Construction the first gene co-expression-based interactome in cattle

Yan Chen [1] , **Yining Liu** [2] , **Min Du** [3] , **Wengang Zhang** [1] , **Xue Gao** [1] , **Lupei Zhang** [1] , **Huijiang Gao** [1] , **Lingyang Xu** [1] , **Junya Li** Corresp., [1] , **Min Zhao** Corresp. [4]

[1] Innovation Team of Cattle Genetics and Breeding, Institute of Animal Science, Chinese Academy of Agricultural Science, Beijing, China
[2] The School of Public Health, Institute for Chemical Carcinogenesis, Guangzhou Medical University, Guangzhou, China
[3] Department of Animal Science, Washington State University, Pullman, USA
[4] School of Engineering, Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Queensland, Australia

Corresponding Authors: Junya Li, Min Zhao
Email address: jl1@iascaas.net.cn, mzhao@usc.edu.au

Integrating genomic information into cattle breeding is an important approach to exploring the molecular mechanism for complex traits related to diary and meat production. To assist with genomic-based selection, a reference map of interactome is needed to fully understand genotype-phenotype relationships. To this end we constructed a co-expression analysis of 92 tissues and this represents the first systematic exploration of gene-gene relationship in cattle. By using robust WGCNA (Weighted Gene Correlation Network Analysis), we described the gene co-expression network of 13,405 protein-coding genes from the cattle genome. Using the 5,000 genes with majority variations in expression across 92 tissues, we compiled a network with 72,306 co-associations and that provides functional insights into thousands of poorly characterized proteins. Further module identifications found 55 highly organized functional clusters representing diverse cellular activities. To demonstrate the re-use of our interaction for functional genomics analysis, we extracted a sub-network associated with DNA binding genes in cattle. The subnetwork was enriched within regulation of transcription from RNA polymerase II promoter representing central cellular functions. In addition, we identified 28 novel linker genes associated with more than 100 DNA binding genes. Our WGCNA-based co-expression network reconstruction will be a valuable resource for exploring the molecular mechanisms of incompletely characterized proteins and for elucidating larger-scale patterns of functional modulization in the cattle genome.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3178v1 | CC BY 4.0 Open Access | rec: 22 Aug 2017, publ: 22 Aug 2017

Construction the first gene co-expression-based interactome in cattle

Yan Chen[1], Yining Liu[2], Min Du[3], Wengang Zhang[1], Xue Gao[1], Lupei Zhang[1], Huijiang Gao[1], Lingyang Xu[1], Junya Li[1]*, Min Zhao[4]*

[1] Cattle Genetics and Breeding Team, Institute of Animal Science, Chinese Academy of Agricultural Science, No. 2 Yuanmingyuan West Road, Beijing 100193, China;

[2] The School of Public Health, Institute for Chemical Carcinogenesis, Guangzhou Medical University, 195 Dongfengxi Road, Guangzhou 510182, China;

[3] Department of Animal Science, Washington State University, Pullman, WA 99164, USA;

[4] School of Engineering, Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Maroochydore DC, Queensland, 4558, Australia

*Corresponding author:

Min Zhao, email: mzhao@usc.edu.au, Ph: +61 (0)423791085;

Junya Li, email: jl1@iascaas.net.cn, Ph: +86 010 62816065

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3178v1 | CC BY 4.0 Open Access | rec: 22 Aug 2017, publ: 22 Aug 2017

Abstract

Integrating genomic information into cattle breeding is an important approach to exploring the molecular mechanism for complex traits related to diary and meat production. To assist with genomic-based selection, a reference map of interactome is needed to fully understand genotype-phenotype relationships. To this end we constructed a co-expression analysis of 92 tissues and this represents the first systematic exploration of gene-gene relationship in cattle. By using robust WGCNA (Weighted Gene Correlation Network Analysis), we described the gene co-expression network of 13,405 protein-coding genes from the cattle genome. Using the 5,000 genes with majority variations in expression across 92 tissues, we compiled a network with 72,306 co-associations and that provides functional insights into thousands of poorly characterized proteins. Further module identifications found 55 highly organized functional clusters representing diverse cellular activities. To demonstrate the re-use of our interaction for functional genomics analysis, we extracted a sub-network associated with DNA binding genes in cattle. The subnetwork was enriched within regulation of transcription from RNA polymerase II promoter representing central cellular functions. In addition, we identified 28 novel linker genes associated with more than 100 DNA binding genes. Our WGCNA-based co-expression network reconstruction will be a valuable resource for exploring the molecular mechanisms of incompletely characterized proteins and for elucidating larger-scale patterns of functional modulization in the cattle genome.

Keywords:

Co-expression, network, WGCNA, systems biology, functional enrichment, cattle

Introduction

As the importance in dairy and beef production, the genome of the domestic cattle, *Bos taurus*, was sequenced in 2009 using hierarchical and whole-genome shotgun sequencing strategy (Zimin et al. 2009). To associate the genetic variation with phenotypes, the first phase of the 1000 bull genomes project was started to sequence 234 ancestor bulls (Daetwyler et al. 2014). Although more and more efforts for genetic improvement of production efficiency and quality in cattle, majority of previous studies focused on single-gene based genetic breeding (Barabasi & Oltvai 2004). However, most of production traits are complex traits involving multiple genes. The recent development of systems biology-based approach was promising to explore the genome and gene-gene interactions in a global view to understand molecular mechanisms underlying complex traits (Zhao et al. 2014).

An gene-based interactome is the complete set of gene-gene interactions in a particular cell (Barabasi & Oltvai 2004) and these could be direct physical interactions among molecules as well as indirect interactions among genes (such as gene co-expression). The understanding of interactomes are important in systems biology-based studies as they provide a global view of all the possible molecular interactions that a protein can influence (Barabasi & Oltvai 2004). Because of lacking interactome in cattle, the network-based data mining approach are not able to apply to functional discovery for any interesting genes associated with complex traits (Elsik et al. 2016). Recently, a functional proteomic and interactome analysis of the proteins of Angus cattle was presented (Mitra et al. 2014). However, this data is specific for beef tenderness with limited tissues, not useful for other large-scale functional studies. With the development of next-generation sequencing technologies, cumulative expression data across multiple tissues in cattle are now publicly available and may promote the understanding of gene-gene interaction from network approach (Elsik et al. 2016).

