399. LASSO and SVM: an alternative approach to identify associated genome regions for simple and complex traits in cattle

S. Widmer[1*], F.R. Seefried[2], C. Drögemüller[1] and P. von Rohr[2]

[1]Institute of Genetics, University of Bern, Bremgartenstrasse 109a, 3012 Bern, Switzerland;
[2]Qualitas AG, Chamerstrasse 56, 6300 Zug, Switzerland; sarah.widmer@vetsuisse.unibe.ch

Abstract

Variable selection via the LASSO regression analysis followed by a classification of phenotypes into two groups of cases and controls using SVM was performed to identify genomic regions associated with phenotypic traits of interest. This approach was successfully validated by identifying previously known genomic associations for two Mendelian traits and a complex polygenic trait in Swiss cattle populations. Furthermore, new associated regions for the trait multiple birth were found in Holstein. Therefore, this approach proved to be a valuable alternative to identify associated genomic regions for inherited traits in livestock.

Introduction

Support Vector Machine (SVM) is a general-purpose classification method (Cortes and Vapnik 1995) and is used in many fields (James *et al.* 2021). For example, it was able to find SNPs associated with the risk for type 2 diabetes in humans and to perform genotype-based predictions of the affection status (Ban *et al.* 2010). SVM classifies data consisting of different groups by separating hyperplanes. The separating hyperplanes are defined by explanatory variables such as SNP genotype effects. This definition makes SVM suitable to analyse data that can be divided into two groups, e.g.: cases and controls. Although SVM can be used for high-dimensional data, it is important to first identify the subset of relevant explanatory variables. It is necessary to select relevant SNPs to avoid overfitting which might impair classification negatively. Least Absolute Shrinkage and Selection Operator (LASSO) performs variable selection in a linear model using a constraint on the norm of the absolute values of the coefficients (Tibshirani 1996). LASSO regularizes the coefficient estimates, shrinks them towards zero and consequently reduces their variance significantly (James *et al.* 2021). Combining LASSO and SVM together with using raw phenotypes instead of predicted breeding values promises a highly computational efficient method.

In this study, we first validated the proposed method as an alternative approach to detect genomic associations for two Mendelian traits: *CNGB3*-related achromatopsia, also reported as Original Braunvieh Haplotype 1 (OH1) (Häfliger *et al.* 2021) and *TWIST2*-related depigmentation known as belt pattern (Awasthi Mishra *et al.* 2017). Subsequently we applied the workflow on stature in Holstein, a complex but well studied genetic trait (Bouwman *et al.* 2018). Finally, the proposed method was used to find new genomic regions associated with the trait multiple birth in Holstein cattle, representing a still poorly understood polygenic trait.

Materials & methods

Phenotypes. The proposed approach was validated using three different datasets of three different Swiss cattle populations (Table 1). The phenotypic and genotypic data were provided by the Swiss cattle breeding organisations. The definition of cases and controls for both monogenic traits achromatopsia and belt was based on phenotypic observations. Regarding the trait stature, cows in first lactation born between 2016 and 2019 with an age at first calving between 730 and 820 days were selected. From the routine conformation scoring, 500 top / bottom cows for measured sacrum height were used to define the groups of cases and controls, respectively. For multiple birth, dams with at least one multiple birth event (mostly twins) were defined as cases and dams with at least 3 singleton calvings as controls.

DOI: 10.3920/978-90-8686-940-4_399, © S. Widmer *et al.* 2022

Table 1. Final datasets per trait and the sizes of the respective case and control groups.

Trait	Population	No. cases	No. controls	No. SNP
Achromatopsia	Original Braunvieh	8	231	670,140
Belt	Brown Swiss	92	1,644	675,828
Stature	Holstein	500	500	680,502
Multiple birth	Holstein	238	919	683,277

Genotypes. Animals were genotyped under the umbrella of genomic selection using different arrays that include between 9k and 850k SNPs. A standard imputation approach was applied, which led to the final datasets (Table 1).

LASSO. LASSO uses the residual sum of squares plus a penalty term to estimate coefficients and to do variable selection in a linear model (Tibshirani 1996). LASSO was described in detail by James *et al.* 2021. The LASSO analyses were carried out using the R-package glmnet (Friedman *et al.* 2010). The parameter λ was estimated using a ten-fold leave-one-out cross-validation. The cross-validation resulted in two different estimates for λ: λ_{min} is the estimate with lowest cross-validation error. Adding one standard error to λ_{min} leads to λ_{1se}. Both estimates of λ were used in the LASSO analyses for all four datasets. LASSO resulted in a smaller set of SNPs compared to the complete marker-set.

