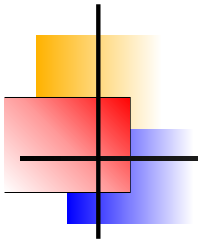
Impact of rare variants on the quality of genomic prediction in dairy cattle

Tomasz Suchocki and Joanna Szyda

Wroclaw University of Environmental and Life Sciences
Department of Genetics
Biostatistics Group



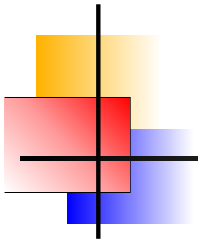
Aim and motivation

Motivation:

- SNP selection based on minor allele frequencies and call rate.
- What with SNPs with very good call rate and very low minor allele frequencies? Will they influence on accuracy of genomic breeding values?

Aim of the study is:

- identification SNP markers with rare allelic variants;
- comparison of accuracy of genomic breeding values for data sets with and without rare allelic variants;
- comparison of Interbull validation procedure for data sets with and without rare allelic variants.



Material

Animals:

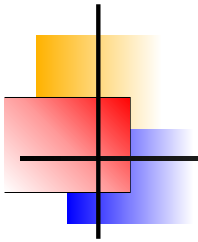
- 5 068 individuals
  - 3 100 proven bulls
  - 1 968 young bulls

Traits:

- Production: milk, fat and protein yield
- Fertility: non-return rate of heifers and cows
- Udder health: somatic cell score

Genotype:

- 46 267 SNPs after selection based on MAF $> 1\%$ and call rate $> 95\%$
- 53 862 SNPs without MAF selection

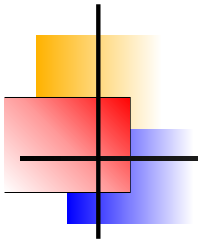


Methods - DGV

SNP effect estimation:

$$y = \mu + Zq + \epsilon,$$

- y - deregressed EBV
- μ - overall mean
- q - random SNP effect $\sim \mathcal{N}(0, I \cdot \frac{\widehat{\sigma}^2_\alpha}{N_{SNP}})$
- $N_{SNP} = 46\ 267$ or $N_{SNP} = 53\ 862$
- $Z = \{-1, 0, 1\}$
- ϵ - error term $\sim \mathcal{N}(0, D \cdot \widehat{\sigma}^2_\epsilon)$



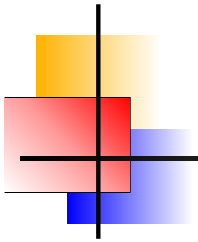
Methods - DGV

SNP effect estimation:

$$y = \mu + Zq + \epsilon,$$

- y - deregressed EBV
- μ - overall mean
- q - random SNP effect $\sim \mathcal{N}(0, I \cdot \frac{\widehat{\sigma}^2_\alpha}{N_{SNP}})$
- $N_{SNP} = 46\ 267$ or $N_{SNP} = 53\ 862$
- $Z = \{-1, 0, 1\}$
- ϵ - error term $\sim \mathcal{N}(0, D \cdot \widehat{\sigma}^2_\epsilon)$

$$DGV = Z \cdot \widehat{q}$$

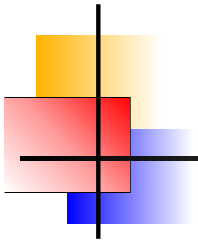


Methods - reliability

Calculation of reliability:

$$Rel = diag\left\{\left(Q - \frac{\widehat{\sigma}^2_\epsilon}{\widehat{\sigma}^2_\alpha} C^{22}\right) Q^{-1}\right\},$$

- C^{22} - inverse of coefficient matrix for MME
- $Q = ZZ^T \frac{1}{p^b_{het}}$
- p^b_{het} - sum over all SNP of heterozygous genotype frequencies in base population



Methods - GEBV validation procedure

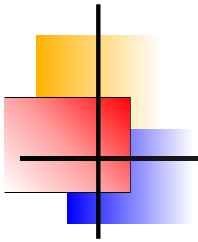
The bias in the national genomic evaluations will be tested using a regression model:

$$\phi_i = b_0 + b_1 GEBV_i + \epsilon_i,$$

where

- ϕ_i is de-regressed predicted genetic merits or daughter deviations from the bulls that have EDC > 20
- *GEBV* - parent averages plus genomic prediction equations