In this study, we hypothesize that the complex genetic traits related to cattle production is reflected by the perturbation of gene-gene co-expressing networks. To this aim, we built the first co-expression based interactome for cattle through integrating expression profiles from 92 tissues from bovine genome database (BGD) (Elsik et al. 2016). In this study, we utilized an established network-based approach, Weighted Gene Co-Expression Network Analysis (WGCNA) (Langfelder & Horvath 2008) , to further identify and characterize a number of functional modules. To demonstrate our reconstructed interactome could provide a new approach for network-based data mining of cattle genetics data, we focused on the DNA-binding genes in cattle and extracted a DNA-binding regulatory network.

Materials & methods

The gene expression data in 92 tissues from bovine genome database

To characterize the gene expression in multiple tissues, the bovine genome database (BGD) collected gene expression data from 92 different tissues from the individual of the reference genome (Elsik et al. 2016). By using RNAseq sequencing and mapping to the reference genome, all the genes in cattle genome was quantified using the FPKM (Fragments Per Kilobase of transcript per Million mapped reads). All the FPKM were further normalized for each expression dataset by using cuffquant and cuffnorm. By using Intermine Web Services API of BovineMine (part of BGD), we downloaded all the normalized FPKM values of the 92 tissues. To further build the co-expression network based on high-quality data, we first removed those non-informative genes with FPKMs in 46 or less tissue samples. After the initial filtering, a list of 13,405 genes with FPKMs were subject to WGCNA analysis.

Weighted Gene Co-Expression Analysis (WGCNA)

WGCNA is a R package to construct gene co-expression networks. By using the package, we first built similarity matrix between all the gene pairs using bi-weight mid-correlation based on normalized FPKMs (Zheng et al. 2014). The expression similarity matrix was further transformed to an adjacency matrix by using the soft thresholding power Beta. By further focusing on the top 5000 genes with more variations across samples, we run the gene co-expression analysis and build the interaction network for all the 5000 genes. To choose a suitable threshold for reconstruction of co-expression network, we adopted a parameter analysis on the Beta value with most approximating scale-free topology of the network (Langfelder & Horvath 2008). As shown in Figure 1, the final optimal Beta value was 4 based on the scale-free topological analysis.

The identification of functional modules

To further identify functional modules in our reconstructed co-expression network with 5000 genes, the adjacency matrix was further transformed to topological overlap matrix (TOM) using WGCNA package. The hierarchical clustering on all the genes were performed to generate a dendrogram. By using dynamic tree cutting, the functional clusters (modules) were obtained from the constructed gene dendrogram. In detail, the cutreeDynamic function in WGCNA package was used to identify the larger module with minimum size of 30 genes as possible. By setting parameter deepSplit from 0 to 4 for the tree cutting, we found the optimal value to generate smaller clusters as more genes as possible. The final deepSplit of 4 was chosen and resulted in 55 modules with average size of 235 genes. Those identified functional modules are illustrated with different colours on the bottom of the Figure 2A. The relationship between modules were further summarized by eigenvalue “eigengene”. Eigengenes are defined as the first principal component of the expression matrix for each identified functional module. Therefore the eigengenes represent the expression profile with weighted genes for each module (Langfelder & Horvath 2007).

Pathway enrichment analysis and network analysis

We performed pathway enrichment analysis on those interested genes by using functional enrichment tools in BGD (Elsik et al. 2016). This online tool includes enrichment in predefined pathways from KEGG and Gene Ontology. The reconstructed co-expression network from WGCNA was visualized using the Cytoscape (version 3.4). The topological centrality analysis was performed by using NetworkAnalyzer in Cytoscape (Shannon et al. 2003). We used degree to represent the sum of the number of connections for each node in a network, and the shortest path represented by the least number of steps from one node to another (Barabasi & Oltvai 2004). By using the sub-network extraction algorithm described in our previous study (Zhao et al. 2015), we built a sub-network to link the 340 DNA binding genes with the other cattle genes. The 340 genes were mapped into the prepared co-expression interactome from WGCNA analysis and the sub-network was extracted according to the shortest paths between the input 340 genes and other genes.

Results

Reconstruction of a scale-free co-expression network from 92 cattle tissues using WGCNA

Network-based data mining is used to explore the behavior of all the gene-gene interactions and the total of these is greater than would be expected from the sum of all the gene functions. However, there is limited information about cattle in the gene-gene interaction database and, for instance, BioGrid (Chatr-Aryamontri et al. 2017) contains only 102 interaction pairs for cattle. To overcome this shortcoming, we used the mature bioinformatics co-expression network approach to reconstruct the functional interactome for cattle. Based on comprehensive transcriptomes with 92 tissue samples covering the majority tissue types in cattle body, we built and mined the gene co-expression network using the WGCNA analysis.

Using 19,064 genes with expression values, we ran a quality control step and removed those genes without expression values in more than half of 92 tissues. This provided a list of 13,405 genes with expression across 92 tissues. However, a large number of these genes were not differentially expressed between samples. Therefore, the data set with 13,405 gene expression was processed further by focusing on the 5,000 most variant genes (Table S1). The remaining 8,405 genes, which showed no or very low changes in expression between samples, were not used for WGCNA analysis. The variability of gene expression data across the 92 samples was measured using a robust method called median absolute deviation (MAD). The 5000 most variant genes were used for analysis in other WGCNA studies (de Jong et al. 2012).

To build a scale-free network, we run a parameter analysis (Figure 1). Briefly, an adjacency function in WGCNA was used to weight between different genes in the hypothesis of following a power law. In detail, the correlation data were transformed to adjacency matrix using the formula: $a_{ij} = (S_{ij}, \beta) = |S_{ij}|^{\beta}$. In the formula, the β represent the exponential parameter for power law distribution. Normally, the β was used to characterize the likeness to a scale-free network. In our data, the co-expression for a pair of gene represent a connection between two genes. In general, the number of connection of all the genes in a scale-free network follow a power law distribution $P(k)$~$k^{-\beta}$. The $P(k)$ in our co-expressing network indicates the probability that a gene is co-expressed with k other genes. By setting the criterion that the co-efficiency of log(k) and log(p(k)) is greater than 0.8, we checked all the possible β values. As shown in Figure 1A, we changed the β value step by step to identify the optimal value that the average connectivity of the network is smooth. The $\beta = 4$ was finally determined based on the diagnosis chart and the average number of co-expressed genes in the final network was 80 (Figure 1B). Using this information, we reconstructed the first and most co-expression network in cattle genome across 92 tissue samples representing the majority of tissue types; this will provide a basis for network-based data mining in cattle genetics and genomics studies.

