Sequence-based GWAS in 180 000 German Holstein cattle reveals new candidate genes for milk production traits

Ana-Marija Križanac[1,2 *], Christian Reimer[2,3], Johannes Heise[4], Zengting Liu[4], Jennie Pryce[5,6], Jörn Bennewitz[7], Georg Thaller[8], Clemens Falker-Gieske[1,2 §], Jens Tetens[1,2 §]

[1]Department of Animal Sciences, University of Goettingen, Burckhardtweg 2, 37077 Göttingen, Germany

[2]Center for Integrated Breeding Research, Department of Animal Sciences, University of Goettingen, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

[3]Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, 31535 Neustadt, Germany

[4]Vereinigte Informationssysteme Tierhaltung w.V. (VIT), 27283 Verden, Germany

[5]Agriculture Victoria Research, AgriBio, Centre for AgriBioscience, Bundoora, Victoria 3083, Australia

[6]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3083, Australia

[7]Institute of Animal Science, University of Hohenheim, 70599 Stuttgart, Germany

[8]Institute of Animal Breeding and Husbandry, Christian-Albrechts-University, 24118 Kiel, Germany

§These authors jointly supervised this work

*Corresponding author

E-mail addresses:

AMK: ana-marija.krizanac@uni-goettingen.de, ORCID: 0000-0002-7446-8031

CR: Christian.Reimer@fli.de, ORCID: 0000-0002-9697-2511

JH: Johannes.Heise@vit.de, ORCID: 0000-0002-7605-7148

ZL: Zengting.Liu@vit.de, ORCID: 0000-0002-7112-0320

JP: jennie.pryce@agriculture.vic.gov.au, ORCID: 0000-0002-1397-1282

JB: j.bennewitz@uni-hohenheim.de, ORCID: 0000-0001-6450-1160

GT: gthaller@tierzucht.uni-kiel.de, ORCID: 0000-0002-6782-2039

CFG: clemens.falker-gieske@uni-goettingen.de, ORCID: 0000-0001-9160-1909

JT: jens.tetens@uni-goettingen.de, ORCID: 0000-0001-5352-464X

Abstract

Background

The use of genome-wide association studies (GWAS) has led to the identification of numerous quantitative trait loci and candidate genes in dairy cattle. To obtain sufficient power of GWAS and to identify quantitative trait nucleotides, whole-genome sequence data is required. Sequence data facilitates the identification of potential causal variants; however, sequencing of whole genomes is still expensive for a large number of animals. Imputation is a quick and efficient way of obtaining sequence data from large datasets. Milk production traits are complex and influenced by many

genetic and environmental factors. Although extensive research has been performed for these traits, with many associations unveiled thus far, due to their crucial economic importance, complex genetic architecture, and the fact that causative variants in cattle are still scarce, there is a need for a better understanding of their genetic background. In this study, we aimed to identify new candidate loci associated with milk production traits in German Holstein cattle, the most important dairy breed in Germany and worldwide. For that purpose, 252,285 cattle were imputed to the sequence level and large-scale GWAS was carried out to identify new association signals.

Results

We confirmed many known and identified 30 previously unreported candidate genes for milk, fat, and protein yield. While all of the genes were functionally associated with the traits, some showed pleiotropic effects as well. Specifically, association with mammary gland development, fatty acid synthesis, metabolism of lipids, or milk production QTLs in other farm animals has been reported. Variants associated with these genes explained a large percentage of genetic variance, compared to random ones.

Conclusions

Our findings proved the power of large samples and sequence-based GWAS in detecting new association signals. In order to fully exploit the power of GWAS, one should aim at very large samples combined with whole-genome sequence data. Although milk production traits in cattle are comprehensively researched, the genetic background of these traits is still not fully understood, with the potential for many new associations to be revealed, as shown in our study. With constantly growing sample sizes, we expect more insights into the genetic architecture of production traits in the future.

Background

Intensive selection for milk production traits enhanced with improved nutrition and management, as well as reproductive technologies and accelerated by genomic selection (reviewed by [1]), has strongly increased milk production over the years [2]. The Holstein breed is dominant in milk production worldwide. The German Holstein population alone comprises 2.4 million cows, with an average milk yield of 10,000 kg per lactation [3]. The breeding goal for German Holstein is balanced and includes many traits that can be grouped into milk production, health, fertility, and longevity [4]. This has not always been the case, and although selection for milk production has been successful in increasing milk yield, it has also been associated with a higher incidence of mastitis, metabolic and reproductive diseases [5]. The relative weight of milk production in total merit indices is decreasing as new traits are continuously added to the breeding goal. However, because production still makes up a substantial part (e.g. 36% in Germany), there is the risk of a further decline in animal health. More extensive knowledge of the genetic architecture of economic traits is needed, especially given that the majority of these traits are complex traits, influenced by many genes and environmental factors.

So far, genome-wide association studies have been successful in the discovery of quantitative trait loci (QTL) and candidate genes (reviewed by [6]), however, only a few causal variants for economically important traits in cattle have been confirmed [7, 8]. In order to be able to detect the underlying causal variant whole-genome sequence (WGS) data and large samples are needed to ensure sufficient power of GWAS [9, 10]. GWAS in cattle is restricted by long-distance linkage disequilibrium (LD) segments [11], due to a small effective population size (N_e) caused by intense selection [12], therefore making it hard to pinpoint the true causal variant which may be hidden

among the many variants in LD. Another source of difficulty in revealing the true associations is the highly polygenic genetic architecture of quantitative traits, i.e. large number of variants with small effects affecting the trait [13]. Genotypes from whole-genome sequences obtained from sequencing the study individuals are limited, especially when large samples are considered. In that case, imputation [14] can be utilized as a method of obtaining the sequence-dense data. Imputation methods exploit LD patterns among the individuals in the sample and reference dataset, with the assumption that apparently unrelated animals inherited haplotype blocks from a common ancestor [15]. Imputation accuracy depends on various factors such as the size of the reference panel, the relationship between the individuals in the reference and sample dataset, imputation software choice and the number of the variants to be imputed [16–18]. In cattle, sequence-level imputation is usually performed in two steps [18], due to higher accuracy obtained when first imputing from a lower to a higher-density SNP chip, and then to sequence level.

To exploit the power of large sample size in detecting novel causative loci, we carried out GWAS for three milk production traits using imputed sequence data. After obtaining GWAS summary statistics with a mixed linear model approach, meta-analysis was utilized to pool the results of different animal groups. Candidate gene search was performed for top variants from GWAS with the lowest *p*-values and functional enrichment analysis was done to confirm the candidate genes. Finally, the percentage of genetic variance explained by the top SNPs was calculated to see which proportion of the variance could be attributed to variants associated with the novel candidate loci.

Methods

Dataset

The dataset for imputation consisted of 252,285 German Holstein cows with 45,613 SNP markers. Animals were mainly genotyped with various low-density SNP genotyping arrays (see Additional file 1: Table S1) and then imputed to 50K level according to the national genetic evaluation procedure [19], or genotyped with various 50K SNP chips (see Additional file 1: Table S1). The dataset was collected during the KuhVision project that aimed to genotype and phenotype German Holstein cows to establish a large-scale female reference population for genomic evaluation. The phenotypes for milk (MY), fat (FY), and protein yield (PY) in kg were obtained in the form of deregressed proofs (DRPs), which are pseudo-phenotypes produced using the special single-step SNP BLUP model for deregressing genomic estimated breeding values (GEBV) [20].

Imputation

The genomic coordinates of the input genotypes were lifted from the previous bovine reference genome assembly UMD 3.1. to the ARS-UCD1.2 with a custom approach. CombineVariants from the Genome Analysis Toolkit (GATK) v. 3.8.1.0 [21] was used to merge the samples by chromosomes and by groups. The sample of 252,285 cows consisting of 30 autosomal and sex chromosome pairs was imputed to sequence level in a two-step imputation approach using the BEAGLE v. 5.2 [22]. The effective population size parameter was set to 1000. The animals were first imputed to high-density (HD) genotype level using the genotype data of 1278 Holstein cows consisting of 585,517 markers [23]. The HD reference panel was phased using BEAGLE v. 5.1 beforehand [24]. In the next step, data were imputed to the WGS level using the multi-breed reference panel from the 1000 Bulls Genome Project Run9 [25]. The reference panel consisted of 5116 cows and bulls of the species *Bos taurus* (see Additional file 1: Table S2). Both imputation steps were performed chromosome-wise, with the samples divided into groups of approximately equal size, due to high computational demand. The imputed files were indexed afterwards with

IndexFeatureFile, GATK v. 4.2.2.0, merged by the sample groups, and split into separate lines due to the presence of multi-allelic variants (SNPs, insertions, and deletions) using BCFtools v. 1.14 [26]. As a quality control, the imputed WGS dataset was filtered using the dosage R-squared parameter, a measure of the estimated squared correlation between estimated and true allele dosage (DR2; [27]). Markers imputed with DR2 < 0.75 were removed with BCFtools. The imputed WGS dataset was annotated with VariantAnnotator from the GATK v. 4.2.2.0 using the Ensembl variation database, release 105 [28] imported from dbSNP [29], to account for SNPs without rsID.

GWAS

Since phenotype measurements were not available for all 252,285 animals, the sample for GWAS consisted of 180,217 WGS-imputed cows with phenotypic observations for MY, FY, and PY. Samples were filtered for minor allele frequency (MAF) > 0.01. Due to memory restrictions of the high-performance computing (HPC) cluster, the samples were divided into 4 groups consisting of ~ 45,000 animals each. GWAS was performed using the GCTA software v. 1.93.2 beta [30] applying a mixed linear model approach (MLMA) for all autosomes. The SNP effects were estimated using the following model:

$$y = Xb + Zu + e$$

where y is a vector of DRPs; b is the fixed effect of the variant tested for the association with each trait; X is a vector containing the genotype score for the tested SNP; u is the vector of polygenic effects with $u \sim N(0, G\sigma^2u)$, where G is genomic relationship matrix (GRM) calculated using 50K SNP genotypes from all chromosomes, and σ^2u is a variance of polygenic effects; Z is the incidence matrix of u; and e is the vector of residual effects with $e \sim N(0, I\sigma^2e)$, with I being an identity matrix and σ^2e residual variance. Bonferroni correction was used to set a genome-wide

significance threshold, corresponding to a *p*-value of 0.05/number of markers. The Manhattan plots were created using RStudio v. 3.6.3 [31]. METAL software [32] for meta-analysis was used to merge the GWAS summary statistics of each of the four animal groups per trait, using the approach that takes into account test statistics and standard errors. To correct for genomic inflation, lambda (λ_{GC}) values were calculated as the median of observed χ^2 test statistics divided by the expected median of χ^2 distribution with one degree of freedom.

Downstream analyses

SnpEff [33] and SnpSift [34] were utilized for functional annotation of genome-wide significant variants and prediction of their effect on genes, as well as the identification of the closest genes. The R packages cluster profiler [35] and DOSE [36] were used to carry out an over-representation analysis (ORA) [37] to determine whether the genes positioned closest to the genome-wide significant variants are enriched in the known biological pathways. ORA was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) [38] database for variants that passed the significance threshold of 0.001/number of markers with enrichKEGG (*q*-value > 0.25). Candidate genes were also investigated manually, through the Animal Quantitative Trait Loci database (Animal QTLdb) [39] and using the publications previously associated with milk production candidate genes, based on the STRING database [40]. A Venn diagram of common candidate genes was created using the R package VennDiagram [41]. The percentage of genetic variance explained by the top 50 genome-wide significant SNPs and 50 random SNPs across all chromosomes was calculated using GCTA's genomic-relatedness-based restricted maximum-likelihood (GREML) approach [42], by fitting the GRMs together in the model with 50K SNP chip variants. The analysis was done for the smaller subset of 45,000 animals due to high

171 computational demand. PLINK v. 1.9 [43] was used to prune the variants in high linkage
172 disequilibrium, based on pairwise R^2 correlation greater than 0.5 (--indep-pairwise 50 5 0.5).

173 Results

174 Imputation

175 To evaluate the genotype liftover quality, we examined the allele frequency (AF) concordance
176 between the imputed WGS dataset and Run9, by plotting the AF from BTA16 of the two datasets
177 against each other. Allele frequencies of imputed variants were congruent with the ones from 1000
178 Bulls Run9, showing the coherence in the frequency for the majority of loci (Figure S1).
179 Imputation quality control was carried out by utilizing the DR2 parameter, built into the BEAGLE
180 software. Markers imputed with DR2 < 0.75 were removed with BCFtools, leaving the 20,737,793
181 markers for further analyses. Then, we checked the DR2 values of known causal variants, such as
182 two variants in the *DGAT1* gene, which were imputed with almost perfect quality (DR2=0.99), as
183 well as rs385640152 in the *GHR* gene with DR2=0.98, and rs211210569 in *MGST1* with DR2=1.
184 After the imputation of 252,285 animals to sequence level, and filtering for DR2 and MAF,
185 17,256,703 variants were left for GWAS.

186 GWAS

187 A large number of variants exceeded the genome-wide significance threshold. GWAS analyses
188 identified 54,032 significant variants for MY, 42,323 for FY, and 35,106 for PY, with the highest
189 number of associations on chromosomes 5, 6, and 14 (Figures 1-3). Low *p*-values were observed
190 for many SNPs, with top variants positioned on bovine chromosome (BTA) 14: rs109050667 (*p* =

191 $7.04x10^{-737}$), rs136630297 ($p = 7.18x10^{-380}$), and rs109050667 ($p = 2.38x10^{-221}$) for MY, FY and
192 PY, respectively.

193 The top 50 variants from each chromosome were chosen for further research (see Additional file
194 2: Tables S1-S3). Considering that significant associations have not been identified on every
195 chromosome and that some chromosomes had less than 50 significant variants, the number of top
196 variants chosen for further investigation differed across chromosomes and traits. For MY, 1012
197 top SNPs were found within or in proximity of 109 genes from 25 chromosomes (see Additional
198 file 1: Table S3). The top candidate genes on chromosomes with the highest number of significant
199 SNPs were *MGST1* and *SLC15A5* on BTA5, *GC*, *NPFFR2*, *ENSBTAG00000049290* and *SLC4A4*
200 on BTA6, *ADCK5*, *CPSF1*, *SLC52A2*, *SLC39A4*, *FBXL6*, *TMEM249* and *SCRT1* on BTA14.

