A single network comprising the majority of genes accurately predicts the phenotypic effects of gene perturbation in *C. elegans*

Insuk Lee, Ben Lehner, Catriona Crombie, Wendy Wong, Andrew G. Fraser, and Edward M. Marcotte

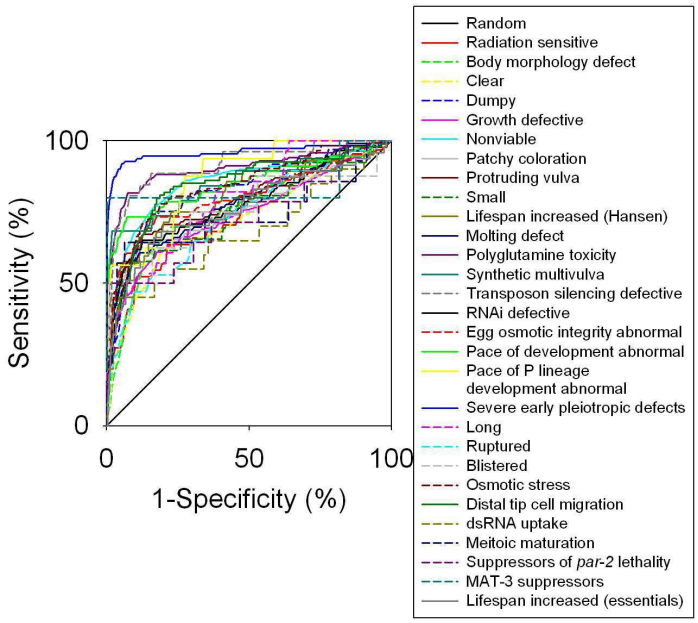
SUPPLEMENTARY FIGURE S1

ROC plots illustrating the Wormnet-based prediction of RNAi phenotypes. For each known phenotype, we analyzed the ability to predict genes conferring the phenotype using leave-one-out analysis. Every gene in the Wormnet was first rank-ordered by the sum of its *LLS* scores to all other genes with the given RNAi phenotype; we then measured the recovery of genes with the given phenotype, calculating true positive rate (TP/(TP+FN)) and false positive rate (FP/(FP+TN)) as a function of rank. In each plot, the diagonal represents no predictive power, curves above the diagonal indicate prediction of the plotted phenotype, with curves farther to the top left of the plot indicating the strongest predictive power. In order to measure rates up to 100%, we employed pseudocounts, assigning a very low *LLS* score (0.00000001) to all unlinked gene pairs in Wormnet (*i.e.*, gene pairs lacking all evidence for functional coupling).

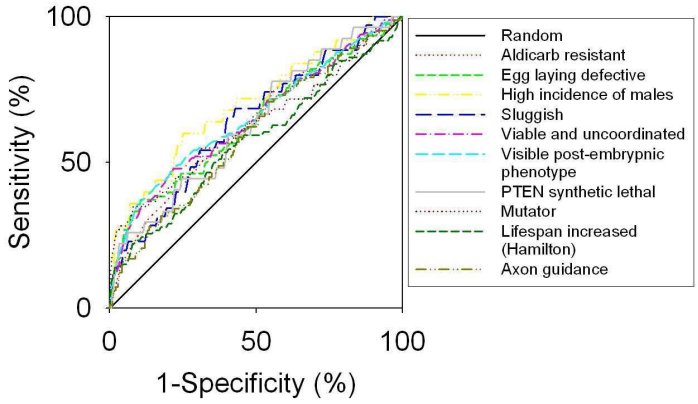
We tested 43 published RNAi phenotypes (see **Supplementary Methods, Table C**), omitting 1 phenotype with counts too low to provide statistical significance (egg size abnormal, 4 genes). Among the 43 tested phenotypes, we found (A) 29 strongly predictable phenotypes, (B) 10 moderately or weakly predictable phenotypes, and (C) 4 predictable at no better than random levels. Strong prediction of phenotypic outcomes indicates that genes sharing the same RNAi phenotype are tightly linked in Wormnet and are considerably closer to each other in the network than to other genes.

Supplementary Figure S1

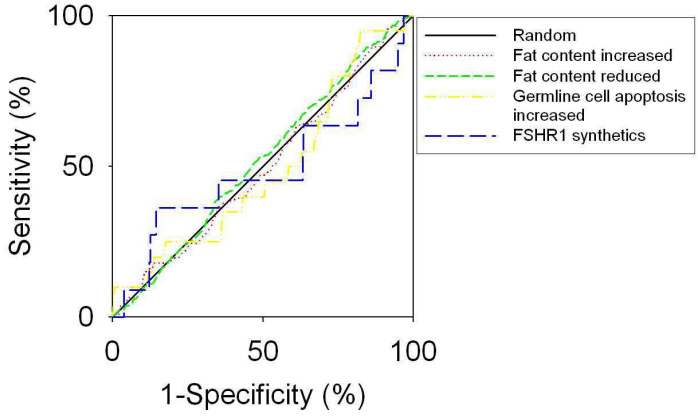
A



B



C



SUPPLEMENTARY TABLE S1

Table S1A

C. elegans DNA microarray mRNA expression data sets analyzed for co-expression, downloaded from the Stanford Microarray Database and Stuart *et al*. [11]. Six subsets of SMD and the non-redundant Stuart *et al.* set showed strong correlations between mRNA co-expression and log likelihood scores (see **Supplementary Methods, Figure A-2**) and were therefore incorporated into the network.

Array group	**Literature sources**	**# experiments**
SMD Aging	Lund J, *et al*. [5]	26
SMD Dauer	Wang, J. and Kim, SK [12]	50
SMD L1	Wang, J. and Kim, SK [12]	44
SMD Developmental stages	Jiang M, *et al*. [2]	26
SMD Germline	Reinke V *et al.* [8]	34
SMD Heat shock	Romagnolo B, *et al.* [9]	40
Stuart nonredundant	Stuart *et al*. [11]	635

Table S1B

C. elegans DNA microarray mRNA expression data sets tested but omitted from the network for insufficient correlation between mRNA co-expression and LLS scores (see **Supplementary Methods, Figure C**).

Array group	**Literature sources**	**# experiments**
GEO heat-stress	McCarroll SA, *et al.* [6]	7
SMD Alzheimer	Link CD, *et al*. [4]	9
SMD EDC treatment	Custodia N, *et al.* [1]	6
SMD ethanol treatment	Kwon JY, *et al.* [3]	7
SMD hypoxia	Shen C, *et al*. [10]	9
SMD sensory ray genes	Portman DS and Emmons SW [7]	7
SMD touch receptor neuron	Zhang Y, *et al.* [13]	6

References for Supplementary Table S1

1. Custodia, N., Won, S.J., Novillo, A., Wieland, M., Li, C., Callard, I.P. (2001) Caenorhabditis elegans as an environmental monitor using DNA microarray analysis. *Ann N Y Acad Sci* **948**: 32-42.
2. Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., Kim, S.K. (2001) Genome-wide analysis of developmental and sex-regulated gene expression profiles in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* **98**(1): 218-23.
3. Kwon, J.Y., Hong, M., Choi, M.S., Kang, S., Duke, K., Kim, S., Lee, S., Lee, J. (2004) Ethanol-response genes and their regulation analyzed by a microarray and comparative genomic approach in the nematode Caenorhabditis elegans. *Genomics* **83**(4): 600-14.
4. Link, C.D., Taft, A., Kapulkin, V., Duke, K., Kim, S., Fei, Q., Wood, D.E., Sahagan, B.G. (2003) Gene expression analysis in a transgenic Caenorhabditis elegans Alzheimer's disease model. *Neurobiol Aging* **24**(3): 397-413.
5. Lund, J., Tedesco, P., Duke, K., Wang, J., Kim, S.K., Johnson, T.E. (2002) Transcriptional profile of aging in C. elegans. *Curr Biol* **12**(18): 1566-73.
6. McCarroll, S.A., Murphy, C.T., Zou, S., Pletcher, S.D., Chin, C.S., Jan, Y.N., Kenyon, C., Bargmann, C.I., Li, H. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet* **36**(2): 197-204.
7. Portman, D.S. Emmons, S.W. (2004) Identification of C. elegans sensory ray genes using whole-genome expression profiling. *Dev Biol* **270**(2): 499-512.
8. Reinke, V., et al. (2000) A global profile of germline gene expression in C. elegans. *Mol Cell* **6**(3): 605-16.
9. Romagnolo, B., Jiang, M., Kiraly, M., Breton, C., Begley, R., Wang, J., Lund, J., Kim, S.K. (2002) Downstream targets of let-60 Ras in Caenorhabditis elegans. *Dev Biol* **247**(1): 127-36.
10. Shen, C., Nettleton, D., Jiang, M., Kim, S.K., Powell-Coffman, J.A. (2005) Roles of the HIF-1 hypoxia-inducible factor during hypoxia response in Caenorhabditis elegans. *J Biol Chem* **280**(21): 20580-8.
11. Stuart, J.M., Segal, E., Koller, D., Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**(5643): 249-55.
12. Wang, J. Kim, S.K. (2003) Global analysis of dauer gene expression in Caenorhabditis elegans. *Development* **130**(8): 1621-34.
13. Zhang, Y., Ma, C., Delohery, T., Nasipak, B., Foat, B.C., Bounoutas, A., Bussemaker, H.J., Kim, S.K., Chalfie, M. (2002) Identification of genes expressed in C. elegans touch receptor neurons. *Nature* **418**(6895): 331-5.

SUPPLEMENTARY TABLE S2

Table S2A
Benchmarking of protein physical or genetic interactions with log likelihood scores using reference gene pairs generated from Gene Ontology biological process annotation.

Data set	# unique genes	# unique gene pairs	Log likelihood score
WI5.literature[1]	131	102	1.76
WI5.scaffold	455	487	0.71
WI5.core1	723	809	0.26
WI5.core2	1114	1264	0.013
WI5.noncore	1823	1865	-0.49
Genetic interactions[2]	772	3663	1.15

1. WI5: Worm Interactome version 5 [7]
2. Collected from Worm base release WS150

Table S2B
Human interactome sets from which worm gene functional linkages were inferred.

Human interactome set	# unique genes	# unique gene pairs
Text mining (Bayesian-ranked co-citation) [9]	1,054	2,013
BIND [1]	1,024	1,572
BIOGRID [11]	2,076	7,079
HPRD [8]	2,689	14,909
Reactome [4]	1,152	22,125
Large-scale yeast 2 hybrid analysis [10]	2,998	6,085

Table S2C
Yeast functional genomics and proteomics data sets from which worm gene functional linkages were inferred. Yeast linkages derive from version 2 of the network in [6], with additional datasets incorporated where cited below.

Yeast data set	# unique genes	# unique gene pairs
Text-mining (by co-citation)	2,111	17,493
mRNA co-expression	1,831	45,252
Gene neighbors	1,301	7,128
Genetic interactions	1,915	10,534
Text-mining (by literature curation)	1,467	7,007
Affinity purification followed by mass spec analysis [3, 5]	1,691	26,153
Rosetta Stone proteins	560	793
Predicted interactions by protein tertiary structures [2]	672	4,201

References for Supplementary Table S2

1. Alfarano, C., et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **33**(Database issue): D418-24.
2. Aloy, P. Russell, R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* **19**(1): 161-2.
3. Gavin, A.C., et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**(7084): 631-6.
4. Joshi-Tope, G., et al. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33**(Database issue): D428-32.
5. Krogan, N.J., et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**(7084): 637-43.
6. Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M. (2004) A probabilistic functional network of yeast genes. *Science* **306**(5701): 1555-8.
7. Li, S., et al. (2004) A map of the interactome network of the metazoan C. elegans. *Science* **303**(5657): 540-3.
8. Peri, S., et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**(10): 2363-71.
9. Ramani, A.K., Bunescu, R.C., Mooney, R.J., Marcotte, E.M. (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol* **6**(5): R40.
10. Rual, J.F., et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**(7062): 1173-8.
11. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**(Database issue): D535-9.

SUPPLEMENTARY TABLE S3

The final contribution of each of the nine data types to the (A) full and (B) core integrated *C. elegans* network, listing the number of unique genes and functional linkages derived from each set. Note that a given linkage may have evidence from more than one dataset. (C) Optimal values of *D* and *T* parameters for the data integration steps, indicating that the linkages derived from different expression datasets were reasonably independent, while substantial redundancy existed among the 9 datasets for the final integration step.

S3A. Complete network

Data set	# unique genes	# unique functional linkages
mRNA co-expression	14,491	287,130
Associalogs from yeast gene network	2,637	56,262
Interologs from human	3,145	30,098
Gene neighbors	2,660	13,645
Co-citation	1,300	5,577
Phylogenetic profiles	649	2,051
Genetic interactions from Wormbase	771	1,690
Worm Interactome version 5	1,165	1,411
Interologs from fly	1,141	910

S3B. Core network

Data set	# unique genes	# unique functional linkages
mRNA co-expression	9,769	64,498
Interologs from human	2,865	27,737
Associalogs from yeast gene network	2,177	22,825
Co-citation	1,288	4,252
Gene neighbors	1,132	3,584
Genetic interactions from Wormbase	771	1,690
Phylogenetic profiles	560	1,558
Interologs from fly	512	321
Worm Interactome version 5	324	232

S3C. Optimized *D* and *T* free parameters for each data integration step

Integrated set	D (relative dependence)	T (threshold of LLS of individual data set)
Co-expression network	1.1	0.182
Human interologs network	20.0	0
Yeast associalogs network	2.9	0.405
Wormnet (final integration)	$+\infty$	0

SUPPLEMENTARY METHODS

A single network comprising the majority of genes accurately predicts the phenotypic effects of gene perturbation in *C. elegans*

Insuk Lee, Ben Lehner, Catriona Crombie, Wendy Wong, Andrew G. Fraser, and Edward M. Marcotte

Caenorhabditis elegans proteome

This study is based on 19,735 predicted protein coding genes (excludes 2,685 alternative spliced products) of *C. elegans* (downloaded from WormBase Release WS140 (*1*) on March 2005). All linkages and calculations of genome coverage are based on this gene set.

Overview of *C. elegans* probabilistic functional gene network (Wormnet version 1.0) construction

Different types of functional and comparative genomics data have quite different value for reconstructing pathways in a metazoan. The different data sets are also often accompanied by distinct internal measures of confidence. To integrate these data into a composite network, we first evaluated each data set using a common scoring scheme, allowing the relative merits of each to be measured prior to integration weighted according to their scores. Specifically, using the log likelihood score (*LLS*) scheme described in (*2*), we estimated functional coupling between each pair of genes, defined as the likelihood of participating in the same pathway, then integrated the gene-gene linkages into the final network. The resulting network therefore represents a unified model of coupling between *C. elegans* genes as estimated from the currently available large-scale, predominantly systematically collected, data.

In this scheme, $LLS = \ln\left(\frac{P(L \mid E)/P(\neg L \mid E)}{P(L)/P(\neg L)}\right)$,

where P(L|E) and P(¬L|E) are the frequencies of linkages (L) observed in the given experiment (E) between annotated genes operating in the *same* pathway and in *different* pathways, respectively, while P(L) and P(¬L) represent the prior expectations (*i.e.*, the

total frequency of linkages between all annotated *C. elegans* genes operating in the *same* pathway and operating in *different* pathways, respectively). Scores greater than zero indicate the dataset tends to link genes in the same pathway, with higher scores indicating more confident linkages and stronger support for the genes operating in the same pathway.

To obtain accurate estimates of dataset accuracy, we employed 0.632 bootstrapping (*3*, *4*) for all *LLS* evaluations. 0.632 bootstrapping has been shown to provide a robust estimate of classifier accuracy, generally out-performing cross-validation (*5*), especially for very small datasets (e.g., see (*6*)). The data evaluation and integration strategy we describe is therefore appropriate even for more poorly annotated genomes. Unlike cross-validation, which uses sampling without replacement for constructing test and training datasets, 0.632 bootstrapping employs sampling with replacement, constructing the training set from data sampled with replacement and the test set from the remaining data that weren't sampled. Each linkage has a probability of 1-1/*n* of not being sampled, resulting in ~63.2% of the data in the training set and ~36.8% in the test set (*7*). The overall *LLS* is the weighted average of results on the two sets, equal to $0.632*LLS_{test} + (1-0.632)*LLS_{train}$, calculated as the average over 10 repeated sampling trials.

