


RESEARCH ARTICLE

Open Access



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

Kathryn M. Tiplady<sup>1,2\*</sup> , Thomas J. Lopdell<sup>1</sup>, Edwardo Reynolds<sup>2</sup>, Richard G. Sherlock<sup>1</sup>, Michael Keehan<sup>1</sup>, Thomas J.J. Johnson<sup>1</sup>, Jennie E. Pryce<sup>3,4</sup>, Stephen R. Davis<sup>1</sup>, Richard J. Spelman<sup>1</sup>, Bevin L. Harris<sup>1</sup>, Dorian J. Garrick<sup>2</sup> and Mathew D. Littlejohn<sup>1,2</sup>

## Abstract

**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.

**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.

**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.

## 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

\*Correspondence: Kathryn.Sanders@lic.co.nz

<sup>1</sup> Research and Development, Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand  
Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

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 [1–4]. Recent research includes studies of milk composition traits that are relevant to manufacturing traits [5–7], individual fatty acids and milk proteins [8, 9], and indirect traits that are related to energy status [10, 11], pregnancy and fertility [12–14], methane emissions [15–17] and bovine tuberculosis [18].

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 [19]. Studies have reported moderate to high heritability estimates for a range of FT-MIR predicted traits, including fatty acids [20–22], milk proteins [9, 23], cheese-making and milk-coagulation properties [24–26], and lactoferrin concentrations [27, 28]. 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 [29–32]. 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 [33–37], and individual fatty acid and protein fractions [38–40], there are comparatively few studies reporting GWAS results for individual FT-MIR wavenumber phenotypes [41–43]. Two such GWAS were conducted on medium density SNP-chip (~50 k markers) genotypes for a subset of wavenumbers, which were identified either by clustering analysis [41], or by using phenotypic correlation structures and heritability estimates within each breed [43]. 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 [42].

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 [41]. 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 [32, 42]. Wang et al. [32] observed that the *DGATI* 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 [32]. Similar effects were also observed by Benedet et al. [42], 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 *DGATI* 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 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, identifying 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.

**Methods**

**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 1645 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 [44]. Records were included in the present study if they passed outlier removal based on the 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; 5374 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 1544 cows from other breeds or crosses. Individual FT-MIR wavenumbers were subjected to piecewise direct standardization [45], with standardization coefficients evaluated from 16 weeks of reference sample calibration data collected across six Bentley instruments as in Tiplady et al. [44].

**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 [46], comprising:

$$y_{ijkl} = \mu + parity_j + dim_k + HD_l + \sum \alpha_m brd_{im} + \sum \delta_n het_{in} + anml_i + e_{ijkl}, \tag{1}$$

where  $y_{ijkl}$  is a test-day phenotype (e.g. absorbance for one wavenumber) for the  $i$  th individual in parity class  $j$  within the days in milk class  $k$  and the herd-by-test date group  $l$ ;  $\mu$  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  $l$ ;  $\alpha_m$  are breed linear regression coefficients for Holstein (HOL), Friesian (FR) and Jersey (JE) proportions and  $brd_{im}$  are the corresponding breed proportions for individual  $i$ ;  $\delta_n$  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  $anml_i \sim N(0, I\sigma_{anml}^2)$ ; and  $e_{ijkl}$  is the random error effect with  $e_{ijkl} \sim N(0, I\sigma_e^2)$ , where  $I$  is an identity matrix and  $\sigma_{anml}^2$  and  $\sigma_e^2$  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.  $\hat{y}_{i(adj)} = anml_i + \bar{e}_{ij}$ .

**Genotypes and imputation**

Animals were genotyped on Illumina BovineHD (HD;  $N=138$ ;  $\sim 777$  k SNP), Illumina BovineSNP50k (50 k;  $N=4087$ ;  $\sim 53$  k SNP), and/or custom GeneSeek Genomic Profiler LDv3 BeadChip (GGP;  $N=33,976$ ;  $\sim 26$  k 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. [47]. 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, 50 k 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) [48], and GATK (v3.2) [49] respectively, as previously described [35, 36, 47]. The pedigree-based breed composition of sequenced animals comprised 138 Holstein-Friesians, 99 Jerseys, 316 Holstein-Friesian  $\times$  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 [50]. 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 [51] 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  $\geq 0.95$ , and the proportion of Mendelian inconsistencies observed between parent-offspring pairs of genotypes was lower than 0.005. The 50 k 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 50 k 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 ( $N=3389$ ), 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 ( $MAF_{overall}$ ) satisfied the criteria:  $|MAF_{ref} - MAF_{overall}| / MAF_{ref} < 0.4$ . 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  $30 \times$  depth coverage were removed if the concordance between genotype and sequencing calls was  $\leq 0.7$ . Likewise, for GGP and 50 k references, any SNPs that were shared with the BovineHD panel were removed if the concordance between genotype calls from each panel was  $\leq 0.7$ ; and for the HD reference, any SNPs that were shared with the BovineSNP50 panel were removed if the genotypic concordance between panels was  $\leq 0.7$ .

### Imputation

All imputation steps were carried out ignoring pedigree information using Beagle 4.0 [50]. Imputation of animals to GGP, 50 k 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.

### Genome-wide association studies

Separate GWAS were conducted using the Bolt-LMM software [52] 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 50 k 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  $-\log_{10}(6.2e-13) = 12.21$ . 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  $\sigma_t^2$  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  $\sigma_g^2$  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 signals.

### Gene expression phenotypes and eQTL identification

Gene expression phenotypes and the resulting eQTL were generated as part of a previously described study [36]. 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) [53], 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 [36]. 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 (A) matrix. Resulting  $\chi^2$  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.

#### 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) [54] 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  $1.5 \times$  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  $2 \times$  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  $2.5 \times$  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. [36]. 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.

#### 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) [55] with default parameters.

## Results

### Sequence-based genome-wide association analysis

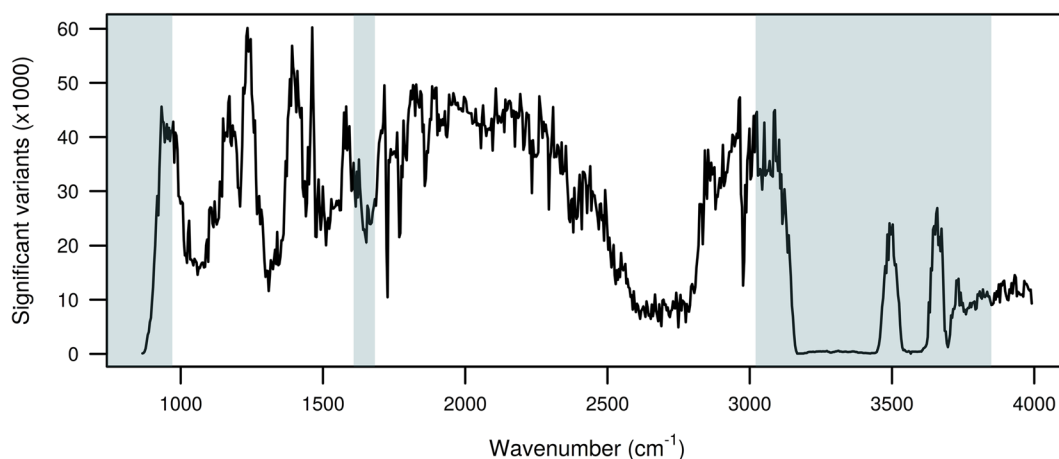
The first-round pre-iteration (base) GWAS, including 17,873,880 imputed sequence variants, resulted in significant associations for 37,779 variants for FP, 17,159 variants for LP, and 36,067 variants for PP. The number of significant associations for individual FT-MIR wavenumbers ranged from 50 to 60,242, 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. 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. [44]. 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 associations observed, 17.0% were positioned within the first 3 Mbp of chromosome 14, which encompasses the *DGATI* gene that has been widely reported as impacting many milk composition traits [56, 57]. 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.

