J. Dairy Sci. 102
https://doi.org/10.3168/jds.2018-15890

The use of mid-infrared spectra to map genes affecting milk composition

A. Benedet,[1] P. N. Ho,[2] R. Xiang,[3] S. Bolormaa,[2] M. De Marchi,[1] M. E. Goddard,[2,3] and J. E. Pryce[2,4]*
[1]Department of Agronomy, Food, Natural Resources, Animals and Environment, University of Padova, Legnaro 35020, Padova, Italy
[2]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, Victoria 3083, Australia
[3]Faculty of Veterinary & Agricultural Science, University of Melbourne, Victoria 3010, Australia
[4]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3083, Australia

ABSTRACT

The aim of this study was to investigate the feasibility of using mid-infrared (MIR) spectroscopy analysis of milk samples to increase the power and precision of genome-wide association studies (GWAS) for milk composition and to better distinguish linked quantitative trait loci (QTL). To achieve this goal, we analyzed phenotypic data of milk composition traits, related MIR spectra, and genotypic data comprising 626,777 SNP on 5,202 Holstein, Jersey, and crossbred cows. We performed a conventional GWAS on protein, lactose, fat, and fatty acid concentrations in milk, a GWAS on individual MIR wavenumbers, and a partial least squares regression (PLS), which is equivalent to a multi-trait GWAS, exploiting MIR data simultaneously to predict SNP genotypes. The PLS detected most of the QTL identified using single-trait GWAS, usually with a higher significance value, as well as previously undetected QTL for milk composition. Each QTL tends to have a different pattern of effects across the MIR spectrum and this explains the increased power. Because SNP tracking different QTL tend to have different patterns of effect, it was possible to distinguish closely linked QTL. Overall, the results of this study suggest that using MIR data through either GWAS or PLS analysis applied to genomic data can provide a powerful tool to distinguish milk composition QTL.
Key words: genome-wide association study, mid-infrared spectroscopy, milk trait, dairy cattle

INTRODUCTION

Genome-wide association studies (**GWAS**) have been widely used to identify SNP associated with variation in phenotypic traits in dairy cattle, presumably because they are in linkage disequilibrium (**LD**) with a QTL or causal variant for the trait. For instance, Pryce et al. (2010), Bouwman et al. (2011), and Buitenhuis et al. (2016) successfully identified important genome regions for milk production and composition traits. However, the power and precision of GWAS to identify the SNP closest to the causal variant is limited by the small effect size of most QTL (Hayes et al., 2009). Stringent significance tests are needed to protect against false positives from multiple testing, and external information independent from GWAS is needed to distinguish causal mutations with strong LD, which extends over megabases in cattle (de Roos et al., 2008) and limits the precision with which the QTL is mapped. The large number of small QTL implies that they are densely packed on chromosomes; this fact, combined with long-range LD, can make it difficult to identify the number of QTL in a genomic region. For example, many SNP within 1 Mb of the *DGAT1* gene on chromosome 14 (Chr14:1795425–1804838) have an association with milk fat percentage (e.g., Iso-Touru et al., 2016). This is likely because they are in LD with the causal variant in *DGAT1* but it could also be that other QTL in this region affect fat percentage. Multi-trait GWAS increases the power to detect QTL (Xiang et al., 2017), especially if the effects of a QTL across traits are different from that expected from the correlation between traits. Multi-trait GWAS analysis can also distinguish closely linked QTL if they have different patterns of effect across the traits.

Mid-infrared (**MIR**) spectroscopy is a useful tool to measure the concentration of many milk components, such as fat, protein, lactose, and fatty acids (De Marchi et al., 2011, 2014; Soyeurt et al., 2011) and thus to generate multi-trait data (De Marchi et al., 2014). Furthermore, the MIR absorption at each wavenumber can be considered a trait in its own right. These traits have been shown to be heritable, with estimates of heritability ranging between 0 and 0.63 (Soyeurt et al., 2010; Wang et al., 2016) and are affected differently by individual genes (Wang et al., 2016). Moreover, because

Received October 22, 2018.
Accepted April 12, 2019.
*Corresponding author: jennie.pryce@ecodev.vic.gov.au

of the low cost and routine use through milk recording, it is feasible to obtain MIR spectra on thousands of animals, which could provide significant power in association analysis.

The relationship between SNP and MIR data can be investigated by performing individual GWAS on single spectral wavenumbers, using SNP as predictors (Wang and Bovenhuis, 2018). However, a multi-trait GWAS of all wavenumbers simultaneously would be more powerful. Moreover, an almost equivalent but faster analysis is to use all wavenumbers to predict the genotype at each SNP (a "reverse GWAS"; Rutten et al., 2011). In this paper, we used partial least squares regression (**PLS**) to predict SNP genotypes from MIR spectral data on each cow; PLS was chosen because it is routinely used to predict phenotypes from MIR data (Geladi and Kowalski, 1986). The advantage of PLS is that it uses all wavenumbers simultaneously to predict SNPs and thus could significantly increase the power of identifying informative SNP markers in a limited computational time.

Therefore, the aim of this study was to investigate the feasibility of using MIR information to increase the power and precision of GWAS for milk composition and to distinguish linked QTL from a single QTL linked to multiple SNP. We first present a conventional GWAS on protein, fat, lactose, and fatty acid concentrations in milk; followed by a GWAS on absorption at individual MIR wavenumbers and a reverse multi-trait GWAS using all wavenumbers to predict a SNP genotype; and finally, an analysis to distinguish whether linked SNP are associated with the same or different QTL.

MATERIALS AND METHODS

In this study, multiple approaches were used to analyze phenotypic and genotypic data. First, single-trait GWAS were performed to test the associations between SNPs and milk composition traits or MIR wavenumbers. Second, PLS was applied to predict SNP genotypes using MIR wavenumber as predictors.

Animal Data

The data used in this study were obtained from 5,202 Holstein, Jersey, and crossbred cows between parity 1 and 6 from 20 commercial farms located in New South Wales, Victoria, and Tasmania (Australia) calving in spring 2017. Milk samples were taken 2 to 8 times per cow (4.7 on average) and sent to TasHerd Pty Ltd. (Hadspen, Tasmania, Australia) for prediction of fat, protein, and lactose contents using an infrared spectrometer (model 2000, Bentley Instruments, Chaska, MN). The corresponding spectra were stored for this study. A single spectrum includes 899 data points, with each point representing the absorption of infrared light through the milk sample at a particular wavelength in the 649 to 3,999 cm^{-1} region. Commercial prediction equations, obtained from Bentley Instruments and calibrated using data of Holstein cows with prediction accuracies (R^2) ranging between 0.74 and 0.96, were applied to these spectra to obtain individual fatty acids (expressed as g/dL of milk), including C4:0, C6:0, C8:0, C10:0, C12:0, C14:0, C16:0, C17:0, C18:0, C18:1 *cis*-9, and C20:0. The original data set comprised 24,655 spectra and milk composition traits, in which the mean and standard deviation (SD) of DIM were 146 and 121, respectively. The original data set was edited and outliers were identified and excluded as milk composition traits being greater than means ± 3.5 SD. Then, phenotype records (spectra and milk composition traits) were averaged, resulting in only one observation per cow; this reduced within-cow errors and eliminated the need to fit DIM as a fixed effect in statistical models. It was not possible to include parity as a fixed effect in statistical models because this information was not readily accessible. Several mathematical treatments were applied to milk spectra. Noisy areas induced by water absorption (1,600 to 1,689 and 3,010 to 3,998 cm^{-1}) were first removed (Hewavitharana and van Brakel, 1997). Spectra with a global distance value >3 were considered outliers and excluded (Shenk and Westerhaus 1995). First derivative was applied to the reduced spectra to increase peak resolution. After editing, a reduced spectrum of 598 wavenumbers per sample was used for the analysis.

Genotypes

All cows were genotyped using a BovineSNP50 BeadChip and then imputed to BovineHD (800K) BeadChip (Illumina Inc., San Diego, CA), following the procedures previously described by Kemper et al. (2015). The genotype data were refined according to a range of quality control filters (Erbe et al., 2012; Kemper et al., 2015). After quality check, 626,777 SNP remained for the analysis. The genotypes for each SNP were encoded in the top/top Illumina A/B format and then genotypes were reduced to 0, 1, and 2 copies of the B allele. All SNPs were mapped to the UMD 3.1 build of the bovine genome sequence (https://www.ensembl.org/biomart).

Single-Trait GWAS

The regression model used in the GWAS to test the association of each SNP with each milk composition and wavenumber trait was as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{Wa} + \mathbf{e}, \quad [1]$$

where $\mathbf{y}$ is the vector of phenotypic records of q individuals; $\boldsymbol{\beta}$ is the vector of fixed effects of breed (Holstein, Jersey, crossbred) and herd (1 to 20); $\mathbf{X}$ is a design matrix relating phenotypes to their fixed effects; $\mathbf{u}$ is the vector of animal effects, where $\mathbf{u} \sim N\left(0, \mathbf{G}\sigma_g^2\right)$, $\mathbf{G}$ is the $q \times q$ genomic relationship matrix (**GRM**, based on high-density genotypes) between pairs of individuals, and σ_g^2 is the additive genetic variance; $\mathbf{Z}$ is the incidence matrix; $\mathbf{W}$ is the vector of animal genotypes at SNP_i coded as 0, 1, or 2 (representing the genotypes aa, Aa, or AA) and $\mathbf{a}$ is the effect of the SNP; $\mathbf{e}$ is the vector of residual errors. The GWAS was conducted using GCTA software (Yang et al., 2011). Associations with a $-\log_{10}(P) \geq 5$ were considered statistically significant, corresponding to a maximum false discovery rate of 0.10. Figure 1 shows the quantile-quantile (Q-Q) plots of GWAS results for some specific phenotypes. We observed strong skews in the Q-Q plots, which indicates that more SNP were associated with our phenotypes than expected by chance.

