Comparison of significant single nucleotide polymorphisms selections in GWAS for complex traits

M. Frąszczak¹ · J. Szyda¹,²

Abstract The goal of this study was to compare significant SNP selection approaches in the context of complex traits based on SNP estimates obtained by models: a model fitting a single SNP (M1), a model fitting a single SNP and a random polygenic effect (M2), the nonparametric CAR score (M3), a SNP-BLUP model with random effects of all SNPs fitted simultaneously (M4). There were 46,267 SNPs tested in a population of 2601 Holstein Friesian bulls, four traits (milk and fat yields, somatic cell score, non-return rate for heifers) were considered. The numbers of SNPs selected as significant differed among models. M1 selected a very large number of SNPs, except for a NRH in which no SNPs were significant. M2 and M3 both selected similar and low number of SNPs for each trait. M4 selected more SNPs than M2 and M3. Considering linkage disequilibrium between SNPs, for MY M2 and M3 selected SNPs more highly correlated with each other than in the case of M4, while for FY M3 selection contained more correlated SNPs than M2 and M4. In conclusion, if the research interest is to identify SNPs not only with strong, but also with moderate effects on a complex trait a multiple–SNP model is recommended. Such models are capable of accounting for at least a part of linkage disequilibrium between SNPs through the design matrix of SNP effects. Functional annotation of SNPs significant in M4 reveals good correspondence between selected polymorphisms and functional information as well as with QTL mapping results.

Keywords Complex traits · GWAS · Mixed model · Significance testing

Introduction

For many years genome-wide association studies (GWAS) have been a useful tool for detecting genetic variants associated with traits in human genetics (Visscher et al. 2012). With the advancement of genotyping technology high-density single nucleotide polymorphism (SNP) platforms have also been developed for other species including livestock. A whole bunch of statistical models have been applied to perform GWAS based on SNP chips. These vary from models where each SNP is considered individually to models with effects of all the available SNPs fitted simultaneously. However, the major drawback of interpreting GWAS results is the optimal selection of polymorphisms which are to be claimed as significantly associated with the analyzed trait and poor repeatability of results across methods and data sets. This problem is especially important when analyzing traits with a complex mode of inheritance (summarized by Visscher et al. 2012).

The variety of GWAS models applied to complex traits covers linear regression, penalized regression approaches with various shrinkage priors for SNP effects, like the least absolute shrinkage and selection operator regression (LASSO) introduced by Tibshirani (1996) (e.g., applied by Wu et al. 2009), the elastic net introduced by Zou and Hastie (2005) (e.g., applied and extended by Do et al. 2011), ridge regression introduced by Hoerl and Kennard (1970) (e.g., applied and extended by Zhan and Xu 2012), normal exponential gamma distribution proposed by Hoggart et al. (2008), or by incorporating the
functional information into SNP selection through biological pathways (Braun and Buetow 2011).

Our study was focused on the comparison of significant SNP selection approaches in the context of complex traits. Significance of SNPs selected based on estimates obtained from a mixed model routinely used in genomic evaluation of dairy cattle in Poland was compared to three single SNP models. In particular, the genomic selection model, which considers all polymorphisms simultaneously by applying shrinkage on SNP estimates through fitting a normal distribution with a predefined variance, is compared with (i) a linear shrinkage on SNP estimates through fitting a normal distribution with large effects, and non-return rate for heifers (NRH) as a nonparametric approach. Generally, we were interested in quantifying the differences between models in defining significant SNPs. Our particular focus though was on assessing the validity of the genomic selection model for GWAS purposes, since such a model is routinely evaluated on large and very informative data sets in many countries and, besides selection purposes, could be a potential source of information on single gene effects on traits undergoing selection.

