CONSTRUCTION OF CO-EXPRESSION NETWORK ASSOCIATED WITH MARBLING SCORE IN BOVINE

Dajeong Lim[1,2], Nam-Kuk Kim[1], Hye-Sun Park[1], Jeongsoo Lee[1], Tae-Hun Kim[1], Heebal Kim[2], Seung-Hwan Lee[1*]

[1] Division of Animal Genomics and Bioinformatics, National Institute of Animal Science, Rural Development Administration, Omokchun-dong 564, Kwonsun-gu, Suwon, Korea

[2] Laboratory of Bioinformatics and Population genetics, Seoul National University, Seoul, Korea

[*]Corresponding author (phone: +82-31-290-1607; fax: +82-31-290-1602; e-mail: slee46@korea.kr)

***Abstract*—Identifying candidate genes related to complex traits or diseases and mapping their relationships require a system-level analysis. Our approach that investigate co-expression relationships of genes related to 'marbling score' trait and systemically analyze the network. We observed that our co-expression network has a poer-law connectivity distribution as many other biological networks have. The hub nodes and structure of the result network are constent with the prior information about marbling score. We also performed experimental validation of the hub nodes between high- and low-marbled groups using qRT-PCR. This network-based approach in livestock may an important method for analyzing the complex effects of candidate genes asssociated with complex traits**

***Index Terms*— Co-expression, Gene network, Marbling score, Power-law function**

I. INTRODUCTION

Gene expression data has been successful in understanding of relationships between genes involved in biological mechanisms and predicting targetable genetic components associated with complex traits or, diseases states. Microarray-based analysis provides differential expressed genes (DEGs) that were altered expression in a various environmental condition under the transcriptome level. Until recently, many researchers have made great efforts to investigate the overlap between genes expressed in a biological pathway and the chromosomal region identified by genetic linkage could detect a candidate that turns out to be the casual gene (Mootha, Lepage et al. 2003). Several high-throughput technologies that were combined gene expression and genetic approaches such as Quantitative trait locus (QTL) mapping. It is called 'genetical genomics' that is a powerful tool to elucidate gene and genome expression and its effect in resulting phenotypes (Jansen and Nap 2001) . An integration of genetic and gene expression studies in a system level accelerates the characterization of functional QTL. Several researches also shown that mRNA levels for candidate genes are heritable, thus apply to genetic analysis (Brem, Yvert et al. 2002; Wayne and McIntyre 2002; Schadt, Monks et al. 2003).

Many complex traits such as disease susceptibility, development, and agricultural product quality in animals are controlled by the interactions of several or many QTL combined with environmental influences. Furthermore, patterns of covariation in expression multiple loci can be used to build networks showing relationships between genes and between genes and functional traits. These networks provide information of the genetic control of complex traits and can help determination of causal genes where these effect gene function rather than gene expression (Haley and de Koning 2006).

II. MATERIALS AND METHODS

Our analysis involved in three main steps: (1) finding candidate genes from the Animal QTL database and analyzing the results of microarray experiments from GEO (Gene Expression Omnibus) database; (2) using these genes and co-expression information to construct co-expression network related to 'marbling score' trait; and analyzing the network topology by visualization; (3) confirmation of gene expression results of hub genes using Quantitative real-time PCR (qRT-PCR).

Identification of candidate genes associated with marbling score

To determine candidate genes associated with marbling score within QTL intervals, we obtained genomic positions of 'Marbling score' trait using 'QTL location by bp' information from the Animal QTL database (http://www.genome.iastate.edu/cgi-bin/QTLdb/BT/index). In the GEO database (http://www.ncbi.nlm.nih.gov/geo/), all data from microarray experiments related to bovine were used: GEO series (GSE) 15544, GSE 15342, GSE 13725, GSE 6918, GSE 10695, GSE 12327, GSE 9256, GSE 12688, GSE 11495, GSE 11312, GSE 7360, GSE 9344 and GSE 8442. All arrays were processed to determine the "Robust Multiarray Average (RMA) (Irizarry, Bolstad et al. 2003)" using "affy" software package (Gautier, Cope et al. 2004). Expression values were computed in detail from raw CEL files by applying the RMA model of probe-specific correction for perfect-match probes. These corrected probe values

were then subjected to quantile normalization, and a median polish was applied to compute one expression measure from all probe values. Resulting RMA expression values were log2-transformed. The gene isoforms and genes containing duplicate probes were excluded by using only those with the highest expression among the redundant transcripts. Finally, we used 844 genes of 1260 genes associated with marbling score for the network construction.

