Quantifying Feed Efficiency of Dairy Cattle for Genome-wide Association Analysis

Tingyang Xu[1], Jiangwen Sun[1], Erin E Connor[2], Jinbo Bi[1*]
[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA
[2]Animal Genomics and Improvement Laboratory, USDA, Agricultural Research Service, Beltsville, MD, USA
{tix11001,javon}@engr.uconn.edu, erin.connor@ars.usda.gov, jinbo@engr.uconn.edu

***Abstract*—Improving feed efficiency in dairy production is an important endeavor, as it can reduce feed costs and negative impacts of production on the environment. Feed efficiency is a multivariate phenotype that is characterized by a variety of phenotypic variables, such as dry matter intake, body weight gain, and milk yield. Currently, there is no consensus method for quantifying the feed efficiency of lactating dairy cattle for the purpose of breeding selection. Residual feed intake, which is the difference between actual feed intake and predicted intake, has been one of the commonly used measures for feed efficiency. However, such a measure is heterogeneous showing substantial variation in the cow population and has relatively low heritability (0.01∼0.38). Hence, its utility in breeding selection is limited. In particular, no prior study has utilized genetic data directly in the development of feed efficiency measures. In this paper, we aim to identify cattle clusters with homogeneous feed efficiency features that are ready to link to genetic variants, and thus can have greater utility in breading selection. In order to achieve this goal, we explore a new multi-view clustering method that jointly analyzes two views of data: phenotypic measures and genotypes, and identifies cattle clusters that are characterized by specific phenotypic features and also associated with genetic markers. Using a set of feed efficiency data collected by USDA, three cattle subgroups have been identified by our analysis, and they offer instructive insights into future feed efficiency studies.**

***Keywords*—dairy cow; feed efficiency; genotype-phenotype analysis; residual feed intake; multi-view biclustering**

I. Introduction

There have been ever rising feed costs and concerns about greenhouse gas emissions and nutrient losses to the environment associated with animal production. Identifying the most efficient dairy cattle for milk production is important [1]. The ability to select individuals in animal breeding programs based on genotypic information circumvents the costly process of progeny testing and reduces the generation interval. As genetic selection programs advance, complex phenotypes such as feed efficiency have become the selection target [1], [2]. New challenges thus arise to determine the right phenotypic measure to be used in the selection program. Feed efficiency of cattle has been measured in different ways, such as by dry matter intake (DMI) [3], residual feed intake (RFI) [4], or relative growth rate [5]. The current estimates of their genetic parameters indicate that these measures should respond to selection pressure for improved feed efficiency [4]. However, feed efficiency in dairy cattle is a multivariate phenotype determined by several component traits including the production of milk, milk composition, feed intake, maintenance requirements, and change in body energy reserves. Each of the existing measures is defined by a function of the various traits. To date, genetic selection strategies for improving a single measure such as selection for decreased DMI, or selection for animals that consume less feed than expected based on their energy requirements (i.e., negative RFI) [6], have had difficulties to incorporating the measures into industry-wide selection indices [3].

Characterization of the genetic architecture of feed efficiency typically uses a combined analysis of multiple measures that contribute to the trait. For example, in order to determine the genetic effects on RFI, DMI, change in body energy reserves, and net energy used for milk production, researchers have built Bayesian models for all of the measures. Then, genomic regions with greatest effects for each measure were investigated to identify regions with pleiotropic effects and potential positional candidate genes [3]. However, the existing measures are heterogeneous showing substantial variation in the cattle population. This phenotypic heterogeneity diminishes evidence of genetic association. These measures may also be associated with different genetic effects besides the pleiotropic effects. Hence, identifying more homogeneous components of feed efficiency and examining if genetic variants are associated with individual components could enhance the genetic selection of dairy cattle. In this work, we employ a multi-view bicluster analysis that jointly analyzes multiple feed efficiency measures and genome-wide genetic markers to identify cattle groups that are more homogeneous in feed efficiency features and can readily map to genetic markers.

For a given data matrix, existing biclustering methods seek groupings from the rows and columns of the matrix simultaneously. However, in our problem, each sample is characterized by two data matrices (views) of respective genotypes and phenotypic measures. To identify more homogeneous feed efficiency measures, and to make the measures useful for genetic analysis, we need to integrate all commonly used measures with the genotypic data. Our recently-proposed multi-view biclustering (MVBC) approach [7] is revised and applied in a genome-wide association study (GWAS) to improve feed efficiency quantification. If rows of a data matrix represent cows and columns represent features (measures/markers), our problem can be viewed as performing bi-clustering in each of the phenotypic and genotypic views to identify both row (cow) clusters and column clusters simultaneously, but we require the row clusters from the two views to be the same. The MVBC method allows to select initial features that are hypothesized to influence a cluster according to prior knowledge, which we use to guide the cluster analysis for meaningful interpretation.

