Genome-Wide Prediction of C. elegans Genetic Interactions

Weiwei Zhong, *et al.*

Science **311**, 1481 (2006);
DOI: 10.1126/science.1123287

21. Y. Tagawa, P. O. Kanold, M. Majdan, C. J. Shatz, *Nat. Neurosci.* **8**, 380 (2005).
22. See supporting data on *Science* Online.
23. Z. Zou, F. Li, L. B. Buck, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7724 (2005).
24. K. R. Illig, L. B. Haberly, *J. Comp. Neurol.* **457**, 361 (2003).
25. P. Giraudet, F. Berthommier, M. Chaput, *J. Neurophysiol.* **88**, 829 (2002).
26. D. A. Wilson, *J. Neurophysiol.* **90**, 65 (2003).
27. P. Duchamp-Viret, A. Duchamp, M. A. Chaput, *Eur. J. Neurosci.* **18**, 2690 (2003).
28. D.-Y. Lin, S. Z. Zhang, E. Block, L. C. Katz, *Nature* **434**, 470 (2005).
29. J. Kang, J. Caprio, *J. Neurophysiol.* **74**, 1435 (1995).
30. R. Tabor, E. Yaksi, J. M. Weislogel, R. W. Friedrich, *J. Neurosci.* **24**, 6611 (2004).
31. L. B. Haberly, *Chem. Senses* **26**, 551 (2001).
32. D. A. Wilson, *Chem. Senses* **26**, 577 (2001).
33. D. G. Laing, G. W. Francis, *Physiol. Behav.* **46**, 809 (1989).
34. C. Margot, personal communication.

35. We thank R. Childs and K. Wilson for expert technical assistance and members of the Buck laboratory for helpful discussions. We also thank D. Laing and C. Margot for information on odorant mixture effects in humans, and International Flavors and Fragrances, Inc., for odor chemicals. This project was supported by the Howard Hughes Medical Institute and by grants from the NIH (National Institute on Deafness and Other Communication Disorders) and the Department of Defense (Army Research Office).

Supporting Online Material
www.sciencemag.org/cgi/content/full/311/5766/1477/DC1
Materials and Methods
References

9 January 2006; accepted 1 February 2006
10.1126/science.1124755

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

Weiwei Zhong and Paul W. Sternberg*

To obtain a global view of functional interactions among genes in a metazoan genome, we computationally integrated interactome data, gene expression data, phenotype data, and functional annotation data from three model organisms—*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster*—and predicted genome-wide genetic interactions in *C. elegans*. The resulting genetic interaction network (consisting of 18,183 interactions) provides a framework for system-level understanding of gene functions. We experimentally tested the predicted interactions for two human disease-related genes and identified 14 new modifiers.

An essential part of understanding how a genome specifies the properties of an organism is elucidating interactions among its genes. Such interactions include protein-protein physical interactions as well as gene-gene and protein-gene interactions. One method to identify genetic interactions is by modifier screens (e.g., synthetic lethal screens). However, this process requires easily detectable phenotypes. Metazoan biological processes often involve phenotypes too complex to score in large-scale screens, and thus candidate genes are often tested. Unfortunately, a genome of 20,000 genes has as many as 200 million pairwise combinations, posing a formidable challenge.

Relative to randomly paired genes, functionally interacting genes are more likely to have similar expression patterns and phenotypes; thus, statistically combining these genetic features might lead to reliable predictions of functional interactions. Several such computational approaches have been applied to *S. cerevisiae* (*1*–*5*). In metazoans, which have more genes and more complex genetic interactions, network constructions have focused on either protein-protein interactions (*6*) or a specific biological process (*7*, *8*), and thus they represent a subset of all genetic interactions. Unlike yeast data sets, most metazoan genetic data sets are incomplete (*9*). For example, only 292 *C. elegans* genes have complete annotations of anatomical expression, phenotype, and biological process in Gene Ontology (GO) (*10*) [WormBase WS140 (*11*)]. We therefore decided to incorporate information from multiple organisms. As gene functions are often conserved at the molecular level, we reasoned that if two genes possess features indicating a genetic interaction, their orthologous genes are also likely to be functionally linked. Pooling information across species might enable detection of interactions even if the genetic data for one organism are incomplete.