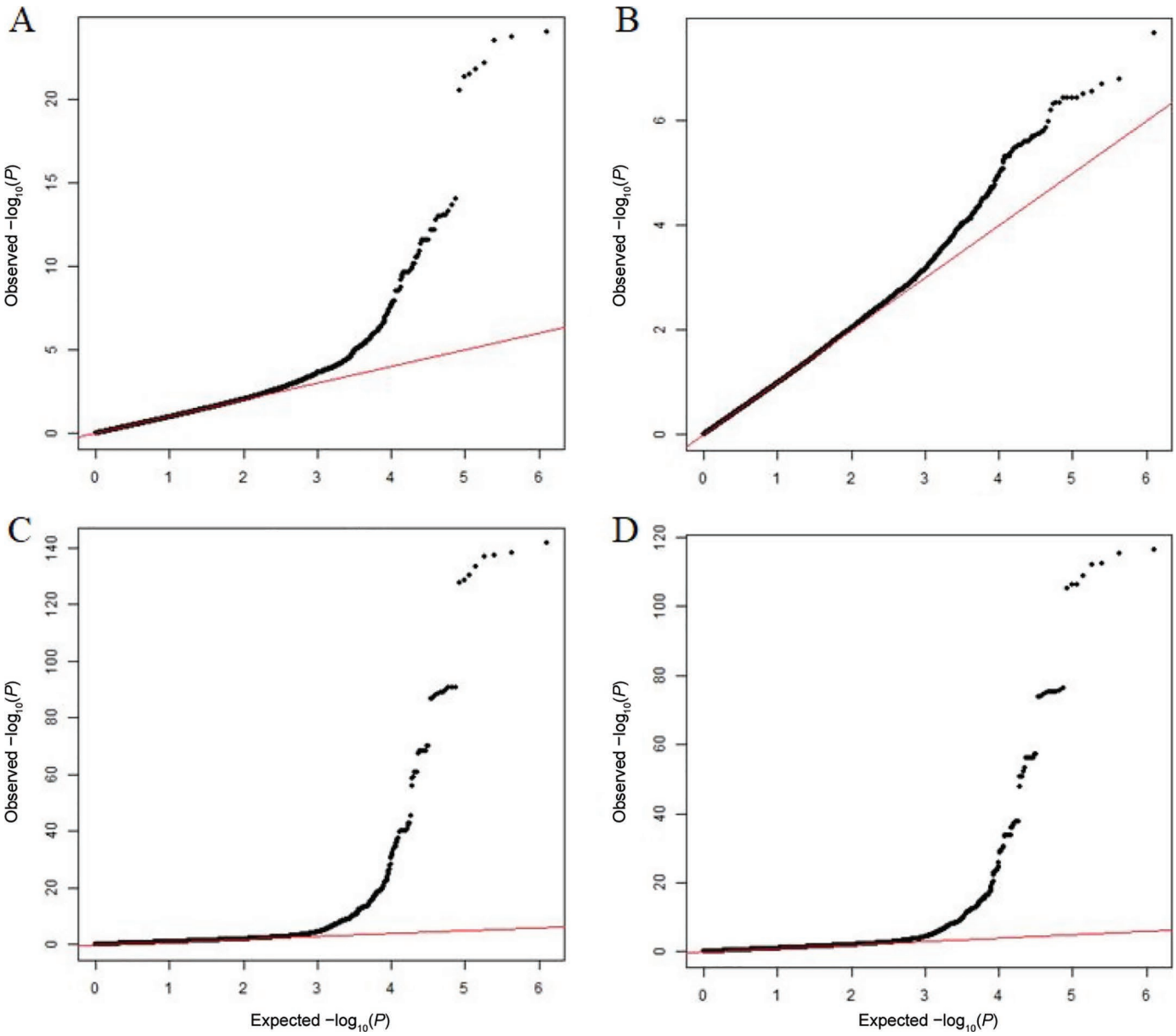

Figure 1. Quantile-quantile (Q-Q) plots of genome-wide association study (GWAS) results of milk protein (A), lactose (B), and fat (C) percentage, and a milk fatty acid (C4:0; D) used as example.

PLS Analysis

The prediction models for SNP genotypes were developed using PLS and implemented in R with the pls package of Mevik and Wehrens (2007), through a 10-fold cross-validation. Partial least squares is similar to a principal component analysis where the data set is transformed into a new projection that represents the entire data set and the most informative components in the new projection are features of the transformed data set. However, PLS considers the dependent variable when constructing its projection, whereas principal component analysis does not (Hempstalk et al., 2015).

The SNP genotypes and MIR spectra were pre-adjusted for the fixed effects of breed and herd using multiple linear regression. Breed (Holstein, Jersey, crossbred) and herd (1 to 20) were considered as class variables in the linear model performed through the R stats package. The residuals from this analysis for SNP and MIR spectra were used as response and explanatory variables, respectively, to calibrate the PLS models. The multiple linear regression adjustment aimed to remove the same possible confounding effects for both response (SNP genotypes) and explanatory (MIR spectra) variables in the PLS analysis. The accuracy measures of the PLS models (developed through a 10-fold cross validation) included the coefficient of determination ($\mathbf{R^2}$) and root mean square error. The optimal number of PLS components was determined based on first local minimum value in root mean squared error.

To investigate the potential contribution of MIR in predicting SNP genotypes, the PLS models were tested in 2 ways. In the first model, only MIR information was used as explanatory variables. The second model included MIR in addition to the first 20 principal components (**PC**; explaining approximately 80% of the total variance) of the GRM of the animals in this study, as explanatory variables. In the last model, only the first 20 PC of GRM were considered. The R^2 of MIR corrected for PC of the GRM ($\mathbf{R^2_{MIRcor}}$) was calculated as the difference between R^2 of the PLS model using MIR and GRM as explanatory variables and R^2 of SNP predicted through only GRM, divided by $1 - R^2$ of SNP predicted through the GRM.

To compare GWAS and PLS results, *P*-values were computed for each R^2 through a chi-squared approximation of the *F*-test in ANOVA, using the following formula: $N\ (R^2/1 - R^2) \sim \chi^2$, where N is the sample size (5,202 cows) of the study, and R^2 is the coefficient of determination for each SNP prediction model. The same approach was used for each R^2_{MIRcor}. Values of $R^2 \geq 0.02$ corresponding to $-\log_{10}(P) \geq 24$ were considered in defining significant peaks in PLS results.

The PLS and linear regression analyses described above were performed using R statistical software version 3.4.3 (R Core Team, 2017).

RESULTS

GWAS on Milk Composition

The GWAS results for protein, lactose, and fat percentages are presented in Figure 2 and 3. The most important SNP [i.e., that were $-\log_{10}(P) \geq 5$] for each trait are reported in Table 1. Because not all the SNP have reference identifications, we have provided SNP position coordinates of the bovine genome (UMD 3.1; https://www.ensembl.org/biomart) to refer to the SNP throughout the text.

Genome-Wide Association Comparison Between Fat Content and Fatty Acids

Figure 3 shows the GWAS results for milk fatty acids and fat percentage. Because the peaks on BTA14 tend to dominate the Manhattan plots, we focused on the bottom of each plot by truncating the y-axis to a maximum value of $-\log_{10}(P)$ at around 20, to highlight the similarities and differences between traits. The most significant SNPs in each peak for each milk fatty acid are listed in Table 1. The observed QTL fell into at least 3 groups. The first group included QTL that affected total fat percentage and the concentration of all fatty acids, such as SNP on BTA5 and BTA14. However, these QTL affected all fatty acids concentrations but not in the same direction. The alleles of *DGAT1* on BTA14 and *MGST1* on BTA5 that increased the concentration of fatty acids synthesized in the mammary gland (de novo fatty acids), decreased the concentrations of long-chain fatty acids. A second group of QTL affected total fat percentage and some fatty acids in the same direction. For example, a SNP at 103.8 Mb on BTA8 and a peak around 31.2–31.9 Mb on BTA20 affected fat percentage and C17:0 or C6:0–C17:0 and C18:1 in the same direction, respectively. This pattern of effects might be caused by an increase in milk volume that dilutes all components (e.g., by *GHR* on BTA20). A third group of QTL affected specific fatty acids. For instance, SNP near *FASN* (51.2–51.3 Mb on BTA19) were associated with C10:0, C12:0, and C14:0, and others near *PAEP* (103.2–103.3 Mb on BTA11) affected the concentration of C18:1.

GWAS on Mid-Infrared Wavenumbers

The most significant SNP for each milk trait were investigated for their associations with MIR spectra.

These SNP tended to be significant for multiple wavenumbers. Consequently, for each SNP, we could draw a profile of their significant effects across spectra (Figures 4, 5 and 6).

The SNP varied widely in their pattern of significance across the spectrum. Figure 4 shows the $-\log_{10}(P)$ of the 598 MIR wavenumbers of 3 different SNP, each associated with different milk composition traits. The first SNP (Chr6:87391848, Figure 4A) was significantly associated with protein percentage and was near *CSN3*. It had significant effects on 66 wavenumbers distributed in 8 peaks across the spectrum (Figure 4A). Significant wavenumbers were observed between 1,212 and 1,387 cm^{-1} and from 1,458 to 1,700 cm^{-1}. Although they had lower significance than those described above, additional wavenumbers were observed in the following part of the spectrum corresponding to 1,876 to 1,898 cm^{-1}, 2,487 to 2,506 cm^{-1}, 2,599 to 2,618 cm^{-1}, 2,857 cm^{-1}, and 2,924 cm^{-1}. The second SNP (Chr28:6491786, Figure 4B) was associated with lactose percentage and was near *KCNK1*. It had a significant effect on few wavenumbers, ranging from 995 to 1,033 cm^{-1} and from 1,163 to 1,167 cm^{-1} (Figure 4B). The third SNP (Chr11:103308330, Figure 4C) was associated with milk fat composition, specifically with C18:1, and close to *PAEP*. It significantly affected 30 wavenumbers, corresponding to the following regions: 958 to 969 cm^{-1}, 1,283 to 1,294 cm^{-1}, and 1,365 to 1,488 cm^{-1} (Figure 4C). These 3 SNP had quite different patterns of significant wavenumbers, which might be expected as they affected different traits.

