J. Dairy Sci. 105:9763–9791
https://doi.org/10.3168/jds.2022-22089

Comparison of the genetic characteristics of directly measured and Fourier-transform mid-infrared-predicted bovine milk fatty acids and proteins

Kathryn M. Tiplady,[1,2]* Thomas J. Lopdell,[1] Richard G. Sherlock,[1] Thomas J. J. Johnson,[1] Richard J. Spelman,[1] Bevin L. Harris,[1] Stephen R. Davis,[1] Mathew D. Littlejohn,[1,2] and Dorian J. Garrick[2]
[1]Research and Development, Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand
[2]School of Agriculture and Environment, Massey University, Ruakura, Hamilton 3240, New Zealand

ABSTRACT

Fourier-transform mid-infrared (FT-MIR) spectroscopy is a high-throughput and inexpensive methodology used to evaluate concentrations of fat and protein in dairy cattle milk samples. The objective of this study was to compare the genetic characteristics of FT-MIR predicted fatty acids and individual milk proteins with those that had been measured directly using gas and liquid chromatography methods. The data used in this study was based on 2,005 milk samples collected from 706 Holstein-Friesian × Jersey animals that were managed in a seasonal, pasture-based dairy system, with milk samples collected across 2 consecutive seasons. Concentrations of fatty acids and protein fractions in milk samples were directly determined by gas chromatography and high-performance liquid chromatography, respectively. Models to predict each directly measured trait based on FT-MIR spectra were developed using partial least squares regression, with spectra from a random selection of half the cows used to train the models, and predictions for the remaining cows used as validation. Variance parameters for each trait and genetic correlations for each pair of measured/predicted traits were estimated from pedigree-based bivariate models using REML procedures. A genome-wide association study was undertaken using imputed whole-genome sequence, and quantitative trait loci (QTL) from directly measured traits were compared with QTL from the corresponding FT-MIR predicted traits. Cross-validation prediction accuracies based on partial least squares for individual and grouped fatty acids ranged from 0.18 to 0.65. Trait prediction accuracies in cross-validation for protein fractions were 0.53, 0.19, and 0.48 for α-casein, β-casein, and κ-casein, 0.31 for α-lactalbumin, 0.68 for β-lactoglobulin, and 0.36 for lactoferrin. Heritability estimates for directly measured traits ranged from 0.07 to 0.55 for fatty acids; and from 0.14 to 0.63 for individual milk proteins. For FT-MIR predicted traits, heritability estimates were mostly higher than for the corresponding measured traits, ranging from 0.14 to 0.46 for fatty acids, and from 0.30 to 0.70 for individual proteins. Genetic correlations between directly measured and FT-MIR predicted protein fractions were consistently above 0.75, with the exceptions of C18:0 and C18:3 *cis*-3, which had genetic correlations of 0.72 and 0.74, respectively. The GWAS identified trait QTL for fatty acids with likely candidates in the *DGAT1*, *CCDC57*, *SCD*, and *GPAT4* genes. Notably, QTL for *SCD* were largely absent in the FT-MIR predicted traits, and QTL for *GPAT4* were absent in directly measured traits. Similarly, for directly measured individual proteins, we identified QTL with likely candidates in the *CSN1S1*, *CSN3*, *PAEP*, and *LTF* genes, but the QTL for *CSN3* and *LTF* were absent in the FT-MIR predicted traits. Our study indicates that genetic correlations between directly measured and FT-MIR predicted fatty acid and protein fractions are typically high, but that phenotypic variation in these traits may be underpinned by differing genetic architecture.

Key words: milk composition, Fourier-transform mid-infrared spectroscopy, genome-wide association study, dairy cattle

INTRODUCTION

Bovine milk is a rich source of dietary nutrients that are important to human health, including proteins, fats, carbohydrates, vitamins, and minerals. The concentrations of these components are determined by genetic factors such as breed and sire, as well as nongenetic factors related to the environment, stage of lactation, feed, and the nutritional status of the animal. Fats are important to human health due to the role they play in growth, development, hormone regulation, and inflammation management. In bovine milk, a typical

Received March 16, 2022.
Accepted July 21, 2022.
*Corresponding author: Kathryn.Sanders@lic.co.nz

fatty acid profile comprises about 70% saturated, 25% monounsaturated, and 5% polyunsaturated fatty acids.

Bovine milk is also a common source of protein, an important nutrient in the human diet because of the role it has in body maintenance, as well as the growth and repair of cells. However, the concentrations of casein and whey proteins in bovine milk differ to that of human milk, with bovine milk protein comprising approximately 80% casein and 20% whey proteins, whereas most of the protein in human milk represents whey proteins. These differences in protein composition are important because casein and whey proteins have different digestibilities and AA profiles. Moreover, the protein profiles have implications for cheese processing and the manufacture of casein supplements.

Fourier-transform mid-infrared (**FT-MIR**) spectroscopy is a method to determine the presence of specific chemical bonds in a composite substance such as milk, and is widely used in the dairy industry to characterize milk composition. The approach involves directing infrared light through a milk sample, leading to interactions between the infrared light and molecules in the milk that cause vibrations and rotational changes in molecular bonds, resulting in the differential absorption of the various infrared light wavelengths. From this process, a spectrum of absorbance values for light wavelengths across the mid-infrared range is generated, which can be used to predict a variety of traits. This is a high-throughput and inexpensive method for predicting milk composition from milk samples and is widely used to reliably quantify concentrations of fat and protein for dairy cattle. This methodology is also of interest for characterizing fat composition, casein, and whey proteins in milk because of the implications these milk components may have for human health and milk processability, and because the FT-MIR spectra are already available from routine milk testing.

Applications using FT-MIR spectral data to predict milk composition traits typically involve using a set of samples with directly measured trait values to develop a calibration equation based on the spectrum of absorbance values, using methods such as partial least squares (**PLS**) regression. The resulting calibration equation can then be applied to future samples to predict trait values as a linear combination of individual wavenumber absorbances from any milk sample with FT-MIR spectral data. The success of using FT-MIR data as a phenotyping tool relies on the strength of the phenotypic correlation between the directly measured trait and the FT-MIR predicted trait. However, the success of using an FT-MIR predicted trait in breeding programs is further dependent on the heritability of the predicted trait, and the genetic correlation between the directly measured and predicted trait.

Previous studies have indicated that FT-MIR spectra can be used to predict fatty acids (Soyeurt et al., 2006; Rutten et al., 2009; Lopez-Villalobos et al., 2014; Bonfatti et al., 2016) and protein fractions in milk (De Marchi et al., 2009; Bonfatti et al., 2011, 2016; Rutten et al., 2011; Soyeurt et al., 2012; McDermott et al., 2016). Moreover, moderate to high heritability estimates have been reported for a range of FT-MIR predicted fatty acids (Rutten et al., 2010; Lopez-Villalobos et al., 2014; Bonfatti et al., 2017b; Narayana et al., 2017; Fleming et al., 2018) and protein fractions (Soyeurt et al., 2007a; Arnould et al., 2009b; Bonfatti et al., 2017b; Sanchez et al., 2017b). Few studies report the genetic correlations between directly measured and FT-MIR predicted fatty acids, or protein fractions, or both, but in those studies the genetic correlations are typically high (Rutten et al., 2010; Bonfatti et al., 2017b).

Several GWAS have been conducted on fatty acids and protein fractions in bovine milk, across a range of genotype densities. This includes studies of directly measured fatty acids using 50k (Bouwman et al., 2011) or high-density (**HD**) genotypes (Buitenhuis et al., 2014; Palombo et al., 2018), and FT-MIR predicted fatty acids using 50k (Cruz et al., 2019; Iung et al., 2019; Freitas et al., 2020), HD (Olsen et al., 2017), or imputed whole-genome sequence (Sanchez et al., 2019) genotypes. Studies of directly measured protein fractions include those using 50k (Schopen et al., 2011; Pegolo et al., 2018) or HD (Buitenhuis et al., 2016; Zhou et al., 2019) genotypes, and studies of FT-MIR predicted protein fractions include those using imputed sequence genotypes (Sanchez et al., 2017b, 2019). Aside from differences in genotype density, the breed composition of animals in these studies also varies. In particular, studies of directly measured fatty acids include Dutch Holstein-Friesians (Bouwman et al., 2011), Danish Holsteins and Jerseys (Buitenhuis et al., 2014), and Italian Simmental and Holsteins (Palombo et al., 2018), whereas studies of FT-MIR predicted fatty acids include Holstein (Cruz et al., 2019; Iung et al., 2019; Freitas et al., 2020), Norwegian Red (Olsen et al., 2017), and Montbéliarde (Sanchez et al., 2019) cows. Studies of directly measured protein fractions in milk include Dutch Holstein-Friesians (Schopen et al., 2011), Italian Brown Swiss cows (Pegolo et al., 2018), and Danish Holsteins and Jerseys (Buitenhuis et al., 2016), whereas studies of FT-MIR predicted protein fractions include Montbéliarde, Normande, and Holstein cows (Sanchez et al., 2017b, 2019). Differences in genotype density and breed composition for GWAS conducted on directly measured and FT-MIR predicted fatty acid and protein traits make it difficult to compare QTL between studies. To date, as far as we are aware, there have been no GWAS that compare QTL for directly

measured fatty acids and protein traits to QTL for the corresponding FT-MIR predicted traits within the same study population.

The objective of this study was to compare the genetic characteristics of directly measured fatty acids and protein fractions to the same traits predicted from FT-MIR spectra. Calibration equations were developed using milk samples from New Zealand crossbred dairy cattle, and pedigree-based models were used to evaluate the (co)variance parameters of each directly measured trait and its corresponding FT-MIR predicted trait. To understand the underlying differences in the genetic architecture of directly measured and FT-MIR predicted traits, we conducted GWAS using imputed whole-genome sequence, and compared QTL from directly measured traits to QTL from the corresponding FT-MIR predicted traits. It was expected that the use of imputed whole-genome sequence genotypes from an F_2 study population would enhance our ability to identify trait QTL and candidate causative mutations, and that using the same data set to conduct GWAS across directly measured and FT-MIR predicted traits would be valuable for determining differences between QTL.

MATERIALS AND METHODS

Ethics Statement

Animal ethics approval for the collection of data used in this study was granted by the Ruakura Animal Ethics Committee (Hamilton, New Zealand; approval numbers 4,232, 4,621, and 10,174), according to the rules and guidelines outlined in the New Zealand Animal Welfare Act 1999.

Study Population/Animals and Milk Samples

Animals included in this study were from an F_2 design crossbreeding experiment with a half-sibling family structure, as previously described (Spelman et al., 2001; Berry et al., 2010). Briefly, 6 F_1 bulls were generated from reciprocal crosses of Holstein-Friesian and Jersey animals that were then mated to high genetic merit F_1 cows. This resulted in a herd of 850 F_2 female progeny, consisting of 2 cohorts produced over consecutive seasons, which were managed in a seasonal, pasture-based dairy system. Because of the phenotypic differences between milk composition for Friesian and Jersey animals, it was expected that the genetic variation exhibited in F_2 animals would typically be higher compared with what would be seen in a study of purebred animals, and that this could assist in the identification of trait QTL.

Measurements of FT-MIR spectra, and fatty acid and protein composition, were evaluated from second lactation milk samples collected at peak-, mid- and late-lactation in the 2003 to 2004 season for cohort 1, and the 2004 to 2005 season for cohort 2. Calving for each cohort took place over ~3 mo between July and October. Samples for each cohort representing peak milk were collected on a daily basis for these cows at 35 d postcalving, whereas mid- and late-lactation samples were collected at a fixed date across the herd within the season. A frequency distribution of the number of samples classified by DIM at the time of sampling has been provided in Appendix A1.

Concentrations of fatty acids were directly determined in milk fat samples by fatty acid methyl ester analysis using GC (MacGibbon and Reynolds, 2011), within 1 of up to 5 batches on a given sample collection day, and were expressed as grams per 100 g of total fat content. In this study, we report an analysis for 17 individual fatty acids and 6 fatty acid groups that were classified based on the degree of saturation and the length of the carbon chain, as follows: (1) SFA (no double bonds); (2) UFA (1 or more double bonds); (3) PUFA (2 or more double bonds); (4) short-chain fatty acids (**SCFA**; 4, 6, or 8 carbons); (5) medium-chain fatty acids (**MCFA**; 10, 12, or 14 carbons); and (6) long-chain fatty acids (**LCFA**; 18 carbons). Milk proteins were determined using HPLC, as described by Palmano and Elgar (2002), and were analyzed within 1 of up to 6 batches on a given sample collection day, and were expressed as grams per liter of total milk volume. Traits were assessed for deviation from normality by visual inspection of normal quantile plots and by evaluating asymmetry according to skewness. With the exception of lactoferrin, all directly measured traits were approximately normally distributed with absolute skewness values less than 1. For lactoferrin, log, square-, and cube-root transformations were applied to determine which transformation minimized skewness. A cube-root transformation was the most effective of those investigated for minimizing skewness and was applied to lactoferrin trait values for all downstream analyses. Frequency distributions of untransformed lactoferrin concentrations and lactoferrin concentrations after applying a cube-root transformation are provided in Appendix A2. Outliers for each fatty acid and protein trait were identified and removed if the trait value was more than 3 standard deviations from the mean for the corresponding season and stage of lactation (peak, mid, late). After removal of outliers, each trait was adjusted to remove batch effects, where batch effects were evaluated from a random effects model with batch nested within season and stage of lactation, using Nelder-Mead optimization as implemented in the lme4 package in R (Bates et al., 2015).

The same milk samples assessed for fatty acid and protein composition were also analyzed on a Foss MilkoScan FT6000 (Foss) instrument, to generate spectral records consisting of 1,060 wavenumbers across the range from 925.66 to 5,010.15 cm^{-1}. Spectral data from regions associated with low signal-to-noise ratios and poor sample measurement repeatability due to the water content in milk were excluded, according to the definitions by Tiplady et al. (2019). Specifically, the excluded low signal-to-noise regions were 649 to 970 cm^{-1}, 1,608 to 1,682 cm^{-1}, and $\geq$3,021 cm^{-1}. This resulted in 542 wavenumbers for use in the development of prediction equations. Outliers in the spectral data were identified using the methodology described in Tiplady et al. (2019). Briefly, the squared Mahalanobis distance between each spectral record and the average spectra were evaluated using the 542 wavenumbers identified as being outside low signal-to-noise regions. The distributions of Mahalanobis distance values for each season were compared and found to be similar, indicating that although the spectra were collected in 2 different seasons, the effect of instrument drift across time was likely to be small. Based on the lowest average information criterion, a logistic distribution with location and scale parameters of 541.7 and 27.3, respectively, had the best fit to the overall Mahalanobis distance values, and based on a P-value of 0.001, 18 outliers were identified and removed. In total, after outlier removal, we had 2,005 samples for 706 animals with FT-MIR spectra and either a fatty acid or protein composition result. Traits varied in the final number of records available for analysis, ranging from 1,686 to 1,977 records, and representing from 699 to 704 animals. The overall mean fat and protein concentrations as predicted from the Foss instrument calibration equation were 5.40 (SD = 0.70) and 3.98 (SD = 0.36), respectively.

Development and Validation of Calibration Equations

Phenotypic calibration equations for each fatty acid and protein fraction were evaluated within a cross-validation framework, whereby records for a random selection of half the animals were assigned to a training dataset, and the remaining records were assigned to a validation dataset. This ensured that validation was cow-independent in that none of the records for animals included in the training dataset were included in the validation dataset. Partial least squares models for each trait were developed using 542 spectral wavenumbers with the caret package in R (Kuhn et al., 2022), based on training data with 10 repeats of 10-fold cross-validation. In addition to the untreated spectra, several mathematical treatments of spectra were assessed using the mdatools package in R (Kucheryavskiy, 2020), including standard normal variate (**SNV**) transformation, multiplicative scatter correction, and first-order Savitzky-Golay derivative (Savitzky and Golay, 1964) treatments. First-derivative treatments were applied to untreated spectra and spectra after SNV or multiplicative scatter correction treatments using a range of window sizes, with up to 1 and 10 points either side. For each trait, the performance of the PLS model was assessed according to the coefficient of determination between actual and predicted phenotypic trait values in the validation dataset $\left(R^2_{cv}\right)$, and the relative prediction error (**RPE**) between actual and predicted trait values in the validation dataset ($\mathbf{RPE}_{cv}$), as described by Lopez-Villalobos et al. (2014).

Genetic Parameters of Traits

Genetic (co)variances of each directly measured trait and its corresponding FT-MIR predicted trait were estimated using a pairwise bivariate repeated measures animal model in ASReml-R (Butler et al., 2009) based on a pedigree comprising 5,943 animals. The model was defined as follows:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix}=\begin{bmatrix}\mathbf{X}_1 & 0\\ 0 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix}+\begin{bmatrix}\mathbf{Z}_1 & 0\\ 0 & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix}+\begin{bmatrix}\mathbf{W}_1 & 0\\ 0 & \mathbf{W}_2\end{bmatrix}\begin{bmatrix}\mathbf{p}_1\\ \mathbf{p}_2\end{bmatrix}+\begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}, \quad [1]$$

where $\mathbf{y}_1$ is a vector of the directly measured fatty acid or protein fraction, $\mathbf{y}_2$ is a vector of the corresponding FT-MIR predicted trait; $\mathbf{X}_1$, $\mathbf{Z}_1$, $\mathbf{W}_1$, $\mathbf{X}_2$, $\mathbf{Z}_2$, and $\mathbf{W}_2$ are design matrices for the fixed, additive genetic and permanent environment effects, respectively, for $\mathbf{y}_1$ and $\mathbf{y}_2$; $\mathbf{b}_1$ and $\mathbf{b}_2$ are vectors of the fixed effect of DIM (represented as 35-day windows from the start of lactation) within season (2003, 2004) for the directly measured and the FT-MIR predicted trait, respectively; $\mathbf{u}_1$ and $\mathbf{u}_2$ are vectors of random additive genetic effects for each trait; $\mathbf{p}_1$ and $\mathbf{p}_2$ are vectors of permanent environment effects for each trait; and $\mathbf{e}_1$ and $\mathbf{e}_2$ are vectors of residuals. The following (co)variance structure for each directly measured ($\mathbf{y}_1$) and FT-MIR predicted ($\mathbf{y}_2$) trait pair is assumed:

$$\text{var}\begin{bmatrix}\mathbf{u}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{G}\otimes\mathbf{A} & 0 & 0\\ 0 & \mathbf{C}\otimes\mathbf{I}_\mathbf{p} & 0\\ 0 & 0 & \mathbf{R}\otimes\mathbf{I}_\mathbf{e}\end{bmatrix},$$

where $\mathbf{u}=\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix}$, $\mathbf{p}=\begin{bmatrix}\mathbf{p}_1\\ \mathbf{p}_2\end{bmatrix}$, and $\mathbf{e}=\begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}$, where $\mathbf{A}$ is the numerator relationship matrix, $\mathbf{I}_\mathbf{p}$ is an identity matrix

of order corresponding to the length of the vector $\mathbf{p}$, $\mathbf{I_e}$ is an identity matrix of order corresponding to the length of the vector $\mathbf{e}$, $\otimes$ is the Kronecker product. Additionally, $\mathbf{G}$, $\mathbf{C}$, and $\mathbf{R}$ are genetic, permanent environment and residual (co)variance matrices, respectively, and are defined as follows:

$$\mathbf{G} = \begin{bmatrix} \sigma^2_{\boldsymbol{u}_1} & \sigma_{\boldsymbol{u}_1\boldsymbol{u}_2} \\ \sigma_{\boldsymbol{u}_1\boldsymbol{u}_2} & \sigma^2_{\boldsymbol{u}_2} \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} \sigma^2_{\boldsymbol{p}_1} & \sigma_{\boldsymbol{p}_1\boldsymbol{p}_2} \\ \sigma_{\boldsymbol{p}_1\boldsymbol{p}_2} & \sigma^2_{\boldsymbol{p}_2} \end{bmatrix},$$

and

$$\mathbf{R} = \begin{bmatrix} \sigma^2_{\boldsymbol{e}_1} & \sigma_{\boldsymbol{e}_1\boldsymbol{e}_2} \\ \sigma_{\boldsymbol{e}_1\boldsymbol{e}_2} & \sigma^2_{\boldsymbol{e}_2} \end{bmatrix}.$$

The heritability and repeatability for each trait were calculated as functions of the estimated (co)variance components based on their parametric definitions of $h^2_i = \frac{\sigma^2_{\boldsymbol{u}_i}}{\sigma^2_{\boldsymbol{u}_i} + \sigma^2_{\boldsymbol{p}_i} + \sigma^2_{\boldsymbol{e}_i}}$ and $t_i = \frac{\sigma^2_{\boldsymbol{u}_i} + \sigma^2_{\boldsymbol{p}_i}}{\sigma^2_{\boldsymbol{u}_i} + \sigma^2_{\boldsymbol{p}_i} + \sigma^2_{\boldsymbol{e}_i}}$, where $i =$ 1 or 2 for traits $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively, and the genetic correlation for each pair of measured/predicted traits was calculated as $r_a = \frac{\sigma_{\boldsymbol{u}_1\boldsymbol{u}_2}}{\sigma_{\boldsymbol{u}_1}\sigma_{\boldsymbol{u}_2}}$. For each bivariate analysis, starting values for additive genetic and residual (co)variances were estimated from single trait models. A range of covariance starting values were iteratively assessed for model convergence, with starting values of $\frac{a(\sigma^2_{\boldsymbol{u}_1} + \sigma^2_{\boldsymbol{u}_2})}{2}$ and $\frac{b(\sigma^2_{\boldsymbol{e}_1} + \sigma^2_{\boldsymbol{e}_2})}{2}$ for additive genetic and residual covariances, respectively, where a and b ranged from 0.1 to 0.9 in increments of 0.1. Among models that converged for each pair of traits, genetic parameter estimates were highly consistent. For traits that had different solutions from different models, the model that minimized the squared sum of the difference between single- and multi-trait model heritability estimates was selected.