For data sets accompanied by intrinsic scores that are continuous (e.g., correlation coefficients between pairs of gene expression vectors), we ranked gene pairs by the scores and calculated log likelihood scores for bins of equal numbers of gene pairs. Those *LLS*s and the corresponding mean of the data intrinsic scores for each bin were used to derive regression models mapping the data intrinsic scores to *LLS* scores, in this manner generating *LLS* scores for both annotated and unannotated gene pairs (see **Figure A**). For integrating evidence from multiple data sets, we used a modification of the weighted sum method (*2*) described in ref. (*8*) to account both for differential quality of each data set and for correlations among the data sets. The weighted sum method was modified to include a parameter, *T*, representing a *LLS* threshold for all data sets being integrated. The total strength of a given functional gene linkage derived from multiple data sets was calculated as the weighted sum (*WS*) of individual scores as

$$WS = L_0 + \sum_{i=1}^{n} \frac{L_i}{D \cdot i} \text{ , for all } L \geq T,$$

where L_0 represents the best *LLS* score among all *LLSs* for that gene pair, *D* is a free parameter for the overall degree of independence among the data sets, and *i* is the order index of the data sets after rank-ordering the *n* remaining *LLS* scores for the given gene pair, starting from the second highest *LLS* score and descending in magnitude. The values of two free parameters (*D* and *T*) are chosen by systematically testing values of D and T in order to maximize overall performance (area under a plot of *LLS* versus gene pairs incorporated in the network) on the Gene Ontology benchmark, selecting a single value of *D* and of *T* for all gene pairs being integrated using these datasets. *D* is a free parameter determining the (linear) decay rate of the weight for secondary evidence. It ranges from 1 to $+\infty$ and captures the relative independence of the data sets, low values of *D* indicating more independence among data sets and higher values indicating less. As the optimal value of *D* approaches $+\infty$, the scheme is equivalent to taking only the single best line of evidence for a linkage (L_0), regardless of which data set it derives from, in this way avoiding overweighting linkages when datasets are highly correlated. (Note that this overweighting applies only to linkages supported by more than one dataset—links from only a single line of evidence are not affected by this scheme regardless of the value of *D*.) We independently explicitly test the performance of a *naïve* Bayesian integration of the *LLS* scores (here, simply the sum of the *LLS* scores for a given gene pair), then select the integration approach maximizing the area under a plot of *LLS* versus gene pairs incorporated in the network.

We first integrated similar classes of data into composite sets (integrating the 6 co-expression data sets into a single set of co-expression linkages, integrating the human PPI data, integrating the worm PPI data, and integrating the yeast-derived linkages), before then integrating the 9 composite sets (see **Table S3A-B**) into an overall network based upon co-citation, co-expression, the Worm Interactome version 5, genetic interactions, gene neighbors, phylogenetic profiles, and conserved interactions transferred from other species (yeast/fly/human). The optimized free parameters, *D* and *T*, for each integration steps are summarized in **Table S3C**.

We note that this approach minimizes the total number of free parameters (such as by not learning weights for all pairs of datasets), making this approach robust to overtraining. In all, <125 free parameters were trained for the reconstruction of the

complete set of several hundred thousand pairwise gene linkages in Wormnet from >20 million experimental observations.

The final network has a total of 384,700 linkages between 16,113 *C. elegans* proteins, covering ~82 % of *C. elegans* proteome; all gene pairs have a higher likelihood of belonging to the same pathway than random chance. To define a model with high confidence and reasonable proteome coverage, we applied a likelihood threshold, keeping only gene pairs linked with a likelihood of being in the same pathway of at least 1.5 fold better than random chance. Using this threshold, we defined the core network, containing 113,829 linkages for 12,357 worm proteins (~63% of the worm proteome).

Reference and benchmark sets

Three different reference annotation sets were used to assess the *C. elegans* functional linkages. The Gene Ontology (GO) annotation downloaded March 2005 from WormBase (*1*) served as the major reference set for training and benchmarking the network. The GO schema lists three hierarchies of function, describing "biological process" (i.e., pathways and systems), "molecular function" (i.e., biochemical activities), and "cellular component" (i.e., subcellular localization). For testing hypotheses of functional coupling between genes, we used the *C. elegans* GO "biological process" annotation, which contains up to 14 different levels of information under the term "biological process" within the hierarchy. We constructed a reference set consisting of gene pairs sharing GO biological process annotation. To optimize annotation specificity and comprehensiveness, we used terms belonging to levels 2 through 10 (terms above level 2 are too general, and terms below level 11 too specific). We sorted all terms by the number of genes annotated (**Figure B**), then excluded the top 5 terms, which account for >78% of total reference set gene pairs, in order to reduce functional bias in the benchmark set. The following terms were omitted: embryonic development, positive regulation of growth rate, growth, locomotory behavior, regulation of transcription (DNA-dependent). The Kyoto-based KEGG database (*9*) provides metabolic and regulatory pathway annotations that are closely related to biological process annotations. A KEGG map for *C. elegans* downloaded on November 2005 was used to generate the second benchmarking set for this study, excluding the 3 most abundant KEGG pathway

annotation terms (oxidative phosphorylation, purine metabolism, and ribosome; accounting for >40% of the linkages) in order to minimize bias. After the above post-processing, there are 786,056 gene pairs sharing annotation from the GO reference set, and 9,406 from KEGG. About half of KEGG gene pairs (5,069 pairs) are shared between the two reference sets. Therefore, the KEGG and GO reference sets for *C. elegans* are fairly independent (*10*). Nonetheless, to ensure complete independence, we removed all GO pairs from the KEGG set, then used the KEGG minus GO set as a 100% independent set for additional tests of the network as well as for comparison to two earlier integrated worm network models (*11, 12*). As an additional benchmark set, we also considered the set of gene pairs sharing GO "cellular component" annotations.

Finally, in order to effectively summarize broad trends of biological functions in the data set, we desire only a few categories of functional annotation. For this purpose, we employed a reference set of functional categories from the clusters of orthologous group (COG) annotation (*13*), which is based on reconstructing homologous groups of proteins in such a manner as to considerably enrich for orthologous proteins within each group, with the functions of genes assigned within 23 broad categories (such as "Transcription" and "Signal Transduction Mechanisms") based on the well-annotated proteins with each COG. We use the recently updated COG collection that includes multicellular eukaryotic genomes (named eukaryotic orthologous groups, or KOG) (*14*). These 23 categories were further collapsed into 12 functional groups for more efficient visualization (see **Table A**). We also constructed a benchmark set from gene pairs sharing KOG annotations.

Inferring gene functional linkages from mRNA expression data

Gene functional linkages were calculated from microarray data of mRNA expression as in (*2*). Expression data are from the Stanford Microarray Database (SMD downloaded on July 2005) (*15*), GEO database (*16*), and published by Stuart *et al*. (*11*). We established an objective criterion for considering DNA microarray datasets: For each collection of DNA microarray data (defined as a set of arrays listed in SMD as from one publication), we considered all gene pairs correlated at the 99% confidence level (by t-test), then tested for the evidence of a relationship between the Pearson correlation

coefficient (PCC) of pairs of genes' expression vectors and the LLS score. Only sets of experiments that showed a positive correlation were analyzed further. An example of a set meeting this criterion is shown in **Figure A-2**. The list of tested datasets meeting this criterion and therefore included in the network calculation is presented in Table S1A. Two examples of datasets failing this criterion are shown in **Figure C**; tested but excluded sets are now listed in **Table S1B**. By this criterion, we selected 6 sets of SMD data containing a total of 220 separate microarray experiments with significant correlations between co-expression and functional associations (see **Table S1A**), as well as the data from Stuart *et al.*. For the Stuart *et al.* data set, we considered only the subset of experiments non-redundant with those available independently from SMD, resulting in 635 separate microarray experiments (the specific experiments analyzed from the Stuart *et al.* dataset are those indicated in their dataset by numerical indices lacking descriptions). We occasionally observed genes with no significant expression dynamics across the experiments (*i.e.*, low variance expression vectors) showing high Pearson correlation coefficients and leading to spurious (not biologically meaningful) linkages. We therefore filtered out such cases by requiring each gene to exhibit significant (typically, >1.2-fold) expression changes in some minimal number of experiments, optimizing both the threshold of expression and minimum number of experiments for each group of expression datasets by recall-precision analysis, maximizing the area under a plot of LLS versus genes included in the network.

Gene functional linkages by physical and genetic interactions between proteins

We incorporated genome-wide yeast two hybrid analyses of *C. elegans* genes from the Worm Interactome database, as well as the published literature set of small scale protein-protein interactions (*17*). We treated subsets of the Worm Interactome Version 5 (literature, scaffold, core1, core2, non-core) separately, providing different confidence scores for the different data subsets, rather than a single averaged confidence score across all interactions of the Worm Interactome set. Genetic interactions (for ~800 genes and ~4000 interactions) were included from WormBase (*1*), derived from >1000 primary publications. Benchmarking of worm protein-protein interactions is summarized in **Table S2A**.

Inferring gene functional linkages from genome context

Functional linkages can be inferred between pairs of genes from comparative analyses of genome sequences. We find the methods of phylogenetic profiling (*18-20*) and gene neighbors (*21-23*) show reasonable performance for metazoan genes. Linkages for each method were derived from analysis of 149 genomes (117 bacteria, 16 archaea, and 16 eukaryotes). Briefly, each *C. elegans* protein sequence was compared to every other sequence using the program BLASTP with default settings (*24*), then the alignment scores analyzed as follows.

Phylogenetic profiles were constructed from these comparisons and analyzed as in (*25*) with the following modifications: We found the profiles derived from organisms of different kingdoms provided considerably different strengths of correlation with gene functional associations. Profiles calculated only from bacterial genomes provided the best range of *LLS* scores; including archaea or eukaryotes in the profiles did not significantly improve performance. Therefore, we inferred gene functional linkages from phylogenetic profiles constructed from only the 117 bacterial genomes. For discretizing BLASTP E-values during the calculation of mutual information between phylogenetic profiles, we employed bins of equal numbers of E-values, rather than equal intervals of E-values, accounting for non-uniform E-value distribution. In previous analyses of phylogenetic profiles, we have observed the best results (measured by recall-precision analysis using LLS scores and protein coverage as measures of precision and recall, respectively) from using 3 E-value bins, and therefore adopted this approach. Gene neighbor linkages were identified as in (*21*) using both bacterial and archaeal genomes.

Inferring gene functional linkages from literature mining

We also identified functional linkages by mining the scientific literature (specifically, Medline abstracts downloaded on December 2004) using the co-citation approach (*26, 27*). We analyzed a set of N = 7,732 Medline abstracts that included the word “elegans” in the abstract for perfect matches to either the systematic names or common names of 19,735 genes of *C. elegans*, scoring gene pairs according to the scheme of (*2*).

Functional linkages derived by transferring conserved gene interactions from other species' interactomes

The assayed subsets of different species' interactomes often complement one another due to differences in bait/prey choices and experimental sensitivity, specificity, and bias. By transferring linkages between species (*i.e.*, the functional linkage equivalent to 'interologs' (conserved physical protein interactions) (*28*), which we term 'associalogs'), we can collect additional gene functional linkages for a given genome. We therefore transferred both physical protein interactions and functional gene linkages from yeast, fly, and human into worm.

For proper identification of worm orthologs of those query genomes, we used INPARANOID (*29*), which reduces false negative ortholog identifications and has proved to be a robust method for identifying functionally equivalent proteins (*30*). Based on the worm and yeast ortholog pairs by INPARANOID, we inferred functional linkages between worm genes based upon linkages in the probabilistic functional gene network (version 2 (*8*) of the network described in (*2*)). We found that transferring linkage information from the individual yeast data sets prior to integration provided better performance (assessed by recall-precision analysis on the *C. elegans* benchmark) than direct transfer of the integrated yeast network linkages. We employed the modified weighted sum method of linkage integration (*2*) described above. Additional functional linkages were inferred from the fly yeast two hybrid network (*31*), as well as from several sets of human protein interaction data (see **Table S2B**). Linkages from the individual human protein interaction sets were first integrated using the weighted sum method into a single set of human-derived linkages before integrating with other datasets. Prior to this integration, individual interactions were either assigned confidence scores according to the hypergeometric probability of occurring at random given the total number of interactions of each partner, calculated as in (*32*), or were assigned single confidence scores for all linkages derived from a single type of experiment, choosing whichever scoring scheme performed better by recall-precision analysis.

Detailed protocol for reconstructing the *C. elegans* gene network

To more clearly define the procedure we employed for generating the network, we provide the full procedure as pseudo-code:

1. Identify worm orthologs of human, yeast, and fly proteins using INPARANOID
2. For worm DNA microarray data
    2.1. For each set of worm DNA microarrays (corresponding to all arrays from a given publication, as defined in SMD)
        2.1.1. Calculate the mean-centered Pearson correlation coefficient (PCC) between all pairs of genes' expression profiles
            2.1.1.1. Calculate (by t-test) the minimum correlation coefficient for 99% confidence given the # of experiments in the set. For further analyses, consider only pairs meeting this criterion.
            2.1.1.2. Evaluate the regression between PCC and the log likelihood score (LLS) of sharing pathway annotations
                2.1.1.2.1. Reject set if no relationship is evident between PCC and LLS
            2.1.1.3. Filter genes considered in the correlation analysis by requiring each gene to exhibit significant expression changes (e.g., >*x*-fold, typically ~1.2-fold) in some minimal # of experiments across the dataset. Optimize these 2 parameters by recall-precision analysis, maximizing the area under a plot of LLS versus # of genes participating in the linkages.
            2.1.1.4. Fit regression (typically sigmoidal) between PCC and LLS, considering only genes passing the optimized filtering criteria (2.1.1.3) and only gene pairs whose correlation exceeds the 99% confidence level (2.1.1.1).
            2.1.1.5. Using regression fit, assign LLS scores to all gene pairs whose correlation exceeds the 99% confidence level, including unannotated gene pairs.
            2.1.1.6. Select minimum LLS threshold from inflection point of regression model. Retain only LLS scores/gene pairs surpassing threshold.
    2.2. Integrate LLS scores from complete collection of sets of DNA microarrays
        2.2.1. Calculate the weighted sum of LLS scores for each gene pair across the analyses of DNA microarray sets
        2.2.2. Optimize the choice of the weighting parameters D and T using recall-precision analysis by maximizing the area under a plot of LLS versus # of genes participating in the linkages. Compare to *naïve* Bayesian integration, and choose from weighted integration versus *naïve* Bayes by recall-precision analysis.
3. For each set of worm protein-protein interaction (PPI) data or genetic interaction data
    3.1. Measure the LLS score for all pairs in the set
    3.2. Assign this LLS score to all interacting pairs in the set, including unannotated pairs
4. For human PPI data
    4.1. For each set of human PPI data, analyze PPI generated by each experimental or computational approach (e.g., yeast two-hybrid, text-mining, etc.) independently

4.1.1. Measure the LLS score for all worm gene pairs corresponding to interacting human proteins in the given data set using the given approach

4.2. Calculate the weighted sum of LLS scores for each gene pair across the sets of human PPI data, optimizing the choice of D and T parameters by recall-precision analysis as in (2.2). Compare to *naïve* Bayesian integration, and choose from weighted integration versus *naïve* Bayes by recall-precision analysis.

4.3. Fit regression between LLS and weighted sum (or *naïve* Bayes sum), then assign LLS scores to all worm gene pairs corresponding to interacting human proteins, including unannotated pairs

5. For worm co-citation, phylogenetic profiles, and gene neighbors data
   5.1. Fit regressions between LLS and data-intrinsic scores (–log(random probability of co-citation), mutual information of phylogenetic profiles, and –log(random probability of being gene neighbors, respectively)
   5.2. Using regression fit(s), assign LLS scores to all co-cited (or co-inherited or co-neighboring) gene pairs, including unannotated gene pairs
6. For fly PPI data
   6.1. Considering worm gene pairs corresponding to interacting fly proteins, fit regression between LLS and fly PPI confidence scores provided with fly PPIs
   6.2. Using regression fit, assign LLS scores to all worm gene pairs corresponding to interacting fly proteins, including unannotated pairs
7. For yeast functional network data
   7.1. Analyze each data type (e.g., DNA microarrays, affinity purification/mass spec, etc.) separately, considering worm gene pairs whose yeast orthologs are linked by the given data type.
      7.1.1. Fit regression between LLS for worm gene pairs and LLS associated with corresponding yeast gene pairs in the yeast network
      7.1.2. Using regression fit, assign LLS scores to all worm gene pairs corresponding to linked yeast genes, including unannotated pairs
   7.2. Integrate yeast-derived linkages by calculating the weighted sum of LLS scores for each gene pair across the set of yeast data types, optimizing the choice of D and T parameters by recall-precision analysis as in (2.2). Compare to *naïve* Bayesian integration, and choose from weighted integration versus *naïve* Bayes by recall-precision analysis.
   7.3. Fit regression between LLS and weighted sum (or *naïve* Bayes sum), then assign LLS scores to all worm gene pairs corresponding to linked yeast genes, including unannotated pairs
8. Integrate all linkages using the weighted sum method, optimizing the choice of D and T parameters by recall-precision analysis as in (2.2). Compare to *naïve* Bayesian integration, and choose from weighted integration versus *naïve* Bayes by recall-precision analysis.

Evaluation of the integrated network model

The final *C. elegans* network model has been assessed extensively using a variety of approaches. First, the accuracy of linkages in the integrated model was evaluated on the

GO annotation benchmark using 0.632 bootstrapping as described above and illustrated in **Figure 1A**. As expected, we observe the LLS scores of linkages to increase with the number of lines of supporting evidence (**Figure D**). The LLS scores decrease reasonably uniformly as a function of network size, as shown in **Figure E**. We observe the linkage scores to be reasonably robust even to large changes in the reference annotation set—for example, limiting the GO reference set to terms in levels 5 – 10, rather than 2 – 10, results in removal of ~36% of all reference set positive examples. However, a plot of recall versus precision using this reduced reference annotation set shows little change in performance from the larger set (**Figure F**).

