Published in final edited form as:
Cell. 2008 August 8; 134(3): 534–545. doi:10.1016/j.cell.2008.07.009.

A protein domain-based interactome network for *C. elegans* early embryogenesis

Mike Boxem[1,2*], **Zoltan Maliga**[3,11], **Niels Klitgord**[1,11], **Na Li**[1,11], **Irma Lemmens**[4,11], **Miyeko Mana**[6,11], **Lorenzo de Lichtervelde**[1], **Joram D. Mul**[1], **Diederik van de Peut**[1], **Maxime Devos**[1], **Nicolas Simonis**[1], **Muhammed A. Yildirim**[1], **Murat Cokol**[5], **Huey-Ling Kao**[6], **Anne-Sophie de Smet**[4], **Haidong Wang**[7], **Anne-Lore Schlaitz**[3], **Tong Hao**[1], **Stuart Milstein**[1], **Changyu Fan**[1], **Mike Tipsword**[3], **Kevin Drew**[6], **Matilde Galli**[8], **Kahn Rhrissorrakrai**[6], **David Drechsel**[3], **Daphne Koller**[7], **Frederick P. Roth**[5], **Lilia M. Iakoucheva**[9], **A. Keith Dunker**[10], **Richard Bonneau**[6], **Kristin C. Gunsalus**[6], **David E. Hill**[1], **Fabio Piano**[6], **Jan Tavernier**[4], **Sander van den Heuvel**[2,8], **Anthony A. Hyman**[3*], and **Marc Vidal**[1*]

[1]*Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA* [2]*Massachusetts General Hospital Cancer Center, Charlestown, MA 02129, USA* [3]*Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany* [4]*Department of Medical Protein Research, VIB, and Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, 9000 Ghent, Belgium* [5]*Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA* [6]*Center for Genomics & Systems Biology, Department of Biology, New York University, New York, NY 10003, USA* [7]*Computer Science Dept., Stanford University, Stanford, CA 94305, USA* [8]*Division of Developmental Biology, Faculty of Science, Utrecht University, 3584 CH Utrecht, The Netherlands* [9]*Laboratory of Statistical Genetics, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA* [10]*Center for Computational Biology and Bioinformatics, Indiana University Schools of Medicine and Informatics, 410 W. 10th Street, Indianapolis, IN 46202, USA*

Summary

Many protein-protein interactions are mediated through independently folding modular domains. Proteome-wide efforts to model protein-protein interaction or “interactome” networks have largely ignored this modular organization of proteins. We developed an experimental strategy to efficiently identify interaction domains and generated a domain-based interactome network for proteins involved in *C. elegans* early embryonic cell divisions. Minimal interacting regions were identified for over 200 proteins, providing important information on their domain organization. Furthermore, our approach increased the sensitivity of the two-hybrid system, resulting in a more complete interactome network. This interactome modeling strategy revealed new insights into *C. elegans* centrosome function and is applicable to other biological processes in this and other organisms.

*Correspondence: marc_vidal@dfci.harvard.edu (M.V.), hyman@mpi-cbg.de (A.A.H.), mboxem@partners.org (M.B.).
[11]These authors contributed equally to this work.

Introduction

Physical interactions between proteins are crucial in most biological processes. Hence, there have been major efforts at systematically identifying protein-protein interactions using yeast two-hybrid (Y2H) and affinity pull-down mass spectrometry (AP/MS) approaches (Formstecher et al., 2005; Gavin et al., 2002; Giot et al., 2003; Ho et al., 2002; Ito et al., 2001; Krogan et al., 2006; Li et al., 2004; Rual et al., 2005; Stelzl et al., 2005; Uetz et al., 2000; Walhout et al., 2000). However, such high-throughput assays typically model interactions between full-length proteins, which fails to reflect that most proteins are composed of multiple distinct domains and motifs (Bornberg-Bauer et al., 2005; Liu and Rost, 2004; Pawson and Nash, 2003). Thus, a more precise description of protein-protein interaction networks requires information on the discrete domains that mediate these interactions. Since current knowledge of protein domains is often limited to sequence conservation, new experimental strategies are required to accurately describe large numbers of interaction domains. The Y2H system is ideally suited to identify binary interactions between proteins, and has been used to define interaction domains of individual proteins. However, domain-based Y2H mapping has not been carried out systematically at the scale of a biological process or the whole proteome.

We decided to test domain-based interactome mapping on 800 proteins required for *C. elegans* early embryogenesis, defined as the first two cell divisions following fertilization. *C. elegans* early embryogenesis is ideally suited for systematic domain-based protein interaction mapping because: (1) most of the proteins involved have been identified (Piano et al., 2002; Sönnichsen et al., 2005; Zipperlen et al., 2001), (2) the proteins are highly conserved in higher eukaryotes, (3) the phenotypic consequences of their inactivation are characterized in detail, and (4) the molecular machines they form have been reasonably well modeled (Gunsalus et al., 2005). Adding domain-based interactome information should bring us closer to the ultimate goal of developing a complete and predictive model of early embryogenesis.

