Investigating associations between milk metabolite profiles and milk traits of Holstein cows

N. Melzer,* D. Wittenburg,* S. Hartwig,† S. Jakubowski,† U. Kesting,† L. Willmitzer,‡ J. Lisec,‡ N. Reinsch,* and D. Repsilber*1

*Research Unit Genetics and Biometry, Leibniz Institute for Farm Animal Biology, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany
†Landeskontrollverband für Leistungs- und Qualitätsprüfung Mecklenburg-Vorpommern e.V. (LKV), Speicherstraße 11, 18273 Güstrow, Germany
‡Max Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany

Received May 17, 2012.
Accepted August 22, 2012.
1Corresponding author: repsilber@fbn-dummerstorf.de

ABSTRACT

In the field of dairy cattle research, it is of great interest to improve the detection and prevention of diseases (e.g., mastitis and ketosis) and monitor specific traits related to the state of health and management. During the standard milk performance test, traditional milk traits are monitored, and quality and quantity are screened. In addition to the standard test, it is also now possible to analyze milk metabolites in a high-throughput manner and to consider them in connection with milk traits to identify functionally important metabolites that can also serve as biomarker candidates. We present a study in which 190 milk metabolites and 14 milk traits of 1,305 Holstein cows on 18 commercial farms were investigated to characterize interrelations of milk metabolites between each other, to milk traits from the milk standard performance test, and to influencing factors such as farm and sire effect (half-sib structure). The effect of influencing factors (e.g., farm) varied among metabolites and traditional milk traits. The investigations of associations between metabolites and milk traits revealed groups of metabolites that show, for example, positive correlations to protein and casein, and negative correlations to lactose and pH. On the other hand, groups of metabolites jointly associated with the investigated milk traits can be identified and functionally discussed. To enable a multivariate investigation, 2 machine learning methods were applied to detect important metabolites that are highly correlated with the investigated traditional milk traits. For somatic cell score, uracil, lactic acid, and 9 other important metabolites were detected. Lactic acid has already been proposed as a biomarker candidate for mastitis (Dohoo and Meek, 1982; Schukken et al., 2003). However, the traditional milk traits used as biomarkers for state of health are not sufficiently sensitive in view of diagnostic efficiency (e.g., SCC; Viguier et al., 2009). Even if, for example, acetone is an accepted biomarker for ketosis (Geishauser et al., 2000; Enjalbert et al., 2001), acetone is increased only when the disease is already acute. Klein et al. (2012) reported that no biomarker was available that shows long-term prognostic potential for ketosis, which may hold true for other diseases. They proposed that the milk ratio of glycerophosphocholine to phosphocholine can be used to indicate risk of keto-

INTRODUCTION

In the last years, metabolomics (e.g., Fiehn, 2002; Krastanov, 2010) has played an important role in several research fields, including plant research (Weckwerth, 2003; Saito and Matsuda, 2010) and clinical oncology research (Spratlin et al., 2009), and has received increasing attention from dairy cattle researchers. In this field, biomarkers (for definition, see Biomarkers Definitions Working Group, 2001) for diseases (e.g., mastitis and ketosis) or for the state of health and management need to be found. Increased milk yield is assumed to cause undesirable side effects, such as an increase in health problems (e.g., Rauw et al., 1998). Diseases often reduce milk yield and also lead to additional costs, such as those due to necessary drug treatment, veterinarian costs, and so on (Ingvartsen et al., 2003).
sis. For mastitis, lactic acid was proposed as a potential biomarker (Farr et al., 2002; Davis et al., 2004). It would also be desirable to replace invasive diagnostics such as monitoring of rumen function (Fievez et al., 2003; Vlaeminck et al., 2005) with noninvasive tests, such as a milk metabolome assay. Cabrita et al. (2003) found that levels of odd-chain fatty acids in milk have the potential to noninvasively monitor rumen function, and it was shown that heptadecanoic acid is a possible biomarker candidate for protein deficiency in the feed. Prognostic markers or biomarker candidates are also sought for other diseases and management problems. Hence, it seems promising to find metabolites that could be used as biomarkers to improve diagnostic tools. However, to our knowledge, only a few studies have been published regarding the association between milk metabolite profiles and milk performance test traits or the correlations among metabolites (e.g., Klein et al., 2010). Mostly, the correlation between single or groups of metabolites and single milk traits of interest have been investigated in the recent literature. To our knowledge, analyses regarding multivariate correlations between sets of metabolites and milk traits from the milk performance test, as well as between metabolites, are lacking.

In this study, we present the results of such analyses for 190 milk metabolites and 14 milk traits from a large number of Holstein cows. Eleven milk traits were measured in the standard milk performance test and 3 additional ratios based on the measured milk traits were investigated to additionally cover the status of energy balance. We investigated the factors influencing metabolite levels and milk traits. Here, farm and day of lactation were analyzed in greater detail, because both affect the metabolic state of the cow. The main focus of our analyses was on relationships between metabolites and the milk traits from milk performance tests. These relationships were analyzed taking a univariate relation analysis approach on the one hand, using the Pearson correlation coefficient (Klein et al., 2010). On the other hand, different multivariate analysis methods (e.g., clustering, random forest) were also applied. In the field of dairy science, principal component analysis is primarily applied as a first unsupervised analysis method (Sugimoto et al., 2012). In other fields (e.g., plant science) typically different multivariate analysis methods are used to explore the data. Sugimoto et al. (2012) give an overview of the current state of the art regarding such methods. In our study, correlation structures between the measured part of the milk metabolome and the milk traits are also reported. Detailed results of statistical tests of influencing factors for all of the investigated metabolites are supplied. An atlas of univariate and multivariate relations between metabolites and milk traits and correlations among milk traits is provided in the supplemental files (available online at http://www.journalofdairyscience.org/). In this context, revealed metabolites or groups of metabolites with a significant association to an investigated milk trait could serve as possible candidates for biomarkers or biosignatures. However, to be able to propose concrete biomarker candidates, a suitable study would also have to include the traits of interest, such as disease data. These types of data were not part of our study. Instead, we used milk traits from the performance tests as surrogates. In the discussion we present, among other things, possible functional backgrounds for some of the associations found for specific important metabolites.

