

A Map of the Interactome Network of the Metazoan *C. elegans*

Siming Li,^{1*} Christopher M. Armstrong,^{1*} Nicolas Bertin,^{1*} Hui Ge,^{1*} Stuart Milstein,^{1*} Mike Boxem,^{1*} Pierre-Olivier Vidalain,^{1*} Jing-Dong J. Han,^{1*} Alban Chesneau,^{1,2*} Tong Hao,¹ Debra S. Goldberg,³ Ning Li,¹ Monica Martinez,¹ Jean-François Rual,^{1,4} Philippe Lamesch,^{1,4} Lai Xu,^{5†} Muneesh Tewari,¹ Sharyl L. Wong,³ Lan V. Zhang,³ Gabriel F. Berriz,³ Laurent Jacotot,^{1‡} Philippe Vaglio,^{1‡} Jérôme Reboul,^{1§} Tomoko Hirozane-Kishikawa,¹ Qianru Li,¹ Harrison W. Gabel,¹ Ahmed Elewa,^{1||} Bridget Baumgartner,⁵ Debra J. Rose,⁶ Haiyuan Yu,⁷ Stephanie Bosak,⁸ Reynaldo Sequerra,⁸ Andrew Fraser,⁹ Susan E. Mango,¹⁰ William M. Saxton,⁶ Susan Strome,⁶ Sander van den Heuvel,¹¹ Fabio Piano,¹² Jean Vandenhoute,⁴ Claude Sardet,² Mark Gerstein,⁷ Lynn Doucette-Stamm,⁸ Kristin C. Gunsalus,¹² J. Wade Harper,^{5†} Michael E. Cusick,¹ Frederick P. Roth,³ David E. Hill,^{1¶} Marc Vidal^{1#}

¹Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA. ²Institut de Génétique Moléculaire, Centre National de la Recherche Scientifique UMR 5535, 1919 Route de Mende, 34293 Montpellier Cedex 5, France. ³Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, MA 02115, USA. ⁴Unité de Recherche en Biologie Moléculaire, Facultés Notre-Dame de la Paix, 61 Rue de Bruxelles, 5000 Namur, Belgium. ⁵Verna and Marrs Department of Biochemistry and Molecular Biology, Program in Cell and Molecular Biology, Biophysics, and Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. ⁶Department of Biology, Indiana University, Jordan Hall 142, 1001 East Third Street, Bloomington, IN 47405, USA. ⁷Department of Molecular Biophysics and Biochemistry and Department of Computer Science, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA. ⁸Agencourt Bioscience Corporation, 100 Cummings Center, Suite 107G, Beverly, MA 01915, USA. ⁹The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. ¹⁰Huntsman Cancer Institute, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112, USA. ¹¹Massachusetts General Hospital Cancer Center, Building 149, 13th Street, Charlestown, MA 02129, USA. ¹²Department of Biology, New York University, 1009 Silver Building, 100 Washington Square East, New York, NY 10003, USA.

*These authors contributed equally to this work.

† ‡ § || ¶ Present addresses: †Department of Pathology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA. ‡Modul-Bio, 232 Boulevard Sainte-Marguerite, 13009 Marseille, France. §INSERM, Unité 119, Institut Paoli Calmettes, 13009 Marseille, France. ||Program in Gene Function and Expression, University of Massachusetts, 55 Lake Avenue, North Worcester, MA 01605, USA. ¶Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA.

#To whom correspondence should be addressed. E-mail: marc_vidal@dfci.harvard.edu.

To initiate studies on how protein-protein interaction (or “interactome”) networks relate to multicellular functions, we have mapped a large fraction of the *Caenorhabditis elegans* interactome network. Starting with a subset of metazoan-specific proteins, more than 4000 interactions were identified from high-throughput yeast two-hybrid (Y2H) screens. Independent co-affinity purification assays experimentally validated the overall quality of this Y2H data set. Together with already described Y2H interactions and interologs predicted in silico, the current version of the Worm Interactome (WI5) map contains ~5500 interactions. Topological and biological features of this interactome network, as well as its integration with phenome and transcriptome data sets, lead to numerous biological hypotheses.