Functional module identification on co-expression network using WGCNA and functional enrichment analyses for the genes in the top five modules

To determine the similarity between genes, the WGCNA consider not only the co-expression coefficients between genes, but also the content of co-expressed gene partners. To this aim, a topological overlap matrix (TOM) was calculated based on the adjacent coefficient and how many shared “friends” between any pairs of co-expressed genes. In this way, all the edges between co-expressed genes were weighted by TOM ranging from 0 to 1, which represent the strength of the communication between the two genes. To identify the clustered co-expressed genes with specific functions, we further conducted module identification using using agglomerative hierarchical clustering based TOM (Figure 2A). Since it was hard to associate small number of genes to specific biological function, we required any functional modules with at least 10 genes.

To validate the potential functions for the modules, we focused on the top five modules with most genes (Table S2). Pathway and gene ontology (GO) enrichment analysis of the chosen modules were performed with BovineMine of BGD. Table 1 shows functionally enriched pathways obtained from BovineMine by setting adjusted P-value < 0.05. We found enriched pathways only for module 1 and module 2. The genes in module 1 were identified as associated with metabolic pathways: there are three genes related to isoleucine degradation. A previous carbon-14 labelling experiment showed that the degradation of valine, leucine, and isoleucine represent a potential source of energy to the mammary gland as well as a source of carbon and alpha-amino nitrogen for the synthesis of nonessential amino acids (Wohlt et al. 1977). The genes from module 2 have extensive roles in extracellular processing and are associated with 15 pathways (Table 1). These pathways are known to be key components in the extracellular signaling system that involve collagen formation and degradation, glycosaminoglycan metabolism and axon guidance (Table 1).

By using the GO enrichment analysis, we further discovered more functional features for the five modules (Table 2). Those genes in module 1 (M1) are mainly metabolism related pathways (all adjusted P-values < 0.05). The components of module 2 (M2) are associated with extracellular structure organization and protein hetero-trimerization and trimerization (adjusted P-values < 0.05). The genes in module 3 (M3) use a microtubule cytoskeleton to organize cell projection (all adjusted P-values < 0.05). The module 4 (M4) is mainly related to pigment cell differentiation and its regulation (two adjusted P-values < 0.05). The genes in module 5 (M5) are enriched for the development of sertoli cells (adjusted P-values < 0.05), which are essential for spermatogenesis. Based on Pearson correlation coefficients, we further explored the relationship between modules. The module eigengenes are further calculated, which provides quantitative assessments for the similarity between the modules (Table S3). As shown in Figure 2B, the top five modules are not clustered together which implies that they have different functions. Combined with our functional results from KEGG pathway and GO, we concluded that the top five modules have distinct and independent functions at the cellular level.

The hub genes in a co-expression based interactome with manageable size

In contrast to the correlation-based network reconstruction, WGCNA considered not only the expression correlation between two genes but also how many co-expressing genes were shared. In WGCNA, the weighted measure TOM was used to reflect the strength of the communication between the two genes and ranged from 0 to 1. In theory, the reconstructed network comprised all the 5000 genes based on the TOM of >0. However, the resulting network is too large for functional genomics analysis. Since our aim was to build a comprehensive interactome covering as many genes with variant expression as possible, we defined three set of the co-expression gene network by using different TOM thresholds. For a TOM >0.01, the resulting co-expression based interactome comprised 4,995 genes with 1,538,522 significant co-expression pairs. With a TOM >0.1, the interactome comprised 4,403 genes with 72,306 significant co-expression pairs and for TOM scores greater than >0.3, there were 2,119 significant co-expression pairs and 1,045 genes.

To visualize the entire network, we used a TOM score >0.1 which covered the about 90% genes in the 5,000 genes but, as seen in Figure 3A, the network is still too large to obtain detail. The diameter of the network is 11 and the average number of neighbors is 32.844. Further network topological analysis revealed that most genes in the reconstructed co-expression network are closely connected. In detail, we found that the probability $P(k)$ for genes with other k co-expressed genes could be fitted to a power law distribution ($P(k){\sim}k^{\beta}$). The estimated β is 1.368 (Figure 3B), which indicate this co-expression network are more closely connected compared to published human protein-protein interaction network with estimated β value of 2.9 (Jin et al. 2013). By further analysis the shortest pathways between all the co-expressed genes, we found the majority of the genes could connected with other genes by co-expressing with three or four more genes (Figure 3C).

In addition, our reconstructed network also helped to identify a number of genes with hundreds of co-expressed genes. In general, these potential hub genes may have central roles for signaling transduction or metabolic transformation. In total, we identified 340 genes with 100 or more co-expressed genes (Table S4) and these genes are involved in fundamental processes: ribonucleotide binding (adjusted P-value = 1.253E-2, 54 genes); RNA binding (54 genes, adjusted P-value = 2.219E-3); RNA polymerase binding (6 genes, adjusted P-value = 4.696E-3); and cyclin-dependent protein kinase (5 genes, adjusted P-value = 1.440E-2). Additionally, there are 20 ATPases (adjusted P-value = 8.199E-3), which may indicates the importance of ATPases in the maintenance of metabolite homeostasis in cattle.

Using the number of connections is the most common way to identify the key genes with important functions (Zhao & Qu 2009). Interestingly, we identified *API5* (apoptosis inhibitor 5) as the gene with highest degree (number of connection = 279). This apoptosis inhibitory protein often prevents apoptosis after growth factor deprivation in humans (Han et al. 2012). As one of the genes with most co-expressed gene partners, *API5* may have critical functions in the cattle development and association with complex genetic traits. Another promising gene is *FBXO11* with hundreds of co-expressed genes in cattle genome (Table S4). As one of gene member of the F-box protein family, *FBXO11* was functioned as a suppressor of p53 function by post-translational modification (Abida et al. 2007). In summary, our reconstructed co-expression network across 92 tissue samples may provide unexplored functional clues for many of the genes with a large number of connections in cattle.

