PEDIGROMICS: A NETWORK-INSPIRED APPROACH TO VISUALISE AND ANALYSE PEDIGREE STRUCTURES

A. Reverter[1], S. Dominik[2], J.B.S. Ferraz[3], L. Corrigan[4] and L.R. Porto-Neto[1]

[1]CSIRO Agriculture & Food, St. Lucia, Brisbane, QLD, 4067 Australia
[2]CSIRO Agriculture & Food, Armidale, NSW 2350 Australia
[3]Universidade de São Paulo (USP), Pirassununga, SP, Brazil
[4]Rennylea Angus, NSW, Australia

SUMMARY

We exploit the power and visual appeal of network theory to visualise and analyse pedigree structures by bringing into the visualization schema elements of the breeding history for a given population. We name the approach "Pedigromics" and illustrate its potential with three datasets from beef cattle populations of varying characteristics. These include tropical cattle from the Beef CRC Legacy Database, a highly curated pedigree from an Angus seedstock producer in Australia, and a large pedigree from Nelore cattle in Brazil. While conceptually very simplistic, we highlight the ability of Pedigromics to identify emerging features of family structures with a wide range of applications from the understanding of the breeding history of a population, to the identification of influential ancestors and to inform mating allocation alternatives for the design of breeding schemes.

INTRODUCTION

With the advent of cheaply available "omics" technologies, the last decade has seen an explosion of molecular data. Whether at the level of genome sequence variants, transcript expression or protein abundance, the wide adoption of these technologies has allowed us to collect tens of thousands of data points from individual samples. As a consequence, new analytical and computational approaches have been developed in order to interrogate the large datasets in an efficient and effective manner. Paramount among these approaches are the numerical algorithms for the inference of gene networks. Two such approaches include WGCNA (Langfelder and Horvath 2008) and PCIT (Reverter and Chan 2008). Once a gene network has been constructed, its optimality should be established by mathematical and biological criteria. The former includes whether or not the connections follow a non-random scale-free distribution by which most genes have a few connections while a few genes, termed hubs, have lots of connections. Biological criteria include the ability of the network to (1) recover known gene-gene interactions and (2) reveal clusters of genes annotated to a biological process of relevance to the subject matter under scrutiny.

Pedigree information and associated metadata (i.e. sex, year of birth, farm of origin, etc.) is ideally suited to be explored using what have now become standard approaches to visualise and analyse gene networks. Individual animals are represented as nodes in the network, and two animals are connected by an edge when a particular kinship exists, e.g. Parent – Offspring.

We name the approach "Pedigromics" and illustrate its potential with three beef cattle datasets: (1) A Brahman and Tropical Composite population from the Beef CRC Legacy Database; (2) A highly curated pedigree from Rennylea, an Angus producer in Australia; and (3) a large pedigree from Brazilian Nelore cattle.

MATERIALS AND METHODS

Beef CRC Dataset. We made use of the Brahman (BR) and Tropical Composite (TC) populations extensively described in the literature by our group and most recently by Raidan *et al.* (2018). For the

present study, the original 2,111 BB and 2,550 TC cattle were further edited to only include animals satisfying the following 3 conditions: (1) having both sire and dam recorded; (2) from the top 20 sires based on number of progeny; and (3) from contemporary groups (CG, including sex, month of birth and property) with at least 10 records and at least two sires represented. These editing criteria resulted in 720 BB and 883 TC cattle used for Pedigromics analyses.

Angus Dataset. We used a pedigree file of 7,551 Angus cattle (4,429 females and 3,122 males) from Rennylea spanning 29 birth years (1989 to 2017). The file contained information from 305 sires averaging 24.5 progeny and ranging from 1 to 338 progeny. Further information used in the Pedigromics visualisation scheme was BreedPlan EBV for weaning weight and the 3,955 animals with SNP genotypes were highlighted.

Nelore Dataset. The dataset contained 51,265 Nelore cattle (25,324 females and 25,941 males) spanning over 50 years to 2017 and with 910 sires averaging 56.3 progeny (range = 1 to 4,613). For the Pedigromics map we captured the progeny (N=31) from the most prolific sire in 2017, namely D4685. We trace the dams of these progenies and all of the male ancestors, i.e. maternal and paternal grandsires, great-grandsires, and so on until exhausting the available genealogy.

In all three cases, we used Cytoscape (available at http://cytoscape.org; Shannon *et al*. 2003), a widely used software program for the visualization and exploration of the networks.

