A gene-transcription factor network associated with residual feed intake based on SNVs/InDels identified in Gir, Girolando and Holstein cattle breeds

L.L. Verardo[1, 2], N.B. Stafuzza[2], D.P. Munari[2], A. Zerlotini[3], T.C.S. Chud[2], D.J. Garrick[4], J.B. Cole[5], J.C.C. Panetto[1], M.A. Machado[1], M.F. Martins[1] & M.V.G.B. Silva[1]

[1]CAPES/Embrapa Gado de Leite, Juiz de Fora, Minas Gerais, Brazil
marcos.vb.silva@embrapa.br (Corresponding Author)
[2]Departamento de Ciências Exatas, Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, São Paulo, Brazil
[3]Embrapa Informática Agropecuária, Campinas, São Paulo, Brazil
[4]Institute of Veterinary, Animal & Biomedical Sciences, Massey University, Hamilton, New Zealand
[5]Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, Maryland, United States of America

Introduction

In tropical cattle production, indicine breeds (*Bos indicus*) are widely used in extensive production systems in tropical areas due to their adaptability to heat and disease. Gir cattle represent the main indicine breed used for dairying in Brazil, where a national dairy breeding program was established in 1985 for this breed (Panetto et al., 2017). This breed is present in more than 80% of Brazilian dairy herds, either as purebreds or composites produced by various crossing schemes with taurine breeds (*Bos taurus*). Girolando is a composite breed resulting from the use of selected bulls that are mostly admixed 5/8 Holstein and 3/8 Gir, but includes animals with breed compositions ranging from 1/4 Holstein and 3/4 Gir to 7/8 Holstein and 1/8 Gir. These composites are known for their robustness, high fertility and high milk production, which has led to annual growth in the use of this breed in tropical production systems (Silva et al., 2017), reinforcing the importance of genetic studies of Gir and Girolando breeds.

Progress in next-generation sequencing (NGS) methodology and in sequence analysis tools has allowed whole genome re-sequencing to become a viable approach to quickly, efficiently and accurately identify genetic variants such as single nucleotide variations (SNVs) and insertions/deletions (InDels). The study of these sequence variants is important for the discovery of causal variants linked to complex traits in animal production (Jiang et al., 2014 and Das et al., 2015). Recently, Stafuzza et al. (2017) identified SNVs and InDels using whole-genome re-sequencing of Gir, Girolando, Guzerat and Holstein breeds. Enrichment analysis revealed that variants in the olfactory transduction pathway were over represented in all four breeds. It has been shown in pigs that the olfactory transduction pathway may be associated with residual feed intake (Do et al., 2014).

Residual feed intake (RFI), first proposed by Koch et al. (1963), is the difference between an observed and predicted feed consumption. Due to its importance, recent attention has been given to using RFI in genomic analyses (Davis et al., 2014 and Dimauro et al., 2016). However, little information is known at the genomic level linking genes to RFI. Analyses of genomic regions associated with economical important traits through gene networks and *in silico* identification of transcription factors (TF) have been shown to provide a better understanding, not only of the genes associated with these traits, but also of the genetic architecture among breeds (Fortes et al., 2010, Verardo et al., 2016).

The aim of this study was to analyze whole-genome re-sequencing data focusing on SNVs and InDels identified in Gir, Girolando and Holstein cattle breeds related to RFI. Thus, genes showing SNVs/InDels in TF binding sites (5' UTR variants) were used to search for TF related to feed intake and to generate a gene-TF network for RFI across two purebred and one admixed dairy cattle breeds. Moreover, we were able to perform a comparative analysis accessing similarities and dissimilarities

at the genomic level across these breeds.

Materials and methods

Whole-genome re-sequencing data

For this study, we used whole-genome re-sequencing data from 10 animals (two Gir, three Girolando and five Holstein) as reported by Stafuzza et al. (2017). In summary, SNVs and InDels were identified based on UMD 3.1 bovine genome assembly. Those variants were then classified according to their potential function (e.g. 5' UTR variant) using the Ensembl Variant Effect Predictor tool (VEP, version 84) (McLaren et al., 2010). Functional enrichment analysis using lists of genes that had variants classified by the VEP tool were submitted for annotation using KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis.

