PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2016, 43(1): 83~89
www.pibb.ac.cn

Research on Co-expression Genes of a Fig Wasp*

CUI Dong-Ya[1,2,3)**], SUN Xu-Bin[1,2)**], WANG Jia-Jia[1,2)], ZHANG Peng[2)], SUN Bao-Fa[4)],
CHEN Xiao-Wei[5)], Robert W. Murphy[6)], HE Shun-Min[2)***], HUANG Da-Wei[1,2)***]

([1)] *School of Life Science, Hebei University, Baoding* 071002, China;
[2)] *Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing* 100101, China;
[3)] *College of Life Science, Yuncheng University, Yuncheng* 044000, China;
[4)] *Laboratory of the Major Diseases Genome and Personalized Medicine, Institute of Genome, Chinese Academy of Sciences, Beijing* 100101, China;
[5)] *Laboratory of Bioinformatics and Noncoding RNA, Institute of Biophysics, Chinese Academy of Sciences, Beijing* 100101, China;
[6)] *Centre for Biodiversity and Conservation Biology, Royal Ontario Museum, Toronto* M5S 2C6, Canada)

Abbreviations: WGCNA, weighted gene co-expression network analysis; DAVID, Database for Annotation, Visualization and Integrated Discovery; FPKM, fragments per kilobase of exon model per million mapped reads; GE, gene expression; CV, coefficient of variation; Pcc, Pearson correlation coefficient; GO, gene ontology.

Abstract Female fig wasps differ phenotypically from conspecific males to the extent that often they cannot be associated with one another. Weighted gene co-expression network analysis (WGCNA) of the genome and transcriptomes of one such fig wasp, *Ceratosolensolmsi*, generated five expression modules, which were flagged as blue, turquoise, brown, green and yellow. These involved two female-biased expression modules and three pupa-biased expression modules, respectively. Gene ontologies indicated three functional enrichment gene sets in modules turquoise and yellow. Two functional enrichment gene sets that participate in cell cycle or have nucleotide binding activityclustered in turquoise module. The functionally enriched gene set in yellow module played roles in cell differentiation, especially in neuron morphogenesis.

Key words co-expression, network, functional enrichment
DOI: 10.16476/j.pibb.2015.0287

Genes that co-participate in biological processes generally display similar expression patterns. Often such gene sets involve functional enrichments, which clustering methods can reveal. Thus, clustering of gene sets facilitates an understanding of the roles of genes play in a biological processes[1].

Weighted gene co-expression network analysis (WGCNA) can cluster together genes that function in concert. One of the most direct methods for identifying gene co-expression networks [2], WGCNA has been successfully applied to the identification of functional enrichment modules in complex diseases[3-8].

Fig wasps are obligate symbiosis of their fig hosts. Female and male fig wasps exhibit extreme morphological dimorphism [9]. Recently, the genome and transcriptomes of *Ceratosolensolmsi* were published [9]. Herein, we employ WGCNA to identify groups of co-expressed functional genes from transcriptomes of male and female *C. solmsi*.

1 Materials and methods

1.1 Differentially expressed genes

Fig wasps has three developmental stages: larva, pupa and adult[10]. Transcriptome data were accessed in the NCBI Short Read Archive(www.ncbi.nlm.nih.gov/sra) under the accession No. SRP029703. The FPKM

*This work was supported by grants from The National Basic Research Program of China (2011CB504605) and The National Natural Science Foundation of China (31090253, 31210103912).
**These authors contributed equally to this work.
***Corresponding author.
HE Shun-Min. Tel: 86-10-64807279, E-mail: heshunmin@gmail.com
HUANG Da-Wei. Tel: 86-10-64807235, E-mail: huangdw@ioz.ac.cn
Received: September 6, 2015 Accepted: December 1, 2015

method (fragments per kb per million reads) [11] was used to calculate the expression of unigenes. Each gene expression (*GE*) value was log2-transformed from its FPKM value. To identify the differentially expressed genes, we used the distribution of the overall *GE* value to remove lowly expressed genes[12]. In total, 7 881 genes with *GE*>0.5 in at least four samples were obtained. Next, we used the coefficient of variation (*CV*) value to identify the differentially expressed genes[13]. Ultimately, we obtained 2 938 genes with *CV* <2.

1.2 Screening and clustering of highly correlated genes

WGCNA was used to describe the correlation patterns among 2 938 genes across the transcriptomic datasets and cluster the genes with highly correlated expression patterns. Firstly, the weights of each of the two genes were calculated using a soft threshold power of 12 because the *R*^2 was greater than 0.7 when applying this threshold[14]. Secondly, two genes with a weight of greater than 0.476 were considered to possess highly correlated expression patterns for 0.94^12 [15]. Thirdly, candidate highly correlated genes were clustered when minModuleSize = 5 and mergeCutHeight = 0.05. Finally, each module of gene expression pattern was visualized using Cluster 3.0 and Treeview[16–17].