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

II. MATERIALS

The datasets used in the present study were collected by the US Department of Agriculture (USDA) from 391 cows in 635 lactations (317 first-lactation heifers and 318 second or greater lactations). For each cow, longitudinal observations of multiple daily traits over lactation periods for multiple years from September 2007 to June 2013 were collected, and genotypes were obtained from Illumina ships (San Diego, CA). To avoid spurious results due to age and correlations (multiple lactations for one cow), only the earliest lactation of each cow was used in our analysis.

Multiple traits and features of each cow were recorded on a daily basis, including daily milk yield (DM) (kg), dry matter intake (DMI) (kg), milk fat (kg), protein (kg), lactose percentage, body weight (BW) (kg), and daily body-weight gain (DG) [3]. Then, nine measures were calculated for each cow from its daily records for each lactation period. These measures included average energy intake (EI), average DMI, average energy-corrected milk yield (ECM), average daily body-weight gain (ADG), average metabolic body weight (MBW), predicted energy intake (pred$_{EI}$), predicted DMI (pred$_{DMI}$), residual feed intake (RFI), and RFI from DMI (RFI$_{DMI}$) [1]. The parity was used to correct the computation for some of the nine measures, such as pred$_{EI}$. The parity was also used in the proposed GWAS analyses as a covariate.

Genotypes from the Illumina BovineHD BEAD chip and the new GeneSeek 140K chip were obtained from USDA's Animal Genomics and Improvement Laboratory where they were evaluated for quality and pedigree conflicts as previously described in [8], [9]. There were in total 777,962 genotypic markers recorded in the data. Among the 391 cows, 58 had a significant amount of missing markers, and hence were excluded from the proposed analyses. This set of single nucleotide polymorphisms (SNPs) was from the high-density chips, and included all SNPs traditionally used in U.S. genomic evaluations, plus other markers in a whole-genome sample.

III. METHOD

We first made an effort to minimize the confounding effect of parity on the cluster analysis. (In other words, cows were grouped together due to the different numbers of lactation (parity) rather than due to the different feed efficiency measures or genetic variants.) Hence, only 333 genotyped cows with their earliest lactation records were used in our analyses. Compared to the number of genetic variants (777,962) in the data, we have a relatively small sample size. Model overfitting is a well known problem associated with an analysis on a sample of small sample size but large amount of variables, and causes spurious clusters even in an unsupervised cluster analysis.

To overcome this problem, we first use a filtering process, i.e., a GWAS, to pre-select candidate markers for use in the subsequent multi-view bicluster analysis. Then, our MVBC method [7] was applied jointly to the nine phenotypic measures and the selected candidate SNPs. Because our method not only finds row clusters (cow clusters) but also column clusters in each view, it selects the phenotypic measures and candidate genetic markers that are used to arrive at the grouping of the cows. The selected phenotypic measures characterize the corresponding cluster of cows, and the selected genetic markers show the potential associations with the cluster. In the last step, we perform another GWAS to test all genetic markers in terms of separating cows in a cluster from those not in it, which we call a case-control study where cases are the cows in the cluster and controls are those not in the cluster. This step is used to further identify genome-wide significant markers associated with the individual clusters by the standards of statistical tests.

A. Filtering genetic variants

For each of the nine phenotypic measures, we performed a GWAS for main effect tests. We tested the effect of each marker to the variance of a phenotypic measure by regressing the measure on the genetic variant together with an offset parameter. This test was carried out using the main effect test function of PLINK [10] with parity as a covariate.

For each phenotypic measure, we chose the 1,000 most significantly associated markers, and took a union of all these markers. There were markers simultaneously associated with multiple measures. For example, variants at SNP positions BovineHD2800005250, BovineHD1800019353, BovineHD1000009886, BovineHD1000009874, BovineHD0800031648, and BovineHD0800030128 were each associated with four different phenotypic measures. Variants at SNP positions BovineHD1000009873 and BovineHD0800031657 were each associated with three different measures. Among the combined markers, 7,490 were unique. We further excluded markers with missing values because in our early simulation studies of the MVBC method, we observed that using imputed values for a large number of markers would mislead the cluster analysis [7]. Eventually, 1,059 markers remained and were used in the subsequent multi-view bicluster analysis.