**MATERIALS AND METHODS**

**Experimental Data**

The experimental data set comprised milk traits and metabolome measurements from 1,305 Friesian Holstein cows held on 18 farms in Mecklenburg-Western Pomerania (northeastern Germany). The milk samples (1 sample per cow, the volume per milk sample ranged from 30 to 40 mL) were taken (during the monthly standard milk performance test) between d 21 and 120 of the first lactation, over 39 test-days (from May to November 2009). To preserve the milk samples, 150 μL of a 5% solution of sodium azide was added per sample. Milk samples were analyzed via infrared spectroscopy (Kombi-Foss FT6000-FC, Foss, Hillerød, Denmark) in the regional institute for standard milk performance testing (LKV, Güstrow, Germany). The following traditional milk traits were recorded: casein (%), protein (%), fat (%), milk yield (measured in kilograms of milk per day, and hereafter called quantity of milk), SCC (1,000/mL), urea (%), lactose (%), pH, SFA (%), unsaturated fatty acids (UFA, %), and acetone (%). The milk samples were frozen in liquid nitrogen. After finishing the milk sampling collection, the milk samples were sent to the Max Planck Institute of Molecular Plant Physiology (Potsdam-Golm, Germany), where the milk metabolite spectra were obtained via GC-MS of the water-soluble phase of each sample (for details on the method, see Lisec et al., 2006). For the GC-MS application, a specific experimental design was created over the milk metabolite measuring days (GC-MS batches), which was as balanced as possible concerning the following factors to enable an unbiased correction and analyses respectively: sire effect (half-sib structure), farm, and test-day (day of milk sampling); for details, see Melzer et al. (2010). During the preparation of milk metabolite profiles, the experimental design was slightly changed due to the laboratory restrictions,
resulting in 47 GC-MS batches, not 34 as planned. Most of the preplanned block design was maintained. In the GC-MS, the following terms were recorded for each milk sample: the molecule retention time (GC step), the mass to charge ratio, and the corresponding intensities of molecule fragments (MS step). The delivered molecule spectra were then processed with the R package TargetSearch (Cuadros-Inostroza et al., 2009; R Development Core Team, 2011). The retention time of each molecule of a milk sample was converted into a retention index based on the retention time standards of fatty acid methyl esters, which were added to each sample during the GC step. Groups of correlating molecule peaks were determined to build a metabolite spectrum. For each metabolite spectrum, the median over all samples was determined at every peak. The spectrum of median peaks was compared with reference spectra from the Golm Metabolome Database (Kopka et al., 2005; http://gmd.mpimp-golm.mpg.de/search.aspx). A metabolite spectrum was accepted if the similarity index between median metabolite spectrum and reference spectrum was >500 (possible range 0 to 1,000; Cuadros-Inostroza et al., 2009). A metabolite was considered for further analyses if it occurred in more than 80% of the samples, and the log2 intensity at the largest peak was recorded as an individual observation. This resulted in 187 metabolites for which a reference was found in the Golm Metabolome Database, and 3 unknown metabolites, for which no reference could be assigned. Finally, nearest neighbor imputation, as implemented in the R package pcaMethods, was applied to impute missing observations, in total 7% (Stacklies et al., 2007). The obtained data matrix was used for a variety of statistical analyses.

**Statistical Analyses**

The following statistical model was fitted to metabolites as well as to milk traits ($y$):

$$y_{ijkl} = a_{hi} \timesstp_j + gld_k + b_1 \times ltp + b_2 \times ltp^2 + se_l + \varepsilon_{ijkl},$$

where $y_{ijkl}$ = observed trait (dependent variable); $a_{hi}$ = farm ($i = 1, \ldots, 18$); $stp_j$ = test-day ($j = 1, \ldots, 39$); $gld_k$ = GC-MS batch ($k = 1, \ldots, 47$); $ltp$ = day of lactation, where $ltp \in \{21, \ldots, 120\}$; $se_l$ = sire effect ($l = 1, \ldots, 214$); and $\varepsilon$ = residuals.

As fixed effects, we considered the interaction $ah \times stp$, farm, and test-day (63 levels), GC-MS batch, and linear and quadratic regression on day of lactation, where $b_1$ and $b_2$ were the regression coefficients. Further, sire effect was considered as a random effect with $se_l \sim N(0, \sigma^2_e)$ and the residual effect $\varepsilon_{ijkl} \sim N(0, \sigma^2_\varepsilon)$. In case of milk traits, the same model was used, but the factor GC-MS batch was not relevant and therefore excluded.

Further, the sire effect accounted for the half-sib structure. Based on the pedigree data received from the Computing Center -IT Solutions for Animal Production (VIT, Verden, Germany), 192 sires could be assigned to cows, and 22 cows had unknown sires.

On average, each sire had 6.1 daughters (range between 1 and 112 cows), each farm had 72.5 cows (range between 36 and 135 cows), and each lactation day had 13.05 cows (range between 1 and 27). The values of lactation day were grouped into 10-d intervals, termed as lactation interval, resulting in the following intervals: lactation d 21–30 = lactation interval 1, lactation d 31–40 = lactation interval 2, and so on, to lactation d 111–120 = lactation interval 10. On average, 130.5 milk samples (ranging from 55 to 183) were analyzed per lactation interval.