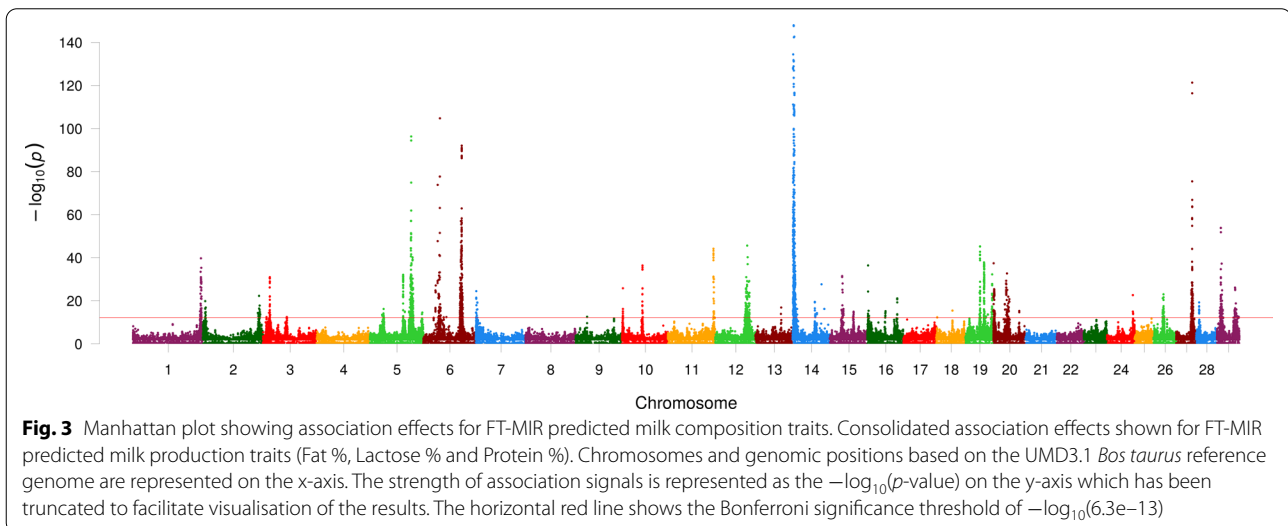
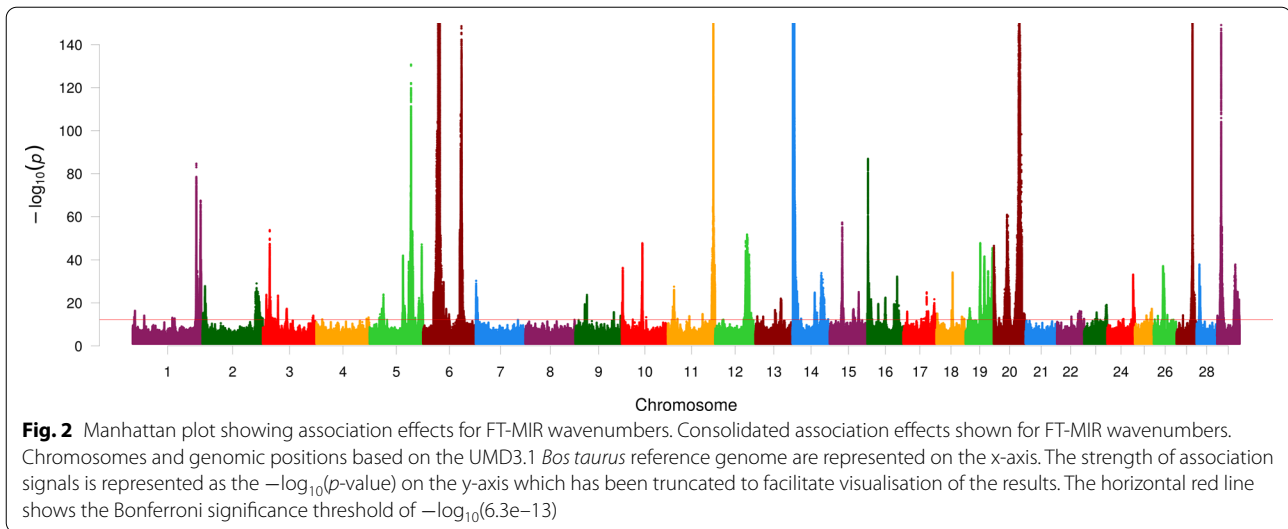
In the base GWAS, individual FT-MIR wavenumber QTL were observed on 27 of the 29 bovine autosomes (Fig. 2) within 450 different 1-Mbp regions. In contrast, QTL for FT-MIR predicted milk composition traits were observed on 25 of the 29 autosomes (Fig. 3) within 246 different 1-Mbp regions. The number 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.

### 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 1. Manhattan plots of a 1-Mbp region centred on the QTL tag



**Fig. 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  $\text{cm}^{-1}$ , from 1608 to 1682  $\text{cm}^{-1}$ , and from 3021 to 3849  $\text{cm}^{-1}$



variant for each of the 31 highly significant wavenumber QTL from the base GWAS are provided (see Additional file 1: Figure S1). Details of the wavenumber QTL classified as moderately significant are in Table S2 (see Additional file 2: Table S2). 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 1), and moderately significant based on the LD with other loci (see Additional file 2: Table S2). Effect sizes and MAF details for the tag SNP of the 31 highly significant wavenumber QTL are in Table S3 (see Additional file 2: Table S3). 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 in Table S4 (see Additional file 2: Table S4). 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 1),  $p$ -values at previous iterations for the phenotype, and  $p$ -values for the corresponding top chromosomal SNP in that iteration are in Table S5 (see Additional file 2: Table S5).

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 2, with details of the QTL classified as moderately significant in Table S6 (see Additional file 2: Table S6). Effect sizes and MAF details for highly significant predicted-trait QTL are in Table S7 (see Additional file 2: Table S7). Details of highly

**Table 1** Peak variants for FT-MIR wavenumbers with highly significant protein sequence association effects

Chr	Position	Tag variant ID	N of hits	Top wvn (cm <sup>-1</sup> )	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	4.3e-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	rs37779402	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  $-\log_{10}(p\text{-value})$  for the effect was greater than  $1.5 \times$  the Bonferroni threshold and the correlation between the tag variant and the protein sequence variant was in the range (0.975, 1]; or the  $-\log_{10}(p\text{-value})$  for the effect was greater than  $2 \times$  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  $-\log_{10}(p\text{-value})$  for the effect was greater than  $2.5 \times$  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:  $-\log_{10}(6.2e-13)$ . N of hits: number of wavenumbers for which the variant was selected as the representative (most significant) tag variant for a peak. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0. L Low impact splice region variant, M Moderate impact missense variant, H High impact splice donor



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

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

significant predicted-trait QTL from iterations subsequent to the base GWAS in Table S8 (see Additional file 2: Table S8).

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 1 and 2). Since this class of variants

was likely to be enriched for annotation errors [58], 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. 4), which suggests that the ABO protein in animals homozygous for the mutation is non-functional.

#### 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 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 (see Additional file 3: Figure S9). Effect sizes and MAF details for all 38 loci with a co-localized trait QTL and eQTL pair are in Table S10 (see Additional file 4: Table S10). For each of these 38 loci, the proportions of phenotypic and genetic variance explained across FT-MIR wavenumber and predicted composition traits are in Table S11 (see Additional file 4: Table S11). 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 in Table S12 (see Additional file 4: Table S12).

Co-localized eQTL were identified for 25 predicted-trait QTL. Details of these trait QTL and eQTL pairs are in Table 4, with effect sizes and MAF details provided in Table S13 (see Additional file 4: Table S13). Further details of the 12 QTL identified in iterations subsequent to the base GWAS are in Table S14 (see Additional file 4: Table S14).

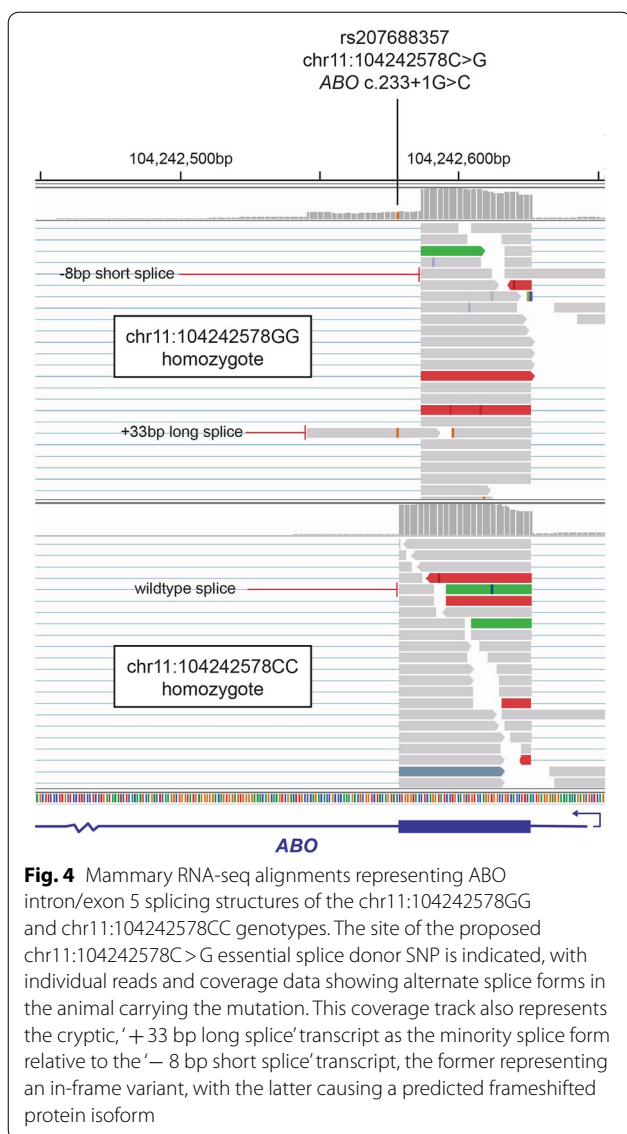
#### Investigation of patterns of FT-MIR wavenumber associations for genes of interest

In total, 70 genes were implicated whereby the tag locus of the wavenumber QTL was in high LD with a PIV (Table 1), or in high LD with the top variant of a co-localized eQTL (Table 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. 5). Significance profiles for all 59 genes are provided in Figure S15 (see Additional file 5: Figure S15).

