Identification of variants associated with divergent feed efficiency groups using multiple RNA-sequencing datasets from dairy and beef cattle

S. Lam¹, F. Miglior¹,², P.A.S. Fonseca¹,³, D. Seymour¹, V. Asselstine¹, L. Brito¹, F. Schenkel¹ & A. Cânovas¹

¹University of Guelph, Department of Animal Biosciences, Centre for Genetic Improvement of Livestock, 50 Stone Road East, N1G 2W1, Guelph, Ontario, Canada slam02@uoguelph.ca (Corresponding author)
²Canadian Dairy Network, 660 Speedvale Avenue West, Suite 102, N1K 1E5, Guelph, Ontario, Canada
³Universidade Federal de Minas Gerais, Departamento de Biologia Geral, 31270-901, Belo Horizonte, Minas Gerais, Brazil

Summary

Improved selection for feed efficiency may lead to economic and environmental benefits for the dairy and beef industry. Recent developments in next-generation sequencing technologies, including RNA-Sequencing (RNA-Seq), has allowed for identification of genetic variants associated with economically interesting traits (i.e. feed efficiency). Public RNA-Seq data from one dairy and two beef cattle studies were used to identify and evaluate genetic variants associated with high and low feed efficiency groups. This approach allowed for identification of genetic variants mapped in genes that play crucial roles in biological processes related to feed efficiency, across all three studies for all high feed efficient animals (SFTPA1, bta-miR-122, STAT3A, SCARB2, and MACF1) and low feed efficiency animals (CEACAM1, CRYL1, SIRT3, IFITM1, STK11, and MYH9). The identified variants show associations with metabolically demanding pathways and organs suggesting their importance to energy metabolism and metabolic pathways. These results emphasize the importance of the integration of high-throughput RNA-Seq data from multiple studies using independent populations which allows for identification of functional variants across multiple breeds, leading to opportunities to improve accuracy in genetic selection strategies for dairy and beef cattle.

Keywords: High-throughput sequencing, RNA-Seq, SNP discovery, functional variants, residual feed intake, cattle

Introduction

Selection for feed efficiency in livestock animals may improve production efficiency and profitability, leading to a lower environmental footprint and economic benefits for livestock production. Residual feed intake (RFI), defined as the difference in actual and predicted feed intake accounting for size and growth of the animal, is currently an impractical method to measure feed efficiency. Therefore, alternative ways to select for feed efficiency are necessary. Feed efficiency is a complex trait regulated by many biological processes and the genetic basis underlying this complex trait is poorly understood. Integration and analysis of biological processes regulating feed efficiency may improve our understanding and current genetic selection strategies for this trait.

Recent developments of high-throughput sequencing technologies, including RNA-
Sequencing (RNA-Seq), have made a revolutionary impact on transcriptome analysis, providing an opportunity to identify genetic variants associated with complex traits. RNA-Seq allows for comparison of the entire transcriptomes to identify differentially expressed genes. Additionally, RNA-Seq data can be used to identify functional variants associated with economically and environmentally important traits (Wickramasinghe et al., 2014). From this, functional genetic variants associated with feed efficiency traits can be identified and used to improve accuracy and selection response in genetic selection for feed efficiency.

Previous studies have found variants associated with feed efficiency to improve our understanding of feed efficiency traits; however, no studies have evaluated RNA-Seq datasets from multiple studies using high-throughput technology to perform whole genome analysis for variants associated with feed efficiency traits. Therefore, the objective of this study was to use high-throughput technology to analyze RNA-Seq data from three different studies on dairy (Salleh et al., 2017) and beef cattle (Tizioto et al., 2015, Tizioto et al., 2016) to determine functional variants fixed within high and low RFI groups for each study and across different dairy and beef cattle breeds. Further validation of the results will be performed using RNA-Seq data from different breeds of Canadian cattle populations (Holstein (dairy) and Angus X Simmental (beef)).