Materials and methods

Data set

The material of our study consists of 2601 Polish Holstein-Friesian bulls genotyped with the Illumina BovineSNP50 BeadChip, which consists of 54,001 SNPs (version 1) and 54,609 SNPs (version 2). The applied SNP selection criteria comprised polymorphism, expressed by the minor allele frequency (MAF), with the minimum MAF of 0.01, and technical quality of a SNP expressed by the minimum call rate of 90 % within the analyzed sample of bulls. After quality control 46,267 SNPs were selected for further analysis.

Four traits undergoing a complex mode of inheritance were considered in the study: somatic cell score (SCS) representing a trait with “pure” polygenic mode of inheritance, milk (MY) and fat (FY) yields representing traits with a polygenic mode of inheritance enhanced by single genes with large effects, and non-return rate for heifers (NRH) as a trait with a very strong environmental component expressed by heritability of 0.02. Deregressed proofs of bulls were used as pseudophenotypes. Deregression, which was performed in order to remove ancestral information from the conventional breeding values of bulls, was based on a method proposed by Jairath et al. (1998). The corresponding conventional breeding values were estimated based on a random regression test day model for MY, FY, and SCS (Strabel and Jamrozik 2006) and based on a lactation model for NRH (Jagusiak and Żarnecki 2006) using phenotypic information corresponding to the routing national evaluation from April 2012. For each trait 2588 (SCS), 2601 (MY and FY), and 2524 (NRH) records were available.

Models for SNP effect estimation

The SNP effects were estimated by four different models. Single SNP models comprising (M1) \( y = \mu + X\beta + \varepsilon \) and (M2) \( y = \mu + X\beta + Z_1\alpha + \varepsilon \) were solved using the ASReml3 software (Gilmour et al. 2009). In the above models \( y \) represents a vector of deregressed breeding values for MY, FY, SCS or NRH; \( \beta \) denotes a fixed SNP effect with a design matrix \( X \); \( \alpha \) denotes a random polygenic effect of a bull with an incidence matrix \( Z_1 \); \( \varepsilon \) is a vector of residual effects; \( \mu \) denotes an overall mean. The CAR score regression (M3) proposed by Zuber and Strimmer (2011) was selected as the third model for its simplicity and computational efficiency of variable ranking in linear regression based on the Mahalanobis-decorrelation of the explanatory variables representing a nonparametric approach. According to Zuber and Strimmer (2011) this approach is very effective computationally and yields prediction errors as well as true and false positive rates that compare favorably with other regression techniques such as elastic net and boosting. The CAR scores \( \omega_i \), which were considered as the SNP selection criterion, are defined as: \( \omega = \mathbf{P}^{-1}/\mathbf{P}_{bb} \), where \( \mathbf{P}_{bb} \) is the marginal correlation vector between deregressed breeding values and SNPs. \( \mathbf{P} \) denotes shrinkage estimator given by: \( \lambda \mathbf{P}_{bb} + (1-\lambda)\mathbf{R}_{\text{empirical}} \), where \( \lambda \) is a shrinkage intensity and \( \mathbf{R}_{\text{empirical}} \) is the empirical non-regularized correlation matrix among SNP genotypes. The CAR criterion was computed using an R package CARE. The genomic selection model (M4) \( y = \mu + Z_g + \varepsilon \), equivalent to the so called SNP-BLUP model, is routinely used for the prediction of direct genomic breeding values for the Polish Holstein-Friesian dairy cattle. The residual effects vector \( \varepsilon \) is distributed as \( N(0, \mathbf{D}\sigma^2_{\varepsilon}) \) where \( \mathbf{D} \) is a diagonal matrix weighted by the effective daughter contribution for each bull and \( \sigma^2_{\varepsilon} \) represents the estimate of total additive genetic variance of a given trait calculated elsewhere for the whole active population of Polish Holstein-Friesian dairy cattle. The residual effects vector \( \varepsilon \) is distributed as \( N(0, \mathbf{D}\sigma^2_{\varepsilon}) \) where \( \mathbf{D} \) is a diagonal matrix weighted by the effective daughter contribution for each bull and \( \sigma^2_{\varepsilon} \) denotes a residual variance; \( \mu \) denotes an overall mean. The CAR score regression (M3) proposed by Zuber and Strimmer (2011) was selected as the third model for its simplicity and computational efficiency of variable ranking in linear regression based on the Mahalanobis-decorrelation of the explanatory variables representing a nonparametric approach. According to Zuber and Strimmer (2011) this approach is very effective computationally and yields prediction errors as well as true and false positive rates that compare favorably with other regression techniques such as elastic net and boosting. The CAR scores \( \omega_i \), which were considered as the SNP selection criterion, are defined as: \( \omega = \mathbf{P}^{-1}/\mathbf{P}_{bb} \), where \( \mathbf{P}_{bb} \) is the marginal correlation vector between deregressed breeding values and SNPs. \( \mathbf{P} \) denotes shrinkage estimator given by: \( \lambda \mathbf{P}_{bb} + (1-\lambda)\mathbf{R}_{\text{empirical}} \), where \( \lambda \) is a shrinkage intensity and \( \mathbf{R}_{\text{empirical}} \) is the empirical non-regularized correlation matrix among SNP genotypes. The CAR criterion was computed using an R package CARE. The genomic selection model (M4) \( y = \mu + Z_g + \varepsilon \), equivalent to the so called SNP-BLUP model, is routinely used for the prediction of direct genomic breeding values for the Polish Holstein-Friesian population (Szyda et al. 2011). Here \( Z_g \) is a design matrix for SNP genotypes, which is parameterized as \( \mathbf{I}_g = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \), \( \mathbf{Z} \) is a vector of random additive SNP effects distributed defined as: \( \mathbf{N} \left( 0, \mathbf{I}_{46267} \right) \), with \( \mathbf{I} \) being an identity matrix.
The SNP effects in M1 were estimated by weighted least squares with the effective number of daughters corresponding to teach observation \( y \) used as a weighting variable. To estimate \( \beta \) in M2 the objective function \( \log f(y | \alpha, \beta, G) + \log f_\kappa(\alpha, G) \), where \( G = \alpha \sigma^2_G \), was used. Differentiating it with respect to \( \beta \) and \( \alpha \) leads to the mixed model equations (Henderson 1984). M3 is a model free procedure in which CAR scores are functions of empirical SNP-pseudophenotype and SNP-SNP correlations. The estimation of parameters of M4 was based on solving the corresponding mixed model equations using the iteration on data technique applying the Gauss-Seidel algorithm with residuals update (Legarra and Misztal 2008).