Figure 5 shows 2 SNP both associated with lactose percentage (Chr28:6491786 and Chr19:61238366). Although they shared a significant spectral region between 995 and 1,029 cm^{-1}, these SNP showed different patterns for the remaining wavenumbers. Conversely, a completely different situation is represented by SNP depicted in Figure 6. The first group of SNP (Figure 6A) were all near *DGAT1*. They had very similar effects on milk composition traits (Table 1) and across the MIR spectrum (Figure 6A). Figure 6B depicts 2 other SNP 3 Mb apart (Chr14:66328304 and Chr14:69890969) but also showing similar effects on milk traits (Table 1) and on the spectrum of wavenumbers (Figure 6B).

SNP Prediction Using PLS

The results described above show that different QTL have different patterns of effect across the MIR spectrum, suggesting that a multi-trait analysis, using all wavenumbers, is likely to be a powerful way to detect QTL affecting milk composition. We examined this proposition by what we describe as a "reverse GWAS," in which we used the MIR data to predict SNP genotype by a PLS analysis. First, we tested whether SNP that affected conventional milk traits could be detected by this new analysis. Table 1, which contains significant SNP for conventional milk traits, also presents

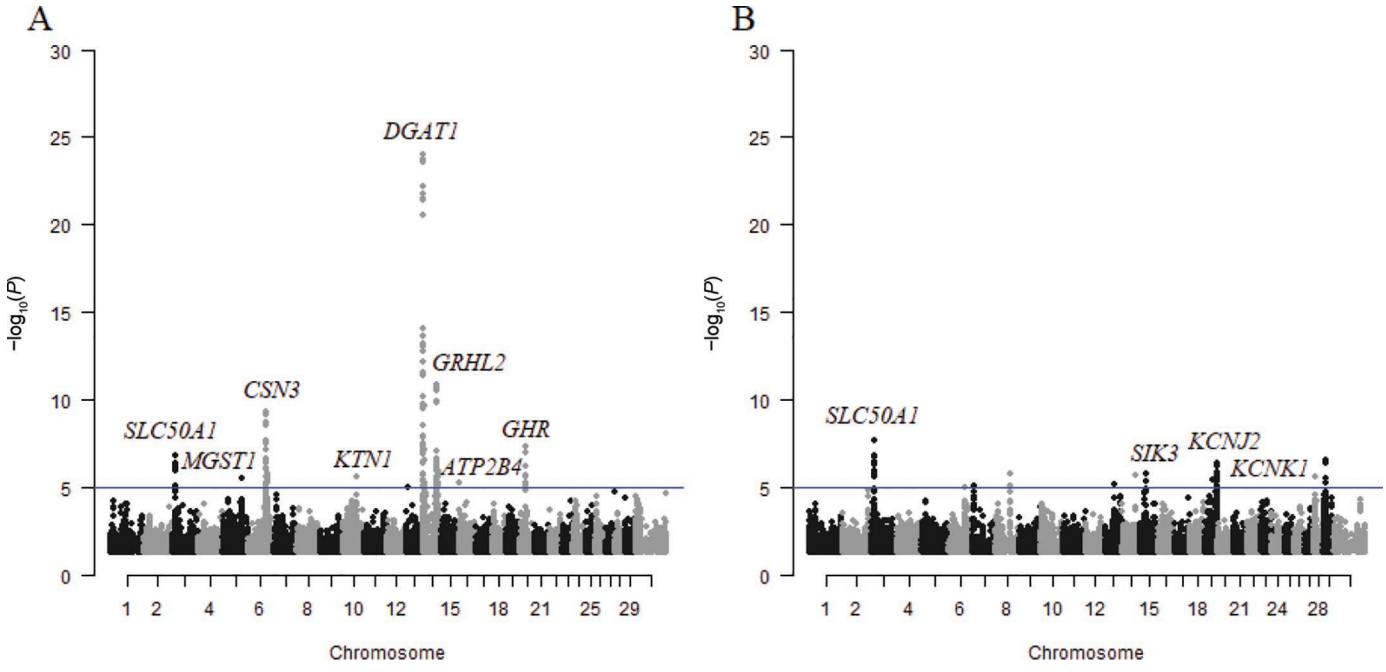

Figure 2. Manhattan plots of genome-wide association study (GWAS) results of milk protein (A) and lactose (B) percentage using SNPs with $P < 0.05$. The horizontal line is $-\log_{10}(P) \geq 5$. Genes close to identified peaks are highlighted.

R^2 values of each SNP obtained through PLS analysis. To compare PLS and the conventional GWAS, the $-\log_{10}(P)$ of each R^2 was computed. This showed that the PLS analysis significantly [$-\log_{10}(P) \geq 5$] predicted all but 4 of the SNP in Table 1. As shown in Figure 7, 7,417 SNPs had $-\log_{10}(P) \geq 24$, highlighting very high and clear significance peaks in specific genome regions (e.g., on BTA6, BTA11, and BTA14). The prediction accuracy of PLS models using only MIR spectra decreased slightly when corrected for population structure through addition of the first 20 PC of GRM (R^2_{MIRcor}) to the model (Table 1). For the SNP in Table 1, the R^2 decreased from 6.5 to 5.4% as a result of adding the PC to the model. Also, the ranking of SNP in Table 1 using $-\log_{10}(P)$ before and after correction using PC from the GRM was similar. Nevertheless, many SNP showed much more significant results in the PLS than in the conventional GWAS. An example of the direct comparison of the significance of effects between the GWAS and PLS is shown in Figure 8. On BTA29, the GWAS results for lactose percentage appeared to be quite flat and not easily distinguishable (Figure 8A). In contrast, stronger differences between SNP were observed in PLS outcomes (Figure 8B). Those SNP significant in the PLS, but not in Table 1, are listed in Table 2. Some SNP close to the genes *LALBA*, *AGPAT6*, and *P2RX4* were identified on BTA5, BTA27, and BTA17, respectively.

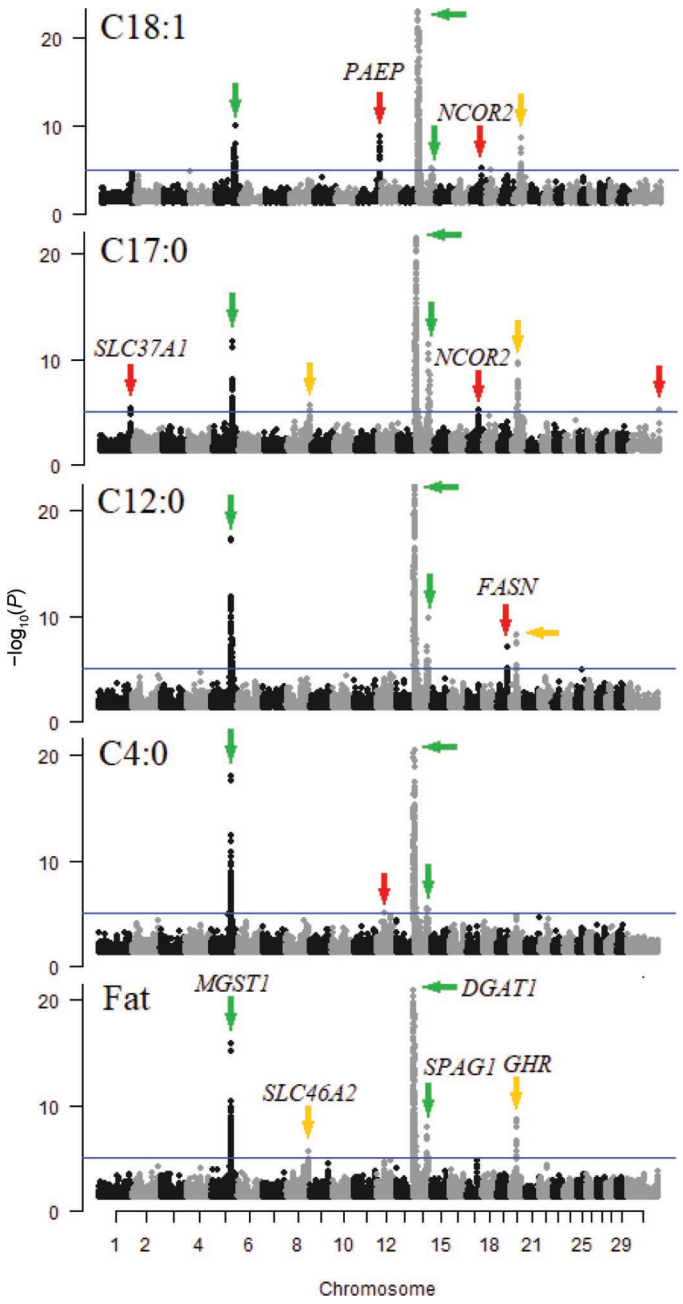

Figure 3. Manhattan plots of genome-wide association study (GWAS) results of fat percentage and milk fatty acids using SNP with $P < 0.05$. The horizontal line is $-\log_{10}(P) \geq 5$. A focus on the bottom of each plot around $-\log_{10}(P) = 20$ is considered to provide a clearer comparison between traits. Arrows of different colors are used to identify similarities and differences between traits: green for common peaks, yellow for common peaks in only a few traits, and red for peaks associated with specific traits.

