mathbf{1}\mu + \begin{bmatrix} 1 & (1-\mathbf{m}_1) \end{bmatrix}\begin{bmatrix} a_1+a_2 \\ a_2-a_1 \end{bmatrix} + \mathbf{e},\ \text{var}\begin{bmatrix} a_1+a_2 \\ a_2-a_1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\sigma_a^2$$

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{0} & \mathbf{1'(1-m_1)} \\ \mathbf{1'1} & \frac{1}{2\sigma_a^2} & \mathbf{1'(1-m_1)} \\ \mathbf{(1-m_1)'1} & \mathbf{0} & \mathbf{(1-m_1)'(1-m_1)}+\frac{1}{2\sigma_a^2} \end{bmatrix}\begin{bmatrix} \mu \\ a_1+a_2 \\ a_2-a_1 \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{1'y} \\ \mathbf{(1-m_1)'y} \end{bmatrix}$$

subtract row 1 from row 2

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{0} & \mathbf{1'(1-m_1)} \\ \mathbf{0} & \frac{1}{2\sigma_a^2} & \mathbf{0} \\ \mathbf{(1-m_1)'1} & \mathbf{0} & \mathbf{(1-m_1)'(1-m_1)}+\frac{1}{2\sigma_a^2} \end{bmatrix}\begin{bmatrix} \mu \\ (a_1+a_2)-\mu \\ a_2-a_1 \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{0} \\ \mathbf{(1-m_1)'y} \end{bmatrix}$$



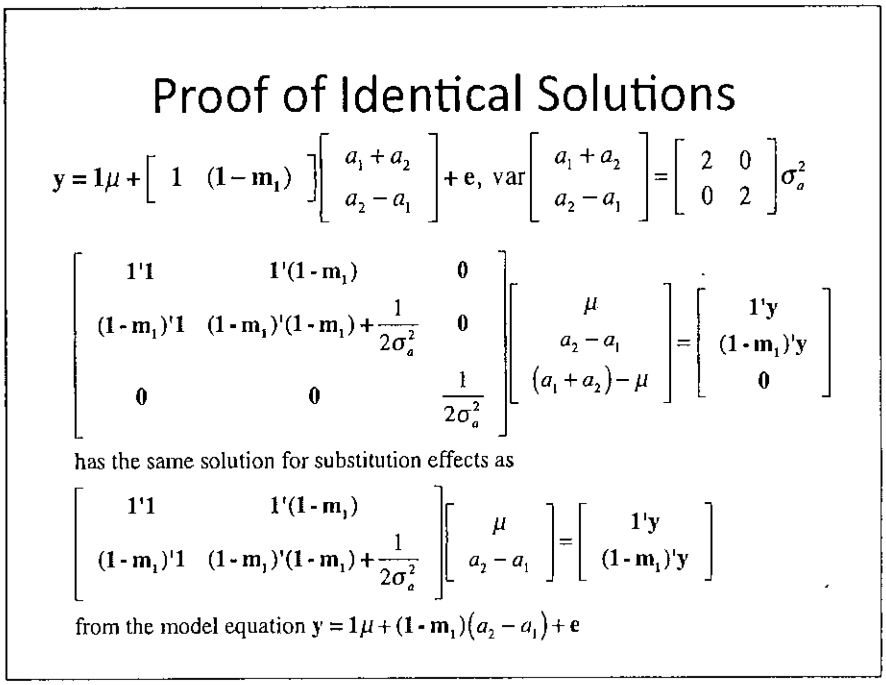
Proof of Identical Solutions

$$\mathbf{y}=\mathbf{1}\mu+\begin{bmatrix}\mathbf{1} & (\mathbf{1}-\mathbf{m}_1)\end{bmatrix}\begin{bmatrix}a_1+a_2\\ a_2-a_1\end{bmatrix}+\mathbf{e},\ \mathrm{var}\begin{bmatrix}a_1+a_2\\ a_2-a_1\end{bmatrix}=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}\sigma_a^2$$

$$\begin{bmatrix}\mathbf{1}'\mathbf{1} & \mathbf{0} & \mathbf{1}'(\mathbf{1}-\mathbf{m}_1)\\ \mathbf{0} & \frac{1}{2\sigma_a^2} & \mathbf{0}\\ (\mathbf{1}-\mathbf{m}_1)'\mathbf{1} & \mathbf{0} & (\mathbf{1}-\mathbf{m}_1)'(\mathbf{1}-\mathbf{m}_1)+\frac{1}{2\sigma_a^2}\end{bmatrix}\begin{bmatrix}\mu\\ (a_1+a_2)-\mu\\ a_2-a_1\end{bmatrix}=\begin{bmatrix}\mathbf{1}'\mathbf{y}\\ \mathbf{0}\\ (\mathbf{1}-\mathbf{m}_1)'\mathbf{y}\end{bmatrix}$$

equation 2 is independent from equations1 and 3

$$\begin{bmatrix}\mathbf{1}'\mathbf{1} & \mathbf{1}'(\mathbf{1}-\mathbf{m}_1) & \mathbf{0}\\ (\mathbf{1}-\mathbf{m}_1)'\mathbf{1} & (\mathbf{1}-\mathbf{m}_1)'(\mathbf{1}-\mathbf{m}_1)+\frac{1}{2\sigma_a^2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \frac{1}{2\sigma_a^2}\end{bmatrix}\begin{bmatrix}\mu\\ a_2-a_1\\ (a_1+a_2)-\mu\end{bmatrix}=\begin{bmatrix}\mathbf{1}'\mathbf{y}\\ (\mathbf{1}-\mathbf{m}_1)'\mathbf{y}\\ \mathbf{0}\end{bmatrix}$$



Proof of Identical Solutions

$$\mathbf{y}=\mathbf{1}\mu+\begin{bmatrix}\mathbf{1} & (\mathbf{1}-\mathbf{m}_1)\end{bmatrix}\begin{bmatrix}a_1+a_2\\ a_2-a_1\end{bmatrix}+\mathbf{e},\ \mathrm{var}\begin{bmatrix}a_1+a_2\\ a_2-a_1\end{bmatrix}=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}\sigma_a^2$$

$$\begin{bmatrix}\mathbf{1}'\mathbf{1} & \mathbf{1}'(\mathbf{1}-\mathbf{m}_1) & \mathbf{0}\\ (\mathbf{1}-\mathbf{m}_1)'\mathbf{1} & (\mathbf{1}-\mathbf{m}_1)'(\mathbf{1}-\mathbf{m}_1)+\frac{1}{2\sigma_a^2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \frac{1}{2\sigma_a^2}\end{bmatrix}\begin{bmatrix}\mu\\ a_2-a_1\\ (a_1+a_2)-\mu\end{bmatrix}=\begin{bmatrix}\mathbf{1}'\mathbf{y}\\ (\mathbf{1}-\mathbf{m}_1)'\mathbf{y}\\ \mathbf{0}\end{bmatrix}$$

has the same solution for substitution effects as

$$\begin{bmatrix}\mathbf{1}'\mathbf{1} & \mathbf{1}'(\mathbf{1}-\mathbf{m}_1)\\ (\mathbf{1}-\mathbf{m}_1)'\mathbf{1} & (\mathbf{1}-\mathbf{m}_1)'(\mathbf{1}-\mathbf{m}_1)+\frac{1}{2\sigma_a^2}\end{bmatrix}\begin{bmatrix}\mu\\ a_2-a_1\end{bmatrix}=\begin{bmatrix}\mathbf{1}'\mathbf{y}\\ (\mathbf{1}-\mathbf{m}_1)'\mathbf{y}\end{bmatrix}$$

from the model equation $\mathbf{y}=\mathbf{1}\mu+(\mathbf{1}-\mathbf{m}_1)(a_2-a_1)+\mathbf{e}$

More Alternatives

Previously $\mathbf{y} = \mathbf{1}(\mu + k_2) + \mathbf{m}_1(a_1 - a_2) + \mathbf{e}$

Note $\mathbf{m}_1$ (and $\mathbf{m}_2$) contain covariate values of 0, 1 or 2

another model with $\mathrm{k}_{12} = (a_1 - a_2)$ is

$$\mathbf{y} = \mathbf{1}(\mu + k_2 + k_{12}) + \mathbf{m}_1(a_1 - a_2) - \mathbf{1}(a_1 - a_2) + \mathbf{e}$$

$$\mathbf{y} = \mathbf{1}(\mu + k_2 + k_{12}) + (\mathbf{m}_1 - \mathbf{1})(a_1 - a_2) + \mathbf{e}$$

whereby the covariate values are now -1, 0 and 1

Computational Alternatives

	covariates
$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e}$	0, 1, 2 and 2, 1, 0
$\mathbf{y} = \mathbf{1}(\mu + k_2) \quad + \mathbf{m}_1(a_1 - a_2) \quad + \mathrm{e}$	0, 1, 2
$\mathbf{y} = \mathbf{1}(\mu + k_2 + k_{12}) + (\mathbf{m}_1 - \mathbf{1})(a_1 - a_2) + \mathrm{e}$	-1, 0, 1
$\mathbf{y} = \mathbf{1}(\mu + k_1) \quad + \mathbf{m}_2(a_2 - a_1) \quad + \mathrm{e}$	2, 1, 0
$\mathbf{y} = \mathbf{1}(\mu + k_1 + k_{21}) + (\mathbf{m}_2 - \mathbf{1})(a_2 - a_1) + \mathrm{e}$	1, 0, -1

All these models have different E[y]
All these models have identical predictions of random effects
Only the first model has the correct PEV for the random effect if e assumed diagonal

Consider the genetic part of var[y]

covariate

M

$\mathbf{m}_1$

$\mathbf{m}_1$ - 1

$\mathbf{m}_2$

$\mathbf{m}_2$ - 1

genetic variance (**ZGZ'**)

$$\mathbf{M} = \begin{bmatrix} \mathbf{m}_1 & \mathbf{m}_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 2 \end{bmatrix}$$

$$\mathrm{var}\left[\mathbf{Ma}\right] = \mathbf{MAM'} = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 2 & 4 \end{bmatrix} \sigma_a^2$$

$$\mathrm{var}\left[\mathbf{m}_1(a_1 - a_2)\right] = 2\sigma_a^2 \mathbf{m}_1 \mathbf{m}_1' = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} 2\sigma_a^2$$

$$\mathrm{var}\left[(\mathbf{m}_1 - 1)(a_1 - a_2)\right] = 2\sigma_a^2 (\mathbf{m}_1 - 1)(\mathbf{m}_1 - 1)' = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} 2\sigma_a^2$$

$$\mathrm{var}\left[\mathbf{m}_2(a_2 - a_1)\right] = 2\sigma_a^2 \mathbf{m}_2 \mathbf{m}_2' = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 4 \end{bmatrix} 2\sigma_a^2$$

$$\mathrm{var}\left[(\mathbf{m}_2 - 1)(a_2 - a_1)\right] = 2\sigma_a^2 (\mathbf{m}_2 - 1)(\mathbf{m}_2 - 1)' = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} 2\sigma_a^2$$

These are typically singular, unless there are more loci than animals

Animal Model Counterpart

Any full rank inverse of the following can be used in place of $\mathbf{A}^{-1}\sigma_a^2$ *in MME to predict animal merit*

$$\sum \mathbf{M}_i \mathbf{M}_i' \sigma_{ai}^2 = \sum \left(m_{1i} m_{1i}' + m_{2i} m_{2i}'\right) \sigma_{ai}^2$$

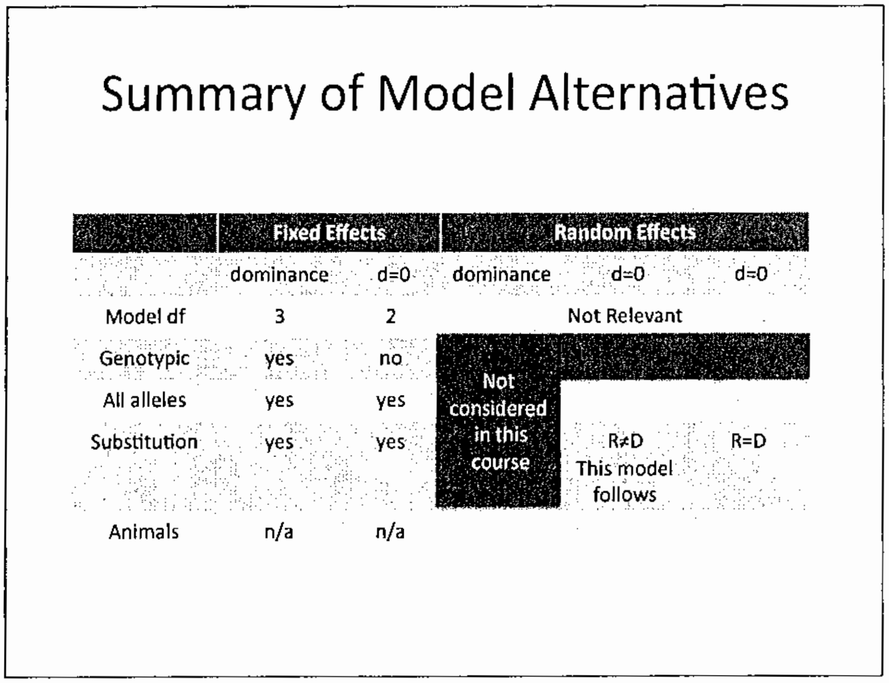
$$\sum m_{1i} m_{1i}' 2\sigma_{ai}^2$$

$$\sum m_{2i} m_{2i}' 2\sigma_{ai}^2$$

$$\sum \left(m_{1i} - 1\right)\left(m_{1i} - 1\right)' 2\sigma_{ai}^2$$

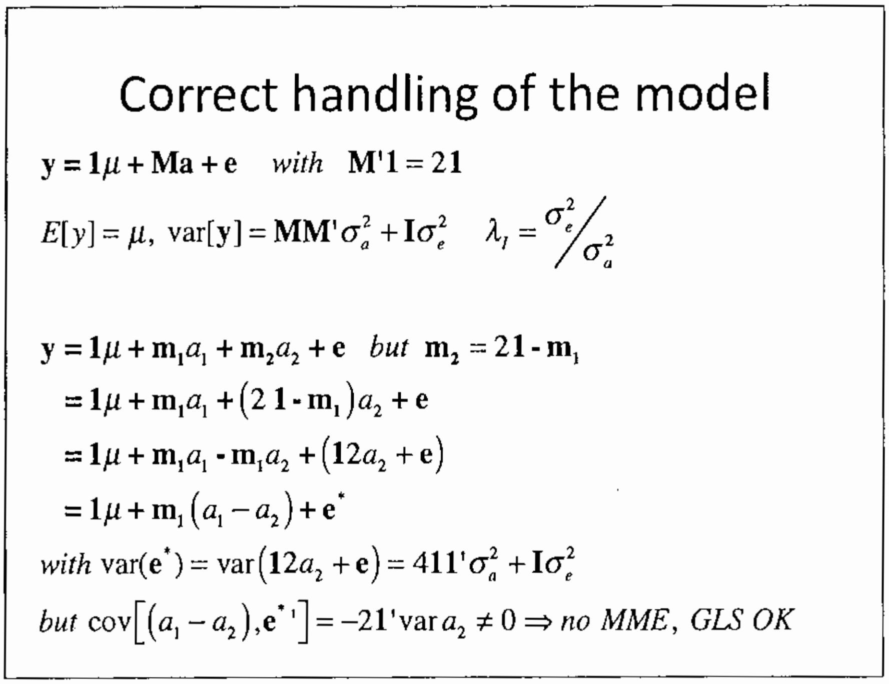
$$\sum \left(m_{2i} - 1\right)\left(m_{2i} - 1\right)' 2\sigma_{ai}^2$$

Only the first can be used for PEV or r^2



Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes			
Substitution	yes	yes		R≠D This model follows	R=D
Animals	n/a	n/a			



Correct handling of the model

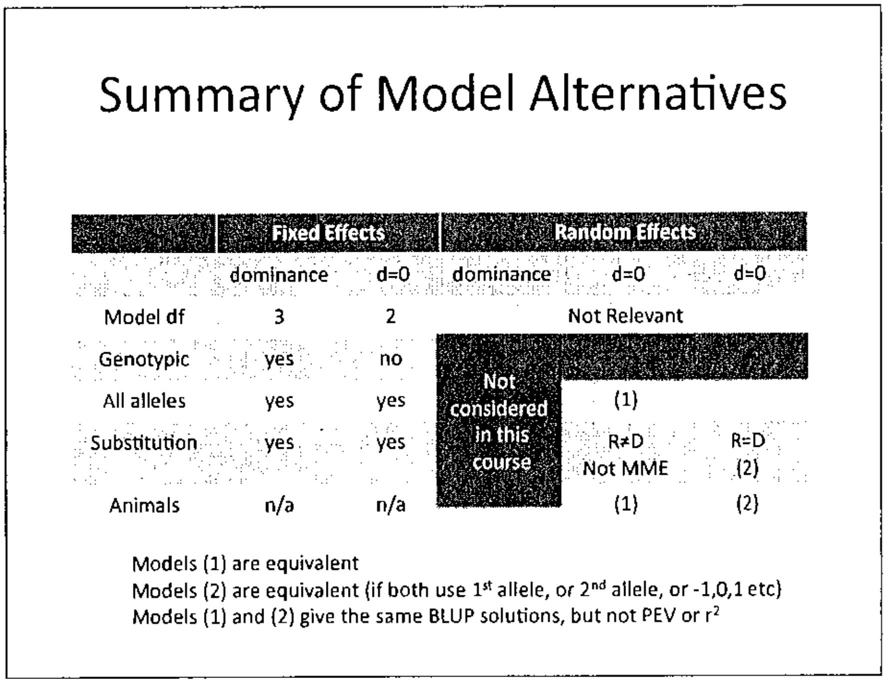
$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e} \quad \textit{with} \quad \mathbf{M'1} = 2\mathbf{1}$$

$$E[y] = \mu,\ \operatorname{var}[\mathbf{y}] = \mathbf{MM'}\sigma_a^2 + \mathbf{I}\sigma_e^2 \quad \lambda_I = {}^{\sigma_e^2}\!\big/_{\sigma_a^2}$$

$$\begin{aligned}\mathbf{y} &= \mathbf{1}\mu + \mathbf{m}_1 a_1 + \mathbf{m}_2 a_2 + \mathbf{e} \quad \textit{but} \ \ \mathbf{m}_2 = 2\mathbf{1} - \mathbf{m}_1 \\ &= \mathbf{1}\mu + \mathbf{m}_1 a_1 + (2\,\mathbf{1} - \mathbf{m}_1)a_2 + \mathbf{e} \\ &= \mathbf{1}\mu + \mathbf{m}_1 a_1 - \mathbf{m}_1 a_2 + (\mathbf{1}2a_2 + \mathbf{e}) \\ &= \mathbf{1}\mu + \mathbf{m}_1(a_1 - a_2) + \mathbf{e}^*\end{aligned}$$

$$\textit{with} \ \operatorname{var}(\mathbf{e}^*) = \operatorname{var}(\mathbf{1}2a_2 + \mathbf{e}) = 4\mathbf{11'}\sigma_a^2 + \mathbf{I}\sigma_e^2$$

$$\textit{but} \ \operatorname{cov}\left[(a_1 - a_2), \mathbf{e}^{*\prime}\right] = -2\mathbf{1'}\operatorname{var} a_2 \neq 0 \Rightarrow \textit{no MME, GLS OK}$$



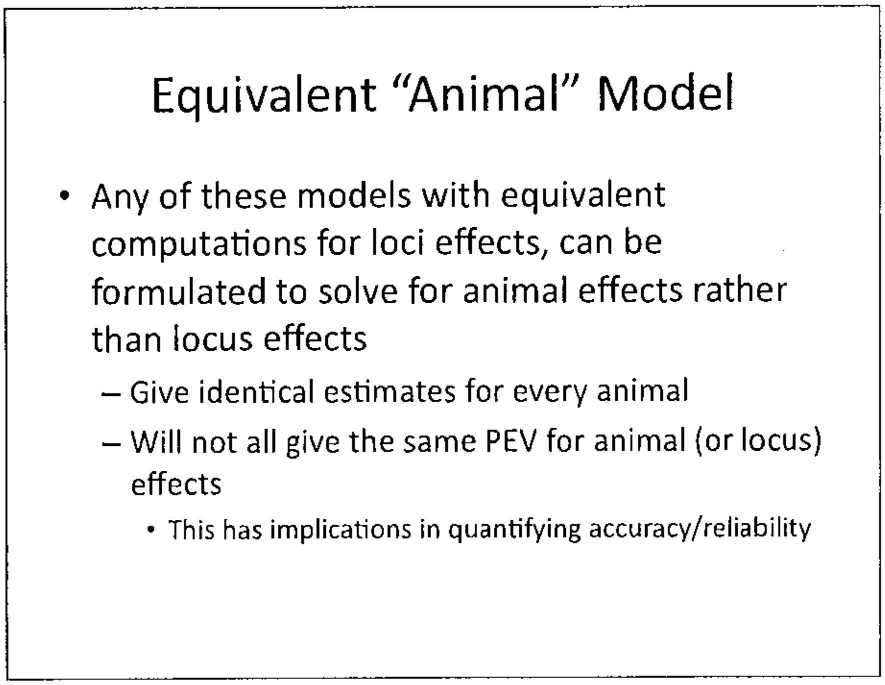
Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes		(1)	
Substitution	yes	yes		R≠D Not MME	R=D (2)
Animals	n/a	n/a		(1)	(2)

Models (1) are equivalent

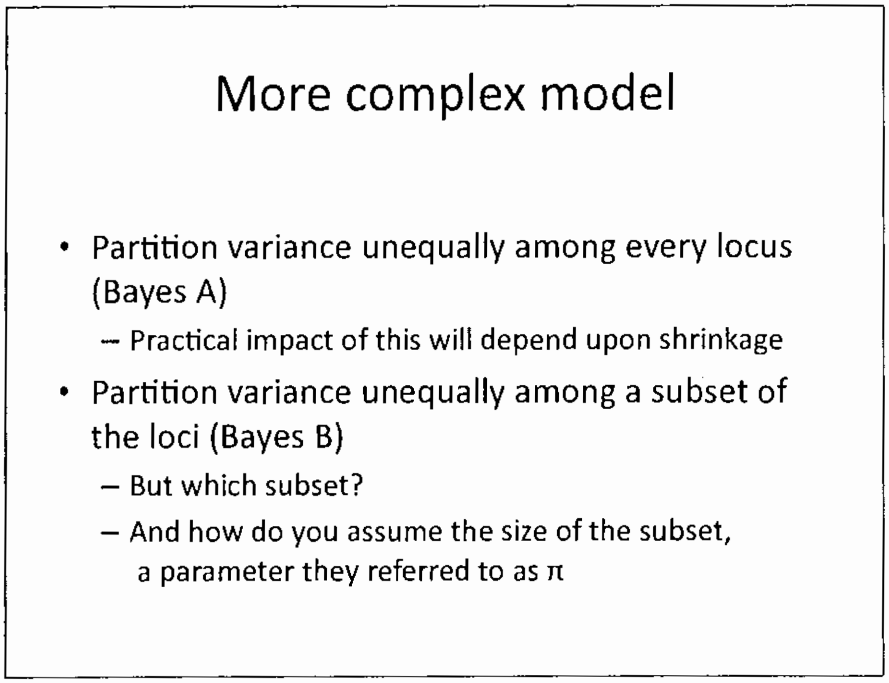
Models (2) are equivalent (if both use 1st allele, or 2nd allele, or -1,0,1 etc)

Models (1) and (2) give the same BLUP solutions, but not PEV or r^2



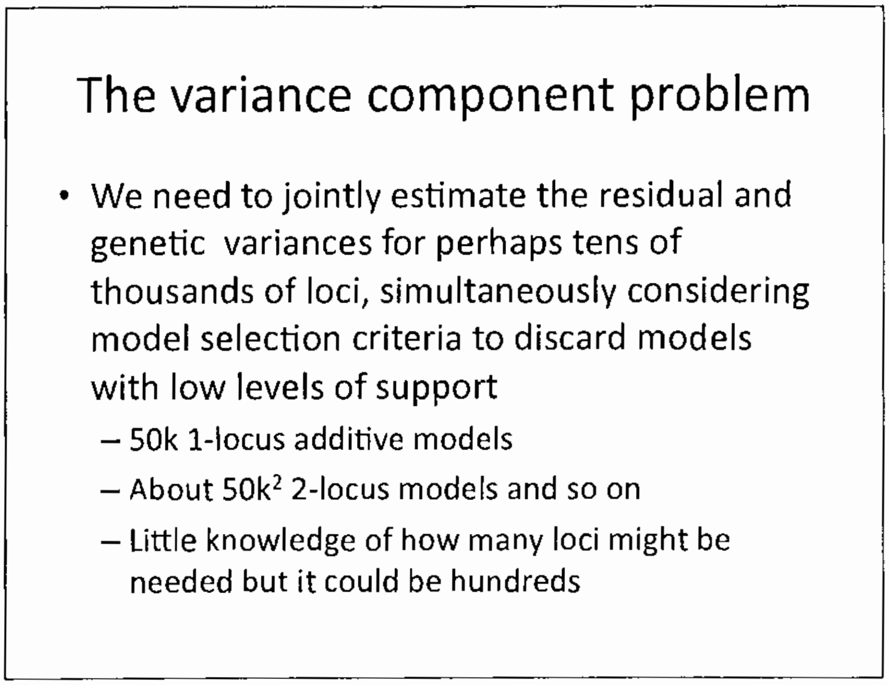
Equivalent "Animal" Model

- Any of these models with equivalent computations for loci effects, can be formulated to solve for animal effects rather than locus effects
  - Give identical estimates for every animal
  - Will not all give the same PEV for animal (or locus) effects
    - This has implications in quantifying accuracy/reliability



More complex model

- Partition variance unequally among every locus (Bayes A)
  - Practical impact of this will depend upon shrinkage
- Partition variance unequally among a subset of the loci (Bayes B)
  - But which subset?
  - And how do you assume the size of the subset, a parameter they referred to as π

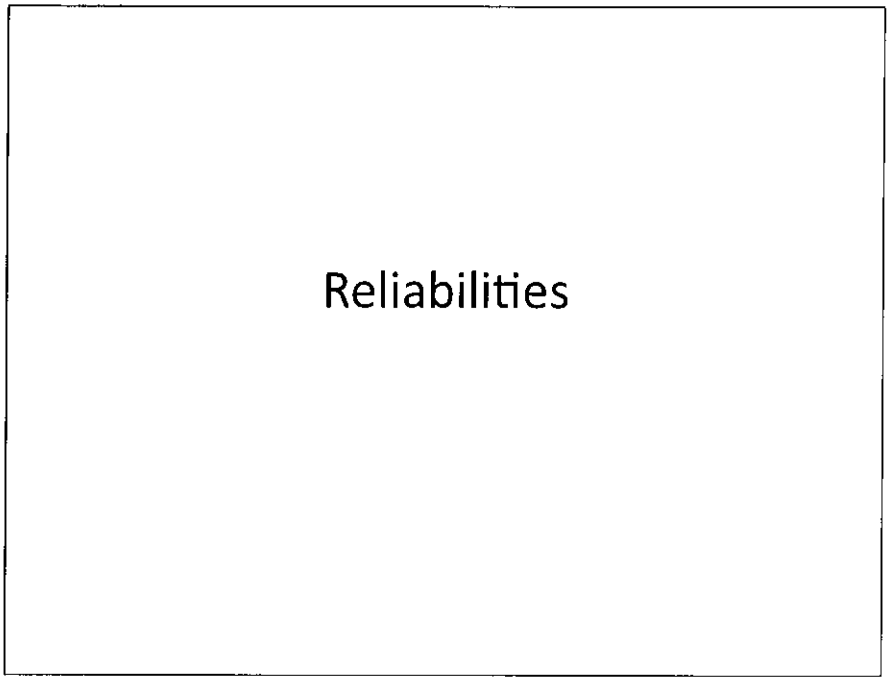


The variance component problem

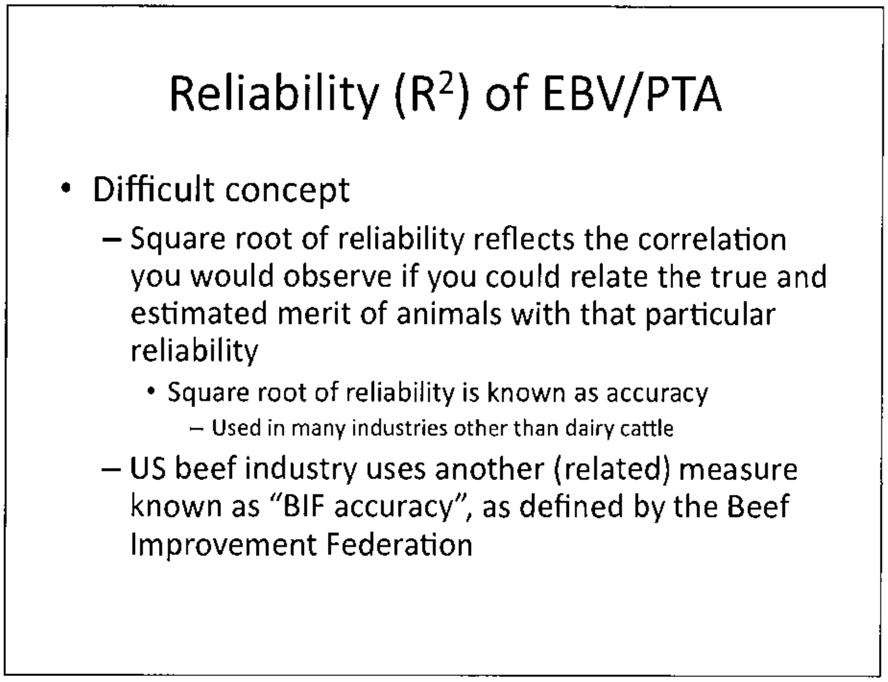
- We need to jointly estimate the residual and genetic variances for perhaps tens of thousands of loci, simultaneously considering model selection criteria to discard models with low levels of support
  - 50k 1-locus additive models
  - About $50k^2$ 2-locus models and so on
  - Little knowledge of how many loci might be needed but it could be hundreds

Fitted Model

- We will use the model that fits a substitution effect for each locus, recognizing that we cannot use the equations for estimating reliabilities
  - Equations are too big anyway
  - Bayesian posteriors can be used for reliability of SNP effects



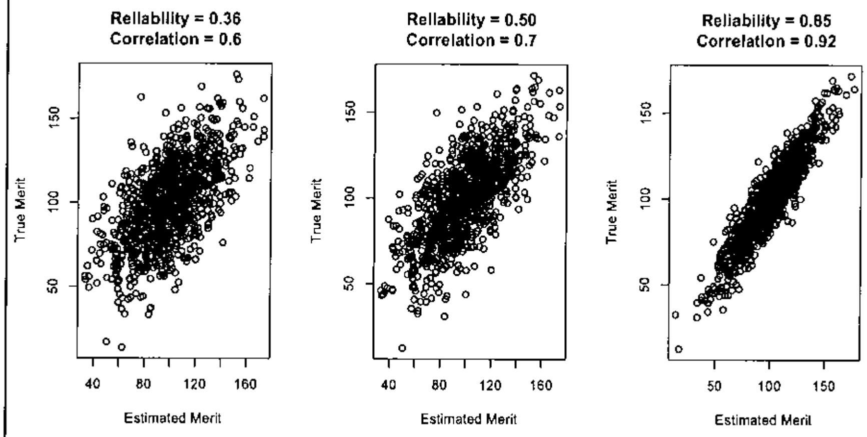
Reliabilities



Reliability (R^2) of EBV/PTA

- Difficult concept
  - Square root of reliability reflects the correlation you would observe if you could relate the true and estimated merit of animals with that particular reliability
    - Square root of reliability is known as accuracy
      - Used in many industries other than dairy cattle
  - US beef industry uses another (related) measure known as "BIF accuracy", as defined by the Beef Improvement Federation

High & Low Reliability



Reliability = 0.36
Correlation = 0.6

Reliability = 0.50
Correlation = 0.7

Reliability = 0.85
Correlation = 0.92

Reliability of non-genotyped Offspring

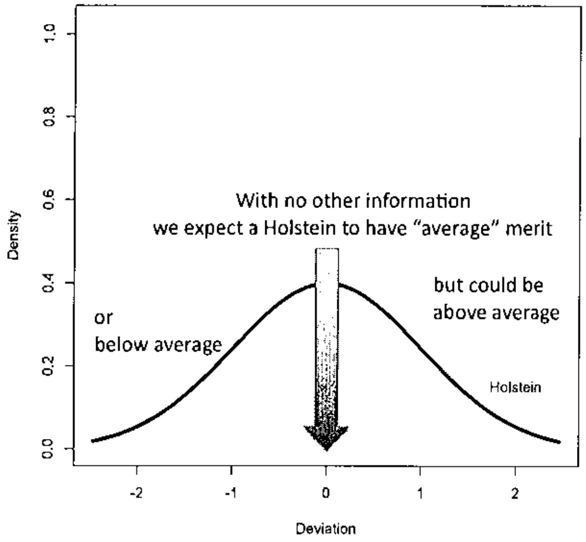
$$R^2_{\text{offspring}} = \frac{R^2_{\text{sire}} + R^2_{dam}}{4}$$

Reliability increases with individual records or offspring

Reliability of individual
with accurate sire is at most 0.25
with accurate sire & dam is at most 0.5

Reliability of EBV/PTA

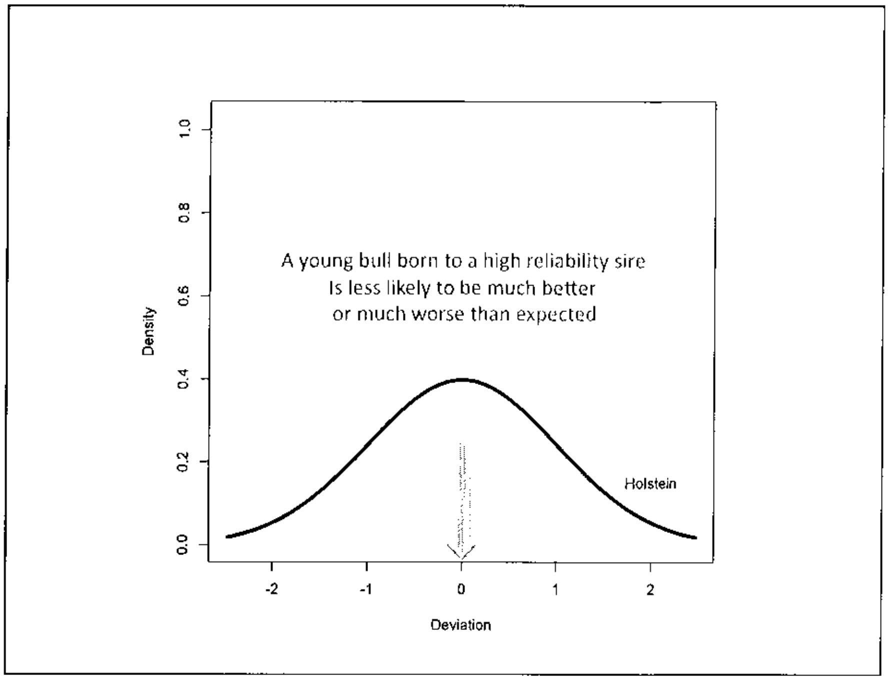
- Unreliability is easier to understand
  - (100-reliability) is the percentage of genetic variation that cannot be explained from knowledge of the pedigree & performance information (or pedigree, performance and genomic information)



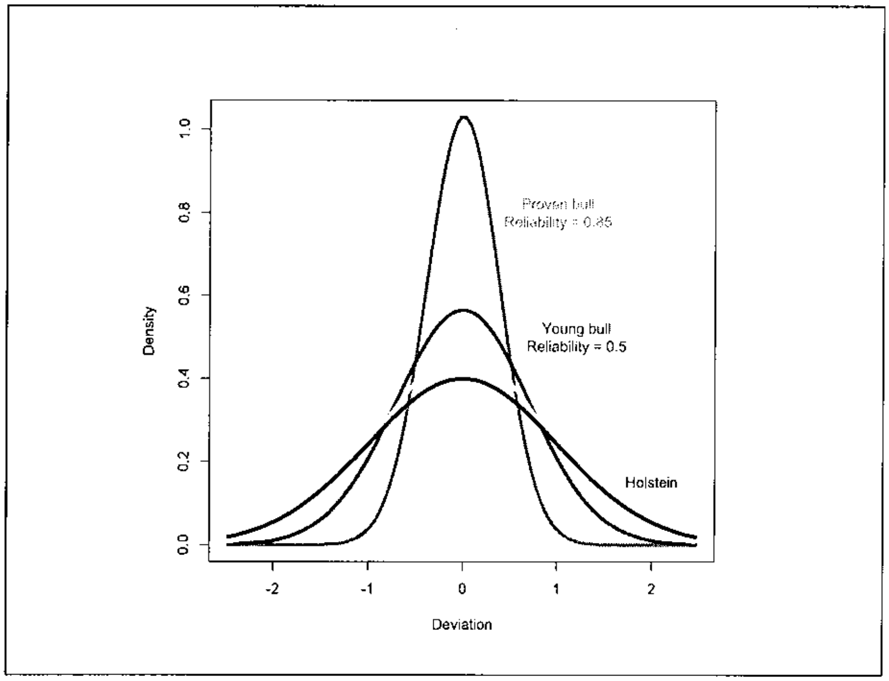
With no other information we expect a Holstein to have "average" merit

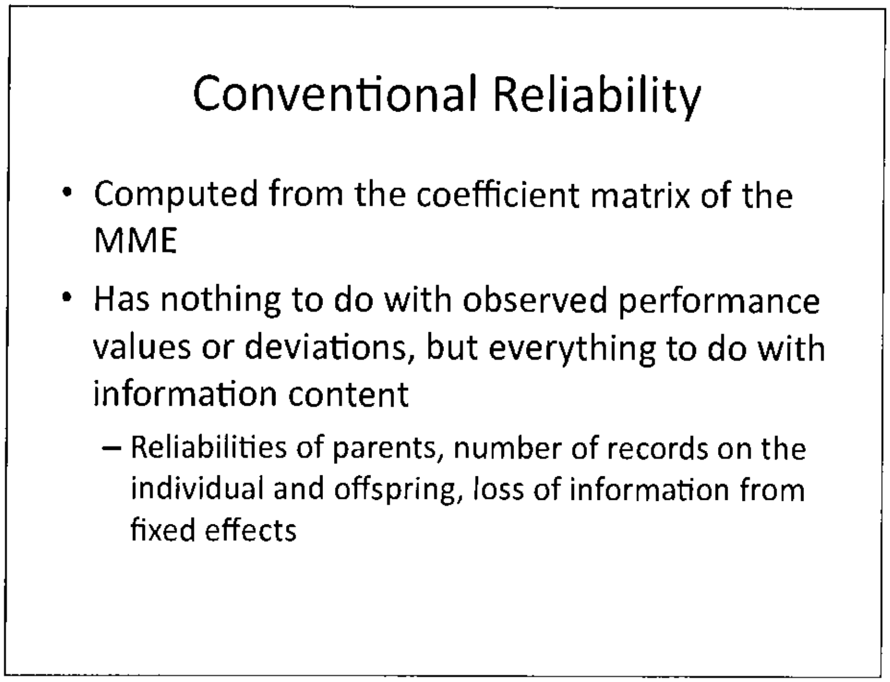
or below average

but could be above average



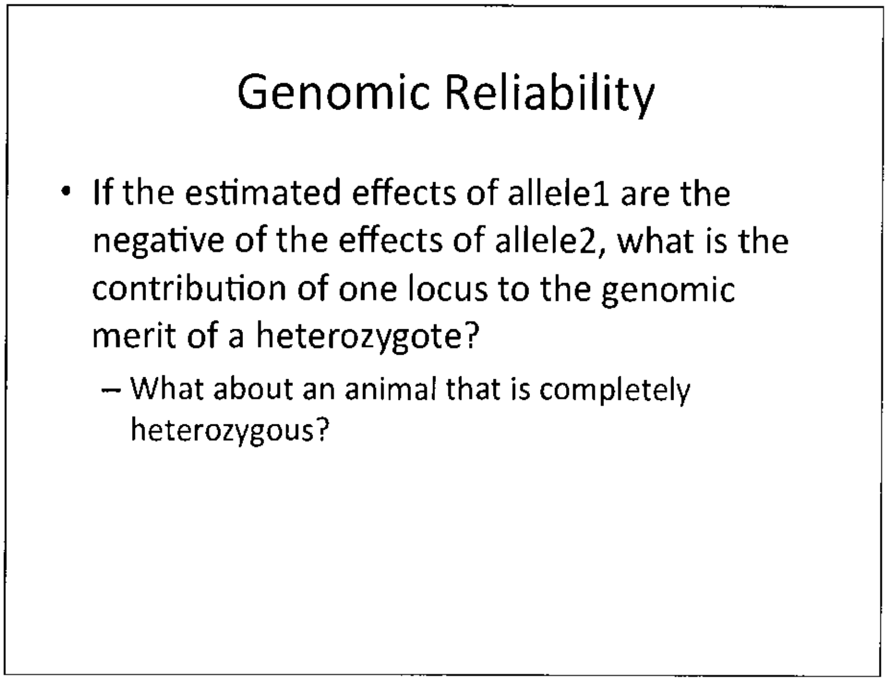
A young bull born to a high reliability sire
Is less likely to be much better
or much worse than expected





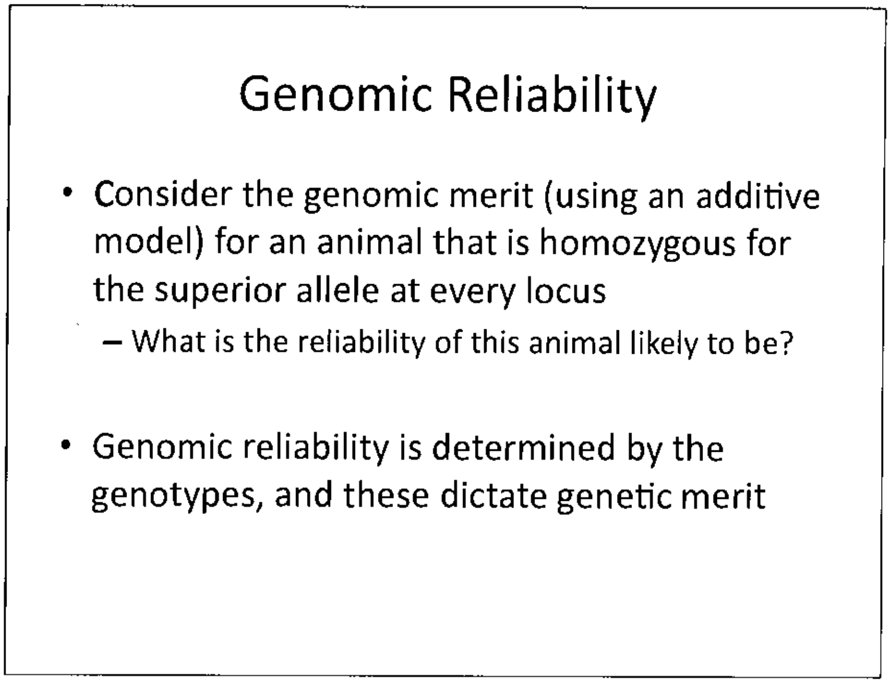
Conventional Reliability

- Computed from the coefficient matrix of the MME
- Has nothing to do with observed performance values or deviations, but everything to do with information content
  - Reliabilities of parents, number of records on the individual and offspring, loss of information from fixed effects



Genomic Reliability

- If the estimated effects of allele1 are the negative of the effects of allele2, what is the contribution of one locus to the genomic merit of a heterozygote?
  - What about an animal that is completely heterozygous?