Genotypes and Imputation

Of the 706 animals with phenotypic data, 685 were genotyped on Illumina BovineHD (HD; n = 12; ~777k SNP) or Illumina BovineSNP50k (50k; n = 685; ~53k SNP) panels, or were genotyped on both. The resultant genotypes were imputed to sequence density as part of a wider set of 153,357 animals, as described previously (Jivanji et al., 2019; Tiplady et al., 2021). Briefly, the imputation process consisted of stepwise imputation of animals to whole-genome sequence genotypes via references of GeneSeek Genomic Profiler, 50k, and HD genotypes. The whole-genome sequence reference consisted of 565 animals, comprised of 138 Holstein-Friesians, 99 Jerseys, 316 Holstein-Friesian × Jersey crossbreeds, and 12 from other breeds or crosses. Notably, the 6 F_1 sires included in our study were included in this whole-genome sequence reference and were sequenced with a target of 60× read-depth coverage. Phasing was undertaken using Beagle 4.0 (Browning and Browning, 2007), based on genotype probabilities, and variants were filtered to remove those where the allelic R^2 for missing genotypes was less than 0.95. Only variants located on *Bos taurus* autosomes were considered, resulting in a sequence reference comprising 19,659,361 segregating variants spanning all 29 autosomes. Imputation was carried out using Beagle 4.0 (Browning and Browning, 2007), ignoring pedigree information, and SNP with allelic $R^2 < 0.7$ were removed after each imputation step. The overall median imputation allelic R^2 for the wider set of 153,357 animals was 0.986, but was 0.992 for the 685 genotyped animals included in this study.

Genome-Wide Association Studies

Before conducting GWAS, adjusted fatty acid and protein phenotypes were generated for directly measured and FT-MIR predicted traits. The generation of the adjusted phenotypes was based on 1 or more samples measured on the same cow, which were fitted to a univariate pedigree-based repeated measures model in ASReml-R (Butler et al., 2009), as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp} + \mathbf{e}, \quad [2]$$

where $\mathbf{y}$ is a vector of the measured or predicted trait, $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}$ are design matrices for the fixed, additive genetic, and permanent environment effects; $\mathbf{b}$ is the fixed effect of DIM (represented as 35-d windows from the start of lactation) within season (2003, 2004) for the trait; $\mathbf{u}$ is a vector of random additive genetic effects with $\mathbf{u} \sim N\left(\mathbf{0}, \mathbf{A}\sigma^2_u\right)$; $\mathbf{p} \sim N\left(\mathbf{0}, \mathbf{I_p}\sigma^2_p\right)$ is a vector of random permanent environment effects; and $\mathbf{e}$ is a vector of random residuals with $\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{I_e}\sigma^2_e\right)$, where $\mathbf{A}$ is the numerator relationship matrix, $\mathbf{I_p}$ is an identity matrix of order corresponding to the length of the vector $\mathbf{p}$, $\mathbf{I_e}$ is an identity matrix of order corresponding to

the length of the vector $\mathbf{e}$, σ_u^2 is the additive genetic variance, σ_p^2 is the permanent environment variance, and σ_e^2 is the residual variance. Adjusted phenotypes used in the GWAS were the average of $\mathbf{y}$ over all observations for a cow minus the relevant fixed effects.

For each directly measured fatty acid or protein trait and its corresponding FT-MIR predicted trait, a GWAS was conducted using Bolt-LMM software (Loh et al., 2015). Before conducting GWAS, a minor allele frequency threshold of 1% based on allele frequencies in the 685-animal study population was applied, resulting in 14,990,779 imputed sequence variants included in each GWAS. To assess the additive effect of each SNP, mixed model association statistics were evaluated under an infinitesimal model. To account for population structure, a genomic relationship matrix based on a subset of 42,374 SNPs was simultaneously fitted. That subset of SNP was derived by applying a minor allele frequency threshold of 1% to the 50k SNP-chip imputation reference (previously described). A leave-one-segment-out approach was used to avoid proximal contamination in the GWAS, whereby a 5-Mbp region flanking the sequence variant of interest was excluded from the set of SNPs used to estimate the genomic relationship matrix.

An adjusted Bonferroni threshold was adopted to determine variants with significant associations for each trait. Because a Bonferroni correction threshold based on all 14,990,779 variants is highly conservative, a modified threshold was evaluated based on the effective number of independent variants, as proposed by Duggal et al. (2008) and implemented in other studies (Zhu et al., 2017; Wang et al., 2019). The effective number of independent variants were identified using a sliding window approach in Plink software (Purcell et al., 2007), with an R^2 threshold of 0.9, a window size of 100 kb and a step size of 5 variants. These criteria resulted in a set of 2,303,435 variants and enabled the calculation of an adjusted Bonferroni threshold which considered all tests across 2,303,435 variants as independent. Based on $\alpha = 0.05$, this resulted in a nominal *P*-value of 4.3e-09 and a corresponding Bonferroni threshold of $-\log_{10}(4.3e\text{-}09) = 8.36$. Whole-genome sequence resolution genotypes within a 1Mbp window were annotated using SnpEff (version 4.3t; build 11-24-2017; Cingolani et al., 2012) and the Ensembl UMD3.1.86 gene annotations to assess the candidacy of QTL identified from the GWAS for each trait. We used a linkage disequilibrium (**LD**)-based approach to prioritize variants, similar to that described by Lopdell et al. (2017) because the association rankings of candidate variants are expected to be affected by phenotyping, genotyping, and imputation errors. Specifically, we identified QTL regions where the most highly associated variant was in high LD ($R^2 > 0.7$) with either a splice region variant, or a moderate or high impact coding variant, according to SnpEff classification.

RESULTS AND DISCUSSION

Trait Prediction Models

Cross-validation prediction model accuracies $\left(R^2_{cv}\right)$ were assessed for untreated spectra, as well as for spectra treated using SNV transformation, multiplicative scatter correction, or first-derivative treatments (Appendix Table A1). Window sizes of 15 data points (7 points either side) had consistently higher R^2_{cv} values, compared to other window sizes, so only these have been presented. Applying treatments to spectral data resulted in marginally higher R^2_{cv} values on average, compared to not treating spectra, and treating spectra with a SNV and first-derivative transformation prior to fitting PLS models resulted in the highest average R^2_{cv} value and was thus used in all further analysis. Descriptive statistics of fatty acid and protein traits, and goodness of fit measures of PLS calibration models (applied to SNV + first-derivative transformed spectra) for training and validation datasets are presented in Table 1.

For individual fatty acids, coefficient of determination values for the validation dataset $\left(R^2_{cv}\right)$ were generally higher for short-chain fatty acids (C4 to C8), ranging from 0.54 to 0.62, compared with medium-chain fatty acids (C10 to C14), which ranged from 0.30 to 0.63, and long-chain fatty acids (C16 to C18), which ranged from 0.18 to 0.57. Concentrations of individual saturated fatty acids were typically higher and had higher average R^2_{cv} values, compared with individual unsaturated fatty acids. For grouped fatty acids, R^2_{cv} values were higher for UFA and SFA groups, compared to PUFA; additionally, for fatty acids grouped by carbon chain length, the highest R^2_{cv} value of 0.65 was observed for SCFA. It is notable that although we found an overall trend of higher R^2_{cv} values coinciding with lower RPE_{cv} values, there were exceptions to this. For example, among individual fatty acids, C16:1 had a particularly low R^2_{cv} of 0.18, but an RPE_{cv} of 0.13, which was comparable to other traits such as C10:0 and C12:0, which had R^2_{cv} values of ~0.60. This highlights the difference between R^2_{cv} and RPE_{cv} as accuracy metrics, the former indicating how well the prediction model explains the variation in the directly measured trait, whereas the latter provides a comparison of how similar the predicted values are to the directly mea-

Table 1. Descriptive statistics of fatty acid and protein traits, and goodness of fit measures of partial least squares calibration models for training and validation data sets

		Trait summary[1]			Training[2]		Validation[3]	
Trait	Description and units	n	Mean	SD	R^2_t	RPE_t	R^2_{cv}	RPE_{cv}
Individual fatty acid								
C4:0	Butyric acid, g/100 g of total fat	1,963	3.90	0.32	0.706	0.043	0.602	0.053
C6:0	Caproic acid, g/100 g of total fat	1,969	2.52	0.19	0.591	0.049	0.542	0.052
C8:0	Caprylic acid, g/100 g of total fat	1,968	1.54	0.18	0.697	0.064	0.622	0.073
C10:0	Capric acid, g/100 g of total fat	1,975	3.51	0.61	0.701	0.094	0.627	0.108
C10:1	Caproleic acid, g/100 g of total fat	1,969	0.31	0.06	0.469	0.151	0.300	0.162
C12:0	Lauric acid, g/100 g of total fat	1,972	3.92	0.74	0.685	0.106	0.590	0.121
C12:1	Lauroleic acid, g/100 g of total fat	1,925	0.13	0.03	0.470	0.169	0.353	0.181
C14:0	Myristic acid, g/100 g of total fat	1,967	11.46	1.17	0.599	0.065	0.491	0.073
C14:1	Myristoleic acid, g/100 g of total fat	1,970	0.75	0.23	0.517	0.211	0.414	0.233
C16:0	Palmitic acid, g/100 g of total fat	1,977	27.64	3.27	0.633	0.073	0.574	0.076
C16:1	Palmitoleic acid, g/100 g of total fat	1,958	1.54	0.22	0.301	0.123	0.184	0.132
C18:0	Stearic acid, g/100 g of total fat	1,968	11.95	2.00	0.544	0.115	0.445	0.124
C18:1 *cis*-7	*cis*-Vaccenic acid, g/100 g of total fat	1,936	4.53	0.70	0.531	0.107	0.411	0.118
C18:1 *cis*-9	Oleic acid, g/100 g of total fat	1,963	17.31	2.55	0.653	0.088	0.569	0.096
C18:2 *cis*-9,*trans*-11	Conjugated linoleic acid, g/100 g of total fat	1,929	0.87	0.25	0.587	0.185	0.498	0.210
C18:2 *cis*-6	Linoleic acid, g/100 g of total fat	1,963	1.20	0.14	0.561	0.078	0.480	0.085
C18:3 *cis*-3	α-Linolenic acid, g/100 g of total fat	1,954	0.80	0.11	0.387	0.112	0.360	0.105
Grouped fatty acid[4]								
SFA	Saturated fatty acids, g/100 g of total fat	1,965	70.59	3.08	0.703	0.024	0.591	0.028
PUFA	Polyunsaturated fatty acids, g/100 g of total fat	1,972	4.16	0.46	0.641	0.065	0.490	0.081
UFA	Unsaturated fatty acids, g/100 g of total fat	1,964	29.42	3.08	0.711	0.057	0.597	0.066
SCFA	Short-chain fatty acids, g/100 g of total fat	1,970	7.96	0.59	0.695	0.041	0.648	0.043
MCFA	Medium-chain fatty acids, g/100 g of total fat	1,969	20.09	2.43	0.659	0.071	0.567	0.080
LCFA	Long-chain fatty acids, g/100 g of total fat	1,974	36.82	4.45	0.609	0.076	0.568	0.079
Individual milk protein								
α-CN	α-Casein, g/L of total volume	1,695	15.79	1.76	0.585	0.072	0.532	0.076
β-CN	β-Casein, g/L of total volume	1,686	14.78	1.84	0.128	0.116	0.190	0.113
κ-CN	κ-Casein, g/L of total volume	1,687	4.24	0.59	0.575	0.087	0.476	0.105
α-LA	α-Lactalbumin, g/L of total volume	1,942	1.21	0.15	0.379	0.099	0.306	0.104
β-LG	β-Lactoglobulin, g/L of total volume	1,959	3.84	0.70	0.773	0.087	0.678	0.104
Lf [5]	Lactoferrin, g/L of total volume	1,936	0.51	0.12	0.411	0.188	0.356	0.194

[1]n = number of samples.

[2]R^2_t = coefficient of determination between actual and predicted trait values in the training dataset; RPE_t = relative prediction error between actual and predicted trait values in the training dataset.

[3]R^2_{cv} = coefficient of determination between actual and predicted trait values in the validation dataset; RPE_{cv} = relative prediction error between actual and predicted trait values in the validation dataset.

[4]SCFA = short-chain fatty acids, sum of C4:0, C6:0, and C8:0; MCFA = medium-chain fatty acids, sum of 10:0, 10:1, 12:0, 12:1, 14:0, and 14:1; LCFA = long-chain fatty acids, sum of C18 fatty acids.

[5]Cube-root transformation of lactoferrin (Lf).

sured trait values. In the present study, most comparisons of accuracy with other studies will be based on R^2_{cv} values because that is the accuracy metric that is most commonly reported; however, the example above shows that other metrics can be valuable for assessing FT-MIR prediction model accuracy.

The R^2_{cv} values we report are consistent with those from previous studies where fatty acids were expressed as a proportion of total fat content, with our values being similar to those reported by Soyeurt et al. (2006), but lower than those reported in other studies (Rutten et al., 2009; Lopez-Villalobos et al., 2014; Bonfatti et al., 2016). In the present study, for grouped SCFA, MCFA, and LCFA, R^2_{cv} values were lower than in other studies (Rutten et al., 2009; Lopez-Villalobos et al., 2014; Bonfatti et al., 2016). Accuracies for fatty acids predicted using FT-MIR spectra were variable in previous studies and were affected by factors such as the production system and the breed composition diversity present in calibration samples, the number of samples used to develop calibration equations, and the variability of fatty acid composition present in the calibration samples. Rutten et al. (2009) demonstrated that increasing the number of observations used in the calibration equations resulted in better predictions for fat composition. Soyeurt et al. (2006, 2011) demonstrated

that prediction accuracy could be improved by increasing the sample size of their study, and by increasing the range of variation present in the fatty acids. Importantly, studies with the highest accuracies were those where the range of fatty acid values present in the validation samples were encompassed within the range of fatty acid values represented in calibration samples.

For individual milk proteins, R^2_{cv} values were generally lower than for fatty acids, ranging from 0.19 for β-CN to 0.68 for β-LG. Notably, although the R^2_{cv} values for β-CN and β-LG were very different, the RPE_{cv} values for these 2 traits were similar (0.11 and 0.10, respectively). The R^2_{cv} values we report for individual milk proteins were typically higher than those reported in previous studies of individual milk proteins, with the exceptions of β-CN and lactoferrin (**Lf**), which were consistently lower here than in other studies (De Marchi et al., 2009; Lopez-Villalobos et al., 2009; Rutten et al., 2011; Soyeurt et al., 2012; Bonfatti et al., 2016; McDermott et al., 2016). Fuentes-Pila et al. (1996) suggest that a RPE of lower than 0.1 is an indicator of satisfactory prediction, a RPE between 0.1 to 0.2 is an indicator of relatively good or acceptable predictions, and a RPE greater than 0.2 is an indicator of unsatisfactory prediction. Based on these criteria, 21 of 23 individual and grouped fatty acids and all 6 protein fractions had good or satisfactory predictions in the validation datasets. Although the guidelines proposed by Fuentes-Pila et al. (1996) are useful as an indicator of prediction acceptability, they are potentially less meaningful when we are considering the value of incorporating FT-MIR predicted traits into animal breeding programs. This is because FT-MIR predictions can provide indicator traits across large numbers of animals at little or no cost, whereas it may be infeasible to directly measure these traits across even a small number of animals. Moreover, when we are considering the potential for incorporating an FT-MIR predicted trait into a breeding program, we are not only interested in the phenotypic correlation between the directly measured and FT-MIR predicted trait, but also the heritability of the FT-MIR predicted trait, and the genetic correlation between the directly measured and FT-MIR predicted trait.

Genetic Parameters of Directly Measured and FT-MIR Predicted Traits

Estimates of variance components for directly measured and FT-MIR predicted fatty acid and protein traits are shown in Table 2 and Appendix Table A2. Heritability estimates (h^2) for the majority of traits were moderate to high, with 17 of the directly measured traits and 20 of the FT-MIR predicted traits having an estimated heritability greater than 0.3. Because this is an F_2 study, genetic variances will include a segregation variance component that would typically inflate these values compared to what would be seen in a study of purebred animals. In general, lower heritability and repeatability estimates were observed for directly measured traits, compared to FT-MIR predicted traits. This was driven by higher total variation $\left(\sigma^2_T\right)$ in the directly measured traits, coupled with a lower magnitude increase in the additive genetic variance component $\left(\sigma^2_u\right)$, compared to the FT-MIR predicted traits. Despite this, the genetic correlations between measured and predicted traits remained high and were mostly greater than 0.75.

Fatty Acid Traits. In fatty acid traits, the lowest heritability estimates were observed for C18:0 and LCFA, with heritability estimates of 0.07 for the directly measured traits, and heritability estimates of 0.14 and 0.15 in the FT-MIR predicted traits, respectively. Although heritability estimates were typically higher in the FT-MIR predicted traits, there were exceptions to this. In particular, C14:1 had an estimated heritability for the measured trait that was substantially higher than that of the FT-MIR predicted trait (0.55 vs. 0.26). Genetic correlations between directly measured and FT-MIR predicted traits ($\boldsymbol{r_a}$) were greater than 0.85 for 18 of 23 individual and grouped fatty acids, and for 11 of these traits, the genetic correlation was greater than 0.95. The lowest genetic correlations were observed for C18:0 ($r_a = 0.72$) and C18:3 *cis*-3 ($r_a = 0.74$). In general, we found a consistent trend for individual and grouped fatty acids, where lower genetic correlations coincided with low R^2_{cv} values.

Although several studies have reported genetic parameter estimates for directly measured or FT-MIR predicted fatty acid traits, or both, these studies vary in the specific individual fatty acids (if any) presented, and whether or not they present parameter estimates for grouped fatty acids. Many studies report genetic parameter estimates for FT-MIR predicted traits only (Soyeurt et al., 2007b; Lopez-Villalobos et al., 2014; Narayana et al., 2017; Fleming et al., 2018), with only 2 studies reporting genetic parameters for both directly measured and FT-MIR predicted traits (Rutten et al., 2010; Bonfatti et al., 2017b). These latter 2 studies also report genetic correlations between directly measured and FT-MIR predicted fatty acids, with Bonfatti et al. (2017b) presenting these for individual and grouped fatty acids, whereas Rutten et al. (2010) presented these for individual fatty acids only.