A gene-gene interaction sub-network related to DNA binding

To demonstrate the application of our reconstructed interactome, we downloaded 614 DNA binding genes from BGD (Table S5). Then, we connected these genes to form a functional network using the method implemented in our previous studies (Zhao et al. 2016a). The resulted sub-network contained 132 genes and 251 interactions (Figure 4A, Table S6). In total, there were 104 genes from our original DNA binding genes, and 28 genes functioned as linker genes to fully connect the DNA binding genes. The degrees of all genes followed a power law distribution $P(k){\sim}k^{-b}$, where b is estimated as 1.388 (Figure 4B) comparing to 1.368 (Figure 3B). Although only 17% of the 614 DNA binding genes are co-expressed, they all formed highly modular structures, which implies coordination in DNA binding-related gene regulation. For example, we found 39 genes were involved in regulation of transcription from RNA polymerase II promoter (adjusted P-value = 2.04E-11). Similarly, there are 36 genes associated with “positive regulation of gene expression” (adjusted P-value = 2.04E-11) and 26 genes associated with “negative regulation of gene expression” (adjusted P-value = 2.43E-5). Taken together, the competitive regulation may be associated with RNA polymerase II promoter regions. With regard to the 28 linker genes, we found only three genes (*AGO4, CAPRIN1, CNOT3*) localized to “P-body” (GO:0000932, adjusted P-value = 0.03) and two genes (*AXIN1* and *CALCOCO1*) that have “armadillo repeat domain binding” (GO:0070016, adjusted P-value = 3.24E-2). Although the majority are not statistically over-represented in any functional modules, their strong co-expression with hundreds of DNA binding regulators may imply their important role in cellular processes.

In summary, by applying the sub-network extraction to the DNA binding genes in cattle, we successfully identified a sub-network with hundreds of DNA binding genes and a number of relevant novel genes. This demonstrated that the use of our reconstructed co-expression interactome is a powerful approach to cluster genes with similar function for network-based data mining in cattle genetics and genomics studies in general.

Discussion and conclusion

The cellular machines can be viewed as the product of thousands of proteins necessary to maintain cellular signalling and respond to extracellular stimulation. The genome-wide gene expression is coordinated in part through networks of protein-protein interactions that assemble functionally related proteins into complexes and organelles. Understanding the architecture of the cattle transcriptome will improve our knowledge of cellular, structural and molecular mechanisms. For instance, those co-expressed genes may have similar biological functions. Ans this co-expression information could be used to elucidating how genome variation and expression contributes to the cattle breeding. Here we present the first co-expression based interactome in cattle. This data will not only enhance network-based characterization of subcellular localization and complex formation, but also provide the basis for network-based mining for specific functional modules.

By using robust co-expression analysis, we characterized a number of interesting genes for further investigation that formed tightly interconnected cluster in our co-expression network. Our further topological analysis revealed 340 highly-connected genes with 100 or more connections that may act as important links in various biological processes. For example, *FBXO11* was identified to play a role in the p53 pathway. Combined with the results from the enrichment analysis of ribonucleotide binding, this gene may be one of the fundamental regulators involved in the suppression of p53 function. The p53 pathway was not only associated with bovine virus-induced leukemogenesis in cattle but is also important in human cancer (Zhao et al. 2016b). Therefore, the identification of p53 inhibitor, *FBXO11*, as a hub gene may provide a feasible approach for the design of molecular inhibitors to prevent p53-related diseases in cattle. Another interesting gene that shows a large connection in cattle co-expression network is *API5*, an apoptosis inhibitor that is involved in the fibroblast growth factor binding. Since cell apoptosis has an important role in vitro-produced beef cattle embryos (Nkadimeng et al. 2016), our result may offer a number of new genes for identifying novel mechanisms of vitro-produced embryos in cattle.

Our additional module analysis identified 55 highly-connected functional modules representing diverse cellular activities. By focusing on the top five modules with the largest number of genes, we characterized some important functions for these modules. For example, there are three genes (*BCKDHA*, *ETFB*, and *PHLDB2*) involving isoleucine degradation in module 1. More interestingly, the biochemical intermediates and final products from the isoleucine degradation pathway are the potential energy source for the mammary gland in cattle (Wohlt et al. 1977).

Moreover, our reconstructed network will serve as a basis for network-based mining as exemplified by the identified sub-network related to DNA binding genes in cattle. This work highlights the importance of a systems biology approach to study largely unexplored transcriptomes by analysing the inherent modularity of the co-expression network concerned with the majority tissue samples. In conclusion, we performed the first systematically co-expression analysis on thousands of genes in cattle genome across 92 tissues. The resulted co-expression pairs connected thousands of genes with similar functions and formed the first cattle interactome for large scale systems biology-based data mining.

Acknowledgements

This work was supported by National Natural Science Foundation of China (31402039), the start-up grant to Dr. Min Zhao, China Scholarship Council (CSC), Cattle Breeding Innovative Research Team of Chinese Academy of Agricultural Sciences (cxgc-ias-03), and National Beef Cattle Industrial Technology System of China's Agriculture and Finance Ministries (CARS-37).

Disclosure of potential Conflict of interest

The authors declare that they have no competing interests.

