


Article

Functional Annotation of *Caenorhabditis elegans* Genes by Analysis of Gene Co-Expression Networks

Wei Liu ^{1,*} , Ling Li ¹, Yiruo He ², Sen Cai ¹, Wenjie Zhao ¹, Hao Zheng ¹, Yuexian Zhong ¹, Shaobo Wang ¹, Yang Zou ¹, Zhenhua Xu ¹, Yu Zhang ¹ and Wei Tu ³

¹ School of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China; 1150561002@fafu.edu.cn (L.L.); cs1159821786@gmail.com (S.C.); zhousifang@mail.com (W.Z.); zhenghao@m.fafu.edu.cn (H.Z.); yuexianz@163.com (Y.Z.); 1160561008@fafu.edu.cn (S.W.); 1170539007@fafu.edu.cn (Y.Z.); 1170561006@fafu.edu.cn (Z.X.); 1170561001@fafu.edu.cn (Y.Z.)

² School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China; andrewhe3@yahoo.com.cn

³ Department of Molecular and Cellular Medicine, Texas A&M Health Science Center, College Station, TX 77843-1114, USA; tuwei1112003@163.com

* Correspondence: weilau@fafu.edu.cn

Received: 3 July 2018; Accepted: 1 August 2018; Published: 3 August 2018



Abstract: *Caenorhabditis elegans* (*C. elegans*) is a well-characterized metazoan, whose transcriptome has been profiled in different tissues, development stages, or other conditions. Large-scale transcriptomes can be reused for gene function annotation through systematic analysis of gene co-expression relationships. We collected 2101 microarray data from National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO), and identified 48 modules of co-expressed genes that correspond to tissues, development stages, and other experimental conditions. These modules provide an overview of the transcriptional organizations that may work under different conditions. By analyzing higher-order module networks, we found that nucleus and plasma membrane modules are more connected than other intracellular modules. Module-based gene function annotation may help to extend the candidate cuticle gene list. A comparison with other published data validates the credibility of our result. Our findings provide a new source for future gene discovery in *C. elegans*.

Keywords: *Caenorhabditis elegans*; transcriptome; gene co-expression network; cuticle; hub gene

1. Introduction

High-throughput transcriptomics technology has been extensively applied to investigate the mechanisms of gene regulation. A promising strategy to find the gene functions of unknown genes is the gene co-expression method, which infers gene functions by similar gene expression patterns. This method has been used to explore the global, temporal, and spatial expression of *Caenorhabditis elegans* and its gene functions [1–5]. Early papers investigating these factors tried to elucidate the transcriptome in *C. elegans* with a relatively small sample size; for example, the 553 samples reported by Kim is the largest sample size to date [1]. Although recent studies have used state-of-the-art tiling arrays or RNA-Seq technologies, they focused on single genes or experimental conditions [6,7]. There are still genes that have been annotated with unknown function. For these genes, little is known about their biological function. To discover the tissue-, temporal-, or experimental condition-specific gene expression, large sample sizes are needed.

Microarray is a mature high-throughput method for genome-wide gene expression profiling. Thousands of microarray data have been deposited in public databases, however, most individual research uses differential expression analysis methods to find significant changes in expression,

ignoring the inherent gene-gene expression correlation. Gene co-expression networks facilitate constructing a global view between genes [8]. Weighted gene co-expression network analysis (WGCNA) groups genes that have similar expression patterns across biological samples. In a gene co-expression network, a module is a subset of genes, whose expression patterns are similar to each other while different from genes in other modules. Usually, these genes are members from the same pathway or biological process. The whole transcriptome can be simplified into several modules, which allows us to look into biosystem components independently. Modules are more stable than individual genes in that the overall function of a module can remain the same while individual gene expression can be changed or replaced by other genes with similar redundant functions [9]. Furthermore, in a module, the importance of a gene can also be delineated by intramodule connectivity, which measures how correlated a gene is with all other module genes [10].

In this research, we applied WGCNA to publicly available *C. elegans* microarray data from different experimental conditions. Genome-scale modules of co-expressed genes with clear functional annotations were identified. Module-based qualitative analysis revealed that modules were associated with diverse biological functions. Module-based gene expression variation analysis suggested potential basal or conditional modules. Five modules that may correlate with molting were identified, and candidate cuticle genes were indicated. Those modules were also compared with previous publications to confirm the validity of our results.

2. Materials and Methods

2.1. Data Acquisition

Microarray datasets were obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database under the platform number GPL200. For simplicity, only Affymetrix *C. elegans* Genome Array data were included. Briefly, 151 datasets with 2101 *C. elegans* samples were downloaded. Detailed information for these datasets is provided in Table S1.

2.2. Weighted Gene Co-Expression Network Analysis

Microarray datasets were obtained from NCBI GEO database under the platform number GPL200. For simplicity, only Affymetrix *C. elegans* Genome Array Raw cel data were processed in Affymetrix Expression Console software (v1.4.1.46; Affymetrix, Inc., Santa Clara, CA, USA) using the MAS5 algorithm. Microarray data analysis was performed using R software (v3.1.2; R Foundation for Statistical Computing, Vienna, Austria) and Bioconductor WGCNA package (v1.51; R Foundation for Statistical Computing, Vienna, Austria; available from https://cran.r-project.org/src/contrib/WGCNA_1.51.tar.gz) [10]. Briefly, signed co-expression networks were constructed on the basis of 14,068 genes mapped from probe sets using the Brainarray Entrez Gene mapping file [11]. For each gene in the gene expression matrix, a pairwise Pearson correlation coefficient is computed, and an adjacency matrix is calculated by raising the correlation matrix to a power [12]. The power of 14 was chosen using the scale-free topology criterion. The weighted network was transformed into a network of topological overlap (TO)—an advanced co-expression measurement that measures not only the correlation of two genes, but also the extent of their shared correlations across the weighted network [12]. Genes were hierarchically clustered on the basis of their TO. Finally, co-expression gene modules were identified by the Dynamic Tree Cut algorithm [13]. Each module was summarized using singular value decomposition so that each module eigengene (ME) represented the first principal component of the module expression profiles [12]. Thus, ME explains the maximum amount of variation of the module expression levels, and is considered the most representative gene expression in a module. To construct the network of modules and identify meta-modules, the same process was applied to the above result. The parameters are power = 3, and minModuleSize = 2. The clustering used the hclust function in the WGCNA package.

