Supporting Online Material for

Genome-wide Prediction of *C. elegans* Genetic Interactions

Weiwei Zhong and Paul W. Sternberg*

*To whom correspondence should be addressed. E-mail: pws@caltech.edu

Published 10 March 2006, *Science* **311**, 1481 (2006)
DOI: 10.1126/science.1123287

This PDF file includes:

Supporting Online Material

DATA SETS

All data used in this study are publicly available open access data.

Reference data sets

We obtained our training set data from WormBase (*1*). For positives, we collected all available yeast-two-hybrid and genetic interaction data in WormBase excluding un-cloned loci and self-interactions. The final positive training set contained 4,687 gene pairs (seven pairs interact both genetically and physically). For negatives, we downloaded the complete 3,296 combinations of *cis* markers from WormBase excluding pairs overlapping with the positive set.

We also constructed a testing set independent of the training set to detect over-fitting problems. For testing set positives, we extracted 5,515 gene pairs from the KEGG pathway database (*2*). We took the KEGG data as our testing set rather than the training set because some KEGG database knowledge was derived from orthologous gene data rather than direct evidence in *C. elegans*. Only three gene pairs in this set overlap with the training set positives, indicating the independency of the two data sets. For negatives, we generated 5,000 random gene pairs with the percentage of annotated genes comparable to that of the positives. To reduce the chance of including interacting pairs in the negatives, we imposed a restriction in the selection process so that any gene pair in the negative set had a minimal distance of six or above in the positive training set (*3*). In other words, if the positives are displayed as a graph with vertices indicating genes and edges indicating interactions, and if a gene pair in the negatives consists of two vertices in the graph, then the shortest path between the two vertices contains at least six edges.

Predictor data sets

We collected multiple data sets and classified them into fourteen predictors (five predictors for each organism, except that the *C. elegans* interactome data set was used as training set, Table S1).

C. elegans data were mainly collected from the frozen WormBase release WS140. We excluded the annotation "wild-type" from our phenotype data, because we believed that the wild-type phenotype only has a significant prediction value if it is observed on null mutants. The biological process data were taken from the gene ontology (GO) (*4*) annotations. A fraction of the *C. elegans* GO processes were annotated electronically by converting the phenotypes curated in WormBase. We excluded these GO annotations to avoid redundancy with the phenotype predictor (a step essential for the naïve Bayesian network approach). Our micorarray data sets included 19 Affymetrix experiments curated in WormBase, 213 SMD microarray experiments also curated in WormBase, 553 experiments published by Kim *et al.* (*5*), and the Gene Recommender results (*6*). Affymetrix data with NP flags were filtered out.

All *D. melanogaster* data were downloaded in January 2005. Gene expression data were obtained from FlyBase (*7*) and the Berkeley *Drosophila* Genome Project (BDGP) (*8*). Both data sets had two attributes: tissue and developmental stage. Two genes were considered to be co-expressed if they were expressed in the same tissue at the same

developmental stage. The phenotype predictor included two data classes in FlyBase: the PHM data that describe tissues in which the phenotype manifested (PHM), and the PHC data that describe the phenotype class (PHC). *Drosophila* microarray data were obtained from Gene Expression Omnibus (GEO) database (*9*). Only the 124 experiments using the same platform (GPL72) were included. To obtain a complete set of *Drosophila* interaction data, we extracted and combined data from FlyBase and the General Repository for Interaction Datasets (GRID) (*10*). The resulting interaction data set included genetic interactions curated from literature by FlyBase and the physical interactions identified from yeast two-hybrid screens (*11*).

We downloaded our *S. cerevisiae* data from multiple databases, including the *Saccharomyces* Genome Database (SGD) (*12*), the Comprehensive Yeast Genome Database (CYPD) developed by the Munich Information Center for Protein Sequences (MIPS) (*13*), GO, and GRID. Our microarray data included 438 experiments obtained from SGD. The interaction data set was a combination of data from GRID, MIPS, and SGD. In addition to pairwise genetic and physical interactions, we converted protein complex data into physical interactions by regarding all components in a complex as interacting with each other.

Sequence data

To map orthologs between species, we obtained the latest sequence data from several databases. Our *C. elegans* sequence version was wormpep140 from WormBase; the *D. melanogaster* sequence version was r4.1 from FlyBase; and our *S. cerevisiae* sequences were downloaded from SGD in January, 2005.

EXPERIMENTAL METHODS

Genetics

The following *C. elegans* strains were used: N2, wild-type; MT2124 (*let-60*(*n1046*)); JT73 (*itr-1*(*sa73*)); MT1212 (*egl-19*(*n582*)); and MT6129 (*egl-19*(*n2368*)). RNAi was performed as previously described (*14*). MRC geneservice *C. elegans* RNAi clones were fed to L4 larvae and progeny were analyzed. RNAi assays were conducted in duplicates in each experiment and at least two independent experiments were carried out.

Microscopy

For pharyngeal pumping rate assays and plate level phenotype observations, animals were observed using a Leica MZ12.5 stereomicroscope. For VPC induction scoring, animals were observed by DIC using a Zeiss Axioplan microscope. Images were taken with a Hamamatsu ORCA-ER digital camera with Openlab 3.1.3 software (Improvision), and were processed with Adobe Photoshop 6.0.

COMPUTATIONAL METHODS

Graph visualization

The cluster visualization in Fig. 1A was obtained using TreeDyn (http://treedyn.org). The graph visualization of the proteasome interactions in Fig. 1B was obtained using GUESS (http://graphexploration.cond.org).

Ortholog mapping

We have considered three methods to map orthologs: reciprocal best BLAST (*15*) hits, Cluster of Orthologous Groups (COG) (*16, 17*), and InParanoid (*18*). The simplest method to identify orthologs is to conduct all-against-all BLAST searches of two genomes and select sequences that are each other's best hit. However, this approach assumes a one-to-one mapping of orthologs between two species. The COG database takes pairwise BLAST searches among multiple genomes and clusters best hits from multiple species to construct ortholog groups (*16*). However, manual adjustment is often required to determine the correct clustering in eukaryotes. Similar to COG, InParanoid is also a many-to-many ortholog mapping method. InParanoid detects reciprocal best hits as well as in-paralogs (paralogs that arise through a gene duplication event after the species split (*18*)). The InParanoid method is fully automatic and provides reliable results. Therefore, we decided to use InParanoid as our ortholog mapping method. The BLAST we used was downloaded from http://www.ncbi.nlm.nih.gov/BLAST/.

Using InParanoid, we were able to construct 3,987 orthologous groups composed of 4,812 *D. melanogaster* genes and 5,010 *C. elegans* genes; we also mapped 1,821 orthologous groups between *S. cerevisiae* and *C. elegans*, covering 2,772 *C. elegans* genes and 2,231 *S. cerevisiae* genes.

We will illustrate the computational strategy for orthologous gene information with the following example. Given two *C. elegans* genes (gene *a* and gene *b*), each having a group of orthologs in another organism (group *A* and group *B*), our strategy for text annotation data (such as phenotype data) was to award the gene pair *ab* a score if any gene in group *A* shared an annotation with any of the genes in group *B*. For microarray data, we computed the Pearson correlation value for all possible gene pairs that comprise a gene in group *A* and a gene in group *B*, and took the maximum Pearson correlation value to score for *ab*.

In brief, our strategy is that if a *C. elegans* gene has multiple orthologs in another species, we combine information from all orthologous genes in that species. We want to emphasize that the multiple orthologs should not to be confused with paralogs, specifically out-paralogs that arise before the species split and that often have different functions. One potential problem of our method is that if there are several *C. elegans* genes in the same orthologous group, all these *C. elegans* genes will artificially appear to interact because the same orthologous information is mapped to all of them. To solve this false positive problem, we applied a filter to exclude genetic interactions among members in the same orthologous group.

Scoring

Below are the technical details in computing the likelihood ratio *L* for various data types.