The effect of an influencing factor on metabolite profiles or on milk traits was studied with the following statistical tests. For the fixed effects, an F-test (ANOVA) was applied in sequence to all traits. For the random effect, a one-sided likelihood ratio test (LRT) was applied. The testing problem was $H_0: \sigma^2_v = 0$ versus the alternative hypothesis $H_A: \sigma^2_v > 0$. The distribution of the LRT statistic under the null hypothesis approximately followed a mixture of $\chi^2$ distributions according to Self and Liang (1987). To investigate a specific influencing factor, metabolite measurements were corrected over all animals for all factors except for the one of interest. The observed $P$-values were corrected, because of multiple testing, using the false discovery rate (FDR) controlling method by Benjamini and Hochberg (1995), fixing the estimated FDR at 5%. We applied the FDR correction as implemented in the R package multtest (Pollard et al., 2010). This correction method was applied to all tests (e.g., for each influencing factor and for milk traits).

To investigate if metabolite intensities differed significantly between levels of an influencing factor of interest, standardized residuals of the linear model [Eq. 1] were used. To compare means pairwise for levels of categorical factors for a metabolite, we applied the Tukey test (Kramer, 1956), using R package DTK (Lau, 2011). Also, for day of lactation, it was possible to prove visually if an increasing or decreasing trend existed for metabolites over lactation days, based on the ANOVA test. For this, the estimated regression coefficients $\hat{b}_1$ and $\hat{b}_2$ from the full linear model [Eq. 1] were applied to the following linear model:

$$\hat{y} = \hat{b}_1 \times ltp + \hat{b}_2 \times ltp^2,$$
to estimate metabolite \( y \). All analyses were implemented using R (R Development Core Team, 2010); analysis scripts are available upon request.

**Multivariate Analyses Investigating Specific Influencing Factors on Metabolites**

To investigate a specific influencing factor, metabolite measurements for all animals were corrected for all influencing factors except for the one of interest according to the linear model [Eq. 1], and the standardized residuals were used in the following analyses.

**Clustering of Metabolite Profiles Regarding Influencing Factors.** For this purpose, the mean over all samples on a specific level of the influencing factor under investigation was taken for each metabolite. The Euclidean distance between vectors of mean metabolite measurements was used to determine similarities between metabolite profiles for the levels of the investigated influencing factor. Hierarchical clustering using the method of average linkage was applied. Two validation criteria were used to evaluate the number of clusters that had to be determined in advance. The silhouette width criterion (Rousseeuw, 1987; Vendramin et al., 2009) provided information about compactness and separation of clusters. The stability of clusters (Hennig, 2007) was calculated using the function clusterboot (R package FPC; Hennig, 2010). This function assesses the clusterwise stability of clustering the data by resampling; the number of bootstrap rounds was 1,000. Therein, the Jaccard coefficient (Jaccard, 1901; Vendramin et al., 2009) was used as a similarity measure.

**Classifying Levels of Influencing Factors.** Linear discriminant analysis (LDA; Fisher, 1936; Hastie et al., 2009, p. 106–112) was used to investigate multivariate relations of standardized metabolite profiles to the influencing factor of interest using R package MASS (Venables and Ripley, 2002). On the one hand, the estimated discriminant function was used to determine important metabolites for specific factors as follows. A 10-fold cross-validation (Hastie et al., 2009, p. 241–249) was implemented. In each cross-validation run, the coefficients of the first linear discriminant function (LD1) were recorded for all metabolites. The coefficients were used as a measure of association between each metabolite and the investigated influencing factor. We defined a metabolite to be important if its coefficient was larger than the 90% quantile of the coefficients for all metabolites in each cross-validation run. On the other hand, the estimated discriminant function was also used to classify new data taking the 10-fold cross-validation approach, to quantify the strength of association and its significance. The precision of prediction was determined as the proportion of correctly classified samples reporting the mean based on 10 runs.

To quantify significance of the observed prediction ability, we applied a resampling approach to randomly destroy the possible association between factor levels and metabolite profiles. The number of resampling rounds was 1,000, and a 10-fold cross-validation was applied in each resampling round. The resampling P-value of the observed prediction precision was determined as usual for a permutation test (Good, 2005) as the relative proportion of resampling rounds with a prediction precision as large or larger than the observed precision of prediction for the original data.

**Analyses of Milk Traits Related to Metabolites**

We studied 11 milk traits measured in the standard milk performance test: casein (%), protein (%), fat (%), quantity of milk (kg), SCC (1,000/mL), urea (%), lactose (%), pH, SFA (%), UFA (%), and acetone (%), and 3 additional milk traits characterizing the status of energy balance: fat:protein ratio, fat:lactose ratio, and energy content of milk. Energy content of milk (E, MJ/kg) was determined as follows (Kirchgessner, 1982, p. 270):

\[
E = 0.39 \times \text{fat }\% + 0.24 \times \text{protein }\% + 0.17 \times \text{lactose }\%.
\]

Somatic cell count was transformed to SCS following Ali and Shook (1980): SCS = log2 (SCC/1,000) + 3.

Metabolites and milk traits were corrected for the influencing factors as modeled in equation [Eq. 1]. The observed residuals were standardized and used in the subsequent analyses presented here. The R function cor.test was applied, using Pearson correlation coefficient, to test correlations between paired samples of metabolites and milk traits, between milk traits, and between metabolites.

The regression methods random forest (RF; Breiman, 2001) and partial least squares (PLS; Wold, 1975; Hastie et al., 2009, p. 80–82) were used to predict milk traits from metabolite profiles. Procedures were used as implemented in the R packages randomForest (Liaw and Wiener, 2002) and mixOmics (Dejean et al., 2011) for PLS. Further, a 10-fold cross-validation was implemented to determine the precision of prediction (\( \rho \)), which is defined as the correlation between predicted and observed values of a milk trait. Additionally, for PLS, it was necessary to determine the number of latent components to achieve a minimal prediction error, which was measured as the mean squared error of prediction. To determine an optimal number
of latent components for prediction in each (outer) cross-validation run, an inner 10-fold cross-validation was implemented. The vip function of the R package mixOmics was used to extract metabolite importance for PLS. In RF, after finishing the (outer) 10-fold cross-validation runs, a resulting mean decrease in accuracy was used as a measure of importance of metabolites. To determine the metabolites important for predicting a milk trait for each prediction method, we used the 90% quantile of the importance measurements of all metabolites in each cross-validation run, and defined a metabolite to be important if its importance measurement was larger than this 90% quantile in each of the 10 cross-validation runs.

