Proc. Assoc. Advmt. Anim. Breed. Genet. 22:289-292

A PIPELINE FOR THE ANALYSIS OF MULTI-OMICS DATA

Loan To Nguyen[1,2], Marina R. S. Fortes[1] and Antonio Reverter[3]

[1]School of Chemistry and Molecular Biosciences, The University of Queensland, Australia,
[2] Faculty of Biotechnology, Vietnam National University of Agriculture, Vietnam,
[3] CSIRO Agriculture & Food, Queensland Bioscience Precinct, Queensland 4067, Australia

SUMMARY

We describe an analytical pipeline to exploit the results from RNA sequencing (RNA-Seq) experiments combining a series of processes from data normalization to network inference. The pipeline makes use of numerical approaches aimed at identifying key regulators via the regulatory impact factor (Reverter *et al.* 2010) metrics. It also employs the partial correlation and an information theory (Reverter and Chan 2008) for the identification of significant edges in the construction of gene co-expression networks. Key nodes in the network include differentially expressed genes, transcription factors, tissue specific genes as well as genes harboring SNPs found to be associated with the phenotype(s) of interest. The pipeline has already been successfully employed in two beef cattle studies, dealing with the onset of puberty and feed efficiency. In the present paper, we describe a pipeline to analyze RNA-Seq data, focus on relevant genes, generate gene co-expression networks and identify emerging clusters within the network to provide new insight about the subject matter under scrutiny.

INTRODUCTION

Gene expression is the process which transferring the information of the gene into the production of a functional product. Genes may be expressed at specific tissue or only at certain physiological state in the animal life cycle. By measuring the abundance of gene products (RNA transcripts) in a tissue at a specific physiological state, the gene expression rate can be evaluated. Using gene expression analysis to identify candidate genes and biomarkers could ultimately enhance the accuracies of genomic prediction for key traits.

RNA sequencing (RNA-Seq) is a next-generation sequencing technique developed in 2008 for the analysis of gene expression across the entire transcriptome (Mortazavi *et al.* 2008; Wang *et al.* 2009). RNA-Seq was first applied in model organisms including Arabidopsis (Lister *et al.* 2008), yeast (Nagalakshmi *et al.* 2008) and mouse (Mortazavi *et al.* 2008), but has rapidly increased its popularity to a number of other organisms including human (Sultan *et al.* 2008) and bovine (Huang and Khatib 2010). RNA-seq is high-throughput and the analysis of large-scale datasets has a wide range of applications, however, every RNA-seq experimental scenario may have different optimal methods for analyses. New approaches are currently being developed (Han *et al*. 2015). Here we provide a step-by-step recipe on how to use the pipeline to analyze RNA-Seq data, focus on relevant genes, generate gene co-expression networks and identify emerging clusters within the network to provide insight about the subject matter under scrutiny. Without entering in detailed numerical intricacies (published elsewhere and cited herein), we discuss the essential principles of the analytical methods of each step in the process.

METHODS

In what follows, we provide a step-by-step recipe on how to exploit RNA-Seq data in order to identify differential expressed genes, key regulatory genes and generate gene co-expression network, in combination with algorithms such as RIF (Reverter *et al.* 2010) and PCIT (Reverter and Chan 2008). Figure 1 provides a schematic of the flow chart for this analytical pipeline.

Generally, the pipeline used for the analysis of multi-omics data requires a series of four steps as follows:

Step 1 – RNA-Seq Experimental Resource. In order to infer differentially expressed genes and gene co-expression networks in our multi-omics pipeline, the following resources are required: 1) the RNA-Seq data comprising at least two experimental conditions; 2) the experiment data conducting at least in two tissues. In the puberty example, the two experimental conditions would be the pre- or post-puberty stages; while the reproductive tissues of interest could include hypothalamus, pituitary, ovaries and uterus, as well as tissues related to the onset of puberty such as liver, fat and muscle. Other experimental setting could include healthy versus disease states, various breeds and/or various time points as conditions.

Step 2 – Normalization via Mixed-Model Equations. The ability of mixed-models in terms of their power to accommodate covariance structures in various forms is well documented in the animal breeding and genetics literature. Similarly, mixed-models are the ideal tool for the normalization of gene expression data (Reverter *et al.* 2005). Aiming for parsimony the simplest model will contain the library as the only fixed effect, and the interaction effect of gene by animal by condition by tissue and the residual as the only random effects:

$$Y = Library + Gene + Gene*Animal*Condition*Tissue + Error$$

The solution of the Gene*Animal*Condition*Tissue (GACT) interaction are used as the normalized mean expression (NME) of each gene in each animal and tissue. However, combinations of lower order gene interactions, such as Gene*Animal, Gene*Condition and Gene*Tissue are also possible. Additionally, the GACT solutions for all the animals from the same condition could be averaged to obtain the NME of each gene in each condition and tissue. The NME values will provide the basis for the computation of differential expression and tissue-specificity.