Genomic Reliability

- Consider the genomic merit (using an additive model) for an animal that is homozygous for the superior allele at every locus
  - What is the reliability of this animal likely to be?
- Genomic reliability is determined by the genotypes, and these dictate genetic merit

Laboratory 1

The objective of this laboratory session is to gain familiarity with the R programming language and the mixed linear models that we will be using in the Bayesian analyses later in the course.

Exercise 1

The lecture notes introduced the equations for generalized least squares (GLS). The GLS equation(s) for the model we discussed in the lecture are

$$\hat{\mathbf{b}}^0 = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\right), \text{ for } \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}.$$

These equations are useful as **V** is typically full rank, but are not practical in many situations where **V** is large. In this example with just the mean fitted as the only fixed effect, the GLS equation will be a scalar form.

In order to form **V**, you will need to know **G** and **R**.

Create a small Hendersonian data set by constructing a vector **y** of phenotypic observations (no more than 6 observations). Create a corresponding **X** matrix to represent the incidence matrix for the fixed effects. This matrix will have as many rows as there are observations in **y**, and as many columns as there are fixed effects in **b**. The minimum configuration for **X** would be a vector of 1's that would correspond to a model that included an overall mean. Other alternatives for **X** might be to include a vector of covariates (eg age of the animal at measurement) or a class variable such as a fixed effect for the sex of the measured animal.

Construct a **G** matrix that will be square and have order equal to the number of animals in the pedigree file. For ease of viewing, the order of **G** should not exceed 6. The **G** matrix is the variance-covariance matrix of the fitted random effects, such as the breeding values. In that case, **G** will be the product of the numerator relationship or **A** matrix, and the scale additive genetic variance. Form **A** for some simple pedigree and assume a value of the additive genetic variance. Note that the pedigree might contain some animals that do not have observed phenotypes, so the length of y may be less than the order of **G**.

Construct an incidence matrix **Z**, that relates the observations in **y** to the corresponding breeding value in **u**. The matrix **Z** may be an identity matrix if all animals in the pedigree have a phenotypic record. More typically, **Z** has as many rows as there are records in y, and as many columns as there are animals in **u** (and therefore the **G** matrix).

Lastly, construct **R**, the variance-covariance matrix for the residual effects, which for independent and identically distributed residual effects will be an identity matrix of order equal to the length of y, multiplied by the scalar residual variance. Recall that the heritability is the ratio of the genetic variance over the phenotypic variance, and the phenotypic variance in this model is the sum of the additive genetic and residual variances, so the values you assume will imply a particular heritability.

Given defined values for all these vectors, matrices and constants, calculate the phenotypic variance-covariance matrix **V**, and then solve the GLS equations to obtain best linear unbiased estimates (BLUEs) of the fixed effects. Use the BLUEs to adjust the phenotypic records and form deviations, that you can then use to compute the best linear unbiased predictions (BLUP) of the random effects as a linear function of these deviations, as described below. Note that this form of obtaining BLUP works with a singular **G** matrix.

The equations to obtain BLUP estimates are

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^0\right).$$

Be sure to save all your steps so you can immediately repeat your calculations with a modified dataset or different parameters. Print out and inspect the results of all your calculations.

Exercise 2

Repeat the same exercise as above, but this time estimate the BLUEs and predict the BLUPs by setting up and solving the mixed model equations. The answers should be identical to those you obtained using GLS. The mixed model equations are shown below.

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}^0 \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Exercise 3

Obtain the variance of the estimated BLUP effects, and the prediction error variance. These values require elements of the inverse of the mixed model coefficient matrix. We will use the following notation

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}$$

and the corresponding partitions of the inverse are

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix}.$$

In relation to random effects, we need only concern ourselves with the $\mathbf{C}^{22}$ partition of the inverse coefficient matrix. Note however that the entire coefficient matrix must be inverted to obtain the partition of interest. From this partition you have the prediction error variance-covariance matrix. That is,

$$\text{var}[\mathbf{u} - \hat{\mathbf{u}}] = \mathbf{C}^{22}$$

$\text{var}[\hat{\mathbf{u}}] = \mathbf{G} - \mathbf{C}^{22}$, and recall that $\text{var}[\mathbf{u}] = \mathbf{G}$.
A common unitfree measure of how well we have estimated the BLUP is the square of the correlation between the true and estimated effect. Since the true effects are not known, this cannot be calculated directly, but is a function of the **G** and $\mathbf{C}^{22}$ matrices. Specifically, $r^2 = \frac{\text{var}[\hat{\mathbf{u}}]}{\text{var}[\mathbf{u}]} = \frac{diag\left[\mathbf{G} - \mathbf{C}^{22}\right]}{diag\left[\mathbf{G}\right]}$ for best linear predictions (BLP) and best linear unbiased predictions (BLUP).

Exercise 4

In many circumstances we are interested in linear combinations of random effects. For example, we might want to know the BLUP and the r^2 of a team of sires rather than an individual. Alternatively, we might be interested in the contrast or difference between one or more alternative sires or teams. To compute these, we need to construct a relevant vector of contrasts that we will denote as **k**. For example, to predict the superiority of sire 1 over sire 2, for $\mathbf{u}' = \begin{bmatrix} u_1 & u_2 & u_3 & u_4 \end{bmatrix}$, we would form $\mathbf{k}' = \begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix}$. To compare a team of the first two sires to the second two sires we would use $\mathbf{k}' = \begin{bmatrix} 0.5 & 0.5 & -0.5 & -0.5 \end{bmatrix}$. Both of these contrasts can be considered simultaneously by stacking them up the rows of **k'** together in a matrix, $\mathbf{K} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0.5 & 0.5 & -0.5 & -0.5 \end{bmatrix}$.
The BLUP of $\mathbf{k'u}$ is simply obtained as $\mathbf{k'}\hat{\mathbf{u}}$, and $\text{var}(\mathbf{k'u}) = \mathbf{k'Gk}$,
$\text{var}(\mathbf{k'}\hat{\mathbf{u}}) = \mathbf{k'}\left[\mathbf{G} - \mathbf{C}^{22}\right]\mathbf{k}$.
Construct some linear combinations, and estimate the prediction error variance and r^2 for these linear combinations.

Useful R commands for this exercise.

array()	used to form a vector
matrix()	used to form a matrix
dim()	used to determine the dimension of an object (eg vector or matrix)
diag()	used to construct a diagonal matrix or extract the diagonal elements of a matrix
t()	transpose a matrix
%*%	used to perform matrix (or matrix-vector) multiplications
solve()	used to solve a set of equations or to obtain the inverse of a matrix
rbind()	used to join objects in different rows
cbind()	used to join objects into columns
?	used for syntax help, e.g., ?solve

Laboratory 2

Consider the dataset in Table 1, from p110 Ben Hayes course notes. We will use this dataset to explore some alternative models for fitting SNP effects. The columns include the allele calls at each marker locus (M1, M2 and M3), followed by the covariate that represent the number of 1 (a1, b1 and c1) or 2 (a2, b2 and c2) alleles at each locus (designated A, B and C).

Animal	phenotype	M1	M2	M3	a1	a2	b1	b2	c1	c2
1	9.68	22	21	11	0	2	1	1	2	0
3	2.29	12	22	22	1	1	0	2	0	2
20	0.81	11	21	12	2	0	1	1	1	1
4	3.42	11	21	11	2	0	1	1	2	0
2	5.69	22	22	22	0	2	0	2	0	2
5	5.92	21	11	11	1	1	2	0	2	0
6	2.82	21	21	22	1	1	1	1	0	2
7	5.07	22	21	22	0	2	1	1	0	2
8	8.92	22	22	11	0	2	0	2	2	0
9	2.4	11	22	12	2	0	0	2	1	1
10	9.01	22	22	11	0	2	0	2	2	0
11	4.24	12	12	21	1	1	1	1	1	1
12	6.35	22	11	12	0	2	2	0	1	1
13	8.92	22	12	11	0	2	1	1	2	0
14	-0.64	11	22	22	2	0	0	2	0	2
15	5.95	21	11	11	1	1	2	0	2	0
16	6.13	12	21	11	1	1	1	1	2	0
17	6.72	21	21	11	1	1	1	1	2	0
18	4.86	12	21	12	1	1	1	1	1	1
19	6.36	22	22	22	0	2	0	2	0	2
21	9.67	22	12	11	0	2	1	1	2	0
22	7.74	22	21	12	0	2	1	1	1	1
23	1.45	11	22	21	2	0	0	2	1	1
24	1.22	11	21	21	2	0	1	1	1	1
25	-0.52	11	22	22	2	0	0	2	0	2

This data first needs to be read into R. The command getwd() will show the working directory. The datafile needs to be located in the working directory. You could either copy it there, navigate to the working directory from the menu options, or change the working directory using the setwd("dirname") command, where dirname is the path to the working directory. The command dir() will show the files in the working directory.

A simple R script will be provided with the following commands to read the datafile.

```
genomicdata <- read.table("BenHayesp110.txt", header=TRUE)
```

will read the text file into a table object in R. Typing the name of the table (ie genomicdata) or using the command print(genomicdata) will display the information if the read.table command was successful. The commands dim() or str() will also provide details of the object if you place the object name between the brackets. The named columns of the table can be accessed using the name of the table, followed by a $ sign, followed by the name of the column. For example,

```
ytmp     <- matrix(genomicdata$phenotype, ncol=1)
Ztmp   <- as.matrix( cbind(genomicdata$a1, genomicdata$a2,
genomicdata$b1, genomicdata$b2,genomicdata$c1,genomicdata$c2))
```

will read in a potential y vector and Z matrix.

We will be fitting some models where rank is an issue for certain analyses. For example, in least squares models, we need to have at least as many animals as we have effects. This is typically not an issue if the fitted effects are treated as random. However, for equivalent models that fit animal effect using SNP genotypes to form relationships, the genomic relationship matrix will not be full rank unless there are at least as many SNP effects fitted as there are animals. For this reason, in different models we will use different subsets of the complete **y** and **Z** vector. The variable nanim sets the number of animals to be used. The following lines will set up the example to use the first thirteen animals in the datafile.

```
nanim   <- 13

y      <- matrix(ytmp[1:nanim])
X      <- matrix(1,nanim)
Z      <- Ztmp[1:nanim,]
neffects <- dim(Z)[2]
nfix    <- dim(X)[2]
nloci   <- neffects/2
istart  <-nfix+1     #these are pointers to assist in extracting subvectors
iend    <-nfix+neffects
```

Example 1: Fitting both alleles at the three loci as random effects using GLS.

The GLS equation(s) for the model we discussed in the lecture are

$$\hat{\mathbf{b}}^0 = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\right), \text{ for } \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}.$$

These equations are useful as **V** is typically full rank, but are not practical in many situations where **V** is large. In this example with just the mean fitted as the only fixed effect, the GLS equation will be a scalar form.

In order to form **V**, you will need to know **G** and **R**. Suppose the residuals are homogeneous and uncorrelated. We will use a residual variance of 1. **R** can be formed using the diag command.

```
R	<-diag(sigmasqe,nanim)
```

The incidence matrix **Z** has 6 columns – one for each of the allelic effects. Suppose the three loci have different variance – say 2, 4 and 3, respectively. Create a **G** matrix of order 6 with columns corresponding to the columns in **Z**. Inspect **V**. You will need to use commands for transpose (eg t(X)), matrix multiplication (eg, X %*% Vinv), and matrix inversion (eg solve (V)). Take advantage of the help facility in R, using commands such as ?solve or ?t() for any commands you are unsure of. Inspect the intermediate calculations and record the subsequent results.

Be sure to save all your steps so you can immediately repeat your calculations with a modified dataset or different parameters.

Estimate the fixed effects by solving the GLS equations. Print out the result(s). The BLUPs of the random effects can then be obtained from selection index principles, but adjusting the phenotypic records with the GLS estimates of the fixed effects (rather then the true values as is required in selection index). That is, solve

$$\hat{\mathbf{a}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^{0}\right).$$

Note that the estimates of the allelic effects sum to zero, even though no such constraint was actively used. This is a feature of mixed models in certain circumstances.

Calculate the substitution effects by forming a contrast vector (k) with order equal to the order of $\hat{\mathbf{a}}$, that contains all zeros except elements 1 and -1 corresponding to the first and second allele at a locus, and then compute the linear function $\mathbf{k}'\hat{\mathbf{a}}$. Record the results. You can align (using cbind()) the three contrast vectors into a matrix **K** whose first column is the **k** vector given above and the second and third columns are the corresponding vectors for computing substitution effects at the second and third loci respectively. In that case, the matrix-vector product **K'a** will compute all three substitution effects at once.

Example 2: Shrinkage of substitution effects.

Modify the three pairs of diagonal elements of **G**, or equivalently, modify the single diagonal element of the nanim by nanim matrix **R** in order to modify the variance ratio lambda, of residual to genetic variance. In an animal model, lambda is $(1\text{-}h^2)/h^2$ which will be 0 if h^2 is 1 and a large number if h^2 is small. For a heritability of 0.25, lambda is 3. In genomic prediction models, the genetic variance is partitioned among all the loci. If there are hundreds of loci, the lambda ratio for each locus will be large. You can simulate this effect by making the diagonal elements or **R** say 10 or 100 times larger than **G**. Compare the estimated substitution effects for varying values of residual variance (in relation to additive variance). Shrinkage is related to the magnitude of the ratio of residual to additive variance. If residual variance is small this ratio will be reduced and the estimates will approach least squares. Inspect the variance ratio for each scenario you attempt.

If order to compute the least squares estimate you will need to form the least squares equations treating allelic effects as fixed. To do this, you need to form a new incidence matrix for fixed effects that includes the old fixed effects (eg the overall mean) as well as the allelic effects. You can do this using cbind(X,Z) to augment the columns of the two incidence matrices. However, this new matrix will not have full column rank so the least squares equations will not be full rank. You should be able to constrain the new equations to full rank by limiting the augmented matrix to include only one column of allelic effects for each locus.

For example, $\mathbf{X}_{new}$ <- cbind(X,Z[,c(1,3,5)]) will use only those three columns. Then the least squares solutions can be obtained from solving the following full rank equations. The first effect in these equations will be an intercept rather than a mean, unless you center the covariates in the **Z** matrix by subtracting 1.

$$\left[\mathbf{X}_{new}'\mathbf{X}_{new}\right]\left[\hat{\mathbf{b}}^0\right]=\left[\mathbf{X}_{new}'\mathbf{y}\right]$$

Modify the constant nanim to alter the number of animals in the datafile that will be used in the calculation. Try larger and smaller values.

What do you conclude about the importance of treating SNP effects as random in terms of shrinkage of estimated effects?

Before continuing, you will want to reset the genetic and residual variances back to their original values.

Example 3: Fitting both alleles at the three loci as random effects using MME.

An alternative approach to estimate random effects is to use the mixed model equations. Rather than requiring the inverse of **V**, the typical form of the mixed model equations requires the inverse of **G** and the inverse of **R**. Its general form is as follows

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{b}}^0 \\ \hat{\mathbf{a}} \end{bmatrix}=\begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

In simple cases where **R** is a scaled identity, only the inverse of **G** is required as the scalar residual variance can be factored out by multiplication. Remember that the inverse of the coefficient matrix will need to be scaled by the residual variance to compute the correct prediction error variances or reliabilities when you use this modified form. Form and solve these simpler mixed model equations, as follows

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\sigma_e^2\mathbf{G}^{-1} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{b}}^0 \\ \hat{\mathbf{a}} \end{bmatrix}=\begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}.$$

You will need to use the commands cbind() and/or rbind() to join two matrices of conformable order by column or by row respectively.

Compare the solutions for the fixed effects and the six random allelic effects to the GLS solutions. They should be identical. If not, check your equations before you proceed.

Extract the prediction error variance-covariance (PEV) matrix ($\text{var}(\hat{\mathbf{a}} - \mathbf{a}) = \mathbf{C22}\sigma_e^2$) of the fitted allelic effects, where **C22** is that submatrix of the inverse of the mixed model equations corresponding to the rows and columns representing random effects (ie $\mathbf{Z'Z}+\sigma_e^2\mathbf{G}^{-1}$ portion of the inverse). Compute $\text{var}(\hat{\mathbf{a}}) = \mathbf{G} - \mathbf{C22}\sigma_e^2$ by subtracting the PEV matrix from the genetic variance-covariance matrix. The reliability of the predictions (squared correlation between true and predicted merit) are obtained by dividing the diagonal elements of **G-C22** σ_e^2 by the diagonal elements of **G**. You might find the R function diag() useful for this purpose. Reliability is used in some industries (eg dairy) to convey the information content in estimated breeding values (EBVs).

Compute the substitution effects by forming relevant contrast vectors as in the previous question.

From the viewpoint of genomic prediction rather than QTL detection, we will be more interested in linear functions of the estimated SNP effects, such as $\mathbf{Z}\hat{\mathbf{a}}$. Compute that linear function for all animals. You may want to plot that estimate of genetic merit against the phenotype using the plot() command, or compute the correlation with phenotype using the cor() function.

We typically have to compute reliabilities of estimated breeding values. The reliability for any arbitrary contrast **k**, can be calculated as linear function of the **G** and **C22** matrices as follows

$$r^2_{\mathbf{k'\hat{a}}} = \frac{\mathbf{diag}\left[\mathbf{k'}\left(\mathbf{G} - \mathbf{C22}\sigma_e^2\right)\mathbf{k}\right]}{\mathbf{diag}\left[\mathbf{k'Gk}\right]}.$$

In mixed models, any linear combination of random effects is estimable, so conformable **k** can contain any elements. One meaningful choice of **k'** is the elements of a row of **Z**, as that contrast estimates the linear combination of random contributions relevant to a particular animal. The reliabilities of all animals can be simultaneously predicted using the entire **Z** matrix in place of **k'** in the above equation. Compute the breeding values of all the animals and their corresponding reliabilities.

Example 4: Directly fitting animal effects using genomic relationships.

Rather than estimating allelic effects at every locus, an equivalent model can be derived that directly solves the animal effects in the appropriate mixed model equations. This formulation of the problem in the usual representation of the mixed model equations will only work when the genomic relationship matrix is full rank. The genomic relationship matrix will not be full rank if there are more animals than loci or if any two animals have identical genotypes.

Reduce nanim to 3 and recompute the quantities in example 2. The animals in the original Hayes datafile have been reordered so that the genomic relationship matrix is full rank for the first three animals.

Form the genomic relationship matrix as **ZGZ'**, and invert it using solve(). Form and solve the mixed model equations, and compute the reliabilities for each animal. In computing the reliabilities, note that the matrix you previously used for **G** should now be replaced by **ZGZ'**. To fit animal effects directly, use the mixed model equations in the form below where the previous incidence matrix for the random effects has been replaced by the matrix **Z**.

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'} \\ \mathbf{X} & \mathbf{I} + \sigma_e^2[\mathbf{ZGZ'}]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}^0 \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{y} \end{bmatrix}$$

Compare your results to the answers you obtained in example 2. They should be identical.

Example 5: Alternative parameterizations fitting substitution effects rather than allelic effects.

Modify the **Z** matrix by reading only columns 1, 3 and 5 (or 2, 4 and 6). This allows you to fit substitution effects rather than both allelic effects. You will also need to appropriately alter the order of **G** and double the genetic variance for substitution effects for each locus compared to allelic effects because $\text{var}(\alpha) = \text{var}(a_1 - a_2) = \text{var}(a_1) + \text{var}(a_2) = 2\,\text{var}(a)$. If you don't recode the new **Z** matrix, you have effectively modified the overall mean and the estimated breeding values will all be altered by a constant compared to the previous questions. This is no problem in real life, as breeding values are typically rescaled to a consistent base after computation and prior to publication of the results.

You may want to further experiment by subtracting 1 from every element of **Z**, so each SNP is coded -1, 0 and 1 rather than 0, 1 or 2.

For the modified incidence matrices, repeat example 1, fitting the GLS equations, example 2, fitting the mixed model equations for substitution effects and example 3, fitting the genomic relationship matrix. These three models are equivalent to each other and should give the same solutions to each other for this parameterization. You should also find that the solutions for substitution effects or animals are the same as you obtained in examples 1-3 except the breeding values may differ by a constant depending upon your parameterization. The fixed effects solutions will not be the same, neither will the prediction error variances or reliabilities of predicted random effects be typically identical.

Bayesian Methods in Genome Association Studies

Rohan L. Fernando

Iowa State University

February, 2010

Outline of Part I

Fundamentals

Bayesian Inference
- Theory
- Computing Posteriors

Outline of Part II

Bayesian Regression Models
- Normal
- Student-t
- Mixture Models

Simulations

Part I

Bayesian Inference: Theory

Bayes Theorem

The conditional probability of X given Y is

$$\Pr(X|Y) = \frac{\Pr(X, Y)}{\Pr(Y)} = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)}$$

where $\Pr(X, Y)$ is the joint probability of X and Y, $\Pr(X)$ is the probability of X, and $\Pr(Y)$ is the probability of Y.

Conditional Probability by Example

Joint distribution of smoking and lung cancer in a hypothetical population of 1,000,000:

		Smoking		
		Yes	No	
Lung Cancer	Yes	42,500	7,500	50,000
	No	207,500	742,500	950,000
		250,000	750,000	

Question: What is the relative frequency of lung cancer among smokers?

Answer: $\frac{42{,}500}{250{,}000} = 0.17$

Conditional Probability by Example

- As explained below, this relative frequency is also the conditional probability of lung cancer given smoking.
  - The frequentist definition of probability of an event is the limiting value of its relative frequency in a large number of trials.
  - Suppose we sample with replacement individuals from the 250,000 smokers and compute the relative frequency of lung cancer incidence.
  - It can be shown that as the sample size goes to infinity, this relative frequency will approach $\frac{42{,}500}{250{,}000} = 0.17$.
- This conditional probability is usually written as $\frac{42{,}500/1{,}000{,}000}{250{,}000/1{,}000{,}000} = 0.17$.
- The ratio in the numerator is joint probability of smoking and lung cancer, and the ratio in the denominator is the marginal probability of smoking.

Meaning of Probability in Bayesian Inference

- In the frequency approach, probability is a limiting frequency
- In Bayesian inference, probabilities are used to quantify your beliefs or knowledge about possible values of parameters
  - What is the probability that $h^2 > 0.5$?
  - What is the probability that milk yield is controlled by more than 100 loci?

Essentials of Bayesian Inference

- Prior probabilities quantify beliefs about parameters before the data are analyzed
- Parameters are related to the data through the model or "likelihood", which is the conditional probability density for the data given the parameters
- The prior and the likelihood are combined using Bayes theorem to obtain posterior probabilities, which are conditional probabilities for the parameters given the data
- Inferences about parameters are based on the posteior

Bayes Theorem in Bayesian Inference

- Let $f(\theta)$ denote the prior probability density for θ
- Let $f(\boldsymbol{y}|\theta)$ denote the likelihood
- Then, the posterior probability of θ is:

$$\begin{aligned} f(\theta|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\theta)f(\theta)}{f(\boldsymbol{y})} \\ &\propto f(\boldsymbol{y}|\theta)f(\theta) \end{aligned}$$

Computing posteriors

- Often no closed form for $f(\theta|\boldsymbol{y})$
- Further, even if computing $f(\theta|\boldsymbol{y})$ is feasible, obtaining $f(\theta_i|\boldsymbol{y})$ would require integrating over many dimensions
- Thus, in many situations, inferences are made using the empirical posterior constructed by drawing samples from $f(\theta|\boldsymbol{y})$
- Gibbs sampler is widely used for drawing samples from posteriors

Gibbs sampler

- Want to draw samples from $f(x_1, x_2, \ldots, x_n)$
- Even though it may be possible to compute $f(x_1, x_2, \ldots, x_n)$, it is difficult to draw samples directly from $f(x_1, x_2, \ldots, x_n)$
- Gibbs:
  - Get valid a starting point $\boldsymbol{x}^0$
  - Draw sample $\boldsymbol{x}^t$ as:

$$\begin{array}{lll} x_1^t & \text{from} & f(x_1|x_2^{t-1}, x_3^{t-1}, \ldots, x_n^{t-1}) \\ x_2^t & \text{from} & f(x_2|x_1^t, x_3^{t-1}, \ldots, x_n^{t-1}) \\ x_3^t & \text{from} & f(x_3|x_1^t, x_2^t, \ldots, x_n^{t-1}) \\ \vdots & & \vdots \\ x_n^t & \text{from} & f(x_n|x_1^t, x_2^t, \ldots, x_{n-1}^t) \end{array}$$

- The sequence $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n$ is a Markov chain with stationary distribution $f(x_1, x_2, \ldots, x_n)$

Inference from Markov chain

Can show that samples obtained from the Markov chain can be used to draw inferences from $f(x_1, x_2, \ldots, x_n)$ provided the chain is:

- Irreducible: can move from any state i to any other state j
- Positive recurrent: return time to any state has finite expectation
- *Markov Chains*, J. R. Norris (1997)

Example

Let $f(\boldsymbol{x})$ be a bivariate normal density with means

$$\mu' = \begin{bmatrix} 1 & 2 \end{bmatrix}$$

and covariance matrix

$$\boldsymbol{V} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2.0 \end{bmatrix}$$

Suppose we do not know how to draw samples from $f(\boldsymbol{x})$, but know how to draw samples from $f(x_i|x_j)$, which is univariate normal with mean:

$$\mu_{i.j} = \mu_i + \frac{v_{ij}}{v_{jj}}(x_j - \mu_j)$$

and variance

$$v_{i.j} = v_{ii} - \frac{v_{ij}^2}{v_{jj}}$$

Gibbs sampler

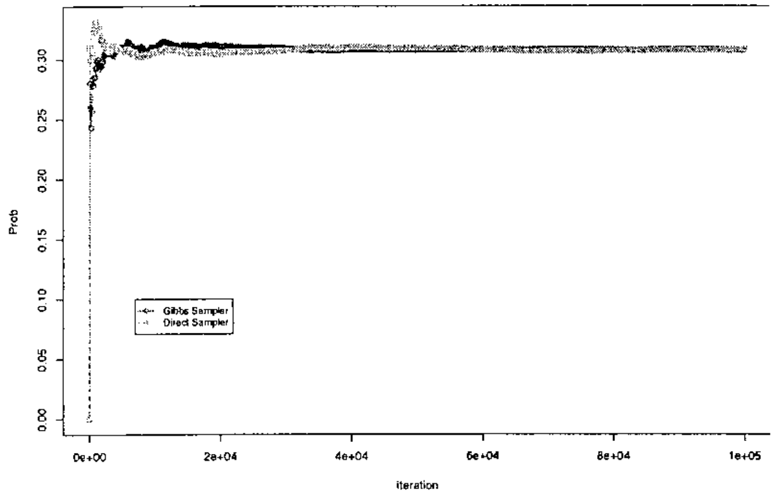
- Gibbs:
  - Start with $\boldsymbol{x}^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
  - Draw sample $\boldsymbol{x}^t$ as:

$$\begin{array}{lll} x_1^t & \text{from} & f(x_1 | x_2^{t-1}) \\ x_2^t & \text{from} & f(x_2 | x_1^t) \end{array}$$

- Use the sequence $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n$ to compute any property of $f(\boldsymbol{x})$, for example

$$\Pr(x_1 > \mu_1 \text{ and } x_2 > \mu_2)$$

MCMC Estimates of $\Pr(x_1 > \mu_1 \text{ and } x_2 > \mu_2)$



Metropolis-Hastings sampler

- Sometimes may not be able to draw samples directly from $f(x_i|\mathbf{x}_{i\_})$
- Convergence of the Gibbs sampler may be too slow
- Metropolis-Hastings (MH) for sampling from $f(x)$:
  - a candidate sample, y, is drawn from a proposal distribution $q(y|x^{t-1})$
  - $$x^t = \begin{cases} y & \text{with probability } \alpha \\ x^{t-1} & \text{with probability } 1-\alpha \end{cases}$$
  - $$\alpha = \min(1, \frac{f(y)q(x^{t-1}|y)}{f(x^{t-1})q(y|x^{t-1})})$$
- The samples from MH is a Markov chain with stationary distribution $f(x)$

Proposal distributions

Two main types:

- Approximations of the target density: $f(x)$
  - Not easy to find approximation that is easy to sample from
  - High acceptance rate is good!
- Random walk type: stay close to the previous sample
  - Generally easy to construct proposal
  - High acceptance rate may indicate that candidate is too close to previous sample
  - Intermediate acceptance rate is good

MH Sampler to Estimate $\Pr(x_1 > \mu_1 \text{ and } x_2 > \mu_2)$

MH Sampler:

- Start with $\boldsymbol{x}^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
- Draw sample $\boldsymbol{x}^t$ as:

$$\begin{aligned} y_1 &= x_1^{t-1} + u_1 \\ y_2 &= x_2^{t-1} + u_2 \end{aligned}$$

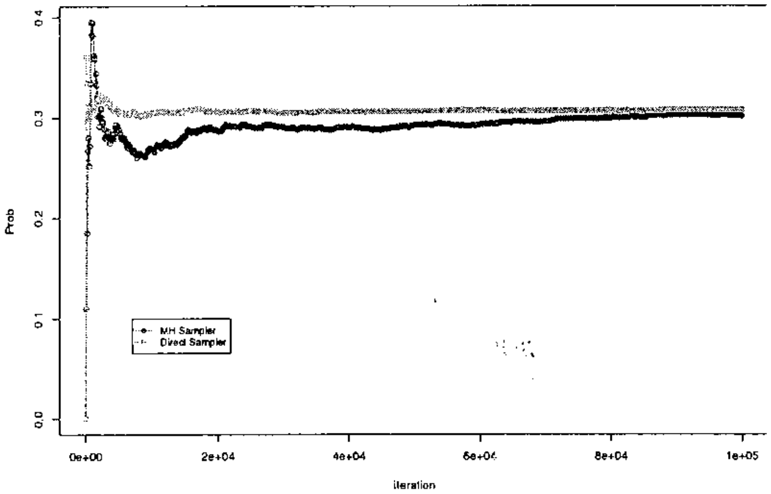
  where u_i is Uniform$(-v_{ii}^{1/2}, v_{ii}^{1/2})$.
- Compute

$$\alpha = min(1, \frac{f(\boldsymbol{y})}{f(\boldsymbol{x}^{t-1})})$$

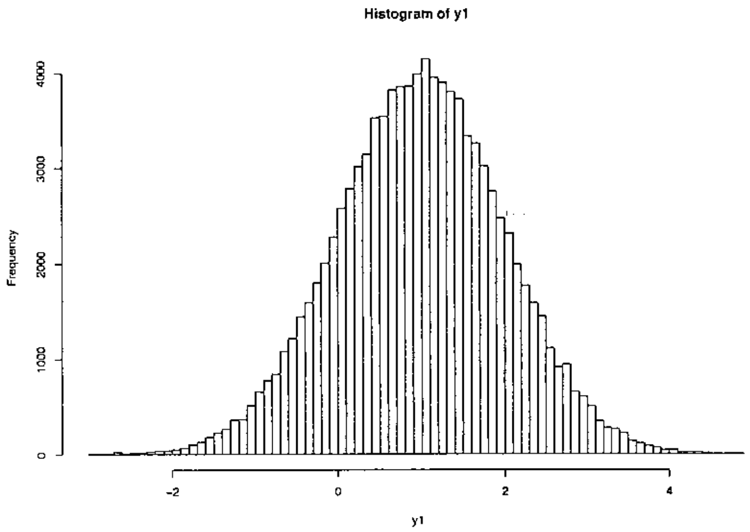
  and

$$\boldsymbol{x}^t = \begin{cases} \boldsymbol{y} & \text{with probability } \alpha \\ \boldsymbol{x}^{t-1} & \text{with probability } 1 - \alpha \end{cases}$$

MCMC Estimates of $\Pr(x_1 > \mu_1 \text{ and } x_2 > \mu_2)$



Distribution of y_1 Sampled Using MH



Part II

Bayesian Inference: Application to Whole Genome Analyses

Model

Model:

$$y_i = \mu + \sum_j X_{ij}\alpha_j + e_i$$

Priors:

- $\mu \propto$ constant (not proper, but posterior is proper)
- $(e_i|\sigma^2_e) \sim$ (iid)N$(0, \sigma^2_e)$; $\sigma^2_e \sim \nu_e S^2_e \chi^{-2}_{\nu_e}$ scaled inverse [illegible]
- Consider several different priors for α_j

Normal

- Prior: $(\alpha_j|\sigma^2_\alpha) \sim$ (iid)N$(0, \sigma^2_\alpha)$; σ^2_α is known
- What is σ^2_α?
- Assume the QTL genotypes are a subset of those available for the analysis
  - Then, the genotypic value of i can be written as:

$$g_i = \mu + \boldsymbol{x}'_i\boldsymbol{\alpha}$$ [illegible]

  - Note that α is common to all i
  - Thus, the variance of g_i comes from $\boldsymbol{x}'_i$ being random
- So, σ^2_α is not the genetic variance at a locus
- If locus j is randomly sampled from all the loci available for analysis:
  - Then, α_j will be a random variable
  - $\sigma^2_\alpha = \text{Var}(\alpha_j)$

Relationship of σ_α^2 to genetic variance

Assume loci with effect on trait are in linkage equilibrium. Then, the additive genetic variance is

$$V_A = \sum_j^k 2p_jq_j\alpha_j^2,$$

where $p_j = 1 - q_j$ is gene frequency at SNP locus j.
Letting $U_j = 2p_jq_j$ and $V_j = \alpha_j^2$,

$$V_A = \sum_j^k U_jV_j$$

For a randomly sampled locus, covariance between U_j and V_j is

$$C_{UV} = \frac{\sum_j U_jV_j}{k} - (\frac{\sum_j U_j}{k})(\frac{\sum_j V_j}{k})$$

Relationship of σ_α^2 to genetic variance

Rearranging the previous expression for C_{UV} gives

$$\sum_j U_jV_j = kC_{UV} + (\sum_j U_j)(\frac{\sum_j V_j}{k})$$

So,

$$V_A = kC_{UV} + (\sum_j 2p_jq_j)(\frac{\sum_j \alpha_j^2}{k})$$

Letting $\sigma_\alpha^2 = \frac{\sum_j \alpha_j^2}{k}$ gives

$$V_A = kC_{UV} + (\sum_j 2p_jq_j)\sigma_\alpha^2$$

and,

$$\sigma_\alpha^2 = \frac{V_A - kC_{UV}}{\sum_j 2p_jq_j}$$

Blocked Gibbs sampler

- Let $\theta' = [\mu, \alpha']$
- Can show that $(\theta|\boldsymbol{y}, \sigma_e^2) \sim \mathrm{N}(\hat{\theta}, \boldsymbol{C}^{-1}\sigma_e^2)$
- $$\hat{\theta} = \boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{y}; \quad \boldsymbol{W} = [\mathbf{1}, \boldsymbol{X}]$$
- $$\boldsymbol{C} = \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\boldsymbol{X} \\ \boldsymbol{X}'\mathbf{1} & \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}\frac{\sigma_e^2}{\sigma_\alpha^2} \end{bmatrix}$$
- Blocked Gibbs sampler
  - García-Cortés and Sorensen (1996, GSE 28:121-126)
  - *Likelihood, Bayesian and MCMC Methods* $\cdots$ (LBMMQG, Sorensen and Gianola, 2002)

Full conditionals for single-site Gibbs

- $(\mu|\boldsymbol{y}, \alpha, \sigma_e^2) \sim \mathrm{N}(\frac{\mathbf{1}'(\boldsymbol{y}-\boldsymbol{X}\alpha)}{n}, \frac{\sigma_e^2}{n})$
- $(\alpha_j|\boldsymbol{y}, \mu, \alpha_{j\_}, \sigma_e^2) \sim \mathrm{N}(\hat{\alpha}_j, \frac{\sigma_e^2}{c_j})$
  - $$\hat{\alpha}_j = \frac{\boldsymbol{x}_j'\boldsymbol{w}}{c_j}$$
  - $$\boldsymbol{w} = \boldsymbol{y} - \mathbf{1}\mu - \sum_{j'\neq j} \boldsymbol{x}_{j'}\alpha_{j'}$$
  - $$c_j = (\boldsymbol{x}_j'\boldsymbol{x}_j + \frac{\sigma_e^2}{\sigma_\alpha^2})$$
- $(\sigma_e^2|\boldsymbol{y}, \mu, \alpha) \sim [(\boldsymbol{y} - \boldsymbol{W}\theta)'(\boldsymbol{y} - \boldsymbol{W}\theta) + \nu_e S_e^2]\chi^{-2}_{(\nu_e+n)}$

Derive: full conditional for α_j

From Bayes' Theorem,

$$f(\alpha_j|\boldsymbol{y}, \mu, \boldsymbol{\alpha}_{j\_}, \sigma_e^2) = \frac{f(\alpha_j, \boldsymbol{y}, \mu, \boldsymbol{\alpha}_{j\_}, \sigma_e^2)}{f(\boldsymbol{y}, \mu, \boldsymbol{\alpha}_{j\_}, \sigma_e^2)}$$

$$\propto f(\boldsymbol{y}|\alpha_j, \mu, \alpha_{j\_}, \sigma_e^2) f(\alpha_j) f(\mu, \boldsymbol{\alpha}_{j\_}, \sigma_e^2)$$

$$\propto (\sigma_e^2)^{-n/2} \exp\{-\frac{(\boldsymbol{w} - \boldsymbol{x}_j\alpha_j)'(\boldsymbol{w} - \boldsymbol{x}_j\alpha_j)}{2\sigma_e^2}\}(\sigma_\alpha^2)^{-1/2} \exp\{-\frac{\alpha_j^2}{2\sigma_\alpha^2}\}$$

where

$$\boldsymbol{w} = \boldsymbol{y} - \mathbf{1}\mu - \sum_{j \neq j'} \boldsymbol{x}_{j'}\alpha_{j'}$$

Derive: full conditional for α_j

The exponential terms in the joint density can be written as:

$$-\frac{1}{2\sigma_e^2}\{\boldsymbol{w}'\boldsymbol{w} - 2\boldsymbol{x}_j'\boldsymbol{w}\alpha_j + [\boldsymbol{x}_j'\boldsymbol{x}_j + \frac{\sigma_e^2}{\sigma_\alpha^2}]\alpha_j^2\}$$

Completing the square in this expression with respect to α_j gives

$$-\frac{1}{2\sigma_e^2}\{c_j(\alpha_j - \hat{\alpha}_j)^2 + \boldsymbol{w}'\boldsymbol{w} - c_j\hat{\alpha}_j{}^2\}$$

where

$$\hat{\alpha}_j = \frac{\boldsymbol{x}_j'\boldsymbol{w}}{c_j}$$

So,

$$f(\alpha_j|\boldsymbol{y}, \mu, \boldsymbol{\alpha}_{j\_}, \sigma_e^2) \propto \exp\{-\frac{(\alpha_j - \hat{\alpha}_j)^2}{2\frac{\sigma_e^2}{c_j}}\}$$

Alternative view of Normal prior

Consider fixed linear model:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\alpha + \boldsymbol{e}$$

This can be also written as

$$\boldsymbol{y} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \end{bmatrix} + \boldsymbol{e}$$

Suppose we observe for each locus:

$$y_j^* = \alpha_j + \epsilon_j$$

Least Squares with Additional Data

Fixed linear model with the additional data:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}^* \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \end{bmatrix} + \begin{bmatrix} \boldsymbol{e} \\ \epsilon \end{bmatrix}$$

[illegible]

OLS Equations:

$$\begin{bmatrix} \mathbf{1}' & \mathbf{0}' \\ \boldsymbol{X}' & \boldsymbol{I}' \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_n\frac{1}{\sigma_e^2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_k\frac{1}{\sigma_\epsilon^2} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{1}' & \mathbf{0}' \\ \boldsymbol{X}' & \boldsymbol{I}' \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_n\frac{1}{\sigma_e^2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_k\frac{1}{\sigma_\epsilon^2} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}^* \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\boldsymbol{X} \\ \boldsymbol{X}'\mathbf{1} & \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}\frac{\sigma_e^2}{\sigma_\epsilon^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{y} + \boldsymbol{y}^*\frac{\sigma_e^2}{\sigma_\epsilon^2} \end{bmatrix}$$

[illegible]

Univariate-t

Prior:

$$(\alpha_j|\sigma_j^2) \sim \mathrm{N}(0, \sigma_j^2)$$

$$\sigma_j^2 \sim \nu_\alpha S_{\nu_\alpha}^2 \chi_{\nu_\alpha}^{-2}$$

Can show that the unconditional distribution for α_j is

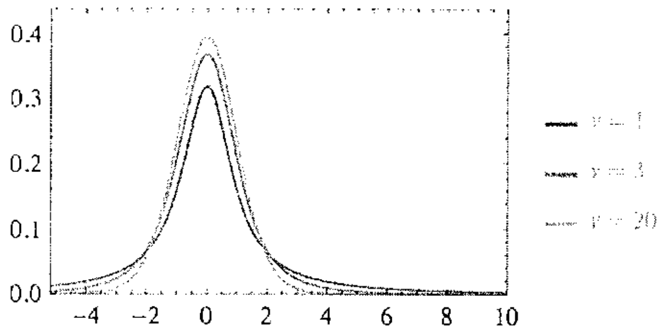
$$\alpha_j \sim (\mathrm{iid}) t(0, S_{\nu_\alpha}^2, \nu_\alpha)$$

(Sorensen and Gianola, 2002, LBMMQG pages 28,60)

This is Bayes-A (Meuwissen et al., 2001; Genetics 157:1819-1829)

Univariate-t

Plots of PDF for typical parameters:



Generated by Wolfram|Alpha (www.wolframalpha.com)

Full conditional for single-site Gibbs

Full conditionals are the same as in the "Normal" model for μ, α_j, and σ_e^2. Let

$$\boldsymbol{\xi} = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2]$$

Full conditional conditional for σ_j^2:

$$f(\sigma_j^2|\boldsymbol{y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\xi}_{j\_}, \sigma_e^2) \propto f(\boldsymbol{y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\xi}, \sigma_e^2)$$

$$\propto f(\boldsymbol{y}|\mu, \boldsymbol{\alpha}, \boldsymbol{\xi}, \sigma_e^2) f(\alpha_j|\sigma_j^2) f(\sigma_j^2) f(\mu, \boldsymbol{\alpha}_{j\_}, \boldsymbol{\xi}_{j\_}\sigma_e^2)$$

$$\propto (\sigma_j^2)^{-1/2} \exp\{-\frac{\alpha_j^2}{2\sigma_j^2}\}(\sigma_j^2)^{-(2+\nu_\alpha)/2} \exp\{\frac{\nu_\alpha S_\alpha^2}{2\sigma_j^2}\}$$

$$\propto (\sigma_j^2)^{-(2+\nu_\alpha+1)/2} \exp\{\frac{\alpha_j^2 + \nu_\alpha S_\alpha^2}{2\sigma_j^2}\}$$

Full conditional for σ_j^2

So,

$$(\sigma_j^2|\boldsymbol{y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\xi}_\_, \sigma_e^2) \sim \tilde{\nu}_\alpha \tilde{S}_\alpha^2 \chi_{\nu_\alpha}^{-2}$$

where

$$\tilde{\nu}_\alpha = \nu_\alpha + 1$$

and

$$\tilde{S}_\alpha^2 = \frac{\alpha_j^2 + \nu_\alpha S_\alpha^2}{\tilde{\nu}_\alpha}$$

Multivariate-t

Prior:

$$(\alpha_j|\sigma^2_\alpha) \sim \text{(iid)N}(0, \sigma^2_\alpha)$$
$$\sigma^2_\alpha \sim \nu_\alpha S^2_{\nu_\alpha} \chi^{-2}_{\nu_\alpha}$$

Can show that the unconditional distribution for $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha} \sim \text{multivariate-}t(\mathbf{0}, \boldsymbol{I}S^2_{\nu_\alpha}, \nu_\alpha)$$

(Sorensen and Gianola, 2002, LBMMQG page 60)

We will see later that this is Bayes-C with $\pi = 0$.

