Published in final edited form as:
Cell Rep. 2016 October 11; 17(3): 917–928. doi:10.1016/j.celrep.2016.09.051.

Gene Function Prediction Based on Developmental Transcriptomes of the Two Sexes in *C. elegans*

Byunghyuk Kim[1], **Bangxia Suo**[1], and **Scott W. Emmons**[1,2,3,*]

[1]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[2]Dominick P. Purpura Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY 10461, USA

SUMMARY

We compare whole-animal RNA-seq transcriptomes for *C. elegans* males and hermaphrodites from the late L3 larval stage to young adulthood. During this interval, male sexual structures develop, including extensive neurogenesis and synaptogenesis that nearly doubles the size of the nervous system. Previous genome-wide expression studies in *C. elegans* have usually focused on only one sex—the hermaphrodite—and there are a relatively large number of genes that remain without meaningful annotation. In the present study, differential expression analysis of the RNA-seq data revealed 1,751 genes expressed at a higher level in the male. By differential expression and co-expression analyses, we identified transcription factors required for differentiation of male genital structures, semen proteins, and candidates for components of synapse function. Comparison with other prediction tools suggests that our dataset can expand gene predictions. The results validate the dataset as a rich resource for future gene discovery in *C. elegans*.

In Brief

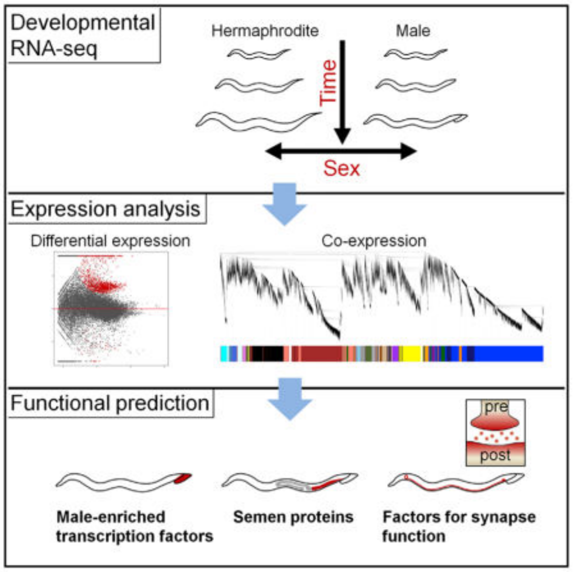
Taking advantage of a unique developmental feature of *C. elegans* males, Kim et al. compare transcriptomes of the two sexes to identify male-specific and nervous-system-enriched genes. The dataset and mining methods used provide a rich resource for future gene discovery.

[*]Correspondence: scott.emmons@einstein.yu.edu.
[3]Lead Contact

ACCESSION NUMBERS
The accession number for the sequence data reported in this paper is NCBI: GSE83216.

SUPPLEMENTAL INFORMATION
Supplemental Information includes four figures and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2016.09.051.

AUTHOR CONTRIBUTIONS
B.K. and S.W.E designed the experiments, analyzed the data, and wrote the paper. B.K. and B.S. conducted the experiments.



INTRODUCTION

Genome-wide transcriptional expression profiles have greatly enhanced our understanding of how genes are regulated and what their roles are in biological systems. One promising strategy to reveal potential functions of unknown genes is a gene co-expression approach, where functions of unknown genes can be inferred from the similarity of their expression profiles to those of genes with known functions. This approach has been applied in global expression analysis of *Caenorhabditis elegans* (Kim et al., 2001) and developed as a computational tool for gene discovery with a large compendium of gene expression profiles (SPELL: http://spell.caltech.edu:3000; Hibbs et al., 2007).

Despite the considerable progress in functional annotation for *C. elegans* genes through global expression and mutational studies as well as many studies focused on single genes, a large fraction of predicted genes still remains without functional assignment. The functions of protein-coding genes can be inferred from similarity of protein structure or amino acid sequence to genes of known function or from mutant phenotype. When this information is available for a gene in *C. elegans*, a functional gene name is assigned following standard nomenclatural rules. Of the approximately 20,000 predicted protein-coding genes in the *C. elegans* genome, 9,761 (48%) have such assigned gene names. The remaining 10,601 genes are named by their position in the genomic sequence (WormBase [WS250]: http://www.wormbase.org). For these genes, little is known about their biological function. One possible explanation for unassigned genes is that males and their behaviors have been largely excluded from prior studies. Thus, genes that act or are expressed predominantly or exclusively in the male may have been overlooked.

Sexually dimorphic gene expression shapes phenotypic differences between the two sexes during development (Ellegren and Parsch, 2007). Temporal gene expression profiles of the two sexes may allow us to predict functional roles of individual genes by providing the timing of action for the gene products during sexual maturation. Up to now, most temporal gene expression studies in *C. elegans* focused on only one sex—the hermaphrodite—whereas some included the male of only one stage, either L4 or adult (Gerstein et al., 2010; Hillier et al., 2009; Jiang et al., 2001; Reinke et al., 2000, 2004; Snoek et al., 2014; Spencer et al., 2011). To our knowledge, only one study, in which microarrays were applied to different larval stages (L2–L4) of both hermaphrodite and a genetically masculinized worm population, has examined temporal gene expression of both sexes, identifying over 300 sex-enriched genes (Thoemke et al., 2005).

When compared to the self-fertilizing hermaphrodite, the *C. elegans* male has a distinct body morphology, a different gonad system, 40 male-specific muscles, and 85 additional neurons that contribute to the copulatory structures and behavior (Emmons, 2014). The morphological differentiation of the male, including the addition of the male-specific cells and the formation of over 8,000 synapses, arises mostly at the third larval stage (L3) and later (Jarrell et al., 2012; Sulston et al., 1980). By contrast, hermaphrodite differentiation is largely complete by this time and primarily involves initiation of gametogenesis. This developmental distinction provides an opportunity to identify genes that are specifically upregulated during the neurogenic and morphogenetic episode associated with male sexual maturation as well as genes that function specifically in adult male reproduction.

Here, we report an expression analysis of the developmental transcriptome ranging from late L3 to adult in the two sexes of *C. elegans* using whole-animal RNA sequencing (RNA-seq). As predicted, among 21,143 genes with detectable expression, a large number are upregulated in the male during this interval. We employed differential expression, unbiased gene correlation analysis, and guilt-by-association approaches to identify transcription factors regulating male morphogenesis, semen components, synaptic components, and putative components for cilia function. From the results, functions for a large number of genes may be inferred, making this dataset a rich source of future gene discovery.

RESULTS

RNA-Seq Transcriptomes of the Two Sexes during Sexual Maturation

To analyze gene expression during sexual maturation in *C. elegans*, we performed RNA-seq for five samples for each sex ranging at 6-hr intervals from late L3 to young adult stages (Figures 1A and S1). Approximately 140 million sequencing reads from the ten samples (8–28 million reads per sample) were mapped to the *C. elegans* genome (WS190). We considered a gene as expressed if there were a total of five or more reads summing across the ten samples. We obtained expression data for a total of 21,143 genes, including 17,967 protein-coding genes covering 88% of ~20,000 predicted protein-coding genes annotated in *C. elegans* (WormBase [WS250]: http://www.wormbase.org). The expression data are made available in the Supplemental Information (Table S1).

Principal-component analysis revealed a robust difference between the sets of genes expressed by the two sexes during sexual maturation. During the late L3 stage, when sexual differentiation is not yet prominent, overall gene expression profiles of the two sexes are similar, but they diverge as worms develop into the adult stage (Figure 1B).

As an initial assessment of the gene sets, we analyzed the data for genes with different overall levels of expression in the two sexes. By comparing two sets of pooled data from the five samples for each sex, we identified 1,751 genes (1,645 protein coding) expressed at a higher level in the male, consistent with the burst of differentiation in the male, whereas 68 genes (66 protein coding) were expressed at a higher level in the hermaphrodite (false discovery rate < 1%; >4-fold change; Figure 1C; Table S2). These are minimum estimates because variation across the time points reduces the statistical power of the comparison. As expected, the male-enriched set included many genes for known sperm proteins (e.g., major sperm protein gene family) and several transcription factors known to be critical for male development (e.g., *egl-5*, *mab-3*, *mab-23*, and *dmd-3*; Emmons, 2014), whereas the hermaphrodite-enriched set included genes that are known to be expressed specifically in the hermaphrodite or in the hermaphrodite-specific cells (e.g., *vit-2* and *cht-3*; Yi and Zarkower, 1999; Mounsey et al., 2002). Gene ontology (GO) term analysis showed that the male-enriched set was overrepresented with the GO terms related to protein kinases and phosphatases, which is a characteristic of sperm-enriched genes (Reinke et al., 2000; Table S3).

