UNIVERSITÄT BERN

Graduate School for Cellular and Biomedical Sciences

University of Bern

Genetic analysis of female fertility focussing on multiple birth events in Swiss cattle

PhD Thesis submitted by

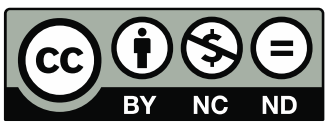
Sarah Widmer

for the degree of

PhD in Computational Biology

Supervisor
Prof. Dr. Cord Drögemüller
Institute of Genetics
Vetsuisse Faculty of the University of Bern

Co-advisor
Dr. Franz R. Seefried
Qualitas AG
Zug, Switzerland



Copyright notice

Accepted by the Faculty of Medicine, the Faculty of Science and the Vetsuisse Faculty of the University of Bern at the request of the Graduate School for Cellular and Biomedical Sciences

Bern, Dean of the Faculty of Medicine

Bern, Dean of the Faculty of Science

Bern, Dean of the Vetsuisse Faculty Bern

I Abstract

In dairy cattle farming, intensive selection for milk yield has led to a decline in female fertility in the last decades, due to unfavourable genetic correlations between milk yield and female fertility. Phenotypes included in current genetic evaluations of fertility are interval and binary traits, calculated from insemination and previous calving date records and deduced from insemination success, respectively. For improved selection, the development of novel phenotypes that describe the physiology of reproduction more precisely would be beneficial. A potential novel phenotype is multiple births. Especially in dairy cattle, multiple birth events are undesirable due to negative impacts on a cow's performance and potential health issues of the dam and the calves.

In the first part of the thesis, I investigated the genetic background of multiple birth events in population studies for the four main Swiss dairy cattle breeds. For this purpose, I designed a breeding value estimation for this novel phenotype in Switzerland. By applying genome-wide association studies (GWAS) on the estimated breeding values, quantitative trait loci (QTL) for multiple births were detected in the three different Swiss dairy cattle populations Holstein, Brown Swiss and Original Braunvieh on chromosomes 11, 15 and 11, respectively. In all populations I identified candidate causal variants affecting the expression of the genes *LHCGR*, *FSHR*, *ID2*, *PRDM11* and *SYT13* by using linkage disequilibrium analysis for fine-mapping.

In the second part of the thesis, I tested alternative methods to identify associated genomic regions, which do not require a complex pipeline of specialised software-tools and massive computing resources. Preliminary work for using machine learning tools in the analysis of binary traits was provided. Thereby, I used the Least Absolute Shrinkage and Selection Operator (Lasso), support vector machine and random forest algorithms for identifying QTL in a case/control approach. The machine learning approaches were validated as promising and efficient alternatives to classical methods. Their application led to the identification of genomic regions showing suggestive associations for multiple births in Holstein cattle.

In future, the machine learning tools random forest, Lasso and support vector machine can offer a low input alternative for GWASs while the availability of data for traits of interest are increasing. The identification of QTL for multiple births improves the understanding of the genetic architecture that underlies our trait of interest and female fertility in general. By developing the breeding value prediction, we set the foundation

for implementing our knowledge in the breeding strategies to avoid multiple births in future. Considering this novel phenotype of female fertility will improve the sustainability of dairy cattle farming.

II Tables of contents, figures and tables

i Table of content

ii List of figures

iii List of tables

1 Introduction

Worldwide there were around 1.53 billion cattle (*Bos taurus* and *Bos indicus*) in the year 2020 [1]. The domestication of cattle was roughly 10,000 years ago [2]. To this day ruminants and especially cattle play an important role in mountain farming, as they can easily digest rough plants in arduously accessible areas. Thereby, feed can be efficiently used that cannot be used for human diets. Especially in Switzerland with a high amount of mountain area, cattle play an important role.

1.1 Swiss cattle population

In Switzerland, there are currently 1.5 million cattle, of which 380,000 go to alpine pastures during the summer months [3]. The number of animals has stagnated since 2020, after a decline in the last decade. Nevertheless, the most important populations, beside crossings, are Brown Swiss and Holstein cattle - two international dairy breeds. In the dairy sector there are currently 530,000 cows registered, excluding rearing heifers and calves [3]. The most important dairy cattle breeds in Switzerland are Holstein, Brown Swiss, Swiss Fleckvieh (originating from the crossing of Holstein x Simmental cattle), Simmental and Original Braunvieh (the founder breed of the modern Brown Swiss population) (Table 1 and Figure 1). In the last 10 years one can observe a clear decrease of Brown Swiss cattle and a slight increase of Holstein animals, while the numbers in the dual-purpose breeds Simmental and Original Braunvieh remain stable [3]. The observation of the Swiss Fleckvieh population is more difficult, as they have only been declared a breed of their own, and had their own herdbook, since 2014.

Table 1: Population size of the main Swiss dairy cattle breeds.

Breed	**Purpose**	**Swiss population[1]**	**Animals in herdbook[2]**
Holstein[3]	dairy	407,260	250,674
Brown Swiss	dairy	253,590	157,361
Swiss Fleckvieh	dual-purpose	138,930	64,749
Simmental	dual-purpose	94,118	23,096
Original Braunvieh	dual-purpose	48,647	13,654

[1] Cut-off date 30th September 2022 [3]

[2] Cut-off date 30th November 2021 [4–6]

[3] Includes Holstein and Red Holstein

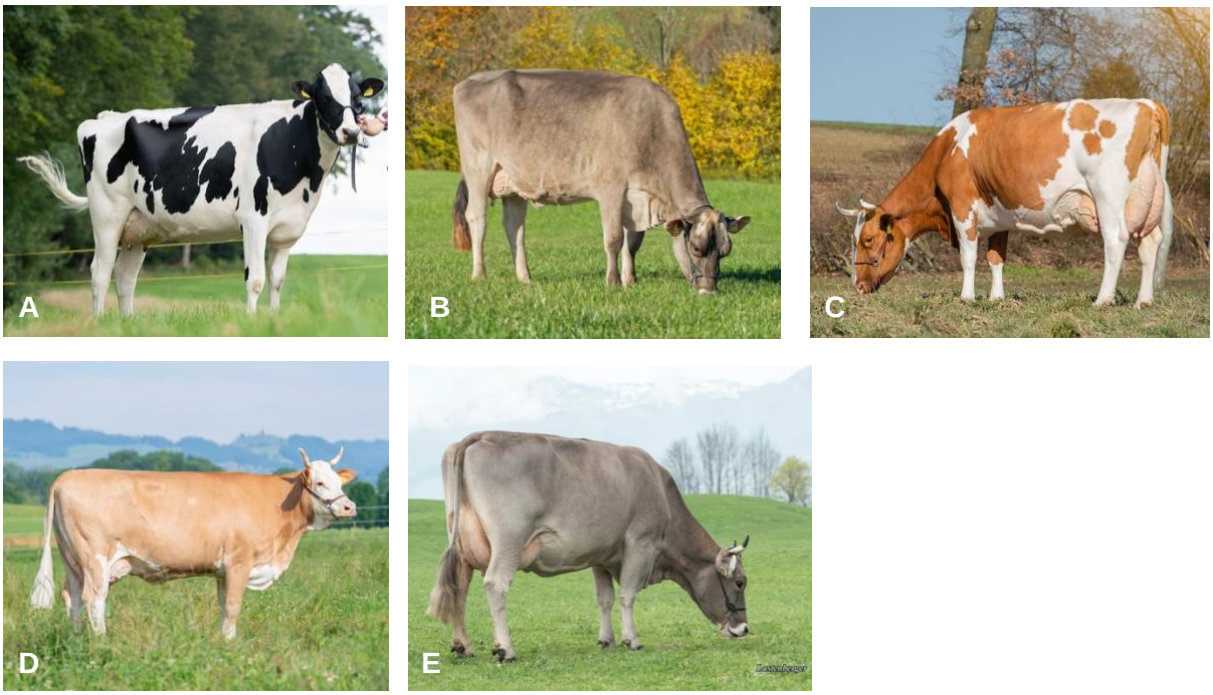


Figure 1: Pictures of cows representing the five most important dairy cattle breeds in Switzerland: (A) Holstein [7], (B) Brown Swiss [8], (C) Swiss Fleckvieh [7], (D) Simmental [7], (E) Original Braunvieh [8].

1.2 Cattle breeding

Although herd books exist since more than 100 years, the real breeding success started in the 1980s when the first breeding value (BV) estimation became available. By applying quantitative genetic approaches, it was possible to compare the animals statistically regarding their own performance [9]. This led to a tremendous increase in performance of highly heritable traits, such as production traits. Especially in traditional breeding schemes, the focus laid on production traits, such as milk yield, milk composition, daily gain and slaughter weight (Figure 2A) [10]. In modern breeding programs the focus changed towards health and fertility traits. In general, the number and extent of recorded phenotypes are increasing population-wide, allowing for better predictions. A second phase of large progress in cattle breeding was the development of genomic selection (GS), which is based on single nucleotide polymorphism (SNP) data [11, 12]. While the theoretical background has been known since 2010 [11], SNP genotyping arrays for livestock have been available since 2008 and the formation of a reference population was necessary for the implementation of GS. This explains the time lag to the introduction of GS in 2012 in Swiss cattle breeding. Another major step

in cattle breeding was the availability of whole genome sequencing (WGS) data. Since 2011, next-generation sequencing technology has made it possible to obtain information about an individual's entire genome in increasingly cost-effective and time-saving ways [13].

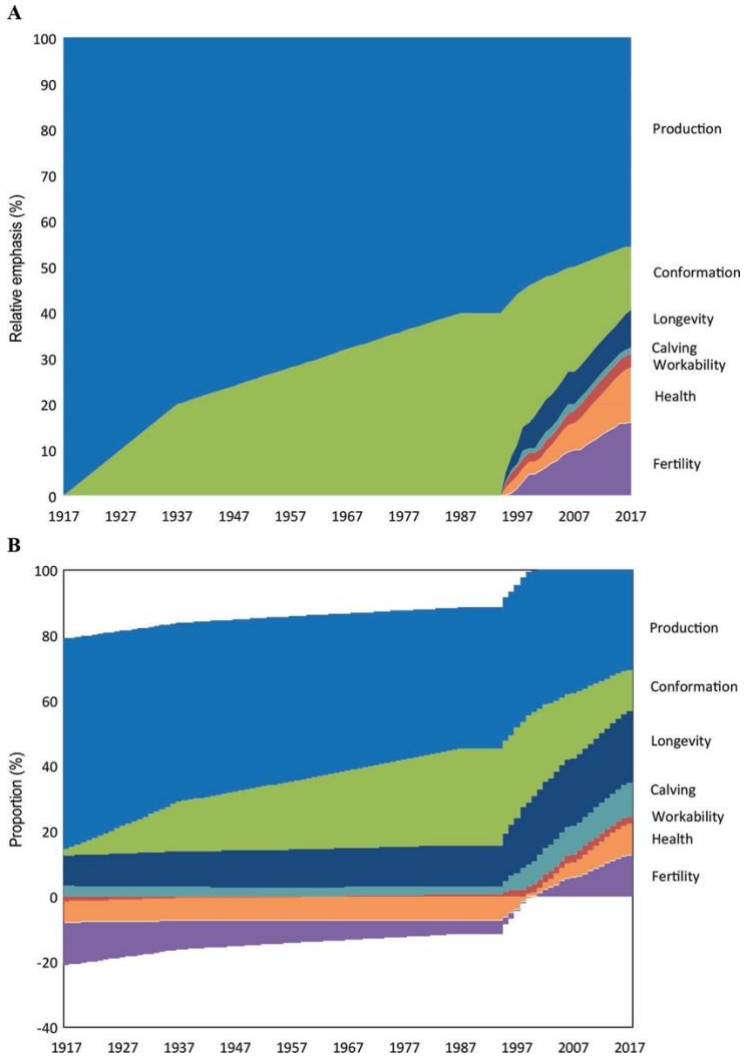


Figure 2: Schematic representation of (A) relative emphasis of traits included in an average selection index over time and (B) proportion of estimated selection response for various trait categories over time (summing to 100%) (from [10]).

1.3 Reproduction and female fertility

In dairy farming, high fertility contributes to herd profitability by achieving more efficient production and maintaining short calving intervals, as well as fewer inseminations and lower veterinary treatment costs. Intensive selection for milk yield in dairy cattle has led to a decline in female fertility, due to unfavourable genetic correlations between

milk yield and female fertility [10, 13]. Also, a study in Swiss Simmental cattle showed that female fertility and milk yield have antagonistic genetic correlations [14]. Improved management practices, the implementation of female fertility traits in the selection scheme and GS have contributed to reversing negative trends in dairy cow fertility. The emphasis on fertility traits in the selection scheme has increased since the 1990s (Figure 2A) [10]. Figure 2B shows that with the inclusion of these low heritable traits in breeding programs, the negative correlations could be counteracted, but further progress is still required.

In recent years, a number of studies have identified recessive alleles segregating in modern cattle populations, including embryonic recessive lethal alleles [e.g. 15–18]. One consequence of the existence of recessive lethal alleles is their impact on female fertility, as females carrying recessive lethal alleles, for e.g. will lose embryos that are homozygous for one of these alleles. In addition, female fertility traits tend to suffer from inbreeding depression [19].

While we are interested in female fertility traits with a special focus on the occurrence of multiple births, male fertility should also be addressed. The Animal Genomics group at ETH in Zurich analysed semen quality measurements and identified variants associated with male fertility traits in the genes *QRICH2*, *WDR19* and *SPATA16* on chromosomes 19, 6, and 1, respectively [20–22].

Phenotypes included in current genetic evaluations of fertility are largely interval and binary traits calculated from inseminations and previous calving records. In Switzerland, BVs for the following female fertility traits are calculated:

- Days to first service
- Interval between first and last insemination heifers
- Interval between first and last insemination cows
- Non-return rate heifers (after 56 days)
- Non-return rate cows (after 56 days)

Additional traits such as calving, health, variation in body condition, and longevity traits will also increase genetic improvement of fertility in future. For improved selection it is important to develop novel phenotypes, which describe the physiology of reproduction more closely and reduce the potential bias of management on observations.

1.4 Multiple births

A potential novel phenotype based on more physiological processes is multiple births. Cattle are usually monoparous, so that a pregnancy usually leads to the birth of a single calf; however, multiple births occur at a rate of 1.02 to 9.6% depending on the population of interest [23–25]. Most multiple births result from multiple ovulations, when several ovulatory follicles mature simultaneously. Therefore, only around 5 to 10% of bovine twins are monozygotic [26, 27].

Especially for dairy cattle, twin and multiple birth events are undesirable due to their negative impacts on health and performance [28–30]. Multiple births lead to decreased cow performance due to longer calving intervals and lower conception rates. In addition, the twin calves show a reduced survival rate and are weaker. Furthermore, during birth, the risk for dystocia, abortions and stillbirths is increased. Similarly, increased occurrences of retained placentas, metabolic disorders, displaced abomasum and ketosis for the dam are observed [28–30]. Previous studies using low marker density or microsatellites showed evidence for quantitative trait loci (QTL) for twin births in North American Holsteins [31–33].

Interestingly, the frequency of twin deliveries varies among human populations. Already in 1909 Weinberg [34] suggested that hereditary twinning is transmitted through the female line only. Natural multiple pregnancies in women leading to dizygotic twins, is heritable and varies between racial groups, suggesting a genetic predisposition (OMIM 276400). A genome-wide association study (GWAS) performed in mothers of spontaneous dizygotic twins identified significant association to the genes *FSHB* and *SMAD3* [35]. These QTL were confirmed, in addition to the detection of association, with variants close to the *PIAS1* and *SKOR1* genes in a GWAS comparing multiples as cases against singletons as controls [36].

Selection for improved fertility using traditional traits benefits from implementation of new phenotypes such as selection against multiple birth events, as these phenotypes describe the physiology of reproduction more closely. The genetic variability of fertility traits, e.g. multiple birth events, proves that it is possible to use these traits in breeding programs and achieve improvement over time. This fact and their high economic importance, point out high relevance of fertility traits in the selection scheme for Swiss cattle. Thereby, genetic improvement of reproductive efficiency using novel

phenotypes can be achieved, which is required for long-term sustainability of the dairy cattle populations.

1.5 Methods for genetic analyses

1.5.1 Classic genetic approaches

The large-scale availability of high throughput SNP array genotyping data from thousands of cattle generated for the purpose of genomic evaluation in Swiss cattle populations enables the investigation of genetic associations using genomic and phenotypic information. GWAS has become a useful tool to reveal the genetic architecture for complex traits [37]. For several traits of female fertility, e.g. non-return rate at 56 days and interval from first to last insemination, QTL were found in the Brown Swiss cattle population [38]. These GWAS for additive effects are commonly performed on pseudo-phenotypes based on de-regressed BVs as response variables. Recently, the accumulation of cows with available phenotypes and genotypes has enabled an increase in the reliability of the estimated BVs.

There are different GWAS methods available, such as the single SNP regression or the window-based Bayes B approach. The single SNP regression tests the association of every single SNP with the response variable under consideration of the genomic relationship matrix [39]. This single marker association analysis has to be corrected for multiple testing, e.g. by Bonferroni correction, due to repeated testing [40]. Important for the interpretation is that the associated SNP does not necessarily have to be the causative one, but could be linked to the causative marker in the associated region. Because fitting one genotype at a time can easily lead to biased results due to population stratification and linkage disequilibrium (LD) [41], we also used the approach to fit subsets of multiple markers simultaneously. This reduced the issue of bias and led to the window-based analysis. We applied the window-based GWAS approach based on a Bayes B algorithm. Window-based association analyses were conducted using a window size of 1 mega base (MB).

Haplotype regression are useful to identify associated haplotypes, which are suitable for subsequent fine-mapping of detected QTL by GWAS. Therefore, using SNP-derived haplotypes allows for association testing within known QTL regions in the populations of interest.

Next-generation sequencing enables us to unravel the whole genomic DNA of an individual in a cost- and time-efficient way. Mining of population-based WGS data in associated genomic regions has identified causative variants for both monogenic Mendelian as well as polygenic traits [15]. Especially the availability of massive WGS data provided by the 1000 Bull Genomes Project (Run 9 includes 5,116 animals) enables to distinguish between common and rare variants [42].

Linkage disequilibrium (LD) analysis was applied to the WGS data to identify candidate causal variants [43]. LD is the correlation between nearby variants, acknowledging the fact that neighbouring SNP on the same chromosome are more frequently inherited together [44]. The co-segregation of loci can be estimated as an LD score [43]. We did not estimate linkage between two SNP, as the association of the best-associated haplotype was predicted. Therefore, the LD between the best associated haplotype from the previous regression analysis with variants from WGS data was calculated.

The workflow used for this work with the classic genetic approaches is shown in Figure 3.

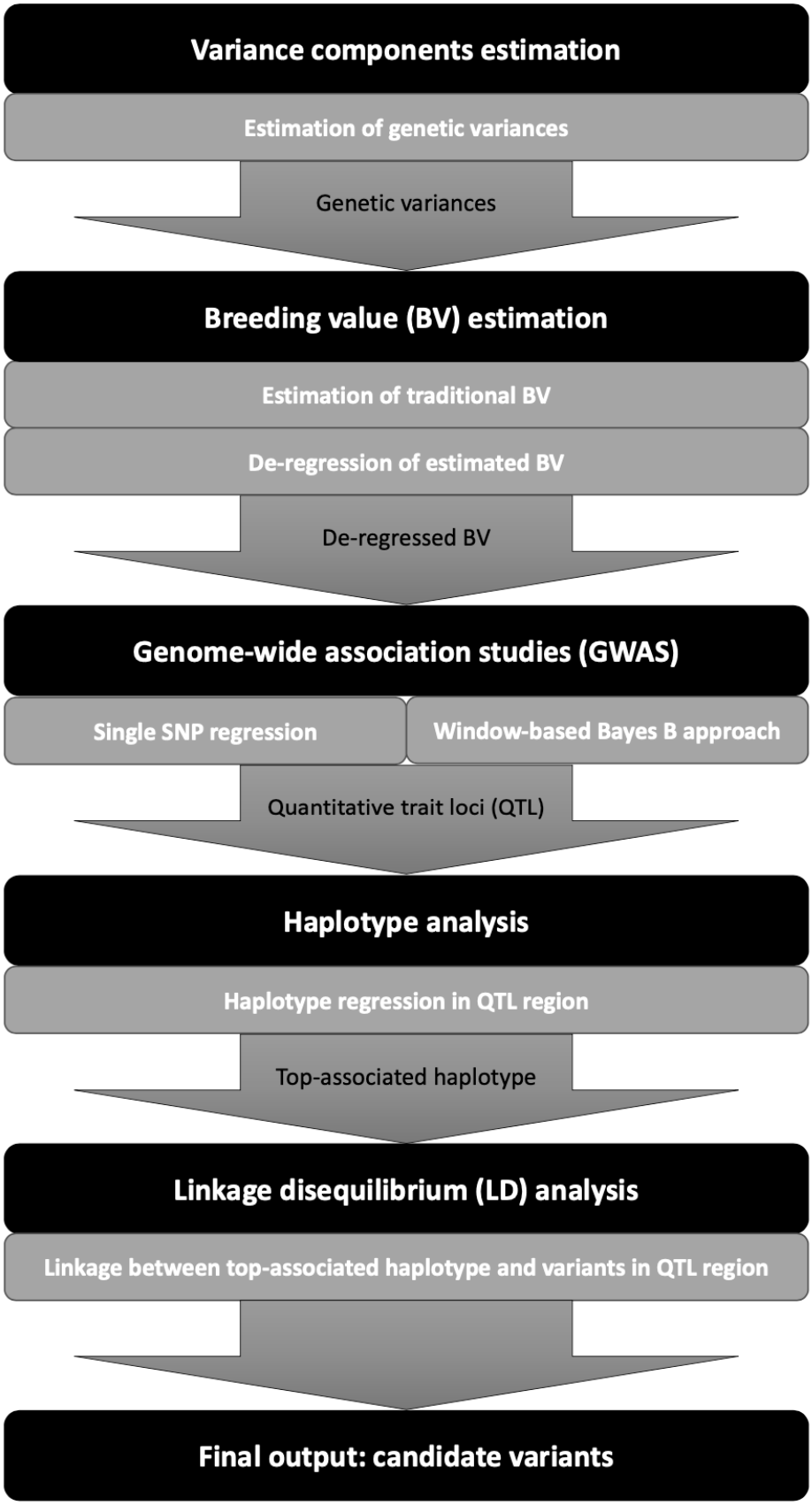


Figure 3: Workflow for the classic genetic approaches applied in this thesis**.**

1.5.2 Machine learning approaches

The classic approaches to identify associated genomic regions are complex and require massive computing resources because they use pseudo-phenotypes based on de-regressed estimated BVs as response variables in the statistical model [45]. The preparation procedure to obtain these pseudo-phenotypes requires a complex pipeline of specialised software-tools, not all of which are publicly available. Hence, there is a considerable need for a simpler process that allows for the identification of genomic regions associated with a trait of interest. An alternative could be machine learning approaches, which can be applied directly to raw phenotypes. The workflow used for this work with machine learning methods is shown in Figure 4.

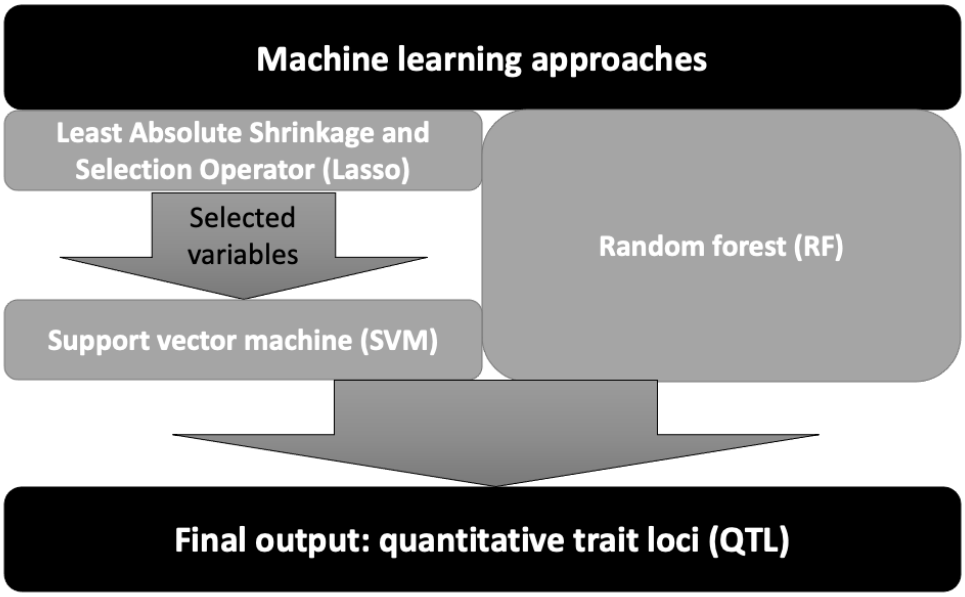


Figure 4: Workflow for the machine learning approaches applied in this thesis**.**

Support vector machine (SVM) classifies data consisting of different groups by separating hyperplanes [46, 47] (Figure 5). These separating hyperplanes are defined by explanatory variables, such as SNP genotypes. Therefore, SVM can be used to analyse data which is divided into two groups as cases and controls. SVM is used in multiple fields, as for example in human medicine to detect SNPs associated with the risk for type 2 diabetes and to predict genotype-based health status [48]. Although SVM can be used for high-dimensional data, it is important to first identify the subset of relevant explanatory variables to avoid the introduction of noise through irrelevant

variables. A high level of noise reduces the quality of separation of the data based on the estimated hyperplane and increases the risk of overfitting. Overfitting impairs the classification.

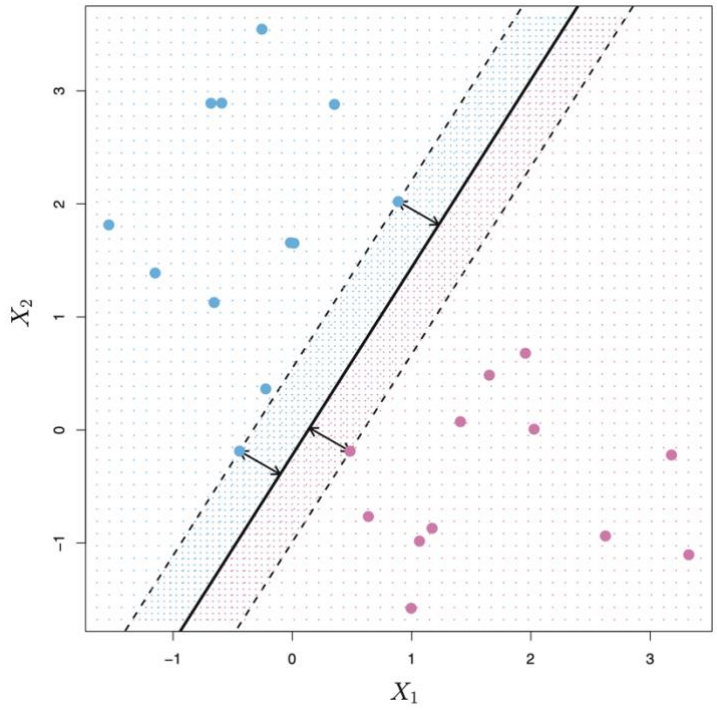


Figure 5: Principle of support vector machine (SVM). There are two classes of observations, shown in blue and in purple. The maximal margin hyperplane is shown as a solid line. The margin is the distance from the solid line to either of the dashed lines. The two blue points and the purple point that lie on the dashed lines are the support vectors, and the distance from those points to the hyperplane is indicated by arrows. The purple and blue grid indicates the decision rule made by a classifier based on this separating hyperplane (from [47]).

The variable selection prior to SVM can be performed with the Least Absolute Shrinkage and Selection Operator (Lasso) algorithm. Lasso performs variable selection in a linear model by using a constraint on the norm of the absolute values of the coefficients, where in our analysis the coefficients are the SNPs [49]. Lasso regularises the coefficient estimates, shrinks them towards zero and consequently reduces their variance significantly [47]. Only SNPs with a non-zero coefficient are regarded as relevant candidate positions for being associated to a given phenotype of interest. Hence, the variable selection procedure imposed by Lasso is used to detect

SNP which are relevant for a phenotype of interest. The remaining SNP can be used to identify QTL. The variable selection is then validated by the SVM classification.

Random forest (RF) is an alternative one-step approach. This machine learning method uses regression trees and bootstrapping where at every tree node a variable selection will be made [50] (Figure 6). The method uses hundreds to thousands of trees, with each tree started with a bootstrap sample of the entire dataset. Bootstrapping represents the repeated sampling and combining of the manyfold of variables. At each node of each tree, a random subset of variables is selected and used as candidate variables to find the best split. As a result, one gets a permutation importance score for each predictor variable which measures the difference in prediction accuracy before and after permuting values of the variable over all trees [51]. RF is already known for application on the analysis of genomic data, e.g. variable selection and genetic association detection as well as prediction and classification of human diseases [52].