Results

Domain-based interactome mapping

To define interaction domains, we developed a Y2H approach based on screening a PCR-generated library of systematically produced protein domains fused to the Gal4p activation domain (AD-Fragment library) (Figure 1). This unbiased approach should identify novel protein interaction domains as well as domains corresponding to computationally defined domain signatures. In addition, using an AD-Fragment library should increase the completeness of interaction networks. Current interactome maps are far from complete, partly due to inherent limitations in the methods used (Venkatesan *et al.*, personal communication). Y2H fusion proteins are frequently incapable of interacting, for example because they do not fold properly in yeast or because the full-length protein is locked in a “closed” conformation that masks potential interaction domains. The use of multiple fragments for each protein in a fragment library increases the probability that at least one fusion product will be capable of interacting in the assay. In addition, false negatives due to underrepresentation of particular proteins can be significantly reduced by using a normalized fragment library as we generate here (Reboul et al., 2003).

We first examined the effect of using a fragment library on specificity and detectability of the Y2H system based on a literature derived set of binary interactions between human proteins (Venkatesan *et al.*, personal communication). Specifically, we tested if the AD-Fragment library approach could recover a higher fraction of 20 literature derived interactions than a full-length clone based approach, while retaining specificity, *i.e.* not identifying interactions between 20 random protein pairs that serve as a negative control. We recovered the 3 literature

derived interactions that we previously found to test positive using full-length constructs (Venkatesan *et al.*, personal communication), as well as 4 additional interactions already described in the literature (Figure 1D). These findings are consistent with the idea that using a fragment library increases the sensitivity of the Y2H system. Importantly, we did not identify any of the 20 randomly selected protein pairs (Figure 1E), suggesting that specificity is not dramatically decreased.

An early embryogenesis interactome domain map

To generate a high quality early embryogenesis AD-Fragment library, we first generated sequence-verified wild-type full-length Gateway (Hartley et al., 2000) entry clones for 681 early embryogenesis proteins (Table S1 and File S1). These clones and an additional 68 full-length PCR products were used as templates in PCR reactions to generate fragments (Figure 1). Most self-folding domains are estimated to be between 100 and 200 residues long (Trifonov and Berezovsky, 2003). We generated all possible fragments up to a size of 800 base pairs (266 residues). In addition, we generated select fragment sizes between 800 base pairs and full length (Figure 1C). Finally, for each ORF we generated three full-length constructs, starting at base pairs 1, 7, and 13, to increase the probability of identifying interactions with (nearly) full-length constructs. In total, we completed 32,158 PCRs for 804 ORFs corresponding to 749 genes, resulting in an average of 40 fragments per ORF (Table S2). PCR fragments were cloned into the Y2H AD vector and pooled to generate the final AD-Fragment library.

As bait proteins, we generated 706 full-length Gal4p DNA binding domain (DB) fusion constructs that do not result in auto-activation of Y2H reporter genes (Walhout and Vidal, 2001a) (Table S2). To obtain the highest coverage possible, the AD-Fragment library should ideally be screened with multiple fusions for each bait protein. As this was not feasible for all ORFs, we tested the benefits of using multiple DB-ORF fusion constructs for two molecular machines: the centrosome and the nuclear pore complex (NPC). For 16 centrosome and 12 NPC proteins (Table S2), we generated 5 additional bait constructs corresponding to the N-terminal and C-terminal fragments spanning ~2/3 of the proteins, and the N-terminal, middle, and C-terminal fragments spanning ~1/3 of the proteins.

All DB-ORF strains were screened against the AD-Fragment library described above, as well as an AD-cDNA library generated from mixed stage *C. elegans* (a kind gift from X. Xin and C. Boone, U. Toronto). To increase the precision of our interaction data set, we eliminated *de novo* autoactivators that arose during the screening process (Vidalain et al., 2004; Walhout and Vidal, 1999), and included only those interactions found in two or more independent yeast colonies. The final data set involves 522 proteins and 755 Y2H interactions between them (Table S3), of which only 92 were previously published or identified by Y2H mapping. Of the 755 interactions, 472 were between early embryogenesis proteins (Figure 2A).

Experimental verification of interactions

To provide an overall estimate of the quality of our data set, we retested a sample of the identified interactions in an independent assay: the Mammalian Protein-Protein Interaction Trap (MAPPIT) (Eyckerman et al., 2001). MAPPIT is based on reconstitution of a JAK/STAT signaling pathway through interaction of a bait protein fused to a receptor lacking STAT binding sites with a prey protein fused to a STAT recruitment domain. Previously we found that MAPPIT recovers 25% ± 4.7% of 40 literature derived interactions between *C. elegans* proteins (Figure 2B) (Simonis *et al.*, personal communication). We tested all pairs for which we had wild-type full-length Gateway clones of both proteins available (355 corresponding to 47% of all interactions). The overall proportion of pairs verified by MAPPIT was 20% ± 2.2%. This represents 80% of the maximum number of interactions expected to test positive using MAPPIT based on the retest rate of the literature derived pairs. Verification by MAPPIT was

only attempted using full-length constructs. This is likely the main reason why interactions originally found with full-length AD-ORF fusions retested at a higher rate than those where only truncated AD-ORF clones were found (29% ± 4.1% and 16% ± 2.4%, respectively).

AD-Fragment library screens increase fraction of detectable interactions

Most interactions between early-embryogenesis proteins (376/472) were found only using the AD-Fragment library. This is likely due to a combination of in-depth screening of a normalized library, and detection of interactions that cannot be detected using full-length constructs. The AD-cDNA library derived interactions enabled us to examine the level of saturation of our AD-Fragment library screens, *i.e.* the fraction of interactions detected out of all interactions that can be identified using the exact Y2H procedure employed here. Out of 96 cDNA derived interactions where both proteins are present in the AD-Fragment library, we recovered 75 (78%) in the AD-Fragment library screens (Figure 2C). This high recovery rate indicates that the AD-Fragment library screens approach saturation.

