THE ROLE OF FOURIER-TRANSFORM MID-INFRARED SPECTROSCOPY IN IMPROVING THE PREDICTION OF NEW AND EXISTING TRAITS IN NEW ZEALAND DAIRY CATTLE

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Animal Science

at Massey University, AL Rae Centre, Hamilton,

New Zealand.

Kathryn Maree Tiplady

2022

Abstract

Bovine milk is a rich source of dietary nutrients that are important to human health. Market demand for bovine milk is driven by its nutritional value, price, processability, and consumer expectations and perceptions about food production systems. The ability to quantify traits associated with milk quality, processability, animal health and environmental impact is critical for selective breeding and thus highly valuable to the dairy industry. However, obtaining direct measurements of such traits can be difficult and expensive. Estimation of major milk components using Fourier-transform mid-infrared (FT-MIR) spectroscopy is common practice, and spectral-based predictions of these traits are already widely used in milk payment and animal evaluation systems. Applications using FT-MIR spectra to predict other traits have increased in popularity over the last decade, and are attractive alternatives to directly measuring phenotypes because the FT-MIR spectra are readily available as a by-product of routine milk testing. The objectives of this thesis were to improve understanding of the phenotypic and genetic characteristics of FT-MIR spectra, and assess the role that such data can play in predicting new traits or improving the prediction of existing traits in New Zealand dairy cattle. We assessed different strategies for improving the quality of spectral data and demonstrated that there are limitations in predicting traits such as pregnancy status, due to confounding effects such as stage of lactation. From a genetics perspective, we reviewed the evolving role of spectral data in the improvement of dairy cattle by selection and discussed opportunities for consolidating spectral datasets with other genomic and molecular data sources. We conducted GWAS on individual FT-MIR wavenumbers and demonstrated that the individual wavenumbers provided stronger association effects and improved power for identifying candidate causal variants, compared to conducting GWAS on FT-MIR predicted traits. We also demonstrated the potential utility of spectral data for predicting and incorporating fatty acids and protein traits into breeding programs, but showed that even when genetic correlations between directly measured and FT-MIR predicted traits were high, the detectable QTL underpinning these traits were not always the same. Although there are many potential applications for FT-MIR spectral datasets, there are still challenges to developing robust prediction equations and understanding the genetic relationships between traits of interest and their FT-MIR predictions. Addressing these challenges will provide opportunities to improve the prediction of new and existing traits in dairy cattle milk production systems and breeding programs into the future.

Acknowledgements

Firstly, I would like to thank my supervisors for their guidance throughout the course of my PhD. Each of them contributed in different ways across the course of writing this thesis, ensuring that I was supported and had what I needed to do the work at hand. In particular, I would like to thank Professors Dorian Garrick and Matt Littlejohn for their encouragement and active involvement in providing feedback and guidance in designing and writing the papers that make up this PhD. I would also like to thank Professor Jennie Pryce for her helpful direction and insights for writing my literature review, and Professor Hugh Blair for his attention to detail and making sure I had everything I needed to complete submission. Alongside my supervisors, I would like to thank my co-authors, colleagues at LIC and fellow students at the AL Rae Centre for their contributions and support. In particular, I would especially like to thank Thomas L for his help with writing, Hieu for his work on developing machine-learning models, Andrew for generously proof-reading my final draft, Swati for her words of encouragement and inspiration to get along to the gym, and my coffee buddy Lorna who checked in on me, listened when I needed to talk, and generously fed me home-baked cookies and almond croissants.

Over the years, LIC has become somewhat of a family for me and I would like to acknowledge their financial support and sponsorship of my PhD. I am particularly grateful to Richard Spelman, Bevin Harris, Steve Davis and Ric Sherlock who were instrumental in providing the opportunity for me to do a PhD and supporting me throughout. Also, I would like to thank Lindsay Burton and the late Rob Jackson who first gave me the opportunity to work at LIC and believed in me back at my job interview in 1994, even though I was very young and knew little of the dairy industry.

I would also like to thank my Mum's extended family who always provided a sense of stability and belonging throughout my childhood years and beyond, my Dad and his wife Vicky for believing in me in the background, and the Stuart family for their unending belief in my ability to do well. In particular, I want to acknowledge the guidance and influence throughout my life of my Auntie Pat and my Auntie Sue, both of whom we sadly lost while I was doing my PhD. These two women always saw the best in me and were exceptional examples of how to look beyond oneself and take care of those around us. They are both sorely missed by us all.

As a child, I was raised predominantly by my Mum who always believed in me, and I want to acknowledge the special support she has given me, particularly over my adolescent and adult years, always there, cheer-leading me from the side, jumping in to help in any way she could, and looking out for and praying for me every step of the way. I will always be grateful for the sacrifices she made to keep a roof over our head when I was young, the strong sense of Christian faith and work ethic she instilled in me, and the unconditional love she has always shown me.

During the course of writing this thesis, there were many sacrifices by those around me, in particular by my two young children, Bertie and Beatrice. There were many times they had to entertain themselves and sacrifice time with me while I worked either at the office or at home. I am incredibly proud of how they adapted to many of the challenges we faced as a family over the last few years, particularly whilst in lockdown due to Covid. I would not have been able to do any of this without the support of my husband's parents, Allan and Jenny Tiplady. I am incredibly grateful for the support they provided with looking after Bertie and Beatrice, and giving them the best school holiday experiences, making it possible for me to forge ahead with my study.

Finally, I would like to thank my husband David for his unending patience and support, and the wisdom and stability he provides for our family. He has been the calm within my chaos on many occasions and has always been there to encourage me, remind me to take care of myself, and to make sure our family was fed when I was immersed in some aspect of completing this work. This thesis has only been made possible by his sacrifices and his love and commitment to our family.

Contents

List of Tables

List of Figures

List of Abbreviations

a_{30}	curd firmness after 30 minutes
a_{60}	curd firmness after 60 minutes
AdaptiveAvgPool	adaptive average pooling
AIC	average information criterion
ANN	artificial neural network
AUC	area under the receiver operating characteristic curve
AY	Ayrshire
BatchNorm	batch normalization
BHB	β-hydroxybutyrate
bp	base pair
BTA	*Bos taurus* autosome
BUN	blood urea nitrogen
CFR	curd firming rate
CH_4	methane
Chr	chromosome
CMS	casein micelle size
CNN	convolutional neural network
CY	cheese yield
DenseNet	dense convolutional network
DIM	days in milk

DMI . . . . . dry matter intake

DNA . . . . . deoxy-ribose nucleic acid

EBV . . . . . estimated breeding value

edQTL . . . . . RNA-editing quantitative trait loci

EN . . . . . elastic net

eQTL . . . . . expression quantitative trait loci

FEI . . . . . Fat Evaluation Index

FR . . . . . Friesian

FT-MIR . . . . . Fourier-transform mid-infrared

GC . . . . . gas chromatography

GRM . . . . . genomic relationship matrix

GWAS . . . . . genome-wide association study

H . . . . . high impact splice donor

HCT . . . . . heat coagulation time

HOL . . . . . Holstein

ICAR . . . . . International Committee for Animal Recording

ISO . . . . . International Organization for Standardization

Iter . . . . . iteration

JE . . . . . Jersey

k_{20} . . . . . curd-firming time

L . . . . . low impact splice region variant

LASSO . . . . . least absolute shrinkage and selection operator

LCFA . . . . . long-chain fatty acids

LD . . . . . linkage disequilibrium

LDA	linear discriminant analysis
LeakyReLU	leaky rectified linear unit
LIC	Livestock Improvement Corporation
LOSO	leave-one-segment-out
M	moderate impact missense variant
MAF	minor allele frequency
Mbp	megabase pair
MCFA	medium-chain fatty acids
MCP	milk coagulation properties
MD	squared Mahalanobis distance
MFA	monounsaturated fatty acids
MIR	mid-infrared
MLP	multilayer perceptron
MS	mass spectroscopy
MUN	milk urea nitrogen
NEFA	non-esterified fatty acids
NH_3	ammonia
NIR	near-infrared
NMR	nuclear magnetic resonance
NN	neural network
NP	non-pregnant
P	pregnant
PAG	pregnancy-associated glycoproteins
PCA	principal components analysis

PDS . . . . . piecewise direct standardization

PIV . . . . . putative impact variant

PKE . . . . . palm kernel extract

PLS . . . . . partial least squares

PLS-DA . . . . . partial least squares discriminant analysis

PUFA . . . . . polyunsaturated fatty acids

QTL . . . . . quantitative trait loci

RCT . . . . . rennet coagulation time

REC . . . . . nutrient recovery

RF . . . . . random forest

RFI . . . . . residual feed intake

RMSE . . . . . root mean square error

RNA . . . . . ribonucleic acid

RNA-seq . . . . . ribonucleic acid sequence

ROC . . . . . receiver operating characteristic

RPS . . . . . retroactive percentile standardization

RR . . . . . ridge regression

SCFA . . . . . short-chain fatty acids

SE . . . . . standard error

SF_6 . . . . . sulphur hexafluoride

SFA . . . . . saturated fatty acids

SNP . . . . . single nucleotide polymorphism

SVM . . . . . support vector machine

TA . . . . . titratable acidity

UFA . . . . . unsaturated fatty acids

UUN . . . . . urinary urea nitrogen

VAL-PAG . . . . . validation using pregnancy-associated glycoproteins dataset

VAL-Test . . . . . validation using test dataset

VFA . . . . . volatile fatty acid

Chapter 1

General introduction

1.1 Introduction

Bovine milk is a rich source of dietary nutrients including proteins, fats, carbohydrates, vitamins and minerals, with their concentrations being influenced by genetic factors such as breed and sire, along with non-genetic factors related to feed, environment, stage of lactation and the nutritional status of the animal. Key drivers for the market demand of bovine milk are its nutritional value, its cost, and its processability into products such as cheese and butter. Consumer expectations and perceptions about food production systems are also becoming increasingly important. In particular, consumers are concerned about the impact that animal production systems have on the environment, and animal health and welfare. It is thus important that the dairy industry achieves profitability within an efficient and sustainable framework where milk quality, animal wellbeing and reduction in the environmental footprint are key priorities.

The ability to quantify relevant traits of interest and incorporate them into dairy cattle breeding programs is of significant financial value to the dairy industry. However, obtaining direct measurements of all these traits of interest can be difficult, time-consuming and expensive. Estimation of major milk components using Fourier-transform mid-infrared (FT-MIR) spectroscopy is common practice, and spectral-based predictions of milk composition are already widely used in dairy cattle milk payment and animal evaluation systems. Applications using FT-MIR spectra to predict other traits are appealing because of the opportunity to obtain indicator traits across large numbers of animals at little or no marginal cost, due to the spectral data already being available as a by-product of routine milk testing.

Applications using FT-MIR spectra to predict traits typically involve using a set of samples with directly measured trait values to develop a calibration equation based on individual FT-MIR wavenumber absorbance values. The resulting calibration equation can then be applied to future samples, to predict trait values as a linear combination of individual wavenumber values from any milk sample with FT-MIR spectral data. The success of using FT-MIR spectra 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 a breeding program is further dependent on the extent of genetic variation present in the trait of interest, the heritability of the predicted trait, and the genetic correlation between the directly measured and predicted trait. Although there are many studies related to the genetics of FT-MIR predictions of milk composition traits, there are relatively few studies of the genetics of the individual FT-MIR wavenumbers. This is despite the individual wavenumbers exhibiting additional genetic signal that is often not observed in FT-MIR predictions of major milk composition traits.

1.2 Thesis objectives and outline

The objectives of this thesis were to improve understanding of the phenotypic and genetic characteristics of FT-MIR spectra, and assess the role that these data can play in predicting new traits or improving the prediction of existing traits in New Zealand dairy cattle. Chapter 2 provides an overview of existing applications for using FT-MIR spectra to predict traits of interest from milk samples; and summarises existing studies of the underlying genetic characteristics of FT-MIR predicted traits and individual FT-MIR wavenumbers. In Chapter 3, strategies for preparing FT-MIR spectral data for downstream analysis are discussed, and methods are compared for standardizing FT-MIR spectra from milk samples collected across a multi-instrument network. Chapter 4 compares strategies for predicting pregnancy status from FT-MIR spectra, including different ways of defining pregnant and non-pregnant cows, and different ways of accounting for stage of lactation in prediction models. The next three chapters focus on the genetics of FT-MIR predicted traits and individual FT-MIR wavenumbers. Chapter 5 provides a review of the evolving role of spectral data in the genetic improvement of dairy cattle, including a discussion of opportunities for consolidating FT-MIR datasets with other genomic and molecular data sources. In Chapter 6, we present a large sequence-based GWAS of individual FT-MIR wavenumbers, and compare the genetic signals we observe from individual FT-MIR wavenumbers to those of FT-MIR predicted major milk composition traits. Chapter 7 brings together the knowledge from all previous chapters to develop prediction equations for a number of fatty acids and protein fractions, and compare the genetic characteristics and QTL underlying directly measured traits to those for corresponding FT-MIR predicted traits. Finally, in Chapter 8, I provide a general discussion to highlight key areas of consideration for the use of FT-MIR spectra to improve dairy cattle trait prediction and advance selective breeding into the future.

1.3 PhD Supervisors

Professor Dorian Garrick

Professor Matt Littlejohn

Professor Jennie Pryce

Professor Hugh Blair

1.4 List of publications

Tiplady, K.M., Sherlock, R.G., Littlejohn, M.D., Pryce, J.E., Davis, S.R., Garrick, D.J., Spelman, R.J. and Harris, B.L., 2019. Strategies for noise reduction and standardization of milk mid-infrared spectra from dairy cattle. ***Journal of dairy science*****, 102(7), pp.6357-6372. https://doi.org/10.3168/jds.2018-16144.**

This work is included as Chapter 3 and has been published as an open access article under the Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Tiplady, K.M., Trinh, M-H., Davis, S.R., Sherlock, R.G., Spelman, R.J., Garrick, D.J. and Harris, B.L., 2022. Pregnancy status predicted using milk mid-infrared spectra from dairy cattle. ***Journal of dairy science*****, 105(4), pp.3615-3632. https://doi.org/10.3168/jds.2021-21516.**

This work is included as Chapter 4 and has been published as an open access article under the Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

Tiplady, K.M., Lopdell, T.J., Littlejohn, M.D., Garrick, D.J., 2020. The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle. ***Journal of Animal Science and Biotechnology*****, 11(1), pp.1-13. https://doi.org/10.1186/s40104-020-00445-2.**

This work is included as Chapter 5 and has been published as an open access article under the Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

Tiplady, K.M., Lopdell, T.J., Reynolds, E., Sherlock, R.G., Keehan, M., Johnson, T.J.J., Pryce, J.E., Davis, S.R., Spelman, R.J., Harris, B.L., Garrick, D.J., Littlejohn, M.D., 2020. Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. ***Genetics Selection Evolution*****, 53(1), pp.1-24. https://doi.org/10.1186/s12711-021-00648-9.**

This work is included as Chapter 6 and has been published as an open access article under the Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

1.5 Publications – *In press*

Tiplady, K.M., Lopdell, T.J., Sherlock, R.G., Johnson, T.J.J., Spelman, R.J., Harris, B.L., Davis, S.R., Littlejohn, M.D. and Garrick, D.J., (in press). Comparison of the genetic characteristics of directly measured and FT-MIR predicted bovine milk fatty acids and proteins. *Journal of Dairy Science*.

This work is included as Chapter 7 and was accepted for publication in July, 2022.

Chapter 2

Literature review

2.1 Fourier-transform infrared spectroscopy

Infrared spectroscopy is a method to determine the presence of specific chemical bonds in a composite substance and is widely used to determine the chemical composition of milk. When infrared light is shone through the milk, interactions between the infrared light and molecules cause vibrations and rotational changes in molecular bonds. When the light frequency matches the frequency of a vibrating bond or group, molecules in the milk absorb some of the light. Fourier-transform infrared spectroscopy simultaneously obtains data across a wide spectral range and transforms the raw results to absorptions for each wavelength using a Fourier-transform function. Subdivisions of the infrared region vary across different sources (Bittante and Cecchinato, 2013), but can be broadly classified into near-infrared (NIR; 14,000-4,000 cm^{-1}), mid-infrared (MIR; 4,000-400 cm^{-1}) and far-infrared (FIR; 400-10 cm^{-1}). Typically, milk composition traits are predicted from wavenumbers in the MIR range of the spectrum because absorbance coefficients are higher in the MIR range and there is a lesser effect of other factors such as water in the MIR region (McParland and Berry, 2016; Williams and Norris, 1987).

2.2 Phenotyping applications

It is common practice to use Fourier-transform mid-infrared (FT-MIR) spectra to estimate major milk components such as fat, protein and lactose for incorporating in milk payment and animal evaluation systems. Applications using FT-MIR spectral data to predict traits typically involve using a small set of samples with measured trait values to develop a calibration equation based on spectral wavenumber data. The resulting calibration equation can then be applied to future samples to predict trait values as a function of individual wavenumber absorbance values from any milk sample with FT-MIR spectral data. Because calibration equations are generally developed on small datasets with more predictors than observations, approaches such as partial least squares (PLS) regression for continuous responses and PLS discriminant analysis (PLS-DA) for binary outcomes are commonly used to reduce the predictors to a smaller set of uncorrelated components, from which least squares regression can be performed. Although PLS and PLS-DA are the most widely-used methods for developing calibration models from FT-MIR spectra, there are a number of studies that employ other approaches such as Bayesian methods (Bonfatti et al., 2017b; El Jabri et al., 2019; Ferragina et al., 2015; Toledo-Alvarado et al., 2018a) or other machine learning algorithms (Brand et al., 2021; Contla Hernández et al., 2021; Denholm et al., 2020; Dórea et al., 2018; Frizzarin et al., 2021a, 2021b; Hempstalk et al., 2015; Pralle et al., 2018). The prediction accuracies for different types of calibration models vary between studies. For example, El Jabri et al. (2019) reported that PLS models outperformed Bayesian models for predicting cheese-making

properties, but Ferragina et al. (2015) showed that Bayesian models outperformed PLS models for predicting fatty acids, whereas Bonfatti et al. (Bonfatti et al., 2017b) demonstrated that differences between prediction accuracies for PLS and Bayesian models varied depending on the trait. Further, multiple studies have reported that neural network approaches outperform PLS models for the prediction of health traits (Contla Hernández et al., 2021), pregnancy (Brand et al., 2021) and dry matter intake (Dórea et al., 2018). In contrast, Frizzarin et al. (2021a) showed that differences between prediction accuracies for PLS-DA and other machine learning approaches varied for milk quality traits, and Frizzarin et al. (2021b) demonstrated that PLS-DA models outperformed other machine learning approaches for the prediction of cow diet.

Several recent reviews outline the use of FT-MIR spectroscopy as a phenotyping strategy (Dann et al., 2018; De Marchi et al., 2014; Egger-Danner et al., 2015; Gengler et al., 2016). Ongoing research includes studies of individual fatty acids and protein fractions (Bonfatti et al., 2017d; Lopez-Villalobos et al., 2014; McDermott et al., 2016), technological properties (Cecchinato et al., 2015; Toffanin et al., 2015; Visentin et al., 2015), and indirect traits related to pregnancy (Brand et al., 2021; Delhez et al., 2020; Lainé et al., 2017; Toledo-Alvarado et al., 2018a), energy status (Grelet et al., 2016; McParland et al., 2015; Mehtiö et al., 2018), feed efficiency (McParland and Berry, 2016; Shetty et al., 2017) and methane emissions (Bittante and Cipolat-Gotet, 2018; Vanlierde et al., 2013; Vanlierde et al., 2015). In the following sections, the use of FT-MIR spectroscopy to predict traits such as individual milk fatty acids and proteins, and traits related to milk technological properties, animal health and the environment will be discussed.

2.2.1 Milk fatty acid composition

Fats and fatty acids are important nutrients in the human diet and have a key role in growth, development, hormone regulation and inflammation management. A typical fatty acid profile in bovine milk is approximately 70% saturated fatty acids (SFA), 25% monounsaturated fatty acids (MFA) and 5% polyunsaturated fatty acids (PUFA). This typical profile is unfavourable because fatty acid profiles with lower levels of saturated fats are more desirable for human health. Fat composition in milk has been a popular target for prediction using FT-MIR data from milk samples. Studies of fatty acid composition both as a percentage of total milk volume and as a percentage of total fat content presented higher accuracies when fatty acids were reported as a percentage of total milk volume (Bonfatti et al., 2016; Rutten et al., 2009; Soyeurt et al., 2006). Studies also showed that prediction accuracies for major fatty acids were more accurate than for minor fatty acids (De Marchi et al., 2011; Maurice-Van Eijndhoven et al., 2013; Rutten et al., 2009; Soyeurt et al., 2006, 2011). Rutten et al. (2009) demonstrated that increasing the number

of observations used in the prediction equations resulted in better predictions for fat composition. Also, compared to a previous study, Soyeurt et al. (2006, 2011) demonstrated higher prediction accuracy with a larger sample size, intentionally selected to provide a wider range of variation in the fatty acids. Overall, accuracies for FT-MIR predicted fatty acids have been variable, and were affected by a number of factors including breed composition, spectra pre-treatments, the number of samples used to develop prediction equations, and the variability of fatty acid composition present in the calibration samples.

2.2.2 Milk protein composition

Proteins are important nutrients in the human diet and have a key role in body maintenance and the growth and repair of cells. Bovine milk is a common source of protein, however, bovine milk and human milk differ in their concentrations of casein and whey proteins. Most of the protein in human milk is from whey, whereas bovine milk protein comprises approximately 80% casein and 20% whey proteins. Casein and whey proteins have different digestibility and amino acid profiles and also have important implications for cheese processing and the manufacture of casein supplements.

Characterization of casein and whey proteins is of value to the dairy industry because of the implications it has for human health and milk processability. Early studies assessed the capability of using FT-MIR spectra to predict protein and casein concentrations (Etzion et al., 2004; Luginbühl, 2002; Sørensen et al., 2003). Subsequent studies predicted the concentrations of lactoferrin (Lopez-Villalobos et al., 2009; Soyeurt et al., 2007a, 2012) and other individual casein and whey proteins (Bonfatti et al., 2016, 2011; De Marchi et al., 2009a; McDermott et al., 2016; Rutten et al., 2011). Across these studies, better accuracies were observed for prediction models with protein fractions expressed as a percentage of total milk volume, compared to when protein fractions were expressed as a percentage of the casein or whey content. The most accurate prediction models were observed for the studies by Bonfatti et al. (2011, 2016), but prediction accuracies varied for different protein fractions.

Overall, the reported accuracies for predicting protein fractions from FT-MIR spectra have varied across studies, with differences being partly due to the diversity of production systems and breed composition represented in calibration samples. Compared to prediction models for fatty acids, prediction models for protein fractions were less accurate and were often well below the R^2_{cv} level of 0.75 prescribed by Soyeurt et al. (2011) as a threshold for them to be useful in breeding programs. This may be potentially limiting to the value of their application in the industry.

2.2.3 Milk coagulation and other technological properties

Milk coagulation and other technological properties have an impact on the processability of milk into certain types of dairy products and are particularly important for cheese-making (De Marchi et al., 2009b; Pretto et al., 2013). Milk coagulation properties (MCP) are also affected by titratable acidity (TA), and pH levels have an influence on colloidal stability (De Marchi et al., 2009b). Multiple studies have assessed the accuracy of using FT-MIR spectra from milk samples to predict MCP and TA. Milk coagulation properties are generally reported in terms of rennet coagulation time (RCT) in minutes and curd firmness 30 or 60 minutes after rennet addition (a_{30}, a_{60}). Studies have shown that the assessment of MCP using RCT provides better prediction results, compared to assessing MCP using a_{30} (Bonfatti et al., 2016; Dal Zotto et al., 2008; De Marchi et al., 2009b, 2013). Low to moderate prediction accuracies were observed for a_{30}, but these were significantly better than those observed for a_{60}. Promising predictions of TA (De Marchi et al., 2009b) and pH (Bonfatti et al., 2016; De Marchi et al., 2009b; Visentin et al., 2015) have also been reported. Predictions of Calcium (Ca), Phosphorus (P), Magnesium (Mg) and Potassium (K) were also carried out in studies by Soyeurt et al. (2009) and Bonfatti et al. (2016). In both studies, low prediction accuracies were observed for K, and moderate to high accuracies were observed for Ca, P and Mg.

Utilising FT-MIR spectra to predict milk coagulation and cheese manufacturing traits is appealing, because it offers a high-throughput, timely and efficient method for generating indicator traits across large numbers of animals. Overall, studies have indicated that there is promising potential to use FT-MIR predictions as proxies for traits related to cheese yield and cheese-making efficiency. The success of this approach is critically dependent on ensuring that sufficient variation in the traits of interest is represented in calibration samples.

2.2.4 Animal health and energy status

Physiological changes across lactation that affect energy balance also influence milk composition and have implications for animal welfare, health and fertility. Predicting health and energy status indicators using FT-MIR spectra has been widely studied. Specifically, FT-MIR predictions of acetone and β-hydroxybutyrate (BHB) have been proposed to potentially breed cows with lower susceptibility to ketosis (van der Drift et al., 2012; van Knegsel et al., 2010), and Renaud et al. (2019) demonstrated that it is possible to evaluate hyperketonemia using a prediction from FT-MIR spectra, which can be used at a herd-level to assist with nutritional management. Grelet et al. (2016) reported promising accuracies for the prediction of BHB (R^2_{cv}=0.71), acetone (R^2_{cv}=0.73)

and citrate contents (R^2_{cv}=0.90), and Mehtiö et al. (2018) reported promising accuracies for the prediction of non-esterified fatty acid (NEFA) concentrations (R^2_{cv}=0.58).

More recently, Luke et al. (2019b) investigated the use of FT-MIR spectra from milk for predicting concentrations of metabolites in serum, using early lactation (between 5 and 49 days in milk) data for spring-calving Holstein-Friesian cows in 4 southeastern Australian dairy herds. Prediction accuracies (R^2_{cv}) in that study based on cow-independent validation were 0.48, 0.61 and 0.91 for BHB, NEFA and urea concentrations in serum, respectively; and R^2_{cv} values based on herd-independent validation were 0.60, 0.45 and 0.35 for BHB, NEFA and urea concentrations in serum, respectively. Ho et al. (2021) extended the analysis of Luke et al. (2019b) to include FT-MIR spectra records for 19 herds, collected across 3 seasons, and reported cow-independent validation R^2_{cv} values of 0.60, 0.42 and 0.87 for BHB, NEFA and urea concentrations, respectively; and R^2_{cv} values based on herd-independent validation of 0.48, 0.35 and 0.69 for BHB, NEFA and urea concentrations, respectively. These results were promising in that the R^2_{cv} values for cow- and herd-independent validation were similar. Ho et al. (2021) also validated prediction models using data from a single year to predict data collected in other years. This resulted in relatively consistent R^2_{cv} values between seasons, however the root mean square error values for these prediction models increased substantially. This finding may have been due to changes in spectral measurement across time or other differences in herd management between seasons.

Moderate to high prediction accuracies for body energy status (McParland et al., 2012, 2015) and feed efficiency traits (McParland and Berry, 2016; Shetty et al., 2017) have been reported in a number of studies. Specifically, McParland et al. (2012) developed prediction equations for energy balance, body energy content and energy intake using a consolidated dataset of Holstein and Holstein-Friesian dairy cows raised in Scotland and Ireland, respectively. Prediction accuracies (R^2_{cv}) based on multiple external validation strategies ranged from 0.22 to 0.48 for energy balance, 0.26 to 0.31 for body energy content and 0.58 to 0.64 for energy intake. In a subsequent study, McParland et al. (2015) presented record-independent validation prediction accuracies of 0.53 and 0.56 for energy balance and energy intake, respectively. A further study based on data from differing production systems in the United Kingdom and Ireland resulted in R^2_{cv} values of 0.61, 0.77 and 0.40 for energy balance, energy intake and residual feed intake (RFI), respectively (McParland and Berry, 2016). Shetty et al. (2017) examined the effectiveness of using FT-MIR data to predict dry matter intake (DMI) and RFI using a number of different validation strategies. Based on cow-independent validation, the R^2_{cv} for DMI was 0.58 for models including milk yield only, which increased to 0.72 when live weight was included in the model. The prediction accuracies of DMI models that only included FT-MIR spectra were lower (R^2_{cv}=0.30), but this increased to

0.82 when milk yield and live weight were also included as predictors. Prediction accuracies for RFI models were lower than for DMI models, with the highest cow-independent RFI prediction accuracies observed for early lactation (R^2_{cv}=0.29), compared to across-lactation or mid- and late-lactation R^2_{cv} values which were 0.20, 0.09 and 0.09, respectively. Shetty et al. (2017) also showed that for DMI, most of the variation was from wavenumbers in FT-MIR spectral regions associated with milk fat, whereas for RFI, most of the variation was from wavenumbers in spectral regions associated with milk protein. Overall, findings from these studies indicate the potential value in predicting energy status indicators from FT-MIR spectra. However, prediction accuracies varied between traits and were dependent on the strategy used for validation. In general, the best results were achieved when the phenotypic variation in the prediction population was captured within the dataset used to evaluate prediction equations, and improvements were made when a diverse range of breeds and production systems were included in the calibration dataset.

Pregnancy results in changes to metabolic and energy requirements and the partitioning of resources to different physiological functions, and has a consequent influence on milk composition (Loker et al., 2009; Penasa et al., 2016). Previous studies have examined the impact of pregnancy on detailed milk composition as determined by FT-MIR spectra (Lainé et al., 2017) and the ability to use FT-MIR spectra to predict conception outcomes (Hempstalk et al., 2015; Ho et al., 2019; Ho and Pryce, 2020) or pregnancy (Brand et al., 2021; Delhez et al., 2020; Toledo-Alvarado et al., 2018a). Lainé et al. (2017) observed that the effect of pregnancy was highly variable between mid-infrared wavenumber regions, and that at the start of pregnancy, for some wavenumbers, the relative effect of pregnancy was higher than for milk yield and fat and protein concentrations. In particular, they observed the highest effects of pregnancy in mid-infrared wavenumbers in the region from 968 to 1,577 cm^{-1}.

Improvements in accuracy from incorporating FT-MIR spectra into the prediction of conception and pregnancy status have varied between studies. Hempstalk et al. (2015) assessed the accuracy of predicting conception status from herd- and cow-level factors as well as FT-MIR spectra using a variety of machine learning algorithms. Overall, their findings were that FT-MIR spectra did not improve the accuracy of conception status predictions, above what was possible from using other herd- and cow-level information. In contrast, Toledo-Alvarado et al. (2018a) found that the incorporation of FT-MIR spectra into pregnancy prediction equations improved prediction accuracy. In that study, they assessed and compared the ability to predict pregnancy from milk components (fat, protein, lactose and casein) or from a single wavenumber or from a full set of FT-MIR spectral wavenumbers. The best predictions of pregnancy were obtained using a full set of FT-MIR spectral wavenumbers. Adjustment for herd and year effects improved predictions

even further, but notably, the incorporation of that information may not be easily implementable for timely predictions. Predictions from a single wavenumber, 1,546 cm^{-1}, had area under the receiver operating characteristic (ROC) curve values of 0.55 to 0.58, whereas predictions based on a full set of FT-MIR spectra had area under the ROC curve values of 0.60 to 0.66.

The relationship between the oestrous cycle and milk composition was assessed by Toledo-Alvarado et al. (2018b). They found that milk composition varied during different phases of the oestrous cycle, with fatty acid profiles and major milk components (fat, protein, lactose and casein) all being significantly affected by oestrous cycle phase. That study demonstrated that FT-MIR spectra may be used as a diagnostic tool to predict oestrus phase, but because routine milk testing in New Zealand is usually conducted only every 2 to 3 months, the practical application of such an approach may be limited. However, if advances in technology enable the use of miniaturized inline spectrometers in milking sheds, FT-MIR measurements would be available on farm and more frequently, and this approach could be useful.

2.2.5 Nitrogen

The environmental impact of nitrogen losses into waterways and ammonia (NH_3) volatization into the atmosphere from dairy production is of key interest to producers and consumers of dairy products. Nitrogen losses result from excess nitrogen in the cow's diet which is excreted in urine and faeces. A large proportion of the nitrogen in urine is in the form of urea, which is a potential source of NH_3 emission into the soil, waterways and atmosphere. The relationship between urinary urea nitrogen (UUN), blood urea nitrogen (BUN) and milk urea nitrogen (MUN) is complex. However, several studies have indicated that it is possible to predict UUN from MUN (Jonker et al., 1998; Kauffman and St-Pierre, 2001; Nousiainen et al., 2004; Zhai et al., 2007). Estimation of protein intake and dry matter intake (DMI) from pasture using MUN may also be possible (Jonker et al., 1998; Nousiainen et al., 2004), and strong relationships have been established between MUN and NH_3 emissions (Burgos et al., 2010; Powell et al., 2011). Notably, Spek et al. (2013) reported on the effect of dietary and animal factors on the excretion of UUN and showed substantial improvements in predictions of UUN when total urine collection was used, instead of spot sampling. Using FT-MIR spectra to predict serum concentrations of urea has been discussed in a previous section, with respect to studies of metabolic profiling and animal health (Ho et al., 2021; Luke et al., 2019b). In general, prediction accuracies were promising, but were variable between studies and were influenced by the validation strategy that was used.

Using FT-MIR spectra to predict MUN or BUN would enable high-throughput nitrogen phenotyping for large numbers of animals. However, there may be difficulties in breeding for low-nitrogen output cows based on FT-MIR predictions because of the large amount of between animal variation and the strong effect of feed composition. Potentially, in breeding for high protein-producing cows, those cows would be making more efficient use of nitrogen, and consequently would have less UUN output. Work is ongoing to determine the role that FT-MIR predictions of MUN and BUN may have in reducing the impact of nitrogen outputs from dairy systems. It is likely that solutions will be multi-faceted and will also include a large emphasis on diet and housing systems.

2.2.6 Methane

The agricultural sector contributed half of New Zealand's gross greenhouse gas emissions in 2020, of which 39.2% was methane emissions from dairy cattle (Ministry for the Environment, NZ, 2022). Reducing methane emissions from dairy production is thus important to ensure that New Zealand achieves its International climate change commitments. Breeding for cows with less impact on the environment is a potential part of the solution, but obtaining direct measurements of methane using methods such as respiration chambers, the sulphur hexafluoride (SF_6) tracer technique or the GreenFeed system is difficult and expensive. Using FT-MIR spectra to generate indicator traits for methane are an attractive alternative because they have the potential to provide methane predictions across large numbers of individuals at very low cost.

Methane predictions from FT-MIR spectra have recently been compared to predictions from milk fatty acids determined by gas chromatography (GC; van Gastelen et al., 2018a). Results indicated that GC-determined milk fatty acids were better predictors of methane, but that combining FT-MIR spectra with other information such as feed intake and stage of lactation improved the predictive ability of the FT-MIR spectra. There are a handful of other studies that have assessed the potential utility of using FT-MIR spectra directly to predict methane. Dehareng et al. (2012) developed FT-MIR spectra prediction models for methane as determined by the SF_6 tracer technique using a small number of animals (n=11), with methane measurements and milk samples collected on a daily basis during a 7-day period. The accuracy to predict methane outputs as assessed by leave-one-out cross-validation was high (R^2_{cv}=0.87), better than that for fatty acids (R^2_{cv}=0.76). Vanlierde et al. (2013) extended those models to include more animals (n=146) of multiple breeds and from different countries, resulting in a 50-group cross-validation accuracy of 0.70. In a subsequent study using the same dataset, Vanlierde et al. (2015) also showed that the accuracy of FT-MIR prediction equations could be improved by including stage of lactation, modelled using linear and quadratic Legendre polynomials.

Predictions of methane outputs using FT-MIR spectra based on respiration chamber measurements have also been studied (Denninger et al., 2020; Vanlierde et al., 2018, 2021; Wang and Bovenhuis, 2019). In a study of 584 respiration chamber measurements, Vanlierde et al. (2018) reported a 5-group cross-validation accuracy of 0.57. In a subsequent study, Vanlierde et al. (2021) combined SF_6 and respiration chambers measurements from previous studies (Vanlierde et al., 2015, 2016, 2018) to develop models that accounted for more of the observed variability in CH_4 emissions. However, when Denninger et al. (2020) applied the CH_4 prediction equation developed by Vanlierde et al. (2016) to an independent dataset, it was not possible to differentiate between low and high emitting cows based on average daily CH_4 emissions measured with either respiration chambers or laser CH_4 detectors. Moreover, in a study by Wang and Bovenhuis (2019) with respiration chamber measurements for 801 dairy cows, prediction accuracy as assessed by random cross-validation was promising (R^2_{cv}=0.49), but the prediction accuracy as assessed by herd-independent validation was poor (R^2_{cv}=0.01). These findings highlight that in some instances, random cross-validation can give an overly optimistic view of the quality of FT-MIR predictions. Overall, further work is required to develop FT-MIR prediction equations for methane that are robust and transferable to independent spectral datasets. Issues related to uncertainties and discrepancies in methane datasets and measurement methods still need to be addressed, but there may be potential to improve prediction accuracy through collaboration and by ensuring that datasets represent a range of breeds, diets and production systems (Hristov et al., 2018; Vanlierde et al., 2018).

2.3 Pre-processing of FT-MIR spectra

Although there are many potentially valuable applications for FT-MIR spectra, the ability to predict traits directly from the spectra and to transfer prediction equations between instruments is hindered by a number of sources of unwanted variation. These sources of variation include scaling and baseline effects in spectral measurement, low repeatability of sample measurement for specific regions of the infrared spectrum influenced by the water content of milk, and also systematic variation between measurements from different instruments and within instruments across time due to factors such as temperature fluctuations and wavelength or detector intensity instability (Wang et al., 1991). Addressing these sources of variation appropriately is important, and would provide an opportunity to increase prediction accuracy and improve the utility of using the FT-MIR spectra in downstream applications.

2.3.1 Pre-processing treatments

Applying pre-processing treatments to FT-MIR spectra before generating prediction models is a widely-used practice. The objective in doing so is to correct signal noise in the spectral profile whilst still retaining important features. Common methods for pre-processing are multiplicative scatter methods (Geladi et al., 1985; Martens et al., 2003) or derivation methods such as the Savitzky-Golay derivative (Savitzky and Golay, 1964). Multiplicative scatter correction is a normalization method that corrects spectra for scaling and baseline effects by comparing the spectra to an expected spectral profile, where the expected profile is based on the overall average of all spectral responses. Derivation methods are based on changes in the spectra across specified window sizes and are intended to smooth the spectra whilst retaining key features of its shape. There are a number of studies where prediction accuracies from models using different pre-treatments of spectra have been compared. The findings of these studies vary, with Soyeurt et al. (2011) and De Marchi et al. (2011) reporting that a 1st derivative treatment provided the best prediction equations for fatty acids; De Marchi et al. (2009b, 2013) reporting that untreated spectra provided the best prediction equations for milk coagulation properties and acidity traits; and Bonfatti et al. (2011) reporting varying results for a range of spectra pre-treatments applied to spectra prior to evaluating protein fraction prediction equations. Overall, there is no consensus about the best pre-processing treatment to apply to spectral data. Each dataset has its own unique characteristics and this will determine the effectiveness of each approach. Notably, even when different pre-processing strategies are examined in a study, authors often only report the best prediction models, and this makes it difficult to compare the effectiveness of different pre-processing strategies (De Marchi et al., 2014).

2.3.2 Noise regions of the FT-MIR spectrum

Water content in milk samples results in high noise levels in some bands of the infrared spectrum. Bands of the spectrum associated with high noise levels due to water absorption are generally reported in the O-H bending (~1,600 to 1,700 cm^{-1}) and O-H stretching bands (>~3,000 cm^{-1}). However, the boundaries of these regions vary between publications: 1,616 to 1,678 cm^{-1}, 3,066 to 3,668 cm^{-1} (Soyeurt et al., 2010); 1,586 to 1,698 cm^{-1}, 3,052 to 3,669 cm^{-1} (Bittante and Cecchinato, 2013); and 1,600 to 1,689 cm^{-1}, 3,008 to 5,010 cm^{-1} (Grelet et al., 2015). Notably, most studies, including those mentioned above do not report wavenumbers lower than 925 cm^{-1}, because the milk samples have been analysed on FOSS instruments (Hillerød, Denmark) which do not report any spectral results from that region.

Although it is common practice to remove spectra from noise regions such as those mentioned above, studies indicate that some wavenumbers within commonly-defined noise regions may still carry valuable information. Wavenumbers in the regions between 1,619 to 1,674 cm^{-1} and 3,073 to 3,667 cm^{-1} have been associated with a polymorphism in the *DGAT1* gene (Wang et al., 2016; Wang and Bovenhuis, 2018), a gene that has been shown to have major impacts on milk composition (Grisart et al., 2002). Similarly, Toledo-Alvarado et al. (2018a) reported a significant association between pregnancy status and the 3,683 cm^{-1} wavenumber. Bittante and Cecchinato (2013) also showed that the transmittance for most FT-MIR wavenumbers in the range from 930 to 5,000 cm^{-1} was heritable. They concluded that, although heritability estimates were often low in the water absorption regions from 1,698 to 1,586 cm^{-1} and 3,052 to 3,669 cm^{-1}, those regions should still be considered for investigation, because they include absorbance peaks for chemical bonds related to non-water milk components.

2.3.3 Identifying and removing outliers

Noise in FT-MIR spectra can result from outliers caused by sample or instrument anomalies. Outliers are generally identified using multivariate approaches such as the squared Mahalanobis distance (MD) which is an indicator of the distance between a spectral record and the average spectral response. Notably, many studies are based on spectra from a single instrument so do not have the complication of differing variance-covariance structures from different instruments. Developing robust strategies for identifying noise regions and detecting outliers when spectra are collected across multiple instruments is important, because covariance structures in high-dimensional datasets can be highly sensitive and susceptible to variance inflation.

2.3.4 Standardization

Standardization of FT-MIR spectra is a methodology used to reduce the impact of variation between instruments or within instruments across time. This variation, existing even between instruments of the same brand can result in prediction errors and bias, and is particularly problematic when applying prediction equations developed on one instrument across a historical database of spectra collected on other instruments (Bonfatti et al., 2017d; Grelet et al., 2015). A widely-used practice to address the issue of variation between instruments and shifts in instruments across time is to adjust trait predictions by instrument correction coefficients, previously evaluated from the analysis of reference samples using the approach outlined by Lynch et al. (2006). However, that method is only applicable when trait-specific reference samples are available. With the growing number of traits predicted from FT-MIR spectra, many of which do not have reference

samples, there is increased interest in standardizing individual FT-MIR spectra wavenumbers directly (Bonfatti et al., 2017a; Grelet et al., 2015, 2017). Recent publications by Grelet et al. (2015) and Bonfatti et al. (2017a) have presented methods for standardizing individual wavenumber absorbance values. Both strategies involved assigning a primary instrument and standardizing spectra from other (secondary) instruments to align them to the spectral response of the primary instrument. Grelet et al. (2015) presented a piecewise direct standardization (PDS) approach based on the method described by Wang et al. (1991). Bonfatti et al. (2017a) presented a retroactive approach (RPS) where percentiles of spectral responses for each wavenumber were used to map the primary/secondary instrument relationships. Both the PDS and RPS approaches successfully reduced prediction errors when transferring prediction equations between instruments for fat composition traits (Bonfatti et al., 2017a; Grelet et al., 2015). Grelet et al. (2017) also demonstrated the effectiveness of the PDS approach for reducing prediction errors for traits with lower quality calibration equations such as methane emissions and cheese yield. Whilst there are clear benefits for standardizing FT-MIR spectra wavenumbers, to date there have been no studies that directly compare the reduction in prediction errors for the two methods when measured across the same dataset.

2.4 The genetics of FT-MIR predicted traits

The success of incorporating FT-MIR predicted traits into breeding programs is dependent on the genetic parameters of measured and predicted traits, and the genetic correlations between the measured and predicted trait values (Bonfatti et al., 2016). The genetic parameters of individual fatty acids and protein fractions (Lopez-Villalobos, 2012) and milk coagulation properties (Bittante et al., 2012) have been recently reviewed. Moderate to high heritability estimates have been reported for many FT-MIR predicted individual fatty acids (Bonfatti et al., 2017d; Lopez-Villalobos et al., 2014; Rutten et al., 2010; Soyeurt et al., 2007b). Moderate to high heritability estimates have also been reported for FT-MIR predicted grouped fatty acids, with consistently higher heritability estimates for saturated fat and short- and medium-chain fatty acid groups, compared to unsaturated fat and long-chain fatty acid groups (Fleming et al., 2018; Hein et al., 2018; Narayana et al., 2017). Two studies also evaluated genetic parameters for both directly measured and FT-MIR fatty acids, and reported genetic correlations between measured and predicted traits that were predominantly above 0.95 (Bonfatti et al., 2017d; Rutten et al., 2010). There are fewer studies reporting genetic parameters for FT-MIR predicted milk proteins (Bonfatti et al., 2017d; Buitenhuis et al., 2016; Sanchez et al., 2017a). Buitenhuis et al. (2016) reported

on individual proteins as a proportion of total protein or whey protein, whereas Bonfatti et al. (2017d) and Sanchez et al. (2017a) reported on individual proteins as a proportion of total protein and as a proportion of total milk volume. Only the study by Bonfatti et al. (2017d) presented genetic parameters for both directly measured and FT-MIR predicted milk traits, including genetic correlations between measured and predicted milk proteins. These were generally moderate to high, ranging from 0.231 for α_{s1}-casein to 0.822 for $\varkappa$-casein.

Moderate to high heritability estimates have been reported for FT-MIR predicted milk coagulation traits (Cecchinato et al., 2009; Costa et al., 2019; Visentin et al., 2017). Of those studies, Cecchinato et al. (2009) was the only one that presented genetic correlations between FT-MIR predicted and measured coagulation traits, which ranged from 0.91 to 0.96 for rennet coagulation time (RCT), and from 0.71 to 0.87 for curd firmness after 30 minutes (a_{30}). Moderate to high heritability estimates have also been reported for FT-MIR predicted minerals (Costa et al., 2019; Sanchez et al., 2018). Heritability estimates in those studies were consistently lower for sodium (0.32 to 0.38) and consistently higher for phosphorus (0.53 to 0.56). In studies of cheese yield and nutrient recovery traits, moderate heritability estimates have been reported (Bittante et al., 2014; Cecchinato et al., 2015). In those studies, heritability estimates for protein nutrient recovery were typically higher than for other traits, ranging from 0.32 to 0.44. Bittante et al. (2014) also presented genetic correlations between measured and FT-MIR predicted cheese yield and nutrient recovery traits, which ranged from 0.76 to 0.98 for cheese yield traits, and from 0.79 to 0.98 for nutrient recovery traits.

Although health, fertility and environment traits are valuable targets for breeding programs, there are relatively few studies of the genetic parameters of FT-MIR predictions for these traits. One such study reported a heritability of 0.16 for FT-MIR predictions of the probability of conception to first mating (MFERT), which was higher than the heritability of 0.05 they observed for traditional fertility traits (van den Berg et al., 2021a). In that study, genetic correlations between MFERT and traditional fertility traits were low to moderate, with the weakest correlation being with pregnancy at the end of the mating season (0.13), and the strongest correlation being with calving to first service (-0.61). Genetic parameter estimates for FT-MIR predicted blood β-hydroxybutyrate (BHB) vary between studies. Belay et al. (2017) presented moderate heritability estimates for FT-MIR predicted blood BHB, ranging from 0.25 to 0.37 across different stages of lactation, and moderate genetic correlations between clinical ketosis and FT-MIR predicted blood BHB (0.47). Heritability estimates for FT-MIR predicted blood BHB were lower in other studies, ranging from 0.04 to 0.09 (van den Berg et al., 2021b; Luke et al., 2019a).

Moderate heritability estimates have also been reported for traits relating to methane, ranging from 0.22 to 0.25 for predicted daily CH_4 emissions and 0.17 to 0.18 for log-transformed predicted CH_4 intensity (Khanal and Tempelman, 2022). For FT-MIR predicted MUN, moderate to high heritability estimates, ranging from 0.38 to 0.59 were reported by Miglior et al. (2007) and Wood et al. (2003), with lower estimates of 0.22 and 0.14 presented in studies by Mitchell et al. (2005) and Stoop et al. (2007), respectively. Of those studies, the only one that reported genetic correlations between wet-chemistry measurements of MUN and FT-MIR predicted MUN was by Mitchell et al. (2005), which were 0.38 and 0.23 in lactations 1 and 2, respectively. These genetic correlations are significantly lower than those reported for fatty acids (0.82 to 0.99; Rutten et al., 2010) and milk processability traits (0.76 to 0.98; Bittante et al., 2014), and indicate that wet-chemistry measurements of MUN and FT-MIR predicted MUN are genetically different traits. However, recently, promising genetic parameter estimates have been reported for FT-MIR predicted BUN, with heritability estimates ranging from 0.08 to 0.13, and genetic correlations between BUN and its FT-MIR prediction ranging from 0.96 to 0.98 (van den Berg et al., 2021b).

Moderate to high heritability estimates across many individual fatty acids and proteins, and traits related to milk processability indicate that these traits have genetic variation that could potentially be exploited for the purposes of animal selection. Moreover, for many of these traits, high genetic correlations between direct measurements and FT-MIR predictions indicate that selection based on FT-MIR predictions could provide favourable genetic gains in the true traits of interest. The potential for incorporating FT-MIR predicted animal health and environmental indicators into breeding programs is less clear due to the lack of studies reporting genetic parameters for these traits. Low to moderate heritability estimates have been reported for FT-MIR predicted blood BHB and probability of conception to first mating (Belay et al., 2017; van den Berg et al., 2021a, 2021b; Luke et al., 2019a). Moderate heritability estimates have also been reported for methane traits (Khanal and Tempelman, 2022), however, there are still issues to be resolved to improve the accuracy and robustness of prediction equations to make them applicable across a broader range of production systems and environments (van Gastelen et al., 2018b; Hristov et al., 2018; Negussie et al., 2017; Vanlierde et al., 2018). For FT-MIR predicted MUN, large differences in heritability estimates between studies indicate that there may be underlying instability in prediction equations, and highlight the importance of developing prediction models that are robust across different breeds and production systems. Notably, promising results have recently been reported for FT-MIR predicted BUN (van den Berg et al., 2021b). More research is required to determine the role that FT-MIR predicted animal health and environment traits could have in improving animal health and reducing methane and nitrogen outputs from dairy systems.

2.5 The genetics of FT-MIR spectra

Although there are many studies reporting genetic parameter estimates of FT-MIR predicted traits, there are relatively few studies reporting genetic parameter estimates for the individual spectral wavenumbers. Within those studies, the transmittance of FT-MIR spectral wavenumbers were moderately to highly heritable across a large proportion of the mid-infrared region (Bittante and Cecchinato, 2013; Rovere et al., 2019; Soyeurt et al., 2010; Wang et al., 2016; Zaalberg et al., 2019). Although heritability estimates were consistently low in regions affected by the water content in milk, estimates greater than 0.2 were still reported across most of the mid-infrared region (Soyeurt et al., 2010; Wang et al., 2016). This indicates that there may be potential to achieve genetic gain through the direct use of FT-MIR spectra for selection, rather than selection based on indirect predictions of the composite production traits, which are themselves a function of the FT-MIR spectral wavenumbers. Previous studies have compared indirect vs direct approaches to using FT-MIR spectra to calculate estimated breeding values (EBV) for traits (Bonfatti et al., 2017c, Dagnachew et al., 2013). The indirect approach is the commonly used method whereby the EBV are evaluated from the FT-MIR predicted trait using a single-trait mixed model. Alternatively, the direct approach evaluates trait EBV as a function of individual FT-MIR wavenumber EBV. Typically, the latter approach involves reducing the dimensionality of the spectra to a smaller subset of latent variables and estimating the variance components of the latent traits in a multivariate model. Latent trait EBV are subsequently back-transformed and used to evaluate predicted trait EBV.

In a study of spectral data from dairy goats, Dagnachew et al. (2013) showed that prediction error variances for EBV were reduced for major milk components when a direct approach was used, compared to using an indirect approach. However, Bonfatti et al. (2017c) showed that differences in prediction accuracies for indirect and direct approaches varied depending on the trait, and in particular were sensitive to the spectral variability captured within the latent variables used to evaluate the trait EBV. A subsequent study by Belay et al. (2018) compared EBV prediction accuracies for indirect and direct approaches using simulated traits with different genetic and residual correlation structures. They showed that a direct approach could be sensitive to whether coefficients relating latent traits to the trait of interest were based on phenotypic or genetic relationships. Notably, in the simulation study by Belay et al. (2018), a simplified model with only two predictors (equivalent to two latent variables) was used, but in practice, more latent variables would be required to effectively capture spectral variation. Indeed, to capture 99% of

spectral variation, Soyeurt et al. (2010) identified that 46 latent variables were required and Bonfatti et al. (2017c) identified that 8 latent variables were required. Bonfatti et al. (2017c) also highlighted that when a principal component analysis (PCA) approach is used to reduce dimensionality in spectral data, latent variables are based on the overall spectral variation, and may not appropriately capture the genetic variation of traits less correlated with fat and protein. Evaluating trait-specific latent variables using a supervised PCA approach such as PLS may address this, however, a large number of latent variables may still be required to appropriately capture the spectral variability associated with the trait. More research is required to understand how the genetic variation present in the spectra can be used to improve prediction of trait EBV.

2.5.1 Genome-wide association studies of FT-MIR spectra wavenumbers

Although there are currently many genome-wide association studies (GWAS) for FT-MIR predicted milk production traits such as fat, protein, and lactose concentrations (Jiang et al., 2010; Kemper et al., 2015b; Littlejohn et al., 2016; Lopdell et al., 2017; Raven et al., 2014), and fatty acids and protein fractions (Cruz et al., 2019; Freitas et al., 2020; Iung et al., 2019; Olsen et al., 2017; Sanchez et al., 2017b, 2019), there have been relatively few GWAS for individual FT-MIR wavenumbers. Two such studies conducted GWAS on medium density SNP-chip (~50k markers) genotypes for a subset of wavenumbers, identified either by clustering analysis (Wang and Bovenhuis, 2018), or by using phenotypic correlation structures and heritability estimates within each breed (Zaalberg et al., 2020). A third study explored relationships between FT-MIR wavenumber phenotypes and a subset of SNP previously implicated in a GWAS of milk composition and fatty acid traits (Benedet et al., 2019). Across those studies, a number of FT-MIR wavenumber QTL were identified. Most of the implicated genomic regions had been previously reported in studies of major milk composition traits, but new regions with potential links to milk components such as phosphorus, orotic acid or citric acid were also identified (Wang and Bovenhuis, 2018). Overall, those studies indicated that there is potential to conduct GWAS on individual wavenumbers to further our understanding of the underlying genetics of milk composition, and that these insights could be used for improving dairy cattle breeding programs.

2.6 Summary

Fourier-transform mid-infrared spectroscopy offers a high-throughput and inexpensive method for predicting milk composition and other novel traits. This includes traits related to milk quality, animal health and the environment. Studies have reported promising prediction accuracies for fatty acids, protein fractions, cheese-making characteristics and energy status. However, prediction accuracies have been variable and were affected by a number of factors including breed composition, spectra pre-treatments, the number of samples used in calibration models, and how well the variability of the validation population was represented in the calibration samples. Recent studies have highlighted the potential to use FT-MIR spectra to predict oestrus phase, pregnancy and environmental traits such as methane and nitrogen outputs. However, more research is required to improve the prediction quality for these traits. Although there have been many studies related to the genetics of FT-MIR predicted traits, there are relatively few studies of the genetics of individual FT-MIR wavenumbers. This is despite the individual wavenumbers exhibiting additional genetic signals that are often not observed in FT-MIR predicted traits. Indications are that individual FT-MIR wavenumbers may provide an additional layer of granularity to assist with establishing causal links between the genome and observed phenotypes to enable the discovery of novel QTL. However, conducting GWAS on such large numbers of phenotypes presents computational challenges due to the size and complexity of the datasets involved. Addressing these challenges promises to enhance our knowledge of milk composition and improve future dairy cattle breeding programs.

Chapter 3

Strategies for noise reduction and standardization of FT-MIR spectra from dairy cattle

Originally published as: Tiplady, K.M., Sherlock, R.G., Littlejohn, M.D., Pryce, J.E., Davis, S.R., Garrick, D.J., Spelman, R.J. and Harris, B.L., 2019. Strategies for noise reduction and standardization of milk mid-infrared spectra from dairy cattle. *Journal of dairy science*, 102(7), pp.6357-6372. https://doi.org/10.3168/jds.2018-16144.

3.1 Interpretive summary

Fourier-transform mid-infrared (FT-MIR) spectra from milk samples are valuable resources because they are routinely available and can be used to predict traits that are difficult or expensive to measure directly. Noise in FT-MIR spectra is problematic because it reduces prediction accuracy. This study develops strategies for reducing the impact of noise and compares methods for standardizing FT-MIR spectra across multiple-instrument networks. Our results demonstrate that standardization using spectra from milk-based reference samples is the most consistent method for reducing prediction errors across time. Implementing this approach will improve the quality of predictions based on FT-MIR spectra for various downstream applications.

3.2 Abstract

The use of Fourier-transform mid-infrared (FT-MIR) spectroscopy is of interest to the dairy industry worldwide for predicting milk composition and other novel traits that are difficult or expensive to measure directly. Although there are many valuable applications for FT-MIR spectra, noise from differences in spectral responses between instruments is problematic, because if ignored, it reduces prediction accuracy. The purpose of this study was to develop strategies for reducing the impact of noise and to compare methods for standardizing FT-MIR spectra, to reduce between-instrument variability in multiple-instrument networks. Noise levels in bands of the infrared spectrum due to the water content of milk were characterised, and a methodology for identifying and removing outliers was developed. Two standardization methods were assessed and compared: piecewise direct standardization (PDS) which related spectra on a primary instrument to spectra on five other (secondary) instruments using identical milk-based reference samples (n=918) analysed across the six instruments; and retroactive percentile standardization (RPS) whereby percentiles of observed spectra from routine milk test samples (n=2,044,094) were used to map and exploit primary and secondary-instrument relationships. Different applications of each method were studied to determine the optimal way to implement each method across time. Industry-standard predictions of milk components from 2,044,094 spectra records were regressed against predictions from spectra before and after standardization using PDS or RPS. The PDS approach resulted in an overall drop in root mean square error between industry-standard predictions and predictions from spectra from 0.190 to 0.071 g/100mL for fat, 0.129 to 0.055 g/100mL for protein and 0.143 to 0.088 g/100mL for lactose. Reductions in prediction error

for RPS were similar but less consistent than those for PDS across time, but similar reductions were achieved when PDS coefficients were updated monthly and separate primary instruments were assigned for North and South Islands. We demonstrate that the PDS approach is the most consistent method for reducing prediction errors across time. We also show that the RPS approach is sensitive to shifts in milk composition, but can be used to reduce prediction errors, provided that secondary-instrument spectra are standardized to a primary instrument with samples of broadly equivalent milk composition. Appropriate implementation of either of these approaches will improve the quality of predictions based on FT-MIR spectra for various downstream applications.
Key words: *Fourier-transform mid-infrared spectra, standardization, trait prediction, milk composition, dairy cattle*

3.3 Introduction

Fourier-transform infrared spectroscopy is a method to determine light absorbance at wavenumbers across the infrared spectrum. Applications using Fourier-transform infrared data from the mid-infrared range have increased in popularity over the last decade for predicting milk composition and other novel traits. De Marchi et al. (2014) comprehensively reviewed the use of Fourier-transform mid-infrared (FT-MIR) spectroscopy and many potential applications for the use of the resulting spectra as a phenotyping tool. Ongoing research includes studies of individual milk proteins and fatty acids (Bonfatti et al., 2017d; Lopez-Villalobos et al., 2014; McDermott et al., 2016) and technological properties (Cecchinato et al., 2015; Toffanin et al., 2015; Visentin et al., 2015). Studies have predicted indirect traits related to pregnancy (Lainé et al., 2017; Toledo-Alvarado et al., 2018a, 2018b), energy status (Grelet et al., 2016; McParland et al., 2015; Mehtiö et al., 2018), efficiency (McParland and Berry, 2016; Shetty et al., 2017) and methane emissions (Bittante and Cipolat-Gotet, 2018; Vanlierde et al., 2013, 2015).

Although there are many valuable applications for FT-MIR spectra, the ability to predict traits directly from the spectra and to transfer calibration equations between instruments is hindered by a number of sources of noise. These sources include noise across bands of the infrared spectrum due to the water content of milk, and noise resulting from spectral outliers caused by sample or instrument anomalies. A third source of noise is the variation between instruments or within instruments across time. This variation, that exists even between instruments of the same brand can result in prediction errors and bias, and is particularly problematic when applying calibration models developed on one instrument across a historical database of spectra collected on different instruments (Bonfatti et al., 2017d; Grelet et al., 2015).

A widely-used practice to address the issue of variation between instruments and shifts in instruments across time, is to adjust trait predictions by instrument correction coefficients, previously evaluated from the analysis of reference samples, using the approach outlined by Lynch et al. (2006). However, that method is only applicable where reference samples are available for a trait. With the growing number of traits predicted from spectra, many of which do not have reference samples, there is increased interest in standardizing individual FT-MIR spectra wavenumbers directly (Bonfatti et al., 2017a; Grelet et al., 2015, 2017). Recent publications by Grelet et al. (2015) and Bonfatti et al. (2017a) present methods for standardizing individual wavenumber absorbance values. Both strategies involve assigning a primary instrument and standardizing spectra from other (secondary) instruments to align them to the spectral response of the primary instrument. Grelet et al. (2015) present a piecewise direct standardization (PDS) approach based on the method described by Wang et al. (1991). Bonfatti et al. (2017a) present a retroactive approach (RPS) where percentiles of spectral responses for each wavenumber are used to map the primary/secondary instrument relationships.

The effectiveness of standardization for reducing prediction errors when transferring calibration models between instruments for fat composition traits has been demonstrated previously (Bonfatti et al., 2017a; Grelet et al., 2015). Grelet et al. (2017) also demonstrated the effectiveness of standardization for reducing prediction errors for traits with lower quality calibration models such as methane emissions and cheese yield. Whilst those studies demonstrate the clear benefits of standardization, to date there have been no studies that directly compare the reduction in prediction errors for the two methods when measured across the same dataset.

The purpose of this study was to develop strategies for reducing the impact of noise on predictions and to compare methods for standardizing FT-MIR spectra from milk samples collected across multiple-instrument networks. Our aims included identifying bands with high noise levels across the mid-infrared spectrum, developing an outlier removal methodology, and quantifying the effect of standardization on milk trait predictions. Standardization methods were compared across the same set of milk samples using industry-standard trait predictions for concentrations of major milk components, and different applications of each method were studied to determine the optimal way to implement each method across time.

3.4 Materials and methods

3.4.1 Ethics statement

All data were generated as part of routine commercial activities and were outside the scope of requiring formal ethics approval.

3.4.2 Instrumentation

Fourier-transform mid-infrared spectra from six Bentley FTS (Chaska, MN, USA) instruments, located at two different centres in New Zealand, within the Livestock Improvement Corporation (LIC) milk testing network were used in this study. Three instruments (A4-A6) were located in Hamilton (North Island) and three instruments (A1-A3) were located in Christchurch (South Island). Each spectra record had absorbance values reported for 899 wavenumbers across the range from 649.03 to 3,998.59 cm^{-1}. This range is referred to broadly as the mid-infrared region throughout this study. However, subdivisions of the infrared region vary across different sources, sometimes defining the range 649.03 to 3,998.59 cm^{-1} as including part of the long-wavelength infrared and short-wavelength infrared regions (Bittante and Cecchinato, 2013).

3.4.3 Milk-based reference samples

Milk-based reference sample sets were used to calibrate instruments weekly to meet the International Committee for Animal Recording (ICAR) requirements, in accordance with relevant standards (ISO 9622:2013). Reference sample calibration sets were prepared by MilkTestNZ (Hamilton, NZ) to comply with ICAR guidelines (ICAR, 2017). Those sets included up to 11 milk-based samples with known concentrations of fat, protein and lactose, as determined by industry-accepted chemical reference methods, in accordance with relevant standards (ISO 1211:2010; ISO 8968-4:2016; ISO 22662:2007). A separate set of samples, designed to be identical in composition was generated for each instrument, with samples in each set reflecting milk component concentrations ranging from ~0.1 to 6.1 g/100mL for fat, ~3.5 to 4.5 g/100mL for protein and ~4.7 to 5 g/100mL for lactose.

In total, 918 milk-based samples from reference sets across 16 weeks from February to May 2018 were included in this study. In each week, there were six sets of samples, designed to be identical, with each set analysed across a different instrument. On average, spectra from only 9.6 of the 11 samples were available for each week. This was because some reference sets included only 10 samples (instead of 11), and because some spectra records were discarded if samples were processed out of sequence during the calibration process.

3.4.4 Noise region identification using reference samples

Noise regions across the mid-infrared spectrum range were identified using spectra generated from the weekly instrument calibration process. During instrument calibration, each reference sample was analysed in duplicate to obtain two spectra records (paired-spectra), which were averaged. For each set of paired-spectra (n=918), the difference in absorbance was calculated for each wavenumber. Paired-spectra represented the same reference sample analysed in duplicate on the same instrument, so the expectation was that the absorbance difference would be zero, under the assumption that there was no other interference in the absorbance signal. Metrics to describe the distribution of absorbance differences for each wavenumber were calculated as follows: the mean of the absolute differences for a wavenumber; the standard deviation of the differences for a wavenumber; and the Wasserstein distance metric, to compare the distribution of differences for a wavenumber to the distribution of differences for the wavenumber with the lowest variance. Wasserstein distance metrics were calculated using the transport package in R (Schuhmacher et al., 2017).

Paired-spectra difference metrics were multiplied by 100 and the distributions of each of the scaled absolute mean, the scaled standard deviation and the scaled Wasserstein distance were approximated with a Cauchy distribution. Location and scale parameters for each fitted Cauchy distribution were estimated using the fitdistrplus package in R (Delignette-Muller and Dutang, 2015), where the location parameter defined the location of the distribution peak, and the scale parameter defined the spread of the distribution. Critical-value thresholds based on these parameters were evaluated for each scaled metric for each of the α-levels 0.05, 0.1 and 0.15. A wavenumber was assigned to belong to a noise region if the scaled metric was above the corresponding critical-value threshold for a specified α-level. Levels of α indicated the probability of falsely assigning a wavenumber to a noise region, with higher α-levels resulting in an increased likelihood of assigning wavenumbers to noise regions.

3.4.5 Identification of a primary instrument

Averaged spectra from milk-based reference samples (n=918) were used to identify a high-performing primary instrument to which the other secondary instruments would be calibrated. For each sample, uncorrected predictions of milk component concentrations were generated by applying industry-accepted calibration equations to the average of FT-MIR wavenumber absorbance values. The intercept, slope, R^2, root mean square error (RMSE) and relative RMSE between uncorrected predictions of milk component concentrations and concentrations determined by chemical reference methods were calculated. Relative RMSE values were calculated as the

ratio of the RMSE to the overall average of the reference values for each milk component. The instrument with the highest average R^2 across milk component concentrations for fat, protein and lactose was designated as the primary instrument (A6). The primary instrument also demonstrated consistently lower relative RMSE values and lower deviations from unity for the slope across all three milk components, compared to all other instruments.

3.4.6 Milk test samples from routine milk testing

In New Zealand, milk testing is currently carried out on ~3 million cows per year, located in ~7,500 herds (LIC and DairyNZ, 2017). Most dairy herds in New Zealand operate as pasture-based, seasonal production systems, with milk testing conducted on a bi-monthly basis so that each cow has 3 to 4 tests per lactation. LIC is one of two milk testing providers in New Zealand and has both FOSS (Hillerød, Denmark) and Bentley instruments in their milk testing network. Samples from the North Island were processed at the Hamilton centre and samples from the South Island were processed at the Christchurch centre. Samples were randomly allocated to instruments at each centre, with approximately half being analysed on Bentley instruments.

Fourier-transform mid-infrared spectra records from 2,109,750 individual milk test samples for 1,533,669 cows across 5,574 herds were included in the dataset. Samples were collected and analysed on Bentley instruments as part of routine milk testing conducted by LIC, over the period from September 2017 to May 2018. Median calving dates were 8th August 2017 for cows with samples in the North Island and 20th August 2017 for cows with samples in the South Island. The median parity of cows was 3 with a range of 1 to 15. Cows were from a mixed-breed population. The breed composition of cows sampled comprised 516,893 Holstein-Friesian, 159,249 Jersey, 762,210 Holstein-Friesian x Jersey and 95,317 other breeds.

Outlier removal for milk test samples

The squared Mahalanobis distance (MD) between industry-standard predictions of milk component concentrations (fat, protein and lactose) were evaluated for each milk test record. Outliers were identified and removed if the MD of milk component predictions had a p-value <0.001 based on a χ^2 distribution with 3 degrees of freedom.

The MD between each spectrum and the average spectra were evaluated after excluding noise regions. Under the assumption that the spectra were distributed as a multivariate normal distribution, the MD values for the spectra were expected to follow a χ^2 distribution with r degrees of freedom, where r is the number of wavenumbers after excluding noise regions. Instrument-specific clustering was present in the MD values, necessitating the calculation of within-instrument MD values for the purpose of outlier removal.

The distribution of best-fit was determined for each set of within-instrument MD values using the fitdistrplus package in R (Delignette-Muller and Dutang, 2015). The distributions considered were the normal, gamma, χ^2, lognormal and logistic distribution. The logistic distribution was identified as the best-fit to within-instrument MD values, based on having the lowest average information criterion (AIC), on average, across instruments. Outliers were identified and removed if the within-instrument MD had a p-value <0.001 based on the logistic distribution of best-fit.

3.4.7 Evaluation of standardization coefficients

Piecewise direct standardization

Coefficient sets to relate primary-instrument spectra to spectra from secondary instruments were generated using a PDS approach (Grelet et al., 2015). Briefly, milk-based reference samples measured across all instruments were used to relate the response for each wavenumber j on the primary instrument to a small window around the same wavenumber on each secondary instrument. Each secondary instrument window included five responses, centred on the wavenumber j. A principal components regression was used to map the relationship between the primary-instrument spectral wavenumber and each corresponding secondary-instrument spectral window:

$$\mathbf{p}_j = \boldsymbol{\beta}_{0j} + \mathbf{S}_j\boldsymbol{\beta}_j \tag{3.1}$$

where $\mathbf{p}_j$ is a vector of average absorbance values from paired-spectra for up to 153 samples, for the jth wavenumber on the primary instrument, $\mathbf{S}_j$=[$s_{(j-2)}$,$s_{(j-1)}$,s_j,$s_{(j+1)}$,$s_{(j+2)}$] is a matrix of the corresponding window on the secondary instrument, $\boldsymbol{\beta}_{0j}$ is an offset term and $\boldsymbol{\beta}_j$ is a vector representing transformation coefficients. These defined a complete standardization coefficient set comprising PDS estimates of $\boldsymbol{\beta}_{0j}$ and $\boldsymbol{\beta}_j$ for wavenumbers j=3 to j=897 (895 wavenumbers), with coefficient sets for j=1, j=2, j=898 and j=899 being undefined.