RESULTS

**Significant Influencing Factors for Metabolites and Milk Traits**

Numbers and percentages of metabolites and traditional milk traits, for which the analyzed influencing factors were significant (based on ANOVA and LRT), are listed in Table 1. The factors showed a higher percentage of significance for milk traits than for metabolites; for example, day of lactation had a significant influence on 45.79% of the metabolites and on 85.71% of the milk traits. More than 85% of all traditional milk traits were significantly influenced by all influencing factors. A detailed list of the observed corrected \( P \)-values can be found in Supplemental Table S1 (available online: http://www.journalofdairyscience.org/).

**Effect of Specific Influencing Factors on Metabolites**

**Clustering of Average Metabolite Profiles.** In Figure 1, we present the observed dendrogram for the influencing factor lactation interval. The lactation intervals can be split into 3 clusters: cluster 1 comprised lactation intervals 1 to 3 (d 21–50), cluster 2 comprised intervals 4 to 8 (d 51–100), and cluster 3 comprised intervals 9 to 10 (d 101–120). The number of clusters was evaluated using the silhouette width criterion, for which the maximum average silhouette width was determined at 0.30. For the analysis involving 3 clusters, the cluster stability criterion was applied, and we observed a Jaccard coefficient >0.68 for all 3 clusters, indicating intermediate stability. The shown dendrogram reproduced the lactation timeline.

![Figure 1](image.png)

For the influencing factor farm, the observed dendrogram is shown in Figure 2a. The chosen number of clusters was 3, which was proved by the silhouette width criterion (average silhouette width = 0.37) and the Jaccard coefficient, which was >0.85 for all 3 clusters, indicating high cluster stability.

**Pairwise Statistical Tests.** We tested how many metabolites showed a significant difference between levels of an investigated influencing factor. For the influencing factor lactation interval, all 45 pairwise comparisons were analyzed and the greatest number of differences was found between lactation intervals 1 and 7 or 8, respectively. Metabolites with the highest numbers of significant differences for all pairwise comparisons of lactation intervals are listed in Table 2.

For the influencing factor farm, 153 pairwise comparisons were investigated, of which 66 showed significant differences between metabolites. The number of significant differences for each farm compared with all other farms is graphically presented in Figure 2b: the largest number of significant differences was found between farms 10 and 15, 16 metabolites in total, whereas no metabolite differed significantly, for example, between farms 6 and 12. The observed number and corresponding relative percentage of significant differences are reported for the metabolites with the highest counts in Table 2.

---

**Table 1.** The observed number of metabolites and milk traits (relative proportions in parentheses), on which influencing factors had a significant effect.

<table>
<thead>
<tr>
<th>Influencing factor</th>
<th>Metabolites</th>
<th>Milk traits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sire</td>
<td>34 (17.80)</td>
<td>13 (92.86)</td>
</tr>
<tr>
<td>Day of lactation</td>
<td>87 (45.79)</td>
<td>12 (85.71)</td>
</tr>
<tr>
<td>Farm</td>
<td>145 (76.32)</td>
<td>13 (92.82)</td>
</tr>
<tr>
<td>Test-day</td>
<td>105 (53.62)</td>
<td>12 (85.71)</td>
</tr>
<tr>
<td>Farm × test-day</td>
<td>159 (83.68)</td>
<td>14 (100.00)</td>
</tr>
<tr>
<td>GC-MS batch</td>
<td>190 (100.00)</td>
<td>14 (100.00)</td>
</tr>
</tbody>
</table>
Classifying Levels of Specific Influencing Factors. We used LDA to determine important metabolites for the prediction of an influencing factor. Important metabolites were derived from LD1 coefficients. For lactation interval, we observed 11 important metabolites, and for farm we found 7 important metabolites. The most important metabolites for both factors are listed in Table 3. Also, we applied LDA to predict the influencing factor of interest from metabolite profiles.

The observed precision of prediction for the original data was significant (resampling $P$-value $\leq 0.001$) for each influencing factor.

**Analyses of Milk Traits**

**Testing Correlation.** Investigating correlations between metabolites, we observed that in total 80% of all pairwise correlations were significant after FDR correction at a level of 5%. The highest correlations were found between ethanolaminephosphate and orotic acid, and between 2-methyl-fumaric acid and itaconic acid ($\rho > 0.9$ for these pairs). Most metabolites were positively correlated. A detailed correlation matrix is available as supplementary material (Supplemental File 2; http://www.journalofdairyscience.org/).

The significant correlations between metabolites and milk traits are presented in Figure 3. In total, 75 metabolites showed no significant correlations with any milk trait. We observed that for casein and protein, metabolites mostly coincided regarding significance and degree of correlation. This finding is further illustrated in Figure 3, where parts 2 and 5 of the correlation structure show that these 2 traits mainly correlated with the same metabolites, whereas Figure 3 part 3 shows less congruence between both milk traits. Further, we observed that metabolites that were clearly positively correlated with SCS were clearly negatively correlated with lactose (Figure 3 part 3). Investigating the correlation structure between milk traits, we found 81% of all pairwise correlations significant but with very different levels of correlation (Figure 4).

Casein and protein, energy and fat, energy and fat:lactose, and fat and fat:lactose were highly correlated ($\rho > 0.9$). A positive correlation was observed, for example, between fat and protein, and a negative correlation was observed, for example, between fat and lactose as well as between urea and casein or protein, respectively.