Gene-TF networks

Aiming to analyze genomic regions related to RFI, we first selected all genes showing SNV/InDel in the 5' UTR region identified in each breed. From these gene sequences, the TFM-Explorer program (http://bioinfo.lifl.fr/TFM/TFME; accessed http://thebiogrid.org) was used to search for locally overrepresented TF binding sites (TFBS) using weight matrices from the JASPAR vertebrate database (Sandelin et al., 2004) to detect all potential TFBS by calculating a score function as described in Touzet & Varré (2007). From that set of genes, we collected sequences 3,000 bp upstream and 300 bp downstream (FASTA format) from the transcription start site, based on the UMD 3.1 bovine genome assembly. These data were used as the input for TFM-Explorer. The given list of TF was fed into Cytoscape (Shannon et al., 2003) using a Biological Networks Gene Ontology tool (BiNGO) plug-in (Maere et al., 2005) to determine which GO terms were significantly overrepresented assuming defaults statistical and multiple testing corrections (P-value < 0.05).

Based on biological processes (e.g., response to nutrient levels and sensory organ development) in conjunction with literature review, we were able to select the main TF related to feed intake (key TF) in each population, from which we constructed a gene–TF network. Aiming to identify putative candidate genes, we used the NetworkAnalyzer tool within Cytoscape. According to the number of TFBS and, consequently, the number of connections/lines (edges) in each node (gene and TF), the genes most connected within the gene–TF network were determined. Genes with more TFBS, for the most representative TF, might have more edges and thus identify a gene–TF network. Finally, the gene–TF networks derived for the three breeds were merged, highlighting the genes and TF in common, providing an overview across them.

Results and discussion

A total of 163 genes among the three breeds were identified as having variants at the 5' UTR region. From these genes, 25 TF for each breed were identified and filtered based on the biological processes overrepresented in BiNGO and on a literature review related to feed intake. We selected the four key TFs (Table 1) to construct a combined gene-TF network (Figure 1). That network highlighted the most connected genes (e.g. *LOC522763*, *TRIM40*, *CHL1*, *CDK4*, *EPN3*, *DHRS7C* and *TBC1D7*) and provides an overview of shared TFs and genes across the breeds. A Venn diagram was built to illustrate the shared genes and TF observed in the gene-TF network (Figure 2). Comparing the number of genes initially identified with those present in the RFI gene-TF network, we found that more than 50% of Gir and more than 40% of Girolando genes set remained in the

network, while only 16% of Holstein genes were present in the network. This may indicate a higher genetic variability of Gir and Girolando breeds for residual feed intake related genes.

Table 1. *Main transcription factors (TF) associated with genes identified in the three breeds (Gir, Girolando and Holstein), their biological process and literature evidences related to residual feed intake.*

TF	Breed	Biological Process (GO)	Literature evidence*
STAT1	Girolando	Response to nutrients	Energy metabolic pathways (Pitroda et al., 2009)
PPARG	Gir	Response to nutrients/Response to nutrients levels	Energy balance and food intake (Cecil et al., 2006)
MAX	Gir, Girolando and Holstein	Response to nutrients levels/Sensory organ development	Glycolysis and carbohydrate metabolism (Kim et al., 2007)
YY1	Holstein	Sensory organ development	Energy metabolism balance (Cunningham et al., 2007)

* The cited literature studies are just a sample from the vast available literature