The heritability, repeatability, and genetic correlation estimates we report in the present study were broadly

Table 2. Variance component estimates for directly measured and Fourier-transform mid-infrared (FT-MIR) predicted fatty acid and protein traits

	Directly measured trait[2]				FT-MIR prediction[3]				
Trait[1]	σ_u^2	σ_T^2	h^2	t	σ_u^2	σ_T^2	h^2	t	r_a
Individual fatty acid (g/100 g of total fat)									
C4:0	0.022	0.069	0.31 (0.12)	0.52 (0.03)	0.014	0.042	0.34 (0.13)	0.57 (0.03)	0.988 (0.014)
C6:0	0.005	0.025	0.20 (0.10)	0.35 (0.03)	0.003	0.011	0.24 (0.11)	0.45 (0.03)	0.925 (0.099)
C8:0	0.005	0.019	0.29 (0.11)	0.44 (0.03)	0.003	0.009	0.33 (0.12)	0.45 (0.03)	0.983 (0.020)
C10:0	0.098	0.241	0.41 (0.14)	0.54 (0.03)	0.057	0.125	0.46 (0.14)	0.52 (0.03)	0.986 (0.027)
C10:1	0.001	0.003	0.33 (0.13)	0.54 (0.03)	3e-4	0.001	0.27 (0.10)	0.42 (0.03)	0.811 (0.124)
C12:0	0.132	0.378	0.35 (0.13)	0.52 (0.03)	0.083	0.197	0.42 (0.14)	0.53 (0.03)	0.996 (0.017)
C12:1	2e-4	0.001	0.24 (0.10)	0.33 (0.03)	1e-4	0.0003	0.25 (0.10)	0.41 (0.03)	0.849 (0.125)
C14:0	0.342	0.997	0.34 (0.14)	0.47 (0.03)	0.161	0.449	0.36 (0.13)	0.41 (0.04)	0.947 (0.043)
C14:1	0.021	0.037	0.55 (0.17)	0.71 (0.02)	0.003	0.012	0.26 (0.11)	0.42 (0.03)	0.866 (0.100)
C16:0	2.187	5.782	0.38 (0.12)	0.58 (0.03)	1.214	3.123	0.39 (0.12)	0.46 (0.03)	0.954 (0.058)
C16:1	0.008	0.043	0.20 (0.10)	0.48 (0.03)	0.002	0.011	0.16 (0.08)	0.38 (0.03)	0.773 (0.173)
C18:0	0.176	2.714	0.07 (0.05)	0.48 (0.02)	0.149	1.034	0.14 (0.08)	0.37 (0.03)	0.718 (0.259)
C18:1 *cis*-7	0.125	0.412	0.30 (0.12)	0.51 (0.03)	0.063	0.193	0.33 (0.12)	0.51 (0.03)	0.947 (0.040)
C18:1 *cis*-9	0.881	3.986	0.22 (0.09)	0.41 (0.03)	0.551	1.955	0.28 (0.11)	0.42 (0.03)	0.986 (0.024)
C18:2 *cis*-9,*trans*-11	0.017	0.048	0.35 (0.13)	0.60 (0.03)	0.010	0.023	0.46 (0.16)	0.62 (0.03)	0.939 (0.047)
C18:2 *cis*-6	0.004	0.013	0.33 (0.12)	0.45 (0.03)	0.002	0.006	0.32 (0.12)	0.44 (0.03)	0.904 (0.077)
C18:3 *cis*-3	0.004	0.009	0.40 (0.13)	0.46 (0.03)	0.001	0.002	0.45 (0.12)	0.51 (0.03)	0.743 (0.144)
Grouped fatty acid (g/100 g of total fat)									
SFA	1.472	6.175	0.24 (0.09)	0.49 (0.03)	1.293	3.469	0.37 (0.14)	0.56 (0.03)	0.977 (0.035)
PUFA	0.078	0.181	0.43 (0.14)	0.57 (0.03)	0.049	0.105	0.46 (0.15)	0.63 (0.03)	0.980 (0.019)
UFA	1.468	6.167	0.24 (0.09)	0.49 (0.03)	1.299	3.474	0.37 (0.14)	0.56 (0.03)	0.975 (0.037)
SCFA	0.037	0.196	0.19 (0.09)	0.40 (0.03)	0.026	0.101	0.26 (0.12)	0.51 (0.03)	0.961 (0.040)
MCFA	1.293	4.206	0.31 (0.12)	0.45 (0.03)	0.797	2.158	0.37 (0.13)	0.46 (0.03)	0.974 (0.040)
LCFA	0.852	11.70	0.07 (0.05)	0.40 (0.03)	0.813	5.301	0.15 (0.08)	0.36 (0.03)	0.925 (0.099)
Individual milk protein (g/L of total volume)									
α-CN	0.579	2.029	0.29 (0.12)	0.45 (0.03)	0.559	1.109	0.50 (0.18)	0.61 (0.03)	0.941 (0.067)
β-CN	0.421	3.105	0.14 (0.07)	0.17 (0.03)	0.204	0.537	0.38 (0.15)	0.65 (0.03)	0.802 (0.222)
κ-CN	0.172	0.315	0.55 (0.18)	0.57 (0.04)	0.083	0.162	0.51 (0.16)	0.68 (0.03)	0.956 (0.050)
α-LA	0.008	0.019	0.42 (0.14)	0.51 (0.03)	0.002	0.005	0.39 (0.14)	0.56 (0.03)	0.755 (0.130)
β-LG	0.282	0.448	0.63 (0.18)	0.80 (0.02)	0.240	0.343	0.70 (0.19)	0.80 (0.02)	0.995 (0.006)
Lf[4]	0.007	0.012	0.59 (0.17)	0.61 (0.03)	0.001	0.003	0.30 (0.12)	0.45 (0.03)	0.771 (0.148)

[1]Trait definitions and units as described in Table 1. Standard errors shown in parentheses.

[2]σ_u^2 = additive genetic variance; σ_T^2 = total variance $\left(\sigma_u^2 + \sigma_p^2 + \sigma_e^2\right)$; h^2 = heritability estimate; t = repeatability estimate.

[3]r_a =genetic correlation between directly measured and FT-MIR predicted trait.

[4]Cube-root transformation of lactoferrin (Lf).

consistent with those from previous studies. For directly measured fatty acids, the heritability estimates we report were typically higher than those reported by Bonfatti et al. (2017b), but lower than those reported by Rutten et al. (2010). For FT-MIR predicted fatty acids, the heritability and repeatability estimates we report for individual and grouped fatty acids were similar to those presented by Lopez-Villalobos et al. (2014), but lower than those presented by Narayana et al. (2017) and higher than those presented in other studies (Soyeurt et al., 2007b; Bonfatti et al., 2017b). Compared with other studies that report genetic correlations between directly measured and FT-MIR predicted fatty acids, the genetic correlations we report were similar, with standard errors of a similar magnitude (Rutten et al., 2010; Bonfatti et al., 2017b). The moderate to high heritability estimates we report, alongside high genetic correlations between directly measured and FT-MIR predicted fatty acid traits, indicate that there is genetic variation in the FT-MIR predicted traits that could potentially be exploited in animal breeding programs, and, in most cases, that selection for an FT-MIR predicted fatty acid trait would be expected to provide genetic gain in the actual fatty acid trait of interest.

Individual Milk Protein Traits. Heritability estimates were moderate to high for nearly all directly measured and FT-MIR predicted individual milk proteins (Table 2). The exception to this was for directly measured β-CN, which had a heritability of 0.14. The highest heritability estimates were for β-LG, with $h^2 = 0.63$ and $h^2 = 0.70$ for directly measured and FT-MIR predicted β-LG, respectively. In general, heritability estimates for measured and FT-MIR predicted proteins were similar. An exception to this was β-CN, which had

heritability estimates for the directly measured and FT-MIR predicted trait of 0.14 and 0.38, respectively. Another exception was Lf, which had an estimated heritability for the measured trait that was substantially higher than that of the FT-MIR predicted trait (0.59 vs. 0.30). With the exceptions of α-LA and Lf, genetic correlations between directly measured and FT-MIR predicted individual milk proteins were greater than 0.8. In general, as we observed for fatty acid traits, we found a trend of low R^2_{cv} values, coinciding with low genetic correlations between directly measured and FT-MIR predicted traits.

There are few studies that report genetic parameters for directly measured or FT-MIR predicted milk proteins, or both, but those studies vary in the breed composition of the cows. Specifically, study populations include Dutch Holstein-Friesians (Schopen et al., 2009), Danish Holsteins and Jerseys (Buitenhuis et al., 2016), Italian Simmentals (Bonfatti et al., 2017b), or French Montbéliarde, Normande, and Holstein cows (Sanchez et al., 2017a). Studies also vary in that some report on individual proteins as a proportion of total protein or whey protein (Schopen et al., 2009; Buitenhuis et al., 2016), whereas other studies report on individual proteins as a proportion of total protein or as a proportion of total milk volume (Bonfatti et al., 2017b; Sanchez et al., 2017a). The heritability estimates we report for directly measured α-CN, β-CN, and κ-CN were lower than those previously reported by Bonfatti et al. (2017b), but the heritability estimates we report for directly measured α-LA and β-LG were substantially higher. In contrast, for FT-MIR predicted α-CN, β-CN, and κ-CN, the heritability estimates we report were consistently higher than those reported by Bonfatti et al. (2017b), but were similar to those reported by Sanchez et al. (2017a).

The only study to report genetic correlations between directly measured and FT-MIR predicted milk proteins that we are aware of is that of Bonfatti et al. (2017b). The genetic correlations that we report were higher than in that study. Specifically, for the protein fractions we studied, genetic correlations ranged from 0.76 for α-LA to 0.995 for β-LG, whereas in Bonfatti et al. (2017b), genetic correlations for these traits ranged from 0.24 for α-LA to 0.48 for β-LG. Interestingly, Bonfatti et al. (2017b) reported moderate heritability estimates for directly measured milk proteins (0.12 to 0.59), but much lower heritability estimates for FT-MIR predicted milk proteins (0.07 to 0.21). In contrast, the heritability estimates we observed for directly measured proteins (0.14 to 0.63) were similar to (and often lower than) the heritability estimates we observed for FT-MIR predicted proteins (0.30 to 0.70). These differences in heritability were likely due to factors related to differences in the breed composition and population structure between the 2 studies (i.e., Italian Simmental cows from herds enrolled in the Italian national milk recording program versus Holstein-Friesian Jersey F_2 cows from a single research herd).

Moderate to high heritability estimates and high genetic correlations between directly measured and FT-MIR predicted milk proteins in our study indicate that indirect selection on FT-MIR predicted milk proteins could be used in animal breeding programs to achieve desired changes to milk protein composition. Moreover, high genetic correlations from pedigree-based models imply that directly measured and FT-MIR predicted traits may have a similar underlying genetic architecture and that genes contributing to the traits are likely to be co-inherited (Lynch and Walsh, 1998). To assess this directly, we conducted GWAS on directly measured traits and their corresponding FT-MIR predictions, and compared the QTL for each trait.

Sequence-Based Genome-Wide Association Analyses

Previously, there have been several GWAS that used a range of genotype densities for fatty acids in bovine milk samples determined by GC (Bouwman et al., 2011; Buitenhuis et al., 2014; Palombo et al., 2018) or fatty acids predicted using FT-MIR spectra (Olsen et al., 2017; Cruz et al., 2019; Iung et al., 2019; Sanchez et al., 2019; Freitas et al., 2020). Similarly, multiple GWAS have been conducted on protein fractions in milk samples determined by HPLC (Schopen et al., 2011; Buitenhuis et al., 2016; Pegolo et al., 2018; Zhou et al., 2019) or FT-MIR predicted protein fractions (Sanchez et al., 2017b, 2019). Each of those studies was conducted using either the directly measured trait (GC-based for fatty acids; HPLC-based for protein fractions) or the FT-MIR predicted trait, though none of these presented comparisons between the GWAS for directly measured and FT-MIR predicted traits. In the present study, we have sought to make these comparisons using imputed whole-genome sequence genotypes from an F_2 study population to enhance our ability to identify trait QTL and candidate causative mutations.

For each of 17 individual fatty acids, 6 grouped fatty acids, and 6 protein traits, GWAS were conducted using 14,990,779 imputed sequence variants. These analyses resulted in the identification of 40,946 variants with significant effects for directly measured traits, and 18,843 variants with significant association effects for the FT-MIR predicted traits. We found more than twice as many variants with significant effects for

directly measured traits, compared with FT-MIR predicted traits, which was largely due to 20,949 variants with significant effects on BTA26 for directly measured traits compared with only 110 variants with significant effects on BTA26 for FT-MIR predicted traits. It was also notable that we detected 3,579 variants with significant effects on BTA22 for directly measured Lf, but no variants with significant effects on BTA22 for FT-MIR predicted traits. Manhattan plots showing the strength of association signals are presented in Figures 1-4 for individual fatty acids, Figure 5 for grouped fatty acids, and Figure 6 for individual protein traits. To assess the candidacy of QTL, relevant protein coding variants that were in high LD ($R^2 > 0.7$) with the most highly associated variant from each peak were identified. The most highly associated variant from each trait QTL and any relevant protein coding variants are shown in Table 3 for directly measured fatty acid and protein traits, and Table 4 for FT-MIR predicted fatty acid and protein traits. Effect sizes and minor allele frequency details for relevant variants and effects are provided in Appendix Table A3 for fatty acids and Appendix Table A4 for protein traits.

Short-Chain Fatty Acids. Prominent peaks were observed on BTA17 for the short-chain fatty acids, C4:0, and C6:0 (Tables 3 and 4; Figure 1). For directly measured and FT-MIR predicted C4:0, these peaks were underpinned by the same QTL at chromosome (**Chr**) 17:53.03 Mbp (rs461037541). A peak of similar magnitude was also observed for FT-MIR predicted C6:0 at a nearby locus (rs207997694), with a less significant peak for directly measured C6:0 at that same locus. Other significant effects were also observed at this locus for directly measured SCFA (*P*-value = 1.2e-14) and FT-MIR predicted SCFA (*P*-value = 7.1e-22). The 2 implicated loci for the peaks on BTA17 were situated between the *AACS* and *BRI3BP* genes, and visual examination revealed several significant variants across both genes. The *AACS* gene codes for the enzyme acetoacetyl-CoA synthetase, which forms an important metabolic link between the ketone body acetoacetate on one hand, and the tricarboxylic acid cycle and fat synthesis on the other (Bergman, 1971). As this gene is highly expressed in both adipose and mammary tissue (NCBI Bioprojects PRJEB4337 and PRJEB2445), *AACS* makes a good candidate for the causal gene underlying fatty acid QTL in this region. Knutsen et al. (2018) also reported an effect for C4:0 fatty acids in this region and suggested that the QTL may be the result of a regulatory effect.

Medium-Chain Fatty Acids. Significant effects were observed on BTA11, BTA19, and BTA26 for medium-chain fatty acids (Tables 3 and 4; Figure 2). The peak on BTA11 was underpinned by a Chr11:103.30 locus (rs41255687) and was observed for FT-MIR predicted C12:1, but was absent for directly measured C12:1. This locus was in high LD ($R^2 > 0.98$) with 2 missense mutations in the *PAEP* gene, which encodes the major whey protein, β-LG. One of the missense mutations reported (rs109625649; V134A) is a variant that distinguishes the 'A' and 'B' haplotypes of β-LG (Caroli et al., 2009), where the 'A' haplotype is known to be associated with higher levels of β-LG expression. The *PAEP* locus has also been linked to FT-MIR wavenumbers characterized by carboxylic C=O bond stretching (Tiplady et al., 2021). This type of bond is found in both fats and proteins, strongly suggesting that the peak observed for the FT-MIR predicted phenotype is a false positive due to contamination of the signal by varying levels of β-LG expression.

Several QTL were identified for directly measured and FT-MIR predicted medium-chain fatty acids (C10:0, C12:0, C14:0) on BTA19 that were in high LD ($R^2 > 0.97$), with a missense mutation (rs41921160) in the *CCDC57* gene (Tables 3 and 4; Figure 2). Significant effects were also observed in this region for FT-MIR predicted C8:0 (*P*-value = 8.9e-10; Figure 1), and directly measured (*P*-value = 1.4e-13) and FT-MIR predicted MCFA (*P*-value = 9.2e-13; Figure 5). The *CCDC57* encodes a coiled-coil domain-containing protein that is expressed in the bovine mammary gland (Medrano et al., 2010). Previous studies have implicated the same or a nearby locus to the one reported here as having a significant association for fatty acids (Bouwman et al., 2014; Knutsen et al., 2018; Palombo et al., 2018) and fat composition (Tribout et al., 2020) in bovine milk. Significant effects have also been reported at a nearby locus for several FT-MIR wavenumbers, characterized by carboxylic C=O bond stretching (Tiplady et al., 2021). Bouwman et al. (2014) examined this region in depth using HD genotypes and identified 2 possible genes underlying an effect for C14:0, *CCDC57* and *FASN*. The missense mutation we have highlighted (rs41921160) is located within the same region as the most highly associated variants in the study by Bouwman et al. (2014), and was in perfect LD with the set of 8 intronic HD variants with the most highly associated effects. On closer examination of the association effects for C10:0, C12:0, and C14:0 in our study, we determined that alongside the most highly associated variants in the QTL peaks, there were 47 other imputed whole-genome sequence variants between 51,306,219 and 51,330,072 bp that were in perfect LD with one another (including the missense variant rs41921160), with only marginally less significant *P*-values. A small cluster of association effects near to or in the *FASN* gene were also observed, with the most significant of these for directly measured C14:0 being at 51,380,689 bp, but the *P*-value for that

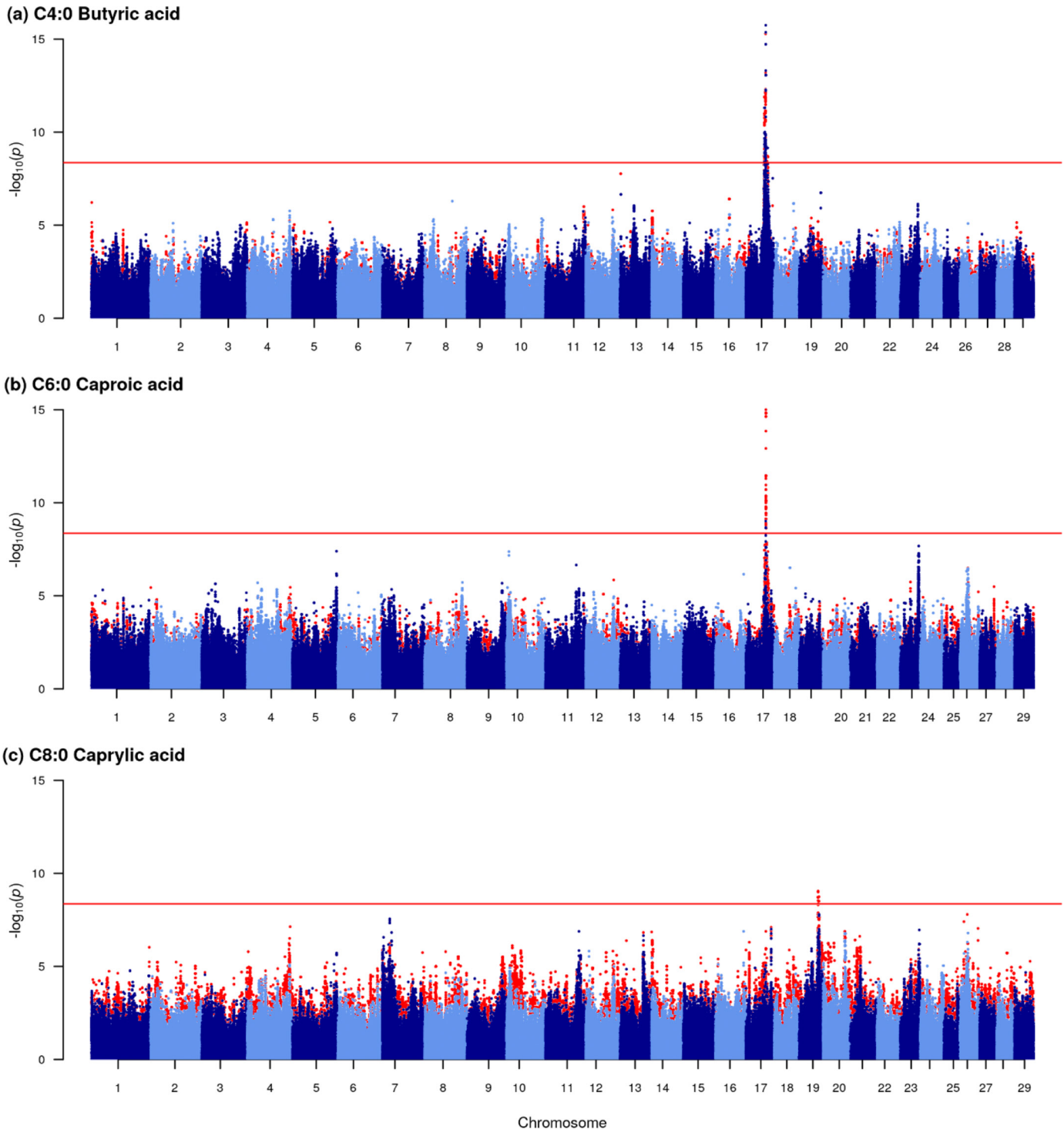

Figure 1. Manhattan plots showing association effects for directly measured (GC-based) and Fourier-transform mid-infrared (FT-MIR) predicted individual short-chain fatty acid traits. Dark and light blue data points represent association signals for GC-based traits and red data points represent association signals for FT-MIR predicted traits. Chromosomes and genomic position based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals is represented as the $-\log_{10}$(*P*-value) on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}$(4.3e-09).

