Application of different statisical techniques for the selection of significant SNPs on the 50K chip

Magdalena Frąszczak, Joanna Szyda

Wrocław University of Life and Environmental Sciences, Institute of Genetics, Biostatistics Group

1 Materials and Methods

- dataset: Polish Holstein-Friesian dairy cattle. Genotypes were obtained by the use of Illumina BovineSNP50 Genotyping BeadChip, which consist 54001 SNPs, but to analysis were selected 46267 SNPs with maf>0.01
- considered traits: milk yield (MY)(2509 bulls), fat yield (FY) (2601 bulls) and non return rate for heifers (NRJ) (2504 bulls)
- selection of significant SNPs in based on
  - genomic model (M3) (all of the SNPs fitted simultaneously)
  - a single step model in Asreml with (M1) and without (M2) a random effect

$$(M1)\ \mathbf{y_1} = \mu + \mathbf{X}\beta + \mathbf{Z}\alpha + \epsilon,$$

$$(M2)\ \mathbf{y_2} = \mu + \mathbf{X}\beta + \epsilon,$$

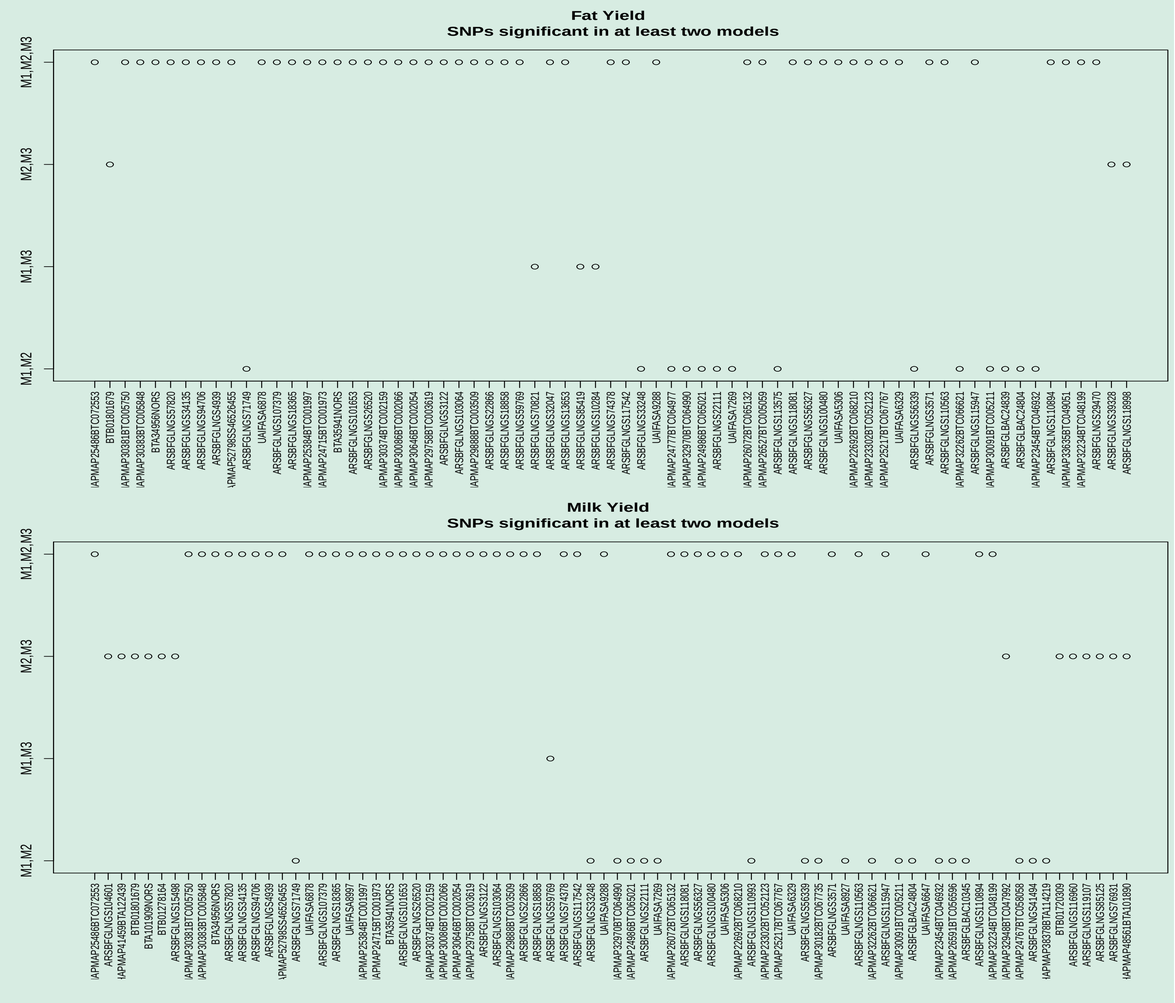
$$(M3)\ \mathbf{y_3} = \mathbf{X_b}\mathbf{b} + \mathbf{Z_g}\mathbf{g} + \epsilon,$$

where:
$\mathbf{y}$ - vector of deregressed breeding values,
μ - an overall mean,
β - vector of the SNP with a design matrix $\mathbf{X} = \{-1, 0, 1\}$ (for homozygous, heterozygous and homozygous SNP genotypes respectively),
α - a random polygenic effect of a bull with an incidence matrix $\mathbf{Z}$, where α is distributed $N(0, (A)\sigma_\alpha^2)$, $\mathbf{A}$ - a relationship matrix, σ_α^2 - an additive polygenic variance,
ϵ - vector of residual effects with distribution $N(0, \mathbf{R}\sigma_\epsilon^2)$, where $\mathbf{R}$ - diagonal matrix weighted by effective daughter contribution for each bull and σ_ϵ^2 -residual variance.

