J. Dairy Sci. 107:9615–9627
https://doi.org/10.3168/jds.2023-24319

Exploring a Bayesian sparse factor model-based strategy for the genetic analysis of thousands of mid-infrared spectra traits for animal breeding

Yansen Chen,[1]* Hadi Atashi,[1,2] Jiayi Qu,[3] Pauline Delhez,[1] Daniel Runcie,[4] Hélène Soyeurt,[1] and Nicolas Gengler[1]
[1]TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium
[2]Department of Animal Science, Shiraz University, 71441-13131 Shiraz, Iran
[3]Department of Animal Science, University of California Davis, Davis, CA 95616
[4]Department of Plant Sciences, University of California Davis, Davis, CA 95616

ABSTRACT

With the rapid development of animal phenomics and deep phenotyping, we can obtain thousands of traditional (but also molecular) phenotypes per individual. However, there is still a lack of exploration regarding how to handle this huge amount of data in the context of animal breeding, presenting a challenge that we are likely to encounter more and more in the future. This study aimed to (1) explore the use of the mega-scale linear mixed model (MegaLMM), a factor model-based approach that is able to simultaneously estimate (co)variance components and genetic parameters in the context of thousands of milk traits, hereafter called thousand-trait (TT) models; (2) compare the phenotype values and genomic breeding value (**u**) predictions for focal traits (i.e., traits that are targeted for prediction, compared with secondary traits that are helping to evaluate), from single-trait (ST) and TT models, respectively; (3) propose a new approximate method of GEBV (**U**) prediction with TT models and MegaLMM. We used a total of 3,421 milk mid-infrared (MIR) spectra wavepoints (called secondary traits) and 3 focal traits (average fat percentage [AFP], average methane production [ACH4], and average SCS [ASCS]) collected on 3,302 first-parity Holstein cows. The 3,421 milk MIR wavepoint traits were composed of 311 wavepoints in 11 classes (months in lactation). Genotyping information of 564,439 SNPs was available for all animals and was used to calculate the genomic relationship matrix. The MegaLMM was implemented in the framework of the Bayesian sparse factor model and solved through Gibbs sampling (Markov chain Monte Carlo). The heritabilities of the studied 3,421 milk MIR wavepoints gradually increased and then decreased in units of 311 wavepoints throughout the lactation. The genetic and phenotypic correlations between the first 311 wavepoints and the other 3,110 wavepoints were low. The accuracies of phenotype predictions from the ST model were lower than those from the TT model for AFP (0.51 vs. 0.93), ACH4 (0.30 vs. 0.86), and ASCS (0.14 vs. 0.33). The same trend was observed for the accuracies of **u** predictions for AFP (0.59 vs. 0.86), ACH4 (0.47 vs. 0.78), and ASCS (0.39 vs. 0.59). The average correlation between **U** predicted from the TT model and the new approximate method was 0.90. The new approximate method used for estimating **U** in MegaLMM will enhance the suitability of MegaLMM for applications in animal breeding. This study conducted an initial investigation into the application of thousands of traits in animal breeding and showed that the TT model is beneficial for the prediction of focal traits (phenotype and breeding values), especially for difficult-to-measure traits (e.g., ACH4).

Key words: phenomics, MegaLMM, methane, milk mid-infrared

INTRODUCTION

With the rapid development of high-throughput phenotyping (**HTP**) technologies (e.g., remote sensing, cameras, spectrometric analyses) fostering the novel field of phenomics, researchers encounter substantial volumes of phenotypic data (Silva et al., 2021). Moreover, efforts are currently being made to link the phenotypic expression of traits to a diverse range of molecular and biological mechanisms, often called molecular phenotypes based on the metabolome, the proteome, and the transcriptome (Suravajhala et al., 2016). Although these molecular phenotypes can be used to predict some traditional traits for animal or plant management, they have rarely been used directly for breeding. However, these molecular phenotypes may contain more information that has not been used by the predicted traits. For example, milk mid-

Received October 17, 2023.
Accepted June 10, 2024.
*Corresponding author: yansen.chen@uliege.be

The list of standard abbreviations for JDS is available at adsa.org/jds-abbreviations-24. Nonstandard abbreviations are available in the Notes.

infrared (**MIR**) spectra have traditionally been used to predict the content of fat, protein, lactose, and urea, and have been demonstrated to be useful for predicting novel traits (e.g., methane production). However, milk MIR spectra information is still only extracted on a trait-by-trait phenotypic basis.

Breeding plays an important role in the production, reproduction, and disease resistance of animals and plants (Bernardo, 2020; Brito et al., 2021). For example, more than half of the increase in protein yield of US Holstein cows in the past 50 yr comes from genetic improvements (Cole et al., 2020). Molecular phenotypes may be more effective in helping breeders improve their associated traits (e.g., methane emissions). However, extracting relevant information from molecular phenotypes using the traditional approach is challenging. Therefore, the new challenge is how to incorporate a large number of molecular phenotypes into the breeding programs.

In animal breeding, selection index is used to combine multiple traits into an overall index for measuring animal genetic value (Cole et al., 2021). However, most applications include only several or dozens of traits at the same time, as opposed to the thousands of traits that are currently available. The simultaneous genetic analysis of thousands of traits is a major challenge, even with advances in computing.

Transformation algorithms to simplify solving multi-trait (**MT**) models have been proposed many years ago (Jensen and Mao, 1988). Canonical transformation has traditionally been the most used approach to solve MT mixed model equations in animal breeding (Ducrocq and Chapuis, 1997). Runcie et al. (2021) recently introduced a novel mega-scale linear mixed model (**MegaLMM**) which has been tested in the context of plant breeding. The MegaLMM approach re-parameterizes the MT linear mixed model into a Bayesian sparse factor model (Runcie et al., 2021). Factor analysis and canonical transformation are both techniques used in multivariate analysis. Although they are related, they serve different purposes and offer distinct advantages depending on the context of the analysis. Factor analysis helps in identifying underlying factors or latent variables that explain the patterns of correlations among potentially many observed traits. It also aims to reduce the dimensionality of the data by uncovering the common sources of variation, something that canonical transformation, on the other hand, does not do natively. Factor analysis is also more suitable for modeling complex relationships among traits by capturing shared variance among them. Based on these elements, the Bayesian sparse factor model (Runcie et al., 2021), as implemented in MegaLMM, is an interesting alternative to the use of canonical transformation because it seems to be an adequate solution for resolving the methodological challenges of canonical transformation. Qu et al. (2023) extended MegaLMM to mega-scale Bayesian regression methods, which were used in genome prediction and genome-wide association studies in plant breeding. To our knowledge, the MegaLMM method has not been evaluated in animal breeding.

Milk MIR spectra, which represent the absorbance of hundreds or thousands of individual wavepoints, are widely used to predict milk composition and phenotypes linked to animal health, efficiency, emissions, resilience, and even milk processability (Gengler et al., 2016; Grelet et al., 2021; Shadpour et al., 2022). The MIR spectra can be collected routinely during milk composition analyses, making them available at low cost. Beyond the traditional use of milk MIR spectra to predict phenotypes, researchers performed genetic analyses for milk MIR wavepoints with single-trait (**ST**) models, which cannot provide a direct view of the overall genetic structure (Rovere et al., 2019; Du et al., 2020; Tiplady et al., 2021). Several authors (e.g., Soyeurt et al., 2010; Bonfatti et al., 2017) conducted genetic analysis of principal components (**PC**) derived from phenotypic (co)variances based on principal component analysis of milk MIR spectra. However, they selected the number of PC based on the phenotypic variance explained by PC, which can result in some genetic information being lost (Chen et al., 2023b). Even if these studies used MT analysis to partially evaluate genetic correlations among milk MIR PC, the genetic (co)variances among original MIR spectra could only be partially evaluated. Rovere et al. (2019) suggested that studying the genetic correlations between milk MIR wavepoints at different time points and between milk MIR wavepoints and economic and environmental traits will be beneficial for integrating milk MIR spectra into genetic evaluations. The high correlation between milk MIR wavepoints requires direct simultaneous genetic analysis to have a better view of their genetic structures and help us to add milk MIR spectra to genetic evaluation.

With the sustainable and balanced development of animal production, an increasing number of traits (e.g., feed efficiency, methane emissions) are being included in animal breeding programs. The aims of this study were to (1) explore simultaneous estimates of (co)variance components and genetic parameters of thousands of milk MIR traits with the MegaLMM; (2) compare the phenotype value and genomic breeding value (**u**) predictions for focal traits from ST and thousand-trait (**TT**) models, respectively; (3) propose a new approximate method for estimating genomic breeding values (**U**) using MegaLMM. This study will provide a preliminary demonstration and reference for the application of thousands of traits in animal breeding.