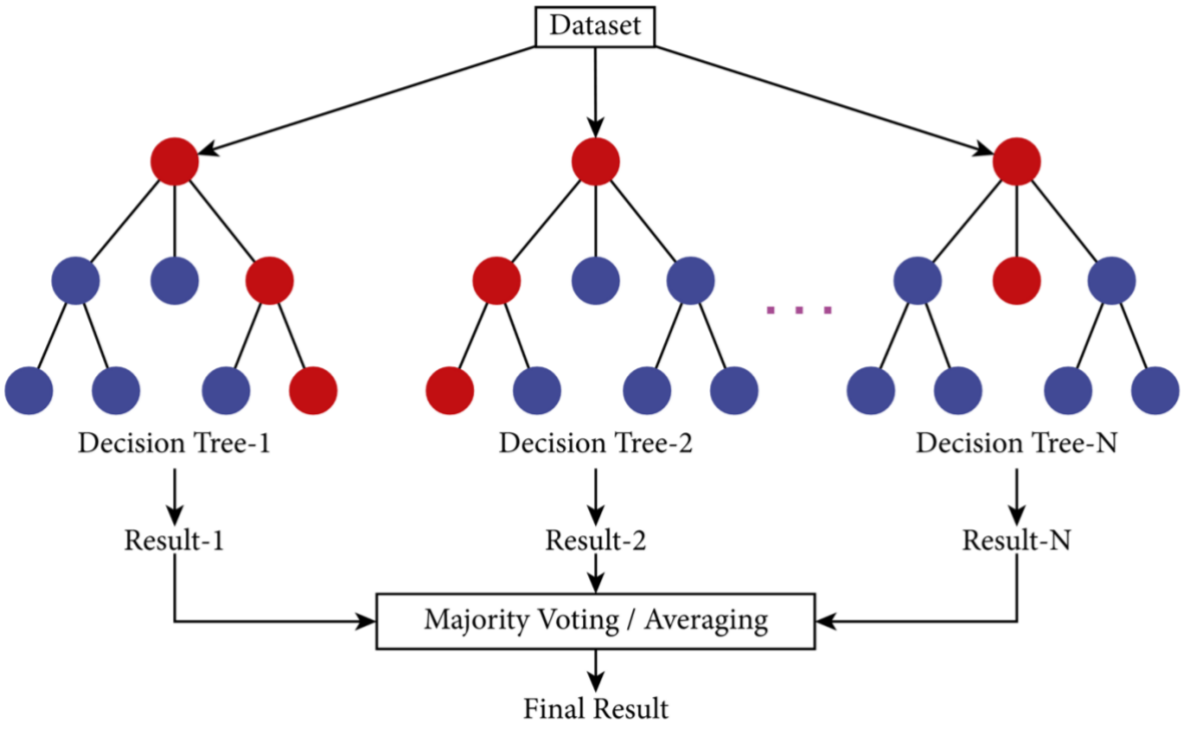


Figure 6: Illustration of random forest trees (from [53]).

All these machine learning tools can be applied directly to raw phenotypes. I have used these methods in a case/control design regarding cows having multiple births records.

2 Hypothesis and aim

The general hypothesis of this thesis was that multiple birth events are heritable. Furthermore, the genetic architecture of multiple births will be investigated based on different types of association studies.

The aims of this study are to estimate BVs for multiple birth events in Swiss cattle populations, to detect QTL with additive effects using GWAS and to unravel causative genomic variants using WGS data.

The analyses were carried out for the four major Swiss dairy cattle populations:

- Holstein
- Brown Swiss
- Original Braunvieh
- Simmental

Furthermore, novel machine learning algorithms were applied to validate alternative approaches and to identify QTL directly by using raw phenotypes.

Finally, the goal was to explore the possibilities of including the phenotype multiple birth as female fertility trait in the breeding programme to improve animal health and welfare.

3 Results

3.1 Results table of content

3.2 Publications

A major QTL at the *LHCGR*/*FSHR* locus for multiple birth in Holstein cattle

Journal:	Genetics Selection Evolution
Manuscript status:	published
Contributions:	phenotyping data preparation, data analyses, visualization of the results, writing original draft and revisions
Displayed version:	published version
DOI:	https://doi.org/10.1186/s12711-021-00650-1

Widmer *et al. Genet Sel Evol* (2021) 53:57
https://doi.org/10.1186/s12711-021-00650-1

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

A major QTL at the *LHCGR/FSHR* locus for multiple birth in Holstein cattle

Sarah Widmer[1], Franz R. Seefried[2], Peter von Rohr[2], Irene M. Häfliger[1], Mirjam Spengeler[2] and Cord Drögemüller[1*]

Abstract

Background: Twin and multiple births are rare in cattle and have a negative impact on the performance and health of cows and calves. Therefore, selection against multiple birth would be desirable in dairy cattle breeds such as Holstein. We applied different methods to decipher the genetic architecture of this trait using de-regressed breeding values for maternal multiple birth of ~2500 Holstein individuals to perform genome-wide association analyses using ~600 K imputed single nucleotide polymorphisms (SNPs).

Results: In the population studied, we found no significant genetic trend over time of the estimated breeding values for multiple birth, which indicates that this trait has not been selected for in the past. In addition to several suggestive non-significant quantitative trait loci (QTL) on different chromosomes, we identified a major QTL on chromosome 11 for maternal multiple birth that explains ~16% of the total genetic variance. Using a haplotype-based approach, this QTL was fine-mapped to a 70-kb window on chromosome 11 between 31.00 and 31.07 Mb that harbors two functional candidate genes (*LHCGR* and *FSHR*). Analysis of whole-genome sequence data by linkage-disequilibrium estimation revealed a regulatory variant in the 5′-region of *LHCGR* as a possible candidate causal variant for the identified major QTL. Furthermore, the identified haplotype showed significant effects on stillbirth and days to first service.

Conclusions: QTL detection and subsequent identification of causal variants in livestock species remain challenging in spite of the availability of large-scale genotype and phenotype data. Here, we report for the first time a major QTL for multiple birth in Holstein cattle and provide evidence for a linked variant in the non-coding region of a functional candidate gene. This discovery, which is a first step towards the understanding of the genetic architecture of this polygenic trait, opens the path for future selection against this undesirable trait, and thus contributes to increased animal health and welfare.

Background

In dairy cattle production, herd profitability is heavily influenced by the number of calves born alive and by the length of calving intervals. Intensive selection for milk yield in dairy cattle has led to a decline in female fertility, due to unfavorable genetic correlations between milk yield and female fertility [1, 2]. Improved management practices and genomic selection have contributed to reversing negative trends in dairy cow fertility, but further progress is still required.

Genome-wide association studies (GWAS) have become a useful tool to partially reveal the genetic architecture of complex traits. For several traits related to female fertility, such as non-return rate at 56 days and interval from first to last insemination, quantitative trait loci (QTL) have been detected e.g. in Holstein [3] and Brown Swiss cattle [4]. Mining of population-based whole-genome sequence (WGS) datasets in the associated genomic regions has been used to identify causative

*Correspondence: cord.droegemueller@vetsuisse.unibe.ch
[1] Institute of Genetics, Vetsuisse Faculty, University of Bern, 3012 Bern, Switzerland
Full list of author information is available at the end of the article

variants for both monogenic Mendelian and polygenic complex traits [5]. The necessary tools, such as GWAS and linkage disequilibrium mapping, are available for the analysis of other female fertility traits such as multiple birth.

Cattle are generally a monotocous species and pregnancies typically result in the birth of singletons. Multiple births are rare, with multiple birth rates (MBR) ranging from 1.02 to 9.6% depending on the breed and study [6–15] and are generally higher in dairy cattle than in beef cattle [6]. Most multiple births result from multiple ovulations when two or more ovulatory follicles mature simultaneously. In cattle, only about 5 to 10% of the twins are monozygotic. Silva del Rio et al. [16] found that 7.5% of the twins were monozygotic and Atteneder [6] determined a rate that ranges from 6.5 to 11.7% depending on the breed analysed. However, to date, this trait has not been analyzed in general or with a genetic model using data from Switzerland.

For dairy cattle, twins and multiple births are undesirable for several reasons. Multiple births are generally associated with increased health problems for both the dam and the calves. Occurrence of remained placenta, metabolic disorders, displaced abomasum, and ketosis are some of the direct negative effects on the dam [17–21]. The impact of multiple births on fertility is indicated by its effect on the subsequent calving interval and conception rate [18, 22–24]. Calves from multiple births also have a higher risk of mortality and deficiency syndrome; the risk of abortion, dystocia and stillbirth also increases with multiple births [6, 17, 18, 20, 22–26]. All these factors result in higher costs for the farmers. Thus, selection against this trait could improve fertility and as well as profitability of dairy operations.

There are different non-genetic factors which might influence MBR. Several studies have shown that the parity of the cow influences MBR significantly, as well as other environmental factors such as season and herd [6–11, 26, 27]. Cows in parity one (0.70 to 1.63%) have significantly lower MBR than multiparous cows (2.87 to 7.35%), and MBR was found to increase until the 3rd to 5th parity and then to remain stable in later lactations. Two studies have analyzed the association between milk yield and MBR. While one study found no association [12], the second suggested an association of higher MBR with higher milk yield [13]. Calving season seems to be an important factor that affects the variation in MBR, with different studies reporting a higher MBR for births occurring in the summer months [6–10, 26]. These results suggested that the rate of multiple ovulations at conceptus is higher in late summer and fall, although another study reported that MBR is highest between the end of summer and fall [27]. A positive phenotypic trend for MBR over time has been observed in several studies [8, 10, 11, 14, 26–28], which indicates that MBR is associated with other traits under selection. Reports on the genetic trend for MBR are conflicting, with one study showing a negative trend over time [7] and another one finding no association between MBR and time [13]. Overall, MBR is increasing in most of the dairy cattle populations analyzed to date. Atteneder [6] suggested that the age at first calving affects MBR, with older dams having a higher rate. A heritability of 0.011 to 0.160 was estimated for MBR, which indicates a low but non-zero genetic contribution [7, 9, 10, 12, 14, 15, 27, 29, 30]. Estimations from linear models were lower than those from threshold models. Furthermore, Atteneder [6] showed that the maternal heritability for MBR was higher than the direct heritability (0.017 to 0.063 vs. 0.001 to 0.005) [6]. Hence, it is likely that multiple loci are influencing the trait, which have not been all identified.

Different QTL mapping studies have analyzed maternal multiple birth in North American Holstein, Israelian Holstein or Norwegian cattle using family-based microsatellite interval mapping approaches [15, 31–38] and have identified several QTL on 13 chromosomes, depending on the population examined and the study [15, 36–38]. For North American Holsteins, paternal half-sib families were analyzed using single-marker analysis or combined linkage-linkage disequilibrium approaches resulting in the detection of QTL on eight chromosomes [31, 34]. One study analyzed the Maremmana beef breed in Italy by using a single-trait linear mixed effect model and an animal threshold model that included the number of calves born per cow as the phenotype [14], and detected one significant SNP on chromosome 24 using a GWAS based on 54 k SNP data. The analyses, which we present here based on large-scale phenotype and genotype data and linear mixed models to analyze multiple birth in cattle, are the first in the field.

In this study, our aim was to conduct a comprehensive genetic analysis of the trait multiple birth in Swiss Holstein cattle. To estimate breeding values, we used phenotypic data for single and multiple birth cases that have been recorded over several decades through the national animal recording database and combined them with pedigree data. In addition, we used large-scale genotype data that were obtained during routine genomic selection of males and females to identify associated QTL by GWAS and haplotype regression analysis. A fine-mapping approach was carried out to define a critical genomic region and potential candidate causal variants. Finally, the detected association was validated by evaluating the effect of the identified haplotype on available birth and fertility traits.

Methods

Phenotypes

Large-scale phenotypic recording of birth records was carried out routinely through the Swiss national animal recording database between 2006 and 2018. Here, we focused on data from only one breeding organization: swissherdbook (Zollikofen, Switzerland). The raw dataset contained 3,977,467 birth records mainly from the Holstein and Simmental breeds. Data analysis and preparation for breeding value estimation were performed with an inhouse software written in R [39] using RStudio [40]. Birth records resulting from embryo transfer were removed. After data validation and preparation, 971,613 records (235,053 on Holstein, 190,243 on Simmental, 536,932 on Swiss Fleckvieh (Holstein × Simmental), 6726 on Monbéliarde, 1511 on Normande, 1088 on Pinzgauer, and 60 on Evolène) including a multiple birth code were available for the genetic analyses of the discrete trait multiple birth. The overall MBR was 3.56%. Further details on the final dataset are in Table 1.

Estimation of variance components

A mixed linear model was fitted to the phenotypic data described as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wh} + \mathbf{Z_d mb_d} + \mathbf{Z_m mb_m} + \boldsymbol{\epsilon}, \quad (1)$$

where $\mathbf{b}$ represents the vector of the fixed effects, $\mathbf{h}$ is the vector of the random herd-year effect, $\mathbf{mb_d}$ and $\mathbf{mb_m}$ represent the direct (calf) and the maternal (dam) genetic effects, respectively, and $\boldsymbol{\epsilon}$ represents the residual. The fixed effects for parity, season, semen sexing, and the covariate of age of dam at birth were considered. The number of records per level of fixed effect are shown in Table 1. $\mathbf{X}$, $\mathbf{W}$, $\mathbf{Z_d}$ and $\mathbf{Z_m}$ are the design matrices for the fixed ($\mathbf{X}$) and random ($\mathbf{W}$, $\mathbf{Z_d}$, and $\mathbf{Z_m}$) effects. Model (1) is based on a previous unpublished study that analyzed a similar dataset. The components in the vector of observations ($\mathbf{y}$) were encoded with 1 as single birth, 2 for twin or triplet births. The dataset was filtered to exclude from the analysis all the records from herds that had less than 260 records per herd, and all records in herd-year classes that had less than five records per herd-year class, which resulted in a final dataset of 167,703 records. Variance components were estimated using the software-program vce [41].

Table 1 Solutions for fixed effects of the estimated breeding values based on the final dataset

Fixed factor	Level	Number of observations per level[a]	Solution for effect
Parity	1	229,669	0.305
	2	229,034	0.356
	3	169,509	0.368
	4	123,887	0.374
	5+	219,514	0.379
Sexed semen	No	920,157	0.354
	Yes	51,456	0.346
Season of birth	Spring	186,693	0.346
	Summer	185,133	0.364
	Fall	313,158	0.354
	Winter	286,629	0.351

[a] Based on the final dataset of 971,613 records

Prediction of breeding values

Breeding values were predicted with the MiX99 program [42] using the mixed linear effects model (1) shown above. The dataset was filtered to exclude all records from herds that had less than 260 records but without setting a minimal number of observations per herd-year levels. In total, 971,613 records were used for breeding value prediction. The required variances and covariances were based on the values estimated in the previous step. Reliabilities of breeding values were estimated based on the approach by Tier and Meyer [43].

Breeding values were standardized to have a mean of 100 and a standard deviation of 12. We defined base animals, the 8- to 10-year old Holstein sires (Red Holstein and Holstein), which is similar to the definition of base animals for the calving ease trait used in Switzerland.

Estimated breeding values (EBV) for the direct (mbd) and maternal (mbm) multiple birth traits were de-regressed according to Garrick et al. [44] (Table 2). De-regressed EBV were used in the association analyses if the corresponding EBV reliability was ≥ 0.35, which left 728 and 2540 animals for the mbd and mbm traits, respectively.

Genotypes

Routine SNP genotype data generated for genomic selection were available for ~60,000 animals. Animals were genotyped using several routinely available array chips that include between 3 and 150 k SNPs. The available genotype archive was used in a two-step imputation approach and was imputed first to a density of 150 k.

Table 2 Statistics of the de-regressed breeding values (BV) for the direct (mbd) and maternal (mbm) multiple birth traits

Trait	Min de-regressed BV	Max de-regressed BV	Mean de-regressed BV (sd)	Number of observations
mbd	− 359.479	119.458	3.453 (35.320)	881
mbm	− 117.343	307.131	2.852 (40.060)	3220

Subsequently, imputation to HD-density was carried out using 150 k data. The reference dataset for the 150 k array included 1688 Holstein and 1511 Simmental animals and the reference dataset for the HD-density array included 703 Holstein and 663 Simmental cattle. FImpute software was used with default parameters for both steps [45]. In each step, SNPs with a minor allele frequency (MAF) lower than 1% were removed from the dataset. The final marker set included 114,657 and 691,222 SNPs for each density (150 k and HD), respectively. SNPs were filtered using the following thresholds: MAF higher than 0.01 and an SNP call rate higher than 0.99 in the genotype data from the reference population. Filtering was done separately for the Simmental and Holstein populations. The marker set used in both imputation steps was created by selecting SNPs that met the criteria in at least one of the two breeds, Simmental and Holstein. The current ASR-UCD1.2 cow assembly was used as the reference genome during imputation. The mean distance between SNPs in the final HD-density dataset with 691,222 SNPs was 3595 bp.

Association studies

Single SNP regression

Genome-wide single marker association studies were carried out using the mixed model approach and the software snp1101 [46]. Only data from animals with a Holstein pedigree-based gene proportion higher than 0.6 were included in all analyses. After calculating the genomic relationship for the animals used in the single-marker association analyses, it was fitted in the model to correct for population stratification [47], as follows:

$$y_i = \mu + \beta g_i + a_i + \varepsilon_i \quad (2)$$

where y_i is the de-regressed EBV of animal i, μ is the overall mean, β is the allele substitution effect, g_i is the SNP genotype of animal i, which was coded as 0, 1, and 2 for SNP genotypes AA, AB, and BB, respectively. a_i is the random additive polygenic effect of animal i with $\mathbf{a} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_a^2\right)$ where $\mathbf{G}$ is the genomic relationship matrix [47] and σ_a^2 is the polygenic additive genetic variance. ε_i is the random residual effect. We used this model to identify variants that were significantly associated with the traits mbd and mbm.

BayesB approach

Fitting one genotype at a time can easily lead to biased results due to stratification and linkage disequilibrium (LD) [48]. Fitting subsets of genotypes at many markers simultaneously can address this paradigm. Window-based association analyses were conducted using the GenSel software package [49] and the BayesB algorithm [50]. The π parameter that represents the proportion of loci with zero effect was estimated beforehand using the same dataset, and a value of 0.989 was set as an a priori starting value. Genomic windows were constructed for 1-Mb segments and the window variance was estimated.

Haplotype analysis

A subsequent fine-mapping approach was used to detect possible causal variants. To identify the haplotype that encompasses the putative causal QTL, we performed a haplotype regression analysis limited to the previously identified genomic windows that explained a significant proportion of the total genetic variance. Following an approach described by Pausch et al. [51], our aim was to identify significantly associated haplotypes within each window. Therefore, we estimated haplotype effects by cutting the haplotypes using different lengths of all odd numbers between 9 and 301 SNPs within each identified segment. By shifting the starting point SNP-wise and using different haplotype lengths, each haplotype within each segment was considered in the analysis. Again, de-regressed EBV were used in the haplotype association analysis as response variables. The most significantly associated haplotype was used in the following steps to identify any candidate genomic variants in the whole-genome sequence data. The observed top-associated haplotype for significantly associated windows was tested for association with routinely recorded fertility and birth traits using the GCTA software package [52] and by fitting the haplotype as a single diplotype. The genomic relationship was included in the model to correct for population stratification. In addition, association analyses using de-regressed EBV for routinely available fertility and calving traits were carried out, to test the likely co-association between the most significantly associated haplotypes for twinning and other fertility traits.

Fine-mapping

Using the plink software package [53] and decoded diplotype data derived from the diplotype of the most significantly associated haplotype, WGS data were screened for variants that were in significant LD with the haplotype. A 880-kb window between positions 30,474,126 bp and 31,350,487 bp on chromosome 11 was chosen based on the localization of the highly associated haplotypes. We used WGS data on 4109 animals provided by the 1000 Bull Genomes Project run 8 that includes 1121 Holstein cattle [54]. The data were converted chromosome-wise from VCF to plink input files using the software VCFtools v0.1.16 [55]. Furthermore, the data were converted to binary files and the VCF quality controls were performed with the plink v.1.90b3.46 software [53]. Variants with a missing call rate higher than 0.1, a MAF lower than 0.01, and samples with missing call rate higher than

0.1 were removed. Mendelian errors were analyzed and all samples and variants that had a Mendelian error rate higher 2% were removed. Pedigree records for the 1000 Bull Genomes Project samples were obtained from the swissherdbook database (Zollikofen, Switzerland) and comprised 544 duos and 10 trios for Mendelian analyses. For the LD-based variant search approach, 318 samples of the 1121 whole-genome sequenced Holstein animals for which additional SNP genotype information was available, were used after data filtering and included 4323 variants located in the 880-kb window on chromosome 11.

To obtain supporting evidence for a putative causal role of the variants, an analysis of sequence homology among 20 mammals was performed with the UCSC genome browser. All variants that were in high LD ($r^2 \geq 0.7$) with the identified associated haplotype were analyzed.

Table 3 Estimated raw variance components

Component	Raw value	Standard error
Herd-year	0.992 E−04	0.185 E−04
Animal	0.578 E−04	0.354 E−04
Dam	0.130 E−02	0.134 E−03
Residual	0.036	0.136 E−03
Correlation animal/dam	− 0.158 E−03	0.816 E−04

Based on the used dataset of 167,703 records

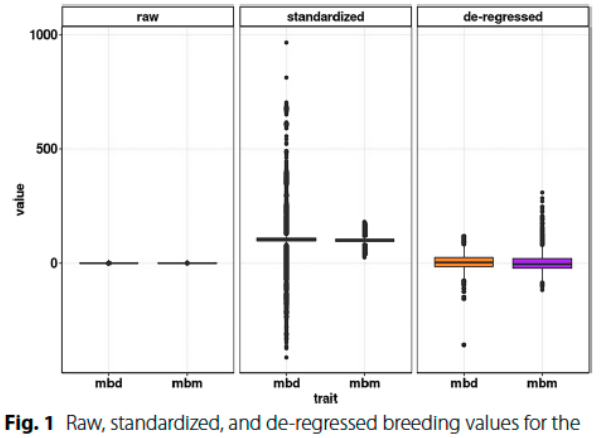


Fig. 1 Raw, standardized, and de-regressed breeding values for the direct (mbd) and maternal (mbm) multiple birth traits. The standard deviation (SD) for the standardized breeding values is 35.959 for mbd and 8.691 for mbm

Results

Estimation of variance components

The raw values of the estimated variance components are in Table 3. The heritability of the direct and the maternal genetic effects were 0.0015 and 0.0348, respectively.

Prediction of breeding values

Breeding values were estimated for the traits mbm and mbd, and these were used in a de-regressed transformation as input for the association analysis. The solutions for the fixed effect of parity, use of sexed semen, and season of birth are in Table 1. We found a significant negative effect of the use of sexed semen on multiple birth rate. Furthermore, the highest prevalence for multiple births was observed in summer, and multiparous cows had a higher incidence for multiple births. The empirical distributions of the raw, standardized, and de-regressed estimated breeding values are shown as boxplots in Fig. 1.

The mean reliability for the standardized breeding values of 1,750,016 animals was 0.144 for mbd and 0.219 for mbm. The mean estimated breeding values grouped by year of birth for all the animals is a measure of the genetic trend of a trait. Interestingly, no clear genetic trend (Fig. 2) was observed for either of the traits analyzed, mbd and mbm, which led us to speculate that they have not changed during the last decades, neither directly due to selection pressure nor indirectly due to any correlated selection response.

Association studies reveal a major QTL on chromosome 11

Single SNP regression

Single SNP regression GWAS models showed 15 significantly associated SNPs at the Bonferroni corrected level of 5% for the mbm trait (Table 4). Four SNPs were significantly associated at the Bonferroni corrected level of 1% (Fig. 3a). Fourteen of these 15 SNPs defined a QTL on chromosome 11 between positions 31,022,855 and 31,337,157 bp (Table 4). An additional significantly associated SNP was identified in a different region on the same chromosome at position 37,136,773 bp. The most highly associated SNP was identified at position 31,004,983 bp. For the second trait analyzed (mbd), we did not observe any significant associations.

BayesB approach

We used a window size of 25 SNPs in the window-based BayesB approach. For the trait mbm, a significantly associated window that explained 15.66% of its genetic variance was identified (Fig. 3b) on chromosome 11 between positions 31,001,894 and 31,995,008 bp (Table 5). The values of $p > 0$ (proportion of models where this window was included, and therefore accounted for more than 0% genetic variance) and of $p >$ Average (proportion of

models where this window accounted for more than the amount of variance that would be explained if every window had the same effect) were both equal to 1. In the BayesB approach, the QTL reached a significance level of 1%, and thus, the significantly associated window on chromosome 11 overlapped with most of the significantly associated markers from single SNP regression. The other five possibly associated windows, each on a different chromosome, were detected below the significance threshold of 5% and explained smaller amounts of genetic variance (< 2%) (Table 5).

Subsequently, we inspected the gene content of the significantly associated QTL region on chromosome 11, and interestingly, we identified only two genes including their regulatory flanking regions in this segment, according to the NCBI Annotation Release 106: *LHCGR* and *FSHR*, which both represent obvious candidate genes for multiple births. The *LHCGR* gene (chr11:30,977,805–31,040,344) encodes the luteinizing hormone/choriogonadotropin receptor and the *FSHR* gene (chr11:31,255,649–32,450,537) encodes the follicle stimulating hormone receptor.

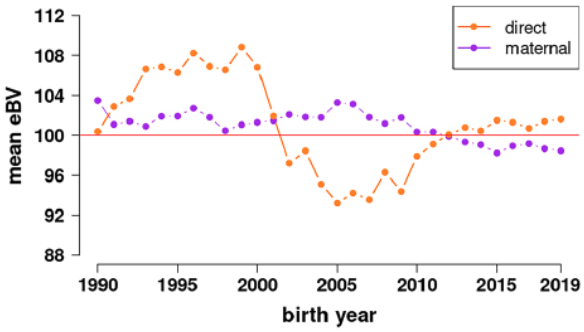


Fig. 2 Genetic trend of the estimated breeding values of multiple birth from 1990 to 2019

Haplotype analysis

We selected a 2-Mb-segment on chromosome 11 that starts at 30.5 Mb for haplotype association analyses based on the association results shown above. As discussed, this region encompasses two candidate genes including their regulatory regions. The 36 top-associated haplotypes according to their significance level from the association test are in Table 6. They encompass 19 to 37 SNPs and are located in the region between 31.00 and 31.07 Mb on chromosome 11. We selected one of the longest haplotypes (chr11:31,003,265–31,074,419) for further analysis, which included 37 SNPs and had a frequency of 0.275 in the entire genotype dataset (Table 6). An estimated additive effect of − 8.618 (± 1.188 standard error) was found for this haplotype, which indicates that it has a negative effect on the trait mbm.

Fine-mapping to the *LHCGR* locus

To unravel the most likely causal variant, we phased all genotyped samples for the chr11:31,003,265–31,074,419 haplotype using a routine imputation pipeline. Then, we merged the WGS data with the individual diplotypes for this haplotype and calculated the pairwise LD (r^2) between the haplotype and all the detected variants in the VCF file for the 880-kb region of the top-associated haplotypes (Fig. 4c). All diplotype genotypes in the WGS samples were identical across all the 36 top-associated

Table 4 Significantly associated markers from single SNP regression models for the mbm trait

Chromosome	Position	SNP	Alleles	Frequency	Variance	p-value
11	31,004,983	rs110112100	G/A	0.539	126.161	4.369 e−09
11	37,136,773	rs41579835	G/T	0.863	101.135	6.365 e−09
11	31,034,069	rs136576573	T/G	0.490	117.056	1.180 e−08
11	31,022,855	NA	A/G	0.490	117.346	1.206 e−08
11	31,037,875	rs42634817	G/A	0.490	114.435	1.771 e−08
11	31,049,877	rs135661502	C/T	0.581	112.918	2.680 e−08
11	31,329,763	rs43677285	C/T	0.633	100.127	5.626 e−08
11	31,330,363	rs43677273	G/T	0.633	100.127	5.626 e−08
11	31,332,310	rs43677261	G/A	0.633	100.127	5.626 e−08
11	31,336,575	rs43677231	A/C	0.365	99.800	5.931 e−08
11	31,337,157	rs43676629	A/G	0.365	99.800	5.931 e−08
11	31,338,042	rs136274250	C/T	0.635	99.800	5.931 e−08
11	31,060,572	rs135937618	C/A	0.581	106.172	6.378 e−08
11	31,044,741	rs137648582	A/G	0.431	103.469	7.291 e−08
11	31,070,514	rs133193362	A/C	0.419	105.105	7.383 e−08

Based on the ASR-UCD1.2/bosTau9 assembly

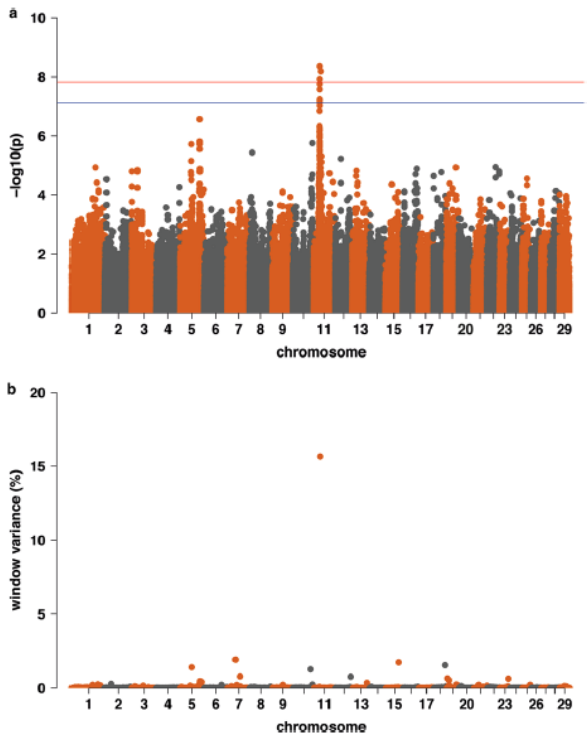


Fig. 3 Manhattan plot of GWAS for the mbm trait. **a** For the single SNP regression, with Bonferroni corrected threshold level of 5% (blue line) and 1% (red line) and **b** for the window-based BayesB approach

haplotypes (Table 6). Six variants located between 31.08 and 31.24 Mb on chromosome 11 showed r^2 values ≥ 0.7 (Table 7), but none was in perfect LD ($r^2 = 1$) with the haplotype (Fig. 4e). The variant chr11:31,089,325C>G had by far the highest r^2 (0.856) (Fig. 4e), which as the other five variants, mapped to the intergenic region between the *LHCGR* and *FSHR* genes, close to the 5′-region of *LHCGR* (Fig. 4b). None of these six variants are located in the top-associated haplotype between 31.00 and 31.07 Mb as described above (Table 7).