Data sets from different sources have different intrinsic error rates and different predictive strengths. A solid statistical model is thus essential for producing reliable predictions. Two types of methods, unions and Bayesian networks, have been used to integrate heterogeneous data [e.g., (*1*, *3*)]. Bayesian network models are preferable to unions because they weight data sets according to their reliability (*3*), but current methods often assume that data sets are independent [naïve Bayesian network, e.g., (*3*, *6*)]. We thus used logistic regression, a classical method for predicting binary outcomes (interaction versus no interaction), which provides performance comparable to that of naïve Bayesian networks (fig. S2) but relaxes the requirement of data independency, as it uses a weighted sum for data integration (*9*).

To calibrate the parameters of the computational system, we constructed a training set for *C. elegans* genetic interactions. Our positives were 1816 genetic interaction pairs curated from the literature (*11*) and 2878 physical interaction pairs identified by yeast two-hybrid screens (*12*). Yeast two-hybrid data may not be as accurate as results of small-scale interaction studies in the literature. Also, two-hybrid data describe physical rather than genetic interactions, which include both direct and indirect interactions. We included the two-hybrid data in our training set because they greatly increased the size of our training set and provided unbiased coverage (the literature-curated data were often biased toward evolutionarily conserved genetic interactions).

Negatives for metazoan genetic interactions are more difficult to define. Proteins with different subcellular localizations have been used as negatives for physical interactions (*3*, *6*); however, genetic interactions (for example, cell signaling pathways) do not require two gene products to colocalize. Genes annotated to function in different pathways have been used as negatives for yeast genetic interactions (*3*, *6*), but in metazoans, knowledge of interactions is so limited that if two genes are not annotated to function in the same pathway, their relation should be considered unknown rather than noninteracting. We reasoned that two genes are less likely to be an interacting pair if a double mutant of these genes exists and there is no reported interaction. We thus took from WormBase 3296 pairs of linked cis markers used in genetic mapping experiments as our negatives. Although it is possible that these cis markers may interact in processes not examined during the mapping experiments, the probability that cis marker pairs are interacting genes should be lower than for other gene pairs.

Our computational algorithm started by mapping orthologous genes with the use of InParanoid (*13*). We then searched each *C. elegans* gene pair as well as its orthologous pairs in *D. melanogaster* and *S. cerevisiae* for five features: identical anatomical expression, phenotype, function annotation (e.g., biological process in GO), microarray coexpression, and the presence of interlogs (i.e., whether the *D. melanogaster* or *S. cerevisiae* orthologous gene pairs interact genetically or physically). We used likelihood ratios (*3*, *4*) to assign a weighted score to each feature. The likelihood ratio is defined as

$$L = \frac{P(v|pos)}{P(v|neg)} \qquad (1)$$

Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA.

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

where $P(v|pos)$ and $P(v|neg)$ are frequencies of gene pairs having the feature v (such as having a similar yeast phenotype) in the positives (*pos*) and the negatives (*neg*), respectively. A value of L greater than 1 indicates that the feature is enriched in interacting gene pairs, with higher scores indicating stronger predictive power of the feature. Gene pairs with stronger Pearson correlations (closer to 1) in microarray results get higher L-scores than pairs with weaker expression correlations; gene pairs that share specific expression, phenotype, or function an-