Transcription Factors for Male Differentiation

In order to assess the possibility for gene discovery based on the differential expression levels, we examined transcription factor genes in more detail. As the male-specific transcription factors noted above were present in the male-enriched set, we sought to identify additional transcription factors expressed at a higher level in the male. By combining the male-enriched set with a list of putative transcription factor genes in the *C. elegans* genome (Reinke et al., 2013), we identified 17 male-enriched transcription factors (Figure 2A). In addition to the four genes mentioned above (the three doublesex/MAB-3 [DM] domain transcription factor genes *mab-3* [Raymond et al., 1998], *mab-23* [Lints and Emmons, 2002], and *dmd-3* [Mason et al., 2008] and the homeobox gene *egl-5* [Ferreira et al., 1999]), the list included an additional transcription factor known to be involved in male development, the COUP (chicken ovalbumin upstream promoter) transcription factor gene *unc-55* (Shan and Walthall, 2008).

To test whether the members of the remaining set are also required for male development, we examined four available mutant males. There were no obvious defects in gross morphology or morphology of the gonad in the mutants males (data not shown). However, two out of four mutants tested showed slightly abnormal ray formation in the male tail. The wild-type male has nine bilateral pairs of rays in its tail; however, in mutant males for two genes *ceh-48* and *T20H4.2*, missing or fusion of the ray structure was frequently observed (Figure 2B), indicating these genes encode putative transcription factors required for male tail development. To further observe the structure of the nervous system in the four mutants, we used a *pkd-2p::gfp* reporter to visualize some male-specific neurons, including RnBs, CEMs, and HOB (Jia and Emmons, 2006). We did not find any observable defects in cell number, cell body location, or process placement of these neurons, except for occasional misplacement of RnB dendrites corresponding to missing or fusion of the rays described above (data not shown).

We also examined expression of all 11 members of the DM domain gene family. DM domain transcription factors are evolutionarily conserved regulators of sexual development in many metazoans (Matson and Zarkower, 2012). Hierarchical clustering analysis divided these genes into two groups: higher expression in male (>2-fold change; *dmd-3*, *dmd-4*, *dmd-6*, *dmd-10*, *dmd-11*, *mab-3*, and *mab-23*) and similar or lower expression in male (<1.3-fold change; *dmd-5*, *dmd-7*, *dmd-8*, and *dmd-9*; Figure S2). Thus, it is possible that the four DM domain genes *dmd-4*, *dmd-6*, *dmd-10*, and *dmd-11* may also control male development along with *mab-3*, *mab-23*, and *dmd-3*.

Unbiased Gene Correlation Analysis Partitioned Genes into Multiple Modules for Diverse Functions

We next treated each time point as a separate data point and determined the co-expression relationships between genes by performing weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008). This method partitions a gene set into modules defined as branches of the co-expressed gene cluster tree selected by a tree-cutting method. By this analysis, the set of 21,143 genes was partitioned into 27 modules with sizes ranging from 105 to 5,747 genes (Figure 3A; Table S1). Six major modules, each containing more than 1,000 genes, were readily associated by enrichment of GO terms with nervous system function (Mod1 and Mod2), semen/sperm development (Mod3), cuticle/hypodermis formation (Mod4), and germline/oocyte development (Mod5 and Mod6; Table 1; Figure 3A). However, each module was also enriched with more than one functional category: Mod1 also contained genes for muscle development, whereas Mod3 contained a set of genes related to neuropeptide signaling. Most of the male- or hermaphrodite-enriched genes described above belonged, respectively, to Mod3 or Mod6 (Table 1), suggesting that most of these sex-enriched genes function in various aspects of gamete development and reproduction.

The temporal gene expression patterns of the genes in the six major modules were well correlated with the presumptive functions of genes in these modules (Figure 3B). Genes in Mod1 and Mod2 showed relatively higher expression in the male compared to the hermaphrodite during late developmental stages from L4 to adult, as expected given the addition of many male-specific neurons and muscles at these stages (Sulston et al., 1980). Genes in Mod3 or Mod5 and Mod6 were highly enriched in late stages of the male or the hermaphrodite, respectively, consistent with timing of sperm or oocyte development. Finally, genes in Mod4 were enriched in larval stages of both sexes, when general body morphogenesis occurs. Notably, genes previously shown to be enriched in specific tissues, including neurons (Spencer et al., 2011), sperm and oocytes (Reinke et al., 2004), and cuticle (Johnstone, 2000; Page and Johnstone, 2007), were preferentially assigned to one or two of the six major modules (Figure 3C), consistent with the notion that each module contains a set of co-regulated genes operating in a specific tissue or for related biological functions.

One large set of predicted genes about which little is known consists of over 280 genes encoding C-type lectin-like domains (*clec* genes). Some *clec* genes have been shown to be upregulated in the male (Thoemke et al., 2005; Miersch and Döring, 2012), others upon infection with pathogens (O'Rourke et al., 2006), whereas one is required for nervous system patterning (Kulkarni et al., 2008). As a validation of module assignment, we analyzed the tissue-specific expression patterns of several *clec* genes. Two *clec* genes selected from Mod1 (*clec-64* and *clec-199*) were expressed in subsets of head neurons in both sexes (Figure 3D). Although genes in Mod1 were expected to have additional expression in the male compared to the hermaphrodite, we did not observe robust expression of the two *clec* genes in the male tail except for occasional faint expression, possibly resulting from incomplete promoter sequences we used. Two *clec* genes selected from Mod3 (*clec-207* and *clec-219*) were expressed in the male-specific gonadal structure called vas deferens (Figure 3E). These results are consistent with module assignment and indicate that the function of unknown genes can be suggested from the modules to which they are assigned. As each module contains many uncharacterized genes, these data should be a rich source for gene discovery.

Putative Semen Protein Genes Identified by a "Guilt-by-Association" Approach

To further identify subsets of genes functioning closely together in a specific biological process, we adopted a correlation-based guilt-by-association approach. In this approach, genes are sought with expression patterns highly correlated with a set of known genes in a pathway or process (probes; see Experimental Procedures). Using this approach, we sought to identify components of semen. We expected that a set of genes in Mod3 might produce semen components. Seminal fluid proteins are important for reproductive success in many organisms. Although studied extensively in insects (reviewed in Avila et al., 2011), less is known about semen components in *C. elegans*, with there being just two studied examples, TRY-5 and PLG-1 (Palopoli et al., 2008; Smith and Stanfield, 2011). In our dataset, *try-5* is placed in Mod3, whereas *plg-1* is not expressed because this gene was disrupted by a retrotransposon insertion in the laboratory strain (Palopoli et al., 2008).

We used four genes (*clec-161*, *clec-197*, *F58A4.1*, and *T26A8.3*) as probes, because they were highly co-regulated in the RNA-seq data (Pearson correlation coefficient > 0.9720) and known to be expressed in the same region of the male gonad, the vas deferens, the presumptive site of semen generation and storage (Thoemke et al., 2005). We found 54 genes in Mod3 co-expressed with these four probe genes (correlation co-efficient > 0.9720; Figure 4A). This gene list included the two *clec* genes we randomly selected from Mod3 (*clec-207* and *clec-219*) and showed above are expressed in the vas deferens (Figures 3D and 3E). None of the 54 genes have previously been functionally characterized, with only six of them having human orthologs; 44 do not have assigned gene names (Table S4).

We further tested the three most highly expressed genes among the 54 co-expressed genes to determine whether their protein products were indeed components of semen. Using promoter::GFP fusion transgenes, we found that all three are expressed in the male gonad as expected: *ins-31* in the seminal vesicle; *F59B2.12* in the vas deferens; and *B0207.5* in the valve region of the vas deferens (Figures 4B and 4C). Expression of these genes was not previously known, except *ins-31*, which was shown to be expressed in the male gonad (Thoemke et al., 2005). There was no detectable expression in any stages of hermaphrodites (data not shown). We examined protein localization for *ins-31* and *F59B2.12* (we did not include *B0207.5* because it is expressed in the valve cells that are not well characterized anatomically) using translational reporter lines. We found that their protein products are localized to the vesicular structures in the seminal vesicle and vas deferens, respectively (Figure 4D). INS-31 protein was also found in coelomocytes in males (data not shown). Both INS-31::sfGFP and F59B2.12:: sfGFP (superfolder GFP) fusion proteins were transferred from males to hermaphrodites during mating (Figures 4E and S3). Thus, these proteins are components of semen.