We performed an analysis of the sequence homology of the six variants identified in high linkage with the associated haplotype between 20 mammalian species using the UCSC genome browser to search for evidence of a putative causal role of these variants. Only one variant showed a homology score higher than zero, but its low value (0.071) indicates a low level of conservation (Table 8). The overall mean homology for the genomic region of the six variants was equal to 0.127.

Since the causative variants for the same trait can differ among breeds of the same species, we analyzed the frequencies of the variant on chromosome 11 in different breeds for all 4109 records in the vcf file provided through the 1000 Bull Genomes Project. The analysis was performed only for the variant in high LD ($r^2 \geq 0.8$) with the top-associated haplotype for all breeds that had at least 50 records. Interestingly, the frequency of this variant (chr11:31,089,325C>G) was highest for the Deutsches Schwarzbuntes Niederungsrind, the founder breed of modern Holstein cattle (Fig. 5), but frequencies higher than 10% were also found in Limousine and Red Dairy cattle. Obviously, this variant segregates in other breeds than Holstein, and also in related and unrelated breeds.

Haplotype effects on routinely recorded fertility and birth traits

To evaluate a possible co-association between the most significantly associated chr11-haplotype for mbm, haplotype effects on routinely available fertility and birth traits were estimated. We found that only two specific traits, days to first service (DFS) and stillbirth maternal (SBM), showed a significant association at the 5% level (Table 9). The associated effects between the chr11:31,003,265–31,074,419 haplotype and these traits were positive. For other birth and fertility related traits, such as non-return rate cow, a suggestive association was observed at a significance level of 15%. The estimated effect for the chr11:31,003,265–31,074,419 haplotype on non-return rate in cows was negative. Taken together, these results suggest a clear effect of the identified haplotype on different female fertility traits.

Discussion

Decreasing female fertility is an acknowledged issue in high-performance dairy production and the occurrence of multiple births has long been an undesired trait. In this study, which is based on large-scale phenotyping and genotyping data, we have detected for the first time genetic factors that contribute to this trait. The estimated breeding values for maternal and direct multiple births were used as phenotypes in association studies. Two alternative GWAS approaches were applied and identified a major QTL on bovine chromosome 11 in a region that harbours two plausible candidate genes, *LHCGR* and *FSHR*, which directly affect the female reproduction cycle. We used the top-associated haplotype to identify variants in high LD with this region in the WGS data of hundreds of Holstein genomes. We found only one variant that was located in the 5′-regulatory region of the *LHCGR* gene and represented a potential causal candidate.

The values of the estimated genetic parameters (variance components of direct and maternal effect for multiple birth) were similar to those in the literature [7, 15]. The development of a procedure to predict breeding values for binary traits, such as multiple birth, is not trivial because the phenotypic observations are not normally distributed. The use of linear mixed models that require

Table 5 Associated genome regions from the BayesB window approach for the mbm trait

Chr	Start position	End position	Number of SNPs	Proportion of explained genetic variance[a]	Cumulative proportion of varG[a]	p > 0	p > Average
11	31,001,894	31,995,008	345	15.66	15.66	1	1
7	39,005,627	39,999,299	266	1.91	17.58	0.942	0.416
15	62,002,607	62,994,714	242	1.73	19.31	0.942	0.404
18	56,015,521	56,987,121	344	1.54	20.85	0.987	0.363
5	59,005,181	59,981,866	173	1.40	22.25	0.850	0.291
10	85,003,025	85,988,287	327	1.26	23.51	0.972	0.351

Chr chromosome

p > 0 = proportion of models where this window was included, and therefore accounted for more than 0% genetic variance

p > Average = proportion of models where this window accounted for more than the amount of variance that would be explained if every window had the same effect

[a] In %

normally distributed phenotypes can lead to acceptable rankings of animals according to estimated breeding values as shown by Negussie et al. [56] but the inferred results might not be valid and predictions do not have the same support as the discrete response variable. These challenges may help explain the extreme outlier values that we obtained here for both traits (mbd and mbm) and the smaller standard deviation of the estimated breeding values for mbm. The potential benefits of using a generalized linear mixed model or a threshold model in the context of this study are the subject of future research. Previous studies have already shown how to use generalized linear mixed models or threshold models for genetic evaluations in a general context as well as for multiple birth analysis [9, 30].

In the recent past, no clear genetic trend for the two studied traits (mbd and mbm) has been observed in the Swiss data in contrast to previous findings in Norwegian cattle [28] that showed a positive trend. Hence, it seems very likely that multiple births have not been under selection in the studied Swiss Holstein population. Thus, the prediction of breeding values could represent an important selection tool for reducing the occurrence of multiple births, which will hopefully lead to improved animal health and welfare. Availability of these results will allow to implement selection programs in the local breeding schemes.

In our dataset, we found only one clear QTL for the maternal trait, which underlines the essential role of the maternal component for the complex multiple birth trait. This QTL on chromosome 11 was detected by two GWAS approaches: single SNP regression and a window-based BayesB approach. Although single marker detection can be useful to detect many associations, only a small fraction of the genetic variance of quantitative traits can be significantly highlighted and identified with this approach [57–59]. The applied BayesB approach, used as a method that fits all the markers as random effects simultaneously, can account for most of the genetic variance [60–62] as well as a window-based approach that captures most of the variability at an associated trait locus [50]. The fact that the same QTL was detected by these two approaches provides strong evidence for the corresponding genomic region. The identified QTL explained 15.66% of the genetic variance of the trait. Furthermore, only one significantly associated SNP was detected by single SNP regression and was located downstream on the same chromosome, which indicates a single possibly false positive association signal. However, since mbm represents a classical polygenic trait, additional QTL of smaller effect might be detected with a larger sample size. The non-significantly associated segments on five other chromosomes observed in the window-based BayesB approach might represent suggestive QTL that need to be confirmed in the future. The theory of a polygenic trait is supported by previous studies, which revealed multiple QTL on various chromosomes [15, 31–38]. Regarding chromosome 11, only combined linkage-linkage disequilibrium analysis for North American Holstein sires revealed a QTL for twinning rate on the same chromosome but located in a different segment [31]. Those results were confirmed later in an additional American Holstein cattle population [34] but they were not validated in the USDA Meat Animal Research Center (USMARC) special herd that was selected for twinning rate [63]. Interestingly in the cattle QTL database, no QTL has been reported in the chromosome 11 region described here for the trait of interest [64]. Furthermore, this supports our assumption that this specific genome region, which is associated with multiple birth, has not been under selection in Holstein cattle, since the allele frequency has not changed significantly over time.

Table 6 Top associated haplotypes from the regression analysis on chromosome 11 for the mbm trait

Start position	End position	Number of SNPs	Frequency	Effect	p-value
31,003,265	31,074,419	37	0.275	− 8.618 (1.188)	5.435 e−13
31,003,265	31,072,259	35	0.275	− 8.618 (1.188)	5.435 e−13
31,004,983	31,076,106	37	0.275	− 8.618 (1.188)	5.435 e−13
31,004,983	31,073,807	35	0.275	− 8.618 (1.188)	5.435 e−13
31,013,207	31,074,419	35	0.275	− 8.618 (1.188)	5.435 e−13
31,013,207	31,072,259	33	0.275	− 8.618 (1.188)	5.435 e−13
31,018,127	31,076,106	35	0.275	− 8.618 (1.188)	5.435 e−13
31,018,127	31,073,807	33	0.275	− 8.618 (1.188)	5.435 e−13
31,020,736	31,074,419	33	0.275	− 8.618 (1.188)	5.435 e−13
31,020,736	31,072,259	31	0.275	− 8.618 (1.188)	5.435 e−13
31,022,855	31,076,106	33	0.275	− 8.618 (1.188)	5.435 e−13
31,022,855	31,073,807	31	0.275	− 8.618 (1.188)	5.435 e−13
31,023,833	31,074,419	31	0.275	− 8.618 (1.188)	5.435 e−13
31,023,833	31,072,259	29	0.275	− 8.618 (1.188)	5.435 e−13
31,034,069	31,076,106	31	0.275	− 8.618 (1.188)	5.435 e−13
31,034,069	31,073,807	29	0.275	− 8.618 (1.188)	5.435 e−13
31,037,875	31,074,419	29	0.275	− 8.618 (1.188)	5.435 e−13
31,037,875	31,072,259	27	0.275	− 8.618 (1.188)	5.435 e−13
31,041,776	31,076,106	29	0.275	− 8.618 (1.188)	5.435 e−13
31,041,776	31,073,807	27	0.275	− 8.618 (1.188)	5.435 e−13
31,043,114	31,074,419	27	0.275	− 8.618 (1.188)	5.435 e−13
31,043,114	31,072,259	25	0.275	− 8.618 (1.188)	5.435 e−13
31,044,741	31,076,106	27	0.275	− 8.618 (1.188)	5.435 e−13
31,044,741	31,073,807	25	0.275	− 8.618 (1.188)	5.435 e−13
31,046,036	31,074,419	25	0.275	− 8.618 (1.188)	5.435 e−13
31,046,036	31,072,259	23	0.275	− 8.618 (1.188)	5.435 e−13
31,049,190	31,076,106	25	0.275	− 8.618 (1.188)	5.435 e−13
31,049,190	31,073,807	23	0.275	− 8.618 (1.188)	5.435 e−13
31,049,877	31,074,419	23	0.275	− 8.618 (1.188)	5.435 e−13
31,049,877	31,072,259	21	0.275	− 8.618 (1.188)	5.435 e−13
31,054,311	31,076,106	23	0.275	− 8.618 (1.188)	5.435 e−13
31,054,311	31,073,807	21	0.275	− 8.618 (1.188)	5.435 e−13
31,055,164	31,074,419	21	0.275	− 8.618 (1.188)	5.435 e−13
31,055,164	31,072,259	19	0.275	− 8.618 (1.188)	5.435 e−13
31,055,946	31,076,106	21	0.275	− 8.618 (1.188)	5.435 e−13
31,055,946	31,073,807	19	0.275	− 8.618 (1.188)	5.435 e−13

In the literature, there is little evidence for a major QTL for multiple birth traits in cattle. Only two studies using data from North American Holstein and Norwegian cattle [32, 36] suggested a positional candidate gene with a possible major impact on the trait, i.e. *IGF1*. The region on chromosome 5 containing *IGF1* does not overlap with the suggestive QTL detected in our analysis. However, one can speculate that a larger and/or global dataset might improve such analyses and reveal additional QTL in Holstein and/or other cattle breeds.

The two genes, *LHCGR* and *FSHR*, which are located in the identified QTL region, are obvious candidate genes for multiple birth since they encode receptors of three essential hormones for female reproduction: luteinizing hormone (LH), choriogonadotropin, and follicle stimulating hormone (FSH). This is the first study that detects these genes as candidates for multiple birth in cattle.

Human chorionic gonadotropin is a hormone that is involved in the maternal recognition of pregnancy and is produced by trophoblast cells that surround the growing embryo, whereas LH is a hormone that is produced by

gonadotropic cells in the anterior pituitary gland under the regulation of the gonadotropin-releasing hormone from the hypothalamus. Mutations in the human *LHCGR* gene that is expressed in the testis and ovary lead to disorders of the development of the male secondary sexual character, including familial male precocious puberty, also known as testotoxicosis, hypogonadotropic hypogonadism, Leydig cell adenoma with precocious puberty, and male pseudohermaphroditism with Leydig cell hypoplasia (OMIM 152790). In females, an acute increase in LH (LH peak) triggers the ovulation by initiating meiosis II in the oocyte at the point of ovulation and leads to follicle rupture and subsequent development of the corpus luteum [65, 66]. Interestingly, the concentrations of LH in the blood and plasma did not differ between selected and unselected bovine females for twin ovulations and dizygotic twins [67]. Therefore, we hypothesize that the identified variant in the 5′-regulatory region of the bovine *LHCGR* gene might alter the expression of the encoded receptor in the ovary cell and thereby influence the ovulation rate explaining the effect on the studied multiple birth trait.

The FSH receptor is a transmembrane receptor that interacts with FSH and represents a G protein-coupled receptor that is expressed in the ovary, testis, and uterus. Mutations in the human *FSHR* gene cause ovarian dysgenesis type 1, and also the ovarian hyperstimulation syndrome (OMIM 136435). In the ovary, the FSH receptor is necessary for follicular development and is expressed on the granulosa cells and during the luteal phase in the secretory endometrium of the uterus. Although this gene can also represent a plausible functional candidate for multiple birth, the genetic findings presented here and the genomic localization of the identified potential causal variant provide stronger support for the *LHCGR* gene. However, an effect on *FSHR* expression cannot be fully ruled out. Interestingly, a study of the Flemish and Dutch human population provided some evidence that the same homologous region on chromosome 2 carrying the candidate genes *FSHR* and *LHCGR* had an effect on multiple birth [68]. In addition, a potentially functional variant in the 5′ untranslated region of *FSHR* has been reported in a single family [69], and two variants in the protein-coding area of *FSHR* have been detected in one woman with two twin pregnancies [70]. However, these findings were rejected by a study including 21 mothers with twins [71]. Recently, comprehensive GWAS studies for twinning rate in humans have become available, which revealed variants associated with the maternal trait in the *FSHB* and the *SMAD3* genes [72], and several QTL for the direct trait [73]. In general, the rate of monozygotic twins in humans is higher than in cattle, which might be a reason for the higher incidence of the direct effect. Our new findings could be used as a reference for other species, including humans.

Unfortunately, LD-based filtering of sequence variants within the critical segment did not filter out a single variant in perfect LD with the top-associated haplotype on chromosome 11. However, several non-coding variants showed high values of LD ($r^2 \geq 0.7$) that formed a 160-kb block. This was not unexpected due to the known relatively long-range LD pattern in cattle populations [74]. It has been shown that a single genomic region can harbor several QTL for a polygenic trait [75]. However, it is more likely that the identified segment that showed LD between the haplotype and several variants was due to LD between adjacent variants, while only one of them is the true causal variant. The effect on gene expression or regulation could also be influenced by multiple variants. The genomic localization suggested an effect of the variants on the regulation and expression of the *LHCGR* gene. Since the candidate variant (chr11:31,089,325C>G) segregates in various breeds at low frequency, it represents an old variation that most likely occurred before the formation of modern breeds. Ideally, a cross-breeding validation analysis should be performed. For most of the breeds analyzed here, neither phenotypes nor genotypes are available to evaluate a possible effect on multiple birth traits. In the data that was available for Brown Swiss and Simmental cattle, this variant showed low minor allele frequencies, which could be the reason why we were not able to confirm an effect in those populations (data not shown). Hence, we speculate that the pattern observed here is similar to that previously reported for the *DGAT1*

(See figure on next page.)

Fig. 4 Graphical representation of the associated QTL region and linkage disequilibrium (LD) analysis results for the mbm trait. **a** Screenshot of the region on bovine chromosome 11 between 30.5 and 32.5 Mb from the NCBI Genome Browser including the localization of the genes in the region. **b** Screenshot of the region on bovine chromosome 11 between 30.95 and 31.5 Mb from the NCBI Genome Browser including the localization of the candidate genes *LHCGR* and *FSHR*. The purple bar shows the localization of the top-associated haplotype. The red star represents the variant with the highest LD score. **c** The resulting haplotypes from the haplotype regression analysis are shown with their localization. The 880-kb region with the top-associated haplotypes from 30.47 to 31.35 Mb used for subsequent fine-mapping is highlighted in red. **d** Heatmap showing the LD in the region between 30.47 and 31.35 Mb on bovine chromosome 11 highlighting the haplotype that was added as an additional variant. The localization and chromosomal orientation of *LHCGR* and *FSHR* are displayed (black arrows). **e** The variants in $r^2 \geq 0.6$ with the top-associated haplotype are highlighted and shown with their position on bovine chromosome 11

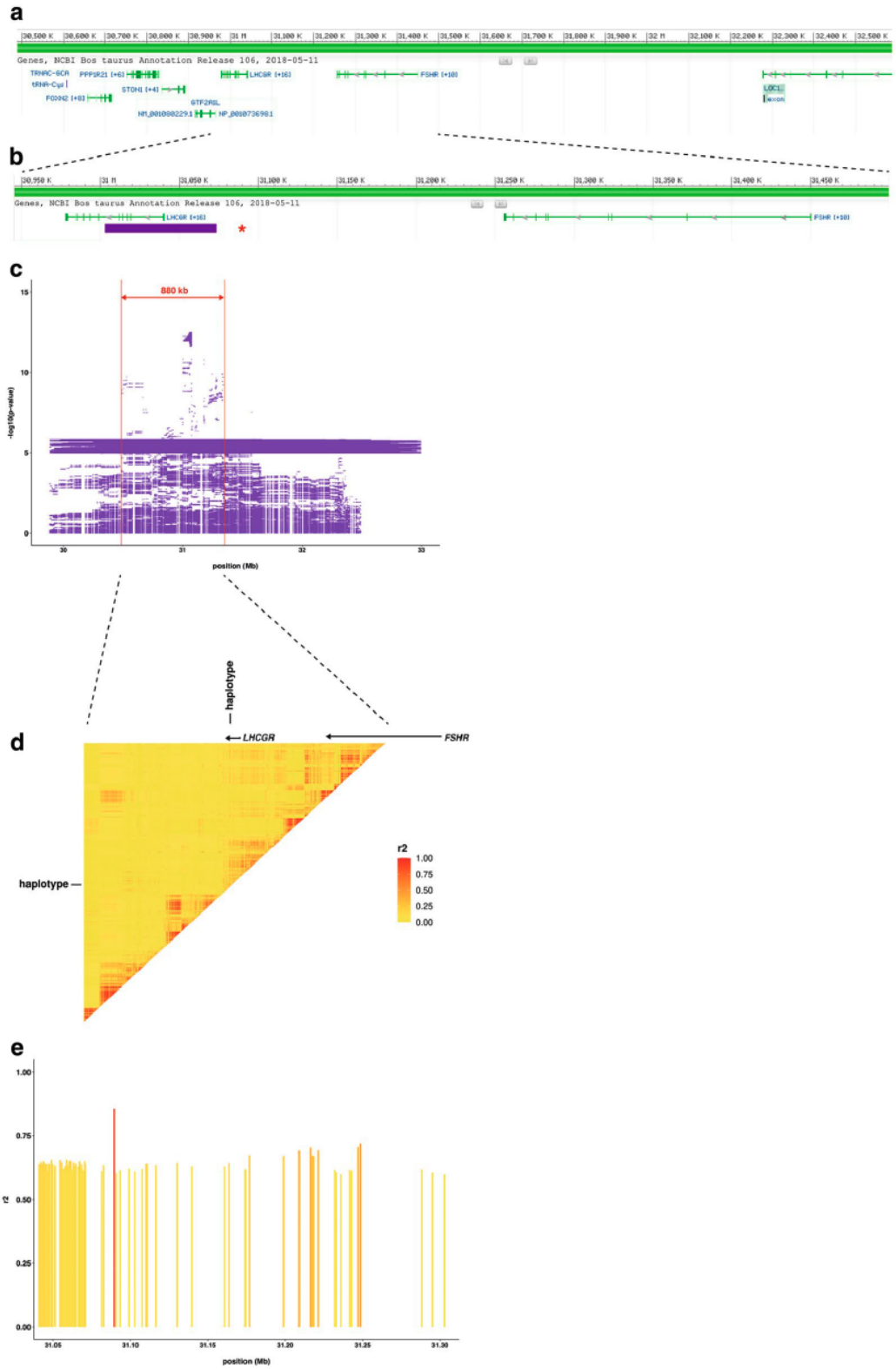


Fig. 4 (See legend on previous page.)

Table 7 Results of the linkage disequilibrium (LD) analysis for the mbm trait

Variant on chromosome 11[a]	Minor allele frequency	LD (r^2)
31,089,325C>G	0.286	0.856
31,216,357C>T	0.334	0.705
31,246,990AGCC>-	0.326	0.707
31,248,462G>A	0.326	0.720
31,248,573C>T	0.333	0.706
31,248,580G>T	0.333	0.706

[a] ASR-UCD1.2/bosTau9 assembly

Table 8 Results of the sequence homology analysis for all the variants with $r^2 \geq 0.7$ from linkage disequilibrium analysis

Variant on chromosome 11[a]	Homolog region for human genome[b]	r^{2c}	Homology score[d]
31,246,990AGCC>-	2:48,957,377	0.707	0.071
31,089,325C>G	2:48,806,473	0.856	0
31,216,357C>T	2:48,930,243	0.705	0
31,248,462G>A	2:48,957,970	0.720	0
31,248,573C>T	2:48,957,971	0.706	0
31,248,580G>T	2:48,957,971	0.706	0

[a] Based on the cattle assembly ASR-UCD1.2/bosTau9

[b] Using the USCS genome browser and human assembly GRCh38/hg38

[c] From a previous linkage disequilibrium analysis with top-associated haplotype

[d] Sequence homology between 20 mammalian species using the UCSC genome browser, score range from 0 to 1

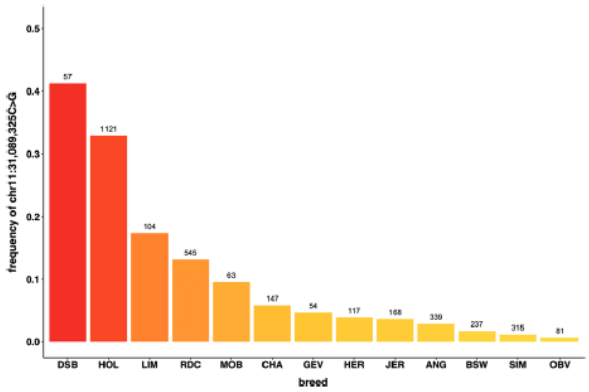


Fig. 5 Frequencies among breeds for the top-linked variant. Calculated from the data of 1000 Bull Genomes Project for the chr11:31,089,235 C>G variant. Breed abbreviations: *ANG* Angus, *BSW* Brown Swiss, *CHA* Charolaise, *DSB* Deutsches Schwarzbuntes Niederungsrind, *GEV* Gelbvieh, *HER* Hereford, *HOL* Holstein, *JER* Jersey, *LIM* Limousine, *MOB* Montbéliarde, *OBV* Original Braunvieh, *RDC* Red Dairy cattle, *SIM* Simmental

p.Lys232Ala mutation that determines fat content in milk in Holstein and Brown Swiss, for which multiple variants showed a similar effect [76]. None of the six variants that had an LD value higher than 0.7 were located in the interval of the top-associated haplotype. They map downstream to this haplotype, but are still located in a block of elevated LD as shown in the presented heatmap (Fig. 4d). The top-associated haplotype identified is part of this LD block. However, to validate these findings, a larger dataset and other approaches such as differential gene expression analysis or targeted genome editing are necessary.

Although the association reported for the SNPs found in this study strongly suggests a close connection to the *LHCGR* gene, a causal analysis based on Mendelian randomization (MR) as described in previous studies is required to make a definitive statement about the causality of the reported candidate location [77, 78]. In an MR analysis, the genetic variants (such as the chr11:31,089,325C>G variant detected here) are used as instrumental variables (IV) to disentangle the possibly confounded relationship between intermediate phenotypes such as *LHCGR* gene expression levels and our trait of interest (mbm). The three key assumptions for a genetic variant to qualify as an IV are shown in Fig. 6 and can be described as follows. First, the genetic variant (e.g. chr11:31,089,325C>G) has to be unrelated to typical confounding factors such as environmental factors, which is considered in our analysis, because the known confounding factors are included as fixed effects in our linear mixed effects model and therefore the influence of the confounding factors on the outcome trait (mbm) is taken into account. In the graph, the unrelatedness is encoded by the missing arrows between the nodes of the confounders and the genetic variant (Fig. 6). Second, the chr11:31,089,325C>G variant has to be associated with the exposure of the intermediate phenotype which in our case corresponds to the postulated effect on *LHCGR* gene expression. This means that the nodes of this genetic variant are connected to the intermediate phenotype. Third, conditional on confounders and exposure, the genetic variant and the outcome (mbm) are independent. Hence, if we know the levels of the exposure and the confounders, the genetic variant does not provide any additional information to the outcome trait (mbm). To test these three assumptions, we would need to have the observed values for the exposure of the intermediate phenotype which in our case are the expression levels of *LHCGR* , in addition to the genotypes of the genetic variant and the recorded events of our outcome trait (mbm). Unfortunately, the data for the intermediate phenotypes are not available in this study. The additional collection of intermediate phenotypes, such as the expression levels of

Table 9 Effect of the top-associated chromosome 11 haplotype on other female fertility and birth traits

Trait group	Trait	Effect[a]	p-value[b]
Female fertility	Days to first service	0.664 (0.277)	0.017
	Interval between first and last insemination heifers	0.653 (0.495)	0.187
	Interval between first and last insemination cows	− 0.091 (0.257)	0.724
	Non-return rate heifers[c]	0.264 (0.350)	0.450
	Non-return rate cows[c]	− 0.436 (0.285)	0.125
Birth	Birth weight direct	− 0.210 (0.438)	0.632
	Birth weight maternal	− 0.687 (0.509)	0.177
	Calving ease direct	0.218 (0.199)	0.274
	Calving ease maternal	− 0.002 (0.298)	0.996
	Gestation length direct	− 0.071 (0.434)	0.870
	Gestation length maternal	− 0.580 (0.511)	0.256
	Stillbirth direct	1.020 (0.772)	0.186
	Stillbirth maternal	1.297 (0.524)	0.013

[a] Standard error of the effect in brackets

[b] Not corrected for multiple comparisons

[c] After 56 days

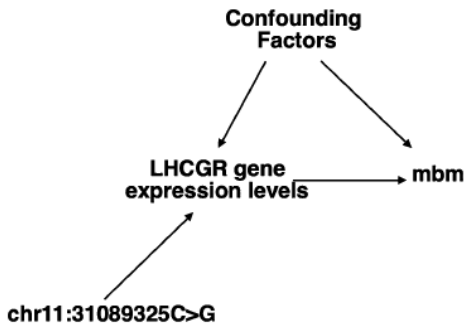


Fig. 6 Direct acyclic graph denoting causality associations. The bovine chr11:31,089,325C>G genetic variant used as an instrumental variable in a Mendelian randomization analysis

LHCGR in order to be able to make a more definite statement about the causes behind the outcome trait of mbm, requires further research.

We compared the sequence homology of the six variants identified in high linkage with the associated haplotype between 20 mammalian species, which resulted in an homology score that is a measure of supporting evidence for a variant. The overall mean of the homology for the whole region for variants with an r^2 value ≥ 0.7 was low, which is expected since our observed segment is an intergenic and non-coding region. Regarding the single variants, we did not observe any variants that showed high homology scores. Therefore, no further information can be obtained from this analysis regarding a potential causal role.

The detected associated effect of the identified haplotype on the studied trait mbm was negative. Hence, female haplotype carriers are expected to have a decreased incidence for multiple birth. Evaluation of the possible effects of the haplotype on other available birth and fertility traits revealed significant effects on days to first service and maternal stillbirth (Table 9). We found a positive significant effect on the female fertility trait, days to first service, which is associated with a larger interval between calving and first insemination, and a negative suggestive effect on non-return rate in cows, which is associated with a decreased insemination success. These two observations might provide supporting evidence for the influence of this genomic segment on female fertility, in general. In addition, we observed a positive effect on the birth trait 'maternal stillbirth', which results in a higher proportion of live birth events. Taken together, these observations support a most likely negative effect of the identified haplotype on female fertility including multiple births.

Conclusions

Our aim was to undertake a comprehensive genetic analysis of multiple births in Swiss Holstein cattle. By analyzing large-scale genotype and phenotype data, we have detected, for the first time, a major QTL for maternal multiple birth on chromosome 11. One non-coding most likely regulatory variant located in the 5′-region of the *LHCGR* gene was linked to the top-associated haplotype. This suggests that selecting against multiple births is possible, but a putative

negative effect on other female fertility traits needs to be considered. These findings improve our understanding of the genetic architecture that underlies multiple birth in mammals and female fertility, in general, but further studies are needed to strengthen the evidence for a causal relationship between the detected *LHCGR* locus and the phenotype.

Acknowledgements
The authors are grateful to the Swiss cattle breeding organizations (swissherdbook and Holstein Switzerland) for providing phenotypic and genomic data as well as the 1000 Bull Genomes Project for providing WGS data. Christine F. Baes is acknowledged for proofreading.