Step 3 – Selection of 'Relevant' Genes. To facilitate the task of generating and analysing the resulting network, only a subset of genes will be used according to the following four categories: differentially expressed (DE) genes, tissue-specific (TS) genes, genes harbouring SNP reported to be associated with a phenotype or phenotypes of relevance, and significant regulators such as transcription factors (TF). Next, we briefly describe the way in which each category is identified.

Differentially expressed (DE) genes. Typically, the contrast of interest will be comprised of the (possibly differential) expression of a given gene in a given tissue across the two (or more) conditions under study. These can be obtained directly from the NME and the statistical inference on the contrast performed based on a number of approaches of which a simple t-test is quite possibly optimal, preferably after correction for multiple testing using either Bonferroni or (preferably) Benjamini and Hochberg methods (both described in (Benjamini and Hochberg 1995)).

Tissue specificity. Similarly, the NME can be used to reveal the expression of each gene in each tissue and then compute the proportion of a gene's total expression in each of the tissues (ie. based on the NME of a gene in a tissue divided by the sum of the NME of the same gene summed across all tissues). This could be done either within or across the two (or more) conditions under study. In doing so, tissue-specific (TS) genes will be identified from those genes whose expression in a given gene is higher than in any other tissue by a particular amount such as fold-based bearing in mind that a gene can be TS for one tissue only. Additionally, using comparative genomics from human studies, we can source the identity of TS genes from the Tissue-specific Gene Expression and Regulation database (TIGER: http://bioinfo.wilmer.jhu.edu/tiger/).

Genes harbouring associated SNP. Today there is a plethora of GWAS in the literature quite possibly studying a condition similar (even identical) to the one in our current study. The results

from these studies can be mined to retrieve the genes surveyed in our RNA-Seq study that are reported to harbour SNP associated with a phenotype or condition similar or preferably identical to the one in our current study.

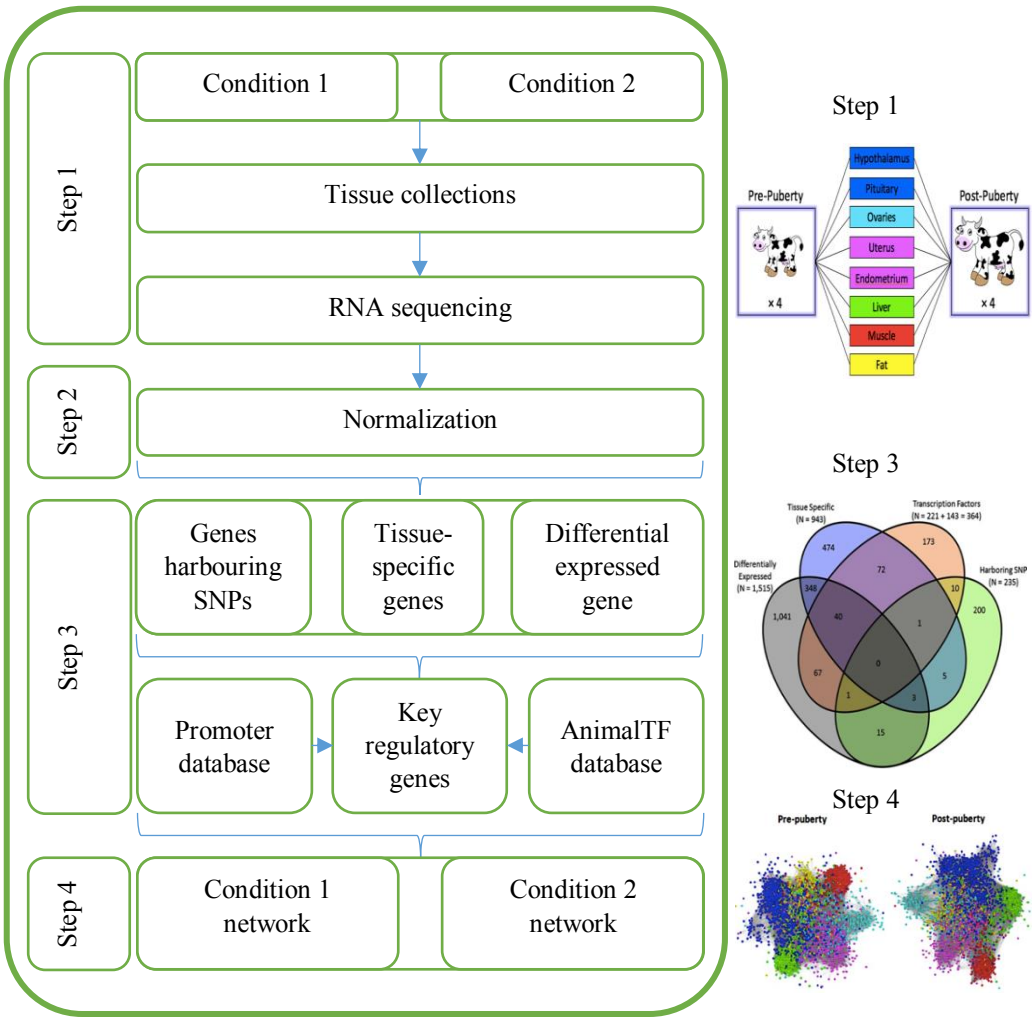

Figure 1. Flow chart of the pipeline for the RNA sequencing analysis (left) and illustrations adapted from Canovas *et al*. (2014) in the context of the onset of puberty in Brangus heifers