To compare our candidate gene list with those obtained using other gene-prediction tools, we utilized a co-expression-based SPELL (Hibbs et al., 2007) and a proteome-scale gene network WormNet (Cho et al., 2014). For direct comparison, we identified the same number of candidate genes in the SPELL database that are most highly correlated with our four semen probe genes (top 54 genes). However, we could obtain only 31 genes from WormNet with the same probe genes, even though WormNet usually gave a list of top 200 genes when we used well-studied genes as probes; this is possibly due to the lack of functional information for the probe genes, as these genes have exclusive expression in males and thus have been understudied previously. None of these candidate genes have previously been related to semen function. Module assignment revealed that all but one gene (98%) yielded by SPELL belong to the highly male-enriched Mod3, whereas only three genes (10%) in WormNet belong to Mod3 (Figure 4F). Our candidate gene list shared 18 genes with the SPELL gene list, but not with the WormNet gene list (Figure 4F; Table S4). The shared gene list contained two semen protein genes (*ins-31* and *F59B2.12*) shown above, suggesting that the number of candidate genes for semen proteins could be reduced by utilizing co-expression data like SPELL. In addition, there exists a considerable number of semen candidate genes exclusively predicted by our approach. This result shows that our dataset can generate a set of candidate genes for semen components.

Candidate Genes for Synapse Formation and Function

Using the same guilt-by-association approach, we sought to identify synaptic proteins. In view of the burst of synaptogenesis during maturation of the male, transcripts of genes involved in synapse formation or function are expected to be elevated in the male-enriched modules (Mod1–Mod3). Previously, a large-scale RNAi screen in *C. elegans* identified over 100 genes required for synapse function based on mutant response to the acetylcholinesterase inhibitor aldicarb (Sieburth et al., 2005); however, there are likely to be many additional synaptic components that remain to be discovered, as proteomic studies have revealed several hundred factors residing in synapses (Bayés and Grant, 2009).

To obtain a probe gene set, we first searched conserved genes for synapse function from the literature (Jin, 2005; Jin and Garner, 2008; Richmond, 2005; Südhof and Rizo, 2011) and selected five genes, including *cpx-1* (complexin), *rab-3* (rab3), *ric-4* (SNAP-25), *unc-13* (Munc13), and *unc-104* (KIF1A/kinesin-3), because they were highly co-regulated in the RNA-seq data (correlation coefficient > 0.9326) and known to be expressed exclusively in most neurons based on WormBase data. Using these five genes as probes, we identified 63 highly co-expressed genes in the dataset (correlation coefficient > 0.9326; Figure 5A; Table S5). Only 16 (25%) out of the 63 genes are previously known to affect synaptic transmission or encode neuropeptides (Table S5). The remaining 47 genes have not been previously associated with synapse formation or function. However, about half of the 63 genes are implicated in synapse function by associated GO terms and domain structure information (channel, neuropeptide-related, synaptic vesicle or dense core vesicle trafficking, signal transduction, and cell adhesion; Figure 5B).

Of the 47 genes not previously associated with synapses, we further focused on 29 genes that have human orthologs (Table S5), because synaptic components in general are likely to be conserved across diverse taxa. The expression patterns of 18 of these genes were previously annotated in WormBase. For ten (out of 11) of the remaining genes, we examined their expression using promoter::GFP fusion transgenes (Figure S4). All 28 examined genes were expressed in neurons and/or muscles (Figure 5C), indicating that these genes act in most relevant tissues for synapse function. In all cases we tested, outside of the male-specific nervous system in the tail, expression was similar in both sexes; hence, these genes are expected to have similar functions in both sexes.

To address further whether these genes are required for synaptic transmission, we tested response of available mutants to aldicarb. As aldicarb perturbs the breakdown of acetylcholine at neuromuscular synapses, aldicarb-resistant or hypersensitive mutants have a decrease or increase in acetylcholine signaling, respectively (Mahoney et al., 2006). Among the 29 genes with human orthologs, mutations were available for 12 genes from the Caenorhabditis Genetics Center, and we used these mutants for an aldicarb assay. We found two resistant mutants, *ckr-1* and *W10C8.5*, and one hypersensitive mutant, *R13A5.9* (Figure 5D), suggesting that these gene products function at least in cholinergic synapses. *ckr-1* and *W10C8.5* encode a cholecystokinin receptor homolog and glutamine synthetase/guanido kinase domain protein, respectively; *R13A5.9* encodes a major facilitator superfamily domain protein. None of these genes were previously implicated in synapse function or found in the previous large-scale RNAi screen using aldicarb (Sieburth et al., 2005). Taken altogether, these results suggest that this set of genes highly correlated in their expression patterns with known synaptic genes may identify previously unrecognized components of synapses or proteins required for synapse formation or function.

To compare our candidate gene list with those obtained using other gene-prediction tools, we obtained candidate gene lists using SPELL and WormNet. For each dataset, we identified the 63 most highly correlated genes with the five synaptic probe genes. All analyses performed well based on their ability to identify previously known genes for synapse function: among the 63 genes, 17% (this study), 27% (SPELL), and 19% (WormNet; Figure 5E). These rates were further increased in shared gene lists between the datasets. For example, we found 14 overlapping genes between our candidate list and the SPELL list, of which six (*nrx-1*, *rbf-1*, *ida-1*, *unc-14*, *mpz-1*, and *gar-2*) have known synaptic function and two (*snt-4* and *DH11.5*; also shared by the WormNet list) are synaptotagmin homologs involved in synaptic vesicle trafficking (total 57%; Figure 5E; Table S5). The remaining six genes include *R13A5.9* for which mutant showed a hypersensitive response to aldicarb in our analysis (Figure 5D). However, two genes for the resistant mutants in the aldicarb assay (*ckr-1* and *W10C8.5*) were predicted by our approach (Table S5). These results show that our dataset can expand a candidate gene list for synapse function.

Prediction of Putative Ciliary Genes

As a final example, we examined the possibility of identifying components of cilia, which are conserved macromolecular structures. In addition to the 60 ciliated neurons present in both sexes, males generate 52 additional ciliated neurons during late-larval maturation (Sulston et al., 1980; White et al., 1986). Therefore, it is expected that transcripts of many ciliary genes will be enriched in the male when compared to the hermaphrodite during this period, and components of cilia may be found by a co-expression approach. For example, three known ciliary genes, *bbs-1*, *che-13*, and *osm-5*, are expressed in most of the ciliated sensory neurons (Efimenko et al., 2005) and are highly co-regulated within Mod1 (correlation coefficient > 0.9226). Forty-six genes are equally co-expressed with these three genes, all of which belonged to Mod1 (Table S6). Only three of these genes (*bbs-9*, *dyf-2*, and *tub-1*) have previously been implicated in ciliary functions (Blacque et al., 2005; Efimenko et al., 2006; Mukhopadhyay et al., 2005). Of the 21 out of the 46 genes with known expression patterns in WormBase, 16 genes (76%) have expression in some of the ciliated sensory neurons. It remains to be determined whether some of the remaining 43 co-expressed genes also function in cilia.

DISCUSSION

In *C. elegans*, most male-specific structures, together with the addition of a large number of synapses, arise during a short interval of late larval development (Emmons, 2014; Sulston et al., 1980). This brief episode of neurogenesis and morphogenesis is expected to be accompanied by relatively rapid changes in overall gene expression (see Figure 1B). This unique developmental feature makes possible the generation of temporal gene expression profiles with great resolution, which provide a template for functional prediction not only of male-specific genes but also genes with functions in both sexes that are upregulated specifically in the male during this period. In this study, we used temporal gene expression profiles to predict gene functions that are sex specific (e.g., male transcription factors and semen components) as well as potentially shared between both sexes (e.g., factors for synapse function and components of cilia). The validation experiments for some candidate genes support our predictions of genetic functions.