### Significant SNP selection

In case of M1, M2, and M4 the Wald test was used to obtain the nominal type I error corresponding to a standard normal distribution. For single SNP models M1 and M2 the resulting \( P \) values were subjected to multiple testing correction for the number of SNP tested \( (N=46,267) \) via Bonferroni’s approach, while for a multiple SNP model M4 a nominal \( P \) value was used as a selection criterion. In M3 the null distribution of the function of CAR scores \( t = \omega \sqrt{\frac{1}{\omega - 1}} - \omega - \omega^2 \) was used for obtaining type I error rates, which follows the Student \( t \) distribution with \( (\omega - 1) \) degrees of freedom estimated by the R package \texttt{fdrtool} (Strimmer 2008). For all models SNPs were selected as significant when \( P \) values associated with their estimates did not exceed the 0.001 threshold.

Genomic annotation of SNPs was performed using SNPchiMp (Nicolazzi et al. 2014) for the identification of SNP positions corresponding to the UMD3.1 bovine genome assembly and Variant Effect Predictor (McLaren et al. 2010) for the identification of genomic positions of SNPs.

### Results

#### The number of significant SNPs selected

The numbers of SNPs selected as significant by different models are presented in Table 1. For MY, FY, and SCS the simplest model M1 always selected a very large number of SNPs ranging between 2242 (SCS) and 3398 (MY), widely exceeding the number of SNPs selected by M2-M4. Although models 2 and 3 markedly differ in modeling and hypothesis testing assumptions, they both select a very similar and low number of SNPs for each trait. The genomic selection model M4 is intermediate in terms of the number of SNPs selected.

Except for M4, traits with different putative inheritance modes resulted in different numbers of significant SNPs. For MY and FY the largest numbers of SNPs were selected at BTA14 — a chromosome harboring DGAT1 gene of high effect on milk production traits. For SCS, a trait with a pure polygenic mode of inheritance, a lower number of SNPs was indicated as significant, with only four SNPs selected by M2. For NRH, a trait with a very weak genetic component (as compared to the environmental based variation) no SNPs were identified by M1-M3. A very different pattern was observed for M4. Since all polymorphisms’ effects are estimated simultaneously with the underlying normal distribution shrinkage, the SNP selection procedure based on the 0.001 threshold chooses 0.1 % of the most significant SNPs regardless of a trait. As a consequence a very similar number of polymorphisms varying between 125 (NRH) and 182 (FY) for each trait was selected. Since, based on the very large number of SNPs selected, M1 does not seem to be a valid model for GWAS on complex traits, in further result description and discussion we confine ourselves to M2-M4.

#### Influence of SNP informativeness on SNP selection

No marked differences between models were observed regarding the information content, expressed by MAF, between SNPs selected as significant by M2–M4. The average MAF of polymorphisms selected for MY by M2 amounted to 0.39, by M3 to 0.30 and M4 to 0.36 and was even more similar for FY with 0.37 (M2), 0.38 (M3) and 0.49 (M4). The average MAF in group of SNPs common between all considering models is 0.40 for MY and 0.38 for FY. On the other hand, considering MY, on average M2 and M3 selected SNPs which were more highly correlated with each other than was the case for M4 since the average LD expressed by pairwise correlation between significant SNPs was equal to 0.18 and 0.17 for M2 and M3, but only 0.08 for the genomic selection model. For FY, the average LD of 0.11 among SNPs selected by M4 was also lower from the corresponding values form M2 and M3 which were equal to 0.17 and 0.21 respectively. The highest average pairwise correlation was observed between significant SNPs which were common for models M2 and M3.

#### Comparison of significant SNP sets between models

Figure 1 presents the percentage of SNPs significant between M2 and M4. For production traits a relatively large proportion of significant SNPs was common among all three models making 24 SNPs common for MY and 41 SNPs...
common for FY, while no common for all models polymorphisms were identified for SCS and NRH. Focusing on the genomic selection model M4 as the reference it is evident from Fig. 2 that more mutual SNPs existed between M4 and M2 than between M4 and M3. In particular, in addition to the polymorphisms common to all three models, the genomic selection model had 12 and 11 additional SNPs in common with M2 for MY and FY respectively, but only one SNP for MY mutual with M3. For MY there are four polymorphisms common for models M2 and M3.

**Genomic SNP annotation**

Genomic annotation was carried out for SNPs which were selected as significant by all models, i.e., M2, M3, and M3 (Table 2). Since M3 indicated no significant polymorphisms for SCS and NRH for those traits no annotation was considered. For FY 22 genes were marked by significant SNPs common between models, while for MY a lower number of eight genes was identified with six of them also significant for FY. All genes are located on BTA14 with varying distances from DGAT1 — a well known candidate gene for milk production traits (Grisart et al. 2002), ranging between 104,646 and 3,768,523 bp (calculated as a distance between corresponding significant SNP positions) (Fig. 3). Most of the SNPs are located within genomic range of the gene, mainly within intronic regions, but rs110323635 and rs41256919 lie in exons of MAPK15 and MAF1 genes respectively.

Although some of the genes may represent spurious associations arising through high LD to DGAT1, those which are located in more distant regions of BTA14 are potential
candidate genes. Special attention is to be focused on EEF1D which according to a recent study of Xie et al. (2014) in dairy cattle exhibits significantly higher expression in mammary gland as in other tissues.