Most interactions were identified exclusively by AD-ORF clones smaller than the full length ORF (Figure 2D). For the AD-Fragment library, a full-length clone was identified for 34% of interactions – significantly less than the 60% expected based on the contents of the AD-Fragment library and the number of times the library was sampled ($p<1\times10^{-5}$). This indicates that we indeed identify interactions that are difficult or impossible to find using full-length clones.

We examined the properties of proteins that were only identified as truncated AD-ORF clones, and found that these proteins are much larger than those for which a full-length clone was observed (average 777 vs. 393 amino acids). We suspect that this is due to larger proteins folding less efficiently in yeast. In addition, although not statistically significant, proteins found as full length were enriched 3.4 fold for the Gene Ontology (GO) term ‘nuclear’, while proteins found only as truncated clones were enriched 4 and 4.6 fold for the GO terms ‘membrane’ and ‘membrane part’ respectively. This fits well with the notion that the Y2H system, which relies on interactions to occur in the nucleus, may have difficulty identifying interactions with membrane proteins.

Although the MAPPIT results already demonstrated the overall quality of the data set, we also examined whether certain protein regions taken out of context of the full-length protein may become promiscuous interactors. A promiscuously interacting fragment would result in a prey protein connected to many different bait proteins. Bait proteins were only tested as full-length constructs and would lack such highly connected promiscuous interactors. We therefore compared the distribution of connectivity of bait and prey proteins (Figure 2E). We also compared the connectivity distribution of prey proteins found as full-length with prey proteins never found as full-length (Figure 2F). In both cases we observed no significant difference (Mann-Whitney U test p-values >0.96 and >0.92 respectively). Thus, the use of fragments does not appear to result in additional promiscuous interactors.

An expanded network of early embryogenesis

We compared our data set with the most recent version of the worm interactome (CCSB-WI8), which contains 108 interactions between early embryogenesis proteins (http://interactome.dfci.harvard.edu/C_elegans) (Simonis *et al*. personal communication). Our screens found 45 of these, and identified an additional 427 interactions between early embryogenesis proteins (Figure 2A), a nearly 5-fold expansion of interactions between early embryogenesis proteins. In addition, the AD-cDNA library screens identified 283 interactions linking early embryogenesis proteins to the rest of the proteome.

We used two different criteria to establish the biological relevance of our data set. First, we found that 52 of our interactions were previously identified in *C. elegans* or as interologs (Matthews et al., 2001; Walhout and Vidal, 2001b) in other organisms (Table S4), as opposed to 4 interactions when the prey names were shuffled. This result supports the overall biological relevance of our interactions.

We next compared the Y2H interactions with the RNAi phenotypes of the corresponding genes. Detailed phenotypic characterizations are available from RNAi experiments for most of the genes involved in early embryogenesis (Sönnichsen et al., 2005). Out of 320 interactions where a phenotypic profile was determined for both binding partners, 55 (17%) belonged to the same functional class (Figure 3A). To determine the significance of this observation, we calculated the phenotypic similarity between each interacting protein pair (Gunsalus et al., 2005). We found a significant enrichment in protein pairs with similar phenotypes, as well as a significant depletion of pairs with low phenotypic correlation (Figure 3B). In addition, interacting protein pairs were more likely to share functional annotations (GO terms), and show similar mRNA expression profiles (Figure 3C,D).

Finally, we examined whether interactions identified only by truncated clones are as biologically relevant as interactions where a full-length clone was identified. We therefore compared the enrichment in shared GO terms, phenotypes, and expression profiles between these subsets of interactions (Figure S2). We restricted the analysis of interactions where only truncated clones were identified to those interactions where a full-length clone was >50% likely to have been identified. Although the numbers that can be examined are low and there were variations, no significant differences were found between the two sets. Therefore, interactions where only truncated AD-ORF clones were found are not dramatically less biologically relevant by these criteria.

Centrosome assembly and nuclear pore complex architecture

We used our domain-based interaction data set to examine interactions within two different molecular machines: the nuclear pore complex (NPC) and the centrosomes. The first is a symmetric molecular array whose structure has been solved at high resolution using conventional methods, whereas centrosomes, apart from the centriole, have no apparent ultrastructural organization. We first examined the results of using multiple DB-ORF fusion constructs for each bait protein. In the entire screen, 37% of full-length DB-ORF fusions yielded interactors. The use of 5 additional bait constructs for 28 centrosome and nuclear pore proteins resulted in the identification of interactors for 23 of these 8 proteins (82%), illustrating that greater coverage can be obtained by using multiple constructs for each bait protein.

Current understanding of NPC architecture is summarized in Figure 4A (adapted from (Alber et al., 2007; Lim and Fahrenkrog, 2006; Schwartz, 2005). Out of 20 known *C. elegans* NPC proteins (Galy et al., 2003), we used the 12 identified as required for early embryogenesis as bait (Table S2). We identified 6 interactions between NPC proteins and 8 interactions between proteins located near the surface of the NPC and the nuclear import/export machinery (Figure 4A). The relatively low number of binary interactions recovered within the core NPC is consistent with a view of the nuclear pore as an assembly of soluble multiprotein sub-complexes refractory to dissection as binary protein interactions. All but one of the 14 interactions identified are consistent with published interactions and EM localization data for proteins within the NPC (Figure 4A) (Alber et al., 2007; Lim and Fahrenkrog, 2006; Schwartz, 2005). Among the core components, the interaction between NPP-7 (NUP-153) and NPP-10 (NUP96) is novel and suggests a mechanism for anchoring the nuclear basket to the nuclear face of the NPC.