Significance profiles for a subset of gene clusters from Fig. 5 are presented in Fig. 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. 5) shown in light grey. Significance profiles varied widely between clusters, but were highly consistent within clusters. The first cluster (Fig. 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 ~1020 to 1180  $\text{cm}^{-1}$ , from ~1200 to 1470  $\text{cm}^{-1}$ , from ~2610 to 2840  $\text{cm}^{-1}$  and from ~2870 to 2980  $\text{cm}^{-1}$ . The second cluster (Fig. 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 1010  $\text{cm}^{-1}$ , from ~1070 to 1560  $\text{cm}^{-1}$ , from ~1700 to 2450  $\text{cm}^{-1}$ , from ~2630 to 2980  $\text{cm}^{-1}$  and from ~3620 to 3680  $\text{cm}^{-1}$ . The third cluster (Fig. 6c) includes a number of genes with significant associations for the FP (*DGATI*, *ABO*, *TGDS*, *GPAT4*, *MGST1*, *MROH1*) and PP (*DGATI*, *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. 6a, b), including from ~910 to 1010  $\text{cm}^{-1}$ , from ~1130 to 1260  $\text{cm}^{-1}$ , from ~1450 to 1500  $\text{cm}^{-1}$ , from ~1700 to 2450  $\text{cm}^{-1}$ , and from 3620 to 3680  $\text{cm}^{-1}$ . Other notable peaks observed for this cluster were from ~1570 to 1700  $\text{cm}^{-1}$ , from ~2820 to 3150  $\text{cm}^{-1}$ , and from ~3460 to 3530  $\text{cm}^{-1}$ .

Significance profiles for gene clusters represented by *CSN3*, *PAEP* and *ANKH* are shown in Fig. 7. The pattern of significance in the profiles represented by *CSN3* and *PAEP* (Fig. 7a and b) were similar, in that a large proportion of the signal was captured within a small part of the mid-infrared range; namely from ~1220 to 1780  $\text{cm}^{-1}$  for the gene cluster represented by *CSN3*, and from ~1350 to ~1650  $\text{cm}^{-1}$  for the gene cluster represented by *PAEP*. Although *ANKH* appeared to be an outlier in the clustering analysis (Fig. 5), a similar pattern was observed with most of the signal captured within three prominent peaks in the range from ~1260 to 1620  $\text{cm}^{-1}$ . Two of these peaks, centred at ~1391  $\text{cm}^{-1}$  and 1582  $\text{cm}^{-1}$  were in common with peaks observed for the *PAEP* profile. From the first cluster (Fig. 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. 7b), the *PAEP* and *CCDC57* genes had significant associations with the FP phenotype, whilst *ANKH* had a significant association with the LP phenotype (Fig. 7c).



## Discussion

### 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 [41, 42], 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 *DGATI* gene. This corroborates findings from previous studies which also observed significant associations between the *DGATI* K232A polymorphism and wavenumbers in the regions from 1619 to 1674  $\text{cm}^{-1}$  and from 3073 to 3667  $\text{cm}^{-1}$  [32, 41].

### 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 *DGATI* genes, respectively, that have been proposed to have major impacts on milk composition [56, 59, 60]. 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 [59]. 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 [36], although the gene has recently also been implicated in the modulation of mammary epithelial cell proliferation [61]. The *PAEP* gene encodes the major whey protein  $\beta$ -lactoglobulin. The variant rs109625649 reported here (V134A) is one of the variants that distinguishes the 'A' and 'B' haplotypes of  $\beta$ -lactoglobulin [62]. The *PAEP* gene also exhibited an eQTL that was significantly correlated with wavenumber 2548  $\text{cm}^{-1}$ , which is concordant with previous reports of *PAEP* promoter variants associated with milk composition [63]. The gene *DGATI* 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 *DGATI* 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 *DGATI* gene on milk production, with a recent study showing that these effects

**Table 3** Peak variants for FT-MIR wavenumbers with co-localized eQTL

Chr	Position	Tag variant ID	Iter	Top wvn (cm <sup>-1</sup> )	N 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	YRMYS5A	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  $-\log_{10}(p\text{-values})$  of the trait QTL and the  $-\log_{10}(p\text{-values})$  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. N of hits: number of wavenumbers for which the variant was selected as the representative (most significant) tag variant for a peak

**Table 4** Peak variants for composite milk production traits with co-localized eQTL

Trait	Chr	Position	Tag variant ID	Iteration	P-value	Gene	Top eQTL variant ID	Top eQTL variant p-value	LD	Pearson	Pearson window (Mbp)
FP	3	34387618	rs109030498	2	6.0e-13	<i>ELAPOR1</i>	rs109030498	3.2e-30	1	0.832	0.5
FP	5	75698283	rs385866519	1	4.0e-19	<i>CSF2RB</i>	rs210641868	9.2e-27	0.910	0.701	1
FP	5	93945738	rs211210569	0	6.7e-106	<i>MGST1</i>	rs209372883	3.2e-43	0.919	0.928	1
FP	10	46483019	rs133089336	0	4.5e-13	<i>USP3</i>	rs208181306	2.0e-31	0.909	0.905	0.5
FP	11	104229609	rs110534892	1	2.6e-14	<i>ABO</i>	rs109750996	3.9e-28	0.944	0.727	1
FP	16	67730371	rs380453838	1	2.6e-19	<i>IVNS1ABP</i>	rs380453838	4.5e-27	1	0.886	0.5
FP	27	36211708	rs209855549	0	9.7e-132	<i>GPAT4</i>	rs209855549	3.7e-21	1	0.819	0.5
LP	1	154122887	rs42167460	0	1.2e-50	<i>SH3BP5</i>	rs380642859	2.6e-32	0.999	0.859	0.5
LP	3	15433518	rs109749506	1	1.3e-20	<i>KRTCAP2</i>	rs133285846	9.7e-09	0.995	0.940	1
LP	3	15545091	rs379353107	0	2.2e-42	<i>SLC50A1</i>	rs379353107	8.7e-16	1	0.806	0.5
LP	3	53994057	rs211488591	2	6.7e-18	<i>LRRRC8C</i>	rs466686834	3.5e-39	0.986	0.753	1
LP	19	43036265	rs210324533	3	9.4e-40	<i>GHDC</i>	rs381442991	1.8e-22	0.945	0.963	0.5
LP	19	61134515	rs41923843	1	1.1e-46	<i>KCNJ2</i>	rs41923843	1.7e-26	1	0.857	0.5
LP	20	58448763	rs134813825	0	3.2e-18	<i>ANKH</i>	rs134813825	2.4e-16	1	0.809	0.5
LP	27	36204066	rs208306200	0	1.9e-21	<i>GPAT4</i>	rs208306200	3.7e-21	1	0.767	0.5
LP	29	9577372	rs380473328	0	2.1e-140	<i>PICALM</i>	rs384691767	2.4e-54	0.996	0.845	0.5
PP	3	15520971	rs109098377	2	7.5e-16	<i>KRTCAP2</i>	rs133285846	9.7e-09	0.989	0.928	0.5
PP	3	15550598	rs380597285	0	1.7e-37	<i>SLC50A1</i>	rs380597285	8.7e-16	1	0.832	0.5
PP	5	75680825	rs208925020	4	8.5e-23	<i>CSF2RB</i>	rs210641868	9.2e-27	0.947	0.871	1
PP	5	93945738	rs211210569	1	3.7e-42	<i>MGST1</i>	rs209372883	3.2e-43	0.919	0.817	0.5
PP	6	87387870	rs382652853	2	2.9e-45	<i>CSN3</i>	rs208009847	9.9e-33	0.963	0.891	0.5
PP	10	46581015	rs109326466	0	4.0e-38	<i>USP3</i>	rs109326466	2.0e-31	1	0.961	0.5
PP	18	2203,325	rs135350753	0	2.1e-13	<i>FA2H</i>	rs137235970	1.9e-27	0.997	0.831	0.5
PP	19	43035006	rs209494359	0	1.6e-40	<i>GHDC</i>	rs381442991	1.8e-22	0.945	0.976	0.5
PP	24	58817202	rs208779762	0	5.7e-26	<i>LMAN1</i>	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  $-\log_{10}(p\text{-values})$  of the trait QTL and the  $-\log_{10}(p\text{-values})$  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. FP Fat %, LP Lactose %, PP Protein %