Next, we investigated the ability of the combination of GWAS results on MIR wavenumbers and PLS results to distinguish whether different SNP were associated with different, or the same, QTL. Table 3 presents results from 4 genome regions, where it was unclear whether one or more QTL occur. If 2 SNP track the same QTL, we expect them both to be in LD with the QTL and therefore most likely in LD with each other. Also, if they track the same QTL, we expect that their effects across the 598 wavenumbers are correlated. For instance, on BTA20, there were 3 SNP from 31.4 to 34.5 Mb. The first 2 (Chr20:31228912 and Chr20:31909478) were in moderately high LD [LD squared coefficient (ρ^2) = 0.49] and their effects across wavenumbers were almost identical, with a coefficient of correlation (r) of 0.996 (Table 3 and Figure 9). However, their LD ρ^2 was close to zero with the third SNP (Chr20:34522480) and the correlations of their effects were 0.872 and 0.864. This suggests that the first 2 SNP track the same QTL but the third SNP (at 34.5 Mb) tracks a different QTL. We tested this conclusion as follows. We used PLS to predict one SNP genotype using another SNP, the MIR data, or both. In this case, the SNP at Chr20:31909478 predicted the genotype of the SNP at Chr20:31228912 and Chr20:34522480, as would be expected from the LD (this R^2 is similar but not equal to the LD ρ^2 because the prediction was tested in the same cross-validation procedure as used for PLS with MIR spectra). Including MIR information in the model did not bring any improvement to the prediction accuracy

Table 1. Significance [$-\log_{10}(P)$] of the most important SNP from peaks significantly [$-\log_{10}(P) \geq 5$] associated with each milk composition trait in genome-wide association studies (GWAS), and R^2 (without or with correction for the genomic relationship matrix, R^2_{MIRcor}) and their significance values for each SNP in partial least squares regression (PLS) analysis[1]

		GWAS[3]													PLS			
SNP	Rsnp ID[2]	F%	P%	L%	C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C17:0	C18:0	C20:0	C18:1	R^2	$-\log_{10}(P)$[4]	R^2_{MIRcor}	$-\log_{10}(P)$[4]
Chr1:144414210	rs109445875											5.38			4.21	0.152	203.89	0.132	172.82
Chr2:17975470	rs134284679	2.13			2.00	1.96	1.99	1.76	1.49	1.55		2.38	5.31	3.68	4.37	0.005	6.34	−0.009	0
Chr3:15518228	—	1.46	6.40	7.67					1.69		3.70					0.035	41.89	0.026	30.77
Chr3:15657861	—	1.48	6.83	5.80					1.89		3.19					0.037	44.82	0.020	23.85
Chr5:81976686	rs109452675	4.18			4.88	5.08	4.93	4.53	4.21	4.33	3.18	1.95	3.12	3.23	2.54	0.007	8.49	−0.001	0
Chr5:93945655	rs134637616	15.14	2.69		18.05	18.55	19.67	19.41	17.34	17.74	13.03	11.02	13.02	17.30	10.07	0.051	61.69	0.031	37.36
Chr5:93945738	—	15.84	5.53		17.55	18.09	19.25	18.83	17.15	16.75	13.97	11.67	11.56	15.31	10.00	0.052	63.40	0.030	35.89
Chr6:87391848	rs109122729		9.30													0.318	—[4]	0.253	—
Chr7:11216500	rs109096668			5.09												0.005	6.80	−0.005	0
Chr8:69093157	rs109160712	1.94	2.43	5.79	1.77	1.85	2.15	2.00	2.51	2.06	2.75					0.006	8.16	−0.002	0
Chr8:103829659	rs109955043	5.56	2.66		3.49	3.29	3.17	3.10	3.44	3.91	5.00	5.64	2.77	3.43	3.34	0.008	9.49	−0.002	0
Chr10:68317330	rs133252628	2.45	5.61					1.31	1.84	1.75	2.34	3.23			1.42	0.004	4.91	−0.006	0
Chr11:103308330	rs109087963	2.58			1.56	1.94	2.85	1.82		4.03	1.70	1.56	1.60	1.41	8.74	0.663	—	0.635	—
Chr12:35353044	rs136314838	4.58	2.45		5.04	4.92	4.76	4.41	3.76	4.32	4.17	3.15	2.55	3.93	2.46	0.002	2.87	−0.004	0
Chr13:12855563	rs109914643		5.03									1.50				0.001	2.09	−0.007	0
Chr13:44868669	rs41692389			5.21												0.017	20.03	0.001	0.78
Chr14:1724688	—	138.06	21.48		115.41	129.09	138.48	147.29	141.63	137.95	98.93	123.42	118.70	130.74	97.81	0.305	—	0.304	—
Chr14:1765055	—	128.39	23.99		105.20	117.57	126.33	135.28	132.15	128.11	94.31	112.74	101.06	112.22	86.30	0.280	—	0.272	—
Chr14:1801116	rs109421300	141.59	23.74		116.28	130.24	140.48	149.96	145.07	140.86	101.45	126.07	118.04	130.13	98.27	0.304	—	0.305	—
Chr14:64004106	rs134918489	4.84			5.44	5.63	6.05	5.47	5.42	5.40	4.36	3.73	2.85	3.00	3.93	0.014	17.09	0.002	2.37
Chr14:65033839	—	7.01	9.89	2.75	4.09	5.25	5.99	5.45	6.61	5.22	7.33	11.35	1.80	2.41	5.25	0.068	83.95	0.046	55.32
Chr14:65083236	—	6.78	10.87	3.77	4.19	5.24	6.38	5.66	7.11	5.60	8.26	9.92		1.63	3.47	0.074	92.30	0.044	53.52
Chr14:65376465	rs134937118	5.11	10.51	5.69	3.85	4.81	5.69	5.44	6.66	4.92	6.49	5.32			2.23	0.049	59.15	0.030	36.57
Chr14:65385787	rs134206557	5.11	10.51	5.69	3.85	4.81	5.69	5.44	6.66	4.92	6.49	5.32			2.23	0.048	58.22	0.032	38.57
Chr14:66328304	—	7.94	10.68	2.87	5.36	6.83	8.54	8.07	9.76	7.42	8.85	10.54	1.64	2.56	4.00	0.075	93.28	0.049	59.87
Chr14:69890969	rs109408130	5.06	5.84		2.51	3.37	4.69	4.36	5.00	3.75	3.97	8.45	2.53	2.66	5.08	0.040	48.85	0.033	39.77
Chr15:27964533	rs135375350		3.22	5.75								1.75				0.017	20.10	0.014	17.49
Chr15:74829183	rs42364316	2.71			3.43	3.15	3.16	2.96	2.41	2.99	2.38	2.99	4.08	5.37	2.37	0.001	2.04	−0.007	0
Chr16:1301628	rs136372327	2.84	5.26						1.89	2.00	3.08	1.97				0.009	10.76	−0.001	0
Chr17:53620581	rs136686539	3.66			1.35	1.82	1.99	2.87	3.53	3.12	1.94	5.16	3.85	3.10	5.12	0.004	5.21	0.001	1.09
Chr18:23510437	rs110846786	3.41			2.46	2.30	2.25	2.22	1.99	2.22	2.11	4.59	3.03	3.13	5.09	0.005	6.46	−0.003	0
Chr18:23511080	rs109605785	3.41			2.46	2.30	2.25	2.22	1.99	2.22	2.11	4.59	3.03	3.13	5.09	0.005	7.02	0.001	1.32
Chr19:42961399	—		2.01	5.45								1.38	2.65	2.51		0.031	37.31	0.017	20.66
Chr19:51386735	rs137372738	3.12			3.25	3.63	4.47	6.54	7.10	7.19	2.39	1.62	1.80	2.09		0.065	80.13	0.034	40.42
Chr19:61238366	rs134657867			6.35							1.58					0.030	36.26	0.010	12.40
Chr20:31228912	rs133057950	8.39	6.99		4.31	5.25	6.02	6.73	7.54	7.04	6.45	9.58	3.23	3.68	7.37	0.016	19.22	0.011	14.10
Chr20:31281218	—	8.65	7.37		4.49	5.37	6.03	6.66	7.41	7.13	7.13	9.58	2.73	3.42	6.92	0.014	17.65	0.011	13.77
Chr20:31909478	—	7.87	7.01		4.70	6.15	7.16	7.53	8.18	6.83	5.30	7.99	3.03	2.99	8.62	0.029	35.53	0.026	31.40
Chr20:58219913	rs136723088	3.15			2.39	2.29	3.02	2.69	2.20	2.77	2.89	4.24	3.46	5.23	3.06	0.065	80.38	0.040	48.17
Chr28:6491786	rs109994328			5.60												0.023	28.17	0.006	8.06
Chr28:16400073	rs133255644	1.30		2.50	1.75	1.78	2.23	2.46	1.93	1.54		2.23	5.10	5.07	1.83	0.004	5.09	−0.001	0
Chr29:9543947	rs42162236			6.57												0.015	18.25	−0.003	0
Chr29:14597080	rs43629620			5.53												0.001	1.53	−0.001	0
Chr30:148276690	rs136716877	3.69	2.40	1.53	2.40	2.41	2.67	2.44	2.54	3.19	4.35	5.23			3.69	0.001	1.62	−0.010	0

[1]A color-coded version of this table is available as Supplemental Table S1 (https://doi.org/10.3168/jds.2018-15890), in which cells are colored to highlight differences between significance values, from the lowest value (green) to the highest value (red); white or empty cells indicate $P > 0.05$.

[2]Rsnp ID = reference SNP identification (https://www.ncbi.nlm.nih.gov/snp).

[3]F% = fat percentage; P% = protein percentage; L% = lactose percentage; C4:0 to C20:0 are expressed in g/dL of milk.

[4]Dash indicates significance of $P = 0$.

of the first SNP but it increased the R^2 of the second SNP. A possible explanation for these results may be that the 2 SNP at 31.4 Mb were tracking the same QTL, but the SNP at 34.5 Mb was tracking a different QTL. The other results in Table 2 support 2 QTL on each region investigated.