The corresponding Pearson correlation matrix for investigations between metabolites and milk traits, and among milk traits can be found in Supplemental File 3 (http://www.journalofdairyscience.org/).

**Important Metabolites with Multivariate Relations to Milk Traits.** The observed precisions for predicting the investigated milk traits from metabolites using RF and PLS are presented in Figure 5. Both methods resulted in very similar precisions for the prediction of most milk traits. The highest mean value of prediction precision was observed for SCS ($\rho = 78\%$, RF). Casein and protein had similar mean prediction precisions above 60%. For all other milk traits, mean prediction precision was observed between 17 and 41%.
Saturated fatty acids had the lowest prediction precision.

For predicting SCS using RF or PLS, a very large importance was observed for one metabolite (uracil), a few metabolites showed small importance, and most had almost zero importance for this milk trait. This behavior was stable over the cross-validation runs. For fat, more metabolites showed greater importance for both prediction methods. The results of metabolite importance measurements varied strongly over the cross-validation runs. This finding was also observed for the remaining milk traits. The most important metabolites for each milk trait that coincided in both prediction methods are listed in Table 4. Comparing the detected important metabolites via RF and PLS (Table 4) to metabolites that showed significantly large correlation values for the investigated milk traits (Figure 3) using the univariate evaluation resulted in good congruence. The important metabolites are listed separately (Supplemental File 4; http://www.journalofdairyscience.org/) for each prediction method and with respect to their importance for the investigated milk traits. Comparing both statistical learning methods, more important metabolites were observed with PLS than with RF for most milk traits. For casein and protein, we found almost the same set of important metabolites with both prediction methods. In total, we found 14 important metabolites that were detected by both prediction methods for casein and protein. In addition, 13 important metabolites coincided between casein and protein, and 11 important metabolites coincided between protein and fat prediction. The most important metabolites for SCS, fat, and protein prediction are listed in Table 4.

Table 2. The observed number of metabolites with the most significant differences between levels of an investigated factor of interest

<table>
<thead>
<tr>
<th>Influencing factor</th>
<th>Metabolite</th>
<th>No. of observations (relative proportion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lactation interval</td>
<td>Glucopyranoside, 1-O-methyl-α</td>
<td>26 (57.78)</td>
</tr>
<tr>
<td>Glucosamine, N-acetyl</td>
<td></td>
<td>26 (57.78)</td>
</tr>
<tr>
<td>Ribose-5-phosphate</td>
<td>25 (55.56)</td>
<td></td>
</tr>
<tr>
<td>Gluconic acid-6-phosphate</td>
<td></td>
<td>25 (55.56)</td>
</tr>
<tr>
<td>Fructose-6-phosphate</td>
<td>24 (53.33)</td>
<td></td>
</tr>
<tr>
<td>Phospho-6-phosphate</td>
<td>21 (46.67)</td>
<td></td>
</tr>
<tr>
<td>Sarcosine</td>
<td>20 (44.44)</td>
<td></td>
</tr>
<tr>
<td>Galactosamine, N-acetyl</td>
<td></td>
<td>18 (40.00)</td>
</tr>
<tr>
<td>Arabinose</td>
<td>17 (37.78)</td>
<td></td>
</tr>
<tr>
<td>Gluconic acid</td>
<td>17 (37.78)</td>
<td></td>
</tr>
<tr>
<td>Farm</td>
<td>Benzoic acid</td>
<td>35 (22.88)</td>
</tr>
<tr>
<td>Kynurenine</td>
<td>17 (11.11)</td>
<td></td>
</tr>
<tr>
<td>1,3-Dihydroxyacetone</td>
<td>11 (7.19)</td>
<td></td>
</tr>
<tr>
<td>Butanoic acid</td>
<td>10 (6.54)</td>
<td></td>
</tr>
<tr>
<td>Pyridoxal</td>
<td>10 (6.54)</td>
<td></td>
</tr>
<tr>
<td>Thiazole, 4-methyl-5-hydroxyethyl</td>
<td></td>
<td>10 (6.54)</td>
</tr>
<tr>
<td>Arabinose</td>
<td>8 (5.23)</td>
<td></td>
</tr>
<tr>
<td>Phosphatidinic acid</td>
<td>8 (5.23)</td>
<td></td>
</tr>
</tbody>
</table>

1The total number of possible pairwise comparisons per metabolite for each influencing factor is given in brackets.

Table 3. Important metabolites detected for the factors lactation interval and farm, using a 10-fold cross-validation in a linear discriminant analysis

<table>
<thead>
<tr>
<th>Lactation interval</th>
<th>Farm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adipic acid, 2-amino</td>
<td>2-Piperidinecarboxylic acid</td>
</tr>
<tr>
<td>Arabinose</td>
<td>Adipic acid, 2-amino-</td>
</tr>
<tr>
<td>Arginine [−NH₃]</td>
<td>Aspartic acid</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Butyric acid, 2,4-diamino- , DL-</td>
</tr>
<tr>
<td>Gluconic acid</td>
<td>Galactosamine, N-acetyl</td>
</tr>
<tr>
<td>Glycerol</td>
<td>Glycerol-2-phosphate</td>
</tr>
<tr>
<td>Orotic acid</td>
<td>Ribose-5-phosphate</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Unknown 315800</td>
</tr>
</tbody>
</table>

1Coefficient of the linear discriminant function (LD1) was used as a measure of association between metabolites and investigated influencing factor. A metabolite was declared important if the corresponding coefficient was above the 90% quantile in each cross-validation run. Metabolites in bold were detected in the univariate analysis (pairwise statistical test).
protein; for example, arabitol, 2-amino adipic acid, and pyroglutamic acid. Urea is an example of disagreement between RF and PLS, for which 5 metabolites were identified as important by each method, but only aspartic acid and 2-amino-adipic acid coincided between them. In the following, we only consider metabolites detected as important by both prediction methods.

