Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network

Nicolas Simonis[1,2,11], Jean-François Rual[1,2,10,11], Anne-Ruxandra Carvunis[1–3,11], Murat Tasan[4,11], Irma Lemmens[5,11], Tomoko Hirozane-Kishikawa[1,2], Tong Hao[1,2], Julie M Sahalie[1,2], Kavitha Venkatesan[1,2,10], Fana Gebreab[1,2], Sebiha Cevik[1,2,10], Niels Klitgord[1,2,10], Changyu Fan[1,2], Pascal Braun[1,2], Ning Li[1,2,10], Nono Ayivi-Guedehoussou[1,2,10], Elizabeth Dann[1,2], Nicolas Bertin[1,2,10], David Szeto[1,2,10], Amélie Dricot[1,2], Muhammed A Yildirim[1,2,6], Chenwei Lin[1,2], Anne-Sophie de Smet[5], Huey-Ling Kao[7], Christophe Simon[1,2,10], Alex Smolyar[1,2], Jin Sook Ahn[1,2], Muneesh Tewari[1,2,10], Mike Boxem[1,2,8,10], Stuart Milstein[1,2,10], Haiyuan Yu[1,2], Matija Dreze[1,2,9], Jean Vandenhaute[9], Kristin C Gunsalus[7], Michael E Cusick[1,2], David E Hill[1,2], Jan Tavernier[5], Frederick P Roth[1,4] & Marc Vidal[1,2]

To provide accurate biological hypotheses and elucidate global properties of cellular networks, systematic identification of protein-protein interactions must meet high quality standards. We present an expanded *C. elegans* protein-protein interaction network, or 'interactome' map, derived from testing a matrix of ~10,000 × ~10,000 proteins using a highly specific, high-throughput yeast two-hybrid system. Through a new empirical quality control framework, we show that the resulting data set (Worm Interactome 2007, or WI-2007) was similar in quality to low-throughput data curated from the literature. We filtered previous interaction data sets and integrated them with WI-2007 to generate a high-confidence consolidated map (Worm Interactome version 8, or WI8). This work allowed us to estimate the size of the worm interactome at ~116,000 interactions. Comparison with other types of functional genomic data shows the complementarity of distinct experimental approaches in predicting different functional relationships between genes or proteins.

The interactome of an organism is the network formed by the complete set of binary physical interactions that can occur between all proteins. Low-throughput protein-protein interaction experiments are of considerable value in understanding cellular processes at the molecular level. However, the development of high-throughput approaches can substantially increase the pace and scale of discovery, while permitting the implementation of standardized and systematic quality control. Initial steps toward binary interactome mapping in metazoans have been undertaken[1–5], and the resulting partial interactome maps have (i) provided insights into the organization of biological networks, (ii) assisted in determining functions of many proteins and complexes and (iii) identified hundreds of connections to proteins associated with human diseases.

High-throughput interactome mapping is particularly needed for *C. elegans*, a widely used model organism for which the set of protein-protein interactions derived from small-scale experiments and accessible in public databases is limited to less than 500. The first proteome-scale version of the Worm Interactome (WI5)[3] combined several sources of protein-protein interaction data: literature-curated interactions, yeast two-hybrid (Y2H) 'module' maps each devoted to a specific biological process[1,6–11], 'interolog'

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. [2]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [3]Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble (TIMC-IMAG), Unité Mixte de Recherche 5525 Centre National de la Recherche Scientifique (CNRS), Faculté de Médecine, Université Joseph Fourier, 38706 La Tronche Cedex, France. [4]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, Massachusetts 02115, USA. [5]Department of Medical Protein Research, Vlaams Instituut voor Biotechnologie, and Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, 3 Albert Baertsoenkaai, 9000 Ghent, Belgium. [6]Division of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, Massachusetts 02138, USA. [7]Center for Genomics and Systems Biology, Department of Biology, New York University, 100 Washington Square East, New York, New York 10003, USA. [8]Massachusetts General Hospital Center for Cancer Research, Building 149, 13th Street, Charlestown, Massachusetts 02129, USA. [9]Unité de Recherche en Biologie Moléculaire, Facultés Notre-Dame de la Paix, 61 Rue de Bruxelles, 5000 Namur, Belgium. [10]Present addresses: Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA (J.-F.R.), Novartis Institutes for Biomedical Research Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA (K.V.), School of Biomolecular and Biomedical Science, University College Dublin, Belfield, Dublin 4, Ireland (S.C.), Bioinformatics Program, Boston University, 705 Commonwealth Avenue, Boston, Massachusetts 02215, USA (N.K.), Wyeth Pharmaceuticals Inc., 35 Cambridgepark Drive, Cambridge, Massachusetts 02140, USA (N.L.), Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA (N.A.-G.), RIKEN Omics Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan (N.B., C.S.), University of California San Francisco School of Medicine, 500 Parnassus Avenue, San Francisco, California 94143, USA (D.S.), Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA (M. Tewari), Utrecht University, Kruytgebouw N309, 8 Padualaan, 3584 CH Utrecht, The Netherlands (M.B.) and Alnylam Pharmaceutical, 300 Third Street, Cambridge, Massachusetts 02142, USA (S.M.). [11]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu), F.P.R. (fritz_roth@hms.harvard.edu), J.T. (jan.tavernier@ugent.be) or D.E.H. (david_hill@dfci.harvard.edu).

RECEIVED 4 JUNE; ACCEPTED 29 OCTOBER; PUBLISHED ONLINE 14 DECEMBER 2008; DOI:10.1038/NMETH.1279

interactions—that is, predicted pairs of interactors whose respective orthologs interact in another organism[12]—and lastly, Y2H interactions derived from a high-throughput screen performed with ~2,000 metazoan proteins as baits[3] (WI-2004). WI5 represents a key resource for formulating biological hypotheses and investigating the properties of the *C. elegans* interaction network. However, WI5 includes nonbinary interactions derived from the literature, interologs not experimentally confirmed, and some lower-confidence Y2H interactions.