MATERIALS AND METHODS

Data

Phenotypic Data. All milk samples were collected by Elevéo (Awé groupe, Ciney, Belgium) from January 2012 to December 2017 during the official milk recording in the Walloon Region of Belgium. These milk samples were chosen because they could be associated with the genotyped cow population. Milk samples were analyzed by MilkoScan FT6000 spectrometers (FOSS) and Fossomatic FC (FOSS, flow cytometry) to generate predicted fat percentage (**FP**), SCC, and MIR spectra (1,060 wavepoints). Methane emission (g/d) of each animal associated with a milk sample was predicted based on milk MIR spectra, milk yield, breed, and parity as described by Vanlierde et al. (2021). The coefficient of determination and root mean square error of 5-fold cross-validation for the methane equation were 0.68 and 57 g/d methane, respectively (Vanlierde et al., 2021). The SCS was calculated by the following formula: SCS = $\log_2$ (SCC/100,000) + 3. The FP and SCC were limited from 1.5% to 9% and from 10,000 to 10,000,000 cells/mL of milk, respectively, keeping the limits that are used in routine genetic evaluation systems of Holstein cows (Vanderick et al., 2022). For methane, editing was only based on the plausibility of values, and 21 records predicted to be less than 0 were set to missing values. The plausibility value is the individual predicted value within the range of the average value in the same lactation month group ± 3 SD. Each herd had to have a minimum of 10 records for each DIM. We exclusively considered the first-parity records observed from 5 to 365 DIM. Consequently, each animal should contribute to a total of 11 records for each analyzed trait, otherwise the animals had missing records.

The MIR spectra were selected because they can be obtained routinely and linked to multiple traits in dairy cows as described in the introduction. The selection of FP, methane, and SCS was guided by their varying link to MIR spectra. The estimation of FP is done routinely using milk MIR spectra with a very high coefficient of determination close to 1. As previously explained SCC is not obtained by milk MIR spectra. The link between SCC and spectral data is weaker even if a certain overlap of their reaction to mastitis was found, as cows suffering from mastitis show both more SCC and also changing milk composition (Bruckmaier et al., 2004) detectable milk MIR. Another reason for selecting methane is that it is expensive to directly measure it in dairy cows.

A total of 27,855 records (MIR, FP, methane, and SCS) were collected on 3,302 Holstein cows distributed in 74 herds used in this study. The MIR spectra (1,060 wavepoints) were first processed using the first derivative and then standardized by subtracting the corresponding mean and dividing by the corresponding SD (Delhez et al., 2020). The 311 milk MIR wavepoints from 3 distinct regions (933–1,589 cm^{-1}, 1,704–1,809 cm^{-1}, and 2,553–2,981 cm^{-1}) were retained. The spectral regions were selected based on the experience of our extended research team (e.g., Grelet et al., 2021), and these regions are highly related to major elements of milk composition (e.g., fatty acid, protein; Soyeurt et al., 2006; Grewal et al., 2018).

Genotypic Data. Genotypic data of the 3,302 animals were extracted from the routine genetic evaluation system of Holstein cows in the Walloon Region of Belgium. The animals were genotyped by the 50K chip (Illumina, San Diego, CA) and were imputed to high-density using FImpute V2.2 software (Sargolzaei et al., 2014) with a reference population of 4,352 high-density individuals (1,046 bulls and 3,288 cows). Quality control measures for SNPs were conducted following the criteria outlined in Chen et al. (2023a). Ultimately, 564,439 SNPs, distributed across 29 chromosomes, were retained from the initial 730,539 SNPs.

Preprocessing of Data

For each test-day record, we organized the 311 milk MIR wavepoints into a single row. To categorize DIM (5–365 d), we divided the wavepoints into 11 classes, each spanning 31 d, except for the final 2 classes, which encompassed 41 d each. Each animal had a maximum of 1 recording in each class, as the recording scheme consists of 4 to 6 wk intervals. In the case that 2 records qualified for a class, the one that was closer to the center point of the class was used. As a result, each animal had 3,421 (311 × 11) milk MIR wavepoints, referred to as unique traits. It needs to be mentioned that from this point, each trait × DIM lactation class combination will also be called a secondary trait. For FP, methane, and SCS, we computed averages of each trait across the 11 classes, condensing them into single traits known as focal traits: average fat percentage (**AFP**), average methane production (**ACH4**), and average SCS (**ASCS**). The raw numbers of records for all traits are given in Supplemental File S1 (see Notes).

Missing values of secondary (3,421) and focal (3) traits in the 11 classes were imputed by the best prediction method (VanRaden, 1997), which is a standard method used in DHI with the following formula:

$$\hat{y}_i = \mu + \mathbf{c}\,'\mathbf{V}^{-1}\mathrm{t},$$

where $\hat{y}_i$ is the imputed value of trait i, which falls into one of the defined 11 classes, μ is the average of trait i (including classes with missing values), $\mathbf{c}$ is the vector

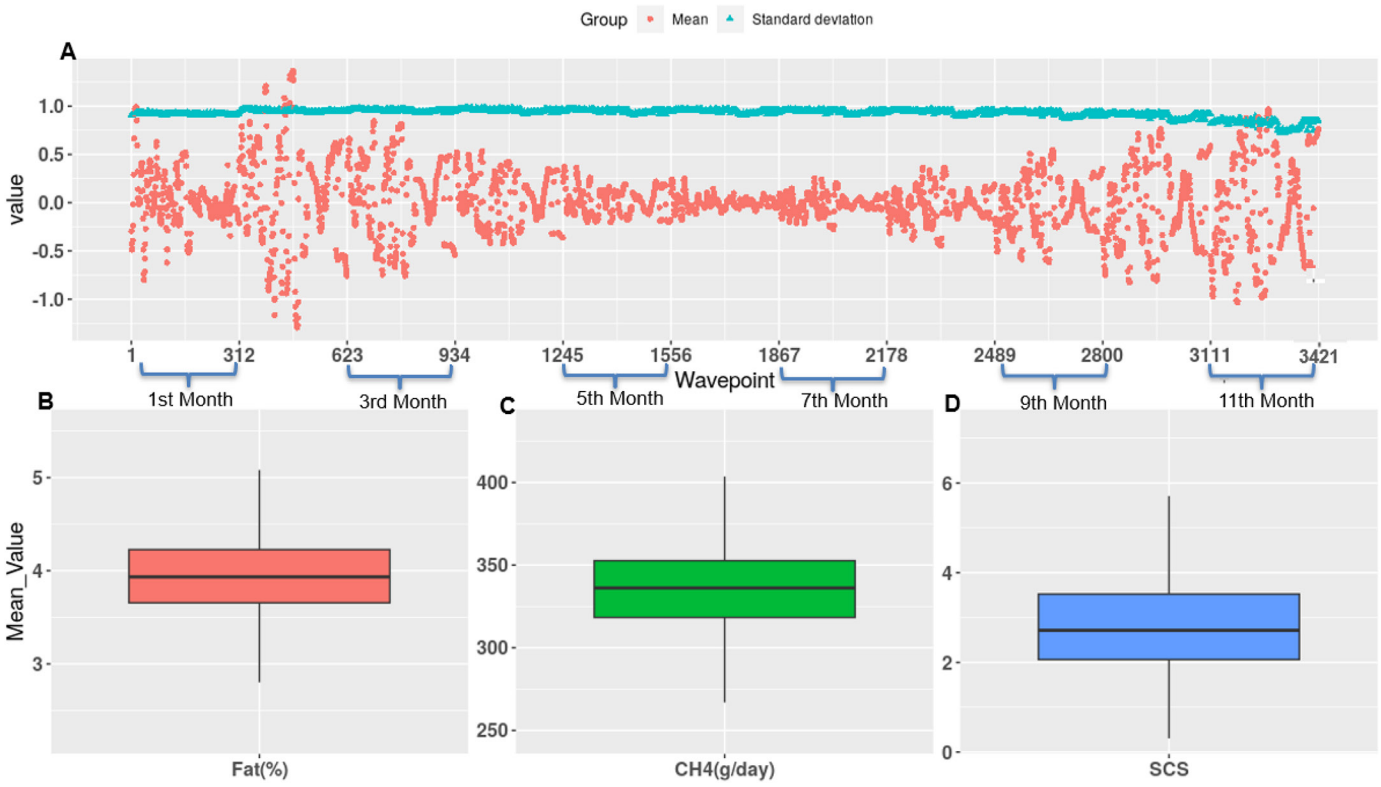

Figure 1. Description of all studied traits in this study. (A) Description of 3,421 observed milk mid-infrared wavepoints across 11 mo, with each month featuring 311 consistent wave points. (B) Description of average fat percentage (AFP) within the first parity (11 mo). (C) Description of average methane production (ACH4) within the first parity. (D) Description of average SCS (ASCS) within the first parity (n = 3,302). The lower, middle, and upper edges of the box represent the first quartile, median, and third quartile values of the trait, respectively; the lower and upper ends of whiskers are the minimum and maximum values of the trait.

of (co)variances between missing and observed values, **V** is the observed (co)variance matrix between observed values, and t is the observed deviations of trait *i* in observed classes. In this study, all data, after imputation, were treated as observed values.