Full conditional for σ^2_α

We will see later that

$$(\sigma^2_\alpha|\boldsymbol{y}, \mu, \boldsymbol{\alpha}, \sigma^2_e) \sim \tilde{\nu}_\alpha \tilde{S}^2_\alpha \chi^{-2}_{\nu_\alpha}$$

where

$$\tilde{\nu}_\alpha = \nu_\alpha + k$$

and

$$\tilde{S}^2_\alpha = \frac{\boldsymbol{\alpha}'\boldsymbol{\alpha} + \nu_\alpha S^2_\alpha}{\tilde{\nu}_\alpha}$$

Spike and univariate-t

Prior:

$$(\alpha_j|\pi,\sigma_j^2)\begin{cases}\sim \mathrm{N}(0,\sigma_j^2) & \text{probability}\,(1-\pi),\\ = 0 & \text{probability}\,\pi\end{cases}$$

and

$$(\sigma_j^2|\nu_\alpha,S_\alpha^2)\sim\nu_\alpha S_\alpha^2\chi_{\nu_\alpha}^{-2}$$

Thus,

$$(\alpha_j|\pi)(\text{iid})\begin{cases}\sim \text{univariate-}t(0,S_\alpha^2,\nu_\alpha) & \text{probability}\,(1-\pi),\\ = 0 & \text{probability}\,\pi\end{cases}$$

This is Bayes-B (Meuwissen et al., 2001; Genetics 157:1819-1829)

Notation for sampling from mixture

The indicator variable δ_j is defined as

$$\delta_j = 1 \Rightarrow (\alpha_j|\sigma_j^2)\sim\mathrm{N}(0,\sigma_j^2)$$

and

$$\delta_j = 0 \Rightarrow (\alpha_j|\sigma_j^2) = 0$$

Sampling strategy in MHG (2001)

- Sampling σ_e^2 and μ are as under the Normal prior.
- MHG proposed to use a Metropolis-Hastings sampler to draw samples for σ_j^2 and α_j jointly from their full-conditional distribution.
- First, σ_j^2 is sampled from

  $$f(\sigma_j^2|\boldsymbol{y}, \mu, \boldsymbol{\alpha}_{j\_}, \boldsymbol{\xi}_{\_}, \sigma_e^2)$$

  using MH with prior as proposal.
- Then, α_j is sampled from its full-conditional, which is identical to that under the Normal prior

MH acceptance probability when prior is used as proposal

Suppose we want to sample θ from $f(\theta|\boldsymbol{y})$ using the MH with its prior as proposal. Then, the MH acceptance probability becomes:

$$\alpha = min(1, \frac{f(\theta_{can}|\boldsymbol{y})f(\theta^{t-1})}{f(\theta^{t-1}|\boldsymbol{y})f(\theta_{can})}$$

where $f(\theta)$ is the prior for θ. Using Bayes' theorem, the target density can be written as:

$$f(\theta|\boldsymbol{y}) \propto f(\boldsymbol{y}|\theta)f(\theta)$$

Then, the acceptance probability becomes

$$\alpha = min(1, \frac{f(\boldsymbol{y}|\theta_{can})f(\theta_{can})f(\theta^{t-1})}{f(\boldsymbol{y}|\theta^{t-1})f(\theta^{t-1})f(\theta_{can})}$$

Sampling σ_j^2

Thus when the prior for σ_j^2 is used as the proposal, the MH acceptance probability becomes

$$\alpha = \min(1, \frac{f(\boldsymbol{y}|\sigma_{can}^2, \boldsymbol{\theta}_{j\_})}{f(\boldsymbol{y}|\sigma_j^2, \boldsymbol{\theta}_{j\_})})$$

where σ_{can}^2 is used to denote the candidate value for σ_j^2, and $\boldsymbol{\theta}_{j\_}$ all the other parameters. It can be shown that, α_j depends on $\boldsymbol{y}$ only through $r_j = \boldsymbol{x}_j'\boldsymbol{w}$ (page 30). Thus

$$f(\boldsymbol{y}|\sigma_j^2, \boldsymbol{\theta}_{j\_}) \propto f(r_j|\sigma_j^2, \boldsymbol{\theta}_{j\_})$$

"Likelihood" for σ_j^2

Recall that

$$\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{1}\mu - \sum_{j' \neq j} \boldsymbol{x}_{j'}\alpha_{j'} = \boldsymbol{x}_j\alpha_j + \boldsymbol{e}$$

Then,

$$\mathrm{E}(\boldsymbol{w}|\sigma_j^2, \boldsymbol{\theta}_{j\_}) = \boldsymbol{0}$$

When $\delta = 1$:

$$\mathrm{Var}(\boldsymbol{w}|\delta_j = 1, \sigma_j^2, \boldsymbol{\theta}_{j\_}) = \boldsymbol{x}_j\boldsymbol{x}_j'\sigma_j^2 + \boldsymbol{I}\sigma_e^2$$

and $\delta = 0$:

$$\mathrm{Var}(\boldsymbol{w}|\delta_j = 0, \sigma_j^2, \boldsymbol{\theta}_{j\_}) = \boldsymbol{I}\sigma_e^2$$

"Likelihood" for σ_j^2

So,

$$\mathrm{E}(r_j|\sigma_j^2, \boldsymbol{\theta}_{j\_}) = 0$$

and

$$\mathrm{Var}(r_j|\delta_j = 1, \sigma_j^2, \boldsymbol{\theta}_{j\_}) = (\mathbf{x}_j'\mathbf{x}_j)^2\sigma_j^2 + \mathbf{x}_j'\mathbf{x}_j\sigma_e^2 = v_1$$

$$\mathrm{Var}(r_j|\delta_j = 0, \sigma_j^2, \boldsymbol{\theta}_{j\_}) = \mathbf{x}_j'\mathbf{x}_j\sigma_e^2 = v_0$$

So,

$$f(r_j|\delta_j, \sigma_j^2, \boldsymbol{\theta}_{j\_}) \propto (v_\delta)^{-1/2}\exp\{-\frac{r_j^2}{2v_\delta}\}$$

Alternative View of Prior in BayesB

- How much information is being added by the prior?
- BayesB is identical to ML with additional data!
- Can “see” how much additional data in BayesB prior.

Maximum Likelihood with Additional Data

- Suppose at locus j, $\delta_j = 1$, and we observe additional data:

$$\boldsymbol{u}_j \sim N(\boldsymbol{0}, \boldsymbol{I}_q \sigma_j^2)$$

- Assume that only unknown is σ_j^2
- So, adjust phenotypes as:

$$\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{1}\mu - \sum_{j' \neq j} \boldsymbol{x}_{j'} \alpha_{j'}$$

- Likelihood:

$$L(\sigma_j^2; \boldsymbol{w}, \boldsymbol{u}_j) = L(\sigma_j^2; \hat{\alpha}_j, \boldsymbol{u}_j)$$

Likelihood with Additional Data

- $$L(\sigma_j^2; \hat{\alpha}_j, \boldsymbol{u}_j) \propto f_1(\hat{\alpha}_j | \sigma_j^2) \times f_2(\boldsymbol{u}_j | \sigma_j^2)$$
- $$\begin{aligned} f_2(\boldsymbol{u}_j | \sigma_j^2) &\propto (\sigma_j^2)^{-q/2} \exp[\frac{-\boldsymbol{u}_j' \boldsymbol{u}_j}{2\sigma_j^2}] \\ &\propto (\sigma_j^2)^{-[\nu/2+1]} \exp[\frac{-\nu S^2}{2\sigma_j^2}] \end{aligned}$$
- $\nu = q - 2, S^2 = \frac{\boldsymbol{u}_j' \boldsymbol{u}_j}{\nu}$

Alternative algorithm for spike and univariate-t

Rather than use the prior as the proposal for sampling σ_j^2, we

- sample $\delta_j = 1$ with probability 0.5
- when $\delta = 1$, sample σ_j^2 from a scaled inverse chi-squared distribution with
  - scale parameter = $\sigma_j^{2(t-1)}/2$ and 4 degrees of freedom when $\delta_j^{(t-1)} = 1$, and
  - scale parameter = S_α^2 and 4 degrees of freedom when $\delta_j^{(t-1)} = 0$

Multivariate-*t* mixture

Prior:

$$(\alpha_j|\pi, \sigma_\alpha^2) \begin{cases} \sim \mathrm{N}(0, \sigma_\alpha^2) & \text{probability}\,(1-\pi), \\ = 0 & \text{probability}\,\pi \end{cases}$$

and

$$(\sigma_\alpha^2|\nu_\alpha, S_\alpha^2) \sim \nu_\alpha S_\alpha^2 \chi_{\nu_\alpha}^{-2}$$

Further,

$$\pi \sim \mathrm{Uniform}(0,1)$$

- The α_j variables with their corresponding $\delta_j = 1$ will follow a multivariate-*t* distribution.
- This is what we have called Bayes-Cπ

Full conditionals for single-site Gibbs

Full-conditional distributions for μ, α, and σ_e^2 are as with the Normal prior.
Full-conditional for δ_j:

$$\Pr(\delta_j|\boldsymbol{y},\mu,\boldsymbol{\alpha}_{-j},\boldsymbol{\delta}_{-j},\sigma_\alpha^2,\sigma_e^2,\pi) = \Pr(\delta_j|r_j,\boldsymbol{\theta}_{j\_})$$

$$\Pr(\delta_j|r_j,\boldsymbol{\theta}_{j\_}) = \frac{f(\delta_j,r_j|\boldsymbol{\theta}_{j\_})}{f(r_j|\boldsymbol{\theta}_{j\_})}$$

$$= \frac{f(r_j|\delta_j,\boldsymbol{\theta}_{j\_})\Pr(\delta_j|\pi)}{f(r_j|\delta_j=0,\boldsymbol{\theta}_{j\_})\pi + f(r_j|\delta_j=1,\boldsymbol{\theta}_{j\_})(1-\pi)}$$

Full conditional forσ_α^2

This can be written as

$$f(\sigma_\alpha^2|\boldsymbol{y},\mu,\boldsymbol{\alpha},\boldsymbol{\delta},\sigma_e^2) \propto f(\boldsymbol{y}|\sigma_\alpha^2,\mu,\boldsymbol{\alpha},\boldsymbol{\delta},\sigma_e^2)f(\sigma_\alpha^2,\mu,\boldsymbol{\alpha},\boldsymbol{\delta},\sigma_e^2)$$

But, can see that

$$f(\boldsymbol{y}|\sigma_\alpha^2,\mu,\boldsymbol{\alpha},\boldsymbol{\delta},\sigma_e^2) \propto f(\boldsymbol{y}|\mu,\boldsymbol{\alpha},\boldsymbol{\delta},\sigma_e^2)$$

So,

$$f(\sigma_\alpha^2|\boldsymbol{y},\mu,\boldsymbol{\alpha},\boldsymbol{\delta},\sigma_e^2) \propto f(\sigma_\alpha^2,\mu,\boldsymbol{\alpha},\boldsymbol{\delta},\sigma_e^2)$$

Note that σ_α^2 appears only in $f(\alpha|\sigma_\alpha^2)$ and $f(\sigma_\alpha^2)$:

$$f(\alpha|\sigma_\alpha^2) \propto (\sigma_\alpha^2)^{-k/2}\exp\{-\frac{\boldsymbol{\alpha}'\boldsymbol{\alpha}}{2\sigma_\alpha^2}\}$$

and

$$f(\sigma_\alpha^2) \propto (\sigma_\alpha^2)^{-(\nu_\alpha+2)/2}\exp\{\frac{\nu_\alpha S_\alpha^2}{2\sigma_\alpha^2}\}$$

Full conditional for σ^2_α

Combining these two densities gives:

$$f(\sigma^2_\alpha|\boldsymbol{y},\mu,\boldsymbol{\alpha},\boldsymbol{\delta},\sigma^2_e) \propto (\sigma^2_\alpha)^{-(k+\nu_\alpha+2)/2}\exp\{\frac{\boldsymbol{\alpha}'\boldsymbol{\alpha}+\nu_\alpha S^2_\alpha}{2\sigma^2_\alpha}\}$$

So,

$$(\sigma^2_\alpha|\boldsymbol{y},\mu,\boldsymbol{\alpha},\boldsymbol{\delta},\sigma^2_e) \sim \tilde{\nu}_\alpha\tilde{S}^2_\alpha\chi^{-2}_{\tilde{\nu}_\alpha}$$

where

$$\tilde{\nu}_\alpha = k+\nu_\alpha$$

and

$$\tilde{S}^2_\alpha = \frac{\boldsymbol{\alpha}'\boldsymbol{\alpha}+\nu_\alpha S^2_\alpha}{\tilde{\nu}_\alpha}$$

Hyper parameter: S^2_α

If σ^2 is distributed as a scaled, inverse chi-square random variable with scale parameter S^2 and degrees of freedom ν

$$\mathrm{E}(\sigma^2) = \frac{\nu S^2}{\nu-2}$$

Recall that under some assumptions

$$\sigma^2_\alpha = \frac{V_a}{\sum_j 2p_jq_j}$$

So, we take

$$S^2_\alpha = \frac{(\nu_\alpha-2)V_a}{\nu_\alpha k(1-\pi)2\overline{pq}}$$

Full conditional for π

Using Bayes' theorem,

$$f(\pi|\boldsymbol{\delta}, \mu, \boldsymbol{\alpha}, \sigma^2_{\alpha}, \sigma^2_e, \mathbf{y}) \propto f(\mathbf{y}|\pi, \boldsymbol{\delta}, \mu, \boldsymbol{\alpha}, \sigma^2_{\alpha}, \sigma^2_e) f(\pi, \boldsymbol{\delta}, \mu, \boldsymbol{\alpha}, \sigma^2_{\alpha}, \sigma^2_e)$$

But,

- Conditional on δ the likelihood is free of π
- Further, π only appears in probability of the vector of bernoulli variables: δ

Thus,

$$f(\pi|\boldsymbol{\delta}, \mu, \boldsymbol{\alpha}, \sigma^2_{\alpha}, \sigma^2_e, \mathbf{y}) = \pi^{(k-m)}(1-\pi)^m$$

where $m = \delta'\delta$, and k is the number of markers. Thus, π is sampled from a beta distribution with $a = k - m + 1$ and $b = m + 1$.

BayesCπ with Unknown S^2_{α}

- Prior for S^2_{α}: Gamma(a,b)

$$f(S^2_{\alpha}|a, b) \propto b^a (S^2_{\alpha})^{a-1} \exp\{-bS^2_{\alpha}\}$$

- Using Bayes theorem,

$$f(S^2_{\alpha}|\boldsymbol{\delta}, \mu, \boldsymbol{\alpha}, \sigma^2_{\alpha}, \sigma^2_e, \mathbf{y}) \propto f(\mathbf{y}|S^2_{\alpha}, \sigma^2_{\alpha}, \ldots) f(S^2_{\alpha}, \sigma^2 \ldots)$$

- Given $\mu, \boldsymbol{\alpha}$, and σ^2_e, $f(\mathbf{y}|S^2_{\alpha}, \sigma^2_{\alpha}, \ldots)$ does not depend on S^2_{α}.
- In $f(S^2_{\alpha}, \sigma^2 \ldots)$, S^2_{α} is only in $f(S^2_{\alpha}|a, b)$ and $f(\sigma^2_{\alpha}|S^2_{\alpha}, \nu_{\alpha})$

BayesCπ with Unknown S_α^2

- Prior for S_α^2: Gamma(a,b)

$$f(S_\alpha^2|a,b) \propto b^a (S_\alpha^2)^{a-1} \exp\{-bS_\alpha^2\}$$

- Prior for σ_α^2:

$$f(\sigma_\alpha^2) \propto (\sigma_\alpha^2)^{-(\nu_\alpha+2)/2} \exp\{\frac{\nu_\alpha S_\alpha^2}{2\sigma_\alpha^2}\}$$

- Combining these gives:

$$f(S_\alpha^2|\sigma_\alpha^2, \mathbf{y}, \ldots) \propto S_\alpha^{2(a-1+\nu/2)} \exp\{-S_\alpha^2(\frac{\nu_\alpha}{2\sigma_\alpha^2} + b)\}$$

BayesCπ with Unknown S_α^2

So, $f(S_\alpha^2|a,b)$ is Gamma(a*,b*), where

$$a* = a + \nu_\alpha/2$$

and

$$b* = b + \frac{\nu_\alpha}{2\sigma_\alpha^2}$$

Simulation I

- 2000 unlinked loci in LE
- 10 of these are QTL: $\pi = 0.995$
- $h^2 = 0.5$
- Locus effects estimated from 250 individuals

Results for Bayes-B

Correlations between true and predicted additive genotypic values estimated from 32 replications

π	S^2	Correlation
0.995	0.2	0.91 (0.009)
0.8	0.2	0.86 (0.009)
0.0	0.2	0.80 (0.013)
0.995	2.0	0.90 (0.007)
0.8	2.0	0.77 (0.009)
0.0	2.0	0.35 (0.022)

Simulation II

- 2000 unlinked loci with Q loci having effect on trait
- N is the size of training data set
- Heritability = 0.5
- Validation in an independent data set with 1000 individuals
- Bayes-B and Bayes-Cπ with $\pi = 0.5$

Results

Results from 15 replications

				Corr($g, \hat{g}$)	
N	Q	π	$\hat{\pi}$	Bayes-Cπ	Bayes-B
2000	10	0.995	0.994	0.995	0.937
2000	200	0.90	0.899	0.866	0.834
2000	1900	0.05	0.202	0.613	0.571
4000	1900	0.05	0.096	0.763	0.722

Simulation III

- Genotypes: 50k SNPs from 1086 Purebred Angus animals, ISU
- Phenotypes:
  - QTL simulated from 50 randomly sampled SNPs
  - substitution effect sampled from N(0,σ^2_α)
  - $\sigma^2_\alpha = \frac{\sigma^2_g}{502\overline{pq}}$
  - $h^2 = 0.25$
- QTL were included in the marker panel
- Marker effects were estimated for 50k SNPs

Validation

- Genotypes: 50k SNPs from 984 crossbred animals, CMP
- Additive genetic merit (g_i) computed from the 50 QTL
- Additive genetic merit predicted ($\hat{g}_i$) using estimated effects for 50k SNP panel

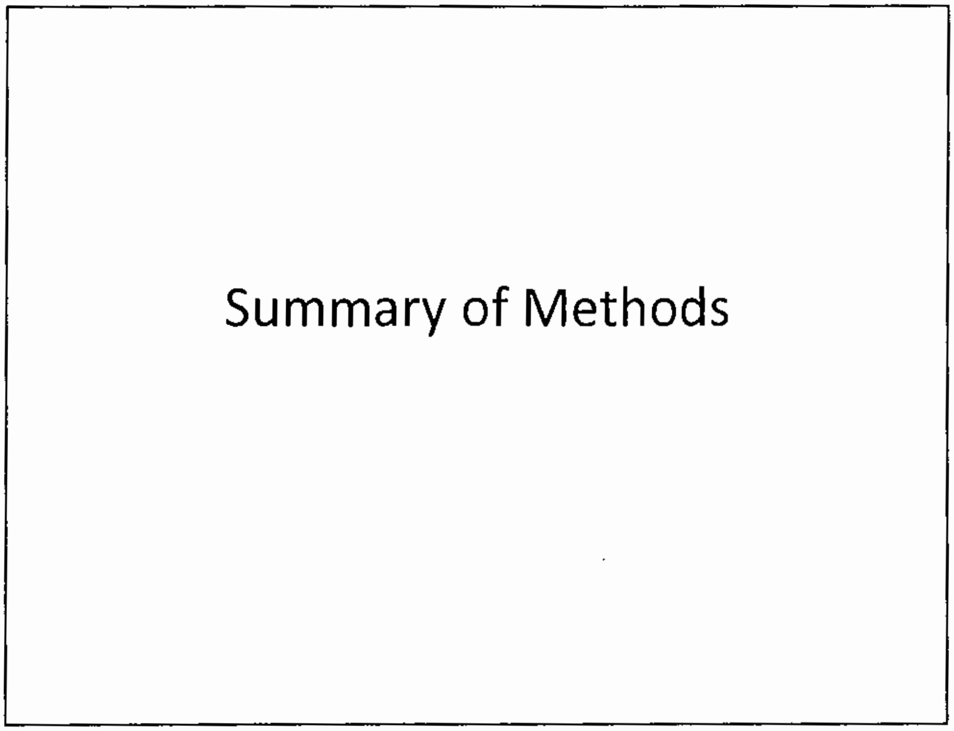
Results

Correlations between g_i and $\hat{g}_i$ estimated from 3 replications

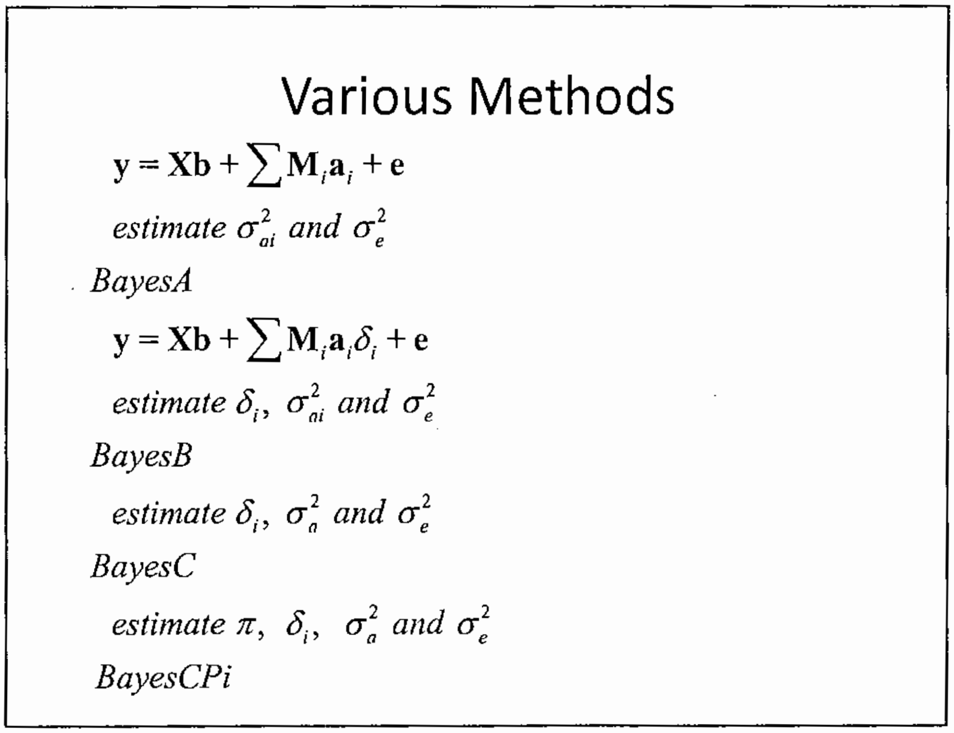
	Correlation	
π	Bayes-B	Bayes-C
0.999	0.86	0.86
0.25	0.70	0.26

BayesCπ:

- $\hat{\pi} = 0.999$
- Correlation = 0.86



Summary of Methods



Various Methods

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

estimate σ^2_{ai} *and* σ^2_e

BayesA

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

estimate δ_i, σ^2_{ai} *and* σ^2_e

BayesB

estimate δ_i, σ^2_a *and* σ^2_e

BayesC

estimate π, δ_i, σ^2_a *and* σ^2_e

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

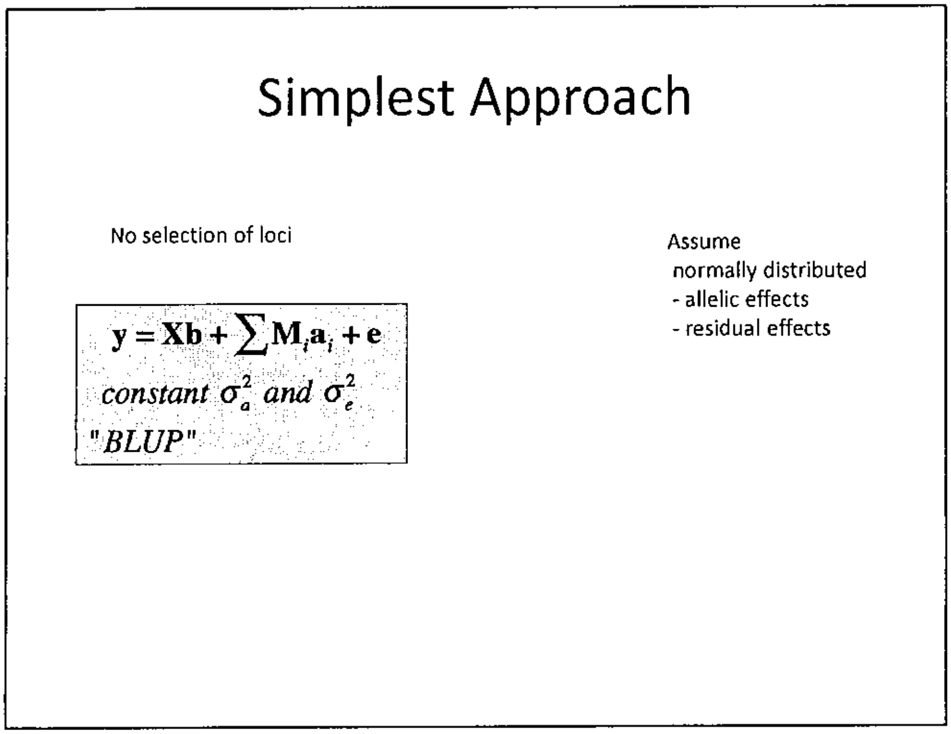
dorian@iastate.edu

Animal Science

1858 150 2008

Iowa State University

Animal Breeding & Genetics

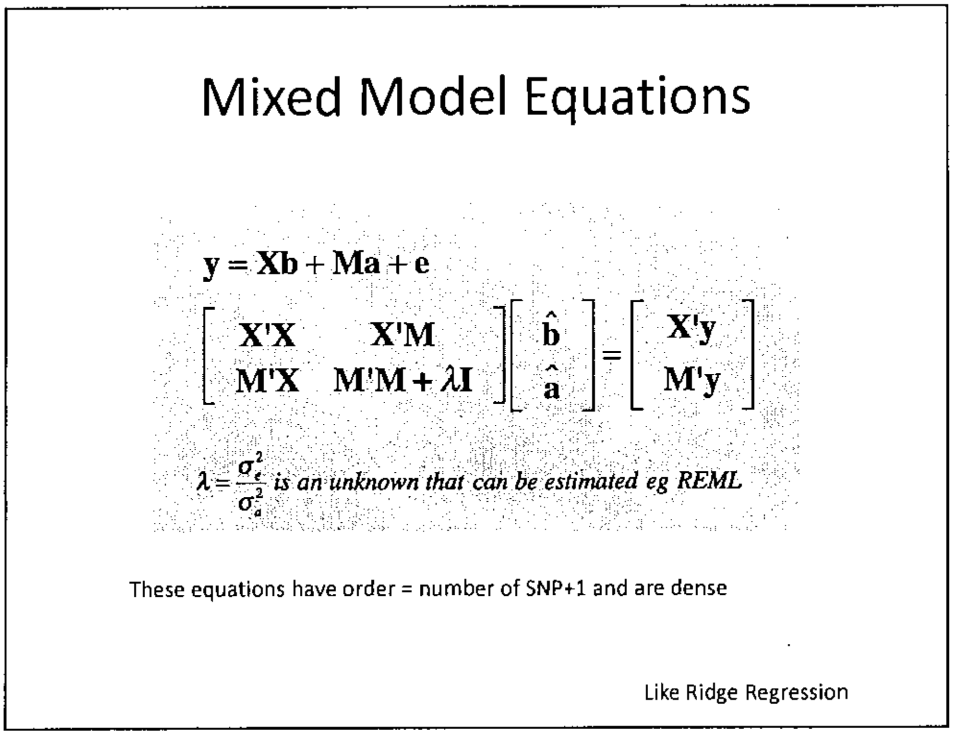


Simplest Approach

No selection of loci

$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$
$$\textit{constant } \sigma_a^2 \textit{ and } \sigma_e^2$$
$$\textit{"BLUP"}$$

Assume
normally distributed
- allelic effects
- residual effects



Mixed Model Equations

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

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

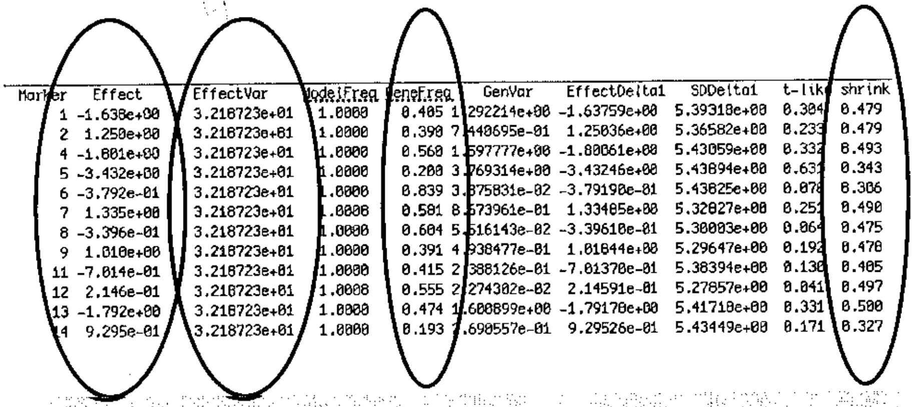
$$\lambda = \frac{\sigma_e^2}{\sigma_a^2} \textit{ is an unknown that can be estimated eg REML}$$

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

Like Ridge Regression

[illegible]

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 0.078  0.306
    7  1.335e+00  3.218723e+01  1.0008    0.581 8.573961e-01  1.33405e+00 5.32027e+00 0.251  0.490
    8 -3.396e-01  3.218723e+01  1.0000    0.604 5.516143e-02 -3.39610e-01 5.30003e+00 0.064  0.475
    9  1.010e+00  3.218723e+01  1.0000    0.391 4.938477e-01  1.01044e+00 5.29647e+00 0.192  0.470
   11 -7.014e-01  3.218723e+01  1.0000    0.415 2.388126e-01 -7.01370e-01 5.38394e+00 0.130  0.405
   12  2.146e-01  3.218723e+01  1.0008    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 -1.79170e+00 5.41710e+00 0.331  0.500
   14  9.295e-01  3.218723e+01  1.0000    0.193 7.690557e-01  9.29526e-01 5.43449e+00 0.171  0.327
```

$\hat{\mathbf{a}}$ σ_a^2 $2pq$ $Shrinkage = \dfrac{BLUP\ estimate}{OLS\ estimate}$

Equivalent Model (All SNPs)

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

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

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

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

Current method using genomic G instead of pedigree A $\mathbf{G} = \sum M_i M_i{}'$

Analytical Methods

No selection of loci

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

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

$"BLUP"$

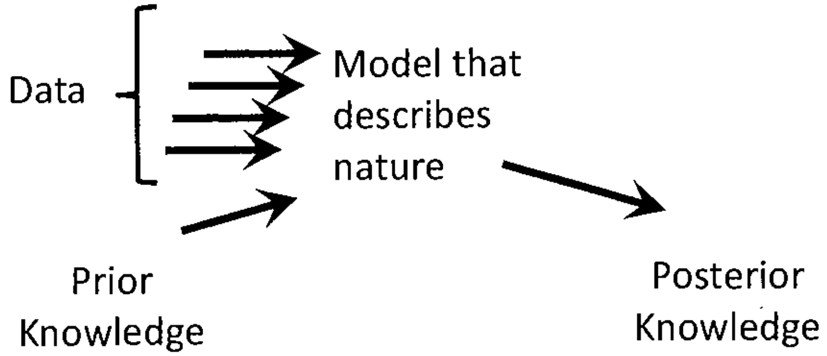
$SNP - specific\ \sigma_{ai}^2\ and\ \sigma_e^2$

$BayesA$

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

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

$$\text{so} \quad S_{v_a}^2 = \frac{(v_a-2)V_a}{v_a k2\overline{p}(1-\overline{p})} \quad \text{for } k \text{ SNP}$$

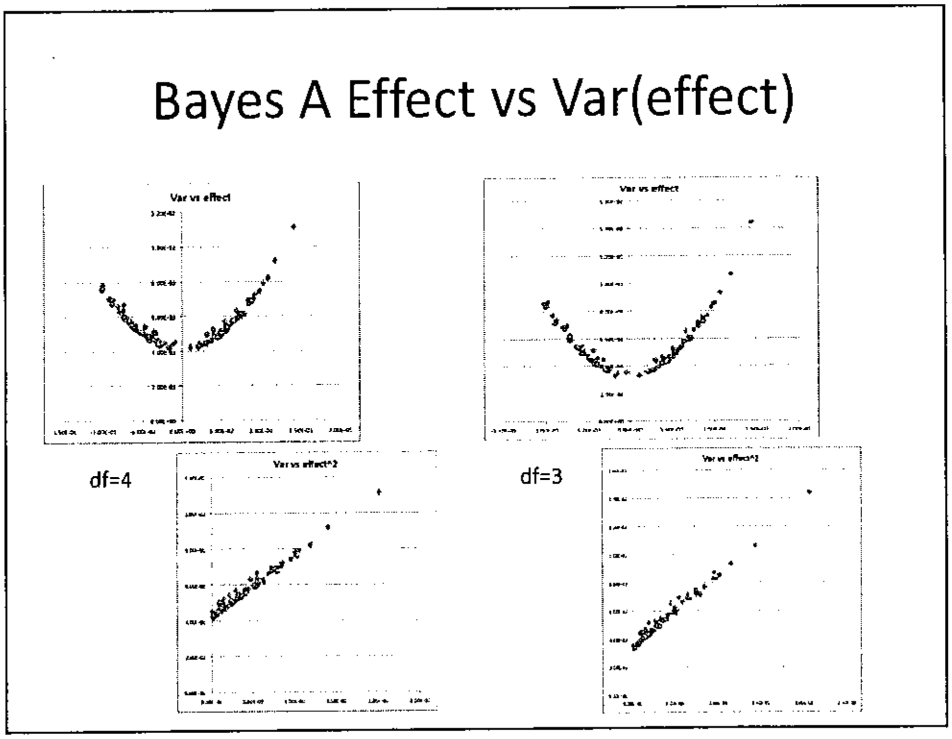
8,300 Holstein Bulls w/50k

```
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   3.708718e+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    0.280 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    0.501 6.670957e-01  1.17062e+00  5.50165e+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  0.540
   9  5.559e-01   3.567120e+01   1.0000    0.391 1.471572e-01  5.55940e-01  5.20357e+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 1.046104e-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 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 σ_a^2 *and* σ_e^2

"*BLUP*"

estimate σ_{ai}^2 *and* σ_e^2

BayesA

Mixture Models (model selection)

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

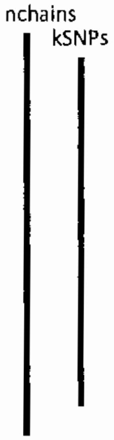
estimate δ_i, σ_{ai}^2 *and* σ_e^2

BayesB (*known* π)

π = *fraction loci with no effect*

Meuwissen, Hayes & Goddard (2001)

Mixture Models


nchains

kSNPs

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

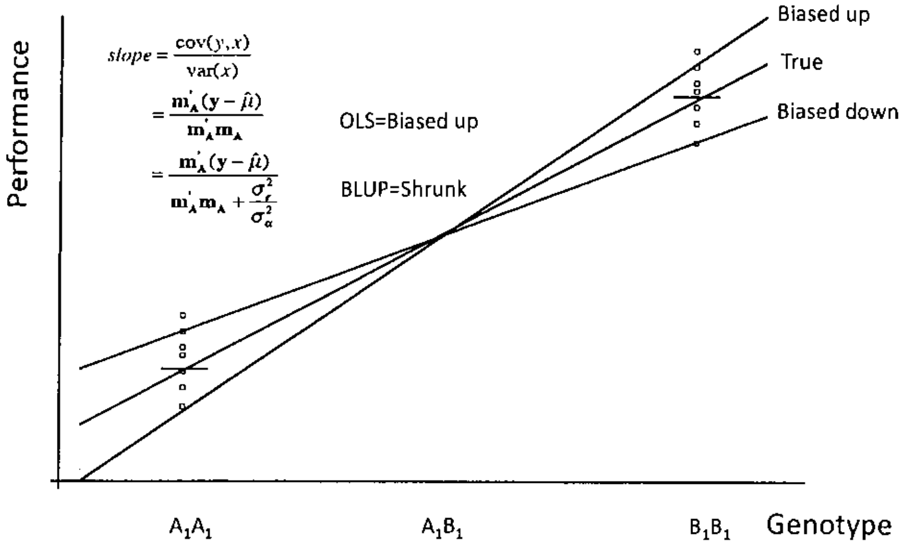
$$\delta_i = 1 \quad L_1 = L(\mathbf{Xb} + \mathbf{M}_i \mathbf{a}_i + \mathbf{e}) \;\; given \;\; (1-\pi)$$

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

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

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

Shrinkage Estimation



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

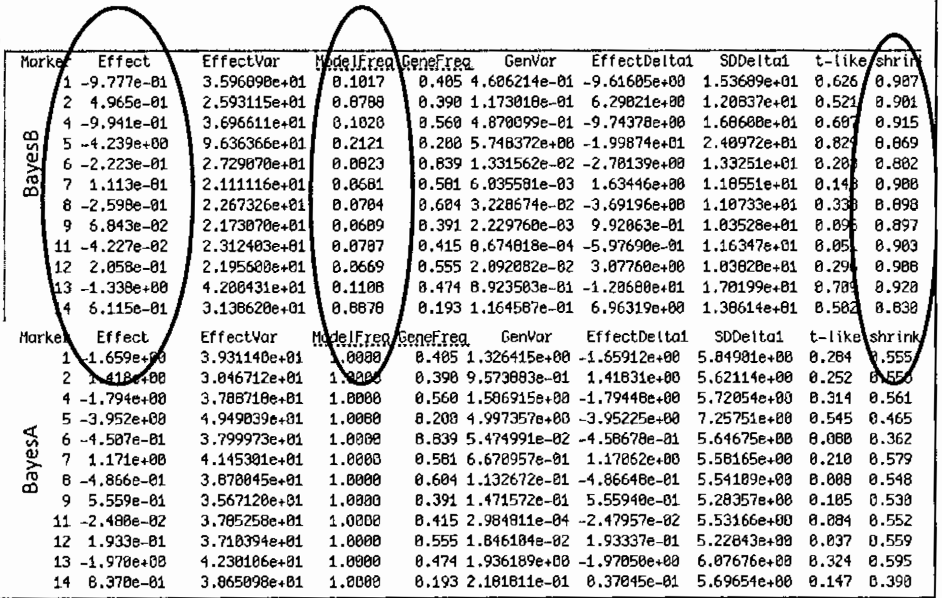
OLS=Biased up

BLUP=Shrunk

Bayesian Estimation

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

Bayes A vs B marker effects



BayesB

```
Marker    Effect         EffectVar   ModelFreq GeneFreq       GenVar    EffectDelta1     SDDelta1  t-like shrink
   1 -9.777e-01     3.596090e+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  6.29021e+00  1.20837e+01  0.521  0.901
   4 -9.941e-01     3.696611e+01     0.1020     0.560 4.870099e-01 -9.74378e+00  1.60600e+01  0.607  0.915
   5 -4.239e+00     9.636366e+01     0.2121     0.200 5.740372e+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     0.501 6.035501e-03  1.63446e+00  1.10551e+01  0.148  0.900
   8 -2.598e-01     2.267326e+01     0.0704     0.604 3.220674e-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  9.92063e-01  1.03528e+01  0.096  0.897
  11 -4.227e-02     2.312403e+01     0.0707     0.415 0.674018e-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  1.03020e+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     0.0870     0.193 1.164587e-01  6.96319e+00  1.30614e+01  0.502  0.830
```