1.3 Annotation and functional speculation

Fig wasp genes were not well annotated. Therefore, the well-annotated genes of *Drosophila melanogaster* were used to imply the functions of fig wasp genes. *Drosophila* protein sequences were downloaded from Flybase. Our co-expressed genes were aligned to those in Flybase by using BLASTP[18]. Functional annotations had the highest similar sequences of fig wasp genes to those in Flybase. The fly gene IDs, which were mapped by fig wasp genes, were put into Annotation, Visualization and Integrated Discovery (DAVID)[19] to deduce enrichment levels of the latter.

2 Results

2.1 Highly correlated genes grouped into five co-expression modules

Using a gene expression value (*GE*) of > 0.5 [12] and a coefficient of variation (*CV*) of < 2 [13], 2 938 differently expressed genes were obtained. The weight-values between each two genes were calculated using WGCNA. For better screening of the highly correlated co-expression genes, 0.476 was taken as a co-expression threshold value. This resulted in 336 genes being chosen. The Cytoscape v2.6.3 [20] network divided 336 genes into four parts (Figure 1). The features of the network were show in Figure S1～S4 and Table S1. Interestingly, the 336 genes clustered into five modules using WGCNA clustering function (Figure 2). Depicted in colors, these consisted of 54 (green), 62 (yellow), 63 (brown), 76 (blue) and 81 (turquoise) genes.

The turquoise and blue modules clustered together, as did green and brown. Yellow module did not cluster with any other module (Figure 1, 2). To confirm that co-expressed genes had similar levels of expression, we visualized them in heatmaps using Cluster 3.0 and Treeview [16–17]. Genes clustered in brown, green and yellow modules were highly expressed in females, and genes in blue and turquoise module were expressed in pupa (Figure S5).

2.2 Gene ontology

In total, 336 co-expressed genes clustered into three co-expression networks and five modules. These results suggested that each module of genes should have unique, enriched functions. Our annotation of fig wasp gene functions by using BLASTP against the genome of *Drosophila* discovered 251 genes(Table S2). These genes were then analyzed for the functional enrichments using DAVID [19]. Three annotation clusters with enrichment scores of > 3 were chosen and we terms with Benjamini values of < 0.05 in each annotation cluster to be candidate functional terms.

We used the fold change (c) to summarize the degree of functional enrichment by calculating formula (1) as follows:

$$c=\frac{a}{b}/\frac{m}{n} \quad (1)$$

a: Number of enriched genes of a given module in an annotation function term. b: Number of all enriched genes in an annotation function term. m: Number of all genes in a given module. n: Number of all candidate genes (336).

Gene sets with clear function enrichments had c values of greater than 2 and gene sets likely to have functional enrichments had c values between 1 and 2.

GO results showed that three annotation-clusters involved biological processes and molecular functions (Table S3). Genes that played roles in cell cycle functions exhibited the greatest amount of enrichment and these belonged to annotation cluster 1, and