(Co)variance Component Estimation

The ST model was used to estimate the variance components for each trait; TT model was used to estimate the (co)variance components for thousands of traits. The base model was as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \quad [1]$$

where **y** is the vector of traits (3,421 milk MIR wavepoints or 3,421 milk MIR wavepoints plus one of the focal traits or only one of the focal traits), **b** is the vector of fixed effects (age of calving group, herd-year of calving group, and year of calving group-month of calving). The age of calving was divided into 7 classes (<25, 25–26, 27–28, 29–30, 31–32, 33–34, ≥35 mo), and the year of calving was divided into 2 classes (2011–2014, 2015–2017). **u** and **e** are the random additive genetic and residual effects, respectively, and **X** and **Z** are the corresponding incidence matrixes. For ST models, the distributional assumptions of **u** and **e** were $\mathrm{u} \sim \mathrm{N}\left(0,\ \mathbf{G}\sigma_g^2\right)$ and $\mathrm{e} \sim \mathrm{N}\left(0,\ \mathbf{I}\sigma_e^2\right)$, where **G** is the genomic relationship matrix of the first method described by VanRaden (2008), σ_g^2 was the additive genetic variance, **I** was an identity matrix, and σ_e^2 is the residual variance. Genomic BLUP was used for the ST models.

The variance components of ST models were estimated by BLUPF90+ (ver. 2.48, Misztal et al., 2014) with average information REML (**AI-REML**) and with MegaLMM (ver. 0.9.4, Runcie et al., 2021) R package with Markov chain Monte Carlo (**MCMC**). The (co) variance components of the TT models (3,421 milk MIR wavepoints or 3,421 milk MIR wavepoints plus one of the focal traits) were estimated by MegaLMM (ver. 0.9.4, Runcie et al., 2021) in R package with MCMC. An MCMC chain of 50,000 was run with the first 10,000 iterations discarded as burn-in, and each 1 sample of 50

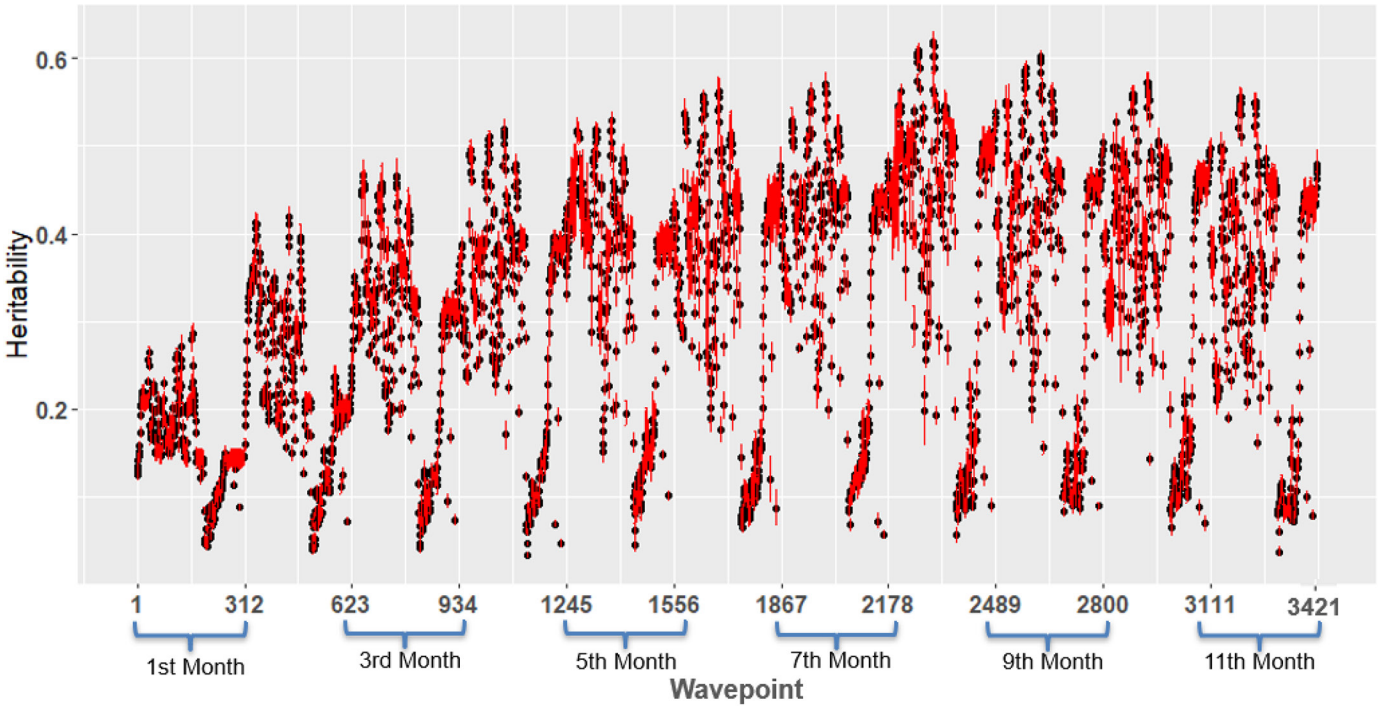

Figure 2. Heritabilities of 311 milk mid-infrared wavepoints over 11 mo of lactation, estimated from the thousand-trait model. The red lines indicate the SD of the 800 saved MCMC samples.

iterations was saved. The posterior mean of (co)variance of the TT model was calculated from 800 saved samples. The convergence of the MegaLMM method was assessed by visual inspection of trace plots (example figures in Supplemental File S2, see Notes).

The MegaLMM was implemented in the framework of the Bayesian sparse factor model and solved through MCMC (Runcie et al., 2021). In this study, the number of factors used in the MegaLMM was fixed at 500. Equation 1 can be rewritten and transformed for a factor model as the following:

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\Lambda} + \mathbf{Y}_\mathbf{R}, \quad [2]$$

$$\text{with } \mathbf{F} = \mathbf{Z}_\mathbf{F}\mathbf{U}_\mathbf{F} + \mathbf{E}_\mathbf{F},$$

$$\mathbf{Y}_\mathbf{R} = \mathbf{X}_\mathbf{R}\mathbf{B}_\mathbf{R} + \mathbf{Z}_\mathbf{R}\mathbf{U}_\mathbf{R} + \mathbf{E}_\mathbf{R},$$

where $\mathbf{Y}$ is a $n \times t$ matrix of observations for n animals and t traits (3,302 × 3,421 or 3,302 × 3,422), $\mathbf{F}$ is a $n \times k$ matrix of latent factor trait records (3,302 × 500), $\boldsymbol{\Lambda}$ is a $k \times t$ matrix of factor loadings (500 × 3,421 or 500 × 3,422), $\mathbf{Y}_\mathbf{R}$ is an $n \times t$ matrix of uncorrelated residual values (3,302 × 3,421 or 3,302 × 3,422). The $\mathbf{U}_\mathbf{F}$ and $\mathbf{E}_\mathbf{F}$ are matrixes of additive and residual effects for $\mathbf{F}$; $\mathbf{B}_\mathbf{R}$, $\mathbf{U}_\mathbf{R}$, and $\mathbf{E}_\mathbf{R}$ are matrixes of fixed (same as in Equation 1), additive, and residual effects for $\mathbf{Y}_\mathbf{R}$, $\mathbf{Z}_\mathbf{F}$, $\mathbf{X}_\mathbf{R}$, $\mathbf{Z}_\mathbf{R}$ are the corresponding incidence matrixes.

Assuming that all correlations of traits in $\mathbf{Y}$ were explained by $\mathbf{F}$, $\mathbf{Y}_\mathbf{R}$ would be uncorrelated. The MegaLMM sampling at each iteration of MCMC can be obtained simultaneously in parallel across $\mathbf{F}$ and $\mathbf{Y}_\mathbf{R}$, which leads to the simultaneous analysis of thousands of traits. The horseshoe prior distribution was used for $\boldsymbol{\Lambda}$, and the priors' distribution for other parameters was the same as used in Runcie et al. (2021).

Phenotypic (co)variance components of the original thousands of traits were equal to $\boldsymbol{\Lambda}'$ × (co)variance component of $\mathbf{F} \times \boldsymbol{\Lambda}$ plus the estimated variance component of $\mathbf{Y}_\mathbf{R}$. Solutions for the original thousands of traits were obtained by back-solving the factor part and adding the residual.