Discussion

We observed that for complex traits approaches which model the genetic component by a single SNP — here represented by M1, which are very common and successful in GWAS for disease traits, have methodological problems. Since such models do not account for the polygenic background differentiating between SNPs in high LD, especially in the neighborhood of genes with strong effects is problematic and the multiple testing correction is not sufficient to remove the SNP effect upward bias due to correlated SNPs. An intensive post-processing of SNP estimates and/or corresponding P values regarding LD is then required.

Table 2  Annotation of SNPs significant in M2, M3, and M4. Genes significant for both MY and FY are marked in bold. For significant genes all SNPs located within intron/exon regions are listed

| Gene acronym | Gene name                                      | Ensemble ID         | BTA | SNP                | SNP annotation       | SNP position |
|--------------|-----------------------------------------------|---------------------|-----|--------------------|----------------------|--------------|
| Fat yield    |                                               |                     |     |                    |                      |              |
| FOXH1        | Forkhead box H1                               | ENSBTAG000000004761| 14  | rs109146371        | 3390 bp upstream     | 1,651,311    |
| CYHR1        | Cysteine/histidine-rich 1                     | ENSBTAG000000035254| 14  | rs109968515        | intron               | 1,675,278    |
| VPS28        | Homologous to vacuolar protein sorting 28     | ENSBTAG000000026320| 14  | rs17870736         | intron               | 1,696,470    |
| DGAT1        | Diacylglycerol O-acyltransferase 1            | ENSBTAG000000026356| 14  | rs109421300        | intron               | 1,801,116    |
| MAF1         | Homologous to RNA polymerase III-inhibiting protein | ENSBTAG00000012242| 14  | rs142656919        | exon – missense variant | 1,923,292    |
| SPATC1       | Spermatogenesis and centriole associated 1    | ENSBTAG000000026350| 14  | rs14629750         | 4472 bp upstream     | 2,002,873    |
| PLEC         | Plectin                                       | ENSBTAG00000011922 | 14  | rs109350371        | intron               | 2,054,457    |
| MAPK15       | Mitogen-activated protein kinase 15           | ENSBTAG000000019864| 14  | rs110323635        | exon – missense variant | 2,239,085    |
| EEF1D        | Elongation factor 1-delta                     | ENSBTAG00000014643 | 14  | rs109661298        | intron               | 2,319,504    |
| ZC3H3        | Zinc finger CCCH-type containing 3            | ENSBTAG00000021472 | 14  | rs109617015        | intron               | 2,386,688    |
| RHPN1        | Rhophilin, Rho GTPase binding protein 1        | ENSBTAG00000002104 | 14  | rs109529219        | intron               | 2,468,020    |
|              | Novel gene                                    | ENSBTAG00000003606 | 14  | rs110199901        | intron               | 2,524,432    |
| LY6K         | Lymphocyte antigen 6 complex, locus K         | ENSBTAG00000000158 | 14  | rs110174651        | 297 bp downstream     | 2,754,909    |
| LY6D         | Lymphocyte antigen 6 complex, locus D         | ENSBTAG000000034498| 14  | rs110237430        | 978 bp downstream     | 2,803,998    |
| LYPD2        | LY6/PLAUR domain containing 2                | ENSBTAG000000016210| 14  | rs109476486        | 2653 bp upstream     | 2,826,632    |
| BAI1         | Brain-specific angiogenesis inhibitor 1       | ENSBTAG00000006385 | 14  | rs109545018        | intron               | 3,006,509    |
| GPR20        | G protein-coupled receptor 20                | ENSBTAG000000015985| 14  | rs110411273        | 71 bp downstream      | 3,640,788    |
| TSNARE1      | t-SNARE domain containing 1                  | ENSBTAG00000009974 | 14  | rs109875744        | intron               | 3,078,843    |
|              |                                               |                      |     | rs110888717        | intron               | 3,117,493    |
|              |                                               |                      |     | rs109295487        | intron               | 3,137,184    |
| PTK2         | PTK2 protein tyrosine kinase 2                | ENSBTAG00000009578  | 14  | rs109470279        | 87 bp downstream      | 3,885,798    |