1) Annotated text data

This category includes data sets such as expression, phenotype and GO annotation data. For this type of data, we search for exact string match of gene annotations. *v* is the term usage frequency (percentage of annotated genes associated with the term) of the overlapping annotation.

In practice, due to the limited size of our training set, L was computed to a range of v values rather than one value. That is,

$$L = \frac{P(v_1 \le predictor < v_2 \mid pos)}{P(v_1 \le predictor < v_2 \mid neg)}$$

Two rules were applied to determine v_1 and v_2. First, there were at least 10 data points in each bin, that is, there must be at least 10 gene pairs in the positive training set and 10 gene pairs in the negative training set with values between v_1 and v_2. Second, there were at most 100 bins for each data set, that is, let v_{min} and v_{max} denote the minimum and the maximum possible value of the data set, then the value of $(v_2 - v_1)$ should be multiples of 1% of $(v_{max} - v_{min})$.

We will illustrate the computation with an example of the *C. elegans* GO process data. First, we selected gene pairs from the training set that have GO annotations for both genes. This resulted in 2,404 gene pairs in the positive training set and 2,094 pairs in the negative training set. Then for each of these gene pairs, we examined whether the two genes shared the same GO annotation, and if they did, what annotation term(s) was shared. We further examined the specificity of the shared GO term by computing how many genes among all annotated genes in the genome were associated with that GO term (term usage frequency). 35 gene pairs in the positive training set and 10 pairs in the negative training set shared GO term(s) between the two genes in each pair and the term usage frequencies for the shared GO terms were within 0.24-0.48%. The resulting L score is (35/2404)/(10/2094)=3.05. Therefore, a score of 3.05 is rewarded to a gene pair if the two genes share the same GO term and if the GO term has a usage frequency between 0.24-0.48%.

The resulting scoring schemes for all annotated data are listed in Fig. S1A with L plotted against the mean of its corresponding v_1 and v_2.

2) Microarray data

All data were centered and normalized, and Pearson correlations were calculated and used as v to compute L. The computational method is the same as that used for the annotated text data. The results are displayed in Figure S1B.

3) Interaction data

This category includes the Gene Recommender data, *D. melanogaster* and *S. cerevisiae* interaction data. Although the Gene Recommender data were microarray results, they were in a format similar to those of interaction data sets (*i.e.* pairs of genes). Thus we put them in this category simply to illustrate the computational method. The value v for this type of data is a binary value (interact/non-interact). Since there are extremely few interlogs for gene pairs in our negative training set, we computed L using the expected rather than the observed probability for $P(predictor = v \mid neg)$. Let I denote the number of gene pairs in the *C. elegans* genome that have interlogs in the other species, T denote the total number of *C. elegans* genes that have orthologs in that species, and N be the size of the negative training set, then $T(T\text{-}1)/2$ represents the number of all possible combinations of gene pairs and $\frac{I}{T(T-1)/2}$ represents the probability of getting a pair

with interlogs from random selection. Thus the expected probability P for having interlogs in the negative training set is

$$P = \frac{I}{T(T-1)/2} N .$$

The resulting scores are displayed in Fig. S1C. Unsurprisingly, the *Drosophila* genetic interaction data have the highest score among all predictors (Fig. S1C).

4) Manual checkpoints

We imposed several manual checkpoints in the process to ensure that our computation did not contradict the biological meaning of the data. The first manual inspection is to enforce that $L \geq 1$ for positive predictors. The basis for our genetic interaction prediction is the hypothesis that two genes are likely to interact if their orthologs interact, or if they or their orthologs are expressed in the same cell, have the same phenotype, or have the same GO annotation. The hypothesis is validated by the fact that these features indeed increased the likelihood of two genes being an interacting pair (Fig. S1). However, a few data points had $L<1$. In these cases, we manually adjusted the minimum likelihood ratios to 1 for all positive predictors. For example, if two genes are expressed in the same *C. elegans* cell group, this feature cannot mean that the two genes are less likely to interact ($L<1$), regardless of the generality of the cell group term.

The second checkpoint examines the relationship of L and v. For the text annotation data where v is term usage frequency, we expect that the lower the term usage frequency, the more specific the term, and thus the higher L should be. Our analysis revealed that this expectation was true for all data sets with the exception of the *C. elegans* phenotype data (Fig. S1A). The exception is probably due to the fact that, unlike other data sets that were professionally curated, the *C. elegans* phenotype data were directly contributed by authors who published the phenotypes. We therefore changed our scoring scheme to compute L for each phenotype. For example, we computed L for gene pairs with both genes having the Egl (egg laying defect) phenotype, L(Egl-Egl). Similarly, we computed L(Lvl-Lvl) (Lvl: larval lethal), and so on. For phenotypes where there were not enough data in the training set to compute, we used L(same phenotype=true), which is the likelihood ratio of interaction if two genes simply have the same phenotype.

For microarray data, a common understanding is that two genes are more likely to interact if their expressions are strongly correlated. Therefore, we expect a higher value of L as the Pearson correlation value (v) approaches 1 (strong correlation) and a lower L value as the correlation approaches 0 (no correlation). With this inspection rule applied, we excluded the *C. elegans* Affymetrix data set and the *D. melanogaster* microarray data set since they failed to display any relationship of L and v (Fig. S1B).

Score integration

For different data sets within the same predictor (Table S1), we took the maximum L of all data sets to consolidate them into one predictor score. For example, if two genes have the same *C. elegans* expression annotated in cell, cell group, and anatomy, the final L for the predictor "*C. elegans* expression" is the maximum value of all three L scores, $L_{Ce\ \text{expression}} = \max\{L_{\text{cell}}, L_{\text{cell group}}, L_{\text{anatomy}}\}$.

The following two methods were tested to integrate L from different predictors.

1) naïve Bayesian network

The naïve Bayesian network model assumes that all predictors are independent, thus, the overall likelihood ratio L is

$$L = \prod_{i=1}^{n} L_i$$

where L_i is the likelihood ratio for the i^{th} predictor. L is the final output.

2) Logistic regression

Let O be the odds of a gene pair being an interacting pair given that the predictor value is v, then according to Bayes' rule, L is proportional to O as shown below.

$$\begin{aligned} O &= \frac{P(pos \mid predictor = v)}{P(neg \mid predictor = v)} \\ &= \frac{P(predictor = v \mid pos) \times P(pos) / P(predictor = v)}{P(predictor = v \mid neg) \times P(neg) / P(predictor = v)} \\ &= L \times \frac{P(pos)}{P(neg)} \end{aligned}$$

Here, $P(pos)$ and $P(neg)$ denote the frequency of interacting pairs and non-interacting pairs among all possible gene combinations, and can be considered constants. Logistic regression uses a weighted sum to integrate scores into overall odds:

$$\ln O = \ln \frac{p}{1-p} = c + \sum_{i=1}^{n} a_i \ln L_i$$

where c and a_i are constants determined by fitting of the training set, and p is used as the final output. Gene pairs with no annotation data were assigned $\ln L_i = 0$ for that predictor.

When calibrating logistic regression parameters for *D. melanogaster* and *S. cerevisae* predictors, we believe that it is more appropriate if we only include gene pairs from the training set with both genes having orthologs in that organism. Ideally, we would like to use gene pairs that have both *D. melanogaster* and *S. cerevisae* orthologs to conduct one regression for all the predictors. However, with our current training set size, this approach would leave us with only a few hundred data points in the training set to conduct a regression for as many as 14 predictors. Therefore, we conducted a serial logistic regression. We first used logistic regression to integrate predictors in each organism, and then combined scores from the three organisms. The serial logistic regression strategy increased the number of usable entries in the training set for each fitting process. For example, we can include gene pairs that have no *S. cerevisae* orthologs for the regression of *D. melanogaster* predictors. With a bigger training set and fewer predictors for each process, we can model the data better for each organism and thus improve the overall accuracy. The resulting score integration scheme is (predictor name indicates the ln L value of that predictor):

$$\ln\frac{p}{1-p} = 0.8\text{Ce} + 0.8\text{Dm} + 0.8\text{Sc} + 0.02 \quad \text{where}$$

Ce = 0.8Ce_expression + Ce_phenotype + 0.6Ce_process + 0.2Ce_microarray + 0.2

Dm = 0.3Dm_expression + 0.2Dm_phenotype + 0.7Dm_process + 0.4Dm_interaction + 0.1

Sc = 0.4Sc_localization + 0.6Sc_phenotype + 0.3Sc_process + 0.6Sc_interaction +1.1Sc_microarray − 0.3.