Five time-based criteria were used to restrict the samples included for evaluating coefficient sets. An overall coefficient set was evaluated based on all samples (PDS:Overall). For each of k=1 to 16 weeks, coefficient sets were evaluated: using samples in week k only (PDS:Weekly), using samples from all other weeks, except week k (PDS:AllOtherWks); using samples from all weeks in the same calendar month as week k, but excluding week k (PDS:Monthly); and using the last w weeks of samples prior to week k, where w=1 to 8 (PDS:RollingWks). These coefficient sets allowed different applications of the PDS method across time, and defined different values for $\mathbf{p}_j$ and $\mathbf{S}_j$ in equation 3.1).

Retroactive percentile standardization

Coefficient sets to relate primary-instrument spectra to spectra from secondary instruments were assessed using the RPS approach outlined by Bonfatti et al. (2017a). Briefly, standardization coefficients were calculated using linear regression to map the absorbance percentiles for each wavenumber from the primary instrument to the corresponding absorbance percentiles from each secondary instrument.

Three separate RPS coefficient sets were constructed from milk test samples: using spectra from all milk test samples to evaluate an overall coefficient set (RPS:Overall); using spectra from milk test samples in each month to evaluate monthly coefficient sets (RPS:Monthly); and using spectra from milk test samples in each month, with a different primary instrument used for spectra from South Island samples, to evaluate coefficient sets for each island in each month (RPS:Monthly^{Is}). To evaluate the RPS:Monthly^{Is} coefficient sets, an alternative South Island instrument (A1) was designated as the primary instrument for standardizing South Island spectra. This instrument was selected from South Island instruments following the methodology described for designating an overall primary instrument.

3.4.8 Assessment of standardization strategies

Assessment of standardization strategies was undertaken in two stages: i) assessment of PDS on spectra from milk-based reference samples; and ii) assessment of PDS and RPS on spectra from milk test samples.

Assessment of PDS on milk-based reference samples

The process for assessment of PDS strategies is shown in Fig. 3.1. From a total of 918 milk-based reference samples, 153 were analysed in duplicate on the primary instrument to obtain primary instrument spectra. Corresponding samples, designed to be identical, were analysed in duplicate on the five secondary instruments (n=765) to obtain spectra for each secondary instrument. Calibration models for concentrations of fat, protein and lactose were developed by regressing the average absorbance values from primary instrument paired-spectra against component concentrations, previously determined by chemical reference methods. In these models, wavenumbers from noise regions were excluded and each partial least squares regression was conducted using the plsr package in R (Mevik and Wehrens, 2007). For each model, the number of components to minimise the RMSE of prediction was identified and subsequently employed in model applications.

Secondary-instrument reference sample spectra were standardized using PDS:Weekly, PDS:AllOtherWks, PDS:Monthly and PDS:RollingWks coefficient sets. Calibration models developed from primary instrument spectra were then applied to primary-instrument spectra, and unstandardized and standardized secondary-instrument spectra. This resulted in a set of predicted traits from each of the primary-instrument spectra, the unstandardized secondary-instrument spectra, and each of the standardized secondary-instrument spectral datasets.

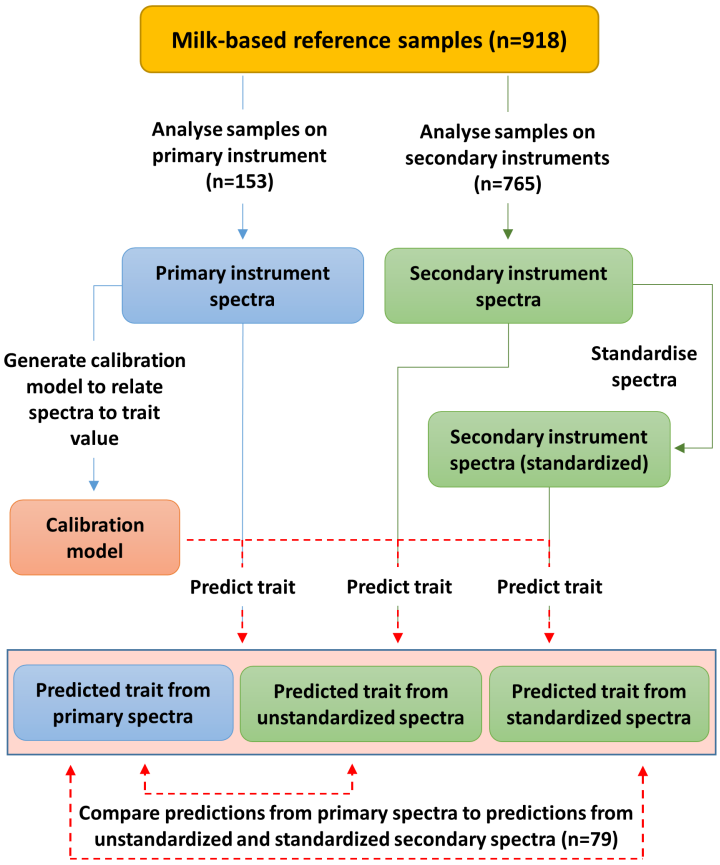

Figure 3.1: Summary of process for assessment of piecewise direct standardization (PDS) strategies on 79 sets of identical milk-based reference samples.

Primary-instrument trait predictions were regressed on predictions from unstandardized and standardized secondary-instrument spectra, and the R^2, intercept, slope, RMSE and relative RMSE were evaluated for each strategy. Relative RMSE values were calculated as the ratio of the RMSE to the overall average of the reference values for each milk component. Because the PDS:RollingWks strategies were dependent on having up to 8 weeks of spectra from previous weeks available, coefficient sets for w=8 were only estimable for weeks k=9 to 16. Therefore, to ensure that standardization strategies were compared across the same period, comparisons between primary and secondary-instrument spectra were restricted to weeks k=9 to 16 across all strategies (n=79).

Assessment of PDS and RPS on milk test samples

The process for assessment of PDS and RPS strategies is shown in Fig. 3.2. Milk test samples (n=2,044,094) were analysed on one of six instruments, and sample spectra were standardized using PDS:Overall, RPS:Overall, RPS:Monthly and RPS:Monthly^{Is} coefficient sets. Industry-standard calibration equations were used to predict individual milk component concentrations (fat, protein and lactose) from unstandardized and standardized spectra. This resulted in a set of predicted traits from unstandardized spectra and a set of predicted traits from each of the standardized sets of spectra. Traits predicted from unstandardized spectra were also adjusted by instrument-specific calibration coefficients (previously evaluated from the weekly calibration process), to obtain the industry-standard prediction for each trait, in accordance with ICAR requirements and relevant milk testing standards (ISO 9622:2013). Industry-standard trait predictions were regressed on predictions from unstandardized and standardized spectra and the overall R^2, intercept, slope, RMSE and relative RMSE were evaluated for each strategy. Relative RMSE values were calculated as the ratio of the RMSE to the overall average of the industry-standard prediction for each milk component. Relative RMSE values for each individual milk component were compared overall and by individual instrument and month.

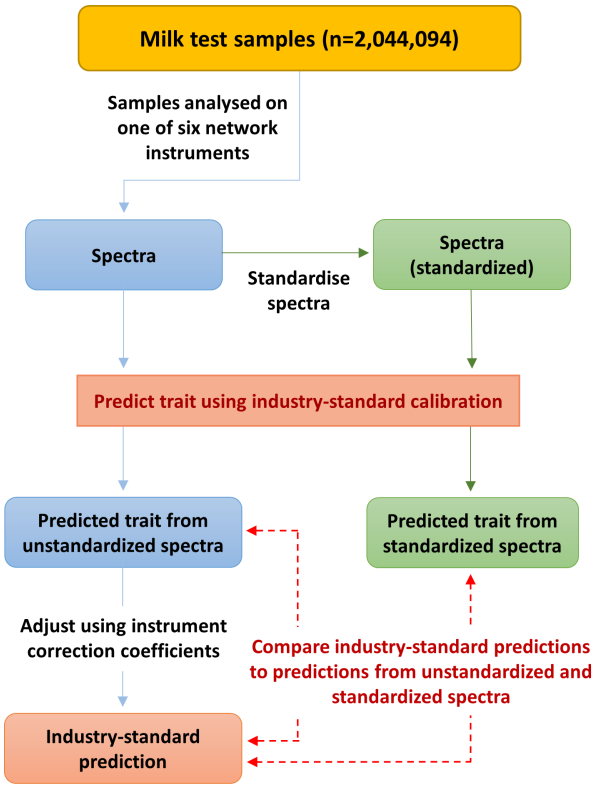

Figure 3.2: Summary of process for assessment of piecewise direct standardization (PDS) and retroactive percentile standardization (RPS) strategies on 2,044,094 milk test samples.

3.5 Results and discussion

3.5.1 Noise region identification using reference samples

Profiles of absorbance differences for paired-spectra records across the mid-infrared spectrum are presented in Fig. 3.3. Six separate regions are identified in Fig. 3.3(a) based on noise levels observed across the spectrum, with close up views of regions (i), (iii) and (v) shown in Figs. 3.3(b), 3.3(c) and 3.3(d), respectively. Representative distributions of absorbance differences for individual wavenumbers from each region are presented in Fig. 3.4. Regions of the spectrum below ~950 cm^{-1} and from ~3,000 to 3,700 cm^{-1} had the largest noise levels. Notably, noise observed in the region ~1,600 to 1,700 cm^{-1} was much lower than that in regions (i) or (v).

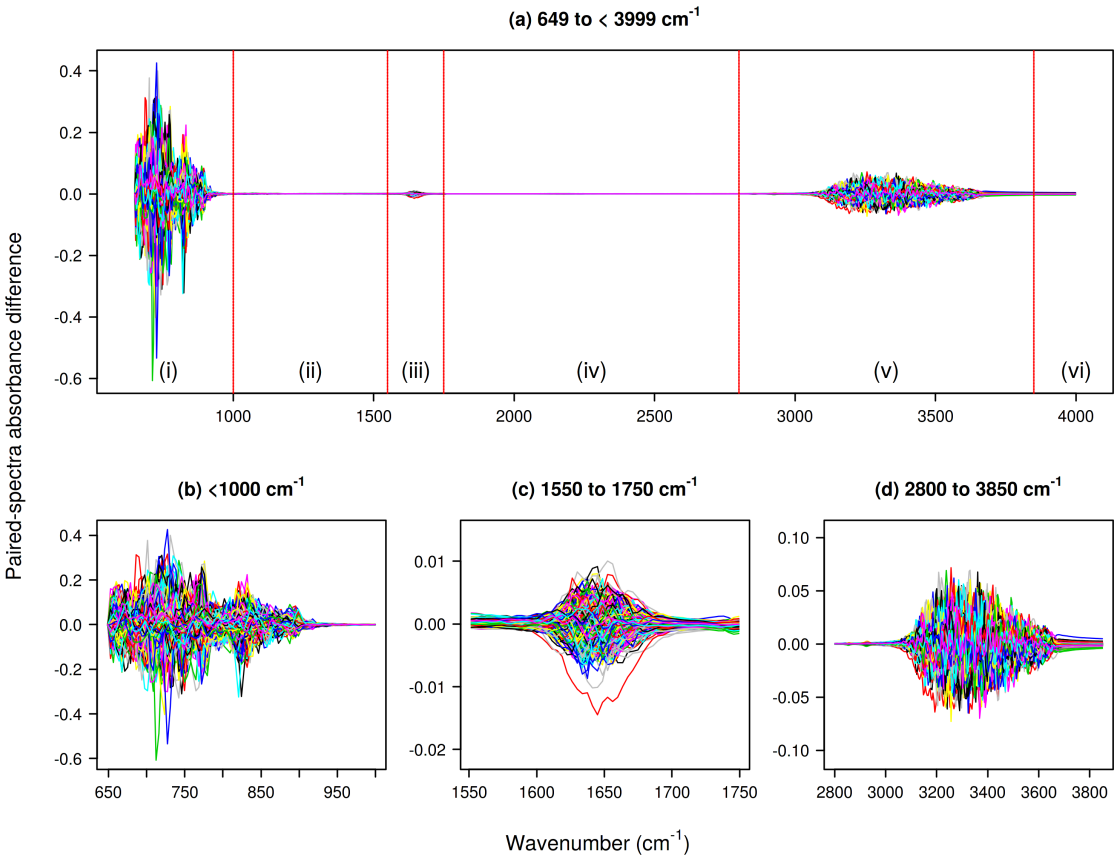

Figure 3.3: Absorbance differences for paired FT-MIR milk spectra wavenumbers where absorbance = $Log(1/T)$ and T=transmittance. Based on paired-spectra (n=918) from reference samples analysed across six Bentley instruments.

Table 3.1: Noise regions for FT-MIR milk spectra as defined by different paired-spectra difference metrics and varying α levels[1]

Scaled paired-spectra difference metric	Noise region number	Noise region wavenumber ranges (cm^{-1})		
		α=0.05	α=0.1	α=0.15
Mean	1	653-940	649-959	649-970
	2	1,626-1,664	1,615-1,675	1,608-1,686
	3	3,074-3,667	3,040-3,737	3,018-3,879
Standard deviation	1	653-944	649-962	649-974
	2	1,623-1,667	1,611-1,679	1,608-1,686
	3	3,070-3,674	3,036-3,801	3,006-3,999
Wasserstein distance	1	653-940	649-955	649-970
	2	1,626-1,664	1,615-1,675	1,608-1,682
	3	3,077-3,663	3,044-3,726	3,021-3,849

[1] The α levels indicate the probability of falsely assigning a wavenumber to a noise region.

Table 3.1 presents noise regions defined by scaled paired-spectra difference metrics. Wavenumbers were assigned to noise regions if the scaled difference metric was above the critical-value threshold from the appropriate Cauchy distribution. Location and scale parameters for Cauchy distributions were 0.0143 and 0.0157 for absolute means; 0.0170 and 0.0187 for standard deviations; and 0.00989 and 0.0163 for Wasserstein distances. For the first and second noise regions, for any given α-level, noise region boundaries defined for each metric differed by up to only seven wavenumbers. The third noise region was the most variable between α-levels and between metrics; the upper limit varying between the standard deviation and Wasserstein distance metrics by 75 wavenumbers for α=0.1 and by 150 wavenumbers for α=0.15. Similar noise regions have been presented in other studies: 1,616 to 1,678 cm^{-1}, 3,066 to 3,668 cm^{-1} (Soyeurt et al., 2010); 1,586 to 1,698 cm^{-1}, 3,052 to 3,669 cm^{-1} (Bittante and Cecchinato, 2013); 1,600 to 1,689 cm^{-1}, 3,008 to 5,010 cm^{-1} (Grelet et al., 2015). Wavenumbers lower than 925 cm^{-1} were not reported in those studies, because they used FOSS instruments (Hillerød, Denmark) that do not report any part of the spectra from this region.

In the present study, most of the region from 649.03 to 925 cm^{-1} had high noise levels (Fig. 3.3). However, the first wavenumber, 649.03 cm^{-1} was an exception and had comparatively low noise levels with a scaled absolute difference mean of 8.92e-02 and scaled difference standard deviation of 1.28e-01. Low noise levels were also observed in the distribution of paired-spectra absorbance differences for 649.03 cm^{-1} (not shown), which was narrower than that for the wavenumber 1,648.7 cm^{-1} (Fig. 3.4(iii)), but wider than that for 3,924 cm^{-1} (Fig. 3.4(vi)). Notably, for α=0.05, the wavenumber 649.03 cm^{-1} was not classified as part of the first noise region for any of the difference metrics (Table 3.1).

The second noise region defined in this study was in the water absorption band of the spectrum affected by O-H bending (~1,600 to 1,700 cm^{-1}), and the third noise region was in the band affected by O-H stretching (>~3,000 cm^{-1}). Although wavenumbers in the O-H bending and O-H stretching bands of the spectrum are often attributed as noise and removed, there is evidence that these regions contain valuable information. Wang et al. (2016) and Wang and Bovenhuis (2018) found wavenumbers in the regions between 1,619 to 1,674 cm^{-1} and 3,073 to 3,667 cm^{-1} that were affected by a polymorphism in the *DGAT1* gene that has major impacts on milk composition (Grisart et al., 2002). Similarly, Toledo-Alvarado et al. (2018a) reported a significant association between cows' pregnancy status and the 3,683 cm^{-1} wavenumber. Bittante and Cecchinato (2013) also showed that the transmittance for most FT-MIR wavenumbers in the range from 930 to 5,000 cm^{-1} was heritable. They concluded that, although heritabilities were often low in the water absorption regions from 1,586 to 1,698 cm^{-1} and 3,052 to 3,669 cm^{-1}, these regions should still be considered for investigation, because they included absorbance peaks for chemical bonds related to non-water milk components.

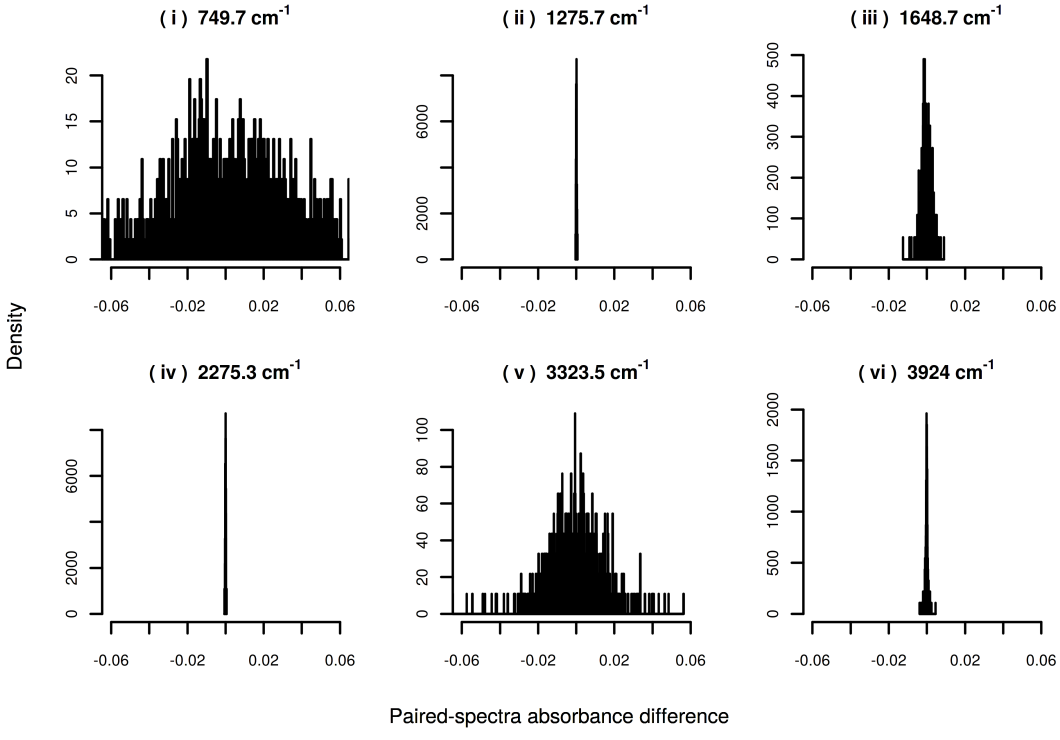

Figure 3.4: Representative distributions of absorbance differences for paired FT-MIR milk spectra at specified wavenumbers where absorbance = $Log(1/T)$ and T=transmittance. Based on paired-spectra (n=918) from reference samples analysed across six Bentley FTS instruments.

The impact of including wavenumbers from noise regions in a study will depend on the specific application. In applications where wavenumbers are considered independently, such as in a single wavenumber genome-wide association study, it is prudent to retain spectra from all wavenumbers in the analysis. However, in applications where wavenumbers are considered in a multivariate manner, such as in the evaluation of principal components or partial least squares regression, the exclusion of noise regions is an important step. For all subsequent applications in the present study, noise regions have been defined according to the Wasserstein distance metric with α=0.15 (649 to 970 cm^{-1}, 1,608 to 1,682 cm^{-1} and 3,021 to 3,849 cm^{-1}). This definition provided boundaries similar to those previously reported (Bittante and Cecchinato, 2013; Grelet et al., 2015; Soyeurt et al., 2010). The resulting spectra with noise regions removed included only 526 of the 899 original wavenumbers.

3.5.2 Outlier removal for milk test samples

From 2,109,750 milk test records, 2,081,455 remained after outlier removal based on the MD for milk component concentrations. The distribution of MD values between each spectra record and the average spectra is presented in Fig. 3.5. Also shown is the curve of the expected χ^2 distribution with 526 degrees of freedom and the corresponding critical-value threshold based on a p-value of 0.001. The distribution of MD values for the spectra was not consistent with the expected χ^2 distribution, due to instrument-specific clustering. Therefore, outlier removal was conducted on within-instrument MD values.

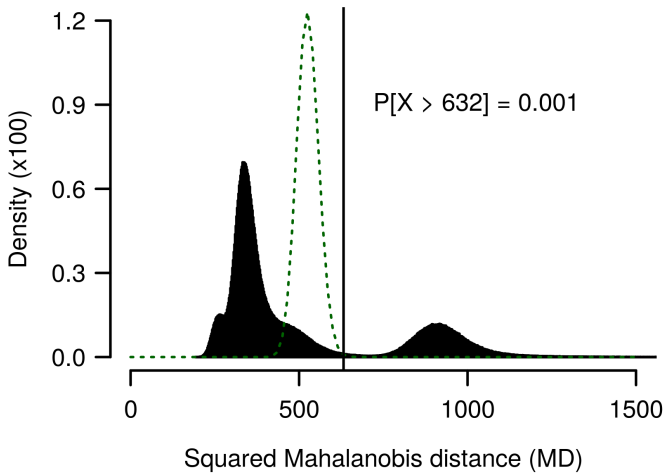

Figure 3.5: Squared Mahalanobis distance distribution (MD) across herd test records (n=2,081,445). The curve of the expected χ^2 distribution with 526 degrees of freedom is also shown with the corresponding critical-value threshold associated with a p-value of 0.001.

Within-instrument MD values were calculated and the logistic distribution of best-fit was determined. Within-instrument MD values are presented in Fig. 3.6. Curves of the best-fit logistic distributions are also shown with outlier thresholds corresponding to a p-value of 0.001. Within-instrument outlier thresholds ranged from 572 to 772. Using these thresholds, 1.79% of records were identified as outliers and removed, leaving 2,044,094 records for analysis.

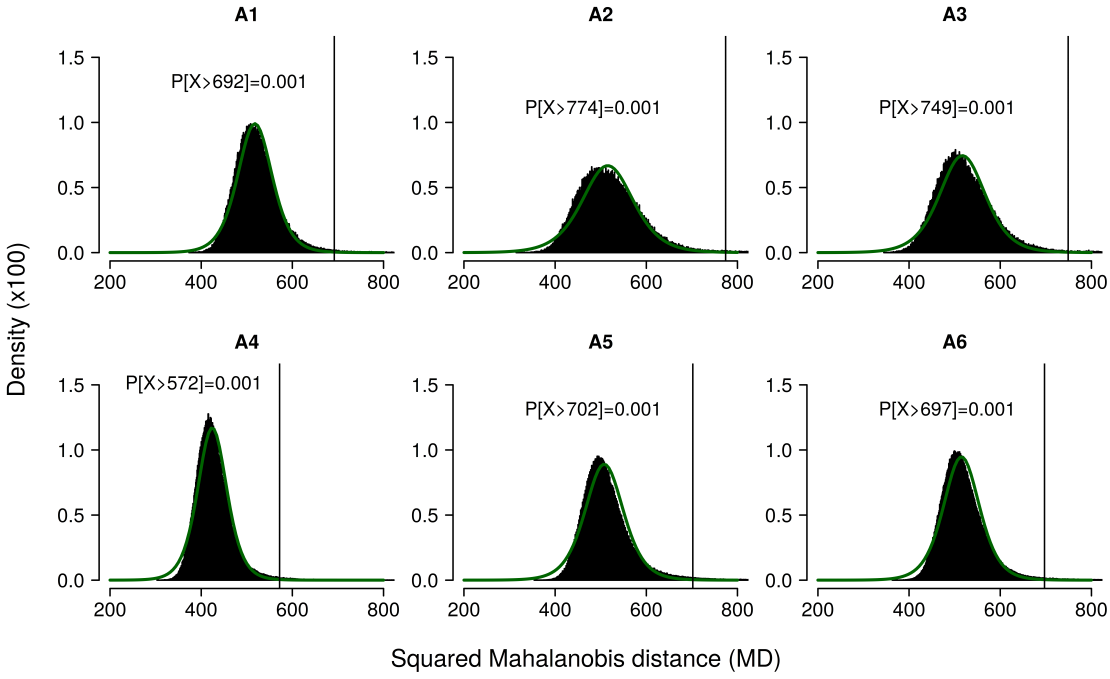

Figure 3.6: Within-instrument squared Mahalanobis distance (MD) distributions and corresponding critical-value thresholds associated with a p-value of 0.001, based on approximated Logistic distributions with location and scale parameters respectively: A1:518.4,29.2; A2:516.3,42.1; A3:517.8,37.0; A4:420.8,20.6; A5:509.1,37.8; and A6:515.8,31.2. (A1: n=191,655; A2: n=200,612; A3: n=157,691; A4: n=430,940; A5: n=461,130; A6: n=639,427).

In studies of FT-MIR spectra, outlier removal using MD values is a common approach. However, many studies use spectra from only a single instrument and do not have the complexity of differing variance structures between instruments. Results in this study demonstrate the importance of accounting for instrument-specific variance structures when applying multivariate outlier identification methods. Failure to do so and applying a threshold based on a χ^2 distribution with 526 degrees of freedom would have resulted in removing a large proportion of records from one instrument and not removing anomalies from others.

3.5.3 Assessment of PDS on milk-based reference samples

Root mean square errors between primary and secondary-instrument predictions from unstandardized and standardized reference sample spectra are presented in Table 3.2. Each of the PDS strategies resulted in lower RMSE values across milk component concentrations, compared to the RMSE values from unstandardized spectra. The PDS:Weekly strategy resulted in the lowest RMSE values with a reduction in RMSE from 0.222 to 0.022 g/100mL for fat, 0.265 to 0.020 g/100mL for protein and 0.299 to 0.010 g/100mL for lactose. A recent study reported a comparable RMSE between primary and secondary-instrument predictions after standardization of 0.016 g/100mL for fat (Grelet et al., 2015). That study used a similar approach to the PDS:Weekly strategy, in that the spectra for each week was also included in the records used to evaluate standardization coefficients. This approach is likely to underestimate prediction errors when coefficients are applied to different spectral datasets. The PDS:AllOtherWks and PDS:Monthly strategies were structured to ensure that for any given week, the spectra for the week being assessed was independent of the spectra used to evaluate the coefficients being applied to that week. Of these two strategies, the PDS:AllOtherWks strategy resulted in the lowest RMSE values, namely 0.059 g/100mL for fat, 0.051 g/100mL for protein and 0.053 g/100mL for lactose (Table 3.2). These RMSE values equate to a reduction by 73% for fat, 81% for protein and 82% for lactose. These reductions in RMSE were similar to those presented by Grelet et al. (2017) for methane emissions (83%), polyunsaturated fatty acids (86%) and cheese yield (81%).

Table 3.2: Root mean squared errors (RMSE) between primary and secondary-instrument trait predictions from unstandardized and standardized spectra (n=79)[1]

Strategy	**Trait**		
	Fat (g/100mL)	**Protein** (g/100mL)	**Lactose** (g/100mL)
Standardized[2]			
PDS:Weekly	0.022	0.020	0.010
PDS:AllOtherWks	0.059	0.051	0.053
PDS:Monthly	0.080	0.066	0.054
Unstandardized	0.222	0.265	0.299

[1] Standardization conducted using implementations of the piecewise direct standardization (PDS) method. Results shown for weeks k=9 to 16: 8 week validation period from April to May 2018.

[2] For each week k, PDS coefficients evaluated and applied. PDS:Overall: Standardized using PDS coefficients evaluated from all reference samples; PDS:Weekly: Standardized using PDS coefficient sets from reference samples from week k only; PDS:AllOtherWks: Standardized using PDS coefficients evaluated from reference samples from all other weeks, except week k; and PDS:Monthly: Standardized using PDS coefficients evaluated from reference samples from all weeks in the same calendar month as week k, but excluding week k.

Relationships between primary and secondary-instrument predictions from unstandardized and standardized spectra based on PDS:AllOtherWks and PDS:Monthly coefficient sets are presented in Fig. 3.7. After standardization, bias and deviation from unity for slopes consistently decreased and R^2 values consistently increased. Slope deviations from 1 after standardization were < 0.02 for fat and protein concentrations. For lactose, the deviation from 1 prior to standardization was high at 0.75, but was reduced to < 0.17 after standardization. The highest bias levels between primary and secondary-instrument predictions were observed for lactose concentrations. Prior to standardization, the bias was 3.56 g/100mL, but this was reduced to 0.41 g/100mL using the PDS:AllOtherWks strategy.

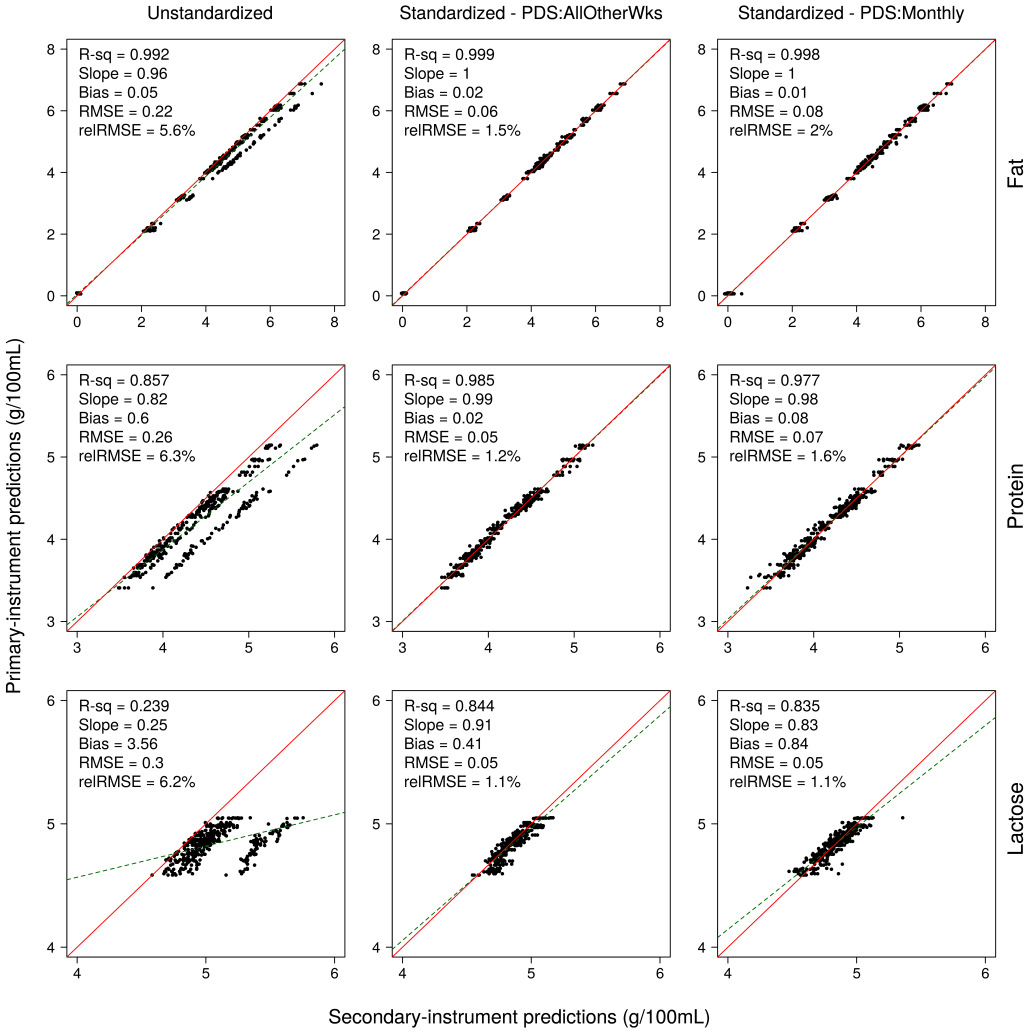

Figure 3.7: Comparison between primary and secondary-instrument predictions based on unstandardized and standardized spectra from implementations of the piecewise direct standardization (PDS) method. Results shown for weeks k=9 to 16: 8 week validation period from April to May 2018 (n=79). For each week k, PDS coefficients evaluated and applied. PDS:AllOtherWks: Standardized using PDS coefficients evaluated from reference samples from all other weeks, except week k; and PDS:Monthly: Standardized using PDS coefficients evaluated from reference samples from all weeks in the same calendar month as week k, but excluding week k. Dotted lines represent the regression line between primary and secondary instruments; continuous line represents y=x.

Figure 3.8 presents relative RMSE values between primary and secondary-instrument predictions from standardized spectra using PDS:RollingWks strategies where w=1 to 8, represents the number of previous weeks of spectra included in standardization coefficient set evaluation. For fat and protein, relative RMSE values between primary and secondary-instrument predictions decreased as additional weeks of historical data were used to evaluate coefficients, but incremental benefits diminished as more weeks were added. For lactose, the relative RMSE values were consistently low at ~1% regardless of the number of previous weeks of data used to evaluate coefficient sets.

Including spectra from a wider date range in coefficient set evaluations resulted in lower prediction errors. This was evident from the lower RMSE values for the PDS:AllOtherWks strategy (Table 3.2), and was confirmed by the trends in relative RMSE for the PDS:RollingWks strategies (Fig. 3.8). However, the present study included reference sample spectra from a 16 week period only. Over a longer time period, we could expect further instrument drift and shifts in the spectra due to instrument maintenance, parts deterioration/replacement and factors such as temperature fluctuations and wavelength or detector intensity instability (Wang et al., 1991). Major shifts in the spectra due to these factors would affect relationships between primary and secondary-instrument spectra and potentially erode gains in prediction accuracy. Monitoring and adjusting for drift and spectra shifts is thus important to ensure that standardization using the PDS approach consistently reduces prediction errors across time.

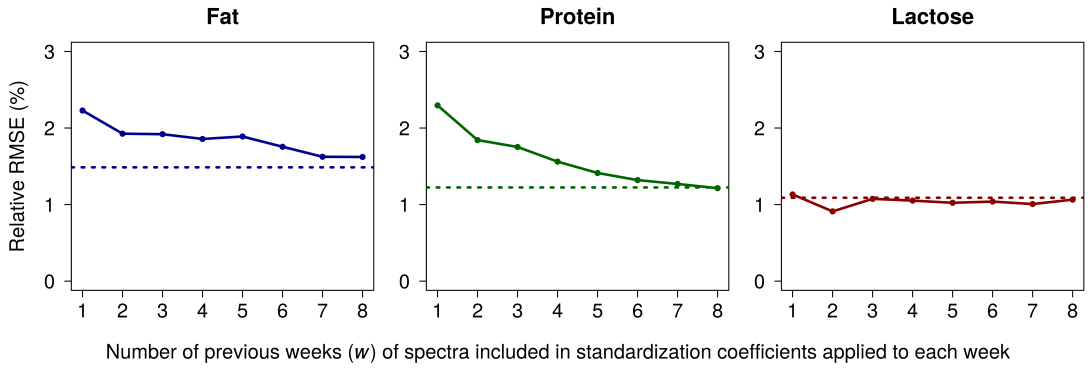

Figure 3.8: Relative root mean squared errors (RMSE) between primary and secondary-instrument predictions for spectra standardized using piecewise direct standardization (PDS). Results shown for weeks k=9 to 16: 8 week validation period from April to May 2018 (n=79). The x-axis represents the number of previous weeks of reference sample spectra included in the evaluation of standardization coefficients applied to each week. Dotted lines represent the relative RMSE from the PDS:AllOtherWks strategy: Each week standardized using PDS coefficients evaluated from reference sample spectra from all other weeks.

Table 3.3: Root mean squared errors (RMSE) between industry standard trait predictions and predictions from unstandardized and standardized spectra (n = 2,044,094)

Strategy	Trait		
	Fat (g/100mL)	**Protein** (g/100mL)	**Lactose** (g/100mL)
Standardized[1]			
PDS:Overall	0.071	0.055	0.088
RPS:Overall	0.156	0.085	0.087
RPS:Monthly	0.142	0.081	0.080
RPS:Monthly^{Is}	0.076	0.049	0.061
Unstandardized	0.190	0.129	0.143

[1] PDS:Overall = standardized using piecewise direct standardization (PDS) coefficients evaluated from all reference samples; RPS:Overall = standardized using retroactive percentile standardization (RPS) coefficients evaluated from all herd test samples; RPS:Monthly = standardized using RPS coefficient sets evaluated monthly from herd test samples; and RPS:Monthly^{Is} standardized using RPS coefficient sets evaluated monthly from herd test samples with a different primary instrument used for South Island samples.

3.5.4 Assessment of PDS and RPS on milk test samples

Root mean square errors between industry-standard predictions of milk components and predictions from unstandardized and standardized spectra are presented in Table 3.3. Standardization resulted in lower RMSE values for all strategies and across all examined milk components. The two most effective strategies for reducing prediction errors were PDS:Overall and RPS:Monthly^{Is}. The RPS strategies that standardized secondary instruments to a common North Island primary instrument (RPS:Overall, RPS:Monthly) did not perform as well at reducing prediction errors when compared to the RPS:Monthly^{Is} strategy that standardized spectra from South Island samples to a South Island instrument.

The PDS:Overall strategy resulted in the lowest RMSE for fat with a drop from 0.190 to 0.071 g/100mL, i.e., a reduction of 63%. The RPS:Monthly^{Is} strategy resulted in the lowest RMSE for protein with a drop from 0.129 to 0.049 g/100mL for protein, i.e., a reduction of 62%. For lactose, standardization using the PDS:Overall strategy resulted in a drop in RMSE from 0.143 g/100mL to 0.088 g/100mL, i.e., a reduction of 38%. This reduction in RMSE was lower than for the RPS:Monthly^{Is} strategy which had a drop in RMSE to 0.061 g/100mL, i.e., a 57% reduction. A likely reason for the PDS strategy not performing as well for lactose is that only a small range of lactose values (~4.7 to 5 g/100mL) were represented in the reference samples used to evaluate PDS coefficients. Wider component concentration ranges in reference samples improves trait

calibrations (Kaylegian et al., 2006). This implies that representation of a wider range of lactose values in reference samples would improve the mapping of relationships between primary and secondary-instrument wavenumbers that have absorbance peaks for lactose. This would improve lactose predictions when using the PDS approach, and also improve predictions for other traits that have spectral signal represented across the same wavenumbers.

Variation in individual instrument prediction accuracy

Figure 3.9 presents relative RMSE values between industry-standard milk component predictions and predictions from unstandardized and standardized spectra, summarised by instrument. Two of the South Island instruments (A2, A3) had consistently high relative RMSE values for unstandardized spectra compared to the other instruments. The PDS:Overall strategy had consistently low RMSE values, even for the A2 and A3 instruments. For all milk components, after standardizing using the RPS:Monthly strategy, relative RMSE values were inflated for South Island instruments. When spectra from South Island samples were standardized to a South Island instrument (RPS:Monthly^{Is}), RMSE values were reduced to similar levels as the PDS:Overall strategy for fat, and were lower for protein and lactose.

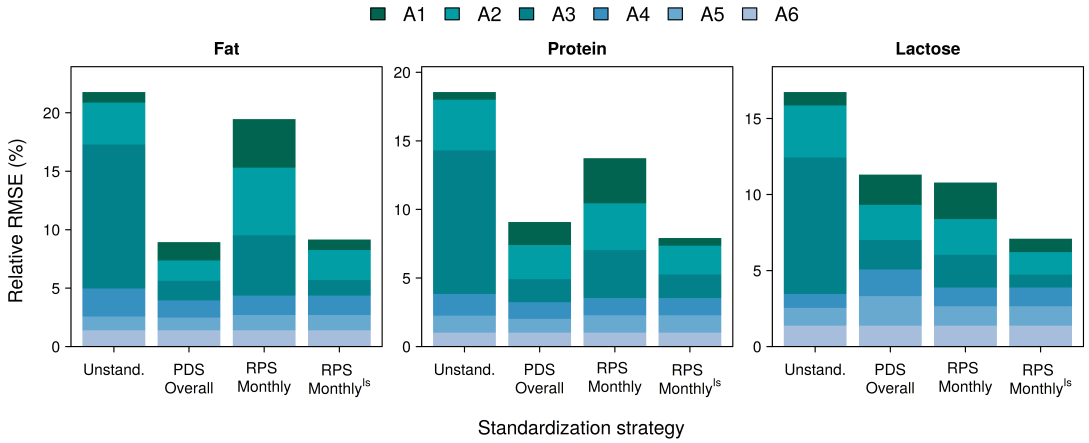

Figure 3.9: Relative root mean squared errors (RMSE) between industry-standard milk component predictions and predictions from unstandardized and standardized spectra, summarised by instrument (n=2,044,094). Standardization conducted using implementations of the piecewise direct standardization (PDS) and retroactive percentile standardization (RPS) methods. Unstand: Unstandardized; PDS Overall: Standardized using PDS coefficients evaluated from all reference samples; RPS Monthly: Standardized using RPS coefficient sets evaluated monthly from herd test samples; and RPS MonthlyIs: Standardized using RPS coefficient sets evaluated monthly from herd test samples with a different primary instrument used for South Island samples.

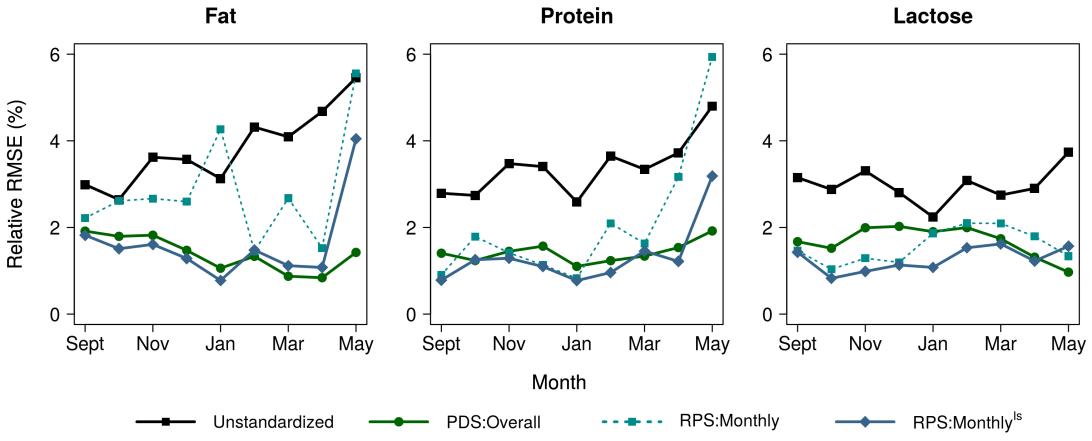

Figure 3.10: Relative root mean squared errors (RMSE) between industry-standard milk component predictions and predictions from unstandardized and standardized spectra, summarised by month (n=2,044,094). PDS:Overall: Standardized using piecewise direct standardization (PDS) coefficients evaluated from all reference samples; RPS:Monthly: Standardized using retroactive percentile standardization (RPS) coefficient sets evaluated monthly from herd test samples; and RPS:MonthlyIs: Standardized using RPS coefficient sets evaluated monthly from herd test samples with a different primary instrument used for South Island samples.

Variation in prediction accuracy over time

Figure 3.10 presents relative RMSE values between industry-standard milk component predictions and predictions from unstandardized and standardized spectra, summarised by month. The PDS:Overall strategy had consistently low relative RMSE values across all months for all milk components. Notably, the PDS:Overall strategy standardizes the full season of herd test spectra using only one set of standardization coefficients, evaluated from reference samples for the 16 week period from February to May 2018.

Consistently low relative RMSE values were also observed for the RPS:Monthly^{Is} strategy, except for a peak for fat and protein in May. On closer examination, these peaks were caused by high relative RMSE values for two of the instruments (A2, A4). The likely cause of this was the drop in overall spectra record numbers included in the estimation of standardization coefficients for May. Between April and May, the number of spectra records dropped by 65% for the North Island, and by 68% for the South Island.

Implementing the RPS strategy with monthly coefficient updates and standardizing to a common North Island primary instrument (RPS:Monthly) was not as effective at reducing prediction errors as the RPS:Monthly^{Is} approach. Using the RPS:Monthly strategy resulted in prediction error peaks for fat in January and May, and protein in April and May. On closer examination, high relative RMSE values for South Island instruments (A1, A2, A3) were underlying these peaks. These were caused by differences in milk component concentrations for

North and South Island samples in these months: average industry-standard predictions for fat in January were 5.10 g/100mL for North Island samples compared to 4.68 g/100mL for South Island samples, i.e., 9% higher; in May, average industry-standard predictions for fat were 5.48 g/100mL for North Island samples compared to 5.91 g/100mL for South Island samples, i.e., 7% lower. Differences between fat concentrations for North and South Island samples were also observed in other months, but were smaller. Similar trends were also observed for protein in April and May.

In general, regional and island differences in milk composition across the season are expected due to genetic factors such as regional breed structure, and other region-specific factors related to feed, management, calving start dates and climate/weather patterns. For this reason, it is unsurprising that differences in milk composition were observed between North and South Island samples, particularly in January and also in April/May. January is at the peak of summer in New Zealand, a time when farmers adopt varying practices to manage feed and maintain body condition score targets. From March onwards, there are also expected to be differences in milk composition as cows are dried off, with drying off on average taking place 2-3 weeks later in the South Island.

The expectation is that milk composition on average across instruments within a centre would be equivalent because milk test samples are randomly allocated to instruments within each milk testing centre. However, overall differences in milk composition for samples processed in the North Island compared to those processed in the South Island are expected. These milk composition differences became problematic when using the RPS:Monthly approach which standardizes spectra from South Island samples to a North Island primary instrument. In doing so, other non-instrument errors and bias were introduced. The risk of correcting for non-instrument factors such as breed and feed when using a retroactive approach was also signalled by Grelet et al. (2017). In the present study, we were able to resolve this by partitioning spectra into subsets and applying standardization within month and with a separate primary instrument assigned for each of the North and South Islands. Partitioning spectra into homogeneous subsets for the purpose of standardization can also be achieved using principal components analysis (PCA) to detect shifts in the PCA scores across time as per Bonfatti et al. (2017a). In their study, Bonfatti et al. (2017a) also demonstrated the importance of standardization within homogeneous sets of spectra and confirmed that small variations in the FT-MIR signal could lead to prediction inaccuracies. They also concluded that the RPS method should be considered as complementary to other classical standardization procedures, and variability in signal across time should be monitored carefully.

3.5.5 Common reference sample sharing between networks

Standardization using the PDS method consistently reduced prediction errors across time, compared to no standardization or standardizing using the RPS method. This was evident in the application of the PDS:AllOtherWks approach applied to reference samples, and also in the PDS:Overall approach applied to milk test samples. Applications of the PDS method are reliant on the analysis of identical reference samples across all instruments in the network. To be able to standardize predictions across multiple networks, including those in other countries requires the sharing of common reference samples. In Europe, reference sample sharing and standardization between instruments in different countries already takes place across the OptiMIR network. This transnational network includes ~65 FT-MIR spectrometers in 25 milk laboratories across five countries, with standardisation data being stored in a common database (European Economic Interest Grouping in the service of dairy farmers. n.d.). Outside Europe, there is little in the way of sharing or analysing common reference samples between countries. Sharing reference samples globally has the potential to enhance collaboration opportunities and maximise the value of FT-MIR spectra. However, the success of this would require the resolution of a number of key issues, such as logistics, sample preservation and integrity, and other biosecurity related risks. Also, it would be ideal if shared reference sample sets were extended to include a broader range of milk component representation as well as a wider range of individual component values. A number of milk components have been confirmed as having absorbance peaks for chemical bonds within specific ranges of the mid-infrared spectrum (De Marchi et al., 2009b; Grelet et al., 2015). If reference samples were extended to capture greater signal diversity across the spectrum, accuracies across a wider range of individual wavenumbers would be improved, and this would result in improved predictions for other new traits.

3.6 Conclusions

In this study, we present strategies for reducing the impact of noise in FT-MIR spectra and compare standardization methods for reducing between-instrument variation. We demonstrate that standardization using a PDS approach gives the most consistent reduction in prediction errors across time. Standardization using the RPS approach can also be highly effective at reducing prediction errors, provided that secondary-instruments are standardized to a primary-instrument with broadly equivalent milk composition. Standardization using PDS is the optimal approach because it is less sensitive to shifts in milk composition and non-instrument errors. However, this method is reliant on having spectra from identical reference samples analysed across all instruments in the network. Where reference sample spectra are unavailable, standardization using the RPS approach can be a suitable alternative. For implementations of either of these standardization methods, instrument drift and other major shifts in the spectra across time should be monitored carefully. Standardization to reduce between-instrument variation will improve the quality of FT-MIR spectra for various downstream applications, including for trait prediction, predicting breeding values and quantifying genetic signals underlying specific FT-MIR spectra wavenumbers.

3.7 Acknowledgements

The authors gratefully acknowledge LIC (Hamilton, New Zealand) herd-testing staff for the processing and analysis of milk samples. Kathryn would also like to thank the wider LIC, R&D team and fellow students for helpful discussions and underlying technical support. The authors also gratefully acknowledge Tod Schilling and Pierre Broutin (Bentley Instruments Inc.) for assistance with accessing FTIR spectra from Bentley instruments. This project is funded by LIC (Hamilton, New Zealand) in association with Massey University (Palmerston North, New Zealand).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Kathryn Maree Tiplady
Name/title of Primary Supervisor:	Professor Dorian Garrick
In which chapter is the manuscript /published work:	Chapter Three

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Tiplady K.M., Sherlock R.G., Littlejohn M.D., Pryce J.E., Davis S.R., Garrick D.J., Spelman R.J. and Harris B.L. Strategies for noise reduction and standardization of milk mid-infrared spectra from dairy cattle. Journal of dairy science. 2019 Jul 1;102(7):6357-72.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Kathryn Tiplady Digitally signed by Kathryn Tiplady Date: 2022.03.23 18:15:50 +13'00'
Date:	
Primary Supervisor's Signature:	Dorian Garrick
Date:	25-Mar-2022

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 4

Pregnancy status predicted using FT-MIR milk spectra from dairy cattle

Originally published as: Tiplady, K.M., Trinh, M-H., Davis, S.R., Sherlock, R.G., Spelman, R.J., Garrick, D.J. and Harris, B.L., 2022. Pregnancy status predicted using milk mid-infrared spectra from dairy cattle. *Journal of dairy science*, 105(4), pp.3615-3632. https://doi.org/10.3168/jds.2021-21516.

4.1 Interpretive summary

Knowledge of pregnancy status is important for effective herd management of dairy cattle. Utilising Fourier-transform mid-infrared data to predict pregnancy is of interest, because alternative methods for determining pregnancy status are costly and/or time-consuming. This study compared pregnancy prediction models based on milk spectra using differing strategies for classifying pregnant and non-pregnant records. We show that in pasture-based seasonal calving herds, confounding between pregnancy status and lactation stage can produce misleading results. For models where the effect of this confounding was reduced, prediction accuracies were not sufficiently high to be used as a sole indicator of pregnancy status for herd management.

4.2 Abstract

Accurate and timely pregnancy diagnosis is an important component of effective herd management in dairy cattle. Predicting pregnancy from Fourier-transform mid-infrared (FT-MIR) spectroscopy data is of particular interest because the data is often already available from routine milk testing. The purpose of this study was to evaluate how well pregnancy status could be predicted in a large dataset of 1,161,436 FT-MIR milk spectra records from 863,982 mixed-breed pasture-based New Zealand dairy cattle managed within seasonal calving systems. Three strategies were assessed for defining the non-pregnant cows when partitioning the records according to pregnancy status in the training population. Two of these used records for cows with a subsequent calving only, whilst the third also included records for cows without a subsequent calving. For each partitioning strategy, partial least squares discriminant analysis (PLS-DA) models were developed, whereby spectra from all the cows in 80% of herds were used to train the models, and predictions on cows in the remaining herds were used for validation. A separate dataset was also used as a secondary validation, whereby pregnancy diagnosis had been assigned according to the presence of pregnancy-associated glycoproteins (PAG) in the milk samples. We examined different ways of accounting for stage of lactation in the prediction models, either by including it as an effect in the prediction model, or by pre-adjusting spectra prior to fitting the model. For a subset of strategies, we also assessed prediction accuracies from deep learning approaches, utilising either the raw spectra or images of spectra. Across all strategies, prediction accuracies were highest for models using the unadjusted spectra as model predictors. Strategies for cows with a subsequent calving performed well in herd-independent validation with sensitivities above 0.79, specificities above 0.91 and area under the receiver operating characteristic curve (AUC) values over 0.91.

However, for these strategies, the specificity to predict non-pregnant cows in the external PAG dataset was poor (0.002 to 0.04). The best performing models were those that included records for cows without a subsequent calving, and used unadjusted spectra and DIM as predictors, with consistent results observed across the training, herd-independent validation and PAG datasets. For the PLS-DA model, sensitivity was 0.71, specificity was 0.54 and AUC values were 0.68 in the PAG dataset; and for an image-based deep learning model, the sensitivity was 0.74, specificity was 0.52 and the AUC value was 0.69. Our results demonstrate that in pasture-based seasonal calving herds, confounding between pregnancy status and spectral changes associated with stage of lactation can inflate prediction accuracies. When the effect of this confounding was reduced, prediction accuracies were not sufficiently high enough to use as a sole indicator of pregnancy status.

Key words: *Fourier-transform mid-infrared spectra, pregnancy prediction, milk composition, dairy cattle, machine learning*

4.3 Introduction

Knowledge of pregnancy status for dairy cattle is an important component of an efficient and productive herd management system. In an ideal seasonal calving system, oestrus is reliably detected during the mating period, so that animals are inseminated and conceive in a timely manner, resulting in a herd average 365-day calving interval. Knowing a cow is pregnant during the mating period avoids wasted re-inseminations, and early identification of non-pregnant cows could provide an opportunity to shorten interbreeding intervals and result in an increase in herd profitability (Ferguson and Galligan, 2011; Giordano et al., 2013). Moreover, knowledge of non-pregnant status beyond the mating period plays a role in herd management and culling decisions. Pregnancy status during the mating period is crudely determined by non-return to oestrus, and later in lactation is ascertained by indirect methods such as those measuring milk progesterone levels, or pregnancy-associated glycoproteins (PAG) in blood or milk, or direct methods such as transrectal palpation and ultrasonography. Direct pregnancy testing is costly and may also require additional animal-handling, and detection of pregnancy status based only on non-return to oestrus is unreliable unless oestrus detection monitoring and recording is of a high standard. Further, in instances of embryonic loss after initial pregnancy establishment, non-pregnant cows may not all return to oestrus due to the extended presence of a corpus luteum (Ricci et al., 2017). For these reasons, a methodology for determining pregnancy status using Fourier-transform mid-infrared (FT-MIR) spectroscopy is of interest, because the data is often already available from routine milk testing at 30- or 60-day intervals.

Pregnancy results in changes to metabolism and energy requirements and leads to a repartitioning of resources to different physiological functions, compared to a non-pregnant lactating animal, and has a consequent influence on milk composition in dairy cattle, particularly in mid to late lactation (Loker et al., 2009; Olori et al., 1997; Penasa et al., 2016). Previous studies have examined the impact of pregnancy stage on detailed milk composition as determined by FT-MIR spectra (Lainé et al., 2017), and have reported the ability to predict conception outcomes (Hempstalk et al., 2015) or pregnancy (Brand et al., 2021; Delhez et al., 2020; Toledo-Alvarado et al., 2018a) from FT-MIR spectra. Improvements in accuracy from incorporating FT-MIR data into pregnancy prediction models vary between studies. Toledo-Alvarado et al. (2018a) assessed and compared the ability to predict pregnancy from milk components (fat, protein, lactose and casein) or from a single wavenumber or a full FT-MIR spectra, using a Bayesian variable selection model. The best predictions of pregnancy in that study were obtained when full FT-MIR spectra were incorporated into prediction models, with area under the receiver operating characteristic curve (AUC) values of around 0.6. Delhez et al. (2020) investigated the potential of FT-MIR to predict pregnancy status of dairy cows with partial least squares discriminant analysis (PLS-DA) using residual FT-MIR spectra, evaluated from the difference between the spectra before and after insemination at a specific stage of lactation; and predicted pregnancy status within lactation stage classes, to account for the effect of lactation stage on milk composition. They found that prediction accuracies for models developed using FT-MIR spectra across different stages of lactation were limited, with AUC values of around 0.6, but that models using data after 150 days of pregnancy had promising prediction accuracies with AUC values of around 0.78. The use of deep learning models to establish pregnancy status were examined by Brand et al. (2021). They compared prediction accuracies between models developed using genetic algorithms for feature selection and network design, and transfer learning models that used a pre-trained Dense Convolutional Network (DenseNet) model. The former of these approaches resulted in high validation accuracies of 0.89, but loss values of 0.18, which were considered too high for useful application in the industry. However, models using transfer learning whereby FT-MIR spectra was converted to grey-scaled images, resulted in accuracy and loss values of 0.97 and 0.08, respectively, indicating that transfer learning can provide pregnancy prediction models with high enough accuracies for industry application.

In previous studies where FT-MIR spectra had been used to predict pregnancy, there were key differences in the manner in which records were selected for inclusion in the analysis, and how records were classified as pregnant or non-pregnant. The purpose of this study is to investigate

pregnancy prediction accuracy from FT-MIR spectra in a dataset of NZ seasonal calving herds, when differing strategies for classifying pregnant or non-pregnant records, broadly similar to those from previous studies, are used across the same dataset. We assess three strategies for partitioning records, two of which use records for cows with a subsequent calving only, whilst the third includes records for cows without a subsequent calving. We examine the impact of different ways of accounting for the effect of stage of lactation in these models, either by including days in milk as a model predictor, or by pre-adjusting the spectra for days in milk; and for a subset of models, we compare prediction accuracies from PLS-DA models to those from alternative models developed using deep learning approaches. Finally, we investigate the relationship between FT-MIR spectra and lactation stage by assessing how well days in milk can be predicted from spectral data.

4.4 Materials and methods

4.4.1 Ethics statement

All data were collected as part of routine on-farm activities and thus did not require formal ethics approval.

4.4.2 Data

Fourier-transform mid-infrared spectra

Fourier-transform mid-infrared spectra were from a wider set of 2,044,094 routine milk test samples for 1,877,456 animals, collected from Bentley FTS (Chaska, MN, USA) instruments by Livestock Improvement Corporation (LIC), as previously described in Tiplady et al. (2019). Briefly, FT-MIR spectra from milk samples analysed between September 2017 and May 2018 were pre-processed to remove outliers and standardized to account for differences between instruments. Outliers were removed according to the squared Mahalanobis distance between each spectrum and the average within-instrument spectrum from each analyser, and standardization was performed using piecewise direct standardization (Grelet et al., 2015). Spectral data consisted of light absorbance values for 899 spectral wavenumbers across the range from 649.03 to 3,998.59 cm^{-1}. These were restricted to exclude wavenumbers within noise regions as defined by Tiplady et al. (2019) (649 to 970 cm^{-1}, 1,608 to 1,682 cm^{-1} and $\geq$ 3,021 cm^{-1}). This resulted in 528 wavenumbers for use in the development of prediction equations. Exclusions were applied to

remove records with high SCC ($\geq$ 400,000 cells/ml) or records where there had been less than 30 samples processed for the herd on a day. Additionally, records were restricted to those for spring-calving animals that calved between June and November, and where the sample took place between 5 and 300 days in milk (DIM). This resulted in a dataset of 1,853,771 spectral records for 1,375,227 cows, across 5,529 herds.

Glycoproteins-based pregnancy diagnosis

Pregnancy-associated glycoproteins (PAG) are macromolecules produced by placental tissue and can provide a good indication of pregnancy (Commun et al., 2016; Green et al., 2005; Sousa et al., 2006), with accurate indication of pregnancy status achievable from PAG in milk samples as early as 25 days after successful AI (Commun et al., 2016; Ricci et al., 2015). Assessment of PAG in milk samples was undertaken at LIC's Animal Health laboratory, with cows assigned as pregnant, not pregnant or unconfirmed. In total, there were 25,493 records among available spectral records for which there was a PAG result, of which 22,235 were assigned as pregnant and 2,032 were assigned as not pregnant. The remaining 1,226 records had an unconfirmed result. Records were restricted to those that were definitively assigned as pregnant or not pregnant, and where the test date of the PAG result was $\geq$ 28 days after the last AI for the cow. This resulted in a dataset of 24,063 records, representing 24,004 cows in 202 herds. At the time of the PAG assessment, the average DIM was 186, ranging from 42 to 299 days; and the average number of days since the last mating was 103 days, ranging from 28 to 222 days.

Consolidation of spectral records with AI and calving data

Records of AI events and those of subsequent calvings were obtained for all cows with spectral records. Validated AI events were assigned where calving took place between 271 and 293 days after an AI event. To reduce the risk of assigning a record incorrectly as pregnant, if there was more than one potential AI date within a 271 and 293 window prior to successful calving, all records for that animal were excluded. Similarly, if there was a subsequent calving but no validated AI event within the 271 and 293 calving window, all records for the animal were removed. The resulting dataset was filtered to exclude all herds with animals that had a pregnancy diagnosis based on PAG, to enable the latter dataset to be used as an external herd-independent validation dataset. This resulted in a final dataset of 1,161,436 records for 863,982 animals, across 5,170 herds for generating and evaluating pregnancy prediction models. The median calving date across these records was 11th August 2017, and the median parity of cows was 3 with a range of 1 to 9.

The breed composition comprised 277,658 cows with $\geq$ 14/16 Holstein or Friesian composition; 87,111 cows with $\geq$ 14/16 Jersey composition, 446,136 cows with $\geq$ 3/16 Holstein-Friesian and $\geq$ 3/16 Jersey composition; and 53,077 cows from other breeds or crosses.

4.4.3 Strategies for classifying pregnancy status

Three different strategies that broadly reflected those from previous studies were used to select and classify records into pregnant and non-pregnant groups. For all of these strategies, records were only defined as pregnant if there was a validated AI event and a subsequent calving (n=700,332 records). The strategies varied in the manner in which non-pregnant records were assigned. Specifically, the strategies were defined as follows: (i) Records prior to the first mating were assigned as non-pregnant (n=164,537); (ii) records after the first mating but prior to the validated AI event were assigned as non-pregnant (n=14,778); and (iii) in addition to non-pregnant records used in (ii), records for cows without a subsequent calving were assigned as non-pregnant (n=197,624). Strategy (i) was similar to that defined in the study by Brand et al. (2021), whereas Strategy (ii) was similar to that defined by Toledo-Alvarado et al. (2018a), except that in their study they only retained records within 90 days after each insemination, and classified records without a subsequent insemination within 90 days as pregnant, and records with a subsequent insemination within 90 days as non-pregnant. Strategy (iii) was similar to that defined in the study by Delhez et al. (2020) in that records were not restricted to those for cows with a subsequent calving, but differed in that our dataset was not restricted to using only a single-spectral record after each insemination.

4.4.4 PLS-DA model development and validation

Animals with confirmed pregnant or non-pregnant status based on PAG formed the basis of an external validation dataset (VAL-PAG). For each pregnancy classification strategy, spectra from the remaining records were partitioned into training and validation datasets. Each training dataset consisted of records for cows from a random sample of 80% of herds, with the remaining spectra assigned to validation (VAL-Test). Random sampling with replacement was conducted to augment the minority class (non-pregnant) to be the same size as the majority class (pregnant) in the training dataset. Partial least squares discriminant analysis (PLS-DA) models were developed from training data with 10 repeats of 10-fold cross-validation using the caret package in R (Kuhn, 2008). For each pregnancy classification strategy, three types of models were evaluated: (a) models using unadjusted spectral wavenumbers as predictors; (b) models using unadjusted spectral wavenumbers but including DIM as a predictor, where DIM was fitted as a class variable

representing 30-day windows from the start of lactation; and (c) models using adjusted spectral wavenumbers as predictors, where the spectra had been pre-adjusted for DIM (30-day window classes) using repeated measures models in ASReml-R (Butler et al., 2009).

To assess the impact of augmenting the data using upsampling of the minority non-pregnant class, a secondary set of models were developed using a downsampled training dataset whereby random sampling was conducted to reduce the majority class (pregnant) to be the same size as the minority class (non-pregnant). Additionally, for models using unadjusted spectral wavenumbers as predictors, we assessed the impact of excluding records classified as pregnant where the test date was within a short time period after a validated successful AI event. Specifically, we evaluated alternative models whereby spectral records classified as pregnant were removed if the test date associated with the record was within 7, 14 or 21 days of a cow's successful AI.

Model performance for each pregnancy classification strategy and model type was assessed according to the sensitivity and specificity of pregnancy prediction, and from the AUC values when the trained model was applied to each of the two validation datasets (VAL-Test and VAL-PAG). Sensitivity was defined as the proportion of pregnant records that were correctly assigned as pregnant by the model; and specificity was defined as the proportion of non-pregnant records that were correctly assigned as non-pregnant by the model. Receiver operating characteristic (ROC) curves represent the relationship between a model's true positive rate (records correctly assigned as pregnant) and the false positive rate (records incorrectly assigned as pregnant), for different classification thresholds. The AUC measures the area under the ROC curve when values of the true positive rate are plotted against values of the false positive rate on a continuous scale, providing a consolidated measure of model performance across all possible classification thresholds. Values of AUC range from 0 to 1, with an AUC value of 0.5 indicating that the model is only able to classify records as well as random allocation of pregnant and non-pregnant status.

4.4.5 Deep learning models

Two different deep learning approaches were developed for a subset of models using training and test datasets as defined by Strategy (iii). The first approach used a multilayer perceptron (MLP) feed-forward artificial neural network to classify pregnancy status based on raw spectra, while the second used a convolutional neural network (CNN). A diagram showing the pipeline for the two approaches is in Fig. 4.1. Both deep learning approaches were implemented with PyTorch (v1.7.1; Paszke et al., 2019), and one-hot encoding was used to transform the categorical predictor DIM, which was classified by 30-day windows from the start of lactation. For the MLP network, we used two fully connected layers with leaky rectified linear unit (LeakyReLU) activation and

batch normalization (BatchNorm) to accelerate convergence speed. When trained with DIM, the obtained one-hot encoded tensor was concatenated with the spectral wavenumbers tensor prior to input to the fully connected layers. The CNN network shared the same fully connected layers design as MLP, but the spectral wavenumbers tensor went through a dense convolutional network architecture, extracting 1,024 features that were then concatenated with one-hot encoded DIM. PyTorch Image Models (Wightman, 2019) were used to generate the original DenseNet121 (Huang et al., 2017) feature extractor layers. The original DenseNet121 architecture was designed for images with 3 channels, 224 rows and 224 columns as input, while the input of our 528 spectral wavenumbers were an image with one channel, one row and 528 columns. To accommodate the difference in the image size of the spectral data, the number of input channels was changed from 3 to 1 in the first convolution layer; and the kernel size and stride of all average pooling layers was changed from (2,2) to (2,1). We applied adaptive average pooling (AdaptiveAvgPool) to the output of our 1D DenseNet121 network to reduce the overall number of parameters, and applied LeakyReLU activation and batch normalization to accelerate the convergence of the stochastic gradient descent. Across both deep learning approaches, networks were trained for 50 epochs on computers equipped with NVIDIA Titan XP or RTX 5000 graphics cards, using a batch size of 1,024 and randomized initial weights. Stochastic gradient descent was used to minimise binary cross entropy loss, with the learning rate starting at 1e-03 and reduced by a factor of 10 at epoch 15, 25 and 35. Validation of each deep learning approach was conducted in the same way as for the PLS-DA models, using test data (VAL-Test) and the separate cow independent dataset whereby pregnancy diagnosis had been assigned according to PAG in the milk sample (VAL-PAG).

4.4.6 Prediction models for stage of lactation

To investigate the relationship between FT-MIR spectra and lactation stage, a partial least squares (PLS) model was developed to predict actual DIM (in days), using the Strategy (iii) dataset. Spectra from a random sample of 80% of herds were assigned as a training dataset (n=724,864) to develop the model with 10 repeats of 10-fold cross-validation, using the caret package in R (Kuhn, 2008). The remaining spectra (n=187,870) were used for validation, comprising 145,014 pregnant and 42,856 non-pregnant records. A secondary set of models were also developed and validated, whereby the model was trained on only records classified as pregnant (n=555,318). For each model, performance was assessed according to the relative root mean square error (RMSE) between actual and predicted DIM, and according to the correlation between actual and predicted DIM in the validation datasets.

