The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Comparison of Bayesian Models for genomic selection using real dairy data

K.L. Verbyla*, P.J. Bowman, B.J. Hayes, H. Raadsma,

M. Khatkar and M.E. Goddard

Prepared for Session 28 EAAP 2009

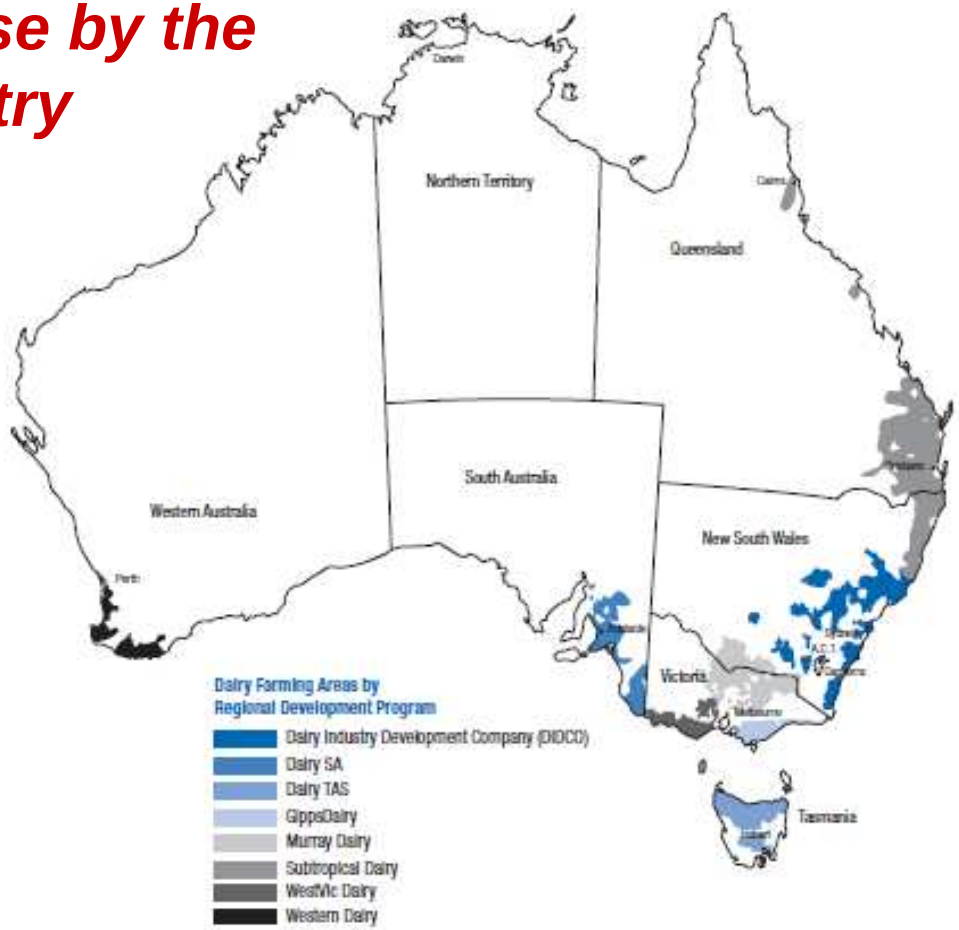
**Email: klara.verbyla@dpi.vic.gov.au*

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Project Aim:
To find the best method for GENOMIC SELECTION for implementation and use by the Australian dairy industry



The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Generally, the methods to predict of GEBV face 2 statistical issues:

- Number of SNP (p) is greater than the number of individuals or records (n) i.e **p>n** problem
  - oversaturated or overparameterised model
- Large number of SNP effects that are zero or close to zero (need a sparse model).

WHAT APPROACH SHOULD BE USED??

- Dimension Reduction (PCA, PLS)
- Machine learning (SVM)
- Shrinkage models – penalised methods exploring sparsity (LASSO)
- Variable Dimension Model approaches (SSVS)
- Variable Selection (reduced set of SNPs)

Project: Tested 20 Methods!

