Genomic prediction within and between dairy cattle breeds with an imputed high density marker panel

Malena Erbe[1], B.J. Hayes[234], P.J. Bowman[23], H. Simianer[1] and M.E. Goddard[235]

[1] Animal Breeding and Genetics Group, Georg-August-University Göttingen
[2] Biosciences Research Division, Department of Primary Industries, Victoria
[3] Dairy Future Cooperative Research Centre, Victoria
[4] La Trobe University, Bundoora
[5] Faculty of Land and Food Resources, University of Melbourne




Introduction

- size of reference set ➔ influence on accuracy of genomic prediction
- large reference set ➔ challenging for small breeds
- alternative: multi-breed reference sets

  ➔ requirements: - QTL segregating in all breeds

  - consistent associations across breeds
- results from 50K data: only limited or no increase in accuracy (Hayes et. al., 2009; Pryce et al., 2011)

Introduction

- now: 777K data available (Illumina Bovine High Density (HD) chip)

- Hypothesis 1:

  accuracy of genomic prediction will increase within breed due to a better LD structure

- Hypothesis 2:

  accuracy of genomic prediction will increase for multi-breed references due to more persistent phases across breeds

Data sets

- 2257 Australian Holstein and 540 Australian Jersey bulls
- phenotypes: DTDs for milk yield, fat yield and protein yield
- genotyped for 50K Illumina SNP Chip
  - ➔ after quality control: 39’745 SNPs
- imputed for 777K Illumina SNP Chip using Beagle (Browning & Browning 2009)
  - ➔ after quality control: 624’213 SNPs

Methods

- different methods available:
  - GBLUP: assuming same variance for each SNP
  - Bayes A/B/… : allowing different variances for SNPs

➔ BayesR: SNP effects from different normal distributions which have different variances

- performed well in our datasets ➔ comparable with or in many cases better than GBLUP

BayesR – Model

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{Zu} + \mathbf{Wg} + \mathbf{e}$$

- **u**: vector of polygenic effects ($\mathbf{u} \sim N\left(0, \mathbf{A}\sigma_u^2\right)$)
- **W**: matrix of genotypes
- **g**: vector of SNP effects ($g_i \sim N\left(0, \sigma_{g_i}^2\right)$)

$$\sigma_{g_i}^2 = \begin{cases} 0 \text{ with probability } p_1 \\ 0.0001 \cdot \sigma_a^2 \text{ with probability } p_2 \\ 0.001 \cdot \sigma_a^2 \text{ with probability } p_3 \\ 0.01 \cdot \sigma_a^2 \text{ with probability } p_4 \end{cases}$$

- **GBV of animal j:** $GBV_j = \hat{u}_j + \mathbf{w}_j' \hat{\mathbf{g}}$

BayesR – Model

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{Zu} + \mathbf{Wg} + \mathbf{e}$$

- **u**: vector of polygenic effects ($\mathbf{u} \sim N\left(0, \mathbf{A}\sigma_u^2\right)$)
- **W**: matrix of genotypes
- **g**: vector of SNP effects ($g_i \sim N\left(0, \sigma_{g_i}^2\right)$)

$$\sigma_{g_i}^2 = \begin{cases} 0 \text{ with probability } p_1 \\ 0.0001 \cdot \sigma_a^2 \text{ with probability } p_2 \\ 0.001 \cdot \sigma_a^2 \text{ with probability } p_3 \\ 0.01 \cdot \sigma_a^2 \text{ with probability } p_4 \end{cases}$$

sampled from Dirichlet distribution

- **GBV of animal j:** $GBV_j = \hat{u}_j + \mathbf{w}_j' \hat{\mathbf{g}}$

Prediction scenario

Scenario	Validation	Reference
Holstein	360 youngest bulls	remaining 1897 bulls
Jersey	86 youngest bulls	remaining 454 bulls

Prediction scenario

Scenario	Validation	Reference
Holstein	360 youngest bulls	remaining 1897 bulls
Jersey	86 youngest bulls	remaining 454 bulls

Purebred reference set

Prediction scenario

Scenario	Validation	Reference
Holstein	360 youngest bulls	remaining 1897 bulls
Jersey	86 youngest bulls	remaining 454 bulls
Combined	360 HF + 86 Jersey bulls	1897+454 = 2351 bulls

Multibreed reference set

Correlations within breed

Correlation between GBVs and phenotypes for validation set

Chip	Reference	Validation	Protein	Fat	Milk
50K	Holstein	Holstein	0.55	0.64	0.62
HD	Holstein	Holstein	0.57	0.65	0.63
50K	Jersey	Jersey	0.42	0.48	0.49
HD	Jersey	Jersey	0.41	0.46	0.48

Correlations with multi-breed reference set

Correlation between GBVs and phenotypes for validation set

Chip	Reference	Validation	Protein	Fat	Milk
50K	Holstein	Holstein	0.55	0.64	0.62
50K	Combined	Holstein	0.56	0.65	0.61
HD	Holstein	Holstein	0.57	0.65	0.63
HD	Combined	Holstein	0.57	0.66	0.62