Authors' contributions
SW carried out the analyses, visualized the results, and drafted the manuscript. FRS co-designed the study, contributed data and data preparation as well as critically revised the manuscript. PVR co-designed the study contributed to the development of breeding value estimation and was a technical advisor. IMH performed bioinformatical methods for data preparation. MS executed the frequencies among breeds calculation and analysis. CD co-designed the study and revised the manuscript. All authors participated in writing the manuscript. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The sequence data for all animals obtained from the 1000 Bull Genomes Project is available at the EVA (www.ebi.ac.uk/eva/). The SNP data analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute of Genetics, Vetsuisse Faculty, University of Bern, 3012 Bern, Switzerland. [2]Qualitas AG, 6300 Zug, Switzerland.

Received: 2 February 2021 Accepted: 25 June 2021
Published online: 03 July 2021

References

1. Miglior F, Fleming A, Malchiodi F, Brito LF, Martin P, Baes CF. A 100-year review: identification and genetic selection of economically important traits in dairy cattle. J Dairy Sci. 2017;100:10251–71.
2. Philipsson J. Genetic aspects of female fertility in dairy cattle. Livest Prod Sci. 1981;8:307–19.
3. Liu A, Wang Y, Sahana G, Zhang Q, Liu L, Lund MS, et al. Genome-wide association studies for female fertility traits in Chinese and Nordic Holsteins. Sci Rep. 2017;7:8487.
4. Frischknecht M, Bapst B, Seefried FR, Signer-Hasler H, Garrick D, Stricker C, et al. Genome-wide association studies of fertility and calving traits in Brown Swiss cattle using imputed whole-genome sequences. BMC Genomics. 2017;18:910.
5. Georges M, Charlier C, Hayes B. Harnessing genomic information for livestock improvement. Nat Rev Genet. 2019;20:135–56.
6. Atteneder V. Analyse der Zwillings- und Mehrlingsgeburten in der Österreichischen Milchviehpopulation. 2007. https://epub.boku.ac.at/obvbokhs/download/pdf/1035817?originalFilename=true. Accessed 18 Dec 2020.
7. Ghavi Hossein-Zadeh N, Nejati-Javaremi A, Miraei-Ashtiani SR, Kohram H. Estimation of variance components and genetic trends for twinning rate in Holstein dairy cattle of Iran. J Dairy Sci. 2009;92:3411–21.
8. Atashi H, Zamiri MJ, Sayadnejad MB. The effect of maternal inbreeding on incidence of twinning, dystocia and stillbirth in Holstein cows of Iran. Iran J Vet Res. 2012;13:93–9.
9. Johanson JM, Berger PJ, Kirkpatrick BW, Dentine MR. Twinning rates for North American Holstein sires. J Dairy Sci. 2001;84:2081–8.
10. Lett BM, Kirkpatrick BW. Short communication: heritability of twinning rate in Holstein cattle. J Dairy Sci. 2018;101:4307–11.
11. Miyake Y-I, Miyoshi K, Moriya H, Matsui M, Haneda S. Studies on the accident rate in single and multiple births in dairy cows. Jpn J Large Anim Clin. 2010;1:5–9.
12. Masuda Y, Baba T, Suzuki M. Genetic analysis of twinning rate and milk yield using a threshold-linear model in Japanese Holsteins. Anim Sci J. 2015;86:31–6.
13. Murillo-Barrantes J, Estrada-König S, Rojas-Campos J, Bolaños-Segura M, Valverde-Altamirano E, Romero-Zúñiga JJ. Factores asociados con partos gemelares en vacas de fincas lecheras especializadas de Costa Rica. Rev Ciencias Vet. 2010;28:7–21.
14. Moioli B, Steri R, Marchitelli C, Catillo G, Buttazzoni L. Genetic parameters and genome-wide associations of twinning rate in a local breed, the Maremmana cattle. Animal. 2017;11:1660–6.
15. Weller JI, Golik M, Seroussi E, Ron M, Ezra E. Detection of quantitative trait loci affecting twinning rate in Israeli Holsteins by the daughter design. J Dairy Sci. 2008;91:2469–74.
16. Silva del Río N, Kirkpatrick BW, Fricke PM. Observed frequency of monozygotic twinning in Holstein dairy cattle. Theriogenology. 2006;66:1292–9.
17. Echternkamp SE, Gregory KE. Effects of twinning on gestation length, retained placenta, and dystocia. J Anim Sci. 1999;77:39–47.
18. Gregory KE, Echternkamp SE, Dickerson GE, Cundiff LV, Koch RM, van Vleck LD. Twinning in cattle: III. Effects of twinning on dystocia, reproductive traits, calf survival, calf growth and cow productivity. J Anim Sci. 1990;68:3133–44.
19. Pardon B, Vertenten G, Cornillie P, Schauvliege S, Gasthuys F, van Loon G, et al. Left abomasal displacement between the uterus and rumen during bovine twin pregnancy. J Vet Sci. 2012;13:437–40.
20. Fricke PM. Twinning in dairy cattle. Prof Anim Sci. 2001;17:61–7.
21. Silva-del-Río N, Fricke PM, Grummer RR. Effects of twin pregnancy and dry period feeding strategy on milk production, energy balance, and metabolic profiles in dairy cows. J Anim Sci. 2010;88:1048–60.
22. Andreu-Vázquez C, Garcia-Ispierto I, Ganau S, Fricke PM, López-Gatius F. Effects of twinning on the subsequent reproductive performance and productive lifespan of high-producing dairy cows. Theriogenology. 2012;78:2061–70.
23. Nielen M, Schukken YH, Scholl DT, Wilbrink HJ, Brand A. Twinning in dairy cattle: a study of risk factors and effects. Theriogenology. 1989;32:845–62.
24. Hossein-Zadeh NG. The effect of twinning on milk yield, dystocia, calf birth weight and open days in Holstein dairy cows of Iran. J Anim Physiol Anim Nutr. 2010;94:780–7.
25. Mee JF, Berry DP, Cromie AR. Risk factors for calving assistance and dystocia in pasture-based Holstein-Friesian heifers and cows in Ireland. Vet J. 2011;187:189–94.
26. Silva Del Río N, Stewart S, Rapnicki P, Chang YM, Fricke PM. An observational analysis of twin births, calf sex ratio, and calf mortality in Holstein dairy cattle. J Dairy Sci. 2007;90:1255–64.
27. Fitzgerald AM, Berry DP, Carthy T, Cromie AR, Ryan DP. Risk factors associated with multiple ovulation and twin birth rate in Irish dairy and beef cattle. J Anim Sci. 2014;92:966–73.
28. Karlsen A, Ruane J, Klemetsdal G, Heringstad B. Twinning rate in Norwegian cattle: frequency, (co)variance components, and genetic trends. J Anim Sci. 2000;78:15–20.
29. Allan MF, Kuehn LA, Cushman RA, Snelling WM, Echternkamp SE, Thallman RM. Confirmation of quantitative trait loci using a low-density single nucleotide polymorphism map for twinning and ovulation rate on bovine chromosome 5. J Anim Sci. 2009;87:46–56.

30. McGovern SP, Weigel DJ, Fessenden BC, Gonzalez-Peña D, Vukasinovic N, McNeel AK, et al. Genomic prediction for twin pregnancies. Animals (Basel). 2021;11:843.
31. Kim ES, Berger PJ, Kirkpatrick BW. Genome-wide scan for bovine twinning rate QTL using linkage disequilibrium. Anim Genet. 2009;40:300–7.
32. Kim ES, Shi X, Cobanoglu O, Weigel K, Berger PJ, Kirkpatrick BW. Refined mapping of twinning-rate quantitative trait loci on bovine chromosome 5 and analysis of insulin-like growth factor-1 as a positional candidate gene. J Anim Sci. 2009;87:835–43.
33. Bierman CD, Kim E, Weigel K, Berger PJ, Kirkpatrick BW. Fine-mapping quantitative trait loci for twinning rate on Bos taurus chromosome 14 in North American Holsteins. J Anim Sci. 2010;88:2556–64.
34. Bierman CD, Kim E, Shi XW, Weigel K, Jeffrey Berger P, Kirkpatrick BW. Validation of whole genome linkage-linkage disequilibrium and association results, and identification of markers to predict genetic merit for twinning. Anim Genet. 2010;41:406–16.
35. Meuwissen THE, Karlsen A, Lien S, Olsaker I, Goddard ME. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics. 2002;161:373–9.
36. Lien S, Karlsen A, Klemetsdal G, Våge DI, Olsaker I, Klungland H, et al. A primary screen of the bovine genome for quantitative trait loci affecting twinning rate. Mamm Genome. 2000;11:877–82.
37. Cobanoglu O, Berger PJ, Kirkpatrick BW. Genome screen for twinning rate QTL in four North American Holstein families. Anim Genet. 2005;36:303–8.
38. Cruickshank J, Dentine MR, Berger PJ, Kirkpatrick BW. Evidence for quantitative trait loci affecting twinning rate in North American Holstein cattle. Anim Genet. 2004;35:206–12.
39. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2020. https://www.R-project.org/. Accessed 18 Dec 2020.
40. RStudio Team. RStudio: integrated development for R. Boston: RStudio Inc. 2016. http://www.rstudio.com/. Accessed 18 Dec 2020.
41. Neumaier A, Groeneveld E. Restricted maximum likelihood estimation of covariances in sparse linear models. Genet Sel Evol. 1998;30:3–26.
42. MiX99 Development Team. MiX99: a software package for solving large mixed model equations. Release 17.11. 2017. Jokioi: Natural Resources Institute Finland (Luke); http://www.luke.fi/mix99. Accessed 18 Dec 2020.
43. Tier B, Meyer K. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. J Anim Breed Genet. 2004;121:77–89.
44. Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. 2009;41:55.
45. Sargolzaei M, Chesnais JP, Schenkel FS. FImpute—an efficient imputation algorithm for dairy cattle populations. J Dairy Sci. 2011;94:421.
46. Sargolzaei M. Ontario Veterinary College, University of Guelph. https://ovc.uoguelph.ca/pathobiology/people/faculty/Mehdi-Sargolzaei. Accessed 31 Mar 2021.
47. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
48. Fernando R, Toosi A, Wolc A, Garrick D, Dekkers J. Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. J Agric Biol Environ Stat. 2017;22:172–93.
49. Putz A. GenSel, GitHub. 2021. https://github.com/austin-putz/GenSel. Accessed 31 Mar 2021.
50. Fernando RL, Garrick DJ. Bayesian methods applied to GWAS. In: Gondro C, van der Werf J, Hayes B, editors. Genome-wide association studies and genomic prediction. New York: Springer; 2013. p. 237–74.
51. Pausch H, Ammermüller S, Wurmser C, Hamann H, Tetens J, Drögemüller C, et al. A nonsense mutation in the *COL7A1* gene causes epidermolysis bullosa in Vorderwald cattle. BMC Genet. 2016;17:149.
52. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
53. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
54. Hayes BJ, Daetwyler HD. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. Annu Rev Anim Biosci. 2019;7:89–102.
55. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
56. Negussie E, Strandén I, Mäntysaari EA. Genetic analysis of liability to clinical mastitis, with somatic cell score and production traits using bivariate threshold–linear and linear–linear models. Livest Sci. 2008;117:52–9.
57. Visscher PM, Yang J, Goddard MEA. A Commentary on 'Common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). Twin Res Hum Genet. 2010;13:517–24.
58. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
59. Maher B. Personal genomes: the case of the missing heritability. Nature. 2008;456:18–21.
60. Fan B, Onteru SK, Du Z-Q, Garrick DJ, Stalder KJ, Rothschild MF. Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. PLoS One. 2011;6:e14726.
61. Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 2010;6:e1001139.
62. Onteru SK, Fan B, Nikkilä MT, Garrick DJ, Stalder KJ, Rothschild MF. Whole-genome association analyses for lifetime reproductive traits in the pig. J Anim Sci. 2011;89:988–95.
63. Kirkpatrick BW, Thallman RM, Kuehn LA. Validation of SNP associations with bovine ovulation and twinning rate. Anim Genet. 2019;50:259–61.
64. Cattle QTL Database. 2020. https://www.animalgenome.org/cgi-bin/QTLdb/BT/traitmap?trait_ID=1078. Accessed 18 Dec 2020.
65. Vinet A, Drouilhet L, Bodin L, Mulsant P, Fabre S, Phocas F. Genetic control of multiple births in low ovulating mammalian species. Mamm Genome. 2012;23:727–40.
66. Qiao J, Han B. Diseases caused by mutations in luteinizing hormone/chorionic gonadotropin receptor. Prog Mol Biol Transl Sci. 2019;161:69–89.
67. Echternkamp SE. Endocrinology of increased ovarian folliculogenesis in cattle selected for twin births. J Anim Sci. 2000;77:1–20.
68. Derom C, Jawaheer D, Chen WV, McBride KL, Xiao X, Amos C, et al. Genome-wide linkage scan for spontaneous DZ twinning. Eur J Hum Genet. 2006;14:117–22.
69. Painter JN, Willemsen G, Nyholt D, Hoekstra C, Duffy DL, Henders AK, et al. A genome wide linkage scan for dizygotic twinning in 525 families of mothers of dizygotic twins. Hum Reprod. 2010;25:1569–80.
70. Al-Hendy A, Moshynska O, Saxena A, Feyles V. Association between mutations of the follicle-stimulating-hormone receptor and repeated twinning. Lancet. 2000;356:914.
71. Montgomery GW, Duffy DL, Hall J, Kudo M, Martin NG, Hsueh AJ. Mutations in the follicle-stimulating hormone receptor and familial dizygotic twinning. Lancet. 2001;357:773–4.
72. Mbarek H, Steinberg S, Nyholt DR, Gordon SD, Miller MB, McRae AF, et al. Identification of common genetic variants influencing spontaneous dizygotic twinning and female fertility. Am J Hum Genet. 2016;98:898–908.
73. Mbarek H, van de Weijer MP, van der Zee MD, Ip HF, Beck JJ, Abdellaoui A, et al. Biological insights into multiple birth: genetic findings from UK Biobank. Eur J Hum Genet. 2019;27:970–9.
74. de Roos APW, Hayes BJ, Spelman RJ, Goddard ME. Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. Genetics. 2008;179:1503–12.
75. Thaller G, Krämer W, Winter A, Kaupe B, Erhardt G, Fries R. Effects of *DGAT1* variants on milk production traits in German cattle breeds. J Anim Sci. 2003;81:1911–8.
76. Winter A, Kramer W, Werner FAO, Kollers S, Kata S, Durstewitz G, et al. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (*DGAT1*) with variation at a quantitative trait locus for milk fat content. Proc Natl Acad Sci USA. 2002;99:9300–5.
77. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. Stat Methods Med Res. 2007;16:309–30.
78. Sheehan NA, Didelez V, Burton PR, Tobin MD. Mendelian randomisation and causal inference in observational epidemiology. PLoS Med. 2008;5:e177.

Publisher's Note

Associated regions for multiple birth in Brown Swiss and Original Braunvieh cattle on chromosomes 15 and 11

Journal:	Animal Genetics
Manuscript status:	published
Contributions:	phenotyping data preparation, data analyses, visualization of the results, writing original draft and revisions
Displayed version:	published version
DOI:	https://doi.org/10.1111/age.13229

Received: 29 March 2022 | Accepted: 4 June 2022

DOI: 10.1111/age.13229

RESEARCH ARTICLE

ANIMAL GENETICS WILEY

Associated regions for multiple birth in Brown Swiss and Original Braunvieh cattle on chromosomes 15 and 11

Sarah Widmer[1] | Franz R. Seefried[2] | Peter von Rohr[2] | Irene M. Häfliger[1] | Mirjam Spengeler[2] | Cord Drögemüller[1]

[1]Vetsuisse Faculty, Institute of Genetics, University of Bern, Bern, Switzerland

[2]Qualitas AG, Zug, Switzerland

Correspondence
Sarah Widmer, Institute of Genetics, Vetsuisse Faculty, University of Bern, Bremgartenstrasse 109a, 3012 Bern, Switzerland.
Email: sarah.widmer@vetsuisse.unibe.ch

Funding information
Open access funding provided by Universitat Bern

Abstract

Twin and multiple births have negative effects on the performance and health of cows and calves. To decipher the genetic architecture of this trait in the two Swiss Brown Swiss cattle populations, we performed various association analyses based on de-regressed breeding values. Genome-wide association analyses were executed using ~600 K imputed SNPs for the maternal multiple birth trait in ~3500 Original Braunvieh and ~7800 Brown Swiss animals. Significantly associated QTL were observed on different chromosomes for both breeds. We have identified on chromosome 11 a QTL that explains ~6% of the total genetic variance of the maternal multiple birth trait in Original Braunvieh. For the Brown Swiss breed, we have discovered a QTL on chromosome 15 that accounts for ~4% of the total genetic variance. For Original Braunvieh, subsequent haplotype analysis revealed a 90-kb window on chromosome 11 at 88 Mb, where a likely regulatory region is located close to the *ID2* gene. In Brown Swiss, a 130-kb window at 75 Mb on chromosome 15 was identified. Analysis of whole-genome sequence data using linkage-disequilibrium estimation revealed possible causal variants for the identified QTL. A presumably regulatory variant in the non-coding 5′ region of the *ID2* gene was strongly associated with the haplotype for Original Braunvieh. In Brown Swiss, an intron variant in *PRDM11*, one 3′ UTR variant in *SYT13* and three intergenic variants 5′ upstream of *SYT13* were identified as candidate variants for the trait multiple birth maternal. In this study, we report for the first time QTL for the trait of multiple births in Original Braunvieh and Brown Swiss cattle. Moreover, our findings are another step towards a better understanding of the complex genetic architecture of this polygenic trait.

KEYWORDS
Bos taurus, marker-assisted selection, ovulation rate, quantitative trait loci, twinning rate

INTRODUCTION

The length of calving intervals and the number of calves born alive have a strong influence on the productivity of dairy cattle. Intensive selection, especially for milk yield in dairy cattle over the last 100 years, has led to a decline in female fertility due to negative genetic correlations between milk yield and female fertility (Miglior et al., 2017). To reverse the negative drifts in dairy cow fertility, producers benefited from the development of genomic selection and advanced management strategies; however, further efforts are needed.

Cattle are usually monoecious, so that a pregnancy usually ends with the birth of a single calf. Multiple births in cattle are rare. The multiple birth rate (MBR) varies between 1.02 and 9.6% depending on breed and study (Johanson et al., 2001; Moioli et al., 2017; Weller et al., 2008). They are generally lower in beef cattle than in dairy cattle (Atteneder, 2007; Lett & Kirkpatrick, 2018). Most of multiple births are due to multiple ovulations, as several ovulatory follicles mature at the same time. Around 5–10% of bovine twins are identical twins (Atteneder, 2007; Silva del Río et al., 2006). So far, however, the trait of multiple births has neither been studied in general nor with a genetic model for Brown Swiss (BS) and Original Braunvieh (OB) breeds.

Generally, multiple births are undesirable in dairy cows for several reasons. In particular, twin and multiple births are linked to increased health problems for the dam and calves. For example, there is a higher risk of ketosis, metabolic disorders, retained placenta, and displaced abomasum (Echternkamp & Gregory, 1999; Gregory et al., 1990; Pardon et al., 2012). In addition, the impact of multiple births on subsequent fertility was observed through their negative influence on conception rate and calving interval (Andreu-Vázquez et al., 2012; Echternkamp & Gregory, 1999). Calves born in multiples have an increased risk for deficiency syndrome and mortality (Atteneder, 2007; Silva Del Río et al., 2007). Furthermore, a higher incidence of dystocia, abortion, and stillbirth has been found in multiple births (Atteneder, 2007; Gregory et al., 1990). In summary, these factors lead to higher production costs for the farmers. Consequently, selection against multiple births could improve fertility and profitability in dairy production.

Several non-genetic factors have been described that could have an influence on MBR such as cow parity, and further environmental factors as the season of birth and the herd (Johanson et al., 2001; Silva Del Río et al., 2007). A significantly higher MBR was observed in multiparous cows (5.22–7.35%) compared to cows in first parity (1.63%) (Johanson et al., 2001). So far, two studies have analysed the relationship between milk yield and MBR. These studies presented contradictory results and thereby do not give a clear picture of the relationship between these two traits (Masuda et al., 2015; Murillo-Barrantes et al., 2010). Most studies reported a higher MBR for births in summer months (Johanson et al., 2001; Silva Del Río et al., 2007). This suggests that the number of multiple ovulations is highest in late summer and autumn. Several studies indicate a positive phenotypic trend for MBR over time (Fitzgerald et al., 2014; Moioli et al., 2017; Silva Del Río et al., 2007), suggesting that MBR is correlated with further selection traits. Regarding the genetic trend, Ghavi Hossein-Zadeh et al. (2009) found a negative trend and Murillo-Barrantes et al. (2010) found no change in the genetic component of MBR over time. In most dairy cattle populations, MBR seems to be increasing at the phenotypic level. The heritability estimates for multiple births range from 0.011 to 0.160, suggesting that a small but non-zero proportion of the observed trait variation can be attributed to genetic factors (Fitzgerald et al., 2014; Johanson et al., 2001; Lett & Kirkpatrick, 2018). Estimated values using a linear model were lower than those using threshold models. Furthermore, an Austrian study reported higher estimates for the heritability of the maternal multiple birth trait (0.017–0.063) compared to the estimates for the heritability of the direct multiple birth trait (0.001–0.005) (Atteneder, 2007).

Maternal multiple birth has been studied with different QTL mapping strategies in Norwegian cattle, as well as in Israelian or North American Holsteins using family-based microsatellite interval mapping approaches (Bierman et al., 2010; Lien et al., 2000; Weller et al., 2008). They detected multiple QTL on 13 different chromosomes, depending on the study conducted and the population examined (Cobanoglu et al., 2005; Lien et al., 2000; Weller et al., 2008). Paternal half-sib families for North American Holstein cattle were analysed using single-marker models or combined linkage–linkage disequilibrium approaches, which resulted in the discovery of multiple QTL on eight chromosomes (Bierman et al., 2010; Kim et al., 2009). The Italian Maremmana beef breed was analysed in a study using a single-trait linear mixed effect model as well as an animal threshold model including the number of calves born per cow as the phenotype. They identified a single significantly associated SNP on chromosome (chr) 24 by a genome-wide association studies (GWAS) based on 54-k SNP data (Moioli et al., 2017). So far, a study has analysed the Holstein population using large-scale data and identified a major QTL for multiple births on chr 11 with a possible effect on the genes *LHCGR* and *FSHR* (Widmer et al., 2021). This QTL region was confirmed in a recent study in the North American Holstein population using whole-genome sequencing (WGS) data (Lett & Kirkpatrick, 2022). It is therefore very likely that multiple births are a polygenic trait where almost no QTL have been identified yet. The study presented in this paper is based on a large-scale dataset of phenotypes and genotypes for the current Swiss dairy cattle breeding populations of BS and OB.

We aimed to perform an extensive genetic analysis of the multiple birth trait in Swiss OB and BS cattle. To predict breeding values, phenotypic data for single and multiple births from 2 decades were used, which were recorded in the national animal database, as well as data from pedigrees. In addition, extensive genotype data obtained from routine genomic selection of male and female animals were used to detect associated QTL using GWAS and haplotype regression analysis. To identify linked genomic regions and candidate causal variants, different fine-mapping procedures were performed. In a final step, we evaluated the effect of the identified haplotype on the routinely recorded birth and fertility traits.

MATERIALS AND METHODS

Phenotypes

Phenotypic data were available on a large scale through the Swiss national animal recording database. In this study we used data collected between 2006 and 2018 from the breeding organisation Braunvieh Schweiz (Zug, Switzerland). The raw dataset contained 5 692 022 birth records including the multiple birth code from BS and OB cattle. Data preparation and analysis for variance component estimation and breeding value estimation were done with internal software written in R (R Core Team, 2020). Birth records subsequent to an embryo transfer were excluded. For the genetic studies of the discrete multiple birth trait 1 367 529 records remained after data preparation and validation. The overall MBR was 7.83%. Further information regarding the final dataset is shown in Table 1.

Variance components estimation

We fitted the following mixed linear model to our phenotypic data described above according to a previous study in a different breed (Widmer et al., 2021):

$$y = \mathrm{Xb} + \mathrm{Wh} + Z_d \mathrm{mb_d} + Z_m \mathrm{mb_m} + \varepsilon$$

where b is the vector of the fixed effects, h represents the vector of the random herd-year effect, $\mathrm{mb_d}$ and $\mathrm{mb_m}$ are the direct (calf) and the maternal (dam) genetic effects respectively, and ε is the residual. The fixed effects of season, parity, use of sexed semen, and number of inseminations leading to pregnancy were taken into account. In Table 1, the numbers of observations per level of each fixed effect are illustrated. X, W, Z_d, and Z_m represent the design matrices for the random (W, Z_d, and Z_m) and fixed (X) effects. The selection of the fixed effects is based on previous unpublished work that analysed a similar dataset in Switzerland. The items in the vector of observations (y) were coded with 1 for single and 2 for multiple birth (twin or triplet). Dataset filtering was performed to eliminate all records from herds with fewer than 260 records per herd in total (<20 per year on average), and all records in herd-year classes with fewer than five records. Finally, this resulted in a dataset of 419 111

TABLE 1 Estimated effects of the factor levels for the fixed effects of the estimated breeding values based on the final dataset

Fixed factor	Level	Number of observations per level [a]	Estimated effect of factor level
Parity	1	351 678	0.212
	2	308 988	0.271
	3	222 356	0.286
	4	168 837	0.295
	5+	315 670	0.302
Use of sexed semen	No	1 312 571	0.269
	Yes	54 958	0.264
Season of birth	Spring	231 908	0.261
	Summer	266 155	0.278
	Fall	519 258	0.270
	Winter	350 208	0.265
Number of inseminations	1	876 021	0.270
	2	324 011	0.267
	3	106 625	0.266
	4	36 382	0.265
	5	13 697	0.256
	6	57 776	0.251
	7+	5017	0.253

[a]Based on the final dataset of 1 367 529 records.

records. Estimation of variance components were conducted for BS and OB together by using the software vce (Neumaier & Groeneveld, 1998).

Breeding value prediction

Breeding values prediction was performed with the MiX99 software (MiX99 Development Team, 2017; https://www.luke.fi/mix99) by applying the mixed linear effects model shown above and run for BS and OB together. Within this analysis, the dataset filtering excluded records from herds with fewer than 260 records but without defining a minimal number of records per herd-year levels. In total, 1 367 529 records were available for breeding value prediction. The estimates from the prior step were the base for the required values of variances and covariances. The estimation of the reliability of all breeding values was based on an approach by Tier and Meyer (2004) using the apax99 program in the MiX99 software. Furthermore, breeding values were standardised to a mean of 100 and a standard deviation of 12.

The de-regression of the estimated breeding values (EBVs) for the direct (mbd) and maternal (mbm) multiple birth traits was performed according to Garrick et al. (2009) (Table 2). For the association analyses, all animals having a de-regressed EBV with a corresponding reliability above 0.35 were selected. The EBVs were not weighted by the reliability for the further steps. The following analyses were carried out separately for the BS and the OB breeds. They were separated using the pedigree-based gene proportion. Minimum required gene proportion was defined at 0.7 for each of the tail populations. Finally, for BS 3792 and 7847 animals remained in the analysis for the traits mbd and mbm respectively. For OB there were 1332 and 3508 animals used for the following analyses for the mbd and mbm traits.

Genotypes

For 97-k animals, routine SNP genotype were available. Imputation was performed as described in a previous study (Widmer et al., 2021) for the SNP data generated for genomic selection. The reference dataset for the 150-k array included 6370 BS and 1516 OB animals and for the high-density array 1686 BS and 421 OB cattle. The final marker set contained 110 510 and 681 178 SNPs for each density (150 k and high density) respectively. SNPs were filtered separately for the BS and OB populations using the following thresholds: minor allele frequency >0.01 and SNP call rate >0.99 in the genotypic data of the reference population. The ASR-UCD1.2 bovine assembly was used as the reference genome for the SNP data.

Association studies

Single SNP regression

Genome-wide single SNP association studies were performed using a mixed model in the software snp1101 (Sargolzaei, 2021). The genomic relationship for the cattle used in the analyses was calculated in order to correct for population stratification in the model (VanRaden, 2008). The aim of this approach was to detect variants that were significantly associated with the studied traits mbd and mbm.

Bayes B approach

To fit one genotype after another can simply lead to biases due to LD and stratification (Fernando et al., 2017). Therefore, fitting subsets of genotypes at the same time can address this paradigm. Window-based association analyses were carried out using GenSel software (Putz, 2021) and the BayesB algorithm (Fernando & Garrick, 2013). First, the proportion of loci with zero effect (known as π parameter) was estimated from the same dataset, and a value of 0.989 was set as the a priori starting value. We estimate the window variance for genomic windows of 1-Mb length.

Haplotype analysis

We used a haplotype regression analysis restricted to the previously identified genomic windows, explaining a significant proportion of the total genetic variance, to detect the haplotype which include the presumed causal QTL. Following a previously described approach by Pausch et al. (2016), we aimed to identify significantly associated haplotypes for both breeds within each window. We intersected haplotypes using all odd numbers

TABLE 2 Summary statistics of the de-regressed breeding values for the direct and maternal multiple birth traits

Trait	Min	Max	Mean	SD	*n* observations
mbd	−50.810	72.971	−0.065	6.441	3293
mbm	−55.453	186.960	0.892	9.268	7284

Abbreviations: mbd, multiple birth direct; mbm, multiple birth maternal.