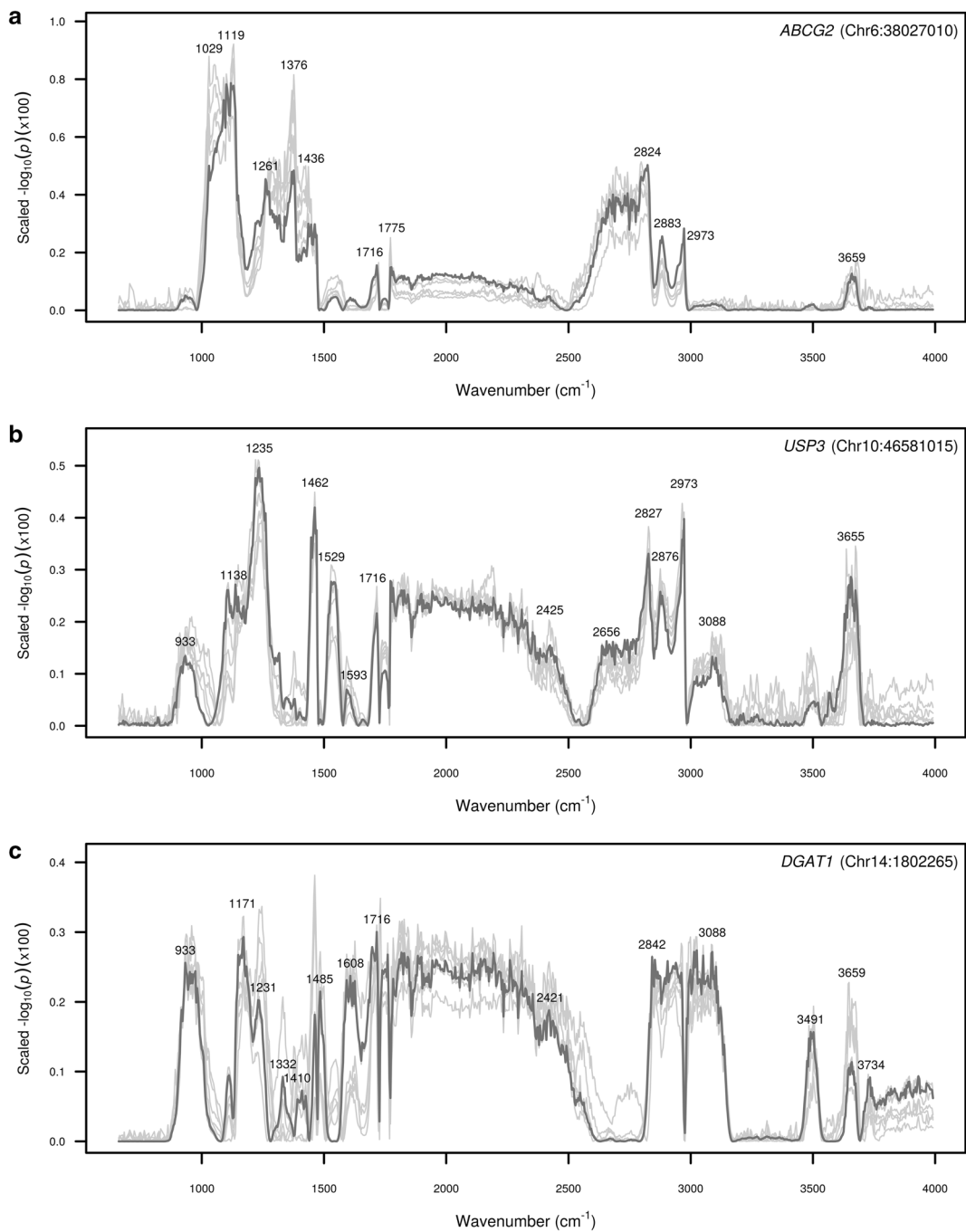
may be due in part to an expression-based mechanism [64].

For the effects observed in the *ABCG2*, *PAEP* and *DGATI* 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) [59], 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 *DGATI* mutation (rs109234250) [56] was 1.5e-2607, whereas the  $p$ -value for the same locus for PP was 1.2e-61.

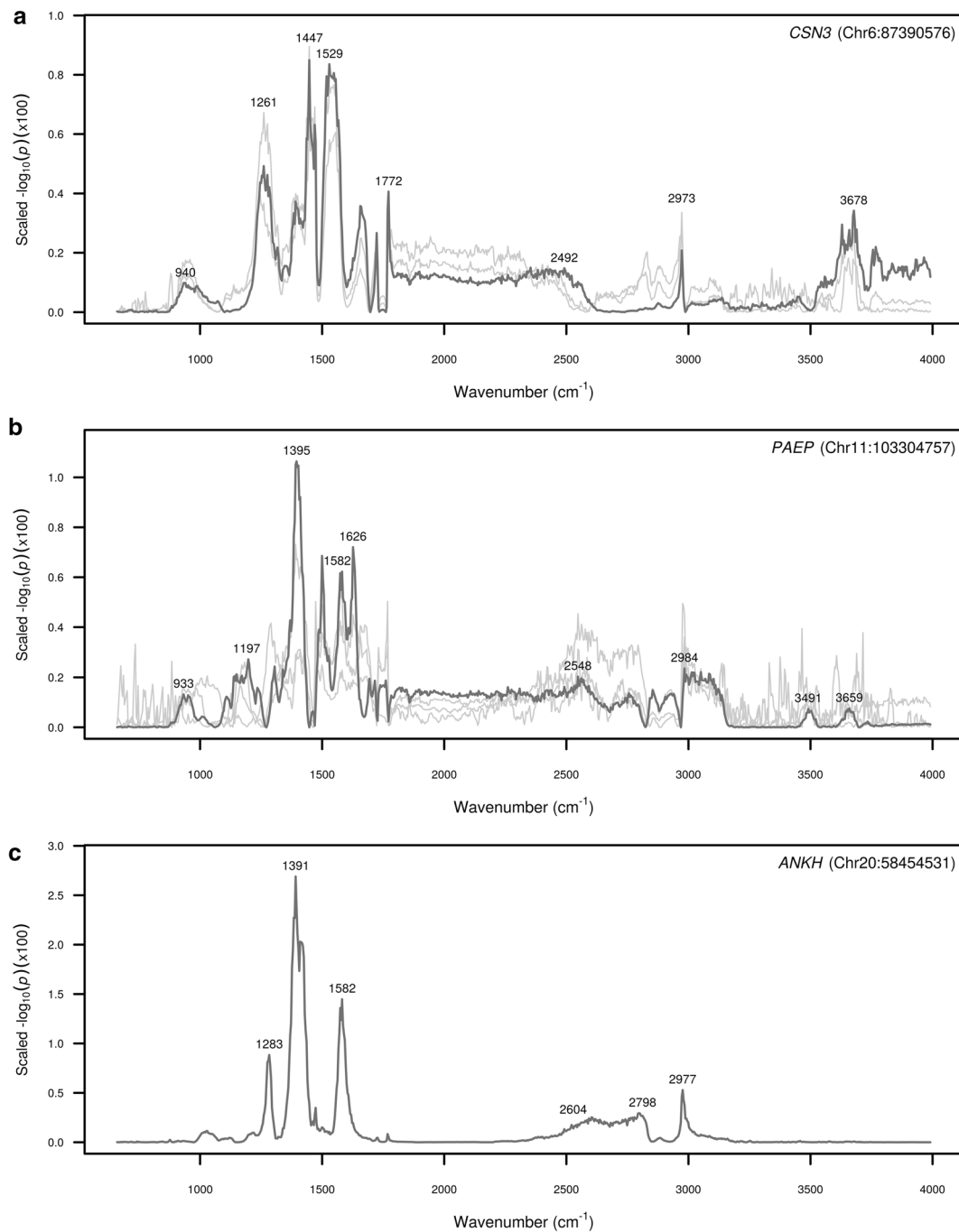
Multiple protein-coding mutations could be attributed to loci with QTL in both wavenumber and milk composition traits, highlighting genes that appear to be novel to the present study (*TGDS* and *DNAJB13*), and genes previously reported in other studies of milk composition traits: *GBA* [37, 65], *MTX1* [66], *SLC50A1* [36, 67], *CSF2RB* [66, 68, 69], *NCF4* [66, 69], *TBC1D22A* [70], *MROH1* [71] and *MTARC1* [36]. 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* [72], *SCAMP3* [66], *THBS3* [66], *CSN1S1*, *CSN2* and *CSN3* [65, 71], *ABO* [73, 74], *CPSF1* [75, 76], *SPAG1* [67], *RNF214* [36], *KAT2A* [36], *STAT5A* [77–79] and *CCDC57* [40, 80]; and QTL highlighting genes that appear novel: *FCRLA*, *WDR97*, *KRT9*, *KRT16*, *KRT17*, *HID1*, *KCNK1* and *NXF1*.