Our updated Worm Interactome map (WI8) implements several techniques and strategies that are critical for generating high-quality protein-protein interaction data on a proteomic scale. First, we expanded the worm interactome map by screening a matrix of ~10,000 × ~10,000 proteins. Second, we developed new standards to deliver a data set of very high quality. These standards involve a highly stringent, high-throughput yeast two-hybrid (HT-Y2H) assay, strict methods for filtering and updating existing data sets, independent measurement of technical quality, and evaluation of biological relevance. Because worm genome annotations are improved frequently, we updated previous protein-protein interaction data according to recent gene models. Finally, we empirically estimated the full size of the *C. elegans* interactome, through the implementation of a new interactome mapping framework based exclusively on protein-protein interaction data[13].

To extend the use of WI8 beyond protein-protein interaction analysis and to place WI8 into broader biological context, we integrated the resulting protein-protein interaction data with complementary data sets, such as physical and genetic interactions from curated literature, our interolog data set (**Supplementary Methods** online), phenotypic profiling data and a coexpression compendium. We also identified tissue localizations and developmental stages in which interacting pairs are most likely to be physiologically relevant whenever anatomical annotation[14] or spatiotemporal expression patterns[15] were available for both proteins.

Our new data set, WI-2007, consists of 1,816 high-confidence, binary, protein-protein interactions. We integrated previously published high-quality *C. elegans* binary protein-protein interactions with WI-2007 into the updated WI8 version of the worm interactome, providing 3,864 high-quality binary physical interactions between 2,528 proteins. WI8 was significantly enriched for functionally linked protein pairs, confirming its biological relevance and demonstrating the value of unbiased, large-scale Y2H screens in inferring protein function.

RESULTS

A new HT-Y2H data set

For this iteration of worm interactome mapping, we implemented a HT-Y2H strategy previously used for human interactome mapping[5]. We tested all open reading frames (ORFs) in the worm ORFeome version 1.1 (ref. 8) against one another (a ~10,000 × ~10,000 matrix), a search space corresponding to ~24% of the total search space for a comprehensive *C. elegans* interactome map, excluding variants due to polymorphism, alternative transcription or alternative splicing (**Fig. 1a**). We also ensured the quality of the new data set by using stringent conditions and controls described previously[5], including low expression of DNA-binding-domain and activation-domain fusion proteins (DB-X and AD-Y), multiple reporter genes to ensure high precision, removal of all a priori and *de novo* DB-X autoactivators, and individual retesting of each positive protein-protein interaction. The resulting set of 1,816 protein-protein interactions between 1,496 proteins (**Fig. 1b**) is called WI-2007.

Characterization of WI-2007

To assess the quality of our new data set and estimate the size of the complete worm interactome, we used a framework we recently developed[13], with a slightly different implementation relevant to the data available in *C. elegans*. This framework empirically measures several parameters to characterize a high-throughput binary protein-protein interaction data set: 'screening completeness', the fraction of the proteome-wide space tested in the experiment; 'precision', the proportion of interactions in the data set that are true biophysical interactions; 'sampling sensitivity', the fraction of all detectable interactions for a particular assay found under the sampling conditions, which corresponds here to the saturation of a single screen; and 'assay sensitivity', the proportion of all biophysical interactions that can be identified by an assay at saturation, as each assay can only detect a fraction of all true biophysical interactions.

To estimate these parameters we performed the following experiments. First, we used the mammalian protein-protein interaction trap technique (MAPPIT) to measure how a random sample of WI-2007 performed in an independent protein interaction detection assay compared to a positive reference set (cePRS-v1, manually recurated interactions from low-throughput studies) and a random reference set (ceRRS-v1, randomly chosen pairs in the search space of WI-2007). Second, we used the overlap between WI-2007 and our previous Y2H study in their common search space to quantify the saturation of our screen. Third, to evaluate the proportion of interactions that can be captured by our Y2H assay, we used the fraction of cePRS-v1 pairs recovered in a pairwise Y2H experiment and in WI-2007, as well as the proportion of widely conserved interologs found in WI-2007. Introducing these measurements into a Monte Carlo simulation (**Supplementary Methods**), we computed the four parameters in our framework, as well as the expected size of the worm interactome. According to this model, the screening completeness was 23.6%, the precision estimate 86% ± 16% (mean and s.d.), the sampling sensitivity 31% ± 8%, the assay sensitivity 16% ± 3% and the size of the worm interactome 115,600 ± 26,400 (**Fig. 1c**).

Given the potential bias in cePRS-v1 and in the set of ultra-conserved interologs toward interactions that are easy to detect, the associated assay sensitivity may be an overestimate. Thus, the predicted interactome size is likely to be a conservative estimate. The strength of this approach is that these calculations rely solely on protein-protein interactions, without depending on functional annotation or other types of genomic or proteomic data. Our estimate provides an endpoint for the worm interactome mapping project and can be used as a reference for evolutionary comparisons between interactome networks from different species.

A combined data set of high-quality binary interactions

To provide a set of integrated, high-quality, binary protein-protein interaction data for *C. elegans*, we employed higher stringency criteria and used updated WormBase (http://www.wormbase.org) gene models to reprocess the raw data from smaller scale Y2H screens encompassing proteins involved in vulval development[1],

protein degradation[6], DNA damage response[7], germline formation[9], TGF-β signaling pathway[11] and RNA interference[10], along with unpublished Y2H interactions (M. Tewari, N.A.-G. and J.S.A.; **Supplementary Methods**). This 'biological processes' subset (BPmaps) contains 554 protein-protein interactions.

WI8 is the union of WI-2004, WI-2007 and BPmaps. The consolidated WI8 network (**Fig. 2** and **Supplementary Table 1** online) contains 3,864 high-quality protein-protein interactions among 2,528 proteins. Approximately 40% of the interactions are newly identified, and the set excludes any lower-confidence interactions from previous studies[3]. The WI8 physical interaction network can be visualized on our website (http://interactome.dfci.harvard.edu/C_elegans/) using N-Browse[16] or VisANT[17].