For SCS, the most important metabolites were uracil and lactic acid. A positive correlation existed between uracil and SCS (70.26%), between lactic acid and SCS (57.80%), and between both these metabolites (62.26%). Comparing important metabolites for the prediction of lactose and SCS, 5 important metabolites coincided: 1,3-dihydroxyacetone, leucine, methionine, phenylalanine, and tyrosine. Between SCS and lactose, we observed a negative correlation ($\rho = -0.28$), which was also mirrored in the observed correlation structure between metabolites and milk traits (Supplemental File 3; http://www.journalofdairyscience.org/). In detail, SCS (lactose) showed the following correlation values for 1,3-dihydroxyacetone, leucine, methionine, phenylalanine, and tyrosine. Between SCS and lactose, we observed a negative correlation ($\rho = -0.28$), which was also mirrored in the observed correlation structure between metabolites and milk traits (Supplemental File 3; http://www.journalofdairyscience.org/).

**Figure 3.** Correlations between metabolites and milk traits; white squares represent nonsignificant correlations. Seventy-five metabolites had no significant correlation to any investigated milk trait and were excluded. Milk traits and metabolites were clustered using hierarchical clustering with Euclidean distance and complete linkage.

**Figure 4.** Correlations between milk traits; white squares represent nonsignificant correlations.
In this context, we found the best coinciding correlation values for casein and protein. Additionally, 1,3-dihydroxyacetone was observed as important for fat, energy, fat:protein, fat:lactose, and SFA. Finally, we did not observe any coinciding important metabolites for fat, UFA, and SFA.

DISCUSSION

More than 2,000 metabolites are thought to exist in cow’s milk (Töpel, 2004, p. 3). In our study, we analyzed 190 milk metabolites in regard to their correlation structure and associations with 14 milk traits in 1,305 Holstein cows. These measurements represent a snapshot of the current metabolic state of this economically important breed. We found significant associations among metabolites, among milk traits, and between metabolites and milk traits, and report the effect of influencing factors on both.

Compared with traditional milk traits, we showed that influencing factors generally had less effect on metabolites. The detected important metabolites for an investigated milk trait using machine learning methods coincided well with the significant metabolites that show the largest correlation values for the investigated milk trait using univariate analysis. Applying such methods allows the user to find metabolite patterns that are similar for different milk traits.

Relevance of Results

Single Metabolite Versus Milk Trait. Influencing factors considered in our study were farm, sire effect, GC-MS batch, day of lactation, and test-day, which had significant effects on some of the 14 milk traits and 190 milk metabolites. Influencing factors were significant for a larger percentage of milk traits compared with metabolites (Table 1). Single metabolite profiles seemed to be only partly affected by influencing factors in our study, whereas complex traits, such as the investigated milk traits, showed larger dependencies. As a possible explanation, it should be considered that complex traits are composed of several molecular traits, so that at least some of these are likely to be affected by the considered influencing factor. On the other hand, metabolites as molecular traits are highly interconnected (and correlated), making it difficult to find simple explanations for the observed selective effect of the influencing factors.

Influence of Lactation Day. A lactation curve was expected to exist for some metabolites, such as for the complex milk trait quantity of milk. Therefore, we used ANOVA to investigate if metabolites showed a significant trend over lactation days. In total, we found 78 metabolites that had such a trend. As we collected milk samples only between d 21 and 120 of lactation, only a part of the possible lactation curve was observed. The

Figure 5. Boxplots of precisions for the prediction of milk traits from metabolite profiles using random forest and partial least squares based on a 10-fold cross-validation. UFA = unsaturated fatty acids. Color version available in the online PDF.
lactation curves for metabolites are probably blurred because all measurements originated from different individuals (1 sample per cow). The cluster analysis (Figure 1) also mirrored a clear change in metabolite profiles during the observed part of lactation. Comparing the lactation intervals for metabolites, we found that the first interval showed the most significant differences compared with the other intervals. This finding could be based on the known course of negative energy balance observed in the early phase of lactation (e.g., Bauman and Currie, 1980; Collard et al., 2000).

In Table 2, we present those metabolites that had the most significant differences over all pairwise comparisons. These metabolites showed the greatest variation over the investigated lactation intervals. A clearer statement of changes and compositions of metabolites could be given if multiple samples from the same cows were taken over the whole lactation.

Influence of Farm. We detected significant differences in metabolite levels between farms (Figure 2b). It is known from the literature that, among other factors (e.g., keeping conditions, milking, and management), feeding has an important influence on milk composition (Sutton and Morant, 1989) and that odd-chain fatty acids in milk have a high potential to discriminate between different diets (e.g., Cabrita et al., 2003). In this context, we identified 2 metabolites, arabitol and 1,3-dihydroxyacetone (Table 2), as being important for discriminating between farms in our data. Both are carbohydrates, implicating their role as a precursor of fat. Considering these findings together with the results of our metabolite clustering of farms (Figure 2a), we assumed that feeding was the most discriminating aspect in our study. In addition, some farms showed very distinct metabolite profiles and thus formed separate clusters. However, it is necessary to consider possible different feeding managements and other influencing aspects for the factor farm in greater detail and to investigate possible relations to our results.

Univariate Versus Multivariate Analysis of Influencing Factors. The multivariate analysis (LDA) revealed additional important metabolites compared with univariate analyses for the investigated influencing factors (Table 3). Important metabolites detected by univariate as well as by multivariate analyses are marked in bold in Table 3 for both influencing factors. Multivariate analyses also consider interactions between metabolite profiles and, hence, can detect additional related features compared with the univariate approach.