Linkage quality was also evaluated on the KEGG annotation set, as described in **Figure 1D**. We also evaluated the network on the KEGG minus GO set, which represents a small, highly biased, but 100% independent subset of reference linkages. This benchmark should be considered to provide a lower bound of accuracy, as all high-confidence linkages confirmed by GO have been removed. Nonetheless, the ranking of network accuracies in **Figure 1D** is generally preserved using this reference (**Figure G**). **Figures H-1 to 5 and Q** present comparisons of Wormnet with 4 previous *C. elegans* gene networks on 5 different annotation sets, including KEGG annotations, KOG annotations, GO cellular component annotations, GO biological process annotations with terms related to protein synthesis removed, and the most recent set of GO biological process annotations. In each test, Wormnet shows considerably increased recall of both genes and linkages, while maintaining an accuracy comparable to the other networks.

Next, we examined functional clusters in the network, with the notion that genes of the same pathway should generally cluster strongly in the overall network. Genes of the core network were clustered by their connectivity and modules were defined as in (*2*), requiring each module to contain at least 3 member genes. In total, we defined 402 modules (median size 8 genes) covering 8,195 *C. elegans* proteins (~42 % of *C. elegans* proteome). The functional coherence of genes in the same module was evaluated as in (*2*). **Figure I** illustrates the high functional coherence of genes in the modules using 12 collapsed KOG gene functional categories (**Table A**; represented by different color codes). The core network is therefore a reliable model of functional associations among

C. elegans genes giving rise to biologically reasonable and functionally coherent estimates of higher level pathway organization in *C. elegans*.

The observation that many RNAi phenotypes are strongly predictable from Wormnet v1 suggests that genes exhibiting similar RNAi phenotypes are also clustered in the network. We explicitly tested this notion for phenotypes of differing specificity. We first examined how genes exhibiting 3 broad categories of RNAi phenotypes (nonviable, growth defective, visible post-embryonic phenotypes (*33*)) were distributed across the modules. Of the 402 modules, 91 showed strong clustering of phenotypes, far greater than random (Z-score = 7.04; $p < 10^{-11}$), with >=25% of the genes in each module sharing a particular broad phenotype, and the remaining modules being dominated by genes with no visible loss-of-function phenotype. The modules that we identify *via* clustering of gene linkages clearly have some capacity for predicting the phenotypic outcome of perturbing gene activity. However, since many datasets used in network construction are still incomplete, the modules that we discover are likely to be correspondingly of limited resolution. To improve our predictions, we thus focused primarily on individual linkages between genes, rather than module membership, as a means to predict phenotypes.

Interestingly, we observe a decay in the probability of genes being essential with increasing distance from other essential genes (**Figure J**), also consistent with the clustering of genes conferring these phenotypes in the network. Similarly, the penetrance of essentiality decreases in a similar fashion (**Figure J**).

As described in the main text, we exhaustively tested Wormnet's ability to predict the genes identified in each of 43 genome-wide RNAi phenotypic screens (**Table C**). Using ROC analysis (**Figure S1**), we find 29 of these screens are strongly predictable from Wormnet, 10 are weakly or moderately predictable, and 4 are predicted at no better than random levels. This trend depends strongly upon including network edges supported only by single lines of evidence (**Figure R**), arguing that while data integration is useful for identifying multiple lines of support for each association, an equally important role is to select confident linkages where only one line of evidence is available.

We next examined topological properties of the network and compared these with properties of a randomized version of the network. **Figure K-1** plots the node degree distribution of the core *C. elegans* gene network. Many network models derived from

complex biological systems are characterized by scale-free degree distributions (*34*). However, the core functional gene network is not scale-free. Instead, we find the degree distribution is well fit (r^2 = 0.99) by a combined power-law/exponential decay model, following a power-law for genes with lower connectivity, then exponential decay for genes with degrees higher than a characteristic threshold (β = 101, **Figure K-1**). Previous protein interaction networks have been observed to be scale-free, although it has been argued that this is simply a consequence of incomplete sampling of the networks (*35*). Wormnet's non-scale free nature may simply be a consequence of the more complete network, or possibly a variation due to its more inclusive linkage type, which spans physical as well as other interactions. One possibility is that it may derive from practical limits on the sizes of typical cellular pathways—this would imply a rough upper bound on pathway size, resulting in systematic under-representation of genes with the highest connectivities.

In further examination of Wormnet topology, the distribution of shortest path lengths between all gene pairs in the network (**Figure K-2)** differs from that of a randomized version of the network that was calculated by randomly swapping edges while maintaining each node's degree (*36*). The real network shows a significantly higher frequency of long paths, indicating that the network exhibits considerable non-random structure. Consistent with the effective capture of pathways and processes in the network, Wormnet shows considerably more clustering than expected at random (**Figure L**).

We tested for representational bias in the core network by ensuring that genes from different functional categories were evenly represented in the network. We measured the retrieval rate of genes from each of 12 functional categories (**Table A**) to test for systematic functional bias among the linkages. **Figure K-3** illustrates that while some bias exists for genes of protein synthesis among high-confidence linkages, such bias is minimal across the other functional categories, which show similar retrieval rates. The final core network includes ~60-90% of the genes from each of the 12 functional categories. We attempted to clarify the contribution to the predictive power of specific GO annotations made by each type of evidence integrated into the network in **Figure M**. To further evaluate the coverage of Wormnet for different biological systems, we

demonstrated that Wormnet effectively captures the 'molecular machines' previously defined by Gunsalus *et al.* (*37*) by testing the ability of Wormnet to predict components of each machine using ROC analysis (**Figure N**).

We asked if genes linked in Wormnet were more likely to be co-expressed in the same tissues. We observed this to be the case (**Figure O**) at levels significantly above random expectation, supporting the observation that Wormnet is capable of making tissue-specific phenotypic predictions, at least in part because genes linked in Wormnet have a higher chance of being expressed in the same tissues.

Finally, we compared the prediction of RNAi phenotypes by Wormnet and four previous networks using a set of 10 RNAi phenotypes reported after all networks were constructed and/or published (**Figure P**). As expected, Wormnet shows greatly increased coverage of genes with each phenotype (accompanied by increased accuracy), largely due to its increased coverage of the proteome over previous networks. The full version of Wormnet shows enhanced performance over the core set, supporting the use of probabilistic linkages for phenotype prediction.

REFERENCES

1. N. Chen *et al.*, *Nucleic Acids Res* **33**, D383 (Jan 1, 2005).
2. I. Lee, S. V. Date, A. T. Adai, E. M. Marcotte, *Science* **306**, 1555 (Nov 26, 2004).
3. B. Efron, R. Tibshirani, *An introduction to the bootstrap*, Monographs on statistics and applied probability (Chapman & Hall, New York, 1993), pp. 439.
4. B. Efron, *J. Am. Stat. Assoc.* **78**, 316 (1983).
5. C. Sima, U. Braga-Neto, E. R. Dougherty, *Bioinformatics* **21**, 1046 (Apr 1, 2005).
6. U. M. Braga-Neto, E. R. Dougherty, *Bioinformatics* **20**, 374 (Feb 12, 2004).
7. I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series in Data Management Systems (Morgan Kaufmann, ed. 2nd, 2005), pp. 560.
8. I. Lee, Z. Li, E. M. Marcotte, *PLoS ONE* **2**, e988 (2007).
9. M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya, *Nucleic Acids Res* **30**, 42 (Jan 1, 2002).
10. P. Bork *et al.*, *Curr Opin Struct Biol* **14**, 292 (Jun, 2004).
11. J. M. Stuart, E. Segal, D. Koller, S. K. Kim, *Science* **302**, 249 (Oct 10, 2003).
12. W. Zhong, P. W. Sternberg, *Science* **311**, 1481 (Mar 10, 2006).
13. R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (Oct 24, 1997).
14. R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (Sep 11, 2003).
15. J. Gollub *et al.*, *Nucleic Acids Res* **31**, 94 (Jan 1, 2003).
16. T. Barrett *et al.*, *Nucleic Acids Res* **33**, D562 (Jan 1, 2005).
17. S. Li *et al.*, *Science* **303**, 540 (Jan 23, 2004).
18. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, *Proc Natl Acad Sci U S A* **96**, 4285 (Apr 13, 1999).
19. M. Huynen, B. Snel, W. Lathe, 3rd, P. Bork, *Genome Res* **10**, 1204 (Aug, 2000).
20. Y. I. Wolf, I. B. Rogozin, A. S. Kondrashov, E. V. Koonin, *Genome Res* **11**, 356 (Mar, 2001).
21. P. M. Bowers *et al.*, *Genome Biol* **5**, R35 (2004).
22. T. Dandekar, B. Snel, M. Huynen, P. Bork, *Trends Biochem Sci* **23**, 324 (Sep, 1998).
23. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, N. Maltsev, *Proc Natl Acad Sci U S A* **96**, 2896 (Mar 16, 1999).
24. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
25. S. V. Date, E. M. Marcotte, *Nat Biotechnol* **21**, 1055 (Sep, 2003).
26. B. J. Stapley, G. Benoit, *Pac Symp Biocomput*, 529 (2000).
27. T. K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, *Nat Genet* **28**, 21 (May, 2001).
28. L. R. Matthews *et al.*, *Genome Res* **11**, 2120 (Dec, 2001).
29. M. Remm, C. E. Storm, E. L. Sonnhammer, *J Mol Biol* **314**, 1041 (Dec 14, 2001).
30. T. Hulsen, M. A. Huynen, J. de Vlieg, P. M. Groenen, *Genome Biol* **7**, R31 (2006).
31. L. Giot *et al.*, *Science* **302**, 1727 (Dec 5, 2003).
32. I. Lee, R. Narayanaswamy, E. M. Marcotte, in *Yeast Gene Analysis* I. Stansfield, Ed. (Elsevier Press 2006).
33. R. S. Kamath *et al.*, *Nature* **421**, 231 (Jan 16, 2003).
34. A. L. Barabasi, R. Albert, *Science* **286**, 509 (Oct 15, 1999).

35. J. D. Han, D. Dupuy, N. Bertin, M. E. Cusick, M. Vidal, *Nat Biotechnol* **23**, 839 (Jul, 2005).
36. R. Milo *et al.*, *Science* **298**, 824 (Oct 25, 2002).
37. K. C. Gunsalus *et al.*, *Nature* **436**, 861 (Aug 11, 2005).
38. J. Wang, S. K. Kim, *Development* **130**, 1621 (Apr, 2003).
39. S. A. McCarroll *et al.*, *Nat Genet* **36**, 197 (Feb, 2004).
40. C. Shen, D. Nettleton, M. Jiang, S. K. Kim, J. A. Powell-Coffman, *J Biol Chem* **280**, 20580 (May 27, 2005).
41. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (Jun 4, 1998).
42. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc Natl Acad Sci U S A* **95**, 14863 (Dec 8, 1998).
43. S. J. McKay *et al.*, *Cold Spring Harb Symp Quant Biol* **68**, 159 (2003).
44. G. van Haaften *et al.*, *Curr Biol* **16**, 1344 (Jul 11, 2006).
45. S. Cho, K. W. Rogers, D. S. Fay, *Curr Biol* **17**, 203 (Feb 6, 2007).
46. C. Schmitz, P. Kinge, H. Hutter, *Proc Natl Acad Sci U S A* **104**, 834 (Jan 16, 2007).
47. T. Lamitina, C. G. Huang, K. Strange, *Proc Natl Acad Sci U S A* **103**, 12173 (Aug 8, 2006).
48. E. J. Cram, H. Shang, J. E. Schwarzbauer, *J Cell Sci* **119**, 4811 (Dec 1, 2006).
49. M. C. Saleh *et al.*, *Nat Cell Biol* **8**, 793 (Aug, 2006).
50. J. A. Govindan, H. Cheng, J. E. Harris, D. Greenstein, *Curr Biol* **16**, 1257 (Jul 11, 2006).
51. J. C. Labbe, A. Pacquelet, T. Marty, M. Gotta, *Genetics* **174**, 285 (Sep, 2006).
52. K. K. Stein, E. S. Davis, T. Hays, A. Golden, *Genetics* **175**, 107 (Jan, 2007).
53. S. P. Curran, G. Ruvkun, *PLoS Genet* **3**, e56 (Apr 6, 2007).
54. D. Wang *et al.*, *Nature* **436**, 593 (Jul 28, 2005).
55. B. Lehner *et al.*, *Genome Biol* **7**, R4 (2006).
56. K. Ashrafi *et al.*, *Nature* **421**, 268 (Jan 16, 2003).
57. N. L. Vastenhouw *et al.*, *Curr Biol* **13**, 1311 (Aug 5, 2003).
58. J. Pothof *et al.*, *Genes Dev* **17**, 443 (Feb 15, 2003).
59. E. A. Nollen *et al.*, *Proc Natl Acad Sci U S A* **101**, 6403 (Apr 27, 2004).
60. G. Lettre *et al.*, *Cell Death Differ* **11**, 1198 (Nov, 2004).
61. G. Poulin, Y. Dong, A. G. Fraser, N. A. Hopper, J. Ahringer, *EMBO J* (Jun 30, 2005).
62. B. Sonnichsen *et al.*, *Nature* **434**, 462 (Mar 24, 2005).
63. D. Sieburth *et al.*, *Nature* **436**, 510 (Jul 28, 2005).
64. B. Hamilton *et al.*, *Genes Dev* **19**, 1544 (Jul 1, 2005).
65. M. Hansen, A. L. Hsu, A. Dillin, C. Kenyon, *PLoS Genet* **1**, 119 (Jul, 2005).
66. A. R. Frand, S. Russel, G. Ruvkun, *PLoS Biol* **3**, e312 (Oct, 2005).
67. J. K. Kim *et al.*, *Science* **308**, 1164 (May 20, 2005).
68. Y. Suzuki, M. Han, *Genes Dev* **20**, 423 (Feb 15, 2006).

FIGURES

Figure A

Regression models derived between the data intrinsic scores for each functional genomic data set and log likelihood scores (*LLS*). The likelihoods of functional association between gene pairs derived from the given experimental or computational lines of evidence was assessed on a reference set of gene pairs derived from the Gene Ontology biological process annotation. Lines of evidence include: (A-1) co-occurrence of two *C. elegans* gene names in across the set of Medline abstracts (*2*), (A-2) mRNA co-expression (as an example of a time series profiling following dauer exit (*38*); see Table S1), (A-3) phylogenetic profiles calculated from 117 bacterial genomes (*25*), (A-4) gene neighbors calculated from 133 archaeal and bacterial genomes, ranked used the scheme of ref. (*21*), (A-5) linkages between orthologs in a yeast probabilistic functional gene network (*8*), (A-6) interologs from the fly yeast two-hybrid based interactome (*31*), (A-7) interologs from human protein interaction data (**Table S2B**). Each filled circle represents a bin of between 500 and 2000 gene pairs, depending on data set.