201 For FY, the top 962 SNPs from 24 chromosomes were located within or close to 108 genes (see
202 Additional file 1: Table S3). The top candidate genes were *MGST1* and *SLC15A5* on BTA5, *GC*,
203 *NPFFR2*, *ENSBTAG00000049290* on BTA6, *CPSF1*, *SLC39A4*, *ADCK5*, *TMEM249*, *SCRT1*,
204 *SLC52A2*, *FBXL6* and *ENSBTAG00000053637* on BTA14.

205 For PY, 1065 top SNPs from 26 chromosomes were located close to or in 172 genes (see
206 Additional file 1: Table S3). The candidate genes associated with the most significant genomic
207 regions were: *GC*, *NPFFR2*, *ENSBTAG00000049290*, and *SLC4A4* on BTA6, *ABCC9*, *ST8SIA1*,
208 *ENSBTAG00000026611* and *CMAS* on BTA5. Many genes were found to be associated with the
209 same traits, as shown on the Venn diagram (Figure 4). The highest number of common candidate
210 genes were found between MY and PY (47). The second highest number of candidate genes was
211 between FY and PY (27), 7 genes were in common for MY and FY, and 23 genes were in common
212 for all three traits (see Additional file 1: Table S4). Lambda values, calculated to assess for false
213 associations were as follows: $\lambda_{MY} = 1.764$, $\lambda_{FY} = 1.898$, and $\lambda_{PY} = 1.928$. The reason for increased

genomic inflation factors was due to the meta-analysis that inflated the *p*-values and therefore the number of genome-wide significant variants. To assess the effect of meta-analysis on inflation we divided the individuals from direct-GWAS summary statistics into smaller groups, running the GWAS for each of these groups again, and merging them into a meta-analysis. The lambda values were higher after merging the animals into meta-analysis compared to direct GWAS summary statistics for the same individuals (Figure 5).

Downstream analyses

SnpEff was used to predict the functional effects of genome-wide significantly associated variants on proteins and to identify the closest genes. The majority of variants were identified in introns (46.41%) or intergenic regions (37.46%). The number of predicted effects was larger than the actual number of variants, due to genes with multiple transcripts and variants which affect multiple genes. A detailed description of the variant effects by type is available in an additional file (see Additional file 1: Table S5). Regarding the variant impact on proteins, a high majority of variants were classified as modifiers (98.38%), and only 0.025% were high-impact variants. Of the 50 top variants which were further investigated, the same high-impact, frameshift variant was found for both PY and MY on BTA16, at 80,129,589 bp, in the *SYT2* gene. One frameshift variant was also found for FY on BTA3, at 7,933,141 bp in the *FCGR2B* gene.

KEGG functional enrichment analysis revealed a large number of over-represented terms. To narrow the list of possible terms, ORA was performed only for genes associated with variants that passed the genome-wide significance threshold of 0.001/number of markers. A list of all over-represented genes and associated pathways is available in an additional file (see Additional file 1: Table S6). The common dot plot of the 20 most significant terms of KEGG analysis for MY, FY,

and PY is shown in Figure 6. The top variants were found in or in proximity to the genes over-represented in 23 pathways, mostly in the PI3K-Akt signaling pathway (Table 1). Two terms were significantly enriched with all three traits.

The percentage of genetic variance explained by 50 top variants, as well as by 50 random variants from all autosomal chromosomes was estimated for all three traits (Table 2). For MY, 1012 top variants from 25 chromosomes explained 8.67% of the variance. Random SNPs from 29 autosomal chromosomes, explained on average 0.78% of the variance. As for FY, 962 top SNPs from 24 chromosomes accounted for 7.04% of the genetic variance, while the random 1450 variants from all chromosomes explained about 0.31%. For PY, 6.66% of the variance was explained by 1065 top SNPs from 26 chromosomes, and 0.37% was due to random 1450 variants. After LD pruning of the top variants for each trait, there were 124, 147, and 182 variants left for MY, FY, and PY, respectively. Genetic variance explained by pruned variants was 10.01% for MY, 6.51% for FY, and 5.17% for PY.

Discussion

Imputation

In this study, we performed a stepwise imputation of 252,285 German Holstein cows from SNP chip up to sequence level, which makes our sample size one of the largest imputed in cattle so far. The stepwise imputation approach seems to improve the imputation accuracy, as previously shown in cattle [18, 44]. Imputation error rate tends to decrease when an intermediate reference panel is used [44], possibly due to a larger choice of possible haplotype matches between WGS and medium-density SNP chip, which are narrowed down when using an HD panel as an intermediate

[18]. In our study, stepwise imputation was done using the Holstein breed HD panel, a subset from van den Berg et al. [23] as an intermediate step, and the WGS panel from 1000 Bulls Genome Consortium, as a second step. The WGS-based panel consisted of various breeds of taurine cattle (see Additional file 1: Table S2). The usage of a multi-breed reference was shown to increase the imputation accuracy in many studies in cows [45–47], especially for low-frequency variants [46]. However, multi-breed panels can be counter-productive if animals in the reference panel are too distant from the sample dataset [48]. The usage of BEAGLE software for imputation can at least partly overcome this issue since its algorithms can prioritize between the closer and the genetically distant individuals in the multi-breed reference panel [49]. Moreover, the 1000 Bulls reference panel consisted of a large number of Holstein animals (~1200) making them the most represented breed in the reference panel (see Additional file 1: Table S2), therefore enabling the reliable imputation of Holsteins even in the presence of genetically distant breeds. Another crucial factor to consider is the value of the N_e parameter [49]. Default Ne in BEAGLE is 1,000,000, however, this corresponds to the human populations for which was it initially developed. Therefore, updating the Ne parameter to smaller values is needed, when working with other, less-diverse populations [49].

To evaluate the accuracy of imputation we used the second category of quality measures [50] based on estimated genotypes (DR2) since SNP array genotyped animals were not whole genome sequenced. Filtering the variants with DR2 < 0.8 is recommended when using DR2, as the imputation error rate increases below this threshold [49] hence we filtered out all the variants with DR2 < 0.75. Known causal variants were left in the dataset after DR2 filtering, and were imputed with near to perfect quality (DR2=0.98 to 1). *DGAT1* causal variants were among the 100 top genome-wide significant variants for all three traits analyzed but were not the top variants. A

possible explanation for this could be the presence of additional variation in the form of a known variable number of tandem repeats (VNTR) in the *DGAT1* region or low imputation accuracy [47, 51, 52]. To assess the liftover quality, AF concordance between the imputed WGS dataset and the Run9 reference panel was examined on the example of chromosome 16, showing the congruence for the majority of variants between the two datasets (see Additional file 1: Figure S1).

GWAS and candidate gene research

After carrying out the GWAS, possible candidate genes were retrieved by searching public databases such as Animal QTLdb and reviewing journal papers on previously reported candidate genes and QTLs. We confirmed many of the previously reported QTLs and candidate genes (see Additional file 1: Table S7 and Additional file 3: Tables S1-S3) such as *DGAT1* and its variants rs109326954 and rs109234250 on BTA14, *MGST1* on BTA5, *ABCG2* on BTA6, *GC* on BTA6 and *GHR* on BTA20, but also discovered new, previously unreported loci (Table 3). There were a large number of genes whose functions have not been reported yet, as well as the ones whose functions could not be associated with milk production or content (see Additional File 4: Table S1). Therefore, these genes were not considered as potential candidate genes. For simplification, we discussed only candidate genes associated with the most significant SNPs, while the list of all associations can be found in Additional file 3: Tables S1-S3. The majority of the most significant variants were intronic (37%) and intergenic (30%) (see Additional file 1: Table S5). Most of the significant variants that were included in candidate gene research were non-coding as well, which is in line with the majority of other GWAS publications [53–55]. Nayeri et al. [55] showed that a large proportion of the most significant variants affecting milk yield and composition traits in Holstein and Jersey cattle were located in non-coding regions of the genome. Both intron and

intergenic variants usually do not code for proteins, making their functional prediction challenging [56]. However, recent research in human studies (reviewed by [53]) and cattle [57] has shown that even the variants in non-coding regions may play an important part in complex traits and diseases, by indirect involvement in gene expression regulation. Known QTNs in livestock are not all coding variants that cause a change in amino acid [6, 58], therefore, variants in non-coding regions can be causal as well [57]. Xiang et al. [59] showed that non-coding variants can contribute substantially to variance in complex traits in cattle.

Due to the large sample sizes in our study, which might contribute to the rise in genomic inflation [60], lambda values were measured before and after performing the meta-analysis. Genomic inflation is a spurious association between a variant and trait, where the relationship between a phenotype and SNP seems to arise from different factors than the true association [61]. These factors include population stratification [62], cryptic relatedness [63], polygenic inheritance [61], or strong association between variant and phenotype [64]. Although some of the genomic inflation in our study might be attributed to the polygenicity of milk production traits [65], and population structure in German Holstein [66], the main source of genomic inflation was the use of meta-analysis software (Figure 5). Similar findings were reported in human studies [67], where large number of individuals are often pooled into the meta-analysis. The use of meta-analysis was inevitable in our case, due to the large samples that our HPC cluster was not able to utilize. MLMA accounted properly for genomic inflation, as the direct GWAS summary statistic had lambda values below 1 (Figure 5), and values up to 1 are usually considered as a threshold for genomic inflation. To prove that inflation was not due to population structure amplification that might arise when pooling the samples into the meta-analysis [68], we divided one of the animal groups on which we obtained summary statistics. After the animals were divided into two groups, GWAS

was run for each of them again. Then, after obtaining the summary statistics, two groups of samples are merged into the meta-analysis. As shown in Figure 5, lambda values for the same samples were increased after pooling into a meta-analysis. Moreover, an increase in the number of animal groups pooled into meta-analysis led to higher genomic inflation (Figure 5). Considering that many factors that could lead to an increase in lambda values were present in our study, including the polygenic nature of the milk production traits, large sample sizes, potential underestimated relationships between animals, and in the end, the use of meta-analysis, we consider the values we obtained on meta-analyzed traits (1.764-1.928) acceptable even though they exceeded the generally accepted threshold of 1.

Candidate genes for milk yield

The novel candidates that appeared to be the most relevant for further validation experiments due to their role in organism will be discussed here, while the list of all novel candidate genes and their roles connected with milk production traits are listed in Table 3. Except for the functional involvement of the candidate genes with milk production traits, and the fact that some of them are reported in other mammal species for the same or similar traits, variants found in candidate genes need experimental validation to be considered causative. For this purpose, prioritization of genome-wide significant variants according to external evolutionary and functional information [59] is suggested as the next step, followed by sequencing and gene editing experiments.

As for the new associations, we identified 9 genes that, to our knowledge, were not previously described in cattle for milk yield or related traits. On chromosome 2, we identified two intergenic variants whose positions fall between the *FEV* and *CDK5R2* genes. While *FEV* was reported earlier [69], *CDK5R2* has not been associated with milk traits in cattle yet. *CDK5R2* (Cyclin

Dependent Kinase 5 Regulatory Subunit 2 (p39)) acts as one the activators of the *CDK5* gene [70] that has numerous important roles in the nervous system [71]. Talouarn et al. [72] identified the variants in the region of this gene to be associated with milk yield in French dairy goats. *CDK5R2* was previously associated with somatic cell count (SCC) in dairy cattle [73], and meat color traits in Nellore cattle [74], and crossbred and purebred pigs [75]. The variant in this gene has been associated with teat length in Chinese Holstein, in the study of Wu et al. [76]. Given the previous association with milk yield in the goat population, as well as with udder-related traits in cattle, this gene presents a strong candidate for further research.

A downstream variant of the *PRDM1* gene on BTA9 at 43,842,866 bp, was significantly associated with MY. *PRDM1* (PR Domain Zinc Finger Protein 1), or *BLIMP1* (B-Lymphocyte-Induced Maturation Protein 1) was described as an essential factor for mammary development in mouse studies [77]. Ahmed et al. [77] discovered that a group of luminal alveolar progenitor cells, expressing *BLIMP1*, were essential for mammary gland development. *BLIMP1* is required for ductal morphogenesis in puberty and alveolar maturation in pregnancy and lactation, with its inactivation causing inadequate milk secretion [77]. In another study [78] *BLIMP1* was described as necessary for the delay of the intestinal epithelium maturation from suckling to adult-type intestinal epithelium, with mutant mice showing growth disturbances and increased mortality.

One intron variant in the *FBXL19* gene was associated with MY. *FBXL19* (F-Box And Leucine Rich Repeat Protein 19) regulates cell migration and proliferation [79, 80] and acts as a major regulator of adipogenesis [81]. Adipose tissue is a source of energy for various organs and tissues, as well as for the mammary gland especially during lactation when it serves as a source for fatty acids synthesis [82]. Adipogenesis is essential for the efficiency of milk production in dairy cows, as well as reproduction [83] making this gene an interesting candidate for milk production traits.

Candidate genes for fat yield

Eight novel candidate genes were associated with FY in our study. The majority were involved with various lipid metabolism functions, therefore, we will describe only a few in the main text, while the description of the rest of the genes and their functions is available in Table 3.

One intergenic region between the *STK25* and *BOK* was significantly associated with fat yield on BTA3. While *BOK*, a member of the family of BCL-2 proteins which are involved in many cellular processes [84], couldn't be linked to milk production traits, *STK25* attracted our attention as a candidate for fat yield and composition. *STK25* (serine/threonine protein kinase 25) belongs to the germinal center kinase III subfamily of Ste20 (sterile 20) proteins that exhibit various cell functions (reviewed in [85]). *STK25* was shown to regulate lipid catabolism in liver cells in humans and the release of non-esterified fatty acids (NEFA) from lipid droplets [86]. High levels of NEFA seem to stimulate the expression of the *CIDEA* gene, and consequently increase fatty acid synthesis *de novo* and milk fat secretion [87]. *CIDEA* was recently described as a regulator of *de novo* fatty acid synthesis in cattle as well [88]. In general, high levels of NEFA are associated with ketosis and fatty liver, poor reproductive performance, and negative energy balance in early lactation (reviewed by [89]). Another study indicated a plausible role of this gene in the regulation of lipid and glucose metabolism in the skeletal muscle of rodents and humans [90]. Altogether, *STK25* seems to have an important role in lipid metabolism and therefore is recommended for further investigation.