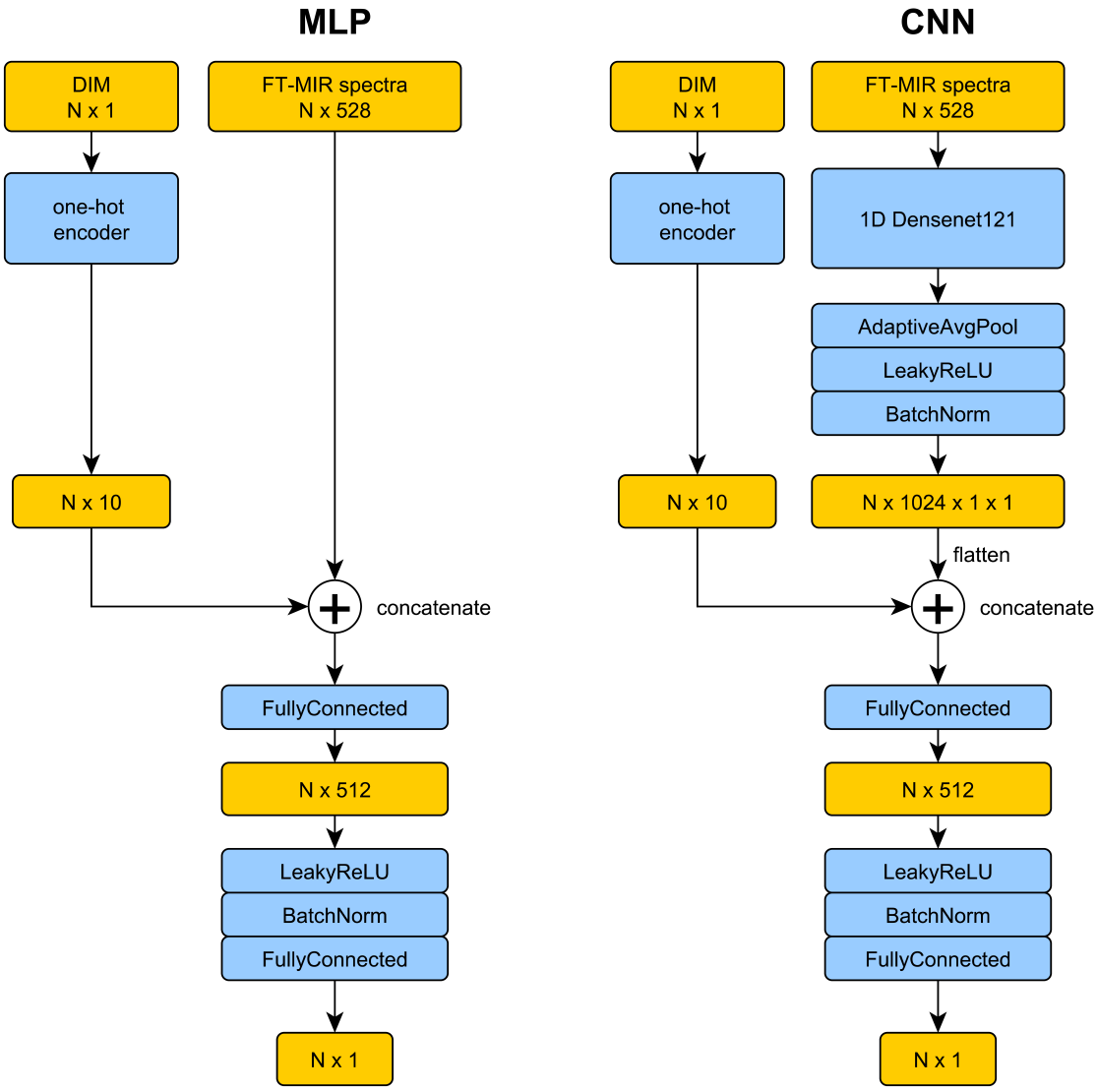

Figure 4.1: Architecture of the Multilayer Perceptron (MLP) feed-forward artificial neural network and the Convolutional Neural Network (CNN) used to classify pregnancy status. DIM: days in milk; N: batch size, representing the number of samples processed before the model is updated. One-hot encoding used to transform the categorical predictor DIM; for the CNN, spectral wavenumbers tensor passed through a dense convolutional network architecture (1D DenseNet121) and adaptive average pooling (AdaptiveAvgPool) applied, followed by Leaky Rectified Linear Unit (LeakyReLU) activation and batch normalization (BatchNorm); for both networks, one-hot encoded DIM tensor concatenated with spectral wavenumbers tensor to form fully connected layers which were passed through a further round of LeakyReLU activation and BatchNorm.

4.5 Results and discussion

4.5.1 Data description

Table 4.1 shows the number of records and cows by pregnancy status for each classification strategy, and the mean DIM values for records in each class. A large difference was observed between the average DIM of non-pregnant and pregnant records for Strategy (i) and Strategy (ii), with values of 55 to 89 for non-pregnant records, compared to 170 to 171 for pregnant records. This difference was smaller for Strategy (iii), where the average DIM for non-pregnant records was 150, compared to 170 to 171 for pregnant records; and in the external PAG validation dataset the average DIM for non-pregnant and pregnant records were 176 and 187, respectively. These differences in the distribution of DIM for records in each pregnancy status group are further demonstrated in Fig. 4.2.

Table 4.1: Total number of records and cows, and descriptive statistics (mean +/- SD) for days in milk (DIM) across pregnancy classification strategies for training and validation datasets

Dataset[1]	**Pregnant**			**Non-pregnant**		
	No. of records	No. of cows	DIM	No. of records	No. of cows	DIM
Strategy (i)						
Training data	552,263	440,083	170 (56.0)	131,056	130,167	55 (20.8)
Test validation	148,069	116,048	171 (56.1)	33,481	33,170	55 (20.5)
Strategy (ii)						
Training data	560,215	444,687	170 (56.0)	11,781	11,746	89 (22.0)
Test validation	140,117	111,407	170 (56.0)	2,997	2,987	89 (22.1)
Strategy (iii)						
Training data	557,440	443,574	170 (55.9)	167,945	141,407	150 (56.5)
Test validation	142,892	112,552	171 (56.6)	44,457	36,912	150 (57.6)
PAG validation (VAL-PAG)	22,117	22,068	187 (36.7)	1,946	1,936	176 (39.3)

[1] For all strategies, records defined as pregnant if there was a validated AI event and a subsequent calving (n=700,332 records). Non-pregnant records defined for each strategy as follows: (i) Records prior to the first mating assigned as non-pregnant (n=164,537); (ii) records after the first mating but prior to the validated AI event assigned as non-pregnant (n=14,778); and (iii) in addition to non-pregnant records used in (ii), records for cows without a subsequent calving assigned as non-pregnant (n=197,624). PAG validation (VAL-PAG): Pregnancy-associated glycoproteins validation dataset.

Figures 4.2(a)-4.2(c) show the DIM distribution of records for the training dataset of each strategy. The distribution of DIM for the VAL-PAG dataset are shown in Fig. 4.2(d). The distributions for Strategy (i) and Strategy (ii) were similar in the early stages of lactation, in that there was a good representation of non-pregnant records, but beyond ~120 days there were few non-pregnant records (Figs. 4.2(a); 4.2(b)). Strategy (iii) differed in that there was representation of both pregnancy classifications across lactation (Fig. 4.2(c)). Similarly, in the VAL-PAG dataset, both pregnancy classifications were well represented across lactation (Fig. 4.2(d)).

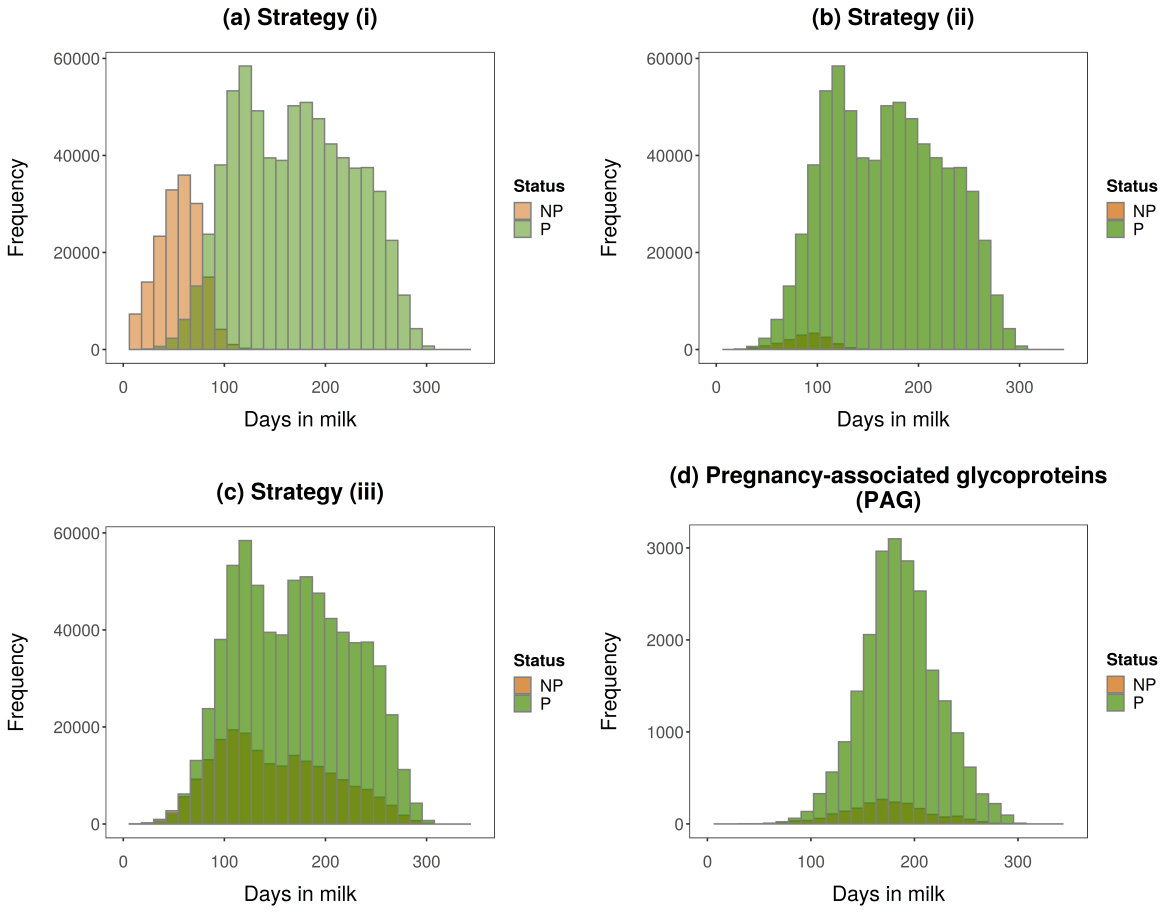

Figure 4.2: Frequency of pregnant (P) and non-pregnant (NP) records across days in milk for training and validation records for (a) Strategy (i); (b) Strategy (ii); (c) Strategy (iii); and (d) Pregnancy-associated glycoproteins (PAG) records. For all strategies, records defined as P if there was a validated AI event and a subsequent calving (n=700,332 records). Records were defined as NP for each strategy as follows: (i) Records prior to the first mating (n=164,537); (ii) records after the first mating but prior to the validated AI event (n=14,778); and (iii) in addition to non-pregnant records used in (ii), records for cows without a subsequent calving were assigned as non-pregnant (n=197,624).

4.5.2 Diagnosis of pregnancy status using PLS-DA models

In this study, we compared three strategies for selecting FT-MIR spectral records for analysis and partitioning records into non-pregnant and pregnant groups. Table 4.2 shows prediction accuracies for PLS-DA models within the training, VAL-Test and VAL-PAG datasets for each strategy and model type. For each strategy, models that used unadjusted FT-MIR spectra as predictors outperformed the prediction accuracy of models that used spectra that had been pre-adjusted for DIM. Boxplots representing prediction probabilities for non-pregnant and pregnant records using unadjusted FT-MIR spectra for each strategy are provided in Fig. 4.3.

Strategy (i) was comparable to the approach used by Brand et al. (2021), where only cows with a subsequent calving were included, with records prior to the first mating assigned as non-pregnant, and records after a validated AI event assigned as pregnant. Brand et al. (2021) reported promising predictive ability to classify pregnancy status in a large dataset of FT-MIR spectra from UK herds using PLS-DA models, with accuracy, sensitivity and specificity values of 0.77, 0.73 and 0.82, respectively. In our study, the model using unadjusted FT-MIR spectra with Strategy (i) data had sensitivity of 0.94, specificity of 0.96, and AUC values of 0.99 for the VAL-Test dataset (Table 4.2), higher than those reported by Brand et al. (2021). However, for this model, the specificity to correctly classify non-pregnant cows in the VAL-PAG dataset was poor (0.002). This lack of consistency in prediction accuracy across the training and validation datasets for the Strategy (i) dataset is clearly demonstrated in Figs. 4.3(a)-4.3(c), where we observed good partitioning between the distribution of prediction probabilities in the training and VAL-Test datasets (Fig. 4.3(a), 4.3(b)), but a tendency to predict non-pregnant records as pregnant in the VAL-PAG dataset (Fig. 4.3(c)). When DIM was included as a predictor for Strategy (i) models, accuracies in the VAL-Test and VAL-PAG datasets were relatively unchanged, compared to fitting FT-MIR spectra alone (Table 4.2). However, pre-adjusting spectra for DIM resulted in prediction accuracies that were more consistent across training and validation datasets, with sensitivity of 0.66, specificity of 0.61 and AUC values of 0.68 in the VAL-Test dataset; and 0.63, 0.42 and 0.53 in the VAL-PAG dataset, respectively. Although the training and VAL-Test accuracies were lower in the model that used spectra pre-adjusted for DIM, the improved consistency in results across training and both validation datasets indicated that using pre-adjusted spectra was at least partially effective at removing some of the confounding effect between stage of lactation and pregnancy status.

Table 4.2: Model performance for partial least squares discriminant analysis (PLS-DA) models with upsampling[1]: Accuracy (Acc), sensitivity (Sens), specificity (Spec) and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test) and pregnancy-associated glycoproteins validation (VAL-PAG) datasets

Classification Strategy[2] and Model[3]	**Training**				**Test validation** (VAL-Test)				**Glycoprotein-based validation (VAL-PAG)**			
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
Strategy (i)												
FT-MIR spectra	0.938	0.932	0.966	0.987	0.941	0.936	0.961	0.987	0.918	0.998	0.002	0.559
FT-MIR spectra + DIM	0.940	0.934	0.966	0.991	0.946	0.940	0.972	0.993	0.917	0.997	0.014	0.544
FT-MIR spectra (pre-adjusted for DIM)	0.672	0.675	0.660	0.723	0.654	0.664	0.606	0.676	0.613	0.630	0.421	0.532
Strategy (ii)												
FT-MIR spectra	0.807	0.805	0.931	0.922	0.801	0.799	0.906	0.914	0.912	0.990	0.018	0.572
FT-MIR spectra + DIM	0.800	0.797	0.943	0.932	0.794	0.791	0.923	0.926	0.907	0.984	0.037	0.585
FT-MIR spectra (pre-adjusted for DIM)	0.716	0.716	0.709	0.770	0.705	0.708	0.604	0.701	0.734	0.776	0.262	0.523
Strategy (iii)												
FT-MIR spectra	0.596	0.594	0.603	0.637	0.599	0.600	0.596	0.636	0.665	0.673	0.571	0.668
FT-MIR spectra + DIM	0.617	0.626	0.588	0.649	0.618	0.628	0.586	0.649	0.697	0.711	0.536	0.677
FT-MIR spectra (pre-adjusted for DIM)	0.568	0.567	0.571	0.596	0.568	0.572	0.554	0.589	0.573	0.569	0.622	0.639

[1] Upsampling undertaken using random sampling with replacement to augment the minority class (non-pregnant) to be the same size as the majority class (pregnant).

[2] For all strategies, records defined as pregnant if there was a validated AI event and a subsequent calving (n=700,332 records). Non-pregnant records defined for each strategy as follows: (i) Records prior to the first mating assigned as non-pregnant (n=164,537); (ii) records after the first mating but prior to the validated AI event assigned as non-pregnant (n=14,778); and (iii) in addition to non-pregnant records used in (ii), records for cows without a subsequent calving assigned as non-pregnant (n=197,624).

[3] FT-MIR spectra models utilise spectral wavenumbers as predictors only; FT-MIR spectra + DIM models utilise spectral wavenumbers and days in milk (30-day window class) as predictors; FT-MIR spectra (pre-adjusted for DIM) models utilise spectral wavenumbers pre-adjusted for days in milk (30-day window class).

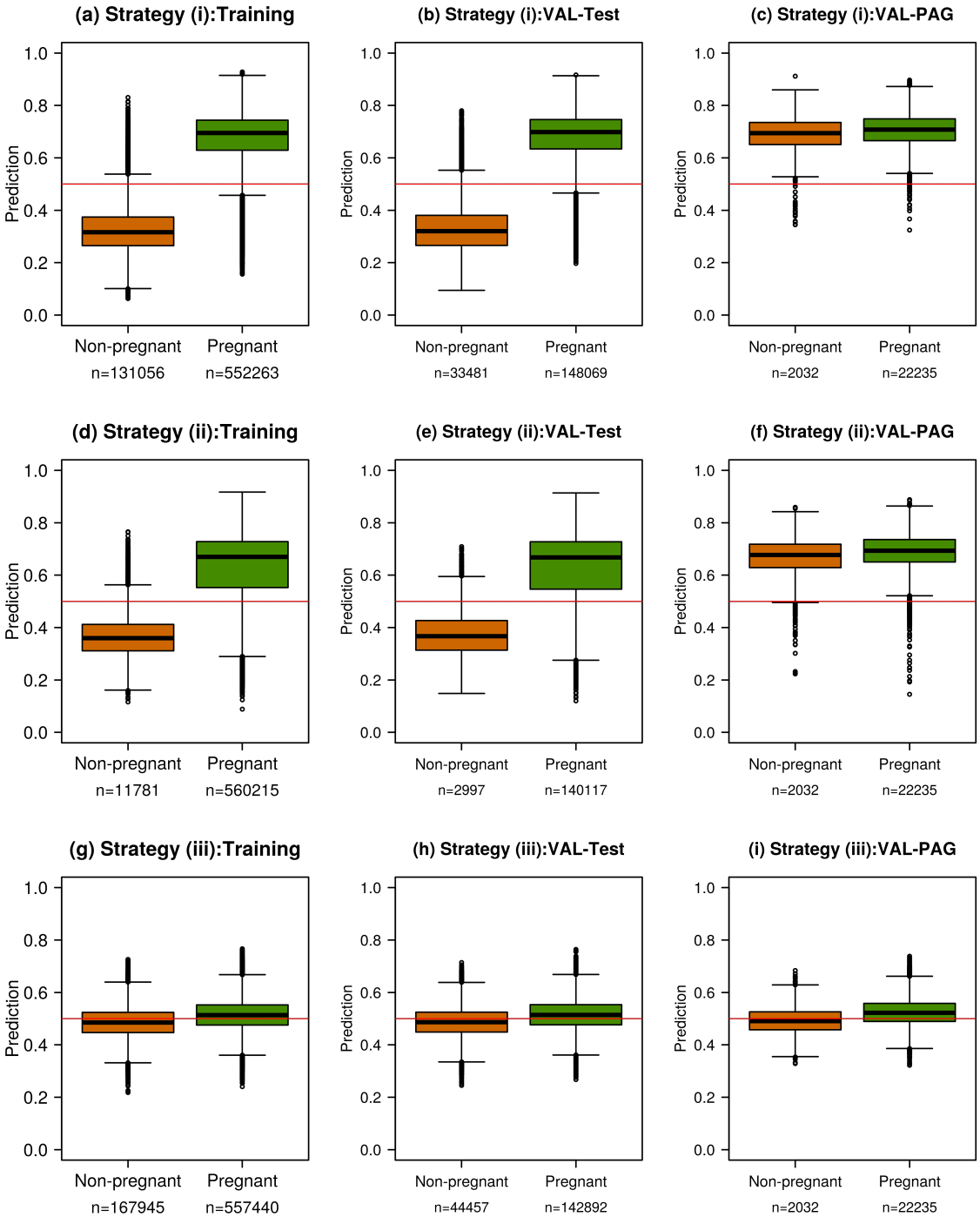

Figure 4.3: Summary of prediction probabilities for non-pregnant and pregnant records for training and validation datasets based on differing strategies for record selection and pregnancy status classification. VAL-Test: herd-independent validation; VAL-PAG: external validation dataset based on pregnancy-associated glycoproteins.

Strategy (ii) for classifying pregnancy status was comparable to the approach used by Toledo-Alvarado et al. (2018a), whereby only cows with a subsequent calving were included, and records between the first mating and a validated AI event were assigned as non-pregnant, and records after a validated AI event were assigned as pregnant. Toledo-Alvarado et al. (2018a) classified pregnancy status using FT-MIR spectra from cattle raised in heterogeneous farming systems in north-eastern Italy using a Bayesian model, and reported cross-validation AUC values of ~0.6 to 0.66. In our study, Strategy (ii) models using unadjusted spectra had comparatively higher AUC values for the VAL-Test dataset (0.91 to 0.93). However, in the VAL-PAG dataset these dropped to between 0.57 and 0.59, and the specificity to correctly classify non-pregnant cows in the VAL-PAG dataset was poor (0.02 to 0.04). Similar to the observations for Strategy (i), we observed good partitioning between the distribution of prediction probabilities in the training and VAL-Test datasets (Fig. 4.3(d), 4.3(e)), but a tendency to predict non-pregnant records as pregnant in the VAL-PAG dataset (Fig. 4.3(f)). A small improvement was observed in the specificity to correctly classify non-pregnant records in the VAL-PAG dataset when spectra were pre-adjusted for stage of lactation, however the AUC value for this model was still low (0.52). The lack of consistency in prediction accuracies across training and validation datasets for Strategy (ii) were similar to those for Strategy (i), indicating a lack of robustness in the models, likely due to confounding between pregnancy status and stage of lactation in the training dataset, and a lack of representation of non-pregnant and pregnant records across lactation.

Prediction accuracies for Strategy (iii) models were relatively consistent across the training and validation datasets for unadjusted and adjusted spectra. For models including unadjusted spectra only, sensitivity was 0.67, specificity was 0.57 and AUC values were 0.67 for the VAL-PAG dataset. Unlike the other two strategies, we did not observe clear partitioning between the distribution of prediction probabilities in the training and VAL-Test datasets (Fig. 4.3(g), 4.3(h)), however, the observed trend was consistent in the VAL-PAG dataset (Fig. 4.3(i)). For models using unadjusted spectra that also included DIM as a predictor, AUC values in the VAL-PAG dataset increased from 0.67 to 0.68. Notably, a decline in prediction accuracy was observed for models that used spectra pre-adjusted for DIM, with overall accuracy and sensitivity dropping to 0.57, and AUC values dropping to 0.64. Strategy (iii) models were comparable to the approach used by Delhez et al. (2020), whereby records for cows were included (and assigned as non-pregnant) if they did not have a subsequent calving. Delhez et al. (2020) classified pregnancy status in a dataset of Australian Holstein cattle using PLS-DA models with FT-MIR spectra as independent predictors, and observed validation AUC values of 0.63 to 0.65. They also examined the effect of

using residual spectra, evaluated as the difference between a non-pregnant record and pregnant record for the same animal, but did not see an improvement in results. However, they did observe an improvement in prediction accuracy for models developed from spectra in different stages of lactation, particularly for spectra recorded after 150 days of lactation, with validation AUC values of 0.76 to 0.78. We also undertook a similar approach for the Strategy (iii) dataset whereby we fitted separate models for different stages of lactation (Appendix Table 4.A.1) and observed a consistent increase in AUC values after 210 days of lactation in the VAL-PAG dataset (0.68 to 0.76), but the overall prediction accuracy after 210 days of lactation remained low (0.55 to 0.64).

Prediction accuracies for PLS-DA models where the majority class (pregnant) was reduced to be the same size as the minority class (non-pregnant) are shown in Appendix Table 4.A.2. Prediction accuracy metrics across all strategies and model types were only marginally different to those presented in Table 4.1. This indicated that augmenting the minority class to be the same size as the majority class did not introduce bias to the results. Moreover, it indicated that the reduced dataset where the majority class was downsampled to be the same size as the minority class, sufficiently captured the extent of the relationships between spectral wavenumbers and pregnancy classification.

Prediction accuracies for PLS-DA models based on unadjusted spectral wavenumbers as predictors are shown in Appendix Table 4.A.3, whereby records classified as pregnant were removed if the test date was within 7, 14 or 21 days of successful AI. The premise for this analysis was that changes in an animal's physiological status might not be detectable in milk composition for some time after pregnancy is established, and that including those records could lower prediction accuracy. Of the total 700,332 pregnant records, 22,338 had a test date within 7 days after a validated AI event; 48,573 had a test date within 14 days after a validated AI event; and 81,581 head a test date within 21 days after a validated AI event. For all strategies, prediction accuracy metrics were relatively unchanged by removing these records (Appendix Table 4.A.3). This indicated that although there may be changes in the physiological status of an animal that are not detectable in milk composition shortly after successful AI, including and classifying those records as pregnant did not impact on pregnancy prediction accuracies, compared to completely ignoring them.

4.5.3 Diagnosis of pregnancy status using deep learning models

Deep learning is a subclass of machine learning that uses neural networks with multiple layers to extract features from data. These neural networks consist of densely interconnected processing nodes arranged into layers, with each node receiving information from nodes in the layer beneath it and sending data to the nodes in the layer above it. The complexity of these networks enable training models to be developed on datasets with multiple connections, making them a good choice for managing high-dimensional datasets such as those presented from FT-MIR spectra. Previous studies have established that it is possible to use artificial neural networks to identify features in spectra relating to pregnancy status (2018), and that this could be extended to predict bovine tuberculosis status of individual cows (Denholm et al., 2020). More recently, Brand et al. (2021) assessed the accuracy of predicting pregnancy status using a deep learning image-based approach with a pre-trained Dense Convolutional Network (DenseNet), compared to a PLS-DA approach. In that study, when a deep learning image-based approach was employed, they observed an increase in sensitivity from 0.77 to 0.88, an increase in specificity from 0.73 to 0.89, and an increase in the AUC value from 0.82 to 0.89.

In this study, we assessed pregnancy status prediction accuracies for two deep learning approaches, and compared these to the accuracies achieved from PLS-DA models. The first approach utilised a simple MLP with one hidden layer, using 4,600 parameters, whilst the second imaged-based CNN approach was significantly more complex with up to 7.4 million parameters. Convolutional neural networks are widely used in the computer vision domain and achieve the best performance when applied on image inputs. This type of architecture can efficiently extract local features, patterns and textures which are very common in natural images generated from optical cameras. In this study, adjacent spectral wavenumbers also present high correlations and thus may contain local patterns that a CNN model can learn and extract. In contrast to the Brand et al. (2021) study, we did not apply transfer learning as we deemed this unnecessary because: a) most pre-trained DenseNet networks are trained on images generated from optical cameras, (e.g: ImageNet), while our spectral data were acquired from a different physical process, namely FT-MIR spectroscopy; and b) we use a large training dataset with more than 1 million samples.

Training and validation accuracy and loss values for deep learning approaches are shown in Fig. 4.4. Accuracies for the VAL-Test dataset stabilized gradually after epochs 15, 25 and 35, corresponding to the reduction of the learning rate. In the case of our CNN trained on adjusted spectra, the model started overfitting from around epoch 10, with an increasing accuracy of the

training dataset while accuracy for the VAL-Test dataset stayed the same. This was despite the usage of batch normalization for regularization. Prediction accuracies for the MLP and CNN approaches within the training, VAL-Test and VAL-PAG datasets are shown in Table 4.3. For each approach, the best performing models used unadjusted spectra and DIM as predictors.

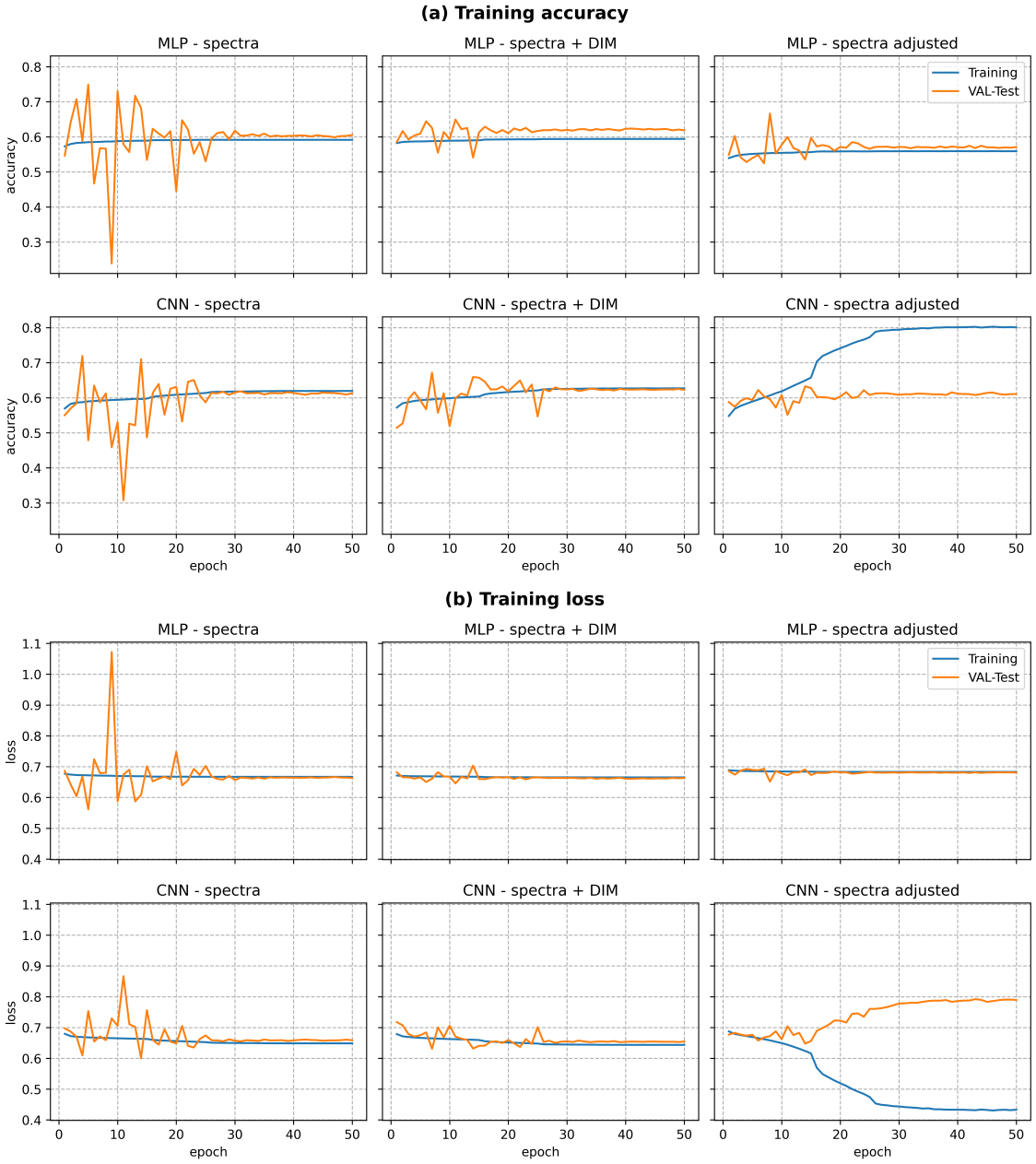

Figure 4.4: Accuracy (a) and loss (b) values for deep learning approaches, assessed using training and herd-independent validation (VAL-Test) datasets.

Table 4.3: Model performance for multilayer perceptron (MLP) and convolutional neural network (CNN) approaches based on Strategy (iii)[1] data: Accuracy (Acc), sensitivity (Sens), specificity (Spec) and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test) and pregnancy-associated glycoproteins validation (VAL-PAG) datasets

Deep learning approach[2] and Model[3]	**Training**				**Test validation (VAL-Test)**				**Glycoprotein-based validation (VAL-PAG)**			
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
MLP approach												
FT-MIR spectra	0.592	0.574	0.611	0.628	0.586	0.580	0.607	0.632	0.664	0.672	0.569	0.669
FT-MIR spectra + DIM	0.594	0.621	0.566	0.631	0.614	0.629	0.564	0.635	0.692	0.709	0.499	0.647
FT-MIR spectra (pre-adjusted for DIM)	0.559	0.554	0.564	0.583	0.562	0.567	0.547	0.581	0.554	0.547	0.636	0.636
CNN approach												
FT-MIR spectra	0.625	0.625	0.625	0.675	0.611	0.620	0.582	0.641	0.684	0.696	0.554	0.676
FT-MIR spectra + DIM	0.645	0.670	0.620	0.700	0.636	0.659	0.563	0.654	0.723	0.741	0.519	0.685
FT-MIR spectra (pre-adjusted for DIM)	0.982	0.975	0.988	0.998	0.668	0.790	0.273	0.551	0.759	0.805	0.266	0.564

[1] Strategy (iii): Records defined as pregnant if there was a validated AI event and a subsequent calving (n=700,332 records). Records after the first mating but prior to a validated AI event assigned as non-pregnant (n=14,778); records for cows without a subsequent calving assigned as non-pregnant (n=197,624).

[2] The multilayer perceptron (MLP) feed-forward artificial neural network classified pregnancy status based on raw spectra; the convolution neural network (CNN) used an image-based approach to classify pregnancy status.

[3] FT-MIR spectra models utilise spectral wavenumbers as predictors only; FT-MIR spectra + DIM models utilise spectral wavenumbers and days in milk (30-day window class) as predictors; FT-MIR spectra (pre-adjusted for DIM) models utilise spectral wavenumbers pre-adjusted for days in milk (30-day window class).

Across all models, prediction accuracies for the MLP approach were similar to those from PLS-DA models, however we observed a marginal increase in prediction accuracy for the models using an image-based CNN approach. For the image-based model that used unadjusted spectra and DIM as predictors, the overall prediction accuracy for the external PAG validation dataset increased from 0.70 to 0.72, sensitivity increased from 0.71 to 0.74, and the AUC value increased from 0.68 to 0.69, but the specificity decreased from 0.54 to 0.52. Notably, gains in prediction accuracy from adopting a deep learning approach were lower than those previously reported by Brand et al. (2021).

4.5.4 Prediction models for stage of lactation

To understand the magnitude of the effect that stage of lactation may be having on pregnancy prediction, we used the Strategy (iii) dataset to develop and validate a PLS model for predicting DIM from FT-MIR spectra. Records from a random sample of 80% of herds were used as a training dataset to develop the prediction model, with the remaining Strategy (iii) records used as herd-independent validation datasets for pregnant and non-pregnant records, respectively. The relationships between predicted and actual DIM values for the training dataset and two validation datasets are shown in Fig. 4.5. Consistently high correlations between actual and predicted DIM were observed across the training and validation datasets when the DIM prediction model was developed across all (pregnant and non-pregnant) records (0.89 to 0.90; Figs. 4.5(a)-4.5(c)). When only records assigned as pregnant were used to develop the DIM prediction model, correlations between actual and predicted DIM remained high (0.89 to 0.90; Figs. 4.5(d)-4.5(f)). Relative RMSE values for the validation dataset of pregnant records dropped marginally from 14.5% to 14.4% when only records assigned as pregnant were used to develop the model (Fig. 4.5(b); 4.5(e)), whereas relative RMSE values for the validation dataset of non-pregnant records increased marginally from 17.3% to 17.6% when only records assigned as pregnant were used to develop the model (Fig. 4.5(c); 4.5(f)). These marginal shifts in validation prediction accuracy when models were developed on all records versus only records assigned as pregnant, highlighted that in pasture-based seasonal calving systems the underlying relationship between FT-MIR spectra and DIM was upheld regardless of pregnancy status.

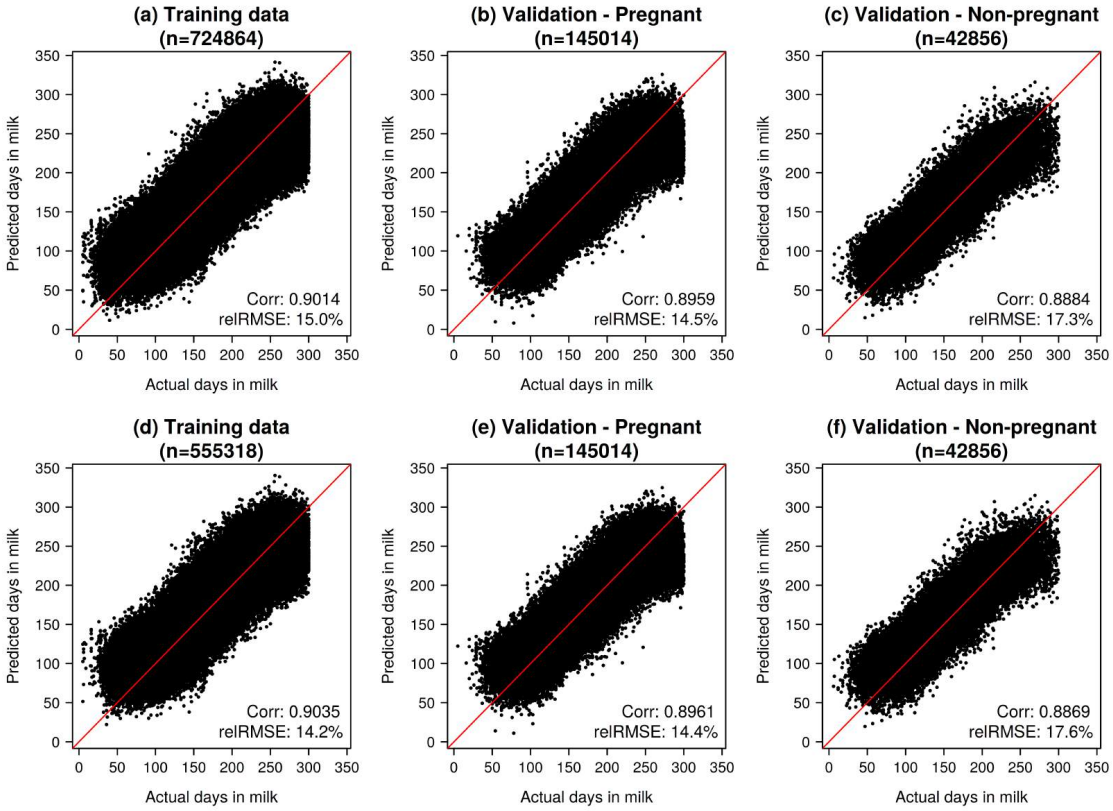

Figure 4.5: Summary of predicted vs actual days in milk (DIM) for training and validation datasets from partial least squares (PLS) prediction models based on pregnant and non-pregnant records (a-c); and prediction models based on pregnant records only (d-f). Continuous red line represents y=x. Corr. is the correlation between actual and predicted DIM; and relRMSE is the relative root mean-square error between actual and predicted DIM.

4.5.5 Confounding between pregnancy status and stage of lactation

Changes in dairy cattle milk composition across lactation are more noticeable in seasonal pasture-based farming systems (compared to non-seasonal systems), where compact calving periods are used so that peak lactation volumes are matched with peak grass growth (Timlin et al., 2021). Although NZ dairy systems are mainly pasture-based, intensification has resulted in widespread use of supplement feed to offset the effect of high-stocking rates and ensure that the nutritional requirements of cows are met. In particular, there has been an increased use of Palm Kernel Extract (PKE) and maize silage as supplements in NZ dairy systems over the last two decades (Ministry for Primary Industries, NZ, 2017). Palm Kernel Extract is associated with an increase in milk fat content (DairyNZ, 2017; van Wyngaard and Meeske, 2017) and changes to milk fatty acid composition (Dias, 2010; Oliveira et al., 2015), and has resulted in the introduction of a Fat Evaluation Index (FEI) by Fonterra in 2018 to assess the suitability of milk for processing

(DairyNZ, 2017). More generally, fatty acids have been shown to change with different dietary systems (Elgersma, 2015) and levels of pasture in the diet (Butler et al., 2011; Couvreur et al., 2006; O'Callaghan et al., 2016; White et al., 2001). Nevertheless, across different diets, as lactation progresses, consistently lower milk volumes (McAuliffe et al., 2016) and higher concentrations of fat and crude protein have been reported (O'Callaghan et al., 2016). This can be problematic for the prediction of indirect traits such as pregnancy status from FT-MIR spectra, particularly when the spectra are from seasonal calving herds, because as lactation progresses, changes in milk composition coincide with the advent of a cow becoming pregnant, and most cows do become pregnant. In seasonal calving pasture-based systems, there are also other changes that are confounded with lactation stage such as climatic changes and the use of dietary supplements to ensure that the energy requirements of cows are met at times when pasture growth is low. It is thus important to ensure that pregnancy prediction models based on FT-MIR spectra include a consistent representation of pregnant and non-pregnant records across all stages of lactation. In this study, the most robust prediction accuracies were achieved when the prediction model was developed on a dataset with a good representation of pregnant and non-pregnant records across lactation. Contrary to this, when we developed prediction models on datasets that did not have a good representation of pregnant and non-pregnant records across lactation, prediction accuracies appeared promising in initial validation, but did not perform well in the external PAG validation dataset. In future research, to improve prediction accuracies for pregnancy status from FT-MIR spectra within a seasonal calving pasture-based context, it is important that careful consideration is given to how models can account for the confounding effects of factors such as lactation stage, feed management and seasonality. Whilst some of this could be addressed by including multiple seasons of data, including other information such as knowledge of feed management and supplementation may also play an important role.

4.5.6 Prediction model validation strategies

This is not the first study to highlight differences in prediction accuracy for FT-MIR predicted traits, depending on the validation dataset/strategy used. Wang and Bovenhuis (2019) observed that using a random cross-validation approach to predict methane (CH_4) emissions from FT-MIR spectra resulted in overoptimistic results. Other studies show that prediction accuracies can be inflated by the split-data strategy used for validation, with cow-independent validation having lower accuracies compared to record-independent cross-validation (Shetty et al., 2017; Smith et al., 2019), and trial- or herd-independent validation having lower accuracies compared to record- or

cow-independent validation (Dórea et al., 2018; Lahart et al., 2019; Luke et al., 2019b). Recently, Bresolin and Dórea (2020) have reviewed the impact of validation strategies on predictive quality for a range of FT-MIR predicted milk composition and animal health traits, and highlight the value of an external validation dataset whereby the external validation uses data from a different herd, trial, or season. In 67 of the 113 studies they reviewed, internal validation (holdout, leave-one-out, k-fold) was performed. Of the 32 papers they reviewed that used an external validation, only 17 conducted validation using an independent dataset based on herd, trial or season, whereas the other 15 used cow-independent validation. In our study, we demonstrate that in some instances, even a herd-independent validation approach can overestimate prediction accuracies, if there is systemic confounding between the trait of interest and other underlying factors in the FT-MIR spectra. Specifically, where there were divergent DIM characteristics between pregnancy status groups, we were not only predicting changes in milk composition due to pregnancy, but also changes due to other factors.

4.6 Conclusions

We have assessed and compared pregnancy prediction accuracy from FT-MIR spectra using different strategies for classifying pregnancy status and accounting for the effect of stage of lactation. We have also compared prediction accuracies from PLS-DA models to alternative models developed using deep learning approaches. We have shown that the ability to predict pregnancy status from FT-MIR spectra is influenced by which records are used, and how these records are partitioned into pregnant and non-pregnant groups. Prediction models developed on datasets without adequate representation of non-pregnant and pregnant status across lactation, led to misleading results, whereby prediction accuracies were high in the training and herd-independent validation dataset, but were not upheld for an external validation dataset where pregnancy status was assigned according to pregnancy-associated glycoproteins in milk samples. This demonstrated that even with herd-independent validation, prediction accuracies can be misleading where there is systematic confounding between pregnancy status and other factors such as stage of lactation. For models where the effect of this confounding was reduced, prediction accuracies were not sufficiently high to be used as a sole indicator of pregnancy status within a seasonal calving herd management context.

4.7 Acknowledgements

The authors would like to acknowledge LIC (Hamilton, New Zealand) herd-testing staff for the processing and analysis of milk samples. Kathryn would also like to acknowledge and thank the wider LIC R&D team and fellow students for underlying technical support and thoughtful discussion, and Tod Schilling (Bentley Instruments Inc., Chaska, USA) and Pierre Broutin (Bentley Instruments Inc., Lille, France) for their help with obtaining FT-MIR spectra from Bentley instruments. With gratitude, we also recognize the use of New Zealand eScience Infrastructure (NeSI) high-performance computing for this research. The funding for this research was provided by Livestock Improvement Corporation (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).

Appendices

4.A Pregnancy status predicted using FT-MIR milk spectra from dairy cattle

Table 4.A.1: Record numbers and model performance for partial least squares discriminant analysis (PLS-DA) models fitted within stage of lactation classes: Accuracy (Acc), sensitivity (Sens), specificity (Spec) and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test) and pregnancy-associated glycoproteins validation (VAL-PAG) datasets

Stage of lactation class	**Training**						**Test validation** (VAL-Test)						**Glycoprotein-based validation (VAL-PAG)**					
	Pregnant	Non-pregnant	Acc	Sens	Spec	AUC	Pregnant	Non-pregnant	Acc	Sens	Spec	AUC	Pregnant	Non-pregnant	Acc	Sens	Spec	AUC
5 to 30 days	127	261	0.956	0.937	0.966	0.989	10	52	0.677	0.700	0.673	0.688	0	0	-	-	-	-
31 to 60 days	4315	4783	0.581	0.582	0.581	0.616	1189	1351	0.543	0.573	0.517	0.554	1	0	-	-	-	-
61 to 90 days	32034	20045	0.576	0.581	0.569	0.608	8420	5583	0.549	0.553	0.541	0.564	77	28	0.590	0.597	0.571	0.666
91 to 120 days	96794	36670	0.582	0.589	0.563	0.607	24176	9913	0.575	0.595	0.525	0.583	661	124	0.508	0.474	0.694	0.648
121 to 150 days	93889	29094	0.602	0.608	0.583	0.632	23618	7399	0.601	0.618	0.549	0.617	2616	346	0.478	0.436	0.792	0.679
151 to 180 days	88318	25074	0.606	0.610	0.593	0.640	21742	6380	0.578	0.581	0.566	0.606	6243	574	0.614	0.615	0.605	0.662
181 to 210 days	92083	23008	0.609	0.612	0.596	0.644	24433	5966	0.598	0.601	0.584	0.626	7104	522	0.656	0.665	0.536	0.640
211 to 240 days	76476	16610	0.618	0.620	0.611	0.660	19683	4233	0.600	0.606	0.572	0.623	3593	220	0.588	0.579	0.736	0.681
240 to 270 days	59902	10199	0.624	0.623	0.633	0.678	15487	2890	0.608	0.610	0.596	0.640	1476	124	0.554	0.539	0.726	0.691
271 to 300 days	13502	2201	0.674	0.670	0.701	0.747	4134	690	0.631	0.638	0.587	0.652	346	8	0.644	0.645	0.625	0.759
Overall	557440	167945	0.604	0.609	0.588	0.639	142892	44457	0.589	0.599	0.555	0.609	22117	1946	0.600	0.595	0.647	0.669

Table 4.A.2: Model performance for partial least squares discriminant analysis (PLS-DA) models with downsampling[1]: Accuracy (Acc), sensitivity (Sens), specificity (Spec) and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test) and pregnancy-associated glycoproteins validation (VAL-PAG) datasets

Classification Strategy[2] and Model[3]	**Training**				**Test validation (VAL-Test)**				**Glycoprotein-based validation (VAL-PAG)**			
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
Strategy (i)												
FT-MIR spectra	0.937	0.930	0.966	0.987	0.940	0.935	0.961	0.987	0.918	0.998	0.002	0.560
FT-MIR spectra + DIM	0.940	0.934	0.966	0.991	0.946	0.940	0.972	0.992	0.917	0.997	0.014	0.544
FT-MIR spectra (pre-adjusted for DIM)	0.671	0.673	0.661	0.723	0.651	0.661	0.608	0.676	0.608	0.624	0.432	0.531
Strategy (ii)												
FT-MIR spectra	0.803	0.800	0.930	0.920	0.798	0.795	0.909	0.912	0.911	0.990	0.020	0.572
FT-MIR spectra + DIM	0.796	0.793	0.945	0.929	0.792	0.789	0.925	0.923	0.907	0.982	0.046	0.584
FT-MIR spectra (pre-adjusted for DIM)	0.701	0.701	0.712	0.763	0.692	0.694	0.616	0.696	0.719	0.757	0.277	0.524
Strategy (iii)												
FT-MIR spectra	0.594	0.592	0.604	0.637	0.598	0.599	0.595	0.636	0.663	0.671	0.578	0.669
FT-MIR spectra + DIM	0.616	0.624	0.591	0.649	0.618	0.627	0.588	0.649	0.694	0.707	0.544	0.679
FT-MIR spectra (pre-adjusted for DIM)	0.566	0.564	0.573	0.596	0.567	0.571	0.554	0.588	0.571	0.566	0.639	0.643

[1] Downsampling undertaken whereby random sampling was conducted to reduce the majority class (pregnant) to be the same size as the minority class (non-pregnant).

[2] For all strategies, records defined as pregnant if there was a validated AI event and a subsequent calving (n=700,332 records). Non-pregnant records defined for each strategy as follows: (i) Records prior to the first mating assigned as non-pregnant (n=164,537); (ii) records after the first mating but prior to the validated AI event assigned as non-pregnant (n=14,778); and (iii) in addition to non-pregnant records used in (ii), records for cows without a subsequent calving assigned as non-pregnant (n=197,624).

[3] FT-MIR spectra models utilise spectral wavenumbers as predictors only; FT-MIR spectra + DIM models utilise spectral wavenumbers and days in milk (30-day window class) as predictors; FT-MIR spectra (pre-adjusted for DIM) models utilise spectral wavenumbers pre-adjusted for days in milk (30-day window class).

Table 4.A.3: Model performance for partial least squares discriminant analysis (PLS-DA) models with FT-MIR spectral wavenumbers as predictors, excluding records classified as pregnant if the test date was within 7, 14 or 21 days after a validated AI event: Accuracy (Acc), sensitivity (Sens), specificity (Spec) and area under the receiver operating characteristic curve (AUC) values within the training, herd-independent validation (VAL-Test) and pregnancy-associated glycoproteins validation (VAL-PAG) datasets

Classification Strategy[1] and Model[2]	**Training**				**Test validation** (VAL-Test)				**Glycoprotein-based validation (VAL-PAG)**			
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
Strategy (i)												
FT-MIR spectra (all recs.)	0.938	0.932	0.966	0.987	0.941	0.936	0.961	0.987	0.918	0.998	0.002	0.559
FT-MIR spectra (excl. recs. within 7 days post-AI)	0.932	0.923	0.972	0.986	0.938	0.930	0.972	0.987	0.918	0.998	0.001	0.556
FT-MIR spectra (excl. recs. within 14 days post-AI)	0.928	0.916	0.974	0.986	0.933	0.923	0.973	0.987	0.917	0.998	0.001	0.556
FT-MIR spectra (excl. recs. within 21 days post-AI)	0.913	0.897	0.982	0.985	0.920	0.907	0.981	0.986	0.917	0.998	0.002	0.556
Strategy (ii)												
FT-MIR spectra	0.807	0.805	0.931	0.922	0.801	0.799	0.906	0.914	0.912	0.990	0.018	0.572
FT-MIR spectra (excl. recs. within 7 days post-AI)	0.803	0.801	0.933	0.920	0.797	0.794	0.929	0.919	0.910	0.988	0.022	0.566
FT-MIR spectra (excl. recs. within 14 days post-AI)	0.797	0.794	0.937	0.920	0.790	0.787	0.935	0.920	0.909	0.987	0.023	0.569
FT-MIR spectra (excl. recs. within 21 days post-AI)	0.794	0.791	0.941	0.918	0.787	0.784	0.942	0.918	0.910	0.988	0.023	0.565
Strategy (iii)												
FT-MIR spectra	0.596	0.594	0.603	0.637	0.599	0.600	0.596	0.636	0.665	0.673	0.571	0.668
FT-MIR spectra (excl. recs. within 7 days post-AI)	0.594	0.592	0.602	0.635	0.601	0.600	0.603	0.642	0.676	0.687	0.548	0.668
FT-MIR spectra (excl. recs. within 14 days post-AI)	0.591	0.588	0.602	0.633	0.598	0.597	0.603	0.639	0.687	0.701	0.532	0.662
FT-MIR spectra (excl. recs. within 21 days post-AI)	0.588	0.585	0.600	0.628	0.596	0.595	0.599	0.636	0.710	0.729	0.496	0.662

[1] For all strategies, records defined as pregnant if there was a validated AI event and a subsequent calving (n=700,332 records). Non-pregnant records defined for each strategy as follows: (i) Records prior to the first mating assigned as non-pregnant (n=164,537); (ii) records after the first mating but prior to the validated AI event assigned as non-pregnant (n=14,778); and (iii) in addition to non-pregnant records used in (ii), records for cows without a subsequent calving assigned as non-pregnant (n=197,624).

[2] FT-MIR spectra models (all recs.): All records included; FT-MIR spectra (excl. recs. within [7,14,21] days post-AI): Records removed if the test date was within 7, 14 or 21 days after a validated AI event. Number of records excluded: Test date within 7 days of validated AI (n=22,338); test date within 14 days of validated AI (n=48,573); and test date within 21 days of validated AI (n=81,581).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Kathryn Maree Tiplady
Name/title of Primary Supervisor:	Professor Dorian Garrick
In which chapter is the manuscript /published work:	Chapter Four

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Tiplady K.M., Trinh M.H., Davis S.R., Sherlock R.G., Spelman R.J., Garrick D.J. and Harris B.L. Pregnancy status predicted using milk mid-infrared spectra from dairy cattle. Journal of Dairy Science. 2022 Feb 16.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Kathryn Tiplady Digitally signed by Kathryn Tiplady Date: 2022.03.23 18:16:32 +13'00'
Date:	
Primary Supervisor's Signature:	Dorian Garrick
Date:	25-Mar-2022

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 5

The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle

Originally published as: Tiplady, K.M., Lopdell, T.J., Littlejohn, M.D., Garrick, D.J., 2020. The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle. *Journal of Animal Science and Biotechnology*, 11(1), pp.1-13. https://doi.org/10.1186/s40104-020-00445-2.

5.1 Abstract

Over the last 100 years, significant advances have been made in the characterisation of milk composition for dairy cattle improvement programs. Technological progress has enabled a shift from labour intensive, on-farm collection and processing of samples that assess yield and fat levels in milk, to large-scale processing of samples through centralised laboratories, with the scope extended to include quantification of other traits. Fourier-transform mid-infrared (FT-MIR) spectroscopy has had a significant role in the transformation of milk composition phenotyping, with spectral-based predictions of major milk components already being widely used in milk payment and animal evaluation systems globally. Increasingly, there is interest in analysing the individual FT-MIR wavenumbers, and in utilising the FT-MIR data to predict other novel traits of importance to breeding programs. This includes traits related to the nutritional value of milk, the processability of milk into products such as cheese, and traits relevant to animal health and the environment. The ability to successfully incorporate these traits into breeding programs is dependent on the heritability of the FT-MIR predicted traits, and the genetic correlations between the FT-MIR predicted and actual trait values. Linking FT-MIR predicted traits to the underlying mutations responsible for their variation can be difficult because the phenotypic expression of these traits are a function of a diverse range of molecular and biological mechanisms that can obscure their genetic basis. The individual FT-MIR wavenumbers give insights into the chemical composition of milk and provide an additional layer of granularity that may assist with establishing causal links between the genome and observed phenotypes. Additionally, there are other molecular phenotypes such as those related to the metabolome, chromatin accessibility, and RNA editing that could improve our understanding of the underlying biological systems controlling traits of interest. Here we review topics of importance to phenotyping and genetic applications of FT-MIR spectral datasets, and discuss opportunities for consolidating FT-MIR datasets with other genomic and molecular data sources to improve future dairy cattle breeding programs.

Key words: *Bovine milk, cattle breeding genetics, Fourier-transform mid-infrared spectroscopy, trait prediction*

5.2 Introduction

Characterisation of milk composition in dairy cattle has a long history of scientific and commercial interest, with many countries establishing formal milk testing programs by the early 1900's (Bayly, 2009; Miglior et al., 2017). Initial selection targets in these programs were yields of milk or fat, which were measured on a small scale from samples taken manually on farm. Over the course of the 20th century, advances in refrigeration and transportation technologies, and the availability of automated on-farm milk meters, resulted in a shift to large-scale collection of samples, processed through centralised laboratories, with the scope extended to include quantification of traits such as protein yield and somatic cell counts. More recently, advances in analytical techniques have led to the widespread use of Fourier-transform mid-infrared (FT-MIR) spectroscopy for phenotyping major milk composition traits for dairy improvement programs.

Fourier-transform mid-infrared spectroscopy uses light from the mid-infrared region to scan milk samples and determine the presence of specific chemical bonds. Results are presented as an absorption profile, consisting of the absorbance values for individual infrared light wavenumbers across the mid-infrared region. Traits are predicted as a function of the individual FT-MIR wavenumber absorbance values, enabling rapid, high-throughput phenotyping of milk traits such as fat and protein yields, at a fraction of the cost of estimating the components using other methods. Increasingly, there is interest in analysing the individual FT-MIR wavenumbers, and in utilising FT-MIR data to predict other novel traits of interest to the industry, because the spectra are already available as a by-product of routine milk testing. Many of these traits are relevant to consumer expectations and concerns about the nutritional quality of milk, and the impact of dairy production systems on animal health and the environment; and are also relevant to farmers as they seek to improve farming systems and select cows based on their productivity, reproductive performance and disease resistance.

Successful phenotyping using FT-MIR data is dependent on the magnitude of the phenotypic correlation between the predicted trait and the trait as measured by a benchmarked standard reference method. The successful incorporation of an FT-MIR predicted trait into a breeding program is further dependent on the heritability of the spectral-based predictions and on the genetic correlation between the spectral-based predictions and the trait as measured by the benchmarked standard (De Marchi et al., 2014; Gengler et al., 2016). Improving our understanding of the genetics underlying the expression of FT-MIR predicted traits of interest is thus highly valuable. Conducting a genome-wide association study (GWAS) is a widely-used practice for identifying

genomic regions that are influencing expression of complex traits, such as those predicted from FT-MIR data. However, linking complex traits, such as those predicted from FT-MIR spectra to specific genetic mechanisms is complex, as the phenotypic expression of traits are a function of a diverse range of molecular and biological mechanisms (Te Pas et al., 2017) that can obscure the underlying causal links between genotypes and phenotypes. These mechanisms may be characterised as a set of intermediate omics measures, including sugars, lipids and amino acids in the metabolome, proteins in the proteome, RNA molecules in the transcriptome and DNA in the genome, all of which interact with environmental factors to ultimately determine what is observed at the phenotypic level (Fig. 5.1).

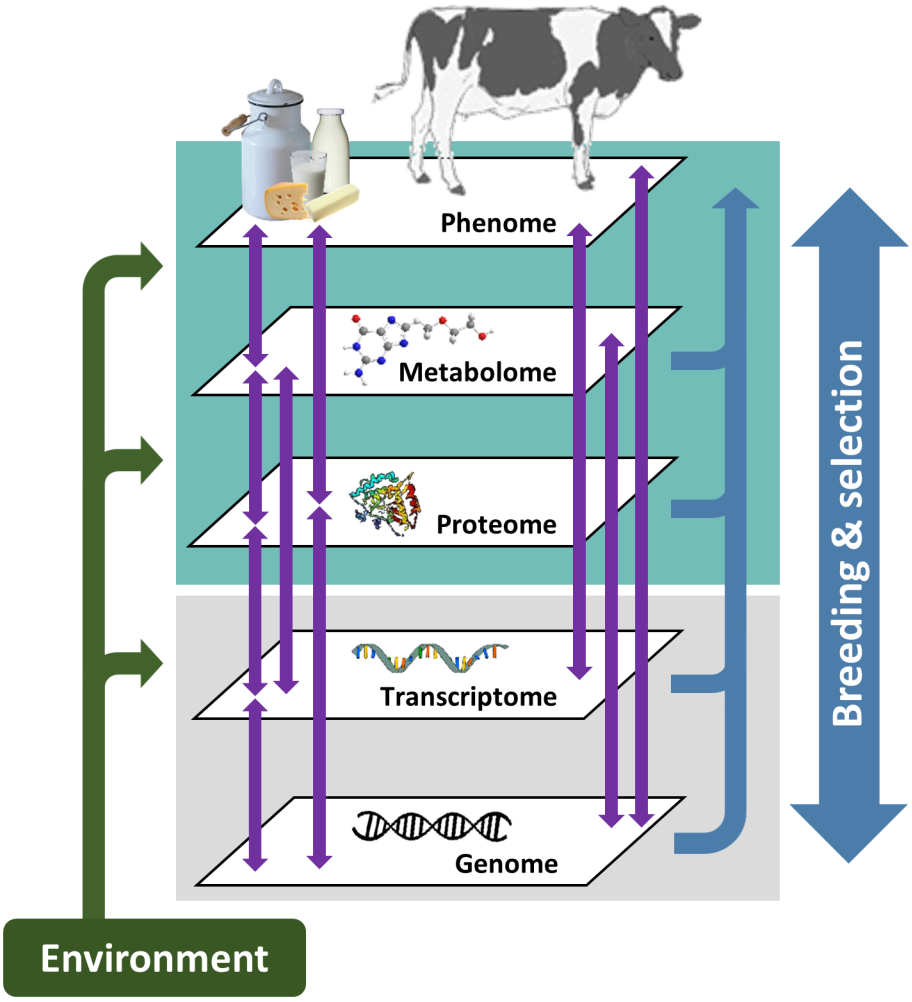

Figure 5.1: Characterisation of the relationships between molecular and biological mechanisms underlying phenotypic trait expression

Establishing causal links between the genome and observed phenotypes may be assisted by employing the individual FT-MIR wavenumbers, and other molecular phenotypes such as those related to the metabolome, chromatin accessibility, transcript levels, and RNA editing. Here we review the shifting role of FT-MIR datasets in dairy cattle improvement as we seek to predict new traits of importance to milk payment and animal evaluation systems. We discuss the broader topics of improving FT-MIR data quality and prediction model accuracy in phenotyping applications; and review existing studies of the genetics of FT-MIR predicted traits and individual FT-MIR wavenumbers. We also discuss opportunities for consolidating FT-MIR spectral datasets with other genomic and molecular data sources, to improve our knowledge of the genetic mechanisms of milk composition and enhance future dairy improvement programs.

5.3 Phenotyping applications of FT-MIR spectra

Fourier-transform mid-infrared spectroscopy uses infrared light to scan a milk sample and determine the presence of specific chemical bonds. As the light passes through the sample, it interacts with the molecules present, causing vibrations and rotational changes in the molecular bonds, resulting in absorption of some of the light. The light absorption is typically represented as an absorption spectrum, consisting of the absorbance values for individual infrared light wavenumbers across the mid-infrared range. Traits of interest are subsequently predicted as a function of the individual FT-MIR wavenumber absorbance values. Utilising FT-MIR data for the prediction of milk composition and other novel traits has been widely studied and recently reviewed (De Marchi et al., 2014, 2018; Egger-Danner et al., 2015; Gengler et al., 2016). Other notable FT-MIR research includes studies of individual fatty acids and milk proteins (Bonfatti et al., 2017d; Lopez-Villalobos et al., 2014), and studies of milk properties related to manufacturing, especially coagulation and other cheese-making properties (Toffanin et al., 2015; Visentin et al., 2015, 2018). Further studies have focussed on traits not directly measurable in milk, including those related to pregnancy (Lainé et al., 2017; Toledo-Alvarado et al., 2018a), energy status (Luke et al., 2019b; McParland et al., 2015), nitrogen outputs (Oliveira et al., 2012) and methane emissions (Bittante and Cipolat-Gotet, 2018; van Gastelen et al., 2018b; Vanlierde et al., 2018). Such applications demonstrate that FT-MIR spectra can be used to predict a wide range of traits, including highly topical traits that are important to animal welfare and the environment. Whilst prediction accuracy is variable across these applications, a number of key principles and findings have been reported for improving spectral data quality and model prediction accuracy.

5.4 FT-MIR data quality and prediction model accuracy

Trait prediction using FT-MIR spectra requires development of a calibration model, typically using a modest set of samples that have corresponding trait values, measured by a benchmarked technique. The most widely-used method for developing calibration models from FT-MIR spectra has been partial least squares (PLS) regression. Fewer studies have employed Bayesian methods to develop calibration models, but no consensus has been attained as to which methodology is best at providing prediction accuracy (El Jabri et al., 2019; Ferragina et al., 2015; Toledo-Alvarado et al., 2018a). This is likely due to the unique set of characteristics of each dataset, and indicates that it is advisable to assess a number of different modelling approaches for any given study. Once a calibration model is developed, the trait of interest can be estimated for any existing spectral absorbance data, or any future milk sample where the FT-MIR spectral data is available. The performance of an FT-MIR calibration model is assessed by how well the model predicts the benchmarked trait measurements in an independent dataset, or within the development dataset using a cross-validation framework. The utility and accuracy of trait predictions from FT-MIR spectra can often be improved by increasing the number of observations used to develop the calibration equation, and by ensuring that a similar extent of the variation in the prediction population is represented in the calibration dataset (McParland et al., 2011, 2012; Rutten et al., 2009; Soyeurt et al., 2011). Prediction accuracy may also be improved by modifying the scale of the trait. For example, higher prediction accuracies have been reported when evaluating fatty acids as a percentage of total milk volume, compared to as a percentage of total fat content (Bonfatti et al., 2016; Rutten et al., 2009; Soyeurt et al., 2006). Similar considerations are important for studies of the concentrations of individual casein and whey proteins (Bonfatti et al., 2011, 2016; De Marchi et al., 2009a; McDermott et al., 2016; Rutten et al., 2011); and in studies related to cheese-making efficiency (Bonfatti et al., 2016; Dal Zotto et al., 2008; De Marchi et al., 2009b, 2013). Other considerations that influence prediction accuracy include: pre-processing treatments to address scaling and baseline effects in spectral data; appropriate management of outliers; low repeatability of sample measurement for specific regions of the infrared spectrum affected by the water content in milk; and managing systematic instrument variation due to factors such as temperature fluctuations and wavelength or detector intensity instability (Wang et al., 1991).

5.4.1 Pre-processing

Pre-processing treatments are commonly applied to FT-MIR spectra before generating a calibration model. The objective of pre-processing is to retain important discriminatory features of the spectra, but address baseline and scaling effects caused by light scattering, that can erode prediction accuracy. Baseline effects are additive and represent a baseline offset in the spectral response, whereas scaling effects are multiplicative and scale the spectral results by a given factor. One common group of methods for pre-processing are the multiplicative scatter methods (Geladi et al., 1985; Martens et al., 2003). Multiplicative scatter correction is a normalization method that corrects spectra for scaling and baseline effects by comparing each spectrum to an expected spectral profile. Another family of techniques are the derivation methods, such as the Savitzky-Golay derivative (Savitzky and Golay, 1964). Derivation methods are based on changes in the spectrum across specified window sizes, and are intended to smooth the spectrum whilst retaining key features of its shape.

Overall, there is no consensus about the best pre-processing treatment to apply to FT-MIR spectra. For example, some studies report that pre-processing spectra provides no significant gains to model prediction accuracy (De Marchi et al., 2009b, 2013), whilst others observe better predictions after pre-processing (De Marchi et al., 2011; Soyeurt et al., 2011), and several studies report mixed results (Bonfatti et al., 2011; Rutten et al., 2011). This is likely because of the unique characteristics of each dataset, indicating that in the development of a new calibration, it is advisable to compare a number of approaches to determine their effectiveness. Notably, even when different pre-processing strategies are examined in a study, authors often only report the best model, making it difficult to compare the effectiveness of other pre-processing strategies (De Marchi et al., 2014).

5.4.2 Outliers and removal of low signal-to-noise regions of the mid-infrared spectrum

Outliers in FT-MIR datasets are often identified using a squared Mahalanobis distance (MD) metric, where the MD is a multivariate indicator of the distance between a spectral record and the average spectral response. Many studies are based on spectra from a single instrument, and are therefore not required to account for the different variance-covariance structures of measurements from different instruments. In a study of spectra from 66 instruments, Grelet et al. (2015) showed considerable variability in the spectral responses of the instruments, while we have also observed that the distribution of MD values can be heterogeneous across instruments (Tiplady et al., 2019).

These results highlight the need to apply MD thresholds within instrument for the purpose of outlier removal.

Bands of the infrared spectrum with low repeatability of sample measurement due to the water content in milk are typically reported in the O-H bending (~1,600 to 1,700 cm^{-1}) and O-H stretching bands (>~3,000 cm^{-1}). These regions have low signal-to-noise ratios, with varying boundaries reported across publications: 1,600 to 1,700 cm^{-1} and 3,040 to 3,470 cm^{-1} (Bonfatti et al., 2011); 1,586 to 1,698 cm^{-1} and 3,052 to 3,669 cm^{-1} (Bittante and Cecchinato, 2013); 1,600 to 1,689 cm^{-1} and 3,008 to 5,010 cm^{-1} (Grelet et al., 2015). Although it is common practice to remove spectra from low signal-to-noise ratio regions, some studies indicate that there may be wavenumbers within these regions that carry valuable information. For example, Wang et al. (2016) and Wang and Bovenhuis (2018) identified wavenumbers in these regions that are affected by a polymorphism in the *DGAT1* gene, and Toledo-Alvarado et al. (2018a) identified a significant association between the 3,683 cm^{-1} wavenumber and pregnancy status. More generally, Bittante and Cecchinato (2013) showed that the transmittance of individual spectra wavenumbers had moderate to high heritability across most of the mid-infrared region and highlighted that absorbance peaks for non-water milk components were present in low signal-to-noise ratio regions and should be considered for investigation. The findings of these studies indicate that a prudent approach to removal of wavenumbers in low signal-to-noise ratio regions should be taken, retaining spectra from all regions in applications where the wavenumbers are considered independently, but removing them in applications where wavenumbers are considered in a multivariate manner (Tiplady et al., 2019).

5.4.3 Managing systematic instrument variation

The instrument calibration approach outlined by Lynch et al. (2006) has been widely used to standardize instrument predictions for major milk composition traits and reduce the impact of systematic variation between and within instruments across time. With this approach, a small set of reference samples are analysed through the instrument, where the reference samples have also been measured for traits of interest using benchmarked standards, such as the Rose-Gottlieb method for fat determination and the Kjeldahl method for protein determination. For these samples, unadjusted trait predictions are made from the spectral data, and instrument-specific correction coefficients are evaluated by comparing the unadjusted predictions to the measured trait values according to the benchmarked standard. A limitation of this approach is that it can only be used to adjust predictions for traits with pre-evaluated correction coefficients. More recent standardisation strategies have instead proposed calibrating the individual wavenumbers (Bonfatti

et al., 2017a; Grelet et al., 2015, 2017; Tiplady et al., 2019), allowing the correction of any trait predicted as a function of the spectral wavenumbers. Studies have shown that standardising individual wavenumbers can effectively reduce prediction errors when transferring calibration models between instruments for fat composition traits (Bonfatti et al., 2017a; Grelet et al., 2015), as well as for calibration models for traits that are more difficult to predict reliably such as methane emissions and cheese yield (Grelet et al., 2017). Tiplady et al. (2019) showed that the most consistent standardisation approach for reducing prediction errors relies on analysing identical reference samples across all instruments, as outlined by Grelet et al. (2015). Ideally, global reference sample sharing would be established, facilitating standardisation across instruments in different countries. That would enable the consolidation of spectral data collected on different instruments, and improve accuracy when applying calibration models developed on one instrument to spectral data collected on other instruments. Global reference sample sharing, however, is reliant on resolving issues related to sample preservation, and on adherence to the bio-security legislation of different countries. Instrument manufacturers such as FOSS (Hillerød, Denmark) and Bentley (Chaska, MN) have started to offer alternative standardisation procedures. The FOSS procedure uses a liquid equaliser with a known spectral response to adjust spectral results (Winning et al., 2014), whereas the Bentley procedure uses a polystyrene film to adjust for interferometer laser frequency shifts across time (Gupta et al., 1995), and infrared flow cell information to adjust for shifts in absorbance measurement (Parsons and Lyder, 2018). While these within-instrument standardisation procedures offer promise for automatic spectral standardisation, there have been no independent studies to validate their effectiveness for standardisation of milk samples collected across or within networks.