Construction of co-expression network

In co-expression networks, we refer to nodes as genes those degrees that indicate the number of links connected by a node. For unweighted networks, the correlation matrix contains binary information (connected = 1, unconnected = 0). We extracted expression values of 844 genes and evaluated pair-wise correlations between the gene expression profiles of each pair of genes using Pearson correlation coefficients. In order to minimize noise in the gene expression dataset, an edge between two nodes in present if their absolute correlation coefficient exceeds a thresthold r =0.7. We identified key drivers (i.e., hub genes) using network topology. To explore the relationship of nodes in the co-expression network, the following measurements were used to reveal hub genes that play important roles in the network: (1) node degree (or connectivity); (2) the betweenness centrality (BC); (3) the edge BC; and (4) the closeness centrality (CC) (Hwang, Son et al. 2008). The degree of a node is the number of connections or edges the node has to other nodes. The degree distribution of a network has a generalized power-law form $p(k) \sim k^{-r}$, which is the defining property of scale-free network (Barabasi and Albert 1999). The genes of highly connected nodes to nodes with few connections (hubs) play an important role as a local property in a network (Barabasi and Oltvai 2004). A node with high BC has great influence over what flows in the network that may play major roles as a global property since the BC is a useful indicator for detecting bottleneck in a network. For node k BC is the fraction of number of shortest paths that pass through each node (Brandes 2001). We calculated BC as global properties according to all nodes of the network. The edge BC is defined in the same method as BC that an edge is central if it is included in many of the shortest paths connected nodes. The CC use information about the average shorted distance to the other nodes, which is calculated a node is 1/average distance to all other node. The genes with high CC have the ability to contact any node of the network in the shortest possible path. From the results of network topology analysis, we select the high degree nodes and high centrality (BC and CC) nodes as the key drivers that are most associated with our interest trait in a network.

Confirmation of gene expression results by Quantitative real-time PCR (qRT-PCR)

We determine the weather association with expression levels and intramuscular fat content in m. *longissimus* tissue of Korean cattle (Hanwoo). Twelve steers of each group with low-marbled group (9.54±1.35%) and high-marbled group (20.84±1.52%) were used in this study for real-time PCR and statistical analyses. Total RNA was prepared from each tissue sample (100 mg) with TRIzol reagent (Invitrogen Life Technologies, USA) and then purified using RNeasy MinElute Clean-up kit (Qiazen, Valencia, CA, USA). RNA concentration was measured with a NanoDrop ND-1000 spectrophotometer (Thermo scientific, USA). The RNA purity (A_{260}/A_{280}) was over 1.90. For cDNA synthesis, 2 µg RNA was reverse transcribed in a 20 µl reaction volume using random primers (Promega, Madison, WI, USA) and reverse transcriptase (SuperScript II Reverse Transcriptase, Invitrogen Life Technologies). Reactions were incubated at 65°C for 5 min, 42°C for 50 min, and then 70°C for 15 min to inactivate the reverse transcriptase. Real-time PCR was performed using the 2X Power SYBR Green PCR Master mix (Applied Biosystems, USA) with the 7500 Real Time PCR system (Applied Biosystems) using 10 pM of each primer. The PCR was run for 2 min at 50°C and 10 min at 95°C, followed by 40 cycles of 95°C for 10 s, and then 60°C for 1 min. Following amplification, a melting curve analysis was performed to verify the specificity of the reactions. The end point used in the real-time PCR quantification, Ct, was defined as the PCR threshold cycle number. A regression model was used to examine the association between gene expression value and intramuscular fat content by lm function in R. This resulted in the following equation:

Expression = μ + IMF + Age + residual

where Expression is a normalized gene expression value and µ is an overall mean, IMF is intramuscular fat content of each animal and Age is slaughtering age (months) as a covariate and also mRNA level of the beta-actin (ß-actin), ribosomal protein, large, P0 (RPLP0) gene was introduced as a covariate (Hocquette and Brandstetter 2002).

III. RESULTS AND DISCUSSION

Construction of co-expression network

We constructed co-expression network associated with marbling score. The nodes represent candidate genes obtained from the animal QTL database and microarray data, and the links between nodes represent the association between expression profiles. The network comprises 844 nodes: 216 isolated nodes and 668 nodes in 10 clusters, with largest cluster containing 643 nodes. These clustered 643 nodes are connected via 4,344 interactions, which correspond to an effective mean degree of 2.16. Degree is the number of nearest neighbors of a node and effective mean degree is the average degree of all nodes except isolated nodes. The 643 nodes of the network are shown in Figure 1(a).