Figure A

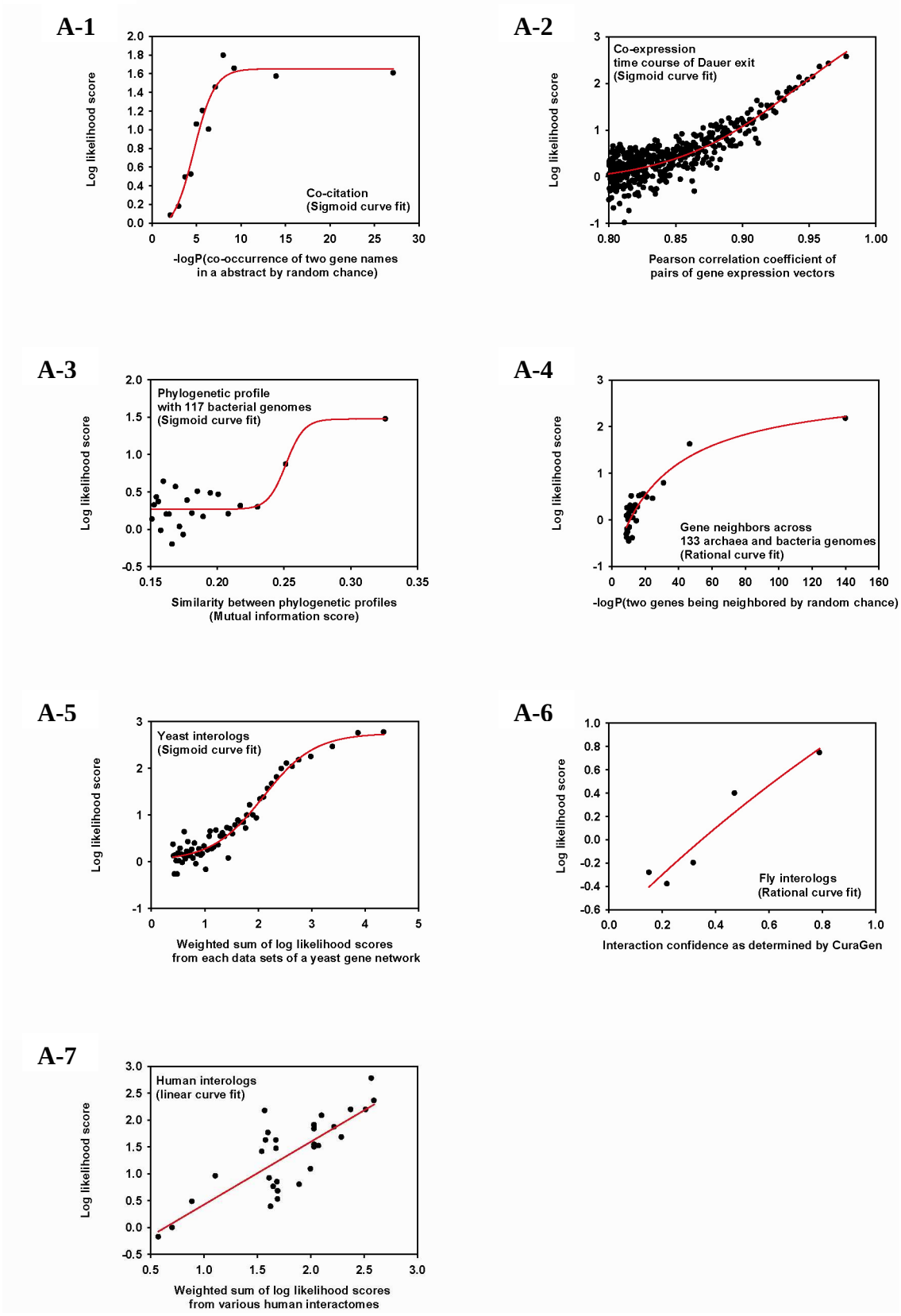

A-1
Log likelihood score
Co-citation
(Sigmoid curve fit)
-logP(co-occurrence of two gene names
in a abstract by random chance)
A-2
Co-expression
time course of Dauer exit
(Sigmoid curve fit)
Log likelihood score
Pearson correlation coefficient of
pairs of gene expression vectors
A-3
Phylogenetic profile
with 117 bacterial genomes
(Sigmoid curve fit)
Log likelihood score
Similarity between phylogenetic profiles
(Mutual information score)
A-4
Log likelihood score
Gene neighbors across
133 archaea and bacteria genomes
(Rational curve fit)
-logP(two genes being neighbored by random chance)
A-5
Yeast interologs
(Sigmoid curve fit)
Log likelihood score
Weighted sum of log likelihood scores
from each data sets of a yeast gene network
A-6
Log likelihood score
Fly interologs
(Rational curve fit)
Interaction confidence as determined by CuraGen
A-7
Human interologs
(linear curve fit)
Log likelihood score
Weighted sum of log likelihood scores
from various human interactomes

Figure B

The GO biological process reference set was constructed from terms in GO level 2 to 10, removing terms annotating excessive genes. This plot shows the number of reference set gene pairs contributed from each GO annotation, ranked by abundance. The top 5 terms account for >78% of the reference set gene pairs and were therefore omitted to remove excess bias in the reference set.

Figure B

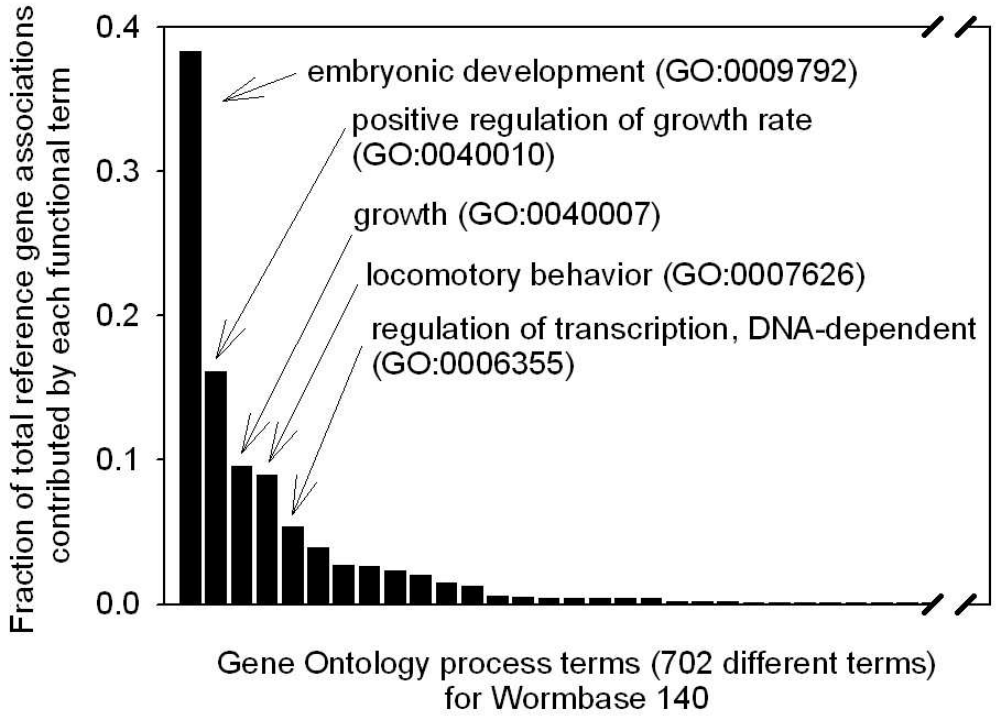

Fraction of total reference gene associations contributed by each functional term
0.4
0.3
0.2
0.1
0.0
embryonic development (GO:0009792)
positive regulation of growth rate (GO:0040010)
growth (GO:0040007)
locomotory behavior (GO:0007626)
regulation of transcription, DNA-dependent (GO:0006355)
Gene Ontology process terms (702 different terms) for Wormbase 140

Figure C

Two examples of DNA microarray datasets that failed the test for inclusion in the network. For the GEO heat shock data set *(39)*, we considered all gene pairs correlated > 0.87 (*i.e.*, the 99% confidence level for a sample size of 7 arrays) and observed no elevated LLS score. Likewise, for the SMD hypoxia data set (*40*), we considered all gene pairs with correlation coefficient > 0.80 (*i.e.*, the 99% confidence level for a sample size of 9 arrays) and observed no elevated LLS score. Contrast these cases with the SMD Dauer example in **Figure A-2**, which shows excellent correlation between PCC and LLS and which was therefore included.

Figure C

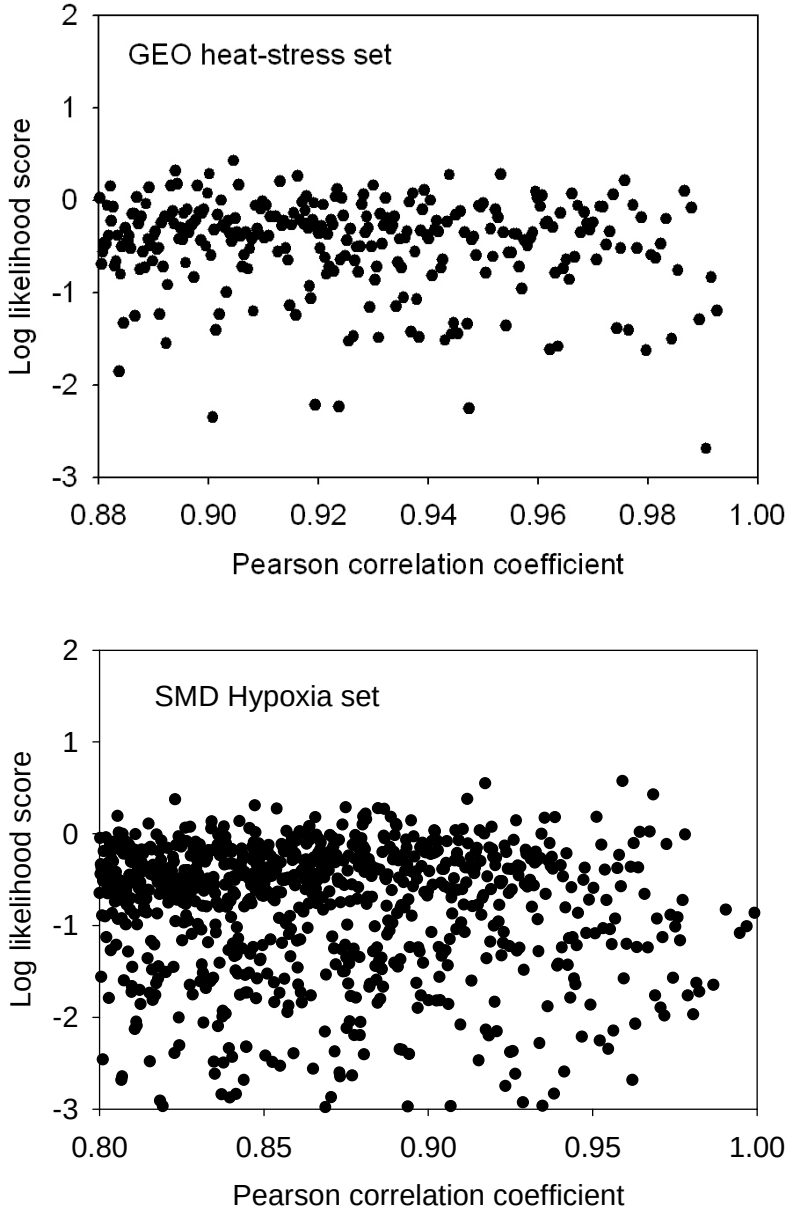
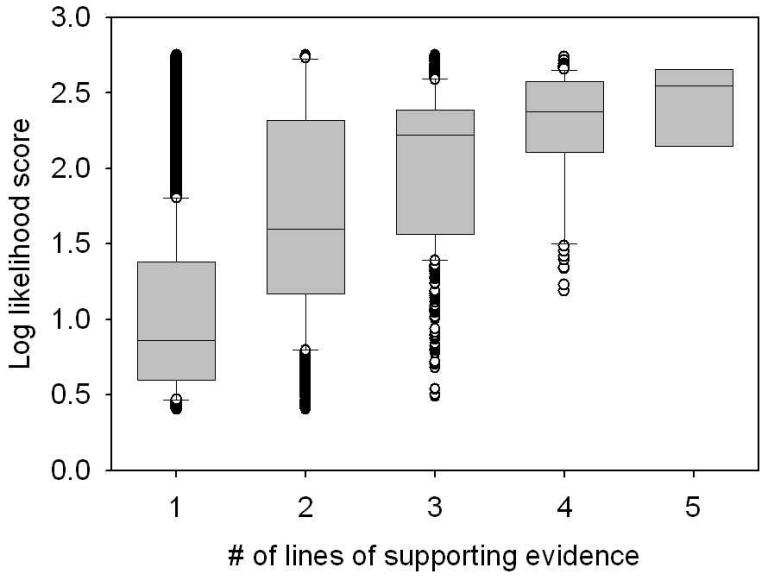

GEO heat-stress set
Log likelihood score
2
1
0
-1
-2
-3
0.88
0.90
0.92
0.94
0.96
0.98
1.00
Pearson correlation coefficient
SMD Hypoxia set
Log likelihood score
2
1
0
-1
-2
-3
0.80
0.85
0.90
0.95
1.00
Pearson correlation coefficient

Figure D

The number of lines of evidence for a linkage correlates with the LLS of the linkage, as shown by plotting the distribution of LLS scores for linkages with different numbers of lines of supporting evidence. Each distribution is summarized as a standard bar-and-whiskers plot, with the central horizontal line indicating the median LLS score and the boundaries of the box indicating the first and third quartiles of the distribution. We see up to 5 lines of evidence for individual links—such links are measurably more accurate on average.

Figure D

For the core network:



For the full network:

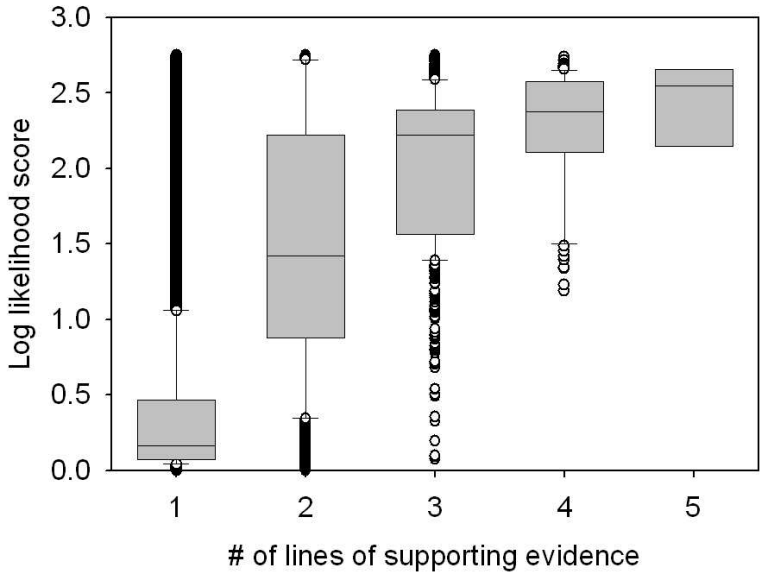


Figure E

The distribution of LLS scores in the network is apparent in a plot of network size as a function of LLS score. The core network represents the top 113,829 linkages and captures the strongest available linkages, following which linkage scores decline towards LLS = 0, with the full network consisting of 384,700 linkages (cumulative LLS > 0).

Figure E

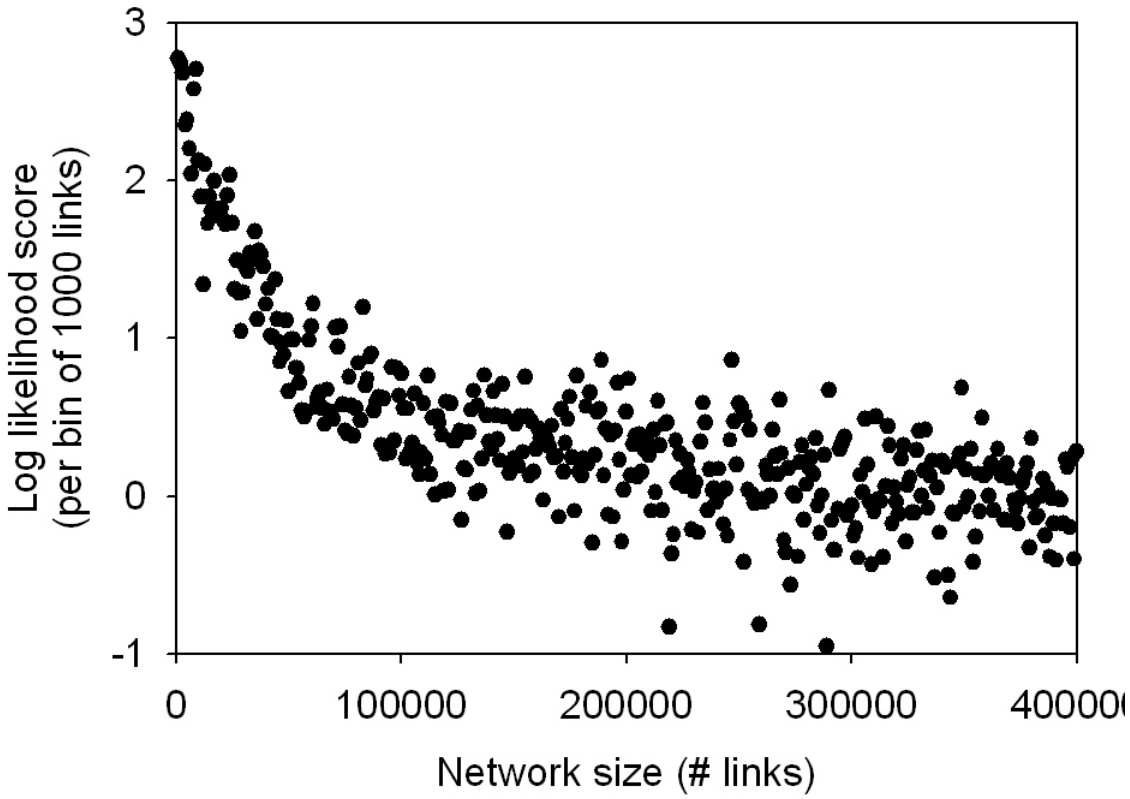

Log likelihood score
(per bin of 1000 links)
3
2
1
0
-1
0
100000
200000
300000
Network size (# links)

Figure F

To demonstrate the robustness of the training algorithm employed, we tested the Wormnet on a new reference set representing only GO terms between level 5 and 10 (506,517 positive example gene pairs), which removes ~280,000 positive example gene pairs from the original reference set (786,056 positive example gene pairs). Even after removing ~36% of all reference examples, there was no major change in the performance as measured by recall-precision analysis.