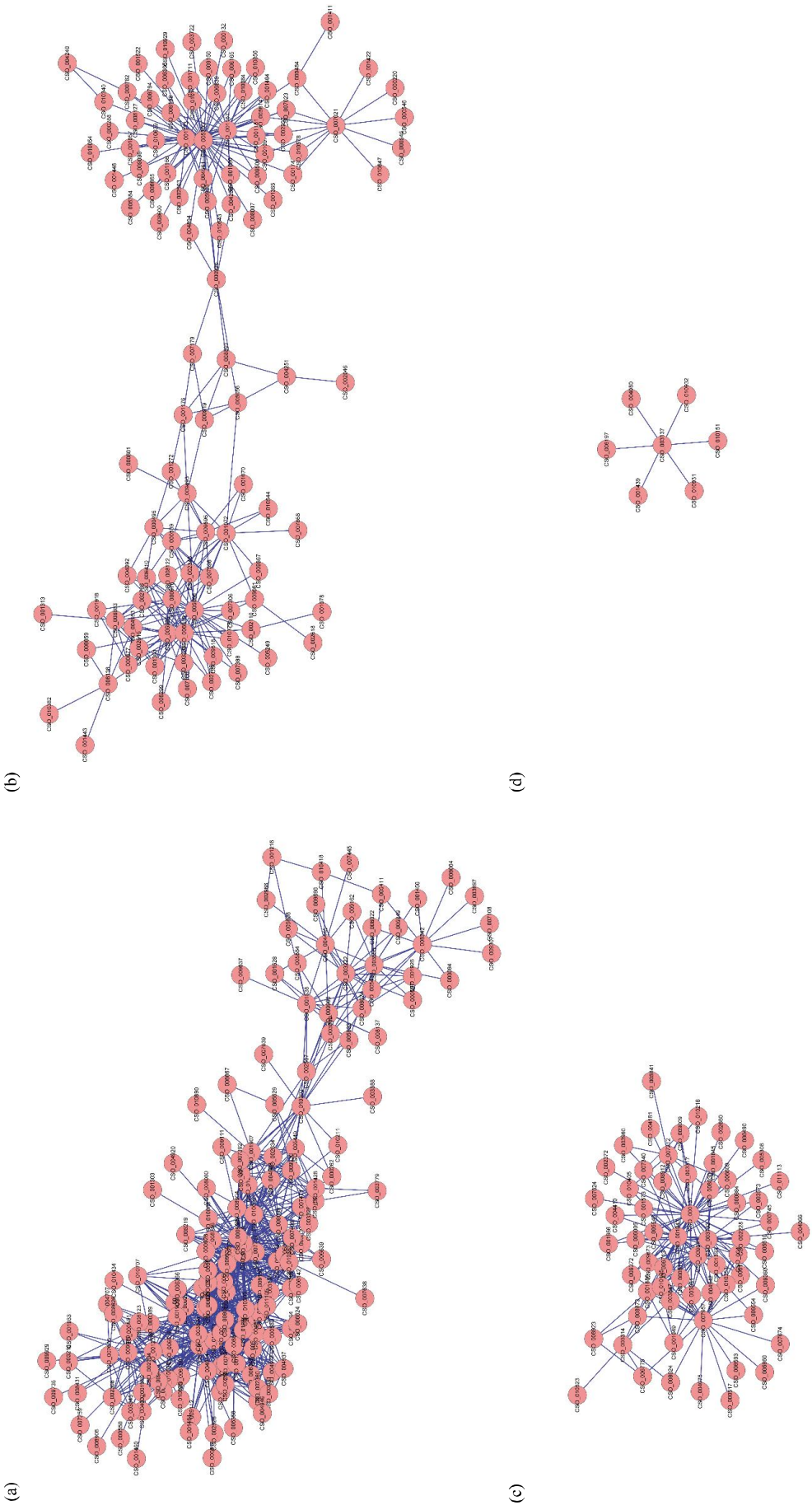

Fig. 1 Highly correlated genes' co-expression network

(a) Divided into blue and turquoise module (connected with Figure 2). (b) Divided into brown and yellow module. (c) Belong to yellow module. (d) Belong to brown module.

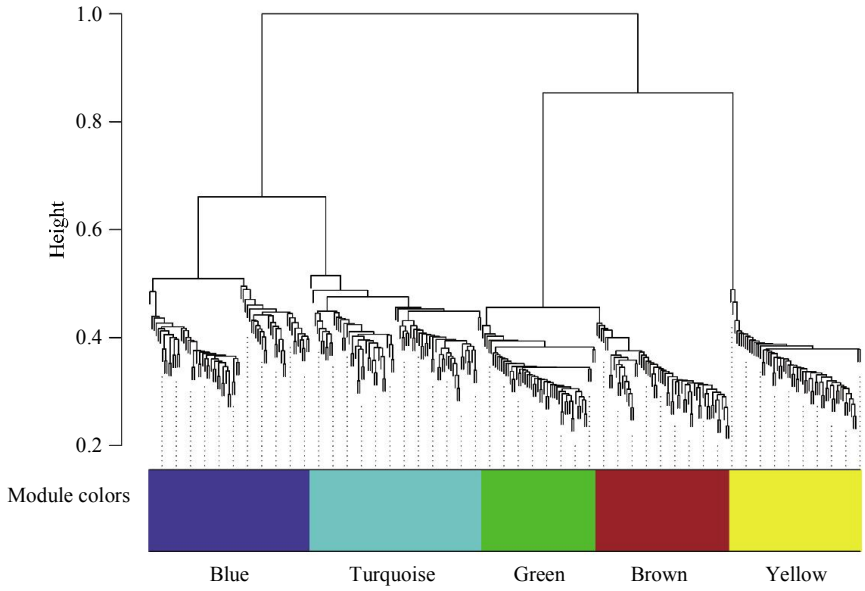

Fig. 2 Highly correlation genes clustered into five expression modules

Blue and turquoise module genes co-expressed together, green and brown module genes co-expressed together(connected with Figure 1).

modules turquoise and blue. Annotation cluster 2 contained genes enriched in nucleotide binding activity. They also exhibited the greatest levels of enrichment in modules turquoise and blue. Finally, annotation cluster 3 was comprised of genes enriched in morphogenesis. These genes tended to cluster into modules yellow and green (Figure 3).

