How can we best use DNA data in the selection of cattle?

5/1/09 **Mike Goddard**
University of Melbourne and Victorian DPI

Victoria
The Place To Be

Introduction

Background

What is Genomic selection and how does it work?

Implementation of genomic selection

Use of genomic selection by farmers

Future

Victoria
The Place To Be

2

Background

Some traits controlled by one gene
Eg red vs black coat colour

Most traits controlled by many genes and by environmental effects

We don't know the genes that control marbling, fertility, growth rate etc

Victoria
The Place To Be
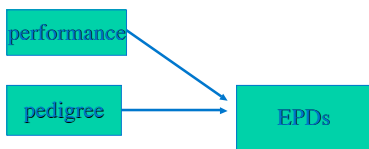
3

Background

Traditional approach

Estimate the combined effect of all genes on a trait by using performance records of the animal and his/her relatives
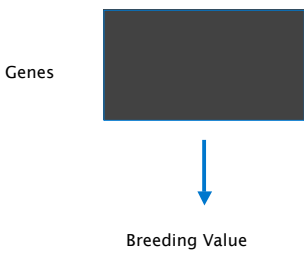
→ EPDs

Victoria
The Place To Be

4

Existing method of calculation of EPD

Victoria
The Place To Be

5

Traditional Genetic Improvement

Genes

Breeding Value

Victoria
The Place To Be

6

1

Traditional Genetic Improvement

Very successful

Australian Profit Index for Holstein bulls over 16 years

Victoria
The Place To Be

7
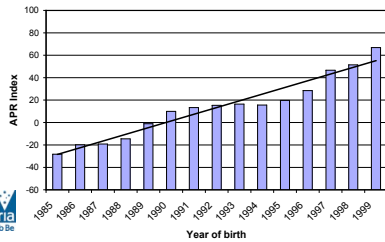
Genetic variation is due to differences in DNA sequence

Genes are a sequence of DNA eg AGTCTAG

Genetic differences are due to differences in DNA sequence
eg AGTCTAG
AGTGTAG

Victoria
The Place To Be

8

Adding DNA sequence data to traditional breeding

Assay DNA sequence variation
+
phenotypes and pedigrees

--> more accurate EBVs
at a younger age

Victoria
The Place To Be

9

Background

Genetic differences are due to naturally occurring variations in DNA sequence

CTGAATGGT
CTGCATGGT

Single nucleotide polymorphism (SNP)

Victoria
The Place To Be

10

Background

SNPs that cause change in performance are not known

>100,000 SNPs scattered throughout the genome are known

Use them as marker for functional SNPs

There is a known SNP near to every gene affecting a trait

Victoria
The Place To Be

11

DNA markers are located close to genes for economic traits

C	+	common
G	-	common
C	-	rare
G	+	rare

Victoria
The Place To Be

12

Background

Use the markers (SNPs) to predict the breeding value for a trait

Victoria
The Place To Be
13

Adding DNA sequence data to calculate more accurate EPDs

phenotype

pedigree → EPDs

DNA

Victoria
The Place To Be
14
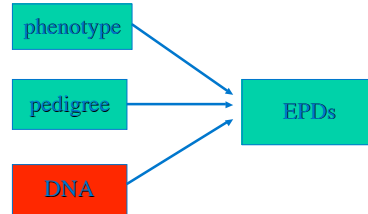
Overview of genomic selection
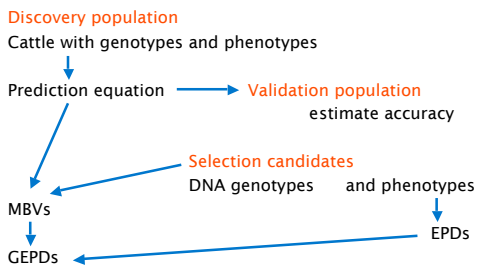
Discovery population
Cattle with genotypes and phenotypes

Prediction equation → **Validation population**
estimate accuracy

Selection candidates
DNA genotypes   and phenotypes

MBVs

GEPDs ← EPDs

Victoria
The Place To Be
15

This talk

Genetic architecture of economic traits

Discovery

Validation

Implementation

Future

Victoria
The Place To Be
16

Genetic Architecture of economic traits

How many genes affect a trait?

How big are their effects?

How frequent are the + alleles?