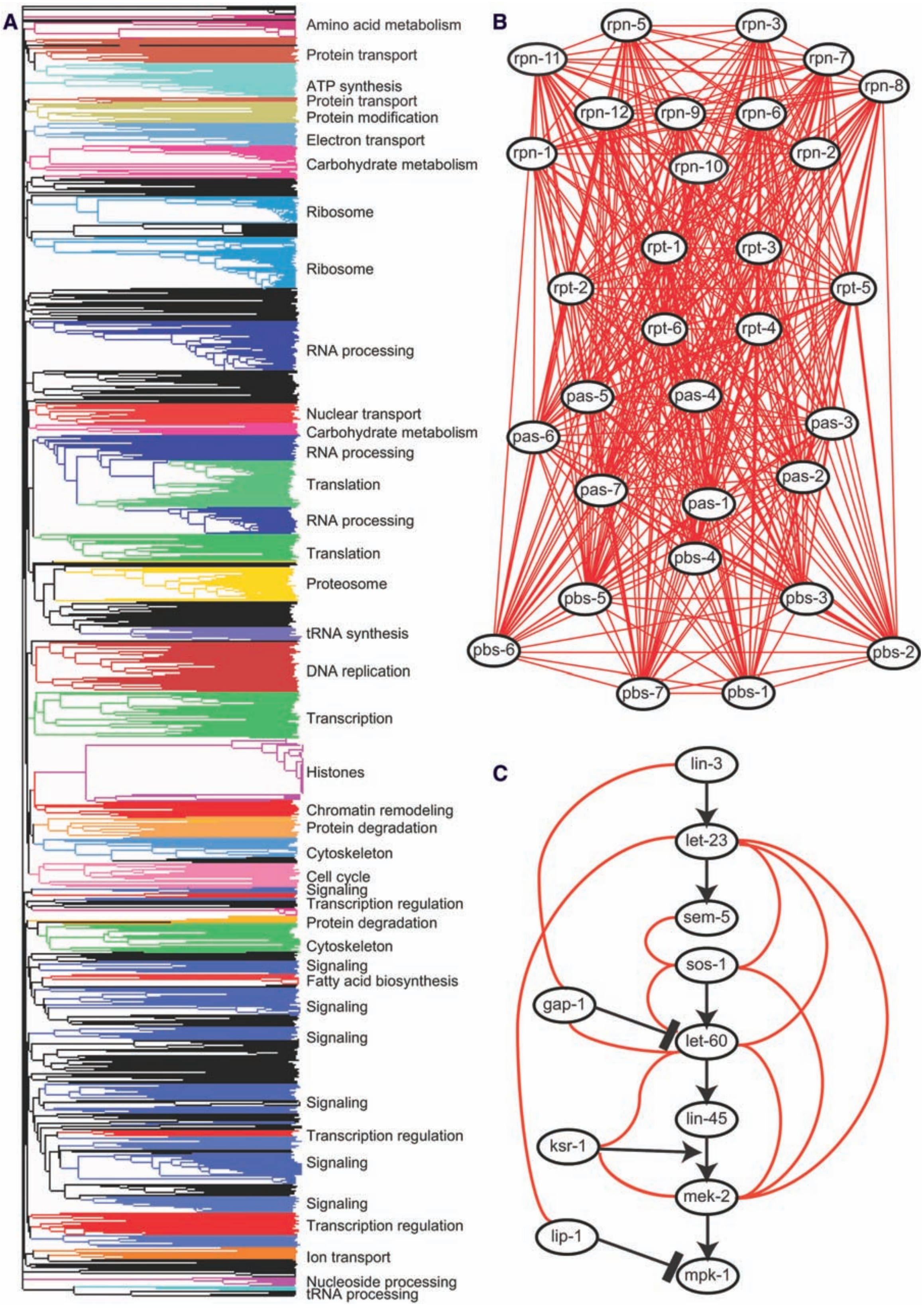

Fig. 1. Predicted genetic interactions. (**A**) Tree representation of genes clustered on the basis of their distances in the predicted network. Clusters were manually inspected. A cluster is colored and annotated if genes in the cluster share a common function; clusters with no common function and clusters with unknown functions are black. (**B** and **C**) Local views of the proteasome complex (B) and the epidermal growth factor receptor–ras–MAPK pathway (C). Red lines indicate predicted interactions (cutoff 0.9). Black lines indicate known genetic interactions, with arrows for activation and bars for inhibition.

notations get higher *L*-scores than pairs that share general annotations or share no annotations (fig. S1). Scores were integrated to estimate the overall probability of the two *C. elegans* genes interacting by

$$\ln \frac{p}{1-p} = c + \sum_{i=1}^{n} a_i \ln L_i \qquad (2)$$

where L_i is the likelihood ratio for the *i*th predictor, *c* and a_i are constants determined by fitting of the training set with the software R (www.R-project.org), and *p* is the final output score that varies between 0 and 1, with 1 indicating a genetically interacting pair and 0 indicating no interaction.