5.5 The genetics of FT-MIR predicted traits

Predictions of major milk composition traits from FT-MIR spectra are already widely incorporated into dairy improvement programs. Other FT-MIR predicted traits that could be of interest to industry improvement programs include milk fatty acids and protein fractions, and traits that form proxy indicators for milk processability properties, and animal health and environmental outcomes. The accuracy of FT-MIR predictions is an important indicator of their utility, but for breeding purposes, the critical parameters are the extent of genetic variation in the benchmarked trait, the heritability of the FT-MIR predictions, and the genetic correlations between the FT-MIR predictions and the benchmarked trait.

5.5.1 Milk fatty acid and protein composition traits

Heritability estimates for FT-MIR predicted individual and grouped fatty acids, and their genetic correlations with gas chromatography (GC) based measurements are shown in Table 5.1. Where available, standard errors are shown in brackets. For individual milk fatty acids, heritability estimates ranged from 0.05 to 0.54 (Lopez-Villalobos et al., 2014; Rutten et al., 2010; Soyeurt et al., 2007b). Heritability estimates for grouped fatty acids ranged from 0.11 to 0.51 (Fleming et al., 2018; Hein et al., 2018; Lopez-Villalobos et al., 2014; Narayana et al., 2017), with the lowest heritability estimates reported by Hein et al. (2018). In the studies by Fleming et al. (2018) and Narayana et al. (2017), heritability estimates were consistently higher for saturated fat and short- and medium-chain fatty acid groups, compared to unsaturated fat and long-chain fatty acid groups. Rutten et al. (2010) was the only study to report genetic correlations between FT-MIR predicted and GC-based fatty acids. These genetic correlations were high, ranging from 0.82 to 0.99.

Fewer studies exist of the genetic parameter estimates of FT-MIR predicted individual milk proteins. Sanchez et al. (2017a) reported moderate to high heritability estimates (0.25 to 0.72) for a number of FT-MIR predicted milk protein contents/fractions (not shown), with especially high estimates for β-lactoglobulin (0.61 to 0.86). Moderate heritability estimates for FT-MIR predicted lactoferrin, ranging from 0.16 to 0.22 have also been reported (Arnould et al., 2009b; Lopez-Villalobos et al., 2009; Soyeurt et al., 2007a). These studies quantify the useful extent of genetic variation in FT-MIR predicted fatty acids and individual milk proteins, and suggest that these predicted traits could be incorporated into cattle improvement programs to change the fatty acid profile and the protein composition of bovine milk.

5.5.2 Milk processability traits

Heritability estimates and genetic correlations between measured and FT-MIR predicted milk processability traits are shown in Table 5.2. Where available, standard errors are shown in brackets. For coagulation traits, heritability estimates ranged from 0.16 to 0.43 (Cecchinato et al., 2009; Costa et al., 2019; Visentin et al., 2017). Cecchinato et al. (2009) was the only study reporting genetic correlations between FT-MIR predicted and measured coagulation traits (not shown). Those ranged from 0.91 to 0.96 for rennet coagulation time (RCT), and from 0.71 to 0.87 for curd firmness after 30 minutes (a_{30}). Heritability estimates for FT-MIR predicted minerals ranged from 0.32 to 0.56, with phosphorus having the highest estimated heritability, and sodium having the lowest estimated heritability in both studies presented (Costa et al., 2019; Sanchez et al., 2018). Heritability estimates for nutrient recovery traits were typically higher than for

cheese yield traits (Bittante et al., 2014; Cecchinato et al., 2015). Bittante et al. (2014) was the only study reporting genetic correlations between FT-MIR predicted and measured cheese yield and nutrient recovery traits. These ranged from 0.76 to 0.98 for cheese yield traits, and from 0.79 to 0.98 for nutrient recovery traits. Overall, these studies show that many FT-MIR predicted processability traits are heritable, and that sufficient variation exists to use FT-MIR predicted traits to change milk processing and cheese-making characteristics in cattle improvement programs.

5.5.3 Animal health traits

Health and fertility traits are valuable targets for breeding programs and selection for these traits would be considerably enhanced if they could be reliably predicted from FT-MIR spectra. A recent review by Bastin et al. (2016) across a wide range of FT-MIR predicted traits related to fertility, mastitis, ketosis and other disease traits highlighted that more research is required to understand the relationships between health and fertility indicators and FT-MIR predicted traits, and to estimate the genetic parameters of these traits. Since then, Belay et al. (2017) have reported moderate heritability estimates for FT-MIR predicted blood β-hydroxybutyrate (BHB), ranging from 0.25 to 0.37 across different stages of lactation, and moderate genetic correlations between clinical ketosis and the FT-MIR predicted BHB (0.47). More research is required in this area to realise the value that FT-MIR spectra might add to animal health breeding goals.

5.5.4 Environment traits

Despite increasing interest in FT-MIR predictions of environmental traits related to methane (CH_4) and nitrogen outputs from dairy systems, there have been few reports of the genetic parameter estimates of these FT-MIR predicted traits, or of the genetic correlations between measured and FT-MIR predicted trait values. Kandel et al. (2017) report moderate heritability estimates, ranging from 0.22 to 0.25 for predicted daily CH_4 emissions and 0.17 to 0.18 for log-transformed predicted CH_4 intensity. There is, therefore, some potential for the future incorporation of FT-MIR predicted methane traits into breeding programs. However, there are still issues to be resolved to address uncertainties and discrepancies in methane datasets and measurement methods, and to improve the accuracy and robustness of prediction equations to make them applicable across a broader range of production systems and environments (van Gastelen et al., 2018b; Hristov et al., 2018; Negussie et al., 2017; Vanlierde et al., 2018).

Table 5.1: Heritability estimates for FT-MIR predicted fatty acids (h^2), and their genetic correlations (r_a) with GC-based[1] fatty acids

Individual fatty acids[2]	Lopez-Villalobos et al. (2014)	Soyeurt et al. (2007b)	Rutten et al. (2010)	
	h^2 (SE)	h^2 (SE)	h^2 (SE)	r_a (SE)
C4:0	0.38 (0.03)	–	0.42 (0.09)	0.94 (0.03)
C6:0	0.32 (0.03)	–	0.35 (0.09)	0.97 (0.02)
C8:0	0.29 (0.03)	–	0.38 (0.09)	0.99 (0.01)
C10:0	0.17 (0.02)	–	0.46 (0.10)	0.98 (0.01)
C10:1	0.30 (0.02)	–	–	–
C12:0	0.16 (0.02)	0.29 (0.02)	0.54 (0.11)	0.97 (0.02)
C12:1	0.41 (0.03)	–	–	–
C14:0	0.19 (0.02)	0.31 (0.03)	0.50 (0.10)	0.99 (0.01)
C14:1	0.27 (0.01)	–	–	–
C15:0	0.22 (0.02)	–	–	–
C16:0	0.29 (0.02)	0.38 (0.02)	0.30 (0.09)	0.86 (0.07)
C16:1	0.30 (0.02)	–	–	–
C17:0	0.41 (0.03)	–	–	–
C17:1	0.14 (0.02)	–	–	–
C18:0	0.26 (0.02)	0.30 (0.02)	0.52 (0.10)	0.82 (0.08)
C18:1	0.43 (0.03)	0.05 (0.01)	–	–
C18:1 *cis*-9	0.22 (0.02)	–	0.25 (0.08)	0.93 (0.05)
C18:1 *trans*-11	0.27 (0.03)	–	–	–
C18:2 *cis*-9,*cis*-12	0.45 (0.03)	0.20 (0.02)	–	–
C18:2 *cis*-9,*trans*-11	0.41 (0.03)	–	–	–
C20:0	0.38 (0.03)	–	–	–
C20:1 *cis*-11	0.37 (0.03)	–	–	–
C22:0	0.35 (0.03)	–	–	–

Grouped fatty acids[3]	Lopez-Villalobos et al. (2014)	Hein et al. (2007b)	Fleming et al. (2007b)	Narayana et al. (2007b)
	h^2 (SE)	h^2 (SE)	h^2 (SE)	r_a (SE)
SCFA	0.39 (0.03)	0.16	0.42	0.24
MCFA	0.30 (0.03)	0.12	0.50	0.32
LCFA	0.50 (0.03)	0.11	0.26	0.23
SFA	0.46 (0.03)	0.15	0.51	0.33
UFA	0.48 (0.03)	–	0.26	0.21
PUFA	0.42 (0.03)	–	–	–

[1] GC-based: Gas chromotography based.

[2] Individual fatty acids: All fatty acids expressed as a % of the total fatty acids.

[3] SCFA = Short-chain fatty acids; MCFA = Medium-chain fatty acids; LCFA = Long-chain fatty acids; SFA = Saturated fatty acids; UFA = Unsaturated fatty acids; PUFA = Polyunsaturated fatty acids.

Table 5.2: Heritability estimates for FT-MIR predicted milk processability traits (h^2), and their genetic correlations (r_a) with measured traits

Trait[1]	Visentin et al. (2017)	Cecchinato et al. (2009)	Costa et al. (2019)	Sanchez et al. (2018)
	h^2 (SE)	h^2 range[2] (SE)	h^2 (SE)	h^2 (SE)
Coagulation traits				
RCT, min	0.28 (0.01)	0.30–0.34 (0.08)	0.35 (0.05)	–
k_{20}, min	0.43 (0.02)	–	0.43 (0.03)	–
a_{30}, mm	0.36 (0.02)	0.22-0.27 (0.07)	0.39 (0.03)	–
a_{60}, mm	0.27 (0.01)	–	–	–
HCT, min	0.16 (0.01)	–	–	–
CMS, nm	0.31 (0.02)	–	–	–
Acidity				
pH, units	0.27 (0.01)	–	–	0.37 (0.01)
Minerals, mg/kg milk				
Calcium	–	–	0.45 (0.02)	0.50 (0.01)
Phosphorus	–	–	0.53 (0.03)	0.56 (0.01)
Magnesium	–	–	0.47 (0.03)	0.52 (0.01)
Potassium	–	–	0.45 (0.03)	0.53 (0.01)
Sodium	–	–	0.38 (0.03)	0.32 (0.01)

	Sanchez et al. (2018)	Bittante et al. (2014)		Cecchinato et al. (2015)
	h^2 (SE)	h^2 (SE)	r_a	h^2 range[3]
Cheese yield, %				
CY_{CURD}	0.38 (0.01)	0.21 (0.09)	0.97	0.18–0.33
CY_{SOLIDS}	0.39 (0.01)	0.22 (0.08)	0.98	0.18–0.28
CY_{WATER}	–	0.18 (0.05)	0.76	0.14–0.29
Nutrient recovery, %				
$REC_{PROTEIN}$	–	0.44 (0.09)	0.88	0.32–0.41
REC_{FAT}	–	0.28 (0.07)	0.79	0.15–0.33
REC_{ENERGY}	–	0.21 (0.07)	0.96	0.19–0.30
REC_{SOLIDS}	–	0.24 (0.08)	0.98	0.17–0.29

[1] RCT = Rennet coagulation time; k_{20} = curd-firming time; a_{30} = curd firmness after 30 min; a_{60} = curd firmness after 60 min; HCT = heat coagulation time; CMS = casein micelle size; CY = cheese yield: weight of fresh curd, curd solids, and curd as a percentage of weight of milk processed; REC = nutrient recovery: protein, fat, energy and solids of the curd as a percentage of the protein, fat, energy and solids of the milk processed.

[2] Range of estimates from 4 subsets of data used to validate calibration equations.

[3] Range of estimates from 3 different breeds.

Milk urea nitrogen (MUN) concentrations are routinely predicted using FT-MIR spectroscopy (Gengler et al., 2016), however, there are few studies of the genetic parameters of FT-MIR predicted MUN and its relationship with other production traits (Miglior et al., 2007). Amongst those studies, moderate to high heritability estimates, ranging from 0.38 to 0.59 were reported by Miglior et al. (2007) and Wood et al. (2003), with lower estimates of 0.22 and 0.14 reported in studies by Mitchell et al. (2005) and Stoop et al. (2007), respectively. Mitchell et al. (2005) was the only study reporting genetic correlations between wet-chemistry direct measurements of MUN and FT-MIR predicted MUN, which were 0.38 and 0.23 in lactations 1 and 2, respectively. These genetic correlations are significantly lower than those reported for fatty acids (0.82 to 0.99; Rutten et al., 2010) and milk processability traits (0.76 to 0.98; Bittante et al., 2014), and indicate that wet-chemistry measurements of MUN and FT-MIR predicted MUN are genetically different traits. Large differences in heritability estimates across studies of FT-MIR predicted MUN indicate that there may be underlying instability in prediction equations. This highlights the importance of developing prediction models that are robust across different breeds and production systems. Research is ongoing to determine the role that FT-MIR predicted MUN could have in reducing nitrogen outputs from dairy systems.

5.6 The genetics of individual FT-MIR wavenumbers

In contrast to the prevalence of studies reporting genetic parameter estimates of FT-MIR predicted traits, there are relatively few studies reporting genetic parameter estimates for the individual spectra wavenumbers. Nevertheless, the transmittance of FT-MIR spectral wavenumbers is moderately to highly heritable across a large proportion of the mid-infrared region (Bittante and Cecchinato, 2013; Rovere et al., 2019; Soyeurt et al., 2010; Wang et al., 2016; Zaalberg et al., 2019). Although heritability estimates were consistently low in water absorption regions across all studies, estimates greater than 0.2 were reported across most of the mid-infrared region in studies by Soyeurt et al. (2010) and Wang et al. (2016). This indicates that genetic gain may be obtained by directly selecting on a linear function of estimated breeding values (EBV) for individual FT-MIR wavenumbers; rather than indirect selection as currently practised on EBV of composite indicator traits which are linear functions of individual FT-MIR wavenumber absorbance values. Recent studies have confirmed this, showing that the accuracies of breeding value predictions estimated directly from FT-MIR spectra can be higher than for breeding value predictions estimated indirectly from the FT-MIR predicted composite traits (Belay et al., 2018; Bonfatti et al., 2017c; Dagnachew et al., 2013). Estimating breeding values directly from FT-MIR spectra requires that spectral data is routinely stored, rather than just the spectral based predictions of milk components, and that has not historically been the case in most dairy nations.

5.7 GWAS of individual FT-MIR wavenumbers

Many GWAS have been published in the last decade for FT-MIR predicted major milk production traits (Jiang et al., 2010; Kemper et al., 2015b; Littlejohn et al., 2016; Lopdell et al., 2017; Raven et al., 2014), and for fatty acid and protein fractions (Bouwman et al., 2011; Buitenhuis et al., 2014, 2016; Li et al., 2014; Sanchez et al., 2016). However, only two studies report GWAS results for individual FT-MIR wavenumbers. In a study of 1,748 Dutch Holsteins across 50,688 SNP, Wang and Bovenhuis (2018) conducted a GWAS on a subset of 50 wavenumbers, selected using a clustering approach to capture more than 95% of the phenotypic variation. In that study, significant associations between individual wavenumbers and over 20 genomic regions were identified. While most of these genomic regions had already been reported for having significant associations with other milk production traits, three new regions were identified. In a larger study of 5,202 Holstein, Jersey and crossbred cows across 626,777 SNP, Benedet et al. (2019) used a PLS approach to associate genotypes to spectral data, and showed that FT-MIR spectra could be used to increase the power of a GWAS, and assist with distinguishing milk composition QTL. The studies by Wang and Bovenhuis (2018) and Benedet et al. (2019) both demonstrate that there are genetic signals in the individual FT-MIR wavenumbers that we do not observe in the currently-used portfolio of composite FT-MIR predicted traits. This confirms that the individual FT-MIR wavenumbers can provide an additional layer of granularity to assist with establishing causal links between the genome and observed phenotypes. Notably, both studies use relatively low numbers of animals compared to recent GWAS published for other traits, and applying these methodologies to larger datasets, with higher genotype densities, promises to increase the power of these approaches. This should enable the discovery of QTL with smaller effect sizes in addition to novel QTL characterised by lower minor allele frequencies than those QTL discovered with datasets numbering only thousands of animals.

5.7.1 Computational challenges

Over the last two decades, the scope of genomic resources available for GWAS has increased, both in terms of the number of genotyped individuals, and in terms of variant density. Developing strategies for managing GWAS on large numbers of densely genotyped individuals is an active area of research, as we look to generate new, more efficient algorithms that will enable the processing of these datasets within acceptable timeframes and computational limits of RAM and CPU. The importance of efficient algorithms is further highlighted when we conduct GWAS across large numbers of FT-MIR predicted traits and the individual FT-MIR wavenumbers.

Existing mixed-linear model-based methods for conducting GWAS, such as GCTA-MLMA (Yang et al., 2011) primarily run in $O(mn^2)$ or $O(m^2n)$ time per trait, where m is the number of variants and n is the number of animals. These models become prohibitively slow as the numbers of genotyped individuals and variants increase (Loh et al., 2015). The ever-increasing cohort sizes of densely-genotyped individuals frequently requires subsampling to use these methods within acceptable computation constraints. This has spurred the development of faster, more memory-efficient algorithms and software. One software package, Bolt-LMM (Loh et al., 2015, 2018) runs in approximately $O(mn^{1.5})$ time; however, it makes assumptions that are only valid for larger sample sizes. Recent versions are capable of running the entire UK biobank data set (n=459k) in a few days on a single computational node (Loh et al., 2018). Another algorithm, fastGWA (Jiang et al., 2019c), available as a recent enhancement of the GCTA software package, provides further reductions in algorithmic complexity, running in approximately O(mn) time. These improvements mean that it is capable of running n=400k UK biobank samples in around 20 minutes, compared to 22 hours for BOLT-LMM on the same hardware. Developments such as this make GWAS across sizeable populations with large numbers of FT-MIR phenotypes feasible.

5.8 Consolidating FT-MIR spectra with other omics data sources for QTL mapping

After conducting a GWAS, it is useful to identify the candidate genes and mutations underlying genomic loci with signal for a trait of interest. This can aid marker-assisted selection and improve our understanding of the biological pathways regulating the trait. Moreover, it has been shown that genomic prediction can be improved by including variants close to the causative mutations (van den Berg et al., 2016). Software such as Ensembl's Variant Effect Predictor (McLaren et al., 2016) is commonly used to identify candidate causal variants that have protein-coding or loss-of-function effects, with the expectation that these variants are more likely to impact the trait than other variants. However, recent studies in both humans (Lee et al., 2015; Maurano et al., 2012) and dairy cattle (Lopdell et al., 2017; Pausch et al., 2016) have highlighted the prevalence of QTL underpinned by expression-based mechanisms, and demonstrate that the majority of variance for at least some traits can be explained by non-coding variants located in regulatory elements. These variants are typically identified by considering the expression levels of genes as phenotypes, and using these data for genetic mapping studies in an approach known as expression QTL (eQTL) analysis.

5.8.1 Expression-based phenotypes

Assuming a causality chain hypothesis, as illustrated in Fig. 5.1, observation of an eQTL, co-located with a QTL for an FT-MIR predicted trait can inform on the mechanism of the trait of interest. This methodology can also be used to identify mechanisms underlying QTL observed for individual FT-MIR wavenumbers. A strong correlation between the variant effects for the two QTL (expression and FT-MIR related) suggest a shared underlying genetic architecture regulating both, while a weak correlation suggests that the two QTL, though co-located, do not co-segregate, and therefore represent distinct genetic signals with different causal variants.

Similar to eQTL analysis, a range of additional omics data sources can be used for QTL mapping, and the resulting QTL could be applied to identify causative genes for FT-MIR predicted traits and individual FT-MIR wavenumbers. The factors yielding these omics data sources can occur before or after mRNA transcription. Factors acting before transcription can help unravel causative regulatory variants by highlighting actively-transcribed regions of the genome, and the variants that sit within them. One of these factors is chromatin accessibility. Transcriptionally active genes, as well as active regulatory elements (such as enhancers), are found in regions of open chromatin (euchromatin); whereas inactive regions of the genome are typically much more densely compacted into a structure known as heterochromatin. Genome features found in euchromatin are therefore more accessible to transcription factors and other factors involved in gene expression, and so are more likely to influence traits of interest compared to factors located in inactive regions. Methods to assay chromatin accessibility include ChIP-seq (O'Neill and Turner, 2003), DNase-seq (Boyle et al., 2008), and ATAC-seq (Buenrostro et al., 2015).

Other factors acting during or after transcription provide intermediate phenotypes that can aid in understanding the underlying biological control of these traits (Kemper et al., 2016). One such factor is RNA-editing, i.e., direct enzymatic conversion of bases within the mRNA transcripts, with the most common form of editing in vertebrates being the conversion of adenosine nucleotides into inosine (Savva et al., 2012). Biologically, RNA editing is involved in protection against dsRNA viruses (Liddicoat et al., 2015) and in adaptation to different environmental conditions (Garrett and Rosenthal, 2012), and therefore has potential relevance to variation in animal health and in providing for animal adaptability to changing environments. RNA-editing QTL (edQTL) were initially identified in Drosophila (Ramaswami et al., 2015), followed soon after by mice (Gu et al., 2016) and humans (Park et al., 2017). Recently, edQTL were reported for the first time within the bovine mammary gland (Lopdell et al., 2019a), and subsequently used to characterise candidate causative genes underlying a milk yield QTL at the *CSF2RB/NCF4* locus (Lopdell

et al., 2019b). That study highlighted the manner in which intermediate molecular phenotypes can be used to investigate mechanisms underlying FT-MIR predicted trait QTL, and exemplifies how other similarly novel molecular phenotypes can be applied.

5.8.2 Metabolomics

Absorbance levels for individual FT-MIR wavenumbers provide insights into the presence of particular chemical bonds in the sample and accordingly provide information as to the chemical composition of a milk sample. Analysing the chemical composition of a sample in more detail, using methodologies such as nuclear magnetic resonance (NMR) spectroscopy or mass spectroscopy (MS), yields the metabolome, i.e., a more complete set of all small molecules present in a tissue sample. Metabolomics can provide detailed information about enzymatic activity in the pathways that exist between gene expression and FT-MIR predicted traits, providing a near-terminal link in the chain of causality. For example, rumen volatile fatty acid (VFA) levels can provide information on measuring and controlling methane production (Knapp et al., 2014). Levels of VFAs in the rumen could therefore provide a proxy measurement for methane production. Identifying QTL that underlie variation in the concentrations of these metabolites could complement genetic signals identified using FT-MIR wavenumbers and FT-MIR based methane trait predictions, and facilitate selection of low-methane emitting animals.

5.9 Conclusions

Over the last 100 years, milk composition phenotyping for dairy cattle has evolved from manual on-farm methods for determining yield and fat levels in milk, to high-tech analysis at centralised laboratories, with many novel FT-MIR predicted traits now being considered for incorporation into improvement programs. Multiple studies have demonstrated that the accuracy of FT-MIR predictions are strongly influenced by how well the variation in the prediction population is represented in the calibration population. Trait prediction accuracy is also strongly affected by how well instrument-specific measurement differences are accounted for, particularly when transferring calibration equations developed on one instrument to spectra collected on other instruments. Utilising FT-MIR data to generate proxies for novel traits has grown in popularity, however, compared to FT-MIR predictions of major milk components, there are relatively few studies of the genetics of other FT-MIR predicted traits, and even fewer of the genetics of the individual wavenumbers. This is despite the individual wavenumbers exhibiting additional genetic signal that is often not observed in FT-MIR predictions of major milk composition traits. Integrating results

from GWAS applied to FT-MIR predicted traits and GWAS applied to individual wavenumbers with other molecular datasets could improve our understanding of the underlying biological systems controlling traits of interest. However, integration of these data sources also brings computational challenges due to the size and complexity of the datasets involved. Resolving the challenges of effectively integrating FT-MIR datasets with other omics data sources will require a mix of both bioinformatics and molecular biology approaches. Successfully consolidating these approaches promises to improve our knowledge of milk composition and enable the future enhancement of animal breeding programs.

5.10 Declarations

5.10.1 Ethics approval and consent to participate

Not applicable

5.10.2 Acknowledgements

We acknowledge staff in the Research & Development group of Livestock Improvement Corporation (LIC; Hamilton, NZ) and fellow students based at the Massey University, AL Rae campus (Hamilton, NZ) for their support throughout the time of writing this manuscript. The authors would also like to thank the reviewers for their constructive feedback, which helped to significantly improve the manuscript.

5.10.3 Funding

This research was funded by Livestock Improvement Corporation (LIC) and the New Zealand Ministry for Primary Industries, through the Sustainable Food & Fibre Futures programme.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Kathryn Maree Tiplady
Name/title of Primary Supervisor:	Professor Dorian Garrick
In which chapter is the manuscript /published work:	Chapter Five

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Tiplady K.M., Lopdell T.J., Littlejohn M.D. and Garrick D.J., 2020. The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle. Journal of Animal Science and Biotechnology. 2020 Dec;11(1):1-3.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Kathryn Tiplady Digitally signed by Kathryn Tiplady Date: 2022.03.23 18:21:52 +13'00'
Date:	
Primary Supervisor's Signature:	Dorian Garrick
Date:	25-Mar-2022

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 6

Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle

Originally published as: Tiplady, K.M., Lopdell, T.J., Reynolds, E., Sherlock, R.G., Keehan, M., Johnson, T.J.J., Pryce, J.E., Davis, S.R., Spelman, R.J., Harris, B.L., Garrick, D.J., Littlejohn, M.D., 2020. Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. *Genetics Selection Evolution*, 53(1), pp.1-24. https://doi.org/10.1186/s12711-021-00648-9.

6.1 Abstract

6.1.1 Background

Fourier-transform mid-infrared (FT-MIR) spectroscopy provides a high-throughput and inexpensive method for predicting milk composition and other novel traits from milk samples. While there have been many genome-wide association studies (GWAS) conducted on FT-MIR predicted traits, there have been few GWAS for individual FT-MIR wavenumbers. Using imputed whole-genome sequence for 38,085 mixed-breed New Zealand dairy cattle, we conducted GWAS on 895 individual FT-MIR wavenumber phenotypes, and assessed the value of these direct phenotypes for identifying candidate causal genes and variants, and improving our understanding of the physico-chemical properties of milk.

6.1.2 Results

Separate GWAS conducted for each of 895 individual FT-MIR wavenumber phenotypes, identified 450 1-Mbp genomic regions with significant FT-MIR wavenumber QTL, compared to 246 1-Mbp genomic regions with QTL identified for FT-MIR predicted milk composition traits. Use of mammary RNA-seq data and gene annotation information identified 38 co-localized and co-segregating expression QTL (eQTL), and 31 protein-sequence mutations for FT-MIR wavenumber phenotypes, the latter including a null mutation in the *ABO* gene that has a potential role in changing milk oligosaccharide profiles. For the candidate causative genes implicated in these analyses, we examined the strength of association between relevant loci and each wavenumber across the mid-infrared spectrum. This revealed shared association patterns for groups of genomically-distant loci, highlighting clusters of loci linked through their biological roles in lactation and their presumed impacts on the chemical composition of milk.

6.1.3 Conclusions

This study demonstrates the utility of FT-MIR wavenumber phenotypes for improving our understanding of milk composition, presenting a larger number of QTL and putative causative genes and variants than found from FT-MIR predicted composition traits. Examining patterns of significance across the mid-infrared spectrum for loci of interest further highlighted commonalities of association, which likely reflects the physico-chemical properties of milk constituents.

6.2 Background

Fourier-transform mid-infrared (FT-MIR) spectroscopy is a high-throughput and inexpensive method for predicting milk composition. The FT-MIR methodology determines the presence of specific chemical bonds in milk by measuring the absorbance of infrared light as the light interacts with molecules in the sample. Data from FT-MIR spectroscopy comprises a spectrum of absorbance values across the mid-infrared range that are readily available through routine milk testing. This technology is widely used to estimate the concentrations of major milk components such as fat and protein for incorporation into milk payment and animal evaluation systems. Over the last decade, there has been increased interest in using FT-MIR data to predict other milk composition and novel traits. Applications of FT-MIR spectroscopy as a phenotyping tool have been widely studied and reviewed (De Marchi et al., 2014, 2018; Egger-Danner et al., 2015; Gengler et al., 2016). Recent research includes studies of milk composition traits that are relevant to manufacturing traits (Toffanin et al., 2015; Visentin et al., 2015, 2018), individual fatty acids and milk proteins (Bonfatti et al., 2017b; Sanchez et al., 2017a), and indirect traits that are related to energy status (Luke et al., 2019b; McParland et al., 2015), pregnancy and fertility (Ho et al., 2019; Lainé et al., 2017; Toledo-Alvarado et al., 2018a), methane emissions (Bittante and Cipolat-Gotet, 2018; van Gastelen et al., 2018b; Vanlierde et al., 2018) and bovine tuberculosis (Denholm et al., 2020).

Successful utilisation of FT-MIR data as a phenotyping tool depends on the strength of the phenotypic correlation between the predicted trait, and the trait as measured by a benchmarked standard; and successful incorporation of FT-MIR predicted traits into breeding programmes further depends on the heritability of the FT-MIR predicted trait, and the genetic correlation between the FT-MIR prediction and the benchmarked trait (Tiplady et al., 2020). Studies have reported moderate to high heritability estimates for a range of FT-MIR predicted traits, including fatty acids (Hein et al., 2018; Lopez-Villalobos et al., 2014; Rutten et al., 2010), milk proteins (Bonfatti et al., 2017d; Sanchez et al., 2017a), cheese-making and milk-coagulation properties (Cecchinato et al., 2015; Poulsen et al., 2015; Visentin et al., 2017), and lactoferrin concentrations (Lopez-Villalobos et al., 2009; Soyeurt et al., 2007a). Studies of individual FT-MIR spectra wavenumbers show that across most of the mid-infrared region, absorbances of individual FT-MIR spectra wavenumbers are moderately to highly heritable (Bittante and Cecchinato, 2013; Rovere et al., 2019; Soyeurt et al., 2010; Wang et al., 2016). This suggests that there is potential for achieving genetic gain through the direct use of FT-MIR spectra for selection, rather than

selection on FT-MIR predicted milk composition traits, which are themselves a function of the absorbance spectra at various wavenumbers.

Although there have been many genome-wide association studies (GWAS) for FT-MIR predicted milk composition traits such as fat, protein, and lactose concentrations (Jiang et al., 2010; Kemper et al., 2015b; Littlejohn et al., 2016; Lopdell et al., 2017; Raven et al., 2014), and individual fatty acid and protein fractions (Bouwman et al., 2011; Buitenhuis et al., 2016; Li et al., 2014), there are comparatively few studies reporting GWAS results for individual FT-MIR wavenumber phenotypes (Benedet et al., 2019; Wang and Bovenhuis, 2018; Zaalberg et al., 2020). Two such GWAS were conducted on medium density SNP-chip (~50k markers) genotypes for a subset of wavenumbers, which were identified either by clustering analysis (Wang and Bovenhuis, 2018), or by using phenotypic correlation structures and heritability estimates within each breed (Zaalberg et al., 2020). A third study explored relationships between FT-MIR wavenumber phenotypes and a subset of SNPs that had previously been implicated in a GWAS of milk composition and fatty acid traits (Benedet et al., 2019). Across these studies, a number of FT-MIR wavenumber QTL were identified. Most of the detected genomic regions had been previously reported in studies of major milk composition traits, but new regions with potential links to milk contents such as phosphorus, orotic acid or citric acid were identified (Wang and Bovenhuis, 2018). Thus, these findings have demonstrated that it is possible to identify genomic regions that are specifically related to individual FT-MIR wavenumber phenotypes.

Previous studies have examined the effects of variants in individual genes and their encoded proteins on FT-MIR wavenumber phenotypes (Benedet et al., 2019; Wang et al., 2016). Wang et al. (2016) observed that the *DGAT1* K232A polymorphism had highly significant effects on wavenumbers associated with carboxylic and ester C=O bond stretching, triglyceride ester linkage C-O stretching and alkyl C-H stretching. In that same study, a polymorphism in the *CSN3* gene had effects on wavenumbers that coincided with amide II, amide III and phosphate bands, and a polymorphism in the *PAEP* gene had effects on wavenumbers in a mid-infrared band that was attributed to C-N stretching (Wang et al., 2016). Similar effects were also observed by Benedet et al. (2019), with an additional absorption band associated with unsaturated fatty acids that was reported for a polymorphism near *CSN3*. Across those studies, association patterns varied widely for loci in different genes, with *DGAT1* having highly significant effects across many wavenumbers, while *PAEP* had significant effects across fewer wavenumbers that were concentrated within a small number of spectral bands. Assessing association patterns across the mid-infrared spectrum for a wider range of loci could improve our understanding of the impact that different genes have

on the molecular structure of milk. Moreover, comparing these association patterns could provide insights into commonalities in the way genes influence milk composition and how these impacts are detected.

The purpose of the current study was to investigate the underlying genetics of dairy cattle milk composition, by conducting GWAS on 895 individual FT-MIR wavenumber phenotypes, and comparing these results to GWAS conducted on three FT-MIR predicted major milk composition traits. We report the use of a much larger sample (n=38,085) than previous such studies and at a higher genomic resolution, with imputed whole-genome sequence consisting of 17,873,880 variants. We further report molecular dissection of these signals through the use of variant annotation information and a large mammary RNA-seq resource, and identification of candidate causative genes and variants for a substantial number of loci. Finally, we evaluated patterns of significance across the mid-infrared range for different loci, highlighting clusters of QTL that are broadly defined by the biochemical properties of the molecules that they encode.

6.3 Methods

6.3.1 Study population, animals and milk samples

In total, 100,571 FT-MIR spectra records from individual milk test samples for 38,085 multi-breed and crossbred cows across 1,645 herds were included for analysis. This dataset was a subset of a wider set of 2,044,094 FT-MIR spectra records analysed on six Bentley FTS (Chaska, MN, USA) instruments as part of routine milk testing conducted by Livestock Improvement Corporation (LIC), over the period from September 2017 to May 2018 (Tiplady et al., 2019). Records were included in the present study if they passed outlier removal based on the squared Mahalanobis distance between each spectrum and the average within-instrument spectra for each analyser, and had imputed sequence available for the cow from which the milk sample was taken. The pedigree-based breed composition of cows comprised 11,235 cows with $\geq$ 14/16 Holstein (HOL) or Friesian (FR) genetics; 5,374 cows with $\geq$ 14/16 Jersey (JE) genetics; 19,915 crossbred cows with HOL-FR ($\geq$ 3/16) and JE ($\geq$ 3/16) genetics only; 17 cows with $\geq$ 14/16 Ayrshire (AY) genetics; and 1,544 cows from other breeds or crosses. Individual FT-MIR wavenumbers were subjected to piecewise direct standardization (Grelet et al., 2015), with standardization coefficients evaluated from 16 weeks of reference sample calibration data collected across six Bentley instruments as in Tiplady et al. (2019).

6.3.2 Pre-adjustment of individual FT-MIR wavenumber and predicted milk composition phenotypes

Prior to conducting GWAS, adjusted cow phenotypes were generated for 895 individual FT-MIR wavenumbers and three FT-MIR predicted milk composition traits. Adjusted phenotypes were generated from one or more test-day samples on the same cow by fitting repeated measures models in ASReml-R (Butler et al., 2009), comprising:

$$y_{ijkl} = \mu + parity_j + dim_k + HD_l + \sum \alpha_m brd_{im} + \sum \delta_n het_{in} + anml_i + e_{ijkl} \quad (6.1)$$

where y_{ijkl} is a test-day phenotype (e.g., absorbance for one wavenumber) for the ith individual in parity class j within the days in milk class k and the herd-by-test date group l; μ is the overall mean; $parity_j$ is the fixed effect for parity j (5 classes: 1, 2, 3, 4, $\geq$ 5); dim_k is the fixed effect for the days in milk class k (9 intervals of 30 days each from the start of lactation); HD_l is the fixed effect for the herd by test day class l; α_m are breed linear regression coefficients for Holstein (HOL), Friesian (FR) and Jersey (JE) proportions and brd_{im} are the corresponding breed proportions for individual i; δ_n are heterosis linear regression coefficients between breeds (FRxJE, FRxHOL, JExHOL, FRxAY, JExAY, AYxHOL) and het_{in} are the corresponding heterosis proportions for individual i, according to sire and dam breed proportions; $anml_i$ is the random animal effect with $anml_i \sim N(0, I\sigma^2_{anml})$; and e_{ijkl} is the random error effect with $e_{ijkl} \sim N(0, I\sigma^2_e)$, where I is an identity matrix and σ^2_{anml} and σ^2_e are the variances of the independent and identically distributed animal and error variances, respectively. Adjusted phenotypes were evaluated for individual i as y minus all the relevant fixed effects averaged over all observations for a cow, or equivalently, the sum of the prediction of $anml_i$ and the average of the predicted error terms for all test-day records for the animal, i.e., $\hat{y}_{i(adj)} = anml_i + \bar{e}_{ij}$.

6.3.3 Genotypes and imputation

Animals were genotyped on Illumina BovineHD (HD; n=138; ~777k SNP), Illumina BovineSNP50k (50k; n=4,087; ~53k SNP), and/or custom GeneSeek Genomic Profiler LDv3 BeadChip (GGP; n=33,976; ~26k SNP) panels, with the resultant genotypes imputed to sequence density as part of a wider set of 153,357 animals, as described by Jivanji et al. (2019). More detailed descriptions of SNP-chip data handling and imputation criteria are given below, and as a summary, this process consisted of step-wise imputation of animals to whole-genome sequence genotypes via references

of GGP, 50k and HD genotypes. Whole-genome sequences for 565 animals had been mapped and called from the UMD3.1 *Bos taurus* reference genome using BWA-MEM (v0.78-r455) (Li and Durbin, 2009), and GATK (v3.2) (DePristo et al., 2011) respectively, as previously described (Jivanji et al., 2019; Littlejohn et al., 2016; Lopdell et al., 2017). The pedigree-based breed composition of sequenced animals comprised 138 Holstein-Friesians, 99 Jerseys, 316 Holstein-Friesian×Jersey crossbreeds and 12 from other breeds or crosses. Only variants located on *Bos taurus* autosomes were considered, and phasing with genotype probabilities was undertaken using Beagle 4.0 (Browning and Browning, 2007). Variants were filtered to remove those for which the allelic R^2, defined as the estimated squared correlation between the most likely allele dosage and the true allele dosage (Browning and Browning, 2009) for missing genotypes was less than 0.95. This resulted in a sequence reference comprising 19,659,361 segregating variants spanning all 29 bovine autosomes.

SNP-chip imputation references

The reference sets for SNP-chip panels used at each imputation step were generated based on a uniform set of criteria. Genotypes were eligible for inclusion in a reference if the sample call rate was ≥ 0.95, and the proportion of Mendelian inconsistences observed between parent-offspring pairs of genotypes was lower than 0.005. The 50k reference included eligible Illumina BovineSNP50 BeadChip genotypes for all males, and females that were a dam of a genotyped sire or had at least five recorded progeny (46,621 SNPs; 10,786 animals). The GGP reference included eligible GGP LD BeadChip genotypes for all males, and females that had recorded progeny (20,846 SNPs; 11,872 animals). Additional 50k reference SNPs that were not on the GGP panel were also included as a background scaffold, resulting in a reference with 57,493 SNPs across 11,872 animals. The HD reference included all available Illumina BovineHD BeadChip genotypes, predominantly from widely-used sires and/or sequenced animals (n=3,389), with 675,321 SNPs remaining after eligibility filters were applied.

For all references, SNPs that were monomorphic or had a batch call rate lower than 0.9 were excluded. Quality checks were made to ensure that allele frequencies in the reference population reflected those in the wider population. That is, for SNPs with a count of more than 1000 minor alleles in the overall population, the relationship between the minor allele frequency (MAF) in the reference population (MAF_{ref}) and the MAF in the overall population ($\text{MAF}_{overall}$) satisfied the criteria: $|\text{MAF}_{ref}\text{-MAF}_{overall}|/\text{MAF}_{ref} < 0.4$. This resulted in the removal of 12 SNPs from the Illumina BovineSNP50 BeadChip, and three SNPs from the GGP LDv3 BeadChip. In addition,

for all references, SNPs that were in common with sequence variants with more than 30x depth coverage were removed if the concordance between genotype and sequencing calls was ≤ 0.7. Likewise, for GGP and 50k references, any SNPs that were shared with the BovineHD panel were removed if the concordance between genotype calls from each panel was ≤ 0.7; and for the HD reference, any SNPs that were shared with the BovineSNP50 panel were removed if the genotypic concordance between panels was ≤ 0.7.

Imputation

All imputation steps were carried out ignoring pedigree information using Beagle 4.0 (Browning and Browning, 2007). Imputation of animals to GGP, 50k and HD references was carried out using default parameters, except for window sizes which were adjusted to ensure that whole chromosomes were imputed as one window. After each imputation step, SNPs with an allelic $R^2 < 0.7$ were removed. Imputation to the sequence level was carried out by using default parameters except for window sizes which were set at 50,000 SNPs. The overall median imputation allelic R^2 for the wider set of 153,357 animals was 0.986, the same value for the set of 38,085 animals included in this study.

6.3.4 Genome-wide association studies

Separate GWAS were conducted using the Bolt-LMM software (Loh et al., 2015) for each of the 898 pre-adjusted phenotypes that included the 895 FT-MIR wavenumber phenotypes and three FT-MIR predicted milk composition traits, namely, fat, lactose and protein concentrations (FP, LP, and PP). In total, 17,873,880 imputed sequence variants were included in each GWAS after applying a MAF threshold of 0.1%, based on allele frequencies in the study population of 38,085 animals. Mixed model association statistics were evaluated under an infinitesimal model (as defined by the Bolt-LMM software) to assess the additive effect of each SNP. A genomic relationship matrix (GRM) based on a subset of 43,851 SNPs was simultaneously fitted to account for population structure. That subset of SNPs was derived by filtering the 50k SNP-chip imputation reference (previously described) to exclude SNPs with a MAF lower than 0.1%. To avoid proximal contamination, a leave-one-segment-out (LOSO) approach was used in the GWAS, with segments of 5 Mbp used to subdivide the autosomes. A conservative Bonferroni significance threshold was used, which considered all tests across the 898 traits and 17,873,880 variants as independent. Based on a genome-wide threshold of $\alpha = 0.01$, the nominal p-value was 6.2e-13 and the corresponding Bonferroni threshold was $-\log_{10}(6.2\text{e-}13) = 12.21$. The proportion of

phenotypic variance explained by each SNP was evaluated as $\frac{2pqa^2}{\sigma_t^2}$ where p is the frequency of the minor allele, q = 1 - p, a is the estimated allele substitution effect, and σ_t^2 is the total phenotypic variance. Similarly, the proportion of genetic variance accounted for by each SNP was evaluated as $\frac{2pqa^2}{\sigma_g^2}$ where σ_g^2 is the estimated genetic variance according to SNP-based estimates generated by the Bolt-LMM software.

To distinguish between multiple QTL segregating within the same region of a chromosome, an iterative conditional approach was undertaken for each phenotype. After running an initial GWAS that we refer to as the 'base GWAS', chromosomes with a significant p-value based on the Bonferroni threshold were identified; and for each of these chromosomes, the most significant variant was identified and added to the set of covariates included in the next iteration. These subsequent iterations were only conducted on chromosomes that retained significant effects, whereby the process was repeated until these analyses ceased to highlight significant effects. For each of these iterations, the set of 43,851 SNPs representing genomic relationships continued to be fitted (using the LOSO approach) to account for population structure. These analyses resulted in a list of variants for each phenotype that aimed at capturing all the significant association analysis signal.

6.3.5 Gene expression phenotypes and eQTL identification

Gene expression phenotypes and the resulting eQTL were generated as part of a previously described study (Lopdell et al., 2017). Briefly, tissue from 411 cows was used to conduct high-depth mammary RNA-seq, yielding approximately 89 million read pairs per sample. Reads were mapped to the UMD3.1 *Bos taurus* reference genome using the Tophat2 program (version 2.0.12) (Kim et al., 2013), and filtered to remove outliers based on a principal components analysis of the gene expression values. Additional filters were applied to remove animals with excessively low call rates, and those with genotypes that were not concordant with sire or dam genotypes. This resulted in a dataset containing 357 animals, 62 of which were in common with the 38,085 animals in the current study. Transformed gene expression phenotypes for genes overlapping 1-Mb windows of whole-genome sequences were used to identify significant eQTL (Lopdell et al., 2017). Genetic impacts on gene expression were evaluated by fitting a generalised least-squares model that assessed the relationship between genotype and transformed gene expression phenotypes, with covariances between animals accounted for by the numerator relationship ($\boldsymbol{A}$) matrix. Resulting χ^2 statistics with 1 degree of freedom were used to identify eQTL p-values. The Bonferroni significance threshold had been set at $-\log_{10}$(2.53e-07), based on $\alpha = 0.05$, corrected for 197,338 tests.

6.3.6 Identification of protein-coding variants and co-localized eQTL

Whole-genome sequence resolution genotypes within a 1-Mbp window were annotated using the SnpEff software (version 4.1d; build 2015-04-13) (Cingolani et al., 2012) and Ensembl UMD3.1.78 gene annotations, to assess the candidacy of each wavenumber and predicted-trait QTL from the iterative GWAS. To focus on the most plausible candidates, variants in QTL regions were filtered to include only those in high linkage disequilibrium (LD) ($R^2 > 0.9$) with a putative impact variant (PIV), where we have defined a PIV as being a splice region variant, or a moderate or high impact coding variant, according to the SnpEff classification. For variants in QTL regions that met these criteria, emphasis was placed on those with 'highly significant' effects. That is, the correlation between the PIV and the QTL was in the range (0.975, 1] and the $-\log_{10}(p\text{-value})$ for the effect was greater than 1.5x the Bonferroni threshold; or the correlation between the PIV and the QTL was in the range (0.95, 0.975] and the $-\log_{10}(p\text{-value})$ for the effect was greater than 2x the Bonferroni threshold; or the correlation between the PIV and the QTL was in the range (0.925, 0.95] and the $-\log_{10}(p\text{-value})$ for the effect was greater than 2.5x the Bonferroni threshold. All other variants in QTL regions where the correlation between the PIV and the QTL was higher than 0.9, and the $-\log_{10}(p\text{-value})$ for the effect was greater than the Bonferroni threshold, were classified as 'moderately significant'.

Wavenumber and predicted-trait QTL were scrutinized to identify co-localized eQTL, following the methodology of Lopdell et al. (2017). This approach compares association statistics from the trait QTL to association statistics from variants in the same interval for an eQTL mapping to the same general locus, with the expectation that trait QTL underpinned by eQTL will have common top-associated variants, and/or will have similar patterns of association across the wider spectrum of variants within that interval. Briefly, for each QTL from the iterative GWAS, any significant, pre-computed eQTL within the same 1-Mbp window were identified. To identify cases where trait and expression QTL shared the same top-associated variant, LD criteria were used to highlight tag variants that, at $R^2 > 0.9$, were linked to the most significant, co-localized eQTL variant. To assess commonalities of association within the broader interval (i.e., beyond pairwise analysis of the top-associated trait QTL/eQTL tag variants), Pearson correlation coefficients between the log-scaled p-values of the trait QTL and all eQTL within the interval of interest were computed. To account for regional differences in LD structure, Pearson correlation coefficients were evaluated across the entire 1-Mbp region of interest, and a smaller 500-kbp region, with the strongest correlation used to assess the relationship between the trait and expression QTL p-values. Trait QTL were filtered to those for which the Pearson correlation from either window was higher than 0.7.

6.3.7 FT-MIR wavenumber association effect patterns for genes of interest

After conducting GWAS across FT-MIR wavenumbers, wavenumber QTL that were in strong LD with a PIV, or had a co-localized eQTL (as described in detail above) were identified. In cases where there were multiple candidate genes implicated for a QTL, the gene with a PIV in highest LD with the QTL was selected as representative of the locus. Where multiple loci were implicated for the same gene, the variant in highest LD with either the corresponding PIV or the top variant of the eQTL was used. For the identified genes, the $-\log_{10}(p\text{-values})$ for the representative tag variant were compiled across FT-MIR wavenumbers, creating significance 'profiles' that allowed patterns of association across the mid-infrared region to be compared between loci. To facilitate these comparisons and account for differences in p-value magnitudes between loci, the $-\log_{10}(p\text{-values})$ were scaled to sum to unity. Differences between scaled significance profiles for loci were evaluated based on the Euclidean distance between corresponding points on the profiles for pairs of genes, and clustering of the distances based on the largest pairwise dissimilarity across elements was performed using the hclust function in R (v4.0.2) (R Core Team, 2020) with default parameters.

6.4 Results

6.4.1 Sequence-based genome-wide association analysis

The first-round pre-iteration (base) GWAS, including 17,873,880 imputed sequence variants, resulted in significant associations for 37,779 variants for FP, 17,159 variants for LP, and 36,067 variants for PP. The number of significant associations for individual FT-MIR wavenumbers ranged from 50 to 60,242, with a mean and median of 24,505 and 25,895 variants, respectively. For 18 of the 895 individual wavenumber phenotypes, the Bolt-LMM GWAS did not converge, due to insufficient genetic variation in the trait. Among the remaining wavenumbers, 830 had at least one significant association in the base GWAS. The numbers of significant variants in the base GWAS for individual wavenumbers across the mid-infrared range are shown in Fig. 6.1. Regions of the spectrum associated with low signal-to-noise ratios and poor sample measurement repeatability, due to the water content in milk are shaded in blue, according to the definitions in Tiplady et al. (2019). Significant associations were identified across most of the spectrum, including within regions that were commonly associated with low signal-to-noise ratios. Among the significant

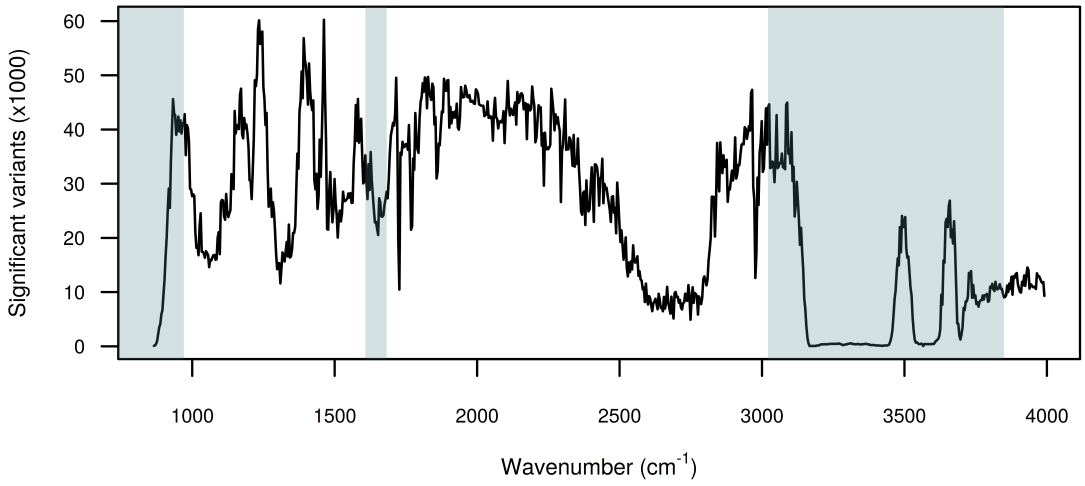

Figure 6.1: Number of significant variants from GWAS for each individual FT-MIR wavenumber. Noise regions (blue) with low repeatability are defined as from 649 to 970 cm^{-1}, from 1,608 to 1,682 cm^{-1}, and from 3,021 to 3,849 cm^{-1}

associations observed, 17.0% were positioned within the first 3 Mbp of chromosome 14, which encompasses the *DGAT1* gene that has been widely reported as impacting many milk composition traits (Grisart et al., 2002; Schennink et al., 2007). For the FP and PP phenotypes, the proportion of significant associations that were positioned within the first 3 Mbp of chromosome 14 were 16.5% and 13.6%, respectively. None of the significant associations for the LP phenotype localized to that region.

In the base GWAS, individual FT-MIR wavenumber QTL were observed on 27 of the 29 bovine autosomes (Fig. 6.2) within 450 different 1-Mbp regions. In contrast, QTL for FT-MIR predicted milk composition traits were observed on 25 of the 29 autosomes (Fig. 6.3) within 246 different 1-Mbp regions. The number of iterations required after the base GWAS until the analyses ceased to highlight significant effects for the FT-MIR wavenumber phenotypes ranged from 0 to 10, with an average of 3.9. For the FT-MIR predicted milk composition traits, FP, LP and PP, the number of iterations required after the base GWAS was 6, 8 and 7, respectively. For the FT-MIR wavenumber phenotypes, all significant signals were captured by no more than 68 tag variants, with the mean and median number of tag variants required to capture the signal for an individual wavenumber being 26 and 29, respectively. For FT-MIR predictions of FP, LP and PP, all significant signals were captured by 55, 72 and 86 tag variants, respectively.

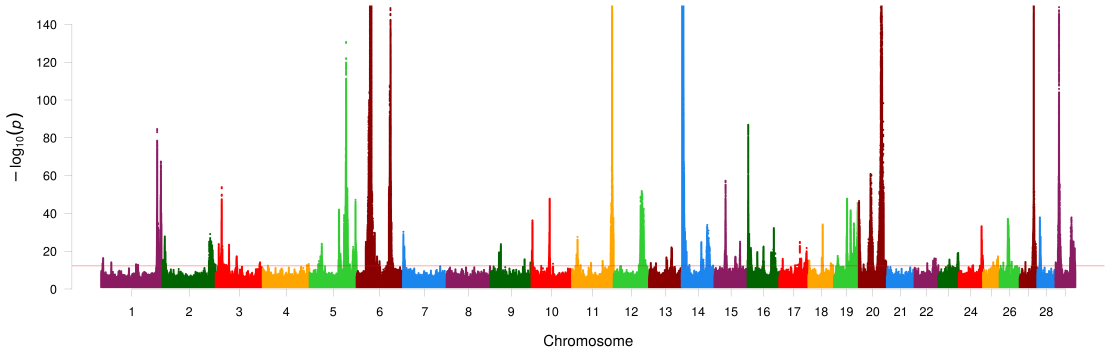

Figure 6.2: Manhattan plot showing association effects for FT-MIR wavenumbers. Consolidated association effects shown for FT-MIR wavenumbers. Chromosomes and genomic positions based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals are represented as the -$\log_{10}$(p-value) on the y-axis which has been truncated to facilitate visualisation of the results. The horizontal red line shows the Bonferroni significance threshold of -$\log_{10}$(6.2e–13)

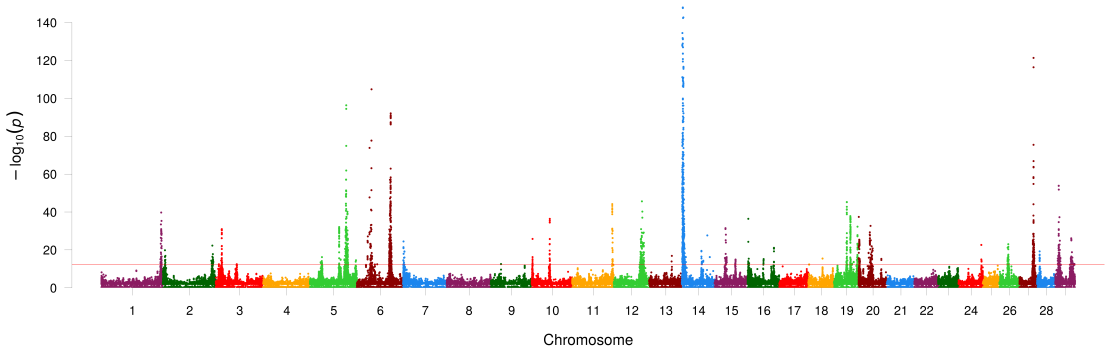

Figure 6.3: Manhattan plot showing association effects for FT-MIR predicted milk composition traits. Consolidated association effects shown for FT-MIR predicted milk production traits (Fat %, Lactose % and Protein %). Chromosomes and genomic positions based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals are represented as the -$\log_{10}$(p-value) on the y-axis which has been truncated to facilitate visualisation of the results. The horizontal red line shows the Bonferroni significance threshold of -$\log_{10}$(6.2e–13)

6.4.2 Identification of candidate causative variants

To identify candidate causative variants for wavenumber and predicted-trait QTL, we used functional annotation to find PIV in strong LD ($R^2 > 0.9$) with trait QTL from the GWAS iterations. Those criteria yielded 42 1-Mbp regions, encompassing 55 effects with a PIV for at least one FT-MIR wavenumber. Based on our categorisation of signals into moderately and highly significant groups, 31 of the 55 wavenumber QTL were classified as highly significant. Details of these 31 effects are in Table 6.1. Manhattan plots of a 1-Mbp region centred on the QTL tag variant for each of the 31 highly significant wavenumber QTL from the base GWAS are provided in Additional file 1: Fig. S1 of the original paper (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00648-9#additional-information). Details of the wavenumber QTL classified as

moderately significant are provided in Appendix 6.A.1. Note that there are three effects where the locus has been identified as highly significant based on the LD with one or more other loci (Table 6.1), and moderately significant based on the LD with other loci (Appendix 6.A.1). Effect sizes and MAF details for the tag SNP of the 31 highly significant wavenumber QTL are provided in Appendix 6.A.2. For each of these 31 QTL, the proportions of phenotypic and genetic variance that they account for across FT-MIR wavenumber and predicted composition traits are provided in Additional file 2: Table S4 of the original paper (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00648-9#additional-information). Of the 31 highly significant wavenumber QTL, 14 were identified in the base GWAS (Iteration 0). For the 17 highly significant wavenumber QTL identified in subsequent GWAS iterations after the base GWAS (Table 6.1), p-values at previous iterations for the phenotype, and p-values for the corresponding top chromosomal SNP in that iteration are provided in Additional file 2: Table S5 of the original paper (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00648-9#additional-information).

For predicted composition traits, 27 effects with a PIV were identified within 15 1-Mbp regions. Of the 27 predicted-trait QTL, 18 were classified as highly significant. Details of these effects are in Table 6.2, with details of the QTL classified as moderately significant provided in Appendix 6.A.3. Effect sizes and MAF details for highly significant predicted-trait QTL are provided in Appendix 6.A.4. Details of highly significant predicted-trait QTL from iterations subsequent to the base GWAS are provided in Additional file 2: Table S8 of the original paper (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00648-9#additional-information).

Of all candidate protein coding mutations identified, we were particularly interested in those identified as having a high impact according to the SnpEff classification, in which variants that are expected to strongly disrupt or ablate gene function could *a priori* be considered as excellent candidates for these QTL. Three such PIV from the wavenumber and predicted-trait QTL fit this definition, comprising frameshift mutations in the *FCGR2B* or *KCNH4* genes, and a splice donor mutation in the *ABO* gene (Tables 6.1 and 6.2). Since this class of variants was likely to be enriched for annotation errors (MacArthur et al., 2012), we manually visualized mammary RNA-seq alignments for these mutations to help confirm their predicted impacts as disruptive of coding sequences. Although the *FCGR2B* rs381714237 variant was represented in the RNA-seq reads, the mutation appeared to be intronic. Annotation of the *KCNH4* mutation appeared similarly dubious, with limited evidence suggesting that it was localized in a mammary-expressed

exon. The *ABO* rs207688357 mutation was clearly localized in the donor site of the splice junction of intron/exon 5, with animals that carried the mutation showing activation of cryptic alternative splice sites. These alternative transcripts comprised an 8-bp contraction, or 33-bp expansion of exon 5 (splicing at chr11:104242578 and chr11:1042425462 respectively, Fig. 6.4), which suggests that the ABO protein in animals homozygous for the mutation is non-functional.

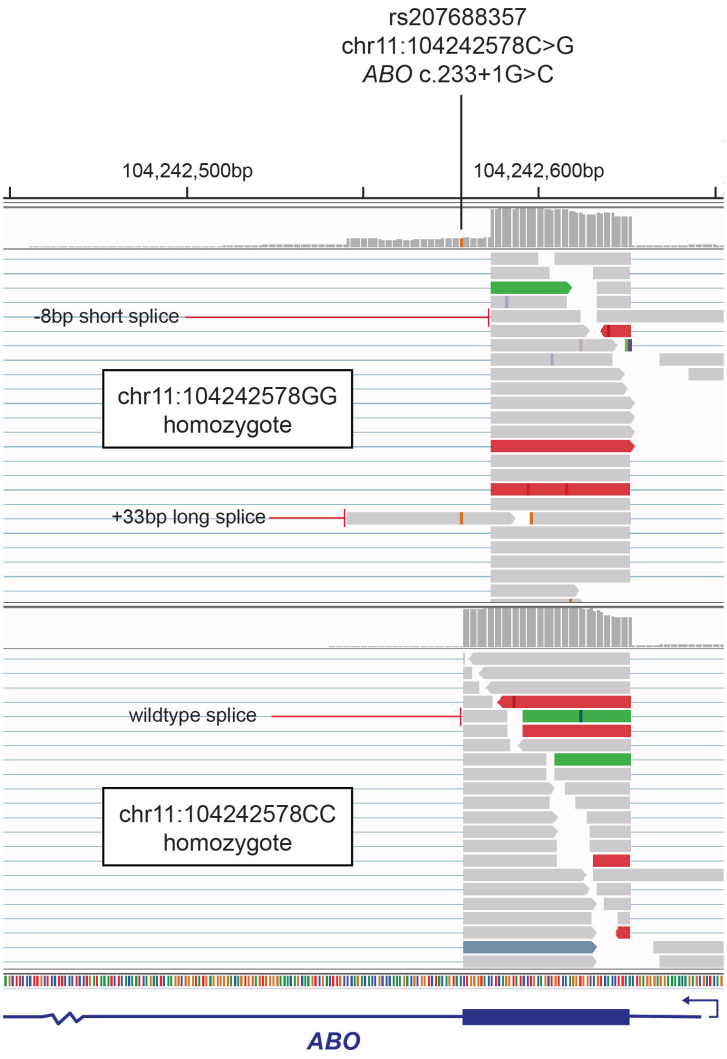

Figure 6.4: Mammary RNA-seq alignments representing *ABO* intron/exon 5 splicing structures of the chr11:104242578GG and chr11:104242578CC genotypes. The site of the proposed chr11:104242578C>G essential splice donor SNP is indicated, with individual reads and coverage data showing alternate splice forms in the animal carrying the mutation. This coverage track also represents the cryptic, '+33 bp long splice' transcript as the minority splice form relative to the '- 8 bp short splice' transcript, the former representing an in-frame variant, with the latter causing a predicted frame-shifted protein isoform

Table 6.1: Peak variants for FT-MIR wavenumbers with highly significant protein sequence association effects

Chr	Position	Tag variant ID	No. of hits	Top wvn cm^{-1}	Iter	*P*-value	Protein coding variant ID	LD	Gene	Impact	Description
3	7908611	rs137763930	11	940	1	6.7e-20	rs110560331	0.976	*FCRLA*	L	c.233-3T>C
3	7931694	rs211402696	20	1462	2	1.2e-23	rs381714237	0.989	*FCGR2B*	H	c.899dupC
3	15411459	rs134900385	6	1022	1	4.3e-19	rs382689947	0.994	*FAM189B*	M	c.1237T>C
3	15411459	rs134900385	6	1022	1	4.3e-19	rs134844772	0.990	*GBA*	M	c.1080C>A
3	15411459	rs134900385	6	1022	1	4.3 e-19	rs132659643	0.999	*HCN3*	M	c.1699A>G
3	15411459	rs134900385	6	1022	1	4.3e-19	rs109330809	0.990	*MTX1*	L	c.508-6T>C
3	15517871	rs109328483	6	1007	1	4.4e-19	rs136761456	0.992	*SCAMP3*	M	c.151G>C
3	15517871	rs109328483	6	1007	1	4.4e-19	rs43706482	0.994	*THBS3*	L	c.2075-3T>C
3	15550598	rs380597285	327	1462	0	1.3e-54	rs109816684	0.994	*SLC50A1*	L	c.282+7G>A
5	75729880	rs384734208	50	1466	1	5.0e-47	rs207628090	0.930	*CSF2RB*	M	c.41T>C
5	75758989	rs210094995	2	1447	0	5.8e-40	rs210937722	0.926	*NCF4*	M	c.841G>C
5	118246868	rs136859160	308	1261	0	3.0e-44	rs456403270	0.937	*TBC1D22A*	M	c.1063C>T
6	38027010	rs43702337	455	1119	0	7.3e-948	rs43702337	1	*ABCG2*	M	c.1742A>C
6	87181619	rs43703011	17	3633	2	2.5e-22	rs43703011	1	*CSN2*	M	c.245C>A
6	87274397	rs378808772	3	1283	2	9.9e-51	rs43703010	0.974	*CSN1S1*	M	c.620A>G
6	87390576	rs43703015	18	1473	1	4.0e-108	rs43703015	1	*CSN3*	M	c.470T>C
11	103304757	rs109625649	329	1593	0	1.2e-134	rs109625649	1	*PAEP*	M	c.401T>C
11	104242578	rs207688357	11	1462	0	5.5e-33	rs207688357	1	*ABO*	H	c.233+1G>C
12	69612955	rs383509255	132	1716	0	6.4e-45	rs208744187	0.950	*TGDS*	M	c.204A>C
14	1726650	rs133611586	6	3514	1	1.6e-75	.	0.992	*WDR97*	L	c.2656-5_2656-4insG
14	1732043	rs437406031	384	2846	1	6.3e-42	rs450710918	0.990	*ENS..39978*	M	c.352G>A
14	1732043	rs437406031	384	2846	1	6.3e-42	rs476736066	0.997	*MROH1*	M	c.3549G>C
14	1755742	rs384226556	5	2656	0	4.0e-20	rs209542297	0.9998	*CPSF1*	L	c.4287T>C
14	1802265	rs109234250	310	1716	0	1.5e-2607	rs109234250	1	*DGAT1*	M	c.694G>A
14	1802265	rs109234250	310	1716	0	1.5e-2607	rs134364612	0.999	*SLC52A2*	M	c.724A>G
14	66328304	rs446084949	19	1029	1	2.7e-20	rs446084949	1	*SPAG1*	M	c.2044G>A
15	28347165	rs210034037	5	1537	0	7.7e-35	rs208325660	0.999	*RNF214*	M	c.314G>A
15	53940444	rs382926661	23	1205	1	4.2e-19	rs380220394	0.993	*DNAJB13*	L	c.69-4T>C
16	24977696	rs111027377	62	2742	2	4.8e-25	rs109896036	0.988	*MTARC1*	L	c.628-5C>T
16	24977696	rs111027377	62	2742	2	4.8e-25	rs110899826	0.988	*MTARC1*	M	c.581C>G
19	42428366	rs209808022	4	1250	1	3.1e-25	rs209302038	0.991	*KRT9*	M	c.196C>T
19	42488389	rs379667889	8	1447	0	7.8e-34	rs209756857	0.969	*KRT42*	L	c.57+7C>T
19	42488389	rs379667889	8	1447	0	7.8e-34	rs383013355	0.963	*KRT16*	M	c.896A>G
19	42488389	rs379667889	8	1447	0	7.8e-34	rs208923483	0.966	*KRT17*	M	c.146G>C
19	42488389	rs379667889	8	1447	0	7.8e-34	rs385937063	0.966	*KRT17*	L	c.1233C>T
19	43036265	rs210324533	11	1029	1	5.3e-43	rs207799702	0.944	*KAT2A*	L	c.700-7C>G
19	43036265	rs210324533	11	1029	1	5.3e-43	rs209410283	0.945	*KCNH4*	M	c.408C>G
19	43036265	rs210324533	11	1029	1	5.3e-43	rs377779402	0.945	*KCNH4*	H	c.2663+2T>C
19	43053995	rs481837688	24	1212	1	6.6e-26	rs481837688	1	*STAT5A*	M	c.2305C>A
19	51303887	rs41921224	65	1499	0	1.9e-35	rs41921160	0.993	*CCDC57*	M	c.1907T>C
19	57087981	rs41920620	6	1216	0	1.8e-21	rs469721022	0.999	*HID1*	L	c.1147-7G>C
28	6559147	rs133101552	3	1261	0	8.6e-23	rs133101552	1	*KCNK1*	M	c.934C>A
29	41821270	rs207854419	14	1257	1	4.6e-30	rs384900272	0.998	*NXF1*	M	c.1555G>A

Peak variants and association effects for FT-MIR wavenumbers classified as highly significant. Highly significant effects are classified such that: the -$\log_{10}$(p-value) for the effect was greater than 1.5 x the Bonferroni threshold and the correlation between the tag variant and the protein sequence variant was in the range (0.975, 1]; or the -$\log_{10}$(p-value) for the effect was greater than 2 x the Bonferroni threshold and the correlation between the tag variant and the protein sequence variant was in the range (0.95, 0.975]; or the -$\log_{10}$(p-value) for the effect was greater than 2.5 x the Bonferroni threshold and the correlation between the tag variant and the protein sequence variant was in the range (0.925,0.95]. Bonferroni threshold: -$\log_{10}$(6.2e-13). No. of hits = number of wavenumbers for which the variant was selected as the representative (most significant) tag variant for a peak. Iterations (Iter) are defined relative to the base GWAS, with the base GWAS represented as iteration 0.
Abbreviations: L = Low impact splice region variant; M = Moderate impact missense variant; H = High impact splice donor.