To further understand biological processes it is important to consider protein functions in the context of complex molecular networks. The study of such networks requires the availability of proteome-wide protein-protein interaction, or “interactome,” maps. The yeast *Saccharomyces cerevisiae* has been used to develop a eukaryotic unicellular interactome map (1–6). *Caenorhabditis elegans* is an ideal model to study how protein networks relate to multicellularity. Here we investigate its interactome network with a high-throughput yeast two-hybrid system (HT-Y2H).

As Y2H baits, we selected a set of 3,024 worm predicted proteins that relate directly or indirectly to multicellular functions (7). Gateway-cloned open reading frames (ORFs) were available in the *C. elegans* ORFeome 1.1 (8) for 1,978 of these selected proteins. Of these, 81 auto-activated the Y2H *GAL1::HIS3* reporter gene as Gal4 DNA binding

domain fusions (DB-X), and 24 others conferred toxicity to yeast cells. The remaining 1,873 baits were screened against two different Gal4 activation domain libraries (AD-wrmcDNA and AD-ORFeome1.0), each with distinct yet complementary advantages (7).

We maximized the specificity of the Y2H system by applying stringent experimental and bioinformatic criteria (fig. S1). To eliminate interactions that originated from non-specific promoter activation, DB-X/AD-Y pairs were only considered if they activated at least two out of three different Gal4-responsive promoters. Positives were subsequently retested in fresh yeast cells and their AD-Y identities were determined with Interaction Sequence Tags (ISTs) obtained by sequencing the corresponding PCR products (9). The AD-Y reading frame was verified for each IST to avoid the recovery of out-of-frame peptides. In total, ~16,000 ISTs were obtained.

Having applied those criteria, we subdivided the interactions into three confidence classes (fig. S1): those that were found at least three times independently and for which the AD-Y junction is in frame ("Core-1," 858 interactions), those in frame found fewer than three times and that passed the retest ("Core-2," 1,299 interactions), and all other Y2H interactions found in our screens ("Non-Core," 1,892 interactions). The "Core" data set ("Core-1" and "Core-2") contains 2,157 high-confidence interactions between 502 DB-X baits and 1,039 AD-Y preys. After collapsing 22 interactions that occur in both DB-X/AD-Y and DB-Y/AD-X configurations, a total of 2,135 unique interactions are obtained (table S1). The "Non-Core" data set contains 1,892 interactions between 531 DB-X baits and 1,395 AD-Y preys. Altogether, Core and Non-Core constitute the "First-Pass" data set with a total of 4,027 distinct interactions. Out of 2,783 and 1,505 interactions found with AD-wrmcDNA and AD-ORFeome1.0 respectively, 239 interactions were identified with both libraries.

To estimate the coverage of the HT-Y2H data sets, the baits screened here were manually searched for known interactors in WormPD (10). This search gave rise to 108 interactions, referred to as the "literature" data set (table S1). The Core and Non-Core data sets recapitulated 8 and 2 interactions in this benchmark data set, respectively. Thus, our overall rate of coverage for the First-Pass data set is ~10% [(8+2)/108].

In order to evaluate the accuracy of the HT-Y2H data sets, we reasoned that interactions detected in two different binding assays are unlikely to be experimental false positives. A representative sample of Y2H interaction pairs from each of these three subsets (33 for Core-1, 62 for Core-2 and 48 for Non-Core) was randomly selected, and tested in a co-affinity purification (co-AP) glutathione-S-transferase (GST) pull-down assay (Fig. 1). Bait and prey ORFs were transiently

transfected into 293T cells as GST-bait and Myc-prey fusions, respectively. For potential interaction pairs where both proteins were expressed at detectable levels, the co-AP success rates were: 14/17 (82%) for Core-1, 17/29 (59%) for Core-2, and 8/23 (35%) for Non-Core (table S2). These data demonstrate that our three data sets contain a large proportion of highly reliable interactions and corroborate their expected relative qualities.