Figure 4B illustrates current understanding of centrosome assembly during the first cell division of *C. elegans*, based primarily on a genetic hierarchy of localization dependencies (Oegema and Hyman, 2006). Centrosome assembly starts with duplication of the centriole, which requires sequential and dynamic recruitment of SPD-2, ZYG-1, and SAS-4, SAS-5, SAS-6 (Dammermann et al., 2008; Delattre et al., 2006; Pelletier et al., 2006). The Polo kinase PLK-1 is also localized to the centriole in a SPD-2 dependent manner (Kemp et al., 2004), although its role in centrosome function is less well understood. Following centriole duplication, the pericentriolar material (PCM) is assembled, a process that is critically dependent on SPD-5, a coiled-coil protein required to recruit all known effector components to the PCM (Dammermann et al., 2004; Hamill et al., 2002). Surprisingly, the only protein known to interact with SPD-5 to date is RSA-2, the centrosome targeting subunit of a protein phosphatase 2A (PP2A) complex (Schlaitz et al., 2007).

We recovered 12 interactions between proteins throughout the centrosome assembly pathway, indicating that this process can be viewed as a set of binary protein-protein interactions that can occur independently of one another. We identified all four previously described direct physical interactions (SAS-5/SAS-6, SPD-5/RSA-2, AIR-1/TPXL-1, and TAC-1/ZYG-9). The remaining intra-centrosomal interactions are novel physical interactions consistent with previous epistatic analyses. The homotypic interactions of SAS-5 and SPD-5 suggest a scaffolding role for these proteins in centriole duplication and PCM assembly, respectively. The binding of both SPD-2 and AIR-1 (the aurora A homolog in *C. elegans*) to SPD-5 provides a testable biochemical model for the genetic requirement of all three proteins for PCM growth. Moreover, both SAS-4 and SPD-2 are required for centriole duplication and bind PLK-1. As SPD-2 is required to target PLK-1 to the centrioles, the role of SPD-2 in centriole duplication might in part be the targeting of PLK-1 to SAS-4.

We also identified two novel interactors of RSA-2: the microtubule-associated proteins TAG-201 and EBP-1. TAG-201 is uncharacterized, while EBP-1 is an evolutionarily conserved protein that binds the growing plus-ends of microtubules. Functional analysis of RSA-2 binding to the microtubule-binding proteins should shed light on how PP2A stabilizes microtubules in mitosis.

Identification and validation of minimal regions of interaction

For each interaction, we defined the minimal region of interaction (MRI) as the smallest region shared by all interacting protein fragments. Our approach was sensitive enough to resolve two independent Ran-binding domains in NPP-9 (Figure 5A). The AD-Fragment library screens defined MRIs in 149 proteins. We observed a small tendency for MRIs to localize toward the C-terminus of proteins (Figure S3). On average, MRIs are 217 amino acids long and correspond to ~39% of their respective fulllength protein (Figure 5B). Only 30 proteins were found solely as full-length fusions (Figure 5B). These proteins were generally small – average length 288 amino acids compared to 565 for all proteins in the AD-Fragment library – and likely consist of a single globular domain that fails to fold properly when truncated. The AD-cDNA derived interactions define MRIs for an additional 134 proteins. However, as the AD-cDNA library contains mostly 5’ deletions, these MRIs are less well refined, with an average length of 400 amino acids, over 67% of their corresponding full-length proteins. Two examples of MRIs that fully encompass a structurally determined binding region are shown in Figure S4, and graphical representations of all MRIs are shown in Figure S5.

To verify the accuracy of the identified MRIs, we first compared them to published interaction domains. For 26 proteins in our data set, interaction domains were present in the literature. For 23 (88%), the MRI identified is consistent with the known interaction site of the *C. elegans* or orthologous protein, demonstrating the accuracy of our approach (Table S4). For three, we found a difference between our MRI and the interaction site of the orthologous human proteins

(Figure 6A). Differences in the MRIs in NPP-7 and NPP-9 and their human counterparts can be explained by evolutionary divergence between the proteins. For example, in our data set IMB-4 binds to the N-terminus of NPP-9, while the mammalian counterpart of IMB-4, Exportin1, binds to a Zinc-finger-rich region located in the center of the NPP-9 homolog RanBP2 (Singh et al., 1999). This region is largely lacking in NPP-9, and motif searches identify only one potential Zinc finger in NPP-9. Interestingly, this region appears subject to rapid evolution, as bovine, mouse, and human RanBP2 have 5, 6, and 8 Zinc fingers, respectively. It is generally assumed that maintaining interactions, especially essential ones, restricts evolutionary drift. These examples indicate that it is possible to maintain an interaction while changing the binding site.