TABLE 3 Estimated raw variance components

Component	Raw value	Standard error
Herd-year	0.88×10^{-4}	0.15×10^{-4}
Animal	0.72×10^{-6}	0.18×10^{-5}
Dam	0.16×10^{-2}	0.74×10^{-4}
Residual	0.037	0.91×10^{-4}
Correlation animal/dam	-0.34×10^{-4}	0.43×10^{-4}

between 9 and 301 SNPs as lengths within the identified segment per breed. We slid the starting point of the haplotype SNP-wise and used different lengths so that we can consider each haplotype for the effect estimation. The de-regressed EBV were again used as response variables for the estimation of haplotype effects. For the following fine-mapping approach to identify any candidate causal variants using whole-genome sequence data, we selected the most significantly associated haplotype (lowest *p*-value).

The top-associated haplotype was analysed regarding associations with routinely recorded and available calving and fertility traits. Therefore, the haplotype was fitted as a single diplotype representing haplotype carrier states (0 = non-carrier, 1 = carrier and 2 = homozygous carrier). To correct for population stratification, the genomic relationship was included in the model in the GCTA software package (Yang et al., 2011).

Additionally, we performed an analysis to investigate a possible co-association between the most significantly associated haplotype for our trait of interest in each breed and further recorded fertility traits. For this test we used routinely available de-regressed EBV for all fertility and birth traits.

Fine-mapping

WGS data were searched for variants in significant LD with the selected most significantly associated haplotype, which was decoded as diplotype using the plink software package (Purcell et al., 2007). For BS, the 1-Mb window between 75 and 76 Mb on chr 15 was selected based on the associated Bayes B window and the localisation of the highly associated haplotypes. For the OB breed, we selected the 1-Mb window between 88 and 89 Mb on chr 11. WGS data provided by the 1000 Bull Genomes Project run 8 were used. This dataset contains 4109 animals and includes 237 BS and 81 OB cattle (Hayes & Daetwyler, 2019). We converted the data chromosome-wise from VCF format to plink input files using the software VCFtools v0.1.16 (Danecek et al., 2011). Subsequently, we transformed the data into binary files and performed VCF quality controls using the plink v.1.90b3.46 software (Purcell et al., 2007). We removed variants with a missing call rate>0.1, a minor allele

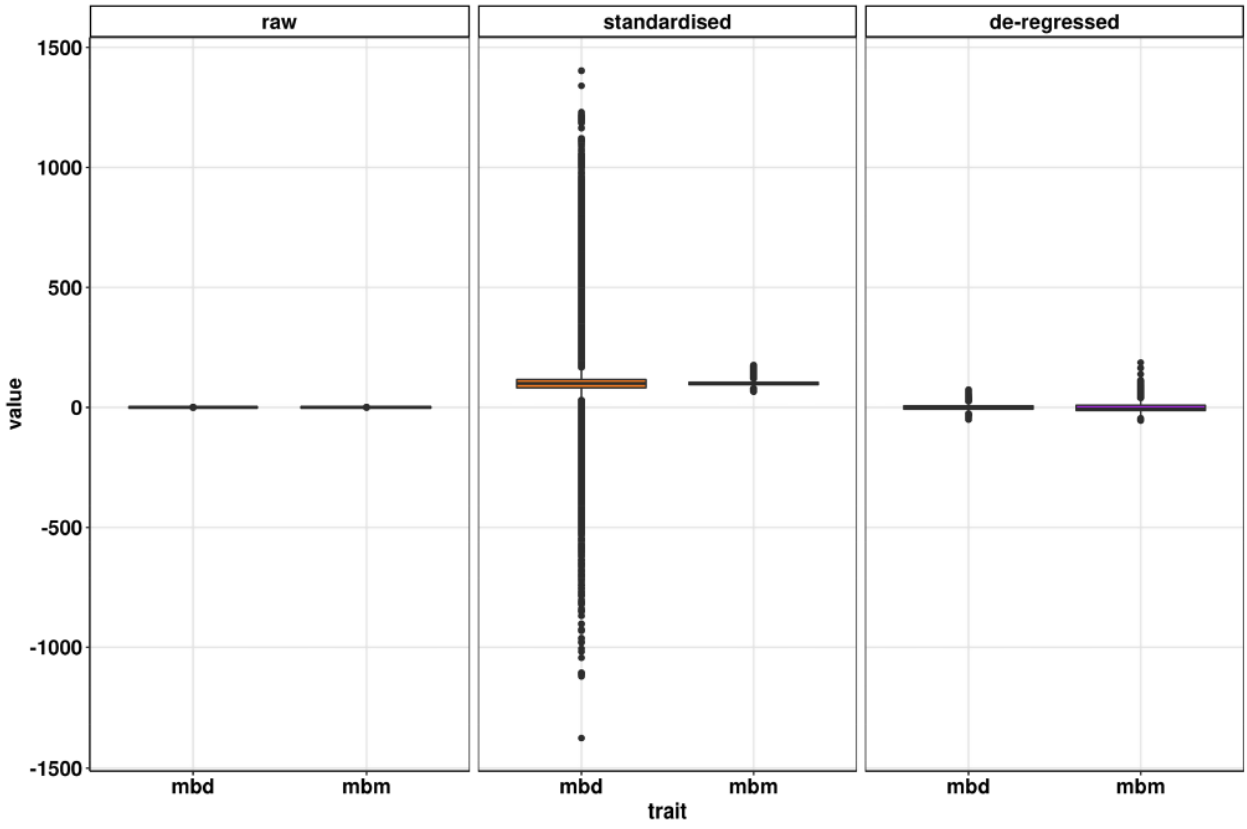


FIGURE 1 Raw, standardised, and de-regressed breeding values for the traits direct (mbd) and maternal (mbm) multiple birth

frequency<0.01, and samples with missing call rate>0.1. In addition. Samples and variants with a Mendelian error rate>2% were excluded. For the animals of the 1000 Bull Genomes Project, pedigree data were provided by the Braunvieh Schweiz (Zug, Switzerland) and included 526 duos and 19 trios that could be used for Mendelian analyses. Of the 237 whole-genome sequenced BS and 81 OB animals, 137 samples were used for LD-based variant detection after quality control. Only for these animals could additional SNP genotype data be obtained: 5933 variants were located in the 1-Mb window on chr 15 for BS and, for the OB breed, the final analysis included 8041 variants in the 1-Mb region on chr 11.

Furthermore, we have looked at the frequency of the top-linked variants across breeds. We used all 4109 samples in the data file provided by the 1000 Bull Genomes Project. We analysed all breeds that contained at least 50 records.

RESULTS

Prediction of breeding values

The raw values resulting from the estimation of the variance components are shown in Table 3. The heritability of the direct (animal) and the maternal (dam) genetic effects was 0.2 10^{-4} and 0.040 respectively.

We estimated breeding values for the traits mbm and mbd. These values were used in a de-regressed alteration as input variable for the association analysis. The estimated effects of the factor levels for the fixed effect of use of sexed semen, parity, season of birth, and number of inseminations leading to pregnancy are shown in Table 1. The use of sexed semen has a significant negative effect on the occurrence of multiple births. In addition, the prevalence of multiple births was lower in primiparous cows and in the summer months, the highest incidence for multiple births was observed. The empirical distributions of the raw, de-regressed, and standardised values of the breeding value prediction can be seen as boxplots in Figure 1.

The reliabilities of the standardised breeding values of 1 905 641 animals had a mean of 0.263 for mbd and 0.264 for mbm, and the standard deviation was 81.663 for mbd and 8.858 for mbm. As a measure of the genetic trend, we grouped the mean estimated breeding values by year of birth of the animals for each trait. Interestingly, no clearly discernible genetic trend was found for the trait mbm in either breed or for the trait mbd in BS (Figure 2). For the trait mbd in OB only, a negative trend was observed for 2000–2010.

A QTL close to the *ID2* gene on chromosome 11 for OB

For OB cattle, the window-based BayesB approach revealed a QTL with a significantly associated window

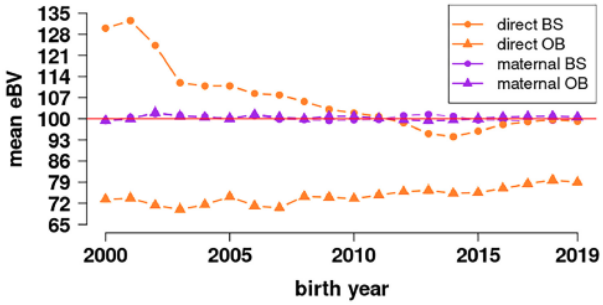


FIGURE 2 Genetic trend of the estimated breeding values of multiple births from 2000 to 2019

that explained 5.88% of the genetic variance of the trait mbm (Figure 3a). The window is located on chr 11 between 88 000 459 and 88 996 149 bp (Table 4). The values of p>0 (proportion of models where this window was included, and thus explained for >0% of the genetic variance) and of p>average (proportion of models where this window explains for more than the amount of variance that would be explained if each window had the same effect) were 1 and 0.980 respectively. Therefore, this QTL reached a significance threshold of 5% in the BayesB approach. The other three possibly associated windows, each located on different chromosomes, explained smaller amounts of genetic variance (<2.5%), and were identified below the significance level of 5% (Table 4).

In the single SNP regression models for the trait mbm, no significantly associated SNPs were found at the Bonferroni-corrected threshold of 5% (Figure 3b). Different suggestive signals can be observed on chr 1, 10, 11 and 22. The significantly associated window on chr 11 from the BayesB analysis overlapped with the suggestive QTL on chr 11. The top SNP is located at position 88 765 206 bp. For the second trait analysed (mbd), we could not observe a significant signal in either of the two GWAS analyses.

We selected a 1-Mb-segment on chr 11 at 88 Mb for haplotype association analyses based on the results from the BayesB analysis shown above. The 16 top-associated haplotypes are listed according to their *p*-value from the association analysis in Table S1. They were located in the region between 88.74 and 88.83 Mb on chr 11 and contained 11 to 15 SNPs. We selected from the most significantly associated haplotypes the longest haplotype (chr11: 88 748 439–88 808 495) for further analysis. This haplotype comprises 15 SNPs and had a frequency of 0.583 in the OB genotype dataset. For this haplotype, we discovered an estimated additive effect of −3.573 (±0.751 standard error), indicating a negative effect on the trait of interest mbm.

LD-analysis revealed six variants located between 88.75 and 88.79 Mb on chr 11 showing r^2 values ≥0.7 with the chr11: 88 748 439–88 808 495 haplotype (Table S1), but none were in perfect LD (r^2 = 1; Figure 4b). By far the highest r^2 value (0.905) was determined for the variant

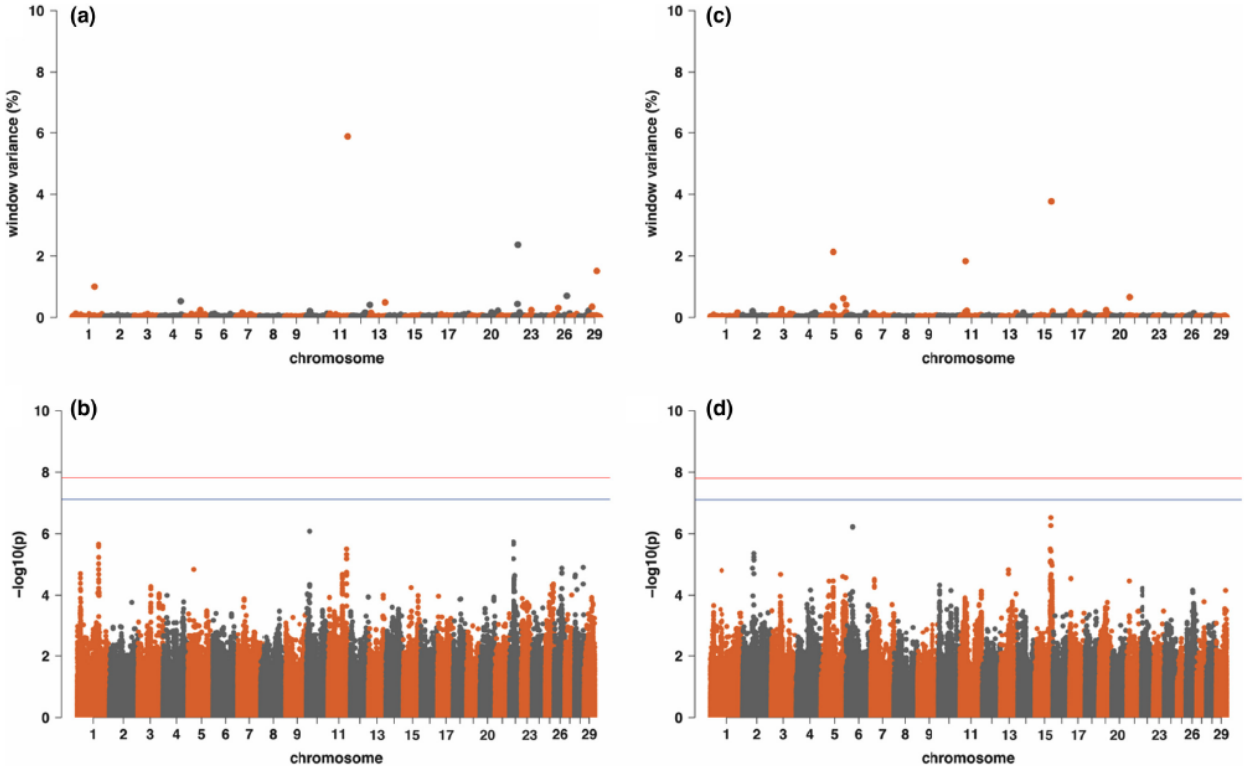


FIGURE 3 Manhattan plot of genome-wide association studies for the trait mbm. (a, c) results of the window-based BayesB approach and (b, d) the single SNP regression with Bonferroni-corrected threshold level of 5% (blue line) and 1% (red line). (a, b) Original Braunvieh, (c, d) Brown Swiss

TABLE 4 Associated genome regions from the BayesB window approach for the trait maternal multiple birth

Breed	Chr	Start position	End position	*n* SNP	Proportion of explained genetic variance (%)	$p>0$	$p>$average
Original Braunvieh	11	88 000 459	88 996 149	366	5.88	1	0.980
	22	24 002 842	24 995 229	224	2.36	0.985	0.754
	19	36 007 788	36 998 151	274	1.51	0.967	0.729
	1	107 022 412	107 999 432	304	1.00	0.980	0.526
Brown Swiss	15	75 001 650	75 995 887	339	3.77	0.997	0.950
	5	57 008 008	57 996 196	196	2.13	0.962	0.664
	11	24 001 475	24 994 135	296	1.83	0.995	0.897

Abbreviations: Chr, chromosome; $p>0$, proportion of models where this window was included, and thus explained for more than 0% of the genetic variance; $p>$average, proportion of models where this window explains for more than the amount of variance that would be explained if each window had the same effect.

chr11: 88 791 842 A>T (Figure 4b), which, like the other five variants, mapped to a non-coding region on chr 11 (Table 5). The closest annotated gene is *ID2*, which codes for the inhibitor of DNA binding 2 and is located between 88 604 880 and 88 607 401 bp (Figure 4a). All these six variants are located within the top-associated haplotype between 88.75 and 88.81 Mb as described above (Table S1).

Since candidate causative variants for the identical trait may differ between breeds of the same species, we have looked at the frequency of the highly linked variant on chr 11 across breeds. The frequency of this top-linked variant (chr11: 88791842 A>T) was high in several of the breeds including BS (Figure S1a). It is obvious that this variant segregates strongly in breeds other than OB, even in unrelated breeds.

A QTL close to the *PRDM11* and *SYT13* genes on chromosome 15 for BS

For the studied BS population, the window-based BayesB approach for BS revealed evidence for a QTL with a significantly associated window that explained 3.77% of the genetic variance of the trait mbm (Figure 3c). The localisation of this window is on

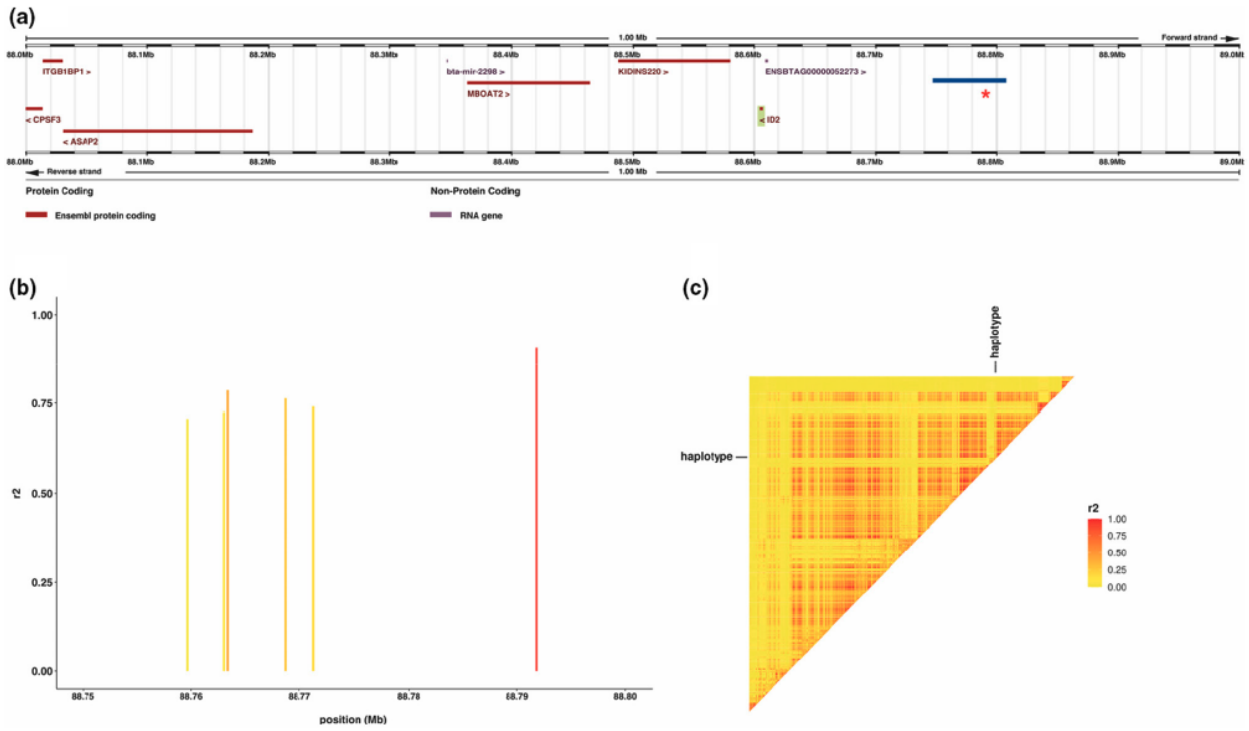


FIGURE 4 Graphical representation of the associated QTL region on bovine chromosome 11 and linkage disequilibrium (LD) analysis results for the mbm trait in Original Braunvieh. (a) Screenshot of the region between 88 and 89 mb from www.ensembl.org including the localisation of the genes in the region. The blue bar shows the localisation of the top-associated haplotype. The red star represents the variant with the highest LD score. (b) The variants in $r^2 \geq 0.7$ with the top-associated haplotype are highlighted and shown with their genomic position. (c) Heatmap showing the LD in the region between 88 and 89 mb highlighting the haplotype that was added as an additional variant

TABLE 5 Overview of the highly associated variants from the linkage disequilibrium analysis for the trait multiple birth maternal

Breed	Chr	Position [a]	Variant [a]	Impact	Associated gene [a]	MAF	LD (r^2)
Original Braunvieh	11	88 791 842	A>T	Intergenic variant	5′ of *ID2*	0.343	0.905
Brown Swiss	15	75 213 046	T>C	Intron variant	*PRDM11*	0.142	0.970
		75 297 912	C>T	3′ UTR variant	*SYT13*	0.142	0.970
		75 399 114	G>T	Intergenic variant	5′ of *SYT13*	0.142	0.970
		75 402 900	G>A	Intergenic variant	5′ of *SYT13*	0.140	0.970
		75 405 408	T>G	Intergenic variant	5′ of *SYT13*	0.142	0.970

Abbreviations: Chr, chromosome; UTR, untranslated region.
[a]ASR-UCD1.2/bosTau9 assembly.

chr 15 between positions 75 001 650 and 75 995 887 bp (Table 4). The values p>0 and p>average reached values of 0.997 and 0.950 respectively. Therefore, the QTL from the BayesB analysis was significant at a threshold of 5%. Two other associated windows, each located on different chromosomes, were identified below the significance level of 5% (Table 4). They explained <2.5% of the genetic variance.

The result from single SNP regression analysis did not show any significantly associated SNPs at the Bonferroni corrected level of 5% for the mbm trait (Figure 3d). Nevertheless, a suggestive signal was observed on chr 15. The best-associated SNP is at position 75 847 291 bp. This overlapped with the significantly associated window on chr 15 from the BayesB analysis. Further suggestive but non-significant regions can be identified on chr 2 and 6 from the single SNP GWAS models. Also, for the BS population, we could not find any significant association signals for the trait mbd in either of the two GWAS analyses.

Based on the results from the BayesB analysis shown above, we used a 1-Mb segment on chr 15 starting at 75 Mb for haplotype association analyses. The six best-associated haplotypes according to their significance level from the association test were located in the region between 75.43 and 75.56 Mb on chr 15 and contained 41–51 SNPs (Table S1). The longest top-associated haplotype (chr15: 75430142–75 561 032) was selected for further analysis, for which we identified an estimated additive effect of 5.018 (±0.686 standard error), which

indicates that it has a positive effect on the trait mbm. This haplotype included 51 SNPs and had a frequency of 0.583 in the entire genotype dataset of the BS population studied.

The analysis of the pairwise LD (r^2) between the top-associated haplotype (chr15: 75 430 142–75 561 032) and all detected variants in the 1-Mb region revealed 60 variants located between 75.09 and 75.96 Mb on chr 15 with r^2 values ≥0.7 (Table S1). Five out of these 60 variants had values of $r^2 \geq 0.95$ (Table 5), but for none was perfect LD ($r^2 = 1$) observed with the haplotype (Figure 5b). Of these five highly associated variants, one maps in an intron of the *PRDM11* gene, one is located in the 3′ UTR of the *SYT13* gene, and three are intergenic variants located 5′ upstream of the *SYT13* gene (Table 5 and Figure 5a). None of the highly associated variants ($r^2 \geq 0.95$) and only 6 of these 36 variants with r^2 values ≥0.7 map into the region of the top-associated haplotype between 75.43 and 75.56 Mb (Table 5 and Table S1).

Additionally, we analysed the occurrence of the five chr 15 variants that were in high LD ($r^2 \geq 0.95$) with the top-associated haplotype across breeds. Interestingly, the frequencies of these five variants were low (<0.1) in all breeds except BS (Figure S1b–f). Four out of five variants segregated only in a small number of breeds. Only the variant chr15: 75 405 408 T>G occurred in all analysed breeds (at low level in OB), with the only exception of Angus where the variant does not segregate.

No detectable haplotype effects on routinely available fertility and calving traits

We estimated effects of the haplotypes for each breed, OB and BS, on routinely accessible fertility and calving traits to analyse a potential co-association between other traits and the most significantly associated haplotypes for mbm. For the top-associated haplotypes in OB (chr11: 88 748 439–88 808 495) and BS (chr15: 75 430 142–75 561 032) we found no significant association at the Bonferroni corrected level of 5% (p-value ≤0.038; Table S1).

DISCUSSION

Impaired female fertility is a well-known problem in high-performance dairy cattle and multiple births have been an undesired trait for a long time. In this study, based on large-scale genotyping and phenotyping data, we identified genetic factors that are highly likely to affect this trait. Using the de-regressed breeding values for the trait of direct and maternal multiple births as phenotypes, two novel QTL were detected in two breeds: a QTL on chr 11 in OB and a QTL on chr 15 in BS. For OB, we found a variant located in the 5′-regulatory region of the *ID2* gene as potential candidate causal variant. In the BS population, we detected five variants that could presumably affect the expression of the *PRDM11* and

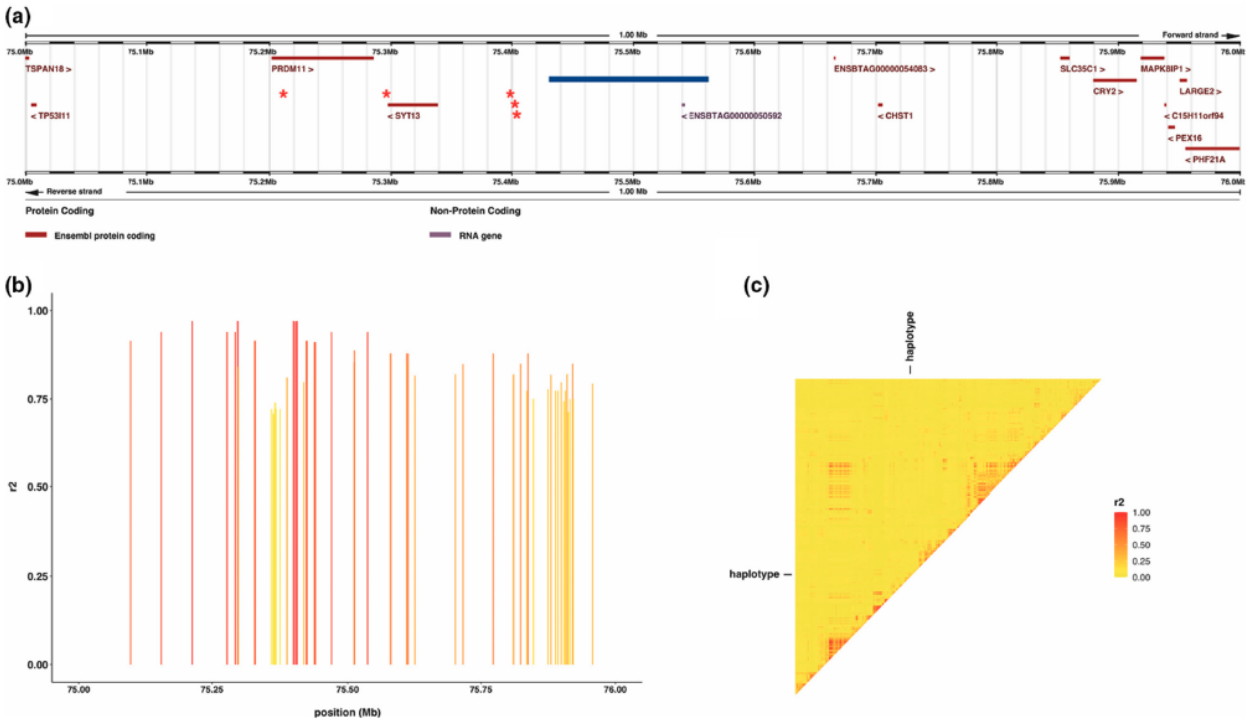


FIGURE 5 Graphical representation of the associated QTL region on bovine chromosome 15 and linkage disequilibrium (LD) analysis results for the mbm trait Brown Swiss. (a) Screenshot of the region between 75 and 76 mb from www.ensembl.org including the localisation of the genes in the region. The blue bar shows the localisation of the top-associated haplotype. The red stars represent the variants with the highest LD scores ($r^2 \geq 0.95$). (b). The variants in $r^2 \geq 0.7$ with the top-associated haplotype are highlighted and shown with their genomic position. (c) Heatmap showing the LD in the region between 75 and 76 mb highlighting the haplotype that was added as an additional variant

SYT13 genes and therefore could potentially be causal candidates.

The genetic parameter estimates (variance components of direct and maternal genetic effect for trait multiple birth) were similar to those reported before in different cattle populations (Ghavi Hossein-Zadeh et al., 2009; Weller et al., 2008; Widmer et al., 2021). Since the phenotypic observations do not have a normal distribution, it is non-trivial to develop a method to predict breeding values for a binary trait such as multiple births. However, the use of linear mixed models can lead to sufficient rankings of animals according to predicted breeding values, as Negussie et al. (2008) have shown, but the results obtained may not be valid and predictions do not have the same evidence as the discrete response variable. These challenges could be the reason for the extreme outliers in the standardised breeding values that we observed for both traits analysed (mbd and mbm). The potential of a threshold model or a generalised linear mixed model in the context of this trait needs further research. Previous studies have already used threshold models or generalised linear mixed models for genetic evaluations and predictions in a general purpose and for the analysis of multiple births (Johanson et al., 2001; McGovern et al., 2021).

In the last 2 decades, no clear genetic trend could be observed for both studied traits (mbd and mbm) in BS and for the trait mbm in OB. OB data for the trait mbd showed a negative trend from 2000 to 2010, in contrast to earlier results in Norwegian cattle where a positive trend was observed (Karlsen et al., 2000). Since the reliability and heritability of the trait mbd are low, the genetic trend could be caused by non-genetic effects. However, it is very likely that multiple births were generally not selected in the Swiss populations studied, neither by direct selection nor by genetic correlation with selected traits. Thus, the development of breeding value estimation could be an important selection tool to reduce the risk of multiple births, hopefully leading to improved animal health and welfare of dairy cattle. Implementing a selection program into the local breeding schemes could be the next step.

In our datasets, we found significant associations for the maternal trait. This shows the important role of the maternal component for the complex trait of multiple births. The QTL on chr 11 for OB and on chr 15 for BS were detected by the window-based BayesB approach. Only suggestive signals were observed in the single SNP regression analysis. As shown in this study, merely a small part of the genetic variance of quantitative traits can be significantly revealed by the identification of a single marker (Maher, 2008; Visscher et al., 2010). The BayesB approach considers all markers simultaneously as random effects and can therefore explain most of the genetic variance (Fan et al., 2011; Hayes et al., 2010). A window-based approach captures the majority of the variability at an associated trait locus (Fernando & Garrick, 2013).