In addition to experimental screens, we also performed *in silico* searches for potentially conserved interactions, or "interologs," whose orthologous pairs are known to interact in (an) other species (9, 11). Starting from a high-confidence yeast interaction data set (7), reciprocal best hit BLAST searches (E-value <10⁻⁶) were performed against the worm predicted proteome. In all, 949 potential worm interologs were identified, constituting the "interologs" data set (7). In addition, the Y2H interactome maps that have been previously generated for individual biological processes (including vulval development, protein degradation, DNA damage response, and germline formation) (9, 12–14) were pooled to define the "scaffold" data set. The HT-Y2H, "literature," "interologs," and "scaffold" data sets were combined into Worm Interactome version 5 (WI5) containing 5,534 interactions and connecting 15% of the *C. elegans* proteome (table S1). WI5 gives rise to a giant network component of 2,898 nodes connected by 5,460 edges (Fig. 2A). Similar to other biological networks (15), the worm interactome network exhibits small-world and scale-free properties (Fig. 2B) (7). This data set also allowed us to analyze whether or not evolutionary recent proteins tend to preferentially interact with each other rather than with ancient proteins. We subdivided the nodes of the network into three classes: 748 proteins with a clear ortholog in yeast ("ancient"), 1,314 proteins with a clear ortholog in *Drosophila*, *Arabidopsis* or humans but not in yeast ("multicellular"), and 836 proteins with no detectable ortholog outside of *C. elegans* ("worm") (7). These three groups seem to connect equally well with each other (Fig. 2C), suggesting that new cellular functions rely on a combination of evolutionarily new and ancient elements, consonant with the classic proposal of evolution as a tinkerer that modifies and adds to pre-existing structures to create new ones (16).

Previous studies have related interactome data with genome-wide expression (transcriptome) and phenotypic profiling (*phenome*) data in *S. cerevisiae* (17). To investigate to what extent different functional genomic assays should correlate in the context of a multicellular organism, we overlapped WI5 with *C. elegans* transcriptome and phenome data sets.

Based on a *C. elegans* transcriptome compendium data set (18), Pearson correlation coefficients (PCCs) were calculated

for gene pairs involved in Y2H interactions and compared to randomized data sets (Fig. 2D). Approximately 150 Core interactions (9.5%) correspond to gene pairs with significantly higher PCCs than expected from random ($P < 0.05$) (table S3). Thus those pairs can be considered as “more biologically likely” since two completely independent approaches point to a functional relationship between the corresponding genes. The remaining pairs are labeled as “without additional evidence.” Indeed, it is important to note that lack of co-expression does not suggest that the corresponding interactions are irrelevant. Indeed, 75% of “literature” pairs, defined as biologically relevant, do not correlate with transcriptome data (Fig. 2D).

We also systematically examined Y2H interactions where both proteins belong to common *C. elegans* expression clusters, or “Topomap mountains” (18). As an example, a highly connected sub-network derived from mountain 29 (Fig. 2E) contains seven proteins (ABU-1, ABU-8, ABU-11, PQN-5, PQN-54, PQN-57 and PQN-71) that share common domains (DUF139 domain and cysteine-rich repeat). Furthermore, these proteins are all expressed in the pharynx (19–21), suggesting that they may act together in pharynx function or development.

For relatively small-scale *S. cerevisiae* and *C. elegans* interactome data sets, physical interactions pointed to genes that share similar phenotypes when knocked out or knocked down (17). To evaluate this idea for the *C. elegans* interactome, we assembled a collection of phenotypic data based on RNAi knock-down experiments from WormBase (7, 22), and calculated the percentage of protein interaction pairs that share embryonic lethal phenotypes for the interaction data sets and their randomized controls and found a two-fold enrichment for the Core and First-Pass data sets (Fig. 2F). Similar correlations were also observed for the maternal sterile phenotype and 4 groups of post-embryonic phenotypes (data not shown). Because protein-protein interactions for which both genes are co-expressed across many conditions and show similar phenotype(s) when knocked down should be considered particularly likely, the global correlations described above illustrate how biological hypotheses can be derived from overlapping interactome, transcriptome and phenome data sets (table S3).