→ **Bayesian Inference**

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Data

Reference Population

• 1098 Holstein Friesian bulls progeny tested ≤ 2003

Validation Population

• 400 Holstein Friesian bulls progeny tested > 2003

Phenotypes

• deregressed breeding values for protein, fat, milk volume, protein%, fat%, fertility, ASI (Australian Selection Index), APR (Australian Profit Ranking) and Overall type.

Genotypes

•39,048 markers

Evaluate methods on

• r(GEBV,ABV)

Correlation of Predicted GEBV with Australian breeding value (ABV)

•MSE

•Regression Coefficient

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Bayesian Models

Statistical Model:

$$y = \mu 1_n + \sum_j^p X_j \beta_j + Zu + e$$

- y is the vector of phenotypes of the trait for n individuals
- μ is the mean
- 1_n is a vector of ones of length n
- X_j is a vector of indicator variables representing the genotypes of the jth marker for all individuals (x_{ij}=0,1,2)
- β_j is the size of the SNP effect associated with marker j
- u is the vector of random polygenic effects of length n (Z is the associated design matrix)

$$u \sim N\left(0, \sigma_u^2 A\right)$$

- e is the residual error

$$e \sim N\left(0, \sigma_e^2 \mathbf{I}_n\right)$$

$$\boldsymbol{GEBV} = \hat{\boldsymbol{u}} + \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Bayes A

Prior Distributions
SNP effects

$$\beta_i \mid \sigma_i^2 \sim N\left(0, \sigma_i^2\right)$$

$$\sigma_i^2 \sim \chi^{-2}(r, S) \sim \gamma^{-1}\left(\frac{r}{2}, \frac{rS}{2}\right)$$

r degrees of freedom and
scale parameter S

SNP EFFECTS
- Normal-inverse scaled chi square (t distribution)
- unequal variance
- assumes that all SNPs have an effect
- Gibbs Sampler

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Bayes BLUP

Prior Distributions
SNP effects

$$\beta_i \mid \sigma_\beta^2 \sim N\left(0, \sigma_\beta^2\right)$$

$$\sigma_\beta^2 \sim \chi^{-2}(r, S) \sim \gamma^{-1}\left(\frac{r}{2}, \frac{rS}{2}\right)$$

r degrees of freedom and
scale parameter S

SNP EFFECTS
- Normal
- equal variance
- Infinitesimal assumptions
- assumes that all SNPs have an effect
- Gibbs Sampler

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Bayes C
Stochastic Search Variable Selection (SSVS)

(George and McCulloch, 1993)

Use latent variable γ_i (0,1)

SAME ASSUMPTIONS AS BAYES B

Prior Distributions
SNP effects

$$\beta_i \mid \gamma_i, \sigma_i^2 \sim (1-\gamma_i) N\left(0, \sigma_i^2/100\right) + \gamma_i N\left(0, \sigma_i^2\right)$$

$$\sigma_i^2 \sim \chi^{-2}(\boldsymbol{r}, \boldsymbol{S})$$

$$\gamma_i \sim \boldsymbol{bernoulli}(\boldsymbol{p}_i)$$

$$1 - \boldsymbol{p}(\gamma_i = 0) = \boldsymbol{p}(\gamma_i = 1) = \boldsymbol{p}_i$$

SNPs with γ_i=0, posterior values limited to values close to 0 (but not removed from the model- NO changing dimensionality) – GIBBS SAMPLER

SNP EFFECTS
- Mixture of two Normal-inverse scaled chi square distributions (t distributions)
- unequal variance
- assumes that a few SNP have an significant effect

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Methods

- Bayes A
  - All SNPs
  - Selected SNPs with weights
  - Selected SNPs without weights
- Bayes BLUP
  - All SNPs
  - Selected SNPs with weights
  - Selected SNPs without weights
- Bayes C
  - All SNPs

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

SNP Pre-selection

Single SNP analysis (ASReml)

$$y = 1_n \mu + X\beta + Z_1 u_1 + Z_2 u_2 + e$$

- X is a vector of indicator variables representing the genotypes of the current SNP marker for all individuals (X_n=0,1,2) and β is the associated effect of the SNP
- u_1 is the random sire effect (Z_1 associated design matrix)

$$u_1 \sim N(0, \sigma^2_{u_1} I)$$

- u_2 is the random maternal grand sire effect (Z_2 associated design matrix)

$$u_2 \sim N(0, \sigma^2_{u_2} A)$$

Fitted with and without weights

-Weights = Number of Effective Records

-SNPs with p-value <0.1 included in predictive set.