$\mathbf{y_3}$ - vector of deregressed EBV,
$\mathbf{b}$ - vector of the fixed effect with a design matrix $\mathbf{X_b}$
$\mathbf{g}$ - vector of random additive SNP effect, with a design matrix $\mathbf{Z_g} = \{-1, 0, 1\}$ (for homozygous, heterozygous and alternative homozygous SNP genotypes respectively). The covariance matrix of $\mathbf{g}$ is distributed $N(0, \mathbf{I}\frac{\hat{\sigma_a^2}}{N_{SNP}})$, where $\mathbf{I}$ - identity matrix, $\hat{\sigma_a^2}$ - additive genetic variance of given trait, N_{SNP} - number of SNPs
ϵ - vector of residual effects with distribution $N(0, \mathbf{R}\sigma_\epsilon^2)$, where $\mathbf{R}$ - diagonal matrix weighted by effective daughter contribution for each bull and σ_ϵ^2 -residual variance.

2 Results and conclusions

SNPs significant in at least two models



NRJ - 0 significant SNPs in one SNP models, in genomic model - 125 significant SNPs

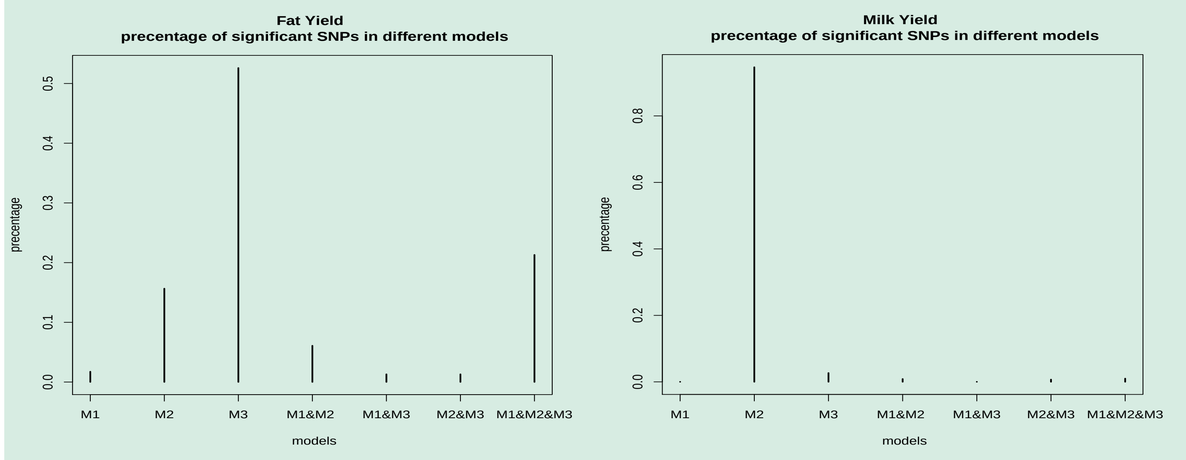
Most of SNPs significant in at least two methods where detected on chromosome 14, as shown in the table below.

	chromosome	0	2	3	5	14	19	20	21	26	28	X
number	FY	1	-	-	1	65	1	1	-	-	-	-
of SNP	MY	1	1	1	4	65	-	-	1	2	1	2

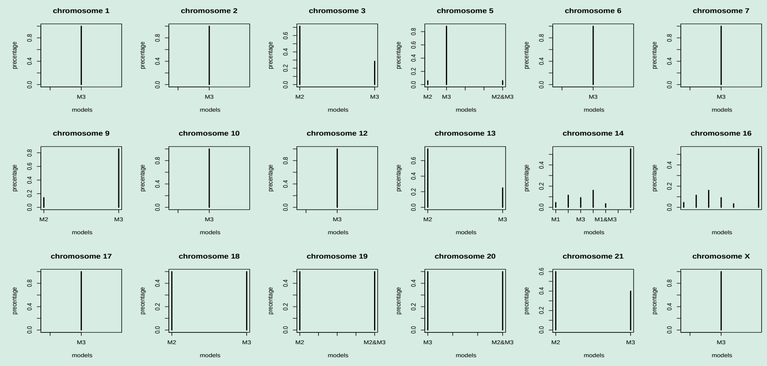
General percentage of significant SNPs depending on chromosome is shown in below table:

chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
FY	0.008	0.039	0.091	0.004	0.073	0.017	0.008	0.004	0.030	0.008	0.004	0.021	0.017	0.378	0.126
MY	0.055	0.053	0.039	0.048	0.041	0.032	0.027	0.059	0.045	0.075	0.033	0.026	0.047	0.068	0.026
chr	16	17	18	19	20	21	22	23	24	25	26	27	28	29	X
FY	0.021	0.004	0.017	0.017	0.008	0.008	0.021	0.004	0.004	0.013	0.017	-	0.013	0.008	0.091
MY	0.015	0.026	0.023	0.026	0.014	0.034	0.024	0.012	0.021	0.026	0.010	0.023	0. 010	0.021	0.020

Graphs presented the percentage of significant SNPs in one, two or three models respectively.



Graphs presented the percentage of significant SNPs in one, two or three models respectively depending on chromosome in case of FY.



- Most of significant SNPs which were detected in only one method are on the separate chromosomes
- LD between SNPs detected on the same chromosome using different methods is < 0.6
- Different SNPs which were detected on the same chromosome in different models are generally in separate factors obtained in PCA (graph below)

Number of significant SNPs in different models depending on factors obtained in PCA for chromosomes 14 and 3 (trait-FY)