In *S. cerevisiae*, two proteins that have many interaction partners in common are more likely to be related biologically (23). We examined the *C. elegans* interactome network for the presence of highly connected neighborhoods by determining the mutual clustering coefficient between proteins in the network (table S4) (23). As an example, we examined the properties of one of the clusters containing such a high-scoring protein pair: VAB-3/C49A1.4 (Fig. 3). VAB-3 and C49A1.4 have strong similarity to the products of the *Drosophila* genes *eyeless* (*ey*) and *eyes absent* (*eya*),

respectively, but not to each other. EY and EYA are components of a conserved network of transcription factors that regulate eye development (24).

VAB-3 and C49A1.4 are part of a highly interconnected subnetwork in WI5 (Fig. 3) with proteins that are known or suspected to be functionally linked to VAB-3 and C49A1.4, or to their respective orthologs in other organisms. These include: (i) EGL-27, which negatively regulates MAB-5 in hermaphrodites (25) and is linked to MAB-5 through C49A1.4; (ii) WRT-2, an interactor of C49A1.4 with similarity to *Drosophila* Hedgehog, which alleviates repression of *eya* expression by *Cubitus interruptus* (26); and (iii) CEH-33 and CEH-35, two of four members of the *sine oculis* homeobox gene family, which is involved in the same *Drosophila* regulatory network of transcription factors as *ey* and *eya* (27). Finally, eight proteins in this cluster are annotated in WormPD as involved in membrane function, suggesting a functional relationship between the *eyeless* transcription network and membrane activity.

Together with interologs and previously described interactions, the Y2H data set provides functional hypotheses for thousands of uncharacterized proteins in the *C. elegans* proteome. Integration with other functional genomic data indicates that the correlation between transcriptome and interactome data, while significant, is lower than what would be expected from observations made in yeast (17). This observation applies to both the Y2H data set described here and well-characterized worm interactions from the literature-derived data set (Fig. 2D). This may occur because, unlike unicellular organisms, metazoans are complicated by the fact that biological processes may occur differently in the organism, across various organs, tissues, or single cells.

Our current interactome map also illustrates how a human interactome project would benefit from an ORFeome cloning project using recombinational cloning systems, such as Gateway (8). Indeed, recombinationally cloned ORFs can be shuffled at will into various expression vectors needed for different types of protein interaction assays, as exemplified by our ability to transfer bait- and prey-encoding ORFs into Myc- and GST-tagged vectors to validate Y2H interactions.

References and Notes

1. E. M. Marcotte et al., *Science* **285**, 751 (1999).
2. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4285 (1999).
3. P. Uetz et al., *Nature* **403**, 623 (2000).
4. T. Ito et al., *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4569 (2001).
5. Y. Ho et al., *Nature* **415**, 180 (2002).
6. A. C. Gavin et al., *Nature* **415**, 141 (2002).
7. See supporting data on Science Online
8. J. Reboul et al., *Nature Genet.* **34**, 35 (2003).
9. A. J. Walhout et al., *Science* **287**, 116 (2000).