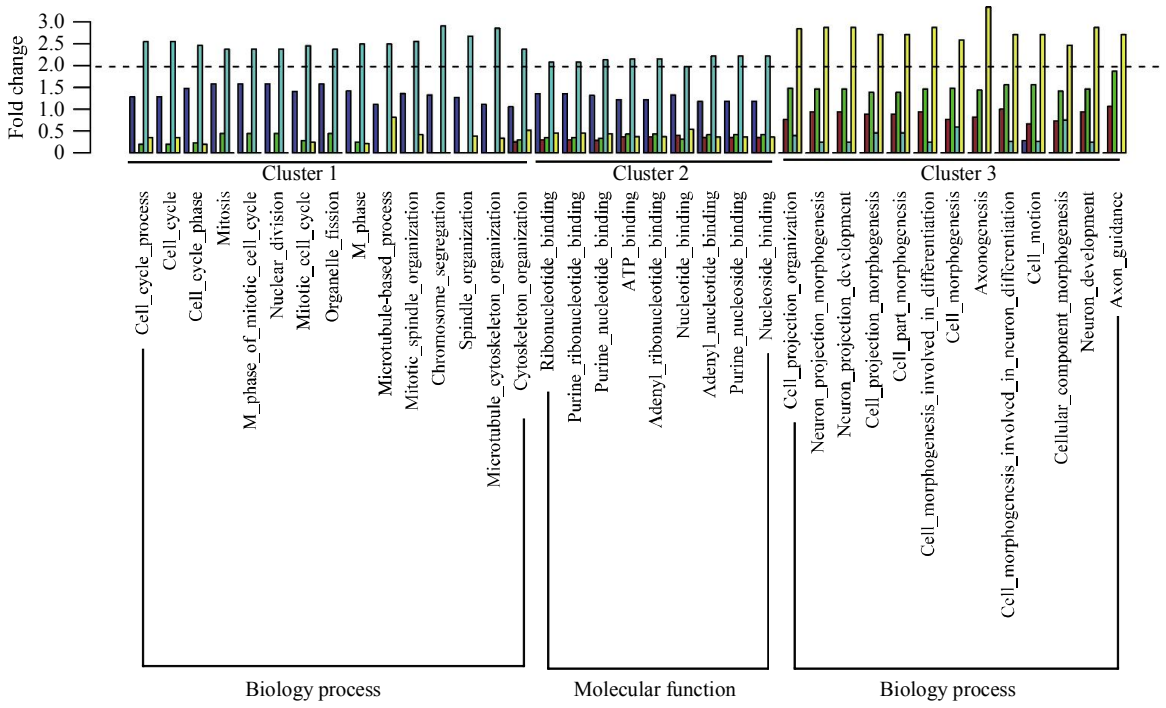

Fig. 3 The enrichment of genes clustered in turquoise and yellow module

Genes of annotation cluster 1 and annotation cluster 2 were enriched biology process and molecular function, respectively. However, genes of annotation cluster 3 were enriched biology process.

Genes in modules turquoise and blue were expressed preferentially in female fig wasps, but not in males (Figure S5). This discovery suggested that these genes might play important roles in the development of females. For example, CSO_000580, which was annotated as FBgn0000063, might have protein kinase activity[21] and play roles in female meiosis[22], female meiosis chromosome segregation [23] and meiotic spindle assembly checkpoint[22]. Thus, this gene might be involved in oogenesis. Another example, CSO_000814, which was annotated by FBgn0004859, might also play roles in oogenesis[24]. These results suggested that clustered highly expressed genes in adult female fig wasps special roles in adult female development.

Genes clustering in modules yellow and green were highly expressed in the pupa stage of the fig wasp. Female and male pupa exhibited the same levels of expression. Thus, these genes might play roles in pupa morphogenesis. For example, CSO_010218, which was annotated by FBgn0030662[25], might be a component of the golgi cisterna membrane and be involved in glucuronosyltransferase activity; it may also function in the chondroitin sulfate biosynthetic process[26]. Ecdysteroid UDP-glucosyltransferases were shown to play a role in the conjugation of ecdysteroids with glucose or galactose[27]. Similarly, CSO_004410 (annotated by FBgn0086680) was found to be a RNA polymerase Ⅱ core promoter in the proximal region of sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription[28]. Further, CSO_004410 might play roles in motor neuron axon guidance [29], regulation of dendrite morphogenesis[30] and brain development[31].

3 Discussion

3.1 Highly co-expressed genes clustered into five expression-modules

The fig wasp is a good organism for studying the co-expression of genes in males and females that exhibit extreme morphological dimorphism. We obtained 2 938 differentially expressed genes from 11 506 annotated genes in RNA-seq data. *GE* values were used to remove lowly expressed genes[12] and *CV* values were employed to identify differentially expressed genes [13]. To ascertain the most relevant genes, WGCNA was used to cluster highly co-expressed genes, those with a weight greater than 0.476 [15], into five expression-modules. Genes in modules turquoise and blue were co-expressed, as were genes in module brown and green. Genes in modules turquoise and blue were highly expressed in female fig wasp only. Further, genes in module green, brown and yellow were expressed mainly in pupae, suggesting involvement in the development of fig wasps. Genes in module yellow did not cluster with genes in modules brown or green, suggesting unique functions of genes in each module. Five differentially expressed gene modules were obtained, and genes function of these modules might be different to each other. The finding was useful for the exploration of fig wasp gene functions.

3.2 Modules with a low-fold change score are revealing

Highly correlated genes can facilitate explorations into the roles genes play in divergent phenotypes. Different modules appear to have different functional enrichments. Modules with a low-fold change score might have unknown functional enrichments. For example, genes in modules blue and turquoise belong to the same co-expression network, but they have different functional enrichment scores and the same situation occurs in modules brown and green. These observations suggest that most genes functions await detailed annotation.

Acknowledgements We are grateful to Dr. GU Hai-Feng for help in constructing the network.

Supplementary material Figures S1 ～S5, Tables S1～S3 are available at paper online(http://www.pibb.ac.cn).