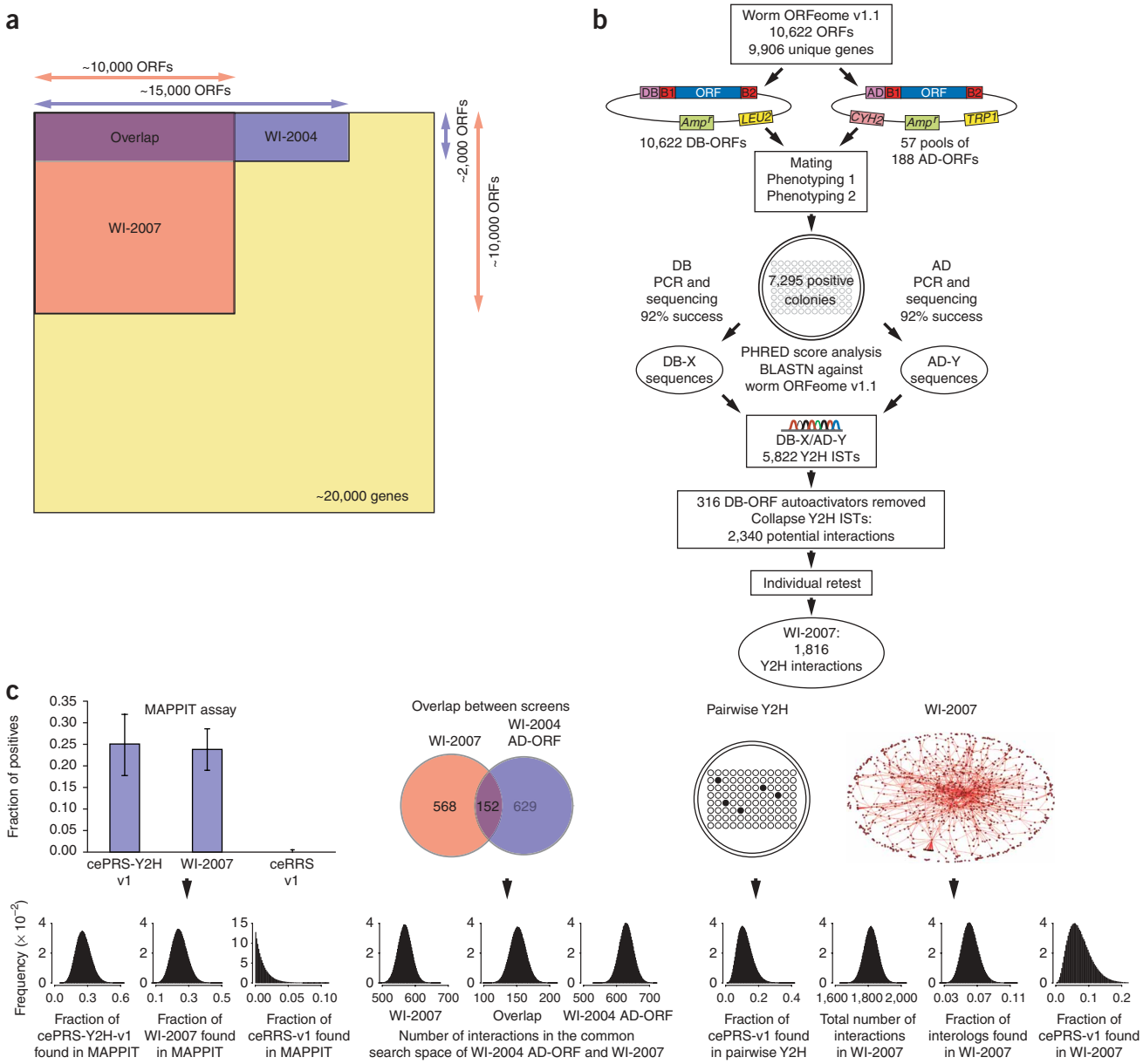

Figure 1 | Construction and characterization of WI-2007. (**a**) Search spaces of WI-2007 and WI-2004 relative to the whole proteome, three times larger for WI-2007 than WI-2004. (**b**) Pipeline used for WI-2007. ORFs from ORFeome v1.1 were transferred into DB and AD vectors by recombinational cloning, then transformed into yeast cells. Each bait was then mated with pools of 188 AD-ORFs. Two rounds of phenotyping were performed to isolate positive colonies, which were used to PCR-amplify DB-ORFs and AD-ORFs for sequencing, leading to the identification of 5,822 interaction sequence tags (ISTs). After excluding autoactivators and collapsing redundant ISTs corresponding to the same, nonoriented protein pair, each interaction was individually retested in an independent Y2H experiment to generate the final WI-2007 data set. (**c**) WI-2007 characterization. Ten measures are shown (left to right): proportions observed in MAPPIT of (i) cePRS-Y2H-v1, (ii) a random sample of WI-2007 and (iii) ceRRS-v1; number of interactions detected in the common search space of WI-2007 and WI-2004 (iv) in WI-2007, (v) in both screens and (vi) in WI-2004 AD-ORF; (vii) proportion of cePRS-v1 detected in an independent pairwise Y2H experiment; (viii) total number of interactions in WI-2007 and proportion recovered in WI-2007 of (ix) ultraconserved interologs and (x) cePRS-v1. The sampling errors on the ten measurements are modeled with beta distributions (bottom row). These distributions are then used in a Monte Carlo simulation to compute precision, sampling sensitivity, assay sensitivity and the total number of interactions in *C. elegans*, along with their associated error bars. Label on *y* axis (frequency) applies to all ten sampling distributions.