Victoria
The Place To Be
17

How many genes?
Results in humans

EG height (Visscher, Nat. Genet. May 2008)

3 studies with <500,000 SNPs and 63,000 people

95 v. sig SNPs → 54 verified
3 genes found by all 3 studies
Effect 0.4 -0.8 cm → 0.1- 0.4% variance explained
Therefore 100s – 1000s of genes for height

Victoria
The Place To Be
18

Results

Distribution of F values for 38,000 SNPs effects on milk protein yield

F- value	Number
0-1	22702
1-5	13047
5-10	2078
10-20	375
20-30	31
>30	11

Results

Fit model for the distribution of a (marker effects)
Mixture of normal distributions (EM algorithm)

76% r ~ N(0,0.00035)
20% r ~ N(0, 0.0014)
4%   r ~ N(0, 0.0032)

How many genes?
Results in dairy cattle

Distribution of F values for 36,000 SNPs effects on milk protein yield

F- value	Number	Predicted
0-1	22702	22898
1-5	13047	12827
5-10	2093	2078
10-20	375	419
20-30	31	29
>30	11	5

Distribution of effects of genes
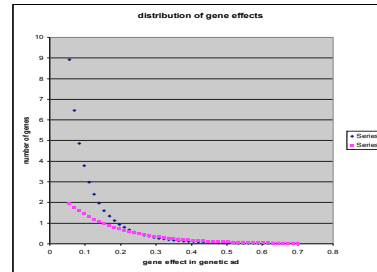
Frequency of neutral mutations
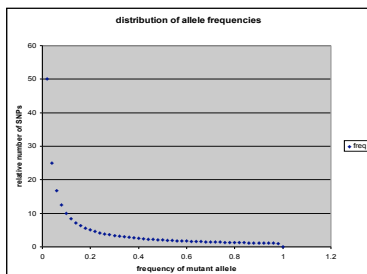
Frequency of mutant alleles after inbreeding

Frequency of mutant alleles affecting economic traits

Selection drives frequency to 0 or 1
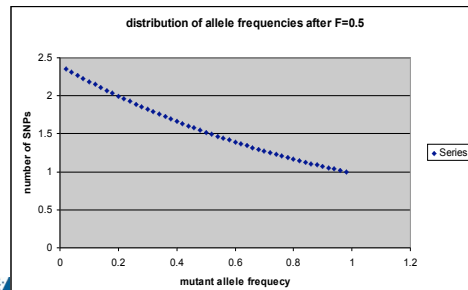
Alleles still segregating are often near 0 or 1 in frequency

25

This talk

Genetic architecture of economic traits

Discovery

Validation

Implementation

Future

26

Discovery

Population with phenotypes and DNA genotypes

Statistical methods to derive prediction equation

Number of animals

Number of SNPs

27

Discovery
Statistical methods

BLUP
   Assume all SNPs have some effect

Bayes B
   Assume only some SNPs have effects and some are larger then others

28

Statistical method

Meuwissen, Hayes and Goddard (2001) proposed genomic selection and tested with simulated data

Table 5.2. Comparing estimated *vs.* true breeding values in progeny with no phenotypic records (from Meuwissen et al. (2001), reproduced with permission from the authors)

	$r_{TBV;EBV}$ + SE	$b_{TBV.EBV}$ + SE
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
Bayes	$0.848 + 0.012$	$0.946 + 0.018$

29

Statistical method

Dairy cattle

Correlation with EBV

BLUP	0.59
Bayes	0.60

30

Statistical method

Choice of methods depends on number of genes for trait compared to number of chromosome segments

($\approx N_eL = 100*30 = 3000$ in Holstein)
($= 1000*30 = 30,000$ across breeds)

Prediction
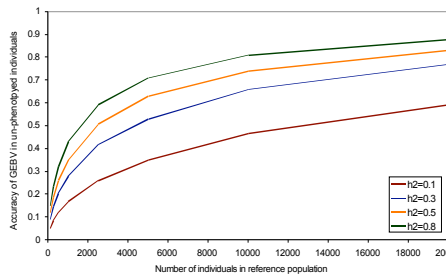Bayes B will be better than BLUP when we have more SNPs and across breed data

31

Number of Cattle in Discovery Population
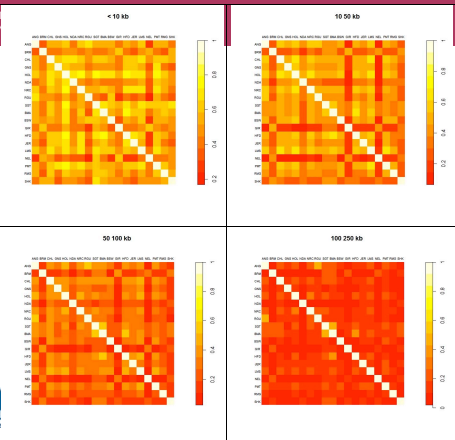
Accuracy of the prediction equation depends on

$$\frac{Th^2}{N_eL}$$

Low traits need more animals

32

Effect of number of animals on accuracy of prediction equation

33

Does the prediction equation work in other breeds?