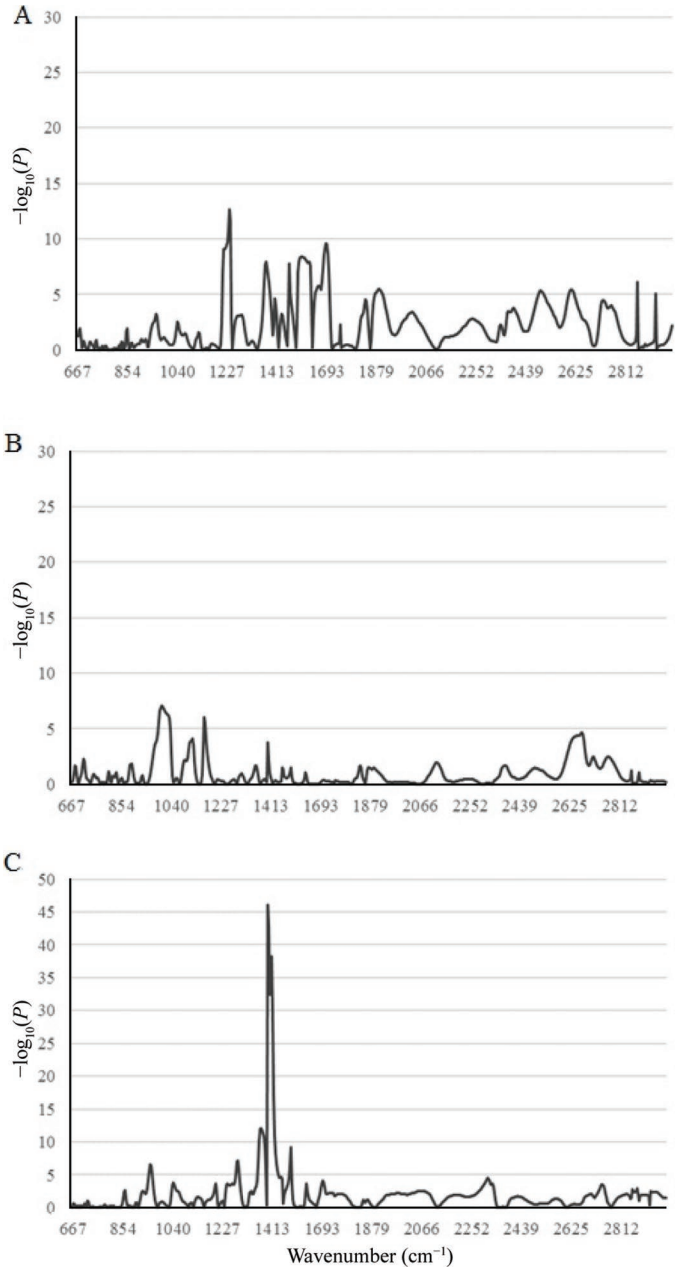

Figure 4. Significance [$-\log_{10}(P)$] pattern of the effect of selected SNP affecting protein percentage (Chr6:87391848; A), lactose percentage (Chr28:6491786; B), and fat chains (Chr11:103308330; C), on infrared wavenumbers in genome-wide association studies.

DISCUSSION

The aim of this study was to test whether MIR spectral data can be used as a tool to improve the power to detect and precisely map QTL associated with milk composition traits. Combining breeds had 2 benefits: it increased sample size and therefore power to detect associations and it decreased LD, increasing mapping precision. Using multi-breed data, we searched for SNP markers close to and in LD with QTL in all breeds. This approach broke down the long-distance LD within a breed, finding SNP that are close to the QTL. Most of the highly significant SNP were segregating in both Holsteins and Jerseys. In the first step of the study, GWAS results on milk composition traits were analyzed. Then, GWAS on single MIR wavenumbers and PLS using all wavenumbers to predict each SNP were performed to investigate differences and similarities between SNP. Finally, the feasibility of combining GWAS and PLS results to distinguish whether linked SNP were associated with the same or different QTL was investigated. Our results showed that PLS combining information from many wavenumbers had more power than single-trait GWAS in identifying informative SNP. Also, the combination of GWAS and PLS results provided information to distinguish QTL that are close together on the chromosome.

GWAS on Milk Composition

Many of the QTL described here have been previously reported and associated with candidate genes, such as *DGAT1*, *MGST1*, and *GHR* for fat and protein percentages (Grisart et al., 2002; Raven et al., 2014; Littlejohn et al., 2016; Nayeri and Stothard, 2016). The QTL on BTA3 at 15 Mb has been previously associated

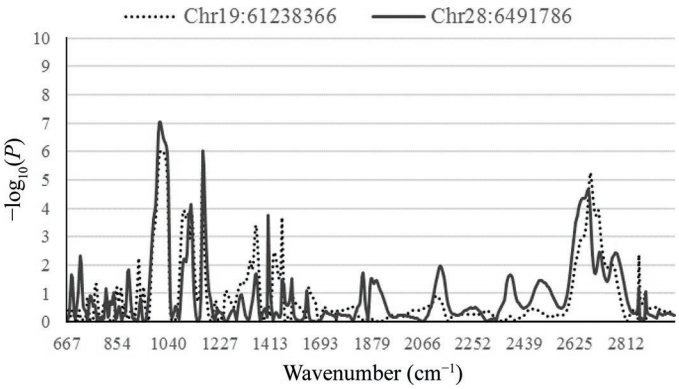

Figure 5. Significance [$-\log_{10}(P)$] pattern of the effect of 2 SNP (Chr19:61238366 and Chr28:6491786) affecting lactose percentage on infrared wavenumbers in genome-wide association studies.

with milk production traits (Xiang et al., 2017) but we found that it specifically affected protein and lactose percentages. The gene *SLC50A1*, which encodes a sugar transporter, is related to lactation in other mammal species (Lopdell et al., 2017) and had several significant SNP nearby. Other authors have suggested *MUC1* as a candidate for this QTL (Nayeri and Stothard, 2016; Raven et al., 2016; Xiang et al., 2017).

Overall, there have been fewer reports of GWAS for lactose percentage than for fat and protein contents. Lactose percentage varies less than fat and protein percentages because lactose is the main osmole regulator in milk (Fox et al., 2015). Consequently, there are limited physiological mechanisms by which it can be altered. One mechanism is to alter the concentration of other osmole regulators such as ions (Fox et al., 2015). We found significant SNP for lactose percentage close to the genes *KCNJ2* (19.6 Mb on BTA19) and *KCNK1*, both of which transport potassium across cell membranes. The *KCNJ2* gene was described as a modulator of milk lactose content, through its function in potassium ion transport in the membranes of secretory cells in mammary glands (Lopdell et al., 2017). We also found significant SNP for lactose percentage near genes involved in lactose synthesis or secretion, including *STAT5B* (42.9 Mb on BTA19) and *SLC50A1* mentioned above. Nayeri and Stothard (2016) reported that *STAT5B* contributes to mammary development, involution, and prolactin signaling pathways, which could explain the significant effects on lactose percentage. Raven et al. (2016) observed a significant relationship between a SNP near *STAT5B* and protein percentage, for which only a tendency was detected in this study. However, any mutation that increases lactose synthesis is expected to increase milk volume and therefore decrease the concentrations of protein and other components (Lopdell et al., 2017).

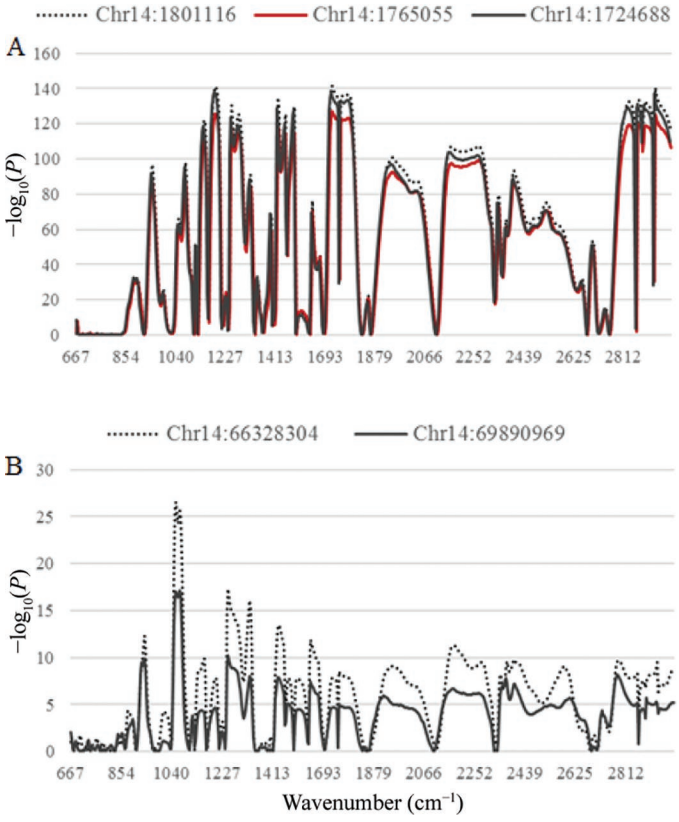

Figure 6. Significance [$-\log_{10}(P)$] pattern of the effect of SNP identified on BTA14 (Chr14:1724688, Chr14:1765055, and Chr14:1801116 in A; Chr14:66328304 and Chr14:69890969 in B) on infrared wavenumbers in genome-wide association studies.

Genome-Wide Association Comparison Between Fat Content and Fatty Acids

The concentration of a fatty acid in milk can be altered by a change in the concentration of all fatty acids or by a change in the proportion of fat made up by each fatty acid (in this study, correlation coefficients between milk fat and fatty acid content ranged from 0.72 to 0.94, with C18:0 and C14:0, respectively). For instance, an allele of the SNP near *GHR* increased milk volume (Blott et al., 2003) and so decreased concentrations of milk components. In contrast, on BTA5 and BTA14, respectively, *MGST1* and *DGAT1* were observed to affect fat content and composition traits. To better investigate their associations with milk fat composition, ratios between milk fatty acids and fat content were analyzed through GWAS (data not shown). Individual alleles of SNP near *MGST1* and *DGAT1* increased the proportion of fat containing de novo fatty acids (C4:0–C16:0) but decreased the proportion of long-chain fatty acids, such as C17:0 and C18:1. Regarding *MGST1*, although its role as a QTL for fat percentage is recognized, its mode of action on milk lipid synthesis and secretion is still unknown (Littlejohn et al., 2016). The enzyme encoded by *DGAT1* is involved in triacylglycerol synthesis by catalyzing the acyl-CoA esterification to diacylglycerol (Buitenhuis et al., 2014). The *DGAT1* gene has been widely studied, showing highly significant associations with several fatty acids (Schennink et al., 2007; Conte et al., 2010; Bouwman et al., 2011). In previous studies (Schennink et al., 2007; Bouwman et al., 2011), the *DGAT1* 232K allele was associated with increased proportions of C6:0, C8:0, C14:0, and C16:0 and a decrease in the fraction of C18:1, confirming our findings.