References

[1] Ficklin S P, Feltus F A. Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. Plant Physiol, 2011, **156**(3): 1244−1256

[2] Allen J D, Xie Y, Chen M, *et al*. Comparing statistical methods for constructing large scale gene networks. PLoS One, 2012, **7** (1): e29348

[3] Fuller T F, Ghazalpour A, Aten J E, *et al*. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm Genome, 2007, **18**(6−7): 463−472

[4] Miller J A, Oldham M C, Geschwind D H. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. J Neurosci, 2008, **28**(6): 1410−1420

[5] Plaisier C L, Horvath S, Huertas-Vazquez A, *et al*. A systems genetics approach implicates USF1, FADS3, and other causal

candidate genes for familial combined hyperlipidemia. PLoS Genet, 2009, **5**(9): e1000642

[6] Kadarmideen H N, Watson-Haigh N S, Andronicos N M. Systems biology of ovine intestinal parasite resistance: disease gene modules and biomarkers. Mol Biosyst, 2011, **7**(1): 235-246

[7] Rosen E Y, Wexler E M, Versano R, *et al*. Functional genomic analyses identify pathways dysregulated by progranulin deficiency, implicating Wnt signaling. Neuron, 2011, **71**(6): 1030-1042

[8] Winden K D, Karsten S L, Bragin A, *et al*. A systems level, functional genomics analysis of chronic epilepsy. PLoS One, 2011, **6**(6): e20763

[9] Xiao J H, Yue Z, Jia L Y, *et al*. Obligate mutualism within a host drives the extreme specialization of a fig wasp genome. Genome Biol, 2013, **14**(12): R141

[10] Jia L Y, Xiao J H, Niu L M, *et al*. Delimitation and description of the immature stages of a pollinating fig wasp, Ceratosolen solmsi marchali Mayr (Hymenoptera: Agaonidae). Bull Entomol Res, 2014, **104**(2): 164-175

[11] Mortazavi A, Williams B A, Mccue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 2008, **5**(7): 621-628

[12] Filteau M, Pavey S A, St-Cyr J, *et al*. Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish. Mol Biol Evol, 2013, **30**(6): 1384-1396

[13] Mao L, Van Hemert J L, Dash S, *et al*. Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics, 2009, **10**: 346

[14] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. Bmc Bioinformatics, 2008, **9**: 559

[15] Zheng Z L, Zhao Y. Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to 'Candidatus Liberibacter asiaticus' infection. BMC Genomics, 2013, **14**: 27

[16] De Hoon M J, Imoto S, Nolan J, *et al*. Open source clustering software. Bioinformatics, 2004, **20**(9): 1453-1454

[17] Eisen M B, Spellman P T, Brown P O, *et al*. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 1998, **95**(25): 14863-14868

[18] Johnson M, Zaretskaya I, Raytselis Y, *et al*. NCBI BLAST: a better web interface. Nucleic Acids Res, 2008, **36** (Web Server issue): W5-9

[19] Huang Da W, Sherman B T, Lempicki R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009, **4**(1): 44-57

[20] Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 2003, **13**(11): 2498-2504

[21] Morrison D K, Murakami M S, Cleghon V. Protein kinases and phosphatases in the *Drosophila* genome. J Cell Biol, 2000, **150**(2): F57-62

[22] Gilliland W D, Hughes S E, Cotitta J L, *et al*. The multiple roles of mps1 in *Drosophila* female meiosis. PLoS Genet, 2007, **3**(7): e113

[23] Gilliland W D, Wayson S M, Hawley R S. The meiotic defects of mutants in the *Drosophila* mps1 gene reveal a critical role of Mps1 in the segregation of achiasmate homologs. Current Biology, 2005, **15**(7): 672-677

[24] Forbes A J, Spradling A C, Ingham P W, *et al*. The role of segment polarity genes during early oogenesis in *Drosophila*. Development, 1996, **122**(10): 3283-3294

[25] Ju T, Brewer K, D'souza A, *et al*. Cloning and expression of human core 1 beta1, 3-galactosyltransferase. J Biol Chem, 2002, **277** (1): 178-186

[26] Wilson I B. Glycosylation of proteins in plants and invertebrates. Curr Opin Struct Biol, 2002, **12**(5): 569-577

[27] Oreilly D R. Baculovirus-encoded ecdysteroid Udp-glucosyltransferases. Insect Biochemistry and Molecular Biology, 1995, **25**(5): 541-550

[28] Junell A, Uvell H, Pick L, *et al*. Isolation of regulators of *Drosophila* immune defense genes by a double interaction screen in yeast. Insect Biochem Mol Biol, 2007, **37**(3): 202-212

[29] Certel S J, Thor S. Specification of *Drosophila* motoneuron identity by the combinatorial action of POU and LIM-HD factors. Development, 2004, **131**(21): 5429-5439

[30] Hasegawa E, Kitada Y, Kaido M, *et al*. Concentric zones, cell migration and neuronal circuits in the *Drosophila* visual center. Development, 2011, **138**(5): 983-993

[31] Meier S, Sprecher S G, Reichert H, *et al*. ventral veins lacking is required for specification of the tritocerebrum in embryonic brain development of *Drosophila*. Mech Dev, 2006, **123**(1): 76-83

榕小蜂基因共表达研究 *

崔东亚[1, 2, 3)**] 孙旭斌[1, 2)**] 王佳佳[1, 2)] 张 鹏[2)] 孙宝发[4)]
陈小伟[5)] Robert W. Murphy[6)] 何顺民[2)***] 黄大卫[1, 2)***]