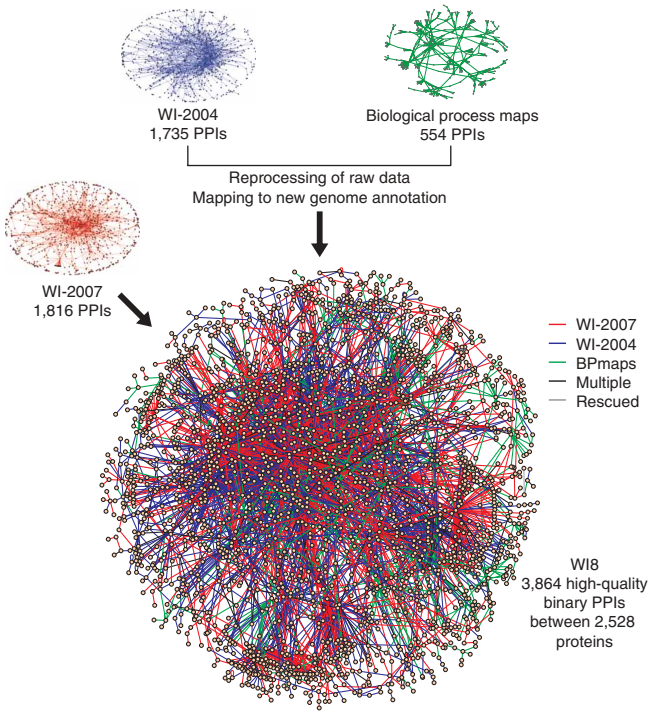

Figure 2 | WI8: an extended, high-quality, protein-protein interaction network. High-quality data on Y2H protein-protein interactions (PPIs) from WI-2007, WI-2004 and diverse medium-throughput biological processes based Y2H maps[1,6–11] were integrated into WI8. The color of the edge indicates the data set of origin: WI-2007, red; WI-2004, blue; biological process maps, green. Edges corresponding to more than one of these evidence types are shown in black, and edges corresponding to 'rescued' interactions—that is, supported by at least two lower-confidence pieces of evidence—in gray. Only the main giant component of the network (connected subgraph that contains the majority of the entire network's nodes) is shown.

Confirmed Y2H interactions may be 'biophysically true' interactions that do not actually occur *in vivo* if the involved proteins are not present at the same time and place within a multicellular organism, or are not present with the proper post-translational modifications. We evaluated the overall biological relevance of WI8 by assessing the degree to which interacting pairs share Gene Ontology annotation terms—that is, can be considered as functionally linked. A Gene Ontology term may be specific or broad, depending on the number of genes to which it is assigned. We therefore defined four different thresholds of functional specificity: less than or equal to 5, 20, 100 and 400 annotated genes per Gene Ontology term. For all three component subsets of WI8, we compared the degree of functional linkage with that of binary interactions derived from the literature (LCI binary; **Supplementary Methods** and **Supplementary Table 2** online), normalizing for protein composition bias of each of these subsets. All data subsets showed a high enrichment for both broad and specific functional linkage (**Fig. 3a**), suggesting high biological relevance. The degree of functional linkage among WI-2007 was similar to or exceeded the literature enrichment at each functional specificity limit tested.

Various interactions in WI8 provide new biological information. For example, EBP-1, a microtubule-binding protein whose homologs are involved in a variety of microtubule-mediated processes[18], interacts with several proteins involved in microtubule dynamics, including UNC-14, a protein required for axon growth and sex myoblast migration[19], VAB-8, a kinesin-like protein required for axon outgrowth and cell migration[20], and RSA-2, a protein specifically required for microtubule outgrowth from centrosomes and for spindle assembly[21].

Integrated functional network

Integration of diverse large-scale data sets was previously used to demonstrate the coordination of interconnected yet distinct molecular machines involved in worm early embryogenesis[22]. Another recent publication[23] describes Bayesian integration of functional linkages into a single network, weighting each type of evidence according to a reference data set (benchmark). Such an approach can be a valuable resource leading to interesting hypotheses, but is highly dependent on the benchmark, which can strongly bias the predictions (**Supplementary Discussion** online). In contrast, we chose to provide an unweighted data set that (i) does not artificially bias the network toward highly-studied proteins, (ii) allows the user to select their own threshold for some types of linkages (for example, correlation coefficient with expression data), (iii) separates each type of experimental evidence and (iv) does not rely on an inevitably biased benchmark.

We integrated WI8 with five different sources of evidence for functional relationships: (i) mRNA coexpression data available in WormBase (**Supplementary Table 3** online); (ii) RNAi phenotypes from RNAiDB[24] (**Supplementary Table 4** online); (iii) genetic interactions curated in WormBase; (iv) interolog interactions and (v) all binary and nonbinary protein-protein interactions from our literature-curated data set (LCI; **Supplementary Methods**). This integrated network involves 178,151 links between 6,176 genes and can be visualized online using N-Browse (http://interactome.dfci.harvard.edu/C_elegans/).

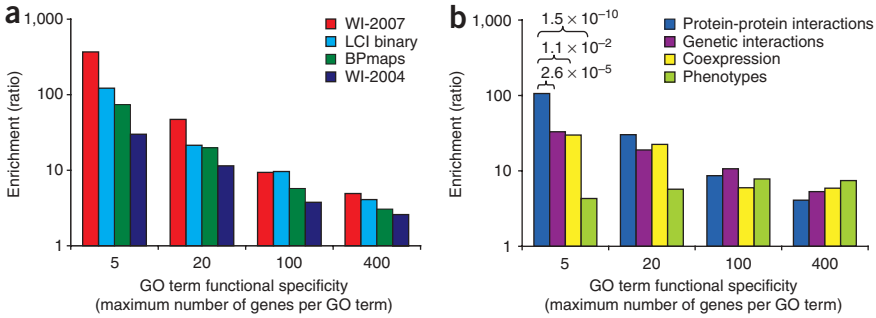

Figure 3 | Biological relevance. Enrichment represents the frequency of functional linkage of protein or gene pairs expressed as a multiple of the value for random pairs and was plotted against functional specificity groupings. The maximum number of genes associated with a particular Gene Ontology (GO) term was used as an estimate of the functional specificity (5, 20, 100 or 400 genes). (**a**) Enrichment for functional relationships in different components of the WI8 data set and in the LCI binary data set. (**b**) Functional relationship enrichments for distinct types of experimental evidence. *P*-values assessing the difference between protein-protein interactions and other types of evidence are shown for very specific Gene Ontology terms (terms with a maximum of five genes).

