Genomic selection in dairy cattle - model comparison for the Polish Holstein-Friesian population

Kacper Żukowski, Joanna Szyda

Institute of Animal Genetics
Wrocław University of Life Sciences

The Aim of Study

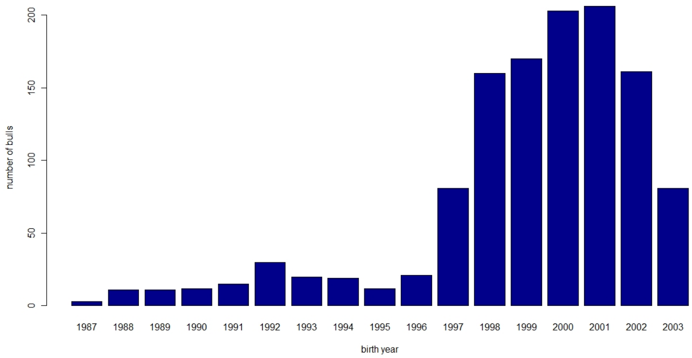
The aim of this study was to use small SNP microarray to estimate the genetic value of Polish Holstein-Friesian bulls based on:

- selection markers to small SNP microarray and
- statistical models.

Animals

- 1216 genotyped Polish Holstein-Friesian bulls;
- 1187 genotyped bulls came from last generation;

Animals



Distribution of genotyped bulls across birth years.

Phenotypic data

Descriptive statistics of EBVs for milk yield, fat yield and protein yield for the reference bulls:

Statistic	Milk yield	Protein yield	Fat yield
minimum	-691.05	-31.84	-40.36
mean	361.82	12.17	9.54
median	360.04	11.96	8.96
maximum	1795.02	54.01	56.19
standard deviation	377.92	10.49	12.91

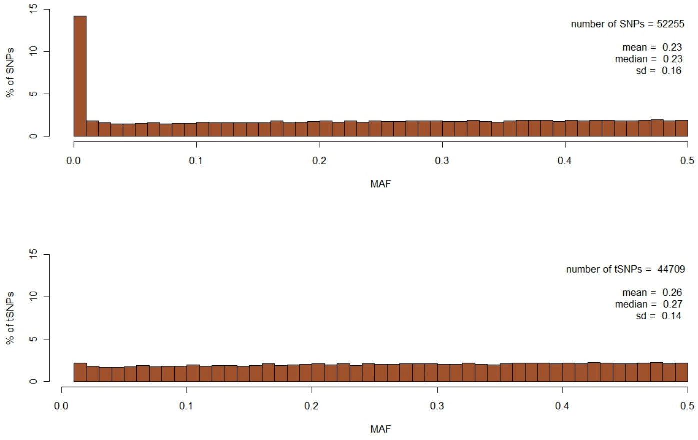
Genotypic data

#SNPs	SNP reduction
54001	Illumina BovineSNP50 BeadChip
−1746	not mapped
−7454	failed frequency test ($MAF > 1\%$)
−135	failed call rate test ($Call\ rate > 90\%$)

Genotypic data

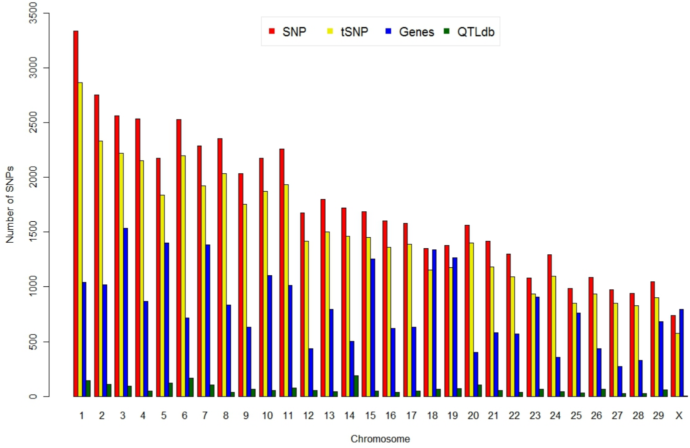
#SNPs	SNP reduction
54001	Illumina BovineSNP50 BeadChip
−1746	not mapped
−7454	failed frequency test ($MAF > 1\%$)
−135	failed call rate test ($Call\ rate > 90\%$)
44709	tSNPs TRAINING DATASET

Animals



Distribution of SNPs across minor allele frequencies (MAF) for selected SNPs (bottom) and all SNPs (top).