B. Identifying homogeneous cow clusters

The MVBC method decomposes each data matrix into a pair of left and right vectors that are both sparse. Let $\mathbf{u}$ and $\mathbf{v}$ be the left and right vectors resulting from the decomposition. Rows in the data matrix corresponding to non-zero components in $\mathbf{u}$ form a row cluster and columns corresponding to non-zero components in $\mathbf{v}$ form a column cluster. To obtain consistent row clusters from the two views, the MVBC method requires the two left vectors $\mathbf{u}$ to have non-zero values at the same positions. By repeating this decomposition on updated data matrices where, specifically in our study, cows already in a row cluster are excluded, the desired number of cow (row) clusters can be obtained.

There are three hyper-parameters in our proposed MVBC method that need to be pre-specified before running the cluster analysis to obtain cow, measure and genotype clusters. We refer to these three hyper-parameters as (1) s_ω controls the size of cow clusters whereas (2) s_{v^1} and (3) s_{v^2} control the numbers of measures and genetic markers that will be used to examine cow similarities. Our early study [7] has shown that this method is more sensitive to the choices of s_{v^1} and s_{v^2} than that of s_ω because naturally, we need to recover the true subspaces (relevant features in each view) before the cows that exhibit similarities in those subspaces can be identified. The early study further suggests to use principal component analysis to help determine appropriate values of s_{v^1} and s_{v^2}. We performed a principal component analysis separately to

the two views of data. We chose the number of principal components that were needed to explain over 90% of the total data variance in each of the views to be the parameter value.

To choose an appropriate value for s_ω, we pre-selected a range of values. For each tested s_ω value, we employed our MVBC method to obtain a cluster solution. We then assessed the validity of the clusters by examining their genetic separability. In other words, we built a logistic regression classifier as a function of the genetic markers to separate cows in the cluster from those not in the cluster. A ten-fold cross validation process was used to evaluate the classification performance and the area under the receiver operating characteristic (ROC) curve (AUC) was used to measure the separability (i.e., classification performance). The average AUC values of all classifiers obtained for each s_ω choice were compared in order to choose a proper value of s_ω that gave the best separability.

C. Case-control GWAS

Each cluster resulting from our analysis corresponded to a binary trait for a cow (either in the cluster (cases labeled by 1) or not in the cluster (controls labeled by 0)). We hence performed another GWAS to test main effects for these binary traits. We tested each genetic marker in our dataset for its effect on differentiating the cases and controls by fitting a logistic regression model based on the genetic marker together with an offset parameter. This step was also carried out by PLINK, which computed p-values for the test of each genetic marker with each binary trait to measure the significance of the association. We then reported the three most significant markers for each cluster.

IV. Results and Discussion

Based on an early investigation of the USDA feed efficiency data [11], we anticipated that 3 to 5 clusters would be appropriate. We repeatedly deployed the MVBC method twice, which produced two clusters and the remaining cows formed the third cluster. In the cluster analysis, the parameters s_{v^1} and s_{v^2} were chosen to be 5 and 125 according to a principal component analysis when obtaining Cluster 1, and to be 4 and 110 when obtaining Cluster 2. The parameter s_ω was chosen to be 180 for Cluster 1 and 60 for Cluster 2 according to the cross validation process. Optionally, our method allows to set the initial choices of $\mathbf{v}^1$ and $\mathbf{v}^2$ (initial phenotypic measures and genetic markers). According to the prior knowledge, RFI and ECM were the most important and widely used feed efficiency metrics. We thus initialized the first cluster with the phenotypic measure RFI, which means to set the initial $\mathbf{v}^1$ to have 1 at the entry corresponding to RFI and 0 at all other eight entries. We initialized the second cluster with the phenotypic measure ECM. We left $\mathbf{v}^2$ uninitialized due to the absence of prior knowledge.

A. Identified clusters

The three clusters resulting from our analysis are summarized as follows. Cluster 1 with the initialization of RFI as feed efficiency metric consisted of 180 cow samples; Cluster 2 with the initialization of ECM consisted of 60 cow samples; and the last cluster contained the remaining 83 samples. Our algorithm automatically selected features from each of the views at the end based on which the samples were grouped into Clusters 1 and 2. It selected phenotypic measures: EI, DMI, MBW, RFI, and RFI_{DMI} for Cluster 1; and ECM, MBW, RFI, and RFI_{DMI} for Cluster 2.