Connectivity for genes in each module was calculated by the `softConnectivity` function. Connectivity is a measurement of the sum of the gene expression correlation with all other genes. Genes with high connectivity in a specific module tend to be hub genes, which may play vital roles in module function. Module stability was tested by the average correlation between the original connectivity and the connectivity from half of the samples that were randomly sampled 1000 times. The process was run for every module.

2.3. Functional Annotation of the Modules

Gene ontology (GO) enrichment for network modules were performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) [14] with the background list of genes on the *C. elegans* genome array. In DAVID, an over-representation of a term is defined as a modified Fisher's exact *p*-value with an adjustment for multiple tests using the Benjamini method. Other enrichments were performed by inputting a gene set into WormBase to find annotated terms that are over-represented using tissue enrichment analysis (TEA) and phenotype enrichment analysis (PEA) [15]. Modular genes enriched within chromosome regions were analyzed by the Positional Gene Enrichment analysis tool [16]. The stage specific module information was found by searching the NCBI PubMed database (www.ncbi.nlm.nih.gov/pubmed/) [17].

For gene expression variation analysis, gene expression relative standard deviation for each gene in a module were calculated and the average values for each module were provided.

2.4. Comparison of Gene Prediction Using Published Data

There are several early papers that tried to elucidate the transcriptome in *C. elegans* [1,3,6,18,19]. The module list and module gene list were compared. As for the WormNet prediction tool, which integrate heterogeneous genomics data into a single gene network for gene function prediction [18], we submitted our candidate gene into the tool, and observed if the predicted functions were similar to our module annotation.

3. Results and Discussion

3.1. A Gene Coexpression Network of *C. elegans* Was Successfully Constructed

A total of 48 co-expressed gene modules were identified (Table 1). A representative network visualization was shown in Figure S1. For simplicity, only the top significant term was recorded. Functional annotation shows that these modules were associated with immune response, RNA processing, proteolysis, translation, signaling, embryo development, ion transport, reproduction, and many other biological processes. The module stability was tested by the correlation between the original connectivity and those calculated by 1000 half-sampled connectivity values for each module [20]. The correlations of connectivity were averaged for each module. All the modules have an average connectivity correlation larger than 0.8 (Figure 1). Among them, skyblue has the lowest module stability, while white has the highest module stability. These results indicate that the relationships between module genes were robust to the exclusion of 50% of the data.

Table 1. Gene ontology (GO) and chromosome annotation of the identified 48 gene co-expression modules in *C. elegans*.

Module (No. of Genes)	Biological Process	Cellular Component	Molecule Function	Chromosome
Antiquewhite4 (35)	Ion transport (6×10^{-7})	Acetylcholine-gated channel complex (9×10^{-5})	Ion channel activity (2×10^{-4})	
Bisque4 (212)	Proteolysis (2×10^{-9})	Membrane raft (8×10^{-10})	Serine-type carboxypeptidase activity (3×10^{-12})	X (2×10^{-5})
Black (1173)	Ion transport (2×10^{-16})	Plasma membrane (5×10^{-14})	Signal transducer activity (1×10^{-12})	X (2×10^{-13})
Blue (930)	Regulation of cell shape (3×10^{-25})	Extrinsic component of cytoplasmic side of plasma membrane (8×10^{-14})	Protein kinase activity (1×10^{-22})	IV (1×10^{-9})
Brown (933)	Embryo development ending in birth or egg hatching (8×10^{-41})	Nucleus (3×10^{-11})	Protein binding (3×10^{-8})	I (1×10^{-15})
Brown4 (137)	Embryo development ending in birth or egg hatching (8×10^{-14})	Cytoplasm (2×10^{-15})	Protein binding (7×10^{-5})	III (4×10^{-3})
Coral1 (153)	Embryo development ending in birth or egg hatching (5×10^{-11})	Nucleus (1×10^{-10})	Protein binding (5×10^{-2})	I (3×10^{-2})
Coral2 (34)	Body morphogenesis (2×10^{-2})	Collagen trimer (2×10^{-25})	Structural constituent of cuticle (8×10^{-26})	
Cyan (219)	Axon guidance (2×10^{-7})	Axon (7×10^{-3})	Protein binding (3×10^{-6})	X (2×10^{-20})
Darkgreen (160)	Reproduction (8×10^{-5})	Cytoplasm (2×10^{-2})	Nucleotide binding (2×10^{-4})	III (1×10^{-6})
Darkgrey (136)	Neuropeptide signaling pathway (1×10^{-9})	Heterotrimeric G-protein complex (5×10^{-5})	Calcium ion binding (4×10^{-4})	X (2×10^{-2})
Darkmagenta (109)		Extracellular space (2×10^{-4})	Iron ion binding (3×10^{-2})	
Darkolivegreen (163)	Embryo development ending in birth or egg hatching (8×10^{-3})	Nucleus (1×10^{-2})	Protein binding (6×10^{-4})	III (8×10^{-3})
Darkorange (367)	Embryo development ending in birth or egg hatching (2×10^{-9})	Mitochondrion (6×10^{-4})		III (2×10^{-9})
Darkorange2 (221)	Innate immune response (4×10^{-4})			V (3×10^{-4})
Darkred (167)	Embryo development ending in birth or egg hatching (8×10^{-11})	P granule (8×10^{-4})		I (2×10^{-3})
Darkseagreen4 (41)	Nematode larval development (1×10^{-7})	Mitochondrion (6×10^{-16})	NADH dehydrogenase (ubiquinone) activity (4×10^{-7})	
Darkslateblue (127)	Endoplasmic reticulum unfolded protein response (1×10^{-4})	Collagen trimer (8×10^{-34})	Structural constituent of cuticle (9×10^{-34})	
Floralwhite (86)		Pseudopodium (3×10^{-3})		IV (1×10^{-4})
Green (602)	Embryo development ending in birth or egg hatching (3×10^{-5})	Nucleolus (6×10^{-8})	RNA binding (2×10^{-6})	I (4×10^{-2})
Greenyellow (389)	Nonmotile primary cilium assembly (1×10^{-29})	Ciliary basal body (2×10^{-19})	G-protein coupled receptor activity (5×10^{-11})	X (7×10^{-4})
Grey60 (522)	Embryo development ending in birth or egg hatching (8×10^{-6})	Nucleus (4×10^{-14})	Protein binding (9×10^{-3})	X (2×10^{-18})

Table 1. Cont.