A significant QTL for medium-chain fatty acids up to C14:0 was identified on BTA19, near the gene *FASN*. Significant effects of *FASN* on this group of fatty acids were confirmed when the proportion between fatty ac-

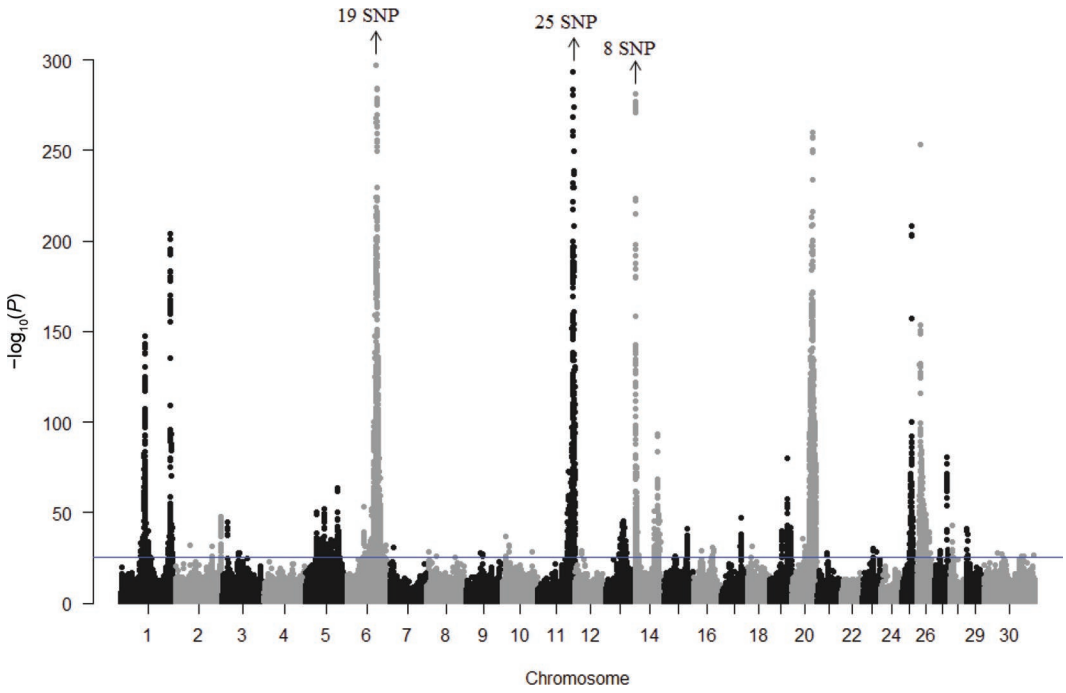

Figure 7. Manhattan plot of partial least squares regression results. The horizontal line is $-\log_{10}(P) \geq 24$, corresponding to $R^2 \geq 0.02$.

ids and fat content was analyzed. These findings were consistent with Stoop et al. (2009) and Bouwman et al. (2011), who also reported the presence of a QTL on BTA19 for medium-chain and de novo fatty acids. This pattern of effects can be explained by the fact that the protein encoded by *FASN* (fatty acid synthase) catalyzes synthesis of fatty acids and so most affects those that are synthesized in the mammary gland (i.e., up to chain length C14).

We found an association between SNP near *PAEP* and C18:1. Although *PAEP* has been identified as a QTL for fat percentage (MacLeod et al., 2016), an association with C18:1 has not previously been described. Previously, a QTL close to *PAEP* associated with C4:0, a trait negatively correlated with C18:1 (Stoop et al., 2008), was reported in Knutsen et al. (2018). As C18:1 in milk arises partly from the mobilization of cow body reserves (Nogalski et al., 2012), the implications for animal health and welfare of including *PAEP* markers in genomic selection should be considered.

Many additional significant SNP not cited above were identified along the genome for short-chain (BTA12) and long-chain (BTA1, BTA2, BTA15, BTA17, BTA18, BTA20, BTA30) fatty acids. Although these SNP affected single or multiple fatty acids (Table 1) within the same group, to our knowledge, they were in regions not previously associated with milk fat composition in the literature.

The MIR prediction equations that we used were calibrated on Holstein data. They may be less accurate when applied to Jersey and crossbred milk samples (Eskildsen et al., 2014). If so, this would reduce the power of the study. Additionally, because we did not include parity in our statistical model, this may also affect the results reported.

GWAS on Mid-Infrared Wavenumbers

The results discussed above showed that individual QTL affect milk composition in different ways. Usually the MIR data are used to predict the concentration of a particular component and then a GWAS for that component is carried out. However, an alternative is to analyze the MIR data directly. Figure 4 shows how individual SNP had quite different patterns of effect. The specific wavenumbers affected by a SNP had a clear relationship to the milk traits associated with that SNP. For example, for the SNP in Figure 4A near the *CSN3* gene, the most significant wavenumbers were observed in ranges assumed to denote carboxylic groups of protein and peptide bonds, respectively (Sivakesava and Irudayaraj, 2002; Dufour, 2009; Soyeurt et al., 2010). Considering SNP in Figure 4B, even in the surrounding locations, the significant spectral regions could be related to the several responses induced by lactose bonds from 1,045 cm^{-1} to 1,250 cm^{-1} (Picque

et al., 1993; Grelet et al., 2015). In Figure 4C, the peak around wavenumber 966 cm^{-1} was previously described to be an absorption band for measuring unsaturated fatty acids (Safar et al., 1994). This was consistent with GWAS results on milk composition traits, which highlighted the significant effect of this SNP on C18:1 (Figure 3 and Table 1). Moreover, the region between 1,365 and 1,488 cm^{-1} has been associated with C–H bending of $–CH_3$ and $–CH_2$ (Grelet et al., 2015). These chemical bonds characterize fat chains, thus, a relationship between this SNP and milk fat composition could be assumed, albeit in the absence of a specific association that allows discrimination between fatty acids. In addition, the SNP was significantly associated with wavenumbers between 1,283 and 1,294 cm^{-1}, which have no apparent association with fat bonds. Thus, this SNP had an association with MIR wavenumbers that is not just a reflection of its effect on one milk component. The pattern of associations is presumably unusual, explaining why the genotype of this SNP was so well predicted from MIR data ($R^2 = 0.66$).

Single nucleotide polymorphisms associated with the same milk trait can be associated with different patterns across the MIR spectrum. In Figure 5, 2 SNP related to lactose percentage shared only a few significant common wavenumbers. If 2 SNP have the same effects on wavenumbers, the correlation between their effects should be equal, or as close as possible, to 1. In this case, even considering only significant effects on these few wavenumbers, a negative correlation of −0.859 was observed. Thus, when all wavenumbers were considered, as done for the other SNP, the correlation between the effects of the 2 SNP decreased to −0.084. These results may suggest that these SNP had significant effects on the same trait but affecting different pathways and components.

The SNPs near *DGAT1* (Figure 6A) had a large association with numerous infrared wavenumbers, which was consistent with findings recently reported in Wang et al. (2016). The fact that 3 SNP around 1.7 to 1.8 Mb on BTA14 (Figure 6A) represented the same QTL was supported by their very close positions and almost identical effects on milk traits; moreover, they were in strong LD ($\rho^2 = 0.87$–0.95). In addition, this assumption was supported by their effect on spectral data, showing the same significance patterns along the spectrum and strong correlations between effects on MIR wavenumbers ($r = 0.9997$–0.9999). Considering this relationship, similar conclusions could be drawn

Table 2. List of SNP at the highest point of each peak identified in the Manhattan plot of partial least squares regression results (Figure 7) and not reported in Table 1

SNP	Rsnp ID[1]	R^2	$-\log_{10}(P)$
Chr1:69733624	rs109186784	0.114	147.33
Chr2:131804601	rs137565772	0.040	47.89
Chr5:31347875	—	0.042	50.44
Chr5:55018535	rs134714163	0.043	51.93
Chr7:16287363	rs17870681	0.025	30.67
Chr8:6443202	rs42335431	0.023	28.04
Chr9:43459255	—	0.023	27.59
Chr10:11445796	rs137713131	0.031	36.88
Chr12:21382137	rs109298045	0.024	29.03
Chr13:49030604	rs42341685	0.037	45.12
Chr15:66146528	rs132923316	0.034	41.28
Chr16:54066099	rs132979921	0.025	30.57
Chr17:56174646	rs137653132	0.039	47.03
Chr18:14524335	rs132988395	0.026	31.56
Chr21:27867971	rs42888536	0.023	27.76
Chr23:30479549	rs137218266	0.025	29.93
Chr25:26308666	rs42072596	0.155	208.12
Chr26:9365588	rs136433913	0.182	253.40
Chr27:36155097	rs110519353	0.066	80.69
Chr29:1997031	rs132805432	0.034	40.89
Chr30:39558858	rs132902510	0.023	27.48

[1]Rsnp ID = reference SNP identification.

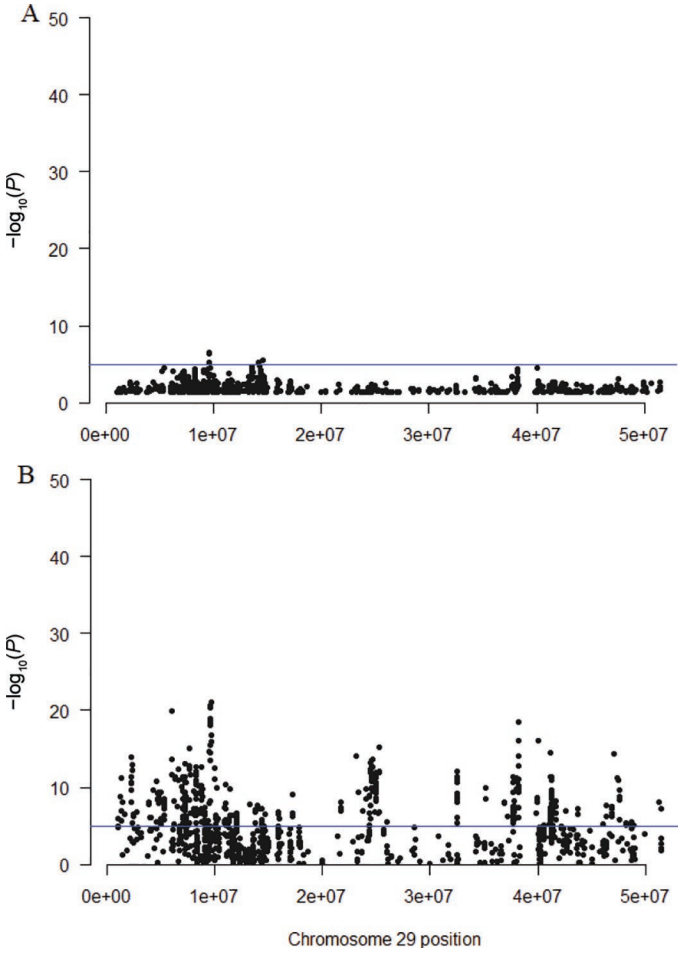

Figure 8. Comparison between SNP in genome-wide association studies results for lactose percentage (A) with $P < 0.05$ and in partial least squares regression results (B). The horizontal line is $-\log_{10}(P) \geq 5$.