Table 6.2: Peak variants for composite milk production traits with highly significant protein sequence association effects

Trait	Chr	Position	Tag variant ID	Iteration	*P*-value	Protein coding variant ID	LD	Gene	Impact	Description
FP	5	75698283	rs385866519	1	4.0e-19	rs207628090	0.979	*CSF2RB*	M	c.41T>C
FP	11	103304757	rs109625649	0	4.3e-46	rs109625649	1	*PAEP*	M	c.401T>C
FP	12	69608900	rs211406918	0	4.2e-33	rs208744187	0.951	*TGDS*	M	c.204A>C
FP	14	1732043	rs437406031	1	7.2e-37	rs450710918	0.990	*ENS..39978*	M	c.352G>A
FP	14	1732043	rs437406031	1	7.2e-37	rs476736066	0.997	*MROH1*	M	c.3549G>C
FP	14	1800439	rs209876151	0	8.9e-2225	rs109326954	0.9999	*DGAT1*	M	c.695C>A
FP	14	1800439	rs209876151	0	8.9e-2225	rs134364612	0.9998	*SLC52A2*	M	c.724A>G
LP	3	15433518	rs109749506	1	1.3e-20	rs382689947	0.995	*TENT5A*	M	c.1237T>C
LP	3	15433518	rs109749506	1	1.3e-20	rs134844772	0.992	*GBA*	M	c.1080C>A
LP	3	15433518	rs109749506	1	1.3e-20	rs109330809	0.992	*MTX1*	L	c.508-6T>C
LP	3	15545091	rs379353107	0	2.2e-42	rs109816684	0.998	*SLC50A1*	L	c.282+7G>A
LP	6	38027010	rs43702337	0	9.0e-717	rs43702337	1	*ABCG2*	M	c.1742A>C
LP	16	24983926	rs110162358	2	1.0e-19	rs109896036	0.999	*MTARC1*	L	c.628-5C>T
LP	16	24983926	rs110162358	2	1.0e-19	rs110899826	0.999	*MTARC1*	M	c.581C>G
LP	19	43036265	rs210324533	3	9.4e-40	rs207799702	0.944	*KAT2A*	L	c.700-7C>G
LP	19	43036265	rs210324533	3	9.4e-40	rs209410283	0.945	*KCNH4*	M	c.408C>G
LP	19	43036265	rs210324533	3	9.4e-40	rs377779402	0.945	*KCNH4*	H	c.2663+2T>C
PP	3	15550598	rs380597285	0	1.7e-37	rs109816684	0.994	*SLC50A1*	L	c.282+7G>A
PP	5	75758989	rs210094995	0	3.3e-34	rs209394772	0.935	*CSF2RB*	M	c.227G>A
PP	5	75758989	rs210094995	0	3.3e-34	rs210937722	0.926	*NCF4*	M	c.841G>C
PP	5	118239754	rs384479185	2	3.9e-32	rs456403270	0.976	*TBC1D22A*	M	c.1063C>T
PP	6	38027010	rs43702337	0	6.4e-115	rs43702337	1	*ABCG2*	M	c.1742A>C
PP	14	1763380	rs135017891	0	5.9e-718	rs135258919	0.999	*HSF1*	M	c.1031T>C
PP	14	1802265	rs109234250	1	1.2e-61	rs109234250	1	*DGAT1*	M	c.694G>A
PP	14	1802265	rs109234250	1	1.2e-61	rs134364612	0.999	*SLC52A2*	M	c.724A>G
PP	15	53940444	rs382926661	1	2.9e-20	rs380220394	0.992	*DNAJB13*	L	c.69-4T>C
PP	19	43035006	rs209494359	0	1.6e-40	rs207799702	0.944	*KAT2A*	L	c.700-7C>G
PP	19	43035006	rs209494359	0	1.6e-40	rs209410283	0.945	*KCNH4*	M	c.408C>G
PP	19	43035006	rs209494359	0	1.6e-40	rs377779402	0.945	*KCNH4*	H	c.2663+2T>C

Peak variants for composite milk production traits with highly significant protein sequence effects whereby: the -$\log_{10}$(p-value) for the effect was greater than 1.5 x the Bonferroni threshold and the correlation between the tag variant and the protein sequence variant was in the range (0.975, 1]; or the -$\log_{10}$(p-value) for the effect was greater than 2 x the Bonferroni threshold and the correlation between the tag variant and the protein sequence variant was in the range (0.95, 0.975]; or the -$\log_{10}$(p-value) for the effect was greater than 2.5 x the Bonferroni threshold and the correlation between the tag variant and the protein sequence variant was in the range (0.925,0.95]. Bonferroni threshold: -$\log_{10}$(6.2e-13). Iterations (Iter) are defined relative to the base GWAS, with the base GWAS represented as iteration 0.
Abbreviations: FP = Fat %; LP = Lactose %; PP = Protein %; L = Low impact splice region variant; M = Moderate impact missense variant; H = High impact splice donor.

6.4.3 Identification of co-localized eQTL

Comparisons of association statistics from trait QTL to those representing mammary eQTL variants in the same interval identified co-localized eQTL for 38 wavenumber QTL (see details in Table 6.3). For 19 of these identified from the base GWAS (Iteration = 0), Manhattan plots are provided for 1-Mbp regions centred on the trait QTL tag variant in Additional file 3: Fig. S9 of the original paper (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00648-9#additional-information). Effect sizes and MAF details for all 38 loci with a co-localized trait QTL and eQTL pair are provided in Appendix 6.A.5. For each of these 38 loci, the proportions of phenotypic and genetic variance explained across FT-MIR wavenumber and predicted composition traits are provided in Additional file 4: Table S11 of the original paper (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00648-9#additional-information). For the 19 trait QTL identified in subsequent GWAS iterations after the base GWAS, p-values at previous iterations for the phenotype, and p-values for the corresponding top chromosomal SNP in that iteration are provided in Additional file 4: Table S12 of the original paper (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00648-9#additional-information).

Co-localized eQTL were identified for 25 predicted-trait QTL. Details of these trait QTL and eQTL pairs are in Table 6.4, with effect sizes and MAF details provided in Appendix 6.A.6. Further details of the 12 QTL identified in iterations subsequent to the base GWAS are provided in Additional file 4: Table S14 of the original paper (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00648-9#additional-information).

6.4.4 Investigation of patterns of FT-MIR wavenumber associations for genes of interest

In total, 70 genes were implicated whereby the tag locus of the wavenumber QTL was in high LD with a PIV (Table 6.1), or in high LD with the top variant of a co-localized eQTL (Table 6.3). In cases where multiple candidate genes were implicated for a QTL, the gene with the PIV in highest LD with the QTL tag variant was used to represent the locus. This resulted in tag loci representing 59 genes, for which scaled significance profiles were generated to represent their association patterns across the mid-infrared region. Clustering analysis based on the largest pairwise dissimilarity between corresponding points on profiles for pairs of genes resulted in > 20 clusters (Fig. 6.5). Significance profiles for all 59 genes are provided in Additional file 5: Fig. S15 of the original paper (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00648-9#additional-information).

Table 6.3: Peak variants for FT-MIR wavenumbers with co-localized eQTL

Chr	Position	Tag variant ID	Iter	Top wvn cm^{-1}	No. of hits	*P*-value	Gene	Top eQTL variant ID	Top eQTL variant *P*-value	LD	Pearson	Pearson window (Mb)
1	5120248	rs42317521	2	2794	55	4.0e-18	*CLDN8*	rs42317521	4.3e-17	1	0.764	0.5
1	144377960	rs208161466	0	2592	22	2.3e-85	*SLC37A1*	rs208161466	4.1e-15	1	0.710	1.0
1	146481250	rs383691757	1	1071	1	4.7e-15	*CSTB*	rs210595016	1.4e-52	0.992	0.938	0.5
1	154125158	rs207836083	0	1130	142	3.6e-68	*SH3BP5*	rs207836083	2.6e-32	1	0.816	0.5
3	15411459	rs134900385	1	1022	6	4.3e-19	*KRTCAP2*	rs133285846	9.7e-09	0.996	0.938	1.0
3	15550598	rs380597285	0	1462	325	1.3e-54	*SLC50A1*	rs380597285	8.7e-16	1	0.854	0.5
3	34387618	rs109030498	1	1466	157	3.5e-25	*ELAPOR1*	rs109030498	3.2e-30	1	0.817	0.5
3	53755929	rs209271975	1	1089	15	8.6e-22	*LRRC8C*	rs466686834	3.5e-39	0.99	0.927	0.5
5	75729880	rs384734208	1	1466	44	5.0e-47	*CSF2RB*	rs210641868	9.2e-27	0.926	0.752	1.0
5	75732526	rs210305241	1	1458	5	4.4e-42	*NCF4*	rs209273109	9.3e-16	0.91	0.813	1.0
5	93945738	rs211210569	0	1171	544	1.8e-131	*MGST1*	rs209372883	3.2e-43	0.919	0.925	1.0
6	46568418	rs210515595	3	1772	5	7.7e-22	*SLC34A2*	rs110805476	2.5e-07	0.979	0.805	0.5
6	87388064	rs379473589	1	1436	17	1.1e-97	*CSN3*	rs208009847	9.9e-33	0.963	0.878	0.5
9	21637056	rs209222932	0	1003	33	9.4e-20	*YRMY5A*	rs209222932	2.8e-36	1	0.634	0.5
9	26534109	rs208123385	0	1462	36	1.9e-24	*RNF217*	rs208173647	1.3e-16	0.986	0.856	0.5
9	87585031	rs110986237	1	1470	6	7.4e-15	*TAB2*	rs110986237	9.5e-12	1	0.851	0.5
9	102874726	rs137238900	0	1768	1	1.0e-14	*MPC1*	rs134094426	6.9e-15	0.969	0.849	0.5
10	46581015	rs109326466	0	1246	15	2.0e-46	*USP3*	rs109326466	2.0e-31	1	0.961	0.5
11	14180010	rs110527112	1	2760	23	3.6e-29	*XDH*	rs207554031	8.8e-26	0.978	0.709	0.5
11	78868975	.	1	1112	10	1.2e-19	*LAPTM4A*	rs110552157	1.3e-40	0.998	0.920	0.5
11	103292402	rs383398415	0	2548	1	3.5e-56	*PAEP*	rs109333988	1.2e-29	0.933	0.956	0.5
11	104229609	rs110534892	0	3648	10	1.2e-21	*ABO*	rs109750996	3.9e-28	0.944	0.803	0.5
14	1754287	rs135443540	0	1085	3	1.6e-39	*DGAT1*	rs137202508	8.9e-42	0.905	0.944	0.5
15	57266467	rs136337092	0	3935	1	2.7e-13	*CAPN5*	rs136208815	9.3e-46	0.997	0.940	0.5
16	66314547	rs42579412	2	1425	1	1.0e-15	*RGL1*	rs42579412	6.3e-14	1	0.727	0.5
16	67730371	rs380453838	1	1757	125	3.8e-21	*IVNS1ABP*	rs380453838	4.5e-27	1	0.876	0.5
18	2203322	rs132899112	1	1466	7	1.4e-15	*FA2H*	rs137235970	1.9e-27	0.998	0.875	0.5
19	33517487	rs434248431	0	1100	23	2.9e-46	*PMP22*	rs434248431	8.6e-38	1	0.832	0.5
19	43036265	rs210324533	1	1029	11	5.3e-43	*GHDC*	rs381442991	1.8e-22	0.945	0.975	0.5
19	57079881	rs381175117	2	1220	9	2.0e-23	*HID1*	rs109407913	1.2e-32	0.936	0.803	0.5
19	61134515	rs41923843	0	1130	45	3.2e-46	*KCNJ2*	rs41923843	1.7e-26	1	0.882	0.5
20	58454531	rs135636613	0	1391	23	4.3e-441	*ANKH*	rs135636613	2.4e-16	1	0.860	0.5
22	53519865	rs109233889	0	1235	7	5.3e-15	*LTF*	rs109233889	1.3e-32	1	0.813	0.5
24	58817202	rs208779762	0	1220	23	6.8e-34	*LMAN1*	rs207893260	1.3e-27	0.958	0.713	1.0
27	36211708	rs209855549	0	1731	157	6.2e-188	*GPAT4*	rs209855549	3.7e-21	1	0.848	0.5
27	41267242	rs109068627	1	2977	23	3.5e-26	*THRB*	rs109068627	1.7e-22	1	0.704	0.5
29	9546217	rs380868305	0	1130	8	4.6e-186	*PICALM*	rs380868305	2.4e-54	1	0.831	0.5
29	44579245	rs439384463	2	1548	3	4.3e-16	*MUS81*	.	3.0e-21	0.924	0.913	0.5

Peak variants for FT-MIR wavenumbers with a co-localized eQTL. Co-localized eQTL are defined such that: the Pearson correlation between the -$\log_{10}$(*p*-values) of the trait QTL and the -$\log_{10}$(*p*-values) of the eQTL is higher than 0.7; and the LD between the tag variant for the trait QTL and the top eQTL variant is higher than 0.9. The Pearson correlation shown is the highest from two different size windows (0.5 Mbp and 1 Mbp), centred on the top tag variant. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0. No. of hits: number of wavenumbers for which the variant was selected as the representative (most significant) tag variant for a peak.

Table 6.4: Peak variants for milk production traits with co-localized eQTL

Trait	Chr	Position	Tag variant ID	Iter	*P*-value	Gene	Top eQTL variant ID	Top eQTL variant *P*-value	LD	Pearson	Pearson window (Mb)
FP	3	34387618	rs109030498	2	6.0e-13	*ELAPOR1*	rs109030498	3.2e-30	1	0.832	0.5
FP	5	75698283	rs385866519	1	4.0e-19	*CSF2RB*	rs210641868	9.2e-27	0.910	0.701	1
FP	5	93945738	rs211210569	0	6.7e-106	*MGST1*	rs209372883	3.2e-43	0.919	0.928	1
FP	10	46483019	rs133089336	0	4.5e-13	*USP3*	rs208181306	2.0e-31	0.909	0.905	0.5
FP	11	104229609	rs110534892	1	2.6e-14	*ABO*	rs109750996	3.9e-28	0.944	0.727	1
FP	16	67730371	rs380453838	1	2.6e-19	*IVNS1ABP*	rs380453838	4.5e-27	1	0.886	0.5
FP	27	36211708	rs209855549	0	9.7e-132	*GPAT4*	rs209855549	3.7e-21	1	0.819	0.5
LP	1	154122887	rs42167460	0	1.2e-50	*SH3BP5*	rs380642859	2.6e-32	0.999	0.859	0.5
LP	3	15433518	rs109749506	1	1.3e-20	*KRTCAP2*	rs133285846	9.7e-09	0.995	0.940	1
LP	3	15545091	rs379353107	0	2.2e-42	*SLC50A1*	rs379353107	8.7e-16	1	0.806	0.5
LP	3	53994057	rs211488591	2	6.7e-18	*LRRC8C*	rs466686834	3.5e-39	0.986	0.753	1
LP	19	43036265	rs210324533	3	9.4e-40	*GHDC*	rs381442991	1.8e-22	0.945	0.963	0.5
LP	19	61134515	rs41923843	1	1.1e-46	*KCNJ2*	rs41923843	1.7e-26	1	0.857	0.5
LP	20	58448763	rs134813825	0	3.2e-18	*ANKH*	rs134813825	2.4e-16	1	0.809	0.5
LP	27	36204066	rs208306200	0	1.9e-21	*GPAT4*	rs208306200	3.7e-21	1	0.767	0.5
LP	29	9577372	rs380473328	0	2.1e-140	*PICALM*	rs384691767	2.4e-54	0.996	0.845	0.5
PP	3	15520971	rs109098377	2	7.5e-16	*KRTCAP2*	rs133285846	9.7e-09	0.989	0.928	0.5
PP	3	15550598	rs380597285	0	1.7e-37	*SLC50A1*	rs380597285	8.7e-16	1	0.832	0.5
PP	5	75680825	rs208925020	4	8.5e-23	*CSF2RB*	rs210641868	9.2e-27	0.947	0.871	1
PP	5	93945738	rs211210569	1	3.7e-42	*MGST1*	rs209372883	3.2e-43	0.919	0.817	0.5
PP	6	87387870	rs382652853	2	2.9e-45	*CSN3*	rs208009847	9.9e-33	0.963	0.891	0.5
PP	10	46581015	rs109326466	0	4.0e-38	*USP3*	rs109326466	2.0e-31	1	0.961	0.5
PP	18	2203325	rs135350753	0	2.1e-13	*FA2H*	rs137235970	1.9e-27	0.997	0.831	0.5
PP	19	43035006	rs209494359	0	1.6e-40	*GHDC*	rs381442991	1.8e-22	0.945	0.976	0.5
PP	24	58817202	rs208779762	0	5.7e-26	*LMAN1*	rs207893260	1.3e-27	0.958	0.737	0.5

Peak variants for composite milk production traits with a co-localized eQTL. Co-localized eQTL are defined such that: the Pearson correlation between the -$\log_{10}$(p-values) of the trait QTL and the -$\log_{10}$(p-values) of the eQTL is higher than 0.7; and the LD between the tag variant for the trait QTL and the top eQTL variant is higher than 0.9. The Pearson correlation shown is the highest from two different size windows (0.5 Mbp and 1 Mbp), centred on the top tag variant. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0.
Abbreviations: FP = Fat %; LP = Lactose %; PP = Protein %.

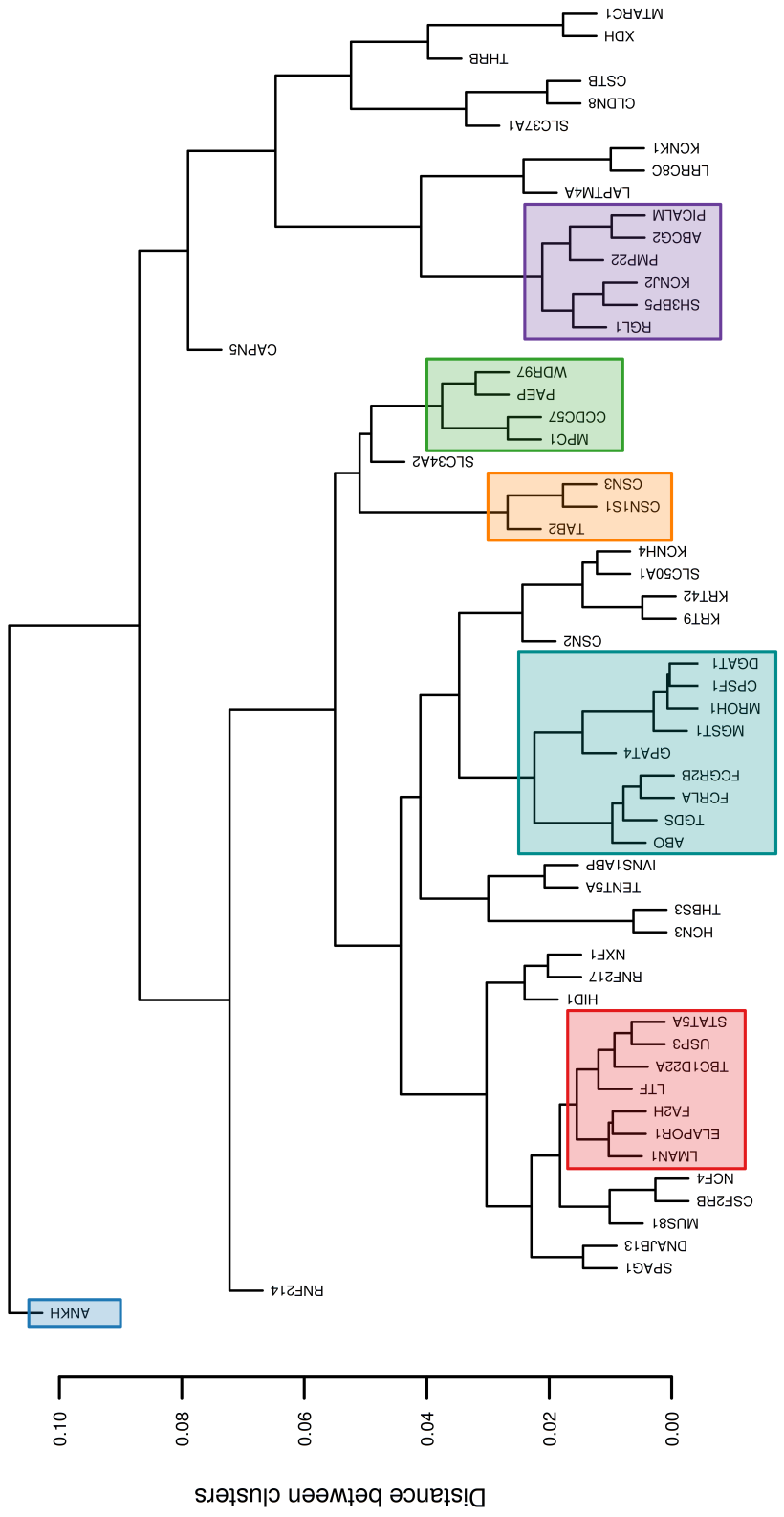

Figure 6.5: Gene clusters for significance profiles of tag variants representing candidate genes. Gene clusters based on the Euclidean distance between pairs of log-scaled p-value profiles across the mid-infrared spectrum for tag variants. Significance profiles for the highlighted gene clusters are presented in Figs. 6.6 and 6.7

Significance profiles for a subset of gene clusters from Fig. 6.5 are presented in Fig. 6.6. For each cluster, the significance profile for the gene with the largest QTL is shown in dark grey with the profiles for other genes within the cluster (according to highlighted clusters in Fig. 6.5) shown in light grey. Significance profiles varied widely between clusters, but were highly consistent within clusters. The first cluster (Fig. 6.6a) includes genes with significant associations for the LP (*ABCG2*, *SH3BP5*, *KCNJ2*, *PICALM*) and PP phenotypes (*ABCG2*). For this cluster of genes, prominent peaks were observed in bands of the mid-infrared spectrum from ~1,020 to 1,180 cm^{-1}, from ~1,200 to 1,470 cm^{-1}, from ~2,610 to 2,840 cm^{-1} and from ~2,870 to 2,980 cm^{-1}. The second cluster (Fig. 6.6b) includes genes with significant associations for the FP (*USP3*, *ELAPOR1*, *TBC1D22A*) and PP (*USP3*, *LMAN1*, *FA2H*, *TBC1D22A*, *STAT5A*) phenotypes, with multiple peaks observed across the mid-infrared spectrum, with the most prominent of these being in the ranges from ~910 to 1,010 cm^{-1}, from ~1,070 to 1,560 cm^{-1}, from ~1,700 to 2,450 cm^{-1}, from ~2,630 to 2,980 cm^{-1} and from ~3,620 to 3,680 cm^{-1}. The third cluster (Fig. 6.6c) includes a number of genes with significant associations for the FP (*DGAT1*, *ABO*, *TGDS*, *GPAT4*, *MGST1*, *MROH1*) and PP (*DGAT1*, *MGST1*) phenotypes. For this cluster of genes, peaks were observed in many bands of the mid-infrared spectrum in common with peaks for *ABCG2* and *USP3* (Fig. 6.6a; 6.6b), including from ~910 to 1,010 cm^{-1}, from ~1,130 to 1,260 cm^{-1}, from ~1,450 to 1,500 cm^{-1}, from ~1,700 to 2,450 cm^{-1}, and from 3,620 to 3,680 cm^{-1}. Other notable peaks observed for this cluster were from ~1,570 to 1,700 cm^{-1}, from ~2,820 to 3,150 cm^{-1}, and from ~3,460 to 3,530 cm^{-1}.

Significance profiles for gene clusters represented by *CSN3*, *PAEP* and *ANKH* are shown in Fig. 6.7. The pattern of significance in the profiles represented by *CSN3* and *PAEP* (Fig. 6.7a and 6.7b) were similar, in that a large proportion of the signal was captured within a small part of the mid-infrared range; namely from ~1,220 to 1,780 cm^{-1} for the gene cluster represented by *CSN3*, and from ~1,350 to ~1,650 cm^{-1} for the gene cluster represented by *PAEP*. Although *ANKH* appeared to be an outlier in the clustering analysis (Fig. 6.5), a similar pattern was observed with most of the signal captured within three prominent peaks in the range from ~1,260 to 1,620 cm^{-1}. Two of these peaks, centred at ~1,391 cm^{-1} and 1,582 cm^{-1} were in common with peaks observed for the *PAEP* profile. From the first cluster (Fig. 6.7a), *CSN3* was the only gene with a significant association for a predicted milk composition trait, namely PP. From the second cluster of genes (Fig. 6.7b), the *PAEP* and *CCDC57* genes had significant associations with the FP phenotype, whilst *ANKH* had a significant association with the LP phenotype (Fig. 6.7c).

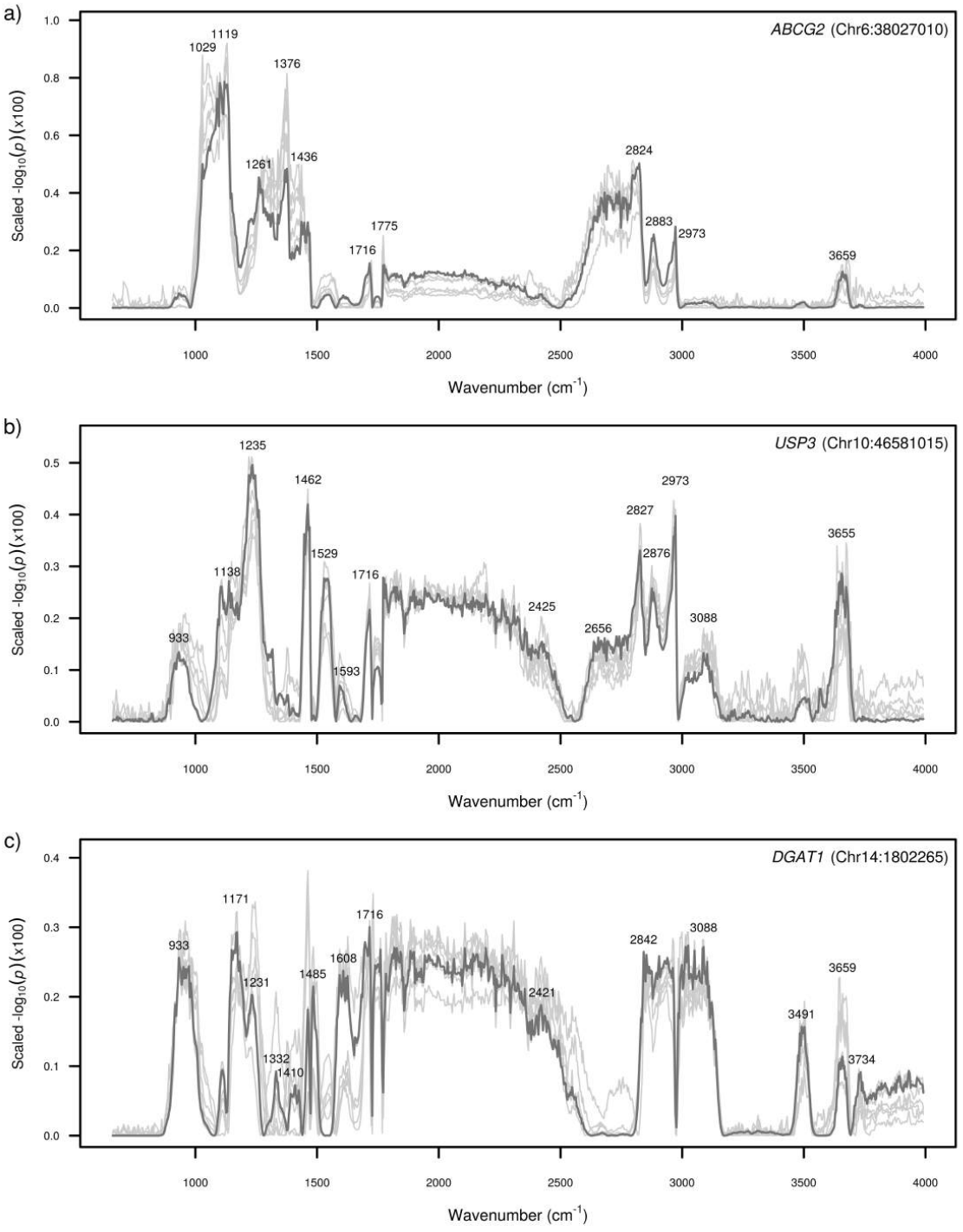

Figure 6.6: Significance profiles across the mid-infrared spectrum for tag variants of candidate genes within gene clusters. Y-axis values represent the strength of association signals as the -$\log_{10}$(p-value) of the effect, scaled to sum to unity across the mid-infrared spectral range. The significance profile for the most highly associated tag variant is shown in dark grey with the profiles for the other genes within the cluster shown in light grey: a) *ABCG2* (Chr6:38027010; dark grey), *SH3BP5* (Chr1:154125158), *RGL1* (Chr16:66314547), *PMP22* (Chr19:33517487), *KCNJ2* (Chr19:61134515), *PICALM* (Chr29:9546217); b) *USP3* (Chr10:46581015; dark grey), *ELAPOR1* (Chr3:34387618), *TBC1D22A* (Chr5:118246868), *FA2H* (Chr18:2203322), *STAT5A* (Chr19:43053995), *LTF* (Chr22:53519865), *LMAN1* (Chr24:58817202); and c) *DGAT1* (Chr14:1802265; dark grey), *FCRLA* (Chr3:7908611), *FCGR2B* (Chr3:7931694), *MGST1* (Chr5:93945738), *ABO* (Chr11:104242578), *TGDS* (Chr12:69612955), *MROH1* (Chr14:1732043), *CPSF1* (Chr14:1755742), *GPAT4* (Chr27:36211708)

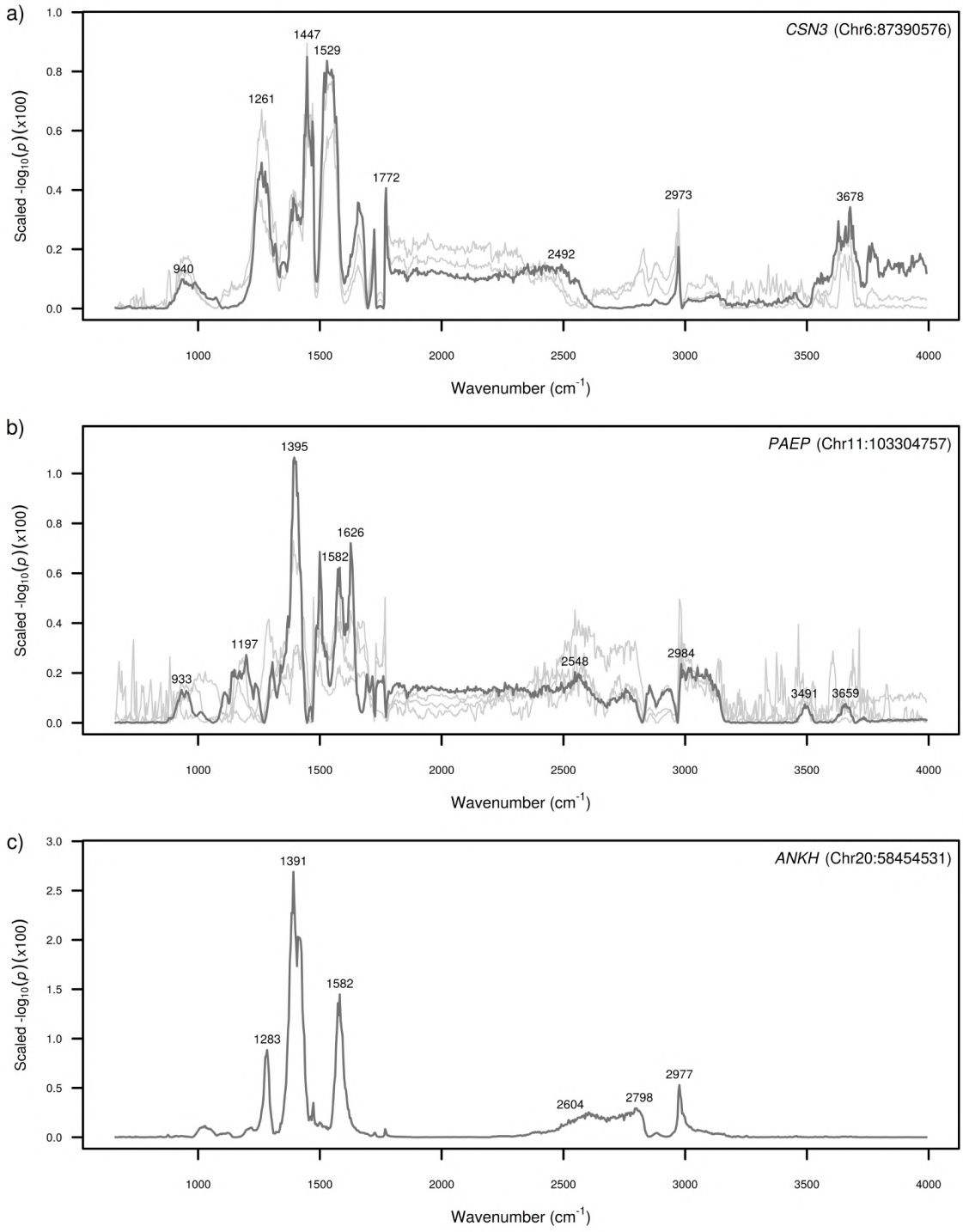

Figure 6.7: Significance profiles across the mid-infrared spectrum for tag variants of candidate genes within gene clusters. Y-axis values represent the strength of association signals as the -$\log_{10}$(p-value) of the effect, scaled to sum to unity across the mid-infrared spectral range. The significance profile for the most highly associated tag variant is shown in dark grey with the profiles for the other genes within the cluster shown in light grey: a) *CSN3* (Chr6:87390576; dark grey), *CSN1S1* (Chr6:87274397), *TAB2* (Chr9:87585031); b) *PAEP* (Chr11:103304757; dark grey), *MPC1* (Chr9:102874726), *WDR97* (Chr14:1726650), *CCDC57* (Chr19:51303887); and c) *ANKH* (Chr20:58454531)

6.5 Discussion

6.5.1 GWAS for FT-MIR wavenumbers

While there have been many GWAS for FT-MIR predicted milk composition traits, there are relatively few studies reporting GWAS results for individual FT-MIR wavenumber phenotypes. This is not withstanding the fact that individual wavenumbers exhibit additional genetic signal, compared to that observed in FT-MIR predictions of major milk composition traits (Wang and Bovenhuis, 2018; Benedet et al., 2019), and that direct analysis of the individual wavenumbers could provide additional granularity to establish causal links between the genome and underlying milk composition. Here, we present the results of GWAS that were conducted across individual FT-MIR wavenumber phenotypes, and the use of an iterative approach to help differentiate multiple, overlapping QTL. In total, wavenumber QTL were observed across 450 1-Mbp genomic regions, whereas predicted-trait QTL were observed across only 246 1-Mbp genomic regions. Notably, many of the observed wavenumber QTL were for wavenumbers within mid-infrared regions that were characterised by low signal-to-noise ratios. Typically, spectral data in these low signal-to-noise regions are discarded from analyses; however, these results indicate that wavenumbers in these regions are potentially informative. The signals that we observed in these noise regions were within several genes, with the highest frequency and strongest signals for variants in the *DGAT1* gene. This corroborates findings from previous studies which also observed significant associations between the *DGAT1* K232A polymorphism and wavenumbers in the regions from 1,619 to 1,674 cm^{-1} and from 3,073 to 3,667 cm^{-1} (Wang et al., 2016; Wang and Bovenhuis, 2018).

6.5.2 Multiple FT-MIR wavenumber QTL observed

In total, 31 wavenumber QTL were identified that we deemed to be 'highly significant' (see Methods for definition). Highly significant QTL were also observed for 12 of these same loci in at least one FT-MIR predicted milk composition trait, whereby the locus was in high LD ($R^2 > 0.9$) with the same PIV. The loci for the three largest of these effects were in perfect LD with missense mutations in the *ABCG2*, *PAEP* and *DGAT1* genes, respectively, that have been proposed to have major impacts on milk composition (Cohen-Zinder et al., 2005; Ganai et al., 2009; Grisart et al., 2002). Notably, the missense variant in the *ABCG2* gene identified here (rs43702337) is the same Y581S variant that was previously reported to be associated with milk yield and composition in Holstein cattle (Cohen-Zinder et al., 2005). The role of

the *ABCG2* mutation in milk composition regulation can be assumed to derive from osmotic impacts due to its function as an efflux transporter (Lopdell et al., 2017), although the gene has recently also been implicated in the modulation of mammary epithelial cell proliferation (Wei et al., 2012). The *PAEP* gene encodes the major whey protein β-lactoglobulin. The variant rs109625649 reported here (V134A) is one of the variants that distinguishes the 'A' and 'B' haplotypes of β-lactoglobulin (Caroli et al., 2009). The *PAEP* gene also exhibited an eQTL that was significantly correlated with wavenumber 2,548 cm^{-1}, which is concordant with previous reports of *PAEP* promotor variants associated with milk composition (Zakizadeh et al., 2012). The gene *DGAT1* encodes diacylglycerol O-acyltransferase 1, which catalyses the final step in triglyceride production and which, given the substantial quantities of fat secreted during milk production, makes *DGAT1* a well-demonstrated and straightforward candidate gene for this effect. The variant rs109234250 (K232A) reported here has been widely ascribed to the effects of the *DGAT1* gene on milk production, with a recent study showing that these effects may be due in part to an expression-based mechanism (Fink et al., 2020).

For the effects observed in the *ABCG2*, *PAEP* and *DGAT1* genes, the *p*-values for the most significant FT-MIR wavenumber were always more significant than the comparable values for any of the milk composition traits. For example, the *p*-value for the most significant wavenumber at the chr6:38027010 locus, the missense mutation in *ABCG2* highlighted above (Y581S, rs43702337) (Cohen-Zinder et al., 2005) was 7.3e-948, whereas the *p*-values for the same variant for LP and PP were 9.0e-717 and 6.4e-115, respectively. Similarly, the *p*-value for the most significant wavenumber at the chr11:103304757 locus, the V134A *PAEP* mutation (rs109625649) was 1.2e-134, whereas the *p*-value for the same variant for FP was 4.3e-46; and the *p*-value for the most significant wavenumber at the chr14:1802265 locus, represented by the K232A *DGAT1* mutation (rs109234250) (Grisart et al., 2002) was 1.5e-2607, whereas the *p*-value for the same locus for PP was 1.2e-61.

Multiple protein-coding mutations could be attributed to loci with QTL in both wavenumber and milk composition traits, highlighting genes that appear to be novel to the present study (*TGDS* and *DNAJB13*), and genes previously reported in other studies of milk composition traits: *GBA* (Jiang et al., 2019b; Raven et al., 2014), *MTX1* (Raven et al., 2016), *SLC50A1* (Jiang et al., 2018; Lopdell et al., 2017), *CSF2RB* (Kemper et al., 2015a; Lopdell et al., 2019b; Raven et al., 2016), *NCF4* (Lopdell et al., 2019b; Raven et al., 2016), *TBC1D22A* (Pausch et al., 2017), *MROH1* (Sanchez et al., 2017b) and *MTARC1* (Lopdell et al., 2017). A number of other QTL that were in strong LD with a PIV were observed in FT-MIR wavenumbers, but not in the FT-MIR predicted milk composition traits. This included QTL highlighting genes that have

been previously reported in other studies of bovine milk composition: *FCGR2B* (Jiang et al., 2019a), *SCAMP3* (Raven et al., 2016), *THBS3* (Raven et al., 2016), *CSN1S1*, *CSN2* and *CSN3* (Jiang et al., 2019b; Sanchez et al., 2017b), *ABO* (Liu et al., 2019; Poulsen et al., 2019), *CPSF1* (Buitenhuis et al., 2014; Cochran et al., 2013), *SPAG1* (Jiang et al., 2018), *RNF214* (Lopdell et al., 2017), *KAT2A* (Lopdell et al., 2017), *STAT5A* (Brym et al., 2004; He et al., 2012; Schennink et al., 2009) and *CCDC57* (Bouwman et al., 2014; Li et al., 2014); and QTL highlighting genes that appear novel: *FCRLA*, *WDR97*, *KRT9*, *KRT16*, *KRT17*, *HID1*, *KCNK1* and *NXF1*.

Although many regions highlighted single mutations that could be considered excellent candidate mutations for a given QTL, other loci presented more complex regions with multiple competing candidates. In some cases, candidate genes at these loci have previously been proposed; however, it is possible that one or more novel genes may explain minor QTL that map to the same positions. For example, the chr3:15.4-15.6 Mbp region which includes the genes *FAM189B*, *GBA*, *HCN3*, *MTX1*, *SCAMP3*, *THBS3* and *SLC50A1*; the chr14:1.7-1.8 Mbp region, which includes the genes *WDR97*, *MROH1*, *CPSF1*, *SLC52A2* and the *DGAT1* K232A amino acid substitution; the chr19:42.4-42.5 Mbp region which includes the genes *KRT9*, *KRT42*, *KRT16* and *KRT17*; and the chr19:43.0-43.1 Mbp region which includes the genes *KAT2A*, *KCNH4* and *STAT5A*. These regions might represent multiple, linked QTL, or instances of single QTL where the LD structure and our relatively simple approach for identifying candidate genes was ineffective at differentiating them. Another possibility is that wavenumbers in these regions detect the presence of multiple chemically-similar compounds, with milk concentrations being influenced by different proteins, such as enzymes or transporters that are encoded by different genes.

6.5.3 Co-localized eQTL suggest widespread regulatory impacts on FT-MIR wavenumbers

Of the 38 significant FT-MIR wavenumber QTL with co-localized eQTL, 18 also had co-localized eQTL that were observed for an FT-MIR predicted milk composition trait. In many cases, the tag variant for the wavenumber QTL was also the top variant for the co-localized eQTL. Genes corresponding to these effects have previously been published in other studies of bovine milk composition: *SH3BP5* (Lopdell et al., 2017), *SLC50A1* (Jiang et al., 2018; Lopdell et al., 2017), *USP3* (Fang et al., 2014; Wang et al., 2019a), *IVNS1ABP* (Lopdell et al., 2017), *KCJN2* (Lopdell et al., 2017), *ANKH* (Lopdell et al., 2017; Sanchez et al., 2017b), *GPAT4* (Littlejohn et al., 2014; Wang et al., 2012) and *PICALM* (Lopdell et al., 2017; Sanchez et al., 2017b). Other cases for which the wavenumber QTL was in high LD ($R^2 > 0.9$) with the top eQTL variant

highlighted genes previously published in other studies of bovine milk composition: *LRRC8C* (Lopdell et al., 2017), *CSF2RB* (Kemper et al., 2015a; Lopdell et al., 2019b; Raven et al., 2016), *MGST1* (Littlejohn et al., 2016; Pausch et al., 2017), *CSN3* (Jiang et al., 2019b; Sanchez et al., 2017b), *ABO* (Liu et al., 2019; Poulsen et al., 2019), *GHDC* (Lopdell et al., 2017; Raven et al., 2016) and *LMAN1* (Pausch et al., 2017; Sanchez et al., 2019); and genes that appear to be novel to the present study: *KRTCAP2* and *FA2H*. Pearson correlations between log-scaled p-values for the trait and expression QTL for the latter two effects were 0.94 and 0.88, respectively, with both displaying very strong LD between the trait QTL and the most highly significant eQTL variant ($R^2 > 0.995$).

Many wavenumber QTL with a co-localized eQTL also had a co-localized eQTL identified for a predicted milk composition trait. In these cases, a common pattern was observed whereby the wavenumber QTL had more highly significant p-values, compared to the p-values for the predicted trait. This was the case for *MGST1*, *ANKH*, *GPAT4* and *PICALM*. Notably, significant wavenumber QTL were detected for several additional milk proteins, with either highly-significant coding variants (*CSN1S1*, *CSN2*, *CSN3*) or a co-localized eQTL (*LTF*). To our surprise, only the *CSN3* eQTL was identified by analysis of the milk composition traits, with a p-value of 2.9e-45 for the PP phenotype, which was less significant than the p-value for the most significant wavenumber (p-value=1.1e-97).

Other wavenumber QTL where a co-localized eQTL was identified within FT-MIR wavenumbers, but not the predicted milk composition traits, included effects that highlighted a number of genes that appear novel to the present study: *CLDN8*, *CSTB*, *TAB2*, *LAPTM4A*, *CAPN5*, *PMP22*, *HID1* and *THRB*; and a number of genes previously reported as having an effect on bovine milk composition: *SLC37A1* (Kemper et al., 2016; Raven et al., 2016), *NCF4* (Lopdell et al., 2019b; Raven et al., 2016), *SLC34A2* (Liu et al., 2013), *TENT5A* (Li et al., 2014), *RNF217* (Jiang et al., 2018), *MPC1* (Sanchez et al., 2019), *XDH* (Ogorevc et al., 2009; Pegolo et al., 2016), *PAEP* (Ganai et al., 2009), *DGAT1* (Grisart et al., 2002), *RGL1* (Yodklaew et al., 2017), *LTF* (Mao et al., 2015; Viale et al., 2017) and *MUS81* (Reynolds et al., 2021). These results underscore the gain in power that is available when using individual FT-MIR wavenumber phenotypes, compared to using predicted milk composition phenotypes which are linear functions of FT-MIR absorbance values.

6.5.4 Candidate causative variants of note

Although we identified a large number of candidate causative variants for FT-MIR wavenumbers and predicted milk composition phenotypes, variants in perfect LD with a tag locus (R^2=1) warrant further discussion. These associations presented missense variants for genes mentioned previously (*ABCG2*, *PAEP* and *DGAT1*), in addition to other genes that have previously been linked to bovine milk composition phenotypes (*CSN2*, *CSN3*, *ABO*, *SPAG1* and *STAT5A*). Of these, the *ABO* exon 5 splice donor mutation (rs207688357; chr11:104242578C>G) is a particularly interesting and seemingly novel candidate causative variant identified through our GWAS of FT-MIR wavenumbers.

The rs207688357 variant was selected as the representative peak tag variant for 11 wavenumbers, with the most significant peak association observed for wavenumber 1,462 cm^{-1}. Visualisation of RNA-seq alignments confirmed that this variant disrupts splicing in carrier and homozygous animals (Fig. 6.4), where the mutation appears to activate two cryptic splice sites. The first and comparatively higher expressed form of these alternative transcripts is a -8-bp frameshifted isoform predicted to lead to premature termination, while the lowly expressed in-frame form is predicted to introduce 11 new amino acids following the 78th residue (due to a +33bp exon 5 extension). In humans, *ABO* has a widely recognised role as encoding the glycosyltransferases that catalyse the synthesis of the oligosaccharide ABO blood group antigens (Kermarrec et al., 1999; Yamamoto et al., 1990). Since both the alternatively spliced forms of bovine *ABO* generated by rs207688357 could be assumed to be non-functional (or at least dysfunctional for the minority in-frame isoform), this mutation would be akin to the human O blood group in homozygotes, where analogous human null alleles generate a non-functional enzyme (Chester and Olsson, 2001). These antigens are best known due to their expression on the surface of red blood cells, although they are also expressed on epithelial cells, as well as appearing as free oligosaccharides in milk (Le Pendu, 2004). This finding suggests a mechanism by which non- or partially-functional bovine *ABO* alleles change carbohydrate structures in milk, therefore presenting differing FT-MIR signals detected by GWAS.

It should also be noted that although we are unaware of other studies proposing the rs207688357 (chr11:104242578) mutation as underlying such effects, other studies have reported genetic associations for bovine milk oligosaccharides for the broader *ABO* locus (Liu et al., 2019; Poulsen et al., 2019). One of these studies proposed an *ABO* p.Arg206Gln (R206Q; chr11:104232763; rs110960674) amino acid substitution present on the Illumina BovineHD chip as a potential causative mutation for this effect (Poulsen et al., 2019). The other study reported associations

with non-coding variants downstream of the *ABO* coding sequence (lead variant chr11:104229609; rs110534892), in this case using imputed sequence-based genotypes (Liu et al., 2019). Both the p.Arg206Gln variant and the non-coding rs110534892 variant are also significant in our population, alongside the rs207688357 splice donor mutation, with peak association observed for the 1,462 cm^{-1} wavenumber phenotype. These alternative candidates are less strongly associated than the rs207688357 splice donor mutation (p-value = 1.1e-23 and 1.8e-28, for the p.Arg206Gln and rs110534892 variants, respectively, compared to 5.5e-33). While these findings might suggest that these variants are simply linked to the functionally more compelling rs207688357 splice donor mutation, LD between the variants and the splice donor mutation is moderate to low (R^2=0.486 and R^2=0.296 for the p.Arg206Gln and rs110534892 variants, respectively). Furthermore, when fitting the rs207688357 splice donor mutation as a covariate in the iterative association analysis of wavenumber 1,462 cm^{-1}, both variants retain residual signal (p-values of 4.2e-04 and 1.4e-07 for the p.Arg206Gln and rs110534892 variants, respectively), which suggests that all three variants might contribute to the oligosaccharide content of milk. In support of this concept, we also note that the non-coding rs110534892 variant proposed by Liu et al. (2019) is in strong LD with the lead variant representing a strong *ABO* eQTL highlighted in our study (R^2=0.944; Table 6.3). By contrast, the splice donor mutation is comparatively modestly associated with *ABO* expression at the whole transcript level (p-value = 9.1e-11 versus 5.9e-27), which suggests that multiple molecular mechanisms (missense, non-sense, and *cis*-regulatory effects) might contribute to oligosaccharide modulation at this locus.

6.5.5 FT-MIR wavenumber association patterns for genes of interest

Although FT-MIR spectroscopy is a valuable tool for predicting a range of milk composition traits, there are limitations to the approach, i.e., it is often unable to detect molecules that are present in small quantities, and does not discriminate well between compounds that are chemically similar. Nevertheless, we have demonstrated that individual FT-MIR wavenumber phenotypes can provide valuable insights for establishing causal links between the genome and milk composition. Having observed patterns of association across multiple FT-MIR wavenumbers for individual loci (i.e., genome positions that appeared to highlight specific subsets of wavenumbers), our aim was to formally detect these patterns of association through cluster analysis. We hypothesised that the identified clusters could be rationalised based on shared biology or the physico-chemical properties of the encoded molecules – given that these signatures would presumably reflect common functions and structures in milk.

The cluster with the largest number of individual attributed loci included genes with prominent roles in the regulation of fat synthesis such as *DGAT1*, *GPAT4*, and *MGST1* (Fig. 6.5). These three loci have been implicated in previous studies of milk fat percentage and fatty acid synthesis (Grisart et al., 2002; Littlejohn et al., 2014, 2016; Pausch et al., 2017; Schennink et al., 2007; Wang et al., 2012). *DGAT1* and *GPAT4* encode acyltransferase enzymes that are responsible for mammary triglyceride synthesis, so it seems likely that the highlighted cluster reflects wavenumbers that are sensitive to changes in milk fat content. Notably, the pattern of the effects observed for *DGAT1* (Fig. 6.6c) was very similar to those reported previously (Wang et al., 2016; Zaalberg et al., 2020). Highly significant effects were observed for the *DGAT1* K232A polymorphism in bands of the spectrum that could be attributed to a number of different chemical bond interactions including: phosphorus compounds (from ~910 to 1,010 cm^{-1}) (Fleming and Williams, 2019), triglyceride ester C-O stretching (from ~1,130 to 1,260 cm^{-1}) (Karoui et al., 2003; Safar et al., 1994), C-H bending vibrations of $–CH_2$ and $–CH_3$ (from ~1,450 to 1,500 cm^{-1}) (Fleming and Williams, 2019; Grelet et al., 2015), C=O stretching in polypeptides within the amide I band of protein (from ~1,600 to 1,700 cm^{-1}) (Safar et al., 1994), carboxylic acid and C=O rotation and stretching of ester groups of fat (from ~1,700 to 1,800 cm^{-1}) (Lefier et al., 1996), and acyl chain C-H stretching (from ~2,820 to 3,150 cm^{-1}) (Karoui et al., 2003).

The cluster that included the *ABCG2* Y581S polymorphism (Fig. 6.5) had highly significant association effects across numerous FT-MIR wavenumbers, with the largest effects concentrated within the regions from ~1,020 to 1,470 cm^{-1} and from ~2,610 to 2,980 cm^{-1} (Fig. 6.6a). Bands of the mid-infrared spectrum related to the largest effects for the *ABCG2* Y581S polymorphism were attributable to hydroxyl groups related to lactose (from ~1,020 to 1,180 cm^{-1}) (Fleming and Williams, 2019; Picque et al., 1993), amide III and phosphate bands (from ~1,200 to 1,390 cm^{-1}) (Hewavitharana and van Brakel, 1997; Safar et al., 1994), C-H bending vibrations for CH_2 and $–CH_3$ (from ~1,410 to 1,470 cm^{-1}) (Fleming and Williams, 2019), overtones and bands of lactose (~2,600 upwards) (Luinge et al., 1993), and C-H stretching vibrations of CH_2 and $–CH_3$ (from ~2,700 to 2,980 cm^{-1}) (Fleming and Williams, 2019). Many of the mid-infrared bands with significant effects were ascribed to chemical bond interactions related to lactose, which is unsurprising, given that *ABCG2* and many of the other genes classified in the same cluster (*SH3BP5*, *PMP22*, *KCNJ2*, and *PICALM*) have been previously associated with lactose phenotypes (Lopdell et al., 2017; Sanchez et al., 2017b). Notably, the strongest association effects for the *ABCG2* Y581S polymorphism were in different bands of the mid-infrared spectrum to the *DGAT1* K232A polymorphism, assumedly reflecting the different roles that these two genes play in altering milk composition.

Three other notable gene clusters were those represented by the *CSN3*, *PAEP* and *ANKH* genes (Fig. 6.7), which had a large proportion of significant signal captured within a small part of the mid-infrared range: *CSN3* (from ~1,220 to 1,780 cm^{-1}), *PAEP* (from ~1,350 to 1,650 cm^{-1}) and *ANKH* (from ~1,260 to 1,620 cm^{-1}). The *CSN3* gene encodes κ-casein, one of the most abundantly expressed proteins in milk. Bound at the aqueous-hydrophobic interface of casein micelles, κ-casein content influences the size of these structures, thereby affecting various coagulation and cheese-making properties (Creamer et al., 1998; Poulsen et al., 2013). The missense mutation reported here at chr6:87390576 (rs43703015) has been associated with milk composition traits and differential expression in mammary tissue (MacLeod et al., 2016). The largest effects for the *CSN3* locus were in spectral regions related to amide III and phosphate bands (from ~1,220 to 1,320 cm^{-1}), C-H stretching vibrations of CH_2 and $-CH_3$ (from ~1,370 to 1,480 cm^{-1}), and N-H bending and C-N stretching in the amide II band (from ~1,490 to 1,590 cm^{-1}) (Garidel and Schott, 2006). Previous studies have reported association effects for *CSN3* in similar bands of the mid-infrared spectrum, with specific wavenumbers coinciding with highly significant association effects observed in our study (Benedet et al., 2019; Wang et al., 2016; Zaalberg et al., 2020). The *ANKH* gene encodes a transmembrane protein involved in pyrophosphate transport regulation, and is associated with lactose concentrations in milk (Lopdell et al., 2017; Sanchez et al., 2017b). Interestingly, *ANKH* and *PAEP* shared a prominent peak for adjacent wavenumbers, 1,391 cm^{-1} and 1,395 cm^{-1}, respectively. These wavenumbers were in a region related to carboxylic acid C=O bond stretching (Fleming and Williams, 2019). Another peak in common between these genes was centred on the 1,582 cm^{-1} wavenumber, also in a region related to carboxylic acid C=O bond stretching (Fleming and Williams, 2019). Association effects in similar bands of the mid-infrared spectrum for *PAEP* have been reported in previous studies (Benedet et al., 2019; Wang et al., 2016; Zaalberg et al., 2020). Although *ANKH* and *PAEP* shared peaks in their significance profiles, it is notable that they also had exclusive peaks. For *ANKH*, a distinct peak was observed in a region related to amide III and phosphate bands (from ~1,270 to 1,290 cm^{-1}) (Hewavitharana and van Brakel, 1997; Safar et al., 1994), and for *PAEP* a distinct peak was observed in a region related to C-NH peptide bonds and N-H stretching and bending vibrations of NH_2 (from ~1,600 to 1,640 cm^{-1}) (Dufour, 2009; Fleming and Williams, 2019), which shows that although commonalities exist, there are also differences in the roles that these genes play in altering milk composition.

6.5.6 Limitations of the present study and future perspectives

In this study, we demonstrated that GWAS conducted on individual FT-MIR wavenumbers can improve power for identifying QTL and candidate causal variants, compared to GWAS conducted on FT-MIR predicted milk composition traits. Although many QTL were successfully identified, several refinements to our approach could be expected to enable the identification of additional QTL. The first of these relates to the approach used in adjusting phenotypes prior to conducting the GWAS. The repeated measures model that we used for adjusting phenotypes included a random effect to capture individual animal variation, but did not use pedigree information to account for covariance between individuals. This means that genetic trend may have been captured in herd by test day effects. A more optimal, but computationally more expensive approach, would have been to fit a repeatability model including the additive relationship matrix, thereby ensuring more accurate partitioning of fixed and random effects. To assess the potential impact of this on the final GWAS results in our study, we generated adjusted phenotypes for FP, LP and PP using a full animal model with an additive relationship matrix, and compared these to the adjusted phenotypes evaluated from the simplified repeated measures model we report. The correlations between the adjusted phenotypes from the two models were all high: 0.983, 0.994 and 0.987 for FP, LP and PP respectively. This implies that although the model that we used may be considered suboptimal, it is likely that the use of this model would have only a very minor impact on the final GWAS results.

Other potential refinements to our approach specifically relate to genomic information and our strategy for identifying QTL. First, our study relied on datasets that were mapped to the UMD3.1 genome, whereas a newer reference genome (ARS-UCD1.2) that has improved sequence continuity and per-base accuracy (Rosen et al., 2020) is now available. Future use of that reference genome might yield additional QTL, as well as reveal additional candidate mutations given the improvements in accompanying transcript annotations. Second, our approach could be extended to account for non-additive QTL. Recently, we conducted non-additive association mapping of growth and development traits in cattle, which highlighted a number of major-effect mutations that had not been identified through application of standard additive models (Reynolds et al., 2021). Although the low MAF variants identified in that study would require larger samples than those explored here, future analyses based on larger populations might be expected to identify similar non-additive effects for FT-MIR wavenumber and predicted milk composition traits. Third, a more sophisticated methodology could be used for the selection of representative variants in each QTL peak. In our approach, we have iteratively taken the top variant from each peak based

on the p-value of the association effect, and fitted this as a covariate in subsequent rounds of GWAS. This approach does not take nonlinear interactions between variants into account, and can lead to the selection of multiple variants in high LD with a single QTL, if that QTL is not itself represented by a single biallelic variant. Alternatively, multiple QTL at a single locus might be best tagged by a single, non-causal variant that captures multiple signals. In both these instances, factors such as imputation or genotyping error may also further compound these issues. To address this, a modified approach could be adopted, whereby gene annotation information and other genomic and molecular data sources are used to assist with variant selection. Finally, although we tried to identify causal variants representing a variety of molecular mechanisms including coding variants (missense and non-sense) and regulatory effects (through integration of mammary eQTL data), these approaches are far from comprehensive, and will still miss many candidates. Improved variant prediction methods, and generation of other functional datasets (e.g., ChIP-seq) could be used to map additional molecular QTL, where integration of those data would enhance fine mapping and identification of candidate variants (Tiplady et al., 2020).

6.6 Conclusions

We conducted a sequence-based GWAS on individual FT-MIR wavenumber phenotypes, and employed gene annotation and mammary tissue gene expression datasets to identify candidate causative genes and variants. Compared to GWAS on predicted milk composition traits, GWAS on individual FT-MIR wavenumbers resulted in stronger association effects, and improved power for identifying candidate causal variants. Although many of the genomic regions with significant associations that we identified in this work have previously been linked to milk composition traits, we report the discovery of several loci that have never previously been linked to milk phenotypes. Examining patterns of significance across wavenumbers in the mid-infrared range for loci of interest provided further insights into the relationships between specific genes and the underlying chemical structure of milk. Leveraging this information and incorporating the candidate causative mutations that we have identified into genomic prediction could result in improved selection of dairy cattle for the ever-growing range of traits of interest to the industry.

6.7 Declarations

6.7.1 Ethics statement

All data were generated as part of routine commercial activities that were outside the scope of activities requiring formal ethics approval. No animals were sacrificed for this study.

6.7.2 Acknowledgements

The authors gratefully acknowledge LIC (Hamilton, New Zealand) herd-testing staff for the processing and analysis of milk samples, and the LIC team for the processing and analysis of genotypes. Kathryn would also like to thank Tracey Monehan (R&D Programme Manager, LIC) for overseeing the funding for this work, and the wider LIC team and fellow students for helpful discussions and underlying technical support. The authors also gratefully acknowledge Tod Schilling (Bentley Instruments Inc., Chaska, USA) and Pierre Broutin (Bentley Instruments Inc., Lille, France) for assistance with accessing FT-MIR spectra from Bentley instruments. Finally, we acknowledge our gratitude for the use of New Zealand eScience Infrastructure (NeSI) high-performance computing.

6.7.3 Funding

This research was co-funded by Livestock Improvement Corporation (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). External funders had no role in the analysis or interpretation of the data, or in writing the manuscript.

6.7.4 Availability of data and materials

Phenotypic data representing individual FT-MIR wavenumbers and FT-MIR predicted milk composition traits has been submitted to the Dryad Digital Repository (Tiplady et al., 2021a; doi:10.5061/dryad.qrfj6q5dj). Genotypes for tag variants representing trait QTL have also been uploaded under the same Dryad submission ID. Relevant eQTL for genes with co-localized trait and expression QTL peaks are available through the Dryad database portal (Lopdell et al., 2018). Whole-genome sequences used for imputation of the genotypes presented in this paper have been deposited in the SRA database (PRJNA656361 Cattle whole genome sequences, 2021). Additional data is available on reasonable request with the permission of Livestock Improvement Corporation, contingent on the execution of an appropriate transfer agreement.

Appendices

6.A Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle

Table 6.A.1: Peak variants for FT-MIR wavenumbers with moderately significant protein-sequence association effects

Chr	Position	Tag variant ID	No. of hits	Top wvn cm^{-1}	Iter	*P*-value	Protein coding variant ID	LD	Gene	Impact	Description
1	143,874,528	rs210853796	1	2544	1	4.50E-13	rs210853796	1.000	*UMODL1*	M	c.1982C>T
1	144,192,782	rs208749212	2	2954	1	8.10E-14	rs208749212	1.000	*TFF1*	M	c.224G>T
3	54,258,151	rs208197051	7	1130	0	5.20E-18	rs209968430	0.970	*ENS..14857*	L	c.301-8G>A
3	54,258,151	rs208197051	7	1130	0	5.20E-18	rs207922427	0.971	*ENS..17670*	M	c.655A>C
3	54,258,151	rs208197051	7	1130	0	5.20E-18	rs209038383	0.972	*GBP5*	M	c.9G>A
5	27,835,741	rs135284663	10	1085	2	3.40E-18	rs110130901	1.000	*KRT7*	M	c.908A>G
5	112,121,312	rs209287637	2	2671	1	5.60E-16	rs209287637	1.000	*TNRC6B*	M	c.797G>A
6	86,335,083	rs462124622	9	1369	3	3.70E-22	rs384429777	0.973	*ENS..03523*	L	c.1014+3A>G
6	86,335,083	rs462124622	9	1369	3	3.70E-22	rs382793163	0.961	*UGT2B10*	M	c.1406G>A
6	87,534,452	rs433620526	1	2973	3	3.80E-16	rs438534592	0.962	*CABS1*	M	c.810C>A
6	88,851,143	rs384262616	2	1276	3	2.80E-28	rs110326785	0.913	*NPFFR2*	M	c.1174G>A
10	46,514,558	rs109205360	18	1235	0	1.70E-48	rs137170204	0.904	*HERC1*	L	c.14079+3G>A
10	46,514,558	rs109205360	18	1235	0	1.70E-48	rs135013504	0.904	*USP3*	L	c.1332+5T>C
12	72,110,595	rs42593480	2	3674	0	1.90E-27	rs42593493	0.941	*ENS..45751*	L	c.1659+4C>T
14	1,802,265	rs109234250	310	1716	0	1.5e-2607	rs135258919	0.903	*HSF1*	M	c.1031T>C
14	1,842,932	rs135134051	1	3189	0	3.80E-15	rs135134051	1.000	*BOP1*	L	c.1215+6G>C
15	52,197,561	rs480295644	1	1555	1	5.30E-14	rs41768351	0.931	*CHRNA10*	M	c.1085G>A
15	53,940,444	rs382926661	23	1205	1	4.20E-19	rs380813700	0.941	*ARHGEF17*	M	c.1756G>A
16	1,455,243	rs380323951	10	1216	1	7.60E-36	rs379734240	0.904	*ZC3H11A*	L	c.503-7A>C
16	60,784,946	rs209235878	6	1451	4	8.40E-17	rs42733251	0.992	*SEC16B*	M	c.1907G>A
17	70,309,223	rs42288202	6	1276	0	1.00E-17	rs42288202	1.000	*HSCB*	L	c.424-3T>C
18	2,203,322	rs132899112	7	1466	1	1.40E-15	rs134184381	0.997	*FA2H*	L	c.271-8T>C
19	33,515,473	rs382520566	3	1399	4	3.30E-15	rs210304540	0.995	*CDRT4*	L	c.-38G>T
19	42,604,860	rs134093156	18	1462	0	1.90E-39	rs209920132	0.918	*ACLY*	L	c.1846-3T>C
19	42,604,860	rs134093156	18	1462	0	1.90E-39	rs209373086	0.919	*JUP*	L	c.1055-4C>G
19	43,036,265	rs210324533	11	1029	1	5.30E-43	rs381010891	0.921	*ZNF385C*	M	c.628C>G
19	51,515,451	rs41925642	22	2984	1	2.50E-18	rs41925642	1.000	*ENS..47973*	M	c.298A>C
25	36,089,539	rs210232064	10	1029	0	2.80E-16	rs210065065	0.915	*PLOD3*	M	c.432G>C
26	21,098,102	rs479414226	9	1138	0	5.20E-20	rs454657689	0.970	*PKD2L1*	M	c.325G>A
26	22,719,393	rs209022793	13	1205	1	1.30E-27	rs110483942	0.902	*GBF1*	M	c.3143C>T
29	44,579,245	rs439384463	3	1548	2	4.30E-16	.	0.924	*DPF2*	M	c.647A>G
29	47,036,875	rs379471283	1	1723	1	6.60E-14	rs208818475	0.988	*TPCN2*	L	c.894C>T

Peak variants of 27 protein-sequence association effects classified as moderately significant for FT-MIR wavenumber phenotypes. Moderately significant effects are those for which the -$\log_{10}$(*p*-value) of the effect was greater than 1 x the Bonferroni threshold of -$\log_{10}$(6.2e-13) and the correlation between the tag variant and the protein-sequence variant was higher than 0.9, but the effect did not meet the criteria of a highly significant effect (see Table 6.1). Effects where the locus has been identified as highly significant based on the LD with one or more other genes (and is also present in Table 6.1) are shaded yellow. No. of hits: number of wavenumbers for which the variant was selected as the representative (most significant) tag variant for a peak. Iterations (Iter) are defined relative to the base GWAS, with the base GWAS represented as iteration 0.
Abbreviations: L = Low impact splice region variant; M = Moderate impact missense variant; H = High impact splice donor.

Table 6.A.2: Minor allele frequencies and allele effects for whole-genome sequence tag variants with a highly significant protein-sequence association effect in at least one FT-MIR wavenumber

Chr	Position	Tag variant ID	Minor allele frequency	Top wavenumber	Beta (x1000)	SE (x1000)	*P*-value
3	7,908,611	rs137763930	0.1518	940	0.4188	0.0459	6.70E-20
3	7,931,694	rs211402696	0.2395	1462	-0.3167	0.0316	1.20E-23
3	15,411,459	rs134900385	0.1263	1022	-0.5463	0.0612	4.30E-19
3	15,517,871	rs109328483	0.1250	1007	-0.3493	0.0391	4.40E-19
3	15,550,598	rs380597285	0.4784	1462	-0.5088	0.0327	1.30E-54
5	75,729,880	rs384734208	0.4177	1466	-0.3909	0.0271	5.00E-47
5	75,758,989	rs210094995	0.1668	1447	-0.4454	0.0337	5.80E-40
5	118,246,868	rs136859160	0.1561	1261	-0.3272	0.0235	3.00E-44
6	38,027,010	rs43702337	0.0042	1119	-19.7003	0.2986	7.3e-948
6	87,181,619	rs43703011	0.2815	3633	0.2972	0.0306	2.50E-22
6	87,274,397	rs378808772	0.3338	1283	0.2153	0.0144	9.90E-51
6	87,390,576	rs43703015	0.3452	1473	-0.3384	0.0153	4.00E-108
11	103,304,757	rs109625649	0.4996	1593	-1.4129	0.0572	1.20E-134
11	104,242,578	rs207688357	0.2629	1462	-0.4434	0.0371	5.50E-33
12	69,612,955	rs383509255	0.2383	1716	-0.4691	0.0334	6.40E-45
14	1,726,650	rs133611586	0.0247	3514	2.8141	0.1530	1.60E-75
14	1,732,043	rs437406031	0.3740	2846	1.3614	0.1003	6.30E-42
14	1,755,742	rs384226556	0.4814	2656	0.0497	0.0054	4.00E-20
14	1,802,265	rs109234250	0.4294	1716	2.8769	0.0263	1.5e-2607
14	66,328,304	rs446084949	0.0069	1029	2.4117	0.2612	2.70E-20
15	28,347,165	rs210034037	0.0799	1537	1.4247	0.1157	7.70E-35
15	53,940,444	rs382926661	0.1019	1205	-0.2217	0.0248	4.20E-19
16	24,977,696	rs111027377	0.3124	2742	-0.0848	0.0082	4.80E-25
19	42,428,366	rs209808022	0.0871	1250	-0.3449	0.0332	3.10E-25
19	42,488,389	rs379667889	0.1000	1447	-0.5064	0.0418	7.80E-34
19	43,036,265	rs210324533	0.0846	1029	1.0710	0.0779	5.30E-43
19	43,053,995	rs481837688	0.1496	1212	0.2164	0.0206	6.60E-26
19	51,303,887	rs41921224	0.2525	1499	-0.4115	0.0331	1.90E-35
19	57,087,981	rs41920620	0.4825	1216	-0.1445	0.0152	1.80E-21
28	6,559,147	rs133101552	0.4215	1261	0.1653	0.0168	8.60E-23
29	41,821,270	rs207854419	0.1077	1257	-0.2990	0.0262	4.60E-30

Table 6.A.3: Peak variants for composite milk production traits with moderately significant protein-sequence association effects

Trait	Chr	Position	Tag variant ID	Iteration	*P*-value	Protein coding variant ID	LD	Gene	Impact	Description
FP	5	118,246,868	rs136859160	2	1.40E-16	rs456403270	0.937	*TBC1D22A*	M	c.1063C>T
FP	14	1,800,439	rs209876151	0	8.9e-2225	rs135258919	0.904	*HSF1*	M	c.1031T>C
FP	19	42,604,653	rs135024837	1	6.70E-19	rs135261291	0.928	*JUP*	L	c.1925-7C>T
FP	19	51,303,887	rs41921224	0	3.80E-15	rs41921160	0.993	*CCDC57*	M	c.1907T>C
LP	5	27,835,741	rs135284663	1	1.60E-15	rs110130901	1.000	*KRT7*	M	c.908A>G
LP	19	43,036,265	rs210324533	3	9.40E-40	rs381010891	0.921	*ZNF385C*	M	c.628C>G
PP	3	15,520,971	rs109098377	2	7.50E-16	rs382689947	0.991	*FAM189B*	M	c.1237T>C
PP	3	15,520,971	rs109098377	2	7.50E-16	rs134844772	0.990	*GBA*	M	c.1080C>A
PP	3	15,520,971	rs109098377	2	7.50E-16	rs109330809	0.990	*MTX1*	L	c.508-6T>C
PP	3	15,520,971	rs109098377	2	7.50E-16	rs136761456	0.991	*SCAMP3*	M	c.151G>C
PP	3	15,520,971	rs109098377	2	7.50E-16	rs43706482	0.993	*THBS3*	L	c.2075-3T>C
PP	14	1,763,380	rs135017891	0	5.9e-718	rs109326954	0.904	*DGAT1*	M	c.695C>A
PP	14	1,855,915	rs379497765	4	7.00E-14	rs476736066	0.999	*MROH1*	M	c.3549G>C
PP	15	53,940,444	rs382926661	1	2.90E-20	rs380813700	0.941	*ARHGEF1'*	M	c.1756G>A
PP	16	60,784,946	rs209235878	4	7.50E-17	rs42733251	0.992	*SEC16B*	M	c.1907G>A
PP	18	2,203,325	rs135350753	0	2.10E-13	rs134184381	0.997	*FA2H*	L	c.271-8T>C
PP	19	43,035,006	rs209494359	0	1.60E-40	rs381010891	0.921	*ZNF385C*	M	c.628C>G
PP	19	43,053,995	rs481837688	2	8.60E-17	rs481837688	1.000	*STAT5A*	M	c.2305C>A

Peak variants of 14 protein-sequence association effects classified as moderately significant for FT-MIR predicted milk composition traits. Moderately significant effects are those where the -$\log_{10}$(p-value) of the effect was greater than 1 x the Bonferroni threshold of -$\log_{10}$(6.2e-13) and the correlation between the tag variant and the protein-sequence variant was higher than 0.9, but the effect did not meet the criteria of a highly significant effect (see Table 6.2). Effects where the locus has been identified as highly significant based on the LD with one or more other genes (and is also present in Table 6.2) are shaded yellow. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0.
Abbreviations: FP = Fat %; LP = Lactose %; PP = Protein %; L = Low impact splice region variant; M = Moderate impact missense variant; H = High impact splice donor.