To further characterize the clusters of cows, we drew a bar plot of the relative mean values of the nine phenotypic measures for each of the three clusters. For each measure, we used the following equation to compute the relative mean values and plotted these values in Fig. 1,

$$\frac{\text{Mean}(Sample\_in\_Cluster) - \text{Mean}(Entire\_sample)}{\text{STD}(Entire\_sample)}$$

where STD refers to the standard deviation of a measure over the entire sample.

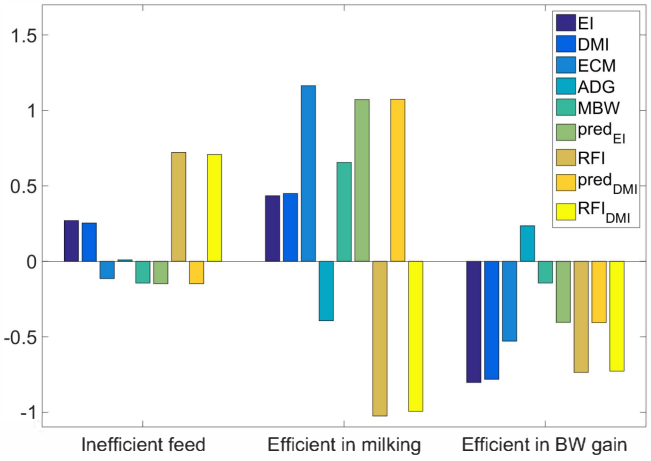

Fig. 1. The characteristics of the three clusters viewed in phenotypic view. The relative mean values of each measure were computed for each cluster.

We observed interesting and instructive patterns in these clusters. The very distinguishable characteristics we observed was that samples in Cluster 1 differed substantially from other clusters on values of RFI and RFI_{DMI} with mean values of 2.27, and 0.77, respectively. The cows in this cluster consumed high energy (mean EI = 56.11), and the energy intake was higher than predicted intake (resulting in positive values of RFI). The RFI was larger than that of other cows, resulting in a positive bar of RFI in Fig. 1. Hence, we name this cluster "Inefficient feed" cluster. In contrast, cows in Cluster 2 (which was the cluster we obtained by initializing with the ECM measure) had a high desirable efficiency in feeding where the RFI (mean = -4.17) and RFI_{DMI} (mean = -1.41) both were negative, meaning that the cows in this cluster consumed less energy than predicted intake, and the negative residual was in a greater magnitude than other cows (showing negative bars in Fig. 1). Their average energy intake (including DMI) was relatively high (mean EI = 63.62) but it did not result in high body weight gain (with a negative bar of ADG). Rather, as RFI was computed based on milk production, cows here had high efficiency transforming consumed energy into dairy production. We name this cluster "Efficient in milk production". As shown in Fig. 1, the last cluster contained the cows that consumed less energy (mean EI = 54.54) than the overall sample and also than the predicted intake (mean RFI = -3.10, RFI_{DMI} = -1.07), but were efficient in transforming the consumed energy into body weight gain (with a positive

bar of ADG but negative bars of RFI and RFI_{DMI}). We hence name the last subgroup "Efficient in BW gain".

We also observed that 125 SNPs and 110 SNPs were respectively selected for the "Inefficient feed" cluster and "Efficient in milk production" cluster. When we used the selected markers to build classifiers to separate the different clusters, the 10 most important SNPs in the classifier for Cluster 1 received weights in the range of $[0.0616, 0.0622]$. The 10 most important SNPs for Cluster 2 received weight in the range of $[0.0667, 0.0703]$, including a SNP marker BovineHD2000003678 that was identified to be genome-wide significantly associated with this cluster (of "Efficient in milk production") as shown in the subsequent GWAS.

B. Genome-wide significant associations

The results from the case-control GWAS with the feed efficiency clusters shows that three SNPs were significantly associated with the identified clusters, and especially the marker *BovineHD2000003678* at a p-value of 9.172e-14 and *BovineHD3000005312* at a p-value of 1.846e-09 associated with the "Efficient in milk production" cluster. There was also a marker *BovineHD1500020572* nominally associated with the identified "Inefficient feed" cluster at a p-value of 9.347e-07. For the cluster formed by the remaining cows, there was lack of significant associations, which was probably partly because it was not a cluster identified by our algorithm.