Table 1 | Overlap between data sets from the integrated functional network

	WI8		LCI		Interologs		Genetic interactions		Phenotypes	
	E	*P*	*E*	*P*	*E*	*P*	*E*	*P*	*E*	*P*
LCI	182.3	1.01×10^{-37}								
Interologs	91.4	1.13×10^{-212}	145.6	5.89×10^{-75}						
Genetic interactions	23.9	1.59×10^{-14}	66.9	1.17×10^{-72}	24.1	6.58×10^{-58}				
Phenotypes	3.0	5.33×10^{-3}	4.6	1.02×10^{-3}	3.0	1.27×10^{-16}	3.3	3.83×10^{-6}		
Coexpression	2.5	1.20×10^{-8}	2.6	3.20×10^{-3}	3.2	5.01×10^{-103}	1.6	1.61×10^{-1}	1.6	1.09×10^{-21}

Enrichment (*E*, expressed as a multiple) and significance (*P*-values) of the overlaps between distinct functional data sets. The enrichment is defined as the number of pairs shared between two data sets divided by the expected random number of shared pairs, and the significance is assessed by Fisher's exact test.

We compared the biological relevance of each type of data from the integrated network by calculating the enrichment for functional linkage, as described before for protein-protein interaction data sets (**Fig. 3b**). All the analyzed data sets showed highly significant enrichment for functional linkage ($P < 2.5 \times 10^{-3}$). Notably, among the analyzed data sets, physical interactions seemed to be the best predictors of highly specific shared Gene Ontology terms, whereas pairs sharing phenotypes showed the highest enrichment for less specific functional linkages. The phenotypic profiles used in this study were gross phenotypes, and more precise phenotypic observations would probably be better predictors for more precise functions but worse predictors for more global functions. Similarly, linkages from expression data were derived from a wide range of experimental conditions; such data could be a better predictor of more specific linkages, if a set of experimental conditions targeting a particular process had been used. This observation reflects how these different data sets address biological questions at different levels, in the same way that sequence and structure similarity are better predictors of whether proteins exert the same enzymatic activity than of whether they belong to the same pathways[25].

Next we examined the overlap between component networks of each type. We observed significant overlap for almost all combinations of component networks (**Table 1**). WI8, LCI, interologs and genetic interactions showed more overlap with one another than coexpression or phenotypic correlation with any other data set. The strong association between the two physical interaction data sets and interologs (LCI and WI8 confirmed 56 and 194 predicted interologs, including 49 and 147 heterodimers, respectively) was expected, and it confirmed that many interactions are conserved during evolution. LCI shared higher overlaps with phenotypically correlated pairs, genetic interactions and interologs than WI8, most likely because lower-throughput assays often test physical interactions that are enriched a priori for a common phenotype or are known to have interacting orthologs. Still, WI8 substantiated 57 pairs of genes with high coexpression among a wide range of experimental conditions, 9 pairs of genes with similar RNAi phenotypic profiles and 14 pairs of genetically interacting genes ("shared edges" section at http://interactome.dfci.harvard.edu/C_elegans/).

Although significant and informative, these overlaps remain relatively low (**Supplementary Table 5** online). This can be explained by lack of 'screening completeness' of most data sets; that is, most of these data sets are not genome or proteome wide. Indeed, more than 60% of genes/proteins in the network (the term 'genes/proteins' is used to reflect the mixed nature of the network, built from links between both genes and proteins) are present in one data set only, whereas less than 5% are present in four data sets or more. Furthermore, most of the screens that have led to the generation of these data sets (including our Y2H screens) are far from saturation and are probably limited by low sampling sensitivity in addition to inherent limitations of each assay; that is, precision and assay sensitivity. Finally, a perfect overlap is not expected because of intrinsic differences in the nature of the biological attributes measured in these data sets.

Module-scale biological networks

Module-scale biological subnetworks can be extracted from the integrated network by selecting 'seed' genes/proteins known to be associated with a specific process and then expanding by selecting neighboring genes/proteins. For example, using as seeds genes/proteins implicated in RNA-binding processes (**Fig. 4a**), nearly all genes/proteins in the expanded set are linked to several RNA-binding genes/proteins and are connected by at least two types of relationships. Most of these linked genes/proteins were thus predicted as functionally related to RNA binding, and several (for example sup-12) were already annotated or predicted by sequence similarity to be associated with RNA binding within WormBase. Other genes/proteins have annotations consistent with RNA binding. For example, T26A8.4 encodes a protein predicted to be part of the CPSF subcomplex of the Polyadenylation Factor I complex through clusters of eukaryotic orthologous groups (KOGs)[26] and is orthologous to yeast Caf120, which is part of the conserved Ccr4-Not transcriptional regulatory complex involved in mRNA initiation, elongation and degradation.

When expanding from a seed set of genes/proteins involved in cell fusion (**Fig. 4b**), almost all added genes/proteins are linked to more than one in the seed set, with many links supported by more than one evidence source. For example, unc-62 had phenotypic correlation with seed genes/proteins nhr-25, lin-29 and ceh-20; physical, genetic and interolog links with ceh-20; and interolog links with mab-5. In contrast to the RNA-binding subnetwork, where most links were physical interactions with few pairs being supported by more than one evidence type, in this example most links were either phenotypic or genetic interactions, and many physical interactions were supported by other evidence.

Notably, WI-2007 contains physical interactions between proteins not previously linked to one another, but at a network distance of two in the integrated network (**Supplementary Fig. 1** online). In the RNA-binding network, for example, star-2 and mec-8, which are known to be indirectly linked through sup-12, were found to directly interact. We found 1,157 new 'triangle closures' of