SUPPORTING RESULTS

Method comparison

1) Naïve Bayesian network *vs.* Logistic regression

We compared these two score integration methods based on their performance on the training set and the testing set (Fig. S2). We define the prediction accuracy as the percentage of true positives among all predictions (correct predictions/all predictions, or 1 − false positive rate); and the prediction sensitivity as the percentage of positives recovered by the prediction (correct predictions/all positives, or 1 − false negative rate). For the training set, at a given sensitivity, the logistic regression method provided slightly higher accuracy, and thus outperformed the naïve Bayesian network model. For the testing set, the two methods provided overall comparable performances, with small differences in performance at different zones (Fig. S2). If we use a cutoff of 0.9 for the logistic regression method and a cutoff of 600 (a value used in yeast studies (*19*)) for the naïve Bayesian network model, then performances of the two methods are similar (Fig. S2).

2) Including penalty scores

Consider this example: if two worm genes both have annotated cell group expression, but the expression patterns do not overlap, should we assign a penalty score ($L < 1$) or a neutral score ($L = 1$) in such case? We computed two scoring schemes using our training sets. In one scheme, we computed the penalty score for all predictors; in the other, we only included positive predictors. We then compared the performance of the two schemes in both the training set and the testing set using the logistic regression model (Fig. S3).

The results showed that if we included penalty scores, the performance was better for the training set, but not for the testing set (Fig. S3). The diminished performance on the testing set indicates that the better performance on the training set is an artifact of over-fitting. We decided not to include penalty predictor scores since they did not improve the prediction performance. One possible explanation for the failure of penalty predictor scores is the limitation of current knowledge (as annotations in the databases). For example, the existing annotated gene expression patterns are likely to be incomplete because some of the gene expression patterns are still unpublished, un-annotated, or undiscovered because of experimental limits in detection. Therefore, if two genes do not have an overlapping expression pattern, it is more accurate to classify their status as unknown.

Cutoff value determination

While one can vary the cutoff value to determine the stringency of predictions, it is desirable to have an estimation of the accuracy and the sensitivity of the genome-wide predictions at different cutoff values. However, the performance data in figures S2 and S3 are only valid for method comparisons and do not reflect the genome-wide performance due to different data compositions: there are many more negatives than positives in the genome whereas there are about equal numbers of the negatives and positives in our training set and the testing set. Lee *et al*. used the correlation of GO annotations and their predictions to evaluate the prediction performance (*20*). We could not use the same approach as we included GO as one of our predictors.

We thus examined the percentage of gene pair combinations covered for each cutoff value in order to estimate the specificity of our predictions and to decide appropriate threshold values (Fig. S4). We focused on the 2,254 genes in the predicted interaction network at the cutoff of 0.9. There are 2,539,131 possible pairwise combinations of these genes. Since our predictors are all positive predictors ($\ln L > 0$), the minimum final score is 0.5. Therefore, if we use 0.5 as cutoff, we get all (100%) possible gene pairs (Fig. S4). As we increase the threshold value, a smaller percentage of the gene pairs will be selected. At the cutoff score of 0.9, 0.72% of all gene pairs is selected (Fig. S4), which we believe is a reasonable specificity.

Power-law distribution of node degrees

Many known networks of protein-protein interactions (*11, 21-25*) and yeast gene-gene interactions (*26*) exhibit a power-law distribution (linear plot on log-log scale) of node degrees (number of adjacent connections). We examined our predicted genetic interaction network (obtained at cutoff 0.9), and were encouraged to discover that it also displays such a feature (Fig. S5).

Profile of predicted interactions

Among the 21,646 genes in *C. elegans*, 9,809 genes have at least one annotation in at least one of our thirteen predictors besides the *C. elegans* microarray data. Thus, the scope of our prediction system is limited to these genes. Among these 9,809 genes, 5,655 genes have orthologs in either *S. cerevisiae* or *D. melanogaster*, 2,127 of which having orthologs in both organisms. Since we depend on cross-species data for our prediction, genes with orthologs are expected to be enriched in our predictions. Our predicted network at cutoff score of 0.9 covered 2,254 genes and 18,183 interactions. We found that as high as 98% of the gene pairs in our predictions have orthologous gene pairs in either fly or yeast (Table S2).

To provide an overview of the properties of our predictions, we also examined other genetic features (Table S2). As a result of our computational method, the predicted gene pairs are enriched with features such as identical expression and phenotypes (Table S2). The top three dominant features of the predicted interacting genes are: their yeast orthologs have identical phenotypes, the yeast ortholog gene products have identical subcellular localizations, and they have yeast interlogs (Table S2). However, because predictors using yeast data are relatively weak predictors (Fig. S1), the highest possible score with these three features combined is 0.8, not sufficient to reach the cutoff of 0.9.

Therefore, additional features are required for genes to be predicted as interacting pairs. These data indicate that the final predictions relied on multiple predictors rather than a few dominant strong predictors.

Experimental testing of *let-60* interactions

Eighty-seven genes were predicted to interact with *let-60*. Twelve of them, *let-23*, *lin-12*, *mek-2*, *egl-15*, *dpy-22*, *sos-1*, *dpl-1*, *ptp-2*, *gap-1*, *ksr-1*, *let-92*, and *sur-6*, are consistent with our training set positives (see WormBase annotation). Five genes not in our training set have been reported in the literature as interacting with *let-60*; these genes are *apx-1* (*27*), *dsl-4* (*27*), *smo-1* (*28*), *hda-1* (*29*), and *mab-5* (*30*). We decided to test the remaining 70 new predictions by RNAi. The MRC geneservice RNAi library has clones for 54 of them, leaving 16 genes untested (C24A1.2, *ceh-9*, *cnb-1*, *daf-2*, *eya-1*, F10F2.1, *igcm-1*, K08F4.2, *psa-1*, *rho-1*, T04D1.4, *wrt-1*, Y111B2A.13, Y38F2AR.9, Y97E10AR.5, and Y48E1B.3). The RNAi results for the 54 genes are listed in Table S3. There are two RNAi clones in the library for the gene Y48G10A.3. We denote them as Y48G10A.3a and Y48G10A.3b. RNAi of five genes (*rpa-1*, C26E6.4, T05H4.6, *cdc-25.2*, and *cls-2*) caused early phenotypes in all progeny (Table S3), blocking any potential vulva phenotype. RNAi results of the remaining 49 genes are discussed in the main text.

As a negative control, we randomly selected 30 genes and applied the same assay. RNAi of four genes (W0102.1, C18E9.6, H37A05.1, F28D1.1) resulted in early phenotypes and we were unable to score VPC induction. Results of the remaining 26 genes are discussed in the main text.

We have also listed phenotypes observed at the level of gross morphology in Table S3. Some phenotypes such as Pvl (protruding vulva) were listed for wild-type animals but not *let-60*(*n1046*) animals because the Muv (multi-vulva) phenotype of *let-60*(*n1046*) blocks the visibility of these phenotypes. No genetic interaction can be detected at the low resolution gross morphology level. Quantifying VPC induction at high resolution allowed us to detect 12 new *let-60* interactors.

Experimental testing of *itr-1* interactions

The second gene we tested was *itr-1*. Sixteen genes were predicted to be *itr-1* interactors. Four of them, *sca-1*, *unc-68*, *gsa-1*, *unc-73*, although not in our training set as *itr-1* interacting genes, have been verified to function together with *itr-1* in neurodegeneration (reviewed by Driscoll and Gerstbrein (*31*)). Our RNAi library has clones for six of the remaining 12 genes and the detailed test results are listed in Table S4. Six genes (*kin-2*, *flr-4*, *trp-2*, *ncx-1*, *trp-4*, and *pmr-1*) remain untested as there was no RNAi clone in the library.