The identified QTL for OB explained 5.88% and the QTL for BS 3.77% of the genetic variance of the trait. Because mbm is a classical polygenic trait, a larger dataset could identify additional QTL with smaller effects. The non-significantly associated signals resulting from the window-based BayesB analysis on three additional chromosomes for OB and two additional chromosomes for BS (Table 4) may be suggestive QTL that need to be investigated in future studies. The theory of a polygenic trait is reinforced by previous studies showing multiple QTL on different chromosomes (Bierman et al., 2010; Kim et al., 2009; Weller et al., 2008; Widmer et al., 2021). Regarding chr 11 and 15, only combined linkage–linkage disequilibrium analysis for North American Holstein sires and ANOVA analysis in Israeli Holsteins showed QTL for twinning rate on the same chromosomes but located in different segments (Kim et al., 2009; Weller et al., 2008). Interestingly, no QTL for the multiple birth trait is reported in the bovine QTL database in the genome regions described in this study (Cattle QTL Database, 2022). There is little evidence in the literature of a major QTL for the multiple birth trait in cattle. In an own study, a major QTL on chr 11 at 31 Mb was detected with a candidate causal variant in the 5′ regulatory region of the *LHCGR* gene in Swiss Holstein (Widmer et al., 2021) and recently confirmed in the North American Holstein population (Lett & Kirkpatrick, 2022). In two older studies with data from Norwegian cattle and North American Holsteins, a candidate gene with a possible major effect on the trait, *IGF1*, was observed (Kim et al., 2009; Lien et al., 2000). The region on chr 5 encompassing *IGF1* does not intersect with the suggestive QTL region identified for BS using the BayesB approach in our analysis. For bovine ovulation rate, a single study provided strong evidence for a major QTL in the region of 13.6 to 14.8 Mb on chr 10 containing three possible candidate genes (*SMAD3*, *SMAD6*, and *IQCH*) (Kirkpatrick & Morris, 2015). Presumably, a global dataset could improve our analyses and uncover additional QTL for the trait of interest.

LD-based filtering of sequence variants located in the QTL regions did not reveal a single variant that had perfect LD with the top-associated haplotypes for either breed. However, for BS on chr 15, a total of 60 variants showed high scores of LD with the haplotype ($r^2 \geq 0.7$) forming an 870-kb block. This long range is to be expected due to the previously described relatively long-range LD pattern in cattle populations (de Roos et al., 2008). As previously shown, a single genomic region can harbour multiple QTL for a polygenic trait (Thaller et al., 2003); however, it is more likely that the identified long segment is due to LD between adjoining variants and only one of them represents the actual causal variant. It could be suspected that one of the five variants in highest LD ($r^2 \geq 0.95$) is the most likely one (Table 5). One of those variants is a variant in the intron of the *PRDM11* gene located between 75 202 012 and

75 285 991 bp on the forward strand on chr 15. This gene codes for the PR-domain containing protein 11 that is predicted to enable chromatin binding activity It is involved in several processes, including negative regulation of cell growth, positive regulation of the apoptotic process in fibroblasts and regulation of transcription, but no role in reproduction has yet been reported (Fog et al., 2015). Different members of the PRDM family have an influence on germ cell specification with PRDM9 being expressed during meiosis (Fog et al., 2012). The other four variants with $r^2 \geq 0.95$ are close to the *SYT13* gene, which is located on chr 15 between 75 297 582 and 75 339 698 bp on the reverse strand, and may have a regulatory influence on its expression. *SYT13* codes for synaptotagmin 13, which belongs to the large family of synaptotagmin proteins whose members function as membrane transporters in multicellular organisms (Quiñones-Frías & Littleton, 2021). *SYT9*, a member of this family, affects the release of follicle stimulating hormone (FSH) in female mice and thus the oestrus cycle and ovulation (Roper et al., 2015). Therefore, we postulate that these five variants are possible candidates for a causal effect on the trait multiple births in BS cattle.

For the OB population, six variants with higher LD ($r^2 \geq 0.7$) were observed in a 40-kb window on chr 11. By far the highest r^2 with 0.905 was observed with the variant chr11: 88791842 A>T, a variant close to the *ID2* gene located between 88 604 880 and 88 607 401 bp on the reverse strand, which encodes for the inhibitor of DNA binding 2. This gene belongs to the inhibitor of DNA binding family whose members are transcriptional regulators. Despite the nearly 200 kb distance to this gene, this non-coding variant could probably have a regulatory effect on *ID2* expression. Such a possible impact on *ID2* is supported by the block of increased LD in this region, as shown by the heatmap presented (Figure 4c). *ID2* is expressed at different levels during the oestrus cycle in the ovarian tissue in mice and may play a role in negatively regulating cell differentiation (Zavareh et al., 2018). In cattle, the expression of maternal *ID2* transcripts decreased from immature to mature oocyte (Thélie et al., 2007). Expression of *ID2* in pigs is influenced by FSH and cumulus oocyte complexes (Verbraak et al., 2011). In birds, *ID2* is sufficient for the expression of the follicle stimulating hormone receptor (*FSHR*) (Johnson & Woods, 2009). Moreover, the expression of *ID2* is significantly decreasing in granulosa cells of preovulatory follicles compared to mid-oestrous follicles in mares (da Silveira et al., 2014). Therefore, we postulate the possibility of an influence of a regulatory variant of *ID2* on ovulation and thus on the occurrence of multiple births. The candidate variant (chr11: 88791842 A>T) segregates in various breeds at high frequency (Figure S1a). Therefore, it could be assumed to represent an old mutation that probably arose before the formation of modern breeds. A validation analysis was performed and we were not able to confirm an association of that specific variant in the Swiss BS, Simmental or Holstein cattle populations (data not shown). Therefore, it is more likely that another variant in the QTL region that is in high LD with chr11: 88791842 A>T is causally effective, or that multiple variants have the same effect on the trait.

Since the proposed candidate causal variants for multiple births in the BS breed (Figure S1b–f) segregate in various breeds at low frequency, they probably occurred before the formation of modern breeds. A validation analysis as described for OB was not possible, because for most of the breeds analysed, neither phenotypes nor genotypes were available to investigate a possible effect on the trait multiple births. Remarkably, none of the five variants that had an LD score >0.95 belonged to the interval of the top-associated haplotype for BS, as they mapped further upstream on chr 15. Interestingly, the haplotype and the variants are not in a block of raised LD as shown in the presented heatmap (Figure 5c), which could explain this phenomenon. The reported variants found in this study strongly suggest a possible functional connection to the genes *SYT13* and *PRDM11* in BS and *ID2* in OB cattle respectively.

The observed associated effect of the specified haplotype on chr 11 on the investigated trait mbm was negative in OB cattle. Female haplotype carriers can therefore be expected to have a lower incidence of multiple births. By contrast, the effect for the haplotype on chr 15 for BS was positive. Consequently, carriers have a higher risk for multiple births. The assessment of the potential impact of haplotypes on other available birth and calving traits showed no significant impact.

CONCLUSIONS

By analysing large-scale genotype and phenotype data in the two Swiss Braunvieh cattle populations, we have identified independent QTL for maternal multiple birth on chr 11 for OB and on chr 15 for BS. Furthermore, we provide evidence for linked variants in both breeds that possibly affect expression of *ID2* as well as *STY13* and *PRDM11* that might impact the occurrence of multiple births. Therefore, these findings improve the understanding of the genetic architecture of this complex trait in cattle and for mammalian reproduction in general. Further studies are desirable, in particular to provide functional evidence of the causal relationship we postulated between the QTL discovered and the phenotype studied.

ACKNOWLEDGMENTS

The authors would like to thank the Swiss cattle breeding organisation Braunvieh Switzerland for providing phenotypic and genomic data, the 1000 Bull Genomes Project for providing WGS data and the intergenomics Consortium for providing genomic data. Open access funding provided by Universitat Bern.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY STATEMENT

Sequence data for all animals, all from the 1000 Bull Genomes Project, are available from EVA (www.ebi.ac.uk/eva/). The SNP and phenotypic data used in this study are available from Braunvieh Schweiz (Zug, Switzerland). The availability of these data, which were used under license for the present study, is subject to restrictions and therefore not publicly available. However, data are available from the authors upon reasonable request and with permission of Braunvieh Schweiz (Zug, Switzerland).

ORCID

Sarah Widmer https://orcid.org/0000-0002-0541-5630
Franz R. Seefried https://orcid.org/0000-0003-4396-2747
Peter von Rohr https://orcid.org/0000-0003-0078-707X
Irene M. Häfliger https://orcid.org/0000-0002-5648-963X
Mirjam Spengeler https://orcid.org/0000-0001-9629-5533
Cord Drögemüller https://orcid.org/0000-0001-9773-522X

REFERENCES

Andreu-Vázquez, C., Garcia-Ispierto, I., Ganau, S., Fricke, P.M. & López-Gatius, F. (2012) Effects of twinning on the subsequent reproductive performance and productive lifespan of high-producing dairy cows. *Theriogenology*, 78, 2061–2070.

Atteneder, V. (2007) *Analyse der Zwillings- und Mehrlingsgeburten in der Österreichischen Milchviehpopulation*. Vienna, Austria: Universitat für Bodenkultur Wien.

Bierman, C.D., Kim, E., Shi, X.W., Weigel, K., Jeffrey Berger, P. & Kirkpatrick, B.W. (2010) Validation of whole genome linkage-linkage disequilibrium and association results, and identification of markers to predict genetic merit for twinning. *Animal Genetics*, 41, 406–416.

Cattle QTL Database (2022). Retrieved on February 22, 2022. https://www.animalgenome.org/cgi-bin/QTLdb/BT/traitmap?trait_ID=1078.

Cobanoglu, O., Berger, P.J. & Kirkpatrick, B.W. (2005) Genome screen for twinning rate QTL in four north American Holstein families. *Animal Genetics*, 36, 303–308.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A. et al. (2011) The variant call format and VCFtools. *Bioinformatics*, 27, 2156–2158.

Echternkamp, S.E. & Gregory, K.E. (1999) Effects of twinning on gestation length, retained placenta, and dystocia. *Journal of Animal Science*, 77, 39–47.

Fan, B., Onteru, S.K., Du, Z.-Q., Garrick, D.J., Stalder, K.J. & Rothschild, M.F. (2011) Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. *PLoS One*, 6, e14726.

Fernando, R.L. & Garrick, D.J. (2013) Bayesian methods applied to GWAS. In: Gondro, C., van der Werf, J. & Hayes, B. (Eds.) *Genome-wide association studies and genomic prediction*. New York, Heidelberg, Dordrecht, London: Springer, pp. 237–274.

Fernando, R., Toosi, A., Wolc, A., Garrick, D. & Dekkers, J. (2017) Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. *Journal of Agricultural, Biological, and Environmental Statistics*, 22, 172–193.

Fitzgerald, A.M., Berry, D.P., Carthy, T., Cromie, A.R. & Ryan, D.P. (2014) Risk factors associated with multiple ovulation and twin birth rate in Irish dairy and beef cattle. *Journal of Animal Science*, 92, 966–973.

Fog, C.K., Asmar, F., Côme, C., Jensen, K.T., Johansen, J.V., Kheir, T.B. et al. (2015) Loss of PRDM11 promotes MYC-driven lymphomagenesis. *Blood*, 125, 1272–1281.

Fog, C.K., Galli, G.G. & Lund, A.H. (2012) PRDM proteins: important players in differentiation and disease. *BioEssays*, 34, 50–60.

Garrick, D.J., Taylor, J.F. & Fernando, R.L. (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*, 41, 55.

Ghavi Hossein-Zadeh, N., Nejati-Javaremi, A., Miraei-Ashtiani, S.R. & Kohram, H. (2009) Estimation of variance components and genetic trends for twinning rate in Holstein dairy cattle of Iran. *Journal of Dairy Science*, 92, 3411–3421.

Gregory, K.E., Echternkamp, S.E., Dickerson, G.E., Cundiff, L.V., Koch, R.M. & van Vleck, L.D. (1990) Twinning in cattle: III. Effects of twinning on dystocia, reproductive traits, calf survival, calf growth and cow productivity. *Journal of Animal Science*, 68, 3133–3144.

Hayes, B.J. & Daetwyler, H.D. (2019) 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual Review of Animal Biosciences*, 7, 89–102.

Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J. & Goddard, M.E. (2010) Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, Milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics*, 6, e1001139.

Johanson, J.M., Berger, P.J., Kirkpatrick, B.W. & Dentine, M.R. (2001) Twinning rates for north American Holstein sires. *Journal of Dairy Science*, 84, 2081–2088.

Johnson, A.L. & Woods, D.C. (2009) Dynamics of avian ovarian follicle development: cellular mechanisms of granulosa cell differentiation. *General and Comparative Endocrinology*, 163, 12–17.

Karlsen, A., Ruane, J., Klemetsdal, G. & Heringstad, B. (2000) Twinning rate in Norwegian cattle: frequency, (co)variance components, and genetic trends. *Journal of Animal Science*, 78, 15–20.

Kim, E.S., Berger, P.J. & Kirkpatrick, B.W. (2009) Genome-wide scan for bovine twinning rate QTL using linkage disequilibrium. *Animal Genetics*, 40, 300–307.

Kirkpatrick, B.W. & Morris, C.A. (2015) A major gene for bovine ovulation rate. *PLoS One*, 10, e0129025.

Lett, B.M. & Kirkpatrick, B.W. (2018) Short communication: heritability of twinning rate in Holstein cattle. *Journal of Dairy Science*, 101, 4307–4311.

Lett, B.M. & Kirkpatrick, B.W. (2022) Identifying genetic variants and pathways influencing daughter averages for twinning in North American Holstein cattle and evaluating the potential for genomic selection. *Journal of Dairy Science*. Available from: https://doi.org/10.3168/jds.2021-21238. Online ahead of print.

Lien, S., Karlsen, A., Klemetsdal, G., Våge, D.I., Olsaker, I., Klungland, H. et al. (2000) A primary screen of the bovine genome for quantitative trait loci affecting twinning rate. *Mammalian Genome*, 11, 877–882.

Maher, B. (2008) Personal genomes: the case of the missing heritability. *Nature*, 456, 18–21.

Masuda, Y., Baba, T. & Suzuki, M. (2015) Genetic analysis of twinning rate and milk yield using a threshold-linear model in Japanese Holsteins. *Animal Science Journal*, 86, 31–36.

McGovern, S.P., Weigel, D.J., Fessenden, B.C., Gonzalez-Peña, D., Vukasinovic, N., McNeel, A.K. et al. (2021) Genomic prediction for twin pregnancies. *Animals*, 11, 843.

Miglior, F., Fleming, A., Malchiodi, F., Brito, L.F., Martin, P. & Baes, C.F. (2017) A 100-year review: identification and genetic selection of economically important traits in dairy cattle. *Journal of Dairy Science*, 100, 10251–10271.

MiX99 Development Team. (2017) *MiX99: a software package for solving large mixed model equations. Release 17.11.* Jokioi: Natural Resources Institute Finland (Luke).

Moioli, B., Steri, R., Marchitelli, C., Catillo, G. & Buttazzoni, L. (2017) Genetic parameters and genome-wide associations of twinning rate in a local breed, the Maremmana cattle. *Animal*, 11, 1660–1666.

Murillo-Barrantes, J., Estrada-König, S., Rojas-Campos, J., Bolaños-Segura, M., Valverde-Altamirano, E. & Romero-Zúñiga, J.J. (2010) Factores asociados con partos gemelares en vacas de fincas lecheras especializadas de Costa Rica. *Rev Ciencias Veterinarias*, 28, 7–21.

Negussie, E., Strandén, I. & Mäntysaari, E.A. (2008) Genetic analysis of liability to clinical mastitis, with somatic cell score and production traits using bivariate threshold–linear and linear–linear models. *Livestock Science*, 117, 52–59.

Neumaier, A. & Groeneveld, E. (1998) Restricted maximum likelihood estimation of covariances in sparse linear models. *Genetics Selection Evolution*, 30, 3–26.

Pardon, B., Vertenten, G., Cornillie, P., Schauvliege, S., Gasthuys, F., van Loon, G. et al. (2012) Left abomasal displacement between the uterus and rumen during bovine twin pregnancy. *Journal of Veterinary Science*, 13, 437–440.

Pausch, H., Ammermüller, S., Wurmser, C., Hamann, H., Tetens, J., Drögemüller, C. et al. (2016) A nonsense mutation in the COL7A1 gene causes epidermolysis bullosa in Vorderwald cattle. *BMC Genetics*, 17, 149.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D. et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81, 559–575.

Putz A. (2021) GenSel, GitHub. https://github.com/austin-putz/GenSel. Accessed 31 March 2021.

Quiñones-Frías, M.C. & Littleton, J.T. (2021) Function of drosophila Synaptotagmins in membrane trafficking at synapses. *Cellular and Molecular Life Sciences*, 78, 4335–4364.

R Core Team. (2020) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

de Roos, A.P.W., Hayes, B.J., Spelman, R.J. & Goddard, M.E. (2008) Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. *Genetics*, 179, 1503–1512.

Roper, L.K., Briguglio, J.S., Evans, C.S., Jackson, M.B. & Chapman, E.R. (2015) Sex-specific regulation of follicle-stimulating hormone secretion by synaptotagmin 9. *Nature Communications*, 6, 8645.

Sargolzaei M. Ontario Veterinary College, University of Guelph. https://ovc.uoguelph.ca/pathobiology/people/faculty/Mehdi-Sargolzaei (2021). Accessed 31 March 2021.

Silva del Río, N., Kirkpatrick, B.W. & Fricke, P.M. (2006) Observed frequency of monozygotic twinning in Holstein dairy cattle. *Theriogenology*, 66, 1292–1299.

Silva Del Río, N., Stewart, S., Rapnicki, P., Chang, Y.M. & Fricke, P.M. (2007) An observational analysis of twin births, calf sex ratio, and calf mortality in Holstein dairy cattle. *Journal of Dairy Science*, 90, 1255–1264.

da Silveira, J.C., Carnevale, E.M., Winger, Q.A. & Bouma, G.J. (2014) Regulation of ACVR1 and ID2 by cell-secreted exosomes during follicle maturation in the mare. *Reproductive Biology and Endocrinology*, 12, 44.

Thaller, G., Krämer, W., Winter, A., Kaupe, B., Erhardt, G. & Fries, R. (2003) Effects of DGAT1 variants on milk production traits in German cattle breeds1. *Journal of Animal Science*, 81, 1911–1918.

Thélie, A., Papillier, P., Pennetier, S., Perreau, C., Traverso, J.M., Uzbekova, S. et al. (2007) Differential regulation of abundance and deadenylation of maternal transcripts during bovine oocyte maturation in vitro and in vivo. *BMC Developmental Biology*, 7, 125.

Tier, B. & Meyer, K. (2004) Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. *Journal of Animal Breeding and Genetics*, 121, 77–89.

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91, 4414–4423.

Verbraak, E.J.C., Van 't Veld, E.M., Groot Koerkamp, M., BAJ, R., van Haeften, T., Stoorvogel, W. et al. (2011) Identification of genes targeted by FSH and oocytes in porcine granulosa cells. *Theriogenology*, 75, 362–376.

Visscher, P.M., Yang, J. & Goddard, M.E. (2010) A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Research and Human Genetics*, 13, 517–524.

Weller, J.I., Golik, M., Seroussi, E., Ron, M. & Ezra, E. (2008) Detection of quantitative trait loci affecting twinning rate in Israeli Holsteins by the daughter design. *Journal of Dairy Science*, 91, 2469–2474.

Widmer, S., Seefried, F.R., von Rohr, P., Häfliger, I.M., Spengeler, M. & Drögemüller, C. (2021) A major QTL at the LHCGR/FSHR locus for multiple birth in Holstein cattle. *Genetics Selection Evolution*, 53, 57.

Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88, 76–82.

Zavareh, S., Gholizadeh, Z. & Lashkarbolouki, T. (2018) Evaluation of changes in the expression of Wnt/β-catenin target genes in mouse reproductive tissues during estrous cycle: an experimental study. *International Journal of Reproductive BioMedicine*, 16, 69–76.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Widmer, S., Seefried, F.R., von Rohr, P., Häfliger, I.M., Spengeler, M. & Drögemüller, C. (2022) Associated regions for multiple birth in Brown Swiss and Original Braunvieh cattle on chromosomes 15 and 11. *Animal Genetics*, 53, 557–569. Available from: https://doi.org/10.1111/age.13229

Least absolute shrinkage and selection operator / support vector machine and random forest: evaluation of alternative approaches to identify associated genomic regions for monogenic and complex traits in cattle

Journal:	Animals
Manuscript status:	submitted
Contributions:	phenotyping and genotyping data preparation, methodology, data analyses, visualization of the results, writing original draft and revisions
Displayed version:	submitted version
DOI:	-

animals

MDPI

Article

Least absolute shrinkage and selection operator / support vector machine and random forest: evaluation of alternative approaches to identify associated genomic regions for monogenic and complex traits in cattle

Sarah Widmer [1], Franz R. Seefried [2], Cord Drögemüller [1,*] and Peter von Rohr [2]

[1] Institute of Genetics, Vetsuisse Faculty, University of Bern, 3012 Bern, Switzerland; sarah.widmer@unibe.ch (S.W.); cord.droegemueller@unibe.ch (C.D.)
[2] Qualitas AG, 6300 Zug, Switzerland; franz.seefried@qualitasag.ch (F.R.S.); peter.vonrohr@qualitasag.ch (P.v.R.)
* Correspondence: cord.droegemueller@unibe.ch

Simple Summary: In animal science, there is a considerable need for a simpler process to identify genomic regions associated with a trait of interest because, given large datasets, classical approaches require massive computational resources and are complex. In this study, we aim to validate machine learning methods (least absolute shrinkage and selection operator, support vector machine and random forest) as an alternative approach to detect genomic associations. In a case-control design, we successfully validated the methods for monogenic traits. For complex inherited traits, we were also able to identify already known associated genomic regions. We conclude that the evaluated machine learning methods are promising and, above all, efficient alternatives to the traditionally used approaches.

Abstract: Classical approaches to identify associated genomic regions such as genome-wide association studies (GWAS) using de-regressed breeding values are complex and require massive computing resources. This is especially true for large datasets in terms of the high number of individuals and/or high density of SNP markers, both increasingly available for livestock species such as cattle. Machine learning tools such as random forest (RF), Least Absolute Shrinkage and Selection Operator (Lasso) and Support Vector Machine (SVM) using raw phenotypes were implemented to provide a simple procedure to identify genomic regions associated with phenotypic traits of interest. While RF directly yields classification results, phenotypes can only be separated into two groups of cases and controls with a SVM after selecting relevant variables by Lasso. These approaches were successfully validated by confirming known genomic associations for two simple Mendelian traits in cattle. For the complex inherited trait stature, some of the previously identified bovine genome regions could be confirmed. Furthermore, suggestive associated regions for the trait multiple birth were found in Holstein cattle including the recently found major QTL on chromosome 11. Based on the presented data, the evaluated machine learning approaches are promising and represent efficient alternatives to traditionally used GWAS approaches.

Keywords: Bos taurus; GWAS; machine learning; quantitative trait loci (QTL); livestock

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

1. Introduction

Classical approaches to identify associated genomic regions are complex and require massive computing resources. These approaches use pseudo-phenotypes based on de-regressed predicted breeding values as responses in a statistical model [1]. The preparation procedure to obtain these pseudo-phenotypes requires a complex pipeline of

specialized software tools which is not publicly available. Hence, there is a considerable need for a simpler process that allows for the identification of genomic regions associated with a trait of interest. We are evaluating a set of three machine learning tools as the building blocks of such an identification process using raw phenotypes, which have already been used in genomic association analyses in cattle using de-regressed breeding values [2–5].

Support vector machine (SVM) is a general-purpose classification method and is used in many fields [6,7]. For example, one 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 [8]. SVM classifies data consisting of different groups by separating hyperplanes [7]. The separating hyperplanes are defined by explanatory variables such as SNP genotype effects. This definition makes SVM suitable to analyze 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. A higher level of noise reduces the quality of separation of the data into cases and controls based on the estimated hyperplane and increases the risk of overfitting which might impair classification negatively. Applying the SVM algorithm on all variables would introduce high levels of noise by irrelevant variables.

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 [9]. Lasso regularizes the coefficient estimates, shrinks them towards zero and consequently reduces their variance significantly [7]. Combining Lasso and SVM results in a computationally very efficient approach, as no prior prediction of breeding values followed by a de-regression procedure is required. The proposed method can be directly applied to raw phenotypes.

An alternative one-step machine learning tool is random forest (RF). This method uses regression trees and bootstrapping where at every tree node a variable selection will be made [10]. From permutations we can estimate an importance score for each variable. This method is already known to be used for the analysis of genomic data, which usually have a low number of animals and a high number of predictor variables [11]. A single study used raw phenotypes regarding digital dermatitis in a case/control approach to find QTL using RF combined with Bayesian regression models in Holstein cattle [12]. The RF method is easily applicable to raw data to find new associated regions.

Multiple birth is a still poorly understood polygenic trait. Besides several older studies in cattle using just a small number of markers and microsatellites, two recent studies analyzed the trait of interest on a population scale in dairy cattle. They found one quantitative trait locus (QTL) for each of the dairy breeds Holstein, Brown Swiss and Original Braunvieh using 600K SNP and whole-genome sequencing data [13,14]. These different QTL were located on various chromosomes. These population analyses provided candidate causal variants altering expression of obvious candidate genes. The major QTL for this trait in Holstein on *Bos taurus* autosome (BTA) 11 containing the *LHCGR* and *FSHR* genes was recently independently confirmed in a different population [15]. Multiple birth is a complex polygenic trait, where the described QTL account only for a small fraction of the genetic variance. Therefore, one expects that further associated regions could be detected.

In cattle breeding, high-density genotyping data of single nucleotide polymorphisms (SNP) from hundreds of thousands of animals are available and are used as genetic markers. Selection based on genetic values predicted by genetic markers has significantly increased the rate of genetic gain in the last decades [16].

In this study, we first validated the proposed methods in a case-control design as an alternative approach to detect genomic associations for two Mendelian traits. The first trait was *CNGB3*-related achromatopsia (OMIA 001365-9913), also reported as Original Braunvieh haplotype 1 (OH1) [17]. *TWIST2*-related depigmentation (OMIA 001469-9913) known as belt pattern was the second Mendelian trait [18]. Subsequently, we applied the

workflow on stature in Holstein, a complex but well studied polygenic trait [19]. Finally, the proposed methods were used to achieve the identification of additional genomic regions associated with the trait multiple birth in Swiss Holstein cattle.

2. Materials and Methods

2.1. Phenotypes

The proposed approaches were validated in a case-control design using three different datasets of three different Swiss cattle populations (Table 1). The datasets were not equal to the references used for the QTL detection [17–19], but comparable for the two Mendelian traits achromatopsia and belt pattern. The phenotypic and genotypic data were provided by the Swiss cattle breeding organizations: swissherdbook (Zollikofen, Switzerland), Holstein Switzerland (Grangeneuve, Switzerland) and Braunvieh Schweiz (Zug, Switzerland). The definition of cases and controls for both monogenic traits achromatopsia and belt was based on veterinary examinations and 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, the 500 cows with the highest and the 500 cows with the lowest measured sacrum height were used to define the groups of cases and controls. The mean of the measured sacrum heights was 158.8 cm and 144.8 cm with a standard deviation of 1.68 cm and 2.11 cm for the group 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 were assigned to the group of controls.

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

Trait	Population	Number of cases	Number of controls	Number of SNPs
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

2.2. Genotypes

Animals were genotyped under the umbrella of routine genomic selection using different commercial genotype arrays that include between 9k and 850k SNPs. A two-step imputation approach (first impute to 150k density and afterwards to high-density) was applied running the FImpute software with default parameters [20]. In each step, SNPs with a minor allele frequency <1% and an SNP call rate <0.99 were removed from the dataset. The threshold for call rate per animal was 0.95. The ASR-UCD1.2 cow assembly was used as the reference genome during imputation. At the end a filtering for minimal allele count of one was carried out for each dataset separately using the software program PLINK [21], which led to the final datasets (Table 1). No filtering was conducted with respect to departure of the Hardy-Weinberg equilibrium in order to keep the rare variants in the dataset.

2.3. Least Absolute Shrinkage and Selection Operator (Lasso)

Lasso uses the residual sum of squares plus a penalty term to get estimates of coefficients in a linear model [9]. The penalty term grows with the number of non-zero coefficient estimates. The magnitude of the penalty term can be scaled with a parameter called "λ". Assuming a value of λ that is large enough, some of the coefficient estimates are forced to be zero. Hence, Lasso can be used to estimate coefficients and to do variable selection in a linear model, simultaneously [7]. The Lasso analyses were carried out using R [22] and the package glmnet [23]. The parameter λ was estimated using a 10-fold cross-validation. For the definition of the training dataset, we used the leave one out method. For the 10-times cross validation we used 1,000 different values for λ evenly distributed between 10^{-10} and 10^{10}. The cross validation resulted in two different estimates for λ. The estimate λ_{min} corresponds to the estimate of λ with the lowest cross-validation error. Adding one standard error to λ_{min} leads to the second estimate of λ called λ_{1se}. Both

estimates of λ were used in the Lasso analyzes for all four datasets. The estimate λ_{1se} is larger than λ_{min} and therefore the SNP coefficients receive a stronger penalty when using λ_{1se} in the Lasso analysis. This leads to a lower number of SNP with a non-zero coefficient remaining in the analysis. Only SNPs with a non-zero coefficient are regarded as relevant candidate positions for being associated to a given phenotype of interest. Hence, the variable selection procedure imposed by Lasso is used to detect SNP which are relevant for a phenotype of interest. Lasso resulted in a smaller set of SNPs compared to the complete marker-set. For the observation of the results, the absolute coefficient values were combined using a sliding window approach whereof each window consisted 50 SNPs. As an importance threshold of the window coefficients, we specified the 5-fold standard deviation of all absolute coefficients.