SOURCE: http://interbull.org



Methods - GEBV validation procedure

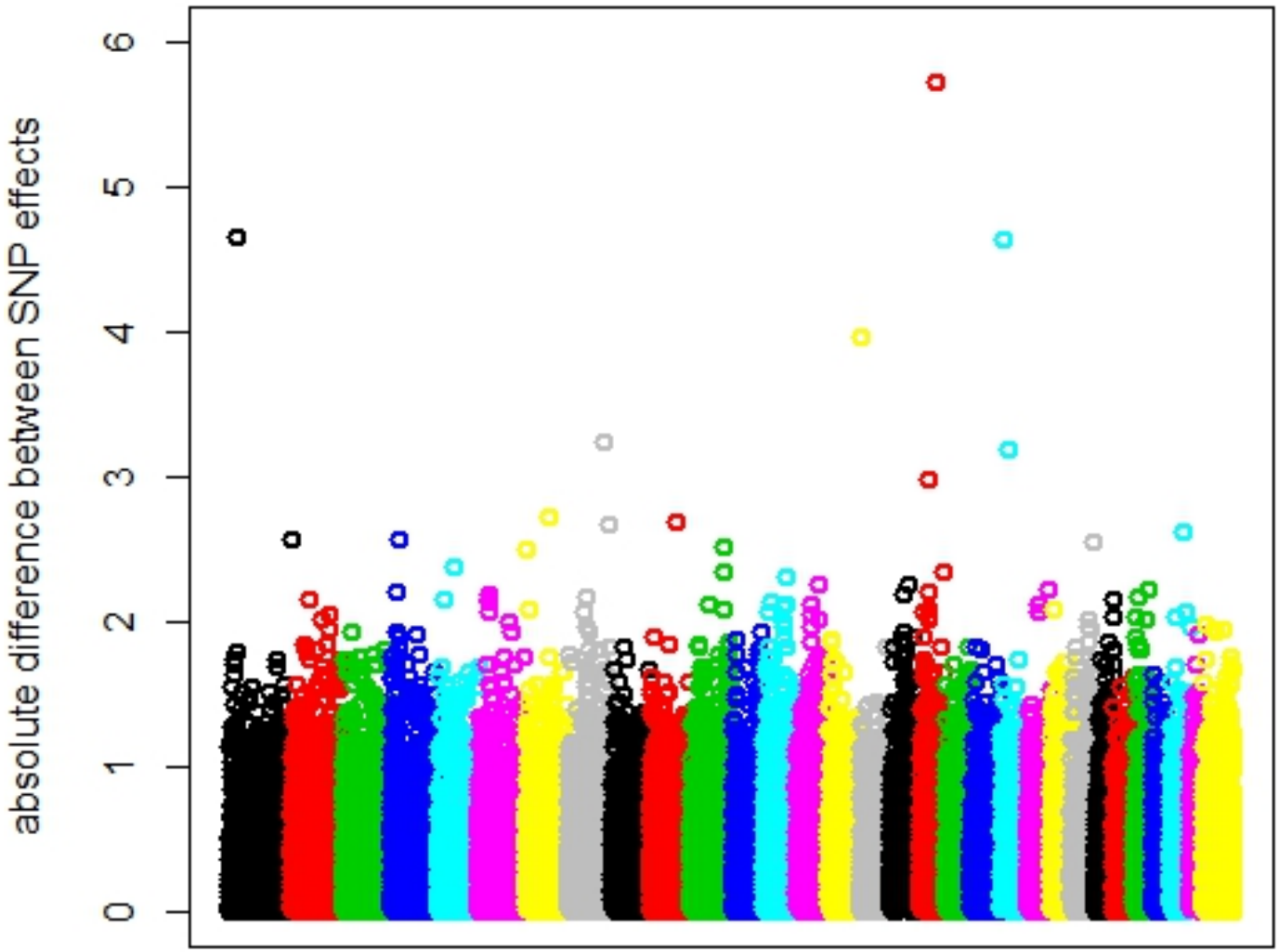
The improvement of the added genomic information to the parental information will be tested using following model:

$$\phi_i = b_0 + b_1 EBV_i + \epsilon_i,$$

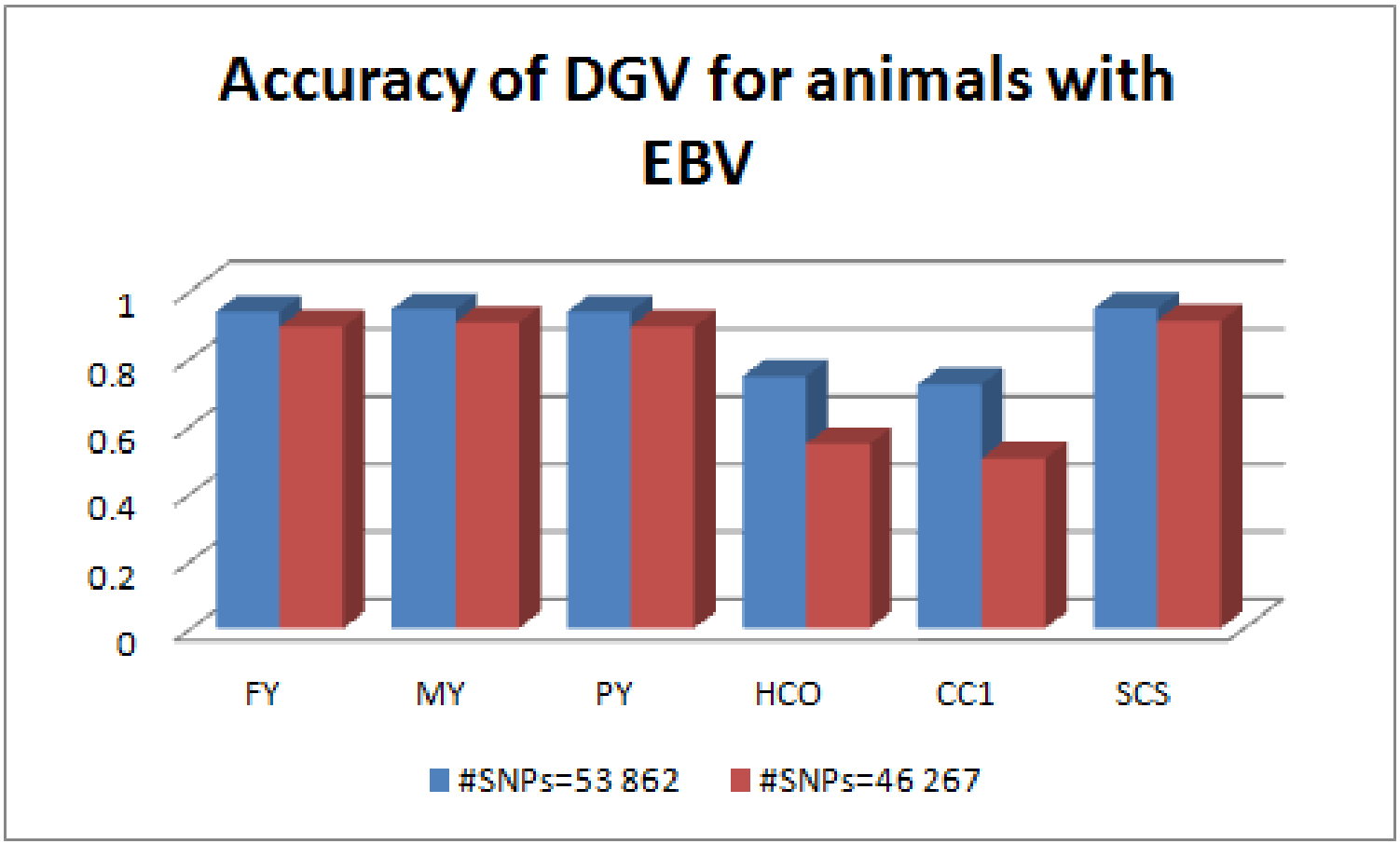
- *EBV* - genetic merit estimates based only on parent averages