On BTA16, one intron variant was found in *KLHL12*, a gene described as essential for the secretion of apolipoprotein B100 (apoB100) very low-density lipoprotein (VLDL) particles in rat hepatoma cell line [91]. ApoB100, a major component of VLDL, is essential for the transport of triglycerides,

the main component of milk fat, from the liver to peripheral tissues [92, 93]. Decreased levels of apoB100 in cattle have been reported in cows with metabolic disorders such as ketosis, milk fever, and displaced abomasum [94, 95]. *KLHL12* (Kelch-like Family Member 12) is a member of the Kelch-like family (KLHL) of proteins with important functions in the ubiquitination of proteins as reviewed by Shi et al. [96]. When it comes to known roles of the *KLHL12*, it has been reported as a negative regulator of the Wnt signaling pathway [97], important for various cell functions in both adult and embryonal tissue homeostasis (reviewed in [98]). It also seems to have a key role in collagen secretion [99]. Everything considered, this gene could potentially affect not only milk production traits due to its role in triglyceride transport but health traits as well, and because of this further validation is needed.

Candidate genes for protein yield

For protein yield, many known candidate genes, as well as the pleiotropic effects of some genes were confirmed (see Additional file 1: Table S7), while 18 genes from 12 chromosomes are reported here, for the first time (Table 3). As for the genes with pleiotropic effects, three intergenic variants were positioned between the *FEV* and *CDK5R2* on BTA2, a gene that we found to be a novel candidate for MY in the previous paragraph. The five variants on BTA3 were located in or downstream of the *STK25* gene, showing the effects of this new candidate gene on both fat and protein yield. Then, on BTA9, 21 variants were located in or in proximity to *PRDM1*, identified in both MY and PY GWAS. The majority of *PRDM1*-associated variants were intergenic, however, variant rs136669229 ($p = 1.009 \times 10^{-10}$) was identified as missense, causing the Valine to Phenylalanine amino acid change, however, there was no difference in protein structure prediction. On BTA16, one intron variant was located within the *KLHL12* gene, whose function is described

in detail in the FY section. Finally, we identified an intron variant in the *FBXL19* gene, our candidate gene for MY.

In the proximity of *STT3A* on BTA9, one variant was significantly associated with PY. *STT3A* (STT3 Oligosaccharyltransferase Complex Catalytic Subunit A) encodes the protein which is a part of the central enzyme complex in glycosylation [84]. Glycosylation is a post-translational protein modification that takes place in the endoplasmic reticulum (ER) and is essential for numerous cellular functions [100]. The two main types of glycosylation are N and O-glycosylation. N-glycosylation consists of the attachment of an oligosaccharide N-Acetylglucosamine to Asparagine residues and occurs in both eukaryotes and prokaryotes [101, 102]. The most important step in N-glycosylation is catalyzed by the oligosaccharyltransferase (OST) complex, consisting of different subunits of which the STT3 subunit is the most important [103]. In the study of human milk lactoferrin glycosylation [104], the expression of *STT3* in milk somatic cells decreased from day 4 to day 15 of lactation, leading to changes in the overall level of glycosylation [104]. Lactoferrin is a milk-derived glycoprotein with many important roles in organism including immunomodulatory and anti-inflammatory, anticancer, and antimicrobial functions [105]. Therefore, the *STT3A* gene might affect the protein yield, and possibly play a role in mastitis given the antibacterial function [105] of lactoferrin.

An intron variant was found within the *RB1* (retinoblastoma 1) gene on BTA12. *RB1* is known for its role in regulating the metabolism of glycolipids in the liver, muscle, and adipose tissues and improving fat and protein metabolism disorders [106]. A study on *RB1* knockout-mouse showed the potential involvement of *RB1* in the gut microbiota and intestinal free fatty acids profiles [107], altogether making this gene a strong candidate for further research.

Downstream analyses

In the KEGG enrichment analysis, a large number of trait-associated genes were found within various pathways and biological processes. However, we restricted our analysis only to genes containing one of the top SNPs (Table 3). The highest number of genes (7) were involved in the PI3K-Akt signaling pathway, one of the most important signaling pathways that affect many biological functions, including cell metabolism, growth, migration, proliferation, and survival [108, 109]. Hou et al. [110] showed that *EEF1D* regulates milk lipid secretion and mammary gland development through interaction with various pathways, including PI3K-Akt. Genes involved in this pathway (Table 1) were all previously reported as candidates for milk production and composition traits (see Additional file 1: Table S7). Other pathways and terms involved with milk composition, synthesis, and secretion or mammary development processes included biosynthesis of amino acids, biosynthesis of cofactors, prolactin signaling pathway [111], ErbB signaling pathway [112], Hedgehog signaling pathway [113], fatty acid metabolism, Hippo signaling pathway [114] and ECM-receptor interaction [115]. The term "biosynthesis of amino acids" included the gene *PKLR*, a known candidate gene for milk yield and composition traits [116], with a role in the regulation of triglyceride levels and fatty acid synthesis [117]. KEGG category "biosynthesis of cofactors" contained three genes, across the three traits. Of these genes, *FLAD1* was associated with MY in our study and was previously reported as a candidate gene for milk calcium content and lactose percentage [118, 119]. Expression of gene *GMPPA*, whose variants were associated with FY, was positively correlated with bovine milk fat globule size in the study of Huang et al. [120]. This pathway was also enriched with *VKORC1L1*, a gene involved in vitamin K metabolism [121], across two traits. Although this gene couldn't be linked to milk production, it was described as a candidate gene for subclinical ketosis in Holstein in the study of Soares et al.

[122]. Interestingly, it has an important role in adipogenesis, with *VKORC1L1* mutated mice having smaller length and weight than wild type [123], making the possible role in milk fat metabolism plausible. Prolactin signaling pathway was enriched with genes *TH*, *STAT5A*, *STAT5B*, and *STAT3* for MY. Prolactin (*PRL*) is a gene well-known for its role in mammary development and lactation in many mammal species, as well as in cattle [111, 124]. *STAT5A*, *STAT5B*, and *STAT3* genes belong to the STAT family of transcription factors that participate in the PRL receptor (*PRLR*) signaling pathway [111] and were previously associated with milk composition traits in GWAS in Holstein cattle [125]. *STAT5A* and *STAT5B* were also enriched in the ErbB signaling pathway for MY. The members of the ErbB family of tyrosine kinase receptors regulate mammary gland development and have an important role in lactation [126]. The Hedgehog signaling pathway is required for normal development of various mammalian organs. Although the research results on its role in mammary gland development have been inconsistent, the latest insights showed that it has an important role in mammary ductal morphogenesis [113]. Gene found in this category for FY included *PTCH1*, a known regulator of mammary ductal morphogenesis [127] that has never been associated with milk production traits in cattle thus far. *SCD*, a gene reported to participate in fatty acid synthesis in Italian Holstein and Simmental GWAS [128] was enriched in the fatty acid metabolism pathway as well for FY, which is in line with the aforementioned findings. Another gene enriched in the term "fatty acid metabolism" for FY was *HSD17B12*, previously reported as a candidate gene for fat yield [125]. The hippo signaling pathway regulates various biological processes in the organism, including potential role in pregnant and lactating mammary gland [114]. This pathway was enriched for the *NKD2*, gene described as a candidate gene for MY, FY, and PY in German Black Pied cattle [129]. Extracellular matrix (ECM) components are involved in mammary gland development processes as reviewed

by Xu et al. [130]. Genes involved in this pathway, *THBS3* and *LAMA5*, were associated with MY. Both were previously reported as milk production and milk composition candidates [55, 131] and were found to be significantly enriched in the PI3K-Akt signaling pathway, as well. Many genes showed pleiotropic effects by involvement with the same terms across the three traits, which is in agreement with our candidate gene analysis, and previous research of other authors, as cited above.

The percentage of trait variance explained by the 50 most significant and 50 random variants from each chromosome, or so-called SNP-based heritability [132] was calculated to see how much of the genetic variance is attributable to top SNPs chosen for candidate gene research. To avoid overestimation of variance previously reported when using GREML [133], a GRM set up from 50K SNP chip data was included in the model, to account for further polygenic variance. Top SNPs explained much more variance than random ones (Table 2) indicating the potential presence of causal variants among those and underpinning the infinitesimal model. To eliminate multiple variants in high LD to each other that represent the same QTL, we pruned out the SNPs taking into account correlations between genotype allele counts [43]. Surprisingly, results differed depending on the trait; for MY, variants that were left after pruning explained more variance than the initial set of top SNPs. We expected that because the pruned variants spread over more QTL and should thus capture more variance. For FY and PY, however, pruned variants explained less variance than top SNPs. This could potentially be related to allelic heterogeneity in *DGAT1* because it can be assumed that the multiple variants capture more segregating variants [52].

Conclusions

After performing large-scale GWAS we identified 30 new candidate genes for three milk production traits; MY (9), FY (8), and PY (18), of which 6 genes (*CDK5R2*, *STK25*, *PRDM1*,

KLHL12, *RNF152*, and *FBXL19*) showed pleiotropic effects. These novel, functionally plausible candidates have not been reported for these traits so far. Variants located within or close to these genes explained a comparatively large proportion of genetic variance. In order to be able to fully exploit the power of GWAS, sequence data of very large samples are required, as shown in our study. Our findings add to existing knowledge of milk production traits architecture and convincingly demonstrate the power of our data set and strategy. Future studies incorporating health traits and their relationship with milk production may leverage the power of this data to add to the improvement of animal welfare.

List of abbreviations

AF – allele frequency

apoB100 – apolipoprotein B100

BTA – *Bos taurus* autosome

CNVR – copy number variation

DR2 – dosage R-squared

DRPs – deregeressed proofs

FY – fat yield

GEBV – genomic estimated breeding values

GREML – genomic-relatedness-based restricted maximum-likelihood

GRM – genomic relationship matrix

GWAS – genome-wide association study

HD – high density

HPC – high-performance computing

KEGG – Kyoto Encyclopedia of Genes and Genomes

LD – linkage disequilibrium

MAF – minor allele frequency

MLMA – mixed linear model approach

MY – milk yield

N_e – effective population size

NEFA – non-esterified fatty acids

ORA – over-representation analysis

PY – protein yield

SCC – somatic cell count

SCS – somatic cell score

VLDL – very low-density lipoprotein

WGS – whole-genome sequence

Declarations

Ethics approval and consent to participate

Not applicable. No live animals or animal material have been used in this study.

Consent for publication

Not applicable.

Availability of data and materials

The SNP chip genotype data and deregressed proofs are not available because they are the property of the national computing center in Germany (Vereinigte Informationssysteme Tierhaltung w.V.). Imputed genotypes and summary statistics will be provided upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is part of the project "QTCC: From Quantitative Trait Correlation to Causation in dairy cattle" and is funded by the Deutsche Forschungsgemeinschaft (DFG) (project number 448536632).

Authors' contributions

AMK performed the imputation, GWAS, and downstream analyses and wrote the paper. CR performed the genotype liftover and participated in genomic inflation analyses. JH provided the 50K SNP chip dataset, JP provided the HD reference dataset, and ZL provided the DRPs and gave useful comments. CFG participated in imputation and downstream analyses. CFG and JT supervised the study and participated in the writing of the paper. JT, JB, and GT conceived and supervised the project. All authors have read and approved the final manuscript.

Acknowledgements

The authors want to thank Iona MacLeod from Agriculture Victoria Research, AgriBio, Centre for AgriBioscience, 5 Ring Road, LaTrobe University, Bundoora, Australia, and Donagh Berry from Teagasc, Animal & Grassland Research and Innovation Centre, Moorepark, Fermoy P61 P302, Co. Cork, Ireland for giving the approval for the use of the HD dataset. We also want to acknowledge the 1000 Bulls Genome Consortium for providing the Run9 WGS dataset.

References

1. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29. doi:10.1093/genetics/157.4.1819.
2. García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, van Tassell CP. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci U S A. 2016;113:E3995-4004. doi:10.1073/pnas.1519061113.
3. German Livestock Association (BRS). BRS Broschüre Deutsche Holsteins Englisch Re-Design R1 01. 2021. https://www.rind-schwein.de/services/files/brs/export/broschueren/P-

2021-7-7-1%20BRS%20Brosch%C3%BCre%20Deutsche%20Holsteins%20Englisch%20Re-Design%20R1%20Einzelseiten.pdf. Accessed 18 Sep 2023.

4. Vereinigte Informationssysteme Tierhaltung w.V. (VIT). Estimation of Breeding Values for Milk Production Traits, Somatic Cell Score, Conformation, Productive Life and Reproduction Traits in German Dairy Cattle. 2023. https://www.vit.de/fileadmin/DE/Zuchtwertschaetzung/Zws_Bes_eng.pdf. Accessed 5 Dec 2023.
5. Fleischer P, Metzner M, Beyerbach M, Hoedemaker M, Klee W. The relationship between milk yield and the incidence of some diseases in dairy cows. J Dairy Sci. 2001;84:2025–35. doi:10.3168/jds.S0022-0302(01)74646-2.
6. Johnsson M, Jungnickel MK. Evidence for and localization of proposed causative variants in cattle and pig genomes. Genet Sel Evol. 2021;53:67. doi:10.1186/s12711-021-00662-x.
7. Grisart B, Farnir F, Karim L, Cambisano N, Kim J-J, Kvasz A, et al. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc Natl Acad Sci U S A. 2004;101:2398–403. doi:10.1073/pnas.0308518100.
8. Blott S, Kim J-J, Moisio S, Schmidt-Küntzel A, Cornet A, Berzi P, et al. Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics. 2003;163:253–66. doi:10.1093/genetics/163.1.253.

9. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101:5–22. doi:10.1016/j.ajhg.2017.06.005.
10. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24. doi:10.1016/j.ajhg.2011.11.029.
11. Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, et al. Extensive genome-wide linkage disequilibrium in cattle. Genome Res. 2000;10:220–7. doi:10.1101/gr.10.2.220.
12. Gibbs RA, Taylor JF, van Tassell CP, Barendse W, Eversole KA, Gill CA, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324:528–32. doi:10.1126/science.1167936.
13. Goddard ME, Kemper KE, MacLeod IM, Chamberlain AJ, Hayes BJ. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. Proc Biol Sci 2016. doi:10.1098/rspb.2016.0569.
14. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165:2213–33. doi:10.1093/genetics/165.4.2213.
15. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406. doi:10.1146/annurev.genom.9.081307.164242.
16. Hozé C, Fouilloux M-N, Venot E, Guillaume F, Dassonneville R, Fritz S, et al. High-density marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol. 2013;45:33. doi:10.1186/1297-9686-45-33.
17. Shi S, Yuan N, Yang M, Du Z, Wang J, Sheng X, et al. Comprehensive Assessment of Genotype Imputation Performance. Hum Hered. 2018;83:107–16. doi:10.1159/000489758.

628 18. van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, Veerkamp
629 RF. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet
630 Sel Evol. 2014;46:41. doi:10.1186/1297-9686-46-41.
631 19. Segelke D, Chen J, Liu Z, Reinhardt F, Thaller G, Reents R. Reliability of genomic prediction
632 for German Holsteins using imputed genotypes from low-density chips. J Dairy Sci.
633 2012;95:5403–11. doi:10.3168/jds.2012-5466.
634 20. Liu Z, Masuda Y. A deregression method for single-step genomic model using all genotype
635 data. 2021:41–51.
636 21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
637 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing
638 data. Genome Res. 2010;20:1297–303. doi:10.1101/gr.107524.110.
639 22. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation
640 Reference Panels. Am J Hum Genet. 2018;103:338–48. doi:10.1016/j.ajhg.2018.07.015.
641 23. van den Berg I, Ho PN, Nguyen TV, Haile-Mariam M, MacLeod IM, Beatson PR, et al.
642 GWAS and genomic prediction of milk urea nitrogen in Australian and New Zealand dairy
643 cattle. Genet Sel Evol. 2022;54:15. doi:10.1186/s12711-022-00707-9.
644 24. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data
645 inference for whole-genome association studies by use of localized haplotype clustering. Am
646 J Hum Genet. 2007;81:1084–97. doi:10.1086/521987.
647 25. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic
648 Traits in Cattle: Applications and Outcomes. Annu Rev Anim Biosci. 2019;7:89–102.
649 doi:10.1146/annurev-animal-020518-115024.

650 26. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of
651 SAMtools and BCFtools. Gigascience 2021. doi:10.1093/gigascience/giab008.
652 27. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase
653 inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84:210–
654 23. doi:10.1016/j.ajhg.2009.01.005.
655 28. Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, et al. Ensembl
656 variation resources. Database (Oxford) 2018. doi:10.1093/database/bay119.
657 29. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the
658 NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.
659 doi:10.1093/nar/29.1.308.
660 30. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait
661 analysis. Am J Hum Genet. 2011;88:76–82. doi:10.1016/j.ajhg.2010.11.011.
662 31. R Core Team. R: A Language and Environment for Statistical Computing 2022. Vienna,
663 Austria.
664 32. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide
665 association scans. Bioinformatics. 2010;26:2190–1. doi:10.1093/bioinformatics/btq340.
666 33. Cingolani P, Platts A, Le Wang L, Coon M, Nguyen T, Wang L, et al. A program for
667 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the
668 genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
669 doi:10.4161/fly.19695.
670 34. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila
671 melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program,
672 SnpSift. Front Genet. 2012;3:35. doi:10.3389/fgene.2012.00035.

673 35. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment
674 tool for interpreting omics data. Innovation (Camb). 2021;2:100141.
675 doi:10.1016/j.xinn.2021.100141.
676 36. Yu G, Wang L-G, Yan G-R, He Q-Y. DOSE: an R/Bioconductor package for disease ontology
677 semantic and enrichment analysis. Bioinformatics. 2015;31:608–9.
678 doi:10.1093/bioinformatics/btu684.
679 37. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and
680 outstanding challenges. PLoS Comput Biol. 2012;8:e1002375.
681 doi:10.1371/journal.pcbi.1002375.
682 38. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res.
683 2000;28:27–30. doi:10.1093/nar/28.1.27.
684 39. Hu Z-L, Fritz ER, Reecy JM. AnimalQTLdb: a livestock QTL database tool set for positional
685 QTL information mining and beyond. Nucleic Acids Res. 2007;35:D604-9.
686 doi:10.1093/nar/gkl946.
687 40. Mering C von, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of
688 predicted functional associations between proteins. Nucleic Acids Res. 2003;31:258–61.
689 doi:10.1093/nar/gkg034.
690 41. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable
691 Venn and Euler diagrams in R. BMC Bioinformatics. 2011;12:35. doi:10.1186/1471-2105-12-
692 35.
693 42. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs
694 explain a large proportion of the heritability for human height. Nat Genet. 2010;42:565–9.
695 doi:10.1038/ng.608.

696 43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool
697 set for whole-genome association and population-based linkage analyses. Am J Hum Genet.
698 2007;81:559–75. doi:10.1086/519795.
699 44. Khatkar MS, Moser G, Hayes BJ, Raadsma HW. Strategies and utility of imputed SNP
700 genotypes for genomic analysis in dairy cattle. BMC Genomics. 2012;13:538.
701 doi:10.1186/1471-2164-13-538.
702 45. Bouwman AC, Veerkamp RF. Consequences of splitting whole-genome sequencing effort
703 over multiple breeds on imputation accuracy. BMC Genet. 2014;15:105. doi:10.1186/s12863-
704 014-0105-8.
705 46. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3
706 (Bethesda). 2011;1:457–70. doi:10.1534/g3.111.001198.
707 47. Pausch H, MacLeod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, Goddard ME.
708 Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal
709 variant detection in cattle. Genet Sel Evol. 2017;49:24. doi:10.1186/s12711-017-0301-x.
710 48. Korkuć P, Arends D, Brockmann GA. Finding the Optimal Imputation Strategy for Small
711 Cattle Populations. Front Genet. 2019;10:52. doi:10.3389/fgene.2019.00052.
712 49. Pook T, Mayer M, Geibel J, Weigend S, Cavero D, Schoen CC, Simianer H. Improving
713 Imputation Quality in BEAGLE for Crop and Livestock Data. G3 (Bethesda). 2020;10:177–
714 88. doi:10.1534/g3.119.400798.
715 50. Stahl K, Gola D, König IR. Assessment of Imputation Quality: Comparison of Phasing and
716 Imputation Algorithms in Real Data. Front Genet. 2021;12:724037.
717 doi:10.3389/fgene.2021.724037.

718 51. Barendse W. The effect of measurement error of phenotypes on genome wide association
719 studies. BMC Genomics. 2011;12:232. doi:10.1186/1471-2164-12-232.
720 52. Kühn C, Thaller G, Winter A, Bininda-Emonds ORP, Kaupe B, Erhardt G, et al. Evidence for
721 multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect
722 on milk fat content in cattle. Genetics. 2004;167:1873–81. doi:10.1534/genetics.103.022749.
723 53. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet.
724 2015;24:R102-10. doi:10.1093/hmg/ddv259.
725 54. Koufariotis L, Chen Y-PP, Bolormaa S, Hayes BJ. Regulatory and coding genome regions are
726 enriched for trait associated variants in dairy and beef cattle. BMC Genomics. 2014;15:436.
727 doi:10.1186/1471-2164-15-436.
728 55. Raven L-A, Cocks BG, Kemper KE, Chamberlain AJ, Vander Jagt CJ, Goddard ME, Hayes
729 BJ. Targeted imputation of sequence variants and gene expression profiling identifies twelve
730 candidate genes associated with lactation volume, composition and calving interval in dairy
731 cattle. Mamm Genome. 2016;27:81–97. doi:10.1007/s00335-015-9613-8.
732 56. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant
733 Effect Predictor. Genome Biol. 2016;17:122. doi:10.1186/s13059-016-0974-4.
734 57. Prowse-Wilkins CP, Lopdell TJ, Xiang R, Vander Jagt CJ, Littlejohn MD, Chamberlain AJ,
735 Goddard ME. Genetic variation in histone modifications and gene expression identifies
736 regulatory variants in the mammary gland of cattle. BMC Genomics. 2022;23:815.
737 doi:10.1186/s12864-022-09002-9.
738 58. Ron M, Weller JI. From QTL to QTN identification in livestock--winning by points rather
739 than knock-out: a review. Anim Genet. 2007;38:429–39. doi:10.1111/j.1365-
740 2052.2007.01640.x.

59. Xiang R, van Berg I den, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, et al. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proc Natl Acad Sci U S A. 2019;116:19398–408. doi:10.1073/pnas.1904159116.

60. Hemani G, Yang J, Vinkhuyzen A, Powell JE, Willemsen G, Hottenga J-J, et al. Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. Am J Hum Genet. 2013;93:865–75. doi:10.1016/j.ajhg.2013.10.005.

61. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 2011;19:807–12. doi:10.1038/ejhg.2011.39.

62. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet. 2003;361:598–604. doi:10.1016/S0140-6736(03)12520-2.

63. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55:997–1004. doi:10.1111/j.0006-341x.1999.00997.x.

64. van den Berg S, Vandenplas J, van Eeuwijk FA, Lopes MS, Veerkamp RF. Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. J Anim Breed Genet. 2019;136:418–29. doi:10.1111/jbg.12419.

65. Da Pimentel ECG, Erbe M, König S, Simianer H. Genome partitioning of genetic variation for milk production and composition traits in holstein cattle. Front Genet. 2011;2:19. doi:10.3389/fgene.2011.00019.

66. Yin T, König S. Genome-wide associations and detection of potential candidate genes for direct genetic and maternal genetic effects influencing dairy cattle body weight at different ages. Genet Sel Evol. 2019;51:4. doi:10.1186/s12711-018-0444-4.

763 67. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association
764 analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat
765 Genet. 2010;42:937–48. doi:10.1038/ng.686.
766 68. Georgiopoulos G, Evangelou E. Power considerations for λ inflation factor in meta-analyses
767 of genome-wide association studies. Genet Res (Camb). 2016;98:e9.
768 doi:10.1017/S0016672316000069.
769 69. Jia R, Fu Y, Xu L, Li H, Li Y, Liu L, et al. Associations between polymorphisms of SLC22A7,
770 NGFR, ARNTL and PPP2R2B genes and Milk production traits in Chinese Holstein. BMC
771 Genom Data. 2021;22:47. doi:10.1186/s12863-021-01002-0.
772 70. Tang D, Yeung J, Lee KY, Matsushita M, Matsui H, Tomizawa K, et al. An isoform of the
773 neuronal cyclin-dependent kinase 5 (Cdk5) activator. J Biol Chem. 1995;270:26897–903.
774 doi:10.1074/jbc.270.45.26897.
775 71. Dhariwala FA, Rajadhyaksha MS. An unusual member of the Cdk family: Cdk5. Cell Mol
776 Neurobiol. 2008;28:351–69. doi:10.1007/s10571-007-9242-1.
777 72. Talouarn E, Bardou P, Palhière I, Oget C, Clément V, Tosser-Klopp G, et al. Genome wide
778 association analysis on semen volume and milk yield using different strategies of imputation
779 to whole genome sequence in French dairy goats. BMC Genet. 2020;21:19.
780 doi:10.1186/s12863-020-0826-9.
781 73. Chen X, Cheng Z, Zhang S, Werling D, Wathes DC. Combining Genome Wide Association
782 Studies and Differential Gene Expression Data Analyses Identifies Candidate Genes Affecting
783 Mastitis Caused by Two Different Pathogens in the Dairy Cow. OJAS. 2015;05:358–93.
784 doi:10.4236/ojas.2015.54040.

785 74. Marín-Garzón NA, Magalhães AFB, Mota LFM, Fonseca LFS, Chardulo LAL, Albuquerque
786 LG. Genome-wide association study identified genomic regions and putative candidate genes
787 affecting meat color traits in Nellore cattle. Meat Sci. 2021;171:108288.
788 doi:10.1016/j.meatsci.2020.108288.
789 75. Heidaritabar M, Huisman A, Krivushin K, Stothard P, Dervishi E, Charagu P, et al. Imputation
790 to whole-genome sequence and its use in genome-wide association studies for pork colour
791 traits in crossbred and purebred pigs. Front Genet. 2022;13:1022681.
792 doi:10.3389/fgene.2022.1022681.
793 76. Wu X, Fang M, Liu L, Wang S, Liu J, Ding X, et al. Genome wide association studies for
794 body conformation traits in the Chinese Holstein cattle population. BMC Genomics.
795 2013;14:897. doi:10.1186/1471-2164-14-897.
796 77. Ahmed MI, Elias S, Mould AW, Bikoff EK, Robertson EJ. The transcriptional repressor
797 Blimp1 is expressed in rare luminal progenitors and is essential for mammary gland
798 development. Development. 2016;143:1663–73. doi:10.1242/dev.136358.
799 78. Muncan V, Heijmans J, Krasinski SD, Büller NV, Wildenberg ME, Meisner S, et al. Blimp1
800 regulates the transition of neonatal to adult intestinal epithelium. Nat Commun. 2011;2:452.
801 doi:10.1038/ncomms1463.
802 79. Zhao J, Mialki RK, Wei J, Coon TA, Zou C, Chen BB, et al. SCF E3 ligase F-box protein
803 complex SCF(FBXL19) regulates cell migration by mediating Rac1 ubiquitination and
804 degradation. FASEB J. 2013;27:2611–9. doi:10.1096/fj.12-223099.
805 80. Wei J, Mialki RK, Dong S, Khoo A, Mallampalli RK, Zhao Y, Zhao J. A new mechanism of
806 RhoA ubiquitination and degradation: roles of SCF(FBXL19) E3 ligase and Erk2. Biochim
807 Biophys Acta. 2013;1833:2757–64. doi:10.1016/j.bbamcr.2013.07.005.