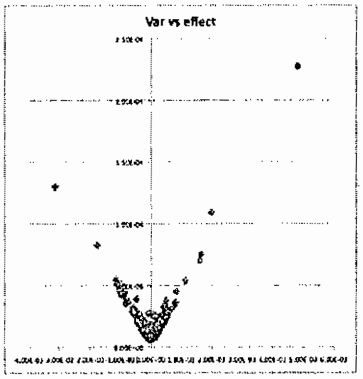
BayesA

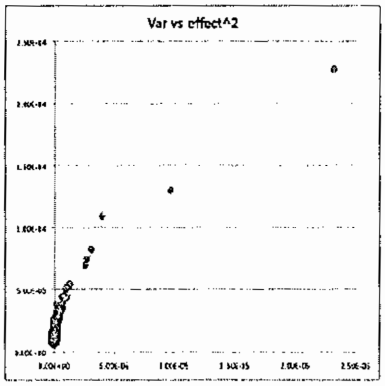
```
Marker    Effect         EffectVar   ModelFreq GeneFreq       GenVar    EffectDelta1     SDDelta1  t-like shrink
   1 -1.659e+00     3.931140e+01     1.0000     0.405 1.326415e+00 -1.65912e+00  5.84901e+00  0.284  0.555
   2  1.418e+00     3.046712e+01     1.0000     0.390 9.573083e-01  1.41831e+00  5.62114e+00  0.252  0.556
   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     1.0080     0.208 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 -4.58678e-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     3.870045e+01     1.0000     0.604 1.132672e-01 -4.86648e-01  5.54109e+00  0.008  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     3.705258e+01     1.0000     0.415 2.984811e-04 -2.47957e-02  5.53166e+00  0.084  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     4.230106e+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  0.37045e-01  5.69654e+00  0.147  0.390
```

Bayes B Effect vs Var(Effect)

df=4

π = 0.99





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 σ_a^2 *and* σ_e^2

"BLUP"

estimate σ_{ai}^2 *and* σ_e^2

BayesA

Mixture Models (model selection)

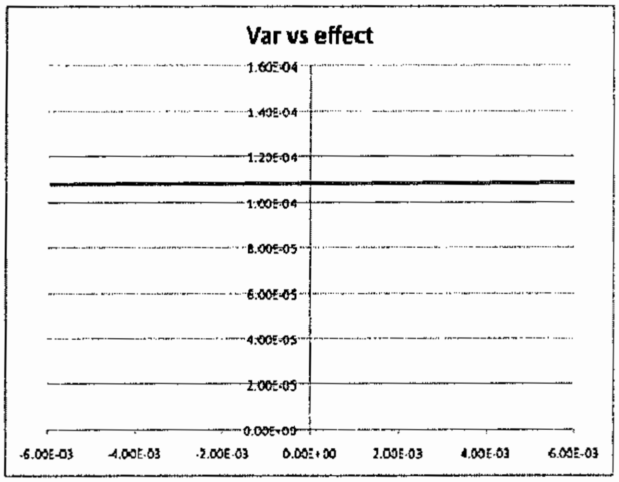
$$\mathbf{y} = \mathbf{Xb} + \sum \mathbf{M}_i\mathbf{a}_i\delta_i + \mathbf{e}$$

estimate δ_i, σ_{ai}^2 *and* σ_e^2

BayesB (known π*)*

π *= fraction loci with no effect*

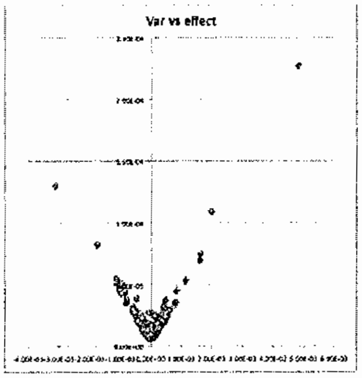
Meuwissen, Hayes & Goddard (2001)

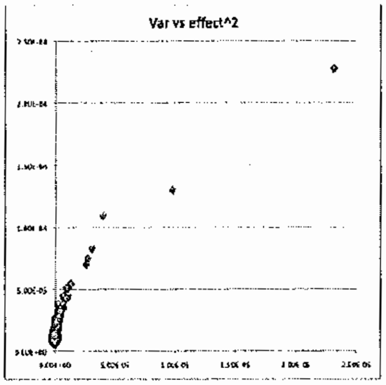


Bayes C0

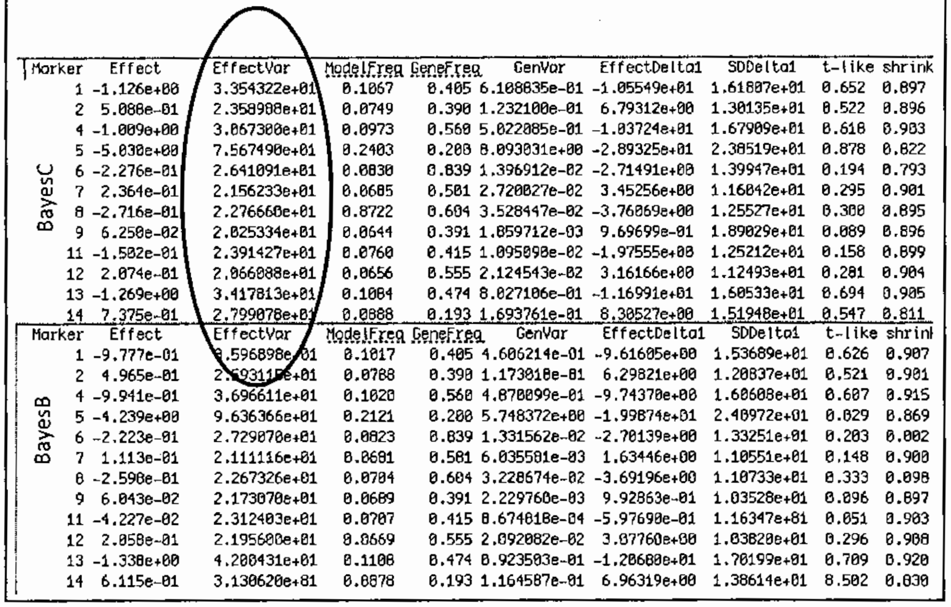
Bayes C (pi>0) or Bayes CPi

Like the following





Bayes C Var(Effect)