|              |                                               |                      |     | rs41624797         | intron               | 3,956,956    |
|              |                                               |                      |     | rs109131748        | intron               | 4,017,201    |
|              |                                               |                      |     | rs110183545        | intron               | 4,043,743    |
| EIF2C2       | Argonaute RISC catalytic component 2         | ENSBTAG00000001579  | 14  | rs109948273        | intron               | 4,103,850    |
|              |                                               |                      |     | rs41576704         | intron               | 4,127,413    |
|              |                                               |                      |     | rs108980964        | intron               | 4,149,375    |
| TRAPPC9      | Trafficking protein particle complex 9        | ENSBTAG00000013955  | 14  | rs109807697        | intron               | 4,240,287    |
|              |                                               |                      |     | rs109248069        | intron               | 4,267,053    |
|              |                                               |                      |     | rs11018678         | intron               | 4,336,714    |
|              |                                               |                      |     | rs110017379        | intron               | 4,364,952    |
|              |                                               |                      |     | rs55617160         | intron               | 4,468,478    |
|              |                                               |                      |     | rs110805364        | intron               | 4,583,344    |
|              |                                               |                      |     | rs110805364        | intron               | 4,583,344    |
| COL22A1      | Collagen, type XXII, alpha 1                  | ENSBTAG000000015374| 14  | rs110444021        | intron               | 5,225,467    |
| Milk yield   |                                               |                      |     | rs110351374        | intron               | 5,428,037    |
| C8orf33      | Chromosome 14 open reading frame 33           | ENSBTAG00000000879  | 14  | rs109752439        | 87 bp downstream      | 1,489,496    |
| FAM135B      | Family with sequence similarity 135, member B | ENSBTAG000000018218| 14  | rs11062450         | intron               | 5,428,037    |
|              |                                               |                      |     | rs110918650        | intron               | 5,462,752    |
|              |                                               |                      |     | rs110501942        | intron               | 5,494,654    |
|              |                                               |                      |     | rs109402117        | intron               | 5,569,639    |
A possible alternative, which is non-intensive computationally, is to use a single SNP model with the genetic background described by an additive polygenic effect (like M2) or a multiple–SNP without any assumptions regarding the inheritance mode (like M3). In our study M2 and M3 for MY and for FY, above and beyond the M2-M3, also the genomic model (M4) selected a similar number of SNPs as the number 55 resulting from the estimator proposed by Hayes and Goddard (2001) of the total number of genes influencing the variance of production traits (M): \( M = N \ln \left( \frac{1}{P} \right) \), where \( P = 1/2N_c \). Note, that here we used the estimated number of heterozygous QTL segregating for production traits as \( N = 10.73 \) (Hayes and Goddard 2001) and the effective population size for dairy cattle as \( N_c = 103 \) (Qanbari et al. 2010).

However, if the research interest is to identify SNPs not only with strong, but also with moderate effects on a complex trait a multiple–SNP model – here represented by M4, is recommended. Such models are capable of accounting for at least a part of LD between SNPs through the design matrix of SNP effects. A fine tuning of SNP selection procedure is needed based on the expected number of SNPs influencing the variability of a given trait, e.g., based on heritability estimates.