Figure F

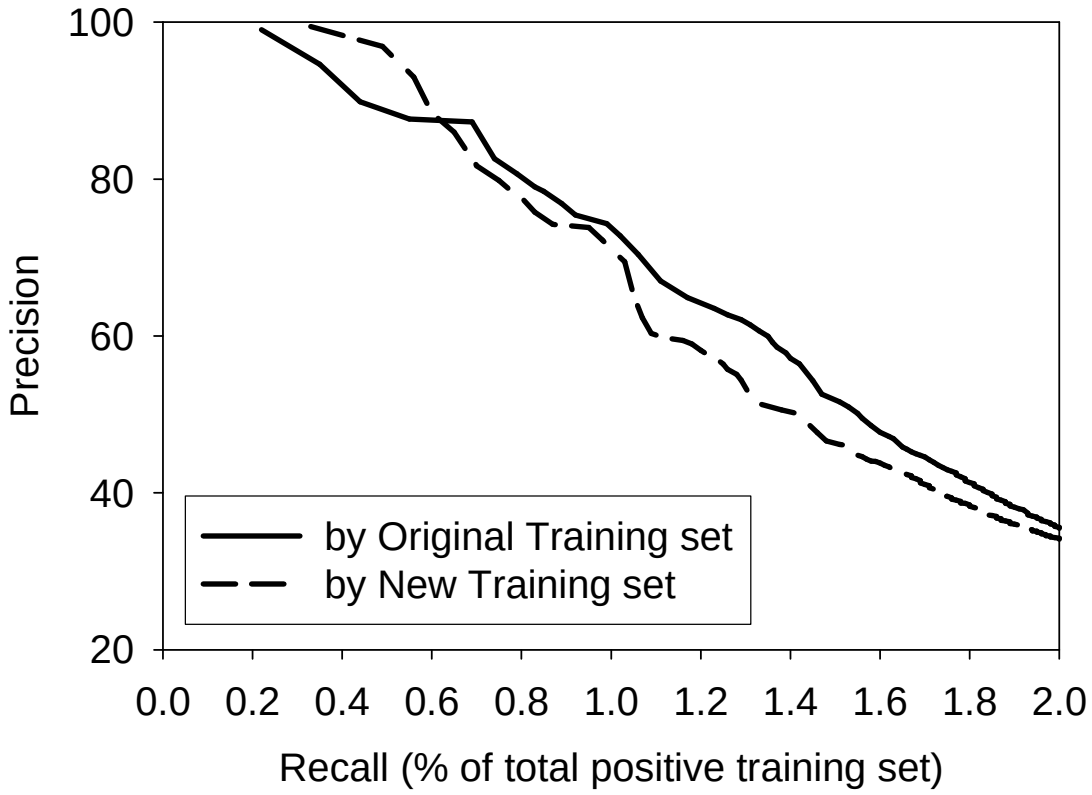

100
80
60
40
20
Precision
by Original Training set
by New Training set
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Recall (% of total positive training set)

Figure G

Evaluation of Wormnet v1 and two earlier worm gene networks (*11, 12*) by comparison to an independent set of pathway relationships composed of gene pairs belonging to the same KEGG pathways, but not sharing GO biological process terms (*i.e.*, a set of pathway linkages completely independent from GO). Values are otherwise calculated as in **Figure 1D**. Note that this benchmark should be considered to provide a lower bound of accuracy, as all high-confidence linkages confirmed by GO have been removed. This removal most affects the highest confidence interactions in the intersection of all three data sets. Nonetheless, the general ranking of network accuracies seen in **Figure 1D** is preserved using this reference set.

Figure G

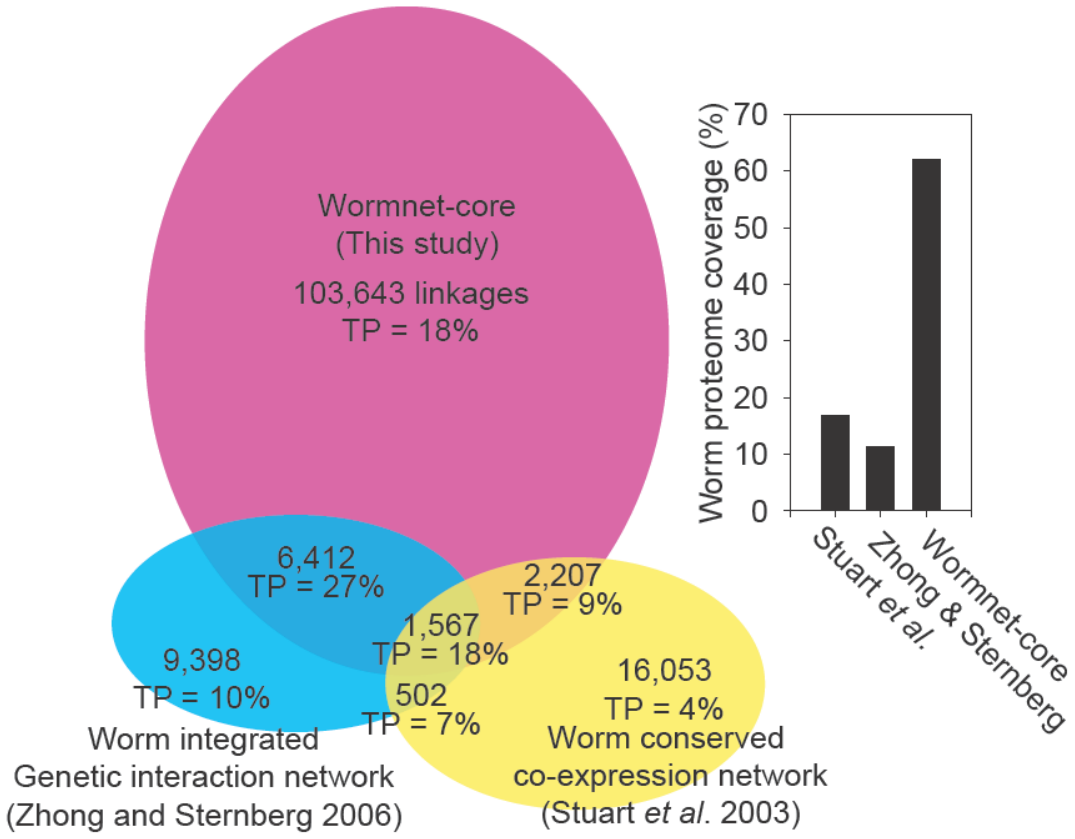
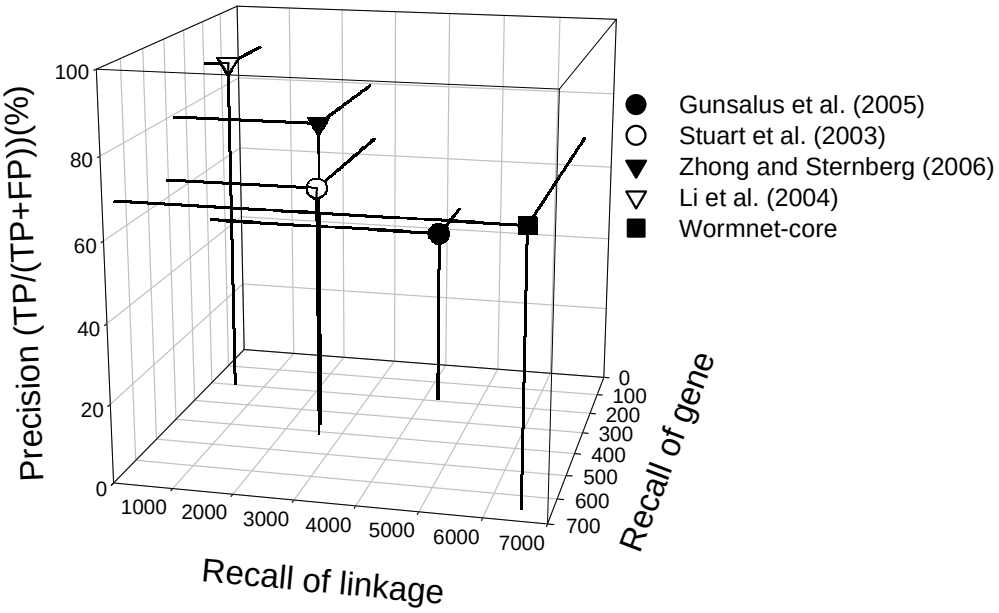

Wormnet-core
(This study)
103,643 linkages
TP = 18%
6,412
TP = 27%
2,207
TP = 9%
1,567
TP = 18%
9,398
TP = 10%
502
TP = 7%
16,053
TP = 4%
Worm integrated
Genetic interaction network
(Zhong and Sternberg 2006)
Worm conserved
co-expression network
(Stuart et al. 2003)
Worm proteome coverage (%)
0
10
20
30
40
50
60
70
Stuart et al.
Zhong & Sternberg
Wormnet-core

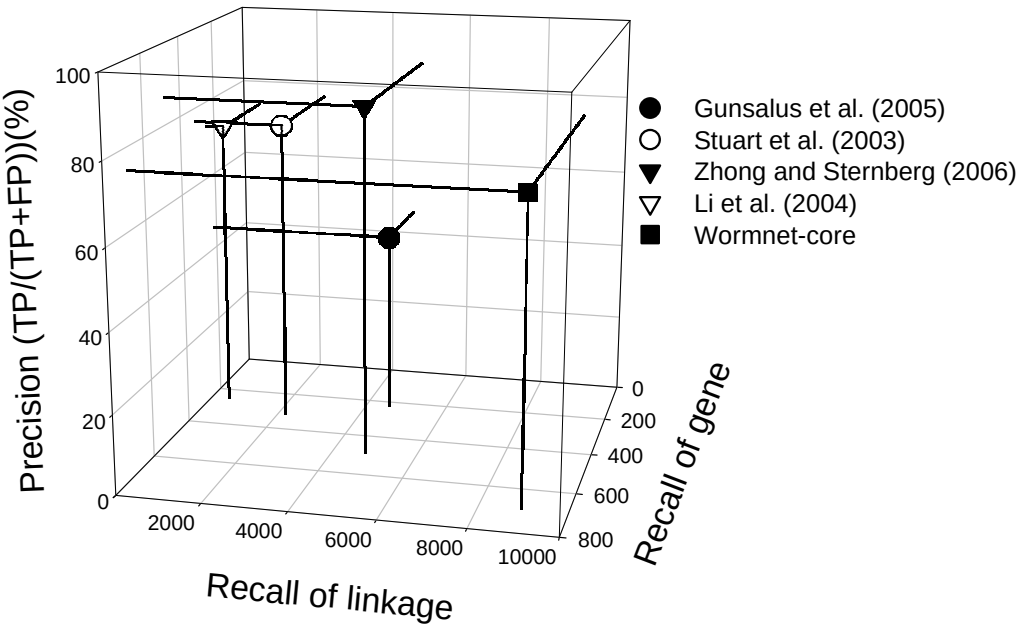
Figure H

Comparative performance of Wormnet and four other *C. elegans* gene networks on five reference linkage benchmark sets: (H-1) KEGG pathways, (H-2) GO cellular component annotations, downloaded March 2005, curated by removing the dominant terms (i.e., that annotate the most genes: plasma membrane, cytoplasm, nucleus), (H-3) GO biological process annotations downloaded March 2007, (H-4) GO biological process annotations downloaded March 2005 with protein biosynthesis-related terms removed (protein biosynthesis, translation, ribosome biosynthesis, rRNA process, etc.), and (H-5) KOG protein function categories. Note that in each test, the 5 network models show comparable ranges of precisions, but differ dramatically in recall of genes and/or linkages, with Wormnet showing higher recall than the other networks. Comparison of the two type of recall for the networks from Li *et al.* and Gunsalus *et al.* indicates a sparse network and a dense network, respectively. Note also that the network of Zhong & Sternberg employs GO 'biological process' terms as data features for calculating the network, and thus the precision of this network on these annotation sets may be an overestimate.

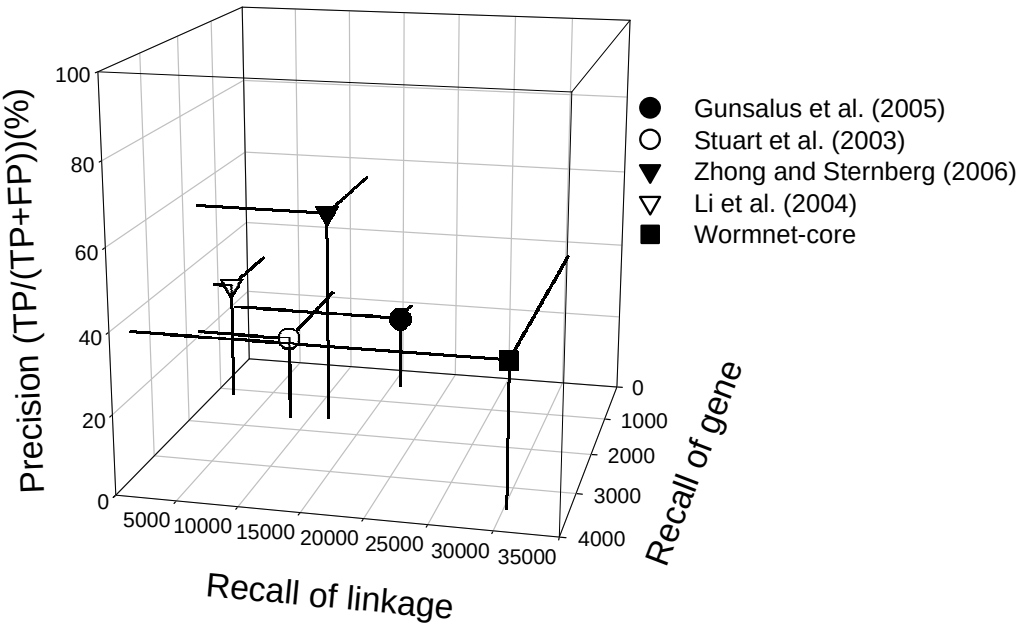
H-1. Testing versus the KEGG reference set



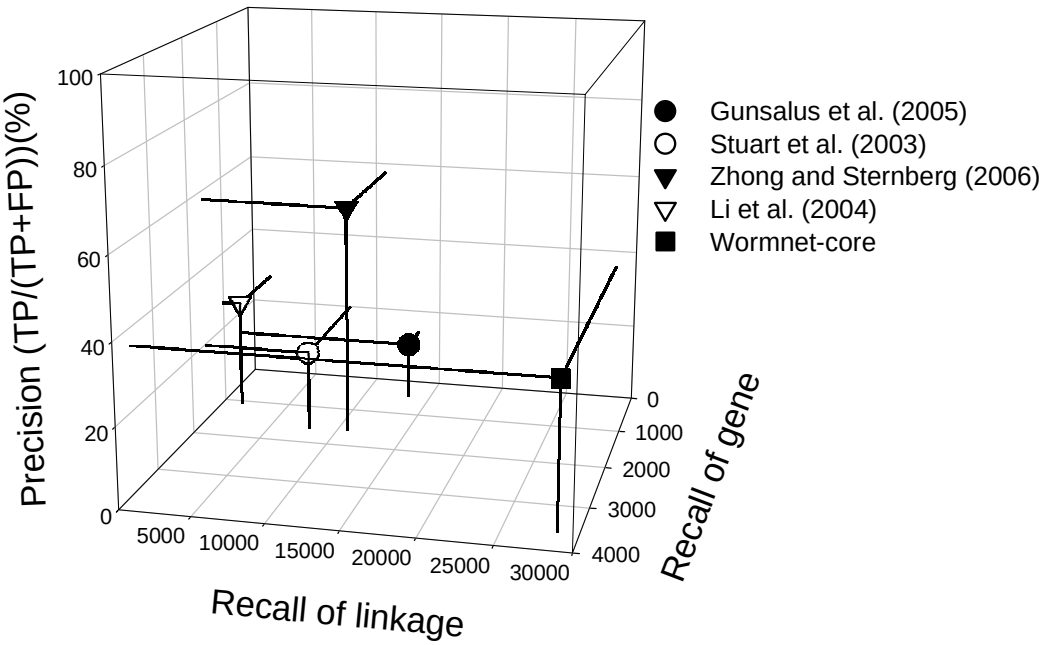
H-2. Testing versus the GO 'cellular component' (March 2007) reference set



H-3. Testing versus the GO 'biological process' (March 2007) reference set



H-4. Testing versus the GO 'biological process' (March 2005) reference set with protein synthesis/ribosome biogenesis-related terms removed



H-5. Testing versus the KOG protein function category reference set

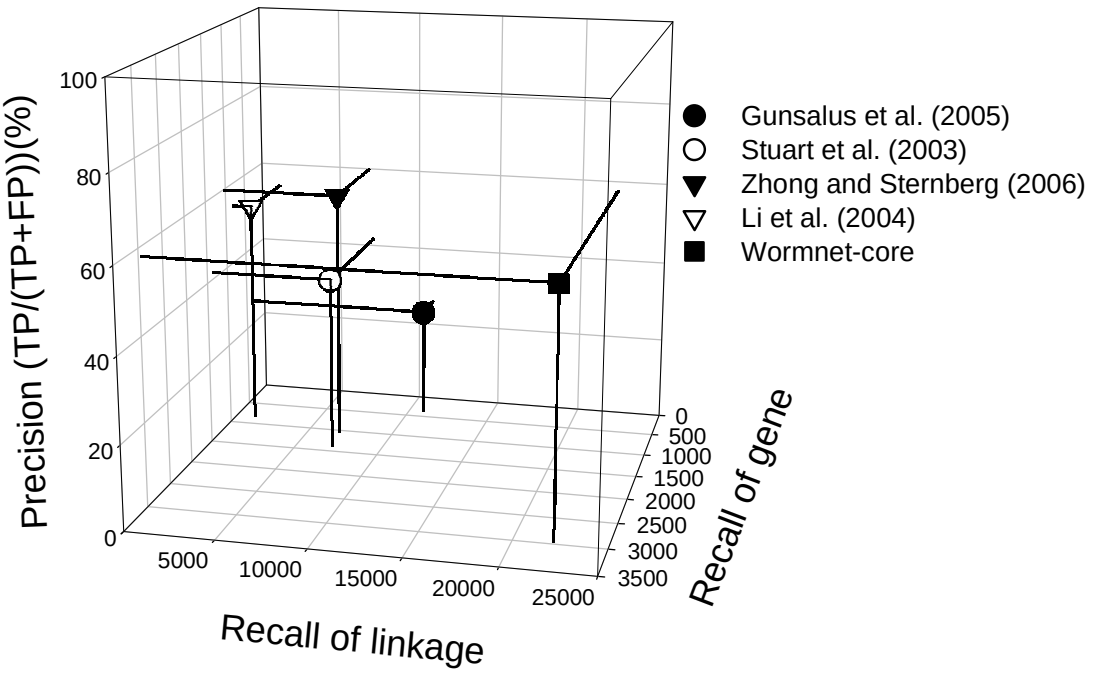


Figure I

For the network of functional modules, we summarize the functions of the genes in each module by plotting the distribution of 12 collapsed KOG functional categories (**Table A**) among the 402 modules, ordered according to the hierarchical clustering tree. The *y* axis indicates the number of genes per cluster in a given functional category, indicated by color. The functional coherence of genes in each cluster is apparent; adjacent modules (sequential along the *x* axis) are often functionally related. The network of functional modules covers 8,195 worm proteins (~42 % of the worm proteome).