Animals



SNP reduction across the genome (release from August 2009).

Genes - genes number from NCBI; QTLdb - QTLs from Cattle QTLdb; http://www.animalgenome.org/cgi-bin/QTLdb/BT/index.

Data Subsets

Subset	#tSNPs	Selection criteria
Subset1	3000	randomly selecting 100 tSNPs from each chromosome
Subset2	2513	uniformly distributed QTLs for each chromosome*
Subset3	2981	uniformly distributed across tSNPs
Subset4	2994	uniformly distributed across base pairs
Subset5	3000	highest estimates for milk yield and $r^2 < 0.8$
Subset6	3000	highest estimates for stature and $r^2 < 0.8$
Subset7	3000	highest estimates for type and $r^2 < 0.8$

*number of QTLs for each chromosome came from QTLdb

Model

Model with fixed SNP effects and a random polygenic effect:

$$Y = \mu + Xb + Z\alpha + \epsilon,$$

where:

- Y – vector of deregressed estimated breeding values;
- μ – the overall mean;
- b – vector of fixed additive SNP effects;
- X – design matrix for SNPs effects the elements given by -1, 0 and 1 for SNP genotype 11, 12(21) and 22;
- α – vector of random additive polygenic effect of genotyped animals, $\alpha \sim \mathcal{N}\left(0, A \cdot \hat{\sigma}^2_\alpha\right)$; $\hat{\sigma}^2_\alpha = 0.33$ for milk yield and 0.29 for protein and fat yield;
- Z – design matrix for random polygenic effect;
- A – polygenic relationship matrix;
- ϵ – vector of random errors, $\epsilon \sim \mathcal{N}\left(0, D \cdot \sigma^2_\epsilon\right)$ and
- D – diagonal matrix with reciprocal of effective daughter contributions (EDC).

Model

Model with a random uncorrelated SNP effects:

$$Y = \mu + Za + \epsilon,$$

where:

- Y – vector of deregressed estimated breeding values;
- μ – the overall mean;
- a – vector of random additive SNP effects assuming $\alpha \sim \mathcal{N}\left(0, I \cdot \frac{\hat{\sigma}_\alpha^2}{\#SNP}\right)$;
  $\hat{\sigma}_\alpha^2 = 0.33$ for milk yield and 0.29 for protein and fat yield;
- Z – design matrix for random SNPs effects the elements given by -1, 0 and 1 for SNP genotype 11, 12(21) and 22;
- I – identity matrix;
- ϵ – vector of random errors, $\epsilon \sim \mathcal{N}\left(0, D \cdot \sigma_\epsilon^2\right)$ and
- D – diagonal matrix with reciprocal of effective daughter contributions (EDC).

Model

Model with a random correlated SNP effects:

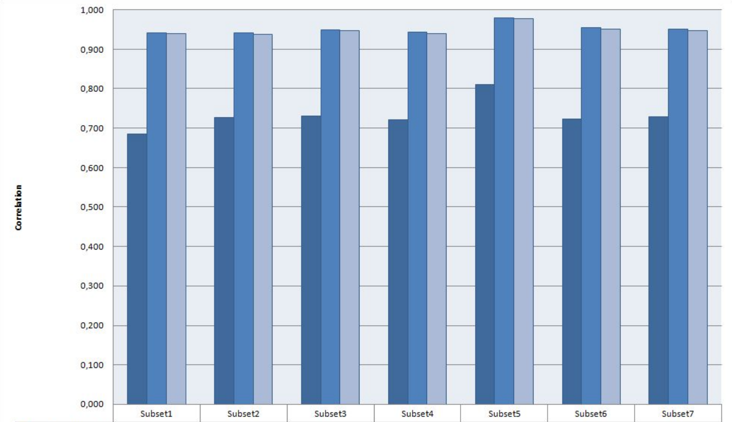
$$Y = \mu + Za + \epsilon,$$

where:

- Y – vector of deregressed estimated breeding values;
- μ – the overall mean;
- a – vector of random additive SNP effects assuming $\alpha \sim \mathcal{N}\left(0, LD \cdot \frac{\hat{\sigma}^2_\alpha}{\#SNP}\right)$;

  $\hat{\sigma}^2_\alpha = 0.33$ for milk yield and 0.29 for protein and fat yield;
- Z – design matrix for random SNPs effects the elements given by -1, 0 and 1 for SNP genotype 11, 12(21) and 22;
- LD – correlation matrix between linked SNPs;
- ϵ – vector of random errors, $\epsilon \sim \mathcal{N}\left(0, D \cdot \sigma^2_\epsilon\right)$ and
- D – diagonal matrix with reciprocal of effective daughter contributions (EDC).

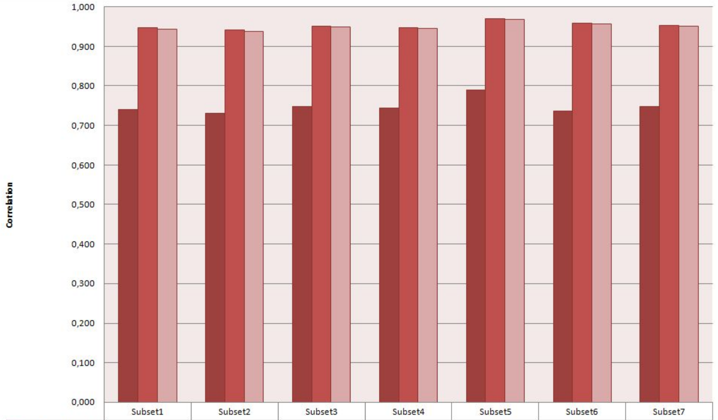
Results, Correlation



	Subset1	Subset2	Subset3	Subset4	Subset5	Subset6	Subset7
f_SNP & r_POLYGEN	0,686	0,726	0,731	0,722	0,811	0,724	0,728
r_SNP(uncorrelated)	0,942	0,942	0,949	0,944	0,979	0,955	0,952
r_SNP(correlated)	0,939	0,938	0,947	0,941	0,978	0,951	0,948

Correlation between EBV and DGV for milk yield.

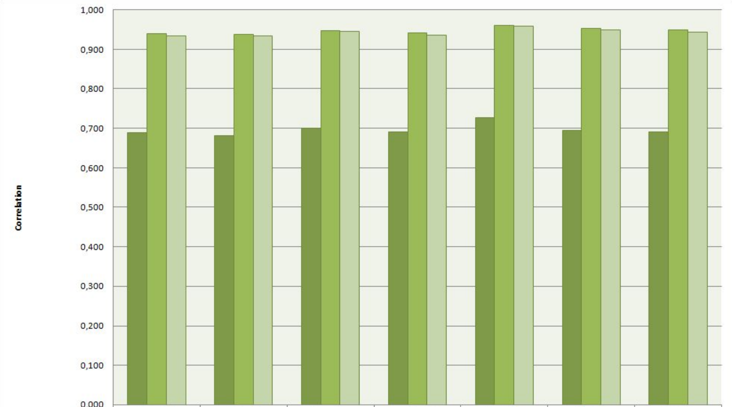
Results, Correlation



	Subset1	Subset2	Subset3	Subset4	Subset5	Subset6	Subset7
f_SNP & r_POLYGEN	0,740	0,731	0,747	0,743	0,790	0,737	0,747
r_SNP(uncorrelated)	0,947	0,941	0,952	0,948	0,970	0,959	0,954
r_SNP(correlated)	0,944	0,939	0,949	0,945	0,968	0,956	0,951

Correlation between EBV and DGV for protein yield.