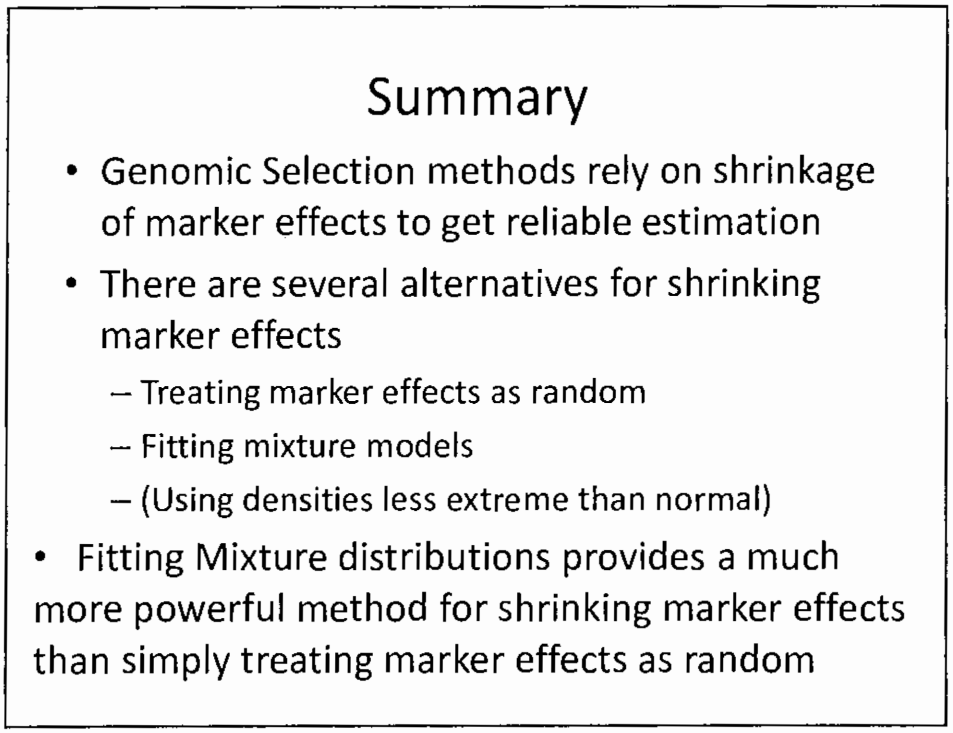


BayesC

```
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 -2.89325e+01 2.38519e+01  0.878  0.822
     6 -2.276e-01  2.641091e+01   0.0830     0.839 1.396912e-02 -2.71491e+00 1.39947e+01  0.194  0.793
     7  2.364e-01  2.156233e+01   0.0685     0.501 2.720027e-02  3.45256e+00 1.16042e+01  0.295  0.901
     8 -2.716e-01  2.276660e+01   0.8722     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 1.89029e+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 8.027106e-01 -1.16991e+01 1.60533e+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

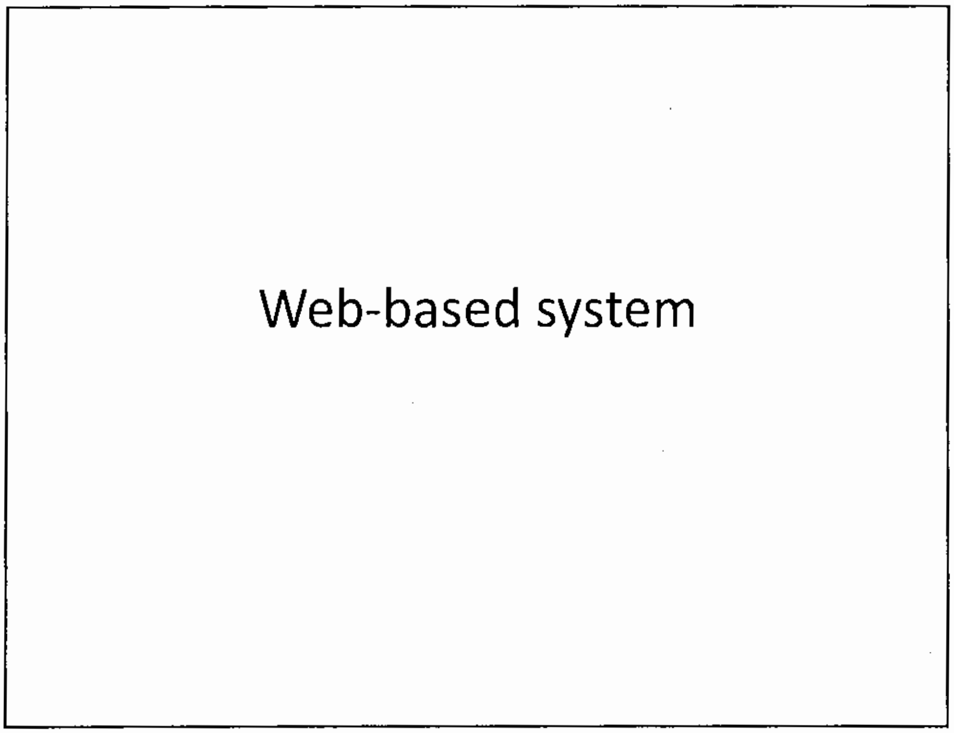
```
Marker   Effect      EffectVar    ModelFreq GeneFreq     GenVar     EffectDelta1   SDDelta1   t-like shrink
     1 -9.777e-01  3.596898e+01   0.1017     0.405 4.606214e-01 -9.61605e+00 1.53689e+01  0.626  0.907
     2  4.965e-01  2.693116e+01   0.0788     0.390 1.173018e-01  6.29821e+00 1.20837e+01  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  0.029  0.869
     6 -2.223e-01  2.729070e+01   0.0823     0.839 1.331562e-02 -2.70139e+00 1.33251e+01  0.203  0.902
     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  6.043e-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 1.03820e+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  3.130620e+01   0.0878     0.193 1.164587e-01  6.96319e+00 1.38614e+01  0.502  0.830
```



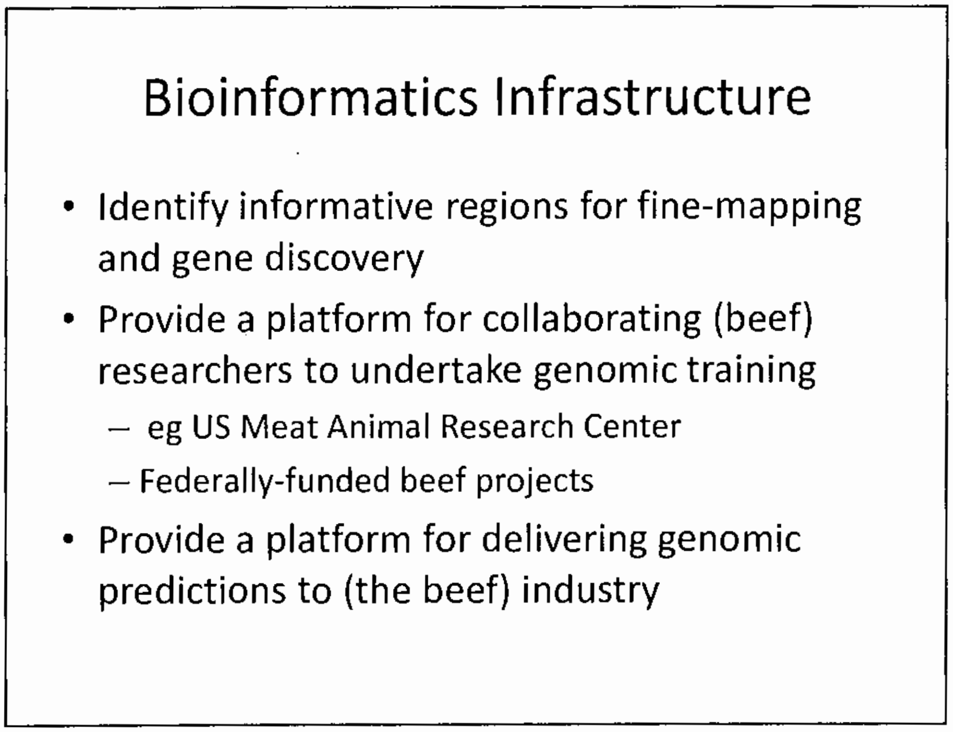
Summary

- 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)

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

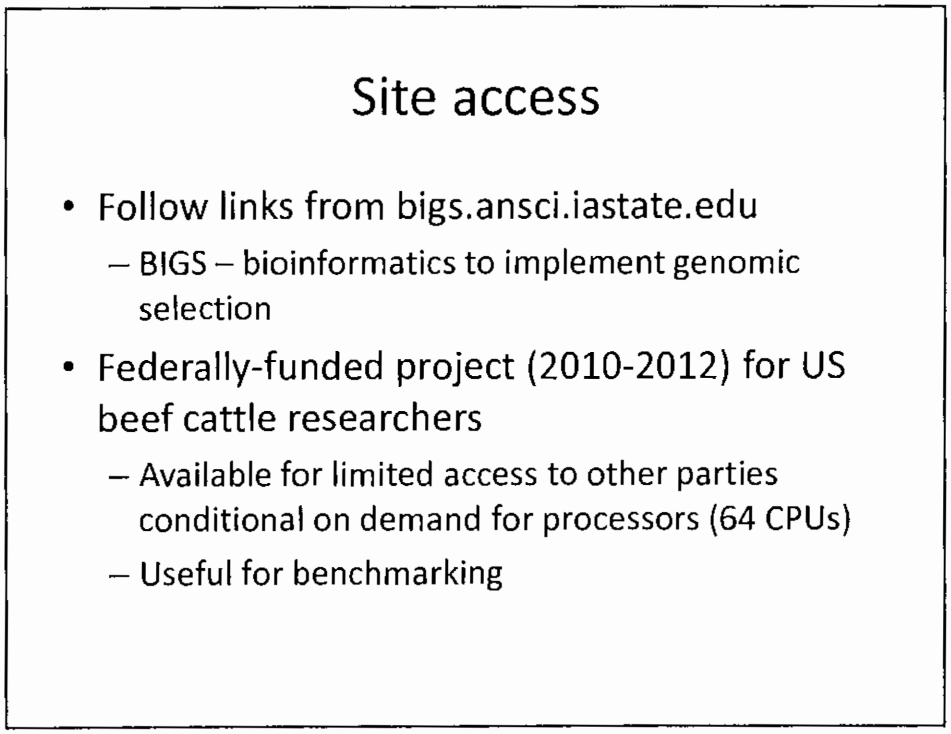


Web-based system



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



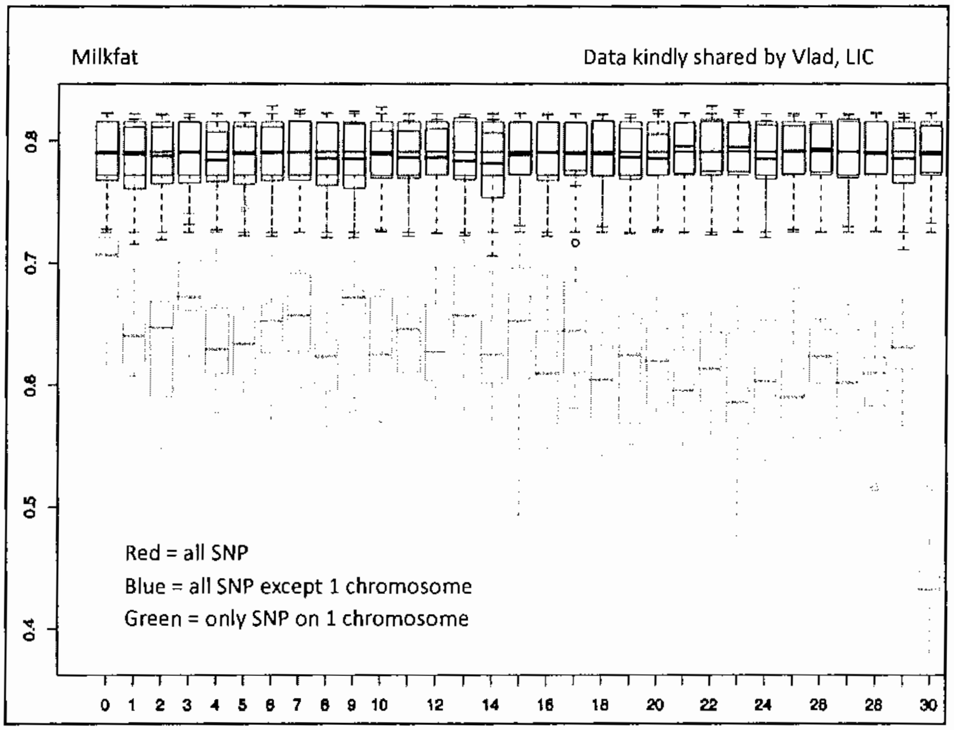
Site access

- 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

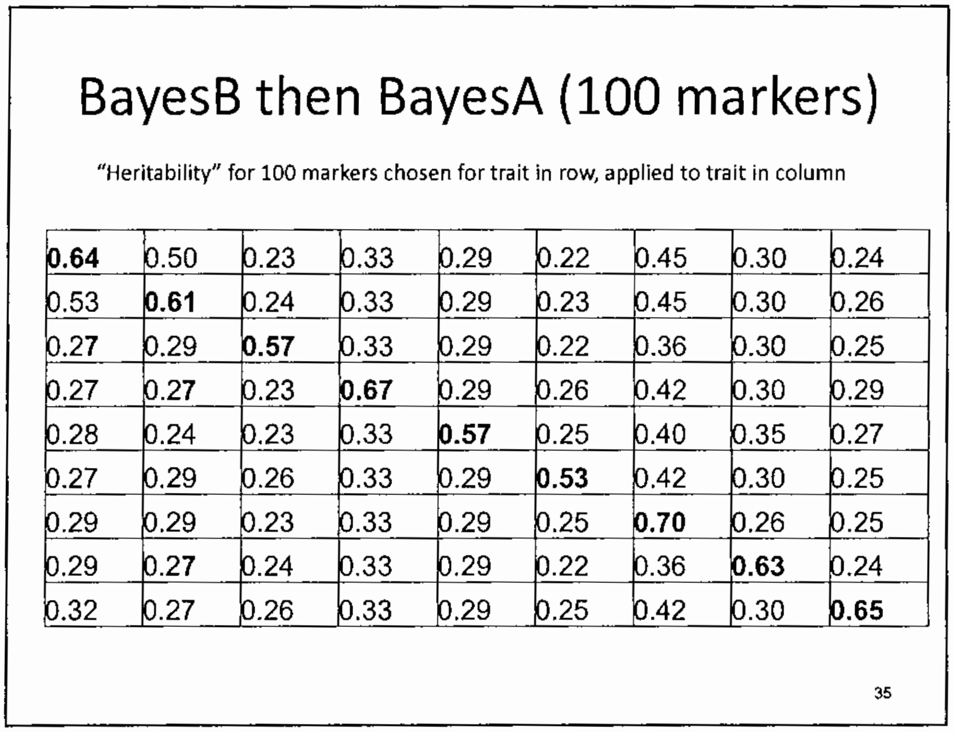
Data kindly shared by Vlad, LIC

Red = all SNP

Blue = all SNP except 1 chromosome

Green = only SNP on 1 chromosome

Problems with Validation

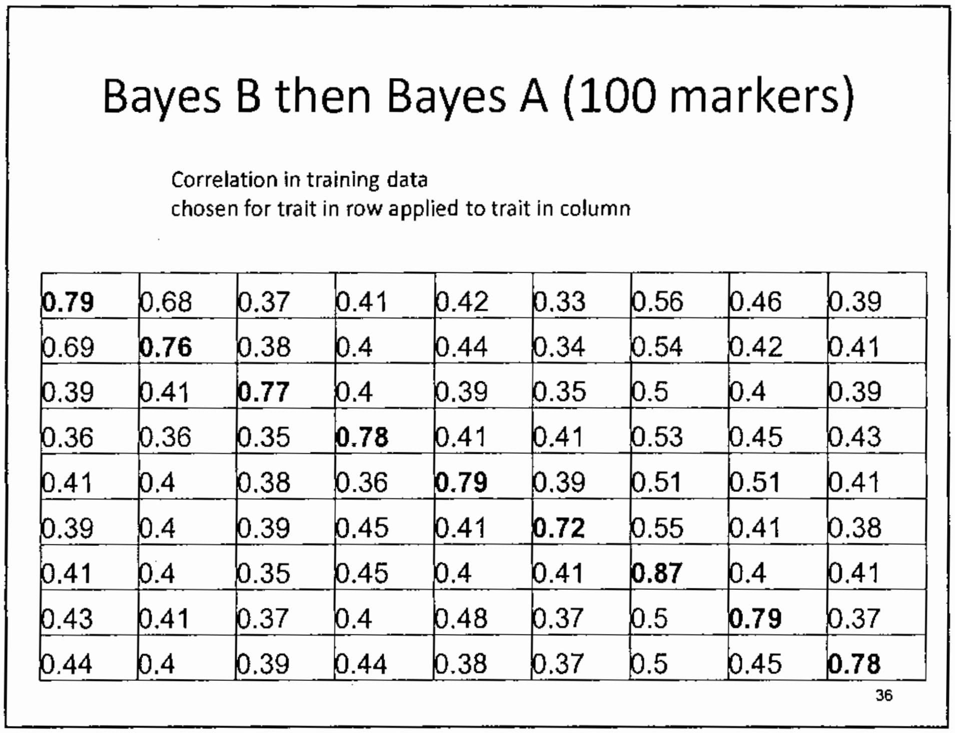


BayesB then BayesA (100 markers)

“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**

35

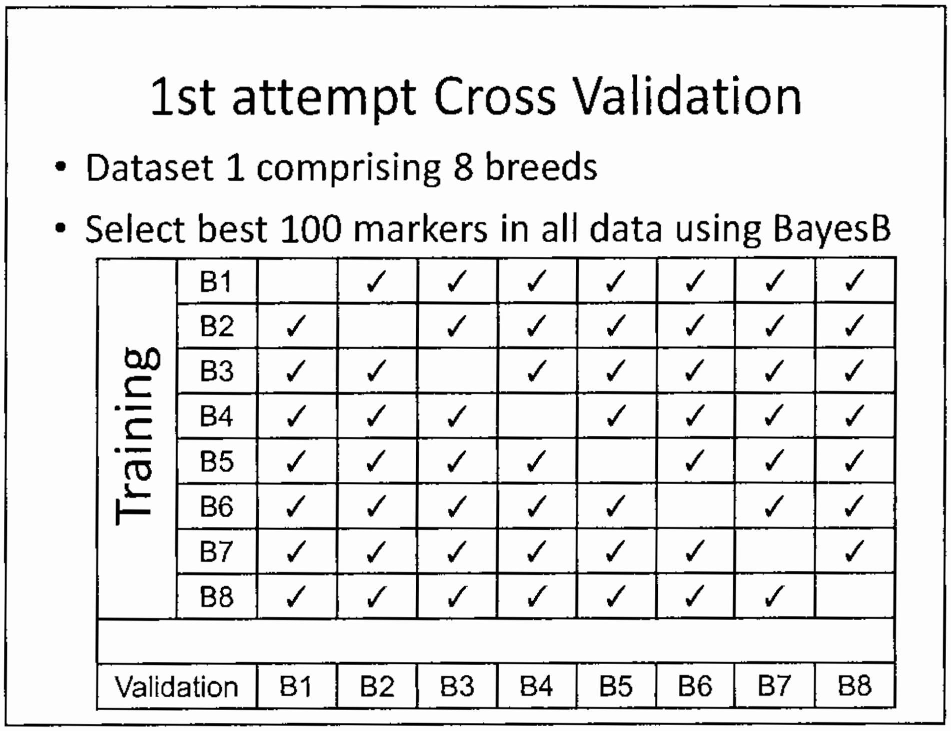


Bayes B then Bayes A (100 markers)

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**

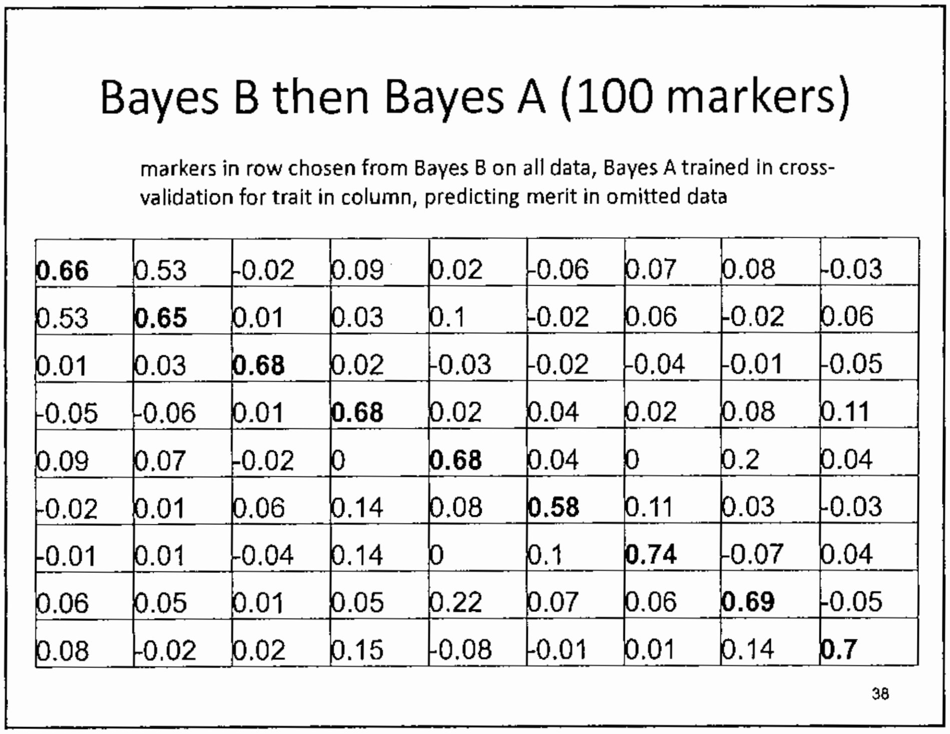
36



1st attempt Cross Validation

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

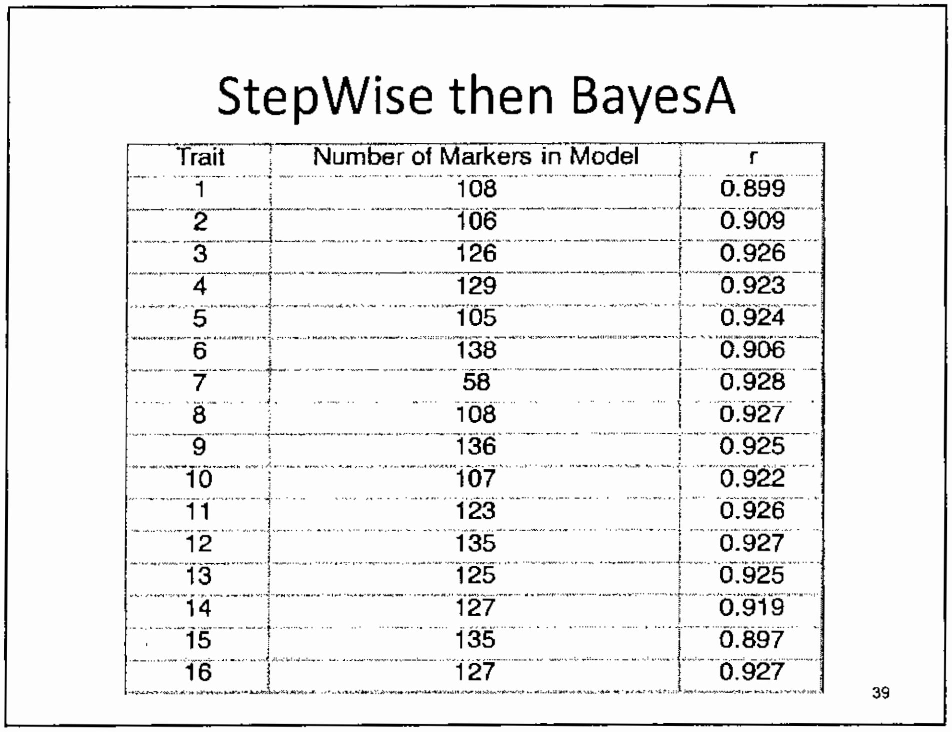
Training	B1		✓	✓	✓	✓	✓	✓	✓
	B2	✓		✓	✓	✓	✓	✓	✓
	B3	✓	✓		✓	✓	✓	✓	✓
	B4	✓	✓	✓		✓	✓	✓	✓
	B5	✓	✓	✓	✓		✓	✓	✓
	B6	✓	✓	✓	✓	✓		✓	✓
	B7	✓	✓	✓	✓	✓	✓		✓
	B8	✓	✓	✓	✓	✓	✓	✓	
Validation		B1	B2	B3	B4	B5	B6	B7	B8



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

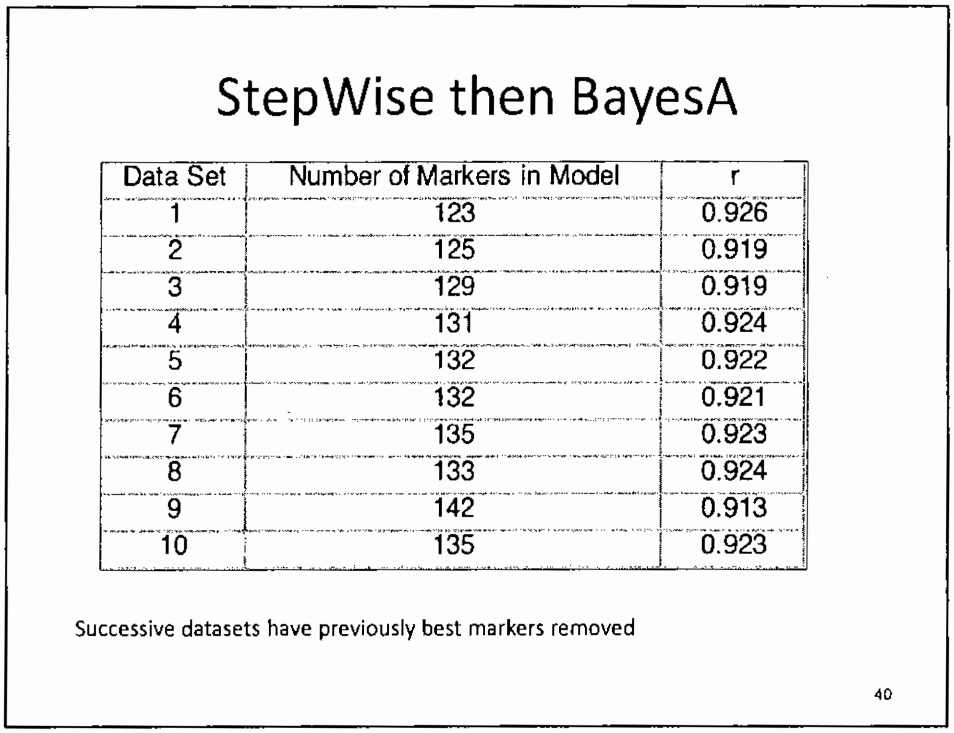
0.66	0.53	-0.02	0.09	0.02	-0.06	0.07	0.08	-0.03
0.53	**0.65**	0.01	0.03	0.1	-0.02	0.06	-0.02	0.06
0.01	0.03	**0.68**	0.02	-0.03	-0.02	-0.04	-0.01	-0.05
-0.05	-0.06	0.01	**0.68**	0.02	0.04	0.02	0.08	0.11
0.09	0.07	-0.02	0	**0.68**	0.04	0	0.2	0.04
-0.02	0.01	0.06	0.14	0.08	**0.58**	0.11	0.03	-0.03
-0.01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

39



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

40

StepWise and BayesA

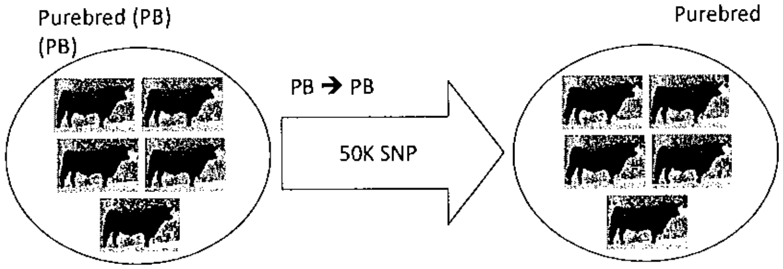
Data Set	Number of Markers in Model	r
Data Set 1	123	0.926
	90	0.880
	50	0.774
	25	0.627
	15	0.530
	10	0.458
Data Set 10	10	0.368

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 → PB

50K SNP

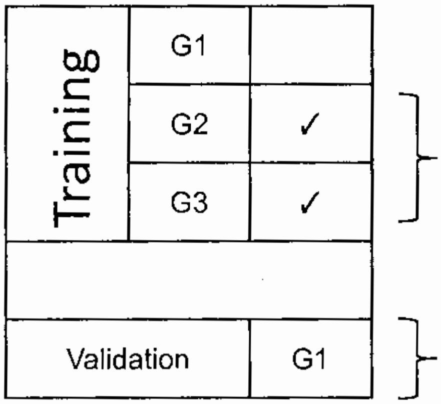
Purebred

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

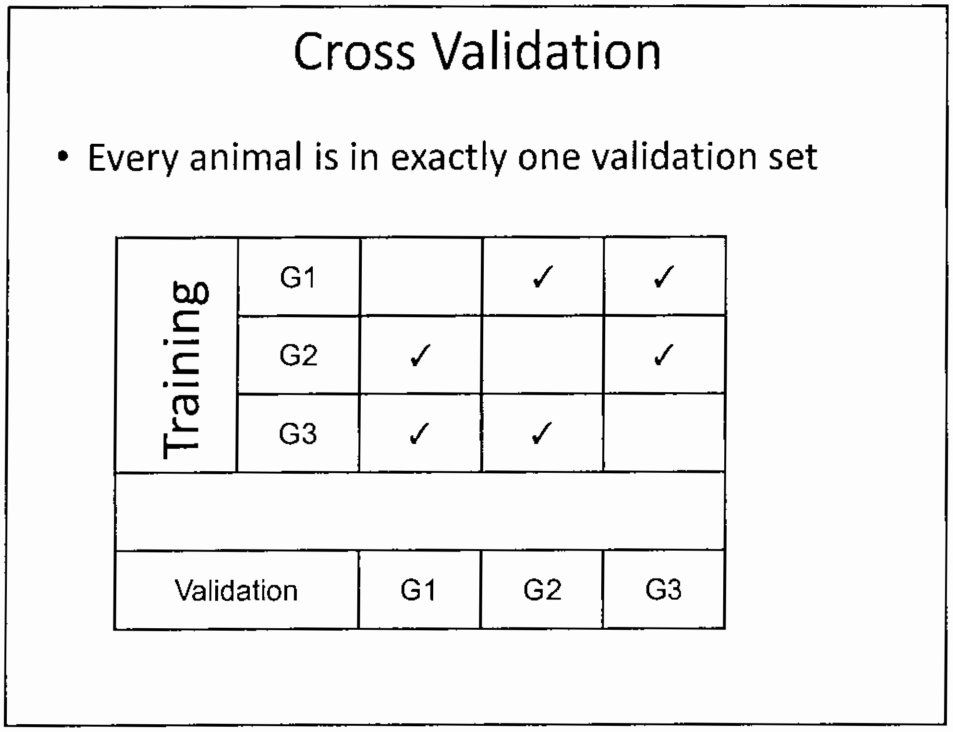
- Partition the dataset (by sire) into say three groups



Training	G1	
	G2	✓
	G3	✓
Validation		G1

Derive g-EPD

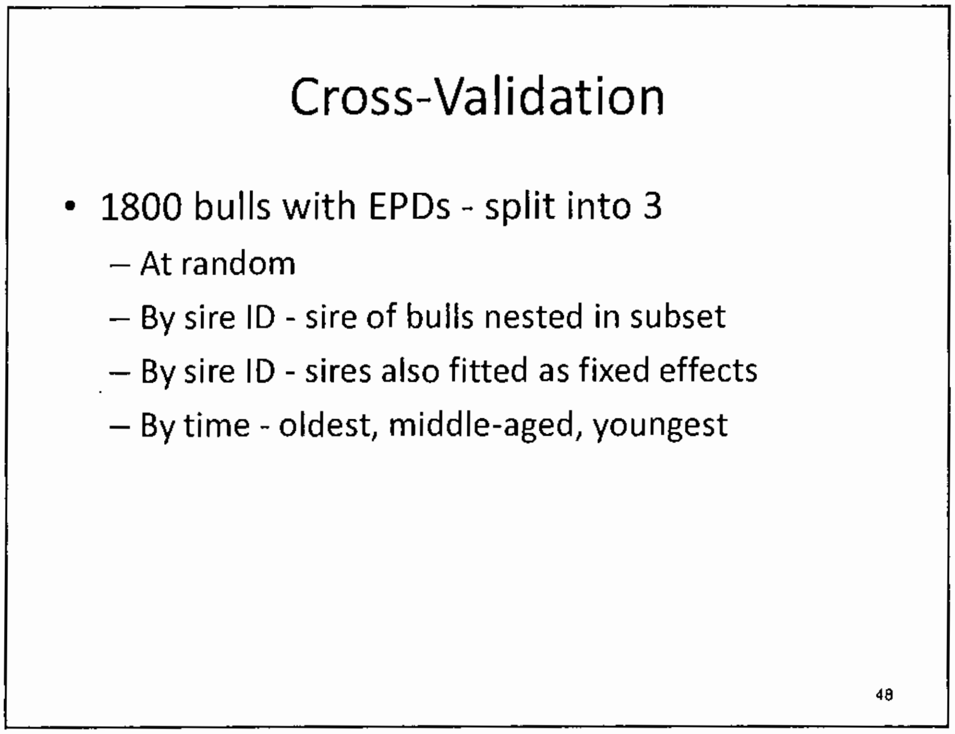
Compute the correlation between predicted genetic merit from g-EPD and observed performance



Cross Validation

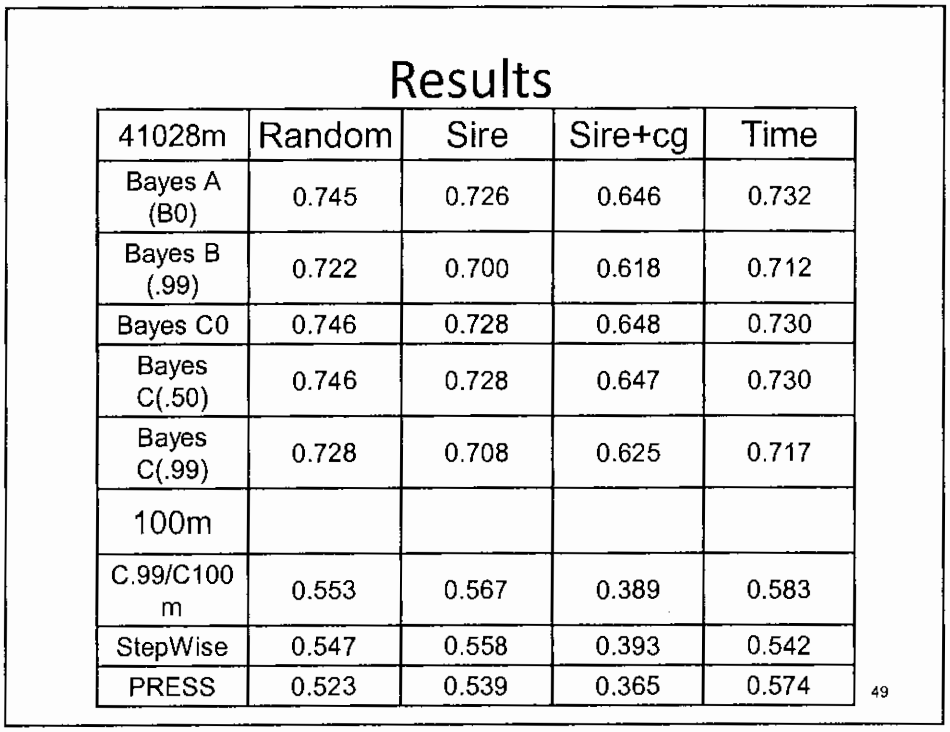
- Every animal is in exactly one validation set

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



Cross-Validation

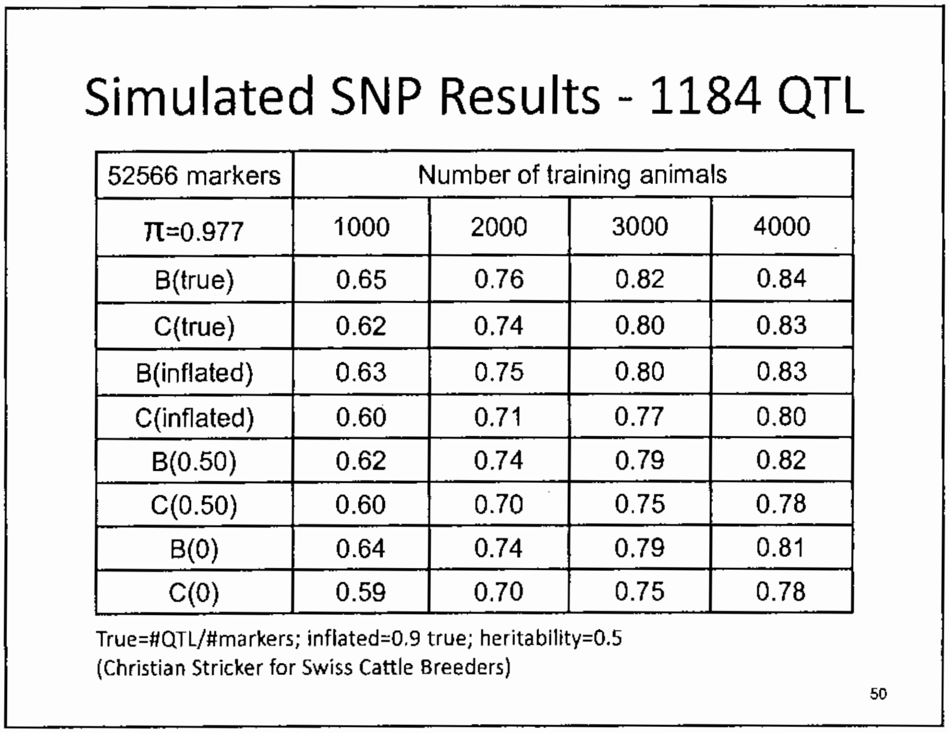
- 1800 bulls with EPDs - split into 3
  - At random
  - By sire ID - sire of bulls nested in subset
  - By sire ID - sires also fitted as fixed effects
  - By time - oldest, middle-aged, youngest



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/C100 m	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

49

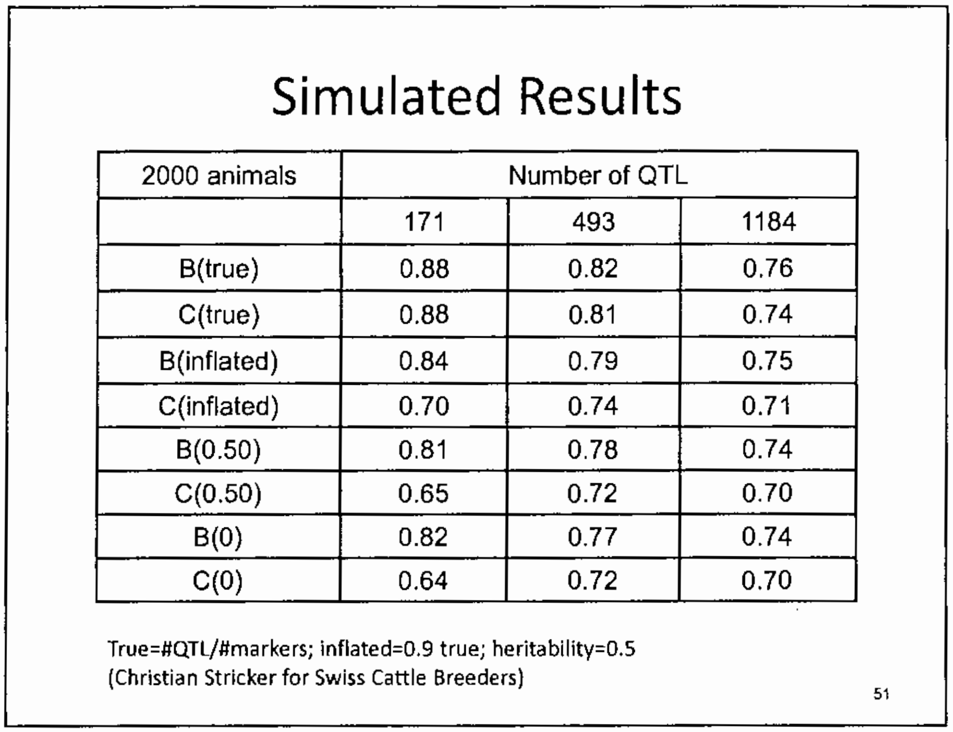


Simulated SNP Results - 1184 QTL

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)

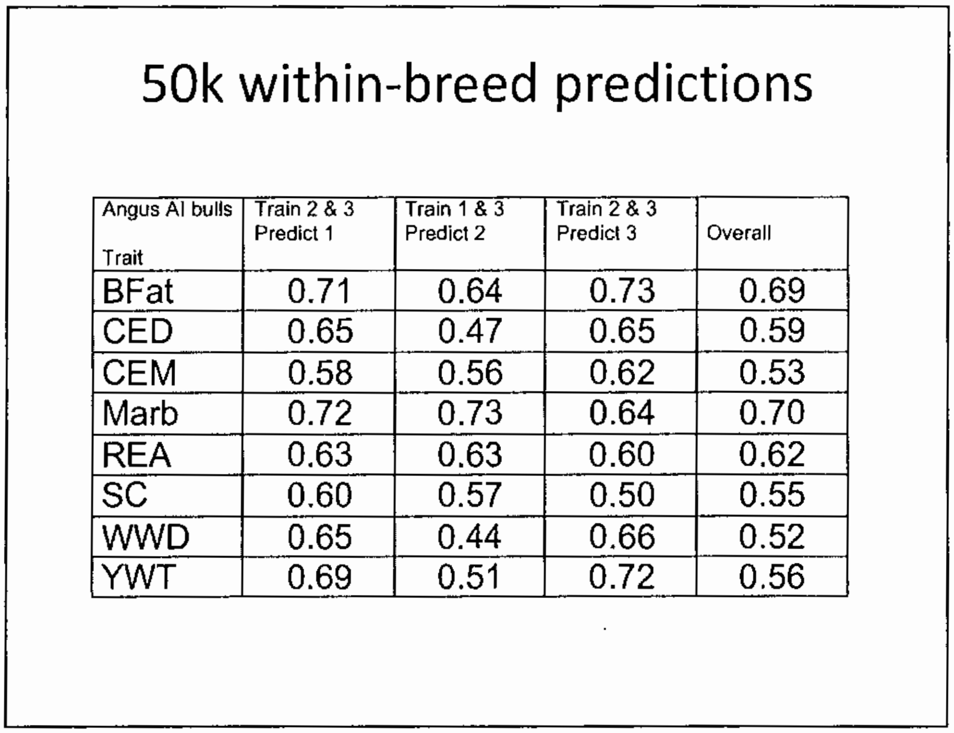
50



Simulated Results

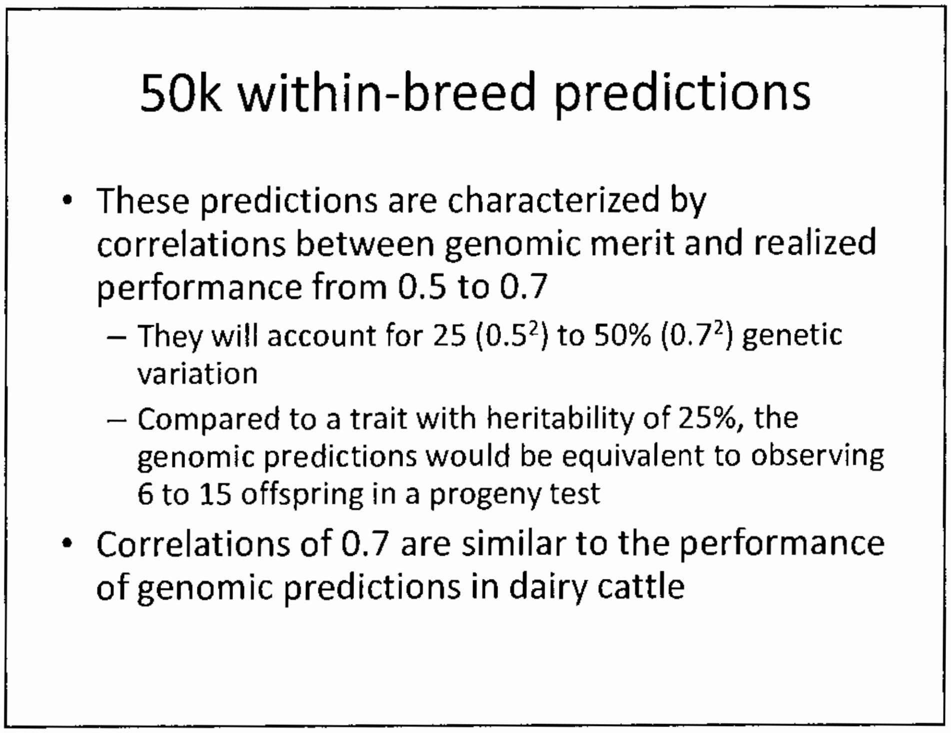
2000 animals	Number of QTL		
	171	493	1184
B(true)	0.88	0.82	0.76
C(true)	0.88	0.81	0.74
B(inflated)	0.84	0.79	0.75
C(inflated)	0.70	0.74	0.71
B(0.50)	0.81	0.78	0.74
C(0.50)	0.65	0.72	0.70
B(0)	0.82	0.77	0.74
C(0)	0.64	0.72	0.70

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



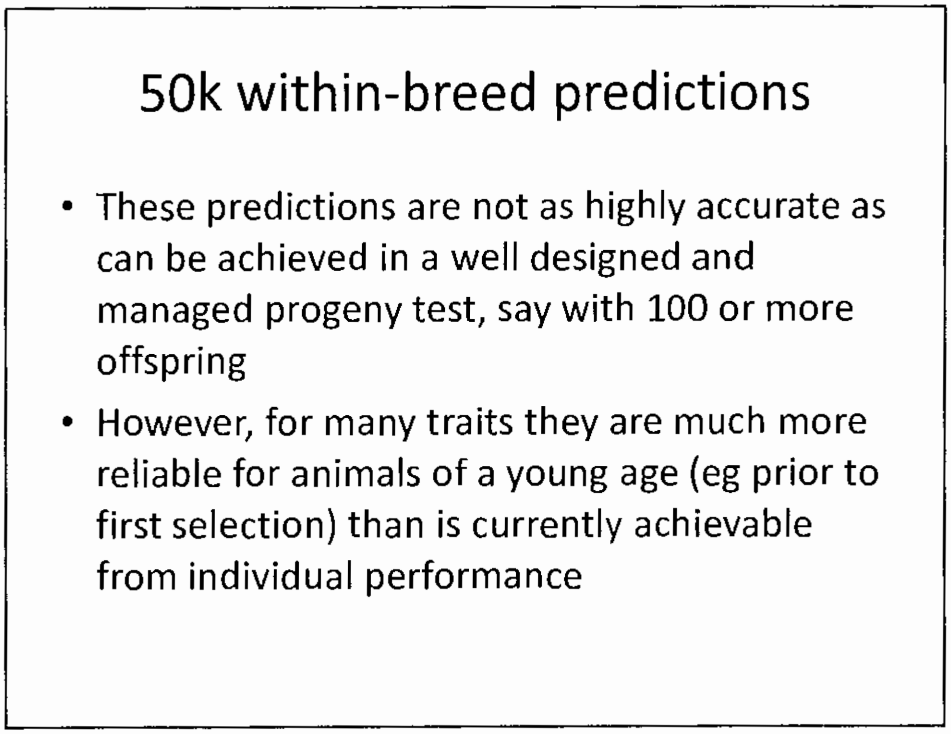
50k within-breed predictions

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
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



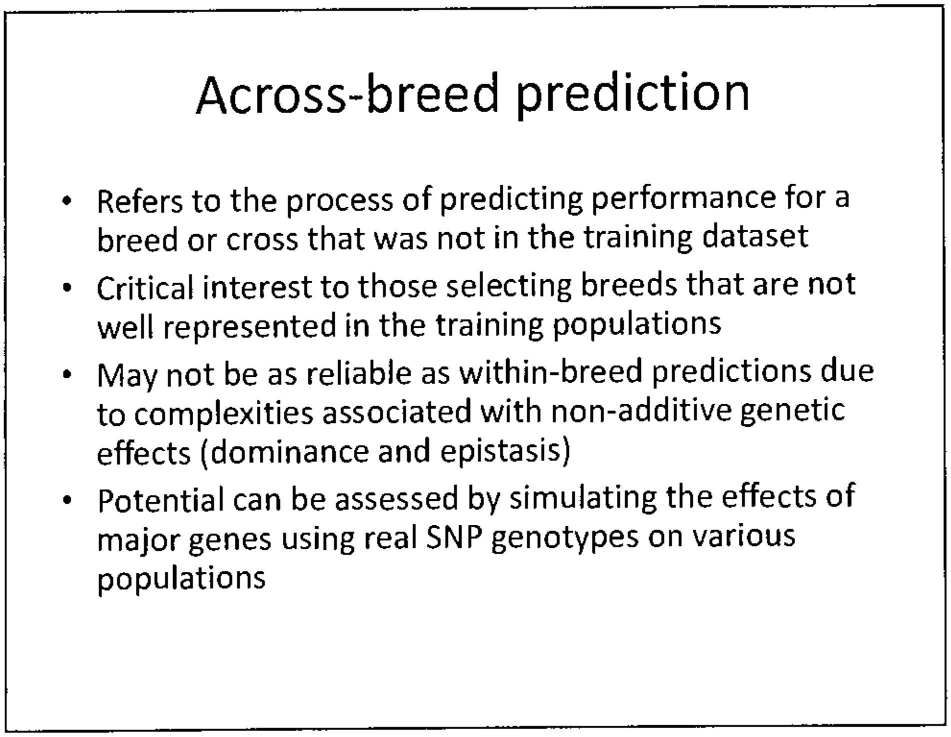
50k within-breed predictions

- 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



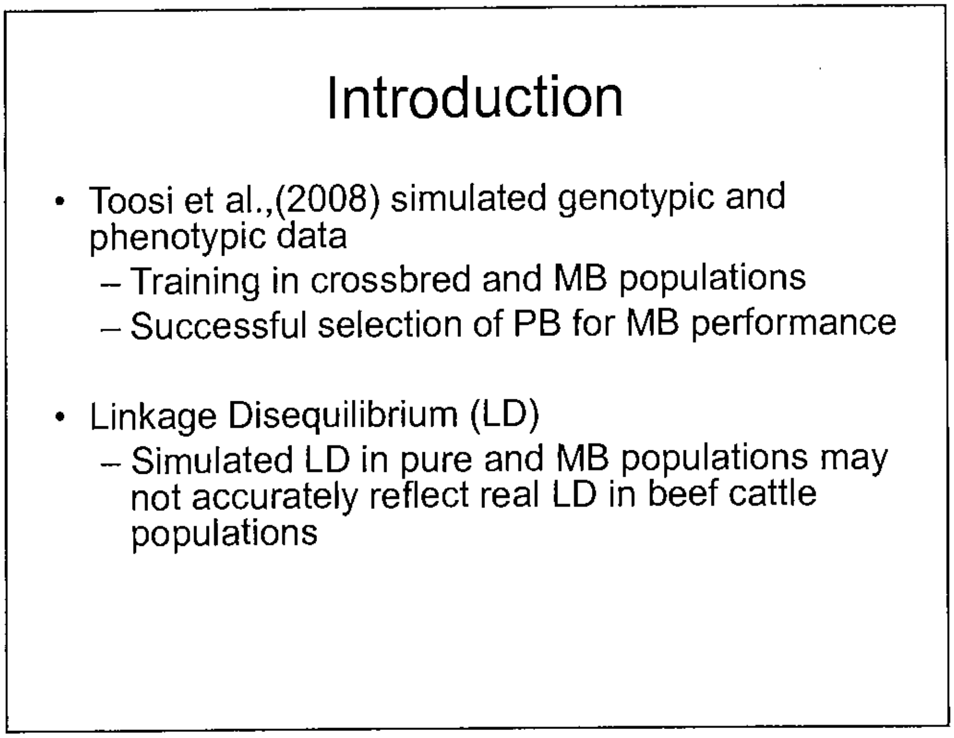