Important Metabolites with Multivariate Relations to Milk Traits. To examine the eligibility of metabolites as possible biomarker candidates, we analyzed how well traditional milk traits could be predicted based on metabolites. To find out which possible combinations of metabolites best explained variation in traditional milk traits, we applied 2 statistical learning approaches (RF and PLS). Figure 5 shows that the milk traits SCS, casein, and protein could be predicted confidently from metabolite profiles, in contrast to the other 11 milk traits. This finding indicates that the metabolites in our data pool included more metabolites relevant for these 3 milk traits. This conclusion was further supported by the fact that a large number

### Table 4. For each milk trait, the observed important metabolites are listed in alphabetical order

<table>
<thead>
<tr>
<th>Milk trait</th>
<th>Important metabolites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetone</td>
<td>Ethanolaminephosphate; Glucaric acid-1,4-lactone; Orotic acid</td>
</tr>
<tr>
<td>Casein</td>
<td>2-Piperidinedecarboxylic acid; Adipic acid, 2-amino-; Alanine; Arabitol; Asparagine;</td>
</tr>
<tr>
<td>Fat</td>
<td>Arabitol; Aspartic acid; Galactitol; Glucaric acid-1,4-lactone; myo-Inositol-1-phosphate; Pyroglutamic acid</td>
</tr>
<tr>
<td>Lactose</td>
<td>1,3-Dihydroxyacetone; Glucaric acid-1,4-lactone; Leucine; Methionine; Phenylalanine;</td>
</tr>
<tr>
<td>pH</td>
<td>Alanine, β-; Glyceraldehyde-2-phosphate; Glyceraldehyde-3-phosphate; Glycine</td>
</tr>
<tr>
<td>Protein</td>
<td>myo-Inositol-1-phosphate; Phosphenolpyruvic acid; Pyroglutamic acid; Spermidine; Thiazole; 4-methyl-5-hydroxyethyl-</td>
</tr>
<tr>
<td>Quantity of milk</td>
<td>Arabitol; Butanoic acid, 2-amino-; Butanoic acid, 4-methylthio-2-oxo-; 2-Piperidinedecarboxylic acid</td>
</tr>
<tr>
<td>SFA</td>
<td>1,3-Dihydroxyacetone; Glycerol</td>
</tr>
<tr>
<td>SCS</td>
<td>1,3-Dihydroxyacetone; Butanoic acid, 2-hydroxy-; Lactic acid, δ-t; Leucine; Methionine; Phenylalanine;</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Tryptophan; Tyrosine; Uracil</td>
</tr>
<tr>
<td>Unsaturated fatty acids</td>
<td>Galactitol; Serine, cyclo-</td>
</tr>
<tr>
<td>Urea</td>
<td>Adipic acid, 2-amino-; Aspartic acid</td>
</tr>
<tr>
<td>Energy</td>
<td>1,3-Dihydroxyacetone; Arabitol; Aspartic acid; myo-Inositol-1-phosphate; Pyroglutamic acid</td>
</tr>
<tr>
<td>Fat:protein ratio</td>
<td>1,3-Dihydroxyacetone; Butanoic acid, 3-hydroxy-; Galactitol; Glyceraldehyde-2-phosphate; Glyceraldehyde-3-phosphate;</td>
</tr>
<tr>
<td>Fat:lactose ratio</td>
<td>1,3-Dihydroxyacetone; Arabitol; Galactitol; Glucaric acid-1,4-lactone; Pyruvic acid; Pyroglutamic acid</td>
</tr>
</tbody>
</table>

*The important metabolites exceeded the 90% quantile for both prediction methods (random forest and partial least squares) in all 10-fold cross-validation runs.*
of important metabolites was detected for these milk traits with both prediction methods (Supplemental File 1; http://www.journalofdairyscience.org/). For casein and protein, we observed 14 important metabolites that coincided between both methods (Table 4). Further, 13 of these important metabolites also coincided between casein and protein. We expected that nearly the same metabolites would be detected for both milk traits, because casein is a main component of milk protein, constituting over 80% of its mass (Töpel, 2004, p. 226; Madureira et al., 2007). The similarity of both milk traits is also mirrored in Figure 3. For casein and protein, the amino acids asparagine and aspartic acid, among others, were detected as important. It is known that amino acids are involved in protein metabolism. It was surprising that only these 2 amino acids were detected as important for the prediction of both milk traits, taking into account that the measured metabolites included 12 proteinogenic amino acids in total. Alanine was found to be important for casein, but was only detected as important by one prediction method for protein. Our further investigations revealed that these 3 metabolites were reproducibly found, in contrast to other metabolites. The remaining 9 amino acids also showed correlations, but other metabolites showed stronger correlations and, therefore, were included for our list instead. Spermidine, a polyamine, was also detected to be important for protein. Spermidine is involved in the production of protein (Sanguansermsri et al., 1974; Bolander and Topper, 1979; Löser, 2000). However, the correlation detected for spermidine and protein was rather low ($\rho = 0.28$), indicating a possible large variation either over the course of lactation or between individuals, as proposed by Motyl et al. (1995).

Somatic cell count is a measure to monitor udder health and is performed as a specific test for mastitis (Schukken et al., 2003; Koivula et al., 2005). In our study, we found uracil as the most important metabolite for predicting SCS, which makes it interesting to investigate this relation further because elevated levels of uracil were already proposed to originate from damaged cells (Bi et al., 2000), and it is conceivable that such damage would occur during infection. The second important metabolite in our study was lactic acid. Farr et al. (2002) and Davis et al. (2004) proposed that lactic acid can be used as a biomarker for mastitis at an early stage. Furthermore, we detected the following amino acids as related to SCS: tyrosine, methionine, phenylalanine, lencine, and tryptophan. Tryptophan has an effect on T-cell proliferation (Frumento et al., 2002; Denis et al., 2007) and, hence, plays a role in the immune system. Additionally, these amino acids, except tryptophan, were also detected as important for the prediction of lactose in our data set. A significant negative correlation was observed between lactose and SCS. This observation agreed with results found in the literature; for example, Klein et al. (2010) and Miglior et al. (2007). It is also known that an increased SCC is associated with an increase in lactose (Harmon, 1994). Corresponding to these findings, correlations between common important metabolites and these traits were negative (compare first and last columns of the heatmap in Figure 3).