To experimentally demonstrate the functional relevance of novel MRIs, we examined the subcellular localization of SAS-5 and RSA-2 MRIs by fusing them to GFP. SAS-5 localizes to centrioles in a SAS-6 dependent manner, while RSA-2 localizes to the PCM in a SPD-5 dependent manner. We generated transgenic lines expressing GFP fusions of the SAS-5 and RSA-2 MRIs responsible for binding to SAS-6 and SPD-5 respectively. The RSA-2 and SAS-5 MRIs accurately recapitulated the localization of the full-length proteins to the PCM and centrioles, respectively (Figure 6B). SAS-5 MRI localization was observed starting at the ~32 cell stage. The recapitulation of subcellular localization by MRIs further demonstrates their relevance *in vivo*.

Comparison of MRIs with computational predictions

Although protein interactions have traditionally been viewed as being between two structured domains, many interactions involve one structured domain and a short, linear amino acid motif (Davey et al., 2006; Puntervoll et al., 2003) typically present in a disordered loop or tail (Fuxreiter et al., 2007; Mohan et al., 2006). To better understand the structural composition of the MRIs delineated, we examined them for overlap with computational domain and structure predictions (Table S5). The predictors used were: Pfam-A and Superfamily, two collections of manually curated domain signatures (Finn et al., 2008; Gough et al., 2001); Pfam-B, a collection of automatically generated domain signatures (Finn et al., 2008); Ginzu, a protocol using orthologous protein sequences to predict the boundaries of globular domains (Chivian et al., 2003); COILS, a coiled-coil prediction algorithm (Lupas et al., 1991); and two different predictors of disordered regions: PONDR VL-XT (Li et al., 1999; Romero et al., 2001) and VSL2 (Obradovic et al., 2005; Peng et al., 2006). We did not observe enrichment of any domain predictions in MRIs compared to the whole proteins (Figure 6C).

We used the overlap between MRIs and the domain predictions to classify our MRIs as known folding region (Pfam-A, Superfamily, structure-based Ginzu), predicted folding region (Pfam-B, coiledcoil, non-structure-based Ginzu), unstructured region (>50% of residues predicted to be disordered), or potential new folding region. As minimal overlap cutoffs for classifying an MRI we used 20%, 40%, 60%, or 80% of the MRI length. Depending on the cutoff chosen, the fraction of novel folding and disordered MRIs ranges from 14% to 38% (Figure 6D). Interactions with peptide motifs are especially difficult to predict, because they appear frequently at random in a protein. Our data should help narrow searches for linear motifs that mediate interactions.

Finally, we compared our experimentally defined MRIs with binding sites predicted by InSite, a recently developed algorithm that predicts protein-protein interaction binding sites based on the domain composition of proteins (Wang et al., 2007). We used InSite to predict Pfam-A binding sites for those interactions where the MRI overlaps with a single Pfam-A domain, and the protein contains more than one Pfam-A domain. For 78 interactions satisfying these criteria, 53 binding site predictions (68%) matched our experimentally defined MRI. Randomly assigning a Pfam-A domain as binding site for each interaction results in a 35% overlap with

our MRIs. The high overlap between binding site predictions and experimentally defined MRIs further highlights the quality of our approach.

Discussion

The use of an AD-Fragment library provides a way to rapidly map interacting regions in proteins and results in a significant increase in sensitivity of the Y2H system. Randomly generated fragment libraries have already been used to map protein interactions of yeast and *Plasmodium falciparum* (Fromont-Racine et al., 1997; Guglielmi et al., 2004; LaCount et al., 2005). For yeast, the library was generated by randomly fragmenting genomic DNA, an approach that is not applicable to higher eukaryotes as only a small fraction of DNA is coding and most genes contain introns. For *Plasmodium*, the library was generated from cDNA. This approach is applicable to higher eukaryotes, but would suffer from variable representation of different gene products and the presence of 5' and 3' untranslated regions. By starting from full-length ORF clones and using PCR to generate the fragments, we created a nearly normalized library in which each ORF is systematically represented be multiple fragments of different sizes.

To our knowledge, our protein domain data set represents the largest effort to date to experimentally identify protein interaction domains for a higher eukaryote. The MRIs that we identified provide structural information for many early embryogenesis proteins. We expect that the MRIs identified can serve as a foundation for future studies, such as high-resolution structural analysis of these protein interactions *in vitro*, or the targeting of individual interactions for disruption. Although the use of an AD-Fragment library alone provided a dramatic increase in knowledge of the protein interactions underlying *C. elegans* early embryogenesis, even greater coverage can be obtained by using multiple bait constructs. The AD-Fragment library will be made available upon request, and can be used by others interested in increasing understanding of early embryogenesis.

Experimental procedures

Generating wild-type entry clones

To generate wild-type entry clones, predicted ORFs for each early embryogenesis gene were PCR-amplified from a mixed stage *C. elegans* cDNA library, and Gateway cloned into entry vector pDonr223. For each ORF, we sequenced up to 6 individual clones. An entry clone was considered wild-type if it contained no mutations or only silent changes within the open reading frame.