Module (No. of Genes)	Biological Process	Cellular Component	Molecule Function	Chromosome
Antiquewhite4 (35)	Ion transport (6×10^{-7})	Acetylcholine-gated channel complex (9×10^{-5})	Ion channel activity (2×10^{-4})	
Bisque4 (212)	Proteolysis (2×10^{-9})	Membrane raft (8×10^{-10})	Serine-type carboxypeptidase activity (3×10^{-12})	X (2×10^{-5})
Black (1173)	Ion transport (2×10^{-16})	Plasma membrane (5×10^{-14})	Signal transducer activity (1×10^{-12})	X (2×10^{-13})
Blue (930)	Regulation of cell shape (3×10^{-25})	Extrinsic component of cytoplasmic side of plasma membrane (8×10^{-14})	Protein kinase activity (1×10^{-22})	IV (1×10^{-9})
Brown (933)	Embryo development ending in birth or egg hatching (8×10^{-41})	Nucleus (3×10^{-11})	Protein binding (3×10^{-8})	I (1×10^{-15})
Brown4 (137)	Embryo development ending in birth or egg hatching (8×10^{-14})	Cytoplasm (2×10^{-15})	Protein binding (7×10^{-5})	III (4×10^{-3})
Coral1 (153)	Embryo development ending in birth or egg hatching (5×10^{-11})	Nucleus (1×10^{-10})	Protein binding (5×10^{-2})	I (3×10^{-2})
Coral2 (34)	Body morphogenesis (2×10^{-2})	Collagen trimer (2×10^{-25})	Structural constituent of cuticle (8×10^{-26})	
Cyan (219)	Axon guidance (2×10^{-7})	Axon (7×10^{-3})	Protein binding (3×10^{-6})	X (2×10^{-20})
Darkgreen (160)	Reproduction (8×10^{-5})	Cytoplasm (2×10^{-2})	Nucleotide binding (2×10^{-4})	III (1×10^{-6})
Darkgrey (136)	Neuropeptide signaling pathway (1×10^{-9})	Heterotrimeric G-protein complex (5×10^{-5})	Calcium ion binding (4×10^{-4})	X (2×10^{-2})
Darkmagenta (109)		Extracellular space (2×10^{-4})	Iron ion binding (3×10^{-2})	
Darkolivegreen (163)	Embryo development ending in birth or egg hatching (8×10^{-3})	Nucleus (1×10^{-2})	Protein binding (6×10^{-4})	III (8×10^{-3})
Darkorange (367)	Embryo development ending in birth or egg hatching (2×10^{-9})	Mitochondrion (6×10^{-4})		III (2×10^{-9})
Darkorange2 (221)	Innate immune response (4×10^{-4})			V (3×10^{-4})
Darkred (167)	Embryo development ending in birth or egg hatching (8×10^{-11})	P granule (8×10^{-4})		I (2×10^{-3})
Darkseagreen4 (41)	Nematode larval development (1×10^{-7})	Mitochondrion (6×10^{-16})	NADH dehydrogenase (ubiquinone) activity (4×10^{-7})	
Darkslateblue (127)	Endoplasmic reticulum unfolded protein response (1×10^{-4})	Collagen trimer (8×10^{-34})	Structural constituent of cuticle (9×10^{-34})	
Floralwhite (86)		Pseudopodium (3×10^{-3})		IV (1×10^{-4})
Green (602)	Embryo development ending in birth or egg hatching (3×10^{-5})	Nucleolus (6×10^{-8})	RNA binding (2×10^{-6})	I (4×10^{-2})
Greenyellow (389)	Nonmotile primary cilium assembly (1×10^{-29})	Ciliary basal body (2×10^{-19})	G-protein coupled receptor activity (5×10^{-11})	X (7×10^{-4})
Grey60 (522)	Embryo development ending in birth or egg hatching (8×10^{-6})	Nucleus (4×10^{-14})	Protein binding (9×10^{-3})	X (2×10^{-18})

Note: Benjamini-adjusted Fisher's exact test p values are given in brackets. NADH, Nicotinamide adenine dinucleotide; UTR, untranslated region; LSU, Large subunit; ATP, Adenosine triphosphate.

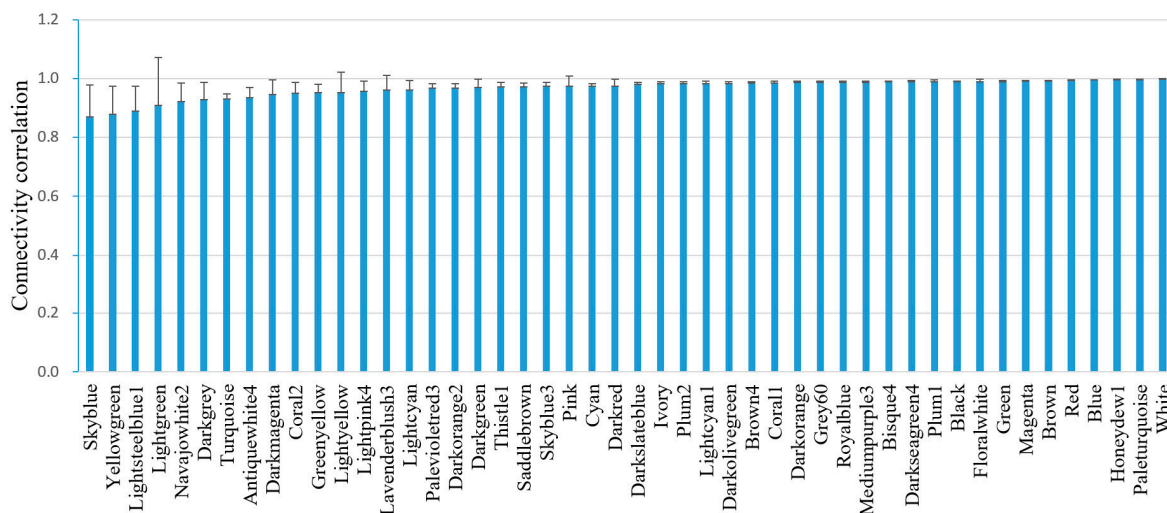


Figure 1. Correlation of intramodule connectivity for each module after 1000 instances of sampling (mean \pm standard deviation (SD)). The connectivity values were calculated by sampling 1050 samples 1000 times randomly.