Prediction of gene functions through co-expression approaches is based on the similarity of gene expression profiles across multiple samples or conditions. It is thought that adding more datasets obtained from different conditions increases the predictive power of the method (D’haeseleer et al., 2000). In our analyses, we used ten samples comprising five different time points for each sex at 6-hr intervals from late L3 to young adult stages. Recently, it has been reported that, during the transition from larval to adult stages, the hermaphrodite shows rapid gene expression shift that was not previously recognized when sampled at hourly intervals (Snoek et al., 2014). Thus, we expect that the predictive power of co-expression analysis can be increased by sampling more-temporal data points during the interval we examined here.

The use of gene transcriptional co-expression analysis to predict gene functions rests on the assumption that subsets of genes that encode proteins primarily or exclusively functioning together in a pathway or molecular complex will be coordinately transcribed. In addition to the sets of genes we identified—transcription factors for male development, semen proteins, and putative synaptic components—many other pathways can be explored with our data. These could include genes for muscle development (Mod1), cuticle and molting (Mod4), ribosome-associated proteins (Mod5), and oocyte development (Mod6). Other pathways and cell structures that involve conserved protein complexes include RNA processing, chromatin, and transcription.

Although our guilt-by-association approach identified several candidate genes with functions possibly related to the functions of the probe gene set, the limited sample size makes it difficult to assess the overall success of the method. We expect that this is the case especially for the synaptic and ciliary candidates, which do not show robust sex-biased expression. We found that three out of 12 mutants were defective in synaptic transmission, but we do not know what the background frequency would be among all genes in the nervous-system-related modules. For ciliary function, we predicted 46 candidate genes, only three of which are known cilia genes and one of which (*unc-104*) may function in cilia (Ou et al., 2007). One way to increase the power of prediction is to utilize existing databases in conjunction with ours, such as SPELL (Hibbs et al., 2007) or WormNet (Cho et al., 2014), as described in Figures 4F and 5E.

We detected significant transcription for 88% of the approximately 20,000 predicted protein-coding genes in the *C. elegans* genome. Of these genes, 8,273 (46%) have assigned functional gene names based on structural or phenotypic information. Interestingly, in the male-enriched set of 1,645 protein-coding genes, only 365 have functional gene names (22%); the remaining 1,280 (78%) have only sequence-based names and usually no listed mutant phenotype. Our modularity analysis partitions the 21,143 protein- and non-protein-coding genes in a highly informative manner, and genes without functional gene names are distributed throughout all the modules. For many of these genes, functions may be inferred from the functions of their nearest neighbors in the co-expression tree.

EXPERIMENTAL PROCEDURES

Synchronized Worm Preparation

To generate worm populations containing large numbers of males, *him-5*(*e1490*) *V* worms were grown at 20°C on standard nematode growth media (NGM) plates with OP50 *E. coli* as a food source. To obtain pure preparations of males and hermaphrodites of a specific age, individual animals were selected by hand from staged, synchronized populations, obtained as follows. Gravid adult hermaphrodites were allowed to lay eggs overnight. The following day, hermaphrodites and any larval worms were gently washed off the agar, leaving eggs adhering to the surface. Newly hatched L1 worms were then harvested at 1-hr intervals by gently washing them off the plate with M9 buffer. The resulting worms were grown at 20°C to a desired time point, and then males or hermaphrodites were individually selected over an interval of 1 hr (up to 600 worms per time point). For NGM agar, M9, and general worm methods, consult Brenner (1974).

To cover a certain developmental stage, worms from three time points were pooled into one sample, resulting in five samples for each sex:

samples M1 and H1: 30, 32, and 34 hr post-hatching (correspond to late L3);

samples M2 and H2: 36, 38, and 40 hr post-hatching (correspond to early L4);

samples M3 and H3: 42, 44, and 46 hr post-hatching (correspond to late L4);

samples M4 and H4: 48, 50, and 52 hr post-hatching (correspond to young adult); and

samples M5 and H5: 54, 56, and 58 hr post-hatching (correspond to young adult).

Each sample was washed three times in M9 buffer and stored at −80°C.

RNA Isolation, RNA Sequencing, and Data Processing

Total RNA was extracted using TRIzol (Invitrogen); DNA was removed with TURBO DNA-free (Life Technologies). Concentration and quality of RNA was determined using the Agilent 2100 Bioanalyzer. For each sample, a total of 1 μg of RNA was submitted to the Genomics Core Facility of the Albert Einstein College of Medicine for library preparation and RNA sequencing. Library preparation followed the protocol described at http://wasp.einstein.yu.edu/index.php/Main_Page, and sequences were determined for 100-bp single-end reads using the Illumina HiSeq2500. Sequence reads were aligned to the *C. elegans* reference genome (WS190) using GSNAP (version 2012-07-12; Wu and Nacu, 2010). Assignment of reads to genes was performed with htseq-count in HTSeq (version 0.5.3p3; Anders et al., 2015). Both of these steps were performed by the Genomics Core Facility. The resulting read count data were normalized using R/Bioconductor package DESeq (Anders and Huber, 2010). After genes with fewer than five reads total across the ten samples were removed, significant expression data were obtained for 21,143 genes. The normalized read counts were utilized for principal-component and differential expression analyses using DESeq or further log-transformed as $\log_2(1 + x)$ for co-expression analysis.

Weighted Gene Co-expression Network Analysis

Weighted gene co-expression network analysis was performed using R package WGCNA (Langfelder and Horvath, 2008). The log-transformed data were used to generate a matrix of the Pearson correlations between all pairs of genes across the samples. A weighted correlation network was created by using the correlation coefficients with the power $\beta = 6$ if a coefficient is more than zero and otherwise the value is zero. The value $\beta = 6$ was chosen as a saturation level for a soft threshold of the correlation matrix, based on the criterion of approximate scale-free topology (Zhang and Horvath, 2005). To minimize effects of noise and spurious connections, the correlation matrix was transformed to the topological overlap matrix using a “TOMsimilarity” function implemented in the WGCNA package (Langfelder and Horvath, 2008). The topological overlap matrix was then used to group highly co-expressed genes by performing average linkage hierarchical clustering. The Dynamic Hybrid Tree Cut algorithm was used to cut the hierarchal clustering tree (Langfelder et al., 2008), and modules were defined as the branches resulting from this tree cutting. The module eigengene, the first principal component of the module, was used to further merge highly correlated modules (correlation coefficient > 0.8).

Guilt-by-Association Approach

The guilt-by-association approach was used to identify genes co-expressed with known genes (probes) for a specific function or expression pattern. To obtain a manageable number of probe genes, we selected probes fulfilling two criteria: (1) known from other work to have a common expression pattern and/or function and (2) highly co-regulated in our data (correlation co-efficient > 0.9). A list of the genes considered to be co-expressed with the probe genes was obtained by finding all genes having correlation coefficients with each of the probe genes that were greater than that of the least-similar pair of probe genes.

Gene Ontology Analysis

Functional annotation for the expressed gene list was performed with two annotation tools: the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 and protein analysis through evolutionary relationships (PANTHER) version 10.0. EASE Score, a modified Fisher Exact p value, was used to calculate enrichment p value in DAVID (Huang et al., 2009). Statistical overrepresentation test was performed to obtain enriched GO terms in PANTHER (Mi et al., 2013). To compare the two annotation tools in generating enriched GO terms, we used "GOTERM_BP_FAT" for DAVID and "GO-Slim Biological Process" for PANTHER as annotation datasets in Table S3. Representative GO terms in Table 1 are significant at $p < 0.05$ in both DAVID and PANTHER.

Comparison of Gene Prediction Using Existing Datasets

Two gene-prediction tools, SPELL (version 2.0.3) and WormNet (version 3) that allow multiple query genes, were used to compare candidate gene lists (Hibbs et al., 2007; Cho et al., 2014). For direct comparison, we obtained the same number of candidate genes that show high scores with the probe (query) genes used in this study. Previously known genes for semen proteins or synapse function were searched in WormBase (http://www.wormbase.org).