Correlations with multi-breed reference set

Correlation between GBVs and phenotypes for validation set

Chip	Reference	Validation	Protein	Fat	Milk
50K	Holstein	Holstein	0.55	0.64	0.62
50K	Combined	Holstein	0.56	0.65	0.61
HD	Holstein	Holstein	0.57	0.65	0.63
HD	Combined	Holstein	0.57	0.66	0.62

Correlations with multi-breed reference set

Correlation between GBVs and phenotypes for validation set

Chip	Reference	Validation	Protein	Fat	Milk
50K	Jersey	Jersey	0.42	0.48	0.49
50K	Combined	Jersey	0.43	0.49	0.45
HD	Jersey	Jersey	0.41	0.46	0.48
HD	Combined	Jersey	0.46	0.49	0.51

Correlations with multi-breed reference set

Correlation between GBVs and phenotypes for validation set

Chip	Reference	Validation	Protein	Fat	Milk
50K	Jersey	Jersey	0.42	0.48	0.49
50K	Combined	Jersey	0.43	0.49	0.45
HD	Jersey	Jersey	0.41	0.46	0.48
HD	Combined	Jersey	0.46	0.49	0.51

SNPs in distributions

How many SNPs were in the different distributions?

(calculated as mean prop. of SNPs in the distribution x total number of SNPs)

e.g. for protein yield

Distribution (Variance)	Combined, 50K	Combined, HD
1st $(0\sigma_a^2)$	34880	619650
2nd $(0.0001\sigma_a^2)$	4820	4478
3rd $(0.001\sigma_a^2)$	36	77
4th $(0.01\sigma_a^2)$	8	8

Conclusions

using 777K instead of 50K

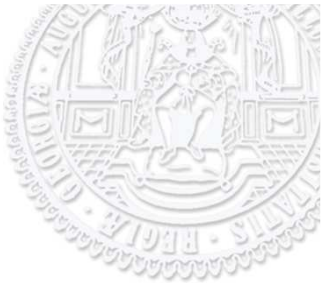
- Hypothesis 1 (accuracy ↑ within breed)?

  ➔ only little support, no significant increase

- Hypothesis 2 (accuracy ↑ in multi-breed situation)?

  ➔ only slight increase in accuracy

Conclusions



using 777K instead of 50K

- Hypothesis 1 (accuracy ↑ within breed)?

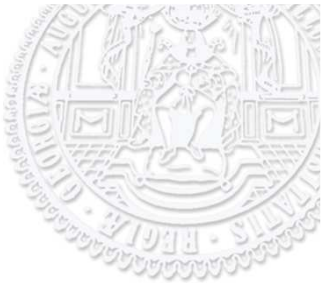
  ➔ only little support, no significant increase

- Hypothesis 2 (accuracy ↑ in multi-breed situation)?

  ➔ only slight increase in accuracy

Why? ➔ low N_e in modern cattle ➔ enough LD even with 50K

Conclusions



using 777K instead of 50K

- Hypothesis 1 (accuracy ↑ within breed)?

  ➔ only little support, no significant increase

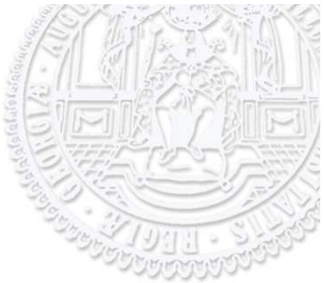
- Hypothesis 2 (accuracy ↑ in multi-breed situation)?

  ➔ only slight increase in accuracy

Why? ➔ low N_e in modern cattle ➔ enough LD even with 50K

➔ breeds not close enough even for HD chip

Conclusions



using 777K instead of 50K

- Hypothesis 1 (accuracy ↑ within breed)?

  ➔ only little support, no significant increase

- Hypothesis 2 (accuracy ↑ in multi-breed situation)?

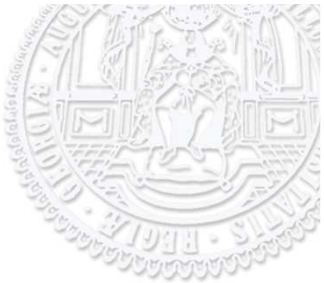
  ➔ only slight increase in accuracy

Why? ➔ low N_e in modern cattle ➔ enough LD even with 50K

➔ breeds not close enough even for HD chip

➔ Jersey data set small ➔ estimation errors, worse imputation accuracy

Conclusions



using 777K instead of 50K

- Hypothesis 1 (accuracy ↑ within breed)?

  ➔ only little support, no significant increase

- Hypothesis 2 (accuracy ↑ in multi-breed situation)?

  ➔ only slight increase in accuracy

Why? ➔ low N_e in modern cattle ➔ enough LD even with 50K

➔ breeds not close enough even for HD chip

➔ Jersey data set small ➔ estimation errors, worse imputation accuracy

➔ still unaccounted genetic variance ➔ MAF

➔ … ???

Acknowledgement

This research was funded by the German Federal Ministry of Education and Research within the AgroClustEr “Synbreed – Synergistic plant and animal breeding” (Funding ID: 0315528C).

Thank you for your attention!