Material and methods

Animal information

Raw RNA-Seq data from three previous studies with NCBI accession numbers GSE92398 (Salleh et al., 2017), PRJEB7696 (Tizioto et al., 2015), and PRJEB15314 (Tizioto et al., 2016) were used to perform the SNP discovery analysis in this study. Results described by Salleh et al., 2017 used a population of 19 Nordic dairy cattle composed of 10 Jersey and 9 Holstein cattle; animals were divided into high and low feed efficiency groups. RNA-Seq analysis were performed from liver biopsies to identify differentially expressed genes and profiles of pathways connecting these genes involved in regulating feed efficiency in dairy cattle (Salleh et al., 2017). Tizioto et al., (2015) and Tizioto et al., (2016) used a population of 20 Nelore steers with divergent phenotypes for high and low feed efficiency groups. RNA-Seq data from liver and muscle (longissimus dorsi) biopsies was used to describe differential gene expression and genome-wide expression profiles in hepatic and longissimus dorsi muscle tissue within the divergent feed efficiency groups (Tizioto et al., 2015, Tizioto et al., 2016).

RNA-Seq analysis

Figure 1 illustrates the workflow used to perform the SNP discovery analysis between high and low feed efficiency groups in both dairy and beef cattle and across breeds. A quality control assessment was performed on the total RNA-Seq reads using FastQC to check the overall base and sequence quality before and after trimming. Reads with low quality scores (Phred<33) were then removed and linker/adapter sequences were trimmed using Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic). Reads were aligned using HISAT2 (Kim et al., 2015) to perform the mapping of reads against the bovine UMC3.1 reference genome (release 90 from Ensembl). The Bcf tools mpileup command was used for SNP discovery (Danecek et al., 2017). Vcftutils.pl varfilter command was used to filter SNPs in a vcf file. The SNP quality control was performed using a Perl script. As described by Cánovas et al., (2010 and 2014), SNPs with at least 2 reads confirming the less frequent
allele, minimum coverage of 10 reads and within three standard deviations of the mean were considered to avoid possible false positives. SNP discovery analysis was performed for each group of animals with divergent feed efficiency profiles to identify those SNPs fixed in each group of high and low RFI. Variant Effect Predictor (VeP) (http://www.ensembl.org/info/docs/tools/vep/index.html) was used for functional analysis and annotation of genomic variants within each previously mentioned group.

**Results and discussion**

The SNP consequence percentage using total SNP identified in both dairy and beef populations for high and low feed efficiency groups is shown in Figure 2. Preliminary results reveal a large number of SNP are segregating between divergent efficiency groups and also shared amongst them in both dairy and beef cattle populations.

Genetic variants identified exclusively in high or low feed efficiency groups when evaluating RNA-Seq data across both dairy and beef cattle populations is shown in Table 1. A total of 1973 unique genetic variants, common across all breeds, were identified in high feed efficient animals. These genetic variants include bta-mir-122, SFTPA1, STAT5A, SCARB2, and MACF1 genes. All genetic variants identified are known to be associated with important functions in energy metabolism, including energy production and transportation, lipid metabolism, and activity in highly metabolically active organs. This includes bta-mir-122, which is the third most expressed miRNA and is highly expressed in bovine liver (Al-Husseini et al., 2016).

A total of 2512 unique genetic variants, common across all breeds, were identified in low feed efficient animals. Among them, SNP uniquely identified in low feed efficient animals in both dairy and beef populations were located in the CEACAM1, CRYLI, IFITM1, SIRT3, STK11, and MYH9 genes. Sirtuin 3 (SIRT3) is localized in the mitochondrial matrix and regulates mitochondrial intermediary metabolism and fatty-acid use suggesting its importance as a metabolic sensor (Hirschey et al., 2010). Preliminary results revealed multiple genetic variants associated with highly metabolically active processes which emphasize the importance in identifying genetic variants involved in regulating feed efficiency. This may assist in improving genetic selection programs and accuracy of genomic prediction in both dairy and beef cattle.

The integration and analyses of RNA-Seq data, using high-throughput technologies such as RNA-Seq, from multiple studies with independent populations may lead to improved understanding of the biology of complex traits by identifying functional genetic variants associated with feed efficiency in cattle. Further studies evaluating multiple datasets from Canadian dairy and beef populations using transcriptomics (RNA-Seq) and other OMICs technologies, will contribute to the current knowledge of the underlying biological processes related to feed efficiency and will therefore lead to improved genetic selection for feed efficiency traits and environmental and economic benefits for the dairy and beef industry.