**Fig. 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  $-\log_{10}(p\text{-value})$  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)



**Fig. 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  $-\log_{10}(p\text{-value})$  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 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)

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* [66, 86], *NCF4* [66, 69], *SLC34A2* [87], *TENT5A* [40], *RNF217* [67], *MPC1* [85], *XDH* [88, 89], *PAEP* [60], *DGAT1* [56], *RGL1* [90], *LTF* [91, 92] and *MUS81* [93]. 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.

#### 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  $1462\text{ cm}^{-1}$ . Visualisation of RNA-seq alignments confirmed that this variant disrupts splicing in carrier and homozygous animals (Fig. 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 78<sup>th</sup> residue (due to a +33 bp 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 [94, 95]. 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 [96]. 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 [97]. 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 [73, 74]. 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 [74]. 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 [72]. 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  $1462\text{ cm}^{-1}$  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  $1462\text{ cm}^{-1}$ , 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. [73] is in strong LD with the lead variant representing a strong *ABO* eQTL highlighted in our study ( $R^2 = 0.944$ ; Table 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.

#### 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 *DGATI*, *GPAT4*, and *MGST1* (Fig. 5). These three loci have been implicated in previous studies of milk fat percentage and fatty acid synthesis [35, 56, 57, 70, 83, 84]. *DGATI* 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 *DGATI* (Fig. 6a) was very similar to those reported previously [32, 43]. Highly significant effects were observed for the *DGATI* K232A polymorphism in bands of the spectrum that could be attributed to a number of different chemical bond interactions including: phosphorus compounds (from  $\sim 910$  to  $1010\text{ cm}^{-1}$ ) [98], triglyceride ester C–O stretching (from  $\sim 1,130$  to  $1260\text{ cm}^{-1}$ ) [99, 100], C–H bending vibrations of  $-\text{CH}_2$  and  $-\text{CH}_3$  (from  $\sim 1450$  to  $1500\text{ cm}^{-1}$ ) [45, 98], C=O stretching in polypeptides within the amide I band of protein (from  $\sim 1600$  to  $1700\text{ cm}^{-1}$ ) [99], carboxylic acid and C=O rotation and stretching of ester groups of fat (from  $\sim 1700$  to  $1800\text{ cm}^{-1}$ ) [101], and acyl chain C–H stretching (from  $\sim 2820$  to  $3150\text{ cm}^{-1}$ ) [100].

The cluster that included the *ABCG2* Y581S polymorphism (Fig. 5) had highly significant association effects across numerous FT-MIR wavenumbers, with the largest effects concentrated within the regions from  $\sim 1020$

to  $1470\text{ cm}^{-1}$  and from  $\sim 2610$  to  $2980\text{ cm}^{-1}$  (Fig. 6b). 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  $\sim 1020$  to  $1180\text{ cm}^{-1}$ ) [98, 102], amide III and phosphate bands (from  $\sim 1200$  to  $1390\text{ cm}^{-1}$ ) [99, 103], C–H bending vibrations for  $\text{CH}_2$  and  $-\text{CH}_3$  (from  $\sim 1410$  to  $1470\text{ cm}^{-1}$ ) [98], overtones and bands of lactose ( $\sim 2600$  upwards) [104], and C–H stretching vibrations of  $\text{CH}_2$  and  $-\text{CH}_3$  (from  $\sim 2700$  to  $2980\text{ cm}^{-1}$ ) [98]. 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 [36, 71]. Notably, the strongest association effects for the *ABCG2* Y581S polymorphism were in different bands of the mid-infrared spectrum to the *DGATI* 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. 7), which had a large proportion of significant signal captured within a small part of the mid-infrared range: *CSN3* (from  $\sim 1220$  to  $1780\text{ cm}^{-1}$ ), *PAEP* (from  $\sim 1350$  to  $1650\text{ cm}^{-1}$ ) and *ANKH* (from  $\sim 1260$  to  $1620\text{ cm}^{-1}$ ). The *CSN3* gene encodes  $\kappa$  casein, one of the most abundantly expressed proteins in milk. Bound at the aqueous-hydrophobic interface of casein micelles,  $\kappa$  casein content influences the size of these structures, thereby affecting various coagulation and cheese-making properties [105, 106]. The missense mutation reported here at chr6:87390576 (rs43703015) has been associated with milk composition traits and differential expression in mammary tissue [107]. The largest effects for the *CSN3* locus were in spectral regions related to amide III and phosphate bands (from  $\sim 1220$  to  $1320\text{ cm}^{-1}$ ), C–H stretching vibrations of  $\text{CH}_2$  and  $-\text{CH}_3$  (from  $\sim 1370$  to  $1480\text{ cm}^{-1}$ ), and N–H bending and C–N stretching in the amide II band (from  $\sim 1490$  to  $1590\text{ cm}^{-1}$ ) [108]. 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 [32, 42, 43]. The *ANKH* gene encodes a transmembrane protein involved in pyrophosphate transport regulation, and is associated with lactose concentrations in milk [36, 71]. Interestingly, *ANKH* and *PAEP* shared a prominent peak for adjacent wavenumbers,  $1391\text{ cm}^{-1}$  and  $1395\text{ cm}^{-1}$ , respectively. These wavenumbers were in a region related to carboxylic acid C=O bond stretching [98]. Another peak in common between these genes was centred on the  $1582\text{ cm}^{-1}$



wavenumber, also in a region related to carboxylic acid C=O bond stretching [98]. Association effects in similar bands of the mid-infrared spectrum for *PAEP* have been reported in previous studies [32, 42, 43]. 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 ~1270 to 1290  $\text{cm}^{-1}$ ) [99, 103], 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  $\text{NH}_2$  (from ~1600 to 1640  $\text{cm}^{-1}$ ) [98, 109], which shows that although commonalities exist, there are also differences in the roles that these genes play in altering milk composition.

#### 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 [110] 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 [93]. 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 nonsense) 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 [19].

#### 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.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12711-021-00648-9>.

**Additional file 1: Figure S1.** Sequence resolution effects for highly significant wavenumber QTL. Effects shown for 14 base GWAS wavenumber QTL in high LD ( $R^2 > 0.9$ ) with a putative impact variant. Putative impact variants are defined as a splice region variant, or a moderate or high impact variant according to the SnpEff classification. 1-Mbp regions centred on the wavenumber QTL are shown. The x-axis represents positions on the UMD 3.1 *Bos taurus* reference genome; the y-axis shows the strength of association signal, represented as the  $-\log_{10}(p\text{-value})$  of the effect for each variant. Effects are coloured based on the predicted effect of the variant on genes, according to the SnpEff classification. The horizontal red line shows the Bonferroni significance threshold of  $-\log_{10}(6.2e-13)$ .

**Additional file 2: Table S2.** 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  $-\log_{10}(p\text{-value})$  of the effect was greater than 1x 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 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 1) are shaded yellow. No. of hits is the number of wavenumbers for which the variant was selected as the representative (most significant) tag variant for a peak. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0. L = Low impact splice region variant; M = Moderate impact missense variant; H = High impact splice donor. **Table S3.** Minor allele frequencies and allele effects for WGS tag variants with a highly significant protein-sequence association effect in at least one FT-MIR wavenumber. **Table S4.** Association statistic profiles for 31 highly significant protein-sequence effects identified in FT-MIR wavenumber phenotypes. For each protein-sequence mutation, the proportion of phenotypic and genetic variance that it accounts for is shown for each of the 895 FT-MIR wavenumbers and three FT-MIR predicted milk composition phenotypes. The genetic variance for each phenotype is the SNP-based estimate evaluated by the Bolt-LMM software. **Table S5.** Chronological profiles across iterations for 17 highly significant protein-sequence association effects observed in FT-MIR wavenumbers, where the association is observed after fitting the top chromosomal variant(s) in previous GWAS iterations and/or the base GWAS as covariates. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0.  $P$ -values at previous iterations for the phenotype, and  $p$ -values for the corresponding top chromosomal SNP in that iteration are provided. **Table S6.** 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  $-\log_{10}(p\text{-value})$  of the effect was greater than 1x 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 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 2) are shaded yellow. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0. FP = Fat %; LP = Lactose %; PP = Protein %; L = Low impact splice region variant; M = Moderate impact missense variant; H = High impact splice donor. **Table S7.** Minor allele frequencies and allele effects for WGS tag variants with a highly significant protein-sequence association effect in at least one FT-MIR predicted milk composition trait. FP = Fat %; LP = Lactose %; PP = Protein % **Table S8.** Chronological profiles across iterations for highly significant protein-sequence association effects observed in FT-MIR predicted milk composition traits, where the association is observed after fitting the top chromosomal variant(s) in previous GWAS iterations and/or the base GWAS as covariates. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0.  $P$ -values at previous iterations for the phenotype, and  $p$ -values for the corresponding top chromosomal SNP in that iteration have been provided. FP = Fat %; LP = Lactose %; PP = Protein %.