There are several modules with same biological process. Ten modules are correlated with embryo development, three modules are correlated with ion transport, two modules are correlated with proteolysis, two modules are correlated with immune response, and two modules are correlated with reproduction. However, these were genes located on different chromosomes or expressed by different tissues. Further, tissue specific gene enrichment analysis revealed that these modules were associated with different tissues (Table 2).

Table 2. Tissue enrichment for the 22 gene co-expression modules with the same GO BP annotation in *C. elegans*.

Module (No. Genes)	Biological Process	Tissue
Grey60 (522)	Embryo development ending in birth or egg hatching (8×10^{-6})	Caap
Darkred (167)	Embryo development ending in birth or egg hatching (8×10^{-11})	Psub1
Brown4 (137)	Embryo development ending in birth or egg hatching (8×10^{-14})	AWB
Darkorange (367)	Embryo development ending in birth or egg hatching (2×10^{-9})	Reproductive system, thermosensory neuron
Green (602)	Embryo development ending in birth or egg hatching (3×10^{-5})	Reproductive system, hermaphrodite distal tip cell
Darkolivegreen (163)	Embryo development ending in birth or egg hatching (8×10^{-3})	Reproductive system, pharyngeal interneuron
Royalblue (292)	Embryo development ending in birth or egg hatching (4×10^{-39})	Reproductive system, Capp
Mediumpurple3 (100)	Embryo development ending in birth or egg hatching (2×10^{-15})	Reproductive system, anal depressor muscle
Brown (933)	Embryo development ending in birth or egg hatching (8×10^{-41})	Reproductive system, Psub1
Coral1 (153)	Embryo development ending in birth or egg hatching (5×10^{-11})	Reproductive system, AVA
Darkorange2 (221)	Innate immune response (4×10^{-4})	AVA, pharyngeal interneuron
Lightcyan1 (98)	Innate immune response (2×10^{-38})	Intestine, outer labial sensillum, PVD
Black (1173)	Ion transport (2×10^{-16})	Ventral nerve cord, FLP, tail
Yellowgreen (104)	Ion transport (4×10^{-5})	Striated muscle

Table 2. Cont.

Module (No. Genes)	Biological Process	Tissue
Antiquewhite4 (35)	Ion transport (6×10^{-7})	Pharyngeal interneuron, retrovesicular ganglion
Paleturquoise (114)	Lipid transport (5×10^{-4})	Cephalic sheath cell, hermaphrodite
Bisque4 (212)	Proteolysis (2×10^{-10})	Intestine, PVD
Navajowhite2 (51)	Proteolysis (6×10^{-3})	Male, reproductive system
Thistle1 (58)	Reproduction (3×10^{-5})	AVA
Darkgreen (160)	Reproduction (3×10^{-5})	Reproductive system

3.2. Gene Expression Variation in Modules

As we have reduced the transcriptome data complexity by gene co-expression modules, we analyzed the gene expression variation at the module level. The gene level relative standard deviation (RSD) of gene expression was calculated, then module level RSD of gene expression was obtained by averaging the RSD of all genes (Figure 2). Sorting modules according to their RSDs, we can observe that the top 10 most stable modules include darkseagreen4 (mitochondrion), mediumpurple3 (mitochondrion), plum1 (mitochondrion), darkorange (mitochondrion), brown4 (embryo development ending in birth or egg hatching), brown (embryo development ending in birth or egg hatching), royalblue (embryo development ending in birth or egg hatching), white (ribosome), coral1 (embryo development ending in birth or egg hatching), and bisque4 (proteolysis). These modules are more related to housekeeping functions. The top 10 most variable modules include lightsteelblue1 (zinc ion binding), ivory (3'-untranslated region (UTR)-mediated mRNA destabilization), darkgrey (heterotrimeric G-protein complex), navajowhite2 (proteolysis), skyblue (nucleosome assembly), palevioletred3 (membrane), lightgreen (striated muscle dense body), lavenderblush3 (cul3-RING ubiquitin ligase complex), coral2 (structural constituent of cuticle), and lightyellow. Those modules are more related with stress response.

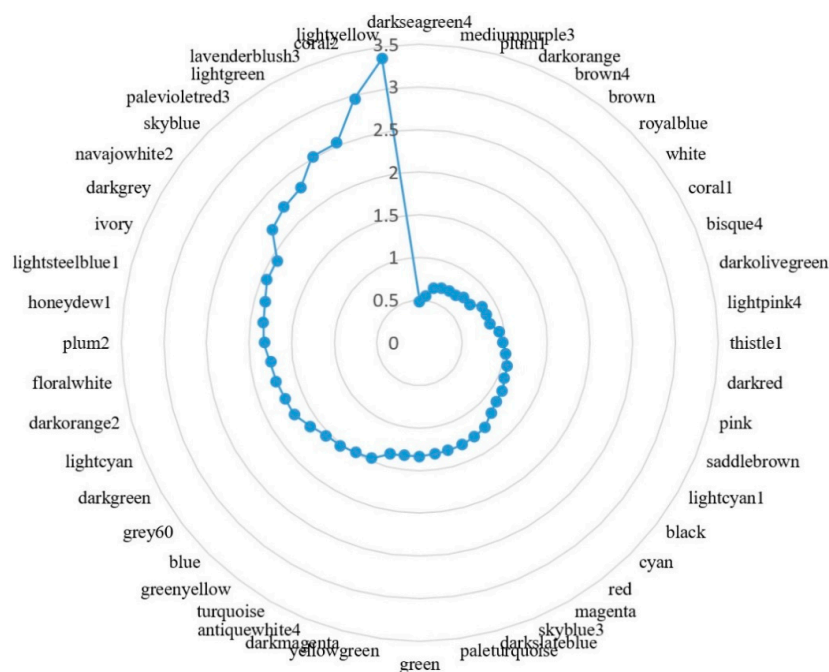


Figure 2. Radar chart showing module-based gene expression variation in *C. elegans*. The relative standard deviation (RSD) for each gene in a module was calculated, then the module gene expression variation was calculated by averaging the RSD of all genes. The blue dots represent the module RSD from highest (3.5) to lowest (0.5), reading anticlockwise.