Figure I

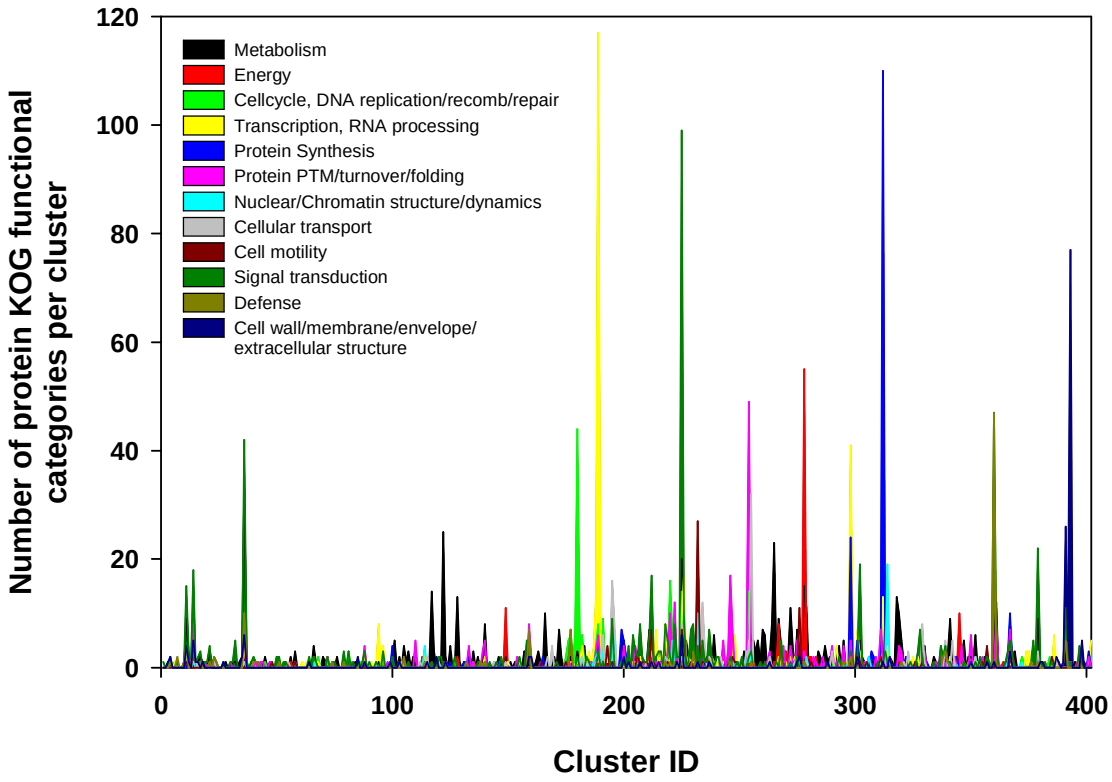

Number of protein KOG functional categories per cluster
Cluster ID
0
20
40
60
80
100
120
0
100
200
300
400
Metabolism
Energy
Cellcycle, DNA replication/recomb/repair
Transcription, RNA processing
Protein Synthesis
Protein PTM/turnover/folding
Nuclear/Chromatin structure/dynamics
Cellular transport
Cell motility
Signal transduction
Defense
Cell wall/membrane/envelope/
extracellular structure

Figure J

Essentiality of genes appears to 'diffuse' across the network. (Left) Based on RNAi phenotype, we categorized genes into two classes, embryonic lethal (emb) and non-embryonic lethal, and plot the % of genes that are emb at 1, 2, and 3 hops from each emb gene (0 hops corresponds to 100%). We find that the probability of being embryonic lethal decays with increasing distance from other embryonic lethal genes in the network. (Right) For the cases where essential genes are linked, we also examined the penetrance of the embryonic lethal RNAi phenotype as it diffuses through the network. We measured the mean % embryonic lethality for lethal genes linked by 1, 2, and 3 hops to a gene with 100% penetrance. The mean penetrance of lethality appears to decay with increasing distance from the 100% penetrant embryonic lethal genes.

Figure J

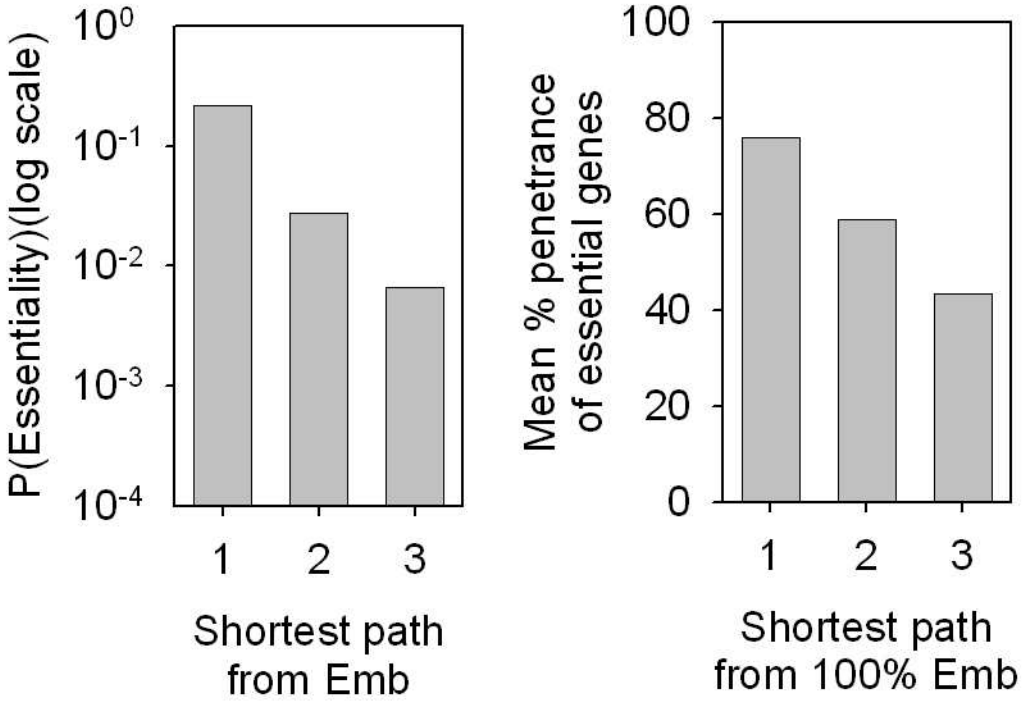
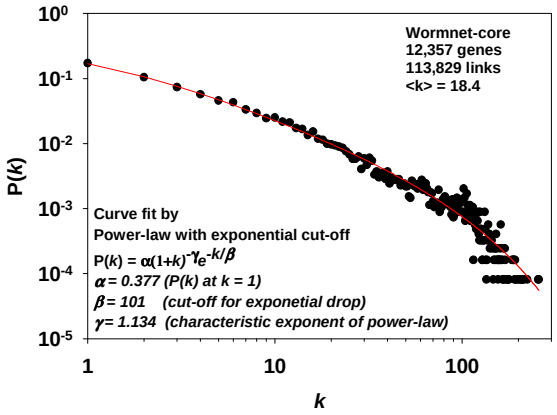

P(Essentiality)(log scale)
10^0
10^-1
10^-2
10^-3
10^-4
1
2
3
Shortest path from Emb
Mean % penetrance of essential genes
100
80
60
40
20
0
1
2
3
Shortest path from 100% Emb

Figure K

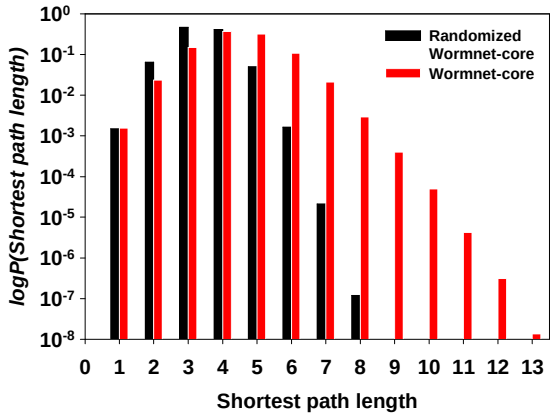
Analyses of topological properties of the core network. (K-1) The network's degree distribution is not scale-free, as shown here with a plot of the probability (*P(k)*) of each degree (*k*), fit by a power-law with exponential cut-off (red curve; $r^2 = 0.99$). (K-2) The distribution of shortest path lengths between all gene pairs in the core network shows higher frequencies of long shortest path lengths than are seen in a randomized version of the network, indicating extensive non-random structure in the actual network. (K-3) The cumulative retrieval rates of genes in each of 12 functional categories (**Table A**) as a function of network size (*i.e.*, rank ordering linkages by *LLS* score and measuring retrieval as a function of score threshold), shows that there is minimal systematic bias for the different gene functions.

Figure K

K-1.



K-2.



K-3.

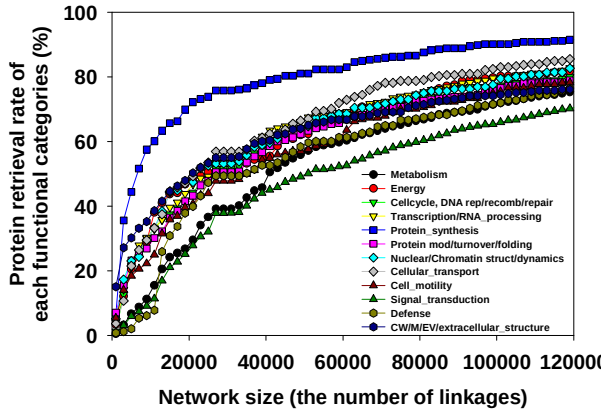


Figure L

Wormnet is considerably more clustered than a randomized version of itself, as seen by plotting the clustering coefficient (as defined by Watz & Strogatz (*41*)) as a function of network size.

Figure L

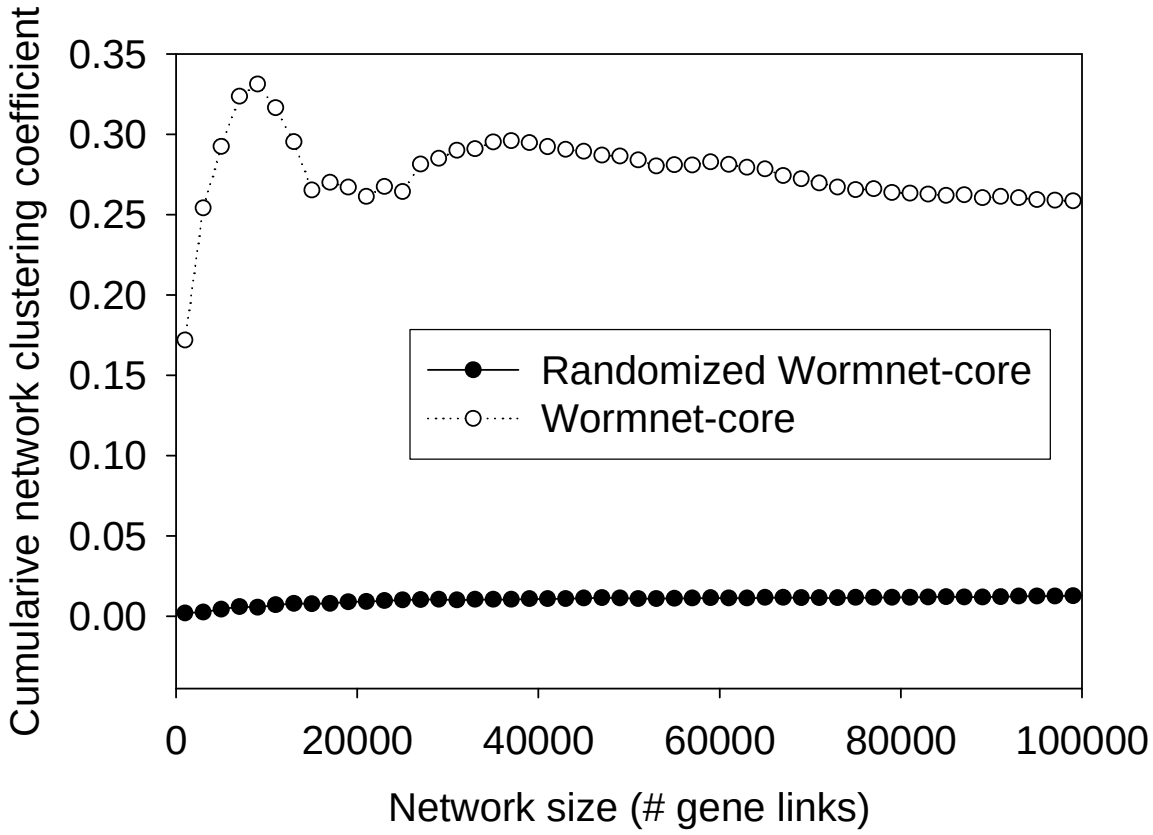

Cumularive network clustering coefficient
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20000
40000
60000
80000
100000
Network size (# gene links)
Randomized Wormnet-core
Wormnet-core

Figure M

We tested the contribution to the predictive power of specific GO annotations made by each type of integrated data. Specifically, we examined how the different lines of evidence contribute to linkages from each GO term, considering the top 290 GO biological process terms with at least one Wormnet link. For each evidence-GO term pair, we calculated the extent of contribution by that evidence towards that GO term as: Score = total true links with the evidence / total possible true links among genes with the GO term.

The matrix of evidence-GO term relationships is shown following hierarchical clustering (*42*) and indicates by increasing red intensity the extent of contribution of a given line of evidence (columns) to a given GO term (rows). For example, the strongest contributions to linkages relating to the GO term "proline biosynthesis" were made by phylogenetic profiles and gene neighbors datasets, while the strongest contribution to "transcription initiation" was made by the yeast and human datasets.

Figure M

Figure N

We examined the extent to which Wormnet captures the 'molecular machines' reported by Gunsalus *et al.* (*37*), testing the ability to recover each Gunsalus machine using ROC analysis (*i.e.*, as for the gene groups in **Figure 3**). Note that none of the RNAi phenotype data used by Gunsalus *et al.* to construct their network was used in the construction of Wormnet. If components of a given machine are clustered in the network, we expect a strongly predictive ROC plot, with a correspondingly large area under the ROC curve (AUC) close to 1.0. By contrast, if the components do not cluster at all, we expect no better than random performance (AUC = 0.5). We observe that the machines are strongly recovered by Wormnet, as represented by standard bar-and-whiskers plots, with the median AUC indicated by the central horizontal line and the boundaries of the box indicating the first and third quartiles. By contrast, randomizing the linkages in Wormnet destroys the signal.