**Additional file 3: Figure S9.** Sequence resolution effects for 19 base GWAS wavenumber QTL with a co-localized expression QTL. 1-Mbp regions centred on the wavenumber QTL are shown. The x-axis represents positions on the UMD 3.1 *Bos taurus* reference genome; the y-axis shows the strength of association signal, represented as the  $-\log_{10}(p\text{-value})$  of the effect for each variant. Effects are coloured based on the predicted effect of the variant on genes, according to the SnpEff classification. The horizontal red line shows the Bonferroni significance threshold of  $-\log_{10}(6.2e-13)$ .

**Additional file 4: Table S10.** Minor allele frequencies and allele effects for WGS tag variants with a significant association effect in FT-MIR wavenumbers and a co-localized eQTL. **Table S11.** Association statistic profiles for 38 loci with a co-localized eQTL for at least one FT-MIR wavenumber phenotype. The proportion of phenotypic and genetic variance accounted for by each locus is shown for each of the 895 FT-MIR wavenumber and three FT-MIR predicted milk composition phenotypes. The genetic variance for each phenotype is the SNP-based estimate evaluated by the Bolt-LMM software. **Table S12.** Chronological profiles across iterations for 19 significant association effects with a co-localized eQTL observed in FT-MIR wavenumbers, where the association is observed after fitting the top chromosomal variant(s) in previous GWAS iterations and/or the base GWAS as covariates. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0.  $P$ -values at previous iterations for the phenotype, and  $p$ -values for the corresponding top chromosomal SNP in that iteration are provided. **Table S13.** Minor allele frequencies and allele effects for WGS tag variants with a significant association effect in at least one FT-MIR predicted milk composition trait and a co-localized eQTL. FP = Fat %; LP = Lactose %; PP = Protein % **Table S14.** Chronological profiles across iterations for 12 significant association effects with a co-localized eQTL observed in FT-MIR predicted milk composition traits, where the association is observed after fitting the top chromosomal variant(s) in previous GWAS iterations and/or the base GWAS as covariates. Iterations are defined relative to the base GWAS, with the base GWAS represented as iteration 0.  $P$ -values at previous iterations for the phenotype, and  $p$ -values for the corresponding top chromosomal SNP in that iteration have been provided. FP = Fat %; LP = Lactose %; PP = Protein %.

**Additional file 5: Figure S15.** Significance profiles of associations between FT-MIR wavenumbers and loci/genes in high LD with a putative impact variant (PIV), or in high LD with the top variant of a co-localized eQTL. A PIV is defined as a splice region variant, or moderate or high impact variant, according to the SnpEff classification. Significance is expressed as the  $-\log_{10}(p\text{-value})$  between each FT-MIR wavenumber and locus/gene of interest.

## 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.

#### Authors' contributions

KMT performed GWAS with assistance and guidance from ER, TJL and MDL; RGS, MK and TJJJ provided genotypes and sequence datasets; KMT carried out sequence imputation with guidance from RGS; TJL provided gene expression and eQTL datasets; SRD assisted with interpretation of milk chemistry results; JEP, RJS, BLH, MDL and DJG were involved in supervising the project; KMT, TJL and MDL wrote the manuscript with input and critique from all other authors. All authors have read and approved the final manuscript.

#### 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.

#### 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 (<https://doi.org/10.5061/dryad.qrfj6q5dj>) [111]. 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 [112]. Whole-genome sequences used for imputation of the genotypes presented in this paper have been deposited in the SRA database [113]. Additional data is available on reasonable request with the permission of Livestock Improvement Corporation, contingent on the execution of an appropriate transfer agreement.

#### Declarations

##### Ethics approval and consent to participate

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.

##### Consent for publication

Not applicable.

##### Competing interests

KMT, TJL, RGS, TJJJ, SRD, RJS, BLH, MDL are employees of Livestock Improvement Corporation (LIC; Hamilton, New Zealand), a commercial provider of bovine germplasm. JEP is an employee of Agriculture Victoria (Bundoora, Australia). All other authors declare that they have no competing interests.

##### Author details

<sup>1</sup>Research and Development, Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand. <sup>2</sup>School of Agriculture, Massey University, Ruakura, Hamilton 3240, New Zealand. <sup>3</sup>School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia. <sup>4</sup>Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC 3083, Australia.

Received: 4 December 2020 Accepted: 22 June 2021

Published online: 20 July 2021

#### References

- De Marchi M, Toffanin V, Cassandro M, Penasa M. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. *J Dairy Sci.* 2014;97:1171–86.
- De Marchi M, Penasa M, Zidi A, Manuelian CL. Invited review: use of infrared technologies for the assessment of dairy products—applications and perspectives. *J Dairy Sci.* 2018;101:10589–604.
- Egger-Danner C, Cole JB, Pryce JE, Gengler N, Heringstad B, Bradley A, et al. Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. *Animal.* 2015;9:191–207.
- Gengler N, Soyeurt H, Dehareng F, Bastin C, Colinet F, Hammami H, et al. Capitalizing on fine milk composition for breeding and management of dairy cows. *J Dairy Sci.* 2016;99:4071–9.
- Toffanin V, De Marchi M, Lopez-Villalobos N, Cassandro M. 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. *Int Dairy J.* 2015;41:68–73.
- Visentin G, McDermott A, McParland S, Berry DP, Kenny OA, Brodtkorb A, et al. Prediction of bovine milk technological traits from mid-infrared spectroscopy analysis in dairy cows. *J Dairy Sci.* 2015;98:6620–9.
- Visentin G, Penasa M, Niero G, Cassandro M, Marchi MD. Phenotypic characterisation of major mineral composition predicted by mid-infrared spectroscopy in cow milk. *Ital J Anim Sci.* 2018;17:549–56.
- Bonfatti V, Tiezzi F, Miglior F, Carnier P. Comparison of Bayesian regression models and partial least squares regression for the development of infrared prediction equations. *J Dairy Sci.* 2017;100:7306–19.
- Sanchez MP, Ferrand M, Gelé M, Pourchet D, Miranda G, Martin P, et al. Short communication: genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande, and Holstein dairy cattle breeds. *J Dairy Sci.* 2017;100:6371–5.
- McParland S, Kennedy E, Lewis E, Moore SG, McCarthy B, O'Donovan M, et al. Genetic parameters of dairy cow energy intake and body energy status predicted using mid-infrared spectrometry of milk. *J Dairy Sci.* 2015;98:1310–20.
- Luke TDW, Rochfort S, Wales WJ, Bonfatti V, Maret L, Pryce JE. Metabolic profiling of early-lactation dairy cows using milk mid-infrared spectra. *J Dairy Sci.* 2019;102:1747–60.
- Lainé A, Bastin C, Grelet C, Hammami H, Colinet FG, Dale LM, et al. Assessing the effect of pregnancy stage on milk composition of dairy cows using mid-infrared spectra. *J Dairy Sci.* 2017;100:2863–76.
- Toledo-Alvarado H, Vazquez AI, De Los CG, Tempelman RJ, Bittante G, Cecchinato A. Diagnosing pregnancy status using infrared spectra and milk composition in dairy cows. *J Dairy Sci.* 2018;101:2496–505.
- Ho PN, Bonfatti V, Luke TDW, Pryce JE. Classifying the fertility of dairy cows using milk mid-infrared spectroscopy. *J Dairy Sci.* 2019;102:10460–70.
- Bittante G, Cipolat-Gotet C. 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. *J Dairy Sci.* 2018;101:7219–35.
- van Gastelen S, Mollenhorst H, Antunes-Fernandes EC, Hetingting KA, van Burgsteden GG, Dijkstra J, et al. Predicting enteric methane emission of dairy cows with milk Fourier-transform infrared spectra and gas chromatography-based milk fatty acid profiles. *J Dairy Sci.* 2018;101:5582–98.
- Vanlierde A, Soyeurt H, Gengler N, Colinet FG, Froidmont E, Kreuzer M, et al. 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. *J Dairy Sci.* 2018;101:7618–24.
- Denholm SJ, Brand W, Mitchell AP, Wells AT, Krzyzelewski T, Smith SL, et al. Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning. *J Dairy Sci.* 2020;103:9355–67.
- Tiplady KM, Lopdell TJ, Littlejohn MD, Garrick DJ. The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle. *J Anim Sci Biotechnol.* 2020;11:39.
- Rutten MJM, Bovenhuis H, van Arendonk JAM. The effect of the number of observations used for Fourier transform infrared model calibration for bovine milk fat composition on the estimated genetic parameters of the predicted data. *J Dairy Sci.* 2010;93:4872–82.
- Lopez-Villalobos N, Spelman RJ, Melis J, Davis SR, Berry SD, Lehnert K, et al. Estimation of genetic and crossbreeding parameters of fatty acid concentrations in milk fat predicted by mid-infrared spectroscopy in New Zealand dairy cattle. *J Dairy Res.* 2014;81:340–9.