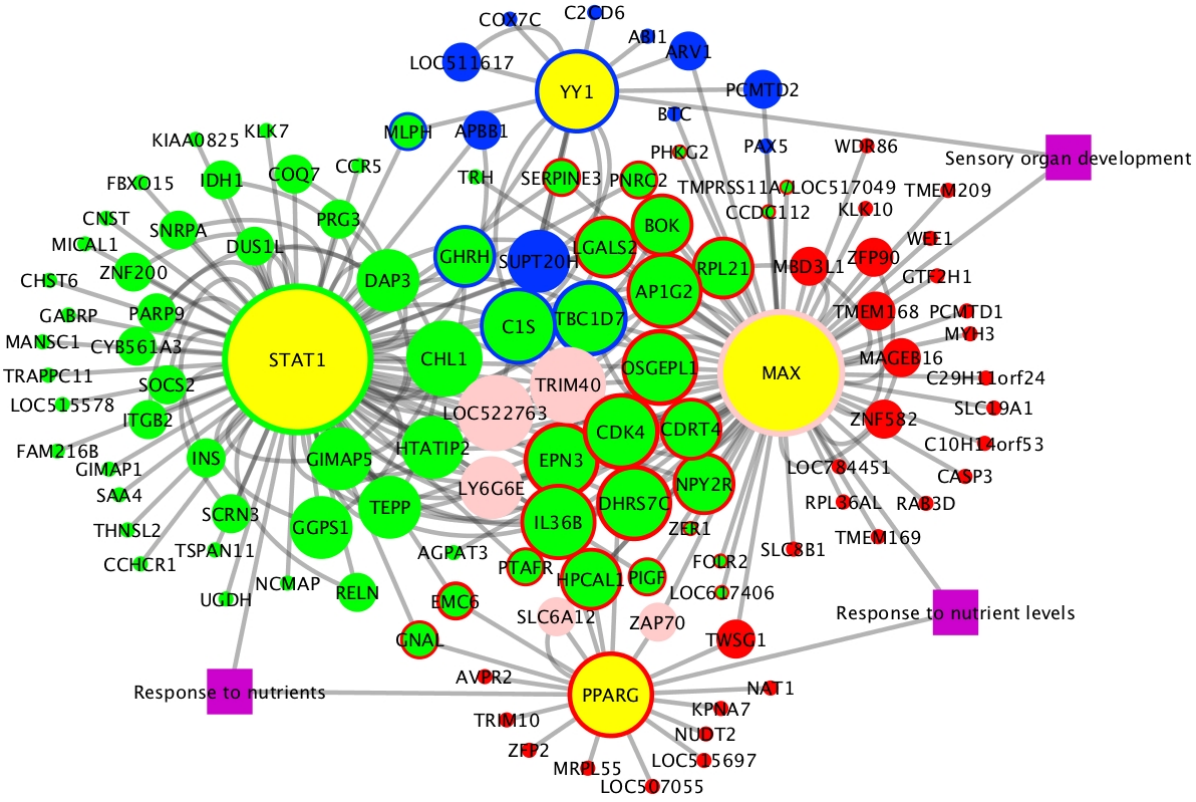

Figure 1. *Gene-transcription factor (TF) network. TF (yellow nodes) and genes showing SNVs and InDels in their 5' UTR region (red nodes are genes observed in Gir; green nodes are genes observed in Girolando; blue nodes are genes observed in Holstein; pink nodes are genes in common to all breeds). Green nodes with blue border are genes in common between Girolando and Holstein breeds while green nodes with red borders are genes in common between Girolando and Gir. The colors of TF borders follow each breed color. The node size corresponds to the network analysis (Cytoscape) score, where bigger nodes represent higher edge density associated with the number of TF-binding sites. Purple square nodes are biological processes (GO terms) associated with residual feed intake.*

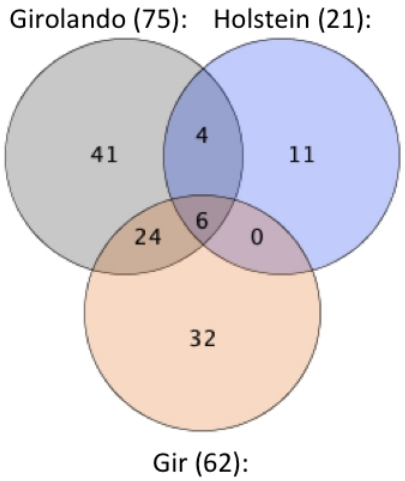

***Figure 2**. Venn diagram showcasing the number of common genes and transcription factors (TF) identified among breeds (Gir, Girolando and Holstein) in the residual feed intake gene-TF network.*

Among the key TF, only *MAX* was in common to all breeds. *MAX* is a MYC-associated factor X gene cited to be trigged with transcripts related with glycolysis and carbohydrate metabolism (Kim et al., 2007). In this study, *MAX* was observed to be enriched with biological processes related to residual feed intake (e.g., response to nutrients levels and sensory organ development). At the gene-TF network, this TF was the second most connected followed by *STAT1* that encodes for the signal transducer and activator of transcription 1 and is suggested to be involved with energy metabolic pathways (Pitroda et al., 2009). This TF was identified in the set of genes from Girolando breed, which showed the higher number of genes what might explain its higher connectivity to this network.