AD-Fragment library generation

Forward and reverse primers with AscI and NotI tails were designed at specific distance intervals across each ORF (75 – 198 base pairs (bp), see Figure 1), and included primers at the start and stop of each ORF. From all possible primer combinations we selected those that create fragments of 800 bp or less. In addition, we selected primer pairs generating two specific fragment sizes between 800 bp and full length (1100 and 1500 bp for ORFs 1000–2000 bp and 1400 and 2000 for ORFs >2000 bp). Finally, we selected the three (nearly) full-length primer pairs starting at positions 1, 7, and 13. Pools of 192 PCR products of similar size were digested with AscI and NotI, and ligated into pPC86-AN (a modified version of pPC86 that contains AscI and NotI sites in-frame with the AD sequence). Nine ORFs contain an AscI or NotI site and PCR fragments containing these sites will be truncated upon digestion. Each ligation yielded >10,000 colonies upon transformation into *E. coli*, while a no insert control yielded <100 colonies. All colonies were washed off each plate, and grown in LB medium for 5 hours before isolating plasmid DNA with a maxiprep kit. All maxipreps were combined to yield the

final AD-Fragment library. To generate AD mating libraries for screening, yeast strain Y8800 was transformed with 30μg of AD-Fragment or 30μg of AD-cDNA library (cDNA library and yeast strains Y8800 and Y8930 a kind gift from X. Xin and C. Boone, U. Toronto). The AD-Fragment library consists of 3.38×10^6 individual colonies and the AD-cDNA of 0.53×10^6 colonies.

Generating Y8930 bait strains

Full-length sequence verified ORFs were transferred to pDest-pPC97 in a Gateway LR reaction. In addition, we cloned 41 full-length ORFs for which no wild-type clone was obtained but a PCR fragment of the right size was generated. Centrosome and NPC Fragment baits were cloned using gap repair. PCR fragments generated during AD-Fragment library creation were further elongated using primers that anneal to the existing AscI and NotI tails. PCR products were transformed into yeast strain Y8930, together with linearized pPC97-AN (a modified version of pPC97 that contains AscI and NotI sites inframe with the DB sequence). All bait strains were plated on Sc –Leu –His plates to eliminate baits able to activate reporter genes in the absence of AD plasmid (auto-activators).

Library screening

Y2H library screens were done using a mating approach (Fromont-Racine et al., 2002). A total of $\sim 6 \times 10^7$ cells of bait yeast and prey library yeast were mixed in equal proportions, and allowed to mate on YEPD for 4 hours before plating on a 15 cm ø Sc –Leu –Trp –His plate. After 4 days of growth at 30°C, colonies were picked for sequence analysis and *de novo* autoactivators were eliminated as described (Vidalain et al., 2004).

Phenotypic comparison

Phenotype correlations between gene pairs range from 0 – 1 (Gunsalus et al., 2005). Fold enrichments were calculated for 4 correlation ranges: 0 – 0.25, 0.25 – 0.5, 0.5 – 0.75, and 0.75 – 1.0. The fold enrichment is the fraction of protein pairs in the interaction network that share a phenotype correlation, relative to the average correlation between all possible pairs of the proteins in the observed interaction network. Significance was calculated using Fisher's exact test.

GO term analysis

Gene Ontology (GO) functional annotations were obtained from the GO database (March 2008 http://www.geneontology.org/). To identify GO terms enriched in one set of proteins, we used Funcassociate (http://llama.med.harvard.edu/cgi/func/funcassociate). To calculate GO term enrichment in protein interactions we used in-house scripts using the R software (http://www.r-project.org). Fisher's exact test was used to calculate significance.

Gene expression profiling comparison

Microarray data from 378 experimental conditions were obtained from WormBase (Table S5). For each pair of genes, the pair-wise Pearson Correlation Coefficient (PCC) was calculated using the R software (http://www.r-project.org), taking into account only the experimental conditions defined for the two genes.

AD-Fragment analysis of human literature derived protein pairs

For the 80 proteins (40 protein pairs), an AD-Fragment library was generated and screened using full-length proteins as described above for *C. elegans* proteins.

Retest by MAPPIT

MAPPIT was performed as described (Eyckerman et al., 2001). Each protein pair is tested in both configurations (bait-prey and prey-bait) and in two independent trials, for a total of four trials. An interaction was scored as positive if at least two of the four trials scored positive.

Generation of GFP-fusion constructs and transgenic lines

Full length *rsa*-2 was cloned into vector TH304 (Green et al., 2008) (C-terminal GFP fusion), *rsa*-2 nucleotides 583 – 1326 was cloned into vector TH315 (Green et al., 2008) (N-terminal S-peptide/GFP fusion), full-length *sas-5* and *sas-5* nucleotides 586 – 1212 were cloned into vectors GFPLAP Gateway (Nterminal S-peptide/GFP fusion) and the newly generated pDest-MB16 (C-terminal GFP fusion). Transgenic lines were generated by microparticle bombardment (Praitis et al., 2001). For SAS-5, the best expressing constructs were selected for imaging.

Comparing MRIs to computational predictions

Pfam-A and Superfamily predictions used scripts available from ftp://ftp.sanger.ac.uk/ and http://www.ebi.ac.uk/interpro/. Coiled-coil and disorder predictions by PONDR VL-XT and VSL2 were performed as described (Li et al., 1999; Lupas et al., 1991; Obradovic et al., 2005; Peng et al., 2006; Romero et al., 2001). Pfam-B predictions used the HMMER2 package (http://hmmer.janelia.org/). Ginzu implements a hierarchically organized combination of sequence based methods (primarily PSI-BLAST, FFAS03 and Pfam) to separate proteins into domains. For comparisons of MRIs to domain predictors, we treated duplicate MRIs with identical start and stops as a single MRI. InSite predictions were performed as previously described (Wang et al., 2007) using 4,542 Y2H interactions and the Pfam-A and Pfam-B domain content of the associated proteins as input.