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Results r(GEBV,ABV)

Method	*Fat*	*Fat%*	*Milk*	*Protein*	*Protein%*
Bayes BLUP - All SNPs	0.528	0.630	0.648	0.613	0.660
Bayes BLUP - Selected SNPs (Unweighted)	0.527	0.689	0.646	0.596	0.678
Bayes BLUP - Selected SNPs (Weighted)	0.543	0.643	0.659	0.610	0.661
Bayes A - All SNPs	0.538	0.700	0.631	0.572	0.645
Bayes A - Selected SNPs - (Unweighted)	0.543	0.712	0.639	0.579	0.667
Bayes A - Selected SNPs - (Weighted)	0.538	0.704	0.635	0.583	0.648
Bayes C	0.557	0.728	0.644	0.588	0.670

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Fat % - DGAT1

Table 1. Effect of the *DGAT1 K232A* Mutation on Sires' Daughter Yield Deviations (DYDs) for Milk Yield and Composition

Trait	$\alpha/2 \pm$ 2std.err.	r^2_{QTL}	P value$_{QTL}$	$r^2_{polygenic}$	r^2_{error}
Milk yield (kgs)	−158 ± 24.5	0.18	5.00E − 35	0.49	0.32
Fat yield (kgs)	5.23 ± 0.9	0.15	1.57E − 29	0.55	0.30
Protein yield (kgs)	−2.82 ± 0.7	0.08	1.70E − 15	0.65	0.26
Fat (%)	0.17 ± 0.012	0.51	4.33E − 122	0.29	0.19
Protein (%)	0.04 ± 0.006	0.14	5.05E − 28	0.66	0.20

(i) $\alpha/2$: QTL allele substitution effect on DYD (≈ halve breeding value), corresponding in the mixed model to the regression coefficient on the number of *K* alleles in the *DGAT1 K232A* genotype, and to $\alpha/2$, where α is defined according to Falconer and Mackay 1996. (ii) r^2_{QTL}: proportion of the trait variance explained by the *DGAT1 K232A* polymorphism. (iii) P value$_{QTL}$: statistical significance of the *DGAT1 K232A* effect. (iv) $r^2_{polygenic}$: proportion of the trait variance explained by the random, polygenic effect in the mixed model. (v) r^2_{error}: proportion of the trait variance unexplained by the model.

- QTL explains > 50% of genetic variance in fat%
- QTL allele is common and acts additively