3.3. Modules That Correlate with Experimental Conditions

Modules can be relatively independent units which perform a biological function. Associating the modular gene expression with experimental conditions may help to discover a module's functioning in specific conditions (Table S2). For example, amide-modified singlewalled carbon nanotube (a-SWCNT) treatment leads to the highest degree of coral2 module gene expression, which is a cuticle module. a-SWCNTs could cause retarded growth, reduced lifespan, and defective embryogenesis in worms [21]; acrylamide treatment induces the highest degree of red module gene expression, which is involved in cell shape regulation. Acrylamide could induce reversible de-phosphorylation of cytokeratins together with reversible filament aggregation [22]. This literature confirms our module GO result. The skyblue3 module has the highest expression in starved L1 wild-type worms, but the module has no significant GO annotation yet. A total of 56 of the 104 skyblue3 modular genes are annotated with hypothetical protein. These module genes may be involved in the starvation response.

3.4. Modules That May Correlate with Molting

C. elegans have a short life cycle of about three days at 22 °C. The cuticle protects *C. elegans* from environmental threats and allows growth by molting. It is synthesized five times, once in the embryo and subsequently at the end of each larval stage prior to molting [23]. Interestingly, we found five modules (coral2, darkslateblue, honeydew1, magenta, and paleturquoise) that are associated with the cuticle. Tissue enrichment analysis revealed that these modules may be associated with different tissue parts at different stages (Table 3). These modules are associated with collagen trimer and structural constituents of the cuticle. The module hub genes include collagen, unfolded protein response activated protein, and protease inhibitor, which may play roles in the cuticle structure [24].

Table 3. Five modules that may correlate with molting in *C. elegans*.

Module	Tissue Enrichment (<i>p</i> Value)	Phenotype Enrichment (<i>p</i> Value)	Stage	Hub Gene
Coral2	Amphid socket cell (2×10^{-4})	Dumpy (3×10^{-5})	Dauer	col-2
Darkslateblue	Gonadal primordium (4×10^{-7})	Dumpy (7×10^{-7})	Embryo	ZK662.2
Honeydew1	Hyp7 syncytium (4×10^{-10})	Blistered (3×10^{-6})	L4	col-138
Magenta	Outer labial sensillum (2×10^{-13})	Molt variant (3×10^{-20}), paralyzed (1×10^{-11})	L1	abu-13
Paleturquoise	Cephalic sheath cell (8×10^{-27})	Pathogen susceptibility increased (2×10^{-6})	Adult	C10G8.4

To seek how these modules affect phenotypes, the transcription factor targeting enrichment was analyzed by WormExp based on curated and high-quality gene expression datasets [25]. These five module genes were submitted to WormExp. Paleturquoise were enriched with PMK-1 targets ($p = 4 \times 10^{-16}$). It has been shown that aging is associated with a decline in the activity of PMK-1 p38 mitogen-activated protein kinase pathway, which regulates innate immunity in *C. elegans* [26]. While, as a key component in barrier integrity, cuticle collagen could sense stress and participate in innate immunity [27]. The lifespan of the *pmk-1* mutant is reduced four-fold by wounding, but the effect is compromised by inhibiting bacterial proliferation [28]. Thus, the mechanism of increased pathogen susceptibility may due to the regulation of cuticle by PMK-1. Magenta and darkslateblue were both enriched with BLMP-1 targets based on genome-wide ChIP-seq in *C. elegans* ($p = 4 \times 10^{-82}$ and 2×10^{-6}) [29]. Indeed, *blmp-1* mutants have a dumpy phenotype, a weak cuticle sensitive to oxidative stress, and show defective distal tip cells (DTC) migration [30].

To check whether the 48 modules were associated with specific chromosome regions, modular genes were subjected to Positional Gene Enrichment analysis. At a stringent *p* value (7×10^{-7}), nine modules were identified to be associated with a specific chromosome region. Four of the molting related modules were associated with a chromosome region, including darkslateblue, honeydew1, magenta, and paleturquoise (Table S3).

3.5. Genes Function Annotation

To demonstrate the application of the gene co-expression module in gene function annotation, the coral2 module were selected as it has the smallest gene number. A total of 16 of the 34 modular genes are known collagen coding genes. Other modular genes encode include Ground-Like, sperm-coating protein (SCP)-Like extracellular protein, Cuticlin-1, and several hypothetical proteins. The hub gene is *col-2*, which is present only in dauer larva [31]. Another modular gene, *cut-1*, has been proven to code for a dauer-specific non-collagenous component of the cuticle [32]. These results suggest the possible dauer-specific role of the coral2 module. Six gene products were annotated with hypothetical protein. Their Entrez GeneIDs are 190357, 179082, 182552, 178567, 184159, and 187146. All these genes encode proteins containing transmembrane helices as predicted by TMPred (data not shown). Three of them contain a signal peptide at the N-terminus as predicted by SignalP (data not shown). These results indicate that those hypothetical proteins may be components of the extracellular cuticle [33].