The list of Gunsalus *et al.* machines tested:
1. Actin (5 genes)
2. APC (6 genes)
3. Chromatin maintenance and nuclear membrane function (41 genes)
4. COPI complex (4 genes)
5. F1F0 ATPase (6 genes)
6. Histones (7 genes)
7. mRNA protein metabolism (35 genes)
8. MT cytoskeleton (13 genes)
9. Oocyte integrity meiosis (47 genes)
10. Polarity (6 genes)
11. Proteasome (26 genes)
12. Ribosome (59 genes)
13. Translation initiation (5 genes)
14. Vacuolar ATPase (9 genes)

Figure N

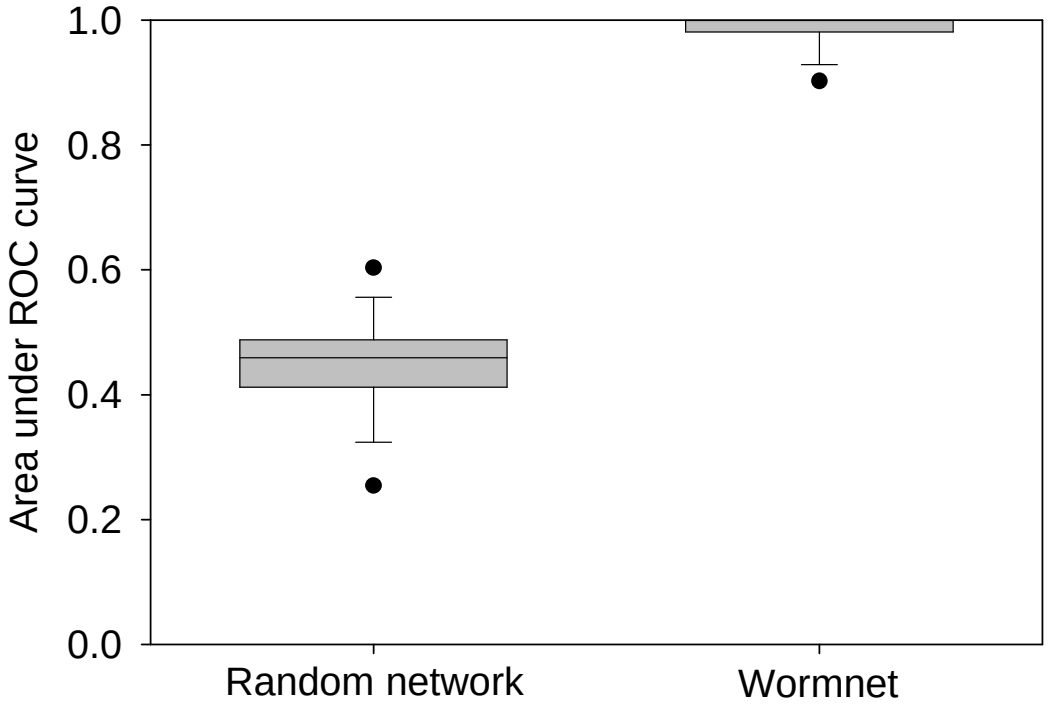


Figure O

Linkages in Wormnet tend to connect genes expressed in the same tissue. We measured how often genes linked in Wormnet are also co-expressed in a given specific tissue, using for this purpose four tissue-specific SAGE libraries derived from specific flow-cytometry-purified GFP-marked cell populations (McKay et al. (*43*); data downloaded from http://elegans.bcgsc.bc.ca). The 4 tissues were purified by McKay et al. (*43*) using either microdissection or flow cytometry on tissue specific promoter::GFP marked cell populations. The depth of SAGE analysis (the number of total tags sequenced) and the number of worm genes identified by at least one sequence tag are as follows:

Gut specific, 54,001 tags, 6,503 worm genes

Neuron specific, 91,752 tags, 8,558 worm genes

Oocyte specific, 160,053 tags, 8618 worm genes

Pharynx specific, 144,788 tags, 8772 worm genes

We measured the enrichment of tissue specific co-expression of two genes in the Wormnet-core using the following measure:

P(tissue specific co-expression) = # of (gene pairs that are linked and co-expressed in the tissue) / # of linked gene pairs.

We observed that genes linked in Wormnet are significantly more co-expressed spatio-temporally in gut, neurons, oocytes, and pharynx than gene pairs from random networks generated with the same number of genes and linkages as Wormnet (error bars indicate +/- s.d. following 10 random trials), with >200% enrichment over random for tissue-specific co-expression in all four tissues.

Figure O

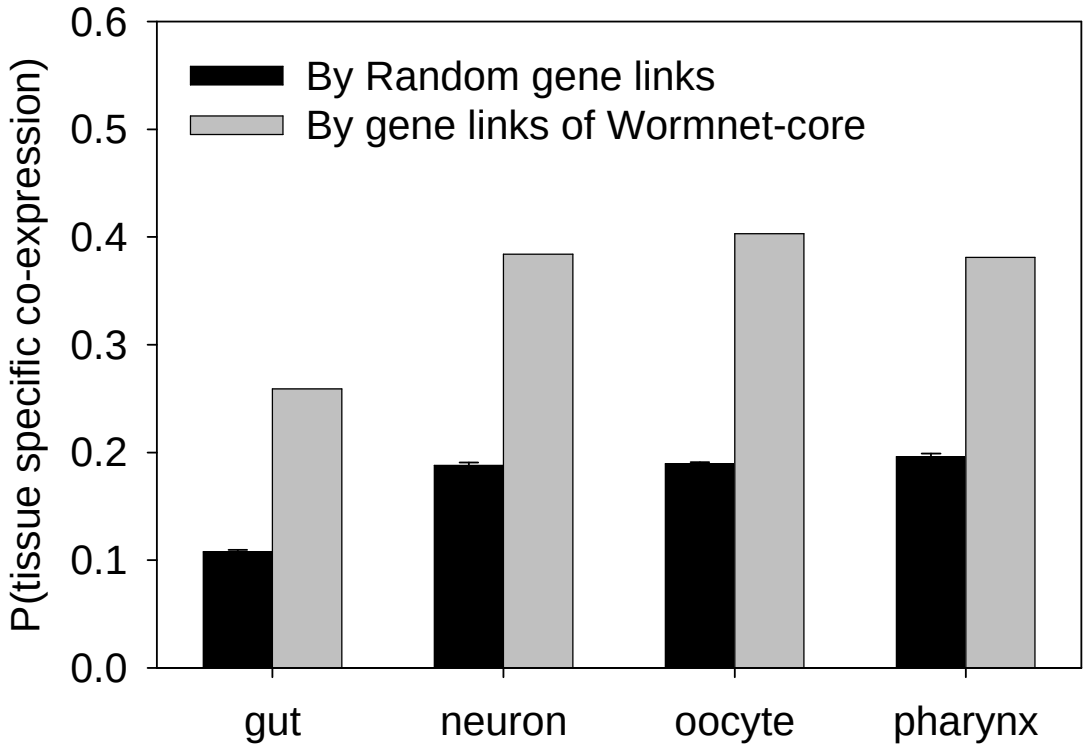

0.6
0.5
0.4
0.3
0.2
0.1
0.0
P(tissue specific co-expression)
By Random gene links
By gene links of Wormnet-core
gut
neuron
oocyte
pharynx

Figure P

Prediction of 10 genome-wide RNAi screens (*44-53*) published after June 2006 (*i*.e., after all networks were constructed) by using Wormnet or four other *C. elegans* gene networks. For each network, the median area under the ROC curve (AUC) for each of the 10 phenotypes was calculated as in **Figures 3** and **S1**, then plotted versus the median fraction of the seed gene sets covered by the network. Error bars indicate the first and third quartiles. Wormnet shows increases in both accuracy and coverage over other networks at predicting RNAi phenotypes, presumably due to its more comprehensive nature. Note also that the full network shows improved performance over the core network, indicating the utility of probabilistic linkages for this purpose.

Figure P

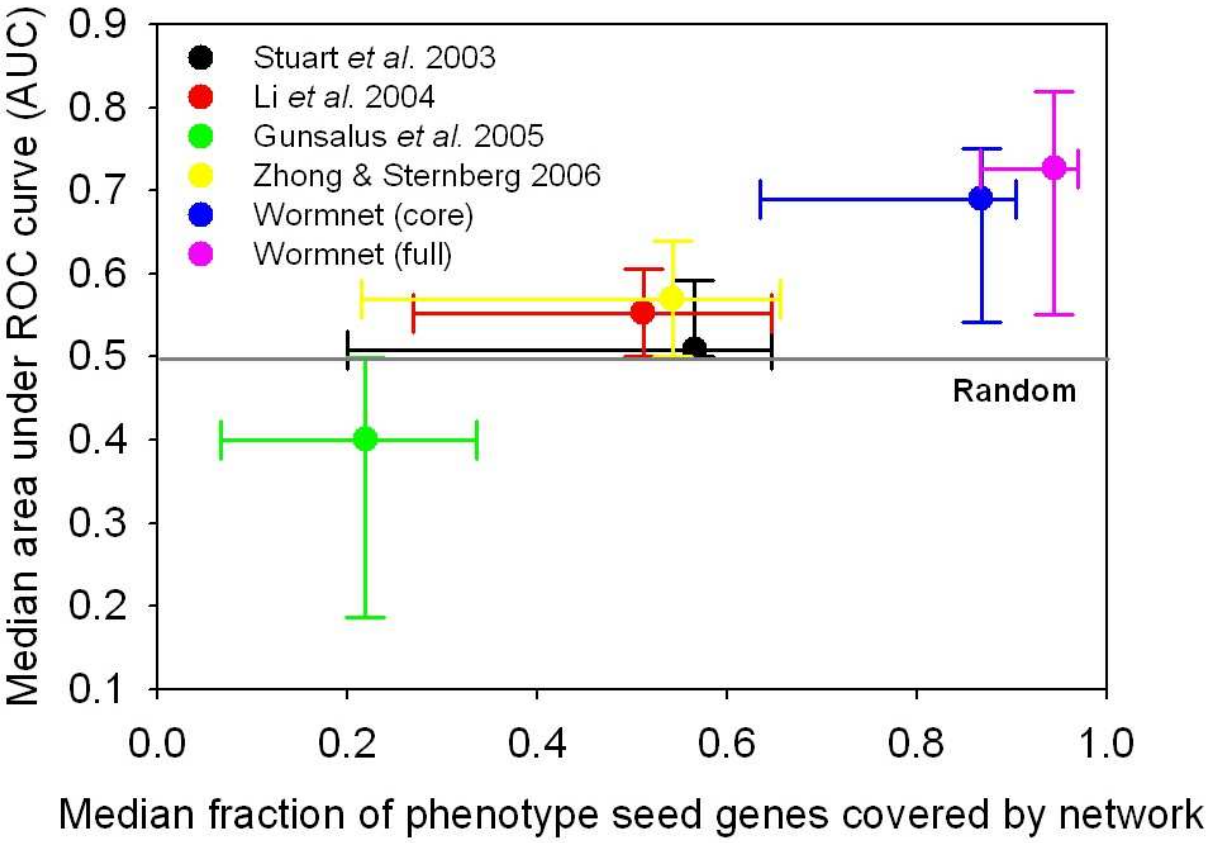

Median area under ROC curve (AUC)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Stuart et al. 2003
Li et al. 2004
Gunsalus et al. 2005
Zhong & Sternberg 2006
Wormnet (core)
Wormnet (full)
Random
0.0
0.2
0.4
0.6
0.8
1.0
Median fraction of phenotype seed genes covered by network

Figure Q

Utility of linkages supported by single lines of evidence. While data integration acts in part to increase support for each given linkage, an important role is the selection of confident linkages with only single lines of evidence. The importance of these latter cases can be seen clearly in a plot of the effects of only including single-evidence (SE) or multiple-evidence (ME) interactions on the accuracy and coverage of Wormnet – without the SE interactions the coverage is massively reduced. Performance of the network of Zhong & Sternberg is included for comparison purposes. Measurements are made on the function benchmark based upon KOG protein function categories.

Figure Q

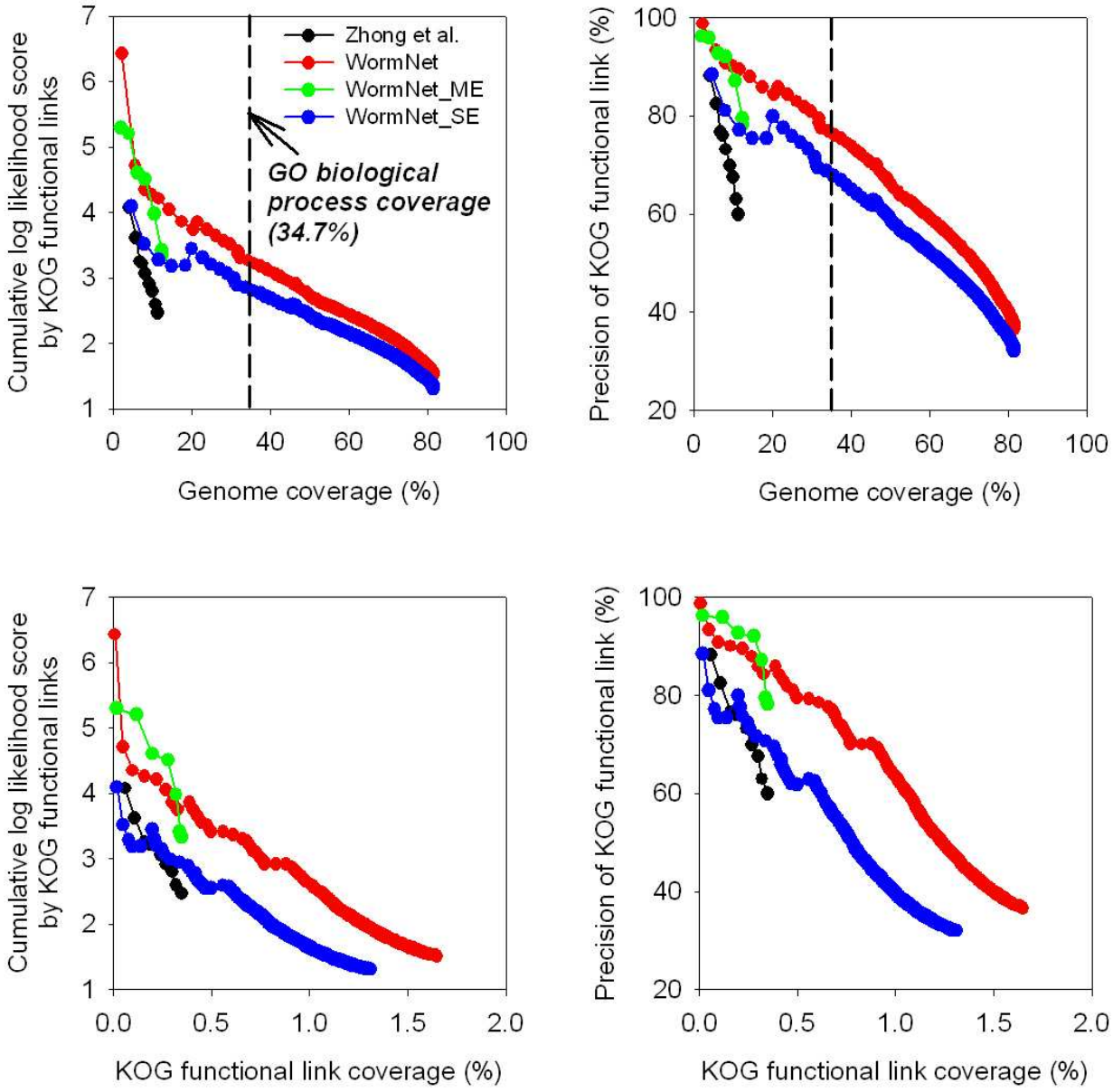
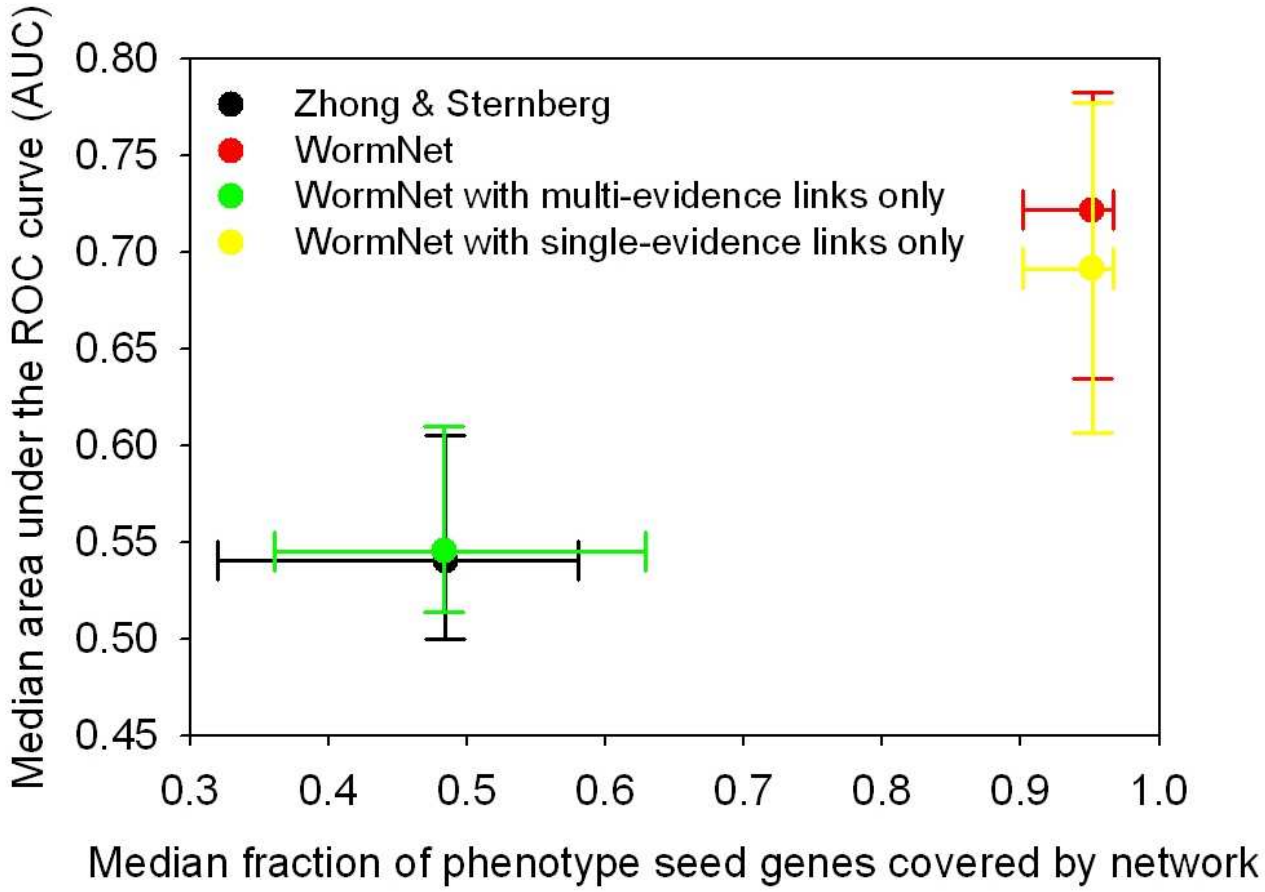

Cumulative log likelihood score by KOG functional links
Zhong et al.
WormNet
WormNet_ME
WormNet_SE
GO biological process coverage (34.7%)
Genome coverage (%)
Precision of KOG functional link (%)
Genome coverage (%)
Cumulative log likelihood score by KOG functional links
KOG functional link coverage (%)
Precision of KOG functional link (%)
KOG functional link coverage (%)

Figure R

The use of single-evidence interactions is also essential for the network's ability to successfully predict the genes associated with RNAi phenotypes. The plots shows a comparative performance of Wormnet, Wormnet subsets containing only multiple evidence or only single evidence linkages, and the network of Zhong and Sternberg on prediction of the complete set of 43 RNAi phenotypes. For each network, the median area under the ROC curve (AUC) for each of the phenotypes (calculated as in **Figures 3** and **S1**) is plotted versus the median fraction of the seed gene sets covered by the network. Error bars indicate the first and third quartiles. Single evidence links are therefore critical for the full predictive power in Wormnet.