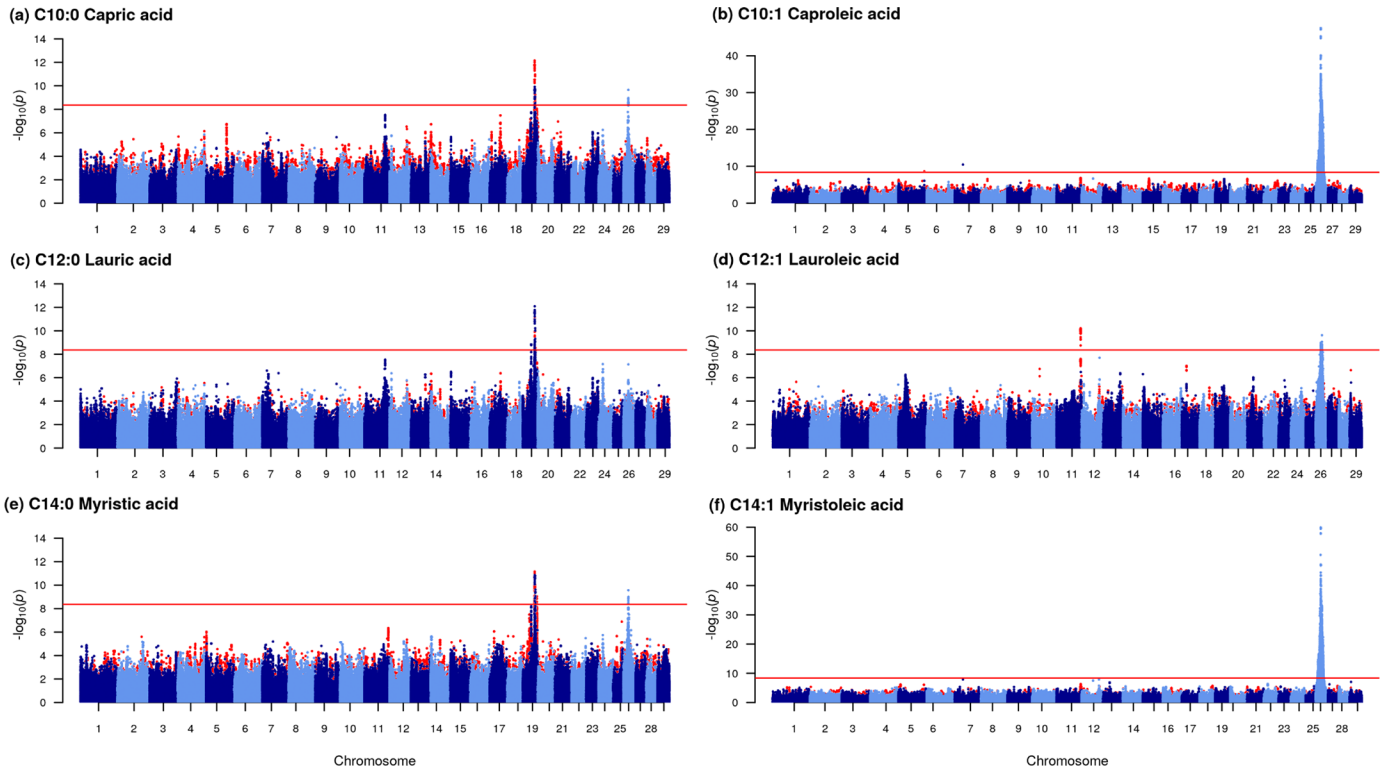

Figure 2. Manhattan plots showing association effects for directly measured (GC-based) and Fourier-transform mid-infrared (FT-MIR) predicted individual medium-chain fatty acid traits. Dark and light blue data points represent association signals for GC-based traits and red data points represent association signals for FT-MIR predicted traits. Chromosomes and genomic position based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals is represented as the $-\log_{10}$(*P*-value) on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}$(4.3e-09).

effect was not significant (*P*-value = 2.4e-07). To assess whether multiple QTL were present in this region, we repeated the GWAS, correcting for the rs136424304 locus by including it as a covariate in the Bolt-LMM model. This resulted in no significant effects remaining in a 1 Mbp region around the Chr19:51.32 locus, and the association effect near the *FASN* gene at 51,380,689 bp, dropping in significance to a *P*-value of 3.9e-02. Although our analysis provides evidence that the effect in this region may be underpinned by a missense mutation in the *CCDC57* gene, the functional candidacy of *FASN* remains and such an effect would need to be confirmed by functional analysis.

Multiple QTL were identified for directly measured medium-chain fatty acids on BTA26 (Table 3; Figure 2). The most significant effects were observed at Chr26: 21.15 Mbp for directly measured C10:1 (rs41255688; *P*-value=1.8e-48) and C14:1 (rs385285356; *P*-value = 6.1e-61). These loci were in high LD ($R^2 = 0.92$) with a splice region variant (rs41255693) in the *SCD* gene. The *SCD* gene was also identified as encompassing other effects with less significant *P*-values for directly measured C10:0, C14:0, SFA, and UFA (Table 3), and FT-MIR predicted UFA (Table 4). Stearoyl-CoA desaturase is an enzyme that plays an important role in biosynthesis of monounsaturated fatty acids (Bernard et al., 2006; Paton and Ntambi, 2009), and has previously been reported in other studies of fatty acids in bovine milk (Mele et al., 2007; Moioli et al., 2007; Schennink et al., 2008; Kgwatalala et al., 2009; Conte et al., 2010; Bouwman et al., 2011). The strong effect we see for directly measured C14:1 in the *SCD* gene is unsurprising, given that C14:0 in milk fat is predominantly derived from de novo synthesis in the mammary gland, meaning that almost all C14:1 *cis*-9 is likely to have been synthesized by stearoyl-CoA desaturase (Bernard et al., 2006). Interestingly, although we found a significant effect for FT-MIR predicted UFA at a nearby locus that was also in high LD with the rs41255693 splice region variant ($R^2 = 0.91$), no other effects were identified within the *SCD* gene for individual FT-MIR predicted fatty acids. A peak for FT-MIR predicted C14:1 was tagged by a nearby Chr26:21 .17 Mbp locus (rs209445650; Table 4). However, the LD between the rs209445650 locus and the splice region variant identified for directly measured fatty acids

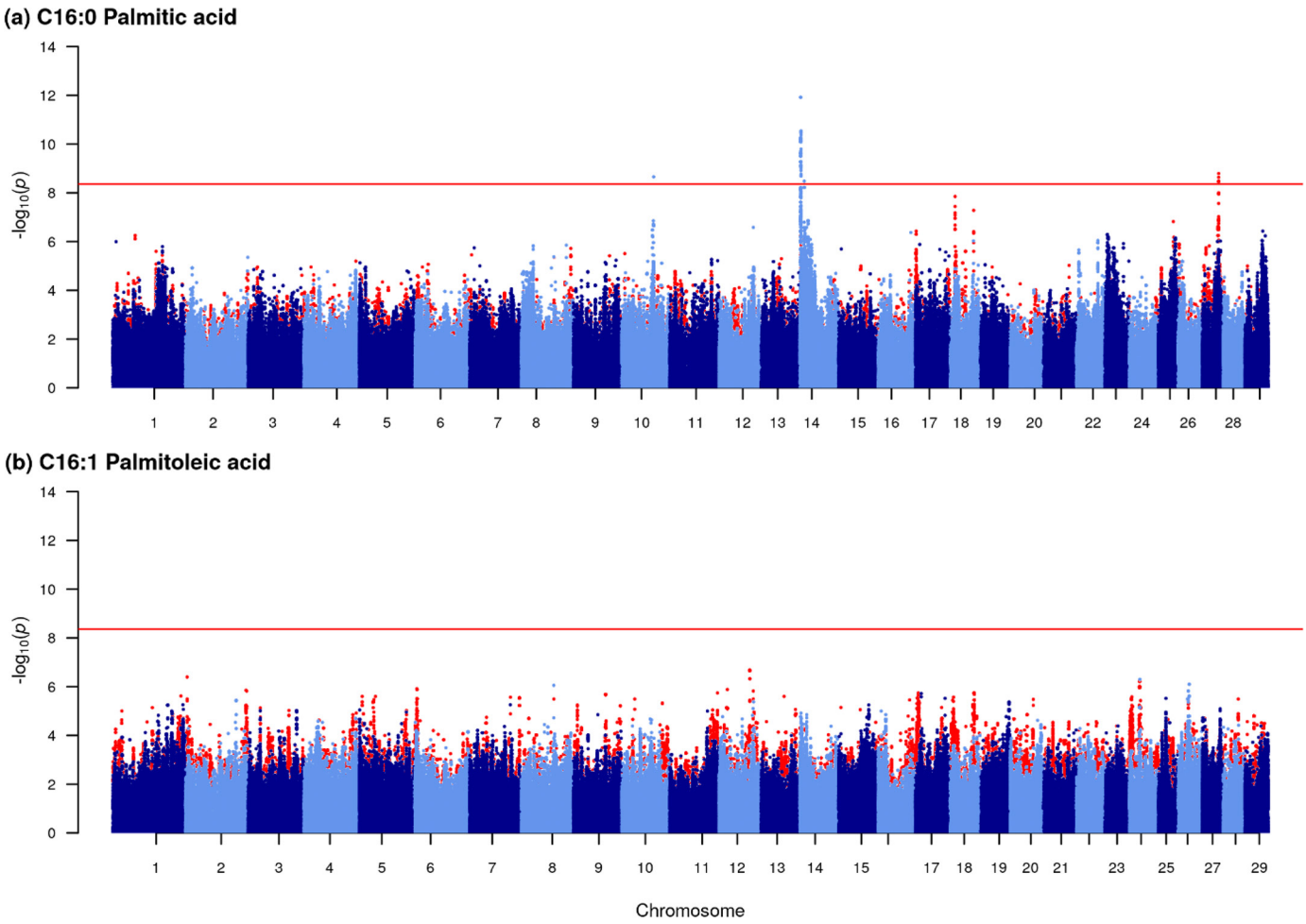

Figure 3. Manhattan plots showing association effects for directly measured (GC-based) and Fourier-transform mid-infrared (FT-MIR) predicted C16 fatty acid traits. Dark and light blue data points represent association signals for GC-based traits and red data points represent association signals for FT-MIR predicted traits. Chromosomes and genomic position based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals is represented as the $-\log_{10}$(*P*-value) on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}$(4.3e-09).

(rs41255693) was moderately low (R^2 = 0 .32). Moreover, in a recent GWAS of individual FT-MIR wavenumbers, there was no evidence of an association effect linked to the *SCD* gene (Tiplady et al., 2021), indicating that changes in milk composition due to this gene may be difficult to detect using FT-MIR spectral data. However, we may also view the absence of FT-MIR predicted trait QTL in the *SCD* gene within the context of trait prediction accuracy. The largest QTL underpinned by *SCD* in directly measured fatty acids were for C10:1 (*P*-value = 1.8e-48) and C14:1 (*P*-value = 6.1e-61). The prediction accuracies for these traits were relatively poor: C10:1 $\left(R^2_{cv} = 0.30; RPE_{cv} = 0.16\right)$ and C14:1 $\left(R^2_{cv} = 0.41; RPE_{cv} = 0.23;\right.$ Table 1). Also, it is notable that for C10:1 and C14:1, the heritability estimates of the FT-MIR predictions were lower than those for direct measurements of these traits. This contrasts with the typical pattern for nearly all other fatty acids where the heritability for the FT-MIR prediction was greater than the heritability for the directly measured trait. In particular, the heritability estimate for directly measured C14:1 was 0.55, whereas the heritability estimate for FT-MIR predicted C14:1 was 0.26 (Table 2). Low prediction accuracy and a substantially lower heritability estimate for FT-MIR predicted C14:1 may in part be explained by C14:1 being at relatively low concentrations in milk samples, particularly compared with saturated fatty acids. Specifically, C14:1 had a mean concentration of 0.75 g/100 g of total fat, compared to mean concentrations of 1.54 to 27.64 g/100 g of total fat for the individual saturated fatty acids included in this study (Table 1). Potentially, it may be possible to improve trait prediction accuracies, herita-

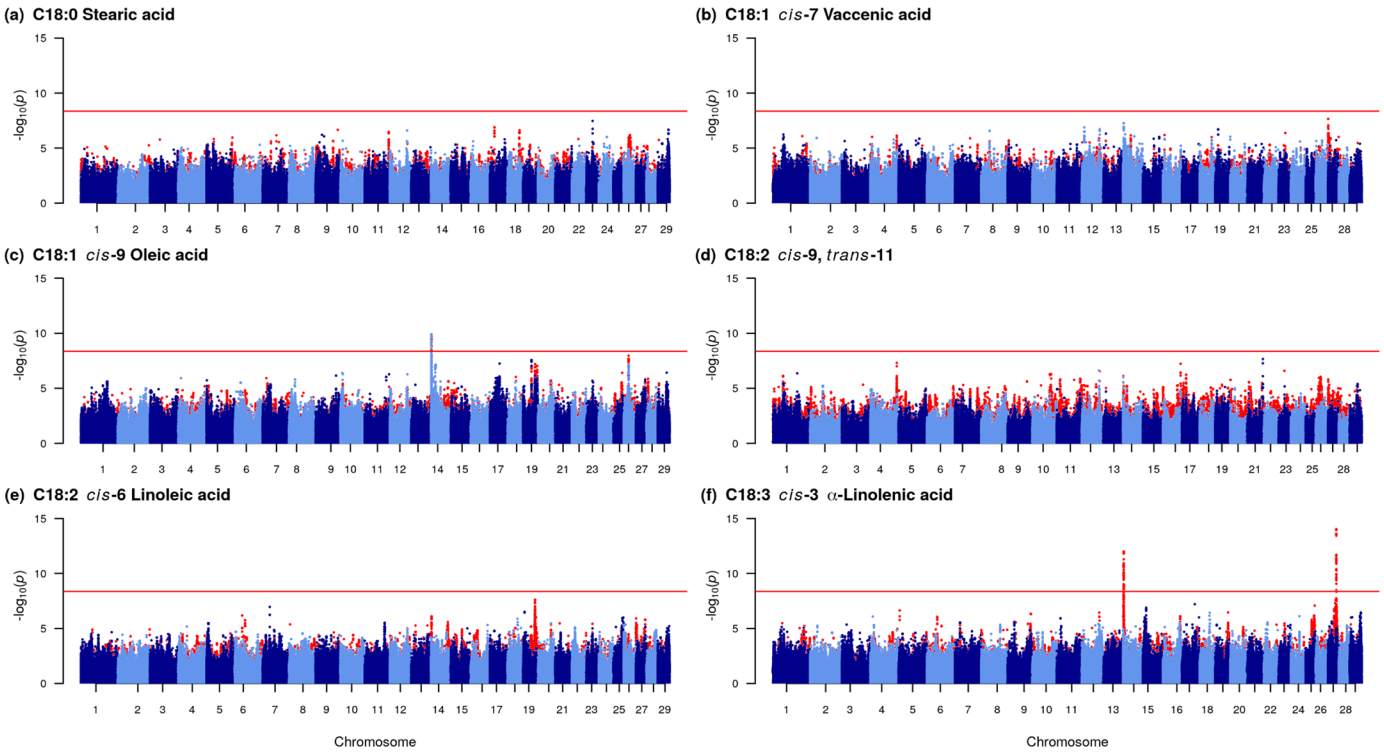

Figure 4. Manhattan plots showing association effects for directly measured (GC-based) and Fourier-transform mid-infrared (FT-MIR) predicted C18 fatty acid traits. Dark and light blue data points represent association signals for GC-based traits and red data points represent association signals for FT-MIR predicted traits. Chromosomes and genomic position based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals is represented as the $-\log_{10}$(*P*-value) on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}$(4.3e-09).

bility estimates, and QTL identification for C14:1 by basing FT-MIR predictions on the ratio of C14:1 to C14:0, as in the study by Arnould et al. (2009a). In that study, they highlight that genetic variation and heritability estimates change throughout lactation for the ratio of C14:1 to C14:0, so it may also be valuable to examine other methods of accounting for stage of lactation such as using Legendre polynomials within random regression models.

One further QTL was observed for directly measured C10:1 and C12:1 on BTA26 at a Chr26:26.46 Mbp locus (rs445758306; Table 3; Figure 2). This locus was in high LD ($R^2 = 0.76$) with a missense mutation in the *ITPRIP* gene (rs379463458). The *ITPRIP* gene modulates intracellular messaging by binding the inositol 1,4,5-triphosphate receptor ITPR. This gene has not previously been reported in GWAS of bovine milk composition, and the potential role it may play in the regulation of bovine milk fatty acids is unclear. An alternative potential candidate gene that the Chr26:26.46 Mbp locus maps close to is *SORCS3*, which encodes the sortilin-related receptor SorCS3. Sortilins are involved in regulating glucose transport into cells in response to insulin (Huang et al., 2013). A potential mechanism by which this gene could influence milk fatty acid concentrations is via changing the supply of glucose available for the pentose phosphate pathway, which in turn provides the nicotinamide adenine dinucleotide phosphate necessary for fatty acid synthesis.