3.6. Higher Order Module Organization

To observe the organization between these modules, the network of modules was also analyzed. These modules can form a higher-order network with 11 meta-modules (Figure 3). Global connectivity analysis shows that the top three highly connected modules are coral1, darkolivegreen, and black, whose cellular component annotations are nucleus (embryo development ending in birth or egg hatching) and plasma membrane (signal transducer activity), respectively. The three least connected modules are coral2, lavenderblush3, and navajowhite2, whose GO annotations are collagen trimer, Cul3-RING ubiquitin ligase complex, and proteolysis (Table S4). These results may indicate more complexity exists in the cell “brain” nucleus and the cell “gatekeeper” plasma membrane. Those modules with more specific intracellular functions are less connected ($p = 2 \times 10^{-7}$, *t* test).

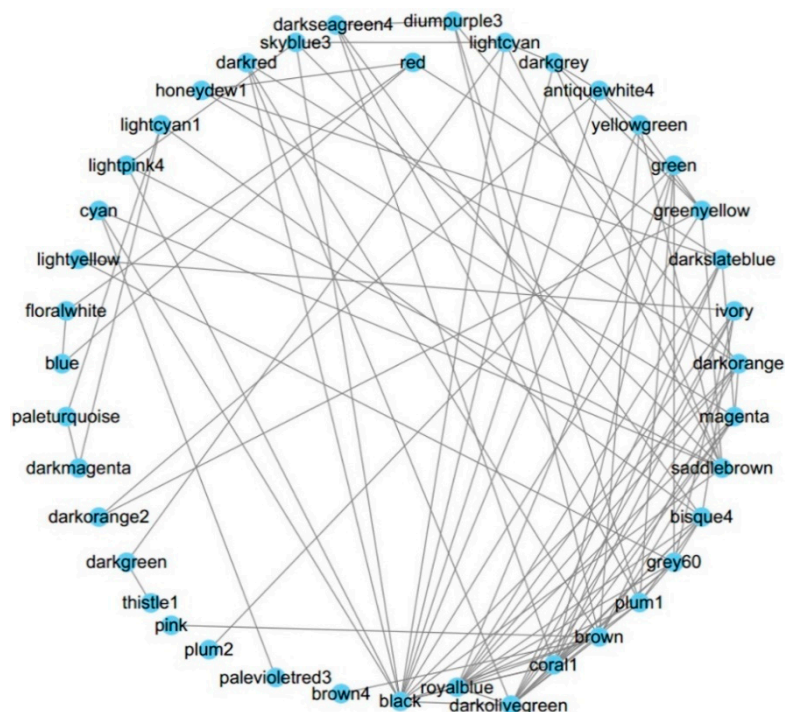


Figure 3. The network of modules. The higher order organization of the 48 modules was analyzed and the top 100 dense connections were visualized.

To check the relationship between these 11 meta-modules, a clustering diagram was plotted showing these modules can be divided into two main branches (Figure S2). The left branch modules

are mainly associated with the GO embryo development ending in birth or egg hatching function, suggesting their distinct expression pattern from other processes.

3.7. Comparison with Previous *C. elegans* Networks

To validate our results, multiple publications results were compared. Kim constructed a gene expression map for *C. elegans*, and identified 43 sets of highly correlated genes [1]. They identified four collagen modules, five germline enriched modules, while we identified five collagen modules, and nine potential germline enriched modules. In their study, 14 of the 43 modules could not be annotated, while in our study only two remained unannotated. However, this may be due to there being fewer annotated genes in the database at that time.

Reinke analyzed the global profile of germline gene expression in *C. elegans* and found that sperm-enriched and germline-intrinsic genes are nearly absent from the X chromosome [3]. We identified nine modules that were annotated as embryo development ending in birth or egg hatching, but only the grey60 module genes were enriched in the X chromosome.

To compare our candidate gene list with those predicted by the online function prediction tools SPELL (v2.0.3; available: <http://spell.caltech.edu:3000>) [19] and WormNet (v3; available: <http://www.functionalnet.org/wormnet>) [18], the coral2 module had six unknown genes submitted [18,19]. Genes 179082 and 187146 had no prediction results in both tools. Genes 190357 and 184159 were predicted as body morphogenesis genes by WormNet. Genes 182552 and 178567 were predicted as genes with sensory perception of chemical stimulus by SPELL, which indicate their potential transmembrane location.

Another recent research finding is the transcriptome analysis of the developmental stages of the two sexes in *C. elegans*, which identified six major WGCNA modules [6]. All six modules were covered by our findings (Table S5). They summarized the modules with similar function, so only six major modules were retrieved. For example, they found only Mod4 with 1421 genes to be associated with the cuticle. We obtained five modules (Table 3) containing 912 genes that are associated with cuticle. When overlapping with those data, 518 shared genes were found ($p < 5 \times 10^{-302}$, hypergeometric test), which comprise about 57% of our module genes.

A table containing all the module genes and their function annotation description information was provided in Table S6. For a better exploration of the module information provided in our analysis, a shiny-based web viewer was developed [34]. The tool site is available here: <http://bioinformatics.fafu.edu.cn/shiny/sample-apps/celegans/>

4. Discussion

Previous *C. elegans* transcriptome studies are limited by their sample size and specific conditions. Although recently scientists began to use state-of-art technologies, such as RNA-Seq and fluorescence activated cell sorting (FACS), to try and analyze the sex-, cell-, and stage-specific gene expression, 10 and 40 samples were profiled in the two studies [2,6]. This showed that a larger sample size may help to more robustly detect the gene co-expression modules in the human brain [13]. Here, we collected a compendium of *C. elegans* transcriptome data, the largest to date, and identified 48 co-expressed gene modules. These modules include genes for embryo development, cuticle, RNA processing, translation, ion transport and other biological processes. After identifying the gene modules, we further detected the associations between modules and experimental conditions, which may help to identify important modules in a specific condition. The five cuticle modules were subjected to gene prediction. Additionally, the higher order module organization analysis helps to illustrate the relationship between these functional units. Finally, our results were compared with previous studies to establish confidence.