2.4. Support Vector Machine (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 [6]. The separating hyperplane has to be estimated using the training data [7]. Since the separating hyperplane can be expressed as a mathematical function, it is often referred to as the classification model. We defined the training dataset 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, R-package e1071 [24] was applied to the SNPs selected by Lasso for the validation of the variable selection. The prediction of animals outside the high and low groups is not considered yet.

2.5. Random forest (RF)

The machine learning method RF combines information from an ensemble of classification and regression trees [10]. Each tree is built using a bootstrap sample of the SNP data set and at each node of the tree the best variable is selected from a random subset of all variables (SNPs). As a result, we get a permutation importance score for each predictor variable which measures the difference in prediction accuracy before and after permuting values of the variable over all trees [25]. If a SNP has no association with the response variable, we expect a permutation importance score of approximately zero.

For RF analyzes the R-package randomForest [26] was applied. We defined the number of trees to grow up to 1000 nodes and used default values for the remaining parameters. As there is no classical significance threshold reported for this method, we used, as for the previous method, the 5-fold standard deviation of all absolute values of importance score.

3. Results

3.1. Least Absolute Shrinkage and Selection Operator (Lasso) and Support Vector Machine (SVM)

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

Table 2. Results of Lasso and SVM analysis for all traits.

Trait	Lasso: number of selected SNPs		SVM: rate of correct prediction (%)		SVM F_1 score	
	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

From the Lasso analysis, window coefficients are plotted as a function of their chromosomal position (Figure 1). Only windows with a high impact on the trait and the colocalization of possible candidate genes were considered (Table 3). For each of the two Mendelian traits, as expected a clear signal was identified on BTA 14 for achromatopsia and BTA 3 for belt, respectively (Figures 1a and 1b). For the complex inherited 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).

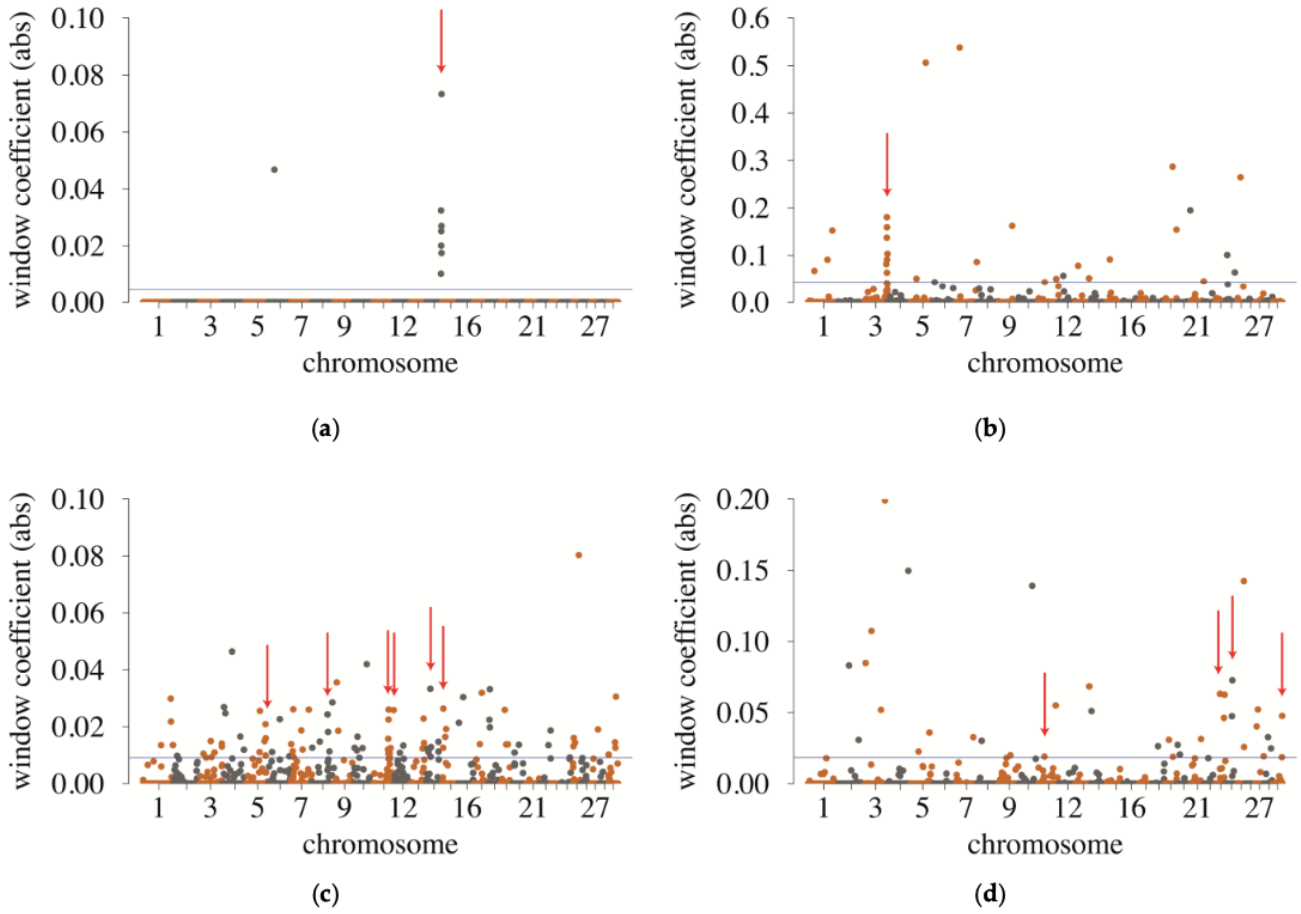


Figure 1. Genome-wide Manhattan plots of the sum of the absolute coefficient values for each window resulting from Lasso analysis. For (a) achromatopsia and (d) multiple birth λ_{min} was used, while for (b) belt and (c) stature λ_{1se} was used. Red arrows are highlighting the QTL from Table 3 and the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

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

Trait	BTA	Start position [2]	End position [2]	Coefficient [3]	Candidate genes [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	*ADCYAP1*, GATA6
	29	48,563,749	50,109,725	0.0671	*IGF2, INS*

[1] For achromatopsia and multiple birth using λ_{min} and for belt and stature using λ_{1se}
[2] Using the ASR-UCD1.2 reference assembly
[3] Sum of all absolute values of the coefficient in the interval (typically more than one window)

3.2. Random forest (RF)

The RF analysis revealed a single clear signal for the Mendelian trait of achromatopsia and belt pattern, as shown in the Manhattan plot (Figure 2a and 2b). For all four traits analyzed, the windows with a high impact on the trait and the colocalization of possible candidate genes are considered as candidate regions (Table 4). For the complex trait of stature, we detected three associated loci on BTA 3, 11 and 18 (Figure 2c). The signals for the polygenic trait of multiple births showed different associated regions, spread over different chromosomes.

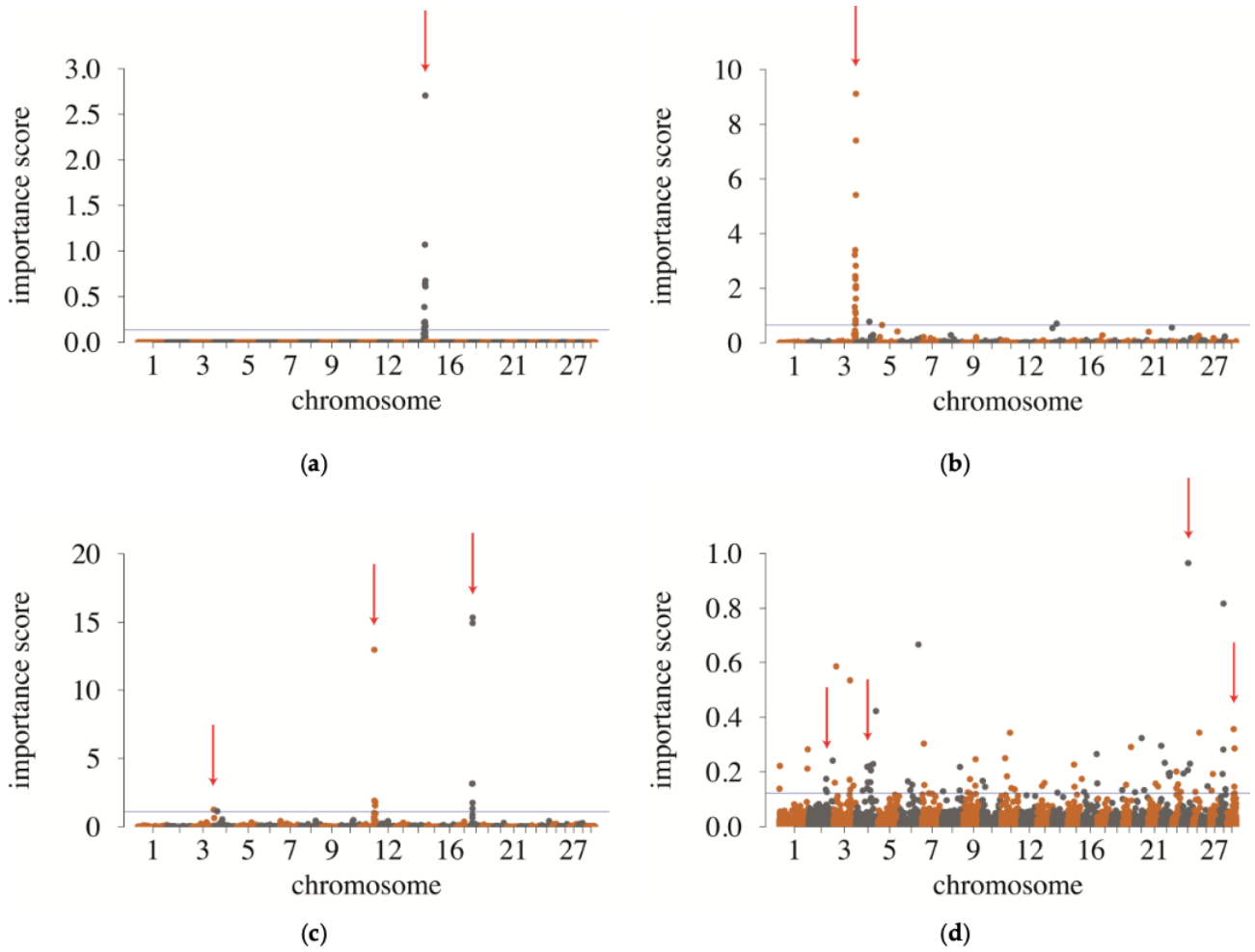


Figure 2. Genome-wide Manhattan plots of the importance scores for each SNP resulting from random forest analysis. Red arrows are highlighting the QTL from Table 4 and the blue lines indicate the values of the 5-fold standard deviation of all importance scores per trait.

Table 4. Identified associated genome regions with selected candidate genes for the four analyzed traits resulting from random forest analysis.

Trait	BTA	Start position [1]	End position [1]	Score [2]	Candidate genes [1]
Achromatopsia	14	75,583,503	78,489,957	6.4913	*CNGB3*
Belt	3	117,018,007	120,845,271	46.1483	*TWIST2*
Stature	3	117,134,264	117,258,257	1.2635	
	11	78,350,965	79,116,575	15.9239	*PUM2*
	18	13,615,771	14,972,698	36.5584	*ANKRD11, PIEZO*
Multiple birth	2	99,106,520	99,377,309	0.1271	*ERBB4*
	4	76,445,916	81,198,776	1.0718	*INHBA*
	24	33,994,132	36,319,363	1.2744	*ADCYAP1, GATA6*
	29	36,535,238	41,035,308	1.3122	*DDB1*

[1] Using the ASR-UCD1.2 reference assembly
[2] Sum of all values of the importance scores in the interval (typically more than one window).

4. Discussion

We successfully validated the machine learning approaches Lasso/SVM and RF by identifying previously known genomic associations for the two Mendelian traits achromatopsia and belt pattern. For the polygenic trait stature, previously identified QTL

were detected partially in the studied population and in addition new potential associated genomic regions were found using raw phenotypes. For the complex trait multiple birth, we found evidence for several additional suggestive associated regions in the studied Holstein cattle population.

For all four analyzed traits, Lasso resulted in a set of SNPs that 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 by Lasso, as expected from the properties. 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 rates of correct predictions from SVM were 100% for the three validation datasets and 90% for multiple births (using λ_{min}). The rate of correct predictions of SVM for both λ-estimates were comparable and hence λ_{1se} (except for achromatopsia and multiple birth) was favored because it yields sparser models. Hence, the validation of SVM was successful.

RF and Lasso easily detect the QTL for Mendelian traits. Regarding the monogenic recessive disorder achromatopsia, both approaches highlighted correctly the region of *CNGB3* on BTA 14 [17], which contains the variant that causes day blindness. For the monogenic dominant inherited belt phenotype, a signal in a window on BTA 3 with *TWIST2* that harbors the copy number variation associated with this depigmentation pattern was observed by both methods [18]. Therefore, the validation for both variable selection approaches was successful for both monogenic traits studied, the recessive traits as well as the trait with dominant inheritance.

The identification of QTL for polygenic traits is crucial. Concerning the trait of 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 [19] and identified by Lasso variable selection. We propose *MMP13* as a potential new candidate gene mapping to the region of a QTL found by Lasso on BTA 15 for stature in cattle, as it influences bone mineralization and growth plate cartilage [27]. Detected by RF analysis, the signals on BTA 3 and 11 are previously shown as QTL [19]. The strong association on BTA 18 in the RF analysis indicates a potentially interesting candidate region as the genes *PIEZO* and *ANKRD11* are annotated, which are associated with bone formation [28,29] and short stature in humans [30]. As it was published from a big meta-analysis [19], there are other strong associations with stature in cattle, such as the *PLAG1* region or the *CCND2* locus, which were not detected by our approaches. Therefore, we can observe a limitation of both methods when testing for association with a polygenic trait.

The signals for the trait of multiple birth are not as strong as the effects observed for stature. As described in the methods section, the case cohort for multiple birth is much smaller than for the trait stature. Additionally, for the trait of multiple birth, the phenotypic data was not precorrected by fixed factors as in the population analysis using linear-mixed models [13,15]. Using Lasso, the very recently published QTL [13,15] in the region of *FSHR* and *LHCGR* genes on BTA 11 was confirmed, which was not the case when using the RF method. Additionally, we reported new suggestive QTL on BTA 2, 4, 23, 24 and 29. Only the associated region on BTA 24 was identified by both variable selection approaches (RF and Lasso) and is harboring the potential candidate genes *GATA6* and *ADCYAP1*, which are associated with normal ovarian function as well as oocyte maturation and number of oocytes resulting from superovulation [31–33]. Annotated potential candidate genes in the suggestive QTL from Lasso analysis on BTA 23 and 29 are *PAQR8*, *GSTA1*, *GSTA2*, *TMEM14A*, *IGF2* and *INS*. These genes could have an impact on multiple births as they encode for a progestin receptor, react on gonadotropin, have an effect on *FSHR* expression or are important for follicle development [34–37]. From the RF analysis, the possibly associated regions on BTA 2, 4 and 29 have the genes *ERBB4*, *INHBA* and *DDB1* annotated. These genes have an effect

on folliculogenesis, follicle development related to hormones or meiotic oocyte maturation [38–40]. Consequently, they could be linked to multiple ovulations leading to multiple births.

For the more difficult polygenic traits, the reduced precision of detecting quantitative genomic regions might be due to the limited size of the dataset. Although machine learning approaches are expected to be able to account for dependency structures in the data, the obtained results might be influenced by the inherent structure caused by the pedigree relationships. Furthermore, the case-control design used in this study is expected to greatly reduce the impact of data dependency caused by pedigree relationships. A more accurate quantification of the influence of the size of the dataset and of the dependency structure in the data is subject to further research.

Comparing both variable selection approaches, RF selects more strictly than Lasso for all four traits analyzed. But as shown for the trait stature, RF identifies only a fraction of known QTL. On the other hand, Lasso might have a higher rate of false positive results, as it detected additional association also for the monogenic traits of achromatopsia and belt pattern. This problem can be addressed by quantifying the Type-1 error in the Lasso analysis, as shown by Arbet et al. [41].

5. Conclusions

Using the combination of Lasso variable selection and SVM classification as well as random forests, we were able to clearly detect the associated genomic regions for monogenic traits in cattle using raw phenotypes. However, it can be concluded that for more complex or highly polygenic traits, the proposed tools require more adaptation and fine-tuning to the specific nature of the analyzed data. For the polygenic trait of multiple birth in cattle, we reported additional suggestive QTL beside the previously identified major QTL. In summary, we propose these machine learning approaches as a promising and efficient alternative to traditional genome-wide association studies (GWAS) approaches used in livestock genomics.

Author Contributions: Conceptualization, S.W., F.R.S., C.D. and P.v.R.; methodology, S.W. and P.v.R.; software, S.W..; resources, F.R.S., C.D. and P.v.R.; writing—original draft preparation, S.W.; writing—review and editing, S.W., F.R.S., C.D. and P.v.R.; visualization, S.W..; supervision, F.R.S., C.D. and P.v.R. All authors have read and agreed to the published version of the manuscript."

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The SNP data analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Acknowledgments: The authors are grateful to the Swiss cattle breeding organizations (swissherdbook, Holstein Switzerland and Braunvieh Schweiz) for providing phenotypic and genotypic data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ostersen, T.; Christensen, O.F.; Henryon, M.; Nielsen, B.; Su, G.; Madsen, P. Deregressed EBV as the Response Variable Yield More Reliable Genomic Predictions than Traditional EBV in Pure-Bred Pigs. *Genet Sel Evol* **2011**, *43*, 38.
2. Waldmann, P.; Mészáros, G.; Gredler, B.; Fuerst, C.; Sölkner, J. Evaluation of the Lasso and the Elastic Net in Genome-Wide Association Studies. *Front Genet* **2013**, *4*, 270.
3. Rafter, P.; Gormley, I.C.; Parnell, A.C.; Naderi, S.; Berry, D.P. The Contribution of Copy Number Variants and Single Nucleotide Polymorphisms to the Additive Genetic Variance of Carcass Traits in Cattle. *Front Genet* **2021**, *12*, 761503.
4. Mokry, F.; Higa, R.; de Alvarenga Mudadu, M.; Oliveira de Lima, A.; Meirelles, S.L.; Barbosa da Silva, M.V.; Cardoso, F.; Morgado de Oliveira, M.; Urbinati, I.; Méo Niciura, S.; et al. Genome-Wide Association Study for Backfat Thickness in Canchim Beef Cattle Using Random Forest Approach. *BMC Genet* **2013**, *14*, 47.

5. Alves, A.A.C.; da Costa, R.M.; Fonseca, L.F.S.; Carvalheiro, R.; Ventura, R.V.; Rosa, G.J. de M.; Albuquerque, L.G. A Random Forest-Based Genome-Wide Scan Reveals Fertility-Related Candidate Genes and Potential Inter-Chromosomal Epistatic Regions Associated With Age at First Calving in Nellore Cattle. *Front Genet* **2022**, *13*, 834724.
6. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach Learn* **1995**, *20*, 273–297.
7. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; 2nd ed.; Springer US: New York, NY, USA, 2021; ISBN 978-1-0716-1417-4.
8. Ban, H.-J.; Heo, J.Y.; Oh, K.-S.; Park, K.-J. Identification of Type 2 Diabetes-Associated Combination of SNPs Using Support Vector Machine. *BMC Genet* **2010**, *11*, 26.
9. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc, B: Stat Methodol* **1996**, *58*, 267–288.
10. Breiman, L. Random Forests. *Mach Learn* **2001**, *45*, 5–32.
11. Chen, X.; Ishwaran, H. Random Forests for Genomic Data Analysis. *Genomics* **2012**, *99*, 323–329.
12. Lai, E.; Danner, A.L.; Famula, T.R.; Oberbauer, A.M. Genome-Wide Association Studies Reveal Susceptibility Loci for Digital Dermatitis in Holstein Cattle. *Animals* **2020**, *10*, 2009.
13. Widmer, S.; Seefried, F.R.; von Rohr, P.; Häfliger, I.M.; Spengeler, M.; Drögemüller, C. A Major QTL at the LHCGR/FSHR Locus for Multiple Birth in Holstein Cattle. *Genet Sel Evol* **2021**, *53*, 57.
14. Widmer, S.; Seefried, F.R.; von Rohr, P.; Häfliger, I.M.; Spengeler, M.; Drögemüller, C. Associated Regions for Multiple Birth in Brown Swiss and Original Braunvieh Cattle on Chromosomes 15 and 11. *Anim Genet* **2022**, *53*, 557–569.
15. Lett, B.M.; Kirkpatrick, B.W. Identifying Genetic Variants and Pathways Influencing Daughter Averages for Twinning in North American Holstein Cattle and Evaluating the Potential for Genomic Selection. *J Dairy Sci* **2022**, *105*, 5972–5984.
16. Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **2001**, *157*, 1819–1829.
17. Häfliger, I.M.; Marchionatti, E.; Stengård, M.; Wolf-Hofstetter, S.; Paris, J.M.; Jacinto, J.G.P.; Watté, C.; Voelter, K.; Occelli, L.M.; Komáromy, A.M.; et al. CNGB3 Missense Variant Causes Recessive Achromatopsia in Original Braunvieh Cattle. *Int J Mol Sci* **2021**, *22*, 12440.
18. Awasthi Mishra, N.; Drögemüller, C.; Jagannathan, V.; Keller, I.; Wüthrich, D.; Bruggmann, R.; Beck, J.; Schütz, E.; Brenig, B.; Demmel, S.; et al. A Structural Variant in the 5′-Flanking Region of the TWIST2 Gene Affects Melanocyte Development in Belted Cattle. *PLoS One* **2017**, *12*, e0180170.
19. Bouwman, A.C.; Daetwyler, H.D.; Chamberlain, A.J.; Ponce, C.H.; Sargolzaei, M.; Schenkel, F.S.; Sahana, G.; Govignon-Gion, A.; Boitard, S.; Dolezal, M.; et al. Meta-Analysis of Genome-Wide Association Studies for Cattle Stature Identifies Common Genes That Regulate Body Size in Mammals. *Nat Genet* **2018**, *50*, 362–367.
20. Sargolzaei, M.; Schenkel, F.; Chesnais, J. FImpute-An Efficient Imputation Algorithm for Dairy Cattle Populations. *J Dairy Sci* **2011**, *94*, 421–421.
21. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **2007**, *81*, 559–575.
22. R Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. Available online: https://www.R-project.org/ (accessed on 18 December 2020).
23. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **2010**, *33*, 1–22.
24. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R Package Version 1.7-9 2021. Available online: https://CRAN.R- 339 project.org/package=e1071 (accessed on 07 November 2022).
25. Nicodemus, K.K.; Malley, J.D.; Strobl, C.; Ziegler, A. The Behaviour of Random Forest Permutation-Based Variable Importance Measures under Predictor Correlation. *BMC Bioinform* **2010**, *11*, 110.
26. Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2*, 18–22.
27. Ståhle-Bäckdahl, M.; Sandstedt, B.; Bruce, K.; Lindahl, A.; Jiménez, M.G.; Vega, J.A.; López-Otín, C. Collagenase-3 (MMP-13) Is Expressed during Human Fetal Ossification and Re-Expressed in Postnatal Bone Remodeling and in Rheumatoid Arthritis. *Lab Invest* **1997**, *76*, 717–728.
28. Sun, W.; Chi, S.; Li, Y.; Ling, S.; Tan, Y.; Xu, Y.; Jiang, F.; Li, J.; Liu, C.; Zhong, G.; et al. The Mechanosensitive Piezo1 Channel Is Required for Bone Formation. *eLife* **2019**, *8*, e47454.
29. Dzamukova, M.; Brunner, T.M.; Miotla-Zarebska, J.; Heinrich, F.; Brylka, L.; Mashreghi, M.-F.; Kusumbe, A.; Kühn, R.; Schinke, T.; Vincent, T.L.; et al. Mechanical Forces Couple Bone Matrix Mineralization with Inhibition of Angiogenesis to Limit Adolescent Bone Growth. *Nat Commun* **2022**, *13*, 3059.
30. Sirmaci, A.; Spiliopoulos, M.; Brancati, F.; Powell, E.; Duman, D.; Abrams, A.; Bademci, G.; Agolini, E.; Guo, S.; Konuk, B.; et al. Mutations in ANKRD11 Cause KBG Syndrome, Characterized by Intellectual Disability, Skeletal Malformations, and Macrodontia. *Am J Hum Genet* **2011**, *89*, 289–294.
31. Bennett, J.; Wu, Y.-G.; Gossen, J.; Zhou, P.; Stocco, C. Loss of GATA-6 and GATA-4 in Granulosa Cells Blocks Folliculogenesis, Ovulation, and Follicle Stimulating Hormone Receptor Expression Leading to Female Infertility. *Endocrinology* **2012**, *153*, 2474–2485.

32. Barberi, M.; di Paolo, V.; Latini, S.; Guglielmo, M.C.; Cecconi, S.; Canipari, R. Expression and Functional Activity of PACAP and Its Receptors on Cumulus Cells: Effects on Oocyte Maturation. *Mol Cell Endocrinol* **2013**, *375*, 79–8.
33. Koppan, M.; Varnagy, A.; Reglodi, D.; Brubel, R.; Nemeth, J.; Tamas, A.; Mark, L.; Bodis, J. Correlation Between Oocyte Number and Follicular Fluid Concentration of Pituitary Adenylate Cyclase-Activating Polypeptide (PACAP) in Women After Superovulation Treatment. *J Mol Neurosci* **2012**, *48*, 617–622.
34. Rabahi, F.; Bruĺé, S.; Sirois, J.; Beckers, J.-F.; Silversides, D.W.; Lussier, J.G. High Expression of Bovine α Glutathione S-Transferase (GSTA1, GSTA2) Subunits Is Mainly Associated with Steroidogenically Active Cells and Regulated by Gonadotropins in Bovine Ovarian Follicles. *Endocrinology* **1999**, *140*, 3507–3517.
35. Liu, Y.; Zhou, Z.; He, X.; Tao, L.; Jiang, Y.; Lan, R.; Hong, Q.; Chu, M. Integrated Analyses of MiRNA-MRNA Expression Profiles of Ovaries Reveal the Crucial Interaction Networks That Regulate the Prolificacy of Goats in the Follicular Phase. *BMC Genom* **2021**, *22*, 812.
36. Bøtkjær, J.A.; Pors, S.E.; Petersen, T.S.; Kristensen, S.G.; Jeppesen, J.V.; Oxvig, C.; Andersen, C.Y. Transcription Profile of the Insulin-like Growth Factor Signaling Pathway during Human Ovarian Follicular Development. *J Assist Reprod Genet* **2019**, *36*, 889–903.
37. van den Hurk, R.; Zhao, J. Formation of Mammalian Oocytes and Their Growth, Differentiation and Maturation within Ovarian Follicles. *Theriogenology* **2005**, *63*, 1717–1751.
38. Veikkolainen, V.; Ali, N.; Doroszko, M.; Kiviniemi, A.; Miinalainen, I.; Ohlsson, C.; Poutanen, M.; Rahman, N.; Elenius, K.; Vainio, S.J.; et al. Erbb4 Regulates the Oocyte Microenvironment during Folliculogenesis. *Hum Mol Genet* **2020**, *29*, 2813–2830.
39. Bao, Y.; Yao, X.; Li, X.; EI-Samahy, M.A.; Yang, H.; Liang, Y.; Liu, Z.; Wang, F. INHBA Transfection Regulates Proliferation, Apoptosis and Hormone Synthesis in Sheep Granulosa Cells. *Theriogenology* **2021**, *175*, 111–122.
40. Yu, C.; Ji, S.-Y.; Sha, Q.-Q.; Sun, Q.-Y.; Fan, H.-Y. CRL4–DCAF1 Ubiquitin E3 Ligase Directs Protein Phosphatase 2A Degradation to Control Oocyte Meiotic Maturation. *Nat Commun* **2015**, *6*, 8017.
41. Arbet, J.; McGue, M.; Chatterjee, S.; Basu, S. Resampling-Based Tests for Lasso in Genome-Wide Association Studies. *BMC Genet* **2017**, *18*, 70.

4 Discussion and outlook

During the work on this thesis, I investigated the genetic background of multiple birth events and designed a BV estimation for this novel phenotype in Switzerland. By applying GWAS, a QTL for multiple births was detected in each of the three different Swiss dairy cattle populations Holstein, Brown Swiss and Original Braunvieh on chromosome 11, 15 and 11, respectively (Table 2). Alltogheter, I identified candidate causal variants affecting the expression of the genes *FSHR*, *LHCGR*, *ID2*, *PRDM11* and *SYT13*. Furthermore, preliminary work was presented for using the machine learning tools Lasso, SVM and RF for identifying QTL on raw phenotypes, such as the binary trait multiple birth events.