22. Hein L, Sørensen LP, Kargo M, Buitenhuis AJ. Genetic analysis of predicted fatty acid profiles of milk from Danish Holstein and Danish Jersey cattle populations. *J Dairy Sci.* 2018;101:2148–57.
23. Bonfatti V, Vicario D, Lugo A, Carnier P. Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle. *J Dairy Sci.* 2017;100:5526–40.
24. Cecchinato A, Albera A, Cipolat-Gobet C, Ferragina A, Bittante G. 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. *J Dairy Sci.* 2015;98:4914–27.
25. Poulsen NA, Buitenhuis AJ, Larsen LB. Phenotypic and genetic associations of milk traits with milk coagulation properties. *J Dairy Sci.* 2015;98:2079–87.
26. Visentin G, McParland S, De Marchi M, McDermott A, Fenelon MA, Penasa M, et al. Processing characteristics of dairy cow milk are moderately heritable. *J Dairy Sci.* 2017;100:6343–55.
27. Soyeurt H, Colinet FG, Arnould VM-R, Dardenne P, Bertozzi C, Renaville R, et al. Genetic variability of lactoferrin content estimated by mid-infrared spectrometry in bovine milk. *J Dairy Sci.* 2007;90:4443–50.
28. Lopez-Villalobos N, Davis S, Beattie EM, Melis J, Berry S, Holroyd S, et al. Breed effects for lactoferrin concentration determined by Fourier transform infrared spectroscopy. *Proc NZ Soc Anim Prod.* 2009;69:60–4.
29. Bittante G, Cecchinato A. Genetic analysis of the Fourier-transform infrared spectra of bovine milk with emphasis on individual wavelengths related to specific chemical bonds. *J Dairy Sci.* 2013;96:5991–6006.
30. Rovere G, de los Campos G, Tempelman RJ, Vazquez AI, Miglior F, Schenkel F, et al. A landscape of the heritability of Fourier-transform infrared spectral wavelengths of milk samples by parity and lactation stage in Holstein cows. *J Dairy Sci.* 2019;102:1354–63.
31. Soyeurt H, Misztal I, Gengler N. Genetic variability of milk components based on mid-infrared spectral data. *J Dairy Sci.* 2010;93:1722–8.
32. Wang Q, Hulzebosch A, Bovenhuis H. Genetic and environmental variation in bovine milk infrared spectra. *J Dairy Sci.* 2016;99:6793–803.
33. Jiang L, Liu J, Sun D, Ma P, Ding X, Yu Y, et al. Genome wide association studies for milk production traits in Chinese Holstein population. *PLoS One.* 2010;5:e13661.
34. Kemper KE, Reich CM, Bowman PJ, van der Jagt CJ, Chamberlain AJ, Mason BA, et al. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genet Sel Evol.* 2015;47:29.
35. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, et al. Sequence-based association analysis reveals an *MGST1* eQTL with pleiotropic effects on bovine milk composition. *Sci Rep.* 2016;6:25376.
36. Lopdell TJ, Tiplady K, Struchalin M, Johnson TJJ, Keehan M, Sherlock R, et al. DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. *BMC Genomics.* 2017;18:968.
37. Raven L-A, Cocks BG, Hayes BJ. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics.* 2014;15:62.
38. Bouwman AC, Bovenhuis H, Visker MHPW, van Arendonk JAM. Genome-wide association of milk fatty acids in Dutch dairy cattle. *BMC Genet.* 2011;12:43.
39. Buitenhuis B, Poulsen NA, Gebreyesus G, Larsen LB. Estimation of genetic parameters and detection of chromosomal regions affecting the major milk proteins and their post translational modifications in Danish Holstein and Danish Jersey cattle. *BMC Genet.* 2016;17:114.
40. Li C, Sun D, Zhang S, Wang S, Wu X, Zhang Q, et al. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. *PLoS One.* 2014;9:e96186.
41. Wang Q, Bovenhuis H. Genome-wide association study for milk infrared wavenumbers. *J Dairy Sci.* 2018;101:2260–72.
42. Benedet A, Ho PN, Xiang R, Bolormaa S, Marchi MD, Goddard ME, et al. The use of mid-infrared spectra to map genes affecting milk composition. *J Dairy Sci.* 2019;102:7189–203.
43. Zaalberg RM, Janss L, Buitenhuis AJ. Genome-wide association study on Fourier transform infrared milk spectra for two Danish dairy cattle breeds. *BMC Genet.* 2020;21:9.
44. Tiplady KM, Sherlock RG, Littlejohn MD, Pryce JE, Davis SR, Garrick DJ, et al. Strategies for noise reduction and standardization of milk mid-infrared spectra from dairy cattle. *J Dairy Sci.* 2019;102:6357–72.
45. Grelet C, Pierna JAF, Dardenne P, Baeten V, Dehareng F. Standardization of milk mid-infrared spectra from a European dairy network. *J Dairy Sci.* 2015;98:2150–60.
46. Butler DG, Cullis BR, Gilmour AR, Gogel BJ. ASReml-R reference manual: mixed models for S language environments. Version 3. Brisbane, Australia, Queensland Department of Primary Industries and Fisheries, NSW Department of Primary Industries; 2009.
47. Jivanji S, Worth G, Lopdell TJ, Yeates A, Couldrey C, Reynolds E, et al. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. *Genet Sel Evol.* 2019;51:62.
48. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754–60.
49. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43:491–8.
50. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 2007;81:1084–97.
51. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet.* 2009;84:210–23.
52. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nat Genet.* 2015;47:284–90.
53. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013;14:R36.
54. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6:80–92.
55. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2020. <https://www.R-project.org/>. Accessed 22 Jul 2020.
56. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, et al. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. *Genome Res.* 2002;12:222–31.
57. Schennink A, Stoop WM, Visker MHPW, Heck JML, Bovenhuis H, van der Poel JJ, et al. *DGAT1* underlies large genetic variation in milk-fat composition of dairy cows. *Anim Genet.* 2007;38:467–73.
58. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science.* 2012;335:823–8.
59. Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee J-H, et al. 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 Res.* 2005;15:936–44.
60. Ganai NA, Bovenhuis H, van Arendonk JAM, Visker MHPW. Novel polymorphisms in the bovine *beta-lactoglobulin* gene and their effects on beta-lactoglobulin protein concentration in milk. *Anim Genet.* 2009;40:127–33.
61. Wei J, Geale PF, Sheehy PA, Williamson P. The impact of *ABCG2* on bovine mammary epithelial cell proliferation. *Anim Biotechnol.* 2012;23:221–4.
62. Caroli AM, Chessa S, Erhardt GJ. Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. *J Dairy Sci.* 2009;92:5335–52.
63. Zakizadeh S, Reissmann M, Miraei-Ashtiani SR, Reinecke P. Polymorphism of beta-lactoglobulin coding and 5'-flanking regions and association with milk production traits. *Biotechnol Biotechnol Equip.* 2012;26:2716–21.
64. Fink T, Lopdell TJ, Tiplady K, Handley R, Johnson TJJ, Spelman RJ, et al. A new mechanism for a familiar mutation—bovine *DGAT1 K232A* modulates gene expression through multi-junction exon splice enhancement. *BMC Genomics.* 2020;21:591.