SVM. SVM separates two groups of data (such as cases and controls) with the hyperplane that maximizes the distance to the training data of the two groups (Cortes and Vapnik 1995). The separating hyperplane is determined by the training data (James *et al.* 2021) that we defined as a random sample of 80% of the complete dataset. The remaining 20% of the data is used as test data to validate the estimated classification model. A linear kernel was used for classification. For SVM analyses R-package e1071 was applied on the SNPs selected by LASSO for the validation of the variable selection (Meyer *et al.* 2021). The prediction of animals outside the high and low groups is not considered yet.

Results

The results of the LASSO analyses of the four datasets are shown in Table 2. Using λ_{1se} instead of λ_{min} led to a lower number of SNPs selected by LASSO for all datasets.

From the LASSO analysis, the absolute coefficient values were combined using a sliding window approach whereof each window consisted 50 SNPs. Window coefficients are plotted as a function of their chromosomal position (Figure 1). Only windows with a higher impact on the trait and the colocalization of possible candidate genes were considered (Table 3). Therefore, the window coefficients must be higher as the 5-fold standard deviation of all absolute coefficients. For each of the two Mendelian traits, as expected

Table 2. Results of LASSO and SVM analysis for all four traits.

	LASSO No. SNP		SVM rate of correct prediction (%)		SVM F_1 score	
Trait	λ_{min}	λ_{1se}	λ_{min}	λ_{1se}	λ_{min}	λ_{1se}
Achromatopsia	8	5	100	100	1	1
Belt	1,243	420	100	100	1	1
Stature	674	393	100	100	1	1
Multiple birth	232	1	89.7	76.3	0.765	0.035

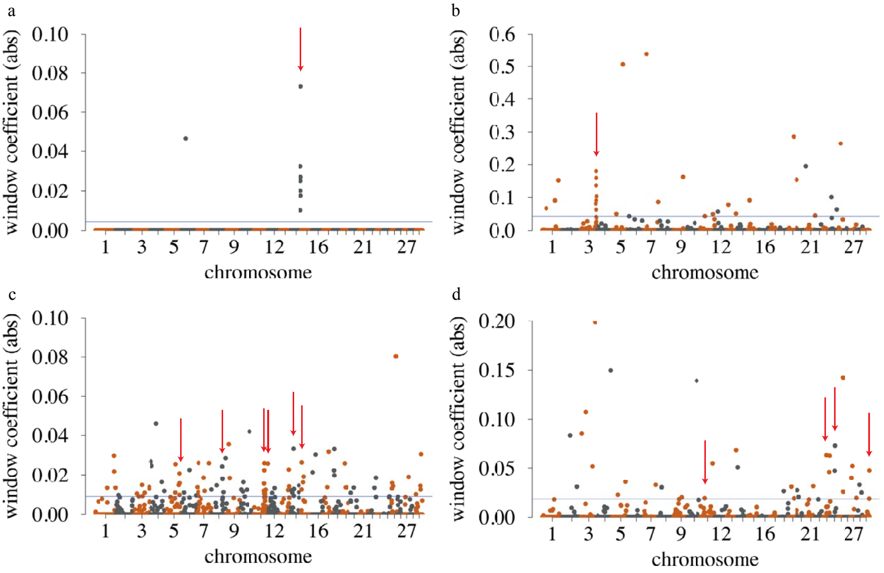

Figure 1. Genome-wide Manhattan plots of the sum of the absolute coefficient values for each window. For (a) achromatopsia and (d) multiple birth λ_{min} was used, while for (b) belt and (c) stature λ_{1se} was used. Red arrows are highlighting the QTL from Table 3 and the blue lines indicate the values of the 5-fold standard deviation of all absolute (abs) window coefficients per trait.

a clear signal was identified on BTA 14 for achromatopsia and BTA 3 for belt, respectively (Figure 1a and 1b). For the trait stature, multiple QTL regions were observed (Figure 1c). Several associated regions were identified for multiple births in Holstein (Figure 1d). The rate of correct predictions from SVM analysis (Table 2) using the test data was 100% (F_1 score=1) for all three validation datasets and for both variants (λ_{1se} and λ_{min}). For the trait of multiple birth, the rate of correct predictions in the test data reached 90% using λ_{min} (F_1 score=0.765).