**Acknowledgements**

We gratefully acknowledge funding by the Efficient Dairy Genome Project, funded by Genome Canada (Ottawa, Canada), Genome Alberta (Calgary, Canada), Ontario Genomics (Toronto, Canada), Alberta Ministry of Agriculture (Edmonton, Canada), Ontario Ministry of Research and Innovation (Toronto, Canada), Ontario Ministry of Agriculture, Food and Rural Affairs (Guelph, Canada), Canadian Dairy Network (Guelph, Canada), GrowSafe Systems
List of References


Figure 1. Workflow used to identify variants within high and low feed efficiency groups in beef and dairy cattle populations separately and combined.
Figure 2. SNP consequence percentage using the list of total SNP identified in both dairy and beef populations for high and low feed efficiency groups.
Table 1. Genetic variants uniquely identified in high and low feed efficiency groups in both dairy and beef cattle populations.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene symbol</th>
<th>SNP Locus</th>
<th>Allele</th>
<th>Consequence of variant</th>
<th>Existing variant identifier</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Feed Efficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSBTAG00000023032</td>
<td>SFTPA1</td>
<td>28:35,843,502</td>
<td>A/G</td>
<td>Synonymous</td>
<td>rs110326717</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000023032</td>
<td>SFTPA1</td>
<td>28:35,843,502</td>
<td>A/G</td>
<td>Upstream gene</td>
<td>rs110326717</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000029851</td>
<td>bta-mir-122</td>
<td>24:58,095,828</td>
<td>C/T</td>
<td>Downstream gene</td>
<td>rs110616983</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000029851</td>
<td>bta-mir-122</td>
<td>24:58,095,813</td>
<td>T/C</td>
<td>Downstream gene</td>
<td>rs109661171</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000029851</td>
<td>bta-mir-122</td>
<td>24:58,095,788</td>
<td>T/C</td>
<td>Downstream gene</td>
<td>rs110420138</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000009496</td>
<td>STAT5A</td>
<td>19:43,039,444</td>
<td>C/T</td>
<td>Synonymous</td>
<td>rs4471054</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000010956</td>
<td>SCARB2</td>
<td>6:92,800,460</td>
<td>A/G</td>
<td>3 prime UTR</td>
<td>rs41256842</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000021024</td>
<td>MACF1</td>
<td>3:107,179,713</td>
<td>T/C</td>
<td>Downstream gene</td>
<td>rs41255164</td>
<td>-</td>
</tr>
<tr>
<td>Low Feed Efficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSBTAG000000005615</td>
<td>CEACAM1</td>
<td>18:51,139,719</td>
<td>G/A</td>
<td>Missense</td>
<td>rs41887907</td>
<td>0.08</td>
</tr>
<tr>
<td>ENSBTAG00000011726</td>
<td>CRY1.1</td>
<td>12:36,204,954</td>
<td>A/G</td>
<td>Synonymous</td>
<td>rs109249868</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000002044</td>
<td>SIRT3</td>
<td>11:107,192,938</td>
<td>T/G</td>
<td>Upstream gene</td>
<td>rs455776433</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000019017</td>
<td>IFITM1</td>
<td>11:107,192,938</td>
<td>T/G</td>
<td>Missense</td>
<td>rs455776433</td>
<td>1.00</td>
</tr>
<tr>
<td>ENSBTAG00000011639</td>
<td>STK11</td>
<td>7:45,287,897</td>
<td>A/G</td>
<td>Synonymous</td>
<td>rs210187107</td>
<td>-</td>
</tr>
<tr>
<td>ENSBTAG00000010402</td>
<td>MYH9</td>
<td>5:75,096,667</td>
<td>C/T</td>
<td>Missense and splice</td>
<td>rs -</td>
<td>1.00</td>
</tr>
</tbody>
</table>

1 allele from reference genome
2 alternative allele from RNA-Seq data
3 Sorting intolerant from tolerant