References

Abida WM, Nikolaev A, Zhao W, Zhang W, and Gu W. 2007. FBXO11 promotes the Neddylation of p53 and inhibits its transcriptional activity. *J Biol Chem* 282:1797-1804. 10.1074/jbc.M609001200

Barabasi AL, and Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5:101-113. 10.1038/nrg1272

Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz BJ, Dolinski K, and Tyers M. 2017. The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 45:D369-D379. 10.1093/nar/gkw1102

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, Liao X, Djari A, Rodriguez SC, Grohs C, Esquerre D, Bouchez O, Rossignol MN, Klopp C, Rocha D, Fritz S, Eggen A, Bowman PJ, Coote D, Chamberlain AJ, Anderson C, VanTassell CP, Hulsegge I, Goddard ME, Guldbrandtsen B, Lund MS, Veerkamp RF, Boichard DA, Fries R, and Hayes BJ. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* 46:858-865. 10.1038/ng.3034

de Jong S, Boks MP, Fuller TF, Strengman E, Janson E, de Kovel CG, Ori AP, Vi N, Mulder F, Blom JD, Glenthoj B, Schubart CD, Cahn W, Kahn RS, Horvath S, and Ophoff RA. 2012. A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PLoS One* 7:e39498. 10.1371/journal.pone.0039498

Elsik CG, Unni DR, Diesh CM, Tayal A, Emery ML, Nguyen HN, and Hagen DE. 2016. Bovine Genome Database: new tools for gleaning function from the Bos taurus genome. *Nucleic Acids Res* 44:D834-839. 10.1093/nar/gkv1077

Han BG, Kim KH, Lee SJ, Jeong KC, Cho JW, Noh KH, Kim TW, Kim SJ, Yoon HJ, Suh SW, Lee S, and Lee BI. 2012. Helical repeat structure of apoptosis inhibitor 5 reveals protein-protein interaction modules. *J Biol Chem* 287:10727-10737. 10.1074/jbc.M111.317594

Jin Y, Turaev D, Weinmaier T, Rattei T, and Makse HA. 2013. The evolutionary dynamics of protein-protein interaction networks inferred from the reconstruction of ancient networks. *PLoS One* 8:e58134. 10.1371/journal.pone.0058134

Langfelder P, and Horvath S. 2007. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 1:54. 10.1186/1752-0509-1-54

Langfelder P, and Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. 10.1186/1471-2105-9-559

Mitra R, Edmonds MD, Sun J, Zhao M, Yu H, Eischen CM, and Zhao Z. 2014. Reproducible combinatorial regulatory networks elucidate novel oncogenic microRNAs in non-small cell lung cancer. *RNA* 20:1356-1368. 10.1261/rna.042754.113

Nkadimeng M, van Marle-Koster E, Netshirovha TR, Nedambale TL, and Lehloenya KC. 2016. 145 the Role of Cell Apoptosis on in Vitro-Produced Beef Cattle Embryos. *Reprod Fertil Dev* 29:181. 10.1071/RDv29n1Ab145

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498-2504. 10.1101/gr.1239303

Wohlt JE, Clark JH, Derrig RG, and Davis CL. 1977. Valine, leucine, and isoleucine metabolism by lactating bovine mammary tissue. *J Dairy Sci* 60:1875-1882. 10.3168/jds.S0022-0302(77)84118-0

Zhao M, Chen L, and Qu H. 2016a. CSGene: a literature-based database for cell senescence genes and its application to identify critical cell aging pathways and associated diseases. *Cell Death Dis* 7:e2053. 10.1038/cddis.2015.414

Zhao M, Kim P, Mitra R, Zhao J, and Zhao Z. 2016b. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res* 44:D1023-1031. 10.1093/nar/gkv1268

Zhao M, Kong L, and Qu H. 2014. A systems biology approach to identify intelligence quotient score-related genomic regions, and pathways relevant to potential therapeutic treatments. *Sci Rep* 4:4176. 10.1038/srep04176

Zhao M, Liu Y, and O'Mara TA. 2015. ECGene: A Literature-Based Knowledgebase of Endometrial Cancer Genes. *Hum Mutat*. 10.1002/humu.22950

Zhao M, and Qu H. 2009. Human liver rate-limiting enzymes influence metabolic flux via branch points and inhibitors. *BMC Genomics* 10 Suppl 3:S31. 10.1186/1471-2164-10-S3-S31

Zheng CH, Yuan L, Sha W, and Sun ZL. 2014. Gene differential coexpression analysis based on biweight correlation and maximum clique. *BMC Bioinformatics* 15 Suppl 15:S3. 10.1186/1471-2105-15-S15-S3

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marcais G, Roberts M, Subramanian P, Yorke JA, and Salzberg SL. 2009. A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol* 10:R42. 10.1186/gb-2009-10-4-r42

Figures

Figure 1 - Determination of power Beta value based on the adjacency matrix using the weighted gene correlation network analysis (WGCNA).

The adjacency matrix from co-expression data was weighted by the power of the correlation data between different genes; i.e., $a_{ij} = |S_{ij}|\beta$. The weighted parameter power Beta value was determined by the scale-free topology criterion. To ensure that the average connectivity of the network is smooth, we chose $\beta = 4$ based on both chart: (A) for topology fitting results and (B) for mean connectivity.

Figure 2 - The WGCNA analysis on the top 5000 genes with most variation across 92 tissues in cattle.

(A) Functional modules are illustrated with different colours. The parameter *deepslip*=4 is set in WGCNA analysis, which providing a high sensitivity to cluster splitting. We additionally required each gene module with 30 or more genes. In total, 4950 genes were grouped into 56 modules which showed with various colours. The top five modules ordered by number of genes were: turquoise with 212 genes; blue with 201 genes; brown with 187 genes; yellow with 162 genes; and green with 155 genes. The grey colour in the left of the figure represents the 50 genes not associated with any module. (B) The relationship tree for all the modules is presented and the top five modules marked in the corresponding number.

Figure 3 - The co-expression network and gene ontology analysis of 340 genes with 100 or more connections.

(A) Co-expression network from WGCNA based on the TOM greater than 0.1; (B) degree distribution for the network; and (C) short path length frequency for the network. The scatterplot (D) shows the gene ontology (GO) cluster representatives for the 340 genes in a two-dimensional space derived by applying multidimensional scaling to a matrix of the GO terms semantic similarities. Bubble colour indicates the corrected P-value of the GO term.

Figure 4 - The sub-network for the DNA binding genes in cattle.

(A) the sub-network extracted for DNA binding genes in cattle; (B) the degree distribution for the network; (C) the short path length frequency for the network.

Tables

Table 1 – The enriched KEGG pathways for the genes in module 1 and 2 from WGCNA analysis.