Long-Chain Fatty Acids. Two QTL were identified on BTA14 for directly measured individual long-chain fatty acids (Table 3; Figures 3 and 4). One of these was at a Chr14:1.80 Mbp (rs385135066) locus that had a significant effect for directly measured C16:0 (*P*-value = 1.2e-12). This locus was in high LD ($R^2 = 0.74$) with missense mutations in the *SLC52A2* and *DGAT1* genes. The other QTL was for directly measured C18:1 *cis*-9 at a Chr14:1.76 Mbp (rs208417762) locus, that was also in high LD ($R^2 = 0.92$) with missense mutations in the *SLC52A2* and *DGAT1* genes. Closer examination of association effects for FT-MIR predicted C16:0 revealed evidence of a peak at this locus, but the peak was marginally below the significance threshold. Notably, in the present study, the identified protein coding mutation in the *SLC52A2* gene (rs134364612) was in perfect LD with the *DGAT1* K232A polymorphism (rs109234250), which has been attributed to changes in bovine milk fat composition (Grisart et al., 2002; Fink et al., 2020) and

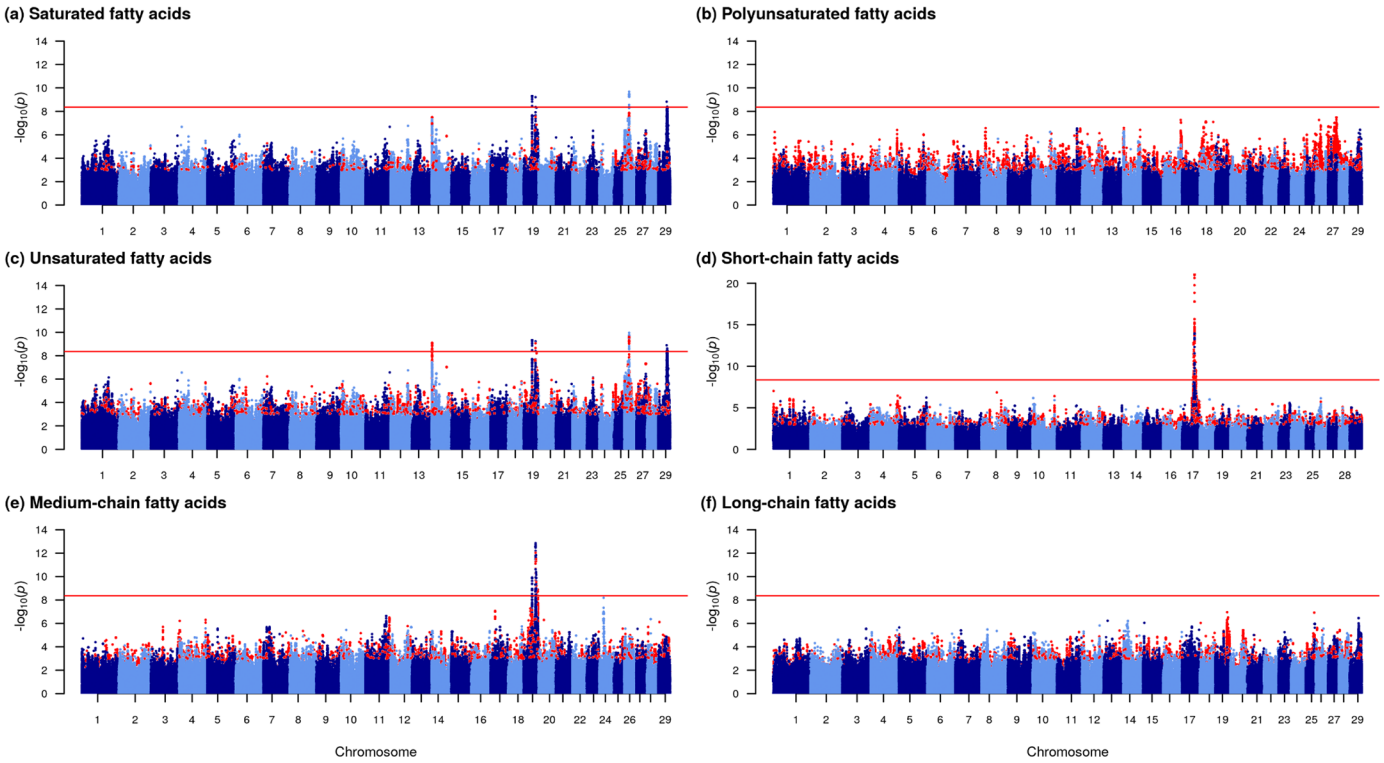

Figure 5. Manhattan plots showing association effects for directly measured (GC-based) and Fourier-transform mid-infrared (FT-MIR) predicted fatty acids classified based on the degree of saturation and the length of the carbon chain. Dark and light blue data points represent association signals for GC-based traits and red data points represent association signals for FT-MIR predicted traits. Chromosomes and genomic position based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals is represented as the $-\log_{10}$(*P*-value) on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}$(4.3e-09).

fatty acids (Bouwman et al., 2011; Buitenhuis et al., 2014; Li et al., 2014). The *DGAT1* gene encodes diacylglycerol O-acyltransferase 1, an enzyme that catalyzes the final step in triglyceride production, thus making this a compelling candidate for the observed effects.

Two further QTL were identified for FT-MIR predicted C16:0 and C18:3 *cis*-3 at Chr27:36.20 Mbp loci that were not in high LD with a splice region variant, or a moderate or high impact coding variant (Table 4; Figures 3 and 4). However, the locus for C18:3 *cis*-3 (rs110950972) was in perfect LD with a 5′ untranslated region (rs208675276) in *GPAT4*, and the locus for C16:0 was also in high LD ($R^2 = 0.997$) with that same 5′ untranslated region. Interestingly, we found no evidence of QTL for the corresponding directly measured traits (Figures 3 and 4). The Chr27:36.20 Mbp loci are situated in the *GPAT4* gene, which encodes the triglyceride synthesis enzyme glycerol-3-phosphate acyltransferase 4. As the milk fat percentage and other QTL at this locus have previously been shown to operate via a mechanism linked to gene expression (Littlejohn et al., 2014), it is not surprising that no significant coding mutations were identified in *GPAT4*.

Other Grouped Fatty Acid Effects. Further significant effects were observed for directly measured SFA and UFA at a Chr19:36.19 Mbp locus (rs110980742), that was in high LD ($R^2 > 0.81$) with 2 missense mutations in the *UTP18* gene (Table 3; Figure 5). This effect was not observed in any other individual or grouped fatty acid traits. The *UTP18* gene is involved in the nucleolar processing of pre-18S ribosomal RNA, and has not previously been reported in GWAS of bovine milk composition. The signal at Chr19:36.19 is close to the locus of the *KCNJ12* gene, which has a similar function to the *KCNJ2* gene that has previously been shown to affect milk phenotypes (Tiplady et al., 2021), although a mechanism by which this gene could affect fatty acids is unclear.

Individual Milk Proteins. Significant effects were observed on BTA6, BTA11, BTA14 and BTA22 for individual milk proteins (Tables 3 and 4; Figure 6). Four QTL were identified on BTA6, 2 of which were for directly measured and FT-MIR predicted α-CN, and the other 2 for directly measured and FT-MIR predicted κ-CN, respectively. The effects for α-CN were observed at a Chr6:87.13 Mbp locus (rs109500363) that

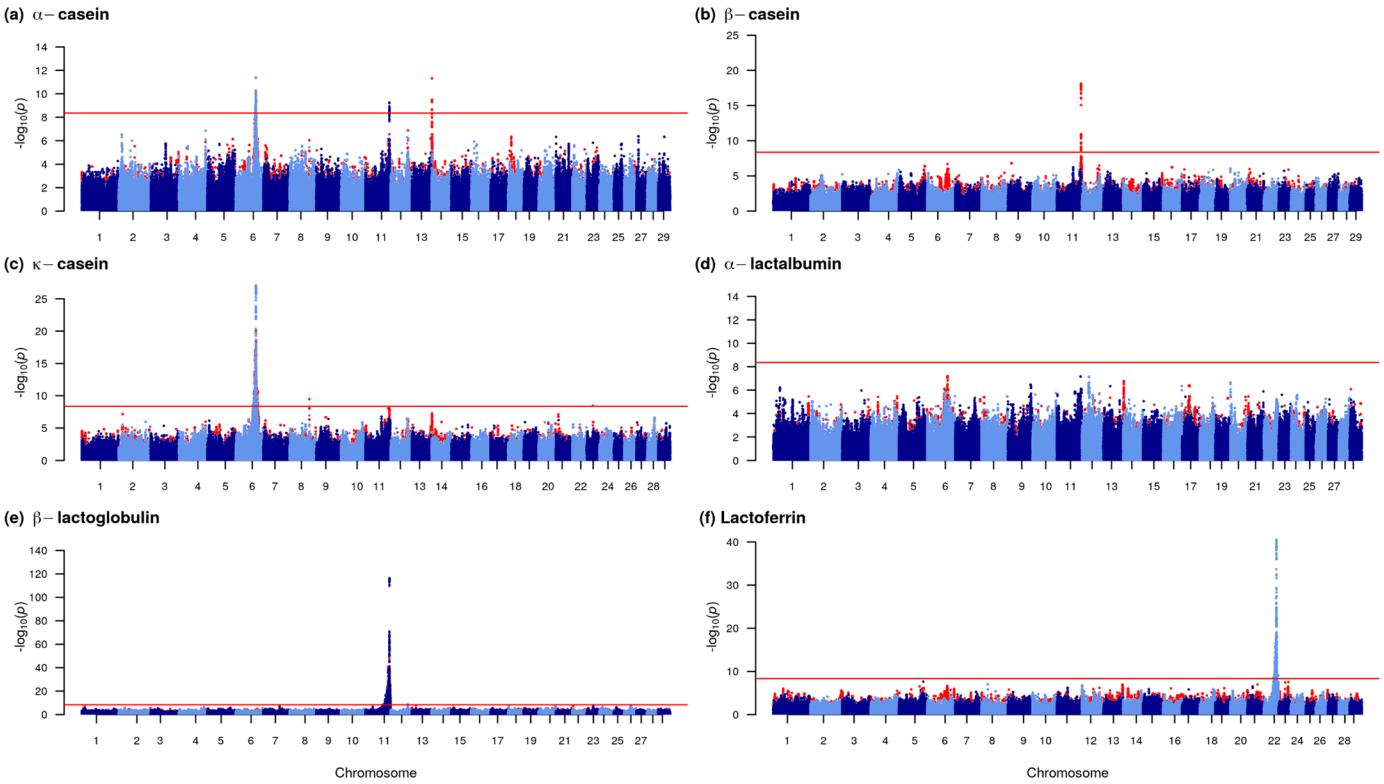

Figure 6. Manhattan plot showing association effects for directly measured (HPLC-based) and Fourier-transform mid-infrared (FT-MIR) predicted proteins. Dark and light blue data points represent association signals for HPLC-based traits and red data points represent association signals for FT-MIR predicted traits. Chromosomes and genomic position based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals is represented as the $-\log_{10}$(*P*-value) on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}$(4.3e-09).

was in high LD (R^2=0.92) with a missense mutation in the *CSN1S1* gene (rs43703010). As the *CSN1S1* gene codes for the α-CN protein (along with *CSN1S2*), it is not surprising that genetic signals affecting α-CN were enriched at this locus. Interestingly, FT-MIR predicted κ-CN also had a significant effect in the same region that was also in high LD (R^2 = 0.79) with rs43703010. The effect for directly measured κ-CN was observed at a Chr6:87.41 Mbp locus (rs110794953), which was in high LD (R^2 > 0.98) with 2 missense mutations in the *CSN3* gene (rs43703015 and rs43703016). The *CSN3* gene encodes κ-CN, an abundantly expressed milk protein. One of the missense mutations reported here (rs43703015) has previously been associated with milk composition traits and differential expression in mammary tissue (MacLeod et al., 2016). Significant effects have also been reported at this locus for a number of FT-MIR wavenumbers characterized by amide III and phosphate bands, C–H stretching vibrations of CH2 and –CH3, and N–H bending and C–N stretching in the amide II band (Tiplady et al., 2021).

Several QTL were identified for individual milk proteins on BTA11 that were in high LD (R^2 > 0.95) with missense mutations in the *PAEP* gene (rs110066229; rs109625649; Tables 3 and 4; Figure 6). Of these, the QTL with the most significant effects were observed for directly measured β-LG (*P*-value = 8.7e-117) and FT-MIR predicted β-LG (*P*-value = 5.4e-116). Smaller association effects were also observed for directly measured α-CN (*P*-value = 5.6e-10) and FT-MIR predicted β-CN (*P*-value = 8.3e-19). One of the implicated missense mutations in the *PAEP* gene was the V134A *PAEP* mutation (rs109625649) that distinguishes the 'A' and 'B' haplotypes of β-LG (previously described). This locus is likely driven by a regulatory effect, with a promoter variant reported to be in LD with the V134A mutation previously reported (Lum et al., 1997) to affect the binding of the Activator Protein-2 transcription factor. An expression QTL (**eQTL**) for *PAEP* was also reported in lactating bovine mammary tissue (Tiplady et al., 2021; Davis et al., 2022).

One further QTL of interest was for directly measured Lf at a Chr22:53.54 Mbp locus (rs43765460; Table 3; Figure 6). The association effect at this locus had a *P*-value of 1.8e-41, but we found no relevant splice region variant, or moderate or high impact coding variant

Table 3. Peak variants for directly measured fatty acid and protein traits with significant association effects[1]

Trait[2]	Chr	Position	Tag variant ID	*P*-value	Protein coding variant ID	LD	Gene	Class	Description
Individual fatty acid (g/100 g of total fat)									
C18:1 *cis*-9	14	1756075	rs208417762	1.3e-10	rs134364612	0.915	*SLC52A2*	Missense	c.724A > G
C18:1 *cis*-9	14	1756075	rs208417762	1.3e-10	rs109234250	0.915	*DGAT1*	Missense	c.694G > A
C16:0	14	1799066	rs385135066	1.2e-12	rs134364612	0.737	*SLC52A2*	Missense	c.724A > G
C16:0	14	1799066	rs385135066	1.2e-12	rs109234250	0.737	*DGAT1*	Missense	c.694G > A
C6:0	17	52971731	rs207997694	9.6e-10	—	—	—	—	—
C4:0	17	53034516	rs461037541	7.2e-18	—	—	—	—	—
C10:0	19	51319673	rs137270097	1.2e-10	rs41921160	0.974	*CCDC57*	Missense	c.1907T > C
C12:0	19	51319673	rs137270097	8.3e-13	rs41921160	0.974	*CCDC57*	Missense	c.1907T > C
C14:0	19	51326050	rs136424304	1.4e-11	rs41921160	0.996	*CCDC57*	Missense	c.1907T > C
C10:0	26	21141279	rs41255696	2.2e-10	rs41255693	0.799	*SCD*	Splice region	c.569C > T
C14:0	26	21141279	rs41255696	2.7e-10	rs41255693	0.799	*SCD*	Splice region	c.569C > T
C10:1	26	21148111	rs41255688	1.8e-48	rs41255693	0.915	*SCD*	Splice region	c.569C > T
C14:1	26	21149680	rs385285356	6.1e-61	rs41255693	0.915	*SCD*	Splice region	c.569C > T
C10:1	26	26458006	rs445758306	2.6e-10	rs379463458	0.761	*ITPRIP*	Missense	c.1301G > A
C12:1	26	26458006	rs445758306	2.4e-10	rs379463458	0.761	*ITPRIP*	Missense	c.1301G > A
Grouped fatty acid (g/100 g of total fat)									
SCFA	17	53034516	rs461037541	1.2e-14	—	—	—	—	—
SFA	19	36187954	rs110980742	5.0e-10	rs210064667	0.816	*UTP18*	Missense	c.85G > A
SFA	19	36187954	rs110980742	5.0e-10	rs382000222	0.848	*UTP18*	Missense	c.79T > A
UFA	19	36187954	rs110980742	4.8e-10	rs210064667	0.816	*UTP18*	Missense	c.85G > A
UFA	19	36187954	rs110980742	4.8e-10	rs382000222	0.848	*UTP18*	Missense	c.79T > A
MCFA	19	51319673	rs137270097	1.4e-13	rs41921160	0.974	*CCDC57*	Missense	c.1907T > C
SFA	26	21149680	rs385285356	2.1e-10	rs41255693	0.915	*SCD*	Splice region	c.569C > T
UFA	26	21149680	rs385285356	1.1e-10	rs41255693	0.915	*SCD*	Splice region	c.569C > T
Individual milk protein (g/L of total volume)									
α-CN	6	87133508	rs109500363	4.3e-12	rs382793163	0.856	*ENSBTAG00000039991*	Missense	c.1406G > A
α-CN	6	87133508	rs109500363	4.3e-12	rs385603965	0.839	*ENSBTAG00000003523*	Missense	c.1378C > T
α-CN	6	87133508	rs109500363	4.3e-12	rs43703010	0.923	*CSN1S1*	Missense	c.620A > G
κ-CN	6	87405588	rs110794953	6.4e-28	rs109739692	0.805	*ODAM*	Missense	c.520G > A
κ-CN	6	87405588	rs110794953	6.4e-28	rs43703015	0.988	*CSN3*	Missense	c.470T > C
κ-CN	6	87405588	rs110794953	6.4e-28	rs43703016	0.985	*CSN3*	Missense	c.506C > A
β-LG	11	103291134	rs110270048	8.7e-117	rs110066229	1	*PAEP*	Missense	c.239G > A
β-LG	11	103291134	rs110270048	8.7e-117	rs109990218	0.997	*PAEP*	Splice region	c.305–5A > T
β-LG	11	103291134	rs110270048	8.7e-117	rs109625649	0.985	*PAEP*	Missense	c.401T > C
α-CN	11	103292575	rs381050299	5.6e-10	rs110066229	0.965	*PAEP*	Missense	c.239G > A
α-CN	11	103292575	rs381050299	5.6e-10	rs109990218	0.962	*PAEP*	Splice region	c.305–5A > T
α-CN	11	103292575	rs381050299	5.6e-10	rs109625649	0.950	*PAEP*	Missense	c.401T > C
Lf[3]	22	53538882	rs43765460	1.8e-41	—	—	—	—	—

[1]Peak variants for directly measured fatty acid traits with significant association effects; Bonferroni threshold: $-\log_{10}(4.3e\text{-}09)$.

[2]Trait definitions and units as described in Table 1.

[3]Cube-root transformation of lactoferrin (Lf).

ascribed to this effect. However, the rs43765460 locus is a synonymous variant in the *LTF* gene. Using our previously published mammary RNA sequence dataset and eQTL mapping methodology (Lopdell et al., 2017; Tiplady et al., 2021), we identified the presence of a co-localized expression-based effect for *LTF* in this region. The rs43765460 locus we identified was in high LD with the top associated eQTL variant for Lf (R^2 = 0.88), and the Pearson correlation between the $-\log_{10}$(*P*-values) of the directly measured Lf QTL, and the $-\log_{10}$(*P*-values) of the Lf eQTL within a 1 Mbp region flanking the rs43765460 variant was 0.92. The *LTF* gene is a major iron-binding protein in milk that is linked to iron homeostasis and plays a key role in immune system response and cell growth. Previous studies have shown that the *LTF* gene is linked to changes in Lf concentrations in bovine milk (Bahar et al., 2011; Pawlik et al., 2014). Notably, there was no evidence of an association effect at or near this locus for FT-MIR predicted Lf (Table 4). Further, in a recent GWAS of individual FT-MIR wavenumbers, there was also no evidence of an association effect linked to the *LTF* gene (Tiplady et al., 2021), indicating that changes in milk composition due to this gene may not be easily detectable using FT-MIR spectral data. However, it is also important to note that prediction accuracies for Lf in the present study were relatively poor (R^2_{cv} = 0.36; RPE_{cv} = 0.19; Table 1), and the heritability estimate for FT-MIR predicted Lf was only 0.30, compared to the heritability estimate for directly measured Lf, which was 0.59 (Table 2). This pattern is similar to that which we observed for C14:1 and the *SCD* gene. That is, the component was in relatively low concentrations in the milk sample, model prediction accuracy

Table 4. Peak variants for Fourier-transform mid-infrared (FT-MIR) predicted fatty acid and protein traits with significant association effects[1]

Trait[2]	Chr	Position	Tag variant ID	*P*-value	Protein coding variant ID	LD	Gene	Class	Description
Individual fatty acid (g/100 g of total fat)									
C12:1	11	103301736	rs41255687	6.3e-11	rs110066229	0.988	*PAEP*	Missense	c.239G > A
C12:1	11	103301736	rs41255687	6.3e-11	rs109625649	0.991	*PAEP*	Missense	c.401T > C
C18:3 *cis*-3	14	2502770	rs137422574	1.0e-12	rs109403601	0.988	*ENSBTAG00000003606*	Missense	c.154C > G
C18:1 *cis*-9	14	2528807	rs110275497	1.3e-10	rs109403601	1	*ENSBTAG00000003606*	Missense	c.154C > G
C6:0	17	52971731	rs207997694	9.9e-16	—	—	—	—	—
C4:0	17	53034516	rs461037541	1.5e-17	—	—	—	—	—
C10:0	19	51314476	rs41922143	7.0e-13	rs41921160	0.989	*CCDC57*	Missense	c.1907T > C
C12:0	19	51314476	rs41922143	3.8e-12	rs41921160	0.989	*CCDC57*	Missense	c.1907T > C
C14:0	19	51314476	rs41922143	7.0e-12	rs41921160	0.989	*CCDC57*	Missense	c.1907T > C
C8:0	19	51326050	rs136424304	8.9e-10	rs41921160	0.996	*CCDC57*	Missense	c.1907T > C
C14:1	26	21174891	rs209445650	1.9e-09	—	—	—	—	—
C10:1	26	25584818	rs210921941	5.8e-10	—	—	—	—	—
C18:3 *cis*-3	27	36200888	rs110950972	9.9e-15	—	—	—	—	—
C16:0	27	36204679	—	1.6e-09	—	—	—	—	—
Grouped fatty acid (g/100 g of total fat)									
UFA	14	2319003	rs110182536	8.1e-10	rs109403601	0.947	*ENSBTAG00000003606*	Missense	c.154C > G
SCFA	17	53034516	rs461037541	7.1e-22	—	—	—	—	—
UFA	19	50919823	rs380534925	8.8e-10	—	—	—	—	—
MCFA	19	51314476	rs41922143	9.2e-13	rs41921160	0.989	*CCDC57*	Missense	c.1907T > C
UFA	26	21138011	rs381655271	2.6e-10	rs41255693	0.914	*SCD*	Splice region	c.569C > T
Individual milk protein (g/L of total volume)									
κ-CN	6	87085918	—	8.2e-21	rs209798512	0.761	*ENSBTAG00000038520*	Missense	c.1623G > C
κ-CN	6	87085918	—	8.2e-21	rs211555767	0.761	*ENSBTAG00000038520*	Missense	c.1301C > T
κ-CN	6	87085918	—	8.2e-21	rs382793163	0.725	*ENSBTAG00000039991*	Missense	c.1406G > A
κ-CN	6	87085918	—	8.2e-21	rs385603965	0.711	*ENSBTAG00000003523*	Missense	c.1378C > T
κ-CN	6	87085918	—	8.2e-21	rs43703010	0.787	*CSN1S1*	Missense	c.620A > G
α-CN	6	87133508	rs109500363	7.0e-11	rs382793163	0.856	*ENSBTAG00000039991*	Missense	c.1406G > A
α-CN	6	87133508	rs109500363	7.0e-11	rs385603965	0.839	*ENSBTAG00000003523*	Missense	c.1378C > T
α-CN	6	87133508	rs109500363	7.0e-11	rs43703010	0.923	*CSN1S1*	Missense	c.620A > G
β-CN	11	103299272	rs110563549	8.3e-19	rs110066229	0.997	*PAEP*	Missense	c.239G > A
β-CN	11	103299272	rs110563549	8.3e-19	rs109625649	0.988	*PAEP*	Missense	c.401T > C
β-LG	11	103299272	rs110563549	5.4e-116	rs110066229	0.997	*PAEP*	Missense	c.239G > A
β-LG	11	103299272	rs110563549	5.4e-116	rs109625649	0.988	*PAEP*	Missense	c.401T > C
α-CN	14	1799066	rs385135066	4.8e-12	rs134364612	0.737	*SLC52A2*	Missense	c.724A > G
α-CN	14	1799066	rs385135066	4.8e-12	rs109234250	0.737	*DGAT1*	Missense	c.694G > A

[1]Peak variants for FT-MIR predicted fatty acid traits with significant association effects; Bonferroni threshold: $-\log_{10}(4.3e\text{-}09)$.
[2]Trait definitions and units as described in Table 1.

was relatively poor, the heritability for the measured trait was substantially higher than for the FT-MIR predicted trait, and a compelling peak was observed for the directly measured trait; however, no corresponding peak was observed for the FT-MIR predicted trait.