808 81. Acharya A, Berry DC, Zhang H, Jiang Y, Jones BT, Hammer RE, et al. miR-26 suppresses
809 adipocyte progenitor differentiation and fat production by targeting Fbxl19. Genes Dev.
810 2019;33:1367–80. doi:10.1101/gad.328955.119.
811 82. McNamara JP. Regulation of Adipose Tissue Metabolism in Support of Lactation1. J Dairy
812 Sci. 1991;74:706–19.
813 83. McNamara JP, Huber K, Kenéz A. A dynamic, mechanistic model of metabolism in adipose
814 tissue of lactating dairy cattle. J Dairy Sci. 2016;99:5649–61. doi:10.3168/jds.2015-9585.
815 84. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards
816 Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc
817 Bioinformatics. 2016;54:1.30.1-1.30.33. doi:10.1002/cpbi.5.
818 85. Pombo CM, Force T, Kyriakis J, Nogueira E, Fidalgo M, Zalvide J. The GCK II and III
819 subfamilies of the STE20 group kinases. Front Biosci. 2007;12:850–9. doi:10.2741/2107.
820 86. Amrutkar M, Kern M, Nuñez-Durán E, Ståhlman M, Cansby E, Chursa U, et al. Protein kinase
821 STK25 controls lipid partitioning in hepatocytes and correlates with liver fat content in
822 humans. Diabetologia. 2016;59:341–53. doi:10.1007/s00125-015-3801-7.
823 87. Sun X, Wang Y, Loor JJ, Bucktrout R, Shu X, Jia H, et al. High expression of cell death-
824 inducing DFFA-like effector a (CIDEA) promotes milk fat content in dairy cows with clinical
825 ketosis. J Dairy Sci. 2019;102:1682–92. doi:10.3168/jds.2018-15439.
826 88. Cheng J, Xu D, Chen L, Guo W, Hu G, Liu J, Fu S. CIDEA Regulates De Novo Fatty Acid
827 Synthesis in Bovine Mammary Epithelial Cells by Targeting the AMPK/PPARγ Axis and
828 Regulating SREBP1. J Agric Food Chem. 2022;70:11324–35. doi:10.1021/acs.jafc.2c05226.

829 89. van Knegsel ATM, van den Brand H, Dijkstra J, Tamminga S, Kemp B. Effect of dietary
830 energy source on energy balance, production, metabolic disorders and reproduction in
831 lactating dairy cattle. Reprod Nutr Dev. 2005;45:665–88. doi:10.1051/rnd:2005059.
832 90. Nerstedt A, Cansby E, Andersson CX, Laakso M, Stančáková A, Blüher M, et al.
833 Serine/threonine protein kinase 25 (STK25): a novel negative regulator of lipid and glucose
834 metabolism in rodent and human skeletal muscle. Diabetologia. 2012;55:1797–807.
835 doi:10.1007/s00125-012-2511-7.
836 91. Butkinaree C, Guo L, Ramkhelawon B, Wanschel A, Brodsky JL, Moore KJ, Fisher EA. A
837 regulator of secretory vesicle size, Kelch-like protein 12, facilitates the secretion of
838 apolipoprotein B100 and very-low-density lipoproteins--brief report. Arterioscler Thromb
839 Vasc Biol. 2014;34:251–4. doi:10.1161/ATVBAHA.113.302728.
840 92. Chan L. Apolipoprotein B, the major protein component of triglyceride-rich and low density
841 lipoproteins. J Biol Chem. 1992;267:25621–4.
842 93. Bauchart D. Lipid absorption and transport in ruminants. J Dairy Sci. 1993;76:3864–81.
843 doi:10.3168/jds.S0022-0302(93)77728-0.
844 94. Oikawa S, Katoh N, Kawawa F, Ono Y. Decreased serum apolipoprotein B-100 and A-I
845 concentrations in cows with ketosis and left displacement of the abomasum. Am J Vet Res.
846 1997;58:121–5.
847 95. Oikawa S, Katoh N. Decreases in serum apolipoprotein B-100 and A-I concentrations in cows
848 with milk fever and downer cows. Can J Vet Res. 2002;66:31–4.
849 96. Shi X, Xiang S, Cao J, Zhu H, Yang B, He Q, Ying M. Kelch-like proteins: Physiological
850 functions and relationships with diseases. Pharmacol Res. 2019;148:104404.
851 doi:10.1016/j.phrs.2019.104404.

852 97. Angers S, Thorpe CJ, Biechele TL, Goldenberg SJ, Zheng N, MacCoss MJ, Moon RT. The
853 KLHL12-Cullin-3 ubiquitin ligase negatively regulates the Wnt-beta-catenin pathway by
854 targeting Dishevelled for degradation. Nat Cell Biol. 2006;8:348–57. doi:10.1038/ncb1381.
855 98. Steinhart Z, Angers S. Wnt signaling in development and tissue homeostasis. Development
856 2018. doi:10.1242/dev.146589.
857 99. Jin L, Pahuja KB, Wickliffe KE, Gorur A, Baumgärtel C, Schekman R, Rape M. Ubiquitin-
858 dependent regulation of COPII coat size and function. Nature. 2012;482:495–500.
859 doi:10.1038/nature10822.
860 100. Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, et al., editors. Essentials
861 of glycobiology. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2022.
862 101. Aebi M. N-linked protein glycosylation in the ER. Biochim Biophys Acta.
863 2013;1833:2430–7. doi:10.1016/j.bbamcr.2013.04.001.
864 102. Schwarz F, Aebi M. Mechanisms and principles of N-linked protein glycosylation. Curr
865 Opin Struct Biol. 2011;21:576–82. doi:10.1016/j.sbi.2011.08.005.
866 103. Kelleher DJ, Gilmore R. An evolving view of the eukaryotic oligosaccharyltransferase.
867 Glycobiology. 2006;16:47R-62R. doi:10.1093/glycob/cwj066.
868 104. Barboza M, Pinzon J, Wickramasinghe S, Froehlich JW, Moeller I, Smilowitz JT, et al.
869 Glycosylation of human milk lactoferrin exhibits dynamic changes during early lactation
870 enhancing its role in pathogenic bacteria-host interactions. Mol Cell Proteomics.
871 2012;11:M111.015248. doi:10.1074/mcp.M111.015248.
872 105. Rascón-Cruz Q, Espinoza-Sánchez EA, Siqueiros-Cendón TS, Nakamura-Bencomo SI,
873 Arévalo-Gallegos S, Iglesias-Figueroa BF. Lactoferrin: A Glycoprotein Involved in

874 Immunomodulation, Anticancer, and Antimicrobial Processes. Molecules 2021.
875 doi:10.3390/molecules26010205.
876 106. Zhou P, Xie W, He S, Sun Y, Meng X, Sun G, Sun X. Ginsenoside Rb1 as an Anti-Diabetic
877 Agent and Its Underlying Mechanism Analysis. Cells 2019. doi:10.3390/cells8030204.
878 107. Zou H, Zhang M, Zhu X, Zhu L, Chen S, Luo M, et al. Ginsenoside Rb1 Improves
879 Metabolic Disorder in High-Fat Diet-Induced Obese Mice Associated With Modulation of Gut
880 Microbiota. Front Microbiol. 2022;13:826487. doi:10.3389/fmicb.2022.826487.
881 108. Hemmings BA, Restuccia DF. PI3K-PKB/Akt pathway. Cold Spring Harb Perspect Biol.
882 2012;4:a011189. doi:10.1101/cshperspect.a011189.
883 109. Vanhaesebroeck B, Guillermet-Guibert J, Graupera M, Bilanges B. The emerging
884 mechanisms of isoform-specific PI3K signalling. Nat Rev Mol Cell Biol. 2010;11:329–41.
885 doi:10.1038/nrm2882.
886 110. Hou Y, Xie Y, Yang S, Han B, Shi L, Bai X, et al. EEF1D facilitates milk lipid synthesis
887 by regulation of PI3K-Akt signaling in mammals. FASEB J. 2021;35:e21455.
888 doi:10.1096/fj.202000682RR.
889 111. Bole-Feysot C, Goffin V, Edery M, Binart N, Kelly PA. Prolactin (PRL) and its receptor:
890 actions, signal transduction pathways and phenotypes observed in PRL receptor knockout
891 mice. Endocr Rev. 1998;19:225–68. doi:10.1210/edrv.19.3.0334.
892 112. Hardy KM, Booth BW, Hendrix MJC, Salomon DS, Strizzi L. ErbB/EGF signaling and
893 EMT in mammary development and breast cancer. J Mammary Gland Biol Neoplasia.
894 2010;15:191–9. doi:10.1007/s10911-010-9172-2.

113. Monkkonen T, Lewis MT. New paradigms for the Hedgehog signaling network in mammary gland development and breast Cancer. Biochim Biophys Acta Rev Cancer. 2017;1868:315–32. doi:10.1016/j.bbcan.2017.06.003.

114. Chen Q, Zhang N, Gray RS, Li H, Ewald AJ, Zahnow CA, Pan D. A temporal requirement for Hippo signaling in mammary gland differentiation, growth, and tumorigenesis. Genes Dev. 2014;28:432–7. doi:10.1101/gad.233676.113.

115. Werb Z, Sympson CJ, Alexander CM, Thomasset N, Lund LR, MacAuley A, et al. Extracellular matrix remodeling and the regulation of epithelial-stromal interactions during differentiation and involution. Kidney Int Suppl. 1996;54:S68-74.

116. Du A, Zhao F, Liu Y, Xu L, Chen K, Sun D, Han B. Genetic polymorphisms of PKLR gene and their associations with milk production traits in Chinese Holstein cows. Front Genet. 2022;13:1002706. doi:10.3389/fgene.2022.1002706.

117. Liu Z, Zhang C, Lee S, Kim W, Klevstig M, Harzandi AM, et al. Pyruvate kinase L/R is a regulator of lipid metabolism and mitochondrial function. Metab Eng. 2019;52:263–72. doi:10.1016/j.ymben.2019.01.001.

118. Atashi H, Bastin C, Wilmot H, Vanderick S, Hubin X, Gengler N. Genome-wide association study for selected cheese-making properties in Dual-Purpose Belgian Blue cows. J Dairy Sci. 2022;105:8972–88. doi:10.3168/jds.2022-21780.

119. Costa A, Schwarzenbacher H, Mészáros G, Fuerst-Waltl B, Fuerst C, Sölkner J, Penasa M. On the genomic regions associated with milk lactose in Fleckvieh cattle. J Dairy Sci. 2019;102:10088–99. doi:10.3168/jds.2019-16663.

120. Huang QX, Yang J, Hu M, Lu W, Zhong K, Wang Y, et al. Milk fat globule membrane proteins are involved in controlling the size of milk fat globules during conjugated linoleic

acid-induced milk fat depression. J Dairy Sci. 2022;105:9179–90. doi:10.3168/jds.2022-22131.

121. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51:D523-D531. doi:10.1093/nar/gkac1052.

122. Soares RAN, Vargas G, Duffield T, Schenkel F, Squires EJ. Genome-wide association study and functional analyses for clinical and subclinical ketosis in Holstein cattle. J Dairy Sci. 2021;104:10076–89. doi:10.3168/jds.2020-20101.

123. Ding Y, Cui J, Wang Q, Shen S, Xu T, Tang H, et al. The Vitamin K Epoxide Reductase Vkorc1l1 Promotes Preadipocyte Differentiation in Mice. Obesity (Silver Spring). 2018;26:1303–11. doi:10.1002/oby.22206.

124. Lacasse P, Ollier S, Lollivier V, Boutinaud M. New insights into the importance of prolactin in dairy ruminants. J Dairy Sci. 2016;99:864–74. doi:10.3168/jds.2015-10035.

125. Pedrosa VB, Schenkel FS, Chen S-Y, Oliveira HR, Casey TM, Melka MG, Brito LF. Genomewide Association Analyses of Lactation Persistency and Milk Production Traits in Holstein Cattle Based on Imputed Whole-Genome Sequence Data. Genes (Basel) 2021. doi:10.3390/genes12111830.

126. Olayioye MA, Neve RM, Lane HA, Hynes NE. The ErbB signaling network: receptor heterodimerization in development and cancer. EMBO J. 2000;19:3159–67. doi:10.1093/emboj/19.13.3159.

127. Moraes RC, Chang H, Harrington N, Landua JD, Prigge JT, Lane TF, et al. Ptch1 is required locally for mammary gland morphogenesis and systemically for ductal elongation. Development. 2009;136:1423–32. doi:10.1242/dev.023994.

940 128. Palombo V, Milanesi M, Sgorlon S, Capomaccio S, Mele M, Nicolazzi E, et al. Genome-
941 wide association study of milk fatty acid composition in Italian Simmental and Italian Holstein
942 cows using single nucleotide polymorphism arrays. J Dairy Sci. 2018;101:11004–19.
943 doi:10.3168/jds.2018-14413.
944 129. Korkuć P, Arends D, May K, König S, Brockmann GA. Genomic Loci Affecting Milk
945 Production in German Black Pied Cattle (DSN). Front Genet. 2021;12:640039.
946 doi:10.3389/fgene.2021.640039.
947 130. Xu R, Boudreau A, Bissell MJ. Tissue architecture and function: dynamic reciprocity via
948 extra- and intra-cellular matrices. Cancer Metastasis Rev. 2009;28:167–76.
949 doi:10.1007/s10555-008-9178-z.
950 131. Nayeri S, Sargolzaei M, Abo-Ismail MK, Miller S, Schenkel F, Moore SS, Stothard P.
951 Genome-wide association study for lactation persistency, female fertility, longevity, and
952 lifetime profit index traits in Holstein dairy cattle. J Dairy Sci. 2017;100:1246–58.
953 doi:10.3168/jds.2016-11770.
954 132. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and
955 interpretation of SNP-based heritability. Nat Genet. 2017;49:1304–10. doi:10.1038/ng.3941.
956 133. Barry C-JS, Walker VM, Cheesman R, Davey Smith G, Morris TT, Davies NM. How to
957 estimate heritability: a guide for genetic epidemiologists. Int J Epidemiol. 2023;52:624–32.
958 doi:10.1093/ije/dyac224.
959 134. Hassel C, Gausserès B, Guzylack-Piriou L, Foucras G. Ductal Macrophages Predominate
960 in the Immune Landscape of the Lactating Mammary Gland. Front Immunol. 2021;12:754661.
961 doi:10.3389/fimmu.2021.754661.