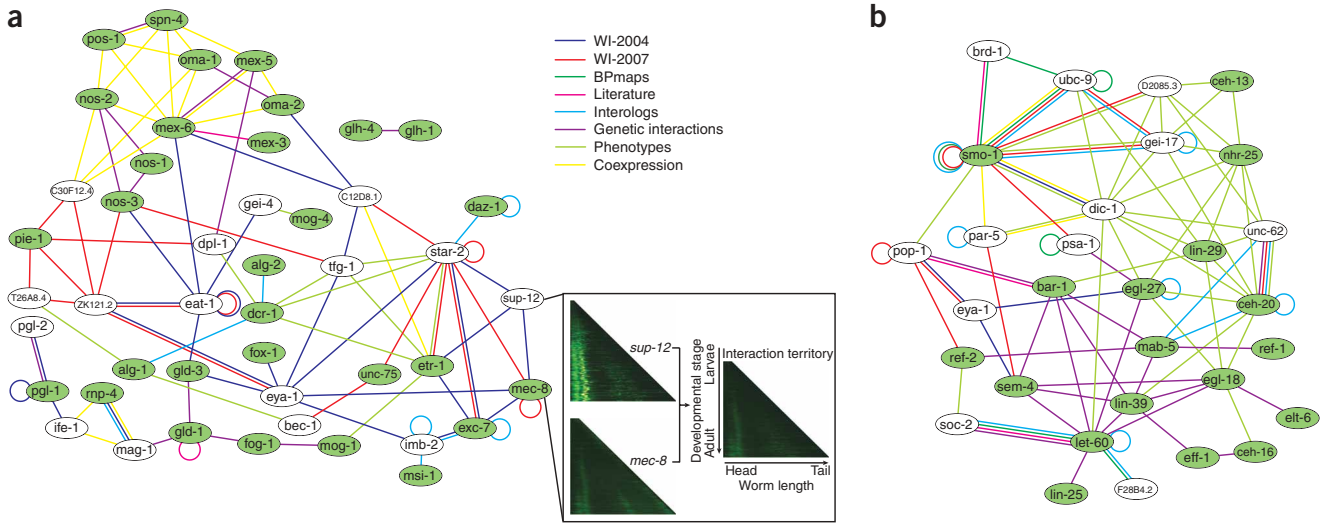

Figure 4 | Examples of multiple-evidence subnetworks. The networks represent relationships among genes/proteins from several evidence sources, color-coded as indicated. Genes and their products are labeled using an unitalicized lower-case version of the standard *C. elegans* three-letter system to reflect the inclusion of links between both proteins and genes. (**a**) Genes/proteins related to RNA binding. Green ellipses are genes/proteins annotated as 'RNA-binding' in WormBook[31]; white ellipses are genes/proteins linked to RNA-binding genes/proteins by at least one protein-protein interaction from WI8 and one other piece of evidence. The inset shows the chronograms of *sup-12* and *mec-8* (left) and their predicted spatiotemporal pattern of interaction (right). The chronograms represent the absolute GFP intensity measured (increasing values coded black-green-yellow-white) using reporter constructs with the indicated promoter, along the worm length (*x* axis) and as a function of development stage (*y* axis)[15]. (**b**) Genes related to cell fusion. Green ellipses are genes/proteins annotated as 'cell fusion' in WormBook[32]; white ellipses are genes/proteins linked to cell fusion genes/proteins by a protein-protein interaction from WI8 and one other type of evidence.

this kind (viewable within the "intersections" and "display" sections of http://interactome.dfci.harvard.edu/C_elegans/).

From 'static' map to spatiotemporal interactome

Spatiotemporal expression patterns for ~2,000 worm genes have recently become available through large-scale studies of worms carrying endogenous promoters driving expression of GFP[14,15]. Examination of the resulting GFP intensity patterns informs the question of where (tissue) and when (developmental stage) promoters are activated. The GFP profiles can be sorted according to developmental stage by worm length and aligned, forming a 'chronogram' representation[15].

We performed computational 'chronogram intersection' of the spatiotemporal expression patterns corresponding to two interacting proteins and used these to infer a potential 'interaction territory' (**Supplementary Figs. 2** and **3** online). We also inferred interaction territories on the basis of explicit anatomical annotations[14] for interacting proteins. We identified 111 common anatomical annotations and generated 69 chronogram intersections for protein-protein interactions from WI8 (viewable within the "localization" section of http://interactome.dfci.harvard.edu/C_elegans/). Examples from the RNA-binding subnetwork (**Fig. 4a**) included common interaction territories for SUP-12 and MEC-8 (**Fig. 4a**, inset), MEC-8 and EXC-7, and MEP-1 and MOG-4 through chronogram intersections, and for 21 more interactions through anatomical annotations. Although this GFP-based technique has limitations related to resolution and coverage, these examples provide a glimpse of how integrating spatiotemporal expression information could eventually allow extraction of tissue-specific subnetworks corresponding to pathways, functional modules or protein complexes, once the technology improves and more data become available.

DISCUSSION

We describe the implementation of an integrated strategy for generating high-confidence networks based on a highly stringent HT-Y2H assay combined with a quality control framework[13], thus achieving a step along the path to completion of the *C. elegans* interactome. Our estimated size of the complete *C. elegans* biophysical interactome is approximately 116,000 interactions, considering only a single protein isoform per gene. Although WI8 provides 3,864 interactions, 96%–97% of the interactome remains untouched because of lack of screening completeness as well as incomplete sampling and assay sensitivity. From the overlap of two independent HT-Y2H screens, we estimate that a single high-throughput screen can capture ~30% of the detectable interactions and thus would need to be repeated several times to reach saturation. Even at saturation, some interactions may not be detectable by Y2H because of intrinsic limitations of the assay—for example, proteins may not be imported into the nucleus, proper folding may not occur because of the fusion with the DNA-binding or activation domains, or interactions may require post-translational modifications or cofactors not present within *S. cerevisiae*. We estimate the proportion of interactions detectable with our HT-Y2H system (assay sensitivity) at approximately 16%.

Several approaches under development, involving optimization of the experimental setup[27] or systematic ORF fragmentation[28], should improve the assay sensitivity in future interactome mapping projects. However, achieving comprehensive mapping of the interactome will require use of various assays with complementary assay sensitivities. For example, experiments conducted in mammalian cells may uncover some interactions missed by Y2H, but fail to find others because some interactions do not occur under the conditions tested[27]. In addition to improving sensitivity, further cloning efforts will have to be undertaken to increase the screening

completeness of future interactome mapping projects. WI8 represents an early milestone toward uncovering the complete interactome network, yet it is to our knowledge the most comprehensive and reliable protein-protein interactions data set available today for *C. elegans*.