Figure R


Median area under the ROC curve (AUC)
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
Zhong & Sternberg
WormNet
WormNet with multi-evidence links only
WormNet with single-evidence links only
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Median fraction of phenotype seed genes covered by network

TABLES

Table A
Twelve functional category-keys collapsed from 23 KOG (Eukaryote clusters of orthologous genes) functional category-keys. These 12 keys were used for data visualization.

Collapsed key	Collapsed key description	Corresponding KOG key(s)
01	Metabolism	G, E, F, H, I, P, Q
02	Energy	C
03	Cell cycle, DNA replication/recombination/repair	L, D
04	Transcription, RNA processing	A, K
05	Protein synthesis	J
06	Protein post-translational modification, turnover, folding	O
07	Nuclear/Chromatin structure/dynamics	B, Y
08	Cellular transport	U
09	Cell motility	N, Z
10	Signal transduction	T
11	Defense	V
12	Cell wall, membrane, envelope, extracellular structure	M, W

Table B. 50 core and 124 non-core interactors tested by RNAi for their ability to suppress the SynMuv phenotype

These genes represent the immediate neighbours of the 6 known suppressors of the synMuv pathway (the genes *zfp-1*, *gfl-1*, *mes-4*, *pqn-28*, *ZK1127.3*, *M03C11.3* (*54*, *55*)) that could be targeted by RNAi using clones from the Ahringer feeding library (*33*).

Table B.

Gene targeted (Wormbase WS140 Public gene name)	Interaction classification
act-1	core
act-2	core
act-3	core
act-4	core
arx-5	core
C04D8.1	core
C09H10.8	core
C14B1.4	core
C17E4.6	core
dac-1	core
daf-12	core
egl-45	core
egr-1	core
epc-1	core
F49E10.5	core
F53F8.1	core
gei-8	core
hda-1	core
hda-4	core
hmp-2	core
hsp-1	core
lin-53	core
lin-59	core
lsm-1	core
M04C9.5	core
mdl-1	core
mep-1	core
mes-3	core
mes-6	core
mom-2	core
mrg-1	core
mys-1	core
npp-9	core
ogt-1	core
pcaf-1	core
pgp-1	core
psa-4	core
R07E5.3	core
R08C7.3	core
ruvb-1	core
ruvb-2	core
set-2	core

spr-1	core
tra-1	core
trr-1	core
unc-55	core
VF13D12L.1	core
Y105E8A.17	core
Y110A7A.16	core
zif-1	core
apr-1	noncore
aat-5	noncore
aat-6	noncore
apl-1	noncore
arl-5	noncore
arx-2	noncore
B0207.6	noncore
B0336.5	noncore
bir-1	noncore
bub-1	noncore
C07A9.2	noncore
C08B11.6	noncore
C08F8.1	noncore
C10C6.6	noncore
C13F10.2	noncore
C13F10.7	noncore
C16C10.1	noncore
C17H11.4	noncore
C26C6.1	noncore
C26F1.3	noncore
C28A5.1	noncore
C28H8.1	noncore
C32D5.3	noncore
C35B8.3	noncore
C35D10.13	noncore
ccf-1	noncore
ccr-4	noncore
clr-1	noncore
cyb-3	noncore
dnc-1	noncore
dnj-5	noncore
dom-6	noncore
E02H1.1	noncore
egl-18	noncore
elt-6	noncore
F01D4.5	noncore
F22E5.9	noncore
F32B6.3	noncore

F33G12.4	noncore
F41H10.6	noncore
F44F4.2	noncore
F46B6.5	noncore
F47B7.7	noncore
F54D10.5	noncore
F55A3.3	noncore
F55A3.7	noncore
F55C5.7	noncore
F55G1.7	noncore
F57C2.3	noncore
F57C9.3	noncore
F58G6.1	noncore
frm-7	noncore
gsk-3	noncore
H19N07.2	noncore
hda-3	noncore
his-1	noncore
his-10	noncore
his-14	noncore
his-26	noncore
his-31	noncore
his-37	noncore
his-38	noncore
his-46	noncore
his-5	noncore
his-64	noncore
his-67	noncore
hmg-1.2	noncore
hum-1	noncore
ima-3	noncore
ire-1	noncore
isw-1	noncore
K03B8.4	noncore
K03H1.7	noncore
K05C4.7	noncore
K06A9.1	noncore
lig-1	noncore
mdf-2	noncore
mdt-18	noncore
mpk-1	noncore
pgp-12	noncore
pgp-13	noncore
pgp-14	noncore
pgp-15	noncore
pgp-2	noncore

pgp-3	noncore
pgp-4	noncore
pgp-6	noncore
pgp-7	noncore
pgp-9	noncore
R02E12.2	noncore
R02F2.7	noncore
R06A4.8	noncore
R07E5.10	noncore
R12E2.10	noncore
R144.4	noncore
rfp-1	noncore
rnp-2	noncore
sax-1	noncore
srj-44	noncore
srt-67	noncore
srw-35	noncore
T01B7.5	noncore
T04C9.1	noncore
T07G12.6	noncore
T09F3.2	noncore
T13F2.2	noncore
T16G12.5	noncore
T21F4.1	noncore
T22C1.5	noncore
T24C4.6	noncore
T24D1.3	noncore
T24F1.2	noncore
tfg-1	noncore
unc-16	noncore
W02F12.6	noncore
Y106G6H.15	noncore
Y39B6A.1	noncore
Y44E3A.6	noncore
Y53C12A.4	noncore
Y5F2A.4	noncore
Y87G2A.10	noncore
ZK1127.10	noncore
ZK863.3	noncore
zyg-9	noncore

Table C.
List of RNAi screens predicted from the network and used to annotate Wormnet v1 modules with enrichment for phenotypes. Predictability indicates ability to recover genes with the marked phenotype in leave-one-out ROC analysis (Figure S1).

Table C.

Phenotype	Predictability	Library screened	Reference
Nonviable	strong	Ahringer	(*33*)
Growth defective (not Nonviable)	strong		
Visible post-embryonic phenotypes (not nonviable/growth defective)	weak		
Dumpy	strong		
Body morphology defect	strong		
Small	strong		
Long	strong		
Clear	strong		
Blistered	strong		
Protruding vulva	strong		
Egg laying abnormal	weak		
Patchy coloration	strong		
High incidence of males	weak		
Ruptured	strong		
Sluggish	weak		
Uncoordinated (not Nonviable)	weak		
Fat content reduced	random	Ahringer	(*56*)
Fat content increased	random		
Transposon silencing defective	strong	Ahringer	(*57*)
Mutator	weak	Ahringer	(*58*)
Polyglutamine toxicity enhanced	strong	Ahringer	(*59*)
Germ line apoptosis increased	random	Ahringer	(*60*)
Synthetic multivulva	strong	Ahringer	(*61*)
Egg osmotic integrity abnormal	strong	Cenix	(*62*)
Egg size abnormal	n/a		
Pace of development abnormal	strong		
Pace of p-lineage development abnormal	strong		
Severe pleiotropic defects	strong		
Aldicarb resistant/synapse function defective	weak	2,072 genes from Ahringer	(*63*)
Lifespan increased (Hamilton)	weak	Ahringer	(*64*)
Lifespan increased (Hansen)	strong	Ahringer	(*65*)
Molting defect	strong	Ahringer	(*66*)
RNA interference defective	strong	Ahringer + Vidal	(*67*)
PTEN(*daf-18*) synthetic lethality	weak	Ahringer	(*68*)
Radiation sensitive	strong	Ahringer	(*44*)
FSHR1 synthetic interactions	random		(*45*)
Axon guidance	weak	4,577 genes from Ahringer	(*46*)
Osmotic stress response	strong	Ahringer	(*47*)
Distal tip cell migration	strong	Ahringer	(*48*)
dsRNA uptake	strong	Ahringer	(*49*)
Meiotic maturation	strong	Ahringer	(*50*)
Suppressors of *par-2* lethality	strong	Ahringer	(*51*)
MAT-3 suppressors	strong	Ahringer	(*52*)
Lifespan increased (essentials, Curran)	strong	2,700 genes from Ahringer	(*53*)

Table D

List of the top 200 genes predicted to increase *C. elegans* lifespan found by using the 29 genes identified by Hansen *et al.* (*65*) as a seed set. Predictions confirmed by the longevity screens of Hamilton *et al.* (*64*) and Curran and Ruvkun (*53*) are indicated.

Table D.

Prediction (Wormbase WS140 Public gene name)	Sum of LLS scores to seed gene set	In Hansen *et al.* seed set?	Confirmed by Hamilton *et al.* or Curran & Ruvkun?
isp-1	13.67		
H28O16.1	10.88		Yes
F58F12.1	10.42		
R53.4	9.95		
atp-2	9.93		Yes
Y82E9BR.3	9.87		
asg-2	9.87		Yes
asg-1	9.87		
T02H6.11	9.69		
Y69A2AR.18	9.3		
F53F4.10	9.14		
W10D5.2	9.12		
T20H4.5	9.12		Yes
tag-99	9.02		
gas-1	9.02		
F27C1.7	8.42	Yes	
T10B10.2	8.37		
VW06B3R.1	8.14		
T24C4.1	8.14		
ZC410.2	8.14		
E04A4.7	8.14		
Y54F10AM.5	8.09		
asb-2	7.77	Yes	
W09C5.8	7.72		Yes
F22D6.4	7.71		
asb-1	7.65		
cyc-1	7.61	Yes	
Y56A3A.19	7.53		Yes
ZC116.2	7.53		
C34B2.8	7.49		
D2030.4	7.4		Yes
Y51H1A.3	7.33		
Y54E10BL.5	7.33		
lpd-5	7.33		
Y63D3A.7	7.33		
C33A12.1	7.33		
Y94H6A.8	7.33		
F59C6.5	7.33		Yes
C16A3.5	7.33		
C25H3.9	7.33		
mdh-1	7.32		
Y37D8A.14	6.81	Yes	
F54D8.2	6.75		
Y71H2AM.5	6.72		

T10E9.7	6.68	Yes	
Y45G12B.1	6.6	Yes	
Y53G8AL.2	6.19		
nuo-1	6.19		Yes
cco-1	6.11	Yes	Yes
Y57G11C.12	6	Yes	
K04G7.4	5.41	Yes	Yes
F43G9.1	5.29		Yes
F25H2.5	4.46		
T22B11.5	4.45		
F36A2.7	4.21		
W02F12.5	3.81		
ZK809.3	3.8		
C37E2.1	3.68		
T08B2.7	3.57		
B0303.3	3.57		
C30F12.7	3.57		
F35G12.2	3.57		
ech-1	3.57		
F56D2.1	3.52		
daf-21	3.16		
F33A8.5	3.12		
rps-0	3.12		
rps-19	2.98		
mev-1	2.94		
C06E7.1	2.78		
unc-97	2.72		
unc-112	2.71		
T26E3.7	2.66		
ctb-1	2.65		
T06D8.5	2.65		
C06E7.3	2.63		
vha-12	2.61		
rps-11	2.61		Yes
tag-32	2.54		
unc-89	2.53		
F23B12.5	2.52		
R05G6.7	2.51		
let-60	2.5		
fum-1	2.47		
rps-2	2.46		
C06H2.1	2.46	Yes	
Y110A7A.12	2.44		
pab-1	2.43		
F55A12.8	2.39		
rab-1	2.38		
unc-52	2.35		Yes
Y105C5B.12	2.35		
rps-4	2.34		
rps-1	2.33		

C16C10.11	2.28		
R04F11.2	2.27		
deb-1	2.27		
Y49A3A.3	2.26		
pab-2	2.22		
C16A3.3	2.21		
rab-7	2.21		
pdk-1	2.19		
C53A5.1	2.19	Yes	
unc-11	2.19		
rps-13	2.16		
rps-7	2.14		
vha-10	2.12		
rps-26	2.08		
rab-5	2.05		
rpl-16	2.03		
rps-23	2.02		
rab-11.1	2.02		
Y48B6A.13	2.01		
B0511.6	2.01		Yes
pas-1	1.98		
K02F2.2	1.95		
rab-6.2	1.94		
rps-8	1.91		
rpl-15	1.89		
unc-10	1.89		
smg-4	1.88		
tba-4	1.87		
egl-45	1.87		Yes
unc-8	1.82		
tag-29	1.81		
tag-207	1.81		
T26C12.1	1.81		
F55F8.3	1.8		
rps-27	1.8		
LLC1.3	1.8		
rab-6.1	1.8		
ran-4	1.79		
rps-24	1.79		
Y46E12BL.2	1.78		
MTCE.26	1.76		
F33A8.6	1.76		
rhi-1	1.75		
rap-1	1.75		
rpl-9	1.74		
app-1	1.73		
rpl-5	1.72		
sri-45	1.72		
rps-9	1.71		
ras-1	1.71		

pat-3	1.7		
rps-15	1.7		
F56D2.6	1.69		
rps-3	1.66		Yes
ran-3	1.66		
rps-10	1.65		
byn-1	1.65		
B0513.9	1.64		
sem-5	1.64		Yes
W02B12.8	1.61		
rfc-2	1.61		
unc-5	1.6		
rps-22	1.6		
C49G7.4	1.58		
ZK430.1	1.58		
rps-18	1.57		
pat-4	1.57	Yes	Yes
pat-6	1.57	Yes	
F13H8.2	1.57		
C18E9.6	1.54		
mel-11	1.53		
cav-1	1.53		
ifg-1	1.53		Yes
inf-1	1.53		
pbs-3	1.52		
F23C8.5	1.52		
eif-3.B	1.51		Yes
acs-17	1.51		
Y40B1A.4	1.5		
ZC373.1	1.5		
F54A3.4	1.5		
atm-1	1.5		
hmp-2	1.5		
eif-3.K	1.5		
Y48G1A.4	1.5		
F54A5.3	1.5		
F59A2.3	1.5		
unc-1	1.49		
rps-30	1.49		
tag-55	1.48		
C32B5.6	1.47		
T21B10.2	1.47		
R04A9.1	1.47		
eif-3.E	1.46		
C04C3.3	1.45		
rpl-24.1	1.45		
rps-5	1.45		
rpl-13	1.43		
rps-17	1.43		
clk-1	1.42		

daf-12	1.42	
daf-1	1.42	
akt-1	1.42	Yes
daf-11	1.42	
daf-16	1.42	
daf-7	1.42	
daf-18	1.42	
age-1	1.42	Yes
akt-2	1.42	
gro-1	1.42	
daf-28	1.42	
clk-2	1.42	
daf-5	1.41	
T25B9.9	1.41	
ZK1127.5	1.4	
K04G2.1	1.4	
rps-21	1.4	
eif-3.F	1.4	Yes
eif-3.D	1.4	