962 135. Michelotti TC, Kisby BR, Flores LS, Tegeler AP, Fokar M, Crasto C, et al. Single-nuclei
963 analysis reveals depot-specific transcriptional heterogeneity and depot-specific cell types in
964 adipose tissue of dairy cows. Front Cell Dev Biol. 2022;10:1025240.
965 doi:10.3389/fcell.2022.1025240.
966 136. Shang P, Li W, Liu G, Zhang J, Li M, Wu L, et al. Identification of lncRNAs and Genes
967 Responsible for Fatness and Fatty Acid Composition Traits between the Tibetan and Yorkshire
968 Pigs. Int J Genomics. 2019;2019:5070975. doi:10.1155/2019/5070975.
969 137. Zhang X, Zhang S, Ma L, Jiang E, Xu H, Chen R, et al. Reduced representation bisulfite
970 sequencing (RRBS) of dairy goat mammary glands reveals DNA methylation profiles of
971 integrated genome-wide and critical milk-related genes. Oncotarget. 2017;8:115326–44.
972 doi:10.18632/oncotarget.23260.
973 138. Igoshin AV, Deniskova TE, Yurchenko AA, Yudin NS, Dotsev AV, Selionova MI, et al.
974 Copy number variants in genomes of local sheep breeds from Russia. Anim Genet.
975 2022;53:119–32. doi:10.1111/age.13163.
976 139. Steri R, Moioli B, Catillo G, Galli A, Buttazzoni L. Genome-wide association study for
977 longevity in the Holstein cattle population. Animal. 2019;13:1350–7.
978 doi:10.1017/S1751731118003191.
979 140. Briand N, Dugail I, Le Lay S. Cavin proteins: New players in the caveolae field. Biochimie.
980 2011;93:71–7. doi:10.1016/j.biochi.2010.03.022.
981 141. Parton RG, Simons K. The multiple faces of caveolae. Nat Rev Mol Cell Biol. 2007;8:185–
982 94. doi:10.1038/nrm2122.

983 142. Liu L, Brown D, McKee M, Lebrasseur NK, Yang D, Albrecht KH, et al. Deletion of
984 Cavin/PTRF causes global loss of caveolae, dyslipidemia, and glucose intolerance. Cell
985 Metab. 2008;8:310–7. doi:10.1016/j.cmet.2008.07.008.
986 143. Morteau O, Gerard C, Lu B, Ghiran S, Rits M, Fujiwara Y, et al. An indispensable role for
987 the chemokine receptor CCR10 in IgA antibody-secreting cell accumulation. J Immunol.
988 2008;181:6309–15. doi:10.4049/jimmunol.181.9.6309.
989 144. Wang W, Soto H, Oldham ER, Buchanan ME, Homey B, Catron D, et al. Identification of
990 a novel chemokine (CCL28), which binds CCR10 (GPR2). J Biol Chem. 2000;275:22313–23.
991 doi:10.1074/jbc.M001461200.
992 145. Illa SK, Mukherjee S, Nath S, Mukherjee A. Genome-Wide Scanning for Signatures of
993 Selection Revealed the Putative Genomic Regions and Candidate Genes Controlling Milk
994 Composition and Coat Color Traits in Sahiwal Cattle. Front Genet. 2021;12:699422.
995 doi:10.3389/fgene.2021.699422.
996 146. Do DN, Bissonnette N, Lacasse P, Miglior F, Sargolzaei M, Zhao X, Ibeagha-Awemu EM.
997 Genome-wide association analysis and pathways enrichment for lactation persistency in
998 Canadian Holstein cattle. J Dairy Sci. 2017;100:1955–70. doi:10.3168/jds.2016-11910.
999 147. Jiang J, Ma L, Prakapenka D, VanRaden PM, Cole JB, Da Y. A Large-Scale Genome-
1000 Wide Association Study in U.S. Holstein Cattle. Front Genet. 2019;10:412.
1001 doi:10.3389/fgene.2019.00412.
1002 148. Le Guillou S, Sdassi N, Laubier J, Passet B, Vilotte M, Castille J, et al. Overexpression of
1003 miR-30b in the developing mouse mammary gland causes a lactation defect and delays
1004 involution. PLoS One. 2012;7:e45727. doi:10.1371/journal.pone.0045727.

149. Silva ÉF, Lopes MS, Lopes PS, Gasparino E. A genome-wide association study for feed efficiency-related traits in a crossbred pig population. Animal. 2019;13:2447–56. doi:10.1017/S1751731119000910.

150. Lee J, Kim Y, Cho E, Cho K, Sa S, Kim Y, et al. Genomic Analysis Using Bayesian Methods under Different Genotyping Platforms in Korean Duroc Pigs. Animals (Basel) 2020. doi:10.3390/ani10050752.

151. Lee J-E, Cho Y-W, Deng C-X, Ge K. MLL3/MLL4-Associated PAGR1 Regulates Adipogenesis by Controlling Induction of C/EBPβ and C/EBPδ. Mol Cell Biol 2020. doi:10.1128/MCB.00209-20.

152. Garcia M, Greco LF, Lock AL, Block E, Santos JEP, Thatcher WW, Staples CR. Supplementation of essential fatty acids to Holstein calves during late uterine life and first month of life alters hepatic fatty acid profile and gene expression. J Dairy Sci. 2016;99:7085–101. doi:10.3168/jds.2015-10472.

153. Huang Z, Wu L-M, Zhang J-L, Sabri A, Wang S-J, Qin G-J, et al. Dual Specificity Phosphatase 12 Regulates Hepatic Lipid Metabolism Through Inhibition of the Lipogenesis and Apoptosis Signal-Regulating Kinase 1 Pathways. Hepatology. 2019;70:1099–118. doi:10.1002/hep.30597.

154. Jiang J, Cole JB, Freebern E, Da Y, VanRaden PM, Ma L. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. Commun Biol. 2019;2:212. doi:10.1038/s42003-019-0454-y.

155. Aloor JJ, Azzam KM, Guardiola JJ, Gowdy KM, Madenspacher JH, Gabor KA, et al. Leucine-rich repeats and calponin homology containing 4 (Lrch4) regulates the innate immune response. J Biol Chem. 2019;294:1997–2008. doi:10.1074/jbc.RA118.004300.

1028 156. Zhang M, Ma Z, Qi H, Cui X, Li R, Gao X. Comparative transcriptomic analysis of
1029 mammary gland tissues reveals the critical role of GPR110 in palmitic acid-stimulated milk
1030 protein and fat synthesis. Br J Nutr. 2023:1–13. doi:10.1017/S0007114523000788.
1031 157. Blaas L, Pucci F, Messal HA, Andersson AB, Josue Ruiz E, Gerling M, et al. Lgr6 labels
1032 a rare population of mammary gland progenitor cells that are able to originate luminal
1033 mammary tumours. Nat Cell Biol. 2016;18:1346–56. doi:10.1038/ncb3434.
1034 158. Di Gerlando R, Sutera AM, Mastrangelo S, Tolone M, Portolano B, Sottile G, et al.
1035 Genome-wide association study between CNVs and milk production traits in Valle del Belice
1036 sheep. PLoS One. 2019;14:e0215204. doi:10.1371/journal.pone.0215204.
1037 159. Sutera AM, Riggio V, Mastrangelo S, Di Gerlando R, Sardina MT, Pong-Wong R, et al.
1038 Genome-wide association studies for milk production traits in Valle del Belice sheep using
1039 repeated measures. Anim Genet. 2019;50:311–4. doi:10.1111/age.12789.
1040 160. Wang R, Shen J, Chen Y, Gao J, Yao J. Fatty acid metabolism-related signature predicts
1041 survival in patients with clear cell renal carcinoma. Aging (Albany NY). 2022;14:9969–79.
1042 doi:10.18632/aging.204433.
1043 161. Bolormaa S, Hayes BJ, van der Werf JHJ, Pethick D, Goddard ME, Daetwyler HD.
1044 Detailed phenotyping identifies genes with pleiotropic effects on body composition. BMC
1045 Genomics. 2016;17:224. doi:10.1186/s12864-016-2538-0.
1046 162. Lu Y, Day FR, Gustafsson S, Buchkovich ML, Na J, Bataille V, et al. New loci for body
1047 fat percentage reveal link between adiposity and cardiometabolic disease risk. Nat Commun.
1048 2016;7:10495. doi:10.1038/ncomms10495.
1049 163. Bowman SR. Investigation of the role of bovine mammary stem cells in the lactation cycle
1050 [PhD thesis]: The University of Sydney; 2016.

1051 164. Niska-Blakie J, Gopinathan L, Low KN, Kien YL, Goh CMF, Caldez MJ, et al. Knockout
1052 of the non-essential gene SUGCT creates diet-linked, age-related microbiome disbalance with
1053 a diabetes-like metabolic syndrome phenotype. Cell Mol Life Sci. 2020;77:3423–39.
1054 doi:10.1007/s00018-019-03359-z.
1055 165. Kweon S-M, Irimia-Dominguez J, Kim G, Fueger PT, Asahina K, Lai KK, et al.
1056 Heterozygous midnolin knockout attenuates severity of nonalcoholic fatty liver disease in
1057 mice fed a Western-style diet high in fat, cholesterol, and fructose. Am J Physiol Gastrointest
1058 Liver Physiol. 2023;325:G147-G157. doi:10.1152/ajpgi.00011.2023.
1059 166. Zhang Y, Li C, Hu C, Wu Q, Cai Y, Xing S, et al. Lin28 enhances de novo fatty acid
1060 synthesis to promote cancer progression via SREBP-1. EMBO Rep. 2019;20:e48115.
1061 doi:10.15252/embr.201948115.
1062 167. Jiang D-X, Zhang J-B, Li M-T, Lin S-Z, Wang Y-Q, Chen Y-W, Fan J-G. Prolyl
1063 endopeptidase gene disruption attenuates high fat diet-induced nonalcoholic fatty liver disease
1064 in mice by improving hepatic steatosis and inflammation. Ann Transl Med. 2020;8:218.
1065 doi:10.21037/atm.2020.01.14.
1066 168. Shin J, Syme C, Wang D, Richer L, Pike GB, Gaudet D, et al. Novel Genetic Locus of
1067 Visceral Fat and Systemic Inflammation. J Clin Endocrinol Metab. 2019;104:3735–42.
1068 doi:10.1210/jc.2018-02656.
1069 169. Zhang Z-G, Zhang H-S, Sun H-L, Liu H-Y, Liu M-Y, Zhou Z. KDM5B promotes breast
1070 cancer cell proliferation and migration via AMPK-mediated lipid metabolism reprogramming.
1071 Exp Cell Res. 2019;379:182–90. doi:10.1016/j.yexcr.2019.04.006.

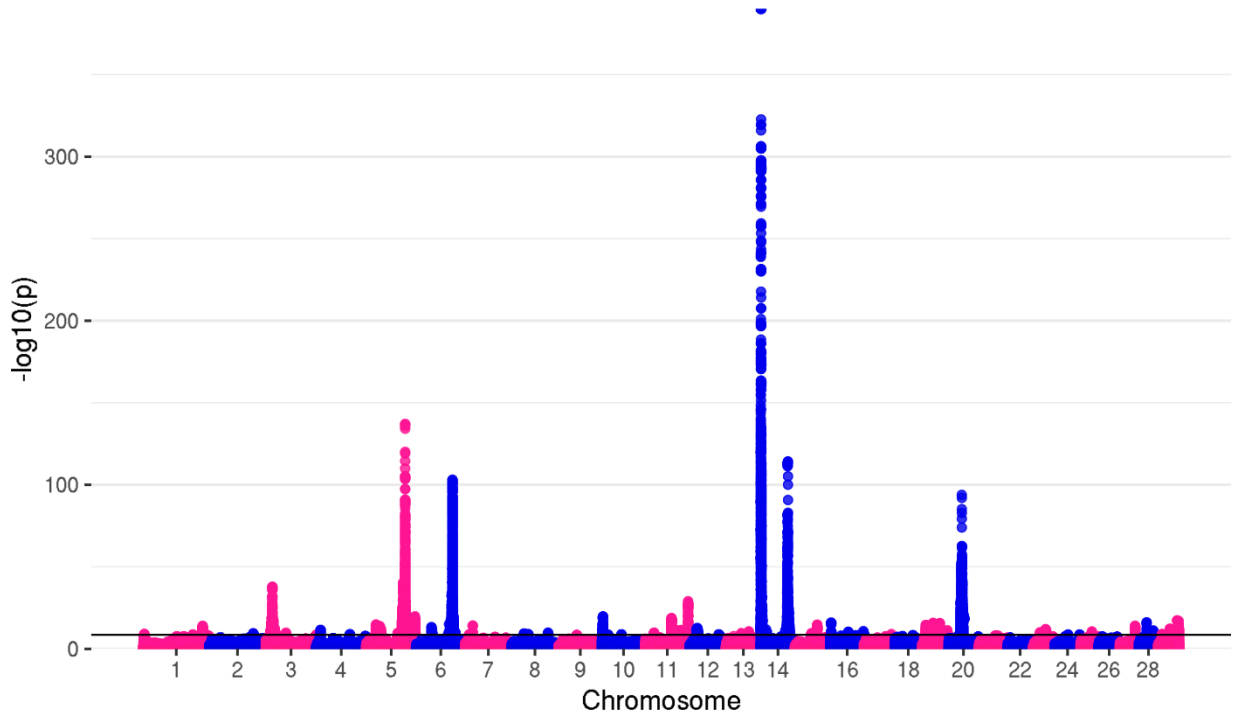
170. Ye JJ, Bian X, Lim J, Medzhitov R. Adiponectin and related C1q/TNF-related proteins bind selectively to anionic phospholipids and sphingolipids. Proc Natl Acad Sci U S A. 2020;117:17381–8. doi:10.1073/pnas.1922270117.