Pathway	# of genes	Q-value
Module 1		
Metabolism	43	6.26E-07
Isoleucine degradation	3	0.04218
Module 2		
Collagen formation	14	4.54E-12
Extracellular matrix organization	21	4.92E-12
Collagen biosynthesis and modifying enzymes	13	1.54E-11
ECM proteoglycans	11	2.85E-10
Collagen degradation	10	7.38E-09
Assembly of collagen fibrils and other multimeric structures	9	3.76E-08
Degradation of the extracellular matrix	12	5.32E-08
Integrin cell surface interactions	11	2.07E-07
NCAM1 interactions	6	8.55E-06
Glycosaminoglycan metabolism	9	0.00409
MET activates PTK2 signaling	5	0.00967
Cooperation of PDCL (PhLP1) and TRiC/CCT in G-protein beta folding	5	0.01919
Non-integrin membrane-ECM interactions	5	0.02361
Axon guidance	14	0.04552

Note: * Q-values: the raw P-values of the hypergeometric test were corrected by Benjamini-Hochberg multiple testing correction.

Table 2 – The enriched biological processes GO terms for the genes in the top five modules from WGCNA analysis.

Modules	GO: Biological process	Q-values
M1	Small molecule metabolic process	0.000971

M1	Carboxylic acid metabolic process	0.00332
M1	Oxoacid metabolic process	0.003628
M1	Organic acid metabolic process	0.005205
M1	Single-organism metabolic process	0.041382
M2	Extracellular matrix organization	0.000392
M2	Extracellular structure organization	0.000427
M2	Protein heterotrimerization	0.000438
M2	Collagen fibril organization	0.000636
M2	Protein trimerization	0.004188
M3	Cell projection organization	0.013119
M3	Microtubule cytoskeleton organization	0.028215
M3	Microtubule-based process	0.045173
M3	Nervous system development	0.04747
M4	Pigment cell differentiation	0.006709
M4	Regulation of pigment cell differentiation	0.008956
M4	Developmental pigmentation	0.024965
M4	Melanocyte differentiation	0.026407
M5	Sertoli cell development	0.00372

Note: * Q-values: the raw P-values of the hypergeometric test were corrected by Benjamini-Hochberg multiple testing correction.

Additional files

Additional file 1 – Table S1. The expression profile for the top 5,000 most variant genes across 92 tissue samples.

Additional file 2 – Table S2. The top five gene modules with most genes in WGCNA analysis.

Additional file 3 – Table S3. The eigengenes for the gene modules from WGCNA analysis.

Additional file 4 – Table S4. The number of connections for all the genes in the co-expression network from WGCNA.

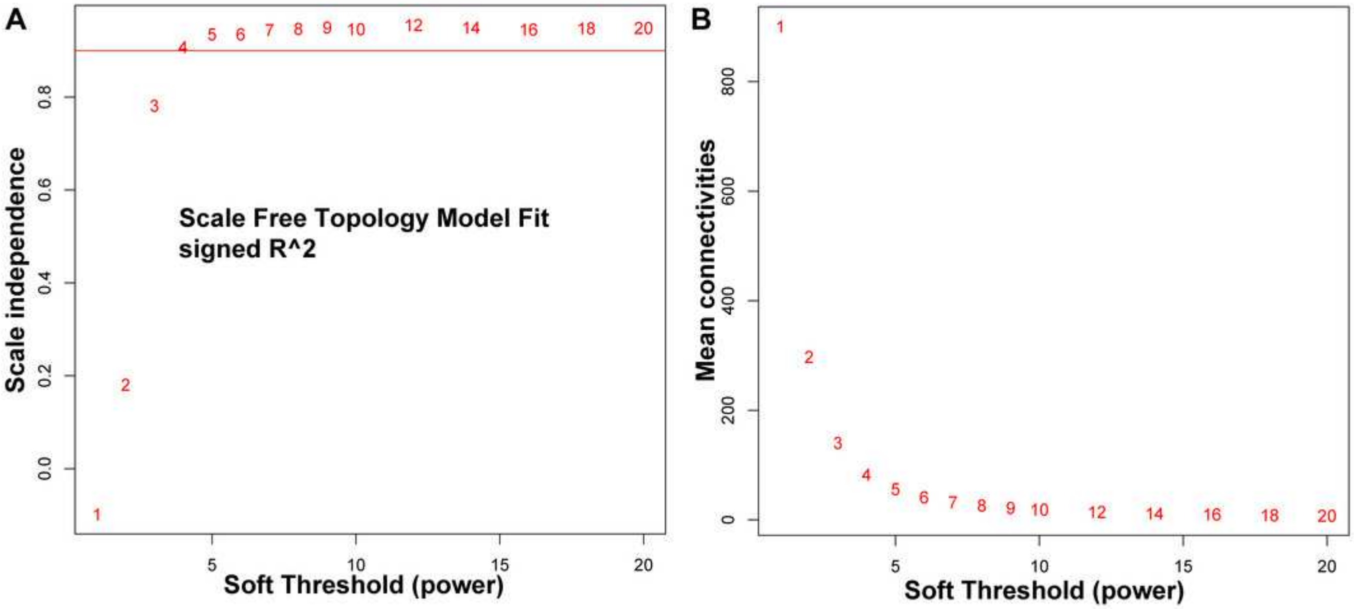
Additional file 5 – Table S5. The gene related to DNA binding in cattle.

Additional file 6 – Table S6. The gene types for the extracted sub-network related to DNA binding.

Figure 1

Determination of power Beta value based on the adjacency matrix using the weighted gene correlation network analysis (WGCNA).

The adjacency matrix from co-expression data was weighted by the power of the correlation data between different genes; i.e., aij = |*Sij*|β. The weighted parameter power Beta value was determined by the scale-free topology criterion. To ensure that the average connectivity of the network is smooth, we chose *β* = 4 based on both chart: (A) for topology fitting results and (B) for mean connectivity.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3178v1 | CC BY 4.0 Open Access | rec: 22 Aug 2017, publ: 22 Aug 2017

Figure 2

The WGCNA analysis on the top 5000 genes with most variation across 92 tissues in cattle.

(A) Functional modules are illustrated with different colours. The parameter *deepslip*=4 is set in WGCNA analysis, which providing a high sensitivity to cluster splitting. We additionally required each gene module with 30 or more genes. In total, 4950 genes were grouped into 56 modules which showed with various colours. The top five modules ordered by number of genes were: turquoise with 212 genes; blue with 201 genes; brown with 187 genes; yellow with 162 genes; and green with 155 genes. The grey colour in the left of the figure represents the 50 genes not associated with any module. (B) The relationship tree for all the modules is presented and the top five modules marked in the corresponding number.