As is the case with the *let-60* interactions, no genetic interaction can be detected at the gross morphology level. In two cases (*egl-19* and *ccb-1*), the Egl (egg-laying defect) phenotype was listed under wild-type but not *itr-1*(*sa73*) animals. This is not an indication of genetic interactions; rather, it is because *itr-1*(*sa73*) animals are sterile, thus masking the occurrence of the Egl phenotype.

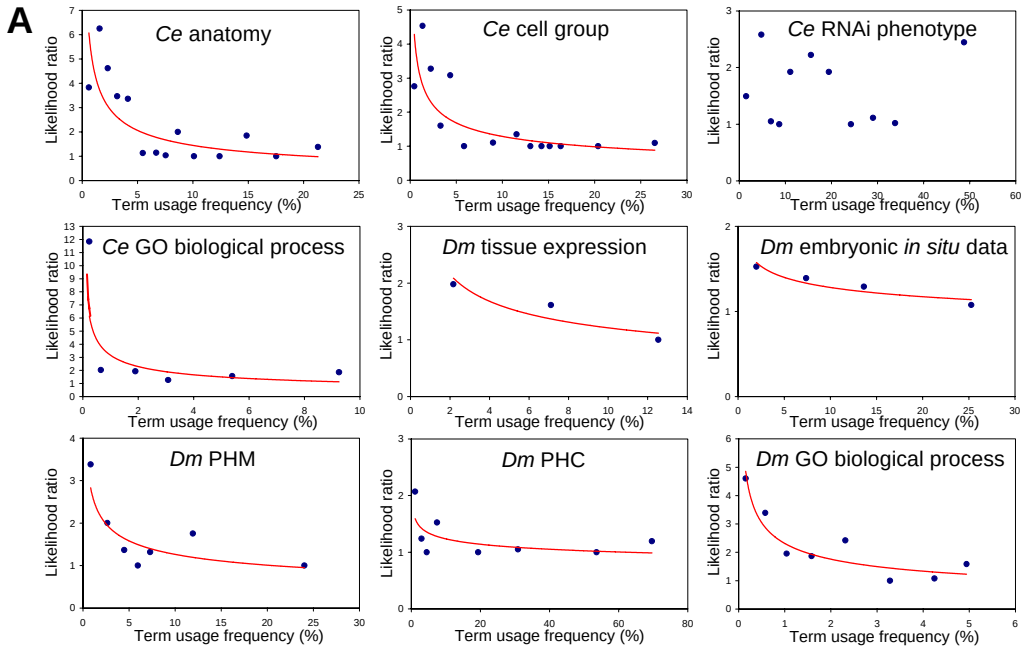
We do not believe that the observed suppression of *egl-19* and *ccb-1* on *itr-1*was a secondary effect resulting from the Egl phenotype. We took two extra precautions when scoring these animals. First, we reduced the strength of RNAi by exposing only one

generation of animals to RNAi bacteria. Instead of plating L4 animals to score their progeny, we put eggs on RNAi bacteria and scored the animals as they grew up. These animals had weaker phenotypes and appeared healthier than those resulting from the two-generation RNAi. Second, we scored young adults that had only a few eggs in them. Unable to lay eggs, Egl animals become bloated as they age. But at the early stage, they are indistinguishable from wild-type animals. The pharyngeal pumping rates listed in Table S4 were scored under these two conditions, and we still observed a significant suppression, indicating that the interaction is not an artifact due to a secondary effect.

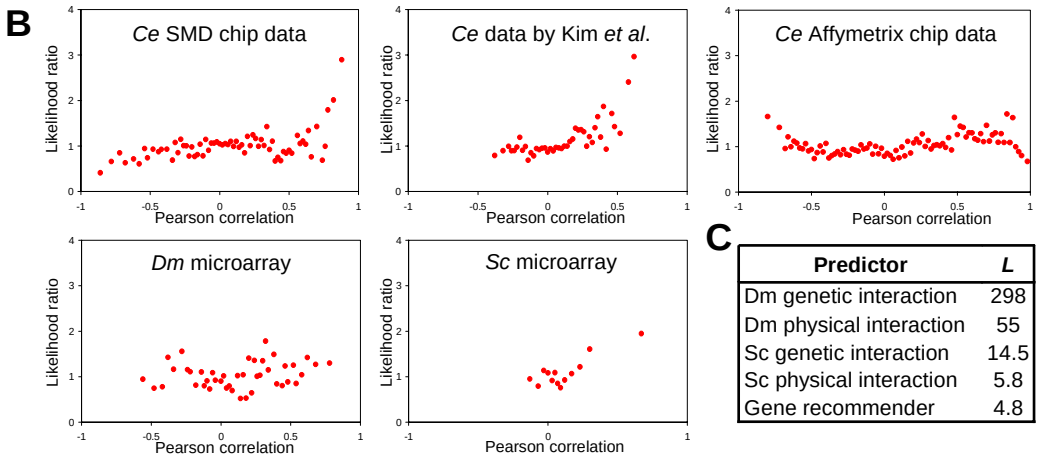
mca-3 RNAi resulted in a small percentage of very sick animals (Sck). These sick animals pumped much slower than the rest of the population. The overall average pumping rates of all animals are listed in Table S4. No significant difference was observed when we compared the pumping rates of sick *mca-3*(*RNAi*) animals against those of sick *itr-1*(*sa73*); *mac-3*(*RNAi*) animals, or the pumping rates of non-sick *mca-3*(*RNAi*) animals against those of non-sick *itr-1*(*sa73*); *mac-3*(*RNAi*) animals.

The predictions of *egl-19* and *ccb-1* as *itr-1* interactors relied on combining several weak predictors such as expression and phenotype from multiple species (Table S5, Fig. S6). There is no known genetic or physical interaction between orthologs of *itr-1* and *egl-19* or *ccb-1* in either fly or yeast.

SUPPORTING FIGURES



	v	*L*		*v*	*L*		*v*	*L*		*v*	*L*		*v*	*L*
	0-3.0	4.8		0-37	1.8		0-11	1.9		0-0.6	1.6		0-0.9	1.9
Ce cell	3.0-13.6	7.2	*Sc* localization	37-51	1.2	*Sc* phenotype	11-16	1.4	*Sc* GO process	0.6-7.1	1.3	*Sc* function	0.9-4.4	2.8
	>13.6	1		>51	1		>16	1		>7.1	1		>4.4	1



C

Predictor	*L*
Dm genetic interaction	298
Dm physical interaction	55
Sc genetic interaction	14.5
Sc physical interaction	5.8
Gene recommender	4.8

Figure S1 Predictor Likelihood ratios. **A**) Likelihood ratios for annotated text data. Red lines in graphs indicate power trendlines. Note that the *C. elegans* (*Ce)* RNAi phenotype data set does not display a trend of increased *L* value as *v* decreases, suggesting that we need an alternative scoring method for this data set. **B**) Likelihood ratios for microarray data. Note that the *Ce* Affymetrix chip data and *Dm* microarray data fail to display a positive correlation of *L* and *v*. They were thus excluded from the predictor data sets. **C**) Likelihood ratios for interaction data. Abbreviations: *Ce*, *C. elegans*; *Dm*, *D. melanogaster*; *Sc*, *S. cerevisiae*.