Genetic Parameters Estimation

The h^2 of each trait equals σ_g^2 divided by total variance $\left(\sigma_g^2 + \sigma_e^2\right)$. The approximated SE of h^2 from the BLUPF90+ program was obtained according to the method of Meyer and Houle (2013). The SD of h^2 from the MegaLMM program was obtained according to the saved MCMC samples. The h^2 was calculated for every 100 saved samples, so each trait received 8 h^2.

Table 1. The heritability of average fat percentage (AFP), average methane production (ACH4), and average SCS (ASCS) within the first parity (11 mo) from single-trait and thousand-trait models in the whole dataset (n = 3,302)[1]

	AFP		ACH4		ASCS	
Model	MegaLMM	AI-REML	MegaLMM	AI-REML	MegaLMM	AI-REML
Single-trait	0.66 (0.00)	0.75 ± 0.02	0.24 (0.00)	0.24 ± 0.03	0.09 (0.00)	0.09 ± 0.02
Thousand-trait	0.62 (0.01)	NA	0.22 (0.02)	NA	0.06 (0.00)	NA

[1]Values in parentheses are the SD of the 800 saved MCMC samples. MegaLMM = mega-scale linear mixed method; AI-REML = average information REML; NA = not applicable.

In the TT model, genetic and phenotypic correlations between each 2 traits were calculated by the following formulas:

$$\text{Genetic correlation} = \frac{\sigma_{g12}}{\sqrt{\sigma^2_{g1} \times \sigma^2_{g2}}}$$

$$\text{Phenotypic correlation} = \frac{\sigma_{p12}}{\sqrt{\sigma^2_{p1} \times \sigma^2_{p2}}},$$

where σ_{g12} and σ_{p12} are the additive genetic and phenotypic covariances between traits 1 and 2, respectively; σ^2_{gt} and σ^2_{pt} are the additive genetic and phenotypic variances of trait t, respectively; σ_{p12} equals additive genetic plus residual covariances between traits 1 and 2; σ^2_{pt} equals additive genetic plus residual variances of trait t.

Phenotype and Genomic Breeding Values Prediction

For phenotype prediction of focal traits by the MegaLMM method, the ST (one of 3 focal traits) and TT (one of the focal traits plus 3,421 milk MIR wavepoints) models were used. A total of 15 herds of the initial 74 herds were randomly selected, then the 3 focal traits of the animals (n = 721) in the selected 15 herds were set as missing values. The Gibbs samplers were used in MegaLMM to predict phenotypes, and more details can be found in Additional File 1 of Runcie et al. (2021). The Pearson correlations between the 3 observed and predicted focal traits of animals (n = 721) from ST and TT were calculated, respectively. The SD of Pearson correlations was calculated based on the saved MCMC samples.

For **u** prediction of focal traits, the ST models were used with the AI-REML and MegaLMM methods; TT models were used only with the MegaLMM. All analyses were done in the partial (with missing values, same as the phenotype prediction of focal traits in the previous paragraph) and whole datasets, respectively. Prediction accuracies of **u** were calculated by the following formula (Legarra and Reverter, 2018):

$$\text{Accuracy} = \sqrt{\frac{\sigma_{u_{pw}}}{\left(1-\bar{f}\right)\sigma^2_{g_p}}},$$

where $\sigma_{u_{pw}}$ is the covariance between **u** of selected animals in the partial and whole datasets, $\bar{f}$ is the average inbreeding coefficient of the selected animals (n = 721), and $\sigma^2_{g_p}$ is the additive genetic variance in the partial dataset. The SD of accuracy from the MegaLMM method was calculated based on the saved MCMC samples.

Approximate Method for Large-Scale Genomic Breeding Value Estimation Using MegaLMM

Because **U** estimation in the animal breeding field is usually based on hundreds of thousands of individuals, an approximate method is needed to calculate **U** of the TT model in larger populations. Similar to current practice in animal breeding, the estimation of (co)variance components and a factor model of loadings can be separated from the estimation of **U**. The following approximate strategy was tested on the same small data et. The new approximate method uses the following procedure:

1. Generate several small datasets that can represent the large dataset. Sampling can be based on various dimensions such as groups, regions, or test years.
2. Estimate the factor loading matrix **Λ** in each small dataset using Equation 2 and MegaLMM:

$$\mathbf{\Lambda} = \text{estimate\_factor\_loading}\ (\mathbf{Y_s}, \mathbf{F}, \mathbf{Y_{Rs}})$$

(estimated individually on the generated small datasets; $\mathbf{Y_s}$ and $\mathbf{Y_{Rs}}$ are the phenotypic and residual values in the small datasets).

3. Compute the transformation matrix **T** for each small dataset:

$$\mathbf{T} = \mathbf{\Lambda}'\left(\mathbf{\Lambda}\mathbf{\Lambda}'\right)^{-1}$$

(a generalized least-square inverse for Equation 2).

Compute the factor loading $\left(\hat{\mathbf{F}}\right)$ and residual $\left(\hat{\mathbf{Y}}_{\mathbf{Rb}}\right)$ to be used in the large dataset:

$$\hat{\mathbf{F}} \approx \mathbf{Y}_{\mathbf{b}}\overline{\mathbf{T}}$$

($\mathbf{Y}_{\mathbf{b}}$ is the phenotypic values in the large dataset; $\overline{\mathbf{T}}$ is the mean of $\mathbf{T}$ from all small datasets)

$$\hat{\mathbf{Y}}_{\mathbf{Rb}} = \mathbf{Y}_{\mathbf{b}} - \hat{\mathbf{F}}\boldsymbol{\Lambda}.$$

Estimate the approximated GEBV $\left(\hat{\mathbf{U}}\right)$:

$$\hat{\mathbf{U}}_{\hat{\mathbf{F}}} = \text{GBLUP\_analysis}\left(\hat{\mathbf{F}}\right)\text{with ST models}$$

$$\hat{\mathbf{U}}_{\hat{\mathbf{Y}}_{\mathbf{Rb}}} = \text{GBLUP\_analysis}\left(\hat{\mathbf{Y}}_{\mathbf{Rb}}\right)\text{with ST models.}$$

Calculate the $\hat{\mathbf{U}}$ of thousands of traits $\left(\hat{\mathbf{U}}_{\mathbf{b}}\right)$:

$$\hat{\mathbf{U}}_{\text{app}} = \hat{\mathbf{U}}_{\hat{\mathbf{F}}}\boldsymbol{\Lambda}$$

$$\hat{\mathbf{U}}_{\text{b}} = \hat{\mathbf{U}}_{\text{app}} + \hat{\mathbf{U}}_{\hat{\mathbf{Y}}_{\mathbf{Rb}}},$$

(after checking whether $\hat{\mathbf{U}}_{\hat{\mathbf{Y}}_{\mathbf{Rb}}}$ contains useful information; if not, it can be removed; $\hat{\mathbf{U}}_{\text{app}}$ = approximate GEBV of all factors).

To initially verify our proposed approximation method, the above small data (ACH4 plus 3,421 milk MIR wavepoints) was used for testing. This data (ACH4 plus 3,421 milk MIR wavepoints) was chosen because ACH4 is difficult to obtain in routine testing and MIR data has a moderate ability to predict ACH4. The $\hat{\mathbf{U}}_{\text{m}}$ from the TT model of thousands of traits were used as the benchmark, and Pearson correlations between $\hat{\mathbf{U}}$ of approximated methods and $\hat{\mathbf{U}}_{\text{m}}$ were used as metrics. The approximate methods considered included 3 methods: the first used only the $\hat{\mathbf{U}}_{\text{app}}$; the second used the $\hat{\mathbf{U}}_{\text{app}}$ plus the $\hat{\mathbf{U}}_{\hat{\mathbf{Y}}_{\mathbf{Rb}}}$; and the third used the $\hat{\mathbf{U}}_{\text{s}}$ from the ST models of all traits. The $\hat{\mathbf{U}}_{\text{m}}$ from these results (MegaLMM method) and the 3 approximated methods were calculated by GBLUP with thousands of ST models (same effects; rrBLUP package, ver. 4.6.2). The difference in Pearson correlation between $\hat{\mathbf{U}}_{\text{m}}$ and $\hat{\mathbf{U}}$ among the 3 approximation methods was tested