TABLE I. MOST SIGNIFICANTLY ASSOCIATED MARKERS FOR THE "INEFFICIENT FEED", "EFFICIENT IN MILK PRODUCTION", AND "EFFICIENT IN WEIGHT GAIN" CLUSTERS

Inefficient feed	p-value
BovineHD1500020572	9.347e-07
BovineHD1800012497	2.286e-06
BovineHD2300012041	2.483e-06
Efficient in milk production	**p-value**
BovineHD2000003678	**9.172e-14**
BovineHD3000005312	**1.846e-09**
BovineHD2800007744	1.656e-07
Efficient in weight gain	**p-value**
BovineHD2600010769	2.381e-06
BovineHD2300012041	3.137e-06
BovineHD2600010660	4.388e-06

V. CONCLUSION

In this paper, we have applied a multi-view sparse clustering approach to the genotype-phenotype association analysis of dairy cattle feed efficiency. We identified markers that might otherwise be difficult to detect without elucidating the phenotypic groups by the proposed method. The proposed analysis contained several steps, the core of which was the application of a matrix-decomposition-based multi-view biclustering algorithm. This algorithm links the phenotypic and genotypic views of the same sample by enforcing the row clusters from both views to be consistent. To the best of our knowledge, our work is among the first approaches that extend a rigorous multi-view analytics to feed efficiency analysis of animals (in particular, dairy cattle). This analysis brought clear insights about different dairy cattle populations that may be efficient in transforming consumed energy into different products (e.g., milk production or meat production by increased weight gain).

There are a few directions for future work. It is possible to extend the MVBC method to the case when missing values are present in any of the views. A simple idea is to recover the missing values in one view based on information from other views (not just imputed from their own view). Although our algorithm is computationally efficient, empirical evaluations on larger feed efficiency datasets might be needed to examine its speed and scalability. A larger sample size will be necessary in the future to validate the cow clusters and their characteristics identified in this study. Additional analysis of identified SNP markers associated with each trait is also needed to provide insight into potential genes and gene pathways contributing to variation among dairy cattle populations.

ACKNOWLEDGMENT

This work was supported by NSF grants DBI-1356655, IIS-1320586, IIS-1447711 and CCF-1514357. Jinbo Bi was also supported by NIH grant R01DA037349 and NSF grant IIS-140720.

REFERENCES

[1] E. Connor, J. Hutchison, H. Norman, K. Olson, C. Van Tassell, J. Leith, and R. Baldwin, "Use of residual feed intake in holsteins during early lactation shows potential to improve feed efficiency through genetic selection," *Journal of animal science*, vol. 91, no. 8, pp. 3978–3988, 2013.

[2] US Department of Agriculture, "National program for the genetic improvement of feed efficiency in beef cattle," vol. http://www.beefusa.org/CMDocs/BeefUSA/Resources/cc2012-Feed-Efficiency-in-Beef-Cattle–Weaber.pdf, 2012.

[3] D. Spurlock, "Genetic architecture and biological basis of feed efficiency in dairy cattle," in *10th World Congress on Genetics Applied to Livestock Production*, 2014.

[4] J. Van Arendonk, G. Nieuwhof, H. Vos, and S. Korver, "Genetic aspects of feed intake and efficiency in lactating dairy heifers," *Livestock Production Science*, vol. 29, no. 4, pp. 263–275, 1991.

[5] D. P. Berry and J. J. Crowley, "Residual intake and body weight gain: a new measure of efficiency in growing cattle," *Journal of Animal Science*, vol. 90, no. 1, pp. 109–115, 2012.

[6] R. M. Koch, L. A. Swiger, and G. K. E., "Efficiency of food use in beef cattle," *Journal of Animal Science*, vol. 22, pp. 486–494, 1963.

[7] J. Sun, J. Lu, T. Xu, and J. Bi, "Multi-view sparse co-clustering via proximal alternating linearized minimization," in *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 757–766.

[8] G. Wiggans, P. Van Raden, and T. Cooper, "The genomic evaluation system in the united states: Past, present, future," *Journal of Dairy Science*, vol. 94, no. 6, pp. 3202–3211, 2011.

[9] G. Wiggans, T. Cooper, D. Null, and P. VanRaden, "Increasing the number of single nucleotide polymorphisms used in genomic evaluations of dairy cattle," *10th World Congress for Genetics Applied Livestock Production*, vol. 301, 2014.

[10] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses." *American journal of human genetics*, vol. 81, no. 3, pp. 559–575, Sep. 2007.

[11] E. Connor, J. Hutchison, and H. Norman, "Estimating feed efficiency of lactating dairy cattle using residual feed intake," *(Chapter 11) In Feed Efficiency in the Beef Industry, Hill, R.A. (ed.), Wiley-Blackwell, NJ*, 2012.