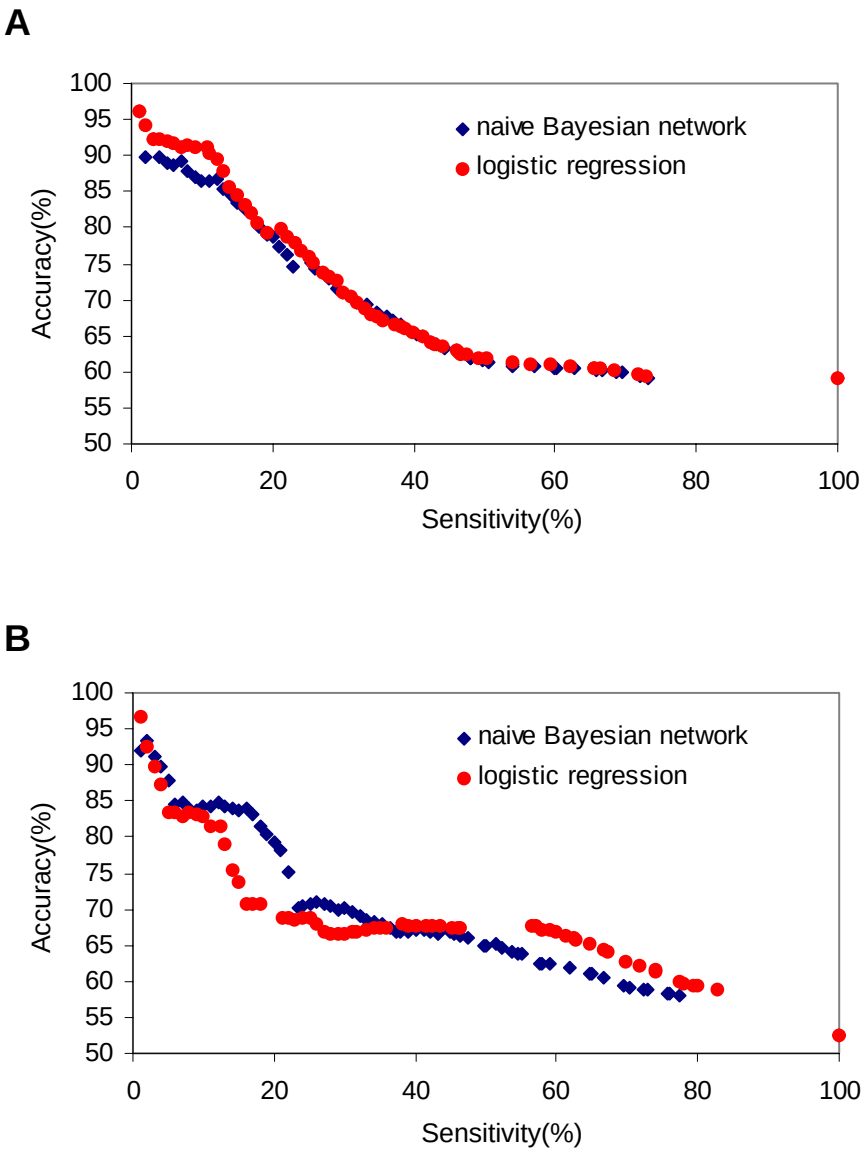


Figure S2 Performance comparison of the naïve Bayesian network model and the logistic regression method. The prediction sensitivity and the accuracy are used to compare the performance of the two methods on the training set data (A) and the testing set data (B). Red dots represent the performance of the naïve Bayesian network method as the threshold value varies. Blue squares represent the performance of the logistic regression method.

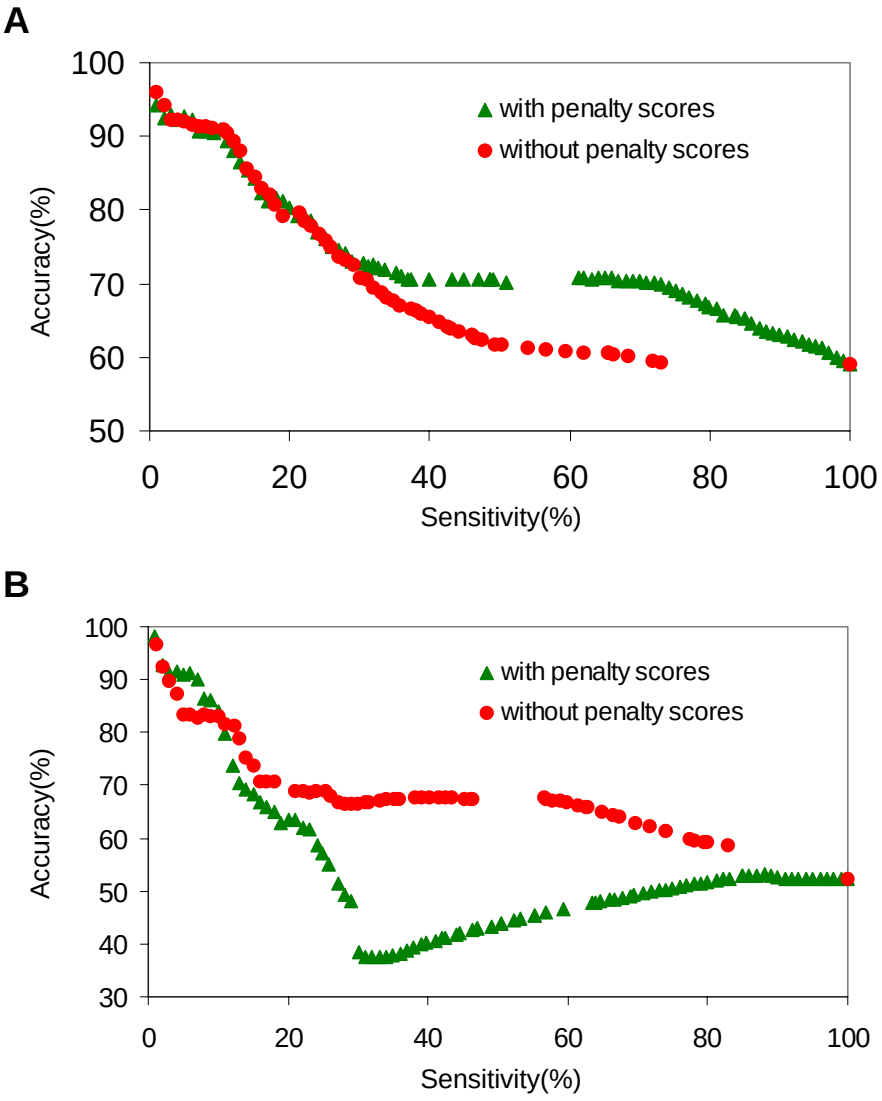


Figure S3 Performance comparison of logistic regression with and without penalty scores. The sensitivity and accuracy plot comparing the performance of the two methods on the training set data (A) and the testing set data (B). Red dots represent the performance of the method without penalty scores. Green triangles represent the performance of the computation with penalty scores.

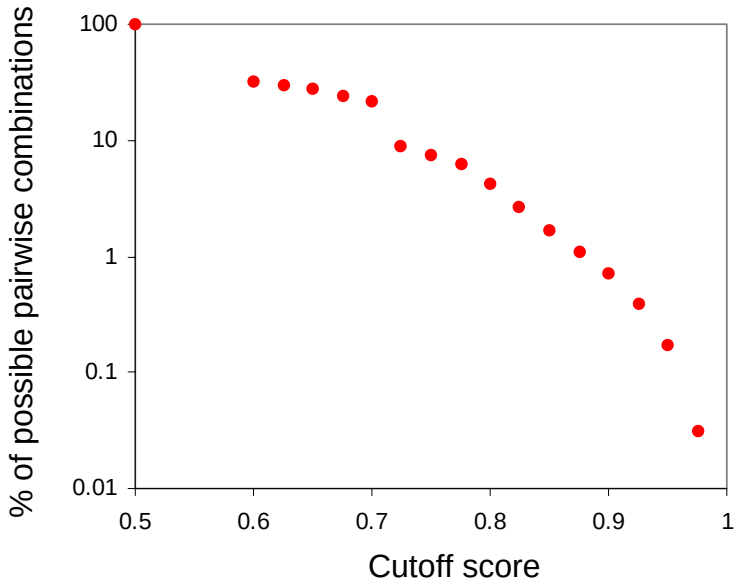


Figure S4 Cutoff values and prediction specificity. X-axis: cutoff score for the predictions; Y-axis: percentage ratio of the number of predicted interactions over the number of all possible combinations of gene pairs, focusing on the 2,254 genes in the predicted interaction network at cutoff 0.9.