Amino acid levels are related to lactose levels because amino acids are precursors of $\alpha$-lactalbumin, which is an important protein to catalyze the synthesis of lactose (Brew et al., 1968). To predict levels of SFA in our data, 1,3-dihydroxyacetone and glycerol were detected as important. Glycerol is important to build milk fat (long chains) in the mammary glands (Luick, 1961).

**Benefits and Constraints of Methods**

Our experimental design involved a large number of Holstein cows (1,305) as a minimal requirement for an additional study investigating associations to genomewide SNP marker data. Multiple measurements per individual cow would be desirable to enable the dissection of different sources of variance. Promising candidates for biomarkers, however, should show high correlations between trait and biomarker, also without multiple measurements per individual, as we have found for some cases of the investigated milk traits.

In investigations concerning the influence of factors on metabolite profiles, the metabolite measurements were always corrected for all factors except the one for which the influence was explored. We could apply a classical LDA because our study design involved sufficiently more samples than features (metabolites); LDA is also appropriate for unbalanced data (Xue and Titterington, 2008). To reveal groupings of the farms regarding mean metabolite profiles, we applied a classical hierarchical clustering involving average linkage. Clustering methods always carry uncertainty regarding the number of clusters. Therefore, it is recommended to apply 2 kinds of criteria, a relative validity criterion and an external criterion (Vendramin et al., 2009).

For the factors investigated in our case, these were silhouette score and Jaccard index. Both criteria concordantly indicated 3 stable clusters for all 18 farms. Results of clustering (Figure 2a) were concordant with results for all pairwise comparisons (Figure 2b). This means that, depending on the similarity between farms in the cluster structure, we observed an increasing number of significant differences between metabolites as the similarity between farms decreased. The same was observed for the influencing factor lactation interval. Based on our results for predicting milk traits from
metabolites, we could observe similar sets of important metabolites using 2 regression methods (RF and PLS). A general threshold for measurements of milk metabolite importance (in our case, we applied the 90% quantiles) cannot be given, because it depends on the investigated trait and measured metabolites. For example, regarding the milk trait protein, a large number of metabolites showed high correlation with this trait, whereas for urea, a much smaller number of relevant metabolites was found. To search for probable links to known biological functions in further analyses, we used the intersection of sets of important metabolites of both RF and PLS. In most cases, we were able to relate information about biological functions of the important metabolites to the respective investigated milk trait (see previous section). One of the detected important metabolites (lactic acid) is already proposed as a potential biomarker in the recent literature. Thus, metabolite profiles seem eligible as new molecular traits and can be investigated as candidate biomarkers. In our study, GC-MS analyses measured only part of the milk metabolome because predominantly short-chain water-soluble metabolites of the energy metabolism were detected. Other techniques that also explicitly monitor soluble metabolites of the energy metabolism were metabolome because predominantly short-chain water-soluble metabolites of the energy metabolism were detected. Other techniques that also explicitly monitor the fatty acids and other fat-soluble metabolites could result in valuable complements.

Outlook

Metabolite profiles proved to be promising new molecular traits eligible as candidate biomarkers or to be used in groups of important metabolites (biosignatures). In this context, it is necessary to further elucidate their physiological role and validate our revealed important metabolites with another data set. Another aspect to be considered is that many of the measured metabolites are used for synthesis of milk components by the alveolar epithelial cells or are involved in intracellular metabolism, which leads to the question of why they are measured in milk. Further, to prove the suitability of metabolites as biomarker candidates, it is necessary to investigate the degree to which they change during the course of individual lactations. Taking multiple samples per individual cow during lactation would help to discriminate technical as well as biological sources of variation.

As far as mastitis is concerned, for example, it could be conceived that SCS, as a traditional biomarker, monitors a different aspect of the disease from a related metabolite (set), as is already known for lactic acid, which specifically shows a relation to the early onset of mastitis (Farr et al., 2002). It could also be beneficial to further investigate the correlation structure (e.g., partial correlations or mutual correlations) of the metabolite profiles to reveal possible associations or functional grouping structures. Such structures could be further related to a priori functional knowledge; for example, metabolite pathways or flux modes. Findings from such investigations could further illuminate the functional basis of candidate biosignatures and help improve learning algorithms that exploit biosignatures that are able to use such a priori information (e.g., Sæbø et al., 2008).

Deeper investigations regarding the influencing factor farm would allow for a comparison of different management systems and explain how these are mirrored in differences in metabolite profiles, probably revealing other important factors. The resulting consequences could lead to improvements in the field of farming.

ACKNOWLEDGMENTS

This study was part of the FUGATO plus project “Bovine Integrative Bioinformatics for Genomic Selection (BoVIBI)” with financial support of the German Federal Ministry of Education and Research (BMBF). The authors thank the participating farm owners and cooperation partners. The contribution of A. Eckardt (Max Planck Institute for Molecular Plant Physiology, Potsdam-Golm, Germany), who established the GC-MS measuring procedure for milk samples and measured the metabolite profiles, is gratefully acknowledged. S. Wolf (LKV, Güstrow, Germany) provided milk performance test data and contributed to discussions during the course of this project. We acknowledge the support of F. Reinhardt (vit Verden, Germany), who assembled the pedigree data. Special thanks to R. Grahl, our colleague at the Leibniz Institute for Farm Animal Biology (Dummerstorf, Germany), who collected blood and milk samples.

REFERENCES