C. elegans Strains

N2 Bristol or CB4088 *him-5*(*e1490*) was used as the wild-type reference strain. The following mutant alleles were used in this study: LGI: *ckr-1*(*ok2502*); *snt-4*(*ok503*); *tom-1*(*ok285*); *T04D3.3*(*ok2924*); *W10C8.5*(*ok2723*); and *Y18D10A.6*(*ok549*), LGII: *C52E12.6*(*ok3724*) and *F57F10.1*(*ok368*), LGIII: *gmeb-3*(*ok3283*); *pros-1*(*ok903*); *unc-119*(*ed3*); *R13A5.9*(*ok3373*); and *T20H4.2*(*ok2547*), LGIV: *C17H12.9*(*ok1395*) and *Y73B6BL.19*(*ok1168*), and LGV: *ric-4*(*md1088*); *sul-2*(*gk187*); *ttn-1*(*ok1018*); and *T11A5.6*(*ok1866*). To examine defects in male tail ray morphology, four mutants, for *ceh-48*, *lst-5*, *gmeb-3*, and *T20H4.2*, were crossed four times into *him-5*(*e1490*) background. These mutants were further crossed with *bxIs14* (*pkd-2p::gfp*; Jia and Emmons, 2006) to visualize male-specific neurons, including RnBs, CEMs, and HOB.

Transgenic Strains

All transgenes were generated using a PCR fusion method (Hobert, 2002). For a transcriptional reporter, a genomic fragment containing the upstream sequence of a gene (~1,000–2,000 bp) was PCR amplified from N2 worms and then fused to GFP or sfGFP. For translational reporters of *ins-31* and *F59B2.12*, a genomic fragment containing upstream sequence and coding region was PCR amplified and fused to GFP or sfGFP at its C terminus. The resulting PCR products were injected into *him-5*(*e1490*) worms at 50~100 ng/μL with co-injection marker pRF4 (*rol-6*(*su1006*)) or *ttx-3p::gfp* to generate the following strains: *bxEx171* (*C29F5.3p::gfp*; amplified region [bp]: −1,102~−4); *bxEx172* (*snt-3p::gfp*; −2,259~−6); *bxEx173* (*dpf-1p::gfp*; −2,058~−6); *bxEx174* (*sul-2p::gfp*; −2,021~−4); *bxEx175* (*DH11.5p::gfp*; −2,005~−1); *bxEx176* (*F27C1.11p::gfp*; −893~−1); *bxEx177* (*M117.1p::gfp*; −2,096~−1); *bxEx178* (*T19A6.4p::gfp*; −999~−1); *bxEx179* (*W10C8.5p::gfp*; −2,065~−1); *bxEx180* (*cka-2p::gfp*; −2,049~−1); *bxEx189* (*clec-64p::gfp*; −873~1,419); *bxEx190* (*clec-207p::gfp*; −1,026~−4); *bxEx191* (*clec-199p::gfp*; −854~−16); *bxEx192* (*clec-219p::gfp*; −2,807~−9); *bxEx193* (*B0207.5p::gfp*; −874~1,265); *bxEx198* (*ins-31p::sfgfp*; −2,035~−1); *bxEx199* (*F59B2.12p::gfp*; −993~−1); *bxEx200* (*ins-31p::ins-31::sfgfp*; −2,035~750); and *bxEx197* (*F59B2.12p::F59B2.12::sfgfp*; −993~3,247).

Microscopy

Worms were mounted on 5% agar pads on glass slides using 10 mM sodium azide (Shaham, 2005). Worms were observed with Nomarski or fluorescence microscopy (Zeiss Axio Imager.A1 or Z2), and images were acquired using a camera (AxioCam; Zeiss) and processed using AxioVision (Zeiss). Figures were prepared using ImageJ software.

Aldicarb-Sensitivity Assay

Aldicarb (1 mM; Sigma-Aldrich) assays were performed using young adult hermaphrodites at least three times as described (Mahoney et al., 2006).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Steven Cook, Hannes Bülow, and Nelson Ramirez-Suarez for comments on the manuscript. Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). Support was from NIH (R01 GM066897) and the G. Harold & Leila Y. Mathers Charitable Foundation.

References

Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. [PubMed: 20979621]

Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31:166–169. [PubMed: 25260700]

Avila FW, Sirot LK, LaFlamme BA, Rubinstein CD, Wolfner MF. Insect seminal fluid proteins: identification and function. Annu Rev Entomol. 2011; 56:21–40. [PubMed: 20868282]

Bayés A, Grant SG. Neuroproteomics: understanding the molecular organization and complexity of the brain. Nat Rev Neurosci. 2009; 10:635–646. [PubMed: 19693028]

Blacque OE, Perens EA, Boroevich KA, Inglis PN, Li C, Warner A, Khattra J, Holt RA, Ou G, Mah AK, et al. Functional genomics of the cilium, a sensory organelle. Curr Biol. 2005; 15:935–941. [PubMed: 15916950]

Brenner S. The genetics of *Caenorhabditis elegans*. Genetics. 1974; 77:71–94. [PubMed: 4366476]

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. [PubMed: 24813450]

D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics. 2000; 16:707–726. [PubMed: 11099257]

Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR, Ruvkun G, Thomas JH, Swoboda P. Analysis of xbx genes in *C. elegans*. Development. 2005; 132:1923–1934. [PubMed: 15790967]

Efimenko E, Blacque OE, Ou G, Haycraft CJ, Yoder BK, Scholey JM, Leroux MR, Swoboda P. Caenorhabditis elegans DYF-2, an orthologue of human WDR19, is a component of the intraflagellar transport machinery in sensory cilia. Mol Biol Cell. 2006; 17:4801–4811. [PubMed: 16957054]

Ellegren H, Parsch J. The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet. 2007; 8:689–698. [PubMed: 17680007]

Emmons SW. The development of sexual dimorphism: studies of the *Caenorhabditis elegans* male. Wiley Interdiscip Rev Dev Biol. 2014; 3:239–262. [PubMed: 25262817]

Ferreira HB, Zhang Y, Zhao C, Emmons SW. Patterning of *Caenorhabditis elegans* posterior structures by the Abdominal-B homolog, *egl-5*. Dev Biol. 1999; 207:215–228. [PubMed: 10049576]

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. modENCODE Consortium. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science. 2010; 330:1775–1787. [PubMed: 21177976]

Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG. Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics. 2007; 23:2692–2699. [PubMed: 17724061]

Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. Genome Res. 2009; 19:657–666. [PubMed: 19181841]

Hobert O. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*. Biotechniques. 2002; 32:728–730. [PubMed: 11962590]

Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4:44–57. [PubMed: 19131956]

Jarrell TA, Wang Y, Bloniarz AE, Brittin CA, Xu M, Thomson JN, Albertson DG, Hall DH, Emmons SW. The connectome of a decision-making neural network. Science. 2012; 337:437–444. [PubMed: 22837521]

Jia L, Emmons SW. Genes that control ray sensory neuron axon development in the *Caenorhabditis elegans* male. Genetics. 2006; 173:1241–1258. [PubMed: 16624900]

Jiang M, Ryu J, Kiraly M, Duke K, Reinke V, Kim SK. Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. Proc Natl Acad Sci USA. 2001; 98:218–223. [PubMed: 11134517]

Jin, Y. Synaptogenesis (WormBook). 2005. http://dx.doi.org/10.1895/wormbook.1.44.1

Jin Y, Garner CC. Molecular mechanisms of presynaptic differentiation. Annu Rev Cell Dev Biol. 2008; 24:237–262. [PubMed: 18588488]

Johnstone IL. Cuticle collagen genes. Expression in *Caenorhabditis elegans*. Trends Genet. 2000; 16:21–27. [PubMed: 10637627]

Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. A gene expression map for *Caenorhabditis elegans*. Science. 2001; 293:2087–2092. [PubMed: 11557892]

Kulkarni G, Li H, Wadsworth WG. CLEC-38, a transmembrane protein with C-type lectin-like domains, negatively regulates UNC-40-mediated axon outgrowth and promotes presynaptic development in *Caenorhabditis elegans*. J Neurosci. 2008; 28:4541–4550. [PubMed: 18434533]

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559. [PubMed: 19114008]

Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2008; 24:719–720. [PubMed: 18024473]

Lints R, Emmons SW. Regulation of sex-specific differentiation and mating behavior in *C. elegans* by a new member of the DM domain transcription factor family. Genes Dev. 2002; 16:2390–2402. [PubMed: 12231628]

Mahoney TR, Luo S, Nonet ML. Analysis of synaptic transmission in *Caenorhabditis elegans* using an aldicarb-sensitivity assay. Nat Protoc. 2006; 1:1772–1777. [PubMed: 17487159]

Mason DA, Rabinowitz JS, Portman DS. *dmd-3*, a doublesex-related gene regulated by *tra-1*, governs sex-specific morphogenesis in *C. elegans*. Development. 2008; 135:2373–2382. [PubMed: 18550714]