Table 6.A.4: Minor allele frequencies and allele effects for whole-genome sequence tag variants with a highly significant protein-sequence association effect in at least one FT-MIR predicted milk composition trait

Chr	Position	Tag variant ID	Minor allele frequency	Trait	Beta	SE	*P*-value
3	15,433,518	rs109749506	0.1263	LP	-0.0181	0.0019	1.30E-20
3	15,545,091	rs379353107	0.4787	LP	0.0165	0.0012	2.20E-42
3	15,550,598	rs380597285	0.4784	PP	-0.0265	0.0021	1.70E-37
5	75,698,283	rs385866519	0.4049	FP	-0.0398	0.0045	4.00E-19
5	75,758,989	rs210094995	0.1668	PP	-0.0337	0.0028	3.30E-34
5	118,239,754	rs384479185	0.1488	PP	-0.0334	0.0028	3.90E-32
6	38,027,010	rs43702337	0.0042	LP	-0.5339	0.0093	9.0e-717
6	38,027,010	rs43702337	0.0042	PP	-0.3652	0.0160	6.40E-115
11	103,304,757	rs109625649	0.4996	FP	0.0720	0.0050	4.30E-46
12	69,608,900	rs211406918	0.2370	FP	0.0709	0.0059	4.20E-33
14	1,732,043	rs437406031	0.3741	FP	0.0782	0.0062	7.20E-37
14	1,763,380	rs135017891	0.4542	PP	-0.1146	0.0020	5.9e-718
14	1,800,439	rs209876151	0.4293	FP	-0.4675	0.0046	8.9e-2225
14	1,802,265	rs109234250	0.4233	PP	-0.1194	0.0072	1.20E-61
15	53,940,444	rs382926661	0.1030	PP	-0.0341	0.0037	2.90E-20
16	24,983,926	rs110162358	0.3098	LP	-0.0113	0.0012	1.00E-19
19	43,035,006	rs209494359	0.0846	PP	-0.0493	0.0037	1.60E-40
19	43,036,265	rs210324533	0.0846	LP	0.0263	0.0020	9.40E-40

Abbreviations: FP = Fat %; LP = Lactose %; PP = Protein %.

Table 6.A.5: Minor allele frequencies and allele effects for whole-genome sequence tag variants with a significant association effect in FT-MIR wavenumbers and a co-localized eQTL

Chr	Position	Tag variant ID	Minor allele frequency	Top wavenumber	Beta (x1000)	SE (x1000)	*P*-value
1	5,120,248	rs42317521	0.2643	2794	0.0925	0.0107	4.00E-18
1	144,377,960	rs208161466	0.2474	2592	-0.0977	0.0050	2.30E-85
1	146,481,250	rs383691757	0.4870	1071	0.4008	0.0512	4.70E-15
1	154,125,158	rs207836083	0.2297	1130	0.586	0.0336	3.60E-68
3	15,411,459	rs134900385	0.1263	1022	-0.5463	0.0612	4.30E-19
3	15,550,598	rs380597285	0.4784	1462	-0.5088	0.0327	1.30E-54
3	34,387,618	rs109030498	0.3017	1466	0.3052	0.0294	3.50E-25
3	53,755,929	rs209271975	0.3108	1089	-0.4860	0.0507	8.60E-22
5	75,729,880	rs384734208	0.4177	1466	-0.3909	0.0271	5.00E-47
5	75,732,526	rs210305241	0.4119	1458	-0.3931	0.0289	4.40E-42
5	93,945,738	rs211210569	0.3393	1171	-1.5185	0.0622	1.80E-131
6	46,568,418	rs210515595	0.3680	1772	-0.0487	0.0051	7.70E-22
6	87,388,064	rs379473589	0.3115	1436	-0.4089	0.0195	1.10E-97
9	21,637,056	rs209222932	0.2412	1003	-0.2801	0.0308	9.40E-20
9	26,534,109	rs208123385	0.2709	1462	0.3742	0.0367	1.90E-24
9	87,585,031	rs110986237	0.4760	1470	-0.1477	0.019	7.40E-15
9	102,874,726	rs137238900	0.1260	1768	0.0653	0.0084	1.00E-14
10	46,581,015	rs109326466	0.2011	1246	0.4317	0.0302	2.00E-46
11	14,180,010	rs110527112	0.3997	2760	0.0937	0.0084	3.60E-29
11	78,868,975	.	0.4700	1112	-0.3273	0.0361	1.20E-19
11	103,292,402	rs383398415	0.4990	2548	-0.0745	0.0047	3.50E-56
11	104,229,609	rs110534892	0.4536	3648	-0.4614	0.0483	1.20E-21
14	1,754,287	rs135443540	0.4938	1085	-0.6561	0.0499	1.60E-39
15	57,266,467	rs136337092	0.3748	3935	-0.2425	0.0332	2.70E-13
16	66,314,547	rs42579412	0.4307	1425	-0.1758	0.0219	1.00E-15
16	67,730,371	rs380453838	0.3392	1757	-0.6245	0.0662	3.80E-21
18	2,203,322	rs132899112	0.2848	1466	-0.2376	0.0298	1.40E-15
19	33,517,487	rs434248431	0.2931	1100	0.6755	0.0473	2.90E-46
19	43,036,265	rs210324533	0.0846	1029	1.0710	0.0779	5.30E-43
19	57,079,881	rs381175117	0.4797	1220	0.1572	0.0158	2.00E-23
19	61,134,515	rs41923843	0.3283	1130	0.4304	0.0302	3.20E-46
20	58,454,531	rs135636613	0.2153	1391	1.4139	0.0315	4.3e-441
22	53,519,865	rs109233889	0.4403	1235	0.1920	0.0246	5.30E-15
24	58,817,202	rs208779762	0.1281	1220	0.3141	0.0259	6.80E-34
27	36,211,708	rs209855549	0.4361	1731	-0.8576	0.0293	6.20E-188
27	41,267,242	rs109068627	0.2187	2977	0.2361	0.0223	3.50E-26
29	9,546,217	rs380868305	0.2688	1130	0.9258	0.0318	4.60E-186
29	44,579,245	rs439384463	0.0473	1548	1.3728	0.1688	4.30E-16

Abbreviations: FP = Fat %; LP = Lactose %; PP = Protein %.

Table 6.A.6: Minor allele frequencies and allele effects for whole-genome sequence tag variants with a significant association effect in at least one FT-MIR predicted milk composition trait and a co-localized eQTL

Chr	Position	Tag variant ID	Minor allele frequency	Trait	Beta	SE	*P*-value
3	34,387,618	rs109030498	0.3034	FP	0.0348	0.0048	6.00E-13
5	75,698,283	rs385866519	0.4049	FP	-0.0398	0.0045	4.00E-19
5	93,945,738	rs211210569	0.3393	FP	-0.1130	0.0052	6.70E-106
10	46,483,019	rs133089336	0.1873	FP	0.0467	0.0064	4.50E-13
11	104,229,609	rs110534892	0.4536	FP	0.0338	0.0044	2.60E-14
16	67,730,371	rs380453838	0.3391	FP	-0.0412	0.0046	2.60E-19
27	36,211,708	rs209855549	0.4361	FP	-0.1223	0.005	9.70E-132
1	154,122,887	rs42167460	0.229	LP	0.0215	0.0014	1.20E-50
3	15,433,518	rs109749506	0.1263	LP	-0.0181	0.0019	1.30E-20
3	15,545,091	rs379353107	0.4787	LP	0.0165	0.0012	2.20E-42
3	53,994,057	rs211488591	0.3084	LP	-0.0112	0.0013	6.70E-18
19	43,036,265	rs210324533	0.0846	LP	0.0263	0.002	9.40E-40
19	61,134,515	rs41923843	0.3283	LP	0.0179	0.0012	1.10E-46
20	58,448,763	rs134813825	0.2146	LP	-0.0128	0.0015	3.20E-18
27	36,204,066	rs208306200	0.4372	LP	0.0114	0.0012	1.90E-21
29	9,577,372	rs380473328	0.2701	LP	0.0342	0.0014	2.10E-140
3	15,520,971	rs109098377	0.1245	PP	0.0270	0.0033	7.50E-16
3	15,550,598	rs380597285	0.4784	PP	-0.0265	0.0021	1.70E-37
5	75,680,825	rs208925020	0.4325	PP	-0.0234	0.0024	8.50E-23
5	93,945,738	rs211210569	0.3306	PP	-0.0328	0.0024	3.70E-42
6	87,387,870	rs382652853	0.3056	PP	-0.0409	0.0029	2.90E-45
10	46,581,015	rs109326466	0.2011	PP	0.0331	0.0026	4.00E-38
18	2,203,325	rs135350753	0.2848	PP	-0.0168	0.0023	2.10E-13
19	43,035,006	rs209494359	0.0846	PP	-0.0493	0.0037	1.60E-40
24	58,817,202	rs208779762	0.1281	PP	0.0324	0.0031	5.70E-26

Abbreviations: FP = Fat %; LP = Lactose %; PP = Protein %.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Kathryn Maree Tiplady
Name/title of Primary Supervisor:	Professor Dorian Garrick
In which chapter is the manuscript /published work:	Chapter Six

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Tiplady K.M., Lopdell T.J., Reynolds E., Sherlock R.G., Keehan M., Johnson T.J., Pryce J.E., Davis S.R., Spelman R.J., Harris B.L. and Garrick D.J. Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. Genetics Selection Evolution. 2021 Dec;53(1):1-24.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Kathryn Tiplady Digitally signed by Kathryn Tiplady Date: 2022.03.23 18:22:29 +13'00'
Date:	
Primary Supervisor's Signature:	Dorian Garrick
Date:	25-Mar-2022

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 7

Comparison of the genetic characteristics of directly measured and FT-MIR predicted bovine milk fatty acids and proteins

Paper accepted for publication *(July 2022)*: Tiplady, K.M., Lopdell, T.J., Sherlock, R.G., Johnson, T.J.J., Spelman, R.J., Harris, B.L., Davis, S.R., Littlejohn, M.D. and Garrick, D.J., (in press). Comparison of the genetic characteristics of directly measured and FT-MIR predicted bovine milk fatty acids and proteins. *Journal of Dairy Science*.

7.1 Interpretive summary

Fourier-transform mid-infrared (FT-MIR) spectroscopy is a high-throughput, and inexpensive methodology commonly used to evaluate concentrations of fat and protein in dairy cattle milk samples. This methodology is also of interest for predicting fatty acids and individual milk proteins. The objective of this study was to compare the genetic characteristics for these predicted traits with those that had been measured directly using gas and liquid chromatography methods. We show that genetic correlations between directly measured and FT-MIR predicted fatty acids and proteins are generally high, but that the underlying genetic architecture is not always the same.

7.2 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 x Jersey animals that were managed in a seasonal, pasture-based dairy system, with milk samples collected across two 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 (PLS) 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 QTL from directly measured traits were compared to QTL from the corresponding FT-MIR predicted traits. Cross-validation prediction accuracies based on PLS 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-, β- 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: *Fourier-transform mid-infrared spectroscopy, milk composition, genome-wide association study, dairy cattle*

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

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 HD genotypes (Buitenhuis et al., 2014; Palombo et al., 2018), and FT-MIR predicted fatty acids using 50k (Cruz et al., 2019; Freitas et al., 2020; Iung et al., 2019), 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 (Pegolo et al., 2018; Schopen et al., 2011) 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; Freitas et al., 2020; Iung et al., 2019), 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 F2 study population would enhance our ability to identify trait QTL and candidate causative mutations, and that using the same dataset to conduct GWAS across directly measured and FT-MIR predicted traits would be valuable for determining differences between QTL.

7.4 Materials and methods

7.4.1 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 4232, 4621 and 10,174, according to the rules and guidelines outlined in the New Zealand Animal Welfare Act 1999.

7.4.2 Study population / animals and milk samples

Animals included in this study were from an F2 design crossbreeding experiment with a half-sibling family structure as previously described (Berry et al., 2010; Spelman et al., 2001). Briefly, six F1 bulls were generated from reciprocal crosses of Holstein-Friesian and Jersey animals that were then mated to high genetic-merit F1 cows. This resulted in a herd of 850 F2 female progeny consisting of two 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 F2 animals would typically be higher compared to 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 2003/04 season for cohort 1 and the 2004/05 season for cohort 2. Calving for each cohort took place over ~3 months between July and October. Samples for each cohort representing peak milk were collected on a daily basis for those cows at 35 days post calving, whilst 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 days in milk at the time of sampling has been provided in Appendix 7.A.1.

Concentrations of fatty acids were directly determined in milk fat samples by fatty acid methyl ester analysis using gas chromatography (GC) (MacGibbon and Reynolds, 2011), within one of up to five batches on a given sample collection day, and were expressed as g/100g 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: (i) saturated fatty acids (SFA; no double bonds); (ii) unsaturated fatty acids (UFA; one or more double bonds); (iii) polyunsaturated fatty acids (PUFA; two or more double bonds); (iv) short-chain fatty acids (SCFA; 4, 6 or 8 carbons); (v) medium-chain fatty acids (MCFA; 10, 12 or 14 carbons); and (vi) long-chain fatty-acids (LCFA; 18 carbons). Milk proteins were determined using high-performance liquid chromatography (HPLC) as described by Palmano

and Elgar (2002) and were analysed within one of up to six batches on a given sample collection day, and were expressed as g/L 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 minimised skewness. A cube-root transformation was the most effective of those investigated for minimising 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 7.A.2. 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 analysed on a Foss MilkoScan FT6000 (FOSS, Hillerød, Denmark) 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 (MD) between each spectral record and the average spectra were evaluated using the 542 wavenumbers identified as being outside noise regions. The distributions of MD values for each season were compared and found to be similar, indicating that although the spectra were collected in two different seasons, the impact 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 MD values, and based on a p-value of 0.001, 18 outliers were identified and removed. In total, after outlier removal, there were 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.

7.4.3 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 (PLS) models for each trait were developed using 542 spectral wavenumbers with the caret package in R (Kuhn, 2008), based on training data with 10 repeats of 10-fold cross-validation. Besides the untreated spectra, several mathematical treatments of spectra were assessed using the mdatools package in R (Kucheryavskiy, 2020): standard normal variate 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 MSC 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 (R^2_{cv}); and the relative prediction error between actual and predicted trait values in the validation dataset (RPE_{cv}), as described by Lopez-Villalobos et al. (2014).

7.4.4 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:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{bmatrix} \quad (7.1)$$

where $\boldsymbol{y}_1$ is a vector of the directly measured fatty acid or protein fraction, $\boldsymbol{y}_2$ is a vector of the corresponding FT-MIR predicted trait; $\boldsymbol{X}_1$, $\boldsymbol{Z}_1$, $\boldsymbol{W}_1$, $\boldsymbol{X}_2$, $\boldsymbol{Z}_2$ and $\boldsymbol{W}_2$ are design matrices for the fixed, additive genetic and permanent environment effects respectively for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$; $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are vectors of the fixed effect of days in milk (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; $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are vectors of random additive genetic effects for each trait; $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are vectors of permanent environment effects for each trait; and $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are vectors of residuals. The following (co)variance structure for each directly measured ($\boldsymbol{y}_1$) and FT-MIR predicted ($\boldsymbol{y}_2$) trait pair is assumed:

$$\text{var}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{p}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G}\bigotimes\boldsymbol{A} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}\bigotimes\boldsymbol{I}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\bigotimes\boldsymbol{I}_e\end{bmatrix}, \text{ where } \boldsymbol{u} = \begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{bmatrix}, \boldsymbol{p} = \begin{bmatrix}\boldsymbol{p}_1\\ \boldsymbol{p}_2\end{bmatrix} \text{ and } \boldsymbol{e} = \begin{bmatrix}\boldsymbol{e}_1\\ \boldsymbol{e}_2\end{bmatrix} \tag{7.2}$$

where $\boldsymbol{A}$ is the numerator relationship matrix, $\boldsymbol{I}_p$ is an identity matrix of order corresponding to the length of the vector $\boldsymbol{p}$, $\boldsymbol{I}_e$ is an identity matrix of order corresponding to the length of the vector $\boldsymbol{e}$, $\bigotimes$ is the Kronecker product; and $\mathbf{G}$, $\mathbf{C}$ and $\mathbf{R}$ are genetic, permanent environment and residual (co)variance matrices, respectively, defined as follows:

$$\mathbf{G} = \begin{bmatrix}\sigma^2_{u_1} & \sigma_{u_1u_2}\\ \sigma_{u_1u_2} & \sigma^2_{u_2}\end{bmatrix}, \mathbf{C} = \begin{bmatrix}\sigma^2_{p_1} & \sigma_{p_1p_2}\\ \sigma_{p_1p_2} & \sigma^2_{p_1}\end{bmatrix} \text{ and } \mathbf{R} = \begin{bmatrix}\sigma^2_{e_1} & \sigma_{e_1e_2}\\ \sigma_{e_1e_2} & \sigma^2_{e_1}\end{bmatrix} \tag{7.3}$$

The heritability and repeatability for each trait were calculated as functions of the estimated (co)variance components based on their parametric definitions of $h^2_i = \dfrac{\sigma^2_{u_i}}{\sigma^2_{u_i} + \sigma^2_{p_i} + \sigma^2_{e_i}}$ and $t_i = \dfrac{\sigma^2_{u_i} + \sigma^2_{p_i}}{\sigma^2_{u_i} + \sigma^2_{p_i} + \sigma^2_{e_i}}$, where $i = 1$ or 2 for traits $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, respectively; and the genetic correlation for each pair of measured/predicted traits was calculated as $r_a = \dfrac{\sigma_{u_1u_2}}{\sigma_{u_1}\sigma_{u_i}}$. 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 $\dfrac{a(\sigma^2_{u_1} + \sigma^2_{u_2})}{2}$ and $\dfrac{b(\sigma^2_{e_1} + \sigma^2_{e_2})}{2}$ for additive genetic and residual covariances, respectively, where a and b ranged from 0.1 to 0.9 in increments of 0.1. Amongst 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 minimised the squared sum of the difference between single- and multi-trait model heritability estimates was selected.

7.4.5 Genotypes and imputation

Of the 706 animals with phenotypic data, 685 were genotyped on Illumina BovineHD (HD; N=12; ~777k SNP) and/or Illumina BovineSNP50k (50k; N=685; ~53k SNP) panels. 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., 2021b). Briefly, the imputation process consisted of stepwise imputation of animals to whole-genome sequence genotypes via references of GGP, 50k and HD genotypes. The whole-genome sequence reference consisted of 565 animals, comprised of 138 Holstein-Friesians, 99 Jerseys, 316 Holstein-Friesian x Jersey crossbreeds, and 12 from other breeds or crosses. Notably, the 6 F1 sires included in our study were included in this whole-genome sequence reference and were sequenced with a target of 60x 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.

7.4.6 Genome-wide association studies

Prior to 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 one 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), comprising:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{W}\boldsymbol{p} + \boldsymbol{e} \tag{7.4}$$

where $\boldsymbol{y}$ is a vector of the measured or predicted trait, $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ are design matrices for the fixed, additive genetic and permanent environment effects; $\boldsymbol{b}$ is the fixed effect of days in milk (represented as 35-day windows from the start of lactation) within season (2003, 2004) for the trait; $\boldsymbol{u}$ is a vector of random additive genetic effects with $\boldsymbol{u}\sim N(0,\boldsymbol{A}\sigma_u^2)$; $\boldsymbol{p}\sim N(0,\boldsymbol{I}_p\sigma_p^2)$ is a vector of random permanent environment effects; and $\boldsymbol{e}$ is a vector of random residuals with $\boldsymbol{e}\sim N(0,\boldsymbol{I}_e\sigma_e^2)$ where $\boldsymbol{A}$ is the numerator relationship matrix, $\boldsymbol{I}_p$ is an identity matrix of order corresponding to

the length of the vector $\boldsymbol{p}$, $\boldsymbol{I}_e$ is an identity matrix of order corresponding to the length of the vector $\boldsymbol{e}$, 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 $\boldsymbol{y}$ 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). Prior to conducting GWAS, a MAF 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 (GRM) based on a subset of 42,374 SNP was simultaneously fitted. That subset of SNP was derived by applying a MAF threshold of 1% to the 50k SNP-chip imputation reference (previously described). A leave-one-segment-out (LOSO) 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 GRM.

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 (Wang et al., 2019b; Zhu et al., 2017). 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 100kb 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 α=0.05, this resulted in a nominal p-value of 4.3e-09 and a corresponding Bonferroni threshold of -$\log_{10}$(4.3e-09)=8.36. Whole-genome sequence resolution genotypes within a 1-Mbp window were annotated using SnpEff (version 4.3t; build 2017-11-24) (Cingolani et al., 2012) and Ensembl UMD3.1.86 gene annotations to assess the candidacy of QTL identified from the GWAS for each trait. We used an LD-based approach to prioritise variants, similar to that described by Lopdell et al. (2017) because the association rankings of candidate variants are expected to be impacted 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.

7.5 Results and discussion

7.5.1 Trait prediction models

Cross-validation prediction model accuracies (R^2_{cv}) were assessed for untreated spectra, as well as for spectra treated using standard normal variate (SNV) transformation, multiplicative scatter correction (MSC) or first-derivative treatments (Appendix 7.A.1). 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 7.1.

For individual fatty acids, coefficient of determination values for the validation dataset (R^2_{cv}) were generally higher for short-chain fatty acids (C4 to C8), ranging from 0.54 to 0.62, compared to 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 to individual unsaturated fatty acids. For grouped fatty acids, R^2_{cv} values were higher for UFA and SFA groups, compared to PUFA; and 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 there was an overall trend of higher R^2_{cv} values coinciding with lower RPE_{cv} values, there were exceptions to this. For example, amongst 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, whilst the latter provides a comparison of how similar the predicted values are to the directly measured 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 also 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 (Bonfatti et al., 2016;

Lopez-Villalobos et al., 2014; Rutten et al., 2009). In the present study, for grouped short-, medium- and long-chain fatty acids, R^2_{cv} values were lower than in other studies (Bonfatti et al., 2016; Lopez-Villalobos et al., 2014; Rutten et al., 2009). 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-casein (β-CN) to 0.69 for β-lactoglobulin (β-LG). Notably, although the R^2_{cv} values for β-CN and β-LG were very different, the RPE_{cv} values for these two 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 than in other studies (Bonfatti et al., 2016; De Marchi et al., 2009a; Lopez-Villalobos et al., 2009; McDermott et al., 2016; Rutten et al., 2011; Soyeurt et al., 2012). Fuentes-Pila et al. (1996) suggest that a relative prediction error (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 six 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.

Table 7.1: Descriptive statistics of fatty acid and protein traits, and goodness of fit measures of PLS calibration models for training and validation datasets

Trait	Description and units	Trait summary			Training		Validation	
		n	Mean	SD	R^2_t	RPE_t	R^2_{cv}	RPE_{cv}
Individual fatty acids								
C4:0	Butyric acid, g/100g of total fat	1963	3.90	0.32	0.706	0.043	0.602	0.053
C6:0	Caproic acid, g/100g of total fat	1969	2.52	0.19	0.591	0.049	0.542	0.052
C8:0	Caprylic acid, g/100g of total fat	1968	1.54	0.18	0.697	0.064	0.622	0.073
C10:0	Capric acid, g/100g of total fat	1975	3.51	0.61	0.701	0.094	0.627	0.108
C10:1	Caproleic acid, g/100g of total fat	1969	0.31	0.06	0.469	0.151	0.300	0.162
C12:0	Lauric acid, g/100g of total fat	1972	3.92	0.74	0.685	0.106	0.590	0.121
C12:1	Lauroleic acid, g/100g of total fat	1925	0.13	0.03	0.47	0.169	0.353	0.181
C14:0	Myristic acid, g/100g of total fat	1967	11.46	1.17	0.599	0.065	0.491	0.073
C14:1	Myristoleic acid, g/100g of total fat	1970	0.75	0.23	0.517	0.211	0.414	0.233
C16:0	Palmitic acid, g/100g of total fat	1977	27.64	3.27	0.633	0.073	0.574	0.076
C16:1	Palmitoleic acid, g/100g of total fat	1958	1.54	0.22	0.301	0.123	0.184	0.132
C18:0	Stearic acid, g/100g of total fat	1968	11.95	2.00	0.544	0.115	0.445	0.124
C18:1 *cis*-7	*cis*-Vaccenic Acid, g/100g of total fat	1936	4.53	0.70	0.531	0.107	0.411	0.118
C18:1 *cis*-9	Oleic acid, g/100g of total fat	1963	17.31	2.55	0.653	0.088	0.569	0.096
C18:2 *cis*-9, trans-11	Conjugated linoleic acid, g/100g of total fat	1929	0.87	0.25	0.587	0.185	0.498	0.210
C18:2 *cis*-6	Linoleic acid, g/100g of total fat	1963	1.20	0.14	0.561	0.078	0.480	0.085
C18:3 *cis*-3	α-linolenic acid, g/100g of total fat	1954	0.80	0.11	0.387	0.112	0.360	0.105
Grouped fatty acids								
SFA	Saturated fatty acids, g/100g of total fat	1965	70.59	3.08	0.703	0.024	0.591	0.028
PUFA	Polyunsaturated fatty acids, g/100g of total fat	1972	4.16	0.46	0.641	0.065	0.490	0.081
UFA	Unsaturated fatty acids, g/100g of total fat	1964	29.42	3.08	0.711	0.057	0.597	0.066
SCFA	Short-chain fatty acids, g/100g of total fat	1970	7.96	0.59	0.695	0.041	0.648	0.043
MCFA	Medium-chain fatty acids, g/100g of total fat	1969	20.09	2.43	0.659	0.071	0.567	0.080
LCFA	Long-chain fatty acids, g/100g of total fat	1974	36.82	4.45	0.609	0.076	0.568	0.079
Individual milk proteins								
α-CN	α-casein, g/L of total volume	1695	15.79	1.76	0.585	0.072	0.532	0.076
β-CN	β-casein, g/L of total volume	1686	14.78	1.84	0.128	0.116	0.190	0.113
ϰ-CN	ϰ-casein, g/L of total volume	1687	4.24	0.59	0.575	0.087	0.476	0.105
α-LA	α-lactalbumin, g/L of total volume	1942	1.21	0.15	0.379	0.099	0.306	0.104
β-LG	β-lactoglobulin, g/L of total volume	1959	3.84	0.70	0.773	0.087	0.678	0.104
Lf [1]	Lactoferrin, g/L of total volume	1936	0.51	0.12	0.411	0.188	0.356	0.194

[1] Cube-root transformation of lactoferrin.
Abbreviations: n=number of samples; SD=standard deviation; 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; 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; 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); SFA=Saturated fatty acids; UFA=Unsaturated fatty acids.

Table 7.2: Variance component estimates for directly measured and FT-MIR predicted fatty acid and protein traits

Trait[1]	Directly measured trait				FT-MIR predicted trait				r_a
	σ^2_u	σ^2_T	h^2	t	σ^2_u	σ^2_T	h^2	t	
Individual fatty acids (g/100g 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)	0.0003	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 *c*9, *t*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 acids (g/100g 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 proteins (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[2]	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 7.1. Standard errors shown in brackets.
[2] Cube-root transformation of lactoferrin.
Abbreviations: n=number of samples; σ^2_u=additive genetic variance; σ^2_T=total variance ($\sigma^2_u+\sigma^2_p+\sigma^2_e$); h^2=heritability estimate; t=repeatability estimate; r_a=genetic correlation between directly measured and FT-MIR predicted trait.

7.5.2 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 7.2 and Appendix 7.A.2. 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 F2 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 (σ^2_T) in the directly measured traits, coupled with a lower magnitude increase in the additive genetic variance component (σ^2_u), 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 individual 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 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, there was a consistent trend for individual and grouped fatty acids, where lower genetic correlations coincided with low R^2_{cv} values.

Whilst there are a number of studies that report genetic parameter estimates for directly measured and/or FT-MIR predicted fatty acid traits, 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 (Fleming et al., 2018; Lopez-Villalobos et al., 2014; Narayana et al., 2017; Soyeurt et al., 2007b), with only two studies reporting genetic parameters for both directly measured and FT-MIR

predicted traits (Bonfatti et al., 2017d; Rutten et al., 2010). These latter two studies also report genetic correlations between directly measured and FT-MIR predicted fatty acids, with Bonfatti et al. (2017d) presenting these for individual and grouped fatty acids, while Rutten et al. (2010) present these for individual fatty acids only.

The heritability, repeatability and genetic correlation estimates we report in the present study were broadly 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. (2017d), 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 (Bonfatti et al., 2017d; Soyeurt et al., 2007b). Compared to 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 (Bonfatti et al., 2017d; Rutten et al., 2010). 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.

7.5.3 Individual milk protein traits

Heritability estimates were moderate to high for nearly all directly measured and FT-MIR predicted individual milk proteins (Table 7.2). 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, there was 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 and/or FT-MIR predicted milk proteins, 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., 2017d), 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 (Buitenhuis et al., 2016; Schopen et al., 2009), whilst other studies report on individual proteins as a proportion of total protein or as a proportion of total milk volume (Bonfatti et al., 2017d; Sanchez et al., 2017a). The heritability estimates we report for directly measured α-, β- and ϰ-CN were lower than those previously reported by Bonfatti et al. (2017d), but the heritability estimates we report for directly measured α-LA and β-LG were substantially higher. In contrast, for FT-MIR predicted α-, β- and ϰ-CN, the heritability estimates we report were consistently higher than those reported by Bonfatti et al. (2017d), 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. (2017d). 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. (2017d), genetic correlations for these traits ranged from 0.24 for α-LA to 0.48 for β-LG. Interestingly, Bonfatti et al. (2017d) 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 two studies (i.e., Italian Simmental cows from herds enrolled in the Italian national milk recording program vs Holstein-Friesian Jersey F2 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.

7.5.4 Sequence-based genome-wide association analyses

Previously, there have been a number of GWAS that used a range of genotype densities for fatty acids in bovine milk samples determined by gas chromatography (GC) (Bouwman et al., 2011; Buitenhuis et al., 2014; Palombo et al., 2018) or fatty acids predicted using FT-MIR spectra (Cruz et al., 2019; Freitas et al., 2020; Iung et al., 2019; Olsen et al., 2017; Sanchez et al., 2019). Similarly, there have been multiple GWAS conducted on protein fractions in milk samples determined by high-performance liquid chromatography (HPLC) (Buitenhuis et al., 2016; Pegolo et al., 2018; Schopen et al., 2011; 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 F2 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. There were more than twice as many variants with significant effects for directly measured traits, compared to FT-MIR predicted traits, which was largely due to 20,949 variants with significant effects on BTA26 for directly measured traits compared to only 110 variants with significant effects on BTA26 for FT-MIR predicted traits. It was also notable that there were 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 Figs. 7.1-7.4 for individual fatty acids, Fig. 7.5 for grouped fatty acids and Fig. 7.6 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 7.3 for directly measured fatty acid and protein traits and Table 7.4 for FT-MIR predicted fatty acid and protein traits. Effect sizes and MAF details for relevant variants and effects are provided in Appendix 7.A.3 for fatty acids and Appendix 7.A.4 for protein traits.

Table 7.3: Peak variants for directly measured fatty acid and protein traits with significant association effects

Trait[1]	Chr	Position	Tag variant ID	*P*-value	Protein coding variant ID	LD	Gene	Class	Description
Individual fatty acids (g/100g 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 acids (g/100g 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 proteins (g/L of total volume)									
α-CN	6	87133508	rs109500363	4.3e-12	rs382793163	0.856	*ENS.. 39991*	Missense	c.1406G>A
α-CN	6	87133508	rs109500363	4.3e-12	rs385603965	0.839	*ENS..03523*	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	0.95	*PAEP*	Missense	c.401T>C
Lf[2]	22	53538882	rs43765460	1.8e-41	.	.	.	.	.

[1] Trait definitions and units as described in Table 7.1.
[2] Cube-root transformation of lactoferrin.

Table 7.4: Peak variants for FT-MIR predicted fatty acid and protein traits with significant association effects

Trait[1]	Chr	Position	Tag variant ID	*P*-value	Protein coding variant ID	LD	Gene	Class	Description
Individual fatty acids (g/100g 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	*ENS..03606*	Missense	c.154C>G
C18:1 *cis*-9	14	2528807	rs110275497	1.3e-10	rs109403601	1	*ENS..03606*	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 acids (g/100g of total fat)									
UFA	14	2319003	rs110182536	8.1e-10	rs109403601	0.947	*ENS..03606*	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 proteins (g/L of total volume)									
κ-CN	6	87085918	.	8.2e-21	rs209798512	0.761	*ENS..38520*	Missense	c.1623G>C
κ-CN	6	87085918	.	8.2e-21	rs211555767	0.761	*ENS..38520*	Missense	c.1301C>T
κ-CN	6	87085918	.	8.2e-21	rs382793163	0.725	*ENS..39991*	Missense	c.1406G>A
κ-CN	6	87085918	.	8.2e-21	rs385603965	0.711	*ENS..03523*	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	*ENS..39991*	Missense	c.1406G>A
α-CN	6	87133508	rs109500363	7.0e-11	rs385603965	0.839	*ENS..03523*	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] Trait definitions and units as described in Table 7.1.

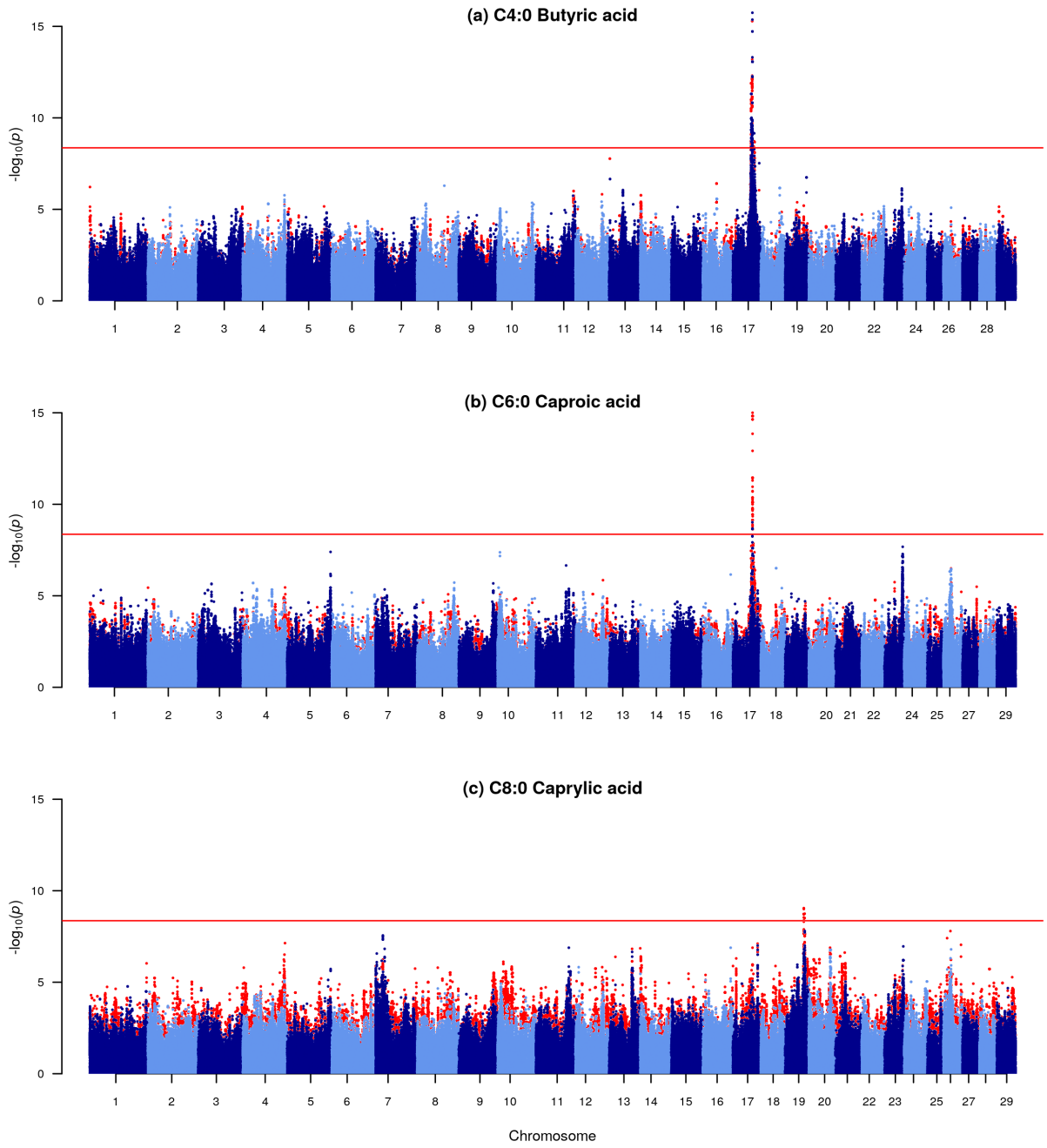

Figure 7.1: Manhattan plots showing association effects for directly measured (GC-based) and FT-MIR predicted individual short-chain fatty acid traits. Dark and light blue data points represent association signals for GC-based traits and red data points represent association signals for FT-MIR predicted 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 are represented as the $-\log_{10}(p\text{-value})$ on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}(4.3\text{e-}09)$. GC=gas chromatography.

Short-chain fatty acids

Prominent peaks were observed on BTA17 for the short-chain fatty acids, C4:0 and C6:0 (Tables 7.3 and 7.4; Fig. 7.1). For directly measured and FT-MIR predicted C4:0, these peaks were underpinned by the same QTL at Chr17:53.03 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 two 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 7.3 and 7.4; Fig. 7.2). 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 two 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 characterised by carboxylic C=O bond stretching (Tiplady et al., 2021b). 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 7.3 and 7.4; Fig. 7.2). Significant effects were also observed in this region for FT-MIR predicted C8:0 (p-value=8.9e-10; Fig. 7.1), and directly measured (p-value=1.4e-13) and FT-MIR predicted MCFA (p-value=9.2e-13; Fig. 7.5). The *CCDC57* gene

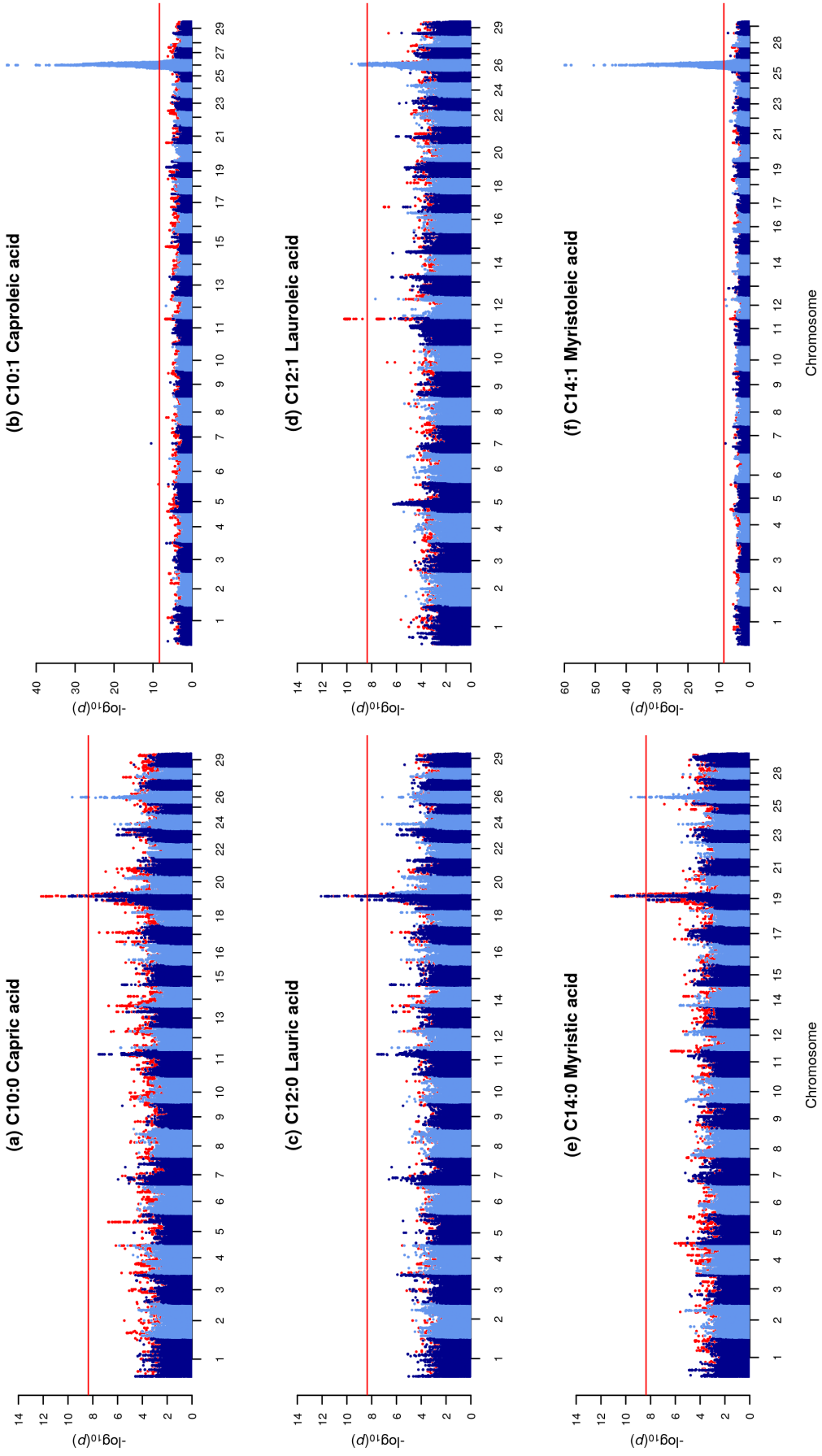

Figure 7.2: Manhattan plots showing association effects for directly measured (GC-based) and FT-MIR predicted individual medium-chain fatty acid traits. Dark and light blue data points represent association signals for GC-based traits and red data points represent association signals for FT-MIR predicted 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 are represented as the $-\log_{10}(p\text{-value})$ on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}(4.3e\text{-}09)$. GC=gas chromatography.

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 a number of FT-MIR wavenumbers characterised by carboxylic C=O bond stretching (Tiplady et al., 2021b). Bouwman et al. (2014) examined this region in depth using HD genotypes and identified two 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 eight 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 being at 51,380,689 bp, but the *p*-value for that 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 to 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 7.3; Fig. 7.2). 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 7.3) and FT-MIR predicted UFA (Table 7.4). 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 (Bouwman et al., 2011; Conte et al.,

2010; Kgwatalala et al., 2009; Mele et al., 2007; Moioli et al., 2007; Schennink et al., 2008). 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 synthesised by stearoyl-CoA desaturase (Bernard et al., 2006). Interestingly, although there was 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 7.4). However, the LD between the rs209445650 locus and the splice region variant identified for directly measured fatty acids (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., 2021b), 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 (R^2_{cv}=0.30; RPE$_{cv}$=0.16) and C14:1 (R^2_{cv}=0.41; RPE$_{cv}$=0.23) (Table 7.1). 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 7.2). 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 to saturated fatty acids. Specifically, C14:1 had a mean concentration of 0.75g/100 g total fat; compared to mean concentrations of 1.54 to 27.64g/100g total fat for the individual saturated fatty acids included in this study (Table 7.1). Potentially, it may be possible to improve trait prediction accuracies, heritability 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.

One further QTL was observed for directly measured C10:1 on BTA26 at a Chr26:26.46 Mbp locus (rs445758306; Table 7.3; Fig. 7.2). 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.

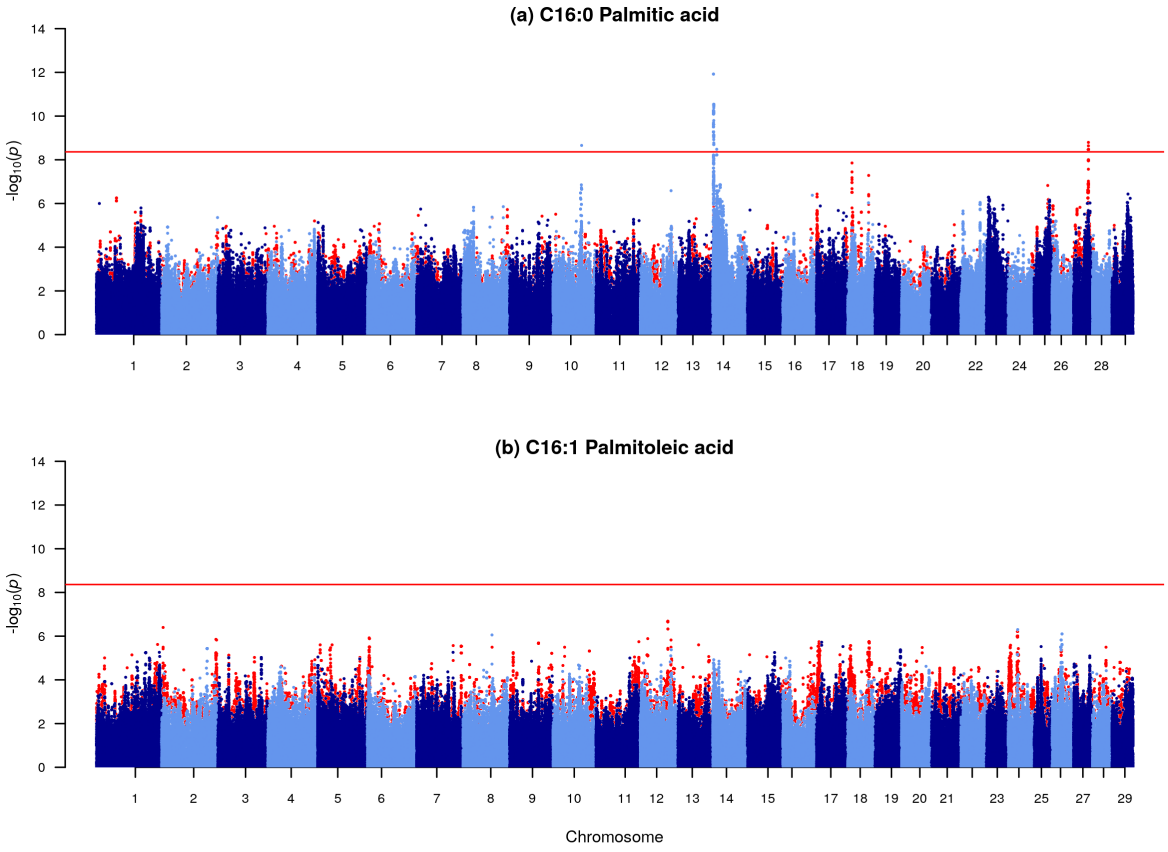

Figure 7.3: Manhattan plot showing association effects for directly measured (GC-based) and FT-MIR predicted C16 fatty acid traits. Dark and light blue data points represent association signals for GC-based traits and red data points represent association signals for FT-MIR predicted traits. Chromosomes and genomic position based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals are represented as the $-\log_{10}(p\text{-value})$ on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}$(4.3e-09). GC=gas chromatography.

Long-chain fatty acids

Two QTL were identified on BTA14 for directly measured individual long-chain fatty acids (Table 7.3; Figs. 7.3 and 7.4). 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 (Fink et al., 2020; Grisart et al., 2002) and 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 catalyses 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 ascribed to any protein coding mutations in genes (Table 7.4; Figs. 7.3 and 7.4). However, the locus for C18:3 *cis*-3 (rs110950972) was in perfect LD with a 5' untranslated region (UTR; rs208675276) in *GPAT4*, and the locus for C16:0 was also in high LD (R^2=0.997) with that same 5' UTR. Interestingly, there was no evidence of QTL on BTA27 for the corresponding directly measured traits (Figs. 7.3 and 7.4). 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 missense mutations in the *UTP18* gene (Table 7.3; Fig. 7.5). 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 impact milk phenotypes (Tiplady et al., 2021b), although a mechanism by which this gene could impact fatty acids is unclear.

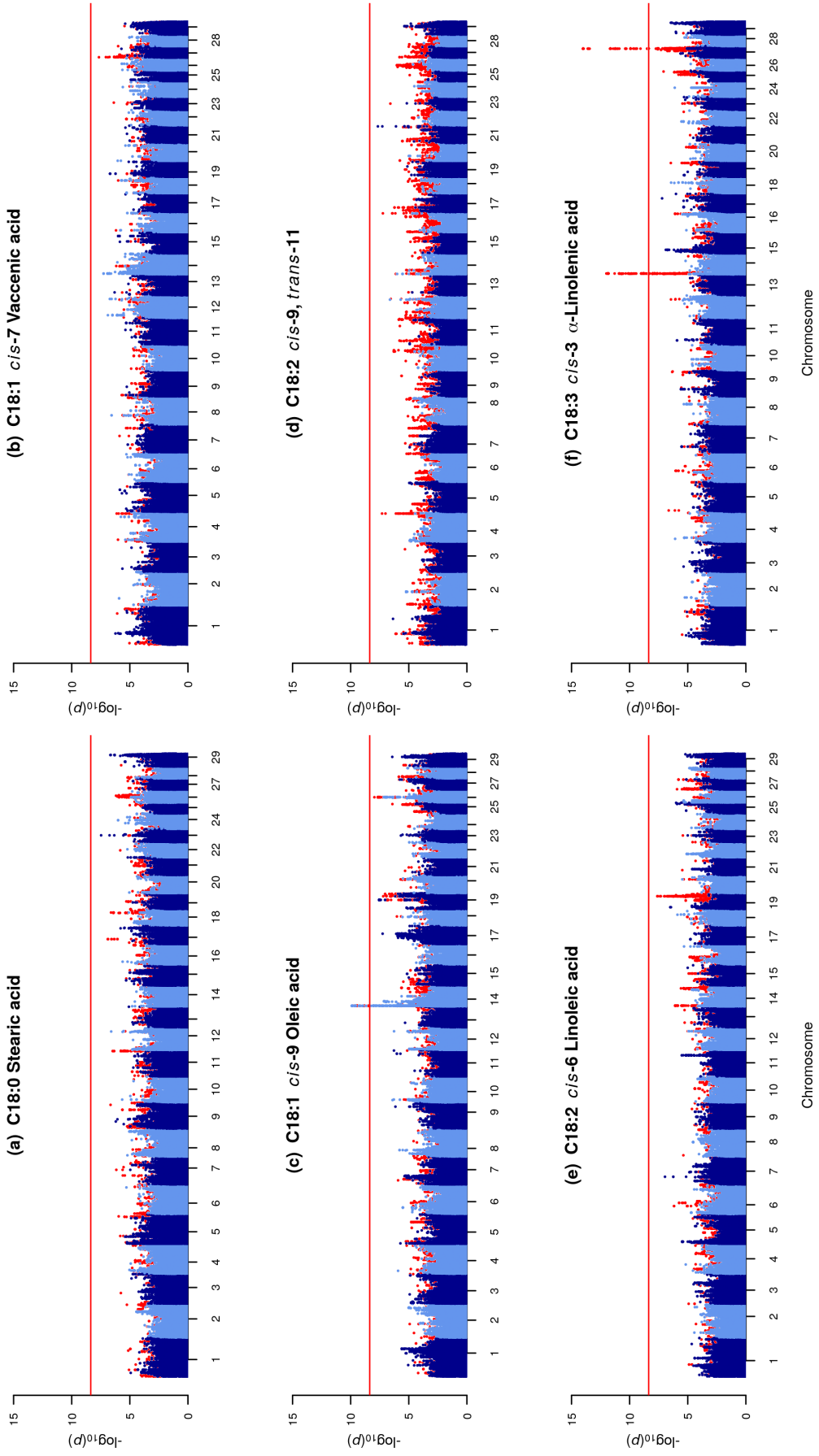

Figure 7.4: Manhattan plot showing association effects for directly measured (GC-based) and FT-MIR predicted C18 fatty acid traits. Dark and light blue data points represent association signals for GC-based traits and red data points represent association signals for FT-MIR predicted traits. Chromosomes and genomic position based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals are represented as the $-\log_{10}(p\text{-value})$ on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}(4.3\text{e-}09)$. GC=gas chromatography.

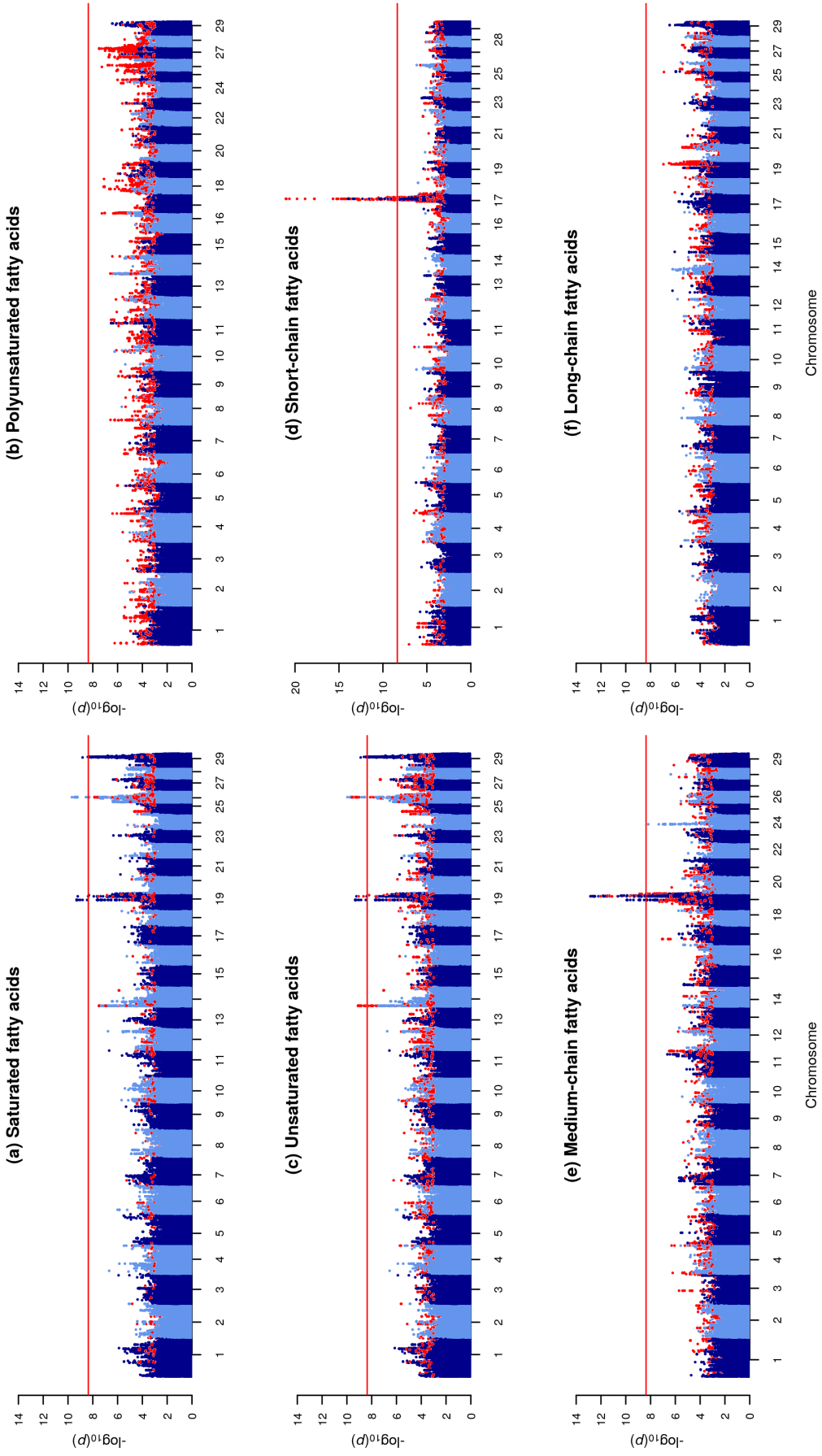

Figure 7.5: Manhattan plot showing association effects for directly measured (GC-based) and FT-MIR predicted fatty acids classified based on the degree of saturation and the length of the carbon chain. Dark and light blue data points represent association signals for GC-based traits and red data points 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 are represented as the $-\log_{10}(p\text{-value})$ on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}(4.3\text{e-}09)$. GC=gas chromatography.

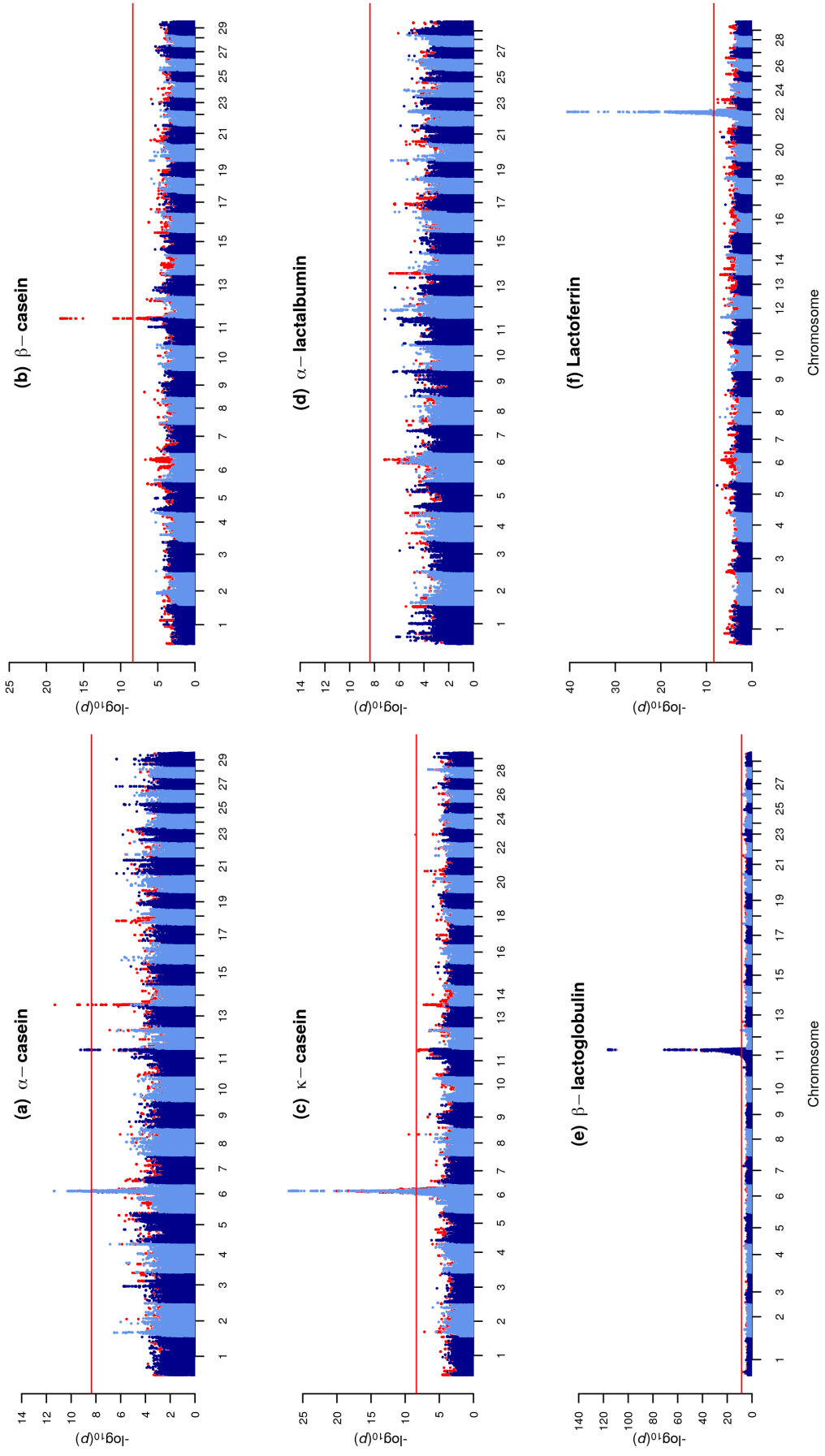

Figure 7.6: Manhattan plots showing association effects for directly measured (HPLC-based) and FT-MIR predicted milk proteins. Dark and light blue data points represent association signals for HPLC-based traits and red data points represent association signals for FT-MIR predicted traits. Chromosomes and genomic position based on the UMD3.1 *Bos taurus* reference genome are represented on the x-axis. The strength of association signals are represented as the $-\log_{10}(p\text{-value})$ on the y-axis. The horizontal red line shows the Bonferroni significance threshold of $-\log_{10}(4.3\text{e-}09)$. HPLC=high-performance liquid chromatography.

Individual milk proteins

Significant effects were observed on BTA6, BTA11, BTA14 and BTA22 for individual milk proteins (Tables 7.3 and 7.4; Fig. 7.6). Four QTL were identified on BTA6, two of which were for directly measured and FT-MIR predicted α-CN and the other two 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 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 two missense mutations in the *CSN3* gene (rs43703015; rs43703016). The *CSN3* gene encodes ϰ-casein, 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 characterised by amide III and phosphate bands, C–H stretching vibrations of CH_2 and $–CH_3$, and N–H bending and C–N stretching in the amide II band (Tiplady et al., 2021b).

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 7.3 and 7.4; Fig. 7.6). 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 eQTL for *PAEP* was also reported in lactating bovine mammary tissue (Davis et al., 2022; Tiplady et al., 2021b).

One further QTL of interest was for directly measured Lf at a Chr22:53.54 Mbp locus (rs43765460; Table 7.3; Fig. 7.6). The association effect at this locus had a p-value of 1.8e-41, but there was no relevant splice region variant or moderate or high impact coding variant 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., 2021b), 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 *LTF* (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 7.4). 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., 2021b), 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 7.1), 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 7.2). This pattern is similar to that which we observed for C14:1 and the *SCD* gene, i.e., the component was in relatively low concentrations in the milk sample, model prediction accuracy 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; but no corresponding peak was observed for the FT-MIR predicted trait.

7.5.5 Perspectives on FT-MIR trait predictions for dairy cattle selection

Utilising 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 is 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 processibility traits. For breeding programs, however, 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, whilst 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; and for 21 of the 29 traits, the genetic gains achievable would be greater than 85% of the gains achievable by direct selection. It is important to note, however that this assumes that there is no true difference in heritability between the directly measured and FT-MIR predicted trait. For traits like 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 (Fig. 7.4). 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 (Fig. 7.6). 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 (Benedet et al., 2019; Tiplady et al., 2021b; Wang and Bovenhuis, 2018; Zaalberg et al., 2020). 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 that 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 SNP.

Instances also arose where a QTL was observed for a directly measured trait, but there was 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 (Fig. 7.2). 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 (Fig. 7.6). 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 to the directly measured trait (Tables 7.1 and 7.2). While 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., 2021b), 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 two 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. But when the mean concentration of Lf is around 0.1g/L and the major whey protein β-LG is at a 20-40 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 to 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 optimising 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 datasets 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 datasets 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 (Hillerød, Denmark) or Bentley (Chaska, MN), 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.

7.5.6 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, prior to 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., 2017b; El Jabri et al., 2019; Frizzarin et al., 2021a) or animal health and feed intake traits (Brand et al., 2021; Contla Hernández et al., 2021; Dórea et al., 2018). 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 to herd-independent validation (Dórea et al., 2018; Lahart et al., 2019; Luke et al., 2019b; 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 datasets 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 non-additive 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 datasets such as mammary eQTL and ChIP-seq datasets could map additional molecular QTL and enhance fine mapping and candidate variant identification (Tiplady et al., 2020).

7.6 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, and the QTL underpinned by *GPAT4* were absent in directly measured 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. Our study highlights the potential value of FT-MIR predictions as indicators for fatty acid and protein fractions in milk, and that selection based on FT-MIR predictions can provide genetic gain in specific milk components of interest. We also highlight that for many traits the genes implicated in phenotypic variation were similar, but that in some instances, phenotypic variation was underpinned by differing genetic architecture and segregation of alleles at QTL.

7.7 Acknowledgements

The authors would like to 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; and staff at Fonterra Research and Development Centre, Palmerston North, for milk analyses. Kathryn would also like to thank the wider LIC R&D team and fellow students for underlying technical support and thoughtful discussion; 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).

Appendices

7.A Comparison of the genetic characteristics of directly measured and FT-MIR predicted bovine milk fatty acids and proteins

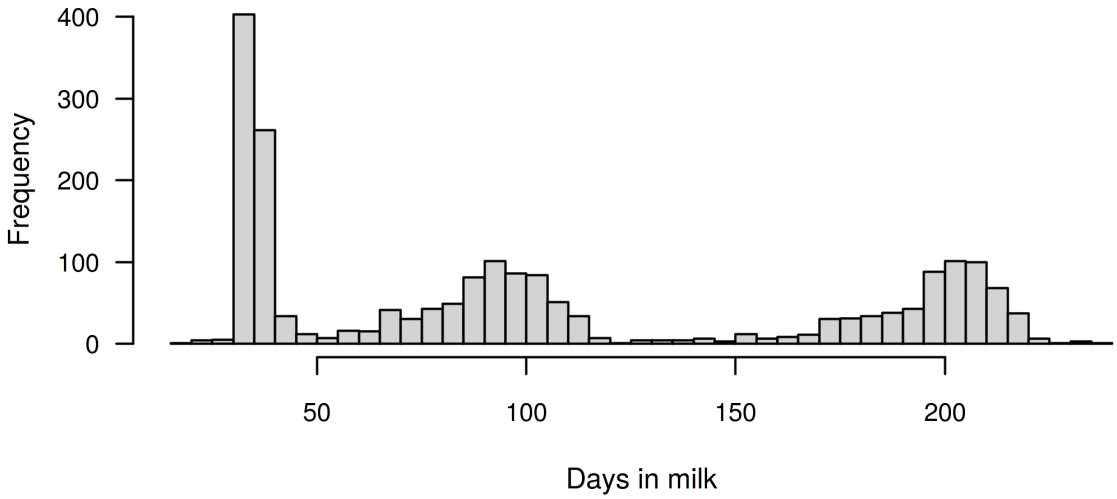

Figure 7.A.1: Frequency distribution of samples across days in milk (n=2,005).

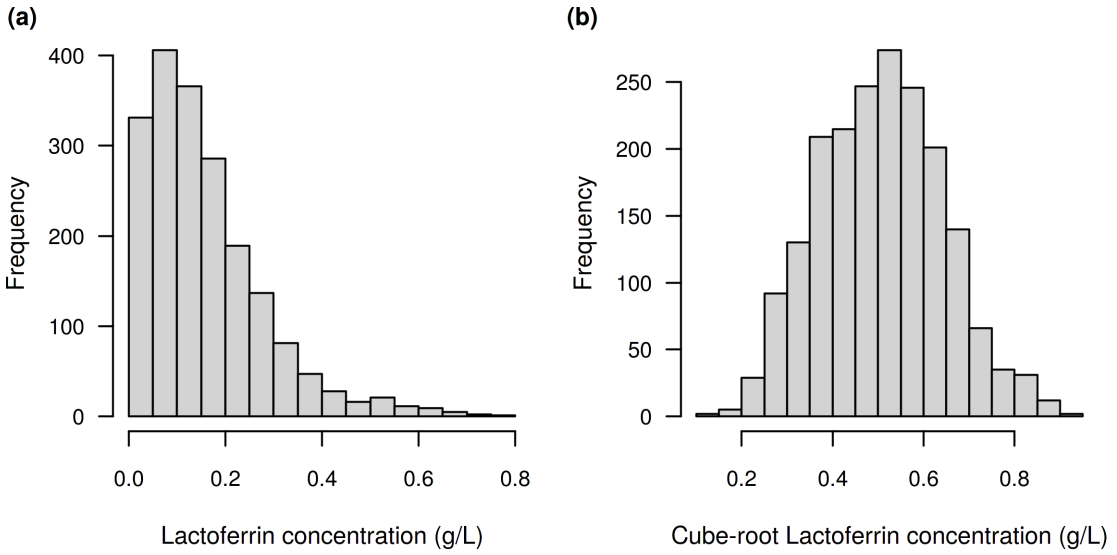

Figure 7.A.2: Frequency distributions of (a) untransformed lactoferrin concentrations and (b) lactoferrin concentrations after cube-root transformation (n=1,936).