We applied a threshold of 0.9, which exceeds the maximum contribution that any single feature can achieve [see also (*9*)]. The resulting genetic interaction network consists of 2254 genes and 18,183 interactions. To explore the predicted network at a system level, we clustered the genes on the basis of their interaction partners (*14*). The clusters revealed a modular network organization; most clusters correlated with protein complexes or biological processes (Fig. 1A). Some functional modules (e.g., the signaling module) contained several clusters, consistent with multiple pathways. Close to 90% of the genes in the network were in one connected component.

To evaluate the performance of the computational system, we compared our predictions with known interactions to investigate the false negative rate. Our system predicted 414 interactions among the 31 proteasome components (Fig. 1B) and many interactions between the mitogen-activated protein kinase (MAPK) signaling pathway genes (Fig. 1C). Two genes, *lin-45* and *mpk-1*, were not connected to the pathway. These false negatives are due to incomplete annotation and incorrect ortholog mapping. For example, *lin-45* has only a sterile phenotype and a neuronal expression annotated in WormBase, although richer phenotype information exists. Also, InParanoid failed to identify orthologs of the two genes in either fly or yeast. Such false negatives will be automatically eliminated when model databases accumulate sufficient data and with better ortholog mappings.

To estimate the false positive rate, we examined all predicted interactors of *let-60*/*ras*, a component of the MAPK pathway. Of 87 predicted interactions for *let-60*, 12 are consistent with our training set, and 5 have been reported in the literature (*9*) although not included in our training set. To uncover novel interactions, we tested 49 of the 70 predictions by RNA interference (RNAi) on a *let-60* gain-of-function mutant, *let-60*(*n1046*) (*15*). *let-60*(*n1046*) animals have a multivulva (Muv) phenotype (Fig. 2A) caused by excess induction of vulval precursor cells (VPCs) (average of 4.3 induced VPCs versus 3.0 for wild type). RNAi against 12 candidate genes significantly affected VPC induction in *let-60* animals (Fig. 2B). Only one candidate, *cdc-42*, affected VPC induction in wild-type animals (table S3), indicating that the observed changes in the *let-60* animals were results of genetic interactions.

As a control, we examined interactions of *let-60*(*n1046*) with 26 randomly selected, low-scoring genes (<0.6) that, like the high-score candidates, have annotated genetic data and have yeast and fly orthologs. Only one such gene significantly affected *let-60*(*n1046*) VPC induction ($P = 0.04$) (Fig. 2C), consistent with the expected false positive rate.

We next experimentally tested the predictions for another essential gene, the inositol-1,4,5-trisphosphate (IP_3) receptor gene *itr-1*. There is no known genetic/physical interlog of *itr-1* with any gene in either fly or yeast, nor are the *C. elegans* data sufficient to predict *itr-1* interactions. All 16 *itr-1* interaction predictions relied on combining several weak predictors