Matson CK, Zarkower D. Sex and thesingular DM domain: insights into sexual regulation, evolution and plasticity. Nat Rev Genet. 2012; 13:163–174. [PubMed: 22310892]

Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013; 8:1551–1566. [PubMed: 23868073]

Miersch C, Döring F. Sex differences in carbohydrate metabolism are linked to gene expression in *Caenorhabditis elegans*. PLoS ONE. 2012; 7:e44748. [PubMed: 22984551]

Mounsey A, Bauer P, Hope IA. Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. Genome Res. 2002; 12:770–775. [PubMed: 11997343]

Mukhopadhyay A, Deplancke B, Walhout AJ, Tissenbaum HA. *C. elegans* tubby regulates life span and fat storage by two independent mechanisms. Cell Metab. 2005; 2:35–42. [PubMed: 16054097]

O'Rourke D, Baban D, Demidova M, Mott R, Hodgkin J. Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of *C. elegans* with *M. nematophilum*. Genome Res. 2006; 16:1005–1016. [PubMed: 16809667]

Ou G, Koga M, Blacque OE, Murayama T, Ohshima Y, Schafer JC, Li C, Yoder BK, Leroux MR, Scholey JM. Sensory ciliogenesis in *Caenorhabditis elegans*: assignment of IFT components into distinct modules based on transport and phenotypic profiles. Mol Biol Cell. 2007; 18:1554–1569. [PubMed: 17314406]

Page, AP., Johnstone, IL. The cuticle. WormBook; 2007. p. 1-15.http://dx.doi.org/10.1895/wormbook.1.138.1

Palopoli MF, Rockman MV, TinMaung A, Ramsay C, Curwen S, Aduna A, Laurita J, Kruglyak L. Molecular basis of the copulatory plug polymorphism in Caenorhabditis elegans. Nature. 2008; 454:1019–1022. [PubMed: 18633349]

Raymond CS, Shamu CE, Shen MM, Seifert KJ, Hirsch B, Hodgkin J, Zarkower D. Evidence for evolutionary conservation of sex-determining genes. Nature. 1998; 391:691–695. [PubMed: 9490411]

Reinke V, Smith HE, Nance J, Wang J, Van Doren C, Begley R, Jones SJ, Davis EB, Scherer S, Ward S, Kim SK. A global profile of germline gene expression in *C. elegans*. Mol Cell. 2000; 6:605–616. [PubMed: 11030340]

Reinke V, Gil IS, Ward S, Kazmer K. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. Development. 2004; 131:311–323. [PubMed: 14668411]

Reinke, V., Krause, M., Okkema, P. Transcriptional regulation of gene expression in *C. elegans*; WormBook. 2013. p. 1-34.http://dx.doi.org/10.1895/wormbook.1.45.2

Richmond, J. Synaptic function; WormBook. 2005. p. 1-14.http://dx.doi.org/10.1895/wormbook.1.69.1

Shaham, S. Methods in Cell Biology. WormBook. 2005. http://dx.doi.org/10.1895/wormbook.1.49.1

Shan G, Walthall WW. Copulation in *C. elegans* males requires a nuclear hormone receptor. Dev Biol. 2008; 322:11–20. [PubMed: 18652814]

Sieburth D, Ch'ng Q, Dybbs M, Tavazoie M, Kennedy S, Wang D, Dupuy D, Rual JF, Hill DE, Vidal M, et al. Systematic analysis of genes required for synapse structure and function. Nature. 2005; 436:510–517. [PubMed: 16049479]

Smith JR, Stanfield GM. TRY-5 is a sperm-activating protease in *Caenorhabditis elegans* seminal fluid. PLoS Genet. 2011; 7:e1002375. [PubMed: 22125495]

Snoek LB, Sterken MG, Volkers RJ, Klatter M, Bosman KJ, Bevers RP, Riksen JA, Smant G, Cossins AR, Kammenga JE. A rapid and massive gene expression shift marking adolescent transition in *C. elegans*. Sci Rep. 2014; 4:3912. [PubMed: 24468752]

Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT, Widmer C, Jo J, et al. A spatial and temporal map of *C. elegans* gene expression. Genome Res. 2011; 21:325–341. [PubMed: 21177967]

Südhof TC, Rizo J. Synaptic vesicle exocytosis. Cold Spring Harb Perspect Biol. 2011; 3:a005637. [PubMed: 22026965]

Sulston JE, Albertson DG, Thomson JN. The *Caenorhabditis elegans* male: postembryonic development of nongonadal structures. Dev Biol. 1980; 78:542–576. [PubMed: 7409314]

Thoemke K, Yi W, Ross JM, Kim S, Reinke V, Zarkower D. Genome-wide analysis of sex-enriched gene expression during *C. elegans* larval development. Dev Biol. 2005; 284:500–508. [PubMed: 15987632]

White JG, Southgate E, Thomson JN, Brenner S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. Philos Trans R Soc Lond B Biol Sci. 1986; 314:1–340. [PubMed: 22462104]

Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26:873–881. [PubMed: 20147302]

Yi W, Zarkower D. Similarity of DNA binding and transcriptional regulation by *Caenorhabditis elegans* MAB-3 and *Drosophila melanogaster* DSX suggests conservation of sex determining mechanisms. Development. 1999; 126:873–881. [PubMed: 9927589]

Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005; 4:e17.

Highlights

- Transcriptomes of *C. elegans* males and hermaphrodites are compared
- Comparison reveals 1,751 genes expressed at a higher level in males
- The male-upregulated set of genes is enriched for previously unannotated genes
- Functional gene clusters defined by gene co-expression analysis are validated

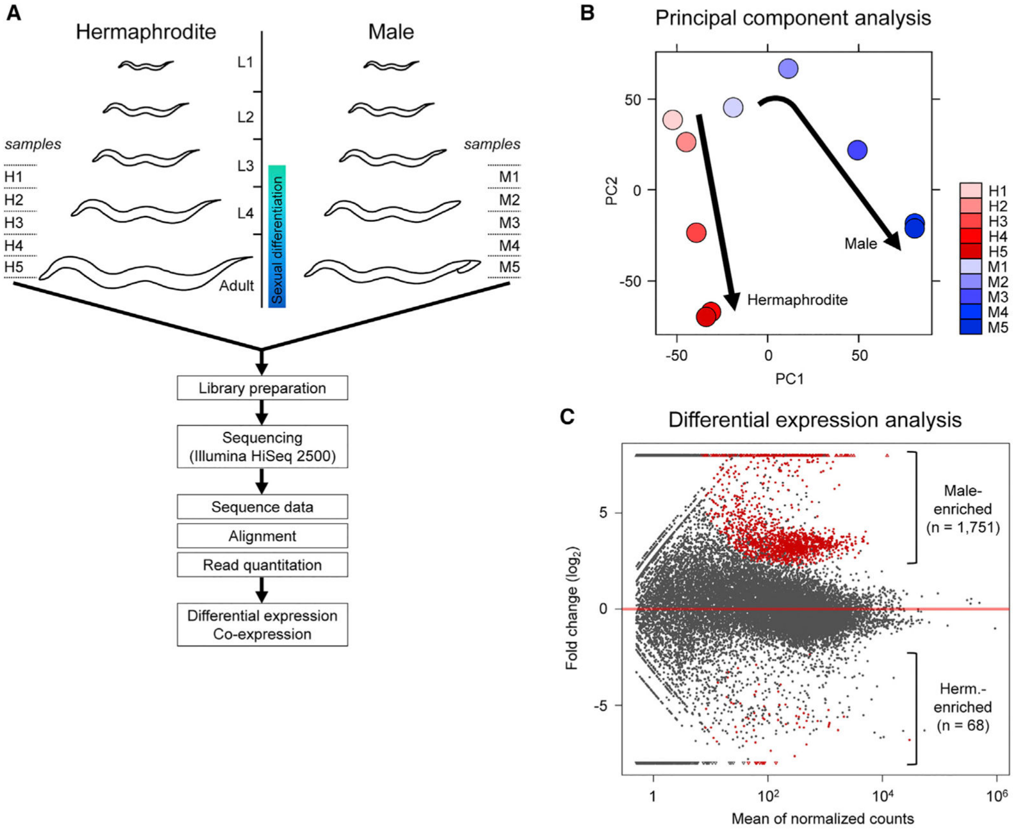
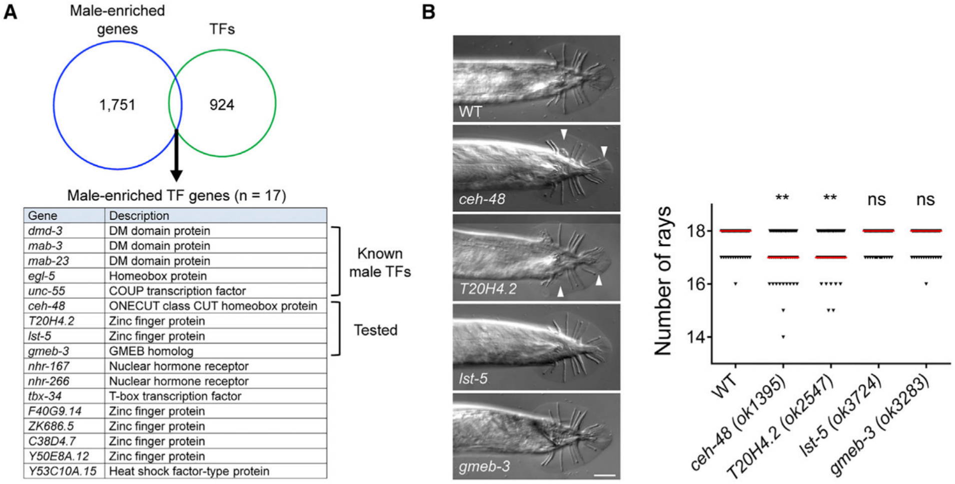