([1)]河北大学生命科学学院，保定 071002；[2)]中国科学院动物研究所系统与进化重点实验室，北京 100101；
[3)]运城学院生命科学系，运城 044000；[4)]中国科学院基因组研究所，北京 100101；[5)]中国科学院生物物理研究所，北京 100101；
[6)] *Centre for Biodiversity and Conservation Biology, Royal Ontario Museum, Toronto* M5S 2C6, Canada)

摘要 榕小蜂的雌雄个体之间存在很大表型差异，以至于很难将雌雄个体彼此联系在一起. 以对叶榕传粉榕小蜂作为材料，利用“加权基因共表达网络分析”软件(WGCNA)，对榕小蜂的基因组和转录组进行分析，结果发现，5 个基因共表达模块，分别用蓝色、蓝绿色、棕色、绿色和黄色标识，其中 2 个模块偏爱在雌蜂中表达，3 个模块偏爱在蛹中表达. 基因本体(GO)分析发现在蓝绿色和黄色表达模块中发现 3 个功能富集的基因集合. 在蓝绿色基因表达模块中发现 2 个基因集合，分别与细胞周期和核苷酸结合活性有关；在黄色基因表达模块中发现 1 个基因结合，与细胞分化有关，尤其是与神经发育有关.

关键词 共表达，网络，功能富集
学科分类号 Q7 **DOI**: 10.16476/j.pibb.2015.0287

* 国家重点基础研究发展计划(2011CB504605)和国家自然科学基金(31090253, 31210103912)资助项目.
** 共同第一作者.
*** 通讯联系人.
何顺民. Tel: 010-64807279, E-mail: heshunmin@gmail.com
黄大卫. Tel: 010-64807235, E-mail: huangdw@ioz.ac.cn
收稿日期：2015-09-06，接受日期：2015-12-01

附　　录

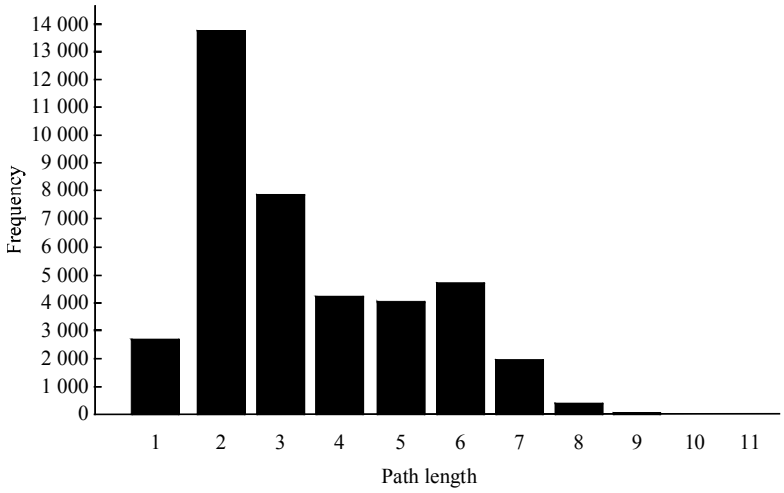

Fig. S1　Frequency distribution of characteristic path length

This figure was output directly by NetworkAnalyzer in Cytoscape v2.6.3. The *x*-axis represents the shortest path length, which represents the expected distance between two connected nodes, and the *y*-axis represents frequency.

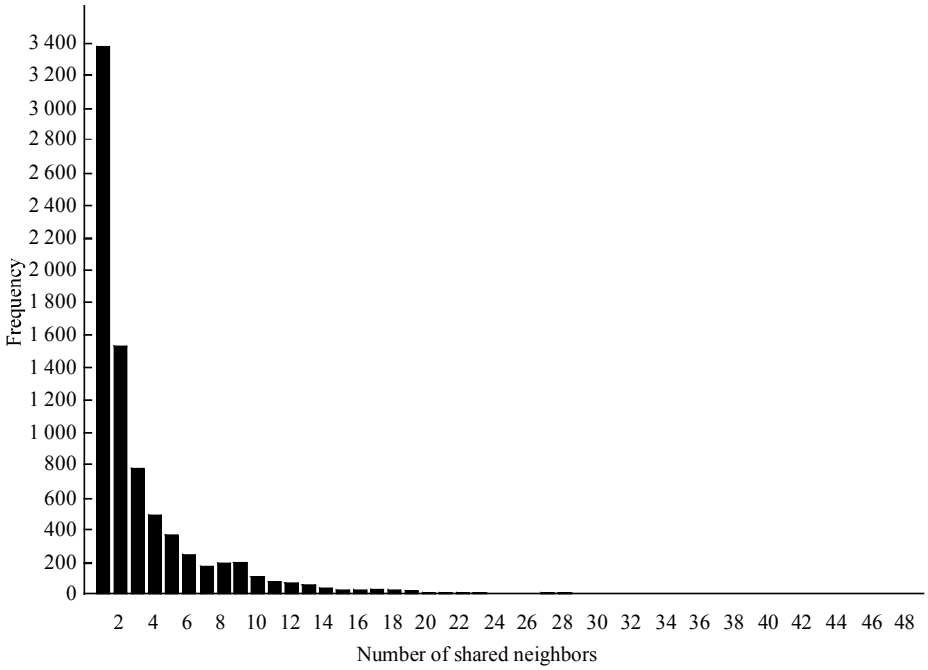

Fig. S2　Frequency distribution of shared neighbors

This figure was output directly by NetworkAnalyzer in Cytoscape v2.6.3. The shared neighbors represent the connections between neighbors. The *x*-axis represents the number of shared neighbors, and the *y*-axis represents frequency.