Perspectives on the Use of FT-MIR Trait Predictions in Dairy Cattle Selection

Utilizing FT-MIR predictions for fatty acids and proteins in milk can provide indicator traits across large numbers of animals at little or no marginal cost, because FT-MIR spectral data are already generated as part of routine milk testing to predict total fat and protein concentrations. The alternative to using FT-MIR trait predictions is to directly measure traits, which may be infeasible across even relatively small numbers of animals. Phenotypic correlations between directly measured and FT-MIR predicted traits provide a useful indication of the utility of FT-MIR trait predictions, particularly for herd management and milk processability traits. However, for breeding programs, we are also interested in the heritability of the FT-MIR predicted trait and the genetic correlation between the directly measured and FT-MIR predicted trait. This is because the heritability of the FT-MIR predicted trait defines the level of genetic variation present, whereas the genetic correlation between the directly measured and FT-MIR predicted trait defines the breeding progress we could expect in the directly measured trait if we were to select animals based on the FT-MIR predicted trait. Specifically, within the context of dairy cattle progeny test schemes, the genetic correlation will limit the relative amount of selection response that will result from using FT-MIR predictions instead of directly measured traits (Rutten et al., 2010). Based on this assumption, the genetic gain from selection using FT-MIR predictions for all traits we have studied would be greater than 70% of the gains achievable by direct selection on these traits; additionally, for 21 of the 29 traits, the genetic gains achievable would be greater than 85% of the gains achievable by direct selection.

However, it is important to note that this assumes that there is no true difference in heritability between the directly measured and FT-MIR predicted trait. For traits such as Lf and C14:1 where the estimated heritability of the direct measurement was lower than the heritability of the FT-MIR prediction, the genetic gain achievable would also be lower.

Although we observed high genetic correlations between directly measured and FT-MIR predicted traits in this study, the QTL underlying each trait were not always the same. An example of this includes where we observed a large association effect within the *GPAT4* gene on BTA27 for FT-MIR predicted C18:3 *cis*-3, but no corresponding association effect was observed for directly measured C18:3 *cis*-3 (Figure 4). Similarly, a large association effect was observed for FT-MIR predicted β-CN within the *PAEP* gene on BTA11, but no corresponding association effect was observed in directly measured β-CN (Figure 6). The presence of QTL with significant effects in an FT-MIR predicted trait only are not entirely surprising, given that FT-MIR predicted traits are a weighted linear function of absorbance values for individual wavenumbers, each of which may be underpinned by multiple genetic signals and QTL (Wang and Bovenhuis, 2018; Benedet et al., 2019; Zaalberg et al., 2020; Tiplady et al., 2021). This means that when a spectral wavenumber is included in a trait prediction equation, multiple genetic signals will also be present, some of which are specifically related to the trait of interest and some that are not. It is important that when FT-MIR predicted traits are used as proxies for other traits, we are mindful of this, particularly when using SNP-based approaches in our estimation of breeding values, whereby the impact will be determined by the relative proportion of genetic variation captured by each SNP and the interaction of additive effects between SNPs.

Instances also arose where a QTL was observed for a directly measured trait, but we found no corresponding QTL observed in the FT-MIR predicted trait. Examples of this include large association effects within the *SCD* gene for directly measured C10:1 and C14:1, but no corresponding association effects for individual FT-MIR predicted fatty acids (Figure 2). Similarly, a large association effect was observed within the *LTF* gene for directly measured Lf, but a corresponding association effect for FT-MIR predicted Lf was absent (Figure 6). In these examples, there was a consistent pattern where we have a component in relatively low concentrations in the milk sample, with relatively poor model prediction accuracies and lower heritability estimates for the FT-MIR predicted trait, compared with the directly measured trait (Tables 1 and 2). Although it might be argued that the failure to detect QTL in the *SCD* and *LTF* genes was because the calibration equations were inadequate for the task of quantifying the milk component targets (C10:1, C14:1, and Lf), it is also notable that in a previous GWAS we conducted on individual FT-MIR wavenumbers (Tiplady et al., 2021), no significant associations were identified between FT-MIR wavenumbers and variants within the *SCD* and *LTF* genes. Potentially, this means that changes in milk composition attributable to these 2 genes may be difficult to quantify directly using FT-MIR wavenumber spectra. For Lf to be detected using FT-MIR spectral data, it needs to provide a unique signal that distinguishes it from other whey proteins in solution that are at much higher concentrations. However, when the mean concentration of Lf is around 0.1 g/L and the major whey protein β-LG is at a 20- to 40-fold higher concentration, it is not surprising that a QTL is seen within the *PAEP* gene and not within the *LTF* gene.

With the growing interest in using FT-MIR spectral data to predict molecules at low concentrations in milk, it is important to understand that the predictive performance of these models may be limited, compared with models for predicting major milk components such as total fat and protein concentrations (Grelet et al., 2021). In the context of the present study, we have shown that for many fatty acids and protein traits, model prediction accuracies are moderate, but that genetic correlations between directly measured and FT-MIR predicted fatty acid and protein fractions are typically high. However, it is also clear that phenotypic variation between directly measured and FT-MIR predicted traits may be underpinned by differing genetic architecture. This may be due to several related factors including the trait of interest being at low concentrations in the milk sample, low prediction model accuracy, or that the trait is not easily detectable using FT-MIR spectroscopy. Improving calibration equations is central to optimizing our use of FT-MIR spectra to generate proxies for traits of interest to the industry such as fatty acids and protein fractions. Collaboration between research groups to generate data sets that include data from a range of herds that capture differences in climate, management systems, diet, and breed composition might improve calibration equations (Grelet et al., 2021). However, a key barrier to consolidating FT-MIR spectral data sets from different research groups is variation in spectral measurements between instruments and within instruments across time. Standardization of individual FT-MIR spectra wavenumbers using reference samples can effectively address these sources of variation (Grelet et al., 2015, 2017; Tiplady et al., 2019); however, outside the European OptiMIR network, reference sample sharing and standardization is not common practice. Other approaches, such as those offered by Foss or Bentley,

are appealing in that they are not reliant on perishable milk samples. However, as far as we are aware, the effectiveness of these procedures has not been independently evaluated. Validation of these within-instrument standardization procedures is important, because if the procedures work well, they could facilitate the consolidation of spectral data from different networks/countries, and assist with the development of better prediction equations and improve trait prediction accuracies.

Study Limitations

In this study, we developed PLS prediction equations and compared the genetic characteristics of directly measured fatty acids and protein fractions to the same traits predicted from FT-MIR spectra. There are several areas of refinement that might improve prediction equations and the identification of QTL. First, before the development of prediction equations, we assessed several mathematical treatments of spectra, but we only assessed the prediction accuracies of those treatments using PLS models. Although PLS is a widely used method for developing calibration models from FT-MIR spectra, it may be possible to develop better prediction models for some traits by employing Bayesian or other machine learning approaches, as demonstrated in other studies of milk composition (Bonfatti et al., 2017a; El Jabri et al., 2019; Frizzarin et al., 2021), or animal health and feed intake traits (Dórea et al., 2018; Brand et al., 2021; Contla Hernández et al., 2021). Second, it is expected that increasing the number of samples in the study and including data from different herds would also improve trait prediction accuracies, particularly for fatty acids and protein fractions at low concentrations in milk samples. Extending the study to include data from different herds would also facilitate a more robust validation strategy. Although the cow-independent validation approach we have used is commonly practiced in studies of FT-MIR spectra trait prediction, it has been shown that record- or cow-independent validation can overinflate prediction accuracies, compared with herd-independent validation (Dórea et al., 2018; Lahart et al., 2019; Luke et al., 2019; Wang and Bovenhuis, 2019). Improving and validating the prediction equations we have developed in this study are important steps for future research to confirm their utility for prediction and use in future breeding programs.

Other potential refinements for the present study relate to genomic information and the strategy for identifying QTL. Specifically, we have used data sets mapped to the UMD3.1 genome; however, it is expected that the improved sequence continuity and per-base accuracy of the ARS-UCD1.2 reference genome (Rosen et al., 2020) may yield a few additional QTL and reveal additional candidate mutations given improvements in accompanying transcript annotations. Also, the approach we used to identify QTL could be extended to account for nonadditive QTL in a similar manner to that outlined in Reynolds et al. (2021). Finally, the approach we used to identify causative genes and variants only considered protein-altering variants as candidates, which we acknowledge is relatively simple and crude, and that many of the identified signals could be underpinned by regulatory effects (e.g., gene expression-based mechanisms). It is expected that integration of other functional data sets such as mammary eQTL and ChIP-seq data sets could map additional molecular QTL and enhance fine mapping and candidate variant identification (Tiplady et al., 2020).

CONCLUSIONS

We developed PLS calibration equations to predict bovine fatty acids and protein fractions in milk samples, and compared the genetic architecture underlying directly measured traits to that of corresponding FT-MIR predicted traits. Low to moderate prediction accuracies were observed, indicating that the potential application of using FT-MIR prediction equations for some traits may be limited. However, for most traits, heritability estimates were moderate to high, indicating that genetic variation exists that could potentially be exploited for the purposes of animal selection. Moreover, high genetic correlations between directly measured and FT-MIR predicted fatty acids and individual milk proteins indicated that selection based on FT-MIR predicted traits could provide high rates of genetic gain in the corresponding trait of interest. Trait QTL for fatty acids were identified with likely candidates in the *DGAT1*, *CCDC57*, *SCD*, and *GPAT4* genes, but QTL underpinned by *SCD* were largely absent in FT-MIR predicted fatty acids. Similarly, likely candidates were identified for directly measured proteins in the *CSN1S1*, *CSN3*, *PAEP*, and *LTF* genes, but the QTL for *CSN3* and *LTF* were absent in corresponding FT-MIR predicted traits. This highlighted that, in some instances, phenotypic variation for directly measured and FT-MIR predicted traits were underpinned by differing genetic architecture and segregation of alleles at QTL.

ACKNOWLEDGMENTS

The authors thank Livestock Improvement Corporation (LIC; Hamilton, New Zealand) farm and technical staff for collecting milk samples, and herd-testing staff

for the processing and analysis of milk samples, as well as the staff at Fonterra Research and Development Centre (Palmerston North, New Zealand) for milk analyses. Kathryn also thanks the wider LIC R&D team and fellow students for underlying technical support and thoughtful discussion, and Tracey Monehan (R&D Programme Manager, LIC) for overseeing the funding for this work. We also gratefully acknowledge the use of New Zealand eScience Infrastructure (NeSI) high-performance computing for this research. This research was funded through BoviQuest, a joint venture between LIC and ViaLactia Biosciences Ltd., a subsidiary (now closed) of Fonterra Cooperative Ltd. (Auckland, New Zealand), LIC (Hamilton, New Zealand), and the New Zealand Ministry for Primary Industries, within the Resilient Dairy Programme through Sustainable Food & Fibre Futures (funding no: PGP06-17006). The authors have not stated any conflicts of interest.

REFERENCES

Arnould, V., N. Gengler, and H. Soyeurt. 2009a. Genetic variability of test-day stearoyl coenzyme-A desaturase 9 activity. J. Dairy Sci. 92:353–354.

Arnould, V. M.-R., H. Soyeurt, N. Gengler, F. G. Colinet, M. V. Georges, C. Bertozzi, D. Portetelle, and R. Renaville. 2009b. Genetic analysis of lactoferrin content in bovine milk. J. Dairy Sci. 92:2151–2158. https://doi.org/10.3168/jds.2008-1255.

Bahar, B., F. O'Halloran, M. J. Callanan, S. McParland, L. Giblin, and T. Sweeney. 2011. Bovine lactoferrin (LTF) gene promoter haplotypes have different basal transcriptional activities. Anim. Genet. 42:270–279. https://doi.org/10.1111/j.1365-2052.2010.02151.x.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67:1–48. https://doi.org/10.18637/jss.v067.i01.

Benedet, A., P. N. Ho, R. Xiang, S. Bolormaa, M. De Marchi, M. E. Goddard, and J. E. Pryce. 2019. The use of mid-infrared spectra to map genes affecting milk composition. J. Dairy Sci. 102:7189–7203. https://doi.org/10.3168/jds.2018-15890.

Bergman, E. N. 1971. Hyperketonemia-ketogenesis and ketone body metabolism. J. Dairy Sci. 54:936–948. https://doi.org/10.3168/jds.S0022-0302(71)85950-7.

Bernard, L., C. Leroux, and Y. Chilliard. 2006. Characterisation and nutritional regulation of the main lipogenic genes in the ruminant lactating mammary gland. Rumin. Physiol. Dig. Metab. Impact Nutr. Gene Expr. Immunol. Stress 295–326.

Berry, S. D., N. Lopez-Villalobos, E. M. Beattie, S. R. Davis, L. F. Adams, N. L. Thomas, A. E. Ankersmit-Udy, A. M. Stanfield, K. Lehnert, H. E. Ward, J. A. Arias, R. J. Spelman, and R. G. Snell. 2010. Mapping a quantitative trait locus for the concentration of β-lactoglobulin in milk, and the effect of β-lactoglobulin genetic variants on the composition of milk from Holstein-Friesian x Jersey crossbred cows. N. Z. Vet. J. 58:1–5. https://doi.org/10.1080/00480169.2010.65053.

Bonfatti, V., L. Degano, A. Menegoz, and P. Carnier. 2016. *Short communication*: Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. J. Dairy Sci. 99:8216–8221. https://doi.org/10.3168/jds.2016-10953.

Bonfatti, V., G. Di Martino, and P. Carnier. 2011. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. J. Dairy Sci. 94:5776–5785. https://doi.org/10.3168/jds.2011-4401.

Bonfatti, V., F. Tiezzi, F. Miglior, and P. Carnier. 2017a. Comparison of Bayesian regression models and partial least squares regression for the development of infrared prediction equations. J. Dairy Sci. 100:7306–7319. https://doi.org/10.3168/jds.2016-12203.

Bonfatti, V., D. Vicario, A. Lugo, and P. Carnier. 2017b. Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle. J. Dairy Sci. 100:5526–5540. https://doi.org/10.3168/jds.2016-11667.

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. https://doi.org/10.1186/1471-2156-12-43.

Bouwman, A. C., M. H. P. W. Visker, J. A. M. van Arendonk, and H. Bovenhuis. 2014. Fine mapping of a quantitative trait locus for bovine milk fat composition on *Bos taurus* autosome 19. J. Dairy Sci. 97:1139–1149. https://doi.org/10.3168/jds.2013-7197.

Brand, W., A. T. Wells, S. L. Smith, S. J. Denholm, E. Wall, and M. P. Coffey. 2021. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. J. Dairy Sci. 104:4980–4990. https://doi.org/10.3168/jds.2020-18367.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097. https://doi.org/10.1086/521987.

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. https://doi.org/10.1186/1471-2164-15-1112.

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. https://doi.org/10.1186/s12863-016-0421-2.

Butler, D. G., B. R. Cullis, A. R. Gilmour, and B. J. Gogel. 2009. ASReml-R Reference Manual: Analysis of Mixed Models for S Language Environments. Queensland Government.

Caroli, A. M., S. Chessa, and G. J. Erhardt. 2009. *Invited review*: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. J. Dairy Sci. 92:5335–5352. https://doi.org/10.3168/jds.2009-2461.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80–92. https://doi.org/10.4161/fly.19695.

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. https://doi.org/10.3168/jds.2009-2581.

Contla Hernández, B., N. Lopez-Villalobos, and M. Vignes. 2021. Identifying health status in grazing dairy cows from milk mid-infrared spectroscopy by using machine learning methods. Animals (Basel) 11:2154. https://doi.org/10.3390/ani11082154.

Cruz, V. A. R., H. R. Oliveira, L. F. Brito, A. Fleming, S. Larmer, F. Miglior, and F. S. Schenkel. 2019. Genome-wide association study for milk fatty acids in Holstein cattle accounting for the *DGAT1* gene effect. Animals (Basel) 9:997. https://doi.org/10.3390/ani9110997.

Davis, S. R., H. E. Ward, V. Kelly, D. Palmer, A. E. Ankersmit-Udy, T. J. Lopdell, S. D. Berry, M. D. Littlejohn, K. Tiplady, L. F. Adams, K. Carnie, A. Burrett, N. Thomas, R. G. Snell, R. J. Spelman, and K. Lehnert. 2022. Screening for phenotypic outliers identifies an unusually low concentration of a β-lactoglobulin B protein isoform in bovine milk caused by a synonymous SNP. Genet. Sel. Evol. 54:22. https://doi.org/10.1186/s12711-022-00711-z.

De Marchi, M., V. Bonfatti, A. Cecchinato, G. Di Martino, and P. Carnier. 2009. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. Ital. J. Anim. Sci. 8(Suppl. 2):399–401. https://doi.org/10.4081/ijas.2009.s2.399.

Dórea, J. R. R., G. J. M. Rosa, K. A. Weld, and L. E. Armentano. 2018. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. J. Dairy Sci. 101:5878–5889. https://doi.org/10.3168/jds.2017-13997.

Duggal, P., E. M. Gillanders, T. N. Holmes, and J. E. Bailey-Wilson. 2008. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. BMC Genomics 9:516. https://doi.org/10.1186/1471-2164-9-516.