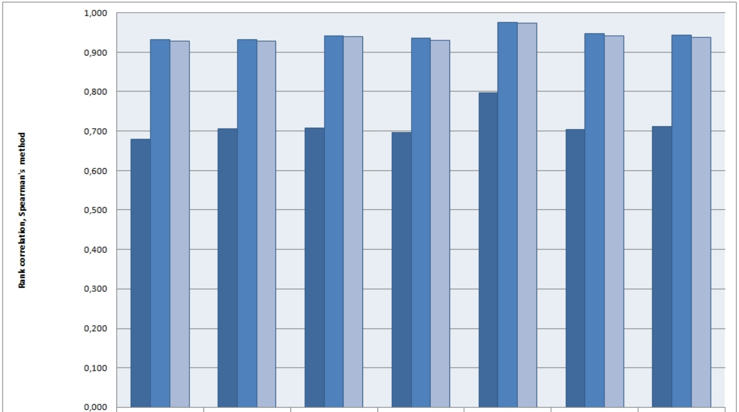
Results, Correlation



	Subset1	Subset2	Subset3	Subset4	Subset5	Subset6	Subset7
f_SNP & r_POLYGEN	0,689	0,682	0,700	0,691	0,727	0,694	0,692
r_SNP(uncorrelated)	0,940	0,937	0,947	0,941	0,962	0,952	0,948
r_SNP(correlated)	0,935	0,934	0,945	0,937	0,958	0,948	0,944

Correlation between EBV and DGV for fat yield.

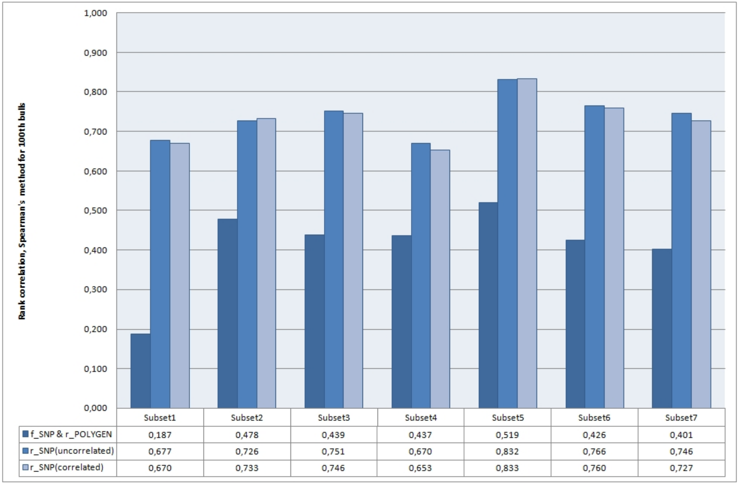
Results, Rank correlation



	Subset1	Subset2	Subset3	Subset4	Subset5	Subset6	Subset7
f_SNP & r_POLYGEN	0,679	0,706	0,709	0,696	0,798	0,703	0,712
r_SNP(uncorrelated)	0,932	0,933	0,942	0,936	0,976	0,947	0,944
r_SNP(correlated)	0,929	0,928	0,939	0,931	0,974	0,943	0,939

Rank correlation between all bulls ranked on highest EBV and DGV for milk yield.

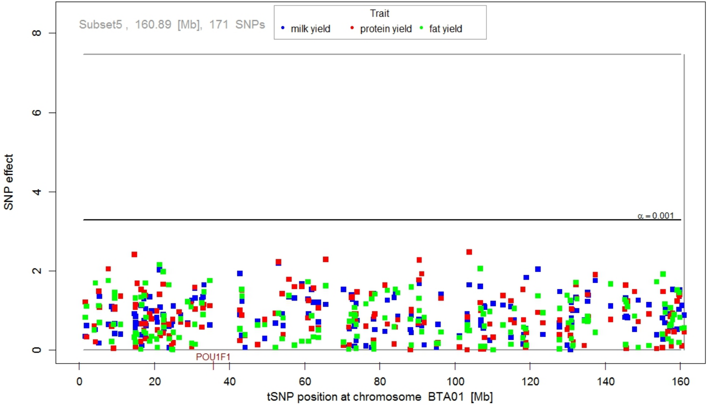
Results, Rank correlation for 100 selected bulls