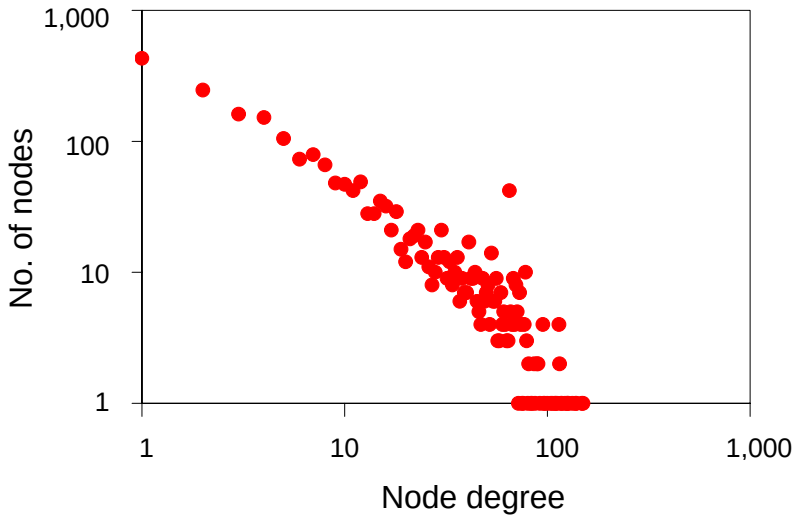


Figure S5 Distribution of node degrees. X- and Y-axis are in log scale.

Gene name: (example: mec-7 or R13A5.9)	itr-1

Search

Rank	Gene Name	Gene Info	Score	Data
1	sca-1	Gene Page at Wormbase	0.9830	Evidence for Prediction
2	unc-68	Gene Page at Wormbase	0.9761	Evidence for Prediction
3	egl-19	Gene Page at Wormbase	0.9640	Evidence for Prediction
4	kin-2	Gene Page at Wormbase	0.9570	Evidence for Prediction
5	flr-4	Gene Page at Wormbase	0.9386	Evidence for Prediction
6	trp-2	Gene Page at Wormbase	0.9138	Evidence for Prediction
7	ncx-1	Gene Page at Wormbase	0.9101	Evidence for Prediction
8	trp-4	Gene Page at Wormbase	0.9068	Evidence for Prediction
9	ccb-1	Gene Page at Wormbase	0.9053	Evidence for Prediction

	itr-1	*ccb-1*
Worm Genes Source database and data collected: WormBase (expression, phenotype, process)	Wormbase	Wormbase
Fly Orthologs Source database and data collected: FlyBase (expression, phenotype, process, interaction) GRID (interaction)	FBgn0010051 Flybase , GRID	FBgn0015608 Flybase , GRID
Yeast Orthologs Source database and data collected: SGD (localization, phenotype, process, interaction) MIPS (localization, phenotype, process, interaction) GRID (interaction)	No yeast ortholog	No yeast ortholog

Feature	Feature Details	Feature Score
Same *C. elegans* expression	[cell group]: pharynx	0.0918
Same *C. elegans* biological process	go:0006816	2.4717
Same *D. melanogaster* biological process	go:0006816	1.2211
total	Naive Bayesian network model	4.401E1
total	Logistic regression model	0.9053

Figure S6 Web interface for searching predicted interactions (simple search). On the first page, users can input the name of the gene, as in our example, *itr-1*, and click on the "Search" button. This brings up the second page of top 100 candidates (or all candidates if there are less than 100) that are predicted to interact with the gene of interest. Note that the scores may be lower than 0.9. From the search result page, users can click on "Evidence for Prediction" to examine the orthologs of the two genes, predictor details and their corresponding log likelihood ratios. As in our example, the evidence used for predicting the *itr-1* and *ccb-1* interaction is that they have the same expression pattern and GO process annotations. Reference links to original data sources (*e.g.* WormBase, FlyBase) are provided.

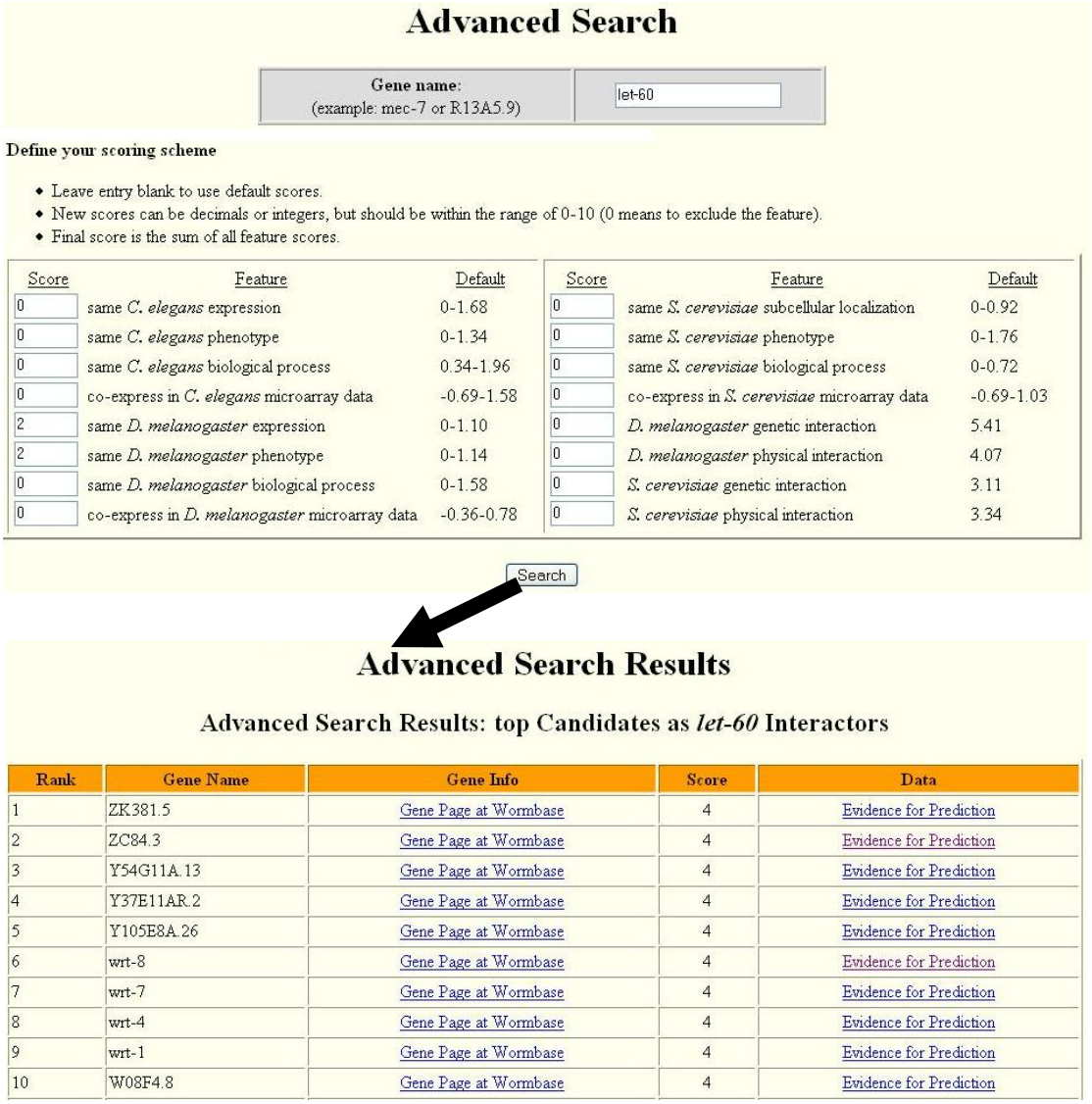


Figure S7 Web interface for cross-species genetic data search (advanced search). On the first page, users can input the name of the gene and specify their own scoring scheme. To customize the scoring scheme, users assign a score (0-10, with 0 indicating "ignorable" and 10 indicating "very important") to each predictor. The example shows how to conduct a query of "*C. elegans* genes whose *D. melanogaster* orthologs have the same expression pattern and the same phenotype as those of the *let-60* ortholog". We simply assign a positive score to the feature "same *D. melanogaster* expression" and the feature "same *D. melanogaster* phenotype", and a score of zero for other features. Similar to the simple search, the second page is the top 100 interactor candidates for the input gene and the third page (not shown) is the evidence used for the prediction.