65. Jiang J, Ma L, Prakapenka D, VanRaden PM, Cole JB, Da Y. A large-scale genome-wide association study in U.S. Holstein cattle. *Front Genet.* 2019;10:412.
66. Raven L-A, Cocks BG, Kemper KE, Chamberlain AJ, Vander Jagt CJ, Goddard ME, et al. Targeted imputation of sequence variants and gene expression profiling identifies twelve candidate genes associated with lactation volume, composition and calving interval in dairy cattle. *Mamm Genome.* 2016;27:81–97.
67. Jiang J, Cole JB, Da Y, VanRaden PM, Ma L. Fast Bayesian fine-mapping of 35 production, reproduction and body conformation traits with imputed sequences of 27K Holstein bulls. *bioRxiv.* 2018. <https://doi.org/10.1101/428227>.
68. Kemper KE, Hayes BJ, Daetwyler HD, Goddard ME. How old are quantitative trait loci and how widely do they segregate? *J Anim Breed Genet.* 2015;132:121–34.
69. Lopdell TJ, Tiplady K, Couldrey C, Johnson TJJ, Keehan M, Davis SR, et al. Multiple QTL underlie milk phenotypes at the *CSF2RB* locus. *Genet Sel Evol.* 2019;51:3.
70. Pausch H, Emmerling R, Gredler-Grandl B, Fries R, Daetwyler HD, Goddard ME. 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.* 2017;18:853.
71. Sanchez M-P, Govignon-Gion A, Croiseau P, Fritz S, Hozé C, Miranda G, et al. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genet Sel Evol.* 2017;49:68.
72. Jiang J, Liu L, Gao Y, Shi L, Li Y, Liang W, et al. Determination of genetic associations between indels in 11 candidate genes and milk composition traits in Chinese Holstein population. *BMC Genet.* 2019;20:48.
73. Liu Z, Wang T, Pryce JE, MacLeod IM, Hayes BJ, Chamberlain AJ, et al. Fine-mapping sequence mutations with a major effect on oligosaccharide content in bovine milk. *Sci Rep.* 2019;9:2137.
74. Poulsen NA, Robinson RC, Barile D, Larsen LB, Buitenhuis B. A genome-wide association study reveals specific transferases as candidate loci for bovine milk oligosaccharides synthesis. *BMC Genomics.* 2019;20:404.
75. Cochran SD, Cole JB, Null DJ, Hansen PJ. Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. *BMC Genet.* 2013;14:49.
76. Buitenhuis B, Janss LLG, Poulsen NA, Larsen LB, Larsen MK, Sørensen P. Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. *BMC Genomics.* 2014;15:1112.
77. Brym P, Kamiński S, Ruś A. New SSCP polymorphism within bovine *STAT5A* gene and its associations with milk performance traits in Black-and-White and Jersey cattle. *J Appl Genet.* 2004;45:445–52.
78. Schennink A, Bovenhuis H, Léon-Kloosterziel KM, van Arendonk JAM, Visker MHPW. Effect of polymorphisms in the *FASN*, *OLR1*, *PPARGC1A*, *PRL* and *STAT5A* genes on bovine milk-fat composition. *Anim Genet.* 2009;40:909–16.
79. He X, Chu MX, Qiao L, He JN, Wang PQ, Feng T, et al. Polymorphisms of *STAT5A* gene and their association with milk production traits in Holstein cows. *Mol Biol Rep.* 2012;39:2901–7.
80. Bouwman AC, Visker MHPW, van Arendonk Johan AM, Bovenhuis H. Fine mapping of a quantitative trait locus for bovine milk fat composition on *Bos taurus* autosome 19. *J Dairy Sci.* 2014;97:1139–49.
81. Fang M, Fu W, Jiang D, Zhang Q, Sun D, Ding X, et al. A multiple-SNP approach for genome-wide association study of milk production traits in Chinese Holstein cattle. *PLoS One.* 2014;9:e99544.
82. Wang D, Ning C, Liu J-F, Zhang Q, Jiang L. Short communication: Replication of genome-wide association studies for milk production traits in Chinese Holstein by an efficient rotated linear mixed model. *J Dairy Sci.* 2019;102:2378–83.
83. Wang X, Wurmser C, Pausch H, Jung S, Reinhardt F, Tetens J, et al. Identification and dissection of four major QTL affecting milk fat content in the German Holstein-Friesian population. *PLoS One.* 2012;7:e40711.
84. Littlejohn MD, Tiplady K, Lopdell T, Law TA, Scott A, Harland C, et al. Expression variants of the lipogenic *AGPAT6* gene affect diverse milk composition phenotypes in *Bos taurus*. *PLoS One.* 2014;9:e85757.
85. Sanchez M-P, Ramayo-Caldas Y, Wolf V, Laithier C, El Jabri M, Michenet A, et al. Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows. *Genet Sel Evol.* 2019;51:34.
86. Kemper KE, Littlejohn MD, Lopdell T, Hayes BJ, Bennett LE, Williams RP, et al. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. *BMC Genomics.* 2016;17:858.
87. Liu R, Sun D, Wang Y, Yu Y, Zhang Y, Chen H, et al. Fine mapping QTLs affecting milk production traits on BTA6 in Chinese Holstein with SNP markers. *J Integr Agric.* 2013;12:110–7.
88. Ogorevc J, Kunej T, Razzpet A, Dovc P. Database of cattle candidate genes and genetic markers for milk production and mastitis. *Anim Genet.* 2009;40:832–51.
89. Pegolo S, Cecchinato A, Mele M, Conte G, Schiavon S, Bittante G. Effects of candidate gene polymorphisms on the detailed fatty acids profile determined by gas chromatography in bovine milk. *J Dairy Sci.* 2016;99:4558–73.
90. Yodklaew P, Koonawootrittriron S, Elzo MA, Suwanasopee T, Laodim T. Genome-wide association study for lactation characteristics, milk yield and age at first calving in a Thai multibreed dairy cattle population. *Agric Nat Resour.* 2017;51:223–30.
91. Mao Y, Zhu X, Xing S, Zhang M, Zhang H, Wang X, et al. Polymorphisms in the promoter region of the bovine lactoferrin gene influence milk somatic cell score and milk production traits in Chinese Holstein cows. *Res Vet Sci.* 2015;103:107–12.
92. Viale E, Tiezzi F, Maretto F, De Marchi M, Penasa M, Cassandro M. Association of candidate gene polymorphisms with milk technological traits, yield, composition, and somatic cell score in Italian Holstein-Friesian sires. *J Dairy Sci.* 2017;100:7271–81.
93. Reynolds EGM, Neeley C, Lopdell TJ, Keehan M, Dittmer K, Harland CS, et al. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. *Nat Genet.* 2021. <https://doi.org/10.1038/s41588-021-00872-5>.
94. Yamamoto F, Clausen H, White T, Marken J, Hakomori S. Molecular genetic basis of the histo-blood group ABO system. *Nature.* 1990;345:229–33.
95. Kermarrec N, Roubinet F, Apoil P-A, Blancher A. Comparison of allele O sequences of the human and non-human primate ABO system. *Immunogenetics.* 1999;49:517–26.
96. Chester MA, Olsson ML. The ABO blood group gene: a locus of considerable genetic diversity. *Transfus Med Rev.* 2001;15:177–200.
97. Le Pendu J. Histo-blood group antigen and human milk oligosaccharides. *Adv Exp Med Biol.* 2004;554:135–43.
98. Fleming I, Williams D. Infrared and Raman spectra. In: Fleming I, Williams D, editors. *Spectroscopic methods in organic chemistry.* Cham: Springer International Publishing; 2019. p. 85–121.
99. Safar M, Bertrand D, Robert P, Devaux MF, Genot C. Characterization of edible oils, butters and margarines by Fourier transform infrared spectroscopy with attenuated total reflectance. *J Am Oil Chem Soc.* 1994;71:371.
100. Karoui R, Mazerolles G, Dufour É. Spectroscopic techniques coupled with chemometric tools for structure and texture determinations in dairy products. *Int Dairy J.* 2003;13:607–20.
101. Lefier D, Grappin R, Pochet S. Determination of fat, protein, and lactose in raw milk by Fourier transform infrared spectroscopy and by analysis with a conventional filter-based milk analyzer. *J AOAC Int.* 1996;79:711–7.
102. Picque D, Lefier D, Grappin R, Corrieu G. Monitoring of fermentation by infrared spectrometry: Alcoholic and lactic fermentations. *Anal Chim Acta.* 1993;279:67–72.
103. Hewavitharana AK, van Brakel B. Fourier transform infrared spectrometric method for the rapid determination of casein in raw milk. *Analyst.* 1997;122:701–4.
104. Luinge HJ, Hop E, Lutz ETG, van Hemert JA, de Jong EAM. Determination of the fat, protein and lactose content of milk using Fourier transform infrared spectrometry. *Anal Chim Acta.* 1993;284:419–33.
105. Creamer LK, Plowman JE, Liddell MJ, Smith MH, Hill JP. Micelle stability: kappa-casein structure and function. *J Dairy Sci.* 1998;81:3004–12.
106. Poulsen NA, Bertelsen HP, Jensen HB, Gustavsson F, Glantz M, Månsson HL, et al. The occurrence of noncoagulating milk and the association of bovine milk coagulation properties with genetic variants of the caseins in 3 Scandinavian dairy breeds. *J Dairy Sci.* 2013;96:4830–42.
107. MacLeod I, Bowman P, Vander Jagt C, Haile-Mariam M, Kemper K, Chamberlain A, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics.* 2016;17:144.



108. Garidel P, Schott H. Fourier-transform midinfrared spectroscopy for analysis and screening of liquid protein formulations. Part 1: understanding infrared spectroscopy of proteins. *BioProcess Int.* 2006;4:40–6.
109. Dufour É. Chapter 1—Principles of infrared spectroscopy. In: Sun DW, editor. *Infrared spectroscopy for food quality analysis and control.* San Diego: Academic Press; 2009. p. 1–27.
110. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience.* 2020;9:giaa021.
111. Tiplady KM, Lopdell TJ, Reynolds E, Sherlock RG, Keehan M, Johnson TJJ, et al. Data from: Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. *Dryad Digital Repository*; 2021. <https://doi.org/10.5061/dryad.qrfj6q5dj>. Accessed 10 Feb 2021.
112. Lopdell TJ, Tiplady K, Struchalin M, Johnson TJJ, Keehan M, Sherlock R, et al. 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>. Accessed 28 Nov 2018.
113. PRJNA656361 Cattle whole genome sequences. 2021. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA656361>. Accessed 11 Aug 2020.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