	Subset1	Subset2	Subset3	Subset4	Subset5	Subset6	Subset7
f_SNP & r_POLYGEN	0,187	0,478	0,439	0,437	0,519	0,426	0,401
r_SNP(uncorrelated)	0,677	0,726	0,751	0,670	0,832	0,766	0,746
r_SNP(correlated)	0,670	0,733	0,746	0,653	0,833	0,760	0,727

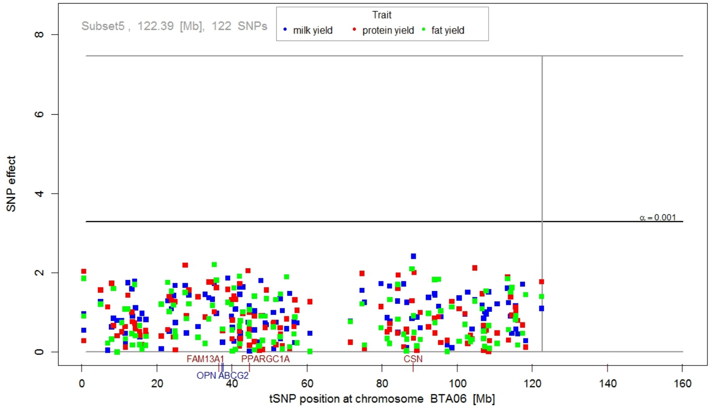
Rank correlation between 100 bulls ranked on highest EBV and DGV for milk yield.

Results, SNP effect across genome



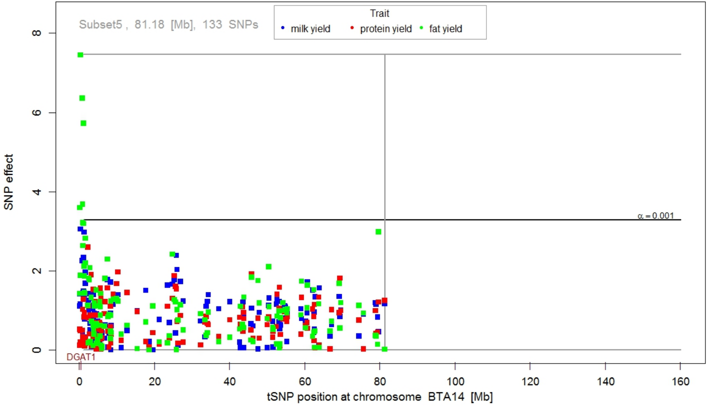
SNP effects for model with uncorrelated SNP effects and Subset5 for all production traits on chromosome BTA01.

Results, SNP effect across genome



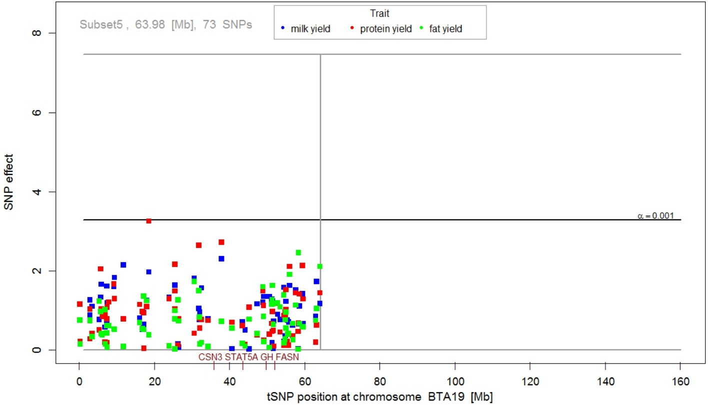
SNP effects for model with uncorrelated SNP effects and Subset5 for all production traits on chromosome BTA06.

Results, SNP effect across genome



SNP effects for model with uncorrelated SNP effects and Subset5 for all production traits on chromosome BTA14.

Results, SNP effect across genome



SNP effects for model with uncorrelated SNP effects and Subset5 for all production traits on chromosome BTA19.

Conclusions

- similar results for model with uncorrelated and correlated random SNP effects;
- highly significant SNP effect connected to DGAT1 position on BTA14;
- highest accuracy was model with random uncorrelated SNP effects and
- for Subset5;
- and also for Subset2 with the smallest number of SNPs.

THANK YOU!