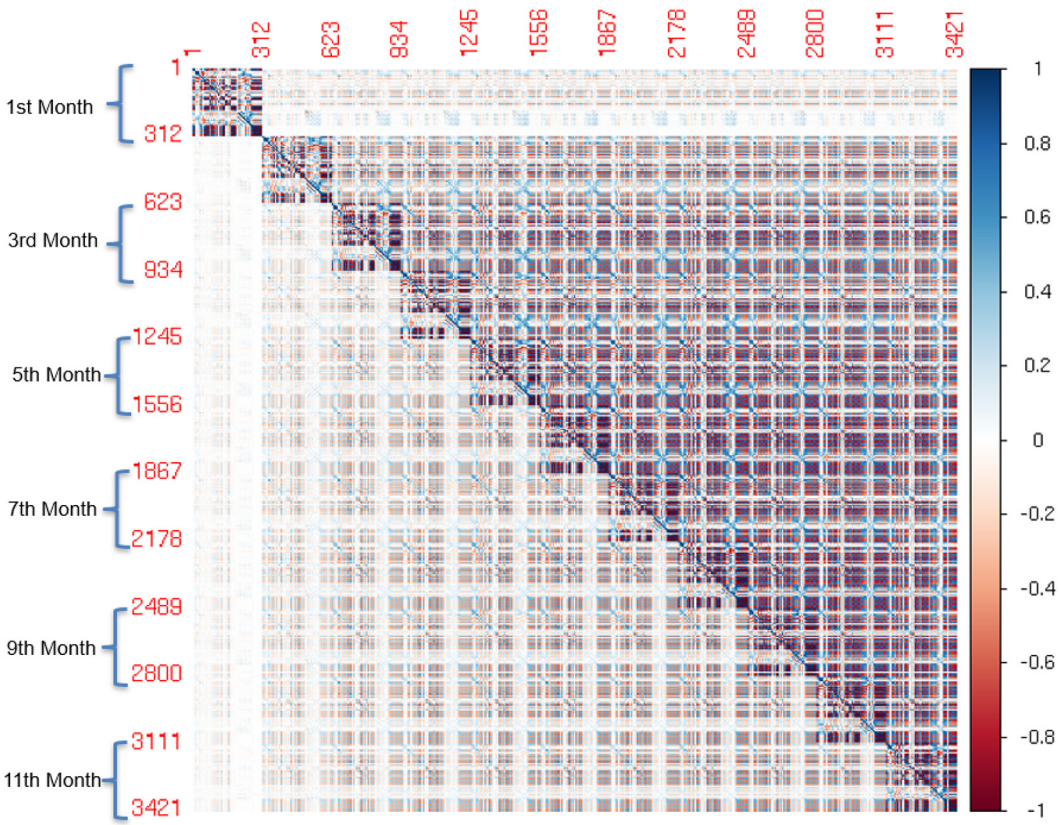

Figure 3. Genetic correlations (above the diagonal) and phenotypic correlations (below the diagonal) among 3,421 milk mid-infrared wavepoints (collected over 11 mo, with each month featuring 311 consistent wavepoints) from the thousand-trait model.

pairwise by the Wilcoxon test. All data preparation and processing were done using R (ver. 4.1.2, https://www.r-project.org/).

RESULTS

Descriptive Statistics and Heritability

Figure 1 shows the description of the studied traits after imputation. The mean of milk MIR wavepoints values varied greatly in early and late lactation and was close to 0 in mid-lactation. The SD of milk MIR wavepoints values were close to 1, except for the last month in milk (335–365 d). The means (SD) of AFP, ACH4, and ASCS were 3.95% (0.44), 335.27 g/d (26.05), and 2.91 (1.16), respectively.

The h^2 of the 3,421 milk MIR wavepoints estimated from the TT model were shown in Figure 2 and ranged from 0.034 (wavepoint 1,132) to 0.619 (wavepoint 2,308). The h^2 of the 311 milk MIR wavepoints gradually increased and then decreased throughout the lactation period. Among the 311 milk MIR wavepoints, the h^2 of certain wavepoints exhibited minor fluctuations throughout the first lactation, whereas for others, we found marked changes in h^2. For example, the milk MIR wavepoints 1,132 (lowest h^2) and 2,308 (largest h^2) corresponding to h^2 of the wavepoints in the first unit (1–311) were 0.045 (wavepoint 199) and 0.200 (wavepoint 131), respectively.

The h^2 of the 3 studied focal traits are shown in Table 1. The h^2 of AFP, ACH4, and ASCS were high (0.62–0.75), medium (0.22–0.24), and low (0.06–0.08), respectively. The h^2 of ACH4 and ASCS estimated from the ST model through MegaLMM and AI-REML methods were similar; however, the h^2 of AFP was different. The h^2 of the 3 studied focal traits estimated from the ST model was higher than that estimated from the TT model through the MegaLMM method.

Genetic and Phenotypic Correlations

The genetic (above the diagonal) and phenotypic (below the diagonal) correlations among the 3,421 milk MIR wavepoints are shown in Figure 3. The genetic correlations of the 3,421 milk MIR wavepoints were higher than their phenotypic correlations. The genetic and phenotypic correlations of milk MIR wavepoints were higher within a month of lactation compared with other months of lactation. The genetic and phenotypic correlations between the first 311 milk MIR wavepoints and the other 3,110 milk MIR wavepoints were low.

The absolute values of genetic correlations between the 3 focal traits and the 3,421 milk MIR wavepoints were higher than the phenotypic correlations (Figure 4). The genetic and phenotypic correlations between AFP and the first 311 milk MIR wavepoints were lower than correlations between AFP and the other 3,110 milk MIR wavepoints. A similar situation was observed for ACH4 and milk MIR wavepoints; however, the pattern of genetic correlations between the milk MIR wavepoints and AFP, and ACH4 differed. Genetic and phenotypic correlations between ASCS and the 311 milk MIR wavepoints remained relatively consistent across the month of lactation, whereas phenotypic correlations approached 0.

Table 2. Correlation of the average fat percentage (AFP), average methane production (ACH4), and average SCS (ASCS) within the first parity (11 mo) between predicted and observed values from single-trait and thousand-trait models by mega-scale linear mixed method (MegaLMM)[1]

Model	AFP	ACH4	ASCS
Single-trait	0.51 (0.01)	0.30 (0.01)	0.14 (0.01)
Thousand-trait	0.93 (0.00)	0.86 (0.01)	0.33 (0.01)
Increased[2] (%)	81.61	184.98	136.20

[1]Values in parentheses are the SD of the 800 saved MCMC samples.
[2]Increased percentage of correlation from the single-trait model to the thousand-trait model in each trait.

Phenotype and Genomic Breeding Values Prediction

The correlations between MIR-based predicted AFP, ACH4, and ASCS and the predicted values obtained by the ST and TT models are shown in Table 2. As expected, correlations of the 3 focal traits increased in the TT model, ranging from around 82% to 185%. The accuracy of **u** prediction of AFP, ACH4, and ASCS from the TT model also increased, from around 47% to 65% (Table 3). In addition, the accuracies of **u** prediction from the ST model through MegaLMM and AI-REML methods for AFP (0.59 vs. 0.57), ACH4 (0.47 vs. 0.47), and ASCS (0.39 vs. 0.40) were similar.

Approximate Method of Genomic Breeding Values Estimation Using MegaLMM

Figure 5A shows the approximate method for applying the TT model by MegaLMM to large populations. The average (SD) correlations between $\hat{\mathbf{U}}_m$ and $\hat{\mathbf{U}}_{app}$, between $\hat{\mathbf{U}}_m$ and $\hat{\mathbf{U}}_b$, between $\hat{\mathbf{U}}_m$ and $\hat{\mathbf{U}}_s$ were 0.90 (0.03), 0.90 (0.03), and 0.82 (0.09), respectively; the results of ACH4 were 0.88, 0.88, and 0.73, respectively. The correlations of the first 2 approximate methods were similar and were both significantly larger than those obtained using the last method (ST model; Figure 5B). The average (SD) h^2 of traits in low (<0.89, bottom 25%) and high (>0.92, top 25%) correlations between $\hat{\mathbf{U}}_m$ and $\hat{\mathbf{U}}_{app}$ were 0.19 (0.11) and 0.39 (0.11), respectively. The h^2 of traits in the low-correlation group was significantly lower than the h^2 of traits in the high-correlation group (Figure 5C). In this