The other two TF illustrated in the gene-TF network are associated with genes presenting variants identified in Gir (*PPARG*) and Holstein (*YY1*). *PPARG* gene encodes a peroxisome proliferator-activated receptor γ gene suggested to be involved with energy balance and feed intake (Cecil et al., 2006), and well associated with response to nutrients levels according to our biological process analyses. *YY1* is a transcription factor identified to participate on energy metabolism balance complexed with peroxisome-proliferator-activated receptor co-activator (*PGC-1α*), which is involved on mitochondrial oxidative control to maintain the energy balance in response to nutrients (Finck & Kelly, 2006; Cunningham et al., 2007).

Among the highlighted genes, we observed genes well connected with the presented key TF that were in common to all three breeds or between two breeds (e.g., *TRIM40*, *TBC1D7* and *EPN3*). *TRIM40* encodes a tripartite motif containing protein 40. Studies indicate that *TRIM40* may be an important factor regulating mucosal growth in the bovine rumen (Connor et al., 2013) and could be directly related to the digestive tract. This gene is located on BTA23 (28.62 Mb - 28.63 Mb) and positioned at the quantitative trait locus (QTL) for RFI previously identified by Sherman et al. (2009), making it a candidate gene for this trait. In this study, in all three breeds exhibited SNVs/InDels in its promoter region.

The *TBC1D7* gene encodes a TBC1 domain family member 7 protein and has been proposed to contribute to nutrient signaling from mammalian target of rapamycin (mTOR) and cell growth (Jewell and Guan, 2013). Signaling pathways involving mTOR have been studied in dairy cattle showing the relation with nutrients provision and protein synthesis rates in mammary cells (Appuhamy et al., 2014). In this study, variants in the promoter region of *TBC1D7* gene were identified in Girolando and Holstein animals and it was well connected with *MAX* and *STAT1* transcription factors at the gene-TF network. Another well-connected gene with these TF was *EPN3*,

which encodes the epsin 3 protein. It is suggested that this protein cooperates with others bi-layer binding proteins with curvature sensing/generating properties in the specialized traffic and membrane remodeling processes typical of gastric parietal cells (Ko et al., 2010). This gene is located on BTA19 (36.81 Mb - 36.82 Mb) close or within previously identified QTLs for RFI (Sherman et al., 2009; Abo-Ismail et al., 2014), making this gene a candidate gene for this trait. In this study, *EPN3* was identified in Gir and Girolando animals.

The results observed in this study should be carefully interpreted due to the small sample size used for whole-genome re-sequencing. However, the gene-TF network provides important genomic information to investigate genetic mechanisms underlying phenotypic differences and similarities among these breeds. Our results highlighted candidate genes (e.g., *TRIM40*, *TBC1D7* and *EPN3*) and TF (*STAT1*, *PPARG*, *MAX* and *YY1*) that might have a role in explaining variation in residual feed intake of cattle. Moreover, in comparison with Holstein, the Gir and Girolando breeds showed more enriched genes in the RFI gene-TF network, suggesting that these latter breeds may have greater genetic variability for this trait. In this study, the Gir breed composition associated with RFI seems to be conserved in Girolando animals, justifying the higher presence of highlighted genes in the gene-TF network observed for indicine and crossbreed cattle. Further *in vitro* functional studies should be conducted considering the identified SNVs/InDels from the highlighted genes in order to improve our knowledge of the relevance of SNVs/InDels in explaining variation for RFI in each breed.

Acknowledgements

LLV would like to thank CAPES/PNPD and FAPESP for the scholarship support. MVGBS was supported by Embrapa (Brazil) SEG 02.13.05.011.00.00 “Detecting signatures of selection using next generation sequencing data ”, CNPq/Universal 456450/2014-9 “Identification of signatures of selection in cattle using next generation sequencing data”, and MCTI/CNPq/INCT-Ciência Animal, FAPEMIG CVZ PPM 00606/16 “Detecting signatures of selection using next generation sequencing data” appropriated projects.

References

Abo-Ismail MK et al., 2014. Single nucleotide polymorphisms for feed efficiency and performance in crossbred beef cattle. BMC Genetics, 15(1): 14.