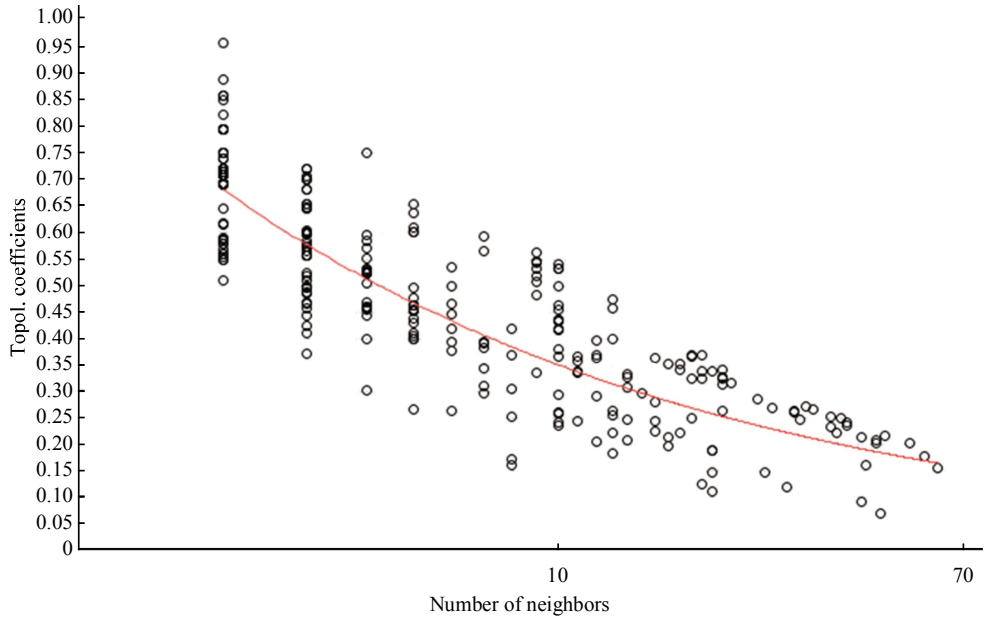

Fig. S3 Topological coefficient of neighbors

This figure was output directly by NetworkAnalyzer in Cytoscape v2.6.3. The topological coefficient is a relative measure of the extent to which a node shares neighbors with other nodes. The *x*-axis represents the number of neighbors, and the *y*-axis represents the topological coefficient.

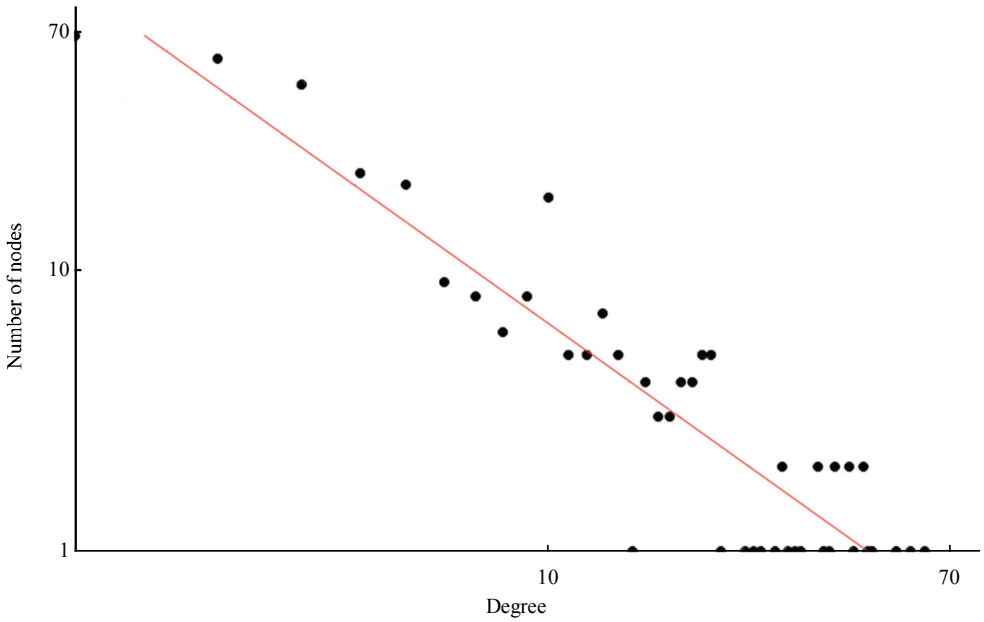

Fig. S4 Node degree distribution

This figure was output directly by NetworkAnalyzer in Cytoscape v2.6.3. The number of nodes (genes) plotted as a property of their degree (number of connections with other nodes) shows a power-law like distribution, which indicates a scale-free network topology.