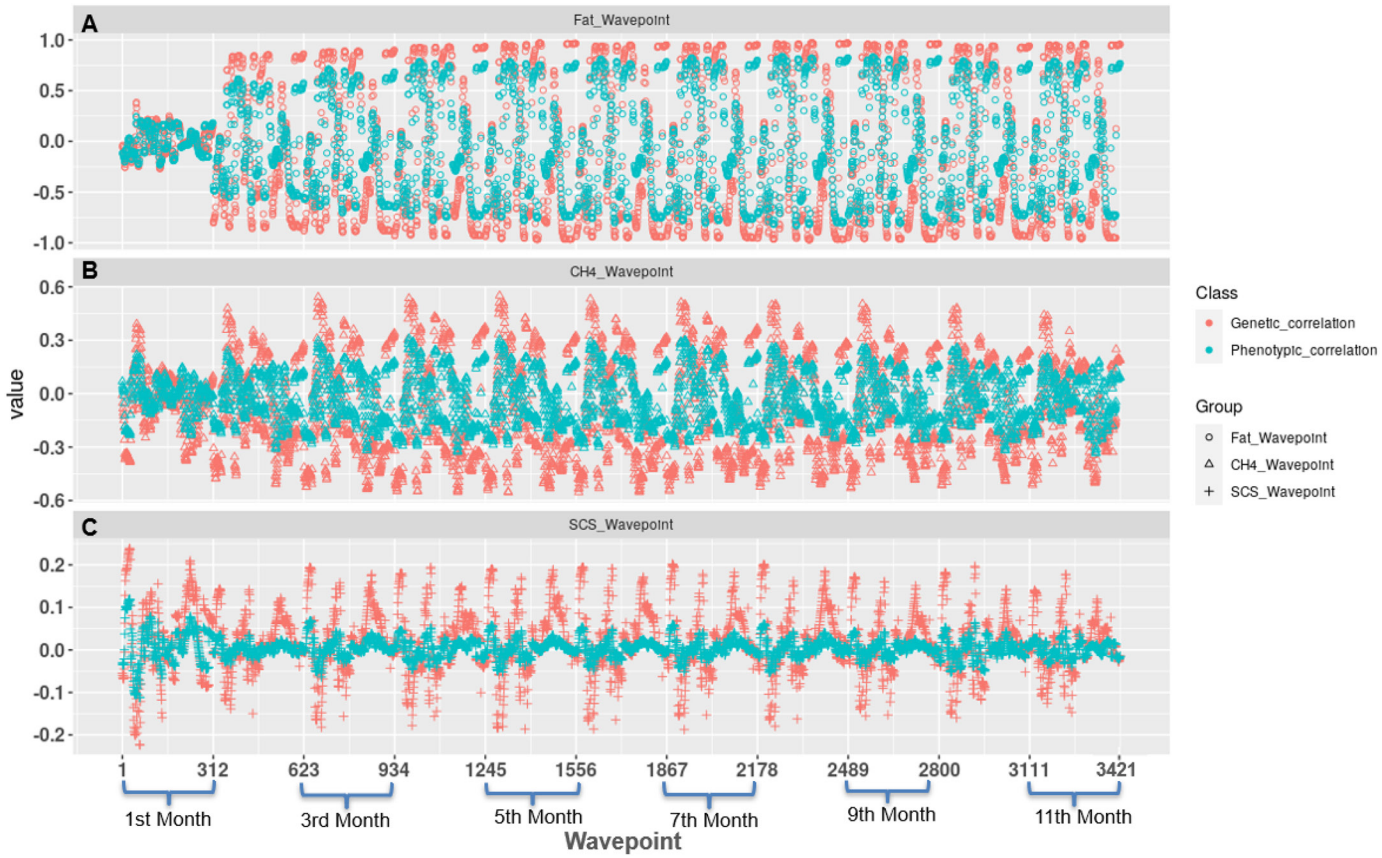

Figure 4. Genetic and phenotypic correlations between 3 focal traits and 3,421 milk mid-infrared (MIR) wavepoints (collected over 11 mo, with each month featuring 311 consistent wavepoints) from the thousand-trait model. (A) Genetic and phenotypic correlations between average fat percentage (AFP) within the first parity (11 mo) against MIR wavepoints. (B) Genetic and phenotypic correlations between average methane production (ACH4) and MIR wavepoints. (C) Genetic and phenotypic correlations between average SCS (ASCS) and MIR wavepoints.

test with a small dataset, the initial MegaLMM took 127 h to estimate $\hat{\mathbf{U}}_{\mathrm{m}}$, whereas the approximate method took 1 h and 25 min to estimate $\hat{\mathbf{U}}_{\mathrm{app}}$ and 10 h to estimate $\hat{\mathbf{U}}_{\mathrm{b}}$.

DISCUSSION

With the advancement of phenomics, thousands of traits per animal are routinely measured. However, the simultaneous genetic analysis of thousands of traits poses a substantial challenge for animal breeding, particularly when routine analyses are necessary. We estimated the (co)variance compositions among thousands of traits in routine recording, alongside the calculation of their genetic parameters. Our results show the benefits of predicting focal traits (phenotype and **u**) using the TT model, especially for difficult-to-measure traits (e.g., ACH4 emissions). In addition, we proposed an approximated method to estimate **U** that is suitable for large datasets using MegaLMM.

Our results suggested that the same wavepoint should not be considered as an identical trait at different time points (Figure 1). The findings of Rovere et al. (2019) support our results even though they analyzed the same milk MIR wavepoint separately at different times using ST models. However, differences in the h^2 patterns of milk MIR wavepoints were observed between our study and that of Rovere et al. (2019). This discrepancy may be attributed to the use of the TT model in the current study. Most prior studies have traditionally treated milk MIR wavepoints as the same trait across different time points (Zaalberg et al., 2019, 2020; Du et al., 2020; Tiplady et al., 2021). This practice can introduce bias into genetic analysis and potentially overlook the identification of relevant candidate genes.

The h^2 of AFP, ACH4, and ASCS from this study align with those of other studies involving big datasets with Walloon region Holstein cows (Paiva et al., 2022; Kandel et al., 2017; Atashi et al., 2023), underscoring the reliability of results obtained through MegaLMM. Paiva et al. (2022) identified a maximum h^2 of 0.54 for daily fat percentage in Walloon region Holstein cows, a value that is in close agreement with our MegaLMM results.

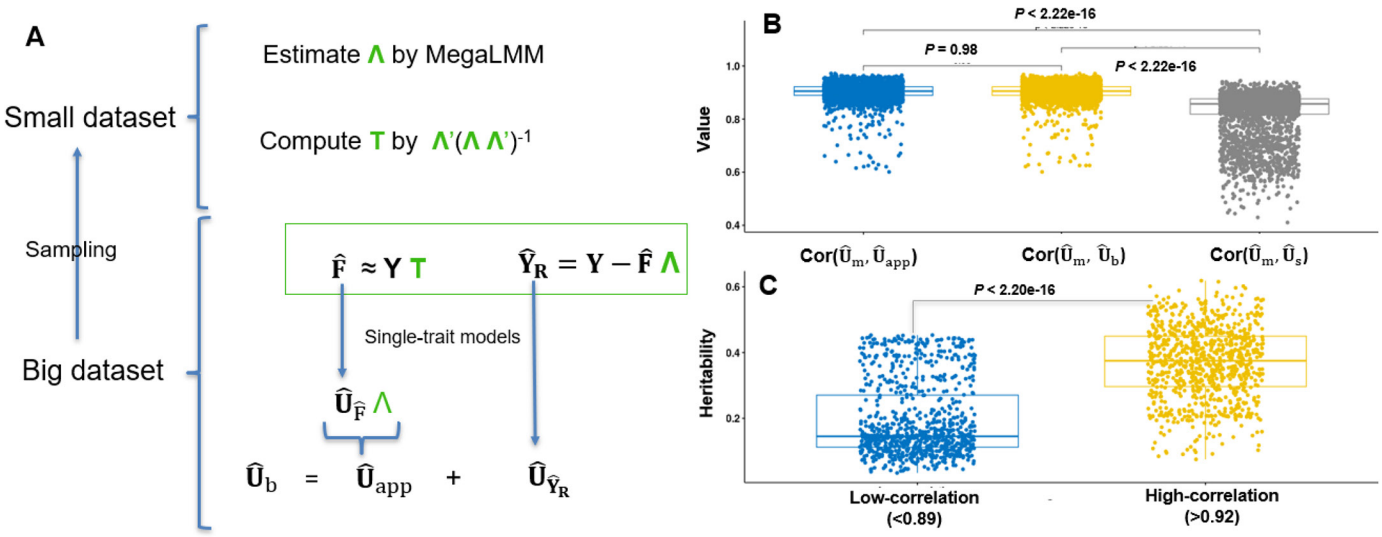

Figure 5. Overview of newly proposed approximation methods for big datasets and their illustration to small datasets. (A) Approximate method of GEBV $(\hat{\mathbf{U}})$ estimation in a big dataset with MegaLMM. (B) Correlations between $\hat{\mathbf{U}}_m$ of thousand-trait model and $\hat{\mathbf{U}}_{app}\hat{\mathbf{U}}_b$ of proposed approximated methods, $\hat{\mathbf{U}}_s$ of single-trait models. (C) Distribution of heritability of bottom (n = 875) and top (n = 875) 25% correlations between $\hat{\mathbf{U}}_m$ and $\hat{\mathbf{U}}_{app}$ ($\mathbf{\Lambda}$ and $\mathbf{T}$ estimated in a small dataset). The lower, middle, and upper edges of the box represent the first quartile, median, and third quartile values of the trait, respectively; the lower and upper ends of whiskers are the minimum and maximum values of the trait. The dots in panel B represent the correlation values of estimated genomic breeding values by different methods; the dots in panel C represent the heritability of traits.