El Jabri, M., M.-P. Sanchez, P. Trossat, C. Laithier, V. Wolf, P. Grosperrin, E. Beuvier, O. Rolet-Répécaud, S. Gavoye, Y. Gaüzère, O. Belysheva, E. Notz, D. Boichard, and A. Delacroix-Buchet. 2019. Comparison of Bayesian and partial least squares regression methods for mid-infrared prediction of cheese-making properties in Montbéliarde cows. J. Dairy Sci. 102:6943–6958. https://doi.org/10.3168/jds.2019-16320.

Fink, T., T. J. Lopdell, K. Tiplady, R. Handley, T. J. J. Johnson, R. J. Spelman, S. R. Davis, R. G. Snell, and M. D. Littlejohn. 2020. A new mechanism for a familiar mutation—Bovine *DGAT1* K232A modulates gene expression through multi-junction exon splice enhancement. BMC Genomics 21:591. https://doi.org/10.1186/s12864-020-07004-z.

Fleming, A., F. S. Schenkel, F. Malchiodi, R. A. Ali, B. Mallard, M. Sargolzaei, J. Jamrozik, J. Johnston, and F. Miglior. 2018. Genetic correlations of mid-infrared-predicted milk fatty acid groups with milk production traits. J. Dairy Sci. 101:4295–4306. https://doi.org/10.3168/jds.2017-14089.

Freitas, P. H. F., H. R. Oliveira, F. F. Silva, A. Fleming, F. Miglior, F. S. Schenkel, and L. F. Brito. 2020. Genomic analyses for predicted milk fatty acid composition throughout lactation in North American Holstein cattle. J. Dairy Sci. 103:6318–6331. https://doi.org/10.3168/jds.2019-17628.

Frizzarin, M., I. C. Gormley, D. P. Berry, T. B. Murphy, A. Casa, A. Lynch, and S. McParland. 2021. Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. J. Dairy Sci. 104:7438–7447. https://doi.org/10.3168/jds.2020-19576.

Fuentes-Pila, J., M. A. DeLorenzo, D. K. Beede, C. R. Staples, and J. B. Holter. 1996. Evaluation of equations based on animal factors to predict intake of lactating Holstein cows. J. Dairy Sci. 79:1562–1571. https://doi.org/10.3168/jds.S0022-0302(96)76518-9.

Grelet, C., P. Dardenne, H. Soyeurt, J. A. Fernandez, A. Vanlierde, F. Stevens, N. Gengler, and F. Dehareng. 2021. Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions. Methods 186:97–111. https://doi.org/10.1016/j.ymeth.2020.07.012.

Grelet, C., J. A. Fernández 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. https://doi.org/10.3168/jds.2014-8764.

Grelet, C., J. A. F. Pierna, P. Dardenne, H. Soyeurt, A. Vanlierde, F. Colinet, C. Bastin, N. Gengler, V. Baeten, and F. Dehareng. 2017. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. J. Dairy Sci. 100:7910–7921. https://doi.org/10.3168/jds.2017-12720.

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. https://doi.org/10.1101/gr.224202.

Huang, G., D. Buckler-Pena, T. Nauta, M. Singh, A. Asmar, J. Shi, J. Y. Kim, and K. V. Kandror. 2013. Insulin responsiveness of glucose transporter 4 in 3T3–L1 cells depends on the presence of sortilin. Mol. Biol. Cell 24:3115–3122. https://doi.org/10.1091/mbc.e12-10-0765.

Iung, L. H. S., J. Petrini, J. Ramírez-Díaz, M. Salvian, G. A. Rovadoscki, F. Pilonetto, B. D. Dauria, P. F. Machado, L. L. Coutinho, G. R. Wiggans, and G. B. Mourão. 2019. Genome-wide association study for milk production traits in a Brazilian Holstein population. J. Dairy Sci. 102:5305–5314. https://doi.org/10.3168/jds.2018-14811.

Jivanji, S., G. Worth, T. J. Lopdell, A. Yeates, C. Couldrey, E. Reynolds, K. Tiplady, L. McNaughton, T. J. J. Johnson, S. R. Davis, B. Harris, R. Spelman, R. G. Snell, D. Garrick, and M. D. Littlejohn. 2019. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. Genet. Sel. Evol. 51:62. https://doi.org/10.1186/s12711-019-0506-2.

Kgwatalala, P. M., E. M. Ibeagha-Awemu, J. F. Hayes, and X. Zhao. 2009. *Stearoyl-CoA desaturase 1* 3′UTR SNPs and their influence on milk fatty acid composition of Canadian Holstein cows. J. Anim. Breed. Genet. 126:394–403. https://doi.org/10.1111/j.1439-0388.2008.00796.x.

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. https://doi.org/10.1038/s41598-018-20476-0.

Kucheryavskiy, S. 2020. mdatools—R package for chemometrics. Chemom. Intell. Lab. Syst. 198:103937.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, and T. Hunt. 2022. Caret: Classification and Regression Training.

Lahart, B., S. McParland, E. Kennedy, T. M. Boland, T. Condon, M. Williams, N. Galvin, B. McCarthy, and F. Buckley. 2019. Predicting the dry matter intake of grazing dairy cows using infrared reflectance spectroscopy analysis. J. Dairy Sci. 102:8907–8918. https://doi.org/10.3168/jds.2019-16363.

Li, C., D. Sun, S. Zhang, S. Wang, X. Wu, Q. Zhang, L. Liu, Y. Li, and L. Qiao. 2014. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. PLoS One 9:e96186. https://doi.org/10.1371/journal.pone.0096186.

Littlejohn, M. D., K. Tiplady, T. Lopdell, T. A. Law, A. Scott, C. Harland, R. Sherlock, K. Henty, V. Obolonkin, K. Lehnert, A. MacGibbon, R. J. Spelman, S. R. Davis, and R. G. Snell. 2014. Expression variants of the lipogenic *AGPAT6* gene affect diverse milk composition phenotypes in *Bos taurus*. PLoS One 9:e85757. https://doi.org/10.1371/journal.pone.0085757.

Loh, P.-R., G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, and A. L. Price. 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. 47:284–290. https://doi.org/10.1038/ng.3190.

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. DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. BMC Genomics 18:968. https://doi.org/10.1186/s12864-017-4320-3.

Lopez-Villalobos, N., S. R. Davis, E. M. Beattie, J. Melis, S. Berry, S. E. Holroyd, R. J. Spelman, and R. G. Snell. 2009. Breed effects for lactoferrin concentration determined by Fourier transform infrared spectroscopy. Proc. N.Z. Soc. Anim. Prod. 69:60–64.

Lopez-Villalobos, N., R. J. Spelman, J. Melis, S. R. Davis, S. D. Berry, K. Lehnert, S. E. Holroyd, A. K. H. MacGibbon, and R. G. Snell. 2014. Estimation of genetic and crossbreeding parameters of fatty acid concentrations in milk fat predicted by mid-infrared spectroscopy in New Zealand dairy cattle. J. Dairy Res. 81:340–349. https://doi.org/10.1017/S0022029914000272.

Luke, T. D. W., S. Rochfort, W. J. Wales, V. Bonfatti, L. Marett, and J. E. Pryce. 2019. Metabolic profiling of early-lactation dairy cows using milk mid-infrared spectra. J. Dairy Sci. 102:1747–1760. https://doi.org/10.3168/jds.2018-15103.

Lum, L. S., P. Dovč, and J. F. Medrano. 1997. Polymorphisms of bovine β-lactoglobulin promoter and differences in the binding affinity of activator protein-2 transcription factor. J. Dairy Sci. 80:1389–1397. https://doi.org/10.3168/jds.S0022-0302(97)76068-5.

Lynch, M., and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates.

MacGibbon, A. K. M., and M. A. Reynolds. 2011. Milk lipids. Anal. Methods.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, 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. https://doi.org/10.1186/s12864-016-2443-6.

McDermott, A., G. Visentin, M. De Marchi, D. P. Berry, M. A. Fenelon, P. M. O'Connor, O. A. Kenny, and S. McParland. 2016. Prediction of individual milk proteins including free amino acids in bovine milk using mid-infrared spectroscopy and their correlations with milk processing characteristics. J. Dairy Sci. 99:3171–3182. https://doi.org/10.3168/jds.2015-9747.

Medrano, J., G. Rincon, and A. Islas-Trejo. 2010. Comparative analysis of bovine milk and mammary gland transcriptome using RNA-Seq. 9th World Congr. Genet. Appl. Livest. Prod. Leipz. Ger. 852.

Mele, M., G. Conte, B. Castiglioni, S. Chessa, N. P. P. Macciotta, A. Serra, A. Buccioni, G. Pagnacco, and P. Secchiari. 2007. Stearoyl-Coenzyme A desaturase gene polymorphism and milk fatty acid composition in Italian Holsteins. J. Dairy Sci. 90:4458–4465. https://doi.org/10.3168/jds.2006-617.

Moioli, B., G. Contarini, A. Avalli, G. Catillo, L. Orrù, G. De Matteis, G. Masoero, and F. Napolitano. 2007. *Short Communication*: Effect of stearoyl-Coenzyme A desaturase polymorphism on fatty acid composition of milk. J. Dairy Sci. 90:3553–3558. https://doi.org/10.3168/jds.2006-855.

Narayana, S. G., F. S. Schenkel, A. Fleming, A. Koeck, F. Malchiodi, J. Jamrozik, J. Johnston, M. Sargolzaei, and F. Miglior. 2017. Genetic analysis of groups of mid-infrared predicted fatty acids in milk. J. Dairy Sci. 100:4731–4744. https://doi.org/10.3168/jds.2016-12244.

Olsen, H. G., T. M. Knutsen, A. Kohler, M. Svendsen, L. Gidskehaug, H. Grove, T. Nome, M. Sodeland, K. K. Sundsaasen, M. P. Kent, H. Martens, and S. Lien. 2017. Genome-wide association mapping for milk fat composition and fine mapping of a QTL for de novo synthesis of milk fatty acids on bovine chromosome 13. Genet. Sel. Evol. 49:20. https://doi.org/10.1186/s12711-017-0294-5.

Palmano, K. P., and D. F. Elgar. 2002. Detection and quantitation of lactoferrin in bovine whey samples by reversed-phase high-performance liquid chromatography on polystyrene–divinylbenzene. J. Chromatogr. A 947:307–311. https://doi.org/10.1016/S0021-9673(01)01563-1.

Palombo, V., M. Milanesi, S. Sgorlon, S. Capomaccio, M. Mele, E. Nicolazzi, P. Ajmone-Marsan, F. Pilla, B. Stefanon, and M. D'Andrea. 2018. Genome-wide association study of milk fatty acid composition in Italian Simmental and Italian Holstein cows using single nucleotide polymorphism arrays. J. Dairy Sci. 101:11004–11019. https://doi.org/10.3168/jds.2018-14413.

Paton, C. M., and J. M. Ntambi. 2009. Biochemical and physiological function of stearoyl-CoA desaturase. Am. J. Physiol. Endocrinol. Metab. 297:E28–E37. https://doi.org/10.1152/ajpendo.90897.2008.

Pawlik, A., G. Sender, M. Sobczyńska, A. Korwin-Kossakowska, H. Lassa, and J. Oprządek. 2014. Lactoferrin gene variants, their expression in the udder and mastitis susceptibility in dairy cattle. Anim. Prod. Sci. 55:999–1004. https://doi.org/10.1071/AN13389.

Pegolo, S., N. Mach, Y. Ramayo-Caldas, S. Schiavon, G. Bittante, and A. Cecchinato. 2018. Integration of GWAS, pathway and network analyses reveals novel mechanistic insights into the synthesis of milk proteins in dairy cows. Sci. Rep. 8:566. https://doi.org/10.1038/s41598-017-18916-4.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575. https://doi.org/10.1086/519795.

Reynolds, E. G. M., C. Neeley, T. J. Lopdell, M. Keehan, K. Dittmer, C. S. Harland, C. Couldrey, T. J. J. Johnson, K. Tiplady, G. Worth, M. Walker, S. R. Davis, R. G. Sherlock, K. Carnie, B. L. Harris, C. Charlier, M. Georges, R. J. Spelman, D. J. Garrick, and M. D. Littlejohn. 2021. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. Nat. Genet. 53:949–954. https://doi.org/10.1038/s41588-021-00872-5.

Rosen, B. D., D. M. Bickhart, R. D. Schnabel, S. Koren, C. G. Elsik, E. Tseng, T. N. Rowan, W. Y. Low, A. Zimin, C. Couldrey, R. Hall, W. Li, A. Rhie, J. Ghurye, S. D. McKay, F. Thibaud-Nissen, J. Hoffman, B. M. Murdoch, W. M. Snelling, T. G. McDaneld, J. A. Hammond, J. C. Schwartz, W. Nandolo, D. E. Hagen, C. Dreischer, S. J. Schultheiss, S. G. Schroeder, A. M. Phillippy, J. B. Cole, C. P. Van Tassell, G. Liu, T. P. L. Smith, and J. F. Medrano. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9:giaa021. https://doi.org/10.1093/gigascience/giaa021.

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011. Predicting bovine milk protein composition based on Fourier transform infrared spectra. J. Dairy Sci. 94:5683–5690. https://doi.org/10.3168/jds.2011-4520.

Rutten, M. J. M., H. Bovenhuis, K. A. Hettinga, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J. Dairy Sci. 92:6202–6209. https://doi.org/10.3168/jds.2009-2456.

Rutten, M. J. M., H. Bovenhuis, and J. A. M. van Arendonk. 2010. The effect of the number of observations used for Fourier transform infrared model calibration for bovine milk fat composition on the estimated genetic parameters of the predicted data. J. Dairy Sci. 93:4872–4882. https://doi.org/10.3168/jds.2010-3157.

Sanchez, M. P., M. Ferrand, M. Gelé, D. Pourchet, G. Miranda, P. Martin, M. Brochard, and D. Boichard. 2017a. Short communication: Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande, and Holstein dairy cattle breeds. J. Dairy Sci. 100:6371–6375. https://doi.org/10.3168/jds.2017-12663.

Sanchez, M.-P., A. Govignon-Gion, P. Croiseau, S. Fritz, C. Hozé, G. Miranda, P. Martin, A. Barbat-Leterrier, R. Letaïef, D. Rocha, M. Brochard, M. Boussaha, and D. Boichard. 2017b. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet. Sel. Evol. 49:68. https://doi.org/10.1186/s12711-017-0344-z.

Sanchez, M.-P., Y. Ramayo-Caldas, V. Wolf, C. Laithier, M. El Jabri, A. Michenet, M. Boussaha, S. Taussat, S. Fritz, A. Delacroix-Buchet, M. Brochard, and D. Boichard. 2019. Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows. Genet. Sel. Evol. 51:34. https://doi.org/10.1186/s12711-019-0473-7.

Savitzky, A., and M. J. E. Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36:1627–1639. https://doi.org/10.1021/ac60214a047.

Schennink, A., J. M. L. Heck, H. Bovenhuis, M. H. P. W. Visker, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2008. Milk fatty acid unsaturation: Genetic parameters and effects of stearoyl-CoA desaturase (*SCD1*) and acyl CoA: Diacylglycerol acyltransferase 1 (*DGAT1*). J. Dairy Sci. 91:2135–2143. https://doi.org/10.3168/jds.2007-0825.

Schopen, G. C. B., J. M. L. Heck, H. Bovenhuis, M. H. P. W. Visker, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2009. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. J. Dairy Sci. 92:1182–1191. https://doi.org/10.3168/jds.2008-1281.

Schopen, G. C. B., M. H. P. W. Visker, P. D. Koks, E. Mullaart, J. A. M. van Arendonk, and H. Bovenhuis. 2011. Whole-genome association study for milk protein composition in dairy cattle. J. Dairy Sci. 94:3148–3158. https://doi.org/10.3168/jds.2010-4030.

Soyeurt, H., C. Bastin, F. G. Colinet, V. M.-R. Arnould, D. P. Berry, E. Wall, F. Dehareng, H. N. Nguyen, P. Dardenne, J. Schefers, J. Vandenplas, K. Weigel, M. Coffey, L. Théron, J. Detilleux, E. Reding, N. Gengler, and S. McParland. 2012. Mid-infrared prediction of lactoferrin content in bovine milk: Potential indi-

cator of mastitis. Animal 6:1830–1838. https://doi.org/10.1017/S1751731112000791.
Soyeurt, H., F. G. Colinet, V. M.-R. Arnould, P. Dardenne, C. Bertozzi, R. Renaville, D. Portetelle, and N. Gengler. 2007a. Genetic variability of lactoferrin content estimated by mid-infrared spectrometry in bovine milk. J. Dairy Sci. 90:4443–4450. https://doi.org/10.3168/jds.2006-827.
Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89:3690–3695. https://doi.org/10.3168/jds.S0022-0302(06)72409-2.
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. https://doi.org/10.3168/jds.2010-3408.
Soyeurt, H., A. Gillon, S. Vanderick, P. Mayeres, C. Bertozzi, and N. Gengler. 2007b. Estimation of heritability and genetic correlations for the major fatty acids in bovine milk. J. Dairy Sci. 90:4435–4442. https://doi.org/10.3168/jds.2007-0054.
Spelman, R., F. Miller, J. Hooper, M. Thielen, and D. Garrick. 2001. Experimental design for QTL trial involving New Zealand Friesian and Jersey breeds. Pages 393–396 in Proceedings of the Association for the Advancement of Animal Breeding and Genetics.
Tiplady, K. M., T. J. Lopdell, M. D. Littlejohn, and D. J. Garrick. 2020. The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle. J. Anim. Sci. Biotechnol. 11:39. https://doi.org/10.1186/s40104-020-00445-2.
Tiplady, K. M., T. J. Lopdell, E. Reynolds, R. G. Sherlock, M. Keehan, T. J. J. Johnson, J. E. Pryce, S. R. Davis, R. J. Spelman, B. L. Harris, D. J. Garrick, and M. D. Littlejohn. 2021. Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. Genet. Sel. Evol. 53:62. https://doi.org/10.1186/s12711-021-00648-9.
Tiplady, K. M., R. G. Sherlock, M. D. Littlejohn, J. E. Pryce, S. R. Davis, D. J. Garrick, R. J. Spelman, and B. L. Harris. 2019. Strategies for noise reduction and standardization of milk mid-infrared spectra from dairy cattle. J. Dairy Sci. 102:6357–6372. https://doi.org/10.3168/jds.2018-16144.
Tribout, T., P. Croiseau, R. Lefebvre, A. Barbat, M. Boussaha, S. Fritz, D. Boichard, C. Hoze, and M.-P. Sanchez. 2020. Confirmed effects of candidate variants for milk production, udder health, and udder morphology in dairy cattle. Genet. Sel. Evol. 52:55. https://doi.org/10.1186/s12711-020-00575-1.
Wang, Q., and H. Bovenhuis. 2018. Genome-wide association study for milk infrared wavenumbers. J. Dairy Sci. 101:2260–2272. https://doi.org/10.3168/jds.2017-13457.
Wang, Q., and H. Bovenhuis. 2019. Validation strategy can result in an overoptimistic view of the ability of milk infrared spectra to predict methane emission of dairy cattle. J. Dairy Sci. 102:6288–6295. https://doi.org/10.3168/jds.2018-15684.
Wang, Z., B. Zhu, H. Niu, W. Zhang, L. Xu, L. Xu, Y. Chen, L. Zhang, X. Gao, H. Gao, S. Zhang, L. Xu, and J. Li. 2019. Genome wide association study identifies SNPs associated with fatty acid composition in Chinese Wagyu cattle. J. Anim. Sci. Biotechnol. 10:27. https://doi.org/10.1186/s40104-019-0322-0.
Zaalberg, R. M., L. Janss, and A. J. Buitenhuis. 2020. Genome-wide association study on Fourier transform infrared milk spectra for two Danish dairy cattle breeds. BMC Genet. 21:9. https://doi.org/10.1186/s12863-020-0810-4.
Zhou, C., C. Li, W. Cai, S. Liu, H. Yin, S. Shi, Q. Zhang, and S. Zhang. 2019. Genome-wide association study for milk protein composition traits in a Chinese Holstein population using a single-step approach. Front. Genet. 10:72. https://doi.org/10.3389/fgene.2019.00072.
Zhu, B., H. Niu, W. Zhang, Z. Wang, Y. Liang, L. Guan, P. Guo, Y. Chen, L. Zhang, Y. Guo, H. Ni, X. Gao, H. Gao, L. Xu, and J. Li. 2017. Genome wide association study and genomic prediction for fatty acid composition in Chinese Simmental beef cattle using high density SNP array. BMC Genomics 18:464. https://doi.org/10.1186/s12864-017-3847-7.