Appuhamy JADRN et al., 2014. Effects of AMP-activated protein kinase (AMPK) signaling and essential amino acids on mammalian target of rapamycin (mTOR) signaling and protein synthesis rates in mammary cells. J. Dairy Sci., 97(1): 419-429.

Cecil JE et al., 2006. Energy balance and food intake: the role of PPARγ gene polymorphisms. Physiol. Behav., 88(3): 227-233.

Connor EE et al., 2013. Gene expression in bovine rumen epithelium during weaning identifies molecular regulators of rumen development and growth. Funct. Integr. Genomics, 13(1): 133-142.

Cunningham JT et al., 2007. mTOR controls mitochondrial oxidative function through a YY1–PGC-1α transcriptional complex. Nature, 450: 736-740.

Das A et al., 2015. Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. BMC Genomics, 16: 1043.

Davis SR et al., 2014. Residual feed intake of lactating Holstein-Friesian cows predicted from high-density genotypes and phenotyping of growing heifers. J. Dairy Sci., 97(3): 1436-1445.

Dimauro C et al., 2016. Use of multivariate statistical analyses to preselect SNP markers for GWAS on residual feed intake in dairy cattle. J. Anim. Sci., 94(supplement5): 155-155.

Do DN et al., 2014 Genome-wide association and pathway analysis of feed efficiency in pigs reveal candidate genes and pathways for residual feed intake. Front. Genet., 5: 307.

Finck BN & Kelly DP, 2006. PGC-1 coactivators: inducible regulators of energy metabolism in health and disease. J. Clin. Invest. 116: 615–622.

Fortes MR et al., 2010. Association weight matrix for the genetic dissection of puberty in beef cattle. Proc. Natl. Acad. Sci., 107(31): 13642–13647.

Jewell JL and Guan K-L, 2013. Nutrient signaling to mTOR and cell growth. Trends Biochem. Sci. 38(5): 233-242.

Jiang L et al., 2014. Targeted resequencing of GWAS loci reveals novel genetic variants for milk production traits. BMC Genomics, 15: 1105.

Kim JW et al., 2007. Hypoxia-inducible factor 1 and dysregulated c-Myc cooperatively induce vascular endothelial growth factor and metabolic switches hexokinase 2 and pyruvate dehydrogenase kinase 1. Mol. Cell Biol., 27: 7381–93.

Ko G et al., 2010. Selective high-level expression of epsin 3 in gastric parietal cells, where it is localized at endocytic sites of apical canaliculi. Proc. Natl. Acad. Sci., 107(50): 21511-21516.

Koch RM et al., 1963. Efficiency of feed use in beef cattle. J. Anim. Sci. 22: 486–494.

Maere S et al., 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21(16): 3448–3449.

McLaren W et al., 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics, 26(16): 2069-2070.

Panetto JCC et al., 2017. Programa Nacional de Melhoramento do Gir leiteiro: Sumário de Touros: Resultado do Teste de Progênie – 8ª Prova de Pré-seleção de Touros. Juiz de Fora: Embrapa Gado de Leite, 2017. 96 p. (Embrapa Gado de Leite. Documentos, 202).

Pitroda SP et al., 2009. STAT1-dependent expression of energy metabolic pathways links tumour growth and radioresistance to the Warburg effect. BMC Medicine, 7(1): 68.

Sandelin A et al., 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res, 32:D91–D94.

Shannon P et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 13(11): 2498–2504.

Sherman EL, 2009. Fine mapping quantitative trait loci for feed intake and feed efficiency in beef cattle. J. Anim. Sci. 87(1): 37-45.

Silva MVGB et al., 2017. Programa de melhoramento genético da raça Girolando - Sumário de Touros - Resultado do Teste de Progênie Junho/2017. Juiz de Fora: Embrapa Gado de Leite, 56 p. (Embrapa Gado de Leite. Documentos, 203).

Stafuzza NB et al., 2017. Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gir, Girolando and Holstein cattle breeds. PloS One, 12(3): e0173954.

Touzet H & Varré JS, 2007. Efficient and accurate P-value computation for position weight matrices. Algorithms Mol. Biol., 2:15.

Verardo LL et al., 2016. After genome-wide association studies: Gene networks elucidating candidate genes divergences for number of teats across two pig populations. J. Anim. Sci., 94(4): 1446-58.