Analysis of network measures

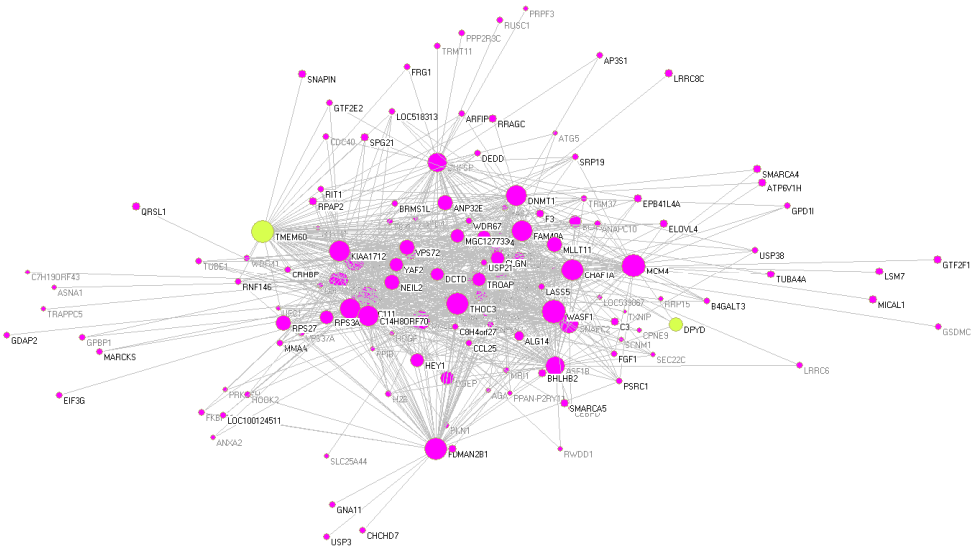
The network follows a power-law ($D(k) \sim k^{-r}$) degree distribution (Figure 1(b)), where r is the degree exponent and ~

indicates ‘proportional to’. Thus, our network has characteristics of scale-free networks whose degree distribution approximates a power law. Highly connected nodes are statistically more significant in a scale-free network than in a random graph. Most of biological networks were characterized by a small number of highly connected nodes, while most other nodes have few connections (Barabasi and Oltvai 2004). The highly connected nodes act as hubs that mediate interactions between other nodes in the network. The hub nodes and nodes with a large BC are summarized in Table 1. The BC is indicator as the global central node. The effect of removing nodes of a large BC is similar to that of removing hub nodes because nodes with a large BC has very correlated fashion that of hub nodes (Son, Kim et al. 2004). They are not hub nodes, they imply that a site of relatively more between all other sites. This means that sites are advantageously located to act as intermediaries. Therefore, we confirmed that hub and large-BC nodes are located in the core to a topological center of the network by calculating the CC.

Table 1. Hub nodes and nodes with large betweenness centrality (BC)

gene	gene description	Hub node	large BC node
TMEM60	transmembrane protein 60	Yes	Yes
CHAF1A	chromatin assembly factor 1, subunit A (p150)	Yes	Yes
MCM4	minichromosome maintenance complex component 4	Yes	Yes
FDX1L	ferredoxin-1-like protein	Yes	Yes
MAEL	maelstrom homolog (Drosophila)	Yes	
HINT1	histidine triad nucleotide binding protein 1	Yes	
DPYD	dihydropyrimidine dehydrogenase		Yes
ELOVL4	elongation of very long chain fatty acids-like		Yes

(a)



(b)

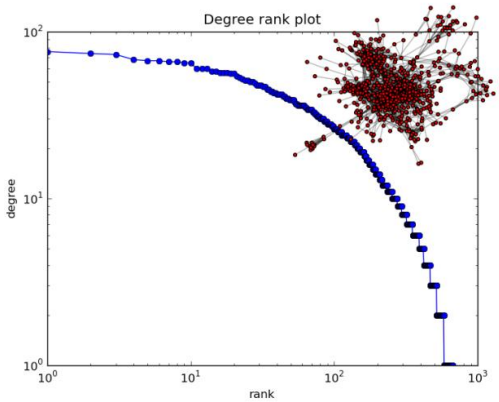


Figure 1. Topological view of co-expression network associated with marbling score. (a) Co-expression network of the marbling score. The edge indicates expression correlation above a threshold (0.7) between the nodes. The node represents candidate genes related to marbling score. Light green represents genes for which there is a consistent result from network analysis and experimental validation. (b) The degree distribution *D(k)* of the network follows a power-law distribution.