50k within-breed predictions

- 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



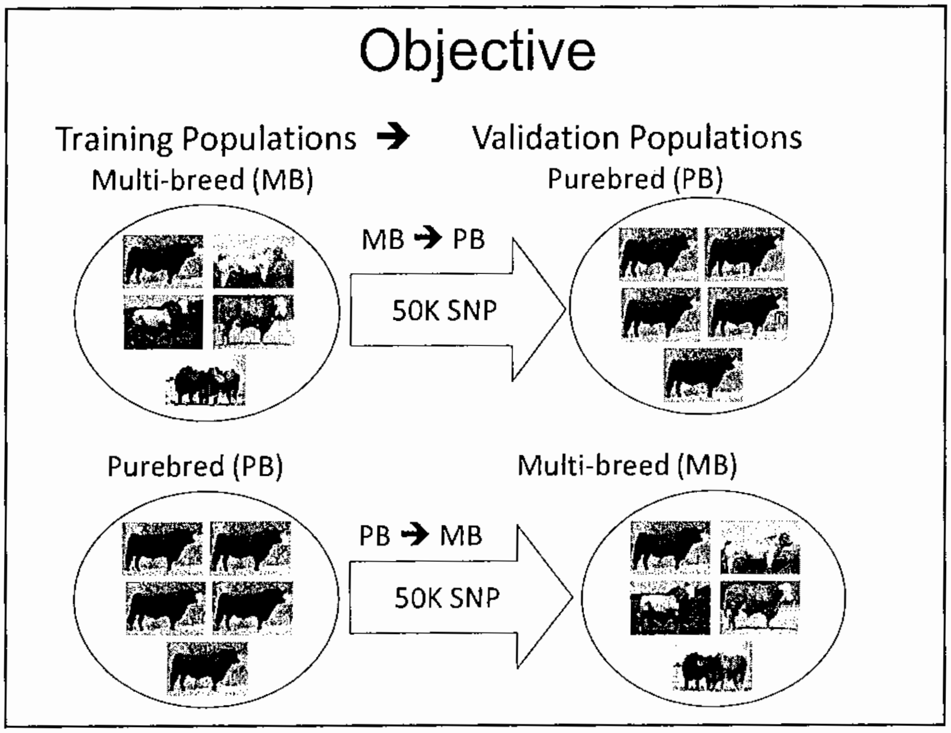
Across-breed prediction

- 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)

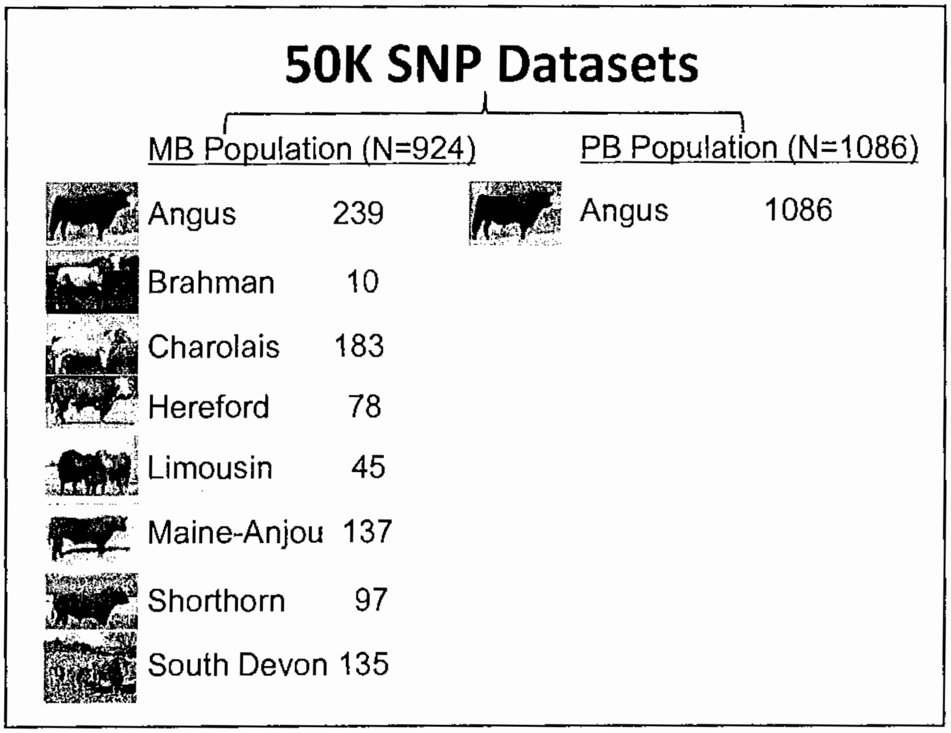
MB → PB

50K SNP

Purebred (PB) → Multi-breed (MB)

PB → MB

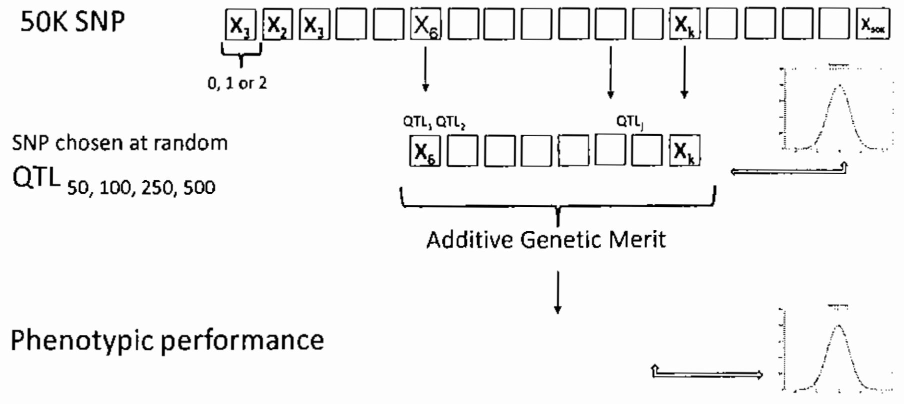
50K SNP



50K SNP Datasets

MB Population (N=924)		PB Population (N=1086)	
Angus	239	Angus	1086
Brahman	10		
Charolais	183		
Hereford	78		
Limousin	45		
Maine-Anjou	137		
Shorthorn	97		
South Devon	135		

Simulation of Additive Genetic Merit and Phenotypic Performance



50K SNP

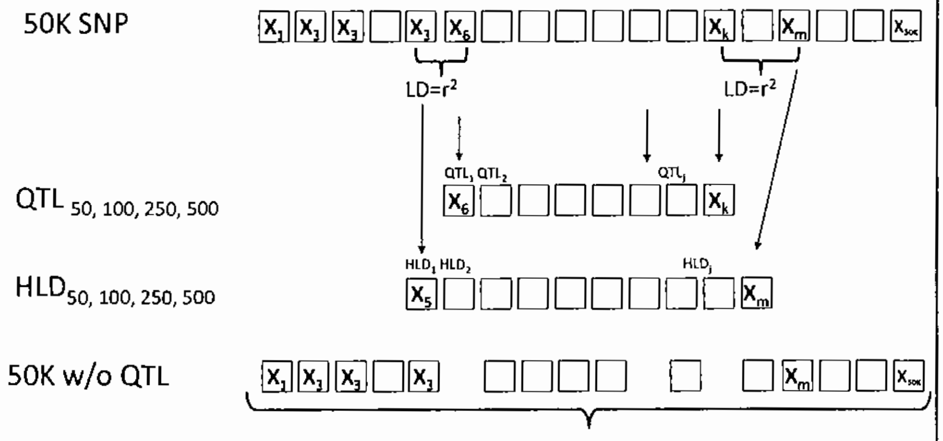
SNP chosen at random

QTL 50, 100, 250, 500

Additive Genetic Merit

Phenotypic performance

Marker Panels



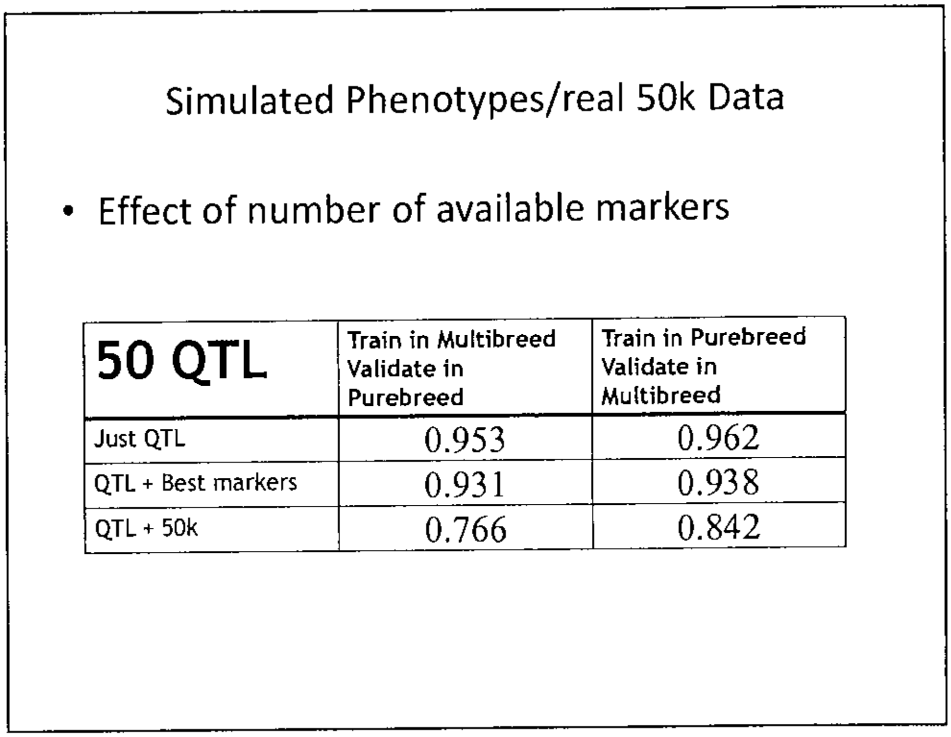
50K SNP

QTL 50, 100, 250, 500

HLD50, 100, 250, 500

50K w/o QTL

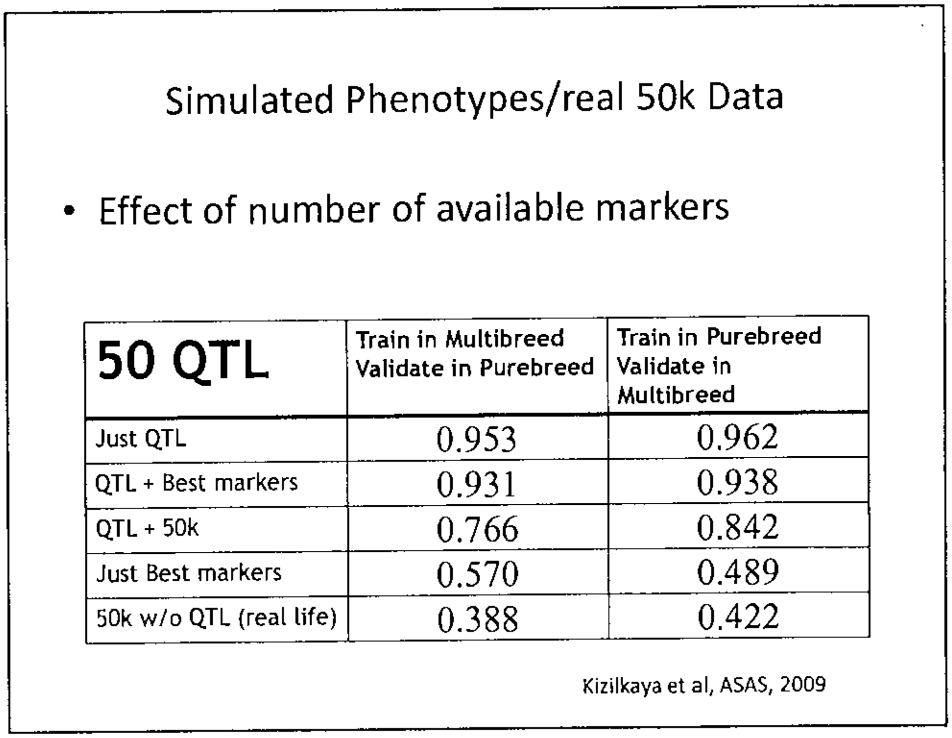
Bayesian Analysis



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

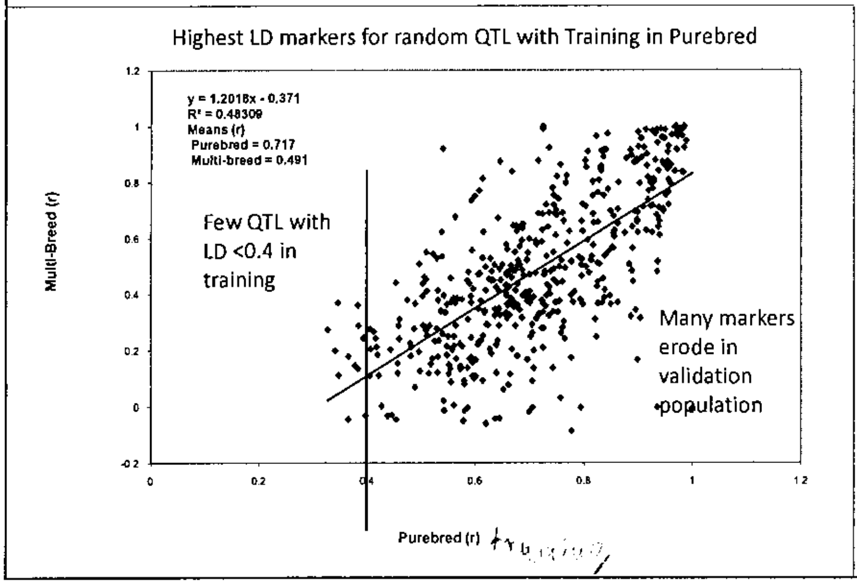
50 QTL	Train in Multibreed Validate in Purebreed	Train in Purebreed Validate in Multibreed
Just QTL	0.953	0.962
QTL + Best markers	0.931	0.938
QTL + 50k	0.766	0.842
Just Best markers	0.570	0.489
50k w/o QTL (real life)	0.388	0.422

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

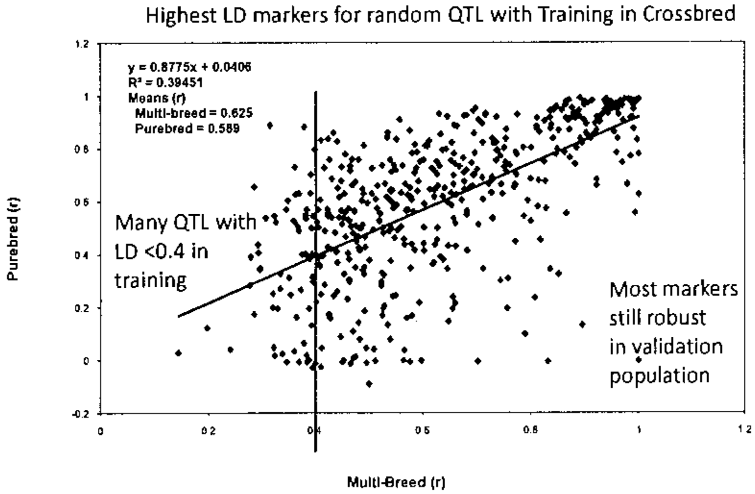


Highest LD markers for random QTL with Training in Purebred

Few QTL with LD <0.4 in training

Many markers erode in validation population

Purebred or Crossbred



Highest LD markers for random QTL with Training in 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

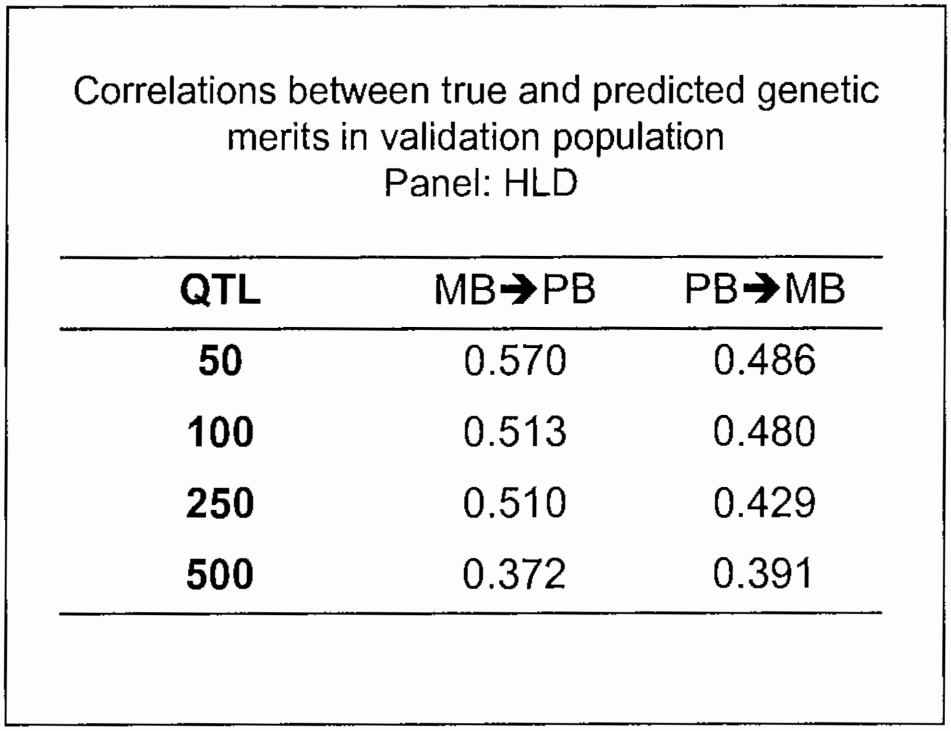
QTL	MB→PB	PB→MB
50	0.953	0.962
100	0.938	0.941
250	0.840	0.853
500	0.720	0.786

Simulated Phenotypes/real 50k Data

- Effect of number of QTL

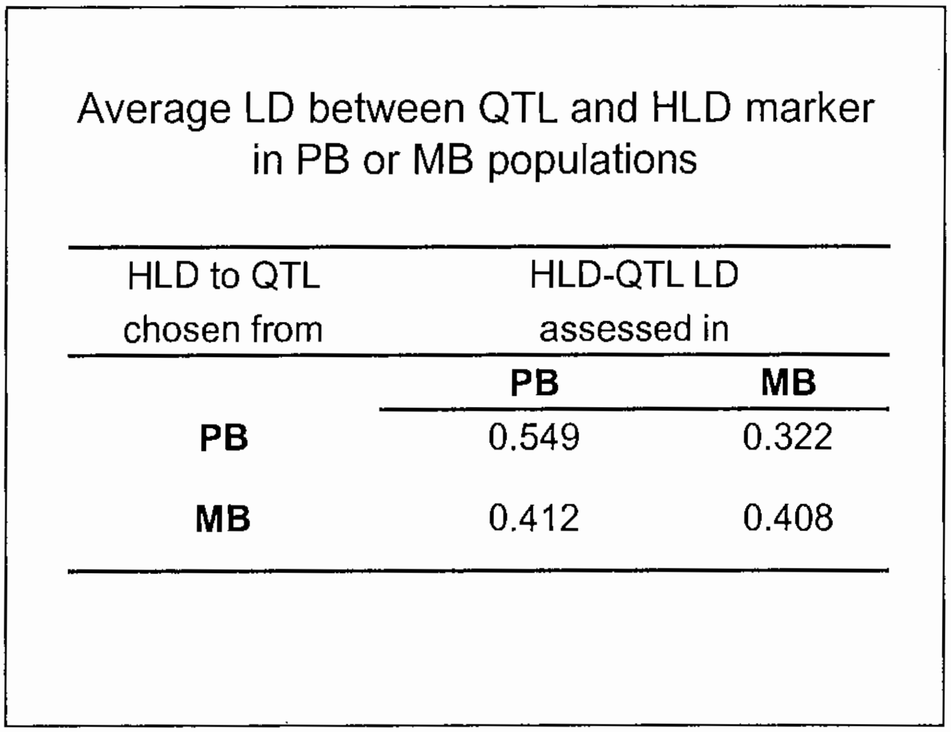
50k w/o QTL	Train in Multibreed Validate in Purebreed	Train in Purebreed Validate in Multibreed
50 QTL	0.388	0.422
100 QTL	0.289	0.308
250 QTL	0.247	0.276
500 QTL	0.200	0.299

- 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



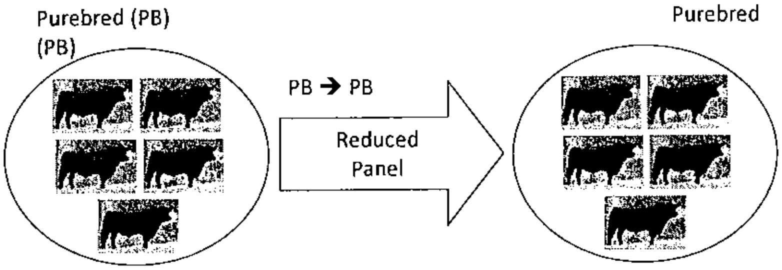
Average LD between QTL and HLD marker in PB or MB populations

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



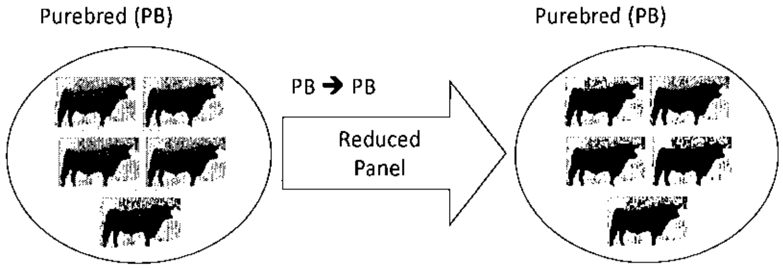
Purebred (PB)
(PB)

PB → PB

Reduced Panel

Purebred

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

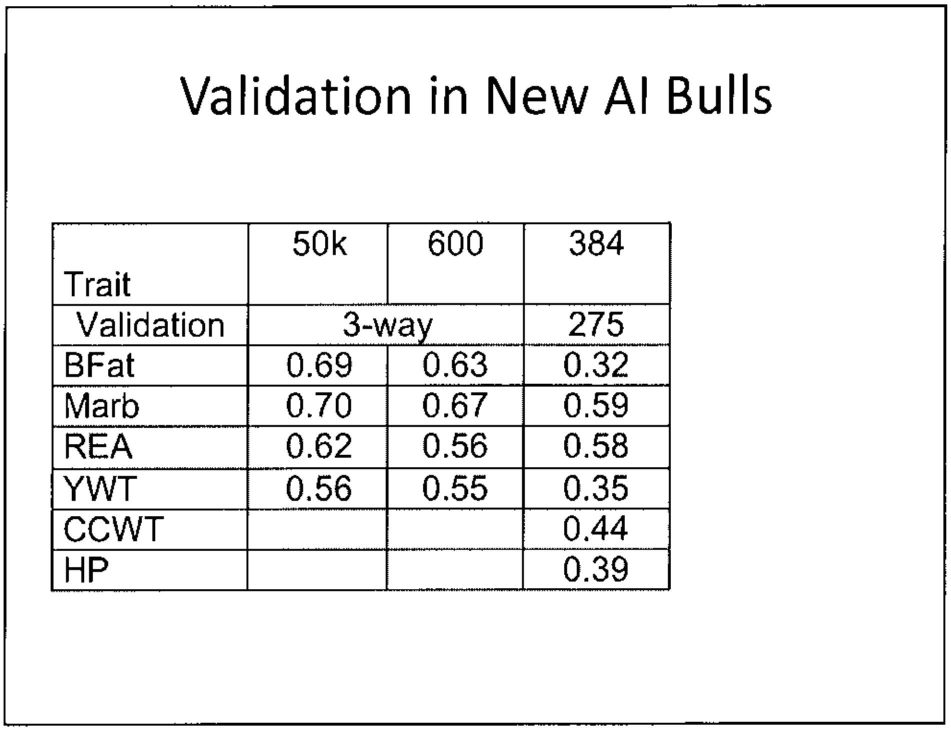
384 SNP Panels

- 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

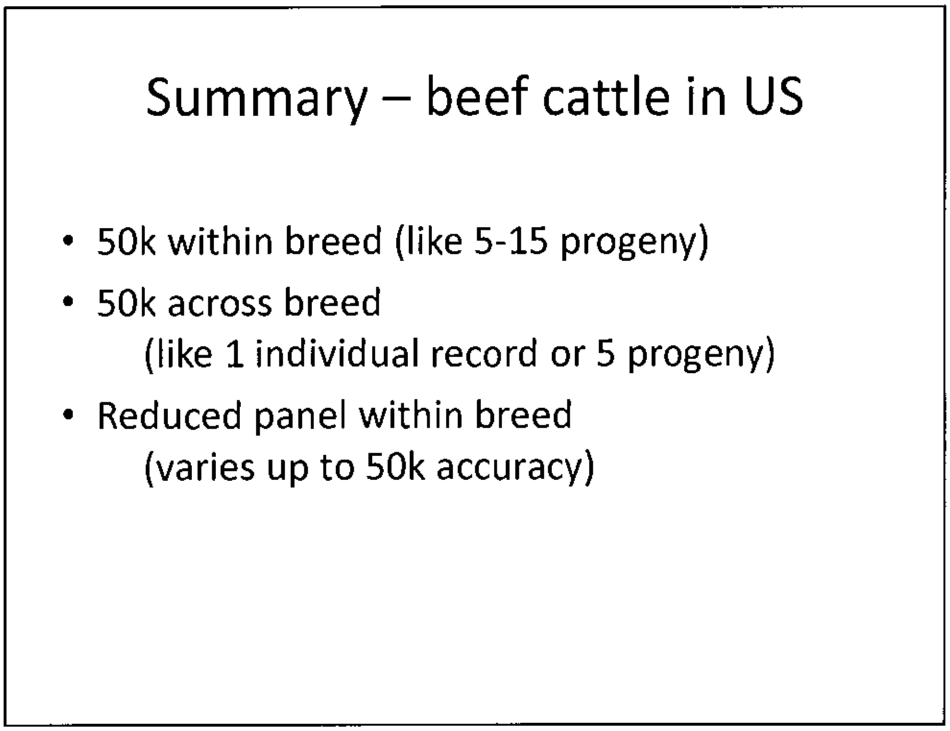
Validation in 698 steers with carcass phenotypes

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



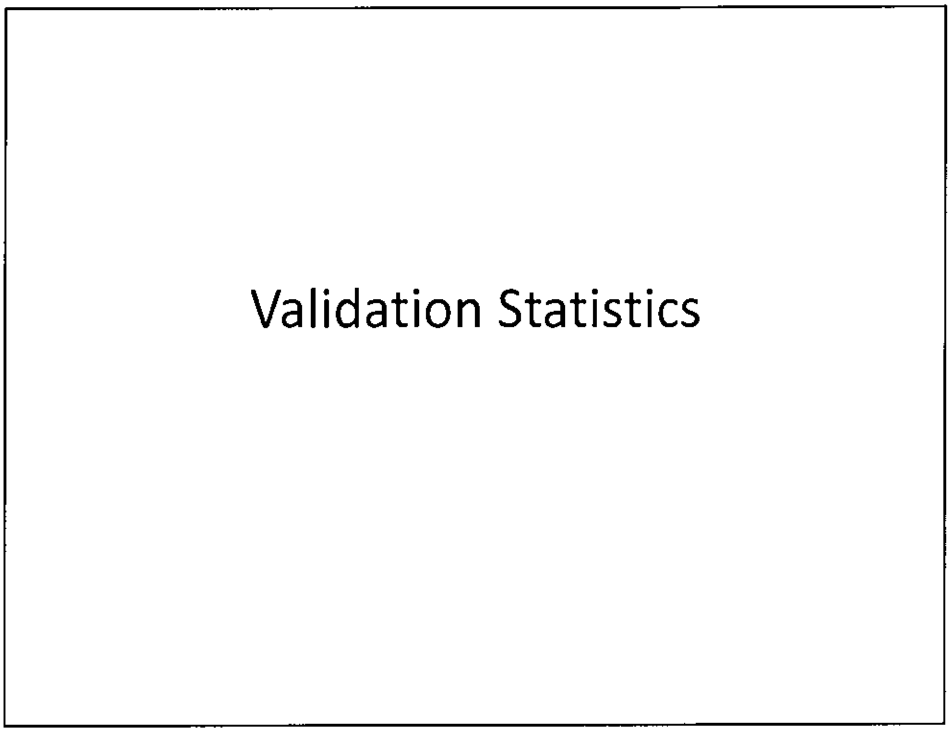
Validation in New AI Bulls

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

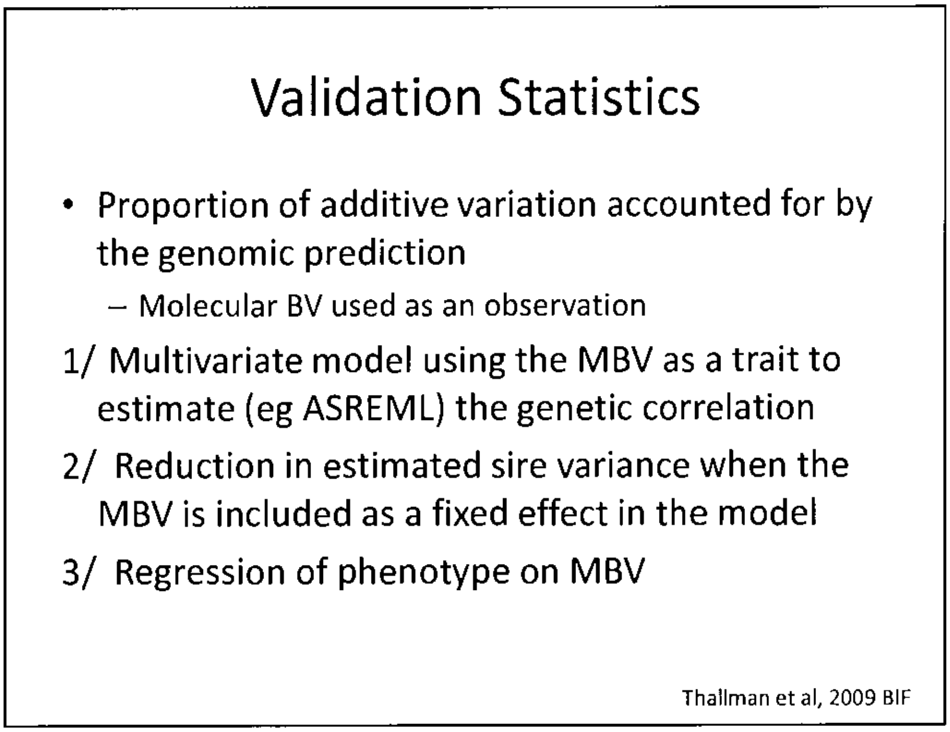


Summary – beef cattle in US

- 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

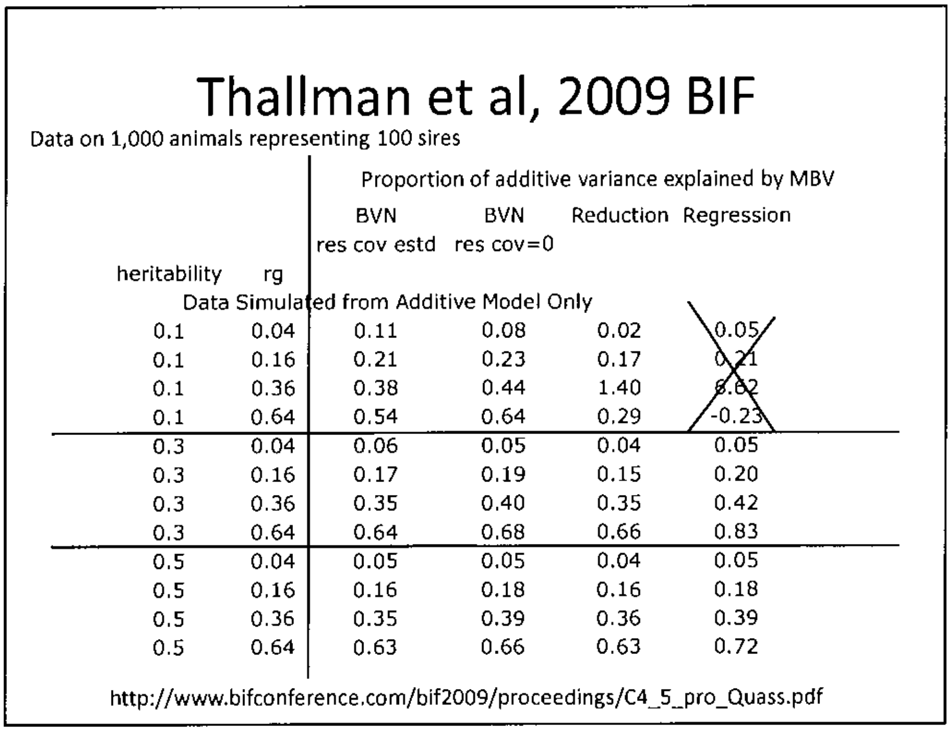
- Proportion of additive variation accounted for by the genomic prediction
  - Molecular BV used as an observation

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

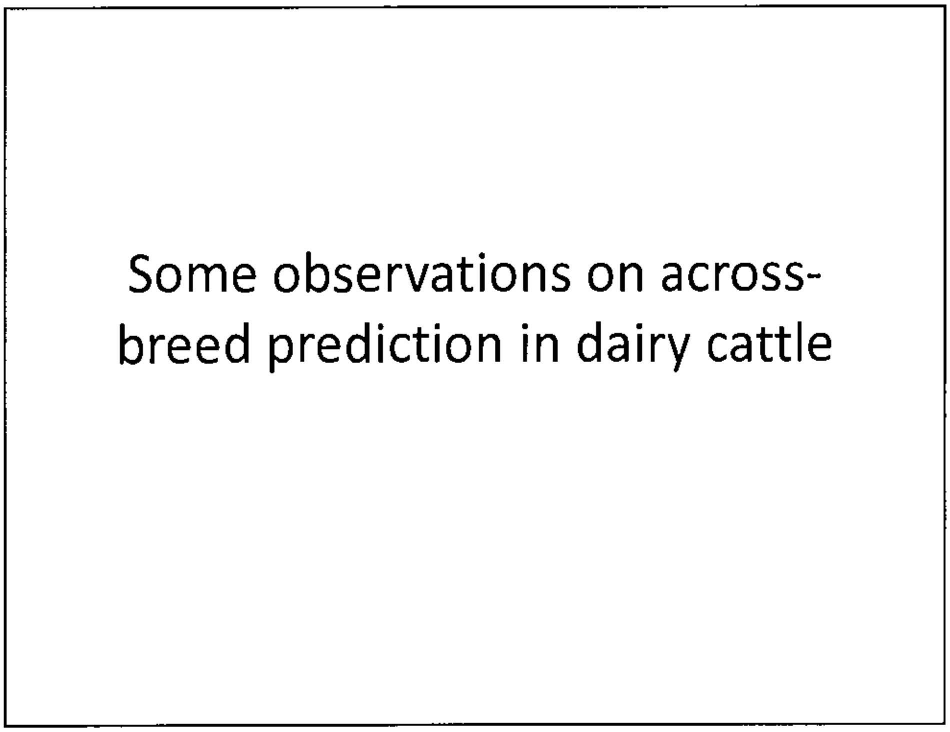


Thallman et al, 2009 BIF

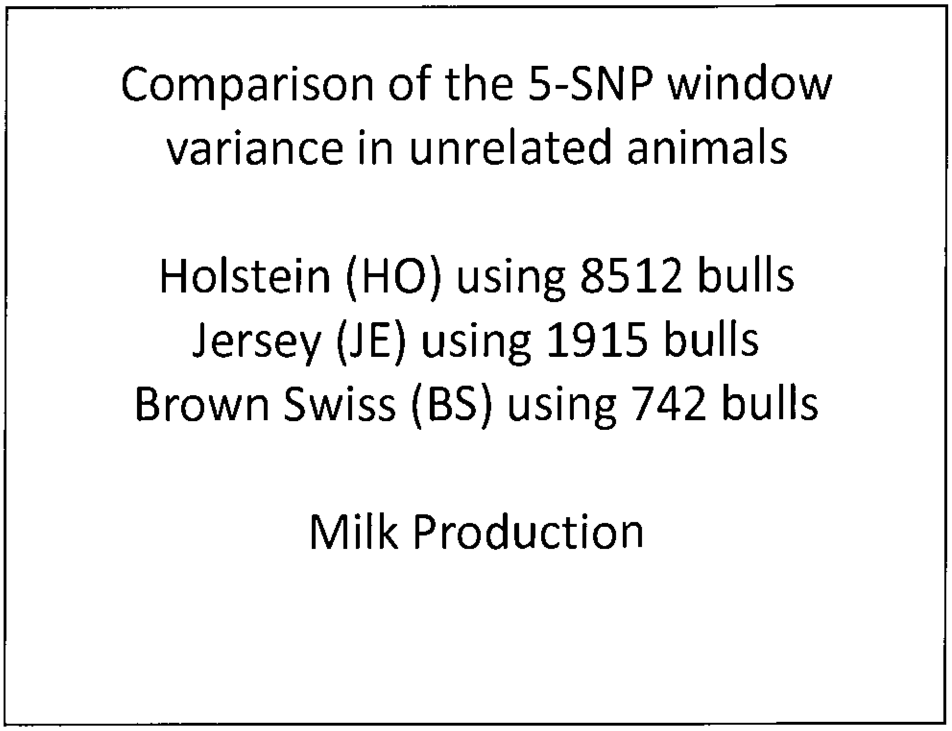
Data on 1,000 animals representing 100 sires

		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	~~0.62~~
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



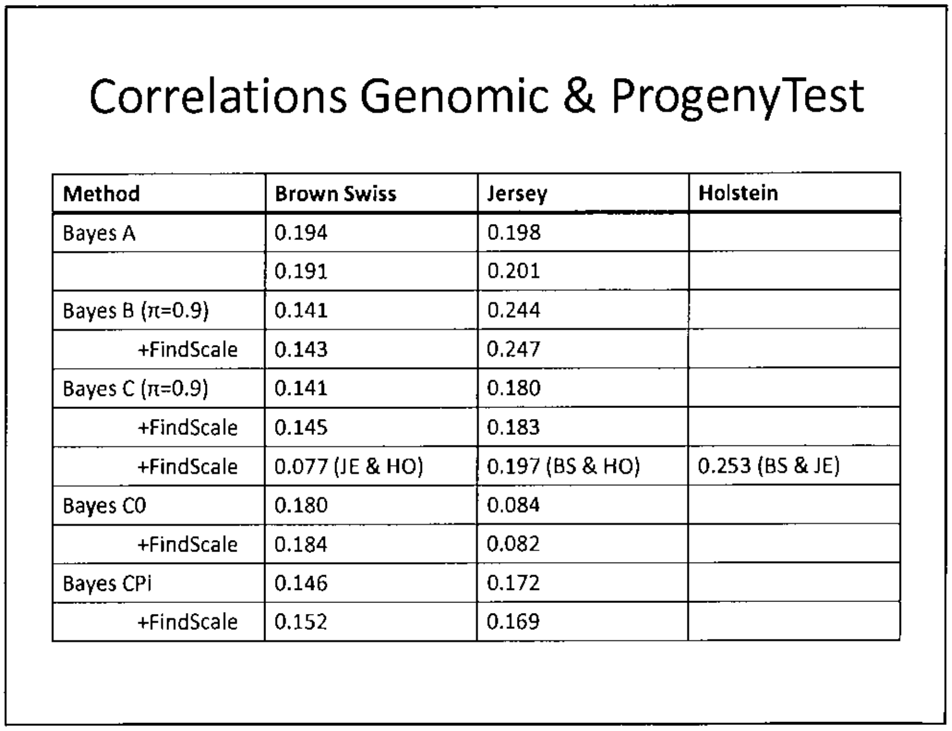
Some observations on across-breed prediction in dairy cattle



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

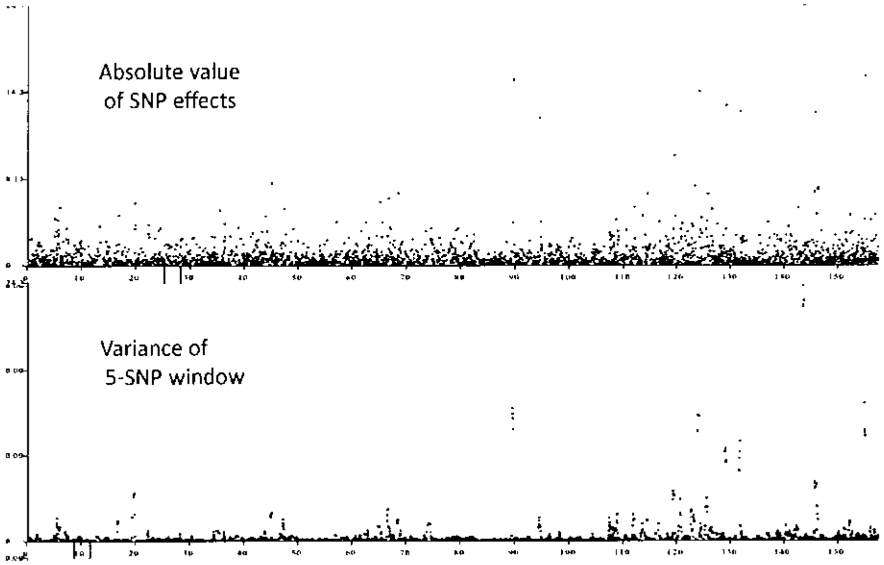
Milk Production



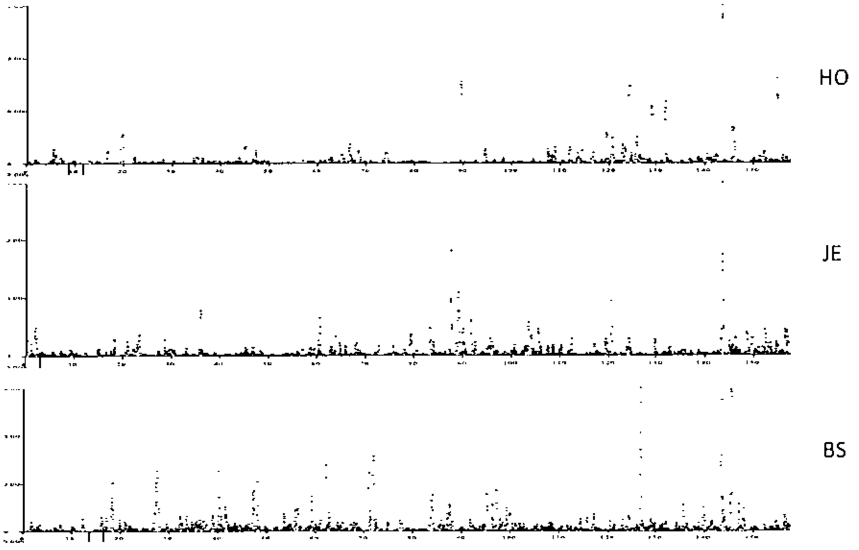
Correlations Genomic & ProgenyTest

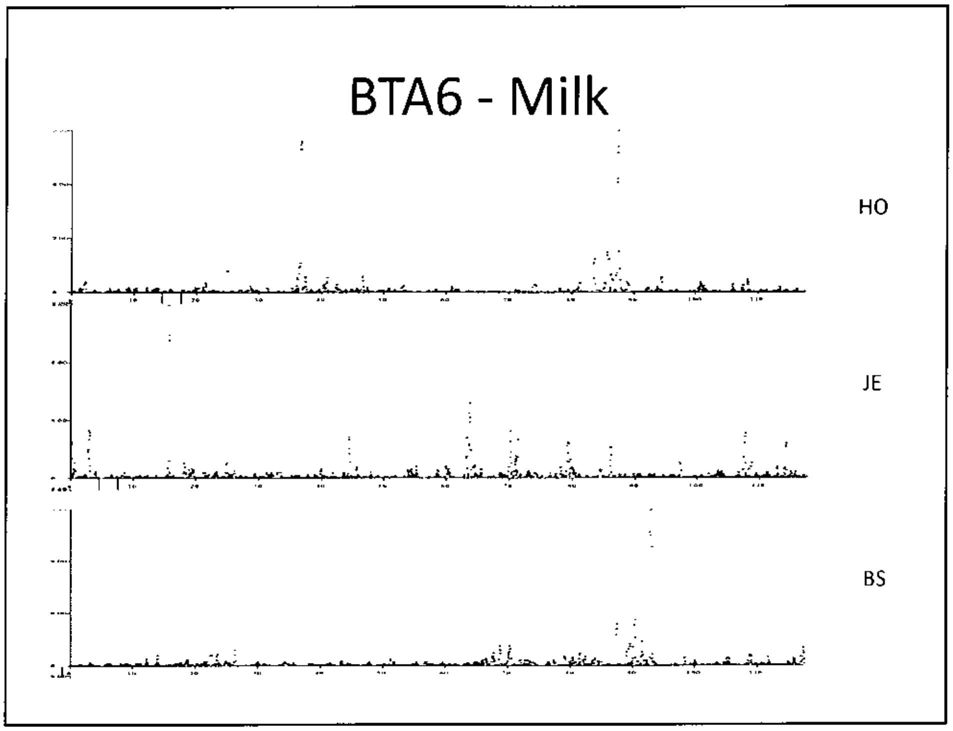
Method	Brown Swiss	Jersey	Holstein
Bayes A	0.194	0.198	
	0.191	0.201	
Bayes B (π=0.9)	0.141	0.244	
+FindScale	0.143	0.247	
Bayes C (π=0.9)	0.141	0.180	
+FindScale	0.145	0.183	
+FindScale	0.077 (JE & HO)	0.197 (BS & HO)	0.253 (BS & JE)
Bayes C0	0.180	0.084	
+FindScale	0.184	0.082	
Bayes CPi	0.146	0.172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

Holstein BTA1 Milk

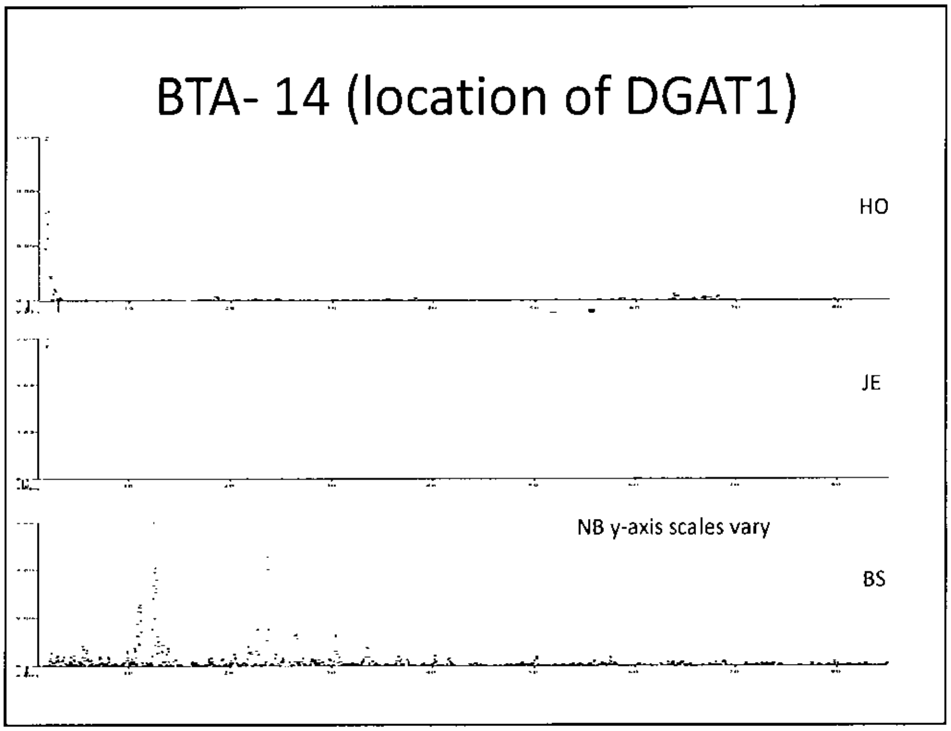


BTA1 - Milk



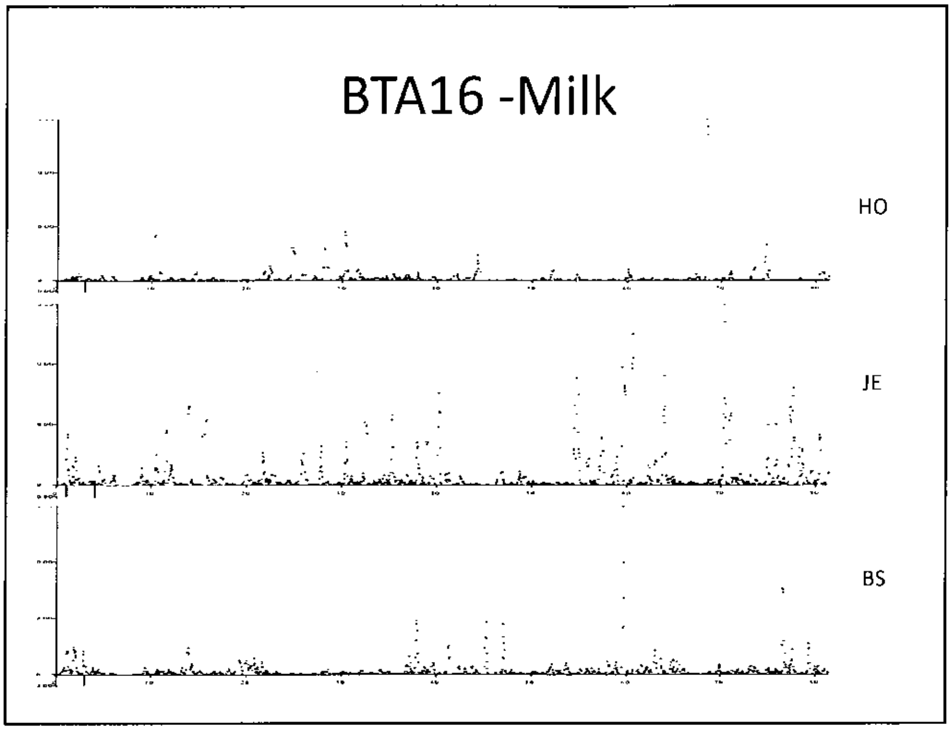


BTA6 - Milk

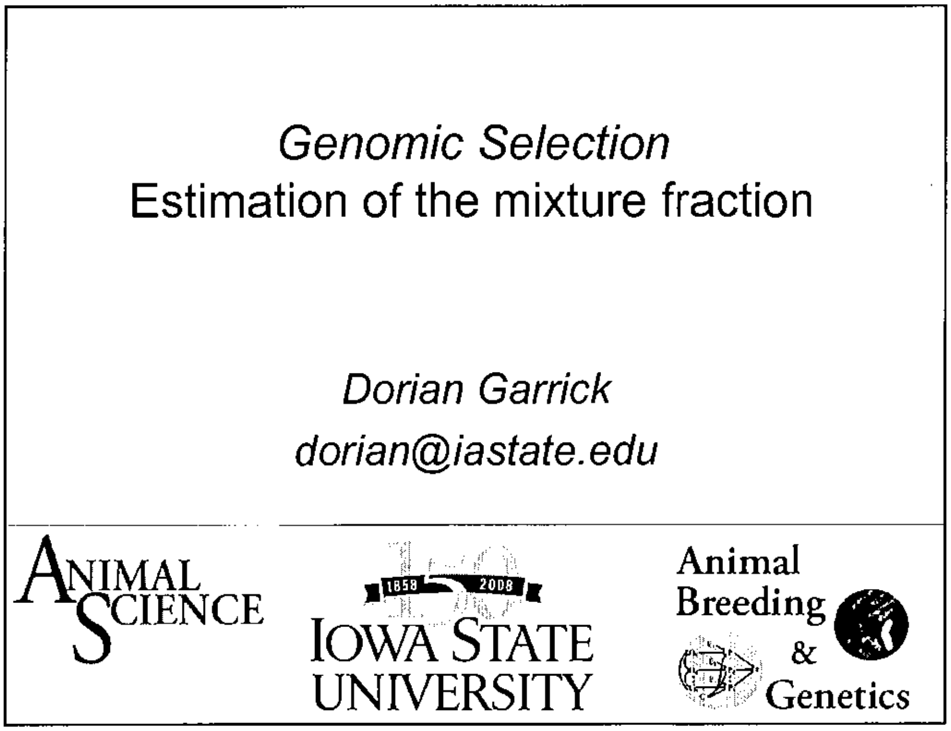


BTA- 14 (location of DGAT1)

NB y-axis scales vary



BTA16 -Milk



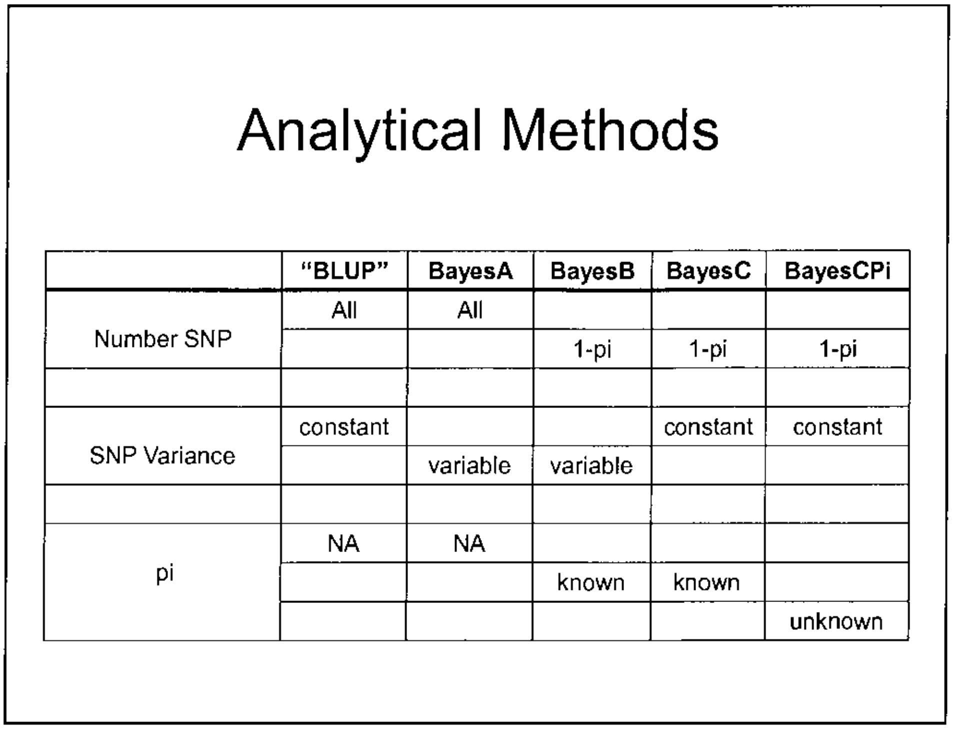
Genomic Selection
Estimation of the mixture fraction

Dorian Garrick
dorian@iastate.edu

ANIMAL SCIENCE

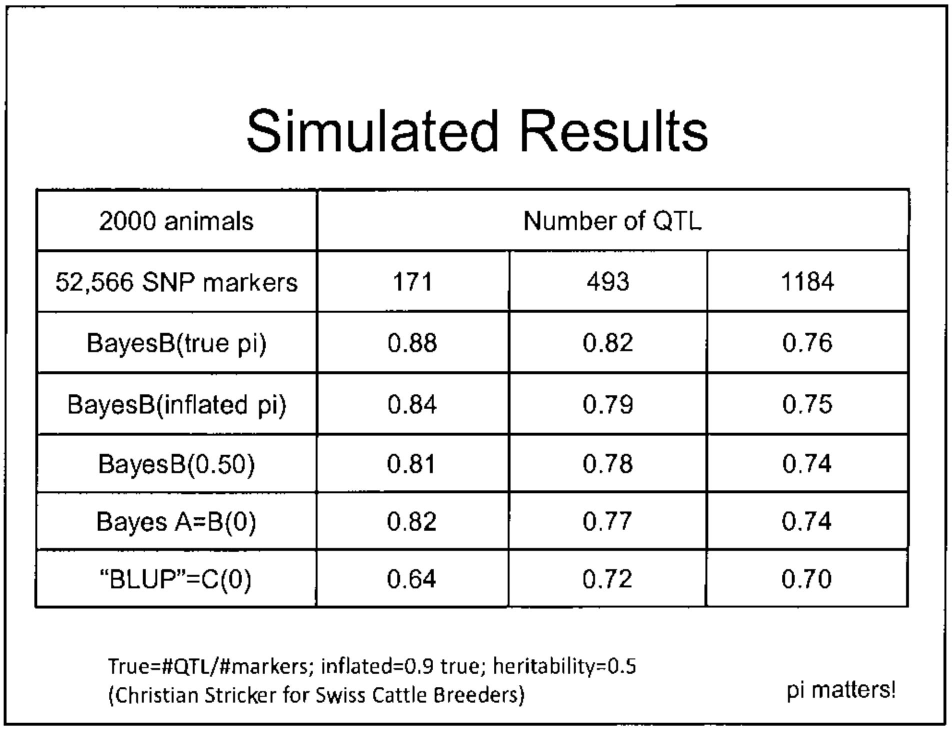
1858 2008
IOWA STATE UNIVERSITY

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

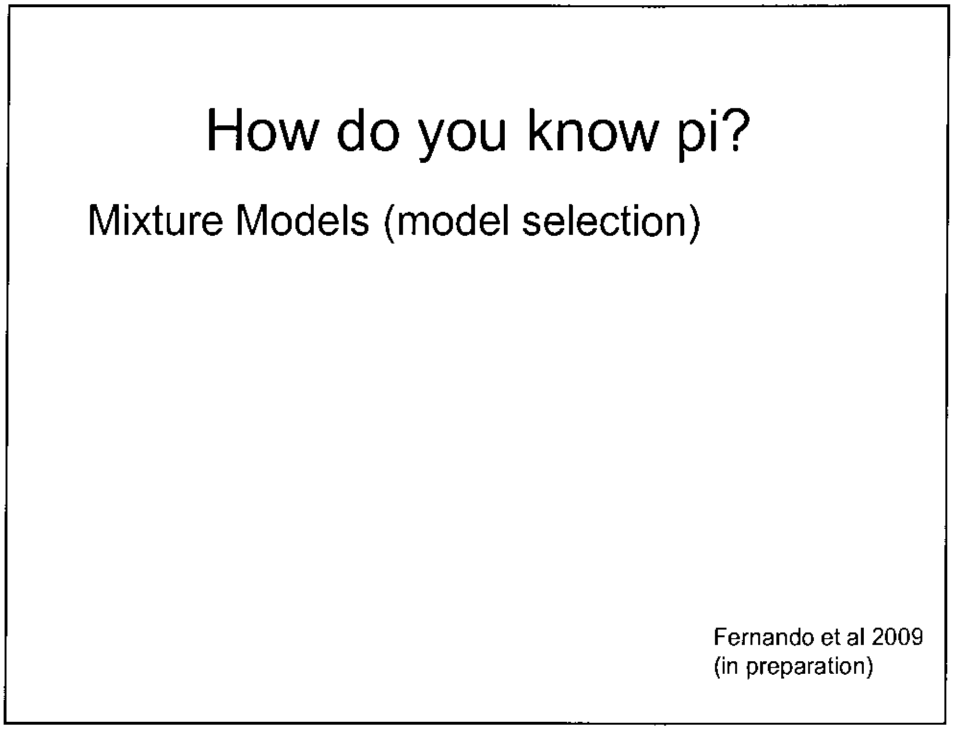


Simulated Results

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)

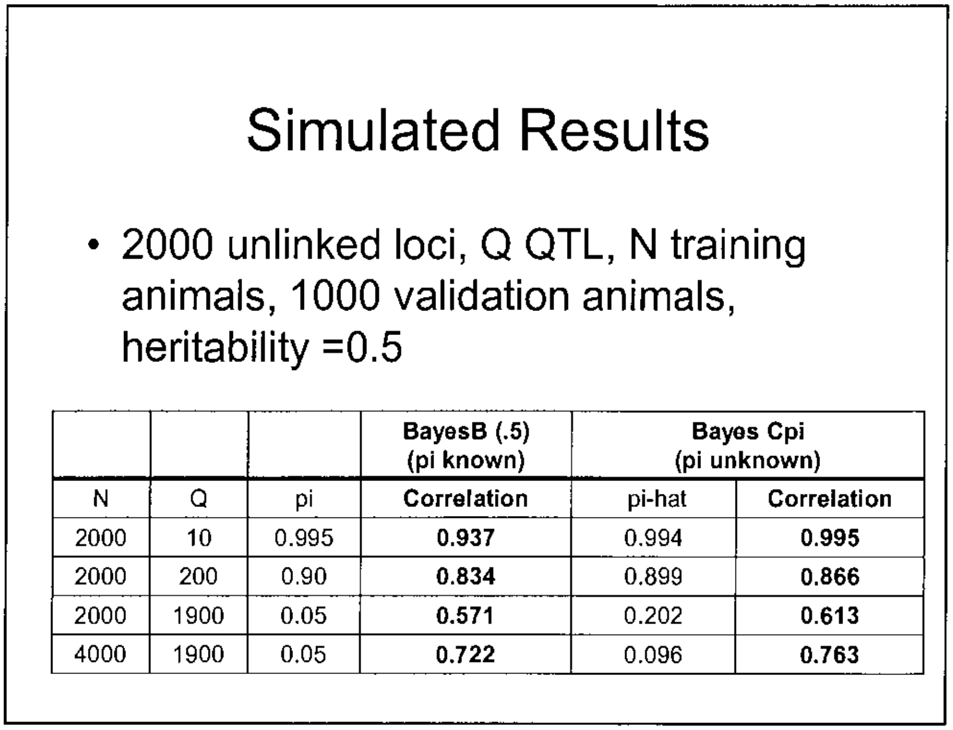
pi matters!



How do you know pi?

Mixture Models (model selection)

Fernando et al 2009
(in preparation)



Simulated Results

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

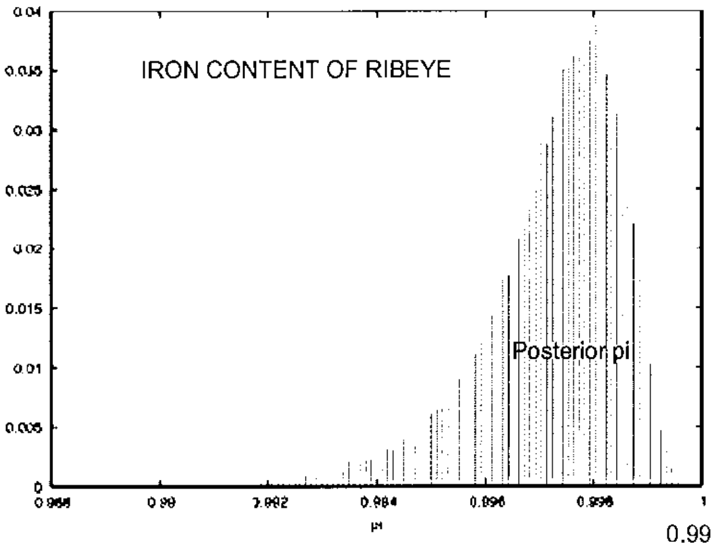
			BayesB (.5) (pi known)	**Bayes Cpi (pi unknown)**	
N	Q	pi	**Correlation**	pi-hat	**Correlation**
2000	10	0.995	**0.937**	0.994	**0.995**
2000	200	0.90	**0.834**	0.899	**0.866**