10. M. C. Costanzo et al., *Nucleic Acids Res.* **29**, 75 (2001).
11. L. R. Matthews et al., *Genome Res.* **11**, 2120 (2001).
12. A. Davy et al., *EMBO Rep.* **2**, 821 (2001).
13. S. J. Boulton et al., *Science* **295**, 127 (2002).
14. A. J. Walhout et al., *Curr. Biol.* **12**, 1952 (2002).
15. S. H. Strogatz, *Nature* **410**, 268 (2001).
16. F. Jacob, *Science* **196**, 1161 (1977).
17. H. Ge, A. J. Walhout, M. Vidal, *Trends Genet.* **19**, 551 (2003).
18. S. K. Kim et al., *Science* **293**, 2087 (2001).
19. J. Gaudet, S. E. Mango, *Science* **295**, 821 (2002).
20. M. Hanazawa, M. Mochii, N. Ueno, Y. Kohara, Y. Iino, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 8686 (2001).
21. F. Urano et al., *J. Cell Biol.* **158**, 639 (2002).
22. L. Stein, P. Sternberg, R. Durbin, J. Thierry-Mieg, J. Spieth, *Nucleic Acids Res.* **29**, 82 (2001).
23. D. S. Goldberg, F. P. Roth, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4372 (2003).
24. S. Wawersik, R. L. Maas, *Hum. Mol. Genet.* **9**, 917 (2000).
25. Q. Ch'ng, C. Kenyon, *Development* **126**, 3303 (1999).
26. K. S. Pappu et al., *Development* **130**, 3053 (2003).
27. C. Dozier, H. Kagoshima, G. Niklaus, G. Cassata, T. R. Burglin, *Dev. Biol.* **236**, 289 (2001).
28. We thank members of M.V.'s laboratory for their input and help; C. Boone, G. Achaz and D. Allinger for discussions; the sequencing staff at Agencourt Biosciences for technical assistance; the ORFeome meeting participants for their input; C. McCowan, T. Clingingsmith and C. You for administrative assistance; C. Fraughton for lab support. This work was supported by a grant from NHGRI and NIGMS) awarded to M.V. Other support includes an NSF award (K.C.G.), NIGMS grants (S.V., S.E.M., J.W.H.), a Department of Defense Predoctoral Fellowship (B.B.), an award from the Ligue Nationale Contre Le Cancer (équipe labélisée) (C.S., A.C.), an institutional HHMI grant (F.P.R., G.F.B) and Fellowships from EMBO (P.O.V.), NSF (D.S.G), Ryan, Milton (S.L.W.) and Fu (L.V.Z.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/1091403/DC1

Material and Methods

Fig. S1

Table S1 to S5

References

11 September 2003; accepted 1 December 2003

Published online 2 January 2004; 10.1126/science.1091403

Include this information when citing this paper.

Fig. 1. Co-affinity purification (co-AP) assays. Shown are 10 examples from the Core-1, Core-2 and Non-Core data sets.

The upper panels show Myc-tagged prey expression after affinity purification on glutathione-sepharose, demonstrating binding to GST-bait. The middle and bottom panels show expression of Myc-prey and GST-bait, respectively. The lanes alternate between extracts expressing GST-bait proteins (+) and GST alone (-). ORF pairs are identified in table S1 with the lane number corresponding to the order in which they appear in the table.

Fig. 2. Analysis of the WI5 network. (A) Nodes (representing proteins) are colored according to their phylogenetic class: ancient (red), multicellular (yellow), and worm (blue). Edges represent protein-protein interactions. The inset highlights a small part of the network. (B) The proportion of proteins, $P(k)$, with different numbers of interacting partners, k , is shown for *C. elegans* proteins used as baits or preys and for *S. cerevisiae* proteins. (C) The pie charts show the proportion of interacting preys found in Y2H screens that fall into each phylogenetic class. Also shown is the distribution of all preys found and all preys searched in the AD-ORFeome1.0 library. (D) Overlap with Topomap (see text) (18), Pearson correlation coefficients (PCCs) were calculated and graphed for each pair of proteins in the interaction data sets and their corresponding randomized data sets. The area to the right corresponds to interactions that show a significant relationship to expression profiling data ($P < 0.05$). (E) Interactions between proteins in Topomap mountain 29 (18). The dash-circled proteins belong to the same paralogous family (sharing more than 80% homology) and are thus collapsed into one set of interactions. (F) Proportion of interaction pairs where both are embryonic lethal ($P < 10^{-7}$).

Fig. 3. Graphical representation of a highly interconnected sub-network around VAB-3 and C49A1.4. Biological functional classes were obtained from WormPD (10).