Key regulators. In order to identify the regulators (not necessarily TF) present among the genes surveyed in our RNA-Seq study, we mine to the Animal Transcription Factor Database (http://www.bioguo.org/AnimalTFDB/). Among these, we define as significant or "key" regulators those with statistically significant RIF metrics (using DE, TS and SNP harbouring genes as targets) and/or those with binding motif in the promoter region of DE, TS and/or SNP harbouring genes. In more detail, RIF comprises a set of two metrics designing to evaluate the regulatory power of molecules by exploring their differential connectivity to other influential genes (eg. those differentially expressed) in two contrasting conditions of interest (eg. pre- and post-puberty).

Step 4 – Network Inference and Visualisation Analysis.
For the network inference, we use the DE, TS, key TF and SNP harbouring genes as nodes and significant connections are identified using the partial correlation and information theory (PCIT) algorithm either through the original FORTRAN90 source code (Reverter and Chan 2008) or through an R package (Watson-Haigh *et al.* 2010). The PCIT exploits the twin concepts of partial correlation and mutual information. In brief, PCIT ascertain the significance of a given correlation between 2 entities (e.g., genes or network nodes) after accounting for all other genes in the dataset. Importantly, the output from PCIT can be viewed with Cytoscape (Shannon *et al.* 2003), a software program for analysing and visualizing gene co-expression network. In order to characterize network features, many Cytoscape plug-ins are available. Of these plug-ins, we recommend MCODE (Bader and Hogue 2003) to identify highly interconnected gene clusters, and BINGO (Maere *et al.* 2005) to determine which Gene Ontology terms are significantly overrepresented in a set of clustered genes. Hopefully, these clusters may have biological significance within the context of the phenotype under study.
One final process in the analysis of the resulting network is to identify the best trio of TF among those spanning the majority of the network topology. To this end, we search for TF with lots of connections in the network but few in common as these indicate redundancy.

CONCLUSIONS

The biological complexity and the rapid accumulation of publicly data arise the need to develop efficient tools for large-scale multi-dimensional data analysis. We conclude that the proposed analytical pipeline is a useful procedure providing an opportunity screen and identify key regulatory genes as well as generate regulatory networks with predictive power for the phenotype under investigation. Therefore, it may also be a significant tool for integrating different RNA-seq dataset and different levels omics data in order to investigate the complexity of biological subjects.

REFERENCES

Bader G. and Hogue C. (2003). *BMC Bioinformatics.* **4**: 2.
Benjamini Y. and Hochberg Y. (1995). *J Royal Stat Soc Series B (Methodological).* **57**:289.
Canovas A., Reverter A., DeAtley K.L., Ashley R.L., Colgrave M.L. *et al.* (2014) *PLoS One.* **9**: e102551.
Han Y., Gao S., Muegge K., Zhang W. and Zhou B. (2015) *Bioinform Biol Insights.* **9**: 29.
Huang W. and Khatib H. (2010) *BMC Genomics.* **11**: 1.
Lister R., O'Malley R.C., Tonti-Filippini J., Gregory B.D., Berry C.C. *et al.* (2008) *Cell.* **133**: 523.
Maere S., Heymans K. and Kuiper M. (2005). *Bioinformatics.* **21:** 3448.
Mortazavi A., Williams B.A., McCue K., Schaeffer L. and Wold B. (2008) *Nat Methods.* **5**: 621.
Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D. *et al.* (2008) *Science.* **320**: 1344.
Reverter A., Barris W., McWilliam S., Byrne K.A., Wang Y.H. *et al.* (2005) *Bioinformatics.* **21**: 1112.
Reverter A. and Chan E.K.F. (2008) *Bioinformatics.* **24**: 2491.
Reverter A., Hudson N.J., Nagaraj S.H., Perez-Enciso M. and Dalrymple B.P. (2010) *Bioinformatics.* **26**: 896.
Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T. *et al.* (2003) *Genome Res.* **13**: 2498.
Sultan M., Schulz M.H., Richard H., Magen A., Klingenhoff A. *et al.* (2008) *Science.* **321**: 956
Wang Z., Gerstein M. and Snyder M. (2009) *Nat Rev Genet.* **10**: 57.
Watson-Haigh N.S., Kadarmideen H.N. and Reverter A. (2010) *Bioinformatics.* **26**: 411.