2000	1900	0.05	**0.571**	0.202	**0.613**
4000	1900	0.05	**0.722**	0.096	**0.763**

Simulated Results - Real 50k

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

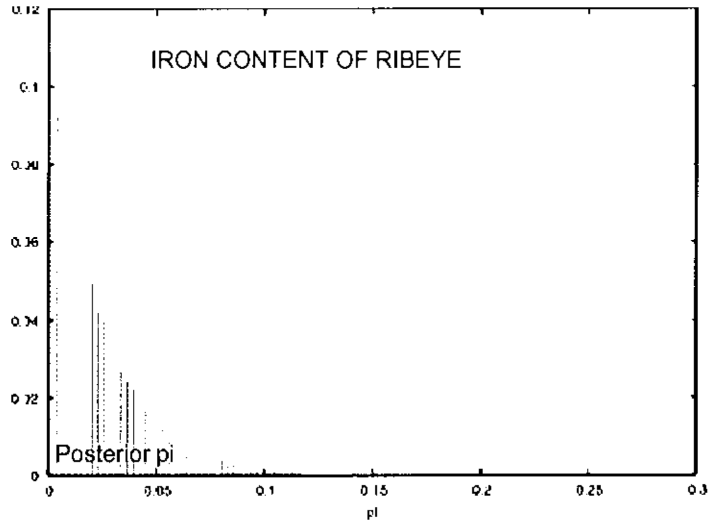
	Correlation True and Predicted Merit		
Assumed pi	Bayes B (pi known)	Bayes C (pi known)	Bayes Cpi (pi unknown)
0.999	0.86	0.86	
0.25	0.70	0.26	
N/A			0.86

50,000 markers (bovine)

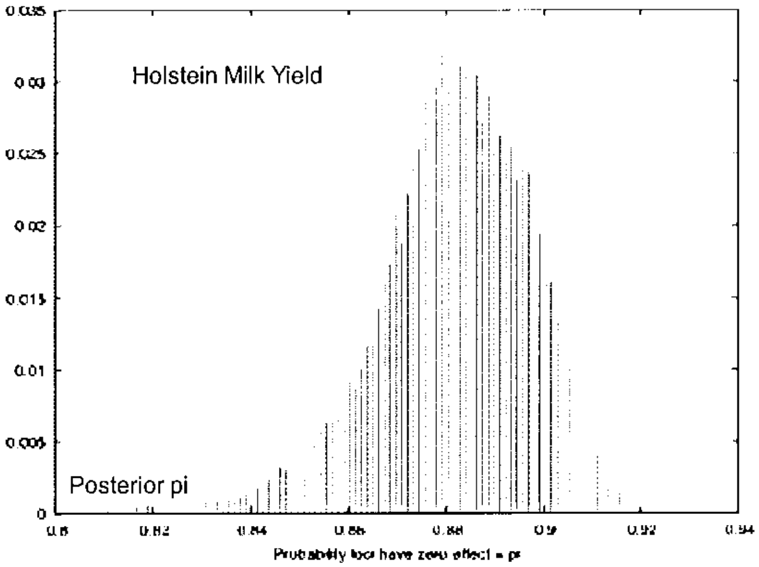


0.998=100 loci

“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

1858 2008
IOWA STATE UNIVERSITY

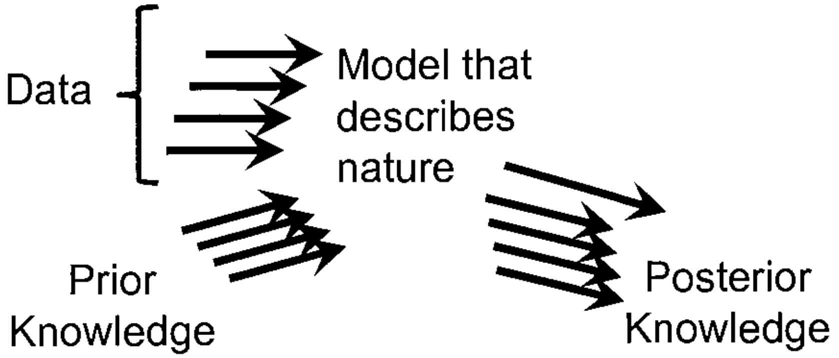
Animal Breeding & Genetics

Bayes A

Meuwissen, Hayes & Goddard (2001)

Sorensen & Gianola, 2002

BayesA/B not Bayesian Methods



Data

Model that describes nature

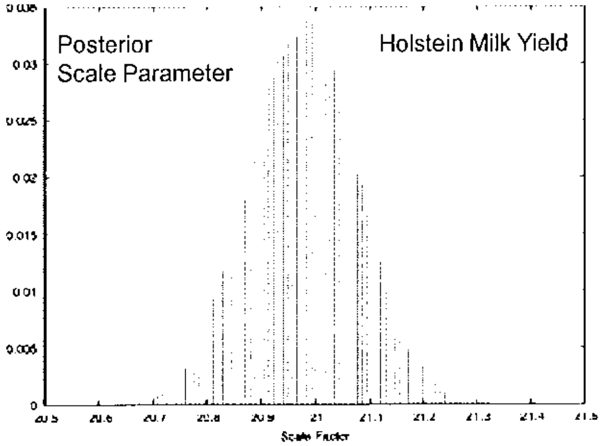
Prior Knowledge

Posterior Knowledge

Gianola et al "Bayesian Alphabet" 2009

But they work very well in practice!

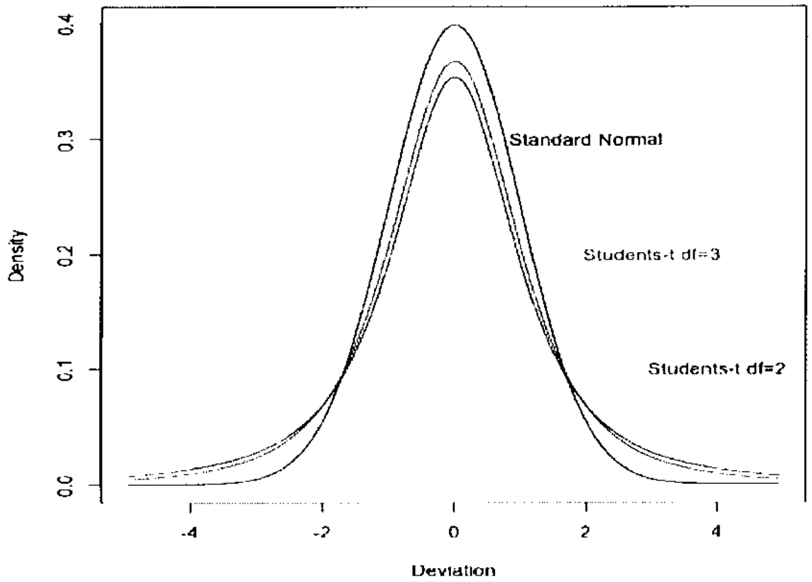
Bayes A on 8,300 bulls



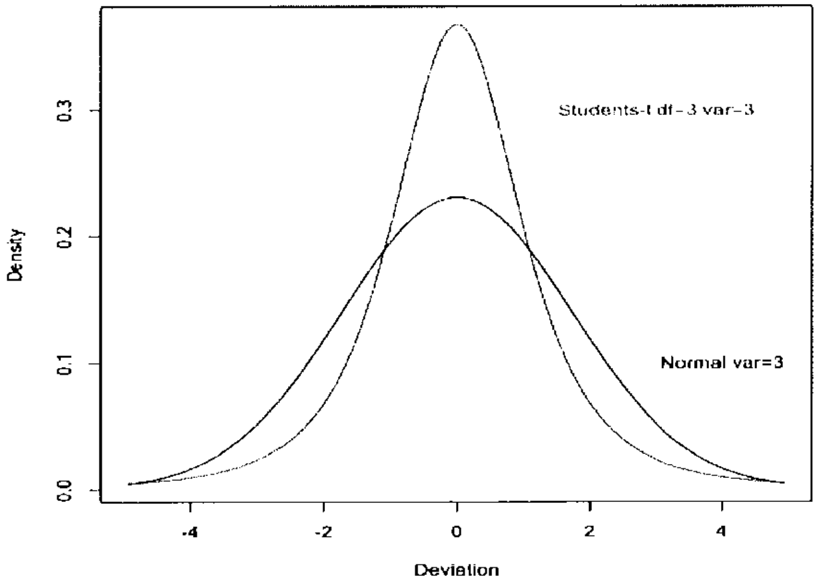
$$S_{v_a}^2 = \frac{\left(v_a - 2\right)V_a}{v_a k 2\overline{p}(1-\overline{p})} = \frac{(4-2)\times 646100}{4\times 43043\times 0.36} = 20.85$$

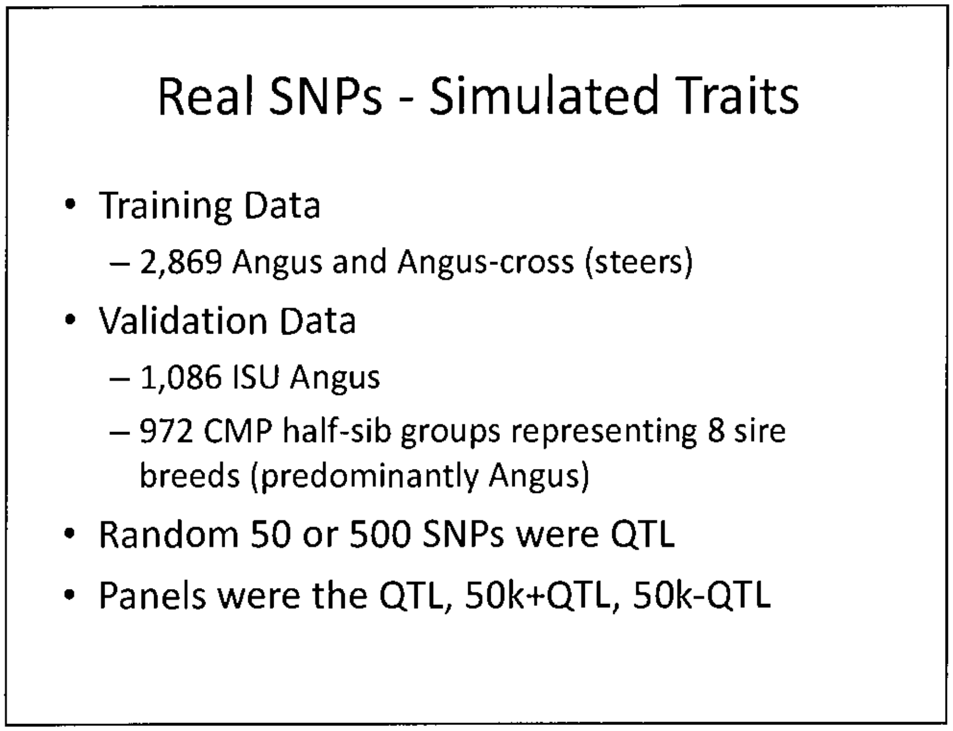
Alternative Distributions (to the normal)

Students-*t* Distributions



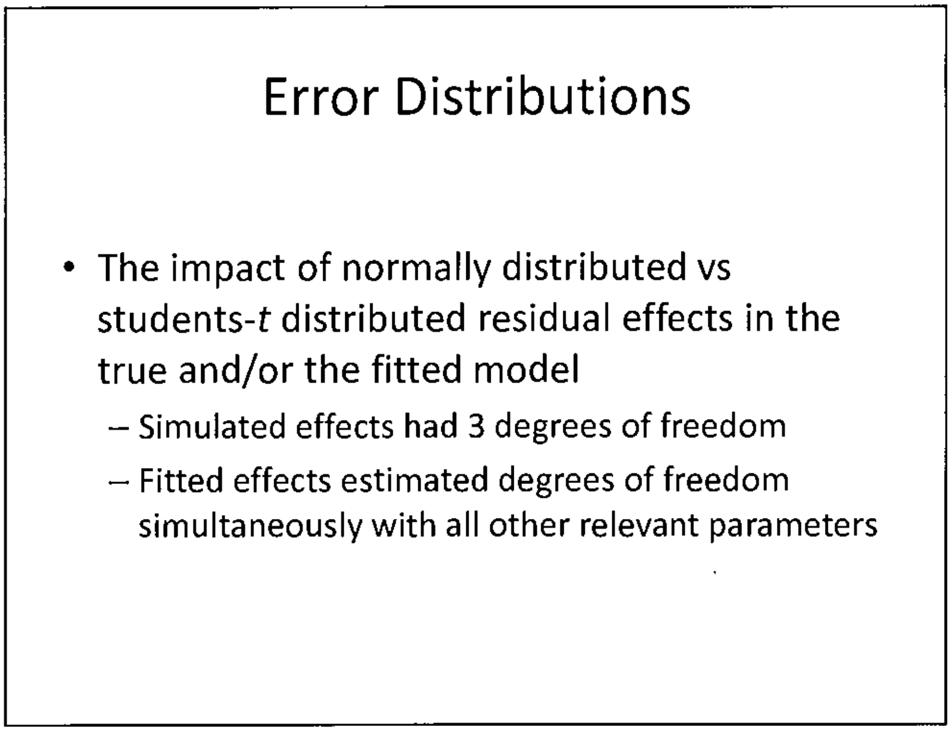
At Constant Variance





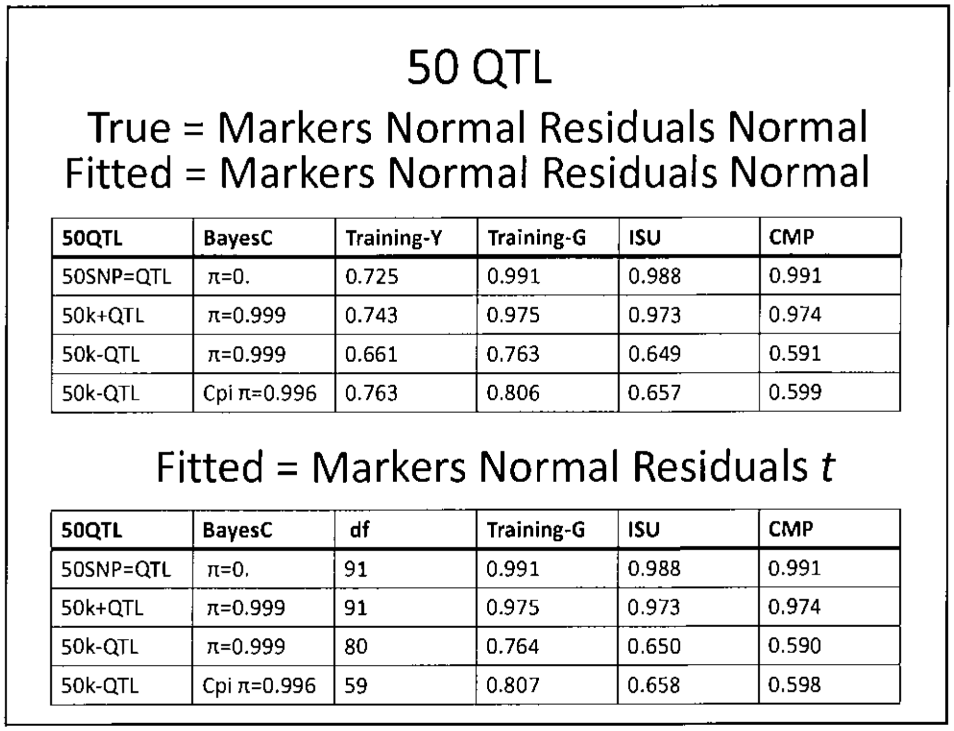
Real SNPs - Simulated Traits

- 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



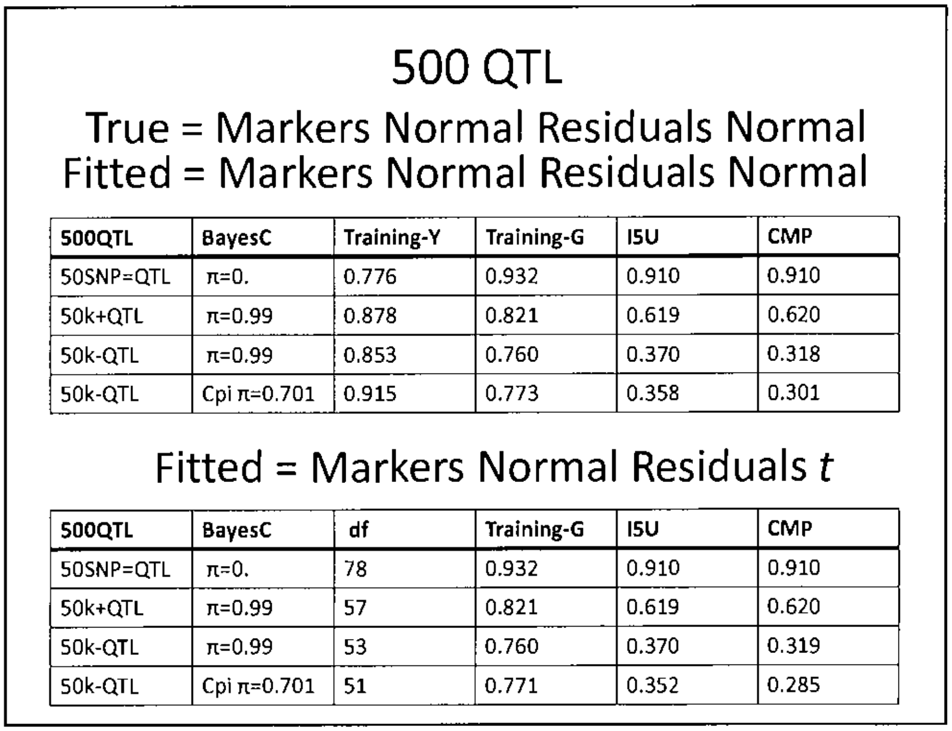
50 QTL

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



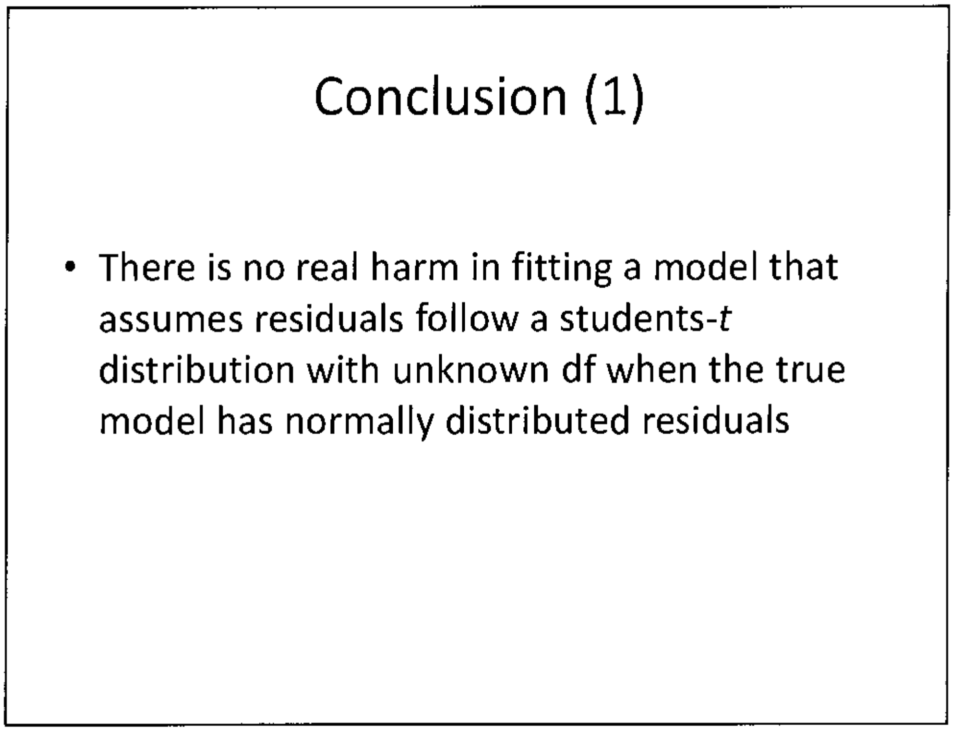
500 QTL

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

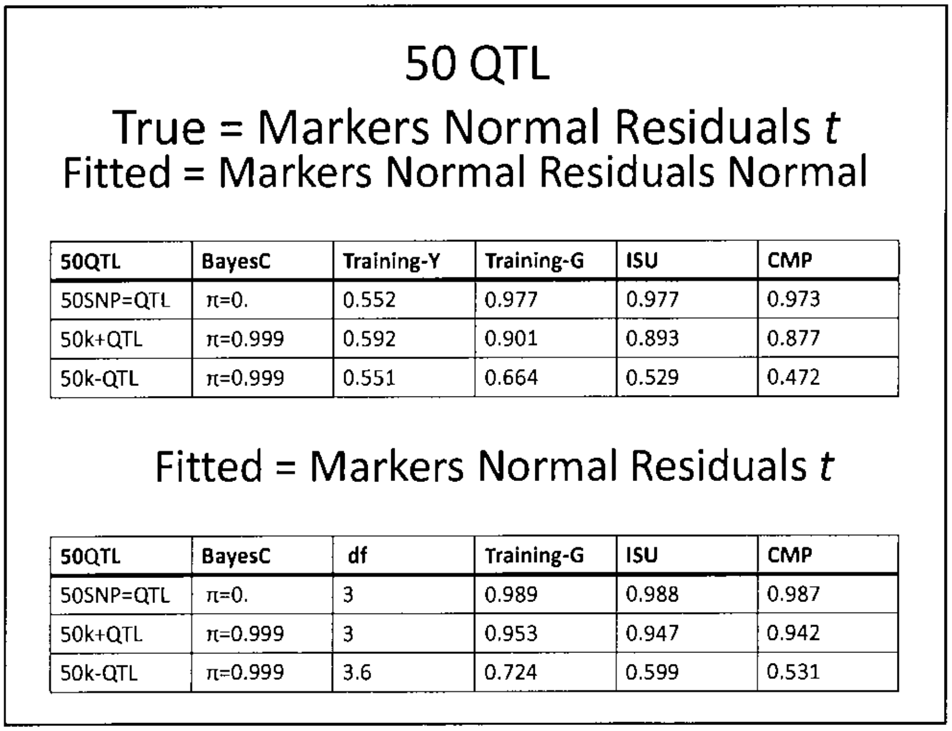
Fitted = Markers Normal Residuals *t*

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
505NP=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

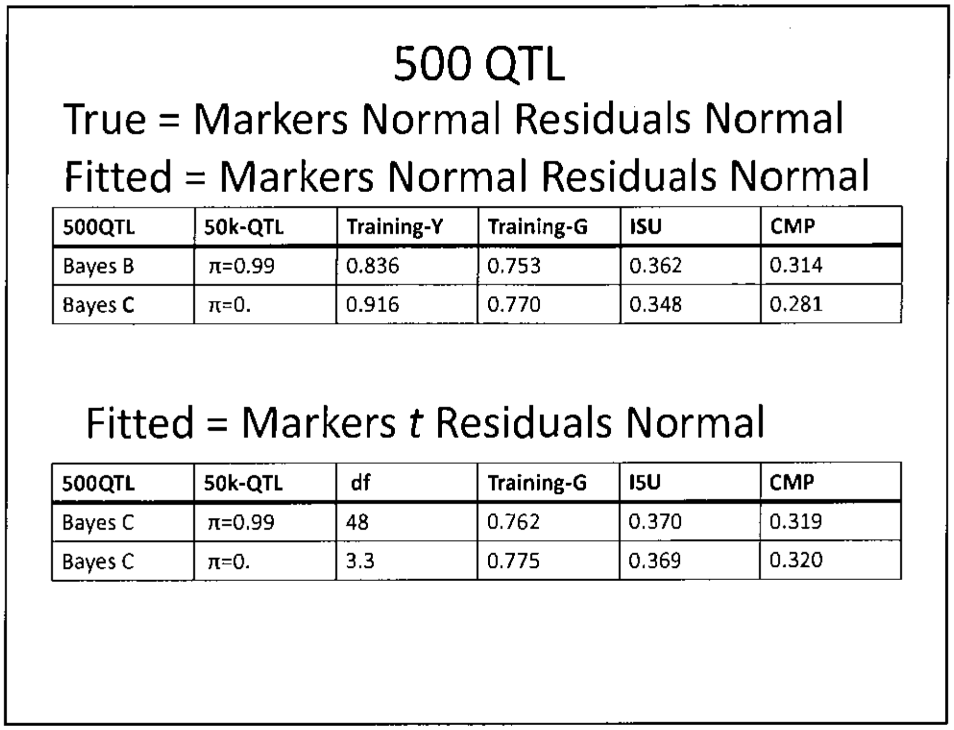
True = Markers Normal Residuals Normal

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

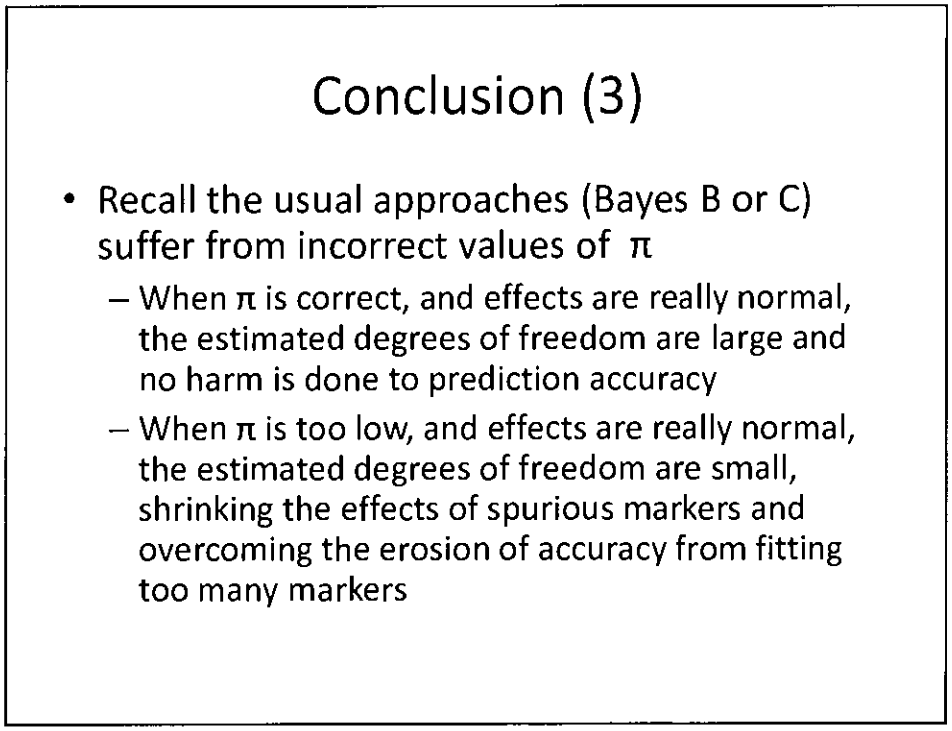
True = Markers Normal Residuals Normal

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

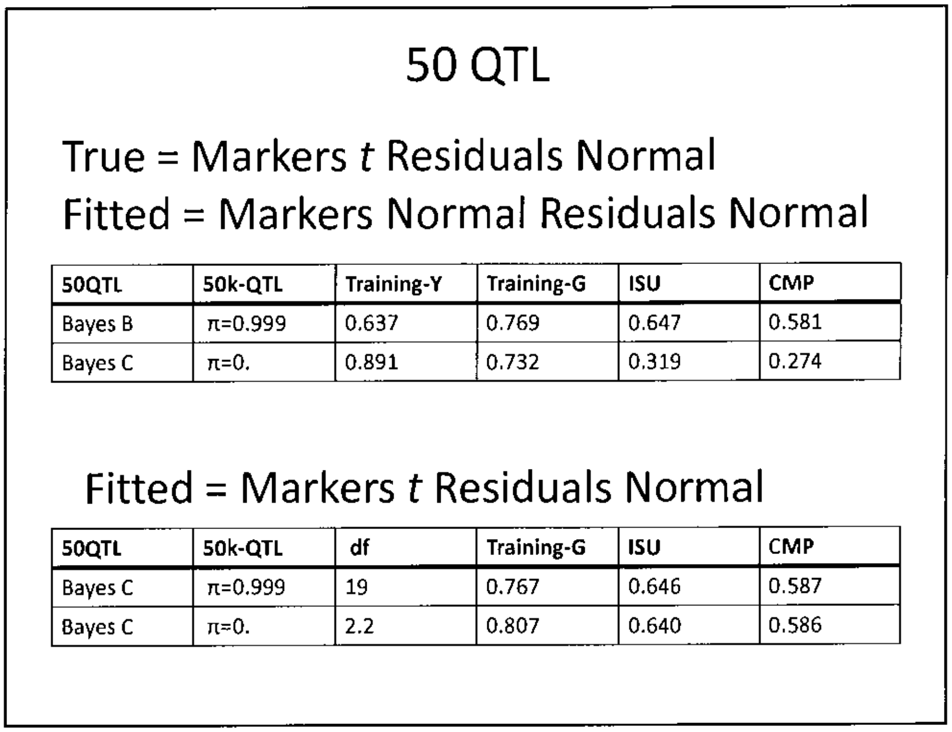
Fitted = Markers *t* Residuals Normal

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



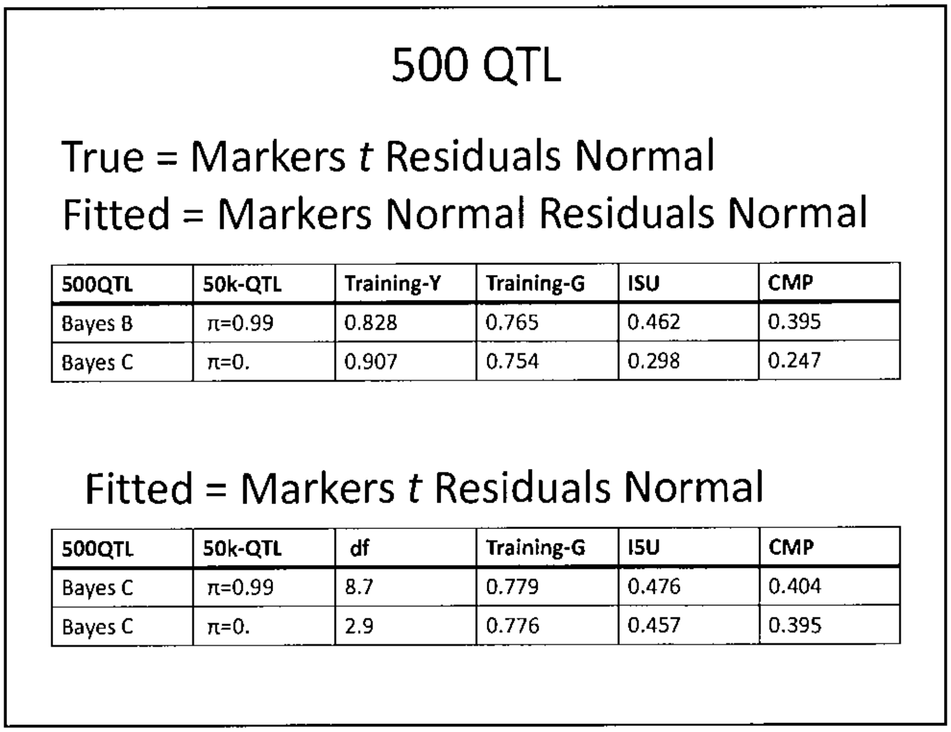
50 QTL

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

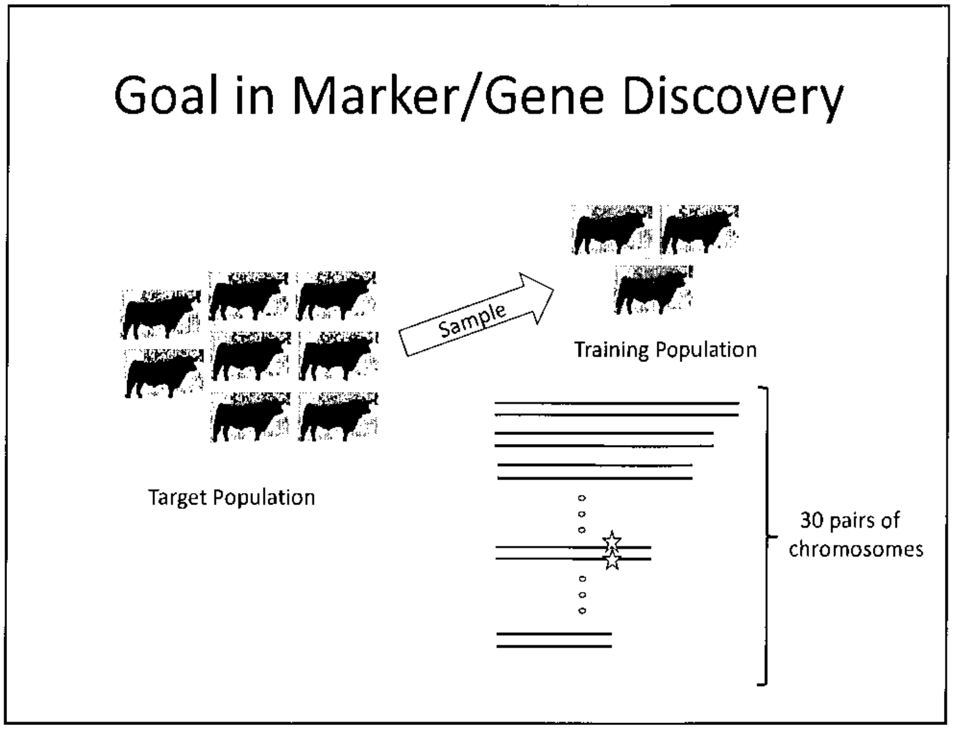
Fitted = Markers *t* Residuals Normal

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



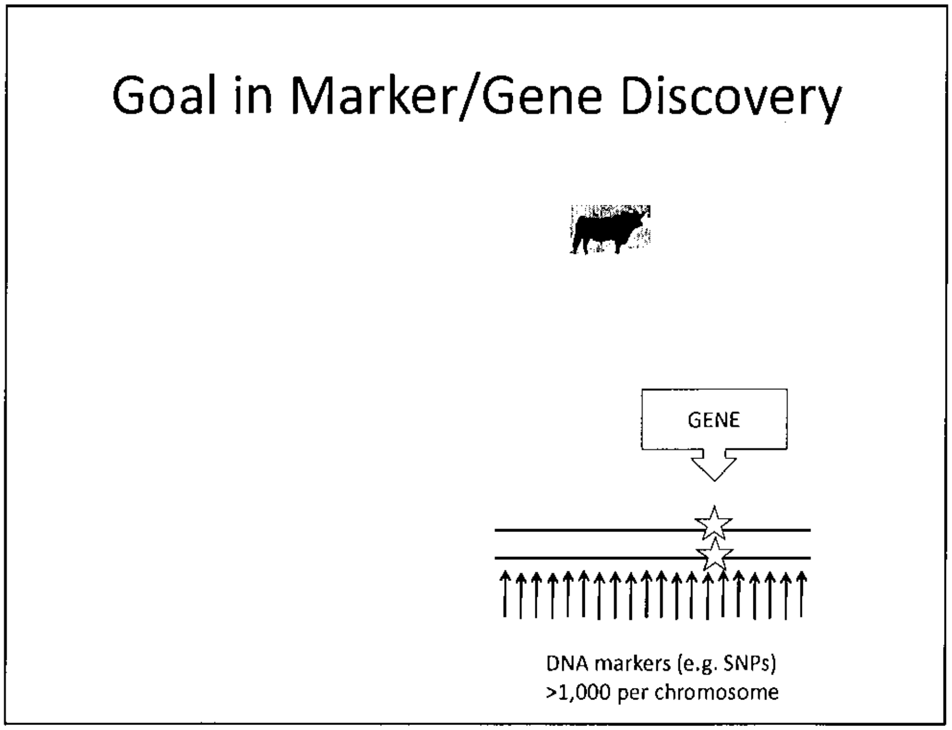
Goal in Marker/Gene Discovery

Target Population

Sample

Training Population

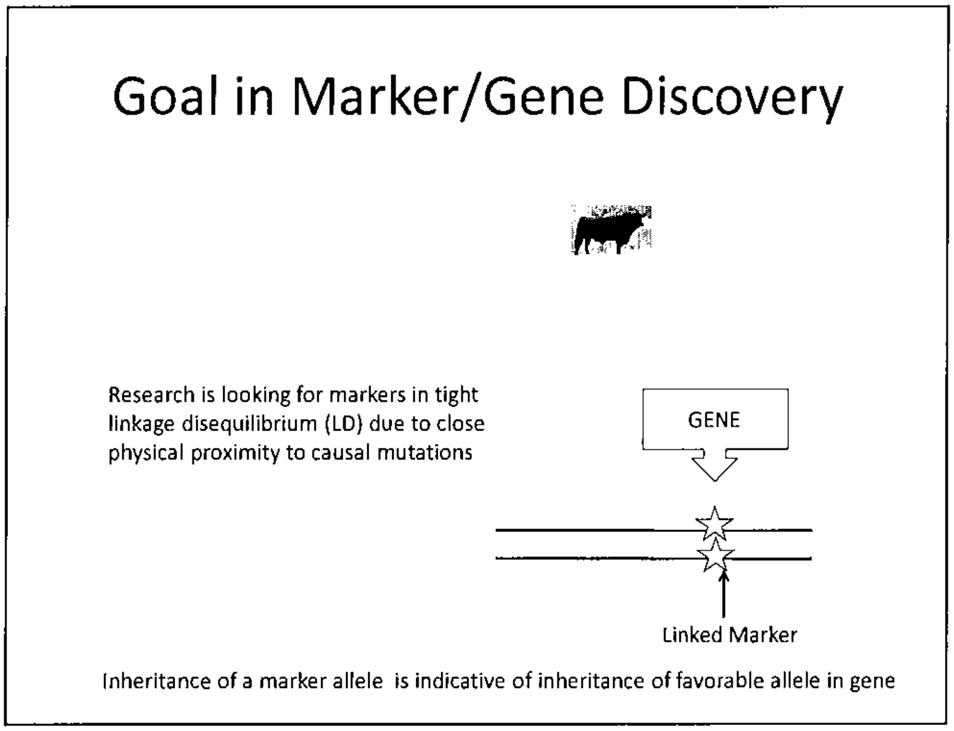
30 pairs of chromosomes



Goal in Marker/Gene Discovery

GENE

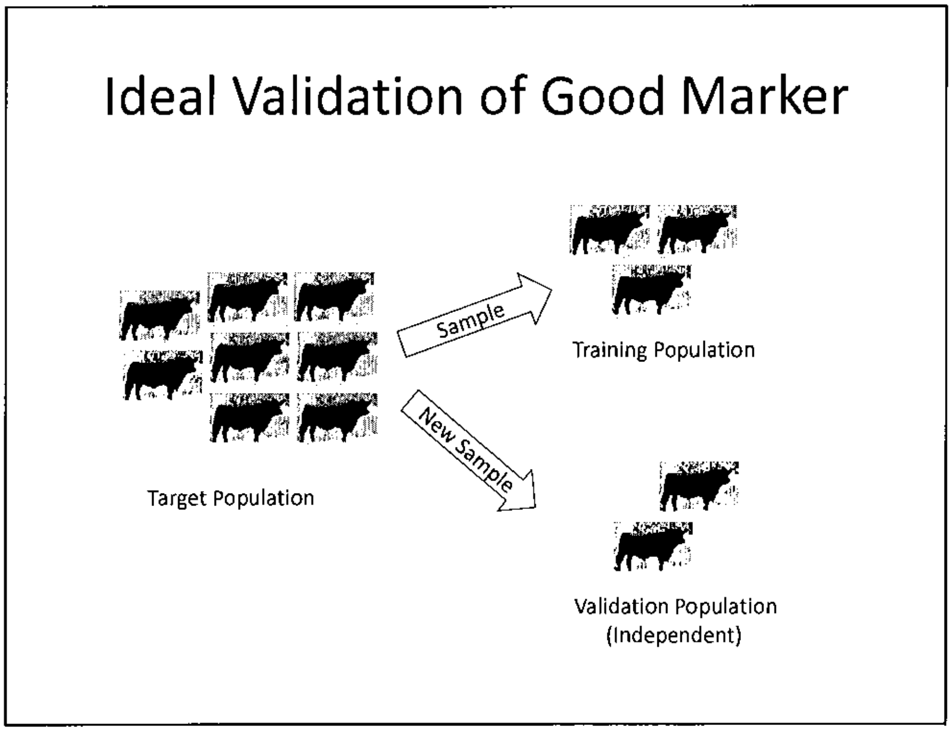
DNA markers (e.g. SNPs)
>1,000 per chromosome



Goal in Marker/Gene Discovery

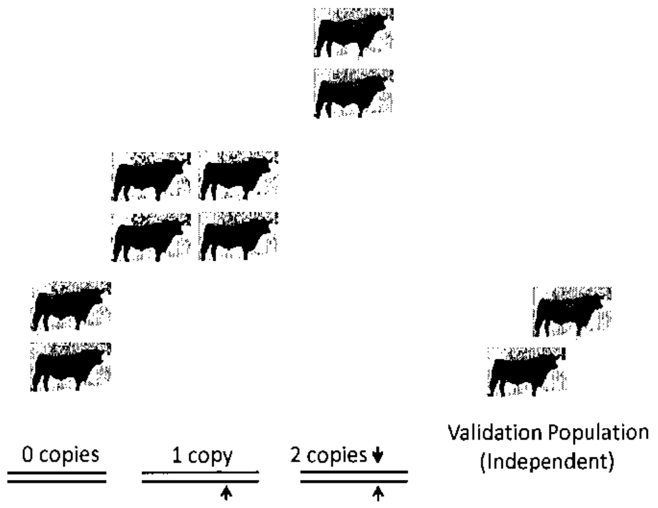
Research is looking for markers in tight linkage disequilibrium (LD) due to close physical proximity to causal mutations

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

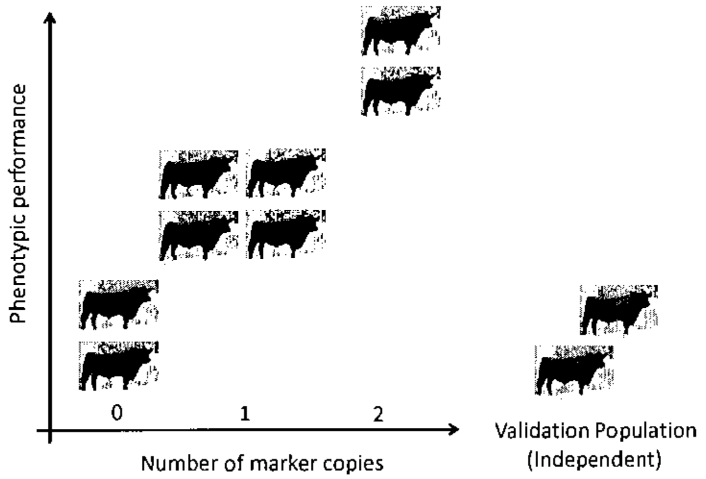


Ideal Validation of Good Marker

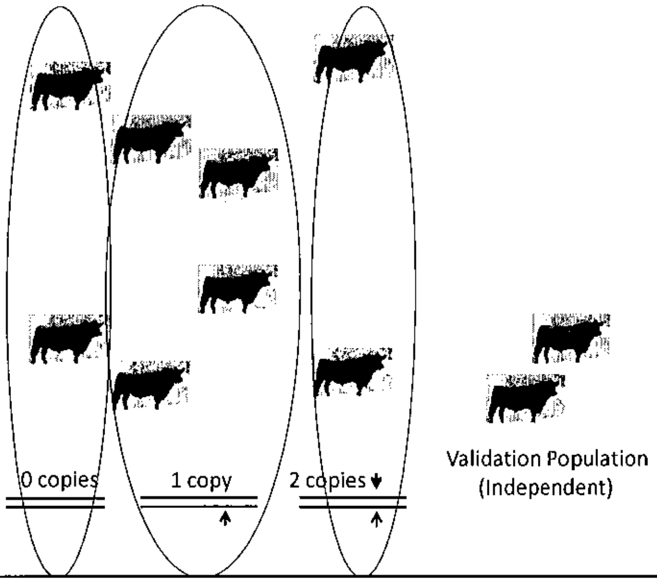
Ideal Validation of Good Marker



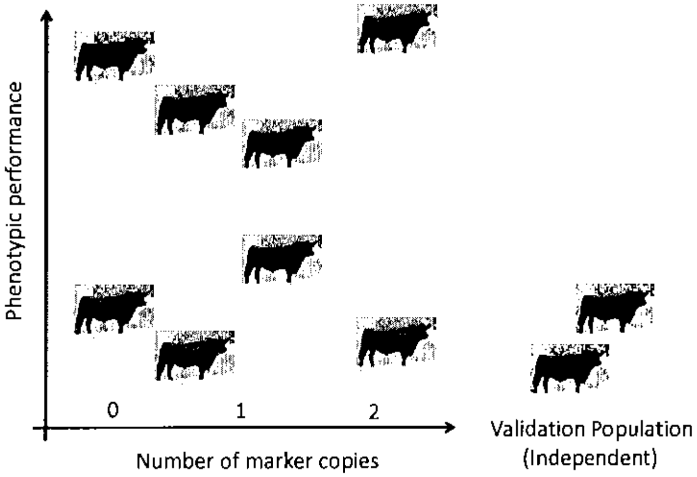
Ideal Validation of Good Marker

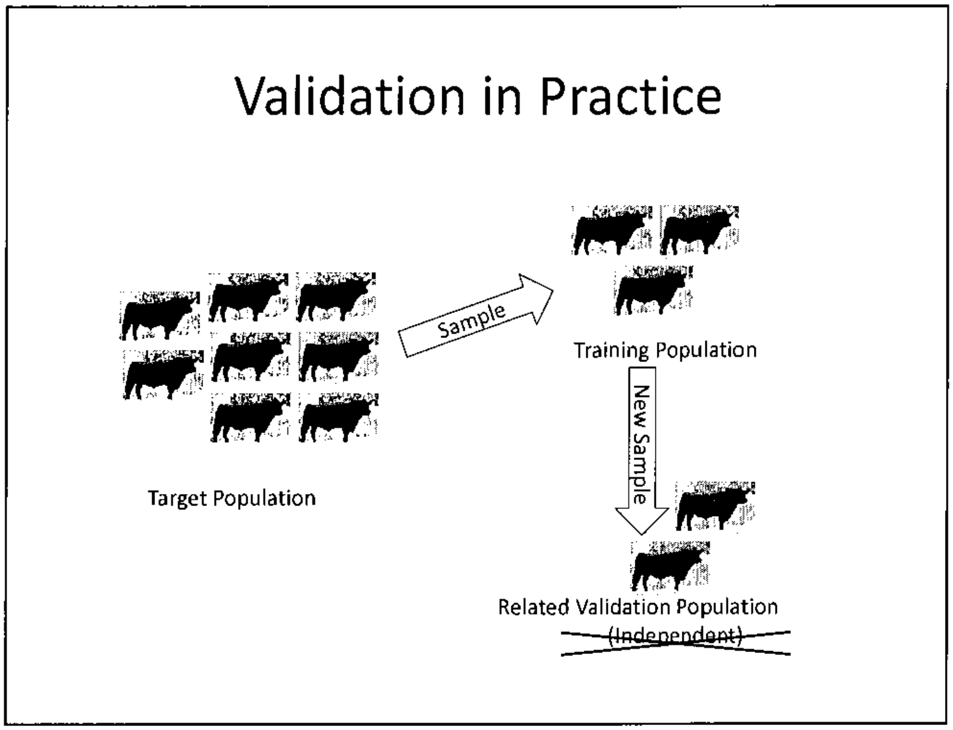


Ideal Failed Validation of Bad Marker



Ideal Failed Validation of Bad Marker





Validation in Practice

Sample

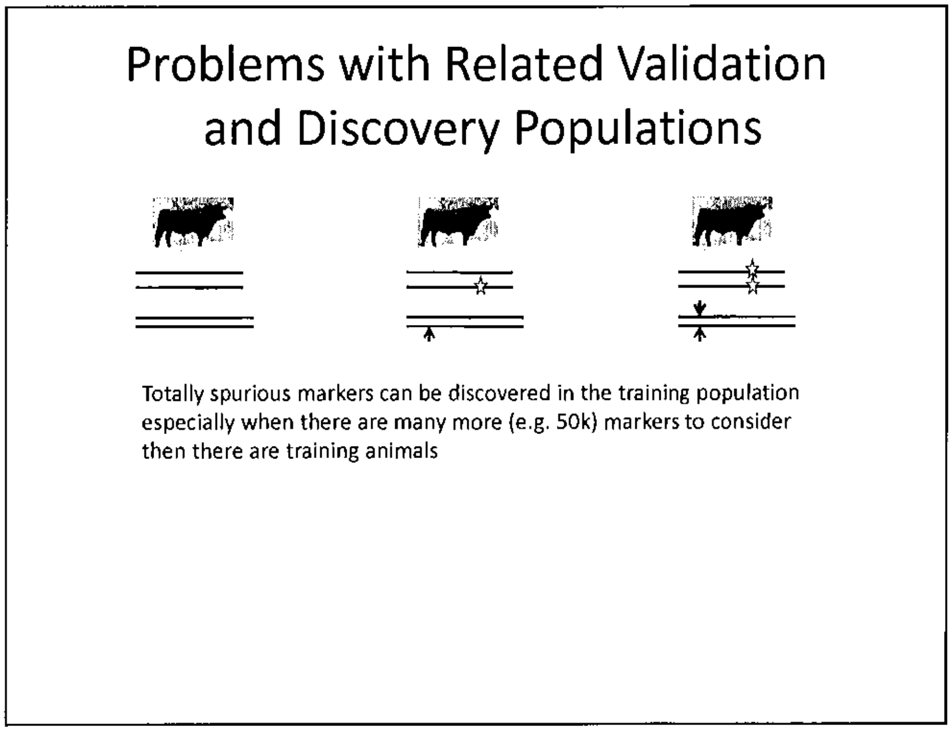
Training Population

New Sample

Target Population

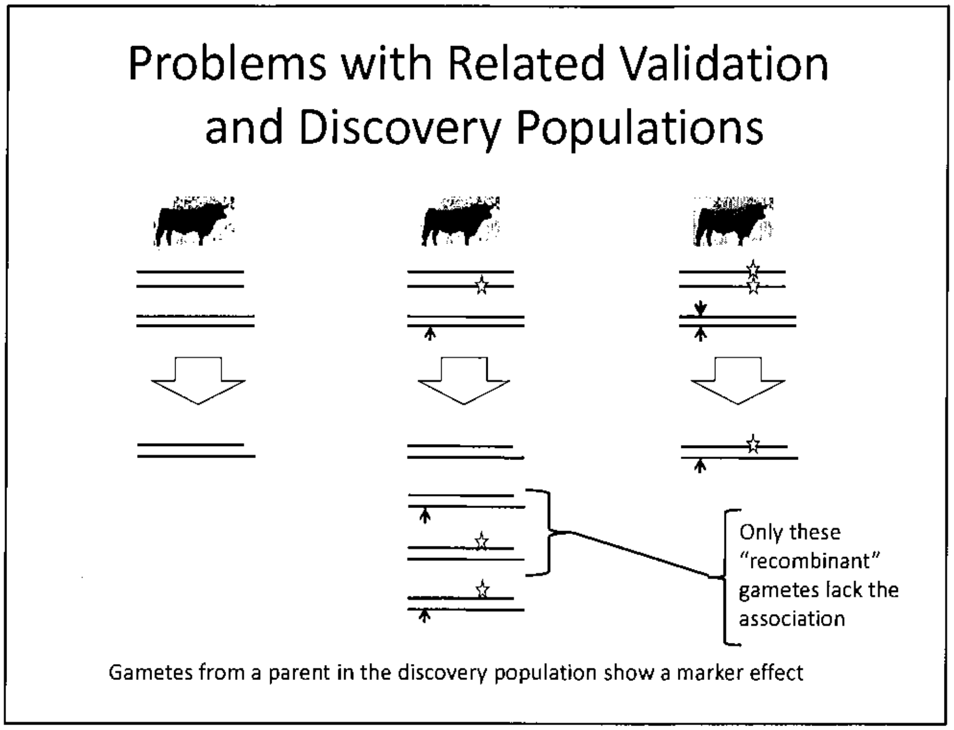
Related Validation Population

~~(Independent)~~



Problems with Related Validation and Discovery Populations

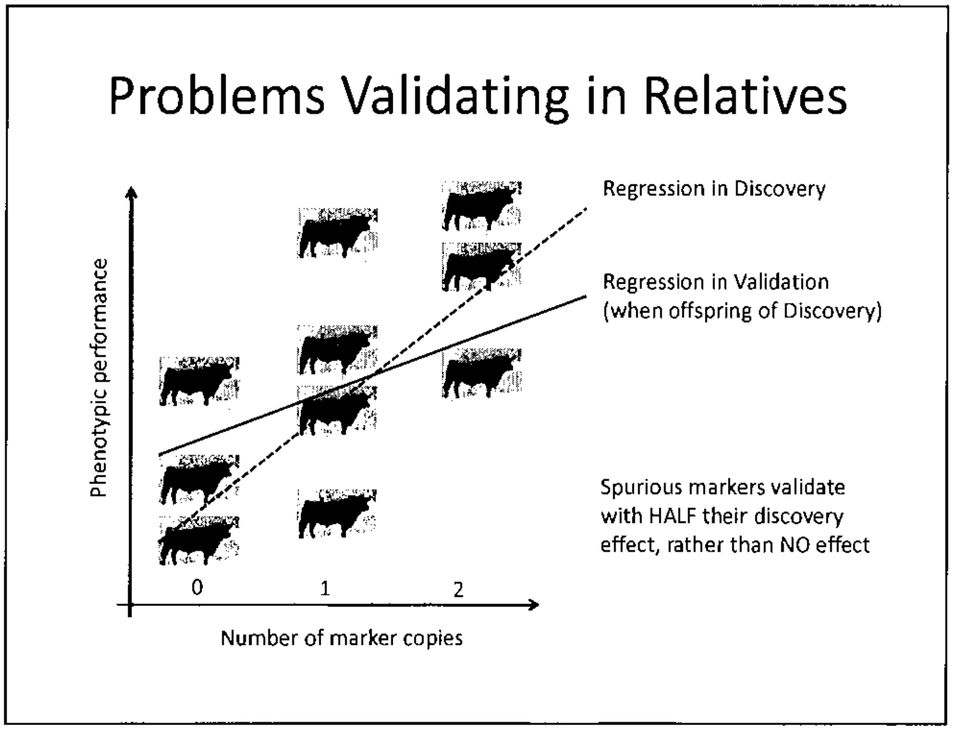
Totally spurious markers can be discovered in the training population especially when there are many more (e.g. 50k) markers to consider then there are training animals



Problems with Related Validation and Discovery Populations

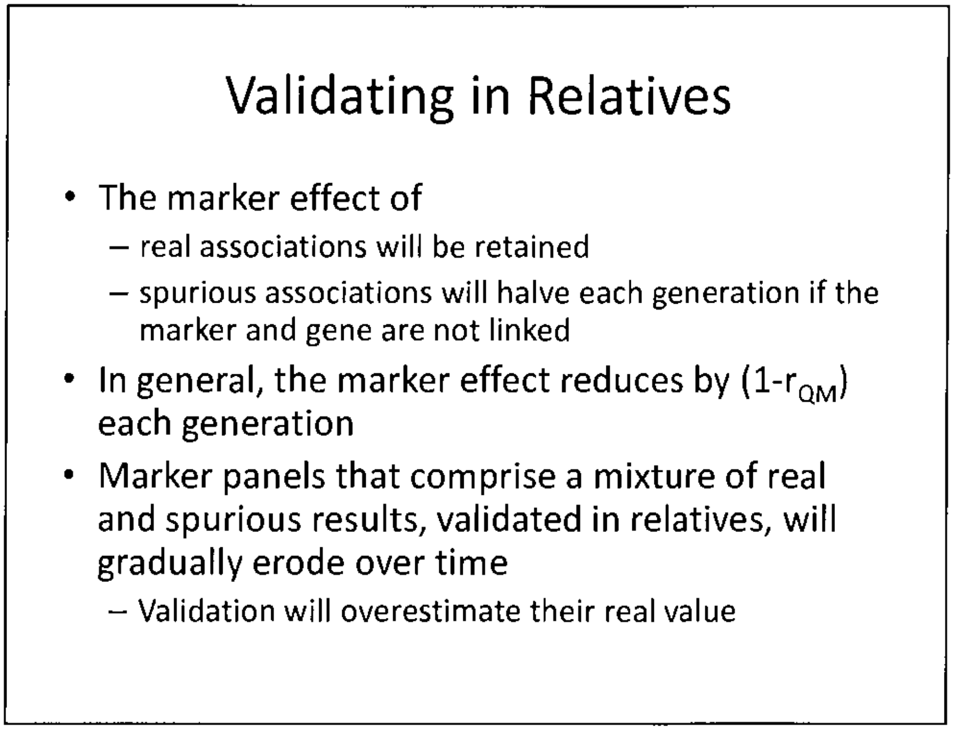
Only these "recombinant" gametes lack the association

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



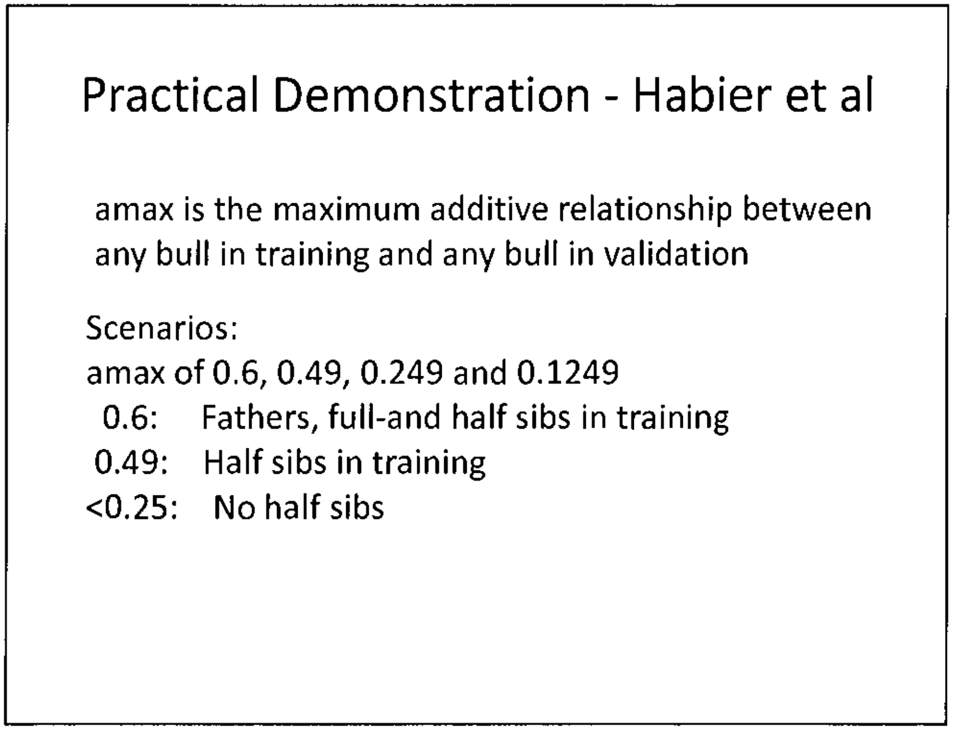
Problems Validating in Relatives

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-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

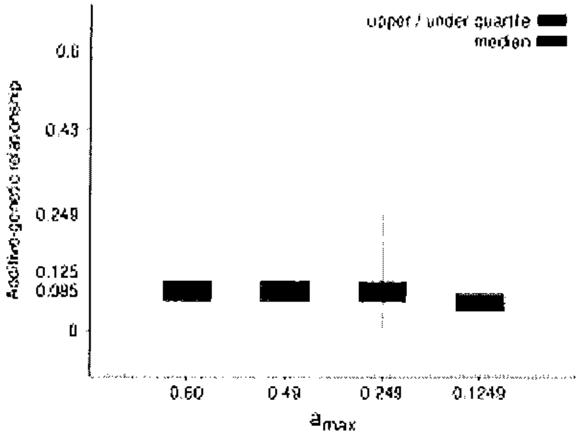


Practical Demonstration - Habier et al

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

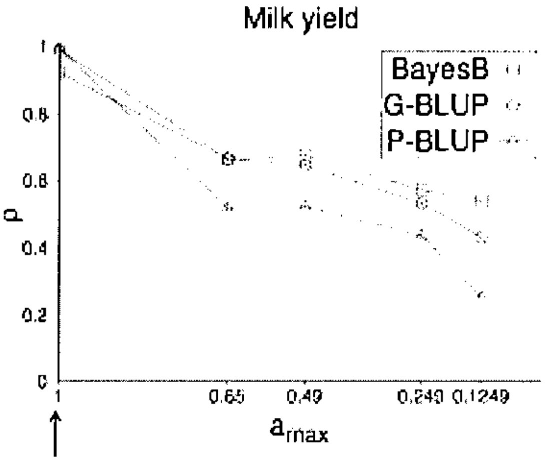
Scenarios:
amax of 0.6, 0.49, 0.249 and 0.1249
0.6: Fathers, full-and half sibs in training
0.49: Half sibs in training
<0.25: No half sibs

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



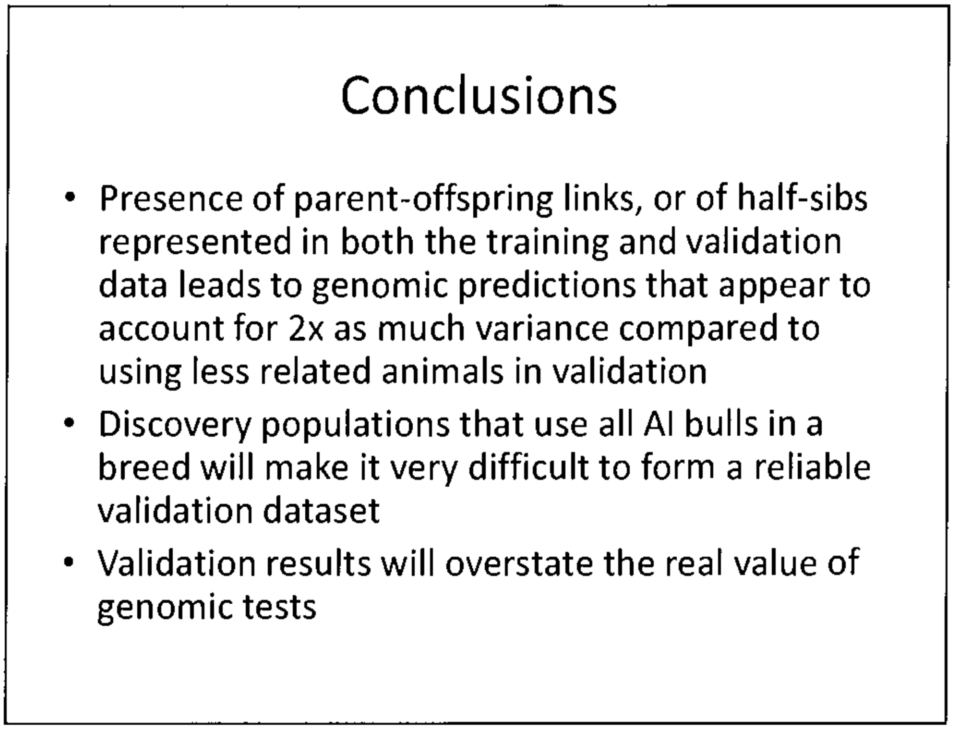
r=0.7 50% variance

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

Faculty of Agriculture and Nutritional Science

C|A|U
Christian-Albrechts-University of Kiel
Institute of Animal Breeding and Husbandry

Genomic Selection using Low-Density SNPs

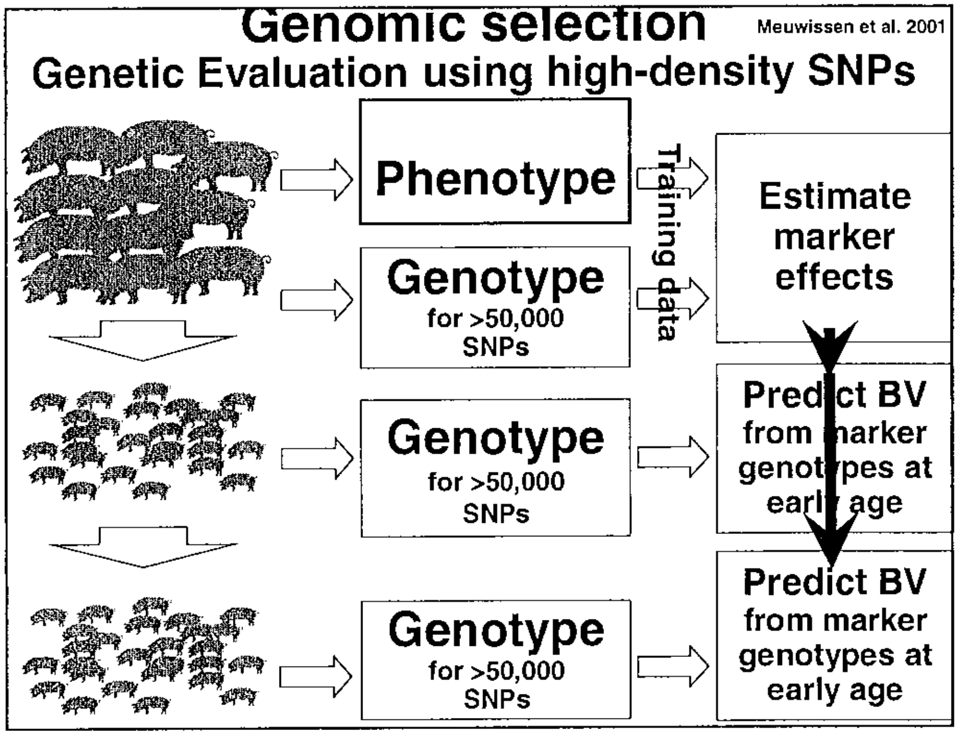
David Habier
Napapan Pyiasatian
Jack Dekkers
Rohan Fernando

***Habier et al. 2009 Genetics** 182: 343 - 353*

Animal Breeding & Genetics

Animal Science CIAG

Iowa State University