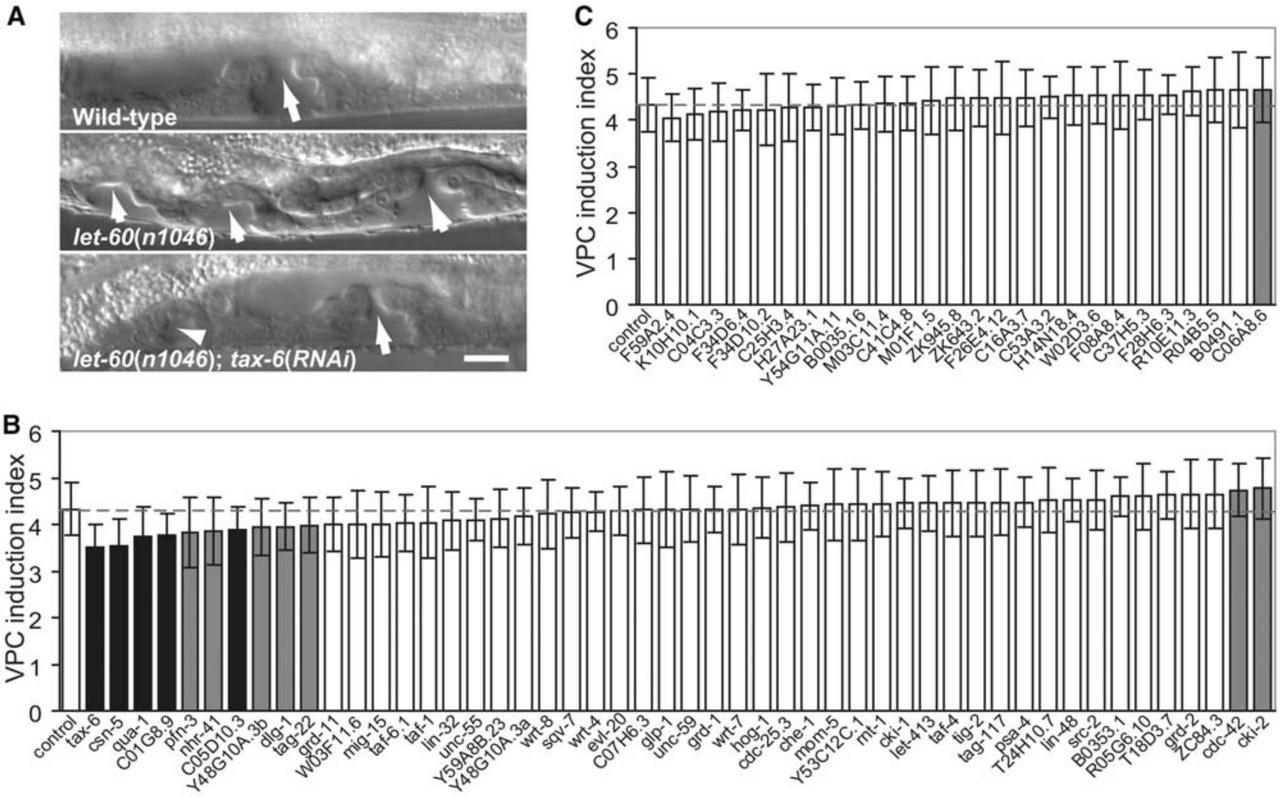

Fig. 2. *let-60* genetic interactions. (**A**) Nomarski images of the developing vulvae of wild-type, *let-60*(*n1046*), and *let-60*(*n1046*); *tax-6*(*RNAi*) L4 larvae, showing different degrees of VPC induction. Arrows indicate vulval invaginations. Scale bar, 10 μm. (**B** and **C**) VPC induction index of *let-60*(*n1046*) animals in response to RNAi of predicted interacting genes (B) and randomly selected low-score genes (C). Bars and error bars represent means and SD, respectively; $n \geq 20$ for each data point. Gray horizontal dashed lines indicate the average induction index under control conditions (RNAi with vector-alone bacteria). Data sets that have significant differences from controls are in black ($P < 0.01$) or gray ($P < 0.05$) based on Student's *t* test (two-tailed, unequal variance).

(*9*). Although not in the training set, 4 of 16 predicted *itr-1* interactions can be confirmed by published results (*9*). Of the remaining 12, we had RNAi clones against six genes, and we tested them on the weak loss-of-function mutant *itr-1*(*sa73*) (*16*). *itr-1*(*sa73*) animals have slow pharyngeal pumping (*17*) (181 ± 12 pumps per minute versus 212 ± 18 for wild type, $n > 20$). Because RNAi against several candidates affected pumping rates in both the wild-type strain and the *itr-1*(*sa73*) strain, we used normalized rates [the ratio of the *itr-1*(*sa73*) and wild-type rates] to distinguish genetic interactions from additive effects. The ratio will be ~0.85 (181/212) if the RNAi has the same effect on both strains. RNAi of two genes, *egl-19* and *ccb-1*, significantly suppressed the *itr-1*(*sa73*) pumping defect ($P <$ 0.001, Fig. 3A). By contrast, RNAi of nine randomly selected low-score genes did not affect *itr-1*(*sa73*) pumping (Fig. 3A). We confirmed the *egl-19* and *itr-1* interaction by performing *itr-1* RNAi on *egl-19* mutants. An *egl-19* gain-of-function mutant, *egl-19*(*n2368*), showed a much stronger phenotype in response to *itr-1* RNAi relative to a weak loss-of-function mutant, *egl-19*(*n582*). *itr-1* RNAi caused sterility, constipation, and reduced body size in all animals (Fig. 3B), but *egl-19*(*n2368*); *itr-1*(*RNAi*) animals became completely paralyzed (Fig. 3B), with a slower pharyngeal pumping rate than that of the *itr-1*(*RNAi*) or *egl-19*(*n582*); *itr-1*(*RNAi*) animals (Fig. 3C).