Figure 1. Developmental Transcriptome of the Two Sexes in *C. elegans*

(A) Schematic of RNA-seq procedure for obtaining whole-animal transcriptomes from late L3 to young adult stages for hermaphrodite (samples H1–H5) and male (samples M1–M5). For general morphologies of sampled worms, see Figure S1.

(B) Principal-component analysis of expression patterns for the five developmental time points of the two sexes. The progression of developmental stages is shown as an arrow for each sex.

(C) Scatterplot for expression of 21,143 genes showing differentially expressed genes. For this analysis, the five time point samples are treated as replicates (n = 5). Red dots indicate statistical significance at a false discovery rate < 1%. This is expected to be an underestimate of the actual number of differentially expressed genes, as it treats developmental changes as measurement noise, reducing the statistical significance of differences between the sexes.



Gene	Description
dmd-3	DM domain protein
mab-3	DM domain protein
mab-23	DM domain protein
egl-5	Homeobox protein
unc-55	COUP transcription factor
ceh-48	ONECUT class CUT homeobox protein
T20H4.2	Zinc finger protein
lst-5	Zinc finger protein
gmeb-3	GMEB homolog
nhr-167	Nuclear hormone receptor
nhr-266	Nuclear hormone receptor
tbx-34	T-box transcription factor
F40G9.14	Zinc finger protein
ZK686.5	Zinc finger protein
C38D4.7	Zinc finger protein
Y50E8A.12	Zinc finger protein
Y53C10A.15	Heat shock factor-type protein

Figure 2. Identification of Putative Transcription Factors Regulating Male Differentiation

(A) List of male-enriched transcription factors (n = 17) identified by using male-enriched genes from the RNA-seq data and known transcription factors in *C. elegans* (Reinke et al., 2013). Five transcription factors already known to act in male development and four genes tested using mutants in this study are indicated.

(B) Morphology of rays in four mutant males. Defects in ray morphology, including missing and fusion of rays, are indicated (arrowhead). Number of rays in mutants was counted and compared to wild-type (WT) (n = 50). Expected normal number of rays is 18. Red line represents the median. **p < 0.01; ns, not significant (by Mann-Whitney test). The scale bar represents 20 μm.

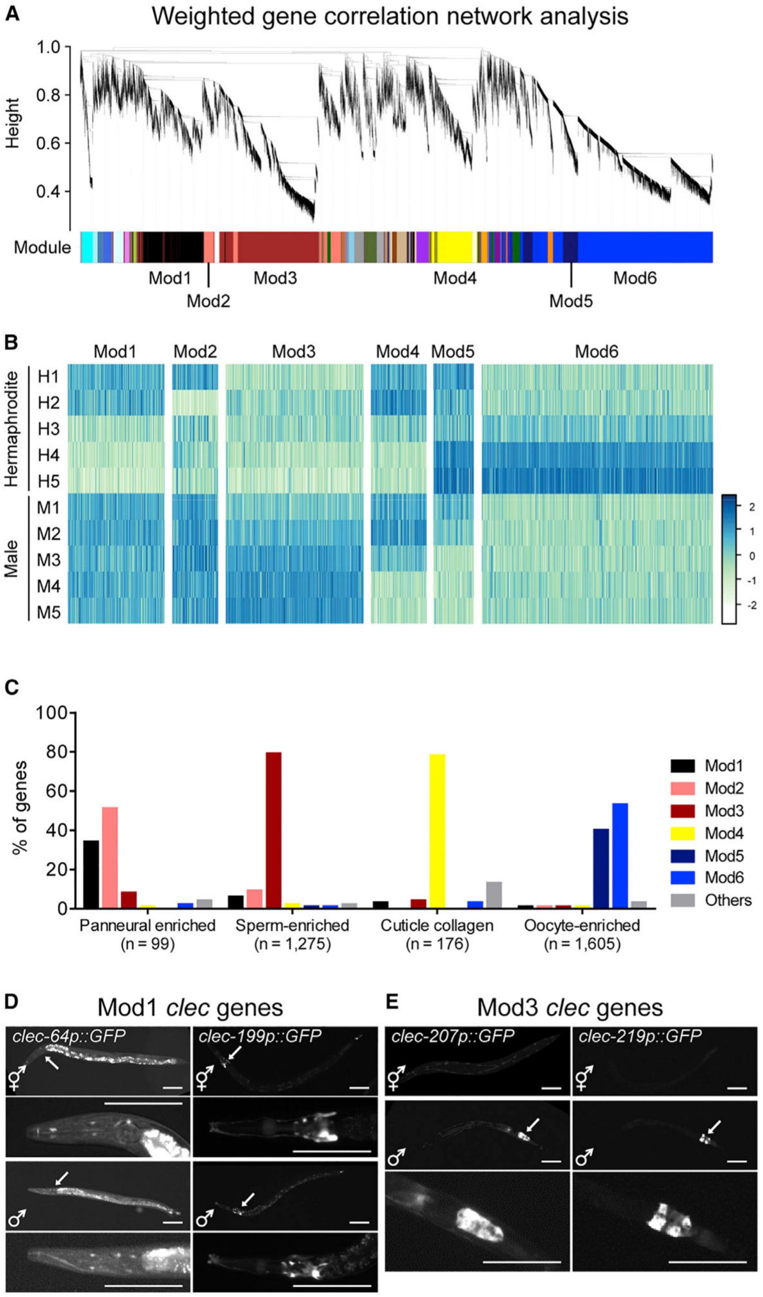


Figure 3. Gene Correlation Network Analysis

(A) Hierarchical cluster dendrogram groups 21,143 genes into distinct co-expression modules identified using weighted gene correlation network analysis. Six major modules are indicated as colored boxes (Mod1, black; Mod2, salmon; Mod3, brown; Mod4, yellow; Mod5, midnight blue; Mod6, blue).

(B) Heatmap showing relative expression of genes in the six major modules across ten samples of two sexes. The color values are the standardized *Z* scores of gene expression.

(C) Percentage of genes that are previously known to be enriched in specific cells or tissues across different modules. Previously known gene lists include for panneural enriched

(Spencer et al., 2011), sperm and oocyte enriched (Reinke et al., 2004), and cuticle collagen enriched (Page and Johnstone, 2007). Number of genes was normalized in order to avoid the effect of different module sizes (y axis).

(D and E) Tissue-specific expression of reporter transgenes for two C-type lectin domain (*clec*) genes chosen from Mod1 (D) and Mod3 (E) in both sexes. Arrow indicates expression in head neurons in (D) or in vas deferens of male in (E). The magnified view of the arrowed region is shown at the bottom of each image. In all images, anterior is left and posterior is right. The scale bars represent 100 μm.