Genomic selection

Meuwissen et al. 2001

Genetic Evaluation using high-density SNPs

Phenotype

Genotype for >50,000 SNPs

Training data

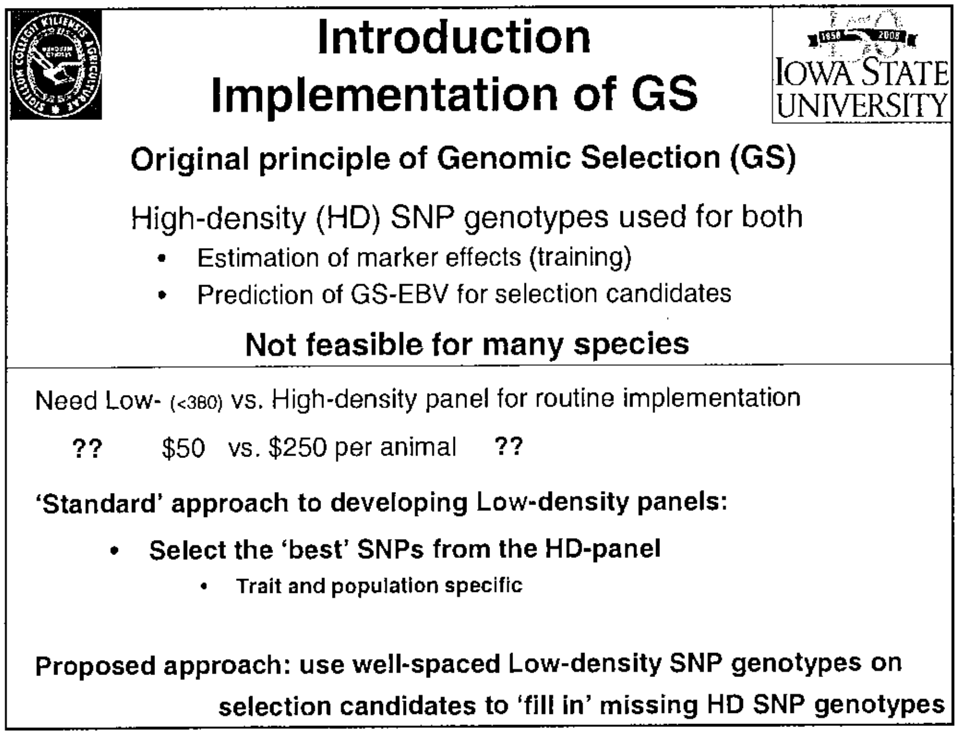
Estimate marker effects

Genotype for >50,000 SNPs

Predict BV from marker genotypes at early age

Genotype for >50,000 SNPs

Predict BV from marker genotypes at early age



Introduction Implementation of GS

Original principle of Genomic Selection (GS)

High-density (HD) SNP genotypes used for both

- Estimation of marker effects (training)
- Prediction of GS-EBV for selection candidates

Not feasible for many species

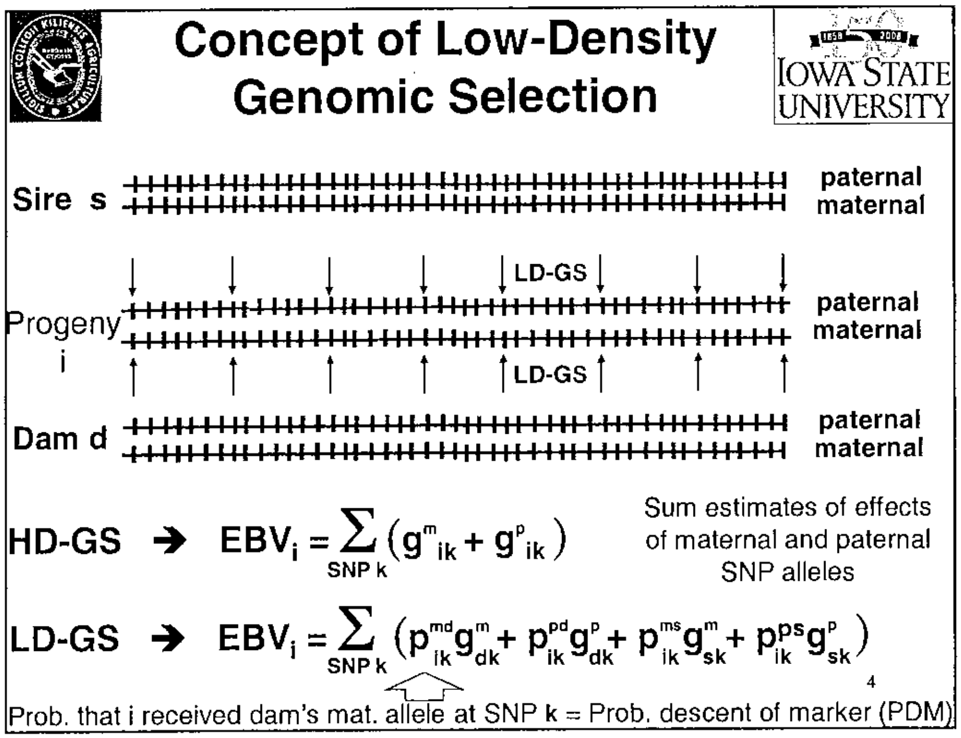
Need Low- (<380) vs. High-density panel for routine implementation

?? $50 vs. $250 per animal ??

'Standard' approach to developing Low-density panels:

- **Select the 'best' SNPs from the HD-panel**
  - **Trait and population specific**

Proposed approach: use well-spaced Low-density SNP genotypes on selection candidates to 'fill in' missing HD SNP genotypes



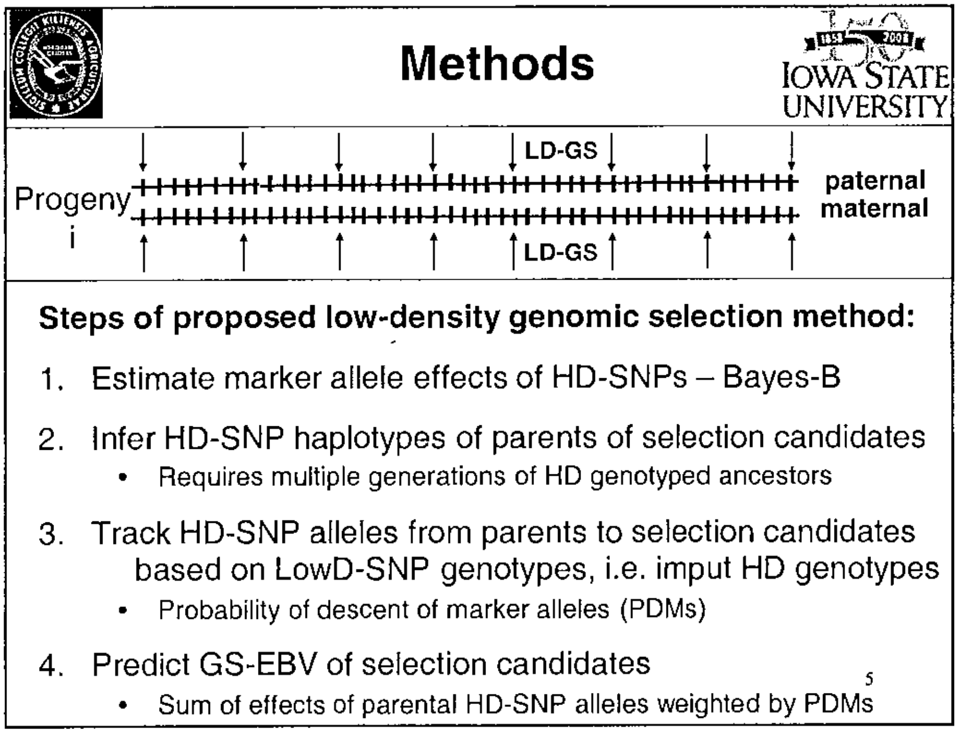
Concept of Low-Density Genomic Selection

HD-GS → $EBV_i = \sum_{SNP\ k} (g^m_{ik} + g^p_{ik})$ Sum estimates of effects of maternal and paternal SNP alleles

LD-GS → $EBV_i = \sum_{SNP\ k} (p^{md}_{ik} g^m_{dk} + p^{pd}_{ik} g^p_{dk} + p^{ms}_{ik} g^m_{sk} + p^{ps}_{ik} g^p_{sk})$

4

Prob. that i received dam's mat. allele at SNP **k** = Prob. descent of marker (PDM)



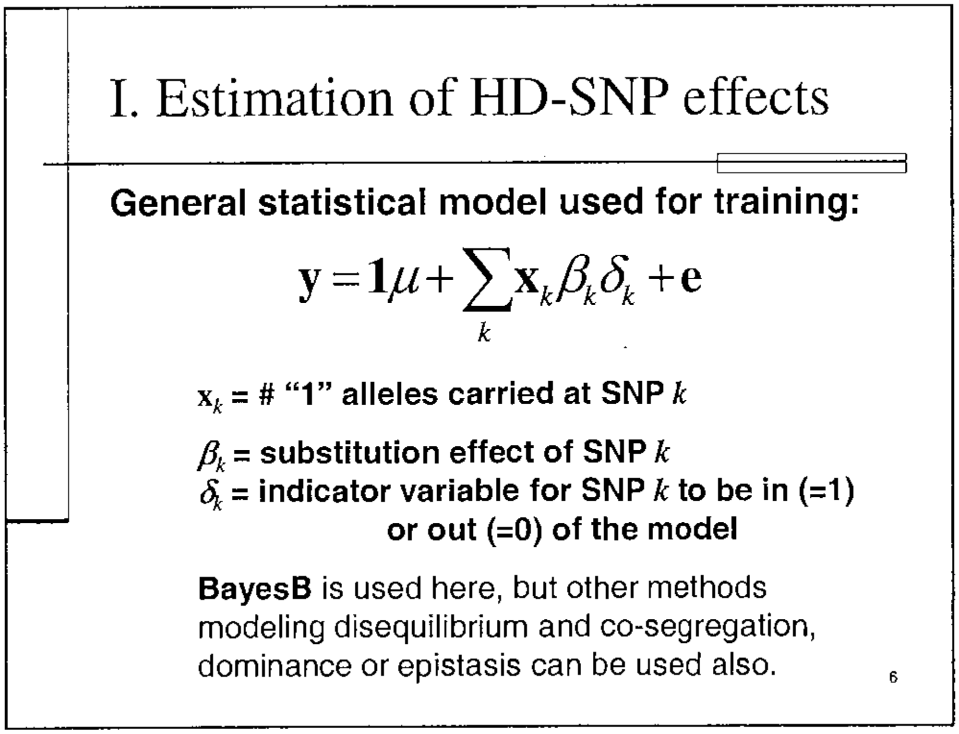
Methods

LD-GS
Progeny i
paternal
maternal
LD-GS

Steps of proposed low-density genomic selection method:

1. Estimate marker allele effects of HD-SNPs – Bayes-B
2. Infer HD-SNP haplotypes of parents of selection candidates
   - Requires multiple generations of HD genotyped ancestors
3. Track HD-SNP alleles from parents to selection candidates based on LowD-SNP genotypes, i.e. imput HD genotypes
   - Probability of descent of marker alleles (PDMs)
4. Predict GS-EBV of selection candidates
   - Sum of effects of parental HD-SNP alleles weighted by PDMs

5



I. Estimation of HD-SNP effects

General statistical model used for training:

$$\mathbf{y} = \mathbf{1}\mu + \sum_k \mathbf{x}_k \beta_k \delta_k + \mathbf{e}$$

x_k = # "1" alleles carried at SNP k

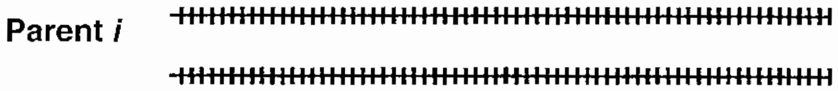
β_k = substitution effect of SNP k
δ_k = indicator variable for SNP k to be in (=1) or out (=0) of the model

BayesB is used here, but other methods modeling disequilibrium and co-segregation, dominance or epistasis can be used also.

6

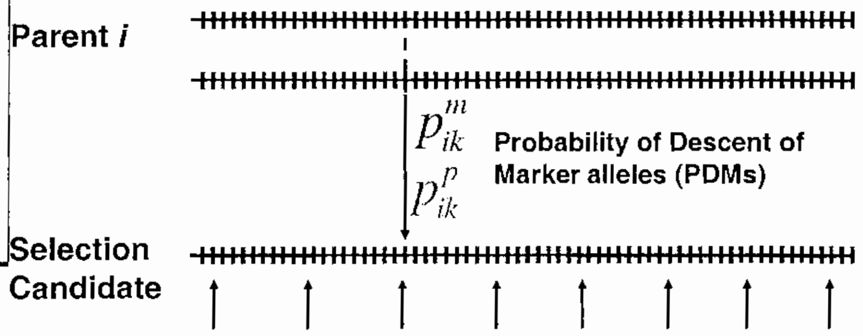
II. Infer HD-SNP haplotypes of parents

To track chromosomal segments from parents to progeny, haplotypes must be inferred for parents



x_{ik}^{m}, x_{ik}^{p} = maternal and paternal allele states of parent *i* at SNP *k*

III. Track HD-SNP alleles – impute HD genotypes



Probability of Descent of Marker alleles (PDMs)

Genotyped for evenly-spaced LD-SNPs

Estimation of PDMs

- MCMC sampling:
  - Joint probabilities of sampled allele origins for adjacent ELD-SNP pairs were estimated
  - Information from all ELD-SNPs is utilized
  - Haplotype phases of HD-genotyped ancestors assumed known

IV. Prediction of GEBVs

- ELD-SNP genotyped selection candidates:

$$GEBV_{ELD} = \sum_{k}^{loci} \left(\hat{x}_k^m + \hat{x}_k^p \right) \hat{b}_k$$

Generation after training: $\hat{x}_k^m = p_k^m * x_k^m \quad \hat{x}_k^p = p_k^p * x_k^p$

Later generations: $\hat{x}_k^m = p_k^m * \hat{x}_k^m \quad \hat{x}_k^p = p_k^p * \hat{x}_k^p$

- HD genotyped parents: $GEBV_{HD} = \sum_{k}^{loci} X\hat{b}_k$

$$= \sum_{k}^{loci} \left(x_k^m + x_k^p \right) \hat{b}_k$$

Tested by Simulation

IOWA STATE UNIVERSITY

	Population	Genome
Generation -1060	Random Mating (N_e=500)	10 chromosomes of 1 M 20,000 SNPs ; 500 QTL
Generation -60	Random Mating (N_e=100)	1,000 SNPs selected after 1060 gener.
Generation -10	Population Growth (N=100 to N=1000)	HD SNP spacing ~ 1 cM LD SNPs at 10 or 20 cM
Generation 1-3	50 males x 500 females (N=1000)	Trait h^2 = 0.5

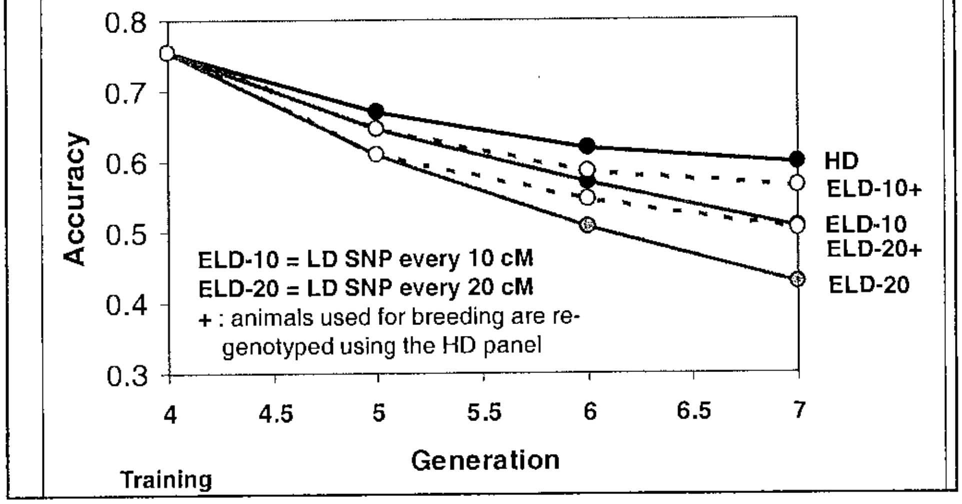
Pedigree recording and genotyping starts

Generation 4	Training data (N=1000)	Bayes-B (Meuwissen et al. '01)
Generation 4-7	10 males x 100 females	→ GS-EBV using HighD SNPs → GS-EBV using LowD SNPs

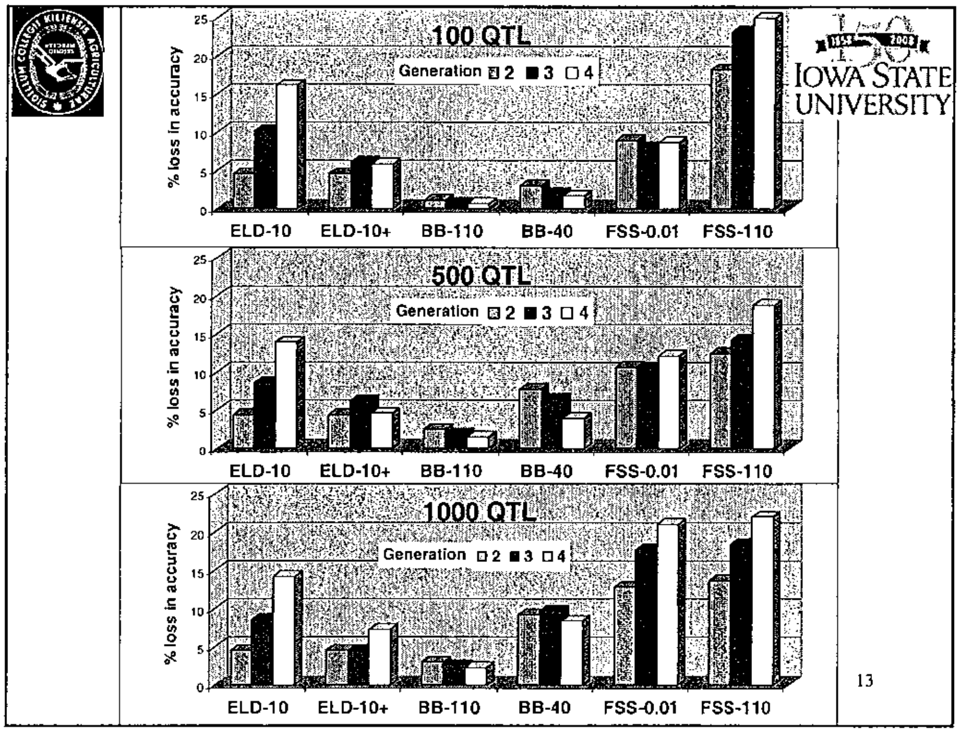
Results

IOWA STATE UNIVERSITY

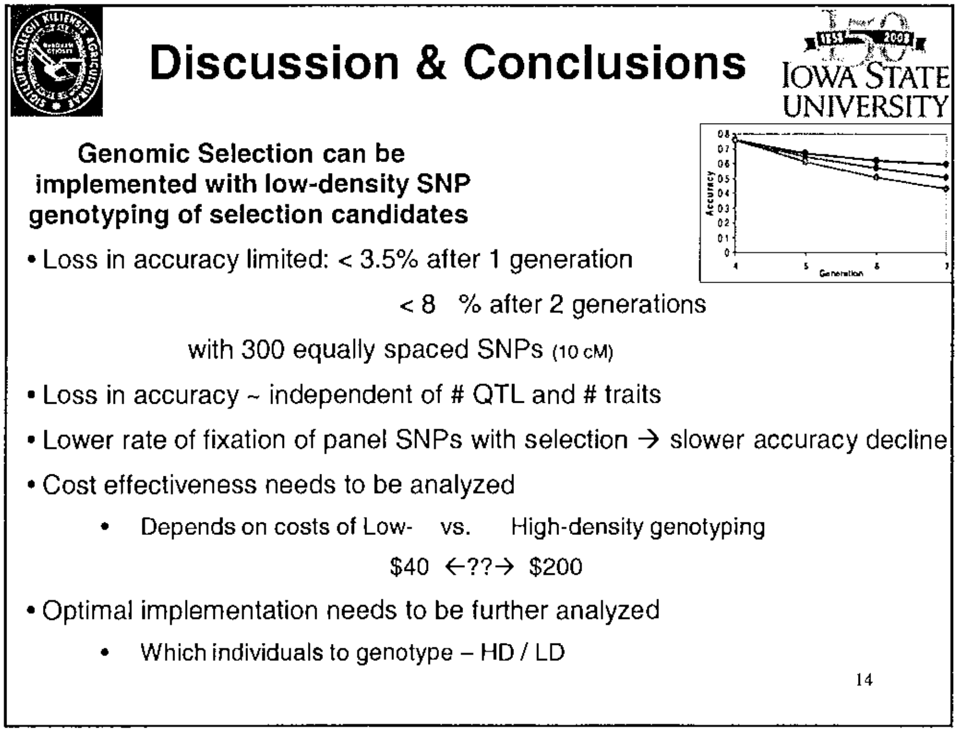
Accuracy of GS-EBV based on High- and Low-Density SNP genotyping (20 Replicates) 500 QTL (~220 MAF>0.01)



ELD-10 = LD SNP every 10 cM
ELD-20 = LD SNP every 20 cM
+ : animals used for breeding are re-genotyped using the HD panel



13



Discussion & Conclusions

IOWA STATE UNIVERSITY

Genomic Selection can be implemented with low-density SNP genotyping of selection candidates

- Loss in accuracy limited: < 3.5% after 1 generation

  < 8 % after 2 generations

  with 300 equally spaced SNPs (10 cM)
- Loss in accuracy ~ independent of # QTL and # traits
- Lower rate of fixation of panel SNPs with selection → slower accuracy decline
- Cost effectiveness needs to be analyzed
  - Depends on costs of Low- vs. High-density genotyping

    $40 ←??→ $200
- Optimal implementation needs to be further analyzed
  - Which individuals to genotype – HD / LD

14