The structure of a gene network reveals 7 biological sub-graphs underlying eQTLs in pig

L. Liaubet[*], N. Villa-Vialaneix[†], A. Gamot*, F. Rossi[‡], P. Cherel[§], M. SanCristobal*

Introduction

Integrative and system biology is a very promising tool for deciphering the biological and genetic mechanisms underlying complex traits. Transcriptomic analyses, in combination with genomic polymorphism, for instance, can give interesting insight on the genetic control of gene expression (eQTL studies). When hundreds of genes are detected with a link between their expression and some genetic polymorphisms (eQTL), the following question raises: what are the biological underlying functions? One tool is the use of a gene network, displaying interactions between genes with a genetic control (having at least an eQTL). There exist several possibilities for inferring a gene network: literature mining (using softwares such as Ingenuity) or inference from gene expression data. Although the first framework is a useful tool, it has some limitations: there is still a serious problem of lack of annotation in the pig genome, and a bias in information provided by Ingenuity (literature mainly devoted to Human, Mouse and Rat). We will hence explore in this work the inference of gene network from expression data. One simple method of inference was focused on, that has proven useful: Gaussian networks (Schäfer and Strimmer 2005). The following problem to be faced is the interpretation of such a "large" network (more than 100 genes). The aim of this study is to propose an adequate method for deciphering the structure of large gene networks. With the use of a good clustering of graph, the structure of one graph can be highlighted, and can reveal several sub graphs, each corresponding to particular biological functions.

Material and methods

eQTL Data. 56 half sib pigs were produced from an F2 intercross between two production sire lines: FH016 (Pietrain type) and FH019 (Synthetic line from Duroc, Hampshire and Large White founders, France Hybrides SA, St. Jean de Braye, France). These animals where produced by three sows mated with the same boar from one to three successive litters. *Longissimus dorsi* muscles RNA were extracted as described by Lobjois et al. (2007). The

[*] INRA Toulouse, UMR444 Laboratoire de Génétique Cellulaire, BP52627, 31326 Castanet Tolosan cedex, France
[†] Institut de Mathématiques de Toulouse, Université de Toulouse and IUT STID de Carcassonne, Université de Perpignan, France
[‡] TELECOM ParisTech
[§] France Hybride, Hendrix Genetics

9K micro-array (GEO accession number GPL3729) used in this work was previously described (Bonnet et al., 2008), as well as the microarray Nylon cDNA hybridization and quantification (Ferre et al., 2007). Genomic DNA was extracted from piglet tails and 170 microsatellites loci spanning the 18 autosomes with an average spacing of 17 cM were selected based on informativity on F1 on animals and genotyped. Analyses were done by SAGA LICOR software. F2, F1 and F0 animals were all genotyped and Mendelian segregation was checked. Custom genetic maps were reconstructed with CRIMAP software (Green, 1992). Additive genetic effect was fitted as an animal model, using pedigree structure to setup animal relationship matrix, and QTL effect was fitted using Identical By Descent (IBD) relationship matrix for the given genome position. IBD relationship matrices were estimated each 4 cM along linkage groups using software package LOKI 2.5 (Heath, 1997) and variance components were estimated using Residual Maximum Likelihood (REML) method with ASREML 2.0 software (Gilmour et al., 2006).

Graph inference. The inference of the interaction graph between the 272 genes having at least one eQTL was undergone according to Schäfer and Strimmer (2005), using the Gaussian model framework. Using this approach, an edge is drawn between 2 nodes (genes) when the partial correlation between these is higher than a given threshold (to be tuned). The matrix of partial correlation between the 272 genes was computed from the expression data by a bootstrap method which improves the stability of the result.

Simulations. Prior to this, a simple simulation study was performed, in order to assess the effect of sample size of the estimation of the (partial) correlation coefficients: 100 replicates were drawn, consisting of 2 correlated Gaussian variables, from which an estimate of the correlation coefficient was computed. The same scheme was used with 3 Gaussian variables to estimate the partial correlation between pairs of genes.

Graph clustering. A clustering of nodes was applied to the graph: following the idea of Reichardt and Bornholdt (2006) or Villa et al. (2010), a quality measure designed for graphs, the modularity (see Newman and Givran (2004)), has been optimized by a simulated annealing algorithm.

Biological validation. The WebGestalt software (Zhang et al., 2005) provided the functional Gene Ontology terms for each gene and gave a statistical analysis of the results. Ingenuity Pathways Analysis (IPA, https://analysis.ingenuity.com/pa/) was used to explore biological relevance of the gene network.

Results and discussion

Boxplots of estimates of correlation coefficients of the simulation study are presented in Figure 1, for varying sample size *n*. The same patterns were obtained with the partial correlation (not shown). The estimation variability is really too high for *n* lower than 30. In that case, the forthcoming graph inference would lead to unstable and hence invalid results. Our data consisting in 56 observations seems adequate for further graph analysis.