Discussion

For all four analysed traits, LASSO resulted in a set of SNPs which is considerably smaller compared to the complete marker-set. This forms the basis for the identification of associated genomic regions using raw phenotypes. Using λ_{1se} instead of λ_{min} resulted in a lower number of SNPs selected, as expected from the properties of LASSO. The resulting set of SNPs was used in an SVM classification to group the data into cases and controls and to validate previous results. The rate of correct predictions of SVM for both λ-estimates were comparable and hence λ_{1se} (except for achromatopsia and multiple birth) was favoured because it yields sparser models. Regarding the monogenic recessive disorder achromatopsia, LASSO selected correctly the region of *CNGB3* on BTA 14 (Häfliger *et al.* 2021). For the monogenic dominant inherited belt phenotype, a signal in a window on BTA 3 with *TWIST2* was observed (Awasthi Mishra *et al.* 2017). Concerning stature, two regions on BTA 11 and the signals on BTA 5, 8 and 14 were previously reported as QTL for bovine height in a meta-analysis using the 1000 Bull Genomes project data (Bouwman *et al.* 2018). We propose *MMP13* as a new candidate gene mapping to the region of a QTL found on BTA 15 for stature in cattle, as it influences bone mineralization and growth plate cartilage (Ståhle-Bäckdahl *et al.* 1997). Finally, regarding multiple birth in Holstein, we found multiple associated regions with genes related to the reproduction cycle and oocyte maturation (Table 3). The QTL in the region of *FSHR* and *LHCGR* genes on BTA 11 agrees well with recent findings of GWAS and classical fine mapping approaches

Table 3. Identified associated genome regions with selected candidate genes for the four analysed traits.[1]

Trait	BTA	Start position[2]	End position[2]	Coefficient[3]	Candidate gene[2]
Achromatopsia	14	75,583,503	78,489,957	0.2055	*CNGB3*
Belt	3	117,018,007	118,619,571	0.7183	*TWIST2*
Stature	5	104,712,189	105,942,817	0.0467	*CCND2*
	8	80,544,099	83,316,183	0.0551	
	11	79,311,390	79,594,907	0.0347	
	11	105,098,263	105,220,577	0.0258	
	14	19,996,921	23,379,474	0.0688	*PLAG1*
	15	4,836,530	5,916,553	0.0389	*MMP13*
Multiple birth	11	31,139,464	31,340,099	0.0236	*FSHR, LHCGR*
	23	23,927,635	25,017,848	0.0777	*GSTA1, GSTA2, PAQR8, TMEM14A*
	24	33,994,132	36,319,363	0.1255	*GATA6, ENOSF1, ADCYAP1*
	29	48,563,749	50,109,725	0.0671	*IGF2, INS, CTSD*

[1] For achromatopsia and multiple birth using λ_{min} and for belt and stature using λ_{1se}.

[2] Using the ARS-UCD1.2 reference assembly.

[3] Sum of all absolute values of the coefficient in the interval (typically more than one window).

using predicted breeding values (Widmer *et al.* 2021). Dichotomizing the quantitative trait by selecting extreme phenotypes as cases and controls helped here to unravel the potential genetic factors important for the expression of the phenotypes of interest. The rates of correct predictions from SVM were 100% for the three validation datasets and 90% for multiple birth (using λ_{min}), respectively. Hence, the validation of the herein applied alternative methods was successful.

In conclusion, using the combination of LASSO variable selection and SVM classification, we were able to detect associated genome regions for simple and complex inherited traits in cattle using raw phenotypes. Therefore, we propose this approach as a valuable and efficient alternative to GWAS based on predicted breeding values.

References

Awasthi Mishra N., Drögemüller C., Jagannathan V., Keller I. *et al.* (2017) PLoS One 12(6):e0180170. https://doi.org/10.1371/journal.pone.0180170.

Ban H.-J., Heo J. Y., Oh K.-S. and Park K.-J. (2010) BMC Genet 11:26. https://doi.org/10.1186/1471-2156-11-26.

Bouwman A. C., Daetwyler H. D., Chamberlain A. J., Ponce C. H. *et al.* (2018) Nat Genet 50(3):362–367. https://doi.org/10.1038/s41588-018-0056-5.

Cortes C., and Vapnik V. (1995) Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018

Friedman J.H., Hastie T., and Tibshirani R. (2010) J Stat Softw 33(1):1-22. https://doi.org/10.18637/jss.v033.i01.

Häfliger I. M., Marchionatti E., Stengård M., Wolf-Hofstetter S. *et al.* (2021) Int J Mol Sci 22(22):12440. https://doi.org/10.3390/ijms222212440.

James G., Witten D., Hastie T., and Tibshirani R. (2021) An Introduction to Statistical Learning. Springer New York, New York, USA.

Meyer D., Dimitriadou E., Hornik K., Weingessel A. and Leisch F. (2021) e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. Available at: https://CRAN.R-project.org/package=e1071.

Ståhle-Bäckdahl M., Sandstedt B., Bruce K., Lindahl A. *et al.* (1997) Lab Invest 76(5):717–728.

Tibshirani R. (1996) J R Stat Soc Series B Stat Methodol 58(1):267–288.

Widmer S., Seefried F. R., von Rohr P., Häfliger I. M. *et al.* (2021) Genet Sel Evol 53(1):57. https://doi.org/10.1186/s12711-021-00650-1.