We thus verified 29/87 of the predicted *let-60* interactions and 6/16 of the predicted *itr-1* interactions. Excluding untested predictions, the prediction accuracy is 44% (29/66) for *let-60* and 60% (6/10) for *itr-1*. The actual prediction accuracy should be higher because not all genes are sensitive to RNAi and only one phenotype and one mutant allele were examined.

Our data are publicly available at http://tenaya.caltech.edu:8000/predict. Users can search for predicted interactions for any gene and see the evidence (fig. S6). The system serves as a cross-species genetic data search engine (fig. S7). Unlike manual searches of multiple databases, our computational approach provides statistical analysis to combine weak evidence and prioritize results, as well as fast and systematic data mining.

Our newly discovered genetic interactions may also have biological implications. For example, *tax-6* encodes the A-subunit of calcineurin, an upstream regulator of heterotrimeric guanine nucleotide–binding protein (G protein) signaling (*18*), and could regulate *let-60* activity in VPC induction through the G_q pathway, which is known to promote VPC induction (*19*). ITR-1 responds to the second messenger IP_3 to induce intracellular Ca^{2+} release. *egl-19* and *ccb-1* both encode L-type voltage-gated Ca^{2+} channels. The surprising antagonistic effect of *itr-1* and these genes suggests that they have different sites of action in pharyngeal pumping regulation: *itr-1* may mainly function in muscle or excitatory neurons, whereas *egl-19* and *ccb-1* may be required for inhibitory neuron function.

Both *let-60* and *itr-1* have been the subject of a number of modifier screens [e.g., (*20*, *21*)] conducted at low resolution to discover qualitative differences (e.g., Muv suppressed to wild type). Our quantitative analysis allowed us to detect interactions that would likely be missed by such screens. Quantitative assays are labor-intensive enough to be impractical for genome-wide screens. By computationally prioritizing candidates, it becomes feasible to effectively discover genetic interactions.