→ Major violation of BLUP assumptions

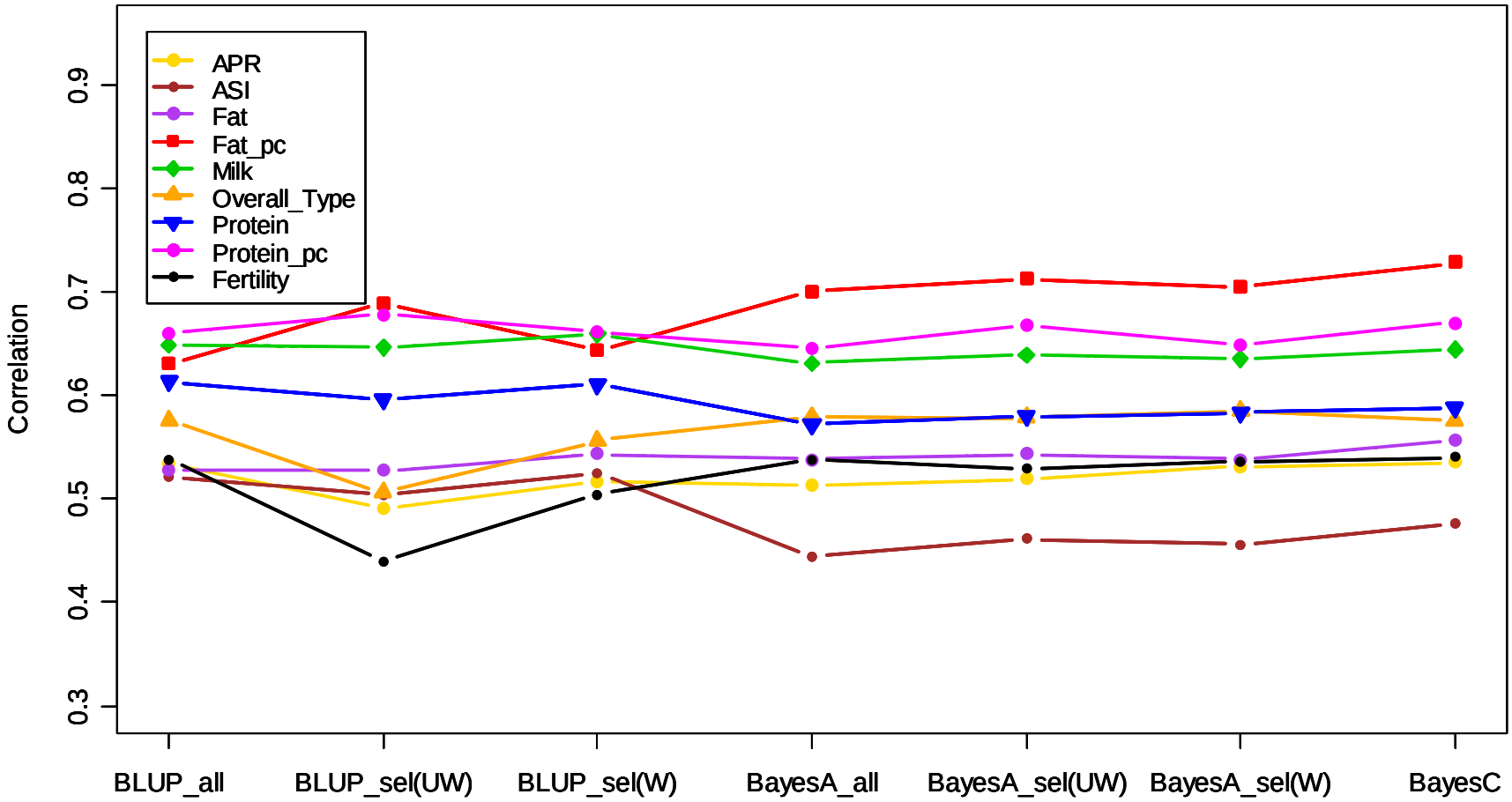
Grisart et al. (2002)

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Results r(GEBV,ABV)



The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

RESULTS - IN CONTEXT:

A SUBSET OF ALL METHODS USING ALL SNPs

Method	*Average Correlation Across all traits*
Bayes BLUP	*0.589*
Bayes A	*0.578*
Bayes C	*0.597*
LASSO	0.595
SVR	0.587
GBLUP	0.588
PLS	0.592

The Cooperative Research Centre for BEEF GENETIC TECHNOLOGIES

State Government Victoria Department of Primary Industries

THE UNIVERSITY OF MELBOURNE Australia

Conclusions

- Only small differences in accuracy and bias of GEBV from different methods
- Method by trait interaction. Better results when priors matched “real” distribution of QTL effects

  → *best method is trait dependent!*
- Pre-selection of SNP neither reduces or increases the accuracy of predicted GEBV
- Bayesian BLUP performs as well as or better than the other methods EXCEPT for traits with QTL that explain large amount of genetic variance eg. Fat % with DGAT1
- Still a need to find a method that produces equally accurate GEBV across traits with different genetic architecture.

Recognitions

Personally:

- Ph.D Supervisors: Mike Goddard, Ben Hayes and Richard Huggins.
- Melbourne University, Beef CRC and DPI Victoria for their ongoing support.
- Marie Curie EST Sabretrain Fellowship Supervisors: Roel Veerkamp, Mario Calus and Han Mulder (Wageningen University and Research)

Project:

B. Hayes

H. Raadsma

P. Bowman

M. Khatkar

M.E. Goddard

All other people involved in the project