ORCIDS

Kathryn M. Tiplady https://orcid.org/0000-0002-3307-9208
Thomas J. Lopdell https://orcid.org/0000-0002-7684-4870
Richard G. Sherlock https://orcid.org/0000-0001-7603-9444
Thomas J. J. Johnson https://orcid.org/0000-0003-1045-456X
Richard J. Spelman https://orcid.org/0000-0002-7968-1392
Bevin L. Harris https://orcid.org/0000-0003-0844-7539
Stephen R. Davis https://orcid.org/0000-0002-4942-1055
Mathew D. Littlejohn https://orcid.org/0000-0001-9044-047X
Dorian J. Garrick https://orcid.org/0000-0001-8640-5372

APPENDIX

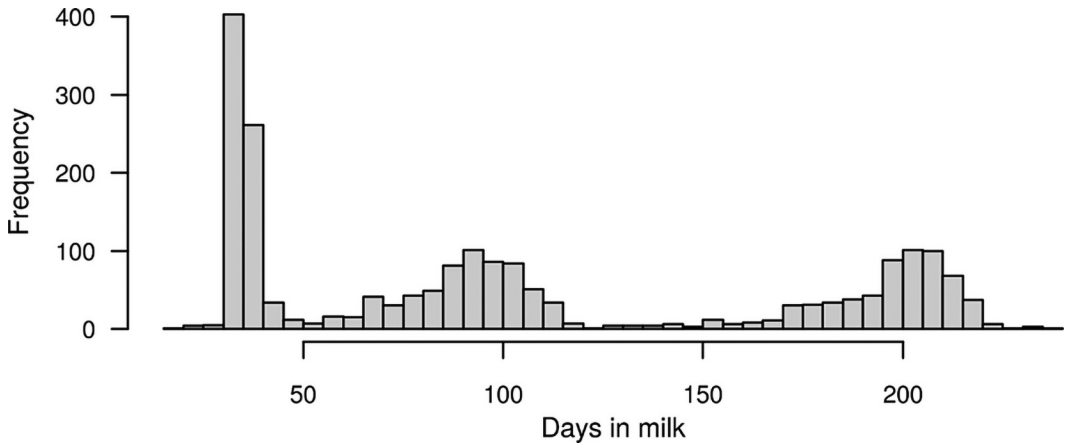

Figure A1. Frequency distribution of samples across DIM (n = 2,005).

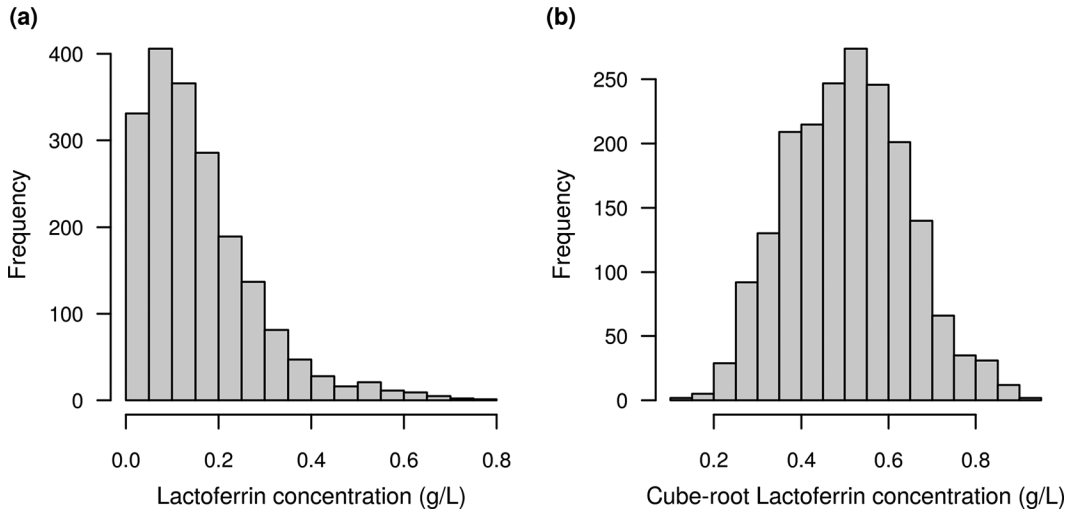

Figure A2. Frequency distributions of (a) untransformed lactoferrin concentrations and (b) lactoferrin concentrations after cube-root transformation (n = 1,936).

Appendix Table A1. Goodness of fit $\left(R^2_{cv}\right)$ of partial least squares calibration models for untreated and pretreated spectra based on cow-independent validation

	Spectral pretreatment[2]					
Trait[1]	Untreated	First derivative	MSC	MSC + first	SNV	SNV + first
Individual fatty acid (g/100 g of total fat)						
C4:0	0.627	0.617	0.625	0.602	0.623	0.602
C6:0	0.534	0.544	0.533	0.548	0.534	0.542
C8:0	0.622	0.610	0.625	0.622	0.628	0.622
C10:0	0.627	0.622	0.642	0.627	0.641	0.627
C10:1	0.344	0.360	0.353	0.365	0.348	0.360
C12:0	0.590	0.587	0.594	0.590	0.596	0.590
C12:1	0.321	0.352	0.323	0.353	0.326	0.353
C14:0	0.494	0.492	0.498	0.499	0.501	0.491
C14:1	0.408	0.413	0.418	0.416	0.412	0.414
C16:0	0.600	0.573	0.603	0.578	0.612	0.574
C16:1	0.205	0.226	0.209	0.182	0.212	0.184
C18:0	0.466	0.452	0.446	0.447	0.475	0.445
C18:1 *cis*-7	0.403	0.408	0.431	0.409	0.444	0.411
C18:1 *cis*-9	0.562	0.553	0.554	0.565	0.554	0.569
C18:2 *cis*-9,*trans*-11	0.475	0.497	0.508	0.497	0.508	0.498
C18:2 *cis*-6	0.431	0.465	0.451	0.480	0.455	0.480
C18:3 *cis*-3	0.356	0.364	0.356	0.351	0.356	0.360
Grouped fatty acid (g/100 g of total fat)						
SFA	0.587	0.590	0.601	0.595	0.598	0.591
PUFA	0.449	0.471	0.482	0.470	0.477	0.490
UFA	0.588	0.587	0.601	0.593	0.595	0.597
SCFA	0.655	0.653	0.652	0.647	0.651	0.648
MCFA	0.539	0.566	0.553	0.564	0.557	0.567
LCFA	0.561	0.567	0.563	0.569	0.568	0.568
Individual milk protein (g/L of total volume)						
α-CN	0.476	0.528	0.458	0.534	0.460	0.532
β-CN	0.193	0.185	0.184	0.188	0.184	0.190
κ-CN	0.467	0.486	0.449	0.471	0.452	0.476
α-LA	0.324	0.307	0.324	0.306	0.322	0.306
β-LG	0.660	0.686	0.667	0.675	0.661	0.678
Lf[3]	0.347	0.344	0.356	0.355	0.356	0.356
Mean R^2_{cv}	0.472	0.479	0.477	0.479	0.479	0.480

[1]Trait definitions and units as described in Table 1.

[2]Untreated = untreated spectral data; first derivative = spectra pretreated with a first-order Savitzky-Golay derivative with a window of 7 data points either side; MSC = spectra pretreated with multiplicative scatter correction; MSC + first = spectra pretreated with MSC, followed by first-derivative transformation; SNV = spectra pretreated with a standard normal variate transformation; SNV + first = spectra pretreated with SNV followed by first-derivative transformation.

[3]Cube-root transformation of lactoferrin (Lf).

Appendix Table A2. Variance component estimates for directly measured and Fourier-transform mid-infrared (FT-MIR) predicted fatty acid and protein traits

	Directly measured trait[2]				FT-MIR prediction			
Trait[1]	σ_u^2	σ_p^2	σ_e^2	σ_T^2	σ_u^2	σ_p^2	σ_e^2	σ_T^2
Individual fatty acid (g/100 g of total fat)								
C4:0	0.022 (0.009)	0.014 (0.007)	0.033 (0.001)	0.069 (0.004)	0.014 (0.006)	0.009 (0.005)	0.018 (0.001)	0.042 (0.002)
C6:0	0.005 (0.003)	0.004 (0.002)	0.016 (0.001)	0.025 (0.001)	0.003 (0.001)	0.002 (0.001)	0.006 (2e-4)	0.011 (0.001)
C8:0	0.005 (0.002)	0.003 (0.002)	0.011 (4e-4)	0.019 (0.001)	0.003 (0.001)	0.001 (0.001)	0.005 (2e-4)	0.009 (5e-4)
C10:0	0.098 (0.039)	0.032 (0.028)	0.110 (0.004)	0.241 (0.015)	0.057 (0.020)	0.008 (0.014)	0.060 (0.002)	0.125 (0.008)
C10:1	0.001 (4e-4)	0.001 (3e-4)	0.001 (1e-4)	0.003 (2e-4)	0.0003 (1e-4)	0.0002 (1e-4)	0.001 (0.00003)	0.001 (1e-4)
C12:0	0.132 (0.055)	0.064 (0.04)	0.182 (0.007)	0.378 (0.021)	0.083 (0.031)	0.020 (0.022)	0.094 (0.004)	0.197 (0.012)
C12:1	2e-4 (1e-4)	7e-5 (1e-4)	5e-4 (2e-5)	0.001 (3e-5)	1e-4 (3e-5)	4e-5 (2e-5)	2e-4 (1e-5)	3e-4 (1e-5)
C14:0	0.342 (0.154)	0.122 (0.111)	0.532 (0.021)	0.997 (0.058)	0.161 (0.065)	0.022 (0.046)	0.266 (0.011)	0.449 (0.025)
C14:1	0.021 (0.007)	0.006 (0.005)	0.011 (4e-4)	0.037 (0.003)	0.003 (0.001)	0.002 (0.001)	0.007 (3e-4)	0.012 (0.001)
C16:0	2.187 (0.804)	1.145 (0.579)	2.451 (0.098)	5.782 (0.327)	1.214 (0.424)	0.209 (0.302)	1.700 (0.068)	3.123 (0.169)
C16:1	0.008 (0.004)	0.012 (0.003)	0.022 (0.001)	0.043 (0.002)	0.002 (0.001)	0.003 (0.001)	0.007 (3e-4)	0.011 (5e-4)
C18:0	0.176 (0.130)	1.137 (0.135)	1.400 (0.056)	2.714 (0.108)	0.149 (0.087)	0.232 (0.070)	0.653 (0.026)	1.034 (0.044)
C18:1 *cis*-7	0.125 (0.053)	0.084 (0.039)	0.202 (0.008)	0.412 (0.022)	0.063 (0.026)	0.036 (0.019)	0.095 (0.004)	0.193 (0.011)
C18:1 *cis*-9	0.881 (0.379)	0.770 (0.288)	2.335 (0.093)	3.986 (0.180)	0.551 (0.234)	0.264 (0.172)	1.140 (0.046)	1.955 (0.097)
C18:2 *cis*-9,*trans*-11	0.017 (0.007)	0.012 (0.005)	0.019 (0.001)	0.048 (0.003)	0.010 (0.004)	0.004 (0.003)	0.009 (4e-4)	0.023 (0.002)
C18:2 *cis*-6	0.004 (0.002)	0.001 (0.001)	0.007 (3e-4)	0.013 (0.001)	0.002 (0.001)	0.001 (5e-4)	0.003 (1e-4)	0.006 (3e-4)
C18:3 *cis*-3	0.004 (0.001)	5e-4 (0.001)	0.005 (2e-4)	0.009 (5e-4)	0.001 (3e-4)	1e-4 (2e-4)	0.001 (4e-5)	0.002 (1e-4)
Grouped fatty acid (g/100 g of total fat)								
SFA	1.472 (0.626)	1.541 (0.478)	3.162 (0.126)	6.175 (0.291)	1.293 (0.530)	0.646 (0.381)	1.530 (0.062)	3.469 (0.206)
PUFA	0.078 (0.029)	0.026 (0.021)	0.077 (0.003)	0.181 (0.011)	0.049 (0.018)	0.017 (0.013)	0.039 (0.002)	0.105 (0.007)
UFA	1.468 (0.626)	1.544 (0.478)	3.156 (0.126)	6.167 (0.291)	1.299 (0.531)	0.640 (0.382)	1.535 (0.062)	3.474 (0.206)
SCFA	0.037 (0.020)	0.041 (0.015)	0.117 (0.005)	0.196 (0.009)	0.026 (0.013)	0.025 (0.010)	0.050 (0.002)	0.101 (0.005)
MCFA	1.293 (0.564)	0.610 (0.410)	2.302 (0.092)	4.206 (0.223)	0.797 (0.311)	0.203 (0.222)	1.158 (0.046)	2.158 (0.121)
LCFA	0.852 (0.546)	3.787 (0.549)	7.060 (0.282)	11.699 (0.445)	0.813 (0.445)	1.102 (0.355)	3.386 (0.137)	5.301 (0.223)
Individual milk protein (g/L of total volume)								
α-CN	0.579 (0.260)	0.337 (0.191)	1.112 (0.049)	2.029 (0.108)	0.559 (0.230)	0.122 (0.162)	0.428 (0.019)	1.109 (0.082)
β-CN	0.421 (0.241)	0.097 (0.193)	2.586 (0.115)	3.105 (0.126)	0.204 (0.091)	0.146 (0.066)	0.187 (0.008)	0.537 (0.035)
κ-CN	0.172 (0.067)	0.007 (0.047)	0.136 (0.006)	0.315 (0.024)	0.083 (0.031)	0.027 (0.022)	0.052 (0.002)	0.162 (0.012)
α-LA	0.008 (0.003)	0.002 (0.002)	0.009 (4e-4)	0.019 (0.001)	0.002 (0.001)	0.001 (5e-4)	0.002 (9e-5)	0.005 (3e-4)
β-LG	0.282 (0.103)	0.076 (0.072)	0.09 (0.004)	0.448 (0.037)	0.240 (0.084)	0.034 (0.058)	0.07 (0.003)	0.343 (0.03)
Lf[3]	0.007 (0.003)	2e-4 (0.002)	0.005 (2e-4)	0.0122 (0.001)	0.001 (4e-4)	0.001 (0.0003)	0.0018 (7e-5)	0.003 (2e-4)

[1]Trait definitions and units as described in Table 1. Standard errors shown in parentheses.

[2]σ_u^2 = additive genetic variance; σ_p^2 = permanent environment variance; σ_e^2 = residual variance; σ_T^2 = total variance $\left(\sigma_u^2 + \sigma_p^2 + \sigma_e^2\right)$.

[3]Cube-root transformation of lactoferrin (Lf).

Appendix Table A3. Effect sizes and minor allele frequency details for fatty acid traits with a significant association effect

Chr[1]	Position	Tag variant ID	Minor allele frequency	Trait[2]	Trait type	Beta	SE	*P*-value
Individual fatty acid (g/100 g of total fat)								
14	1756075	rs208417762	0.311	C18:1 *cis*-9	Measured	0.682	0.106	1.3e-10
14	1799066	rs385135066	0.238	C16:0	Measured	−1.039	0.146	1.2e-12
14	1799066	rs385135066	0.238	C16:0	Measured	−1.039	0.146	1.2e-12
17	52971731	rs207997694	0.085	C6:0	Measured	0.081	0.013	9.6e-10
17	53034516	rs461037541	0.083	C4:0	Measured	0.208	0.024	7.2e-18
19	51319673	rs137270097	0.265	C10:0	Measured	0.162	0.025	1.2e-10
19	51319673	rs137270097	0.263	C12:0	Measured	0.239	0.033	8.3e-13
19	51326050	rs136424304	0.262	C14:0	Measured	0.338	0.050	1.4e-11
26	21141279	rs41255696	0.476	C10:0	Measured	−0.145	0.023	2.2e-10
26	21141279	rs41255696	0.475	C14:0	Measured	−0.288	0.046	2.7e-10
26	21148111	rs41255688	0.493	C10:1	Measured	−0.037	0.003	1.8e-48
26	21149680	rs385285356	0.496	C14:1	Measured	−0.136	0.008	6.1e-61
26	26458006	rs445758306	0.318	C10:1	Measured	−0.017	0.003	2.6e-10
26	26458006	rs445758306	0.308	C12:1	Measured	−0.008	0.001	2.4e-10
11	103301736	rs41255687	0.420	C12:1	Predicted	−0.005	0.001	6.3e-11
14	2502770	rs137422574	0.414	C18:3 *cis*-3	Predicted	0.016	0.002	1.0e-12
14	2528807	rs110275497	0.415	C18:1 *cis*-9	Predicted	0.429	0.067	1.3e-10
17	52971731	rs207997694	0.085	C6:0	Predicted	0.068	0.009	9.9e-16
17	53034516	rs461037541	0.083	C4:0	Predicted	0.150	0.018	1.5e-17
19	51314476	rs41922143	0.262	C10:0	Predicted	0.134	0.019	7.0e-13
19	51314476	rs41922143	0.260	C12:0	Predicted	0.158	0.023	3.8e-12
19	51314476	rs41922143	0.264	C14:0	Predicted	0.230	0.033	7.0e-12
19	51326050	rs136424304	0.261	C8:0	Predicted	0.032	0.005	8.9e-10
26	21174891	rs209445650	0.452	C14:1	Predicted	0.029	0.005	1.9e-09
26	25584818	rs210921941	0.485	C10:1	Predicted	−0.010	0.002	5.8e-10
27	36200888	rs110950972	0.455	C18:3 *cis*-3	Predicted	0.017	0.002	9.9e-15
27	36204679	—	0.464	C16:0	Predicted	−0.485	0.080	1.6e-09
Grouped fatty acid (g/100 g of total fat)								
17	53034516	rs461037541	0.081	SCFA	Measured	0.304	0.039	1.2e-14
19	36187954	rs110980742	0.260	SFA	Measured	−0.927	0.149	5.0e-10
19	36187954	rs110980742	0.259	UFA	Measured	0.933	0.150	4.8e-10
19	51319673	rs137270097	0.265	MCFA	Measured	0.791	0.107	1.4e-13
26	21149680	rs385285356	0.495	SFA	Measured	0.818	0.129	2.1e-10
26	21149680	rs385285356	0.495	UFA	Measured	−0.832	0.129	1.1e-10
14	2319003	rs110182536	0.408	UFA	Predicted	0.593	0.097	8.1e-10
17	53034516	rs461037541	0.081	SCFA	Predicted	0.275	0.029	7.1e-22
19	50919823	rs380534925	0.171	UFA	Predicted	−0.825	0.135	8.8e-10
19	51314476	rs41922143	0.262	MCFA	Predicted	0.544	0.076	9.2e-13
26	21138011	rs381655271	0.493	UFA	Predicted	−0.628	0.099	2.6e-10

[1]Chr = chromosome.

[2]Trait definitions and units as described in Table 1.

Appendix Table A4. Effect sizes and minor allele frequency details for protein traits with a significant association effect

Chr[1]	Position	Tag variant ID	Minor allele frequency	Trait[2]	Trait type	Beta	SE	*P*-value
6	87133508	rs109500363	0.329	α-CN	Measured	0.659	0.095	4.3e-12
6	87405588	rs110794953	0.450	κ-CN	Measured	−0.412	0.038	6.4e-28
11	103291134	rs110270048	0.421	β-LG	Measured	0.838	0.036	8.7e-117
11	103292575	rs381050299	0.455	α-CN	Measured	−0.540	0.087	5.6e-10
22	53538882	rs43765460	0.457	Lf[3]	Measured	−0.072	0.005	1.8e-41
6	87085918	—	0.361	κ-CN	Predicted	0.256	0.027	8.2e-21
6	87133508	rs109500363	0.329	α-CN	Predicted	0.461	0.071	7.0e-11
11	103299272	rs110563549	0.440	β-CN	Predicted	−0.429	0.048	8.3e-19
11	103299272	rs110563549	0.420	β-LG	Predicted	0.728	0.032	5.4e-116
14	1799066	rs385135066	0.237	α-CN	Predicted	−0.527	0.076	4.8e-12

[1]Chr = chromosome.

[2]Trait definitions and units as described in Table 1.

[3]Cube-root transformation of lactoferrin (Lf).