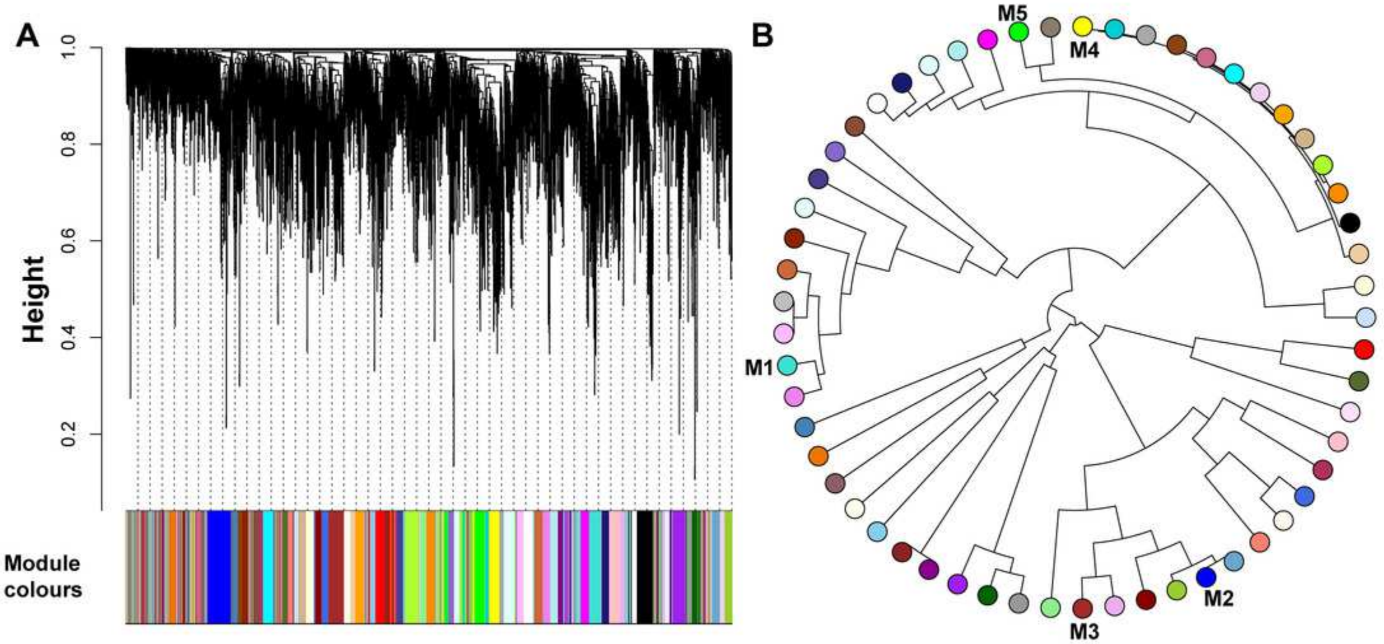
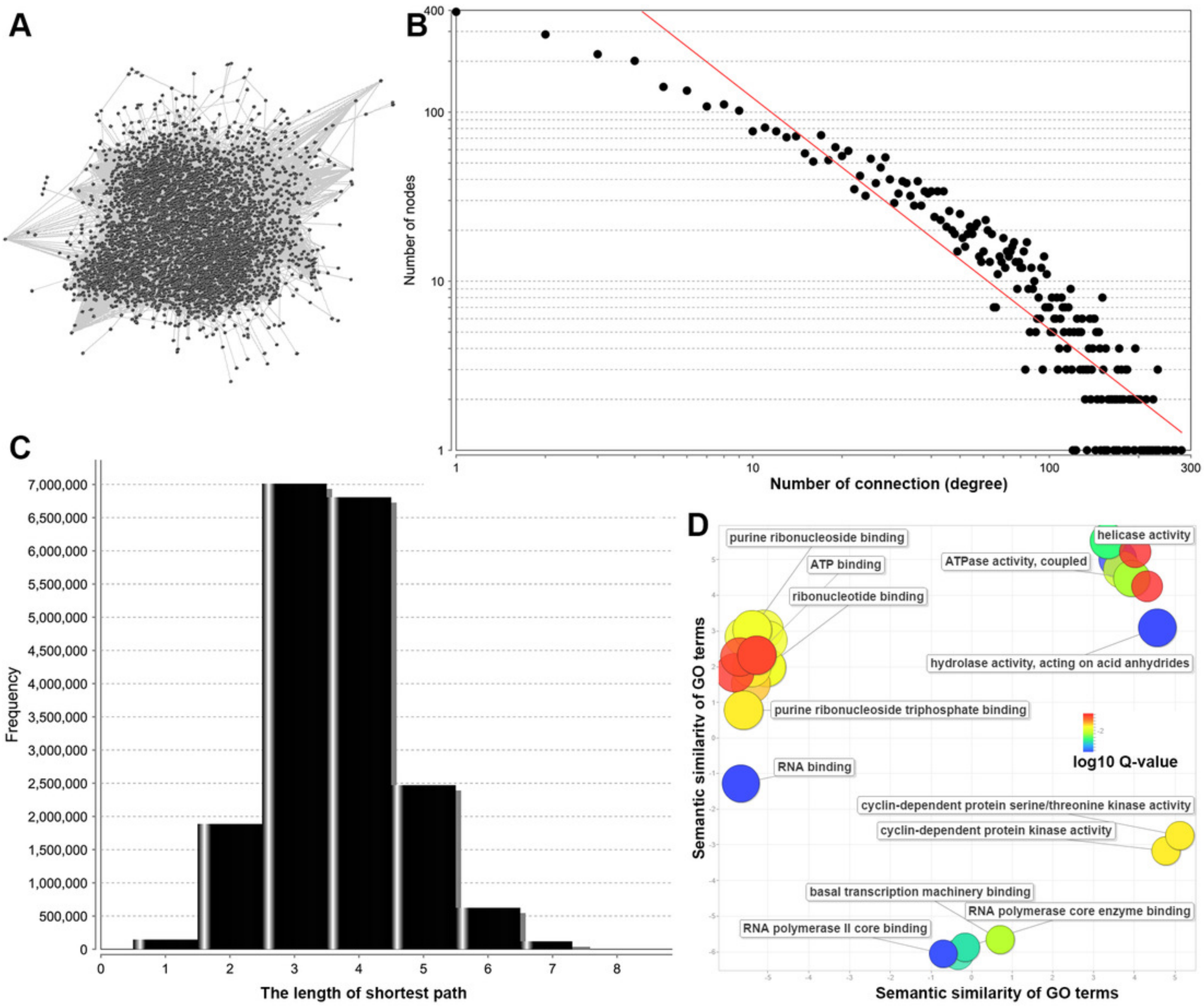


Figure 3

The co-expression network and gene ontology analysis of 340 genes with 100 or more connections.