Table 2: The identified quantitative trait loci and the associated candidate variants from population analyses for multiple births in Swiss dairy cattle.

Population	Chr	QTL (Mb)	Candidate variant[1]	Variant ID	Impact of variant	Associated genes	AF (%)	Ref
Holstein	11	31 - 32	31,089,325C>G	rs386084479	intergenic	*LHCGR*, *FSHR*	28.6	[54]
Brown Swiss	15	75 - 76	75,213,046T>C	rs382040594	intron	*PRDM11*	14.2	[55]
			75,297,912C>T	rs380985793	3′ UTR	*SYT13*	14.2	
			75,399,114G>T	rs521527753	intergenic	*SYT13*	14.2	
			75,402,900G>A	rs721175231	intergenic	*SYT13*	14.0	
			75,405,408T>G	rs42930921	intergenic	*SYT13*	14.2	
Original Braunvieh	11	88 - 89	88,791,842 A>T	rs109730673	intergenic	*ID2*	34.3	[55]

Chr = chromosome, QTL = quantitative trait loci, Mb = mega base, ID = identification, AF = allele frequency, Ref = reference, UTR = untranslated region

[1] Based on the reference sequence ARS-UCD1.2

As one can observe in Table 2, the identified QTL for multiple births were detected in different genome regions for each breed. Consequently, the candidate variants were annotated to different genes. Furthermore, also suggestive associations, identified through the population analysis, showed no overlap across the populations [54, 55]. Consequently, one sees clear breed specific genetic associations for multiple births.

Even though the associated regions varied between the breeds, the proposed candidate variants segregate in various international cattle breeds [54, 55]. Therefore, we assume that ancient variants are associated with multiple birth events, which were developed before the separation of modern cattle breeds.

Our identified QTL and candidate variants are novel and relate to the female reproductive cycle. The QTL on chromosome 11 and the associated genes *LHCGR* and *FSHR* in Holstein were confirmed by a study using north-American Holstein data [56]. As reported, most of the twins are dizygotic [26, 27] and the impact of the female fertility on the occurence of multiple ovulations is high. The effect of the sire on multiple births is low. This is reflected in our results, where solely significant QTL were detected for the maternal trait of multiple births and none for the direct trait, which shows the effect of the dam on the trait [54, 55]. Furthermore, the genes *LHCGR* and *FSHR* encode receptors of three essential hormones for female reproduction: luteinizing hormone (LH), choriogonadotropin, and follicle stimulating hormone (FSH), which shows that both genes are important regulators of the female reproduction cycle. In Brown Swiss, the identified QTL and candidate variants on chromosome 15 are associated with the genes *PRDM11* and *STY13*. Neither gene is known to have a direct effect on reproduction, but other members of their protein family influence the germ cell specification or the release of FSH in female mice and consequently the oestrus cycle and ovulation [57, 58]. The QTL on chromosome 11 in Original Braunvieh is associated with the gene *ID2*, which is expressed at different levels during the oestrus cycle in the ovarian tissue in mice and may play a role in negatively regulating cell differentiation [59]. Consequently, all QTL are associated with female reproduction.

All identified candidate causal variants are either intergenic, intronic or UTR (untranslated region) variants, and therefore could have a regulatory effect on the associated genes. Unfortunately, the precise effects of the variants are not predictable yet. Recently, there have been efforts for animal genomes to understand and estimate the effects of regulatory variants in detail [60]; however, further progress is needed in the future. To truly prove the hypothesised causality, gene expression data for cows carrying the different genotypes of the identified marker would be necessary and recommendable. Within the scope of this thesis, we were not able to collect this data due to limited resources. A possible approach to collect expression data would be to identify animals carrying the different genotypes of the candidate variants,

superovulate these females, sample follicles, oocytes, ovary and granulosa cells of the animals and measure the expression level of the different genes. A possible statistical method to test causality is the Mendelian randomization analysis [61, 62]. In this analysis, the genetic variants (candidate causal variants) are used as instrumental variables to unravel a possible relationship between intermediate phenotypes, such as gene expression levels, and the trait of interest.

In addition, the GWAS approach, which is underpinned by a Bayes B algorithm, and each of the machine learning methods Lasso and RF provided suggestive QTL for the trait of multiple birth events. The potentially associated regions detected by machine learning approaches are linked to some candidate genes, e.g. such as *ADCYAP1*, *PAQR8* and *IGF2*. These encode for a progestin receptor or are known to affect follicle development and oocyte maturation [63–65]. These are promising findings that can improve our understanding of the genetic architecture of the polygenic trait of multiple birth and need to be confirmed in future analyses.

Comparing the population analysis based on de-regressed proofs in linear mixed models and the case-control design with machine learning algorithms, the identification of QTL using linear mixed models leads to clearer results. The used pseudo-phenotypes resulted from the de-regression according to Garrick et al. (2009) [66] of estimated BV for multiple births. Thereby, the BVs are already corrected for the fixed effects parity of the dam, season and use of sexed semen and the random herd-year effect. Accordingly, the dataset used for the GWAS studies is more appropriate, as it is corrected for a manyfold of environmental effects and provides a better base to estimate the genetic effects affecting the trait of interest. This could lead to a better dataset to decipher the genetic causes and to identify QTL.

For more accurate results, better phenotype recordings are needed and more intense genotyping of female breeding animals. This could increase the possibilities for the use of raw phenotypes and the detection of significant results for multiple births. Efforts to obtain more genotyped cows are ongoing, especially with regards to genomic selection. As the risk for giving birth to twins and multiples increases with age and parity of the dam [67] and the genotyped cows are getting older, we can expect more available data in the future, especially in terms of cases. This will increase the power of the case-control analyses using raw phenotypes, such as proposed by our study

using the machine learning tools RF, Lasso and SVM and hopefully to the identification of QTL in the next few years.

In general, data quality and quantity are crucial for identifying significant associations. We observed this during our analysis for the Simmental breed: We were not able to find associated genome regions, as the power of the analysis was too low, because the amount of genotyped animals in this dual-purpose Swiss breed is very limited. In addition, the breeders of Simmental animals are not genotyping individuals proactively, as they attach more importance to the traditional breeding program with progeny testing. Especially in local and small breeds, a high proportion of the population has to be genotyped, in order to perform statistically sound genetic association analyses [68], especially for low heritable traits. Regarding the available phenotyping data, it would be beneficial to have additional information. Potential confounding with the observed phenotypes can be caused by the lack of information regarding veterinary and hormonal treatments prior to inseminations. Hence, these environmental effects cannot be included in the model of BV estimation and not be measured. The key factor to solve this issue is a closer cooperation between veterinarians, farmers and breeding organisations. Efforts to improve health data collections are being continuously increased, but not completed yet.

Multiple birth events are a binary trait which can cause challenges regarding the choice of statistical evaluation. Especially for linear models, which assume a normal distribution of the response variable as default. Regarding the BV estimation, an overestimation of the genetic variances of binary traits cannot be denied. When we compare our results and predictions of genetic variances for the Swiss population, they obviously do not differ compared to previous estimations [25, 67, 69]. Therefore, it was decided that the effect of the binary trait in our analyses can be neglected. Nevertheless, for future studies the use of alternative models, such as a generalised linear mixed model or a threshold model, could be interesting for our trait of interest [25, 70].

Comparing the two variable selection approaches from the machine learning tools, RF selects more strictly than Lasso in the example of the binary traits. While RF identifies a small fraction of known QTL, Lasso shows a higher rate of false positive results. Therefore, if we want to use the combined Lasso/SVM approach, we need to implement a type-1-error-control as shown in another application of Lasso [71]. For

future work on the trait of multiple births and other projects, the machine learning methods need to be further developed.

Multiple births are heritable, but on a low level. In the studies presented, a heritability ranging from 3% to 4% was estimated, depending on the cattle population. Due to this low heritability, the detection of QTL is challenging compared with highly heritable traits. The decoding of the genetic architecture of polygenic traits is also discerning, as many associations with small effects have to be uncovered. It is particularly challenging when not all environmental effects are registered and included in the model, as given in our case and mentioned above. This would explain the fact that we were able to explain only a small fraction of the genetic variance with our tools. The significantly associated QTL identified in our GWAS analyses explain about 4 - 16% of the genetic variance in the respective breed [54, 55]. An international dataset of phenotypes that is as accurate as possible, could lead to higher power of the analyses and detection of a larger proportion of associated genomic regions, especially in the global Holstein breed. Nevertheless, this paragraph shows clearly the limitations of this work, due to the complexity of the phenotype and the associated data quality issues.

In the presented studies, the autosomal chromosomes were analysed. The X-chromosome was not considered due to technical difficulties. Firstly, we used the genotyping data from the routine genomic BVs estimation, which does not include the sex chromosomes. Secondly, for the GWAS studies, BVs for sires were also predicted and used as response variables. The inclusion of the X-chromosome would have needed serious adjustment on the algorithms used for the analyses, as the X-chromosome is haploid in the males and diploid in the females. The consequence is a lack of knowledge regarding the contribution of the X-chromosome to our trait of interest. For male fertility traits a contribution was previously shown [72, 73]. Regarding female fertility a former study detected an effect on prolongation of the pregnancy of the gene *FOXP3* on the X-chromosome [74], but no effect on ovulation is reported yet.

As reported for the QTL and the associated haplotype on chromosome 11 in Holsteins [54], the genomic regions have an effect on multiple births; however, also influence female fertility in general. Due to the fact that most multiple birth events occur due to multiple ovulations, the whole reproduction cycle is influenced by these candidate regions. Especially when essential hormones and their expression are involved as in the main QTL for *LHCGR*/*FSHR* [54]. Especially regarding the potential

implementation in a breeding scheme more thorough analyses should be conducted. For example, impacts of the candidate variants/haplotypes or genetic correlations on other phenotypes could be investigated.

In comparison to other species, in cattle breeding we have extensive phenotyping and genotyping data available from large half-sib families. Therefore, cattle can serve as a model organism for other livestock species and humans. In addition, compared to humans, the availability of data is simplified because data protection guidelines are less strict. The potential of model organisms can be demonstrated by the QTL region identified in the Holstein breed carrying the candidate genes *FSHR* and *LHCGR*. The same QTL was found in a study of the Flemish and Dutch human population, identifying evidence for an association of the homologous region on human chromosome 2 with the occurrence of spontaneous dizygotic twins [75].

In conclusion, the presented thesis summarizes the comprehensive studies on the female fertility trait multiple births in the Swiss dairy breeds Holstein, Brown Swiss and Original Braunvieh. Therein the polygenic trait was analysed with different linear mixed models and led to the identification of QTL and associated candidate causative variants. Additionally, I showed the potential of new machine learning tools for the identification of associations between genomic regions and phenotype records in cattle. In future, the machine learning tools RF, Lasso and SVM can offer a low input alternative for genomic association studies. The detection of QTL for multiple birth events improves our understanding of the genetic architecture underlying this female fertility trait. By developing the BV estimation, we set the foundation for an implementation of our knowledge in a breeding program. Considering this novel phenotype will allow for breeding against multiple births in future and improve the sustainability of dairy cattle farming.

5 Acknowledgments

During my Master's thesis, the joy of research in animal genetics awoke in me. Prof. Dr. Cord Drögemüller invited me to the Institute of Genetics at the University of Bern and offered me a PhD position in the field of cattle genetics, which is dear to my heart. Thank you Cord for your amazing support, your trust, your patience and your great mentoring.

A big thanks goes to my project partners Franz Seefried and Peter von Rohr of Qualitas AG. You supported me the whole time, gave me the best methodical advises and provided the necessary data sets. Additionally, I want to thank Franz for being my co-supervisor.

A big thanks also to the co-authors of my publications, Mirjam Spengeler and Irene Häfliger for providing important information and data for the scientific work.

My work was only possible due to the good collaboration with our partners: the breeding associations and the whole genetics team of Qualitas AG.

I would like to acknowledge the thesis committee: Prof. Dr. Carlo Largiadèr for being my mentor and Prof. Dr. Mathew Littlejohn for being my external co-referee. Thank you all for your interesting inputs during the time of my PhD.

Further thanks go to the people who proof-read my thesis. I appreciate your contribution.

To the whole team of the Institute of Genetics, thanks for a funny, inspiring and motivating working environment. Even though the pandemic changed everything for some time, it was a pleasure to be a part of the team.

Finally, a big “Thank you!” to my family and my friends. You supported me the whole time. Mum and dad, thanks for your patience and consideration during these three years.

6 Curriculum vitae

Removed due to data privacy reasons.

7 List of publications

7.1 Publications in peer-reviewed scientific journals

Widmer, S., Seefried, F.R., von Rohr, P., Häfliger, I.M., Spengeler, M. and Drögemüller, C. **2021**. A major QTL at the *LHCGR*/*FSHR* locus for multiple birth in Holstein cattle. *Genetics Selection Evolution*. https://doi.org/10.1186/s12711-021-00650-1

Widmer, S., Seefried, F.R., von Rohr, P., Häfliger, I.M., Spengeler, M. and Drögemüller, C. **2022**. Associated regions for multiple birth in Brown Swiss and Original Braunvieh cattle on chromosomes 15 and 11. *Animal Genetics*. https://doi.org/10.1111/age.13229

Widmer, S., Seefried, F.R., Drögemüller, C. and von Rohr, P. 2022. Lasso/SVM and random forest: evaluation of alternative approaches to identify associated genomic regions for monogenic and complex traits in cattle. *Animals*. Submitted.

7.2 Conference attendances and invited talks

7.2.1 Oral presentations at conferences

Widmer, S., Seefried, F.R., von Rohr, P., Häfliger, I.M., Spengeler, M. and Drögemüller, C. QTL for multiple birth in Brown Swiss and Original Braunvieh cattle on chromosome 15 and 11. 73rd Annual Meeting of the European Federation of Animal Science (EAAP), Porto, PRT, September 5-9, **2022.**

Widmer, S., Seefried, F.R., Drögemüller, C. and von Rohr, P. LASSO and SVM: an alternative approach to identify associated genome regions for simple and complex traits in cattle. 12th World Congress on Genetics Applied to Livestock Production (WCGALP), Rotterdam, NLD, July 3-8, **2022.**

Widmer, S., Seefried, F.R., Flury C. and Drögemüller, C. A major QTL for brachygnathia inferior in Brown Swiss cattle. American Dairy Science Association Annual Meeting, Kansas City, MO, USA, June 19-22, **2022.**

Widmer, S., Seefried, F.R., von Rohr, P., Häfliger, I.M., Spengeler, M. and Drögemüller, C. A major QTL at *LHCGR* for multiple birth in Holstein cattle. 72nd

Annual Meeting of the European Federation of Animal Science (EAAP), Davos, CHE, August 30 - September 3, **2021.**

Widmer, S., von Rohr, P., Drögemüller, C. and Seefried, F.R. Genetic analysis of multiple birth events in cattle. 71st Annual Meeting of the European Federation of Animal Science (EAAP), virtual meeting, December 1-4, **2020.**

7.2.2 Poster presentations at conferences

Widmer, S., Seefried, F.R., Flury C. and Drögemüller, C. Brachygnathia inferior in Brown Swiss Cattle: A Simple Mendelian Trait?. Mendel Genetics Conference, Brno, CZE, July 20-23, **2022.**

7.2.3 Invited talks

A major QTL at the *LHCGR*/*FSHR* locus for multiple birth in Holstein cattle. Science at Lunch, Vetsuisse Faculty, University of Bern, November 16 **2021.**

Update Zwillingsprojekt. Forschungsausschuss Association Swiss cattle breeders (ASR), online, November 6 **2020.**

8 References

1. FAO. FAOSTAT. 2022. https://www.fao.org/faostat/en/#data. Accessed 28 Oct 2022.

2. Pitt D, Sevane N, Nicolazzi EL, MacHugh DE, Park SDE, Colli L, et al. Domestication of cattle: Two or three events? Evol Appl. 2019;12:123–36.

3. Identitas AG. Tierstatistik Rinder Identitas. 2022. https://tierstatistik.identitas.ch/de/cattle.html. Accessed 28 Oct 2022.

4. Holstein Switzerland. Geschäftsbericht 2021. Holstein Switzerland. 2022. https://www.holstein.ch/wp-content/uploads/rapport_de_gestion2021-WEB.pdf. Accessed 28 Oct 2022.

5. Swissherdbook. Geschäftsbericht 2021. 2022. https://www.swissherdbook.ch/fileadmin/04_Publikationen/04.7_Geschaeftsbericht/D_shb_GB_2021_WEB.pdf. Accessed 28 Oct 2022.

6. Braunvieh Schweiz. Geschäftsbericht 2021. 2022. https://homepage.braunvieh.ch/wp-content/uploads/2022/03/Geschaeftsbericht-2021-D.pdf. Accessed 28 Oct 2022.

7. Swissherdbook. Homepage swissherdbook. 2022. https://www.swissherdbook.ch/. Accessed 28 Oct 2022.

8. Braunvieh Schweiz. Pictures of Brown Swiss and Original Braunvieh cows. Copyright by Konrad Lustenberger. 2022.

9. Falconer DS, Mackay TF. Introduction to quantitative genetics. 4th edition. Harlow, Essex, UK: Pearson Prentice Hall; 1996.

10. Miglior F, Fleming A, Malchiodi F, Brito LF, Martin P, Baes CF. A 100-Year Review: Identification and genetic selection of economically important traits in dairy cattle. J Dairy Sci. 2017;100:10251–71.

11. Goddard ME, Hayes BJ, Meuwissen THE. Genomic selection in livestock populations. Genet Res (Camb). 2010;92:413–21.

12. Wiggans GR, Carrillo JA. Genomic selection in United States dairy cattle. Front Genet. 2022;13:994466.

13. Philipsson J. Genetic aspects of female fertility in dairy cattle. Livest Prod Sci. 1981;8:307–19.

14. Hodel F, Moll J, Kuenzi N. Analysis of fertility in Swiss Simmental cattle — Genetic and environmental effects on female fertility. Livest Prod Sci. 1995;41:95–103.

15. Georges M, Charlier C, Hayes B. Harnessing genomic information for livestock improvement. Nat Rev Genet. 2019;20:135–56.

16. Häfliger IM, Seefried FR, Drögemüller C. Reverse Genetic Screen for Deleterious Recessive Variants in the Local Simmental Cattle Population of Switzerland. Animals. 2021;11:3535.

17. Häfliger IM, Seefried FR, Spengeler M, Drögemüller C. Mining massive genomic data of two Swiss Braunvieh cattle populations reveals six novel candidate variants that impair reproductive success. Genet Sel Evol. 2021;53:95.

18. Häfliger IM, Spengeler M, Seefried FR, Drögemüller C. Four novel candidate causal variants for deficient homozygous haplotypes in Holstein cattle. Sci Rep. 2022;12:5435.

19. Goddard M. Fitness Traits in Animal Breeding Programs. In: In: van der Werf, J., Graser, HU., Frankham, R., Gondro, C. (eds) Adaptation and Fitness in Animal Populations: Evolutionary and Breeding Perspectives on Genetic Resource Management. Dordrecht, NLD: Springer Netherlands; 2009. p. 41–52.

20. Hiltpold M, Janett F, Mapel XM, Kadri NK, Fang Z-H, Schwarzenbacher H, et al. A 1-bp deletion in bovine QRICH2 causes low sperm count and immotile sperm with multiple morphological abnormalities. Geneti Sel Evol. 2022;54:18.

21. Hiltpold M, Niu G, Kadri NK, Crysnanto D, Fang Z-H, Spengeler M, et al. Activation of cryptic splicing in bovine WDR19 is associated with reduced semen quality and male fertility. PLoS Genet. 2020;16:e1008804.

22. Hiltpold M, Kadri NK, Janett F, Witschi U, Schmitz-Hsu F, Pausch H. Autosomal recessive loci contribute significantly to quantitative variation of male fertility in a dairy cattle population. BMC Genom. 2021;22:225.

23. Weller JI, Golik M, Seroussi E, Ron M, Ezra E. Detection of Quantitative Trait Loci Affecting Twinning Rate in Israeli Holsteins by the Daughter Design. J Dairy Sci. 2008;91:2469–74.

24. Moioli B, Steri R, Marchitelli C, Catillo G, Buttazzoni L. Genetic parameters and genome-wide associations of twinning rate in a local breed, the Maremmana cattle. Animal. 2017;11:1660–6.

25. Johanson JM, Berger PJ, Kirkpatrick BW, Dentine MR. Twinning Rates for North American Holstein Sires. J Dairy Sci. 2001;84:2081–8.

26. Silva del Río N, Kirkpatrick BW, Fricke PM. Observed frequency of monozygotic twinning in Holstein dairy cattle. Theriogenology. 2006;66:1292–9.

27. Atteneder V. Analyse der Zwillings- und Mehrlingsgeburten in der Österreichischen Milchviehpopulation. Universitat für Bodenkultur Wien; 2007. https://epub.boku.ac.at/obvbokhs/download/pdf/1035817?originalFilename=true. Accessed 18 Dec 2020.

28. Nielen M, Schukken YH, Scholl DT, Wilbrink HJ, Brand A. Twinning in dairy cattle: A study of risk factors and effects. Theriogenology. 1989;32:845–62.

29. Gregory KE, Echternkamp SE, Dickerson GE, Cundiff L v, Koch RM, van Vleck LD. Twinning in cattle: III. Effects of twinning on dystocia, reproductive traits, calf survival, calf growth and cow productivity. J Anim Sci. 1990;68:3133–44.

30. Fricke PM. Twinning in Dairy Cattle. Prof Anim Sci. 2001;17:61–7.

31. Kim ES, Shi X, Cobanoglu O, Weigel K, Berger PJ, Kirkpatrick BW. Refined mapping of twinning-rate quantitative trait loci on bovine chromosome 5 and analysis of insulin-like growth factor-1 as a positional candidate gene1. J Anim Sci. 2009;87:835–43.

32. Kim ES, Berger PJ, Kirkpatrick BW. Genome-wide scan for bovine twinning rate QTL using linkage disequilibrium. Anim Genet. 2009;40:300–7.

33. Bierman CD, Kim E, Shi XW, Weigel K, Jeffrey Berger P, Kirkpatrick BW. Validation of whole genome linkage-linkage disequilibrium and association results, and identification of markers to predict genetic merit for twinning. Anim Genet. 2010;41:406–16.

34. Weinberg W. Zur Bedeutung der Mehrlingsgeburten fuer die Frage der Bestimmung des Geschlechts. Arch Rass Ges Biol. 1909;:28–32.

35. Mbarek H, Steinberg S, Nyholt DR, Gordon SD, Miller MB, McRae AF, et al. Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility. Am J Hum Genet. 2016;98:898–908.

36. Mbarek H, van de Weijer MP, van der Zee MD, Ip HF, Beck JJ, Abdellaoui A, et al. Biological insights into multiple birth: genetic findings from UK Biobank. Eur J Hum Genet. 2019;27:970–9.

37. Teo YY. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. Curr Opin Lipidol. 2008;19:133–43.

38. Frischknecht M, Bapst B, Seefried FR, Signer-Hasler H, Garrick D, Stricker C, et al. Genome-wide association studies of fertility and calving traits in Brown Swiss cattle using imputed whole-genome sequences. BMC Genom. 2017;18:910.

39. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–-4423.

40. Bland JM, Altman DG. Statistics notes: Multiple significance tests: the Bonferroni method. BMJ. 1995;310:170.

41. Fernando R, Toosi A, Wolc A, Garrick D, Dekkers J. Application of Whole-Genome Prediction Methods for Genome-Wide Association Studies: A Bayesian Approach. J Agric Biol Environ Stat. 2017;22:172–93.

42. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annu Rev Anim Biosci. 2019;7:89–102.

43. Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet. 2015;16:275–84.

44. Holloway JW, Prescott SL. Chapter 2 - The Origins of Allergic Disease. In: Robyn E, O'Hehir S, Holgate, AS (eds). Middleton's Allergy Essentials. Elsevier; 2017. p. 29–50.

45. Ostersen T, Christensen OF, Henryon M, Nielsen B, Su G, Madsen P. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genet Sel Evol. 2011;43:38.

46. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.

47. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. 2nd edition. New York, NY, USA: Springer US; 2021.

48. Ban H-J, Heo JY, Oh K-S, Park K-J. Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. BMC Genet. 2010;11:26.

49. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc, B: Stat Methodol. 1996;58:267–88.

50. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.

51. Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinform. 2010;11:110.

52. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012;99:323–9.

53. Khan MY, Qayoom A, Nizami MS, Siddiqui MS, Wasi S, Raazi SMK-R. Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. Complexity. 2021;2021:1–18.

54. Widmer S, Seefried FR, von Rohr P, Häfliger IM, Spengeler M, Drögemüller C. A major QTL at the LHCGR/FSHR locus for multiple birth in Holstein cattle. Genet Sel Evol. 2021;53:57.

55. Widmer S, Seefried FR, von Rohr P, Häfliger IM, Spengeler M, Drögemüller C. Associated regions for multiple birth in Brown Swiss and Original Braunvieh cattle on chromosomes 15 and 11. Anim Genet. 2022;53:557–69.

56. Lett BM, Kirkpatrick BW. Identifying genetic variants and pathways influencing daughter averages for twinning in North American Holstein cattle and evaluating the potential for genomic selection. J Dairy Sci. 2022;105:5972–84.

57. Fog CK, Galli GG, Lund AH. PRDM proteins: Important players in differentiation and disease. BioEssays. 2012;34:50–60.

58. Roper LK, Briguglio JS, Evans CS, Jackson MB, Chapman ER. Sex-specific regulation of follicle-stimulating hormone secretion by synaptotagmin 9. Nat Commun. 2015;6:8645.

59. Zavareh S, Gholizadeh Z, Lashkarbolouki T. Evaluation of changes in the expression of Wnt/β-catenin target genes in mouse reproductive tissues during estrous cycle: An experimental study. Int J Reprod Biomed. 2018;16:69–76.

60. Giuffra E, Tuggle CK, FAANG Consortium. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annu Rev Anim Biosci. 2019;7:65–88.

61. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. Stat Methods Med Res. 2007;16:309–30.

62. Sheehan NA, Didelez V, Burton PR, Tobin MD. Mendelian Randomisation and Causal Inference in Observational Epidemiology. PLoS Med. 2008;5:1205–10.

63. Barberi M, di Paolo V, Latini S, Guglielmo MC, Cecconi S, Canipari R. Expression and functional activity of PACAP and its receptors on cumulus cells: Effects on oocyte maturation. Mol Cell Endocrinol. 2013;375:79–88.

64. Rabahi F, Brûlé S, Sirois J, Beckers J-F, Silversides DW, Lussier JG. High Expression of Bovine α Glutathione S-Transferase (GSTA1, GSTA2) Subunits Is Mainly Associated with Steroidogenically Active Cells and Regulated by Gonadotropins in Bovine Ovarian Follicles. Endocrinology. 1999;140:3507–17.

65. van den Hurk R, Zhao J. Formation of mammalian oocytes and their growth, differentiation and maturation within ovarian follicles. Theriogenology. 2005;63:1717–51.

66. Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. 2009;41:55.

67. Fitzgerald AM, Berry DP, Carthy T, Cromie AR, Ryan DP. Risk factors associated with multiple ovulation and twin birth rate in Irish dairy and beef cattle. J Anim Sci. 2014;92:966–73.

68. Häfliger IM, Seefried FR, Drögemüller C. Successful trio-based reverse genetic screen in an endangered local cattle breed. In: WCGALP 2022. 2022. https://www.wageningenacademic.com/pb-assets/wagen/WCGALP2022/49_004.pdf. Accessed 28 Oct 2022.

69. Lett BM, Kirkpatrick BW. Short communication: Heritability of twinning rate in Holstein cattle. J Dairy Sci. 2018;101:4307–11.

70. McGovern SP, Weigel DJ, Fessenden BC, Gonzalez-Peña D, Vukasinovic N, McNeel AK, et al. Genomic Prediction for Twin Pregnancies. Animals. 2021;11:843.

71. Arbet J, McGue M, Chatterjee S, Basu S. Resampling-based tests for Lasso in genome-wide association studies. BMC Genet. 2017;18:70.

72. Fortes MRS, Porto-Neto LR, Satake N, Nguyen LT, Freitas AC, Melo TP, et al. X chromosome variants are associated with male fertility traits in two bovine populations. Genet Sel Evol. 2020;52:46.

73. Pacheco HA, Rezende FM, Peñagaricano F. Gene mapping and genomic prediction of bull fertility using sex chromosome markers. J Dairy Sci. 2020;103:3304–11.

74. Arishima T, Sasaki S, Isobe T, Ikebata Y, Shimbara S, Ikeda S, et al. Maternal variant in the upstream of FOXP3 gene on the X chromosome is associated with recurrent infertility in Japanese Black cattle. BMC Genet. 2017;18:103.

75. Derom C, Jawaheer D, Chen W v., McBride KL, Xiao X, Amos C, et al. Genome-wide linkage scan for spontaneous DZ twinning. Eur J Hum Genet. 2006;14:117–22.

9 Declaration of Originality

Declaration of Originality

Last name, first name: Widmer, Sarah

Matriculation number: 14-937-551

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations.
All data, tables, figures and text citations which have been reproduced from any other source, including the internet, have been explicitly acknowledged as such.
I am aware that in case of non-compliance, the Senate is entitled to withdraw the doctorate degree awarded to me on the basis of the present thesis, in accordance with the “Statut der Universität Bern (Universitätsstatut; UniSt)”, Art. 69, of 7 June 2011.

Place, date
Heimiswil, 09.11.2022

Signature