To our knowledge, this study presents the first exploration of genetic and phenotypic correlations between milk MIR wavepoints at different time intervals. The results suggested that we need to partition the spectral analysis of the first 35 DIM and the subsequent lactation period (Figure 3). The pattern of correlations between milk MIR wavepoints and AFP and ACH4 supports this point (Figure 4). The patterns between genetic correlations of milk MIR wavepoints and focal traits were stable after DIM 35, which is beneficial for using milk MIR wavepoints as proxies for difficult-to-record (and therefore often missing) traits in genetic selection. For example, Toledo-Alvarado et al. (2022) proposed employing individual milk MIR wavepoints as a proxy for residual feed intake. However, the accuracies of indirect selection for individual milk MIR wavepoints were too low (from 0.0 to 0.1). This is because the genetic correlations between single individual milk MIR wavepoints and residual feed intake were low, ranging from around −0.25 to 0.22. At this level, one must remember that the combination of multiple milk MIR wavepoints generates relevant information with high genetic correlations (around ± 0.55) to ACH4; thus, it may be possible to conduct genetic selection for ACH4 with these milk MIR wavepoints as proxies. This approach can be extended to numerous other traits predicted by milk MIR. Directly using milk MIR wavepoints as traits for genetic selection can avoid developing calibration models for milk MIR predictive traits.

The improved phenotype and **u** prediction accuracy by the TT model for 3 focal traits are shown in Tables 2 and 3. Similar improvements in **u** prediction have been demonstrated in the field of plant breeding using the hundred-trait models (Runcie et al., 2021; Qu et al., 2023). Even though the observed values of AFP and ACH4 in this study were predicted by milk MIR and may not be a completely fair test of prediction accuracy, ASCS observations remained entirely uncorrelated with milk MIR. Furthermore, the ACH4 observations were predicted using multiple types of information (212 milk MIR wavepoints, milk yield, breed, and parity), which are not exactly the same information used in our study. Tiezzi et al. (2022) showed that milk MIR can be used as a covariate to improve the genome prediction of SCS in new environments. The results of the current study showed that incorporating milk MIR data as secondary traits to predict focal traits using MegaLMM enhanced the prediction accuracy of focal traits that are challenging to measure. In the next step, it is worth exploring the impact of incorporating milk MIR from different months into genetic analysis on the prediction accuracy of focal traits (e.g., ACH4 emissions). In addition, the convergence of MegaLMM needs to be considered (e.g., number of iterations), because this method is based on solving the problem of plant breeding, whereas animal breeding has different conditions (e.g., number of samples). For example, in this study, the analysis of ASCS together with milk

Table 3. The genomic breeding values prediction accuracies[1] of average fat percentage (AFP), average methane production (ACH4), and average SCS (ASCS0 within the first parity (11 mo) from single-trait and thousand-trait models[2]

	AFP		ACH4		ASCS	
Model	MegaLMM	AI-REML	MegaLMM	AI-REML	MegaLMM	AI-REML
Single-trait	0.59 (0.00)	0.57	0.47 (0.01)	0.47	0.39 (0.01)	0.40
Thousand-trait	0.86 (0.01)	NA	0.78 (0.03)	NA	0.59 (0.05)	NA
Increased[3] (%)	47.14	NA	64.77	NA	49.98	NA

[1]Formula for prediction accuracy is from Legarra and Reverter (2018).

[2]Values in parentheses are the SD of the 800 saved MCMC samples. MegaLMM = mega-scale linear mixed method; AI-REML = average information REML; NA = not applicable.

[3]Increased percentage of prediction accuracy from the single-trait model to the thousand-trait model in each trait.

MIR may not have completely converged, although the ability to predict ASCS has been improved.

Animal breeding requires the calculation of $\mathbf{U}$ for hundreds of thousands of individuals. We gave an approximate method of calculating $\mathbf{U}$ of the TT model by MegaLMM in larger populations (Figure 5A). Our results only require doing a genetic estimation of $\hat{\mathbf{F}}$, rather than $\hat{\mathbf{F}}$ plus $\hat{\mathbf{Y}}_{\mathbf{Rb}}$ (Figure 5B), which greatly reduces the number of traits that need to be analyzed (here 500 vs. 3,922). The correlations between $\hat{\mathbf{U}}_m$ and $\hat{\mathbf{U}}_{app}$, and between $\hat{\mathbf{U}}_m$, and $\hat{\mathbf{U}}_b$ were similar, whereas both correlations were significantly larger than the correlations between $\hat{\mathbf{U}}_m$ and $\hat{\mathbf{U}}_s$ of the ST model (Figure 5B). The low correlations of some traits in the $\hat{\mathbf{U}}_{app}$ method may be caused by the lower h^2 of these traits (Figure 5C). However, the number of traits for correlations lower than 0.80 was only 47. The approximate methods also greatly reduced the computation time for the initial MegaLMM. Finally, our proposed approximate method converted thousands of traits in a large population into multiple ST models for hundreds of traits, which can be quickly performed in parallel on a high-performance computer.

CONCLUSIONS

This study explored the application of the TT model analyzed by MegaLMM in animal breeding. The results of this study showed that the TT model is beneficial for exploring the phenotypes obtained by HTP technologies, such as the discovery of the pattern of milk MIR wavepoints over time and the report of the genetic correlation of milk MIR wavepoints at different time points. The results of this study also provide an example of the integration of molecular phenotypes obtained by HTP technologies into animal breeding. For example, wavepoint and focal trait analysis directly through the TT model can improve the phenotype and genomic breeding value prediction accuracy of focal traits in new herds. The novel approximate method to predict the TT genomic breeding values will enhance the applicability of MegaLMM in animal breeding. This study provides an example to explore the challenge of an increasingly large number of traits included in animal breeding program.

NOTES

The China Scholarship Council (Beijing) is acknowledged for funding the PhD project of Yansen Chen (no. 201906760007). Yansen Chen acknowledges the support of the Fonds de la Recherche Scientifique (FNRS, Brussels, Belgium) under grant no. T.0095.19 (PDR "DEEPSELECT"). The authors also acknowledge the support of the Walloon Government (Service Public de Wallonie–Direction Générale Opérationnelle Agriculture, Ressources Naturelles et Environnement, SPW-DGARNE; Namur, Belgium) and the use of the computation resources of the Consortium des Équipements de Calcul Intensif (CÉCI) funded by the FNRS under grant no. 2.5020.11. The University of Liège–Gembloux Agro-Bio Tech (Liège, Belgium) supported computations through the technical platform Calcul et Modélisation Informatique (CAMI) of the TERRA Teaching and Research Centre supported by the FNRS under grant no. T.0095.19 (PDR "DEEPSELECT"). Genotyping was facilitated through the support of the FNRS under grant no. J.0174.18 (CDR "PREDICT-2"). D. Runcie was supported by Agriculture and Food Research Initiative grant nos. 2020-67013-30904 and 2018-67015-27957 from the USDA National Institute of Food and Agriculture (NIFA) and by USDA NIFA Hatch project 1010469. Supplemental material for this article is available at https://github.com/Yansen0515/MegaLMM_for_Animal. No human or animal subjects were used, so this analysis did not require approval by an Institutional Animal Care and Use Committee or Institutional Review Board. The authors have not stated any conflicts of interest.

Nonstandard abbreviations used: AI-REML = average information REML; ACH4 = average methane production; AFP = average fat percentage; ASCS = average

SCS; FP = fat percentage; HTP = high-throughput phenotyping; MCMC = Markov chain Monte Carlo; MegaLMM = mega-scale linear mixed model; MIR = mid-infrared; MT = multitrait; NA = not applicable; PC = principal components; ST = single-trait; TT = thousand-trait; $\hat{\mathbf{U}}$ = GEBV of multiple traits from new approximated methods; $\mathbf{U}$ = GEBV of multiple traits; $\mathbf{u}$ = GEBV of a single trait.

REFERENCES

Atashi, H., Y. Chen, H. Wilmot, C. Bastin, S. Vanderick, X. Hubin, and N. Gengler. 2023. Single-step genome-wide association analyses for selected infrared-predicted cheese-making traits in Walloon Holstein cows. J. Dairy Sci. 106:7816–7831. https://doi.org/10.3168/jds.2022-23206.

Bernardo, R. 2020. Reinventing quantitative genetics for plant breeding: Something old, something new, something borrowed, something BLUE. Heredity 125:375–385. https://doi.org/10.1038/s41437-020-0312-1.

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

Brito, L. F., N. Bedere, F. Douhard, H. R. Oliveira, M. Arnal, F. Peñagaricano, A. P. Schinckel, C. F. Baes, and F. Miglior. 2021. Review: Genetic selection of high-yielding dairy cattle toward sustainable farming systems in a rapidly changing world. Animal 15:100292. https://doi.org/10.1016/j.animal.2021.100292.