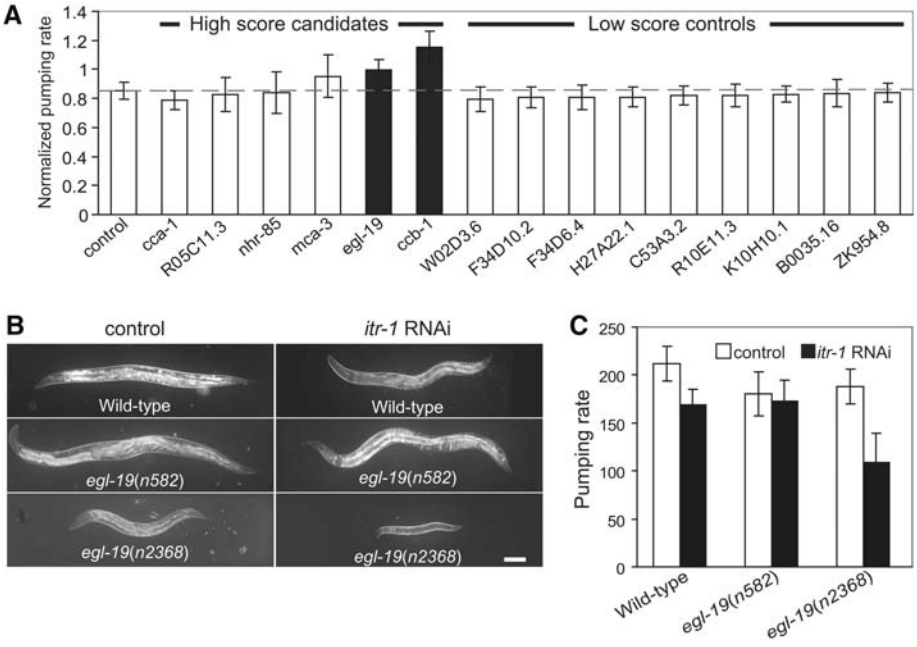

Fig. 3. *itr-1* genetic interactions. (**A**) Normalized pharyngeal pumping rate (*itr-1* rate divided by average wild-type rate) under various RNAi conditions. Bars and error bars represent means and SD, respectively. Under each RNAi condition, we scored the pumping rates of $n \geq 20$ wild-type animals to compute the average wild-type rate and $n \geq 20$ *itr-1* animals for normalized pharyngeal pumping rates. The gray horizontal dashed line indicates the average rate under vector-alone control RNAi; black bars represent data with significant differences from controls ($P < 0.001$). (**B**) Nomarski images of wild-type, loss-of-function *egl-19*(*n582*), and gain-of-function *egl-19*(*n2368*) animals under control and *itr-1* RNAi conditions. Scale bar, 100 μm. (**C**) Pharyngeal pumping rate (pumps per minute) under control and *itr-1* RNAi conditions. Error bars represent SD.

References and Notes

1. E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, D. Eisenberg, *Nature* **402**, 83 (1999).
2. O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8348 (2003).
3. R. Jansen *et al.*, *Science* **302**, 449 (2003).
4. I. Lee, S. V. Date, A. T. Adai, E. M. Marcotte, *Science* **306**, 1555 (2004).
5. S. L. Wong *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15682 (2004).
6. D. R. Rhodes *et al.*, *Nat. Biotechnol.* **23**, 951 (2005).
7. S. J. Boulton *et al.*, *Science* **295**, 127 (2002).
8. K. C. Gunsalus *et al.*, *Nature* **436**, 861 (2005).
9. See supporting material on *Science* Online.
10. M. Ashburner *et al.*, *Nat. Genet.* **25**, 25 (2000).
11. E. M. Schwarz *et al.*, *Nucleic Acids Res.* **34**, D475 (2006).
12. S. Li *et al.*, *Science* **303**, 540 (2004); published online 2 January 2004 (10.1126/science.1091403).
13. K. P. O'Brien, M. Remm, E. L. Sonnhammer, *Nucleic Acids Res.* **33**, D476 (2005).
14. A. Baudot *et al.*, *Bioinformatics* **22**, 248 (2006).
15. G. J. Beitel, S. G. Clark, H. R. Horvitz, *Nature* **348**, 503 (1990).
16. K. Iwasaki, D. W. Liu, J. H. Thomas, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10317 (1995).
17. D. S. Walker, N. J. Gower, S. Ly, G. L. Bradley, H. A. Baylis, *Mol. Biol. Cell* **13**, 1329 (2002).
18. J. Lee *et al.*, *J. Mol. Biol.* **344**, 585 (2004).
19. N. Moghal, L. R. Garcia, L. A. Khan, K. Iwasaki, P. W. Sternberg, *Development* **130**, 4553 (2003).
20. Y. Wu, M. Han, *Genes Dev.* **8**, 147 (1994).
21. T. R. Clandinin, J. A. DeModena, P. W. Sternberg, *Cell* **92**, 523 (1998).

22. We thank the *Caenorhabditis* Genetics Center for strains; WormBase staff for discussions; L. Chen for advice on statistical analysis; B. Perry for maintaining reagents; and C. Van Buskirk, E. Hallem, B. Hwang, M. Kato, E. Kipreos, R. Kishore, A. Petcherski, and E. Schwarz for critical reading of the manuscript. P.W.S. is an investigator and W.Z. is an associate of the Howard Hughes Medical Institute.

Supporting Online Material
www.sciencemag.org/cgi/content/full/311/5766/1481/DC1
Materials and Methods
Figs. S1 to S6
Tables S1 to S5
References

1 December 2005; accepted 26 January 2006
10.1126/science.1123287