Table 7.A.1: Goodness of fit (R^2_{cv}) of PLS calibration models for untreated and pre-treated spectra based on cow-independent validation

Trait[1]	Spectral pre-treatment[3]					
	Untreated	**First-derivative**	**MSC**	**MSC+1st**	**SNV**	**SNV+1st**
Individual fatty acids (g/100g 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 acids (g/100g 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 proteins (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[2]	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 7.1.
[2] Cube-root transformation of lactoferrin.
[3] Untreated=untreated spectral data; First-derivative=spectra pre-treated with a first-order Savitzky-Golay derivative with a window of 7 data points either side; MSC=spectra pre-treated with multiplicative scatter correction; MSC+1st=spectra pre-treated with MSC followed by first-derivative transformation; SNV=spectra pre-treated with a standard normal variate transformation; SNV+1st=spectra pre-treated with SNV followed by first-derivative transformation.

Table 7.A.2: Variance component estimates for directly measured and FT-MIR predicted fatty acid and protein traits

Trait[1]	Directly measured trait				FT-MIR predicted trait			
	σ^2_u	σ^2_p	σ^2_e	σ^2_T	σ^2_u	σ^2_p	σ^2_e	σ^2_T
Individual fatty acids (g/100g 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 *c*-9, *t*-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 acids (g/100g 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 proteins (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[2]	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 7.1. Standard errors shown in brackets.
[2] Cube-root transformation of lactoferrin.
Abbreviations: σ^2_u=additive genetic variance; σ^2_p=permanent environment variance; σ^2_e=residual variance; σ^2_T=total variance ($\sigma^2_u+\sigma^2_p+\sigma^2_e$).

Table 7.A.3: Effect sizes and minor allele frequency details for fatty acid traits with a significant association effect

Chr	Position	Tag variant ID	Minor allele frequency	Trait[1]	Trait type	Beta	SE	*P*-value
Individual fatty acids (g/100g 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 acids (g/100g 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] Trait definitions and units as described in Table 7.1.

Table 7.A.4: Effect sizes and minor allele frequency details for protein traits with a significant association effect

Chr	Position	Tag variant ID	Minor allele frequency	Trait[1]	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[2]	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] Trait definitions and units as described in Table 7.1.
[2] Cube-root transformation of lactoferrin.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Kathryn Maree Tiplady
Name/title of Primary Supervisor:	Professor Dorian Garrick
In which chapter is the manuscript /published work:	Chapter Seven

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

◉ The manuscript is currently under review for publication – please indicate:

- The name of the journal:
Journal of Dairy Science

- The percentage of the manuscript/published work that was contributed by the candidate: 90.00

- Describe the contribution that the candidate has made to the manuscript/published work:
Phenotypic analysis, pedigree-based genetic analysis, genome wide association study, writing the manuscript.

○ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Kathryn Tiplady Digitally signed by Kathryn Tiplady Date: 2022.03.23 18:23:11 +13'00'
Date:	
Primary Supervisor's Signature:	Dorian Garrick
Date:	25-Mar-2022

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 8

General discussion

8.1 Discussion overview

Fourier-transform mid-infrared (FT-MIR) spectroscopy plays a key role in generating phenotypes for major milk composition traits and is a cornerstone of modern dairy cattle milk payment and animal evaluation systems. Increasingly, there has been interest in utilising FT-MIR spectra to predict other novel traits and improve our understanding of milk composition. In this thesis I have examined a range of topics relating to phenotypic and genetic applications of FT-MIR spectra, and the role that these data may play in predicting new traits or improving the prediction of existing traits in dairy cattle. Each chapter includes discussion specific to individual topics. In the following sections, I will consolidate the most important themes and findings from each chapter, and highlight key areas of consideration for the use of FT-MIR spectra to improve dairy cattle trait prediction and advance selective breeding into the future.

8.2 Phenotyping using FT-MIR spectra

Fourier-transform mid-infrared spectroscopy is an analytical technique used to determine the presence of specific chemical bonds in milk samples and is widely used in the dairy industry to quantify major milk components such as fat and protein. Whilst there are many other potentially valuable applications for FT-MIR spectra in the dairy industry, the accuracy of trait prediction can be hindered by a number of sources of unwanted variation. Prior to the development of FT-MIR spectra prediction models, it is important that these sources of variation are addressed to improve the quality of FT-MIR spectral data and prepare it for downstream analysis. In this section, I will discuss methods for improving the quality of FT-MIR spectral data, and will address topics related to the development of prediction models and validation approaches for FT-MIR spectra phenotyping applications.

8.2.1 Pre-processing of FT-MIR spectra to improve data quality

Sources of variation in FT-MIR spectra that affect trait prediction and the transferability of prediction equations to other spectral datasets include: variation in measurement for specific regions of the infrared spectrum due to the water content of milk, and variation in spectral measurements between instruments and within instruments across time. A detailed discussion of these topics has been provided in Chapter 3. In this section, we will discuss important findings related to outlier removal within multi-instrument networks and strategies for managing systematic variation in spectral measurement between and within instruments across time.

Outlier removal within multi-instrument networks

Fourier-transform mid-infrared spectral data is of a high-dimensional nature with a complex correlation structure due to the lack of independence between individual wavenumber spectral responses. Commonly, outlier detection in FT-MIR spectral datasets is conducted using multivariate methods, employing metrics like the squared Mahalanobis distance (MD), which is an indicator of the distance between each spectral record and the average expected spectral response. In Chapter 3, we presented a study where the spectral data were collected from a multi-instrument network. When we evaluated MD values based on the distance between each spectral record and the average spectral response across all instruments, we showed that the distribution of MD values was multimodal and that variance structures for each instrument were not homogeneous. This highlighted that for milk samples analysed within a multi-instrument network, it is important to conduct outlier detection in a manner that takes into account instrument-specific variance structures and is based on the distance of each spectral record from the instrument-specific average spectral response. Failure to do this could result in the removal of a large proportion of records from one instrument and not removing anomalies from others. Within-instrument outlier identification and removal has been employed in Chapter 4 of this thesis prior to the evaluation of prediction models, and in Chapter 6 prior to generating adjusted wavenumber and milk composition phenotypes.

Managing systematic instrument variation

Variation in spectral measurements between instruments and within instruments across time can result in prediction errors and bias. To address these instrument measurement differences, a common approach is to adjust FT-MIR trait predictions by instrument-specific correction coefficients that have been previously evaluated from the analysis of reference samples (Lynch et al., 2006). However, this approach is only possible where reference samples have been analysed on the instrument and a trait-specific set of correction coefficients have been evaluated. Because trait-specific correction coefficients are not always available, there has been an increased interest in standardizing individual FT-MIR spectra wavenumbers directly.

In Chapter 3 we compared two strategies for standardizing individual spectra wavenumbers. The first strategy, piecewise direct standardization (PDS) (Grelet et al., 2015, 2017), utilises milk-based reference samples measured across all instruments in a network to relate the response for each wavenumber on the primary instrument to a small window around the same wavenumber on each secondary instrument. The second strategy, retroactive percentile standardization (RPS), is not

reliant on spectra from milk-based references samples, but instead uses spectral responses for each wavenumber from milk testing samples to map the primary/secondary instrument relationships (Bonfatti et al., 2017a). To assess the performance of these standardization strategies for reducing the prediction errors of major milk composition traits, we applied both strategies to a large dataset of over 2 million spectra records (Chapter 3). We showed that PDS provided the most consistent reduction in prediction errors across time, with reductions of between 38% and 63% for major milk composition traits. We also showed that similar reductions could be achieved using RPS, provided that correction coefficients were updated regularly, and adjustments were made to the approach to account for expected regional differences in milk composition.

Although it has been shown in previous studies (Grelet et al., 2015, 2017) and in Chapter 3 of this thesis that PDS applied to individual wavenumbers can consistently reduce prediction errors for FT-MIR predicted traits, the downside of this approach is that it requires the analysis of identical reference samples across all instruments. This means that to standardize spectra from instruments across multiple countries, global reference sample sharing would be required. Reference sample sharing and standardization between instruments in different countries already takes place across the European OptiMIR network, a transnational network that includes ~69 FT-MIR spectrometers in 29 milk laboratories across seven countries (Belgium, England, Ireland, France, Germany, Luxembourg, Scotland). Outside Europe, however, this reference sample sharing and standardization process is not common practice. For other countries like New Zealand to participate in a process like this, there are barriers related to sample preservation and bio-security issues with sharing milk samples between countries. Instrument manufacturers such as FOSS (Hillerød, Denmark) and Bentley (Chaska, MN) have proposed alternative standardization approaches that are not reliant on perishable milk samples. The FOSS procedure uses a liquid equaliser with a known spectral response to adjust spectral results (Winning et al., 2014), whereas the Bentley procedure uses a polystyrene film to adjust for interferometer laser frequency shifts across time (Gupta et al., 1995), and infrared flow cell information to adjust for shifts in absorbance measurement (Parsons and Lyder, 2018). These within-instrument standardization procedures offer promise for automatic spectral standardization and the sharing of prediction equations between countries because they are not reliant on perishable milk samples. However, there are no independent studies to validate the effectiveness of these procedures. Validating the effectiveness of these standardization procedures is an important area for future research, because if the procedures work well, they could provide a robust way for spectral data from different networks/countries to be consolidated, and lead to more accurate trait prediction.

8.2.2 Trait prediction

Trait prediction based on FT-MIR spectra for major milk components is already widely used to quantify concentrations of fat and protein for dairy cattle. Applications using FT-MIR spectra to predict other traits are appealing because of the opportunity to obtain indicator traits across large numbers of animals at little or no marginal cost, due to the spectral data already being available as a by-product of routine milk testing. The use of FT-MIR spectra as a phenotyping strategy has been widely studied for traits that can be directly measured in milk such as individual fatty acids (Bonfatti et al., 2016; Lopez-Villalobos et al., 2014; Rutten et al., 2009; Soyeurt et al., 2006) and protein fractions (Bonfatti et al., 2016; De Marchi et al., 2009a; Lopez-Villalobos et al., 2009; McDermott et al., 2016; Rutten et al., 2011; Soyeurt et al., 2012), and traits related to milk technological properties (Bittante et al., 2014; Cecchinato et al., 2009; De Marchi et al., 2009b; Toffanin et al., 2015; Visentin et al., 2015). Other studies have focussed on traits not directly measurable in milk, including those related to pregnancy (Brand et al., 2021; Delhez et al., 2020; Lainé et al., 2017; Toledo-Alvarado et al., 2018a), animal health (Belay et al., 2017; Grelet et al., 2016; Ho et al., 2021; Luke et al., 2019b) and the environment (Bittante and Cipolat-Gotet, 2018; Mitchell et al., 2005; Vanlierde et al., 2013, 2015). In this section we will discuss FT-MIR spectra trait prediction within the context of two types of traits that were studied as part of this thesis: fatty acids and protein fractions measured directly in the milk using chromatography methods (Chapter 7); and pregnancy status, indirectly inferred from artificial insemination (AI) and calving information (Chapter 4). We also briefly discuss different types of models used for FT-MIR spectra trait prediction and highlight strategies to account for other on-farm and stage of lactation effects.

Trait predictions for individual fatty acid and protein traits

Prediction models based on FT-MIR spectra for individual fatty acids and protein fractions are typically evaluated using a modest set of samples that have corresponding trait values measured by gas or liquid chromatography. In previous studies, prediction accuracies for these traits have varied, often due to factors including breed composition and differences in farming systems, the number of samples and extent of trait variation present in the calibration model dataset, and the validation strategy used. In Chapter 7, we presented a study where milk fatty acids and protein fractions were predicted using ~2,000 FT-MIR spectra records from crossbred cows raised in a New Zealand seasonal calving pasture-based dairy system. Using a cow-independent validation approach we reported prediction accuracies that were consistent with those from previous studies

of individual fatty acids (Bonfatti et al., 2016; Lopez-Villalobos et al., 2014; Rutten et al., 2009; Soyeurt et al., 2006, 2011) and protein fractions (Bonfatti et al., 2016; De Marchi et al., 2009a; Lopez-Villalobos et al., 2009; McDermott et al., 2016; Rutten et al., 2011; Soyeurt et al., 2012). Based on guidelines outlined by Fuentes-Pila et al. (1996), 21 of 23 individual and grouped fatty acids, and all six protein fractions included in our study had good or satisfactory predictions. This confirmed the findings of previous studies and highlighted the potential of FT-MIR spectra to provide useful proxies for directly measured fatty acids and protein fractions. Validation of the prediction equations developed from the study presented in Chapter 7 is an important step for future research, to confirm their utility for prediction and use in future breeding programs.

Prediction of pregnancy status

Pregnancy status is an appealing target for FT-MIR spectra prediction because of the important role that timely pregnancy diagnosis has in effective herd management of dairy cattle. Previous studies have typically utilised other available data sources such as AI and calving information to infer pregnancy status, thus enabling the use of large datasets for developing and validating FT-MIR prediction models (Brand et al., 2021; Delhez et al., 2020; Toledo-Alvarado et al., 2018a). Notably, however, those studies have important differences in the manner in which records are selected for inclusion in analysis and how records are classified as pregnant or non-pregnant.

In Chapter 4, we developed PLS-DA models to assess the accuracy of using FT-MIR spectra to predict pregnancy status for different strategies of inclusion and classifying records as pregnant and non-pregnant, broadly based on three previous studies (Brand et al., 2021; Delhez et al., 2020; Toledo-Alvarado et al., 2018a). Using a dataset of over 800,000 spectra records from seasonal calving pasture-based herds, we showed that there were large differences in prediction accuracy depending on the strategy used for inclusion and classification of spectra records. When we assigned spectra records as non-pregnant if the milk sample was taken prior to the first mating, and pregnant if the milk sample was taken after successful AI (broadly similar to the definition used by Brand et al., 2021), we encountered problems in prediction because pregnancy status using that definition was highly confounded with stage of lactation. Specifically, based on herd-independent validation, the sensitivity to correctly classify records as pregnant and the specificity to correctly classify records as non-pregnant were above 0.90. However, when prediction equations were applied to an independent pregnancy-associated glycoproteins (PAG) dataset, the specificity to correctly classify non-pregnant records was particularly poor (0.002). Closer examination of pregnancy predictions revealed that when the allocation of pregnancy

status was so highly confounded with stage of lactation, there was a tendency to assign nearly all PAG records as pregnant. An alternative strategy was to assign records as non-pregnant if the milk sample was taken before successful AI or there was no subsequent calving recorded for the cow (broadly similar to the definition used by Delhez et al., 2020). This resulted in a better representation of non-pregnant records across lactation, and the most consistently accurate pregnancy prediction models. Using this strategy, based on herd-independent validation, the sensitivity to correctly classify records as pregnant and the specificity to correctly classify records as non-pregnant were both ~0.60. When the prediction equation was applied to an external PAG dataset, the sensitivity of correctly classifying records as pregnant was 0.67 and the specificity of correctly classifying records as non-pregnant was 0.57. The consistency in validation accuracy for the herd-independent and PAG datasets demonstrated that by including a better representation of non-pregnant records across lactation, the effect of confounding between pregnancy status and stage of lactation was reduced. Notably, however, prediction sensitivity and specificity values were only moderate (~0.60), indicating that FT-MIR predictions from these models are not accurate enough to be used as a sole indicator of pregnancy status.

The prediction accuracies we present in Chapter 4 for classifying pregnancy status are similar to those reported in other studies (Delhez et al., 2020, Khanal and Tempelman, 2022, Toledo-Alvarado et al., 2018a). In contrast, the prediction accuracies reported by Brand et al. (2021) were significantly higher, possible in part due to confounding between pregnant/non-pregnant status and stage of lactation, a point that has also been highlighted by Khanal et al. (2022). Nonetheless, the gains in prediction accuracy Brand et al. (2021) observed from adopting a deep learning approach were notable. Specifically, compared to using a PLS-DA model, when they used a deep learning model the sensitivity to correctly classify records as pregnant increased from 0.73 to 0.90, and the specificity to correctly classify records as non-pregnant increased from 0.82 to 0.92. In the study we presented in Chapter 4, for a subset of modelling strategies, we also assessed pregnancy status prediction accuracies for two types of deep learning models, one using a multilayer perceptron (MLP) feed-forward artificial neural network to classify pregnancy status based on raw spectra, and the other using a convolutional neural network (CNN). We did not observe gains in prediction accuracy of a similar magnitude to Brand et al. (2021), with prediction accuracies for the MLP approach being similar to those from PLS-DA models, and prediction accuracies for an image-based CNN approach being only marginally better than those from PLS-DA models. These findings highlighted that deep learning approaches applied to FT-MIR spectra can provide improvements to prediction accuracy, but that their utility can vary considerably between studies.

Trait prediction models

Trait prediction models developed from FT-MIR spectra are typically based on small datasets with more predictors than observations. For this reason, approaches based on PLS regression are commonly used to reduce the predictors to a smaller set of uncorrelated components, from which least squares regression can be performed. Although PLS regression is the most widely-used method for developing FT-MIR spectra calibration equations, there are studies that employ Bayesian methods (Bonfatti et al., 2017b; El Jabri et al., 2019; Ferragina et al., 2015; Toledo-Alvarado et al., 2018a) or other machine learning algorithms (Brand et al., 2021; Contla Hernández et al., 2021; Denholm et al., 2020; Dórea et al., 2018; Frizzarin et al., 2021a, 2021b; Hempstalk et al., 2015; Pralle et al., 2018).

Typically, FT-MIR spectra prediction models are based on spectral data alone, but there are studies that also include factors to account for other on-farm and stage of lactation effects. Ho et al. (2019) used PLS-DA models to examine the potential of using FT-MIR spectra alongside other FT-MIR derived traits and on-farm data for discriminating cows of good versus poor fertility outcomes. In that study, 10-fold random cross-validation AUC values increased from 0.72 to 0.75 when fertility genomic estimate breeding values and animal genotypes were included in the models. Similarly, Toledo-Alvarado et al. (2018a) reported that cross-validation AUC values for pregnancy status predictions increased from 0.57 to 0.68 when year and herd were fitted alongside FT-MIR spectra. In Chapter 5, we examined different ways of accounting for stage of lactation in pregnancy status prediction models, either by including it as a predictor in the model, or by pre-adjusting spectra for stage of lactation before fitting the model. We also assessed whether the accuracy of prediction could be improved by fitting separate models for different stages of lactation. Our findings showed that prediction accuracies for PLS-DA models that included stage of lactation as a predictor were only marginally different to those for models that used FT-MIR spectra alone as predictors. We also observed that the prediction accuracies for models where the spectra were pre-adjusted for stage of lactation effects were consistently lower than those where the spectra were not pre-adjusted. Interestingly, when we fitted separate models for different stages of lactation, we observed consistently higher cross-validation AUC values after 210 days of lactation (0.68–0.76), compared to up to 210 days of lactation (0.64 to 0.68). This was consistent with findings by Delhez et al. (2020) where they developed separate models for different stages of lactation and found that models using data after 150 days of pregnancy had promising prediction accuracies with AUC values between 0.71 to 0.86, compared to AUC values of 0.63 for up to 150 days of pregnancy.

When using FT-MIR spectra to develop trait prediction models, there are multiple approaches (e.g., PLS, Bayesian, other machine/deep-learning approaches) that can be used. Additionally, knowledge of the trait can be useful when deciding whether to include predictors in the model other than the FT-MIR spectra. Of particular importance is to identify and minimise the impact of any confounding effects related to climate, diet and/or stage of lactation. Including multiple seasons of data across a range of herds may help to address some of these confounding issues, however the incorporation of other information such as knowledge of climatic differences, feed management and supplementation may also play an important role. Using this information, it may be possible to develop models on homogenous sets of herds based on dietary systems and regional classification.

8.2.3 Validation of FT-MIR prediction equations

Validation of an FT-MIR prediction equation is critically important to understanding the level of accuracy we can expect when that equation is applied to other spectral datasets. A common approach to validation is to use record- or cow-independent validation, whereby the prediction model is developed on a subset of the available records and the remaining records are used to validate the accuracy of the model. However, studies have shown that record- or cow-independent validation can overinflate prediction accuracies, compared to herd-independent validation (Dórea et al., 2018; Lahart et al., 2019; Luke et al., 2019b; Wang and Bovenhuis, 2019). This highlights that validation should ideally be conducted using a dataset that is from an independent herd, trial or season. However, a recent review of validation strategies used for FT-MIR predicted traits found that of 113 studies, 67 used internal record-independent validation, 15 used cow-independent validation, and only 17 conducted validation using an independent dataset based on herd, trial or season (Bresolin and Dórea, 2020). Given that it is common practice to use internal record- or cow-independent validation to assess the accuracy of an FT-MIR prediction model, it is important that when the potential utility of a prediction equation is being assessed, that consideration is given to the validation strategy that has been used in the study.

A robust validation strategy should ensure that expected accuracies for FT-MIR predicted traits are not overstated and should ideally be based on herd- or trial-independent validation datasets. However, in the study we presented in Chapter 4, we demonstrated that even when a herd-independent validation is used, estimates of prediction accuracy can be misleading when there is systematic confounding between the trait of interest and other factors such as stage of lactation. Systematic confounding in FT-MIR prediction models due to changes in milk

composition across lactation can be particularly problematic for spectral data from seasonal pasture-based farming systems, due to the use of compact calving periods to ensure that peak lactation volumes are matched with peak grass growth (Timlin et al., 2021). In the New Zealand setting, these issues are further exacerbated by the use of palm kernel extract and maize silage supplements to offset the effect of high stocking rates. This is particularly problematic when FT-MIR spectra are used to predict a trait like pregnancy status, because as lactation progresses, changes in milk composition due to stage of lactation and dietary supplements coincide with the advent of a cow becoming pregnant (Khanal and Tempelman, 2022; Tiplady et al., 2022). In general, when a model is developed to predict trait values using FT-MIR spectra, careful consideration should be given to how the model can account for the effect of confounding factors such as lactation stage, feed management and seasonality.

8.2.4 Machine learning approaches

Partial least squares regression is a machine learning technique that is widely used for developing calibration models for FT-MIR spectral datasets. This technique is a supervised methodology that uses principal component analysis to reduce the large number of correlated wavenumber responses to a smaller set of uncorrelated latent variables from which least squares regression can be performed. Recently, there has also been a high level of interest in the use of other machine learning algorithms to generate prediction models from FT-MIR spectral datasets, including for the prediction of health traits (Contla Hernández et al., 2021; Denholm et al., 2020; Pralle et al., 2018), pregnancy (Brand et al., 2021), milk quality traits (Frizzarin et al., 2021a), dry matter intake (Dórea et al., 2018) and classification of cow diet (Frizzarin et al., 2021b).

Contla et al. (2021) compared a range of machine learning techniques for predicting health status from FT-MIR spectra. They found that a neural network (NN) outperformed PLS-DA and other support vector machine (SVM), random forest (RF) and ensemble approaches for classifying animals based on health status. In studies that compared PLS models to other machine learning methods for evaluating β-hydroxybuterate, Pralle et al. (2018) showed that artificial neural network (ANN) models outperformed PLS, whilst Mota et al. (2021) showed that gradient boosting machine and RF models outperformed PLS. Brand et al. (2021) compared pregnancy status classification accuracies for PLS-DA and two deep learning methods, one developed using genetic algorithms for feature selection and network design, and the other using transfer learning models with a pre-trained Dense Convolutional Network (DenseNet). Prediction accuracies for the pre-trained DenseNet model outperformed those of the PLS-DA model. In the study we presented in Chapter 4, we also compared the prediction performance of a DenseNet model to

that of PLS-DA for pregnancy status classification, and showed that although the DenseNet models outperformed the PLS-DA models, the differences in performance were only marginal and the gains from employing a DenseNet model were much lower than those observed by Brand et al. (2021).

Frizzarin et al. (2021a) compared prediction accuracies for milk quality traits between PLS/PLS-DA models to other approaches including ridge regression (RR), least absolute shrinkage and selection operator (LASSO), elastic net (EN), RF, NN, SVM and boosting decision trees. Results varied for different traits, but for continuous traits, PLS was outperformed by other machine learning approaches, whereas for class traits, PLS-DA and SVM models had the highest prediction accuracies. In a separate study, Frizzarin et al. (2021b) compared prediction accuracies between PLS-DA and a number of other machine learning approaches for discriminating between grass-fed versus nongrass-fed milks. They showed that PLS-DA outperformed all other classification methods including RR, LASSO, EN, RF, SVM and boosting decision trees. Notably, the performance of PLS-DA was very similar to that of linear discriminant analysis (LDA), something I have also observed whilst conducting the work for this thesis. Although I have not published these results, for the pregnancy prediction models presented in Chapter 4, when LDA was used instead of PLS-DA the results were almost identical, even though the run-time for LDA was often 5 to 10 times faster.

Although PLS and PLS-DA models are widely used to develop FT-MIR prediction equations, it is clear that other machine learning approaches have merit and that in some instances, they may be able to provide more accurate trait predictions. Of particular note is the emergence of deep learning approaches which have been recently shown to provide promising predictions for bovine tuberculosis (Denholm et al., 2020) and pregnancy status (Brand et al., 2021) in individual cows. Deep learning is a subclass of machine learning that extracts features from data using neural networks with multiple layers of densely interconnected processing nodes. The complexity of these networks enable training models to be developed on datasets with multiple connections, which make them a good choice for managing high-dimensional datasets such as those presented from FT-MIR spectra. However, applications for deep learning models rely on large datasets, so will not always be a suitable choice. There are no absolute rules as to which type of model is the best one to use for any given study. Partial least squares models are often a good choice due to their relative simplicity, but aside from the choice of machine learning method, input data quality, model optimisation and validation are important, as overfitting complex models can lead to misleading results (Mendez et al., 2019; Shine and Murphy, 2022).

8.3 The genetics of FT-MIR predicted traits

The incorporation of FT-MIR predictions for major milk composition traits into breeding programs has played a significant role in the transformation of dairy cattle milk composition. Whilst the accuracy of FT-MIR predictions is an important indicator of their utility for traits where we are interested in phenotypic values, for breeding purposes, the extent of genetic variation in the benchmarked trait, the heritability of the FT-MIR predicted trait, and the genetic correlation between the benchmarked and FT-MIR predicted trait are also important. This is because the heritability of the FT-MIR predicted trait defines the extent of genetic variation that could potentially be exploited in animal breeding programs, whilst the genetic correlation between the benchmarked or directly measured trait and the FT-MIR predicted trait defines the breeding progress we could expect in the directly measured trait if we were to select for animals based on the FT-MIR predicted trait.

In Chapter 5, we have reviewed the genetics of FT-MIR predicted traits including fatty acids (Bonfatti et al., 2017d; Lopez-Villalobos et al., 2014; Rutten et al., 2010; Soyeurt et al., 2007b) and protein fractions (Arnould et al., 2009b; Bonfatti et al., 2017d; Buitenhuis et al., 2016; Lopez-Villalobos et al., 2009; Sanchez et al., 2017a; Soyeurt et al., 2007a). Moderate to high heritability estimates were reported for many of these traits, but few studies presented genetic correlations between benchmarked or directly measured traits and corresponding FT-MIR predicted traits. For those studies, genetic correlations between directly measured and FT-MIR predicted traits were generally high for fatty acids (0.64 to 0.99) (Bonfatti et al., 2017d; Rutten et al., 2010) and protein fractions (0.57 to 0.94) (Bonfatti et al., 2017d). To further understand the genetic characteristics of directly measured and FT-MIR predicted fatty acids and protein fractions, we conducted a study to evaluate heritability and genetic correlation estimates for 17 individual fatty acids, 6 grouped fatty acids and 6 individual protein fractions (Chapter 7). Heritability estimates for most 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. These results were broadly consistent with those of previous studies (Bonfatti et al., 2017d; Rutten et al., 2010). In general, our findings were that lower heritability estimates were observed for directly measured traits, compared to FT-MIR predicted traits. This was caused by higher total variation in the directly measured traits, but a lower magnitude increase in the additive genetic variance component, compared to the FT-MIR predicted traits. Nevertheless, the genetic correlations between directly measured and FT-MIR predicted traits remained high and were mostly greater than 0.75. Moderate to high heritability estimates across individual fatty acid and protein traits indicate that these traits have genetic variation 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 protein fractions indicate that selection based on FT-MIR predictions for these traits could provide favourable genetic gains in milk fatty acid and protein composition.

In Chapter 5, we also reviewed the genetics of FT-MIR predicted traits related to milk processability (Cecchinato et al., 2009; Costa et al., 2019; Visentin et al., 2017), cheese yield (Bittante et al., 2014; Cecchinato et al., 2015), animal health (Bastin et al., 2016; Belay et al., 2017; van den Berg et al., 2021a, 2021b; Luke et al., 2019a) and the environment (Kandel et al., 2017; Miglior et al., 2007; Mitchell et al., 2005; Stoop et al., 2007; Wood et al., 2003). For milk coagulation traits, moderate heritability estimates and moderate to high genetic correlations ranging from 0.71 to 0.96 were reported (Cecchinato et al., 2009). For cheese yield and nutrient recovery traits, high genetic correlations were reported, ranging from 0.76 to 0.98 (Bittante et al., 2014). With respect to using FT-MIR spectra to provide proxies for animal health traits, results were less clear, and more research is required to understand the relationships between health and fertility indicators and FT-MIR predicted traits, and to realise the value that FT-MIR spectra might add to animal health breeding goals (Bastin et al., 2016). To date, there have been few studies of the genetics of FT-MIR predictions related to methane and nitrogen outputs. For FT-MIR predictions of methane traits, studies have indicated that there is potential to incorporate these into breeding programs, but more research is required to improve the robustness and accuracy of prediction equations and make them suitable to use across a range of production systems (van Gastelen et al., 2018b; Hristov et al., 2018; Negussie et al., 2017; Vanlierde et al., 2018). For FT-MIR predicted MUN, moderate to high heritability estimates have been reported (Mitchell et al., 2005; Miglior et al., 2007; Stoop et al., 2007; Wood et al., 2003). Of those studies, Mitchell et al. (2005) was the only one that reported genetic correlations between direct measurements of MUN and FT-MIR predicted MUN, which were 0.38 and 0.23 in lactations 1 and 2, respectively. These genetic correlations were significantly lower than those observed for fatty acids and milk processibility traits (Bittante et al., 2014; Rutten et al., 2010), and indicate that directly measured MUN and FT-MIR predicted MUN may be genetically different traits. Moreover, the heritability estimates across studies of FT-MIR predicted MUN were highly variable, indicating that underlying instability may be present in these prediction equations. Notably, however, promising genetic parameter estimates have been reported for FT-MIR predicted BUN, with heritability estimates ranging from 0.08 to 0.13 and genetic correlations between BUN and its FT-MIR prediction ranging from 0.96 to 0.98 (van den Berg et al., 2021b). Further research is required to determine the role that FT-MIR predicted nitrogen indicators could have in reducing nitrogen outputs from dairy cattle milk production systems.

8.4 Genome-wide association studies

Genome-wide association studies are commonly used to associate genetic variations with complex dairy cattle traits. Many GWAS have been published in the last decade for FT-MIR predicted major milk production traits (Jiang et al., 2010; Kemper et al., 2015b; Littlejohn et al., 2016; Lopdell et al., 2017; Raven et al., 2014), and for fatty acid and protein fractions (Bouwman et al., 2011; Buitenhuis et al., 2014, 2016; Li et al., 2014; Sanchez et al., 2016). However, there are comparatively few studies that report GWAS results for individual FT-MIR wavenumber phenotypes. Benedet et al. (2019) examined relationships between FT-MIR wavenumber phenotypes and a subset of SNP previously implicated in a GWAS of milk composition and fatty acid traits. Two other studies used medium density SNP-chip genotypes to conduct GWAS on a subset of wavenumbers, identified either by clustering analysis (Wang and Bovenhuis, 2018), or by using phenotypic correlation structures and heritability estimates within each breed (Zaalberg et al., 2020). Across those studies, a number of FT-MIR wavenumber QTL were identified, many of which were in genomic regions that have been previously reported from GWAS for major milk composition traits, but new regions were also identified. These findings highlighted the potential that GWAS on individual FT-MIR wavenumbers may have for the discovery of new QTL, particularly if analyses are conducted using higher density genotypes across larger numbers of animals.

8.4.1 Computational challenges for conducting large-scale GWAS

Computational challenges already exist for conducting GWAS across even a handful of traits, due to the increase in the number of genotyped individuals and the increase in density of available genotypes, including to the whole-genome sequence level. Mixed-linear model-based methods for conducting GWAS such as GCTA-MLMA (Yang et al., 2011) can become prohibitively slow as the number of variants and genotyped individuals increases, and frequently require subsampling to use these methods within acceptable computational constraints. Developing strategies for managing GWAS on large numbers of densely genotyped individuals is an active area of research, and has resulted in the development of more efficient algorithms such as BOLT-LMM (Loh et al., 2015, 2018) and fastGWA (Jiang et al., 2019c) that enable the processing of these datasets within acceptable timeframes and computational constraints. BOLT-LMM software is capable of running the ~400k UK biobank samples in a few days on a single computational node (Loh et al., 2018). The fastGWA software provides further reductions in algorithmic complexity, making it possible to run GWAS on ~400k UK biobank samples in around 20 minutes, compared to 22

hours for BOLT-LMM on the same hardware. The development of these algorithms have made GWAS across large populations of densely genotyped individuals feasible, and are of particular importance when we aim to conduct GWAS across datasets with large numbers of phenotypes.

8.4.2 Sequence-based GWAS of individual FT-MIR wavenumbers

In Chapter 6 we used BOLT-LMM software (Loh et al., 2015) to run sequence-based GWAS for 895 individual FT-MIR wavenumber phenotypes and three FT-MIR predicted major milk composition traits, using records for 38,085 mixed-breed New Zealand dairy cattle. Separate GWAS were conducted for each of 898 phenotypes using 17,873,880 imputed sequence variants, and mixed-model association statistics were evaluated under an infinitesimal model (as defined by BOLT-LMM software) using a genomic relationship matrix (GRM) to account for population structure. To avoid proximal contamination, a leave-one-segment-out (LOSO) approach was used with segments of 5 Mbp used to subdivide the autosomes, and additive effects for each SNP were evaluated against a conservative Bonferroni significance threshold to identify QTL.

Many GWAS for dairy cattle include only a handful of traits, making it possible to manually assess QTL for underlying genes and potential protein-coding effects. However, because our GWAS included such a large number of traits, this was not feasible. To distinguish between multiple QTL segregating within the same region of a chromosome, an iterative conditional methodology was developed whereby the most significant variant was identified in each peak and added to a set of covariates that were used in subsequent GWAS iterations. Subsequent iterations were conducted on chromosomes that retained significant effects, and the process was repeated until these analyses ceased to highlight significant effects. Employing this approach resulted in the identification of a list of variants for each phenotype that aimed to collectively capture all the significant association analysis signal.

Evaluating the list of QTL identified across 898 FT-MIR individual wavenumber and predicted trait phenotypes provided a further challenge, as it was infeasible to manually assess the genes underlying such a large number of effects. To address this challenge, two approaches were used to identify causative genes and variants underlying QTL using methods that informed on potential protein function-based effects and regulatory mechanisms, as described in Lopdell et al. (2017). Potential protein function-based effects were identified using SnpEff (version 4.1d; build 2015-04-13) (Cingolani et al., 2012) and Ensembl UMD3.1.78 gene annotations, and variants in QTL regions were filtered to include only those in high LD ($R^2 > 0.9$) with a splice region variant, or a moderate or high impact coding variant. Causative genes with co-segregating expression QTL (eQTL) were identified using a large previously described RNA sequence resource and eQTL mapping methodology (Lopdell et al., 2017).

Gains in number of QTL and power from individual wavenumber GWAS

In the GWAS of individual FT-MIR wavenumbers presented in Chapter 6, a key area of interest was to assess how well candidate causal genes and variants could be identified using individual FT-MIR wavenumber phenotypes, compared to using FT-MIR predictions of major milk composition traits. Specifically, we were interested in comparing the number of QTL identified, the size of QTL, and the relationship between trait QTL and plausible candidates and co-locating eQTL. In total we identified 450 1-Mbp genomic regions with significant FT-MIR wavenumber QTL, compared to 246 1-Mbp genomic regions with QTL identified for FT-MIR predicted milk composition traits.

Protein function-based effects for wavenumber QTL were identified within 42 1-Mbp regions that encompassed 55 effects with a corresponding splice region variant, or a moderate or high impact coding variant. For FT-MIR predicted milk composition traits, 27 effects were identified with a corresponding splice region variant or a moderate or high impact coding variant within 15 1-Mbp regions. Among the FT-MIR wavenumber QTL observed, those with the largest effects were in perfect LD with missense mutations in the *ABCG2*, *PAEP* and *DGAT1* genes, all of which have been proposed to have major impacts on milk composition (Cohen-Zinder et al., 2005; Ganai et al., 2009; Grisart et al., 2002). Notably, significant association effects were also observed in those genes for FT-MIR predicted milk composition traits, however the p-values for the most significant FT-MIR wavenumber were always more significant than comparable values for any of the FT-MIR predicted milk composition traits. A number of other wavenumber QTL were identified that were in strong LD with either a splice region variant, or a moderate or high impact coding variant, that had no corresponding FT-MIR predicted-trait QTL. This included a previously unreported null mutation in the *ABO* gene that has a potential role in changing milk oligosaccharide profiles (Liu et al., 2019; Poulsen et al., 2019). Other QTL highlighted genes that appeared novel: *FCRLA*, *WDR97*, *KRT9*, *KRT16*, *KRT17*, *HID1*, *KCNK1* and *NXF1*.

Scrutinization of wavenumber and predicted-trait QTL identified 38 wavenumber QTL and 25 predicted-trait QTL that co-located to an eQTL. For the wavenumber QTL, in many cases the tag variant for the wavenumber QTL was also the top variant for the co-locating eQTL. Many of the genes corresponding to these effects have previously been reported in other studies of bovine milk composition, but two genes appeared to be novel: *KRTCAP2* and *FA2H*. In instances where a co-locating eQTL was identified within FT-MIR wavenumbers and within predicted milk composition traits, there was a common pattern, whereby the wavenumber QTL had more highly significant p-values, compared to the p-values for predicted milk composition traits. Other wavenumber QTL where a co-locating eQTL was identified within FT-MIR wavenumbers, but

not within FT-MIR predicted milk composition traits, highlighted a number of genes that have been previously reported and also genes that appeared novel: *CLDN8*, *CSTB*, *TAB2*, *LAPTM4A*, *CAPN5*, *PMP22*, *HID1* and *THRB*.

The results we presented in Chapter 6 underscore the gain in power available when GWAS is conducted on individual FT-MIR wavenumber phenotypes, compared to GWAS using FT-MIR predicted milk composition phenotypes. A greater number of QTL and putative causative genes and variants were identified, and where the QTL were common to FT-MIR wavenumber phenotypes and predicted milk composition phenotypes, the p-values for the most significant FT-MIR wavenumber were always more significant than the comparable values for any of the FT-MIR predicted milk composition traits. Not only that, but in many instances, the QTL identified from wavenumber phenotypes were often in perfect LD with a protein-coding mutation or were the same as the top SNP from eQTL analysis.

8.4.3 Areas of potential improvement for FT-MIR wavenumber GWAS

Conducting a sequence-based GWAS across individual FT-MIR wavenumber phenotypes, and dissecting QTL using variant annotation information and a mammary RNA-seq resource, enabled the identification of candidate causative genes and variants for a substantial number of loci (Chapter 6). Further, through employing an iterative approach, it was possible to distinguish between multiple QTL segregating within the same region of a chromosome. However, there are several areas of refinement to that study that could be expected to enable identification of further QTL. Firstly, 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 additional QTL and reveal additional candidate mutations given improvements in accompanying transcript annotations. Secondly, the approach we used could be extended to account for non-additive QTL, in a similar manner to that outlined in Reynolds et al. (2021). Thirdly, rather than iteratively selecting the top variant from each peak based on the p-value of the association effect, a more sophisticated approach could be used to select the representative variants of each peak by utilising gene annotation information and other genomic and molecular data sources. Finally, improved variant prediction methods and integration of other functional datasets (e.g., ChIP-seq) could be used to enhance fine mapping and candidate variant identification.

8.5 Comparison of QTL for directly measured and FT-MIR predicted fatty acid and protein traits

There have been a number of GWAS conducted on fatty acids in bovine milk samples determined by gas chromatography using a range of genotype densities (Bouwman et al., 2011; Buitenhuis et al., 2014; Palombo et al., 2018). Many GWAS have also been conducted on FT-MIR predicted fatty acids in milk (Cruz et al., 2019; Freitas et al., 2020; Iung et al., 2019; Olsen et al., 2017; Sanchez et al., 2019). Similarly, there have been multiple GWAS conducted on protein fractions in milk samples determined by high-performance liquid chromatography (Buitenhuis et al., 2016; Pegolo et al., 2018; Schopen et al., 2011); and GWAS conducted on FT-MIR predicted protein fractions using whole-genome sequence (Sanchez et al., 2017b, 2019).

In the previously mentioned GWAS for milk fatty acids and protein fractions, the studies focussed on either directly measured or FT-MIR predicted traits, but none conducted GWAS across both directly measured and FT-MIR predicted traits, or made comparisons between the QTL observed for each type of trait. This was a focus for the study outlined in Chapter 7 of this thesis. Of particular interest was whether the high genetic correlations we observed between directly measured and FT-MIR predicted traits were underpinned by a similar genetic architecture. To assess this, we conducted GWAS on direct measurements and FT-MIR predictions for 23 fatty acid and 6 protein traits, and compared the QTL for each pair of measured/predicted traits (Chapter 7). This resulted in the identification of 40,946 variants with significant effects for directly measured traits and 18,843 variants with significant association effects for FT-MIR predicted traits. There were more than twice as many variants with significant effects for directly measured traits, compared to FT-MIR predicted traits, which was largely due to 20,949 variants with significant effects on BTA26 for directly measured traits compared to only 110 variants with significant effects on BTA26 for FT-MIR predicted traits.

To assess the candidacy of QTL, relevant protein coding variants in high LD ($R^2 > 0.7$) with the most highly associated variant from each peak were identified. This resulted in the identification of trait QTL for fatty acids with likely candidates in the *DGAT1*, *CCDC57*, *SCD* and *GPAT4* genes, but the QTL underpinned by *SCD* were 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. For the traits underlying the genetic signals in the *SCD* and *LTF* genes (C10:1, C14:1, Lf), there was a consistent pattern where the milk component was 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 to the directly measured trait. While 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, 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 implies that changes in milk composition attributable to these two genes may be difficult to quantify directly using FT-MIR wavenumber spectra.

Interestingly, in the study presented in Chapter 7, instances also arose where a QTL was observed for an FT-MIR predicted trait, but there was no corresponding QTL observed in the directly measured trait. An example of this was where large association effects were observed within the *DGAT1* and *GPAT4* genes for FT-MIR predicted C18:3 *cis*-3, but corresponding association effects were not observed for directly measured C18:3 *cis*-3. Similarly, a large association effect was observed for FT-MIR predicted β-CN within the *PAEP* gene, but no corresponding association effect was observed in directly measured β-CN. The presence of QTL with significant effects for an FT-MIR predicted trait without any corresponding QTL for the directly measured trait is not entirely surprising. This is because FT-MIR predicted traits are a weighted function of spectral values for individual wavenumbers, each of which is underpinned by multiple genetic signals and QTL (Benedet et al., 2019; Tiplady et al., 2021b; Wang and Bovenhuis, 2018; Zaalberg et al., 2020), some of which will be specifically related to the trait of interest and some that will not. Differences between the QTL for directly measured and FT-MIR predicted traits may be of particular importance when SNP-based approaches are used to estimate 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 SNP.

8.6 Future perspectives

Characterisation of milk composition in dairy cattle has a long history of scientific and commercial interest. Over the last 100 years, advances in refrigeration and transportation technologies, and the availability of automated on-farm milk meters have resulted in a shift from labour intensive, on-farm collection and processing of samples, to large-scale processing of samples through centralised laboratories. More recently, advances in analytical techniques have led to the widespread use of FT-MIR spectroscopy to estimate major milk components such as fat, protein and lactose. Over the last decade, there has been a significant amount of research related to the use of FT-MIR spectra to predict other traits of interest to the industry. In this section I will discuss some of the challenges still faced and highlight areas of opportunity as we look to the future use of FT-MIR spectra in dairy cattle milk production systems.

8.6.1 Managing systematic instrument variation

Due to the success of using FT-MIR spectra for phenotyping major milk composition traits and the availability of the data as a by-product of routine milk testing, there has been a high level of interest in using these data to provide indicators for other traits to the industry. However, to generate accurate FT-MIR trait predictions, there are a number of general principles and caveats that should be considered. In particular, it is important that differences in spectral measurement between instruments are accounted for as these can result in prediction errors and bias. Standardization of individual FT-MIR spectra wavenumbers using reference samples is an approach that can effectively address variation in spectral measurements between instruments and within instruments across time, but a downside to this approach is that it is reliant on the analysis of a common set of reference samples across instruments. In many instances this approach is not feasible or the data is not available for historical spectral datasets. Therefore, an important area of research going forward is to determine how spectral data from different networks/countries can be consolidated. FOSS (Hillerød, Denmark) and Bentley (Chaska, MN) both have within-instrument standardization procedures which could assist with this. However, to date the effectiveness of these procedures has not been independently evaluated. Validation of the effectiveness 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 lead to improved trait prediction accuracies.

8.6.2 Accounting for systematic confounding factors

Systematic confounding in FT-MIR prediction models due to changes in milk composition related to factors such as stage of lactation and the use of dietary supplements can be problematic. This is because of the large effect that different dietary systems (Dias, 2010; Elgersma, 2015; Oliveira et al., 2015) and levels of pasture in the diet (Butler et al., 2011; Couvreur et al., 2006; O'Callaghan et al., 2016; White et al., 2001) have on milk fatty acid composition. Moreover, even across different diets, as lactation progresses, consistently lower milk volumes (McAuliffe et al., 2016) and higher concentrations of fat and crude protein are expected (O'Callaghan et al., 2016). Confounding between FT-MIR spectra and lactation stage can be particularly problematic for the prediction of traits such as pregnancy status in pasture-based seasonal calving herds, because as lactation progresses, changes in dietary supplementation and milk composition coincide with the advent of a cow becoming pregnant (Khanal and Tempelman, 2022; Tiplady et al., 2022). As a general principle, when developing prediction models for an FT-MIR predicted trait, careful consideration should be given to any potential confounding factors, and calibration datasets should be constructed in a manner to minimise the impact they will have on estimates of trait prediction accuracy. Whilst some of the effect of these factors can be mitigated by including multiple seasons of data, integration of other information such as knowledge of feed management and supplementation may also play an important role.

8.6.3 Validation of prediction equations

Validating the accuracy of an FT-MIR trait prediction equation is typically conducted by applying the equation to a dataset of records that have not been included in the development of the prediction equation. Most commonly, validation strategies are based on record- or cow-independent datasets (Bresolin and Dórea, 2020). There is a risk though that these strategies may overinflate estimates of prediction accuracy, compared to validation using data from different herds (Luke et al., 2019b; Wang and Bovenhuis, 2019). Ideally, a robust validation strategy will ensure that expected prediction accuracies for FT-MIR predicted traits are not overstated, and should be evaluated using independent validation datasets from a different herd, trial or season (Bresolin and Dórea, 2020). Moreover, given that the use of record- or cow-independent validation is common practice, when assessing the potential utility of an already existing FT-MIR trait prediction equation, careful consideration should be given to the validation strategy that has been used in the development of the equation.

8.6.4 Machine learning approaches

Partial least squares regression and PLS-DA are the most widely-used methodologies for developing calibration models for FT-MIR spectral datasets. However, there is also an increasing interest in using other machine learning approaches to develop FT-MIR prediction equations. From the research to date, it is clear that other machine learning approaches have the potential to provide more accurate trait predictions (Brand et al., 2021; Frizzarin et al., 2021a; Contla Hernández et al., 2021; Mota et al., 2021; Pralle et al., 2018), however there are no absolute rules as to which model will provide the best predictions for any given study. Due to their relative simplicity, PLS and PLS-DA models are often good choices, but aside from the choice of machine learning method, other factors such as input data quality for the training model and a robust validation strategy are critical to ensuring that models are optimised and that the capability of a prediction model is not overestimated (Mendez et al., 2019; Shine and Murphy, 2022). This is of particular importance for more complex machine learning methods which carry a higher risk of overfitting. Ideally, analysis pipelines should be developed so that a range of different machine learning methods can be assessed with relative ease.

8.6.5 Applications

Applications using FT-MIR spectral data for trait prediction have been widely studied, including for indirect traits related to animal health and the environment, and direct traits such as individual fatty acids and protein fractions. Although there has been strong interest in using FT-MIR spectra to provide proxies for animal health and environment traits, prediction accuracies have varied across studies, and more research is required to realise the value of FT-MIR spectra for the prediction of these traits. Prediction accuracies for fatty acids and protein fractions have also varied, but have generally been improved by increasing the number of observations used to develop prediction equations, and by ensuring that a similar extent of the variation in the prediction population is represented in the calibration dataset. Moreover, moderate to high heritability and high genetic correlations between directly measured and FT-MIR predicted fatty acids and protein fractions indicate that indirect selection on FT-MIR predictions of these traits could be used in animal breeding programs to achieve desired changes to milk composition. This is a promising area for future research, however care should be taken if SNP-based approaches are used to estimate breeding values, because as shown in Chapter 7, the underlying QTL of directly measured and FT-MIR indicator traits may not always be the same.

Although the use of dietary supplements can be problematic when developing FT-MIR trait prediction equations, a growing area of interest is in the use of FT-MIR spectra to specifically target differences in diet, and discriminate between grass-fed and non-grass-fed milks. This interest has been spurred by the growth in demand for grass-fed dairy products which is driven by consumer perceptions of pasture-based production systems being more favourable in terms of health benefits, animal wellbeing and environmental sustainability (Frizzarin et al., 2021b; Joubran et al., 2021). Frizzarin et al. (2021b) showed that it was possible to use FT-MIR spectra to distinguish between milk from cows that were pasture-fed to those that were fed a combination of grass silage, maize silage and concentrates. This is a promising area of research due to the high value of grass-fed dairy products in the consumer market. However, further work is required to improve the prediction equation and validate the transferability of the equation to other spectral datasets.

8.6.6 Frequency and scope of FT-MIR spectra measurements

One of the current limitations of using FT-MIR spectra to predict traits for animal and herd management is that milk testing generally only happens at monthly or bi-monthly intervals. If spectral data were available on a more regular basis, it would enable the monitoring of changes in milk composition across time. This could be particularly helpful for identifying sudden shifts in an animal's physiological status due to health events such as onset of illness or pregnancy loss. Ideally, spectral data would be available on a daily basis, but that would only be possible if inline spectrometers were available for installation into milking sheds. Whilst this is an exciting prospect, it is looking unlikely that any such technology will be available in the near future. This is because the development of miniaturized spectrometers has been slower than expected and there are still a number of issues with portable/miniaturized mid-infrared detectors due to temperature and vibrational sensitivity (Crocombe, 2018). In the absence of inline spectrometers, an alternative possibility is that FT-MIR spectra from daily bulk tank milk samples could be used to monitor shifts in milk composition at a herd level. Whilst this would not provide insights at an individual cow level, it could provide valuable insights into shifts in herd milk composition, and the overall health and wellbeing of a herd across time.

8.7 Conclusions

Fourier-transform mid-infrared spectra plays a key role in generating phenotypes for major milk composition traits and is a cornerstone of modern dairy cattle milk payment and animal evaluation systems. The objectives of this thesis were to add to the understanding of the phenotypic and genetic characteristics of FT-MIR spectra and FT-MIR predicted traits, and to assess the role they may have in further improving dairy cattle milk production systems. A range of topics have been presented, including strategies for improving the quality of FT-MIR spectral data, development and validation of prediction models, and the assessment of the genetic characteristics of FT-MIR predicted traits and individual FT-MIR wavenumbers. Despite there being many potential applications of FT-MIR spectra for trait prediction, there are challenges in developing accurate prediction models due to the complex multivariate structure of spectral data, and due to the presence of other confounding effects such as feed supplementation and stage of lactation. Variation in spectral measurements between instruments can also be problematic, particularly when applying trait prediction equations to spectra collected from a different instrument to the one that was used for developing the prediction equation. International collaboration between research groups to standardize spectral data could facilitate the consolidation of datasets and assist with the development of better prediction equations and improved trait prediction accuracies.

Using individual FT-MIR wavenumbers as phenotypes in GWAS can enhance our understanding of the genetics underlying milk composition, and provide stronger association effects and improved power for identifying candidate causal variants, compared to conducting GWAS on FT-MIR predicted traits. In applications for genetic improvement using FT-MIR predicted traits, it is important to understand the genetic relationships between predicted traits and the true traits of interest. Moreover, it is important to be aware that even when genetic correlations between directly measured and FT-MIR predicted traits are high, the underlying QTL of each trait may not always be the same. This may be particularly important when SNP-based approaches are used to estimate breeding values for FT-MIR predicted traits. Overall, the work presented herein has added to the body of knowledge of phenotyping and genetic applications of FT-MIR spectral datasets. Although there are many potential applications of these data, there are also challenges related to the development and validation of prediction models, and differences in the genetic architecture underlying directly measured and FT-MIR predicted traits. For applications where these challenges can be addressed, FT-MIR spectral datasets have the potential to provide new insights into milk composition, and facilitate the prediction of novel traits that will improve dairy cattle milk production systems and breeding programs into the future.

Bibliography

V. Arnould, N. Gengler, and H. Soyeurt. Environmental and genetic sources of variability of stearoyl Coenzyme-A desaturase 9 activity during and across lactations. 2009a. https://orbi.uliege.be/handle/2268/27550.

V. Arnould, H. Soyeurt, N. Gengler, F. G. Colinet, M. V. Georges, C. Bertozzi, D. Portetelle, and R. Renaville. Genetic analysis of lactoferrin content in bovine milk. *Journal of Dairy Science*, 92(5): 2151–2158, 2009b. doi:10.3168/jds.2008-1255.

B. Bahar, F. O'Halloran, M. J. Callanan, S. McParland, L. Giblin, and T. Sweeney. Bovine lactoferrin (*LTF*) gene promoter haplotypes have different basal transcriptional activities. *Animal Genetics*, 42(3): 270–279, 2011. doi:10.1111/j.1365-2052.2010.02151.x.

C. Bastin, L. Théron, A. Lainé, and N. Gengler. On the role of mid-infrared predicted phenotypes in fertility and health dairy breeding programs. *Journal of Dairy Science*, 99(5):4080–4094, 2016. doi:10.3168/jds.2015-10087.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1 – 48, 2015. doi:10.18637/jss.v067.i01.

C. Bayly. *100 years of herd testing*. Livestock Improvement Corporation Ltd, Newstead, Hamilton, NZ, 2009. ISBN 978-0-473-15126-3.

T. K. Belay, M. Svendsen, Z. M. Kowalski, and T. Ådnøy. Genetic parameters of blood β-hydroxybutyrate predicted from milk infrared spectra and clinical ketosis, and their associations with milk production traits in Norwegian Red cows. *Journal of Dairy Science*, 100(8):6298–6311, 2017. doi:10.3168/jds.2016-12458.

T. K. Belay, B. S. Dagnachew, S. A. Boison, and T. Ådnøy. Prediction accuracy of direct and indirect approaches, and their relationships with prediction ability of calibration models. *Journal of Dairy Science*, 101(7):6174–6189, 2018. doi:10.3168/jds.2017-13322.

A. Benedet, P. N. Ho, R. Xiang, S. Bolormaa, M. De Marchi, M. E. Goddard, and J. E. Pryce. The use of mid-infrared spectra to map genes affecting milk composition. *Journal of Dairy Science*, 102(8): 7189–7203, 2019. doi:10.3168/jds.2018-15890.

E. N. Bergman. Hyperketonemia-ketogenesis and ketone body metabolism. *Journal of Dairy Science*, 54 (6):936–948, 1971. doi:10.3168/jds.S0022-0302(71)85950-7.

L. Bernard, C. Leroux, and Y. Chilliard. Characterisation and nutritional regulation of the main lipogenic genes in the ruminant lactating mammary gland. *Ruminant Physiology: Digestion, metabolism and impact of nutrition on gene expression, immunology and stress*, pages 295–326, 2006. https://hal.inrae.fr/hal-02813288.

S. Berry, N. Lopez-Villalobos, E. Beattie, S. Davis, L. Adams, N. Thomas, A. Ankersmit-Udy, A. Stanfield, K. Lehnert, H. Ward, J. Arias, R. Spelman, and R. Snell. 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. *New Zealand Veterinary Journal*, 58(1):1–5, 2010. doi:10.1080/00480169.2010.65053.

G. Bittante and A. Cecchinato. Genetic analysis of the Fourier-transform infrared spectra of bovine milk with emphasis on individual wavelengths related to specific chemical bonds. *Journal of Dairy Science*, 96(9):5991–6006, 2013. doi:10.3168/jds.2013-6583.

G. Bittante and C. Cipolat-Gotet. Direct and indirect predictions of enteric methane daily production, yield, and intensity per unit of milk and cheese, from fatty acids and milk Fourier-transform infrared spectra. *Journal of Dairy Science*, 101(8):7219–7235, 2018. doi:10.3168/jds.2017-14289.

G. Bittante, M. Penasa, and A. Cecchinato. Invited review: Genetics and modeling of milk coagulation properties. *Journal of Dairy Science*, 95(12):6843–6870, 2012. doi:10.3168/jds.2012-5507.

G. Bittante, A. Ferragina, C. Cipolat-Gotet, and A. Cecchinato. Comparison between genetic parameters of cheese yield and nutrient recovery or whey loss traits measured from individual model cheese-making methods or predicted from unprocessed bovine milk samples using Fourier-transform infrared spectroscopy. *Journal of Dairy Science*, 97(10):6560–6572, 2014. doi:10.3168/jds.2014-8309.

V. Bonfatti, G. Di Martino, and P. Carnier. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. *Journal of Dairy Science*, 94(12):5776–5785, 2011. doi:10.3168/jds.2011-4401.

V. Bonfatti, L. Degano, A. Menegoz, and P. Carnier. Short communication: Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. *Journal of Dairy Science*, 99(10):8216–8221, 2016. doi:10.3168/jds.2016-10953.

V. Bonfatti, A. Fleming, A. Koeck, and F. Miglior. Standardization of milk infrared spectra for the retroactive application of calibration models. *Journal of Dairy Science*, 100(3):2032–2041, 2017a. doi:10.3168/jds.2016-11837.

V. Bonfatti, F. Tiezzi, F. Miglior, and P. Carnier. Comparison of Bayesian regression models and partial least squares regression for the development of infrared prediction equations. *Journal of Dairy Science*, 100(9):7306–7319, 2017b. doi:10.3168/jds.2016-12203.

V. Bonfatti, D. Vicario, L. Degano, A. Lugo, and P. Carnier. Comparison between direct and indirect methods for exploiting Fourier transform spectral information in estimation of breeding values for fine composition and technological properties of milk. *Journal of Dairy Science*, 100(3):2057–2067, 2017c. doi:10.3168/jds.2016-11951.

V. Bonfatti, D. Vicario, A. Lugo, and P. Carnier. 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. *Journal of Dairy Science*, 100(7):5526–5540, 2017d. doi:10.3168/jds.2016-11667.

A. C. Bouwman, H. Bovenhuis, M. H. Visker, and J. A. van Arendonk. Genome-wide association of milk fatty acids in Dutch dairy cattle. *BMC Genetics*, 12(1):43, 2011. doi:10.1186/1471-2156-12-43.

A. C. Bouwman, M. H. Visker, J. A. van Arendonk, and H. Bovenhuis. Fine mapping of a quantitative trait locus for bovine milk fat composition on Bos taurus autosome 19. *Journal of Dairy Science*, 97 (2):1139–1149, 2014. doi:10.3168/jds.2013-7197.

A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, 2008. doi:10.1016/j.cell.2007.12.014.

W. Brand, A. Wells, and M. Coffey. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. *Journal of Dairy Science*, 101(Suppl. 2):347. (Abstr), 2018. https://m.adsa.org/2018/abs/t/74305.

W. Brand, A. T. Wells, S. L. Smith, S. J. Denholm, E. Wall, and M. P. Coffey. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. *Journal of Dairy Science*, 104(4):4980–4990, 2021. doi:10.3168/jds.2020-18367.

T. Bresolin and J. R. Dórea. Infrared spectrometry as a high-throughput phenotyping technology to predict complex traits in livestock systems. *Frontiers in Genetics*, 11:923, 2020. doi:10.3389/fgene.2020.00923.

B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210–223, 2009. doi:10.1016/j.ajhg.2009.01.005.

S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5):1084–1097, 2007. doi:10.1086/521987.

P. Brym, S. Kamiński, and A. Ruść. New SSCP polymorphism within bovine *STAT5A* gene and its associations with milk performance traits in Black-and-White and Jersey cattle. *Journal of Applied Genetics*, 45(4):445–452, 2004. PMID: 15523155.

J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, 109(1):21–29, 2015. doi:10.1002/0471142727.mb2129s109.

B. Buitenhuis, L. L. Janss, N. A. Poulsen, L. B. Larsen, M. K. Larsen, and P. Sørensen. Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. *BMC Genomics*, 15(1):1112, 2014. doi:10.1186/1471-2164-15-1112.

B. Buitenhuis, N. A. Poulsen, G. Gebreyesus, and L. B. Larsen. 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 Genetics*, 17(1):114, 2016. doi:10.1186/s12863-016-0421-2.

S. A. Burgos, N. M. Embertson, Y. Zhao, F. M. Mitloehner, E. J. DePeters, and J. G. Fadel. Prediction of ammonia emission from dairy cattle manure based on milk urea nitrogen: Relation of milk urea nitrogen to ammonia emissions. *Journal of Dairy Science*, 93(6):2377–2386, 2010. doi:10.3168/jds.2009-2415.

D. G. Butler, B. R. Cullis, A. R. Gilmour, B. J. Gogel, and R. Thompson. ASReml-R reference manual: analysis of mixed models for S language environments. Version 3. *The State of Queensland, Department of Primary Industries and Fisheries: Brisbane, Qld*, 2009. https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/ASReml-R-3-Reference-Manual.pdf.

G. Butler, J. H. Nielsen, M. K. Larsen, B. Rehberger, S. Stergiadis, A. Canever, and C. Leifert. The effects of dairy management and processing on quality characteristics of milk and dairy products. *NJAS - Wageningen Journal of Life Sciences*, 58(3):97–102, 2011. doi:10.1016/j.njas.2011.04.002.

A. M. Caroli, S. Chessa, and G. J. Erhardt. Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. *Journal of Dairy Science*, 92(11):5335–5352, 2009. doi:10.3168/jds.2009-2461.

A. Cecchinato, M. De Marchi, L. Gallo, G. Bittante, and P. Carnier. Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. *Journal of Dairy Science*, 92(10):5304–5313, 2009. doi:10.3168/jds.2009-2246.

A. Cecchinato, A. Albera, C. Cipolat-Gotet, A. Ferragina, and G. Bittante. Genetic parameters of cheese yield and curd nutrient recovery or whey loss traits predicted using Fourier-transform infrared spectroscopy of samples collected during milk recording on Holstein, Brown Swiss, and Simmental dairy cows. *Journal of Dairy Science*, 98(7):4914–4927, 2015. doi:10.3168/jds.2014-8599.

M. A. Chester and M. L. Olsson. The ABO blood group gene: A locus of considerable genetic diversity. *Transfusion Medicine Reviews*, 15(3):177–200, 2001. doi:10.1053/tmrv.2001.24591.

P. Cingolani, A. Platts, L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 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*, 6(2):80–92, 2012. doi:10.4161/fly.19695.

S. D. Cochran, J. B. Cole, D. J. Null, and P. J. Hansen. Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. *BMC Genetics*, 14 (1):49, 2013. doi:10.1186/1471-2156-14-49.

M. Cohen-Zinder, E. Seroussi, D. M. Larkin, J. Loor, A. Everts-van der Wind, J.-H. Lee, J. K. Drackley, M. R. Band, A. G. Hernandez, M. Shani, H. A. Lewin, J. I. Weller, and M. Ron. Identification of a missense mutation in the bovine *ABCG2* gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research*, 15(7):936–944, 2005. doi:10.1101/gr.3806705.

L. Commun, K. Velek, J.-B. Barbry, S. Pun, A. Rice, A. Mestek, C. Egli, and S. Leterme. Detection of pregnancy-associated glycoproteins in milk and blood as a test for early pregnancy in dairy cows. *Journal of Veterinary Diagnostic Investigation: Official Publication of the American Association of Veterinary Laboratory Diagnosticians, Inc*, 28(3):207–213, 2016. doi:10.1177/1040638716632815.

G. Conte, M. Mele, S. Chessa, B. Castiglioni, A. Serra, G. Pagnacco, and P. Secchiari. 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. *Journal of Dairy Science*, 93 (2):753–763, 2010. doi:10.3168/jds.2009-2581.

B. Contla Hernández, N. Lopez-Villalobos, and M. Vignes. Identifying health status in grazing dairy cows from milk mid-infrared spectroscopy by using machine learning methods. *Animals*, 11(8):2154, 2021. doi:10.3390/ani11082154.

A. Costa, G. Visentin, M. De Marchi, M. Cassandro, and M. Penasa. Genetic relationships of lactose and freezing point with minerals and coagulation traits predicted from milk mid-infrared spectra in Holstein cows. *Journal of Dairy Science*, 102(8):7217–7225, 2019. doi:10.3168/jds.2018-15378.

S. Couvreur, C. Hurtaud, C. Lopez, L. Delaby, and J. L. Peyraud. The linear relationship between the proportion of fresh grass in the cow diet, milk fatty acid composition, and butter properties. *Journal of Dairy Science*, 89(6):1956–1969, 2006. doi:10.3168/jds.S0022-0302(06)72263-9.

L. K. Creamer, J. E. Plowman, M. J. Liddell, M. H. Smith, and J. P. Hill. Micelle stability: kappa-casein structure and function. *Journal of Dairy Science*, 81(11):3004–3012, 1998. doi:10.3168/jds.S0022-0302(98)75864-3.

R. A. Crocombe. Portable spectroscopy. *Applied Spectroscopy*, 72(12):1701–1751, 2018. doi:10.1177/0003702818809719.

V. A. R. Cruz, H. R. Oliveira, L. F. Brito, A. Fleming, S. Larmer, F. Miglior, and F. S. Schenkel. Genome-wide association study for milk fatty acids in Holstein cattle accounting for the (*DGAT1*) gene effect. *Animals*, 9(11):997, 2019. doi:10.3390/ani9110997.

B. S. Dagnachew, T. H. E. Meuwissen, and T. Ådnøy. Genetic components of milk Fourier-transform infrared spectra used to predict breeding values for milk composition and quality traits in dairy goats. *Journal of Dairy Science*, 96(9):5933–5942, 2013. doi:10.3168/jds.2012-6068.

DairyNZ. Palm Kernel Extract (PKE), 2017. https://www.dairynz.co.nz/feed/supplements/palm-kernel-extract-pke/.

R. Dal Zotto, M. De Marchi, A. Cecchinato, M. Penasa, M. Cassandro, P. Carnier, L. Gallo, and G. Bittante. Reproducibility and repeatability of measures of milk coagulation properties and predictive ability of mid-infrared reflectance spectroscopy. *Journal of Dairy Science*, 91(10):4103–4112, 2008. doi:10.3168/jds.2007-0772.

H. M. Dann, D. M. Barbano, A. Pape, and R. J. Grant. Mid-infrared milk testing for evaluation of health status in dairy cows, 2018. https://ecommons.cornell.edu/handle/1813/59843.

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

M. De Marchi, V. Bonfatti, A. Cecchinato, G. Di Martino, and P. Carnier. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. *Italian Journal of Animal Science*, 8(sup2):399–401, 2009a. doi:10.4081/ijas.2009.s2.399.

M. De Marchi, C. C. Fagan, C. P. O'Donnell, A. Cecchinato, R. Dal Zotto, M. Cassandro, M. Penasa, and G. Bittante. Prediction of coagulation properties, titratable acidity, and pH of bovine milk using mid-infrared spectroscopy. *Journal of Dairy Science*, 92(1):423–432, 2009b. doi:10.3168/jds.2008-1163.

M. De Marchi, M. Penasa, A. Cecchinato, M. Mele, P. Secchiari, and G. Bittante. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. *Animal*, 5 (10):1653–1658, 2011. doi:10.1017/S1751731111000747.

M. De Marchi, V. Toffanin, M. Cassandro, and M. Penasa. Prediction of coagulating and noncoagulating milk samples using mid-infrared spectroscopy. *Journal of Dairy Science*, 96(7):4707–4715, 2013. doi:10.3168/jds.2012-6506.

M. De Marchi, V. Toffanin, M. Cassandro, and M. Penasa. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. *Journal of Dairy Science*, 97(3):1171–1186, 2014. doi:10.3168/jds.2013-6799.

M. De Marchi, M. Penasa, A. Zidi, and C. L. Manuelian. Invited review: Use of infrared technologies for the assessment of dairy products—Applications and perspectives. *Journal of Dairy Science*, 101(12): 10589–10604, 2018. doi:10.3168/jds.2018-15202.

F. Dehareng, C. Delfosse, E. Froidmont, H. Soyeurt, C. Martin, N. Gengler, A. Vanlierde, and P. Dardenne. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. *Animal: An International Journal of Animal Bioscience*, 6(10):1694–1701, 2012. doi:10.1017/S1751731112000456.

P. Delhez, P. N. Ho, N. Gengler, H. Soyeurt, and J. E. Pryce. Diagnosing the pregnancy status of dairy cows: How useful is milk mid-infrared spectroscopy? *Journal of Dairy Science*, 103(4):3264–3274, 2020. doi:10.3168/jds.2019-17473.

M. L. Delignette-Muller and C. Dutang. fitdistrplus: An R package for fitting distributions. *J. Stat. Softw*, 64:1–34, 2015. http://www.jstatsoft.org/v64/i04/.

S. J. Denholm, W. Brand, A. P. Mitchell, A. T. Wells, T. Krzyzelewski, S. L. Smith, E. Wall, and M. P. Coffey. Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning. *Journal of Dairy Science*, 103(10):9355–9367, 2020. doi:10.3168/jds.2020-18328.

T. M. Denninger, A. Schwarm, F. Dohme-Meier, A. Münger, B. Bapst, S. Wegmann, F. Grandl, A. Vanlierde, D. Sorg, S. Ortmann, M. Clauss, and M. Kreuzer. Accuracy of methane emissions predicted from milk mid-infrared spectra and measured by laser methane detectors in Brown Swiss dairy cows. *Journal of Dairy Science*, 103(2):2024–2039, 2020. doi:10.3168/jds.2019-17101.

M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, 2011. doi:10.1038/ng.806.