However, some limitations should be considered in the future studies. Technically, there are several parameters that can be adjusted according to data type and research purpose, but currently no standard guideline has been established for parameter setup. These parameters include multiple

dataset integration, gene expression data normalization methods, similarity measures, and clustering methods. Biologically, the WGCNA presumes that the relationships between genes are approximately linear; nevertheless, the reality is more complex [35]. Gene expression is not the final level in the biosystem, but many other levels of regulations make the interpretation of gene co-expression network difficult. Although most of our modules have a clear functional annotation, we should be cautious when inferring the regulation relationships between genes. The WGCNA-derived gene network is undirected. More types of data, such as mutation, drug treatment, ChIP-Seq, and protein-protein interaction should be integrated to conclude the regulation. Thus, we mainly focus on gene function annotation and provide some potential transcription factor regulation information. In future, new algorithms should be applied to the emerging RNA-Seq data that accumulates. Conserved gene co-expression network can be obtained by comparing network from microarray and RNA-Seq. In sum, our analysis provides the module membership information and an overview of the *C. elegans* transcriptome, facilitating candidate gene screening in future experimental research for biologists.

Supplementary Materials: The followings are available online at <http://www.mdpi.com/2218-273X/8/3/70/s1>, Figure S1: *C. elegans* gene co-expression network. Each node represents a gene, and each line denotes the gene expression correlation between the two nodes. Figure S2: A clustering diagram representing the relationships between meta-modules (the higher order module); Table S1: *C. elegans* dataset used in our study; Table S2: Datasets that have the highest ME value in a specific module; Table S3: Modular genes enriched within chromosome regions by Positional Gene Enrichment analysis in *C. elegans*; Table S4: Module network and connectivity; Table S5: Our module results compared with reference [6]; Table S6: Module genes and connectivity and their function annotation.

Author Contributions: Conceptualization: W.L.; methodology: S.C., W.Z. and H.Z.; formal analysis: Y.Z. (Yuexian Zhong), S.W., and Z.X.; data curation: L.L. and Y.H.; writing—original draft preparation: W.L.; writing—review and editing: Y.Z. (Yang Zou) and W.T.; visualization: Y.Z. (Yu Zhang); funding acquisition: W.L.

Funding: This research was funded by the National Natural Science Foundation of China (no. 81502091) and the Open Project of Key laboratory of Loquat Germplasm Innovation and Utilization, Putian University, Fujian Province (no. 2017003). The article processing charge (APC) was funded by National Natural Science Foundation of China (no. 81502091).

Acknowledgments: We thank scientists who make their microarray data publicly available.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Kim, S.K.; Lund, J.; Kiraly, M.; Duke, K.; Jiang, M.; Stuart, J.M.; Eizinger, A.; Wylie, B.N.; Davidson, G.S. A gene expression map for *Caenorhabditis elegans*. *Science* **2001**, *293*, 2087–2092. [[CrossRef](#)] [[PubMed](#)]
2. Spencer, W.C.; Zeller, G.; Watson, J.D.; Henz, S.R.; Watkins, K.L.; McWhirter, R.D.; Petersen, S.; Sreedharan, V.T.; Widmer, C.; Jo, J.; et al. A spatial and temporal map of *C. elegans* gene expression. *Genome Res.* **2011**, *21*, 325–341. [[CrossRef](#)] [[PubMed](#)]
3. Reinke, V.; Smith, H.E.; Nance, J.; Wang, J.; Van Doren, C.; Begley, R.; Jones, S.J.; Davis, E.B.; Scherer, S.; Ward, S.; et al. A global profile of germline gene expression in *C. elegans*. *Mol. Cell* **2000**, *6*, 605–616. [[CrossRef](#)]
4. McKay, S.J.; Johnsen, R.; Khattra, J.; Asano, J.; Baillie, D.L.; Chan, S.; Dube, N.; Fang, L.; Goszczynski, B.; Ha, E.; et al. Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harb. Symp. Quant. Biol.* **2003**, *68*, 159–169. [[CrossRef](#)] [[PubMed](#)]
5. Hill, A.A.; Hunter, C.P.; Tsung, B.T.; Tucker-Kellogg, G.; Brown, E.L. Genomic analysis of gene expression in *C. elegans*. *Science* **2000**, *290*, 809–812. [[CrossRef](#)] [[PubMed](#)]
6. Kim, B.; Suo, B.; Emmons, S.W. Gene function prediction based on developmental transcriptomes of the two sexes in *C. elegans*. *Cell Rep.* **2016**, *17*, 917–928. [[CrossRef](#)] [[PubMed](#)]
7. Boeck, M.E.; Huynh, C.; Gevirtzman, L.; Thompson, O.A.; Wang, G.; Kasper, D.M.; Reinke, V.; Hillier, L.W.; Waterston, R.H. The time-resolved transcriptome of *C. elegans*. *Genome Res.* **2016**, *26*, 1441–1450. [[CrossRef](#)] [[PubMed](#)]
8. Lee, H.K.; Hsu, A.K.; Sajdak, J.; Qin, J.; Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res.* **2004**, *14*, 1085–1094. [[CrossRef](#)] [[PubMed](#)]