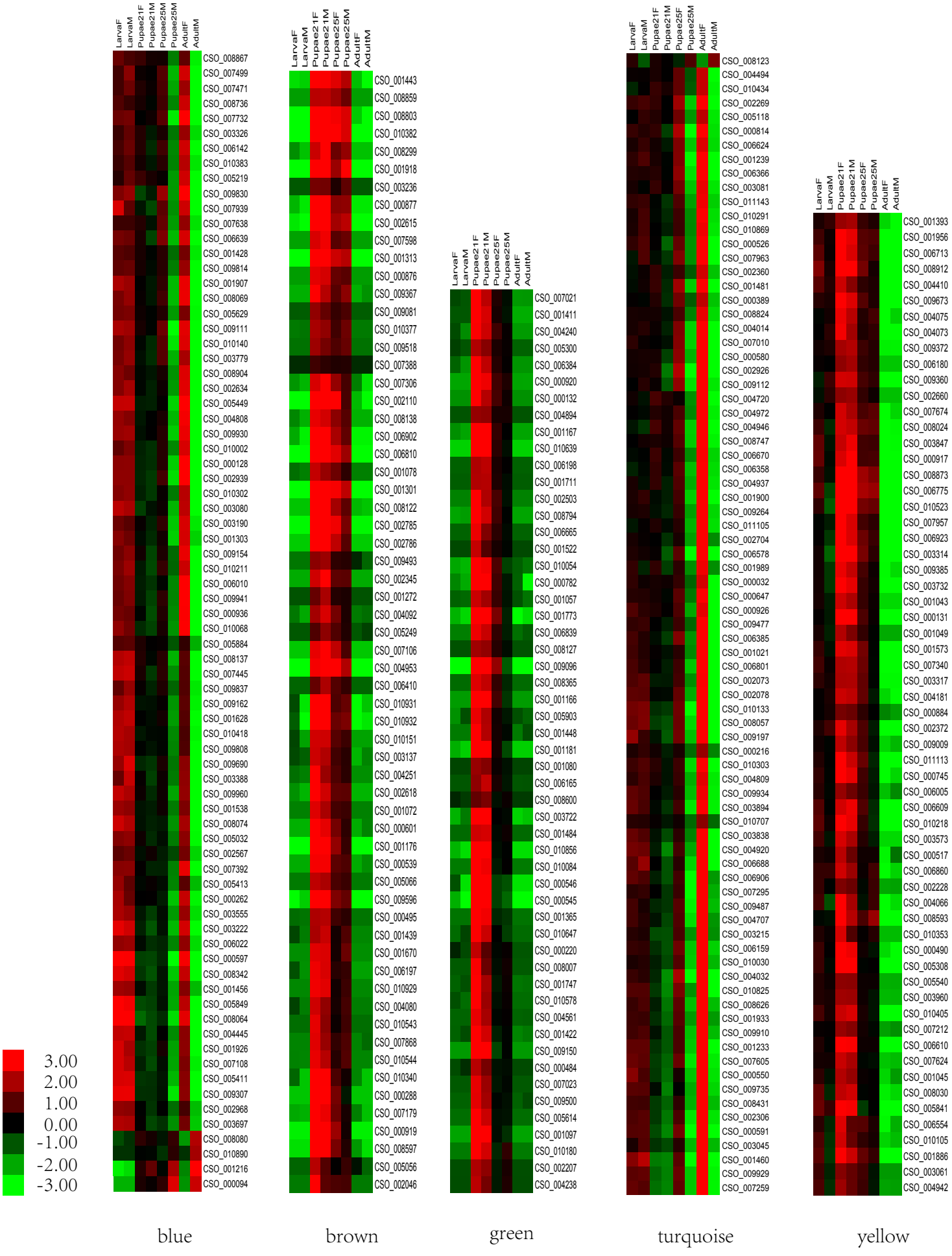

Fig. S5 Heatmap for the expression of genes in five modules

This figure was created using Cluster 3.0 and Treeview. The bar in the lower left shows that the redder colors indicate higher expression, and the more brilliant green colors represent lower expression. The x-axis represents four different developmental stages in the fig wasp (larva, 21-day pupa, 25-day pupa and adults), while the y-axis represents the different genes.

Table S1 Simple parameters of the network

Serial number	Property	Value
1	Clustering coefficient	0.408[1)]
2	Connected components	4
3	Network centralization	0.162[1)]
4	Network heterogeneity	1.310[1)]
5	Characteristic path length	3.462[2)]
6	Network radius	1[2)]
7	Network diameter	10[2)]
8	Avg. number of neighbors	8.214[3)]
9	Number of nodes	336
10	Network density	0.025
11	Shortest path	40306[2)]

Column 1 lists the number of simple parameters. Column 2 lists the names of the basic properties. Column 3 lists the corresponding value of each property in the gene co-expression network. [1)] The parameter associated with scale-free distribution. [2)] The parameter associated with shortest path length. [3)] The parameter associated with the number of neighboring nodes.