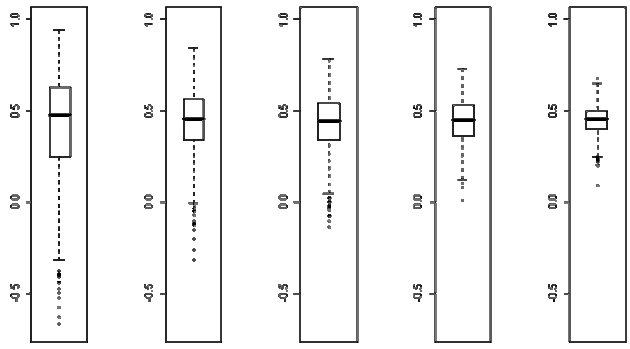

Figure 1: Boxplots of estimates of (partial) correlation coefficient, with n=10, 20, 30, 50 and 100 observations (from left to right)

The inferred graph of the 272 considered genes is presented in Figure 2. The cut-off value of partial correlation between 2 nodes was chosen by the probabilistic framework given in Schäfer and Strimmer K. (2005) given a significance value for the correlation between two genes. The structure of this graph is displayed in Figure 2 after clustering of nodes. Each sub graph corresponds to particular biological function (GO term). Such a clear view was not observed with a graph inferred with the correlation coefficient, instead of partial correlation. A good coherence with bibliography mining networks (Ingenuity) was obtained for each sub graph.

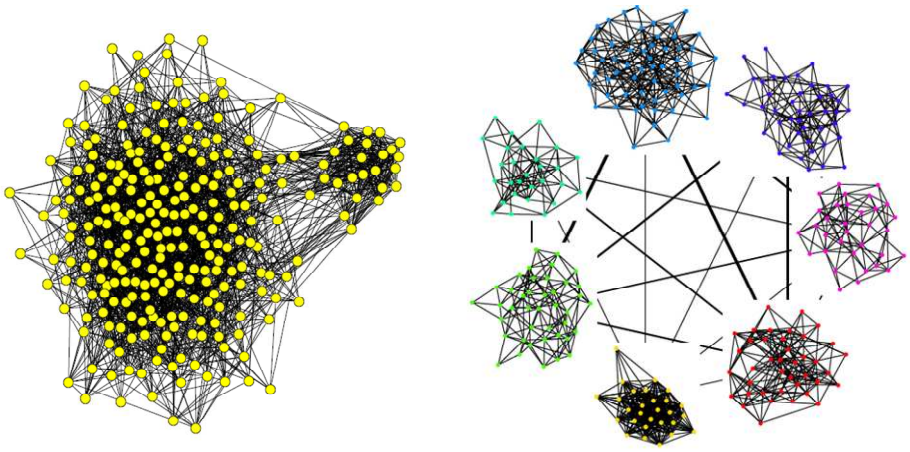

Figure 2: Graph of gene interactions between the top 272 genes having at least one eQTL (left), and structure of the graph after clustering (right). The size of the edges is proportional to the strength of interaction between sub-graphs.

Conclusion

These results show that an adequate combination of statistical methods, namely inference of graph using partial correlation under a Gaussian model, followed by clustering of graph, can lead to a significant improvement of our biological knowledge in the underlying biological functions of a set of genes, such as those having a genetic control (from a eQTL study). This kind of analysis is only valid with a sufficient number of observations.

Acknowledgements. This work was supported the Eadgene network of Excellence.

References

Bonnet, A., Le Cao, K.A., Sancristobal, *et al.* (2008). *Reproduction* 136, 211-24.

Ferre, P.J., Liaubet, L., Concordet, D., *et al.* (2007). *Pharmaceutical Research* 24, 1480-9.

Gilmour, A.R., Gogel, B.J., Cullis, P.R. *et al.* (2006). *ASReml User Guide Release 2.0.*

Green, P. (1992). *Cytogenetics and Cell Genetics* 59, 122-4.

Heath, S.C. (1997). *American Journal of Human Genetics* 61, 748-60.

Lobjois, V., Liaubet, L., SanCristobal, *et al.* (2008). *Animal Genetics* 39, 147-62.

Newman, M.E.J. and Girvan, M. (2004) *Physical Review, E.* 69, 026113.

Schäfer, J. and Strimmer, K. (2005) *Bioinformatics*, 21:754-764.

Reichardt, J. and Bornholdt, S. (2006) *Physical Review* E, 74(016110).

Villa, N., Dkaki, T., Gadat, S., *et al.* (2010) *SciWatch Journal, Hexalog*. To appear.

Zhang, B., Kirov S., Snoddy, J. (2005) *Nucleic Acids Res.* 1;33