Recently, Wang et al. (2012) applied a genomic selection model (single step GBLUP) for GWAS on complex traits based on simulated data and found a good accuracy of prediction of QTL effects through SNPs. As indicated in that study it is important to realize that the selected SNPs do not necessarily represent underlying genes and that a chromosomal region in high LD (which is usually, but not always, equivalent to a region in physical neighborhood of the SNP) should be considered in search for causal mutations. Comparing the average LD of SNPs selected by single gene methods with the set selected by the genomic selection model it is evident that the latter is able to better deal with correlations between particular SNPs which occur through LD. Moreover, as pointed out by Dekkers (2012) a genomic selection model directly accounts for the population structure, not only on an averaged genomewide level (through the additive polygenic covariance, as incorporated into M2), but also at a particular SNP sites (through the design matrix of SNP genotypes, as incorporated into M4). Another very important, practical advantage of using the genomic selection model for GWAS is that in most countries the model and underlying SNP effects are anyway evaluated on large, informative data sets.

**Acknowledgments**  The study was financed by the Polish State Committee for Scientific Research, grant no. N N311 609639. The data set was generated within the framework of the MASinBULL project (currently Genomika Polska).

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

**References**

Braun R, Buetow K (2011) Pathways of distinction analysis: a new technique for multi–SNP analysis of GWAS data. PLoS Genet 7:6. doi: 10.1371/journal.pgen.1002101

Dekkers JCM (2012) Application of genomics tools to animal breeding. Curr Genomics 13:207–212

Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW, Wojcicki A, Eriksson N (2011) Web-based genome-wide association study identifies two Novel Loci and a substantial genetic component for Parkinson’s Disease. PLoS Genet 7:6. doi:10.1371/journal.pgen.1002141

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead

Grilsart B, Coppoeters W, Farin F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, Spelman R, Georges M, Snell R (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res 12:222–231

Hayes B, Goddard ME (2001) The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol 33:209–229

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph

Hoerl AE, Kennard R (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67

Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (2008) Simultaneous analysis of All SNPs in genome-wide and re-sequencing association studies. PLoS Genet 4:7. doi:10.1371/journal.pgen.1000130

Jagusiak W, Żamecki A (2006) Genetic evaluation for fertility traits in Polish Holsteins. Interbull Bulletin 35:37–41

Jairath L, Dekkers JCM, Schaeffer LR, Liu Z, Burnside EB, Kolstad B (1998) Genetic evaluation for herd life in Canada. J Dairy Sci 81:550–562

Legarra A, Misztal I (2008) Technical note: computing strategies in genomewide selection. J Dairy Sci 91:360–366

McLaren W, Pritchard B, Rios D, Chen Y , Flicek P, Cunningham F (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069–2070
Nicolazzi EL, Picciolini M, Strozzi F, Schnabel RD, Lawley C, Pirani A, Brew F, Stella A (2014) SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics 15:123
Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H (2010) The pattern of linkage disequilibrium in German Holstein cattle. Anim Genet 41:346–356
Strabel T, Jamrozik J (2006) Genetic analysis of milk production traits of Polish black and white cattle using large-scale random regression test-day models. J Dairy Sci 89:3152–3163
Strimmer K (2008) fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics 24:1461–1462
Szyda J, Żarnecki A, Suchocki T, Kamiński S (2011) Fitting and validating the genomic evaluation model to Polish Holstein-Friesian cattle. J Appl Genet 52:363–366
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Statist Soc B 58:267–288
Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24
Wang H, Misztal I, Aguilar I, Legarra A, Muir WM (2012) Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res 94:73–83
Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25:714–721. doi:10.1016/j.gene.2013.10.061
Xie Y, Yang S, Cui X, Jiang L, Zhang S, Zhang Q, Zhang Y, Sun D (2014) Identification and expression pattern of two novel alternative splicing variants of EEF1D gene of dairy cattle. Gene 534:189–196
Zhan H, Xu S (2012) Adaptive ridge regression for rare variant detection. PLoS ONE 7(8):e44173. doi:10.1371/journal.pone.0044173
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Statist Soc B 67:301–320
Zuber V, Strimmer K (2011) High-dimensional regression and variable selection using CAR scores. Statist Appl Genet Mol Biol 10:34