Confirmation of gene expression results by Quantitative real-time PCR (qRT-PCR)

We investigated expression levels of ten candidate genes in *m. longissimus* muscle between two distinct intramuscular fat content groups (Table 1). We firstly investigated expression levels of two genes, peroxisome proliferator-activated receptor gamma (PPARG) and CCAAT/enhancer binding protein alpha (C/EBPa) as an indicator for fat accumulation, which are the major transcription factor regulating adipogenesis (MacDougald and Lane 1995). The mRNA expression levels of PPARG and C/EBPa were more highly expressed in the high-marbled group. In present study, we identified two genes, transmembrane protein 60 (TMEM60) and dihydropyrimidine dehydrogenase (DPYD), which were significantly up-regulated according to intramuscular fat content increased ($P<0.05$) (Figure 2).

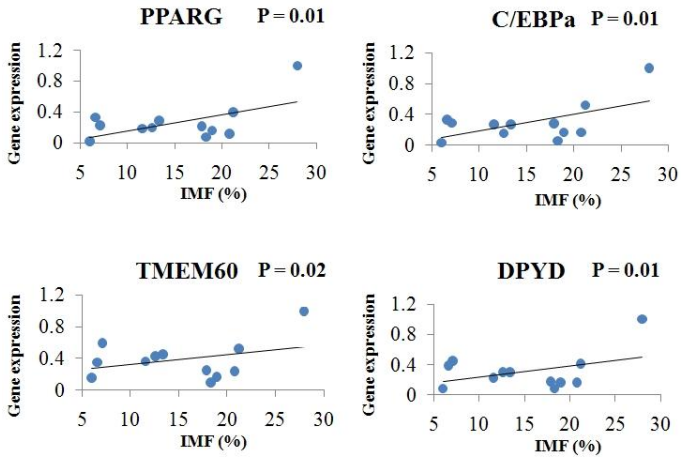


Figure 2. Regression analysis between gene expression value generated by real-time PCR and intramuscular fat (IMF) content. PPARG, peroxisome proliferator-activated receptor gamma; C/EBPa, CCAAT/enhancer binding protein alpha; TMEM60, transmembrane protein 60; DPYD, dihydropyrimidine dehydrogenase.

IV. CONCLUSION

We extracted data related to 'marbling score' trait from Animal QTL database and microarray experiments from the GEO database and subsequently constructed co-expression network using Pearson's correlation matrix that displayed degrees with a power-law distribution, with an exponent of approximately -2. The hub genes were identified and topologically centered with large-degree and BC in the network. We also confirmed that the expression of hub nodes (TMEM60) and nodes with large BC (DPYD) were consistent with network-topology analysis.

ACKNOWLEDGEMENT

This work was supported by Agenda (200901FHT020710399) of National Institute of Animal Science, Rural Development Administration, Republic of Korea.

REFERENCES

Barabasi, A. L. and R. Albert (1999). Emergence of scaling in random networks. *Science* 286(5439): 509-512.
Barabasi, A. L. and Z. N. Oltvai (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5(2): 101-113.
Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* 25(2): 163-177.
Brem, R. B., G. Yvert, et al. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* 296(5568): 752-755.
Gautier, L., L. Cope, et al. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20(3): 307.
Haley, C. and D. J. de Koning (2006). Genetical genomics in livestock: potentials and pitfalls. *Anim Genet* 37 Suppl 1: 10-12.
Hocquette, J. and A. Brandstetter (2002). Common practice in molecular biology may introduce statistical bias and misleading biological interpretation. *The Journal of Nutritional Biochemistry* 13(6): 370-377.
Hwang, S., S. W. Son, et al. (2008). A protein interaction network associated with asthma. *J Theor Biol* 252(4): 722-731.
Irizarry, R. A., B. M. Bolstad, et al. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31(4): e15.
Jansen, R. C. and J. P. Nap (2001). Genetical genomics: the added value from segregation. *Trends Genet* 17(7): 388-391.
MacDougald, O. and M. Lane (1995). Transcriptional regulation of gene expression during adipocyte differentiation. *Annual review of biochemistry* 64(1): 345-373.
Mootha, V. K., P. Lepage, et al. (2003). Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A* 100(2): 605-610.
Schadt, E. E., S. A. Monks, et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422(6929): 297-302.
Son, S., D. Kim, et al. (2004). Response network emerging from simple perturbation. *JOURNAL-KOREAN PHYSICAL SOCIETY* 44(1): 628-632.
Wayne, M. L. and L. M. McIntyre (2002). Combining mapping and arraying: An approach to candidate gene identification. Proc *Natl Acad Sci U S A* 99(23): 14903-14906.