Bruckmaier, R. M., C. E. Ontsouka, and J. W. Blum. 2004. Fractionized milk composition in dairy cows with subclinical mastitis. Vet. Med. (Praha) 49:283–290. https://doi.org/10.17221/5706-VETMED.

Chen, Y., H. Atashi, C. Grelet, R. R. Mota, S. Vanderick, H. HuGplusE Consortium, and N. Gengler. 2023a. Genome-wide association study and functional annotation analyses for nitrogen efficiency index and its composition traits in dairy cattle. J. Dairy Sci. 106:3397–3410. https://doi.org/10.3168/jds.2022-22351.

Chen, Y., P. Delhez, H. Atashi, H. Soyeurt, and N. Gengler. 2023b. Genetic analyses of principal components of milk mid-infrared spectra from Holstein cows. The 74th Annual Meeting of the European Federation of Animal Science. Lyon, France. Accessed Aug. 14, 2024. https://orbi.uliege.be/bitstream/2268/307832/1/EAAP_Yansen_Chen_Genetic_PCs_.pdf.

Cole, J. B., J. W. Dürr, and E. L. Nicolazzi. 2021. Invited review: The future of selection decisions and breeding programs—What are we breeding for, and who decides? J. Dairy Sci. 104:5111–5124. https://doi.org/10.3168/jds.2020-19777.

Cole, J. B., S. A. E. Eaglen, C. Maltecca, H. A. Mulder, and J. E. Pryce. 2020. The future of phenomics in dairy cattle breeding. Anim. Front. 10:37–44. https://doi.org/10.1093/af/vfaa007.

Delhez, P., F. Colinet, S. Vanderick, C. Bertozzi, N. Gengler, and H. Soyeurt. 2020. Predicting milk mid-infrared spectra from first-parity Holstein cows using a test-day mixed model with the perspective of herd management. J. Dairy Sci. 103:6258–6270. https://doi.org/10.3168/jds.2019-17717.

Du, C., L. Nan, L. Yan, Q. Bu, X. Ren, Z. Zhang, A. Sabek, and S. Zhang. 2020. Genetic analysis of milk production traits and mid-infrared spectra in Chinese Holstein population. Animals (Basel) 10:139. https://doi.org/10.3390/ani10010139.

Ducrocq, V., and H. Chapuis. 1997. Generalizing the use of the canonical transformation for the solution of multivariate mixed model equations. Genet. Sel. Evol. 29:205–224. https://doi.org/10.1186/1297-9686-29-2-205.

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

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

Grewal, M. K., T. Huppertz, and T. Vasiljevic. 2018. FTIR fingerprinting of structural changes of milk proteins induced by heat treatment, deamidation and dephosphorylation. Food Hydrocoll. 80:160–167. https://doi.org/10.1016/j.foodhyd.2018.02.010.

Jensen, J., and I. L. Mao. 1988. Transformation algorithms in analysis of single trait and of multitrait models with equal design matrices and one random factor per trait: A review. J. Anim. Sci. 66:2750–2761. https://doi.org/10.2527/jas1988.66112750x.

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

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genet. Sel. Evol. 50:53. https://doi.org/10.1186/s12711-018-0426-6.

Meyer, K., and D. Houle. 2013. Sampling based approximation of confidence intervals for functions of genetic covariance matrices. Pages 523–526 in Proc. Association for the Advancement of Animal Breeding and Genetics. Association for the Advancement of Animal Breeding and Genetics.

Misztal, I., S. Tsuruta, D. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2014. Manual for BLUPF90 family of programs. Accessed Jun. 26, 2023. http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all8.pdf.

Paiva, J. T., R. R. Mota, P. S. Lopes, H. Hammami, S. Vanderick, H. R. Oliveira, R. Veroneze, F. Fonseca e Silva, and N. Gengler. 2022. Random regression test-day models to describe milk production and fatty acid traits in first lactation Walloon Holstein cows. J. Anim. Breed. Genet. 139:398–413. https://doi.org/10.1111/jbg.12673.

Qu, J., D. Runcie, and H. Cheng. 2023. Mega-scale Bayesian regression methods for genome-wide prediction and association studies with thousands of traits. Genetics 223:iyac183. https://doi.org/10.1093/genetics/iyac183.

Rovere, G., G. de los Campos, R. J. Tempelman, A. I. Vazquez, F. Miglior, F. Schenkel, A. Cecchinato, G. Bittante, H. Toledo-Alvarado, and A. Fleming. 2019. 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. 102:1354–1363. https://doi.org/10.3168/jds.2018-15109.

Runcie, D. E., J. Qu, H. Cheng, and L. Crawford. 2021. MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. Genome Biol. 22:213. https://doi.org/10.1186/s13059-021-02416-w.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. https://doi.org/10.1186/1471-2164-15-478.

Shadpour, S., T. C. Chud, D. Hailemariam, G. Plastow, H. R. Oliveira, P. Stothard, J. Lassen, F. Miglior, C. F. Baes, D. Tulpan, and F. S. Schenkel. 2022. Predicting methane emission in Canadian Holstein dairy cattle using milk mid-infrared reflectance spectroscopy and other commonly available predictors via artificial neural networks. J. Dairy Sci. 105:8272–8285. https://doi.org/10.3168/jds.2021-21176.

Silva, F. F., G. Morota, and G. J. M. Rosa. 2021. Editorial: High-throughput phenotyping in the genomic improvement of livestock. Front. Genet. 12:707343. https://doi.org/10.3389/fgene.2021.707343.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89:3690–3695. https://doi.org/10.3168/jds.S0022-0302(06)72409-2.

Soyeurt, H., I. Misztal, and N. Gengler. 2010. Genetic variability of milk components based on mid-infrared spectral data. J. Dairy Sci. 93:1722–1728. https://doi.org/10.3168/jds.2009-2614.

Suravajhala, P., L. J. A. Kogelman, and H. N. Kadarmideen. 2016. Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. Genet. Sel. Evol. 48:38. https://doi.org/10.1186/s12711-016-0217-x.

Tiezzi, F., A. Fleming, and F. Malchiodi. 2022. Use of milk infrared spectral data as environmental covariates in genomic prediction models for production traits in canadian holstein. Animals (Basel) 12:1189. https://doi.org/10.3390/ani12091189.

Tiplady, K. M., T. J. Lopdell, E. Reynolds, R. G. Sherlock, M. Keehan, T. J. Johnson, J. E. Pryce, S. R. Davis, R. J. Spelman, B. L. Harris, D. J. Garrick, and M. D. Littlejohn. 2021. Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. Genet. Sel. Evol. 53:62. https://doi.org/10.1186/s12711-021-00648-9.

Toledo-Alvarado, H. O., R. J. Tempelman, M. Lopez-Cruz, M. J. VandeHaar, J. E. P. Santos, F. Peñagaricano, P. Khanal, and G. de los Campos. 2022. Phenotypic and genetic associations between feed efficiency and FTIR milk-spectra. Pages 252–255 in Proc. 12th World Congress on Genetics Applied to Livestock Production. https://doi.org/10.3920/978-90-8686-940-4_51.

Vanderick, S., R. R. Mota, K. Wijnrocs, and N. Gengler. 2022. Description of the genetic evaluation systems used in the Walloon Region of Belgium. Accessed Oct. 17, 2023. https://www.elinfo.be/indexEN.html.

Vanlierde, A., F. Dehareng, N. Gengler, E. Froidmont, S. McParland, M. Kreuzer, M. Bell, P. Lund, C. Martin, B. Kuhla, and H. Soyeurt. 2021. Improving robustness and accuracy of predicted daily methane emissions of dairy cows using milk mid-infrared spectra. J. Sci. Food Agric. 101:3394–3403. https://doi.org/10.1002/jsfa.10969.

VanRaden, P. M. 1997. Lactation yields and accuracies computed from test day yields and (co)variances by best prediction. J. Dairy Sci. 80:3015–3022. https://doi.org/10.3168/jds.S0022-0302(97)76268-4.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

Zaalberg, R. M., L. Janss, and A. J. Buitenhuis. 2020. Genome-wide association study on Fourier transform infrared milk spectra for two Danish dairy cattle breeds. BMC Genet. 21:9. https://doi.org/10.1186/s12863-020-0810-4.

Zaalberg, R. M., N. Shetty, L. Janss, and A. J. Buitenhuis. 2019. Genetic analysis of Fourier transform infrared milk spectra in Danish Holstein and Danish Jersey. J. Dairy Sci. 102:503–510. https://doi.org/10.3168/jds.2018-14464.

ORCIDS

Yansen Chen https://orcid.org/0000-0002-8593-4384
Hadi Atashi https://orcid.org/0000-0002-6853-6608
Hélène Soyeurt https://orcid.org/0000-0001-9883-9047
Nicolas Gengler https://orcid.org/0000-0002-5981-5509