35
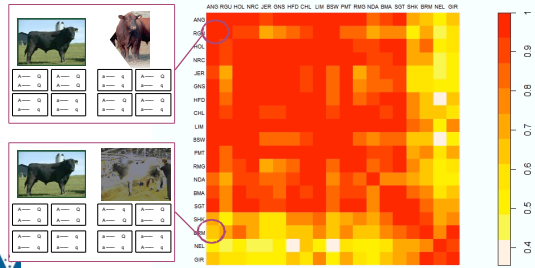
Does a prediction equation work in other breeds?

Holstein equation doesn't work in Jerseys

It is possible to get an equation that works in both Jerseys and Holsteins if both used in discovery population.

Hard to get equation that works in many breeds

Need > 300,000 SNPs

Need separate prediction equations for Bos indicus, Bos taurus and Crosses

36

6

Validation

Prediction equation works better in discovery population than in new population.

Want validation population that is similar to population where prediction equation will be used (selection candidates)

Not satisfactory to randomly split discovery population into discovery and validation

37

Validation

Purpose is to estimate the correlation between the prediction and the true genetic merit.

Answer isn't yes or no but 0.37

Research is working towards a robust method a calculating accuracy from the discovery population, but we're not there yet

38

Validation of Parallele 10k results

	A-tags	Angus PT	JxL	TC
No. of animals	189	396	325	383
No. sig SNP*	12/44	9/89	4/93	8/92
No. same direction	12/12	6/9	1/4	5/8

* p<0.1

39

Reasons for poor validation

· Individual SNP effects are small and so hard to estimate accurately

· LD is different across breeds

· Causal genes have low allele frequency
   – Poor LD with SNPs
   – Contribution to variation varies between breeds and samples

40

Lessons to learn for validation

Use very large discovery populations so SNP effects are estimated accurately

Discover and validate in same breed(s)

Use enough animals in validation

Validation population that represents the target population

41

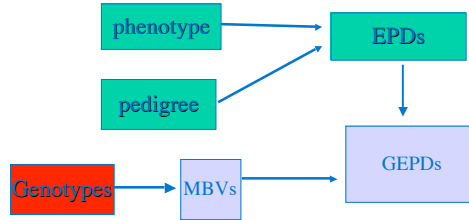## Lessons to learn for validation

Use SNP chip with >300,000 SNPs

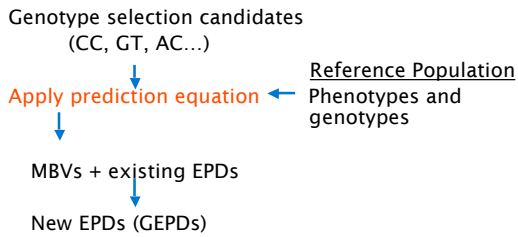Better methods to handle causal genes with low frequency alleles

42

Implementation

43

Adding DNA sequence data to calculate more accurate EPDs

44

Calculation of GEPDs

Genotype selection candidates
(CC, GT, AC…)

Apply prediction equation ← Reference Population
Phenotypes and genotypes

MBVs + existing EPDs

New EPDs (GEPDs)

45

Implementation

Beef Model
    DNA Company supplies EPD centre with MBVs

Dairy model
    DNA lab supplies EPD centre with raw genotypes

46

Implementation

Advantages of dairy model

The prediction equation can be continuously updated using the data from commercial genotyping and estimated from the maximum amount of data

The data stored are stable (MBVs vary with company, panel, time)

Mistakes are less likely

Covariances among all MBVs are not required

47

Future

Genomics technology will continue to develop
– >300,000 SNP chips
– Genome sequencing

DNA profiles will become common

DNA genotypes will be incorporated directly into EPDs

48

Future

Structure of breeding industry may be changed
- Decreased recording
- Cheap bulls
- Selection for novel traits
- Vertical integration
- Pig and poultry model?

49

Summary – Genetic Architecture

There are 100's to 1000s of genes for most traits

Most markers have very small effects

QTL not completely smeared over all DNA

QTL often have a rare allele

50

Summary – Discovery of SNPs

Best method for estimating prediction equation depends on number of QTL

Bayes B best if number of genes small

BLUP best as number of genes approaches N_eL

Bayes B may be better for 300,000 SNP data across breeds

51

Summary – Discovery of SNPs

Need a large population with genotypes and phenotypes (14,000 cattle for trait with h^2=0.3 to get accuracy of 0.7)

Low h^2 means more animals

Try international collaboration to build numbers

Use multiple breeds if prediction equation is to be used in multiple breeds and >300,000 SNPs

52

Summary – Validation of prediction equation

Validate in population similar to selection candidates

Random subset of discovery population not satisfactory

Disappointing results common

Purpose is to estimate correlation of MBV with genetic value

Need resource populations with industry support

53

Summary – Implementation

Beef and Dairy models different

I think Dairy model will deliver more accurate EPDs

54

9