Classifying MRIs by structure

We first searched for MRIs that share more than a certain fraction of residues (20%, 40%, 60%, or 80%) with Pfam-A domains, Superfamily domains, or Ginzu domains with pdbblast or ffas03 evidence. An MRI matching these domains is classified as ‘Known folding region.’ The remaining MRIs were examined for overlap with Pfam-B, coiled-coil, or Ginzu domain predictions not based on pdb or ffas03 at the same cutoff levels for classification as ‘Predicted folding region.’ The remaining MRIs were split into Unstructured (>50% of amino acids predicted to be disordered) or Novel folding region.

Data availability

The website http://interactome.dfci.harvard.edu/fragdb/ provides a searchable interface with details on interacting fragments and domain predictions for all *C. elegans* Y2H interactions for which such information is available. Interactions have also been submitted to the IMEx consortium (ID: MINT-660970) and can be accessed at http://mint.bio.uniroma2.it/mint/search/interaction.do?interactionAc=MINT-6606970.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to X. Xin and C. Boone for sharing of the cDNA library and yeast strains, to Joe Hargitai for unparalleled parallel computing support, to IBM’s World Community Grid (http://www.wcgrid.org), and to M. Cusick for critical reading of the manuscript. Support was provided by the Leukemia Research Foundation to M.B., the W.M. Keck foundation to M.V., the FWO-V to I.L., NIH grants R21RR023114 (M.B. P.I.), R01HG001715 (M.V. P.I.),

R33CA105405 (M.V. P.I.), CA81658 (M.V. P.I.), R21CA113711 (L.M.I. P.I.), U54 CA011295 (J. Nevins, PI; M.V. sub-contract), and CA95281 (S.v.d.H.), US Army Medical Research Acquisition Activity grant W23RYX-3275-N605 (K.C.G.), NYSTAR grant C040066 (K.C.G.), NSF grants MCB 0444818 to L.M.I. and BDI-0345474 to D.K., and grants IUAP-P6:28, UG-GOA12051401 and FWO-G.0031.06 to J.T..

References

Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, et al. The molecular architecture of the nuclear pore complex. Nature 2007;450:695–701. [PubMed: 18046406]

Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J 3rd. The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci 2005;62:435–445. [PubMed: 15719170]

Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2003;53:524–533. [PubMed: 14579342]

Dammermann A, Maddox PS, Desai A, Oegema K. SAS-4 is recruited to a dynamic structure in newly forming centrioles that is stabilized by the gamma-tubulin-mediated addition of centriolar microtubules. J Cell Biol 2008;180:771–785. [PubMed: 18299348]

Dammermann A, Muller-Reichert T, Pelletier L, Habermann B, Desai A, Oegema K. Centriole assembly requires both centriolar and pericentriolar material proteins. Dev Cell 2004;7:815–829. [PubMed: 15572125]

Davey NE, Shields DC, Edwards RJ. SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. Nucleic Acids Res 2006;34:3546–3554. [PubMed: 16855291]

Delattre M, Canard C, Gonczy P. Sequential protein recruitment in *C. elegans* centriole formation. Curr Biol 2006;16:1844–1849. [PubMed: 16979563]

Eyckerman S, Verhee A, der Heyden JV, Lemmens I, Ostade XV, Vandekerckhove J, Tavernier J. Design and application of a cytokine-receptor-based interaction trap. Nat Cell Biol 2001;3:1114–1119. [PubMed: 11781573]

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, et al. The Pfam protein families database. Nucleic Acids Res 2008;36:D281–D288. [PubMed: 18039703]

Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, Trehin A, Reverdy C, Betin V, Maire S, Brun C, et al. Protein interaction mapping: a *Drosophila* case study. Genome Res 2005;15:376–384. [PubMed: 15710747]

Fromont-Racine M, Rain JC, Legrain P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat Genet 1997;16:277–282. [PubMed: 9207794]

Fromont-Racine M, Rain JC, Legrain P. Building protein-protein networks by two-hybrid mating strategy. Methods Enzymol 2002;350:513–524. [PubMed: 12073333]

Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. Bioinformatics 2007;23:950–956. [PubMed: 17387114]

Galy V, Mattaj IW, Askjaer P. *Caenorhabditis elegans* nucleoporins Nup93 and Nup205 determine the limit of nuclear pore complex size exclusion in vivo. Mol Biol Cell 2003;14:5104–5115. [PubMed: 12937276]

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415:141–147. [PubMed: 11805826]

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, et al. A protein interaction map of Drosophila melanogaster. Science 2003;302:1727–1736. [PubMed: 14605208]

Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 2001;313:903–919. [PubMed: 11697912]

Green RA, Audhya A, Pozniakovsky A, Dammermann A, Pemble H, Monen J, Portier N, Hyman A, Desai A, Oegema K. Expression and imaging of fluorescent proteins in the *C. elegans* gonad and early embryo. Methods Cell Biol 2008;85:179–218. [PubMed: 18155464]

Guglielmi B, van Berkum NL, Klapholz B, Bijma T, Boube M, Boschiero C, Bourbon HM, Holstege FC, Werner M. A high resolution protein interaction map of the yeast Mediator complex. Nucleic Acids Res 2004;32:5379–5391. [PubMed: 15477388]

Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, Berriz GF, Bertin N, Huang J, Chuang LS, et al. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. Nature 2005;436:861–865. [PubMed: 16094371]