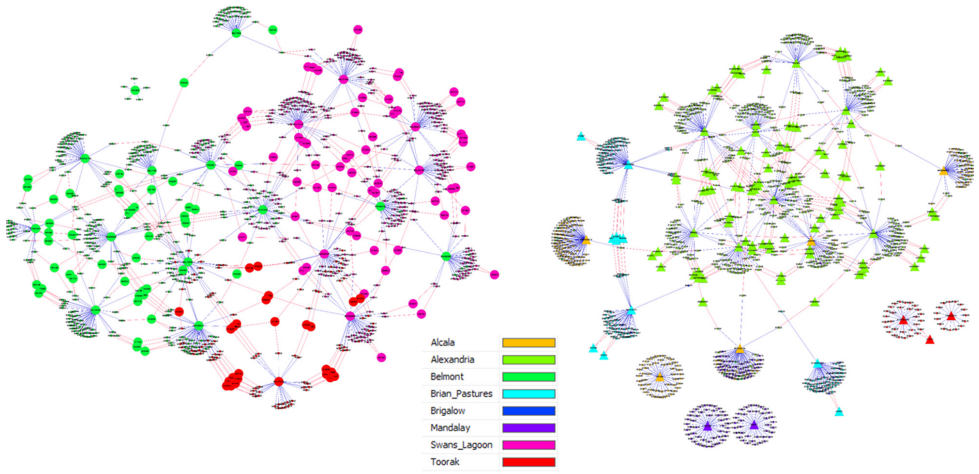

Figure 1. Pedigromics view of the Beef CRC dataset comprised of 1,872 cattle (nodes) connected by 2,108 edges. Two clear clusters emerge corresponding to Brahman (circles, left) and Tropical Composite cattle (triangles, right). Nodes are coloured according to property of origin (see insert). Connections correspond to sire-offspring (black edges) or dam-offspring (red edges). Big and small nodes correspond to parents and offspring, respectively

RESULTS AND DISCUSSION

The Pedigromics view of the Beef CRC (Figure 1), Angus (Figure 2) and Nelore datasets (Figure 3) have been designed to highlight different aspects of each dataset by resorting to a diverse range of criteria for the visualisation scheme. With the Beef CRC dataset (Figure 1), we highlight the power of Pedigromics to quickly identify the presence of two unrelated populations (Brahman and Tropical Composite) while providing a birds eye view of the of cattle across the eight properties.