SUPPORTING TABLES

Table S1 Predictor data and sources

Predictor[a]	Data Set	Source	Coverage (%)[b]
Ce expression	Cell	WormBase (www.wormbase.org)	2.7
	Cell group		9.1
	Anatomy		6.2
Ce phenotype	RNAi phenotype	WormBase	13.6
Ce process	GO biological process	GO (www.geneontology.org)	27.6
Ce microarray	SMD chip data	WormBase	80.8
	Affymetrix chip data	WormBase	78.4
	Data by Kim *et al*.	Kim *et al*. (*5*)	74.7
	Gene Recommender	Stuart *et al*. (*6*)	15.3
Dm expression	Tissue expression	FlyBase (www.flybase.org)	2.8
	Embryonic *in situ* data	BDGP (www.fruitfly.org)	5.5
Dm phenotype	PHM	FlyBase	4.4
	PHC		10.8
Dm process	GO biological process	GO	17.3
Dm microarray	GPL72	GEO (www.ncbi.nlm.nih.gov/geo/)	18.5
Dm interaction	Genetic interaction	FlyBase	2.5
	Physical interaction	GRID (biodata.mshri.on.ca/grid)	11.2
Sc localization	Localization	SGD (www.yeastgenome.org)	10.4
		MIPS	12.3
Sc phenotype	Phenotype	MIPS (mips.gsf.de/genre/proj/yeast)	5.2
	Viability	SGD	12.8
Sc function	GO biological process	GO	11.3
	Function catalogue	MIPS	11.9
Sc interaction	Genetic interaction	SGD, MIPS, GRID	6.0
	Physical interaction		10.4
Sc microarray	SGD data set	SGD	10.2

a. Ce, C. elegans; *Dm, D. melanogaster*; *Sc, S. cerevisiae.*

b. Percentage of *C. elegans* genome (21,646 genes) covered.

Table S2 Profile of predictions[a]

Feature	Positive training set	Negative training set	Predictions
same *Ce* expression	15.3	10.5	7.5
same *Ce* phenotype	12.6	8.9	61.4
same *Ce* GO process	7.8	4.2	39.0
Dm orthologs same expression	2.4	2.5[b]	7.6
Dm orthologs same phenotype	9.6	10.0[b]	22.6
Dm orthologs same GO process	5.1	3.1	44.2
Dm orthologs interact	1.7	0.1	12.2
Sc orthologs same localization	4.1	5.2[b]	73.0
Sc orthologs same phenotype	4.3	6.0[b]	76.6
Sc orthologs same GO process or same MIPS function catalogue	2.0	1.8	58.6
Sc orthologs interact	0.6	0.1	67.0

a. Numbers indicate the percentage of total gene pairs having the feature. Abbreviations: *Ce*, *C. elegans*; *Dm*, *D. melanogaster*; *Sc*, *S. cerevisiae*.

b. Note that a higher percentage in the negative set does not indicate a negative predictor; it was caused because more genes in the negative set were annotated with that feature. If we only include annotated genes, all features are enriched in the positive set (Fig. S1).

Table S3 RNAi testing of *let-60* interactors

RNAi clones	*let-60*(*n1046*)					wild-type		
	phenotype[a]	VPC induction	n	*p* - value[b]	Muv %	phenotype[a]	VPC induction	n
control	Muv	4.3±0.6	30		100		3.0	20
tax-6	Gro	3.5±0.5	20	3.0E-06	60	Gro	3.0	20
csn-5	Emb, Lvl, Gro, Stp	3.6±0.6	20	2.1E-05	60	Emb, Lvl, Gro, Stp	3.0	20
qua-1	Ste, Mlt, Gro, Adl, Clr, Unc	3.8±0.6	20	0.0022	80	Ste, Mlt, Gro, Adl, Clr, Unc	3.0	20
C01G8.9	Ste	3.8±0.4	20	3.5E-04	90	Ste, Pvl	3.0	20
pfn-3		3.8±0.8	20	0.016	70		3.0	20
nhr-41		3.9±0.7	20	0.018	75		3.0	20
C05D10.3		3.9±0.5	20	0.0050	95		3.0	20
Y48G10A.3b		4.0±0.6	20	0.031	85		3.0	20
dlg-1		4.0±0.5	20	0.018	95		3.0	20
tag-22		4.0±0.6	20	0.041	95		3.0	20
grd-11		4.0±0.6	20	0.054	90		3.0	20
W03F11.6		4.0±0.7	30	0.055	83		3.0	10
mig-15	Gro	4.0±0.7	20	0.082	80	Gro	3.0	20
taf-6.1		4.0±0.6	20	0.084	90		3.0	10
taf-1		4.0±0.8	23	0.14	83		3.0	20
lin-32		4.1±0.6	20	0.14	100		3.0	10
unc-55		4.1±0.4	20	0.12	100		3.0	10
Y59A8B.23		4.1±0.6	20	0.24	95		3.0	10
Y48G10A.3a		4.2±0.6	20	0.37	95		3.0	10
wrt-8		4.2±0.7	20	0.58	90		3.0	10
sqv-7		4.3±0.5	30	0.57	97		3.0	10
wrt-4		4.3±0.4	20	0.68	100		3.0	10
evl-20		4.3±0.5	21	0.76	95		3.0	20
C07H6.3		4.3±0.7	24	0.91	88		3.0	10
glp-1	F2 Emb	4.3±0.8	22	0.94	86	F2 Emb	3.0	10
unc-59		4.3±0.7	20	0.97	95		3.0	20
grd-1		4.3±0.5	20	0.96	95		3.0	10
wrt-7		4.3±0.7	20	0.97	90		3.0	10
hog-1		4.4±0.7	20	0.93	90		3.0	10
cdc-25.3		4.4±0.7	30	0.85	90		3.0	10
che-1		4.4±0.5	20	0.67	100		3.0	10
mom-5	Ste, Sck, Clr	4.4±0.8	20	0.67	100	Ste, Sck, Clr	3.0	20
Y53C12C.1		4.4±0.8	21	0.63	95		3.0	10
rnt-1		4.4±0.7	21	0.61	90		3.0	10
cki-1		4.5±0.5	20	0.47	95		3.0	10
let-413		4.5±0.6	20	0.50	95		3.0	20
taf-4		4.5±0.7	20	0.55	90		3.0	10
tig-2		4.5±0.7	20	0.55	90		3.0	10
tag-117		4.5±0.7	20	0.46	95		3.0	10
psa-4		4.5±0.5	22	0.36	100		3.0	10
T24H10.7		4.5±0.7	30	0.27	95		3.0	10
lin-48		4.5±0.5	20	0.21	100		3.0	10