9. Wang, A.; Huang, K.; Shen, Y.; Xue, Z.; Cai, C.; Horvath, S.; Fan, G. Functional modules distinguish human induced pluripotent stem cells from embryonic stem cells. *Stem Cells Dev.* **2011**, *20*, 1937–1950. [[CrossRef](#)] [[PubMed](#)]
10. Langfelder, P.; Horvath, S. Wgcna: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [[CrossRef](#)] [[PubMed](#)]
11. Dai, M.; Wang, P.; Boyd, A.D.; Kostov, G.; Athey, B.; Jones, E.G.; Bunney, W.E.; Myers, R.M.; Speed, T.P.; Akil, H.; et al. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res.* **2005**, *33*, e175. [[CrossRef](#)] [[PubMed](#)]
12. Zhang, B.; Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*, 17. [[CrossRef](#)] [[PubMed](#)]
13. Oldham, M.C.; Konopka, G.; Iwamoto, K.; Langfelder, P.; Kato, T.; Horvath, S.; Geschwind, D.H. Functional organization of the transcriptome in human brain. *Nat. Neurosci.* **2008**, *11*, 1271–1282. [[CrossRef](#)] [[PubMed](#)]
14. Huang da, W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37*, 1–13. [[CrossRef](#)] [[PubMed](#)]
15. Angeles-Albores, D.; Lee, R.Y.N.; Chan, J.; Sternberg, P.W. Tissue enrichment analysis for *C. elegans* genomics. *BMC Bioinform.* **2016**, *17*, 366. [[CrossRef](#)] [[PubMed](#)]
16. De Preter, K.; Barriot, R.; Speleman, F.; Vandesompele, J.; Moreau, Y. Positional gene enrichment analysis of gene sets for high-resolution identification of overrepresented chromosomal regions. *Nucleic Acids Res.* **2008**, *36*, e43. [[CrossRef](#)] [[PubMed](#)]
17. NCBI PubMed. 2003. Available online: <https://www.ncbi.nlm.nih.gov/pubmed> (accessed on 27 July 2018).
18. Cho, A.; Shin, J.; Hwang, S.; Kim, C.; Shim, H.; Kim, H.; Kim, H.; Lee, I. Wormnet v3: A network-assisted hypothesis-generating server for *Caenorhabditis elegans*. *Nucleic Acids Res.* **2014**, *42*, W76–W82. [[CrossRef](#)] [[PubMed](#)]
19. Hibbs, M.A.; Hess, D.C.; Myers, C.L.; Huttenhower, C.; Li, K.; Troyanskaya, O.G. Exploring the functional landscape of gene expression: Directed search of large microarray compendia. *Bioinformatics* **2007**, *23*, 2692–2699. [[CrossRef](#)] [[PubMed](#)]
20. Farber, C.R. Identification of a gene module associated with BMD through the integration of network analysis and genome-wide association data. *J. Bone Miner. Res.* **2010**, *25*, 2359–2367. [[CrossRef](#)] [[PubMed](#)]
21. Chen, P.H.; Hsiao, K.M.; Chou, C.C. Molecular characterization of toxicity mechanism of single-walled carbon nanotubes. *Biomaterials* **2013**, *34*, 5661–5669. [[CrossRef](#)] [[PubMed](#)]
22. Kippenberger, S.; Bernd, A.; Loitsch, S.; Guschel, M.; Muller, J.; Bereiter-Hahn, J.; Kaufmann, R. Signaling of mechanical stretch in human keratinocytes via map kinases. *J. Investig. Dermatol.* **2000**, *114*, 408–412. [[CrossRef](#)] [[PubMed](#)]
23. Page, A.P.; Johnstone, I.L. The Cuticle. 2007, pp. 1–15. Available online: http://www.wormbook.org/chapters/www_cuticle/cuticle.html WormBook (accessed on 27 July 2018).
24. Thierry-Mieg, D.; Thierry-Mieg, J. Aceview: A comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* **2006**, *7*, S12. [[CrossRef](#)] [[PubMed](#)]
25. Yang, W.; Dierking, K.; Schulenburg, H. Wormexp: A web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis. *Bioinformatics* **2016**, *32*, 943–945. [[CrossRef](#)] [[PubMed](#)]
26. Youngman, M.J.; Rogers, Z.N.; Kim, D.H. A decline in p38 mapk signaling underlies immunosenescence in *Caenorhabditis elegans*. *PLoS Genet.* **2011**, *7*, e1002082. [[CrossRef](#)] [[PubMed](#)]
27. Taffoni, C.; Pujol, N. Mechanisms of innate immunity in *C. elegans* epidermis. *Tissue Barriers* **2015**, *3*, e1078432. [[CrossRef](#)] [[PubMed](#)]
28. Tong, A.; Lynn, G.; Ngo, V.; Wong, D.; Moseley, S.L.; Ewbank, J.J.; Goncharov, A.; Wu, Y.-C.; Pujol, N.; Chisholm, A.D. Negative regulation of *Caenorhabditis elegans* epidermal damage responses by death-associated protein kinase. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 1457–1461. [[CrossRef](#)] [[PubMed](#)]
29. Niu, W.; Lu, Z.J.; Zhong, M.; Sarov, M.; Murray, J.I.; Brdlik, C.M.; Janette, J.; Chen, C.; Alves, P.; Preston, E.; et al. Diverse transcription factor binding features revealed by genome-wide chip-seq in *C. elegans*. *Genome Res.* **2011**, *21*, 245–254. [[CrossRef](#)] [[PubMed](#)]
30. Huang, T.-F.; Cho, C.-Y.; Cheng, Y.-T.; Huang, J.-W.; Wu, Y.-Z.; Yeh, A.Y.-C.; Nishiwaki, K.; Chang, S.-C.; Wu, Y.-C. Blmp-1/blimp-1 regulates the spatiotemporal cell migration pattern in *C. elegans*. *PLoS Genet.* **2014**, *10*, e1004428. [[CrossRef](#)] [[PubMed](#)]

31. Cox, G.N.; Fields, C.; Kramer, J.M.; Rosenzweig, B.; Hirsh, D. Sequence comparisons of developmentally regulated collagen genes of *Caenorhabditis elegans*. *Gene* **1989**, *76*, 331–344. [[CrossRef](#)]
32. Sebastiano, M.; Lassandro, F.; Bazzicalupo, P. Cut-1 a *Caenorhabditis elegans* gene coding for a dauer-specific noncollagenous component of the cuticle. *Dev. Biol.* **1991**, *146*, 519–530. [[CrossRef](#)]
33. Frand, A.R.; Russel, S.; Ruvkun, G. Functional genomic analysis of *C. elegans* molting. *PLoS Biol.* **2005**, *3*, e312. [[CrossRef](#)] [[PubMed](#)]
34. Chang, W.; Cheng, J.; Allaire, J.J.; Xie, Y.; McPherson, J. Shiny: Web Application Framework for R. R Foundation for Statistical Computing, Vienna, Austria 2015. R Package Version 0.11. Available online: <http://CRAN.R-project.org/package=shiny> (accessed on 27 July 2018).
35. Uygun, S.; Peng, C.; Lehti-Shiu, M.D.; Last, R.L.; Shiu, S.-H. Utility and limitations of using gene expression data to identify functional associations. *PLoS Comput. Biol.* **2016**, *12*, e1005244. [[CrossRef](#)] [[PubMed](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).