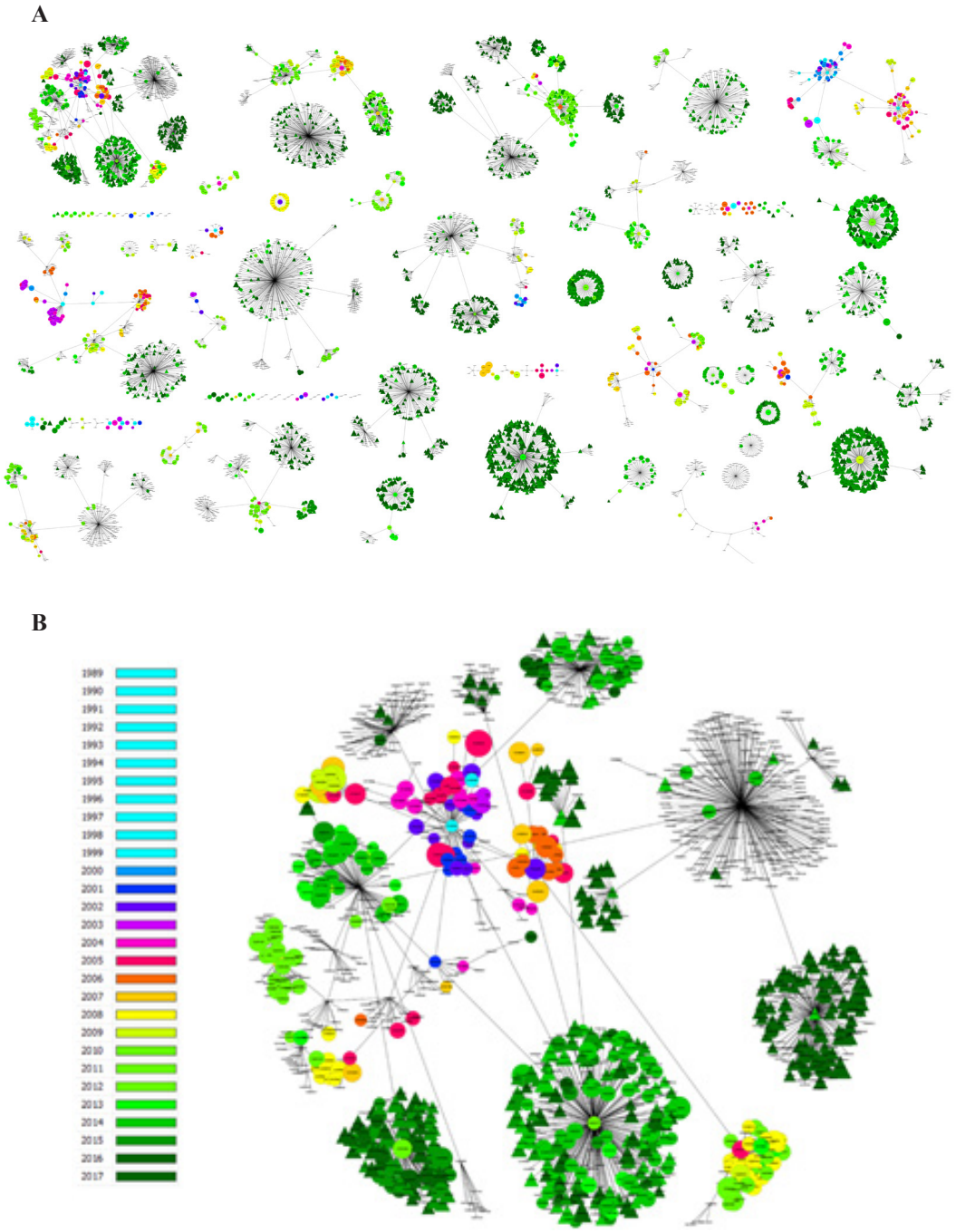

Figure 2. Pedigromics view of the Rennylea Angus dataset for the entire set of 7,551 individuals from 305 sire families (A) and a close up of a single large sire family spanning several years (B). Non-genotyped and genotyped individuals are represented by circles and triangles, respectively. Also, the size of the node has been mapped to the weaning weight EBV

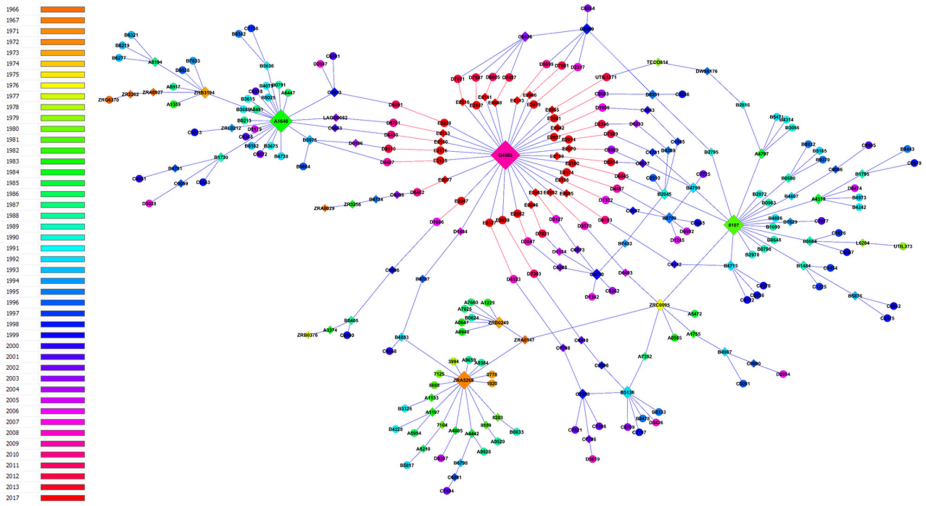

Figure 3. Pedigromics view of the Brazilian Nelore dataset focussing on the ancestors of the 31 progeny of sire D4685 in 2017

With the Angus dataset, we show how the 305 sire families cluster in super-families providing a look similar to constellations in the cosmos. We then zoom into one of these sire super-families so that we can easily trace individual sire families back in time. In addition, by imposing the weaning weight EBV to the size of the nodes, we can immediately see what could be termed as 'heritability at work' by which progeny of a small EBV sire are also small, and vice versa.

Finally, with the Nelore dataset we show how by focusing on the progeny of an influential sire Pedigromics can easily reveal the identity of other influential sires back in time.

CONCLUSIONS

We have presented Pedigromics, a tool to take advantage of network theory and visualisation techniques to analyse pedigree files and reveal features of the population that were not immediately apparent from the flat files. Applied to three distinct datasets allowed us explore different aspects of interest to the history and potential of a breeding population. Further applications of Pedigromics are being explored including assessing conflicts between pedigree and genomic information.

ACKNOWLEDGEMENTS

This work was performed using the legacy database belonging to the Cooperative Research Centre for Beef Genetic Technologies and their core partners including Meat and Livestock Australia. The authors are grateful to Sigrid Lehnert for reviewing this manuscript.

REFERENCES

Langfelder P. and Horvath S. (2008) *BMC Bioinformatics* **9**: 1.

Raidan F.S.S., Porto-Neto L.R., Li Y., Lehnert S.A., Vitezica Z.G. and Reverter A. (2018) *J. Anim. Sci.* **96**: 4028.

Reverter A. and Chan EK. (2008) *Bioinformatics* **24**: 2491.

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin S., Schwikowski B. and Ideker T. (2003) *Genome Res*. **13**: 2498.