F. N. Dias. *Supplementation of palm kernel expeller to grazing dairy farms in New Zealand : a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Animal Science at Massey University, Palmerston North, New Zealand.* PhD thesis, Massey University, Palmerston North, New Zealand, 2010. https://mro.massey.ac.nz/handle/10179/2667.

É. Dufour. Chapter 1 - Principles of Infrared Spectroscopy. In D.-W. Sun, editor, *Infrared Spectroscopy for Food Quality Analysis and Control*, pages 1–27. Academic Press, San Diego, 2009. ISBN 978-0-12-374136-3. doi:10.1016/B978-0-12-374136-3.00001-8.

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

J. R. Dórea, G. J. Rosa, K. A. Weld, and L. E. Armentano. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. *Journal of Dairy Science*, 101(7):5878–5889, 2018. doi:10.3168/jds.2017-13997.

C. Egger-Danner, J. B. Cole, J. E. Pryce, N. Gengler, B. Heringstad, A. Bradley, and K. F. Stock. Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. *Animal*, 9(2):191–207, 2015. doi:10.1017/S1751731114002614.

M. El Jabri, 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. Comparison of Bayesian and partial least squares regression methods for mid-infrared prediction of cheese-making properties in Montbéliarde cows. *Journal of Dairy Science*, 102(8):6943–6958, 2019. doi:10.3168/jds.2019-16320.

A. Elgersma. Grazing increases the unsaturated fatty acid concentration of milk from grass-fed cows: A review of the contributing factors, challenges and future perspectives. *European Journal of Lipid Science and Technology*, 117(9):1345–1369, 2015. doi:10.1002/ejlt.201400469.

Y. Etzion, R. Linker, U. Cogan, and I. Shmulevich. Determination of protein concentration in raw milk by mid-infrared Fourier transform infrared/attenuated total reflectance spectroscopy. *Journal of Dairy Science*, 87(9):2779–2788, 2004. doi:10.3168/jds.S0022-0302(04)73405-0.

European Economic Interest Grouping in the service of dairy farmers. n.d. Standardisation of MIR spectrometers across OptiMIR network. https://www.milkrecording.eu/emr.site/index.php/dairystat/.

M. Fang, W. Fu, D. Jiang, Q. Zhang, D. Sun, X. Ding, and J. Liu. A multiple-SNP approach for genome-wide association study of milk production traits in Chinese Holstein cattle. *PloS one*, 9(8): e99544, 2014. doi:10.1371/journal.pone.0099544.

J. D. Ferguson and D. T. Galligan. The value of pregnancy diagnosis - a revisit to an old art. *Clinical Theriogenology*, 3(4):559–578, 2011. https://www.cabdirect.org/cabdirect/abstract/20123038953.

A. Ferragina, G. de los Campos, A. I. Vazquez, A. Cecchinato, and G. Bittante. Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. *Journal of Dairy Science*, 98(11):8133–8151, 2015. doi:10.3168/jds.2014-9143.

T. Fink, T. J. Lopdell, K. Tiplady, R. Handley, T. J. Johnson, R. J. Spelman, S. R. Davis, R. G. Snell, and M. D. Littlejohn. A new mechanism for a familiar mutation – bovine *DGAT1* K232A modulates gene expression through multi-junction exon splice enhancement. *BMC Genomics*, 21(1):591, 2020. doi:10.1186/s12864-020-07004-z.

A. Fleming, F. S. Schenkel, F. Malchiodi, R. A. Ali, B. Mallard, M. Sargolzaei, J. Jamrozik, J. Johnston, and F. Miglior. Genetic correlations of mid-infrared-predicted milk fatty acid groups with milk production traits. *Journal of Dairy Science*, 101(5):4295–4306, 2018. doi:10.3168/jds.2017-14089.

I. Fleming and D. Williams. Infrared and Raman Spectra. In I. Fleming and D. Williams, editors, *Spectroscopic Methods in Organic Chemistry*, pages 85–121. Springer International Publishing, Cham, 2019. ISBN 978-3-030-18252-6.

P. H. Freitas, H. R. Oliveira, F. F. Silva, A. Fleming, F. Miglior, F. S. Schenkel, and L. F. Brito. Genomic analyses for predicted milk fatty acid composition throughout lactation in North American Holstein cattle. *Journal of Dairy Science*, 103(7):6318–6331, 2020. doi:10.3168/jds.2019-17628.

M. Frizzarin, I. C. Gormley, D. P. Berry, T. B. Murphy, A. Casa, A. Lynch, and S. McParland. Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. *Journal of Dairy Science*, 104(7):7438–7447, 2021a. doi:10.3168/jds.2020-19576.

M. Frizzarin, T. F. O'Callaghan, T. B. Murphy, D. Hennessy, and A. Casa. Application of machine-learning methods to milk mid-infrared spectra for discrimination of cow milk from pasture or total mixed ration diets. *Journal of Dairy Science*, 0(0), 2021b. doi:10.3168/jds.2021-20812.

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

N. A. Ganai, H. Bovenhuis, J. A. van Arendonk, and M. H. Visker. Novel polymorphisms in the bovine beta-lactoglobulin gene and their effects on beta-lactoglobulin protein concentration in milk. *Animal Genetics*, 40(2):127–133, 2009. doi:10.1111/j.1365-2052.2008.01806.x.

P. Garidel and H. Schott. Fourier-transform midinfrared spectroscopy for analysis and screening of liquid protein formulations. Part 1: understanding infrared spectroscopy of proteins. *BioProcess International*, 4:40–46, 2006.

S. Garrett and J. J. C. Rosenthal. RNA editing underlies temperature adaptation in K+ channels from polar octopuses. *Science*, 335(6070):848–851, 2012. doi:10.1126/science.1212795.

P. Geladi, D. MacDougall, and H. Martens. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, 39(3):491–500, 1985. doi:10.1366/0003702854248656.

N. Gengler, H. Soyeurt, F. Dehareng, C. Bastin, F. Colinet, H. Hammami, M.-L. Vanrobays, A. Laine, S. Vanderick, C. Grelet, A. Vanlierde, E. Froidmont, and P. Dardenne. Capitalizing on fine milk composition for breeding and management of dairy cows. *Journal of Dairy Science*, 99(5):4071–4079, 2016. doi:10.3168/jds.2015-10140.

J. O. Giordano, P. M. Fricke, and V. E. Cabrera. Economics of resynchronization strategies including chemical tests to identify nonpregnant cows. *Journal of Dairy Science*, 96(2):949–961, 2013. doi:10.3168/jds.2012-5704.

J. A. Green, T. E. Parks, M. P. Avalle, B. P. Telugu, A. L. McLain, A. J. Peterson, W. McMillan, N. Mathialagan, R. R. Hook, S. Xie, and R. M. Roberts. The establishment of an ELISA for the detection of pregnancy-associated glycoproteins (PAGs) in the serum of pregnant cows and heifers. *Theriogenology*, 63(5):1481–1503, 2005. doi:10.1016/j.theriogenology.2004.07.011.

C. Grelet, J. F. Pierna, P. Dardenne, V. Baeten, and F. Dehareng. Standardization of milk mid-infrared spectra from a European dairy network. *Journal of Dairy Science*, 98(4):2150–2160, 2015. doi:10.3168/jds.2014-8764.

C. Grelet, C. Bastin, M. Gelé, J. B. Davière, M. Johan, R. R. A.Werner, J. F. Pierna, F. G. Colinet, P. Dardenne, and N. Gengler. Development of Fourier transform mid-infrared calibrations to predict acetone, β-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. *Journal of Dairy Science*, 99(6):4816–4825, 2016. doi:10.3168/jds.2015-10477.

C. Grelet, J. F. Pierna, P. Dardenne, H. Soyeurt, A. Vanlierde, F. Colinet, C. Bastin, N. Gengler, V. Baeten, and F. Dehareng. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. *Journal of Dairy Science*, 100(10):7910–7921, 2017. doi:10.3168/jds.2017-12720.

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

B. Grisart, 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. 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 Research*, 12(2):222–231, 2002. doi:10.1101/gr.224202.

T. Gu, D. M. Gatti, A. Srivastava, E. M. Snyder, N. Raghupathy, P. Simecek, K. L. Svenson, I. Dotu, J. H. Chuang, M. P. Keller, and A. D. Attie. Genetic architectures of quantitative variation in RNA editing pathways. *Genetics*, 202(2):787–798, 2016. doi:10.1534/genetics.115.179481.

D. Gupta, L. Wang, L. M. Hanssen, J. Hsia, and R. U. Datla. Standard Reference Materials: Polystyrene films for calibrating the wavelength scale of infrared spectrophotometers - SRM 1921. Technical Report NIST SP 260-122, National Institute of Standards and Technology, Gaithersburg, MD, 1995.

X. He, M. X. Chu, L. Qiao, J. N. He, P. Q. Wang, T. Feng, R. Di, G. L. Cao, L. Fang, and Y. F. An. Polymorphisms of *STAT5A* gene and their association with milk production traits in Holstein cows. *Molecular Biology Reports*, 39(3):2901–2907, 2012. doi:10.1007/s11033-011-1051-4.

L. Hein, L. P. Sørensen, M. Kargo, and A. J. Buitenhuis. Genetic analysis of predicted fatty acid profiles of milk from Danish Holstein and Danish Jersey cattle populations. *Journal of Dairy Science*, 101(3): 2148–2157, 2018. doi:10.3168/jds.2017-13225.

K. Hempstalk, S. McParland, and D. P. Berry. Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. *Journal of Dairy Science*, 98(8): 5262–5273, 2015. doi:10.3168/jds.2014-8984.

A. Hewavitharana and B. van Brakel. Fourier transform infrared spectrometric method for the rapid determination of casein in raw milk. *Analyst*, 122(7):701–704, 1997. doi:10.1039/A700953D.

P. N. Ho and J. E. Pryce. Predicting the likelihood of conception to first insemination of dairy cows using milk mid-infrared spectroscopy. *Journal of Dairy Science*, 103(12):11535–11544, 2020. doi:10.3168/jds.2020-18589.

P. N. Ho, V. Bonfatti, T. D. W. Luke, and J. E. Pryce. Classifying the fertility of dairy cows using milk mid-infrared spectroscopy. *Journal of Dairy Science*, 102(11):10460–10470, 2019. doi:10.3168/jds.2019-16412.

P. N. Ho, T. D. W. Luke, and J. E. Pryce. Validation of milk mid-infrared spectroscopy for predicting the metabolic status of lactating dairy cows in Australia. *Journal of Dairy Science*, 104(4):4467–4477, 2021. doi:10.3168/jds.2020-19603.

A. N. Hristov, E. Kebreab, M. Niu, J. Oh, A. Bannink, A. R. Bayat, T. B. Boland, A. F. Brito, D. P. Casper, L. A. Crompton, J. Dijkstra, M. Eugène, P. C. Garnsworthy, N. Haque, A. L. Hellwing, P. Huhtanen, M. Kreuzer, B. Kuhla, P. Lund, J. Madsen, C. Martin, P. J. Moate, S. Muetzel, C. Muñoz, N. Peiren, J. M. Powell, C. K. Reynolds, A. Schwarm, K. J. Shingfield, T. M. Storlien, M. R. Weisbjerg, D. R. Yáñez-Ruiz, and Z. Yu. Symposium review: Uncertainties in enteric methane inventories, measurement techniques, and prediction models. *Journal of Dairy Science*, 101(7):6655–6674, 2018. doi:10.3168/jds.2017-13536.

G. Huang, D. Buckler-Pena, T. Nauta, M. Singh, A. Asmar, J. Shi, J. Y. Kim, and K. V. Kandror. Insulin responsiveness of glucose transporter 4 in 3T3-L1 cells depends on the presence of sortilin. *Molecular Biology of the Cell*, 24(19):3115–3122, 2013. doi:10.1091/mbc.e12-10-0765.

G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. https://arxiv.org/abs/1608.06993.

ICAR (International Committee for Animal Recording). ICAR Guidelines. Section 12 – Guidelines for Milk Analysis. 2017. www.icar.org/Guidelines/12-Milk-Analysis.pdf.

ISO (International Organisation for Standardization). Milk and milk products — Determination of lactose content by high-performance liquid chromatography (Reference method). *Standard number 22662:2007*, ISO, Geneva, Switzerland, 2007. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/63/36384.html.

ISO (International Organisation for Standardization). Milk — Determination of fat content — Gravimetric method (Reference method). *Standard number 1211:2010*, ISO, Geneva, Switzerland, 2010. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/13/51348.html.

ISO (International Organisation for Standardization). Milk and liquid milk products — Guidelines for the application of mid-infrared spectrometry. *Standard number 9622:2013*, ISO, Geneva, Switzerland, 2013. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/68/56874.html.

ISO (International Organisation for Standardization). Milk and milk products — Determination of nitrogen content — Part 4: Determination of protein and non-protein nitrogen content and true protein content calculation (Reference method). *Standard number 8968-4:2016*, ISO, Geneva, Switzerland, 2016. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/03/60386.html.

L. H. Iung, 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. Genome-wide association study for milk production traits in a Brazilian Holstein population. *Journal of Dairy Science*, 102(6):5305–5314, 2019. doi:10.3168/jds.2018-14811.

J. Jiang, J. B. Cole, Y. Da, P. M. VanRaden, and L. Ma. Fast Bayesian fine-mapping of 35 production, reproduction and body conformation traits with imputed sequences of 27K Holstein bulls. *bioRxiv*, page 428227, 2018. doi:10.1101/428227.

J. Jiang, L. Liu, Y. Gao, L. Shi, Y. Li, W. Liang, and D. Sun. Determination of genetic associations between indels in 11 candidate genes and milk composition traits in Chinese Holstein population. *BMC genetics*, 20(1):48, 2019a. doi:10.1186/s12863-019-0751-y.

J. Jiang, L. Ma, D. Prakapenka, P. M. VanRaden, J. B. Cole, and Y. Da. A large-scale genome-wide association study in US Holstein cattle. *Frontiers in Genetics*, 10:412, 2019b. doi:10.3389/fgene.2019.00412.

L. Jiang, J. Liu, D. Sun, P. Ma, X. Ding, Y. Yu, and Q. Zhang. Genome wide association studies for milk production traits in Chinese Holstein population. *PloS one*, 5(10), 2010. doi:10.1371/journal.pone.0013661.

L. Jiang, Z. Zheng, T. Qi, K. E. Kemper, N. R. Wray, P. M. Visscher, and J. Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, 51(12):1749–1755, 2019c. doi:10.1038/s41588-019-0530-8.

S. Jivanji, G. Worth, T. Lopdell, A. Yeates, C. Couldrey, E. Reynolds, K. Tiplady, L. McNaughton, T. J. Johnson, S. R. Davis, B. Harris, R. Spelman, R. G. Snell, D. Garrick, and M. D. Littlejohn. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. *Genetics Selection Evolution*, 51(1):62, 2019. doi:10.1186/s12711-019-0506-2.

J. S. Jonker, R. A. Kohn, and R. A. Erdman. Using milk urea nitrogen to predict nitrogen excretion and utilization efficiency in lactating dairy cows. *Journal of Dairy Science*, 81(10):2681–2692, 1998. doi:10.3168/jds.S0022-0302(98)75825-4.

A. M. Joubran, K. M. Pierce, N. Garvey, L. Shalloo, and T. F. O'Callaghan. Invited review: A 2020 perspective on pasture-based dairy systems and products. *Journal of Dairy Science*, 104(7):7364–7382, 2021. doi:10.3168/jds.2020-19776.

P. B. Kandel, M. L. Vanrobays, A. Vanlierde, F. Dehareng, E. Froidmont, N. Gengler, and H. Soyeurt. Genetic parameters of mid-infrared methane predictions and their relationships with milk production traits in Holstein cattle. *Journal of Dairy Science*, 100(7):5578–5591, 2017. doi:10.3168/jds.2016-11954.

R. Karoui, G. Mazerolles, and É. Dufour. Spectroscopic techniques coupled with chemometric tools for structure and texture determinations in dairy products. *International Dairy Journal*, 13(8):607–620, 2003. doi:10.1016/S0958-6946(03)00076-1.

A. J. Kauffman and N. R. St-Pierre. The relationship of milk urea nitrogen to urine nitrogen excretion in Holstein and Jersey cows. *Journal of Dairy Science*, 84(10):2284–2294, 2001. doi:10.3168/jds.S0022-0302(01)74675-9.

K. E. Kaylegian, G. E. Houghton, J. M. Lynch, J. R. Fleming, and D. M. Barbano. Calibration of infrared milk analyzers: Modified milk versus producer milk. *Journal of Dairy Science*, 89(8):2817–2832, 2006. doi:10.3168/jds.S0022-0302(06)72555-3.

K. E. Kemper, B. J. Hayes, H. D. Daetwyler, and M. E. Goddard. How old are quantitative trait loci and how widely do they segregate? *Journal of Animal Breeding and Genetics*, 132(2):121–134, 2015a. doi:10.1111/jbg.12152.

K. E. Kemper, C. M. Reich, P. J. Bowman, C. J. vander Jagt, A. J. Chamberlain, B. A. Mason, B. J. Hayes, and M. E. Goddard. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genetics Selection Evolution*, 47(1):29, 2015b. doi:10.1186/s12711-014-0074-4.

K. E. Kemper, M. D. Littlejohn, T. Lopdell, B. J. Hayes, L. E. Bennett, R. P. Williams, X. Q. Xu, P. M. Visscher, M. J. Carrick, and M. E. Goddard. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. *BMC Genomics*, 17(1):858, 2016. doi:10.1186/s12864-016-3175-3.

N. Kermarrec, F. Roubinet, P.-A. Apoil, and A. Blancher. Comparison of allele O sequences of the human and non-human primate ABO system. *Immunogenetics*, 49(6):517–526, 1999. doi:10.1007/s002510050529.

P. Kgwatalala, E. Ibeagha-Awemu, J. Hayes, and X. Zhao. Stearoyl-CoA desaturase 1 3'UTR SNPs and their influence on milk fatty acid composition of Canadian Holstein cows. *Journal of Animal Breeding and Genetics*, 126(5):394–403, 2009. doi:10.1111/j.1439-0388.2008.00796.x.

P. Khanal and R. J. Tempelman. The use of milk Fourier-transform mid-infrared spectroscopy to diagnose pregnancy and determine spectral regional associations with pregnancy in US dairy cows. *Journal of Dairy Science*, 105(4):3209–3221, 2022. doi:10.3168/jds.2021-21079.

D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013. doi:10.1186/gb-2013-14-4-r36.

J. R. Knapp, G. L. Laur, P. A. Vadas, W. P. Weiss, and J. M. Tricarico. Invited review: Enteric methane in dairy cattle production: Quantifying the opportunities and impact of reducing emissions. *Journal of Dairy Science*, 97(6):3231–3261, 2014. doi:10.3168/jds.2013-7234.

T. M. Knutsen, H. G. Olsen, V. Tafintseva, M. Svendsen, A. Kohler, M. P. Kent, and S. Lien. Unravelling genetic variation underlying de novo-synthesis of bovine milk fatty acids. *Scientific Reports*, 8(1):2179, 2018. doi:10.1038/s41598-018-20476-0.

S. Kucheryavskiy. mdatools — R package for chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 198, 2020. doi:10.1016/j.chemolab.2020.103937.

M. Kuhn. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles*, 28(5):1–26, 2008. doi:10.18637/jss.v028.i05.

B. Lahart, S. McParland, E. Kennedy, T. M. Boland, T. Condon, M. Williams, N. Galvin, B. McCarthy, and F. Buckley. Predicting the dry matter intake of grazing dairy cows using infrared reflectance spectroscopy analysis. *Journal of Dairy Science*, 102(10):8907–8918, 2019. doi:10.3168/jds.2019-16363.

A. Lainé, C. Bastin, C. Grelet, H. Hammami, F. G. Colinet, L. M. Dale, A. Gillon, J. Vandenplas, F. Dehareng, and N. Gengler. Assessing the effect of pregnancy stage on milk composition of dairy cows using mid-infrared spectra. *Journal of Dairy Science*, 100(4):2863–2876, 2017. doi:10.3168/jds.2016-11736.

J. Le Pendu. Histo-blood group antigen and human milk oligosaccharides. In L. K. Pickering, A. L. Morrow, G. M. Ruiz-Palacios, and R. J. Schanler, editors, *Protecting Infants through Human Milk*, Advances in Experimental Medicine and Biology, pages 135–143, Boston, MA, 2004. Springer US. ISBN 978-1-4757-4242-8. doi:10.1007/978-1-4757-4242-8_13.

D. Lee, D. U. Gorkin, M. Baker, B. J. Strober, A. L. Asoni, A. S. McCallion, and M. A. Beer. A method to predict the impact of regulatory variants from DNA sequence. *Nature genetics*, 47(8):955–961, 2015. doi:10.1038/ng.3331.

D. Lefier, R. Grappin, and S. Pochet. Determination of fat, protein, and lactose in raw milk by Fourier transform infrared spectroscopy and by analysis with a conventional filter-based milk analyzer. *Journal of AOAC International*, 79(3):711–717, 1996. doi:10.1093/jaoac/79.3.711.

C. Li, D. Sun, S. Zhang, S. Wang, X. Wu, Q. Zhang, L. Liu, Y. Li, and L. Qiao. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. *PloS one*, 9(5):e96186, 2014. doi:10.1371/journal.pone.0096186.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. doi:10.1093/bioinformatics/btp324.

LIC (Livestock Improvement Corporation) and DairyNZ. *New Zealand Dairy Statistics 2016–17*. LIC/DairyNZ, Hamilton, New Zealand, 2017. https://www.dairynz.co.nz/publications/dairy-industry/.

B. J. Liddicoat, R. Piskol, A. M. Chalk, G. Ramaswami, M. Higuchi, J. C. Hartner, J. B. Li, P. H. Seeburg, and C. R. Walkley. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science*, 349(6252):1115–1120, 2015. doi:10.1126/science.aac7049.

M. D. Littlejohn, 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. Expression variants of the lipogenic *AGPAT6* gene affect diverse milk composition phenotypes in Bos taurus. *PloS one*, 9(1):e85757, 2014. doi:10.1371/journal.pone.0085757.

M. D. Littlejohn, K. Tiplady, T. A. Fink, K. Lehnert, T. Lopdell, T. Johnson, C. Couldrey, M. Keehan, R. G. Sherlock, C. Harland, A. Scott, R. G. Snell, S. R. Davis, and R. J. Spelman. Sequence-based Association Analysis Reveals an *MGST1* eQTL with Pleiotropic Effects on Bovine Milk Composition. *Scientific Reports*, 6(1):25376, 2016. doi:10.1038/srep25376.

R. Liu, D.-x. Sun, Y.-c. Wang, Y. Yu, Y. Zhang, H.-y. Chen, Q. Zhang, S.-l. Zhang, and Y. Zhang. Fine mapping QTLs affecting milk production traits on BTA6 in Chinese Holstein with SNP markers. *Journal of Integrative Agriculture*, 12(1):110–117, 2013. doi:10.1016/S2095-3119(13)60211-7.

Z. Liu, T. Wang, J. E. Pryce, I. M. MacLeod, B. J. Hayes, A. J. Chamberlain, C. V. Jagt, C. M. Reich, B. A. Mason, S. Rochfort, and B. G. Cocks. Fine-mapping sequence mutations with a major effect on oligosaccharide content in bovine milk. *Scientific Reports*, 9(1):2137, 2019. doi:10.1038/s41598-019-38488-9.

P.-R. Loh, 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. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, 2015. doi:10.1038/ng.3190.

P.-R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, and A. L. Price. Mixed-model association for biobank-scale datasets. *Nature Genetics*, 50(7):906–908, 2018. doi:10.1038/s41588-018-0144-6.

S. Loker, F. Miglior, J. Bohmanova, J. Jamrozik, and L. R. Schaeffer. Phenotypic analysis of pregnancy effect on milk, fat, and protein yields of Canadian Ayrshire, Jersey, Brown Swiss, and Guernsey breeds. *Journal of Dairy Science*, 92(3):1300–1312, 2009. doi:10.3168/jds.2008-1425.

T. J. Lopdell, K. Tiplady, M. Struchalin, T. J. Johnson, M. Keehan, R. Sherlock, C. Couldrey, S. R. Davis, R. G. Snell, R. J. Spelman, and M. D. Littlejohn. DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. *BMC Genomics*, 18 (1):968, 2017. doi:10.1186/s12864-017-4320-3.

T. J. Lopdell, K. Tiplady, M. Struchalin, T. J. Johnson, M. Keehan, R. Sherlock, C. Couldrey, S. R. Davis, R. G. Snell, R. J. Spelman, and M. D. Littlejohn. Data from: DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. In *Data from: DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content*. Dryad Digital Repository, 2018. http://datadryad.org/stash/dataset/doi:10.5061/dryad.vv469.

T. J. Lopdell, V. Hawkins, C. Couldrey, K. Tiplady, S. R. Davis, B. L. Harris, R. G. Snell, and M. D. Littlejohn. Widespread cis-regulation of RNA editing in a large mammal. *RNA*, 25(3):319–335, 2019a. doi:10.1261/rna.066902.118.

T. J. Lopdell, K. Tiplady, C. Couldrey, T. J. Johnson, M. Keehan, S. R. Davis, B. L. Harris, R. J. Spelman, R. G. Snell, and M. D. Littlejohn. Multiple QTL underlie milk phenotypes at the CSF2RB locus. *Genetics Selection Evolution*, 51(1):3, 2019b. doi:10.1186/s12711-019-0446-x.

N. Lopez-Villalobos. Analysing the genetic basis of milk production traits. *CAB Reviews*, 7(028):1–18, 2012. https://www.cabdirect.org/cabdirect/abstract/20123199701.

N. Lopez-Villalobos, S. R. Davis, E. M. Beattie, J. Melis, S. Berry, S. E. Holroyd, R. J. Spelman, and R. G. Snell. Breed effects for lactoferrin concentration determined by Fourier transform infrared spectroscopy. *Proceedings of the New Zealand Society of Animal Production*, 69:60–64, 2009. https://www.nzsap.org/system/files/proceedings/2009/ab09015.pdf.

N. Lopez-Villalobos, R. J. Spelman, J. Melis, S. R. Davis, S. D. Berry, K. Lehnert, S. E. Holroyd, A. K. MacGibbon, and R. G. Snell. Estimation of genetic and crossbreeding parameters of fatty acid concentrations in milk fat predicted by mid-infrared spectroscopy in New Zealand dairy cattle. *Journal of Dairy Research*, 81(3):340–349, 2014. doi:10.1017/S0022029914000272.

W. Luginbühl. Evaluation of designed calibration samples for casein calibration in Fourier transform infrared analysis of milk. *LWT - Food Science and Technology*, 35(6):554–558, 2002. doi:10.1006/fstl.2002.0902.

H. J. Luinge, E. Hop, E. T. Lutz, J. A. van Hemert, and E. A. de Jong. Determination of the fat, protein and lactose content of milk using Fourier transform infrared spectrometry. *Analytica Chimica Acta*, 284(2):419–433, 1993. doi:10.1016/0003-2670(93)85328-H.

T. D. W. Luke, T. T. T. Nguyen, S. Rochfort, W. J. Wales, C. M. Richardson, M. Abdelsayed, and J. E. Pryce. Genomic prediction of serum biomarkers of health in early lactation. *Journal of Dairy Science*, 102(12):11142–11152, 2019a. doi:10.3168/jds.2019-17127.

T. D. W. Luke, S. Rochfort, W. J. Wales, V. Bonfatti, L. Marett, and J. E. Pryce. Metabolic profiling of early-lactation dairy cows using milk mid-infrared spectra. *Journal of Dairy Science*, 102(2):1747–1760, 2019b. doi:10.3168/jds.2018-15103.

L. S. Lum, P. Dovč, and J. F. Medrano. Polymorphisms of Bovine β-lactoglobulin promoter and differences in the binding affinity of activator protein-2 transcription factor. *Journal of Dairy Science*, 80(7): 1389–1397, 1997. doi:10.3168/jds.S0022-0302(97)76068-5.

J. M. Lynch, D. M. Barbano, M. Schweisthal, and J. R. Fleming. Precalibration evaluation procedures for mid-infrared milk analyzers. *Journal of Dairy Science*, 89(7):2761–2774, 2006. doi:10.3168/jds.S0022-0302(06)72353-0.

M. Lynch and B. Walsh. *Genetics and analysis of quantitative traits*. 1998. Publisher: Sinauer Associates, Sunderland, MA. ISBN: 9780878934812.

D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M.-M. Suner, T. Hunt, I. H. A. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, . G. P. Consortium, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 2012. doi:10.1126/science.1215040.

A. MacGibbon and M. Reynolds. Milk Lipids | Analytical Methods. In *Encyclopedia of Dairy Sciences*, pages 698–703. 2011. ISBN 978-0-12-374407-4. doi:10.1016/B978-0-12-374407-4.00339-3.

I. M. MacLeod, 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. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC genomics*, 17(1):144, 2016. doi:10.1186/s12864-016-2443-6.

Y. Mao, X. Zhu, S. Xing, M. Zhang, H. Zhang, X. Wang, N. Karrow, L. Yang, and Z. Yang. Polymorphisms in the promoter region of the bovine lactoferrin gene influence milk somatic cell score and milk production traits in Chinese Holstein cows. *Research in Veterinary Science*, 103:107–112, 2015. doi:10.1016/j.rvsc.2015.09.021.

H. Martens, J. P. Nielsen, and S. B. Engelsen. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Analytical Chemistry*, 75(3):394–404, 2003. doi:10.1021/ac020194w.

M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, and A. Shafer. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, 2012. doi:10.1126/science.1222794.

M. H. Maurice-Van Eijndhoven, H. Soyeurt, F. Dehareng, and M. P. Calus. Validation of fatty acid predictions in milk using mid-infrared spectrometry across cattle breeds. *Animal*, 7(2):348–354, 2013. doi:10.1017/S1751731112001218.

S. McAuliffe, T. J. Gilliland, and D. Hennessy. Comparison of pasture-based feeding systems and a total mixed ration feeding system on dairy cow milk production. *Sustainable meat and milk production from grasslands*, page 289, 2016. ISBN : 9788217016779.

A. McDermott, G. Visentin, M. De Marchi, D. P. Berry, M. A. Fenelon, P. M. O'Connor, O. A. Kenny, and S. McParland. Prediction of individual milk proteins including free amino acids in bovine milk using mid-infrared spectroscopy and their correlations with milk processing characteristics. *Journal of Dairy Science*, 99(4):3171–3182, 2016. doi:10.3168/jds.2015-9747.

W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17(1):122, 2016. doi:10.1186/s13059-016-0974-4.

S. McParland and D. P. Berry. The potential of Fourier transform infrared spectroscopy of milk samples to predict energy intake and efficiency in dairy cows. *Journal of Dairy Science*, 99(5):4056–4070, 2016. doi:10.3168/jds.2015-10051.

S. McParland, G. Banos, E. Wall, M. P. Coffey, H. Soyeurt, R. F. Veerkamp, and D. P. Berry. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. *Journal of Dairy Science*, 94(7):3651–3661, 2011. doi:10.3168/jds.2010-3965.

S. McParland, G. Banos, B. McCarthy, E. Lewis, M. P. Coffey, B. O'Neill, M. O'Donovan, E. Wall, and D. P. Berry. Validation of mid-infrared spectrometry in milk for predicting body energy status in Holstein-Friesian cows. *Journal of Dairy Science*, 95(12):7225–7235, 2012. doi:10.3168/jds.2012-5406.

S. McParland, E. Kennedy, E. Lewis, S. G. Moore, B. McCarthy, M. O'Donovan, and D. P. Berry. Genetic parameters of dairy cow energy intake and body energy status predicted using mid-infrared spectrometry of milk. *Journal of Dairy Science*, 98(2):1310–1320, 2015. doi:10.3168/jds.2014-8892.

J. Medrano, G. Rincon, and A. Islas-Trejo. Comparative analysis of bovine milk and mammary gland transcriptome using RNA-Seq. *9th World Congress on Genetics applied to Livestock Production; Leipzig, Germany*, 852, 2010.

T. Mehtiö, P. Mäntysaari, T. Kokkonen, S. Kajava, T. Latomäki, L. Nyholm, C. Grelet, T. Pitkänen, E. Mäntysaari, and M. Lidauer. Developing an indicator for body fat mobilisation using mid-infrared spectrometry of milk samples in dairy cows. *Proceedings of the World Congress on Genetics Applied to Livestock Production*, Electronic Poster Session - Biology - Feed Intake and Efficiency 1:225, 2018. https://orbi.uliege.be/handle/2268/224010.

M. Mele, G. Conte, B. Castiglioni, S. Chessa, N. P. P. Macciotta, A. Serra, A. Buccioni, G. Pagnacco, and P. Secchiari. Stearoyl-Coenzyme A Desaturase Gene Polymorphism and Milk Fatty Acid Composition in Italian Holsteins. *Journal of Dairy Science*, 90(9):4458–4465, 2007. doi:10.3168/jds.2006-617.

K. M. Mendez, S. N. Reinke, and D. I. Broadhurst. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*, 15(12):150, 2019. doi:10.1007/s11306-019-1612-4.

B.-H. Mevik and R. Wehrens. The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software*, 18(2):1 – 23, 2007. doi:10.18637/jss.v018.i02.

F. Miglior, A. Sewalem, J. Jamrozik, J. Bohmanova, D. M. Lefebvre, and R. K. Moore. Genetic analysis of milk urea nitrogen and lactose and their relationships with other production traits in Canadian Holstein cattle. *Journal of Dairy Science*, 90(5):2468–2479, 2007. doi:10.3168/jds.2006-487.

F. Miglior, A. Fleming, F. Malchiodi, L. F. Brito, P. Martin, and C. F. Baes. A 100-Year Review: Identification and genetic selection of economically important traits in dairy cattle. *Journal of Dairy Science*, 100(12):10251–10271, 2017. doi:10.3168/jds.2017-12968.

Ministry for Primary Industries, NZ. *Feed use in the NZ dairy industry*. Ministry for Primary Industries (MPI) Technical Paper 2017/53, New Zealand, 2017. https://www.mpi.govt.nz/dmsdocument/20897/direct.

Ministry for the Environment, NZ. *New Zealand's Greenhouse Gas Inventory 1990–2020*. Ministry for the Environment, New Zealand, 2022. https://environment.govt.nz/publications/new-zealands-greenhouse-gas-inventory-1990-2020/.

R. G. Mitchell, G. W. Rogers, C. D. Dechow, J. E. Vallimont, J. B. Cooper, U. Sander-Nielsen, and J. S. Clay. Milk urea nitrogen concentration: heritability and genetic correlations with reproductive performance and disease. *Journal of Dairy Science*, 88(12):4434–4440, 2005. doi:10.3168/jds.S0022-0302(05)73130-1.

B. Moioli, G. Contarini, A. Avalli, G. Catillo, L. Orrù, G. De Matteis, G. Masoero, and F. Napolitano. Short Communication: Effect of Stearoyl-Coenzyme A Desaturase Polymorphism on Fatty Acid Composition of Milk. *Journal of Dairy Science*, 90(7):3553–3558, 2007. doi:10.3168/jds.2006-855.

L. F. M. Mota, S. Pegolo, T. Baba, F. Peñagaricano, G. Morota, G. Bittante, and A. Cecchinato. Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data. *Journal of Dairy Science*, 104(7):8107–8121, 2021. doi:10.3168/jds.2020-19861.

S. G. Narayana, F. S. Schenkel, A. Fleming, A. Koeck, F. Malchiodi, J. Jamrozik, J. Johnston, M. Sargolzaei, and F. Miglior. Genetic analysis of groups of mid-infrared predicted fatty acids in milk. *Journal of Dairy Science*, 100(6):4731–4744, 2017. doi:10.3168/jds.2016-12244.

E. Negussie, Y. de Haas, F. Dehareng, R. J. Dewhurst, J. Dijkstra, N. Gengler, D. P. Morgavi, H. Soyeurt, S. van Gastelen, T. Yan, and F. Biscarini. Invited review: Large-scale indirect measurements for enteric methane emissions in dairy cattle: A review of proxies and their potential for use in management and breeding decisions. *Journal of Dairy Science*, 100(4):2433–2453, 2017. doi:10.3168/jds.2016-12030.

J. Nousiainen, K. J. Shingfield, and P. Huhtanen. Evaluation of milk urea nitrogen as a diagnostic of protein feeding. *Journal of Dairy Science*, 87(2):386–398, 2004. doi:10.3168/jds.S0022-0302(04)73178-1.

J. Ogorevc, T. Kunej, A. Razpet, and P. Dovc. Database of cattle candidate genes and genetic markers for milk production and mastitis. *Animal Genetics*, 40(6):832–851, 2009. doi:10.1111/j.1365-2052.2009.01921.x.

M. C. Oliveira, N. M. Silva, L. P. Bastos, L. M. Fonseca, M. M. Cerqueira, M. O. Leite, and R. S. Conrrado. Fourier transform infrared spectroscopy (FTIR) for MUN analysis in normal and adulterated milk. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, 64(5):1360–1366, 2012. doi:10.1590/S0102-09352012000500037.

R. Oliveira, M. Faria, R. Silva, L. Bezerra, G. Carvalho, A. Pinheiro, J. Simionato, and A. Leão. Fatty acid profile of milk and cheese from dairy cows supplemented a diet with palm kernel Cake. *Molecules*, 20(8):15434–15448, 2015. doi:10.3390/molecules200815434.

V. Olori, S. Brotherstone, W. Hill, and B. McGuirk. Effect of gestation stage on milk yield and composition in Holstein Friesian dairy cattle. *Livestock Production Science*, 52(2):167–176, 1997. doi:10.1016/S0301-6226(97)00126-7.

H. G. Olsen, 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. 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. *Genetics, Selection, Evolution : GSE*, 49, 2017. doi:10.1186/s12711-017-0294-5.

L. P. O'Neill and B. M. Turner. Immunoprecipitation of native chromatin: NChIP. *Methods*, 31(1):76–82, 2003. doi:10.1016/S1046-2023(03)00090-2.

T. F. O'Callaghan, D. Hennessy, S. McAuliffe, K. N. Kilcawley, M. O'Donovan, P. Dillon, R. P. Ross, and C. Stanton. Effect of pasture versus indoor feeding systems on raw milk composition and quality over an entire lactation. *Journal of Dairy Science*, 99(12):9424–9440, 2016. doi:10.3168/jds.2016-10985.

K. P. Palmano and D. F. Elgar. Detection and quantitation of lactoferrin in bovine whey samples by reversed-phase high-performance liquid chromatography on polystyrene–divinylbenzene. *Journal of Chromatography A*, 947(2):307–311, 2002. doi:10.1016/S0021-9673(01)01563-1.

V. Palombo, M. Milanesi, S. Sgorlon, S. Capomaccio, M. Mele, E. Nicolazzi, P. Ajmone-Marsan, F. Pilla, B. Stefanon, and M. D'Andrea. Genome-wide association study of milk fatty acid composition in Italian Simmental and Italian Holstein cows using single nucleotide polymorphism arrays. *Journal of Dairy Science*, 101(12):11004–11019, 2018. doi:10.3168/jds.2018-14413.

E. Park, J. Guo, L. Lin, L. Demirdjian, S. Shen, Y. Xing, and Y. N. Wu. Population and allelic variation of A-to-I RNA editing in human transcriptomes. *Genome Biology*, 18(1):143, 2017. doi:10.1186/s13059-017-1270-7.

C. Parsons and H. Lyder. Determining a size of cell of a transmission spectroscopy device. United States patent US 9,829,378. 2017 Nov 28, 2018.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

C. M. Paton and J. M. Ntambi. Biochemical and physiological function of stearoyl-CoA desaturase. *American Journal of Physiology-Endocrinology and Metabolism*, 297(1):E28–E37, 2009. doi:10.1152/ajpendo.90897.2008.

H. Pausch, R. Emmerling, H. Schwarzenbacher, and R. Fries. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. *Genetics Selection Evolution*, 48(1):14, 2016. doi:10.1186/s12711-016-0190-4.

H. Pausch, R. Emmerling, B. Gredler-Grandl, R. Fries, H. D. Daetwyler, and M. E. Goddard. Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentages in milk at nucleotide resolution. *BMC Genomics*, 18(1):1–11, 2017. doi:10.1186/s12864-017-4263-8.

A. Pawlik, G. Sender, M. Sobczyńska, A. Korwin-Kossakowska, H. Lassa, J. Oprządek, A. Pawlik, G. Sender, M. Sobczyńska, A. Korwin-Kossakowska, H. Lassa, and J. Oprządek. Lactoferrin gene variants, their expression in the udder and mastitis susceptibility in dairy cattle. *Animal Production Science*, 55(8):999–1004, 2014. doi:10.1071/AN13389.

S. Pegolo, A. Cecchinato, M. Mele, G. Conte, S. Schiavon, and G. Bittante. Effects of candidate gene polymorphisms on the detailed fatty acids profile determined by gas chromatography in bovine milk. *Journal of Dairy Science*, 99(6):4558–4573, 2016. doi:10.3168/jds.2015-10420.

S. Pegolo, N. Mach, Y. Ramayo-Caldas, S. Schiavon, G. Bittante, and A. Cecchinato. Integration of GWAS, pathway and network analyses reveals novel mechanistic insights into the synthesis of milk proteins in dairy cows. *Scientific Reports*, 8(1):566, 2018. doi:10.1038/s41598-017-18916-4.

M. Penasa, M. De Marchi, and M. Cassandro. Short communication: Effects of pregnancy on milk yield, composition traits, and coagulation properties of Holstein cows. *Journal of Dairy Science*, 99(6): 4864–4869, 2016. doi:10.3168/jds.2015-10168.

D. Picque, D. Lefier, R. Grappin, and G. Corrieu. Monitoring of fermentation by infrared spectrometry: Alcoholic and lactic fermentations. *Analytica Chimica Acta*, 279(1):67–72, 1993. doi:10.1016/0003-2670(93)85067-T.

N. A. Poulsen, H. P. Bertelsen, H. B. Jensen, F. Gustavsson, M. Glantz, H. L. Månsson, A. Andrén, M. Paulsson, C. Bendixen, A. J. Buitenhuis, and L. B. Larsen. The occurrence of noncoagulating milk and the association of bovine milk coagulation properties with genetic variants of the caseins in 3 Scandinavian dairy breeds. *Journal of Dairy Science*, 96(8):4830–4842, 2013. doi:10.3168/jds.2012-6422.

N. A. Poulsen, A. J. Buitenhuis, and L. B. Larsen. Phenotypic and genetic associations of milk traits with milk coagulation properties. *Journal of Dairy Science*, 98(4):2079–2087, 2015. doi:10.3168/jds.2014-7944.

N. A. Poulsen, R. C. Robinson, D. Barile, L. B. Larsen, and B. Buitenhuis. A genome-wide association study reveals specific transferases as candidate loci for bovine milk oligosaccharides synthesis. *BMC Genomics*, 20(1):404, 2019. doi:10.1186/s12864-019-5786-y.

J. M. Powell, M. A. Wattiaux, and G. A. Broderick. Short communication: Evaluation of milk urea nitrogen as a management tool to reduce ammonia emissions from dairy farms. *Journal of Dairy Science*, 94(9):4690–4694, 2011. doi:10.3168/jds.2011-4476.

R. S. Pralle, K. W. Weigel, and H. M. White. Predicting blood β-hydroxybutyrate using milk Fourier transform infrared spectrum, milk composition, and producer-reported variables with multiple linear regression, partial least squares regression, and artificial neural network. *Journal of Dairy Science*, 101 (5):4378–4387, 2018. doi:10.3168/jds.2017-14076.

D. Pretto, M. De Marchi, M. Penasa, and M. Cassandro. Effect of milk composition and coagulation traits on Grana Padano cheese yield under field conditions. *The Journal of Dairy Research*, 80(1):1–5, 2013. doi:10.1017/S0022029912000453.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, and others. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007. doi:10.1086/519795.

R Core Team. R: A Language and Environment for Statistical Computing. In *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. https://www.R-project.org/.

G. Ramaswami, P. Deng, R. Zhang, M. A. Carbone, T. F. C. Mackay, and J. B. Li. Genetic mapping uncovers cis-regulatory landscape of RNA editing. *Nature communications*, 6(1):1–9, 2015. doi:10.1038/ncomms9194.

L.-A. Raven, B. G. Cocks, and B. J. Hayes. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC genomics*, 15(1):62, 2014. doi:10.1186/1471-2164-15-62.

L.-A. Raven, B. G. Cocks, K. E. Kemper, A. J. Chamberlain, C. J. vander Jagt, M. E. Goddard, and B. J. Hayes. Targeted imputation of sequence variants and gene expression profiling identifies twelve candidate genes associated with lactation volume, composition and calving interval in dairy cattle. *Mammalian Genome*, 27(1-2):81–97, 2016. doi:10.1007/s00335-015-9613-8.

D. L. Renaud, D. F. Kelton, and T. F. Duffield. Short communication: Validation of a test-day milk test for β-hydroxybutyrate for identifying cows with hyperketonemia. *Journal of Dairy Science*, 102(2): 1589–1593, 2019. doi:10.3168/jds.2018-14778.

E. G. Reynolds, C. Neeley, T. J. Lopdell, M. Keehan, K. Dittmer, C. S. Harland, C. Couldrey, T. 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. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. *Nature Genetics*, 53(7):949–954, 2021. doi:10.1038/s41588-021-00872-5.

A. Ricci, P. D. Carvalho, M. C. Amundson, R. H. Fourdraine, L. Vincenti, and P. M. Fricke. Factors associated with pregnancy-associated glycoprotein (PAG) levels in plasma and milk of Holstein cows during early pregnancy and their effect on the accuracy of pregnancy diagnosis. *Journal of Dairy Science*, 98(4):2502–2514, 2015. doi:10.3168/jds.2014-8974.

A. Ricci, P. D. Carvalho, M. C. Amundson, and P. M. Fricke. Characterization of luteal dynamics in lactating Holstein cows for 32 days after synchronization of ovulation and timed artificial insemination. *Journal of Dairy Science*, 100(12):9851–9860, 2017. doi:10.3168/jds.2017-13293.

B. D. Rosen, 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. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, 9(3), 2020. doi:10.1093/gigascience/giaa021.

G. Rovere, G. de los Campos, R. J. Tempelman, A. I. Vazquez, F. Miglior, F. Schenkel, A. Cecchinato, G. Bittante, H. Toledo-Alvarado, and A. Fleming. A landscape of the heritability of Fourier-transform infrared spectral wavelengths of milk samples by parity and lactation stage in Holstein cows. *Journal of Dairy Science*, 102(2):1354–1363, 2019. doi:10.3168/jds.2018-15109.

M. J. Rutten, H. Bovenhuis, K. A. Hettinga, H. J. van Valenberg, and J. A. van Arendonk. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. *Journal of Dairy Science*, 92(12):6202–6209, 2009. doi:10.3168/jds.2009-2456.

M. J. Rutten, H. Bovenhuis, and J. A. van Arendonk. 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. *Journal of Dairy Science*, 93(10):4872–4882, 2010. doi:10.3168/jds.2010-3157.

M. J. Rutten, H. Bovenhuis, J. M. Heck, and J. A. van Arendonk. Predicting bovine milk protein composition based on Fourier transform infrared spectra. *Journal of Dairy Science*, 94(11):5683–5690, 2011. doi:10.3168/jds.2011-4520.

M. Safar, D. Bertrand, P. Robert, M. F. Devaux, and C. Genot. Characterization of edible oils, butters and margarines by Fourier transform infrared spectroscopy with attenuated total reflectance. *Journal of the American Oil Chemists' Society*, 71(4):371, 1994. doi:10.1007/BF02540516.

M.-P. Sanchez, A. Govignon-Gion, M. Ferrand, M. Gelé, D. Pourchet, Y. Amigues, S. Fritz, M. Boussaha, A. Capitan, D. Rocha, G. Miranda, P. Martin, M. Brochard, and D. Boichard. Whole-genome scan to detect quantitative trait loci associated with milk protein composition in 3 French dairy cattle breeds. *Journal of Dairy Science*, 99(10):8203–8215, 2016. doi:10.3168/jds.2016-11437.

M.-P. Sanchez, M. Ferrand, M. Gelé, D. Pourchet, G. Miranda, P. Martin, M. Brochard, and D. Boichard. Short communication: Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande, and Holstein dairy cattle breeds. *Journal of Dairy Science*, 100(8):6371–6375, 2017a. doi:10.3168/jds.2017-12663.

M.-P. Sanchez, 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. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genetics, selection, evolution: GSE*, 49(1):68, 2017b. doi:10.1186/s12711-017-0344-z.

M.-P. Sanchez, M. El Jabri, S. Minéry, V. Wolf, E. Beuvier, C. Laithier, A. Delacroix-Buchet, M. Brochard, and D. Boichard. Genetic parameters for cheese-making properties and milk composition predicted from mid-infrared spectra in a large data set of Montbéliarde cows. *Journal of Dairy Science*, 101(11): 10048–10061, 2018. doi:10.3168/jds.2018-14878.

M.-P. Sanchez, 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. Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows. *Genetics Selection Evolution*, 51(1):34, 2019. doi:10.1186/s12711-019-0473-7.

A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. doi:10.1021/ac60214a047.

Y. A. Savva, L. E. Rieder, and R. A. Reenan. The ADAR protein family. *Genome biology*, 13(12):1, 2012. doi:10.1186/gb-2012-13-12-252.

A. Schennink, W. M. Stoop, M. H. Visker, J. M. Heck, H. Bovenhuis, J. J. van der Poel, H. J. van Valenberg, and J. A. van Arendonk. *DGAT1* underlies large genetic variation in milk-fat composition of dairy cows. *Animal Genetics*, 38(5):467–473, 2007. doi:10.1111/j.1365-2052.2007.01635.x.

A. Schennink, J. M. L. Heck, H. Bovenhuis, M. H. Visker, H. J. van Valenberg, and J. A. van Arendonk. Milk fatty acid unsaturation: genetic parameters and effects of stearoyl-CoA desaturase (*SCD1*) and acyl CoA: diacylglycerol acyltransferase 1 (*DGAT1*). *Journal of dairy science*, 91(5):2135–2143, 2008. doi:10.3168/jds.2007-0825.

A. Schennink, H. Bovenhuis, K. M. Léon-Kloosterziel, J. A. van Arendonk, and M. H. Visker. Effect of polymorphisms in the *FASN*, *OLR1*, *PPARGC1A*, *PRL* and *STAT5A* genes on bovine milk-fat composition. *Animal Genetics*, 40(6):909–916, 2009. doi:10.1111/j.1365-2052.2009.01940.x.

G. C. Schopen, J. M. Heck, H. Bovenhuis, M. H. Visker, H. J. van Valenberg, and J. A. van Arendonk. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. *Journal of Dairy Science*, 92 (3):1182–1191, 2009. doi:10.3168/jds.2008-1281.

G. C. Schopen, M. H. Visker, P. D. Koks, E. Mullaart, J. A. van Arendonk, and H. Bovenhuis. Whole-genome association study for milk protein composition in dairy cattle. *Journal of Dairy Science*, 94(6): 3148–3158, 2011. doi:10.3168/jds.2010-4030.

D. Schuhmacher, B. Bähre, C. Gottschlich, F. Heinemann, and B. Schmitzer. transport: Optimal transport in various forms, 2017. R package version 0.9-4. `https://cran.r-project.org/package=transport`.

N. Shetty, P. Løvendahl, M. S. Lund, and A. J. Buitenhuis. Prediction and validation of residual feed intake and dry matter intake in Danish lactating dairy cows using mid-infrared spectroscopy of milk. *Journal of Dairy Science*, 100(1):253–264, 2017. doi:10.3168/jds.2016-11609.

P. Shine and M. D. Murphy. Over 20 Years of Machine Learning Applications on Dairy Farms: A Comprehensive Mapping Study. *Sensors*, 22(1):52, 2022. doi:10.3390/s22010052.

S. L. Smith, S. J. Denholm, M. P. Coffey, and E. Wall. Energy profiling of dairy cows from routine milk mid-infrared analysis. *Journal of Dairy Science*, 102(12):11169–11179, 2019. doi:10.3168/jds.2018-16112.

N. M. Sousa, A. Ayad, J. F. Beckers, and Z. Gajewski. Pregnancy-associated glycoproteins (PAG) as pregnancy markers in the ruminants. *Journal of Physiology and Pharmacology: An Official Journal of the Polish Physiological Society*, 57 Suppl 8:153–171, 2006. PMID: 17242480.

H. Soyeurt, P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. Estimating fatty acid content in cow milk using mid-infrared spectrometry. *Journal of Dairy Science*, 89(9):3690–3695, 2006. doi:10.3168/jds.S0022-0302(06)72409-2.

H. Soyeurt, F. G. Colinet, V. M.-R. Arnould, P. Dardenne, C. Bertozzi, R. Renaville, D. Portetelle, and N. Gengler. Genetic variability of lactoferrin content estimated by mid-infrared spectrometry in bovine milk. *Journal of Dairy Science*, 90(9):4443–4450, 2007a. doi:10.3168/jds.2006-827.

H. Soyeurt, A. Gillon, S. Vanderick, P. Mayeres, C. Bertozzi, and N. Gengler. Estimation of heritability and genetic correlations for the major fatty acids in bovine milk. *Journal of Dairy Science*, 90(9): 4435–4442, 2007b. doi:10.3168/jds.2007-0054.

H. Soyeurt, D. Bruwier, J.-M. Romnee, N. Gengler, C. Bertozzi, D. Veselko, and P. Dardenne. Potential estimation of major mineral contents in cow milk using mid-infrared spectrometry. *Journal of Dairy Science*, 92(6):2444–2454, 2009. doi:10.3168/jds.2008-1734.

H. Soyeurt, I. Misztal, and N. Gengler. Genetic variability of milk components based on mid-infrared spectral data. *Journal of Dairy Science*, 93(4):1722–1728, 2010. doi:10.3168/jds.2009-2614.

H. Soyeurt, F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *Journal of Dairy Science*, 94(4):1657–1667, 2011. doi:10.3168/jds.2010-3408.

H. Soyeurt, 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. Mid-infrared prediction of lactoferrin content in bovine milk: potential indicator of mastitis. *Animal*, 6(11):1830–1838, 2012. doi:10.1017/S1751731112000791.

J. W. Spek, J. Dijkstra, G. van Duinkerken, W. H. Hendriks, and A. Bannink. Prediction of urinary nitrogen and urinary urea nitrogen excretion by lactating dairy cattle in northwestern Europe and North America: A meta-analysis. *Journal of Dairy Science*, 96(7):4310–4322, 2013. doi:10.3168/jds.2012-6265.

R. Spelman, F. Miller, J. Hooper, M. Thielen, and D. Garrick. Experimental design for QTL trial involving New Zealand Friesian and Jersey breeds. In *Proceedings of the Association for the Advancement of Animal Breeding and Genetics*, volume 14, pages 393–396, 2001. http://www.aaabg.org/livestocklibrary/2001/ab01093.pdf.

W. M. Stoop, H. Bovenhuis, and J. A. van Arendonk. Genetic parameters for milk urea nitrogen in relation to milk production traits. *Journal of Dairy Science*, 90(4):1981–1986, 2007. doi:10.3168/jds.2006-434.

L. K. Sørensen, M. Lund, and B. Juul. Accuracy of Fourier transform infrared spectrometry in determination of casein in dairy cows' milk. *The Journal of Dairy Research*, 70(4):445–452, 2003. doi:10.1017/s0022029903006435.

M. F. Te Pas, O. Madsen, M. P. Calus, and M. A. Smits. The importance of endophenotypes to evaluate the relationship between genotype and external phenotype. *International Journal of Molecular Sciences*, 18(2):472, 2017. doi:10.3390/ijms18020472.

M. Timlin, J. T. Tobin, A. Brodkorb, E. G. Murphy, P. Dillon, D. Hennessy, M. O'Donovan, K. M. Pierce, and T. F. O'Callaghan. The impact of seasonality in pasture-based production systems on milk composition and functionality. *Foods*, 10(3):607, 2021. doi:10.3390/foods10030607.

K. M. Tiplady, R. G. Sherlock, M. D. Littlejohn, J. E. Pryce, S. R. Davis, D. J. Garrick, R. J. Spelman, and B. L. Harris. Strategies for noise reduction and standardization of milk mid-infrared spectra from dairy cattle. *Journal of Dairy Science*, 102(7):6357–6372, 2019. doi:10.3168/jds.2018-16144.

K. M. Tiplady, T. J. Lopdell, M. D. Littlejohn, and D. J. Garrick. The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle. *Journal of Animal Science and Biotechnology*, 11(1):39, 2020. doi:10.1186/s40104-020-00445-2.

K. M. Tiplady, T. J. Lopdell, E. Reynolds, R. G. Sherlock, M. Keehan, T. J. Johnson, J. E. Pryce, S. R. Davis, R. J. Spelman, B. L. Harris, D. J. Garrick, and M. D. Littlejohn. Data from: Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. In *Data from: Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle*. Dryad Digital Repository, 2021a. https://doi.org/10.5061/dryad.qrfj6q5dj.

K. M. Tiplady, T. J. Lopdell, E. Reynolds, R. G. Sherlock, M. Keehan, T. J. Johnson, J. E. Pryce, S. R. Davis, R. J. Spelman, B. L. Harris, D. J. Garrick, and M. D. Littlejohn. Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. *Genetics Selection Evolution*, 53(1):62, 2021b. doi:10.1186/s12711-021-00648-9.

K. M. Tiplady, M.-H. Trinh, S. R. Davis, R. G. Sherlock, R. J. Spelman, D. J. Garrick, and B. L. Harris. Pregnancy status predicted using milk mid-infrared spectra from dairy cattle. *Journal of Dairy Science*, 105(4):3615–3632, 2022. doi:10.3168/jds.2021-21516.

V. Toffanin, M. De Marchi, N. Lopez-Villalobos, and M. Cassandro. Effectiveness of mid-infrared spectroscopy for prediction of the contents of calcium and phosphorus, and titratable acidity of milk and their relationship with milk quality and coagulation properties. *International Dairy Journal*, 41: 68–73, 2015. doi:10.1016/j.idairyj.2014.10.002.

H. Toledo-Alvarado, A. I. Vazquez, G. Campos, R. J. Tempelman, G. Bittante, and A. Cecchinato. Diagnosing pregnancy status using infrared spectra and milk composition in dairy cows. *Journal of Dairy Science*, 101(3):2496–2505, 2018a. doi:10.3168/jds.2017-13647.

H. Toledo-Alvarado, A. I. Vazquez, G. Campos, R. J. Tempelman, G. Gabai, A. Cecchinato, and G. Bittante. Changes in milk characteristics and fatty acid profile during the estrous cycle in dairy cows. *Journal of Dairy Science*, 101(10):9135–9153, 2018b. doi:10.3168/jds.2018-14480.

T. Tribout, P. Croiseau, R. Lefebvre, A. Barbat, M. Boussaha, S. Fritz, D. Boichard, C. Hoze, and M.-P. Sanchez. Confirmed effects of candidate variants for milk production, udder health, and udder morphology in dairy cattle. *Genetics Selection Evolution*, 52(1):55, 2020. doi:10.1186/s12711-020-00575-1.

I. van den Berg, D. Boichard, B. Guldbrandtsen, and M. S. Lund. Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: a simulation study. *G3: Genes, Genomes, Genetics*, 6(8):2553–2561, 2016. doi:10.1534/g3.116.027730.

I. van den Berg, P. N. Ho, M. Haile-Mariam, and J. E. Pryce. Genetic parameters for mid-infrared spectroscopy–predicted fertility. *JDS Communications*, 2(6):361–365, 2021a. doi:10.3168/jdsc.2021-0141.

I. van den Berg, P. N. Ho, T. D. W. Luke, M. Haile-Mariam, S. Bolormaa, and J. E. Pryce. The use of milk mid-infrared spectroscopy to improve genomic prediction accuracy of serum biomarkers. *Journal of Dairy Science*, 104(2):2008–2017, 2021b. doi:10.3168/jds.2020-19468.

S. G. van der Drift, K. J. v. Hulzen, T. G. Teweldemedhn, R. Jorritsma, M. Nielen, and H. C. Heuven. Genetic and nongenetic variation in plasma and milk β-hydroxybutyrate and milk acetone concentrations of early-lactation dairy cows. *Journal of Dairy Science*, 95(11):6781–6787, 2012. doi:10.3168/jds.2012-5640.

S. van Gastelen, E. C. Antunes-Fernandes, K. A. Hettinga, and J. Dijkstra. Short communication: The effect of linseed oil and *DGAT1* K232A polymorphism on the methane emission prediction potential of milk fatty acids. *Journal of Dairy Science*, 101(6):5599–5604, 2018a. doi:10.3168/jds.2017-14131.

S. van Gastelen, H. Mollenhorst, E. C. Antunes-Fernandes, K. A. Hettinga, G. G. van Burgsteden, J. Dijkstra, and J. L. W. Rademaker. Predicting enteric methane emission of dairy cows with milk Fourier-transform infrared spectra and gas chromatography–based milk fatty acid profiles. *Journal of Dairy Science*, 101(6):5582–5598, 2018b. doi:10.3168/jds.2017-13052.

A. T. van Knegsel, S. G. van der Drift, M. Horneman, A. P. de Roos, B. Kemp, and E. A. Graat. Short communication: Ketone body concentration in milk determined by Fourier transform infrared spectroscopy: Value for the detection of hyperketonemia in dairy cows. *Journal of Dairy Science*, 93 (7):3065–3069, 2010. doi:10.3168/jds.2009-2847.

J. D. van Wyngaard and R. Meeske. Palm kernel expeller increases milk fat content when fed to grazing dairy cows. *South African Journal of Animal Science*, 47(2):219–230, 2017. doi:10.4314/sajas.v47i2.14.

A. Vanlierde, F. Dehareng, E. Froidmont, P. Dardenne, P. Kandel, N. Gengler, M. H. Deighton, F. Buckley, E. Lewis, and S. McParland. Prediction of the individual enteric methane emission of dairy cows from milk mid-infrared spectra. *Proc. of the 5th Greenhouse Gasses and Animal Agriculture (GGAA2013*, 4 (2):433, 2013. https://orbi.uliege.be/handle/2268/217000.

A. Vanlierde, M. L. Vanrobays, F. Dehareng, E. Froidmont, H. Soyeurt, S. McParland, E. Lewis, M. H. Deighton, F. Grandl, M. Kreuzer, and B. Gredler. Hot topic: Innovative lactation-stage-dependent prediction of methane emissions from milk mid-infrared spectra. *Journal of Dairy Science*, 98(8): 5740–5747, 2015. doi:10.3168/jds.2014-8436.

A. Vanlierde, M.-L. Vanrobays, N. Gengler, P. Dardenne, E. Froidmont, h. Soyeurt, S. McParland, E. Lewis, M. Deighton, M. Mathot, and F. Dehareng. Milk mid-infrared spectra enable prediction of lactation-stage-dependent methane emissions of dairy cattle within routine population-scale milk recording schemes. *Animal Production Science*, 56:258, 2016. doi:10.1071/AN15590.

A. Vanlierde, H. Soyeurt, N. Gengler, F. G. Colinet, E. Froidmont, M. Kreuzer, F. Grandl, M. Bell, P. Lund, D. W. Olijhoek, M. Eugène, C. Martin, B. Kuhla, and F. Dehareng. Short communication: Development of an equation for estimating methane emissions of dairy cows from milk Fourier transform mid-infrared spectra by using reference data obtained exclusively from respiration chambers. *Journal of Dairy Science*, 101(8):7618–7624, 2018. doi:10.3168/jds.2018-14472.

A. Vanlierde, F. Dehareng, N. Gengler, E. Froidmont, S. McParland, M. Kreuzer, M. Bell, P. Lund, C. Martin, B. Kuhla, and H. Soyeurt. Improving robustness and accuracy of predicted daily methane emissions of dairy cows using milk mid-infrared spectra. *Journal of the Science of Food and Agriculture*, 101(8):3394–3403, 2021. doi:10.1002/jsfa.10969.

E. Viale, F. Tiezzi, F. Maretto, M. De Marchi, M. Penasa, and M. Cassandro. Association of candidate gene polymorphisms with milk technological traits, yield, composition, and somatic cell score in Italian Holstein-Friesian sires. *Journal of Dairy Science*, 100(9):7271–7281, 2017. doi:10.3168/jds.2017-12666.

G. Visentin, A. McDermott, S. McParland, D. P. Berry, O. A. Kenny, A. Brodkorb, M. A. Fenelon, and M. De Marchi. Prediction of bovine milk technological traits from mid-infrared spectroscopy analysis in dairy cows. *Journal of Dairy Science*, 98(9):6620–6629, 2015. doi:10.3168/jds.2015-9323.

G. Visentin, S. McParland, M. De Marchi, A. McDermott, M. A. Fenelon, M. Penasa, and D. P. Berry. Processing characteristics of dairy cow milk are moderately heritable. *Journal of Dairy Science*, 100 (8):6343–6355, 2017. doi:10.3168/jds.2017-12642.

G. Visentin, M. Penasa, G. Niero, M. Cassandro, and M. De Marchi. Phenotypic characterisation of major mineral composition predicted by mid-infrared spectroscopy in cow milk. *Italian Journal of Animal Science*, 17(3):549–556, 2018. doi:10.1080/1828051X.2017.1398055.

D. Wang, C. Ning, J.-F. Liu, Q. Zhang, and L. Jiang. Short communication: Replication of genome-wide association studies for milk production traits in Chinese Holstein by an efficient rotated linear mixed model. *Journal of Dairy Science*, 102(3):2378–2383, 2019a. doi:10.3168/jds.2018-15298.

Q. Wang and H. Bovenhuis. Genome-wide association study for milk infrared wavenumbers. *Journal of Dairy Science*, 101(3):2260–2272, 2018. doi:10.3168/jds.2017-13457.

Q. Wang and H. Bovenhuis. Validation strategy can result in an overoptimistic view of the ability of milk infrared spectra to predict methane emission of dairy cattle. *Journal of Dairy Science*, 102(7): 6288–6295, 2019. doi:10.3168/jds.2018-15684.

Q. Wang, A. Hulzebosch, and H. Bovenhuis. Genetic and environmental variation in bovine milk infrared spectra. *Journal of Dairy Science*, 99(8):6793–6803, 2016. doi:10.3168/jds.2015-10488.

X. Wang, C. Wurmser, H. Pausch, S. Jung, F. Reinhardt, J. Tetens, G. Thaller, and R. Fries. Identification and dissection of four major QTL affecting milk fat content in the German Holstein-Friesian population. *PloS one*, 7(7):e40711, 2012. doi:10.1371/journal.pone.0040711.

Y. Wang, D. J. Veltkamp, and B. R. Kowalski. Multivariate instrument standardization. *Analytical Chemistry*, 63(23):2750–2756, 1991. doi:10.1021/ac00023a016.

Z. Wang, 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. Genome wide association study identifies SNPs associated with fatty acid composition in Chinese Wagyu cattle. *Journal of Animal Science and Biotechnology*, 10(1):27, 2019b. doi:10.1186/s40104-019-0322-0.

J. Wei, P. F. Geale, P. A. Sheehy, and P. Williamson. The impact of *ABCG2* on bovine mammary epithelial cell proliferation. *Animal Biotechnology*, 23(3):221–224, 2012. doi:10.1080/10495398.2012.696567.

S. L. White, J. A. Bertrand, M. R. Wade, S. P. Washburn, J. T. Green, and T. C. Jenkins. Comparison of fatty acid content of milk from Jersey and Holstein cows consuming pasture or a total mixed ration. *Journal of Dairy Science*, 84(10):2295–2301, 2001. doi:10.3168/jds.S0022-0302(01)74676-0.

R. Wightman. PyTorch Image Models. 2019. doi:10.5281/zenodo.4414861. GitHub repository.

P. C. Williams and K. H. Norris. Qualitative applications of near infrared reflectance spectroscopy. *Near-Infrared Technology in the Agricultural and Food Industries. P. Williams and K. Norris, ed. American Association of Cereal Chemists, St. Paul, MN*, pages 241–246, 1987.

H. Winning, K. M. Mulawa, and T. Selberg. Standardization of FT-IR instruments. *White Paper from Foss A/S*, 1(1):7, 2014.

G. M. Wood, P. J. Boettcher, J. Jamrozik, G. B. Jansen, and D. F. Kelton. Estimation of genetic parameters for concentrations of milk urea nitrogen. *Journal of Dairy Science*, 86(7):2462–2469, 2003. doi:10.3168/jds.S0022-0302(03)73840-5.

F.-i. Yamamoto, H. Clausen, T. White, J. Marken, and S.-i. Hakomori. Molecular genetic basis of the histo-blood group ABO system. *Nature*, 345(6272):229–233, 1990. doi:10.1038/345229a0.

J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1):76–82, 2011. doi:10.1016/j.ajhg.2010.11.011.

P. Yodklaew, S. Koonawootrittriron, M. A. Elzo, T. Suwanasopee, and T. Laodim. Genome-wide association study for lactation characteristics, milk yield and age at first calving in a Thai multibreed dairy cattle population. *Agriculture and Natural Resources*, 51(3):223–230, 2017. doi:10.1016/j.anres.2017.04.002.

R. M. Zaalberg, N. Shetty, L. Janss, and A. J. Buitenhuis. Genetic analysis of Fourier transform infrared milk spectra in Danish Holstein and Danish Jersey. *Journal of Dairy Science*, 102(1):503–510, 2019. doi:10.3168/jds.2018-14464.

R. M. Zaalberg, L. Janss, and A. J. Buitenhuis. Genome-wide association study on Fourier transform infrared milk spectra for two Danish dairy cattle breeds. *BMC Genetics*, 21(1):9, 2020. doi:10.1186/s12863-020-0810-4.

S. Zakizadeh, M. Reissmann, S. R. Miraee-Ashtiani, and P. Reinecke. Polymorphism of beta-lactoglobulin coding and 5'-flanking regions and association with milk production traits. *Biotechnology & Biotechnological Equipment*, 26(1):2716–2721, 2012. doi:10.5504/BBEQ.2011.0095.

S. Zhai, J. Liu, Y. Wu, and J. Ye. Predicting urinary nitrogen excretion by milk urea nitrogen in lactating Chinese Holstein cows. *Animal Science Journal*, 78(4):395–399, 2007. doi:10.1111/j.1740-0929.2007.00452.x.

C. Zhou, C. Li, W. Cai, S. Liu, H. Yin, S. Shi, Q. Zhang, and S. Zhang. Genome-wide association study for milk protein composition traits in a Chinese Holstein population using a single-step approach. *Frontiers in Genetics*, 10:72, 2019. doi:10.3389/fgene.2019.00072.

B. Zhu, 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. Genome wide association study and genomic prediction for fatty acid composition in Chinese Simmental beef cattle using high density SNP array. *BMC Genomics*, 18 (1):464, 2017. doi:10.1186/s12864-017-3847-7.