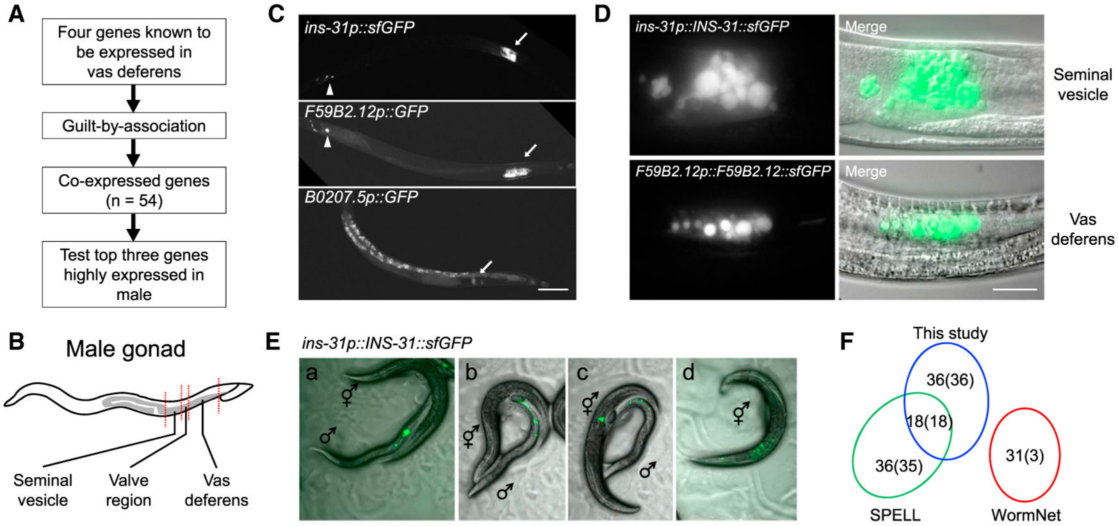


Figure 4. Identification of Semen Protein Genes

(A) Schematic of procedure to identify and validate semen protein genes from the RNA-seq data.

(B) The anatomical location of seminal vesicle, valve region, and vas deferens in male gonad (gray).

(C) Tissue-specific expression of reporter transgenes for three semen candidate genes in male. Arrows indicate expression in seminal vesicle (*ins-31*), vas deferens (*F59B2.12*), and the valve region of vas deferens (*B0207.5*). Arrowhead indicates signal from co-injection marker *ttx-3::GFP*. In all images, anterior is left and posterior is right.

(D) Protein localization identified using translational reporters for *ins-31* and *F59B2.12*. The Nomarski images show localization within the vesicular structures in the seminal vesicle region for INS-31 and in vas deferens for F59B2.12 protein.

(E) Time course images showing transfer of INS-31 from male to hermaphrodite during mating. To facilitate observation of mating behavior, slower-moving *unc-119* mutant hermaphrodites were used. INS-31::sfGFP (green) is visible within the seminal vesicle of a male before spicule insertion (a) and moves into the lumen of vas deferens after insertion until ejaculation (b). During ejaculation, INS-31::sfGFP is transferred into the vulva region of the hermaphrodite (c) and remains diffused in the uterus after mating (d).

(F) A Venn diagram showing the comparison of semen candidate gene lists obtained from this study, SPELL, and WormNet. The numbers of candidate genes and genes belonging to Mod3 (parentheses) are indicated.

The scale bars represent 100 μm in (C) and 20 μm in (D).

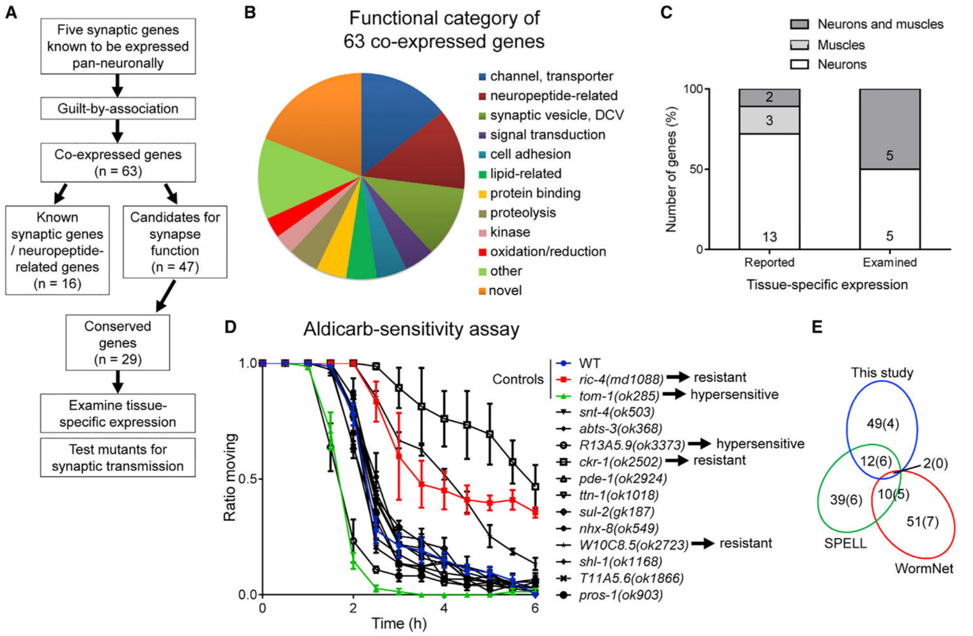


Figure 5. Identification of Putative Genes for Synapse Function

(A) Schematic of procedure to identify and validate synaptic genes from the RNA-seq data.

(B) Functional categories based on gene ontology terms and domain structures of the 63 genes co-expressed with known synaptic genes.

(C) Summary of tissue-specific expression patterns for 28 conserved genes. The expression pattern of reported genes was based on gene expression annotation from WormBase (http://www.wormbase.org). The examined gene expression was identified using the promoter-GFP fusion reporter transgenes.

(D) The time course paralysis response to cholinesterase inhibitor aldicarb (1 mM) of 12 mutant worms. A WT (blue), a known resistant strain (*ric-4*, red), and a known hypersensitive strain (*tom-1*, green) served as controls. The strains tested are indicated on the right. Two resistant strains (*ckr-1*(*ok2502*) and *W10C8.5*(*ok2723*)) and one hypersensitive strain (*R13A5.9*(*ok3373*)) were identified. Error bar represents SEM.

(E) A Venn diagram showing the comparison of synaptic candidate gene lists obtained from this study, SPELL, and WormNet. The numbers of candidate genes and previously known synaptic genes (parentheses) are indicated. Two synaptotagmin homolog genes shared in the three datasets (*snt-4* and *DH11.5*) have not yet been studied but are likely candidates for synapse function.

Table 1

Representative Gene Ontology Terms, Level of Expression, and Number of Sex-Enriched Genes in WGCNA Modules

			Median Expression Level (Interquartile Range)[b]		No. of Sex-Enriched Genes[c]	
Module	No. of Genes	GO Term (Fold Enrichment)[a]	Male	Herm.	Male	Herm.
Mod1	2,462	Synaptic transmission (3.5)	600 (199–1,737)	316 (102–1,022)	26	
		Ion transport (2.4)				
		Muscle development (3.6)				
Mod2	1,166	GPCR signaling (3.1)	126 (59–315)	31 (14–89)	96	
		Cilium (7.5)				
		Synapse (2.5)				
Mod3	3,467	Sperm development (8.9)	864 (185–2,675)	109 (12–392)	1,625	
		Phosphorylation (2.8)				
		Neuropeptide signaling (2.6)				
Mod4	1,421	Cuticle (7.4)	578 (143–1,933)	492 (134–1,714)		
		Molting cycle (2.8)				
		Cell adhesion (3.5)				
Mod5	1,016	Ribosome (15.8)	1,354 (250–5,336)	2,289 (520–8,737)		8
		Growth (2.6)				
		Mitochondrion (4.4)				
Mod6	5,747	Cell cycle (2.4)	993 (76–2,676)	2,028 (280–5,033)		59
		RNA processing (2.2)				
		Transcription (2.6)				
Others	5,864		23 (8–337)	16 (3–310)	4	1

[a] Representative GO terms that are significant at $p < 0.05$ using both DAVID and PANTHER are shown. Fold enrichment scores from DAVID are shown in parentheses.

[b] Expression level is indicated as a median of normalized count reads.

[c] Sex-enriched genes are differentially expressed genes in males ($n = 1,751$) and hermaphrodites ($n = 68$) shown in Figure 1C.