METHODS

Y2H screening. We mated 94 individual *MATα* MaV203 DB-ORF yeast strains, in a 96-well format, with the same *MATa* MaV103 AD-188ORFs mini-library on solid medium containing yeast extract, peptone and dextrose (YPD). Each DB-ORF 96-well plate was individually mated to all AD-ORFs compiled into 57 AD-188ORFs pools. After overnight growth at 30 °C, we transferred the colonies to plates containing synthetic complete (SC) yeast medium lacking leucine, tryptophan and histidine and containing 20 mM 3-aminotriazole (3AT) to select for diploids that showed elevated expression of the *GAL1::HIS3* Y2H marker. The same cells were transferred in parallel onto SC medium lacking leucine (SC–L) and containing 3AT and cycloheximide (SC–L+3AT+CYH). The pAD-dest-CYH vector contains the *CYH2* negative-selection marker, which allows plasmid shuffling on cycloheximide-containing media. This step is crucial to eliminate autoactivators that can arise during Y2H selection. Autoactivators show a $3AT^+/3AT\text{-}CYH^+$ phenotype, whereas genuine positives show a $3AT^+/3AT\text{-}CYH^-$ phenotype in this assay. We picked approximately 180,000 positive colonies from $3AT^+/3AT\text{-}CYH^-$ spots into a second-generation set of 96-well plates for further phenotypic screening.

Scoring Y2H assays. Consolidated and regrown $3AT^+/3AT\text{-}CYH^-$ colonies were transferred to both Sc–L+3AT and Sc–L+3AT+CYH plates to confirm *GAL1::HIS3* transcriptional activity, and to YPD to determine *GAL1::lacZ* transcriptional activity using a β-galactosidase filter assay. We selected colonies that retested $3AT^+/3AT\text{-}CYH^-$ and tested positive at levels equal or higher to that of the control DB-RB/AD-E2F interaction pair in our Y2H control set. Of the original ~180,000 $3AT^+/3AT\text{-}CYH^-$ colonies, 7,295 passed this double phenotypic test and represented Y2H positives. We also systematically tested all DB-ORFs for autoactivation by growth on solid SC-L+3AT medium, identifying all strong autoactivators and removing them from further consideration as baits in Y2H.

Yeast PCR and IST sequencing. We performed PCR amplifications on all Y2H-positive colonies to individually amplify DB-ORFs and AD-ORFs. The products from the PCR were purified and used as templates in a cycle-sequencing reaction to obtain two interaction sequence tags (ISTs) per Y2H positive.

WI-2007 IST analysis. The quality of the ISTs obtained by sequencing was measured by moving a sliding window of 10 base pairs to define the portion of the IST that had an average PHRED (http://www.phrap.com/phred/) score of 10 or higher over at least 10% of their length. We aligned all sequences against the worm ORFeome v1.1 database (http://worfdb.dfci.harvard.edu/) and remapped them to WormBase version WS150. We retained only those 5,822 showing a BLASTN *E*-value $E \leq 10^{-20}$. We collapsed all IST pairs corresponding to the same unordered gene locus pair.

Pairwise Y2H verification. We verified all Y2H interactions by mating fresh individual *MATα* MaV203 DB-ORF yeast cells with their corresponding individual *MATa* MaV103 AD-ORF yeast cells. For genes with more than one clone in the worm ORFeome v1.1, we used the clone with the highest similarity to the IST sequenced in the high-throughput screen for the retest. We tested the resulting diploids for their ability to activate two out of the three Y2H reporter genes. Of the 2,340 potential interactions, 78% (1,816) successfully passed this Y2H retest.

Reference literature data sets: PRS and RRS. To evaluate WI-2007 interactions, we assembled a positive reference set (PRS) and a random reference set (RRS) of binary interactions. We manually recurated physical interactions derived from low-throughput studies in the curated literature, both to ensure high quality and to verify evidence that the interactions were direct and binary, producing the *C. elegans* positive reference set version 1 (cePRS-v1), including 53 worm binary protein-protein interactions. Another 94 pairs selected randomly from the set of ~50,000,000 pairs among proteins represented as clones in the worm ORFeome v1.1 constituted the *C. elegans* random reference set version 1 (ceRRS-v1). To overcome potential biases of MAPPIT compared to Y2H interactions (the two assays may not be completely independent), we selected only the 47 cePRS-v1 pairs that have been detected by Y2H (cePRS-Y2H-v1) to compute the precision.

MAPPIT assay. In this system, the bait is fused to a STAT recruitment-deficient, homodimeric cytokine receptor and the prey protein is fused to functional STAT recruitment sites (gp130). An interaction between bait and prey allows the activation of a ligand-dependent signal transduction pathway, which controls the activation of a luciferase marker. MAPPIT was performed as described[29] with minor changes. We transfected plasmids into human 293T cells in 96-well plates using a calcium phosphate protocol[30]. Transfected cells were cultured for 24 h in Dulbecco's Modified Eagle's Medium supplemented with 10% fetal bovine serum and then stimulated with erythropoietin (R&D Systems) or left untreated for another 24 h, followed by measurement of luciferase activity in triplicate. For details of the use of MAPPIT to evaluate the Y2H data set, see **Supplementary Methods**.

Functional linkage estimation. The enrichment of a particular data set is expressed as an odds ratio—the number of distinct pairs (excluding homomeric interactions) sharing at least one Gene Ontology term (at a given functional specificity threshold) divided by the number of pairs expected at random. Significance of enrichment was calculated using a one-sided Fisher's exact test. We estimated the space of possible gene pairs as all unordered pairs between the genes in the input data set to account for specific biases of each data set, and then restricted this space to pairs in which both genes have one or more annotations at the considered functional specificity level. The number of genes associated with a particular Gene Ontology term was used as an estimate of the functional specificity, and we calculated the enrichments for several functional specificity levels (5, 20, 100 and 400). Differences between enrichments were assessed using an independent, two-sample *t*-test. **Supplementary Figures 4** and **5** online detail the separate branches of the Gene Ontology.