171. Naserkheil M, Bahrami A, Lee D, Mehrban H. Integrating Single-Step GWAS and Bipartite Networks Reconstruction Provides Novel Insights into Yearling Weight and Carcass Traits in Hanwoo Beef Cattle. Animals (Basel) 2020. doi:10.3390/ani10101836.

172. Sutera AM. Comparison of Genome Wide Association Studies for milk production traits in Valle del Belice dairy sheep [PhD Thesis]: Università degli Studi di Palermo; 2018.

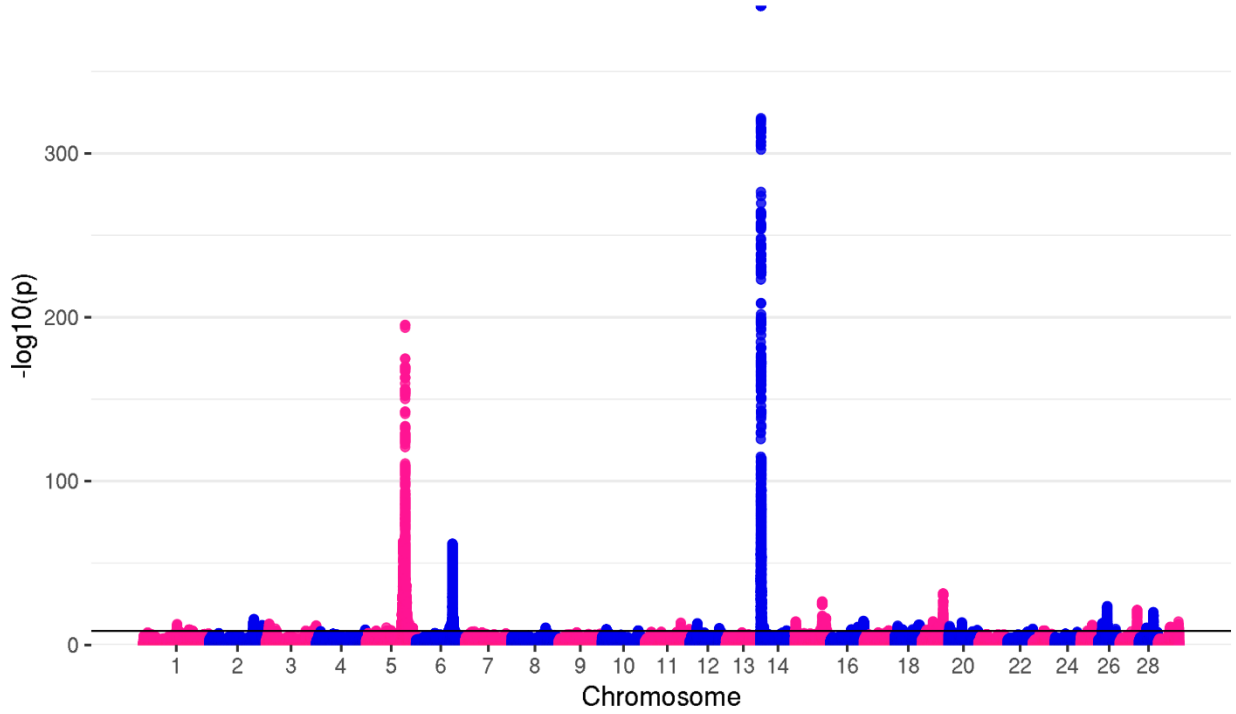
Figures

Figure 1. Manhattan plot for milk yield



The top genome-wide SNP ($p = 7.04x10^{-737}$) for MY was located on BTA14. However, RStudio used for the creation of this plot was not able to show p-values < $10x10^{-325}$, reporting them as "0". Therefore, ylim had to be set lower, to provisional ylim of 390, in order to present all significant variants

Figure 2. Manhattan plot for fat yield



The top genome-wide SNP ($p = 7.18x10^{-380}$) for FY was positioned on BTA14. However, RStudio used for the creation of this plot was not able to show p-values < $10x10^{-325}$, reporting them as "0". Therefore, ylim had to be set lower, to provisional ylim of 390, in order to present all significant variants

Figure 3. Manhattan plot for protein yield

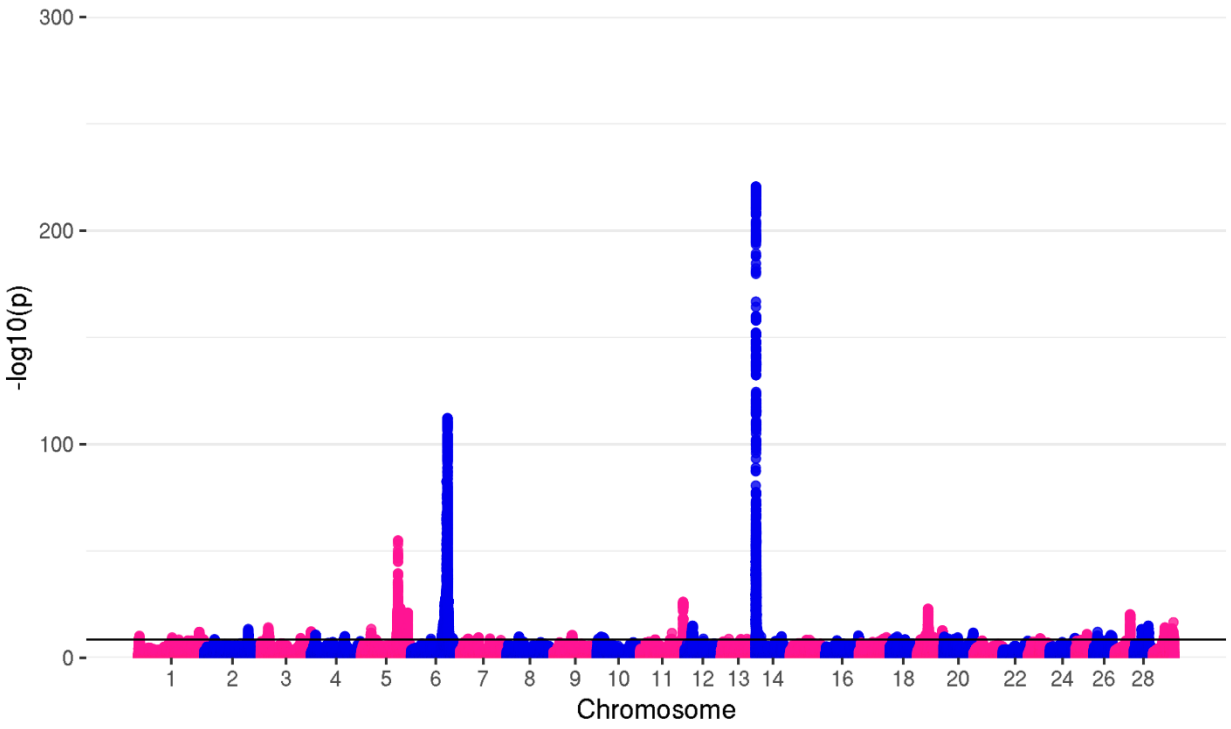


Figure 4. Venn diagram of MY, FY, and PY showing concordant and discordant candidate genes

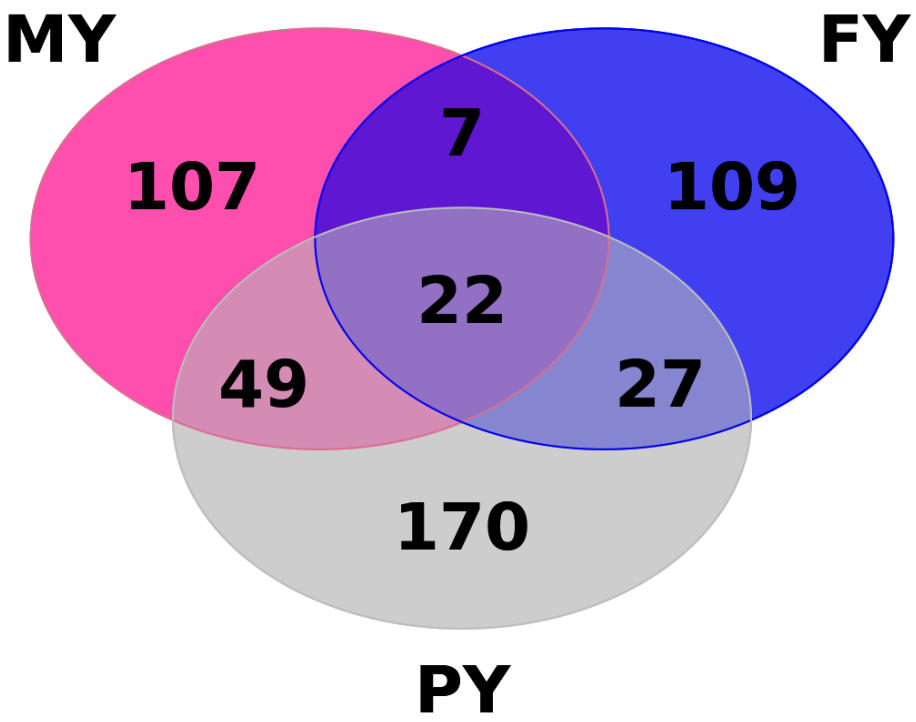
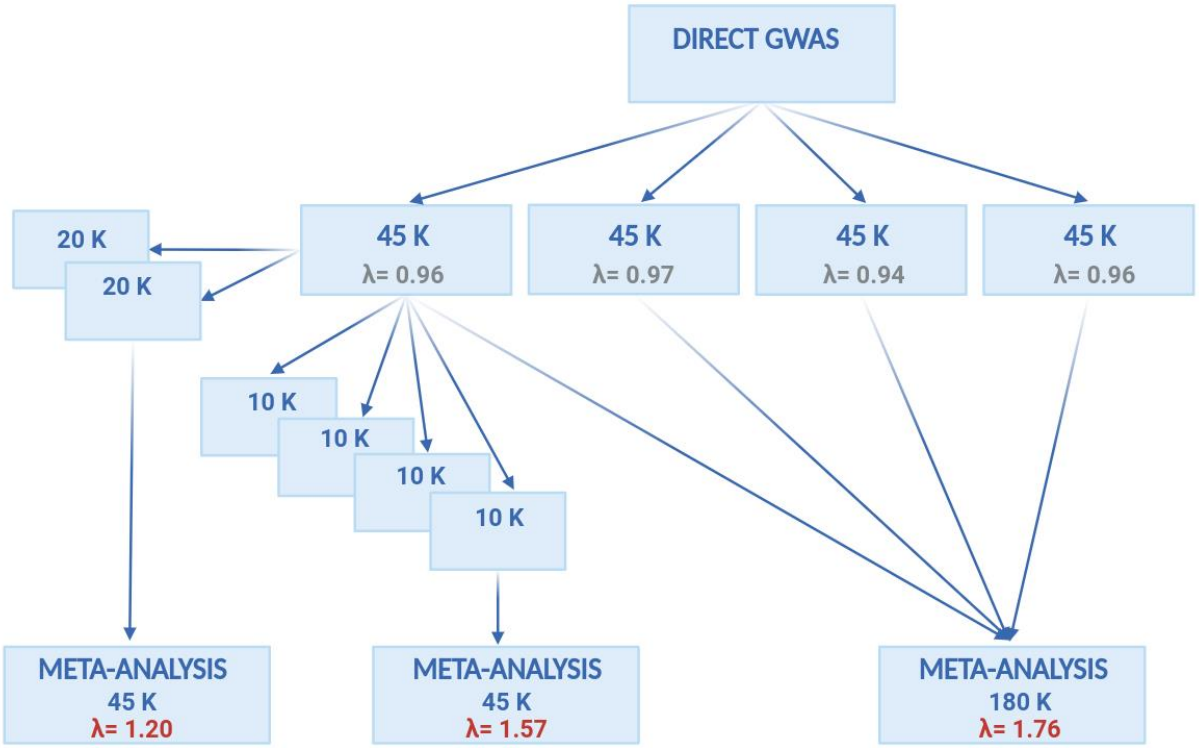
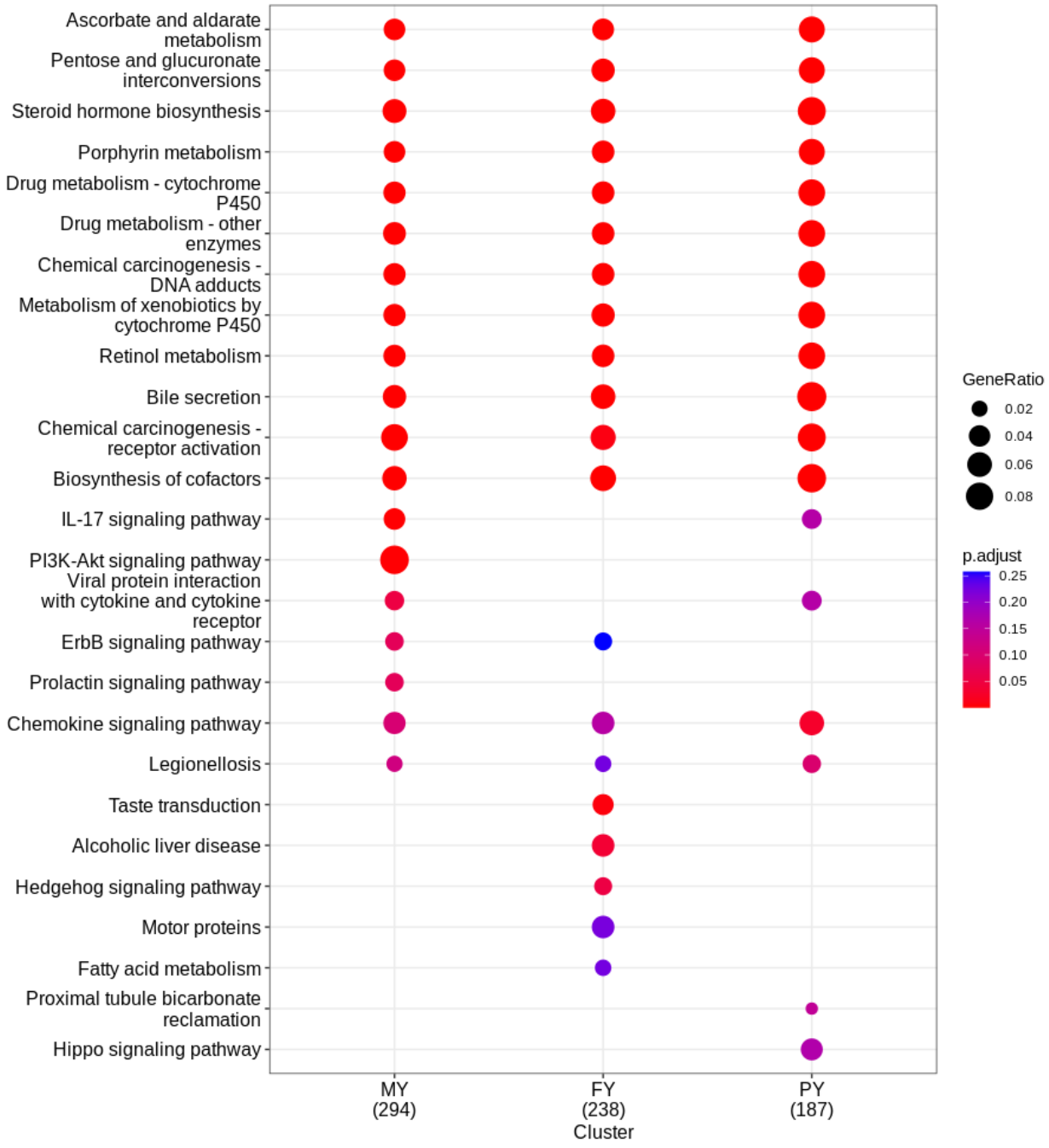


