Identification of Predictor Genes of Feed Efficiency in Beef Cattle by Applying Machine Learning Methods to Multi-Tissue Transcriptome Data

Weihao Chen[1,2], Pâmela A. Alexandre[2], Gabriela Ribeiro[3], Heidge Fukumasu[3], Wei Sun[2], Antonio Reverter[1] and Yutao Li[1]

[1]CSIRO Agriculture & Food, Queensland Bioscience Precinct, 306 Carmody Rd., St. Lucia, Brisbane, QLD 4067, Australia.
[2]College of Animal Science and Technology, Yangzhou University, 88 Daxue Nan Road, Yangzhou, 225009, Jiangsu, CHINA
[3] School of Animal Science and Food Engineering, University of Sao Paulo, Pirassununga, SP, Brazil

Machine learning (ML) methods have shown promising results in identifying candidate genes when applied to large transcriptome datasets. However, no attempt has been made to compare the performance of combining different ML methods together in the prediction of high and low feed efficiency (HFE and LFE) animals. In this study, using RNA-seq data of five tissues from 18 Nellore bulls, we evaluated the prediction accuracies of four analytical methods in classifying animals according to their feed efficiency potential.

Transcriptome Dataset (Alexandre et al, 2015)

- 18 Nellore bulls : 16 ~20 months old, 9 with high and 9 with low feed efficiency (HFE and LFE);
- 5 tissues: adrenal gland, hypothalamus, liver, skeletal muscle and pituitary ;
- 86 RNA libraries sequenced using an Illumina HiSeq2500 equipment (2x100 pb);
- 16,423 genes after QC: 14,158 (adrenal gland), 14,581 (hypothalamus), 12,090 (liver), 11,391 (skeletal muscle) and 13,912 (pituitary). 9,950 genes were common across all 5 tissues.

Identification of Subsets of Genes for Classification of HFE and LFE Animals

- 3-fold Cross-validation scheme;
- Selecting subsets of genes with four methods: Benchmark with edgeR, 3 machine learning (ML) methods: Random Forests (RF), Extreme Gradient Boosting (XGBoost), and combination of RF and XGBoost (RX);
- Classification of HFE and LFE animals with Support Vector Machine (SVM)

Gene Co-expression Network Analysis

- PCIT algorithm (Reverter and Chan, 2008) and Cytoscape Version 3.7.1 (Shannon et al., 2003).

Table 1: Comparison of classification performances (F1-Score) of subsets of genes selected from different methods, when applying SVM. Number of genes used is given in parenthesis.

	Source of Subset Genes				
Tissue	edgeR	RF	XGBoost	RX	Best
Adrenal Gland	0.956 (941)	0.915 (4,993)	0.937 (171)	0.949 (33)	edgeR
Hypothalamus	0.945 (908)	0.947 (4,041)	0.948 (222)	0.951 (33)	RX
Liver	0.886 (486)	0.927 (2,092)	0.897 (227)	0.932 (30)	RX
Muscle	0.940 (950)	0.945 (3,294)	0.924 (193)	0.957 (23)	RX
Pituitary	0.973 (1,625)	0.979 (4,869)	0.958 (180)	0.977 (41)	RF
Average	0.940	0.942	0.933	0.952	RX

FOR FURTHER INFORMATION
Dr Yutao Li
E yutao.li@csiro.au
T +61 7 3214 2465 M 431 555 157

REFERENCES
Alexandre et al. 2015. BMC Genomics. 16: 1073.
Reverter A. and Chan EKF. 2008. J Bioinformatics. 24: 2491–2497.
Shannon et al. 2003. Genome Res. 13: 2498–2504.

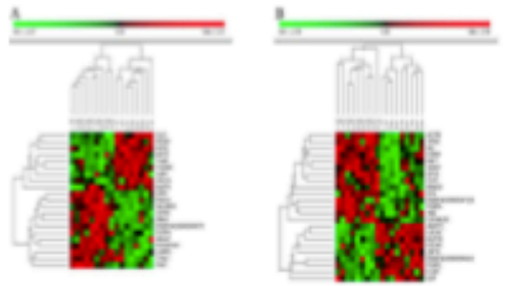

Figure 1: Heatmap of cluster analysis using subsets of genes identified by the RX in muscle (A) and pituitary gland (B). H refers to HFE bulls and L refers to LFE bulls.

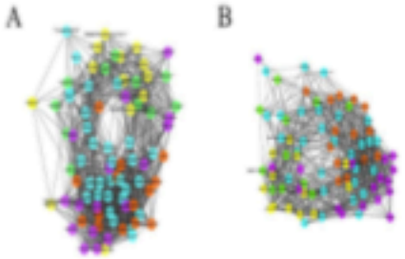

Figure 2: Co-expression networks in LFE (A) and HFE (B), colors are relative to the tissue of maximum expression: yellow represents liver, green represents muscle, orange represents pituitary, purple represents hypothalamus and blue represents adrenal gland.

Table 2. Correlations between ML "Gain" values and network centrality parameters for genes identified by the RX

	Betweenness	Closeness	Clustering	Degree	Neighbourhood	Radiality	Topological
Adrenal Gland	0.10	0.30	0.11	0.30	0.27	0.29	0.03
Hypothalamus	0.23	0.09	-0.14	0.09	-0.04	0.08	-0.13
Liver	0.23	0.25	-0.08	0.29	0.06	0.23	-0.03
Muscle	0.29	-0.01	-0.10	0.00	-0.13	0.02	-0.19
Pituitary	0.21	0.25	-0.11	0.24	-0.09	0.23	-0.13
AVERAGE	0.21	0.17	-0.06	0.18	0.01	0.17	-0.09

Results and Conclusions

1. Of four methods, the two-step ML method combining RF and XGBoost (RX), identified the smallest subsets of potential predictor genes across all tissues with the highest classification accuracy for 9 HFE and 9 LFE animals (Table 1, Figure 1);
2. For genes identified by the RX, there was a correlation between the gene's prediction ranking ("Gain" values) and its relevance to the networks ("Betweenness", Table 2), reflecting a key biological role to the phenotype;
3. When comparing co-expression gene network differences between LFE and HFE groups from the RX (Figure 2), the number of connections between genes with maximum expression in skeletal muscle represented the biggest change between HFE and LFE networks. This indicates more FE related pathways activated in HFE.