Additional methods. Detailed descriptions of the cloning and transformation steps, MAPPIT scoring, WI-2007 characterization through Monte Carlo simulation, reprocessing of BPmaps and WI-2004 data, overlap between component networks, module-scale subnetwork extraction, and chronogram intersections, as well as LCI, interologs, genetic interactions, coexpression, phenotypic similarity and anatomical annotation data sets, are available in **Supplementary Methods**. WI8 is provided with MIMIX specifications as **Supplementary Data 1** online. The integrated functional network is available as **Supplementary Data 2** online.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank F. Piano and members of the Cancer Center for System Biology and the Vidal laboratory for discussions, A. Petcherski from WormBase for assistance with worm genetic interactions, and Z. Hu for VisANT assistance. The worm interactome project was supported by grants from the US National Institutes of Health—R01 HG001715 (M.V. and F.P.R.), R01 HG003224 (F.P.R.), F32 HG004098 (M. Tasan), T32 CA09361 (K.V.)—a University of Ghent grant GOA12051401 (J.T.), and the Fonds Wetenschappelijk Onderzoek – Vlaanderen (FWO-V) G.0031.06 (J.T.). I.L. was supported by a postdoctoral fellowship from the FWO-V. K.C.G. and H.-L.K. were supported by US Department of the Army Award W81XWH-04-1-0307 and the State of New York's Science and Tech Resources contract C040066. M.V. is a Chercheur Qualifié Honoraire from the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium).

AUTHOR CONTRIBUTIONS

J.-F.R., N.S. and A.-R.C. coordinated experiments and data analysis. J.-F.R., T.H.-K., J.M.S., F.G., S.C., P.B., N.L., N.A.-G., E.D., D.S., A.D., C.S., M.V., H.Y., M.B., S.M., M.D., M. Tewari and J.S.A. performed the high-throughput ORF cloning and Y2H screens. I.L., A.-S.d.S., P.B. and J.T. conducted the MAPPIT experiments. N.S., A.-R.C., M. Tasan, T.H., N.K., K.V., C.F., N.B., M.A.Y., C.L., A.S., H.-L.K. and K.C.G. performed the computational analyses. M. Tasan, N.S., C.F., A.-R.C., H.-L.K. and K.C.G. adapted or built the website and visualization tools. N.S., A.-R.C., J.-F.R., M.E.C., J.V., F.P.R. and M.V. wrote the manuscript. M.V. conceived the project. D.E.H., J.T., F.P.R. and M.V. co-directed the project.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Walhout, A.J. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
2. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
3. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
4. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
5. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
6. Davy, A. *et al.* A protein-protein interaction map of the *Caenorhabditis elegans* 26S proteasome. *EMBO Rep.* **2**, 821–828 (2001).
7. Boulton, S.J. *et al.* Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* **295**, 127–131 (2002).
8. Reboul, J. *et al. C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34**, 35–41 (2003).
9. Walhout, A.J. *et al.* Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* **12**, 1952–1958 (2002).
10. Kim, J.K. *et al.* Functional genomic analysis of RNA interference in *C. elegans*. *Science* **308**, 1164–1167 (2005).
11. Tewari, M. *et al.* Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-β signaling network. *Mol. Cell* **13**, 469–482 (2004).
12. Matthews, L.R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs." *Genome Res.* **11**, 2120–2126 (2001).
13. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* advance online publication, doi:10.1038/nmeth.1280 (7 December 2008).
14. Hunt-Newbury, R. *et al.* High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol.* **5**, e237 (2007).
15. Dupuy, D. *et al.* Genome-scale analysis of *in vivo* spatiotemporal promoter activity in *Caenorhabditis elegans*. *Nat. Biotechnol.* **25**, 663–668 (2007).
16. Kao, H.L. & Gunsalus, K.C. Browsing multidimensional molecular networks with the generic network browser (N-Browse). *Curr. Protoc. Bioinformatics* Ch. 9, Unit 9 11 (2008).
17. Hu, Z., Mellor, J., Wu, J. & DeLisi, C. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics* **5**, 17 (2004).
18. Motegi, F., Velarde, N.V., Piano, F. & Sugimoto, A. Two phases of astral microtubule activity during cytokinesis in *C. elegans* embryos. *Dev. Cell* **10**, 509–520 (2006).
19. Branda, C.S. & Stern, M.J. Mechanisms controlling sex myoblast migration in *Caenorhabditis elegans* hermaphrodites. *Dev. Biol.* **226**, 137–151 (2000).
20. Wolf, F.W., Hung, M.S., Wightman, B., Way, J. & Garriga, G. vab-8 is a key regulator of posteriorly directed migrations in *C. elegans* and encodes a novel protein with kinesin motor similarity. *Neuron* **20**, 655–666 (1998).
21. Schlaitz, A.L. *et al.* The *C. elegans* RSA complex localizes protein phosphatase 2A to centrosomes and regulates mitotic spindle assembly. *Cell* **128**, 115–127 (2007).
22. Gunsalus, K.C. *et al.* Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 861–865 (2005).
23. Lee, I. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat. Genet.* **40**, 181–188 (2008).
24. Gunsalus, K.C., Yueh, W.C., MacMenamin, P. & Piano, F. RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Res.* **32**, D406–D410 (2004).
25. Wilson, C.A., Kreychman, J. & Gerstein, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249 (2000).
26. Tatusov, R.L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
27. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* advance online publication, doi:10.1038/nmeth.1281 (7 December 2008).
28. Boxem, M. *et al.* A protein domain-based interactome network for *C. elegans* early embryogenesis. *Cell* **134**, 534–545 (2008).
29. Eyckerman, S. *et al.* Design and application of a cytokine-receptor-based interaction trap. *Nat. Cell Biol.* **3**, 1114–1119 (2001).
30. Lemmens, I., Lievens, S., Eyckerman, S. & Tavernier, J. Reverse MAPPIT detects disruptors of protein-protein interactions in human cells. *Nat. Protoc.* **1**, 92–97 (2006).
31. Lee, M.-H. & Schedl, T. RNA-binding proteins. in *WormBook* (ed. Blumenthal, T.) doi:10.1895/wormbook.1.7.1 (2006).
32. Podbilewicz, B. Cell fusion. in *WormBook* (eds. Kramer, J.M. & Moerman, D.G.) doi:10.1895/wormbook.1.7.1 (2006).