Table 3. Examples of selected SNP and neighboring genes that are in different regions of the genome and that have questionable common QTL presence shown by the squared correlation coefficient (ρ^2) of linkage disequilibrium (LD), coefficients of correlation (r) between effects on spectral wavenumbers, and R^2 of multiple partial least squares regression analyses

SNP[1]								PLS[3] R^2		
y	Rsnp ID[4]	Gene	x	Rsnp ID[4]	Gene	LD (ρ^2)	Corr effect[2] (r)	MIR	SNPx	MIR+SNPx
Chr6:83973635	rs137002240	IV[5]	Chr6:87385233	rs110516603	*CSN3*	0.513	0.722	0.207	0.470	0.476
Chr6:85486799	rs110004470	*TMPRSS11F*	Chr6:87385233	rs110516603	*CSN3*	0.570	0.896	0.190	0.524	0.522
Chr11:103798768	rs111018835	*CCDC187*	Chr11:103308330	rs109087963	*PAEP*	0.289	0.914	0.194	0.286	0.282
Chr11:104258107	rs135261923	*ABO*	Chr11:103308330	rs109087963	*PAEP*	0.173	−0.871	0.142	0.176	0.182
Chr12:72444330	rs134319055	IV	Chr12:70293273	rs137218804	IV	0.017	−0.552	0.010	0.017	0.021
Chr20:31228912	rs133057950	*NNT*	Chr20:31909478	—	*GHR*	0.494	0.996	0.016	0.541	0.538
Chr20:34522480	rs43762676	IV	Chr20:31909478	—	*GHR*	0.174	0.864	0.017	0.147	0.154

[1]Where y = SNP used as dependent variable; x = SNP used as explanatory variable.
[2]SNP effect on ,id-infrared (MIR) wavenumbers.
[3]Where MIR = MIR wavenumbers used as explanatory variable; SNPx = SNP defined as x used as explanatory variable; MIR+SNPx = MIR wavenumbers and SNP defined as x used as explanatory variables.
[4]Rsnp ID = reference SNP identification.
[5]Intergenic variant.

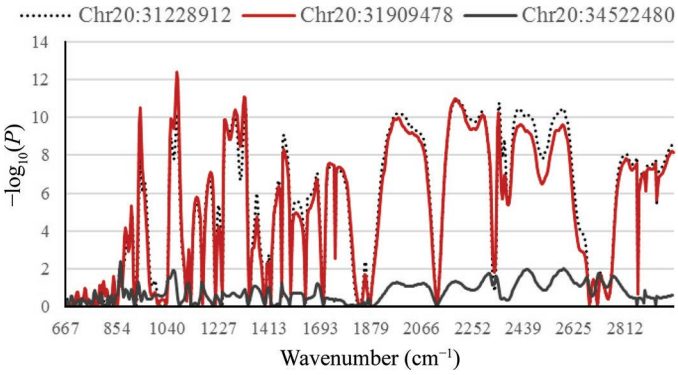

Figure 9. Significance [$-\log_{10}(P)$] of the effect of 3 SNP (Chr20:31228912, Chr20:31909478, and Chr20:34522480) identified on BTA20 on infrared wavenumbers in genome-wide association studies.

for SNP on BTA14 around 66 and 69 Mb (Figure 6B). Although they were 3 Mb apart, these SNP showed the same tendency in influencing milk composition (Table 1), a moderately high LD, and similar effects on wavenumbers ($\rho^2 = 0.58$; r = 0.993). These findings suggest that these 2 SNP might be related to the same QTL. Although Wang and Bovenhuis (2018) performed GWAS on 50 individual MIR wavenumbers, demonstrating their feasibility to detect new genomic regions affecting milk composition, no other study to our knowledge has used GWAS on MIR spectra as a potential tool for discriminating QTL. We found that associations between SNP effects on MIR spectra were confirmed by their significance patterns using GWAS.

SNP Prediction Using PLS

It is apparent that different QTL have different patterns of effect across the MIR spectrum, which suggests that a multi-trait GWAS, using all 598 wavenumbers, should increase the power to detect and map milk composition QTL. To carry out the equivalent of a multi-trait GWAS, we used PLS to predict SNP genotypes from MIR wavenumbers. The aim of the prediction equations was not to obtain accurate predicted genotypes but to test PLS as a tool to better map milk composition QTL.

Table 1 shows that many of the SNP associated with conventional milk composition traits were detected by the PLS analysis. In general, lower *P*-values for the significance test performed using our formula were detected for the PLS compared with the conventional analysis. These results suggest that the PLS method could have greater power to detect milk composition QTL, resulting in more precise QTL mapping. However, the slight reduction in prediction accuracy of the model including both MIR and GRM indicates that the effect

of population stratification might be significant in SNP prediction. On the one hand, these preliminary results suggest that population structure should be accounted for, obtaining more precise results. On the other hand, highly significant SNP detected through PLS confirmed their strong significance even after GRM correction and, overall, there were still more significant *P*-values than with conventional analysis (Table 1).

In addition, the PLS analysis detected QTL that have been previously reported by other authors, but which were not significantly associated with milk traits in our GWAS analysis. This included SNP near *LALBA*, encoding α-LA and related to lactose synthesis, or *AGPAT6*, involved in glycerolipid biosynthesis (Ramakrishnan et al., 2001; Chen et al., 2008; Nayeri and Stothard, 2016). The PLS analysis also found significant SNP that have not been previously reported to be associated with milk composition. For instance, *P2RX4*, a purinergic signaling gene linked to oxidative stress after calving in dairy cows (Seo et al., 2014), was observed. These SNP may be associated with changes in milk composition other than concentrations of fat, protein, lactose, or fatty acids. Indeed, oxidative stress causes milk composition changes (Talukder et al., 2015), with consequent effects on MIR spectral data. Overall, these results suggest the potential of PLS in distinguishing important chromosomal regions for milk composition and identifying some additional SNP not detected through conventional GWAS. Overall, as a proof of concept, our results suggest the potential of PLS in distinguishing important chromosome regions for milk composition and identifying some additional SNP not detected through conventional GWAS.

Testing for Multiple Linked QTL

When many SNP in a region of a chromosome are significantly associated with a trait, it is difficult to tell whether this represents more than one QTL or just one QTL in LD with many SNP. We combined GWAS and PLS analyses to attempt to distinguish between these situations. In the case of SNP on BTA20, we assumed that the SNP at 31.9 Mb, near *GHR*, represented at least one QTL and considered whether the other 2 are tracking the same QTL or an additional one. Because they were in LD, the SNP at 31.9 Mb predicted the genotype of the SNP at 31.2 Mb with $R^2 = 0.541$. When the MIR data were added to the prediction equation, we found no increase in the prediction R^2 (Table 3). If the SNP at 31.2 Mb tracked a QTL with different effects on MIR spectra to the SNP at 31.9 Mb, the use of MIR data should improve the prediction. Because it did not do so, it is likely both SNP tracked the same QTL. In contrast, the PLS prediction R^2 for the SNP at 34.5 Mb was increased by adding the MIR data.

These results are consistent with the interpretation that the SNP at 34.5 Mb was tracking an additional QTL to the *GHR* polymorphism near 31.9 Mb. Similarly, the other results in Table 3 suggest that there were 2 QTL on BTA11 near 103.3 and 104.3 Mb (near *ABO*), but the SNP on BTA6 at 85.5 (*TMPRSS11F*) and 87.4 Mb appeared to be tracking the same QTL or 2 QTL with the same pattern of effects on the MIR spectrum. Using a GWAS alone, it would be difficult to decipher the number of QTL present. These findings suggest that the combination of PLS and GWAS on MIR data can distinguish different, closely linked QTL.

Future Research

Our research has demonstrated that MIR data can be used as a powerful tool to enhance QTL detection and distinguish between multiple QTL. Further work with more animals and genome-sequence data should be considered to increase the power and precision of QTL detection.

CONCLUSIONS

Our results suggest that using MIR data through either GWAS or PLS analysis applied to genomic data can provide a powerful tool to distinguish milk composition QTL. Furthermore, PLS used to predict SNP genotypes showed potential for detecting and mapping significant SNP associated with milk composition, as well as previously undetected QTL for milk composition. Based on these results, using MIR data through GWAS or PLS analysis in genomic investigations can aid in distinguishing milk composition QTL.

ACKNOWLEDGMENTS

The genotype and mid-infrared spectral data were obtained as part of the MIRforProfit project "Integrating very large genomic and milk mid-infrared data to improve profitability of dairy cows," funded by the Australian Government Department of Agriculture (Canberra, Australia) as part of the Rural R&D for Profit programme. The authors thank DairyBio project, funded by Dairy Australia (Melbourne, Australia), the Gardiner Foundation (Melbourne, Australia), and Agriculture Victoria (Melbourne, Australia) for supporting this research. The first author, Anna Benedet, completed this research while on sabbatical at Agriculture Victoria; Fondazione Ing. Aldo Gini (University of Padova, Italy) is gratefully acknowledged for financial

support. Michael E. Goddard and Ruidong Xiang are supported by the Australian Research Council's Discovery Projects funding scheme (DP160101056).