SOURCE: http://interbull.org

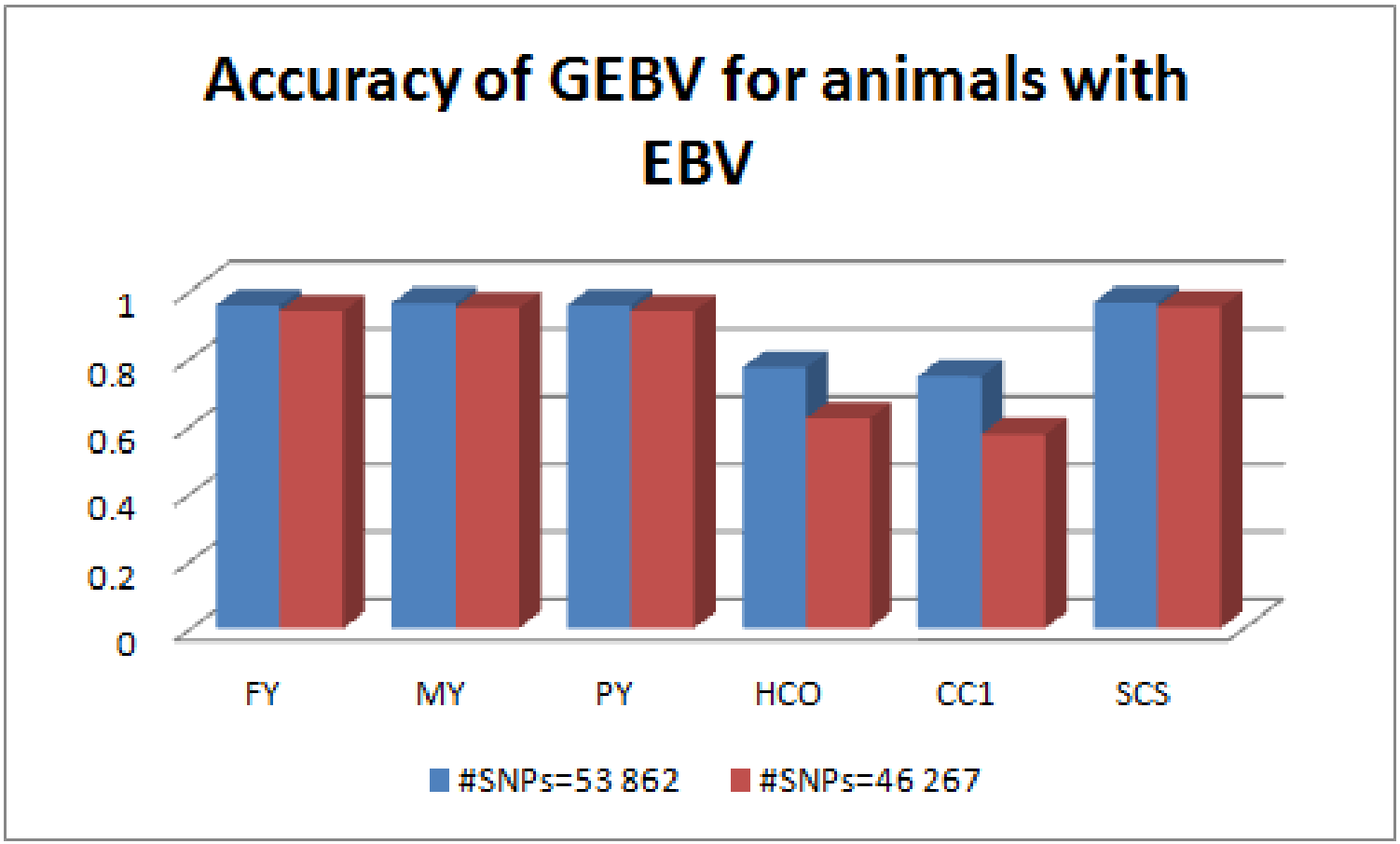
Results - Comparison of common SNPs



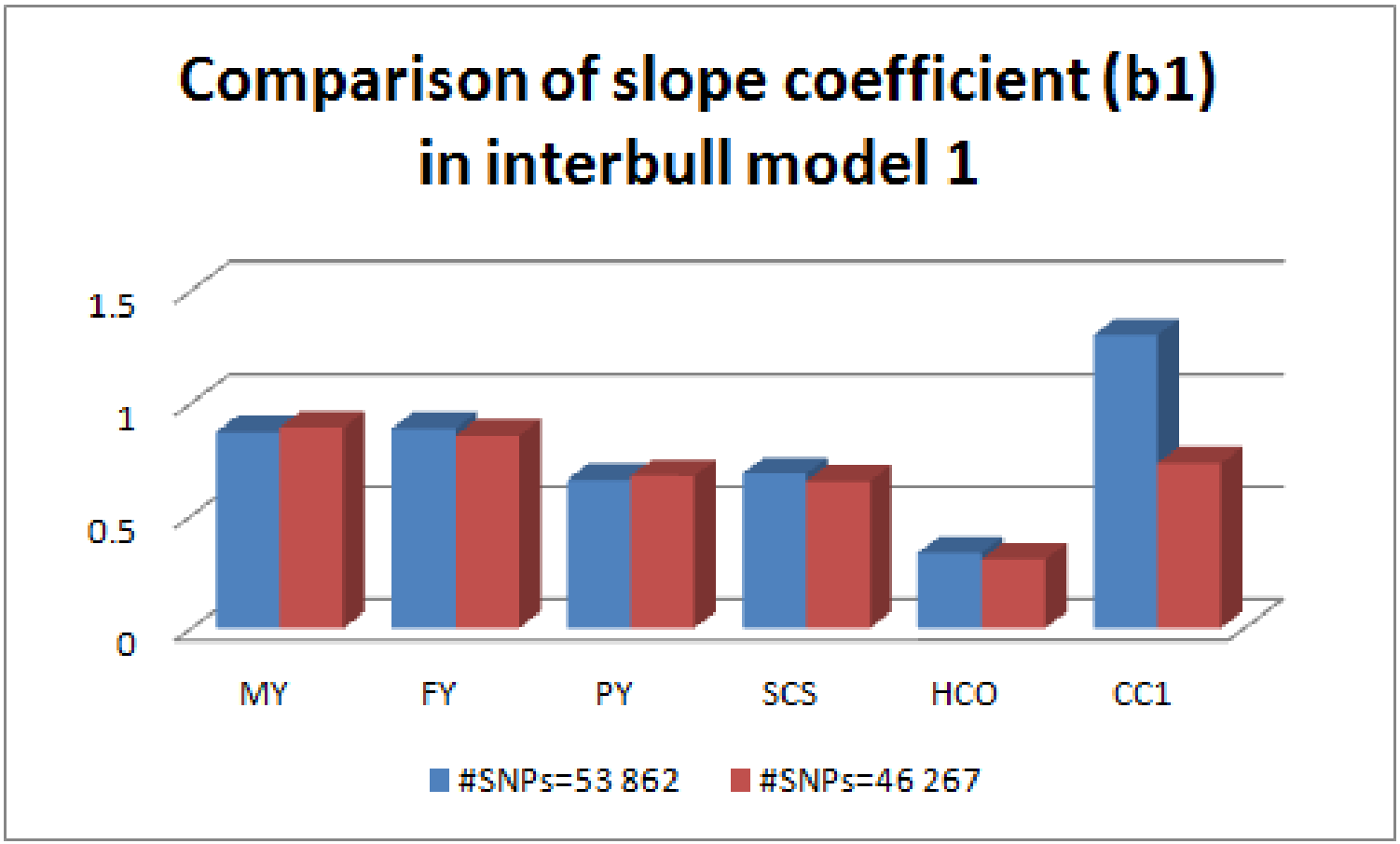
Results - Accuracy



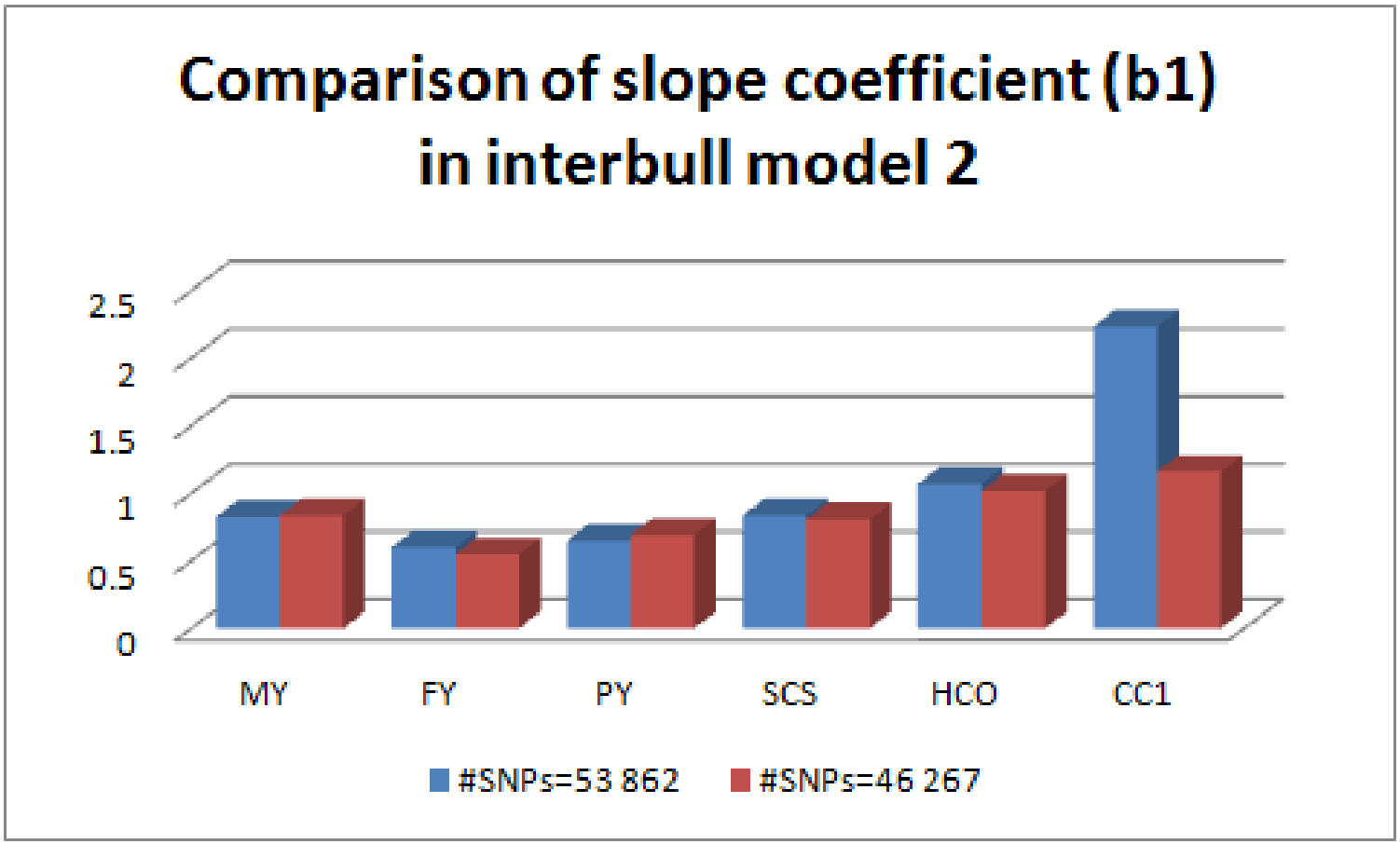
Results - Accuracy