Table S3 RNAi testing of *let-60* interactors (continued)

src-2	4.5±0.6	20	0.29	100		3.0	10
B0353.1	4.6±0.4	20	0.064	100		3.0	10
R05G6.10	4.6±0.7	20	0.17	100		3.0	10
T18D3.7	4.6±0.5	20	0.067	100		3.0	10
grd-2	4.7±0.7	20	0.11	95		3.0	10
ZC84.3	4.7±0.7	20	0.12	100		3.0	10
cdc-42	4.7±0.6	21	0.016	100	Vul, Pvl	2.5±0.4	20
cki-2	4.8±0.7	20	0.020	100		3.0	20
rpa-0	Emb, Lvl (100% penetrance)				Emb, Lvl (100% penetrance)		
C26E6.4	Emb (100% penetrance)				Emb (100% penetrance)		
T05H4.6	Emb, Lvl (100% penetrance)				Emb, Lvl (100% penetrance)		
cdc-25.2	emb, lvl (100% penetrance)				Emb, Lvl (100% penetrance)		
cls-2	emb, lvl (100% penetrance)				Emb, Lvl (100% penetrance)		
RNAi of low-score genes							
F59A2.4	4.1±0.6	20	0.091	90		3.0	10
K10H10.1	4.1±0.5	20	0.16	100		3.0	10
C04C3.3	4.2±0.7	20	0.41	90		3.0	10
F34D6.4	4.2±0.5	22	0.40	100		3.0	10
F34D10.2	4.2±0.5	20	0.51	95		3.0	10
C25H3.4	4.3±0.6	20	0.74	100		3.0	10
H27A23.1	4.3±0.8	20	0.77	90		3.0	10
Y54G11A.11	4.3±0.6	20	0.85	95		3.0	10
B0035.16	4.3±0.9	20	0.97	95		3.0	10
M03C11.4	4.4±0.6	20	0.92	95		3.0	10
C41C4.8	4.4±0.7	20	0.83	95		3.0	10
M01F1.5	4.4±0.7	20	0.62	95		3.0	10
ZK945.8	4.5±0.5	20	0.39	100		3.0	10
ZK643.2	4.5±0.6	20	0.42	100		3.0	10
F26E4.12	4.5±0.7	21	0.43	95		3.0	10
C16A3.7	4.5±0.7	22	0.43	95		3.0	10
C53A3.2	4.5±0.6	21	0.34	100		3.0	10
H14N18.4	4.5±0.6	20	0.28	100		3.0	10
W02D3.6	4.5±0.8	20	0.34	100		3.0	10
F08A8.4	4.5±0.5	20	0.23	100		3.0	10
C37H5.3	4.6±0.5	20	0.16	100		3.0	10
F28H6.3	4.6±0.5	20	0.17	100		3.0	10
R10E11.3	4.6±0.6	20	0.089	100		3.0	10
R04B5.5	4.7±0.7	20	0.093	100		3.0	10
B0491.1	4.7±0.6	20	0.061	100		3.0	10
C06A8.6	4.7±0.5	20	0.043	100		3.0	10
W0102.1 [b]	Lvl (100% penetrance)				Lvl (100% penetrance)		
C18E9.6 [b]	Emb (100% penetrance)				Emb (100% penetrance)		
H37A05.1 [b]	Emb (100% penetrance)				Emb (100% penetrance)		
F28D1.1 [b]	Lvl (100% penetrance)				Lvl (100% penetrance)		

a. Annotated using WormBase terms. Partial phenotypes unless noted otherwise.

b. *p*-value for student t-test against *let-60*(*n1046*) VPC induction under control RNAi.

Table S4 RNAi testing of *itr-1* interactors

RNAi clones	*itr-1*(*sa73*) phenotype[a]	pumps per minute	n	wild-type phenotype[a]	pumps per minute	n	*itr-1* / wild-type [b]	*p*-value[c]
Control	Ste	181±12	23		212±18	55	0.85±0.06	
cca-1		172±15	20		218±16	20	0.79±0.07	0.003
R05C11.3		178±26	20		216±13	20	0.83±0.12	0.4
nhr-85		173±30	22		216±13	21	0.84±0.14	0.7
mca-3	Sck	171±26	22	Sck	180±34	20	0.95±0.15	0.005
egl-19		203±14	27	Egl	203±18	35	1.00±0.07	6.3E-11
ccb-1		190±18	22	Egl	165±15	20	1.15±0.11	7.3E-13
RNAi of low-score genes								
W02D3.6		215±12	20		171±18	18	0.79±0.09	0.02
F34D10.2		204±18	20		165±14	14	0.81±0.07	0.03
F34D6.4		214±15	20		173±18	18	0.81±0.08	0.06
H27A22.1		206±11	20		167±14	14	0.81±0.07	0.04
C53A3.2		209±13	20		171±14	14	0.82±0.07	0.07
R10E11.3		211±15	20		173±17	17	0.82±0.08	0.14
K10H10.1		208±14	20		173±12	12	0.83±0.06	0.24
B0035.16		205±16	20		171±19	19	0.84±0.09	0.53
ZK954.8		205±15	20		172±13	13	0.84±0.06	0.45

a. Phenotypes annotated as in WormBase.

b. Normalized rates were computed by dividing each *itr-1*(*sa73*) rate under an RNAi with the average wild-type rate under that RNAi.

c. *p*-value for student t-test against normalized rate under control RNAi.

Table S5 Evidence for predicting the *egl-19* and *itr-1* interaction

Predictor	Shared features of the two gene	Log likelihood ratio (ln *L*)
Ce expression	[cell group]: M4, neurons, ventral nerve cord, pharynx, M5 [cell anatomy]: ventral cord neuron, M4, pm5	1.8
Ce process	[GO]: ion transport, calcium ion transport	2.5
Ce microarray	[Data by Kim *et al*.]: Pearson correlation=0.2	0.1
Dm phenotype	[PHC]: lethal, viable, flightless	0.6
Dm process	[GO]: calcium ion transport, cation transport	1.2
Final score		**0.96**

REFERENCES

1. E. M. Schwarz *et al.*, *Nucleic Acids Res* **34**, D475 (2006).
2. K. Hashimoto *et al.*, *Glycobiology* (2005).
3. J. S. Bader, A. Chaudhuri, J. M. Rothberg, J. Chant, *Nat Biotechnol* **22**, 78 (2004).
4. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (2000).
5. S. K. Kim *et al.*, *Science* **293**, 2087 (2001).
6. J. M. Stuart, E. Segal, D. Koller, S. K. Kim, *Science* **302**, 249 (2003).
7. R. A. Drysdale, M. A. Crosby, *Nucleic Acids Res* **33**, D390 (2005).
8. P. Tomancak *et al.*, *Genome Biol* **3**, RESEARCH0088 (2002).
9. T. Barrett *et al.*, *Nucleic Acids Res* **33**, D562 (2005).
10. B. J. Breitkreutz, C. Stark, M. Tyers, *Genome Biol* **4**, R23 (2003).
11. L. Giot *et al.*, *Science* **302**, 1727 (2003).
12. R. Balakrishnan *et al.*, *Nucleic Acids Res* **33**, D374 (2005).
13. U. Guldener *et al.*, *Nucleic Acids Res* **33**, D364 (2005).
14. R. S. Kamath *et al.*, *Nature* **421**, 220 (2003).
15. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403 (1990).
16. R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).
17. R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (2003).
18. K. P. O'Brien, M. Remm, E. L. Sonnhammer, *Nucleic Acids Res* **33**, D476 (2005).
19. R. Jansen *et al.*, *Science* **302**, 449 (2003).
20. I. Lee, S. V. Date, A. T. Adai, E. M. Marcotte, *Science* **306**, 1555 (2004).
21. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
22. T. Ito *et al.*, *Proc Natl Acad Sci U S A* **98**, 4569 (2001).
23. Y. Ho *et al.*, *Nature* **415**, 180 (2002).
24. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
25. S. Li *et al.*, *Science* **303**, 540 (2004).
26. A. H. Tong *et al.*, *Science* **303**, 808 (2004).
27. N. Chen, I. Greenwald, *Dev Cell* **6**, 183 (2004).
28. E. R. Leight, D. Glossip, K. Kornfeld, *Development* **132**, 1047 (2005).
29. Z. Chen, M. Han, *Curr Biol* **11**, 1874 (2001).
30. T. R. Clandinin, W. S. Katz, P. W. Sternberg, *Dev Biol* **182**, 150 (1997).
31. M. Driscoll, B. Gerstbrein, *Nat Rev Genet* **4**, 181 (2003).