REFERENCES

Blott, S., J. Kim, S. Moisio, A. Schmidt-Küntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart, D. Johnson, L. Karim, P. Simon, R. Snell, R. Spelman, J. Wong, J. Vilkki, W. Coppieters, and M. Georges. 2003. 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 163:253–266.

Bouwman, A. C., H. Bovenhuis, M. H. P. W. Visker, and J. A. M. Van Arendonk. 2011. Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet. 12:43.

Buitenhuis, B., L. L. G. Janss, N. A. Poulsen, L. B. Larsen, M. K. Larsen, and P. Sørensen. 2014. Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. BMC Genomics 15:1112.

Buitenhuis, B., N. A. Poulsen, G. Gebreyesus, and L. B. Larsen. 2016. Estimation of genetic parameters and detection of chromosomal regions affecting the major milk proteins and their post translational modifications in Danish Holstein and Danish Jersey cattle. BMC Genet. 17:114.

Chen, Y. Q., M. Kuo, S. Li, H. H. Bui, D. A. Peake, P. E. Sanders, S. J. Thibodeaux, S. Chu, Y. Qian, Y. Zhao, D. S. Bredt, D. E. Moller, R. J. Konrad, A. P. Beigneux, S. G. Young, and G. Cao. 2008. AGPAT6 is a novel microsomal glycerol-3-phosphate acyltransferase. J. Biol. Chem. 283:10048–10057.

Conte, G., M. Mele, S. Chessa, B. Castiglioni, A. Serra, G. Pagnacco, and P. Secchiari. 2010. Diacylglycerol acyltransferase 1, stearoyl-CoA desaturase 1, and sterol regulatory element binding protein 1 gene polymorphisms and milk fatty acid composition in Italian Brown cattle. J. Dairy Sci. 93:753–763.

De Marchi, M., M. Penasa, A. Cecchinato, M. Mele, P. Secchiari, and G. Bittante. 2011. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Animal 5:1653–1658.

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97:1171–1186.

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

Dufour, E. 2009. Principles of infrared spectroscopy. Pages 1–27 in Infrared Spectroscopy for Food Quality Analysis and Control. D. W. Sun, ed. Academic Press, San Diego, CA.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114–4129.

Eskildsen, C. E., M. A. Rasmussen, S. B. Engelsen, L. B. Larsen, N. A. Poulsen, and T. Skov. 2014. Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: Understanding predictions of highly collinear reference variables. J. Dairy Sci. 97:7940–7951.

Fox, P. F., T. Uniacke-Lowe, P. L. H. McSweeney, and J. A. O'Mahony. 2015. Dairy Chemistry and Biochemistry. 2nd ed. Springer International Publishing, Basel, Switzerland.

Geladi, P., and B. R. Kowalski. 1986. Partial least squares regression: A tutorial. Anal. Chim. Acta 185:1–17.

Grelet, C., J. A. F. Pierna, P. Dardenne, V. Baeten, and F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. J. Dairy Sci. 98:2150–2160.

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

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443.

Hempstalk, K., S. McParland, and D. P. Berry. 2015. Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. J. Dairy Sci. 98:5262–5273.

Hewavitharana, A. K., and B. van Brakel. 1997. Fourier transform infrared spectrometric method for the rapid determination of casein in raw milk. Analyst (Lond.) 122:701–704.

Iso-Touru, T., G. Sahana, B. Guldbrandtsen, M. S. Lund, and J. Vilkki. 2016. Genome-wide association analysis of milk yield traits in Nordic Red cattle using imputed whole genome sequence variants. BMC Genet. 17:55.

Kemper, K. E., C. M. Reich, P. J. Bowman, C. J. Vander Jagt, A. J. Chamberlain, B. A. Mason, B. J. Hayes, and M. E. Goddard. 2015. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet. Sel. Evol. 47:29.

Knutsen, T. M., H. G. Olsen, V. Tafintseva, M. Svendsen, A. Kohler, M. P. Kent, and S. Lien. 2018. Unravelling genetic variation underlying de novo-synthesis of bovine milk fatty acids. Sci. Rep. 8:2179.

Littlejohn, M. D., K. Tiplady, T. A. Fink, K. Lehnert, T. Lopdell, T. Johnson, C. Couldrey, M. Keehan, R. G. Sherlock, C. Harland, A. Scott, R. G. Snell, S. R. Davis, and R. J. Spelman. 2016. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. Sci. Rep. 6:25376.

Lopdell, T. J., K. Tiplady, M. Struchalin, T. J. J. Johnson, M. Keehan, R. Sherlock, C. Couldrey, S. R. Davis, R. G. Snell, R. J. Spelman, and M. D. Littlejohn. 2017. Open access DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. BMC Genomics 18:968.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes, and M. E. Goddard. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17:144.

Mevik, B.-H., and R. Wehrens. 2007. The pls Package: Principal component and partial least squares regression in R. J. Stat. Softw. 18:1–24.

Nayeri, S., and P. Stothard. 2016. Tissues, metabolic pathways and genes of key importance in lactating dairy cattle. Springer Sci. Rev. 4:49–77.

Nogalski, Z., M. Wronski, M. Sobczuk-Szul, M. Mochol, and P. Pogorzelska. 2012. The effect of body energy reserve mobilization on the fatty acid profile of milk in high-yielding cows. Asian-Australas. J. Anim. Sci. 25:1712–1720.

Picque, D., D. Lefier, R. Grappin, and G. Corrieu. 1993. Monitoring fermentation by infrared spectrometry: Alcoholic and lactic fermentations. Anal. Chim. Acta 279:67–72.

Pryce, J. E., S. Bolormaa, A. J. Chamberlain, P. J. Bowman, K. Savin, M. E. Goddard, and B. J. Hayes. 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J. Dairy Sci. 93:3331–3345.

R Core Team. 2017. R: A language and environment for statistical computing R. Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramakrishnan, B., P. S. Shah, and P. K. Qasba. 2001. α-Lactalbumin (LA) stimulates milk β-1,4-galactosyltransferase I (β4Gal-T1) to transfer glucose from UDP-glucose to N-acetylglucosamine. J. Biol. Chem. 276:37665–37671.

Raven, L. A., B. G. Cocks, M. E. Goddard, J. E. Pryce, and B. J. Hayes. 2014. Genetic variants in mammary development, prolactin signalling and involution pathways explain considerable variation in bovine milk production and milk composition. Genet. Sel. Evol. 46:29.

Raven, L. A., B. G. Cocks, K. E. Kemper, A. J. Chamberlain, C. J. Vander Jagt, M. E. Goddard, and B. J. Hayes. 2016. Targeted

imputation of sequence variants and gene expression profiling identifies twelve candidate genes associated with lactation volume, composition and calving interval in dairy cattle. Mamm. Genome 27:81–97.

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011. Prediction of β-lactoglobulin genotypes based on milk Fourier transform infrared spectra. J. Dairy Sci. 94:4183–4188.

Safar, M., D. Bertrand, P. Robert, M. F. Devaux, and C. Genot. 1994. Characterization of edible oils, butters and margarines by Fourier transform infrared spectroscopy with attenuated total reflectance. J. Am. Oil Chem. Soc. 71:371–377.

Schennink, A., W. M. Stoop, M. H. P. W. Visker, J. M. L. Heck, H. Bovenhuis, H. J. F. Van Valenberg, and J. A. M. Van Arendonk. 2007. DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. Anim. Genet. 38:467–473.

Seo, J., J. S. Osorio, E. Schmitt, M. N. Corrêa, G. Bertoni, E. Trevisi, and J. J. Loor. 2014. Hepatic purinergic signaling gene network expression and its relationship with inflammation and oxidative stress biomarkers in blood from peripartal dairy cattle. J. Dairy Sci. 97:861–873.

Shenk, J. S., and M. O. Westerhaus. 1995. Forage analysis by near infrared spectroscopy. Pages 111–120 in Forages. Vol. II. The Science of Grassland Agriculture. 5th ed. R. F. Barnes, D. A. Miller, and C. J. Nelson, ed., Iowa State University Press, Ames.

Sivakesava, S., and J. Irudayaraj. 2002. Rapid determination of tetracycline in milk by FT-MIR and FT-NIR spectroscopy. J. Dairy Sci. 85:487–493.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94:1657–1667.

Soyeurt, H., I. Misztal, and N. Gengler. 2010. Genetic variability of milk components based on mid-infrared spectral data. J. Dairy Sci. 93:1722–1728.

Stoop, W. M., A. Schennink, M. H. P. W. Visker, E. Mullaart, J. A. M. Van Arendonk, and H. Bovenhuis. 2009. Genome-wide scan for bovine milk-fat composition I. QTL for short and medium chain fatty acids. J. Dairy Sci. 92:4664–4675.

Stoop, W. M., J. A. M. van Arendonk, J. M. L. Heck, H. J. F. van Valenberg, and H. Bovenhuis. 2008. Genetic parameters for major milk fatty acids and milk production traits of Dutch Holstein-Friesians. J. Dairy Sci. 91:385–394.

Talukder, S., K. L. Kerrisk, G. Gabai, A. Fukutomi, and P. Celi. 2015. Changes in milk oxidative stress biomarkers in lactating dairy cows with ovulatory and an-ovulatory oestrous cycles. Anim. Reprod. Sci. 158:86–95.

Wang, Q., and H. Bovenhuis. 2018. Genome-wide association study for milk infrared wavenumbers. J. Dairy Sci. 101:2260–2272.

Wang, Q., A. Hulzebosch, and H. Bovenhuis. 2016. Genetic and environmental variation in bovine milk infrared spectra. J. Dairy Sci. 99:6793–6803.

Xiang, R., I. M. MacLeod, S. Bolormaa, and M. E. Goddard. 2017. Genome-wide comparative analyses of correlated and uncorrelated phenotypes identify major pleiotropic variants in dairy cattle. Sci. Rep. 7:9248.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: A tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88:76–82.