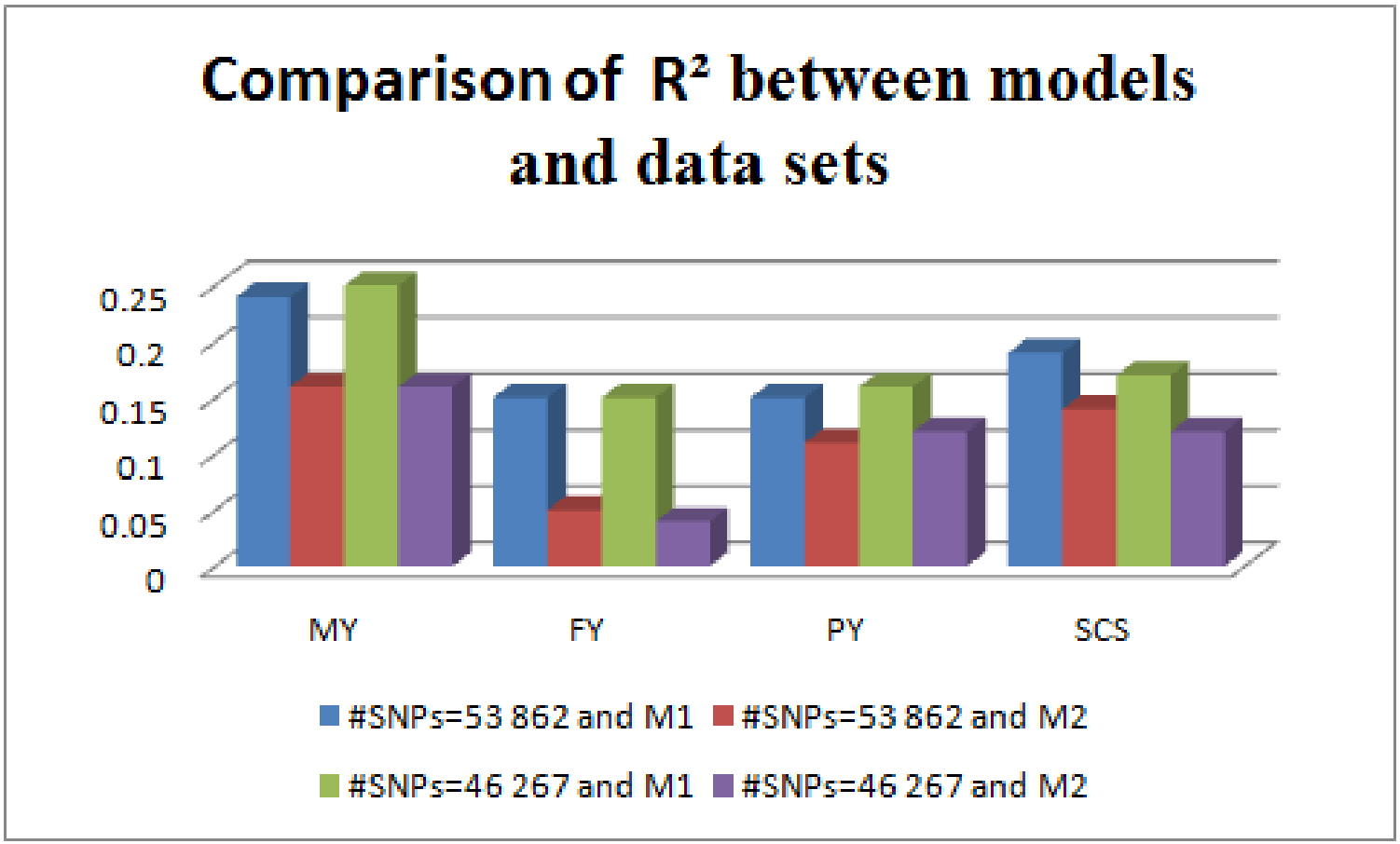
Results - Validation

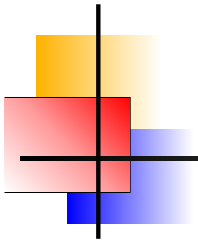


Results - Validation



Results - Validation





Conclusions

- Accuracy of GEBV and DGV is higher for data set with 53 862 SNPs.
- Especially for fertility traits increas of accuracy is high. For CC1 it is 29% for GEBV and 43% for DGV.
- For fat yield, somatic cell score and HCO the parameter b_1 from Interbull validation test is closer to $E(b_1)$ for data set with 53 862 SNPs.
- For other traits better results are for data set with 46 267 SNPs.

Thanks for MASinBULL!

We thank the members of the MASinBULL consortium, who provided the data set used in the analysis.

Thank you for your attention!