Figure 5. Genomic inflation factors of MY measured on direct GWAS summary statistics and after meta-analysis



To check the cause of genomic inflation in meta-analysis summary statistics, one of the animal groups on which we ran direct GWAS was divided into two groups. For each of the two groups, GWAS was run again, and summary statistics were merged into the meta-analysis. Lambda values obtained on meta-analysis summary statistics were higher (λ= 1.20) than ones measured for the same individuals on direct GWAS summary statistics (λ= 0.96). To further check the extent of inflation caused by meta-analysis, the same group of animals was divided again, this time, into four groups. GWAS was run for each of the groups and results were merged into the meta-analysis. Lambda values were even higher this time (λ= 1.57). The figure was created with BioRender.com

Figure 6. Enrich KEGG dot plot of 18 most significant pathways for MY, FY, and PY



Tables

Table 1 Over-represented genes associated with top variants for MY, FY, and PY

Term	MY	FY	PY
PI3K-Akt signaling pathway	*EFNA1, EFNA3, EFNA4, THBS3, LAMA5, GHR, IGF2*		
Biosynthesis of amino acids	*PKLR*		
Maturity onset diabetes of the young	*PKLR*		
Biosynthesis of cofactors	*FLAD1*	*GMPPA, VKORC1L1*	*VKORC1L1*
Chemical carcinogenesis - receptor activation	*MGST1, STAT5B, STAT3, STAT5A*	*MGST1, MIRLET7E*	*RB1, ARRB2*
Metabolism of xenobiotics by cytochrome P450	*MGST1*	*MGST1*	
Chemical carcinogenesis - DNA adducts	*MGST1*	*MGST1*	
Drug metabolism - other enzymes	*MGST1*	*MGST1*	
Drug metabolism - cytochrome P450	*MGST1*	*MGST1*	
Chemokine signaling pathway	*STAT5B, STAT3, CXCL16, CCR10*		*CXCL16, ARRB2*
Prolactin signaling pathway	*STAT5B, STAT3, STAT5A, TH*		
ErbB signaling pathway	*STAT5B, STAT5A*		
Cytokine-cytokine receptor interaction	*CD70, CXCL16, CCR10, GHR*		
Alcoholic liver disease		*TRA2B, SCD, LPIN1*	
Motor proteins		*TUBA4A, DYNLRB2, TUBA1D*	

Viral protein interaction with cytokine and cytokine receptor	*CCR10*		
Hedgehog signaling pathway		*PTCH1*	
Fatty acid metabolism		*HSD17B12, SCD*	
Steroid hormone biosynthesis		*HSD17B12*	
Proximal tubule bicarbonate reclamation			*SLC4A4, SLC9A3*
Bile secretion	*SLC4A4*		*SLC4A4, SLC9A3*
Hippo signaling pathway			*NKD2*
ECM-receptor interaction	*THBS3, LAMA5*		

1114

1115 **Table 2 Genetic variance explained by top and random variants for MY, FY, and PY**

Trait	V_{TOP}	SE_{TOP}	V_{RANDOM}	SE_{RANDOM}
MY	0.086677	0.012530	0.003195	0.002656
			0.015675	0.003461
			0.005690	0.002973
			0.005522	0.002872
			0.008892	0.003040
FY	0.070413	0.010478	0.003675	0.002497
			0.001819	0.002540
			0.004907	0.002939
			0.001645	0.002567
			0.003471	0.002572
PY	0.066613	0.009325	0.003456	0.002645

			0.007485 0.003309 0.001254 0.002862	0.002946 0.002646 0.002492 0.002649

1116 $\mathbf{V_{TOP}}$ = genetic variance explained by top genome-wide significant variants from autosomal chromosomes

1117 $\mathbf{SE_{TOP}}$ = standard error of top variants

1118 $\mathbf{V_{RANDOM}}$ = genetic variance explained by random 50 variants from all autosomal chromosome

1119 $\mathbf{SE_{RANDOM}}$ = standard error of random variants

1120

1121 **Table 3 New candidate genes for milk production traits**

TRAIT	CHR	GENE	FUNCTION
MY	2	*CDK5R2*	Described in the main text
	7	*ADGRE1*	Eight intergenic variants were identified between *VAV1* and *ADGRE1*. *ADGRE1* was found to be highly expressed in the macrophage cells in the lactating murine mammary gland [134]. It was detected in periparturient dairy cows' visceral adipose tissue, in a study by Michelotti et al. [135] that investigated differences between adipose tissue cells in their contribution to the development of metabolic diseases in cattle in the period before and after calving. Association analysis in pigs showed a significant association of this gene with eicosenoic acid content [136]
	9	*PRDM1*	Described in the main text
	16	*CASZ1*	24 intron variants were located within the *CASZ1* gene. In the differential methylation analysis in dairy goats [137] this gene was reported to be downregulated in the lactation period, relative to the dry period. Another study [138] reported copy number variation (CNVR) in the same gene to be connected with milk traits of local sheep breed. In cattle, it has been associated with longevity [139], however, this is the first time that this gene has been associated with milk traits in cattle
	19	*CAVIN1*	Two intergenic variants were located in the proximity of *CAVIN1*, a gene belonging to the group of cavin proteins, that play an important role in caveolae formation [140]. Caveolae are plasma membrane domains with a crucial role in lipid regulation in various cell types [141]. *CAVIN1*

			knockout mice lacked caveolae and exhibited various metabolic disorders including hyperlipidemia and glucose intolerance [142]
	19	*CCR10*	One variant upstream of the *CCR10* gene was found to be significantly associated with milk yield. Experiments on mice lacking *CCR10* [143] showed that *CCR10* is essential for efficient localization and accumulation of IgA antibody-secreting cells in lactating mammary glands. Interestingly, *CCR10* acts as a receptor for *CCL28* [144], known QTL for milk composition traits, lactation persistency [145, 146], fat and protein percentage, and milk yield [147]
	24	*RNF152*	The intergenic variant was located between *RNF152* and *PIGN*. In the study on transgenic mice, *RNF152* was downregulated during involution day 6 [148]. It was also described as a candidate gene for average daily gain and average daily feed intake in crossbred pigs [149], backfat thickness, and other production and growth-related traits in Korean Duroc pigs [150]
	25	*FBXL19*	Described in the main text
	25	*PAGR1*	One variant upstream of *PAGR1*, a gene that has an essential role in adipogenesis [151] was significantly associated with MY
FY	3	*STK25*	Described in the main text
	3	*DUSP12*	Two intergenic variants were positioned close to the *DUSP12* gene on BTA3. Previously, this gene was found to be upregulated in the liver of offspring of dams supplemented with essential fatty acids, compared to dams fed with saturated fatty acids [152]. In another study, *DUSP12* was proposed as a regulator of hepatic lipid metabolism [153], suggesting possible involvement with milk fat composition
	8	*PTCH1*	One intergenic variant was located between *PTCH1* and *ENSBTAG00000049821* on BTA8. *PTCH1* (Patched 1) regulates ductal morphogenesis in mammary epithelium and stroma, and ductal elongation and ovarian hormone responsiveness in the pituitary gland, as shown in the study of Moraes et al. [127]. It was also associated with body depth and strength in the Holstein bulls fine-mapping study [154]. In our case, it was enriched in the Hedgehog signaling pathway
	12	*LRCH1*	Two intron variants on BTA12 were associated with the *LRCH1* gene, which has a role in lipid regulation, including the promotion of lipopolysaccharide (LPS) binding and its delivery to lipid rafts [155]
	16	*KLHL12*	Described in the main text

	16	*LGR6*	On the same chromosome, one intron variant was found in *LGR6*, the gene that was found to be related to lactation in the study of Zhang et al. [156]. Blaas et al. [157] found *LGR6* to be involved with various functions in postnatal mammary gland development, making it a strong candidate for further research
	17	*MED13L*	On BTA17, one intergenic variant was found between *MED13L* and *ENSBTAG00000052624*. While functions of *ENSBTAG00000052624* haven't been described yet, *MED13L* was associated with milk yield and somatic cell score (SCS) in a dairy sheep [158], therefore indicating a similar role in other mammals
	23	*MCCD1*	One variant upstream of gene *MCCD1* was associated with FY, and although little is known about *MCCD1* function, this gene was previously associated with fat and protein percentage in dairy sheep GWAS [159] and with the regulation of fatty acid synthesis in patients with renal cancer [160]
PY	2	*CDK5R2*	Described in the MY section
	3	*FARP2*	On BTA3, 15 variants were found within or downstream of the *FARP2* gene. In multi-trait GWAS on body composition traits [161] *FARP2* was described to be associated with body composition traits and as being able to bind to phospholipids and cytoskeleton
	3	*STK25*	Described in the main text
	3	*CRCT1*	One intergenic variant on BTA3 was positioned between *ENSBTAG00000050431* and *CRCT1*. While little is known about *ENSBTAG00000050431*, *CRCT1* was previously described as a new candidate gene for body fat percentage in humans [162] and might have a role in developing mammary gland, as shown in cattle [163]
	4	*SUGCT*	Within the *SUGCT* gene, three intron variants were found. *SUGCT*-knockout mice exhibited an imbalance in lipid and acylcarnitine metabolism [164]
	7	*MIDN*	One intron variant was located in the *MIDN* gene which has a role in regulating cholesterol/lipid metabolism in the liver [165]
	9	*LIN28B*	The two variants on BTA9 were identified in or in proximity with *LIN28B*. *LIN28A* and its homolog *LIN28B* were reported to enhance *de novo* fatty acid synthesis and metabolic conversion of saturated to unsaturated fatty acids [166]
	9	*PRDM1*	Described in the main text

	9	*PREP*	Next, 14 variants were identified close to the *PREP* (prolyl endopeptidase) gene on BTA9. Previously, it was shown that *PREP* knockout mice exhibited changes in hepatic lipid metabolism [167]
	9	*CRYBG1*	One intron variant on BTA9 was identified on *CRYBG1*, a gene that was shown to participate in the regulation of fat-cell differentiation [168]
	12	*RB1*	Described in the main text
	16	*KDM5B*	Two intron variants on BTA16 were found within *KDM5B*, a gene that was identified as a regulator of lipid metabolism reprogramming in breast cancer cells [169]
	16	*KLHL12*	Described in the main text
	18	*CBLN1*	The intergenic variant was located in proximity to *CBLN1*, a member of the C1q family of proteins that has been reported to have a lipid-binding ability [170]
	23	*ZNF391*	On BTA23, two variants were close to the *ZNF391* gene, previously associated with the marbling score in Hanwoo beef cattle [171] and somatic cell count in dairy cattle [73]. In dairy sheep GWAS [172] this gene was connected with milk traits
	24	*RNF152*	Described in the MY section
	25	*FBXL19*	Described in the main text
	29	*STT3A*	Described in the main text

Additional files

Additional file 1

Format: .pdf

Additional file 1 Table S1

Title: Genotype arrays used for samples genotyping

Additional file 1 Table S2

Title: Composition of breeds of WGS reference panel

Additional file 1 Table S3

Title: Candidate genes associated with the top 50 variants for MY, FY, and PY

Additional file 1 Table S4

Title: Common genes between the three milk production traits

Additional file 1 Table S5

Title: Number of variant effects by type

Additional file 1 Table S6

Title: Functional Enrichment Analysis results for MY, FY, and PY

Additional file 1 Table S7

Title: List of known milk production and composition candidate genes identified in our study

Additional file 1 Figure S1

Format: .png

Title: Concordance between imputed KuhVision AF and 1000 Bulls Run9 AF from BTA16

Additional file 2

Format: .xlsx

Additional file 2 Table S1

Title: List of top variants for MY

Additional file 2 Table S2

Title: List of top variants for FY

Additional file 2 Table S3

Title: List of top variants for PY

Additional file 3

Format: .xlsx

Additional file 3 Table S1

Title: List of all genome-wide significant variants and their effects for MY

Additional file 3 Table S2

Title: List of all genome-wide significant variants and their effects for FY

Additional file 3 Table S3

Title: List of all genome-wide significant variants and their effects for PY

Additional file 4

Format: .xlsx

Additional file 4 Table S1

Title: Genes whose functions couldn't be linked with milk production traits