(A) Co-expression network from WGCNA based on the TOM greater than 0.1; (B) degree distribution for the network; and (C) short path length frequency for the network. The scatterplot (D) shows the gene ontology (GO) cluster representatives for the 340 genes in a two-dimensional space derived by applying multidimensional scaling to a matrix of the GO terms semantic similarities. Bubble colour indicates the corrected P-value of the GO term.

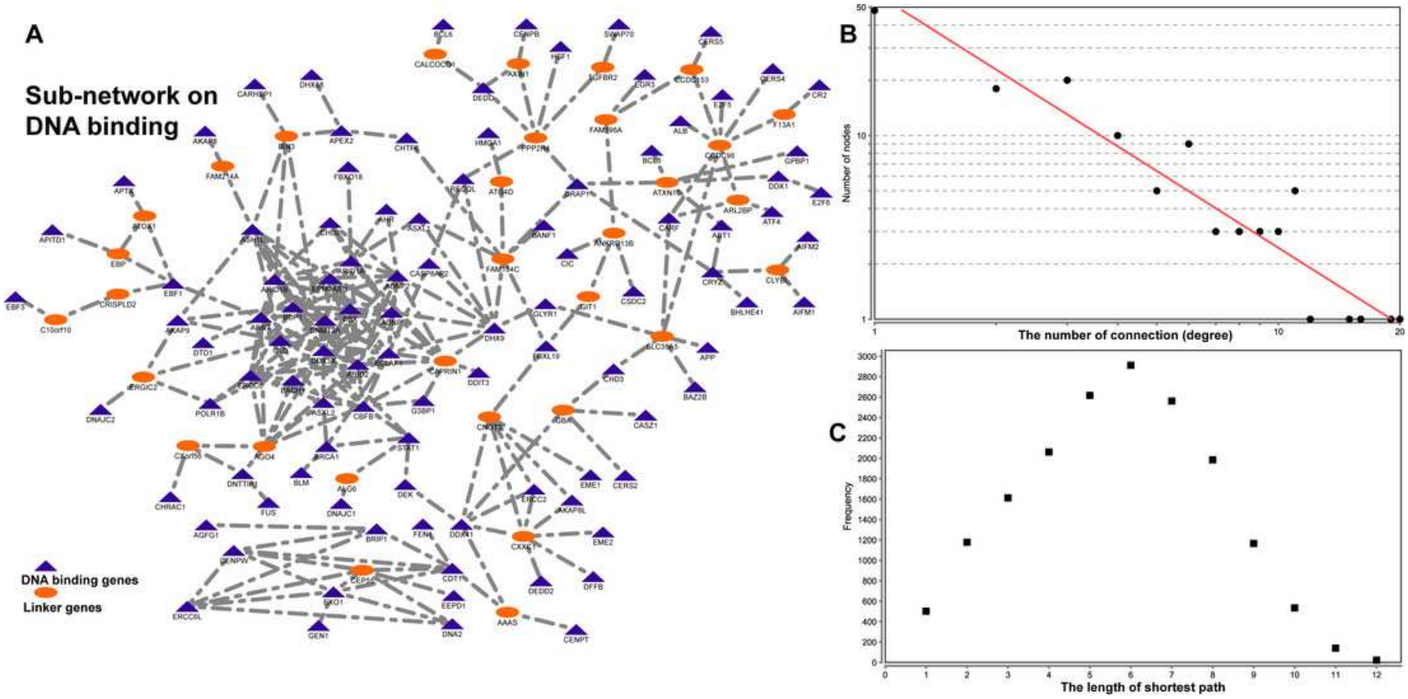


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3178v1 | CC BY 4.0 Open Access | rec: 22 Aug 2017, publ: 22 Aug 2017

Figure 4

The sub-network for the DNA binding genes in cattle.

(A) the sub-network extracted for DNA binding genes in cattle; (B) the degree distribution for the network; (C) the short path length frequency for the network.


A
Sub-network on DNA binding
DNA binding genes
Linker genes
BCL6
CENPB
SWAP70
CERS5
CALCOCO1
AXIN1
HEF1
TGFBR2
EGR3
CCDC153
CERS4
CR2
CARHSP1
DHX4
DEDD
E2F5
F13A1
AKAP8
APEX2
CHTF8
HMGA1
PPP2R4
FAM96A
ALB
FAM214A
BCL6
CCDC98
GPBP1
FBXO18
RECQL
ATG4D
BRAP1
ATXN1
DDX1
E2F6
APTX
ARL2BP
ATF4
APITD1
EBP
EBF1
EBF3
CRISPLD2
C10orf10
AHR
ASXL1
BANF1
ANKRD13B
CARF
ABT1
AIFM2
CASP8AP2
FAM134C
CIC
CRYZ
CLYBL
BHLHE41
AIFM1
AKAP9
DTD1
GLYR1
GIT1
CSDC2
DHX9
FBXL19
SLC35A5
APP
ERGIC2
CAPRIN1
DDIT3
CHD3
BAZ2B
DNAJC2
POLR1B
ASXL2
CBFB
G3BP1
CNOT2
CASZ1
CXorf56
AGO4
BRCA1
STAT1
DNTTIP2
BLM
ALG6
DEK
EME1
CERS2
CHRAC1
FUS
DNAJC1
ERCC2
AKAP8L
AGFG1
BRIP1
FEN1
DDX41
CXXC1
EME2
CENPW
CEP55
CDT1
DEDD2
DFFB
SKG1
EEPD1
ERCC6L
GEN1
DNA2
AAAS
CENPT
B
Number of nodes
The number of connection (degree)
C
Frequency
The length of shortest path