Hamill DR, Severson AF, Carter JC, Bowerman B. Centrosome maturation and mitotic spindle assembly in *C. elegans* require SPD-5, a protein with multiple coiled-coil domains. Dev Cell 2002;3:673–684. [PubMed: 12431374]

Hartley JL, Temple GF, Brasch MA. DNA cloning using in vitro site-specific recombination. Genome Res 2000;10:1788–1795. [PubMed: 11076863]

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 2002;415:180–183. [PubMed: 11805837]

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 2001;98:4569–4574. [PubMed: 11283351]

Kemp CA, Kopish KR, Zipperlen P, Ahringer J, O'Connell KF. Centrosome maturation and duplication in *C. elegans* require the coiled-coil protein SPD-2. Dev Cell 2004;6:511–523. [PubMed: 15068791]

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 2006;440:637–643. [PubMed: 16554755]

LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, et al. A protein interaction network of the malaria parasite *Plasmodium falciparum*. Nature 2005;438:103–107. [PubMed: 16267556]

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, et al. A map of the interactome network of the metazoan *C. elegans*. Science 2004;303:540–543. [PubMed: 14704431]

Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Predicting protein disorder for N-, C-,and internal regions. Genome Informatics 1999;10:30–40. [PubMed: 11072340]

Lim RY, Fahrenkrog B. The nuclear pore complex up close. Curr Opin Cell Biol 2006;18:342–347. [PubMed: 16631361]

Liu J, Rost B. CHOP proteins into structural domain-like fragments. Proteins 2004;55:678–688. [PubMed: 15103630]

Lupas A, Van Dykec M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252:1162–1164.

Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res 2001;11:2120–2126. [PubMed: 11731503]

Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. Analysis of molecular recognition features (MoRFs). J Mol Biol 2006;362:1043–1059. [PubMed: 16935303]

Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins. 2005

Oegema K, Hyman AA. Cell division. WormBook 2006:1–40. [PubMed: 18050484]

Pawson T, Nash P. Assembly of cell regulatory systems through protein interaction domains. Science 2003;300:445–452. [PubMed: 12702867]

Pelletier L, O'Toole E, Schwager A, Hyman AA, Muller-Reichert T. Centriole assembly in *Caenorhabditis elegans*. Nature 2006;444:619–623. [PubMed: 17136092]

Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 2006;7:208. [PubMed: 16618368]

Piano F, Schetter AJ, Morton DG, Gunsalus KC, Reinke V, Kim SK, Kemphues KJ. Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. Curr Biol 2002;12:1959–1964. [PubMed: 12445391]

Praitis V, Casey E, Collar D, Austin J. Creation of low-copy integrated transgenic lines in *Caenorhabditis elegans*. Genetics 2001;157:1217–1226. [PubMed: 11238406]

Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, et al. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 2003;31:3625–3630. [PubMed: 12824381]

Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R, et al. *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. Nat Genet 2003;34:35–41. [PubMed: 12679813]

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins 2001;42:38–48. [PubMed: 11093259]

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005;437:1173–1178. [PubMed: 16189514]

Schlaitz AL, Srayko M, Dammermann A, Quintin S, Wielsch N, MacLeod I, de Robillard Q, Zinke A, Yates JR 3rd, Muller-Reichert T, et al. The *C. elegans* RSA complex localizes protein phosphatase 2A to centrosomes and regulates mitotic spindle assembly. Cell 2007;128:115–127. [PubMed: 17218259]

Schwartz TU. Modularity within the architecture of the nuclear pore complex. Curr Opin Struct Biol 2005;15:221–226. [PubMed: 15837182]

Singh BB, Patel HH, Roepman R, Schick D, Ferreira PA. The zinc finger cluster domain of RanBP2 is a specific docking site for the nuclear export factor, exportin-1. J Biol Chem 1999;274:37370–37378. [PubMed: 10601307]

Sönnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, Brehm M, Alleaume AM, Artelt J, Bettencourt P, Cassin E, et al. Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. Nature 2005;434:462–469. [PubMed: 15791247]

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005;122:957–968. [PubMed: 16169070]

Trifonov EN, Berezovsky IN. Evolutionary aspects of protein structure and folding. Curr Opin Struct Biol 2003;13:110–114. [PubMed: 12581667]

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000;403:623–627. [PubMed: 10688190]

Vidalain PO, Boxem M, Ge H, Li S, Vidal M. Increasing specificity in high-throughput yeast two-hybrid experiments. Methods 2004;32:363–370. [PubMed: 15003598]

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science 2000;287:116–122. [PubMed: 10615043]

Walhout AJ, Vidal M. A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. Genome Res 1999;9:1128–1134. [PubMed: 10568752]

Walhout AJ, Vidal M. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. Methods 2001a;24:297–306. [PubMed: 11403578]

Walhout AJ, Vidal M. Protein interaction maps for model organisms. Nat Rev Mol Cell Biol 2001b;2:55–62. [PubMed: 11413466]

Wang H, Segal E, Ben-Hur A, Li QR, Vidal M, Koller D. InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. Genome biology 2007;8:R192. [PubMed: 17868464]

Zipperlen P, Fraser AG, Kamath RS, Martinez-Campos M, Ahringer J. Roles for 147 embryonic lethal genes on *C.elegans* chromosome I identified by RNA interference and video microscopy. Embo J 2001;20:3984–3992. [PubMed: 11483502]

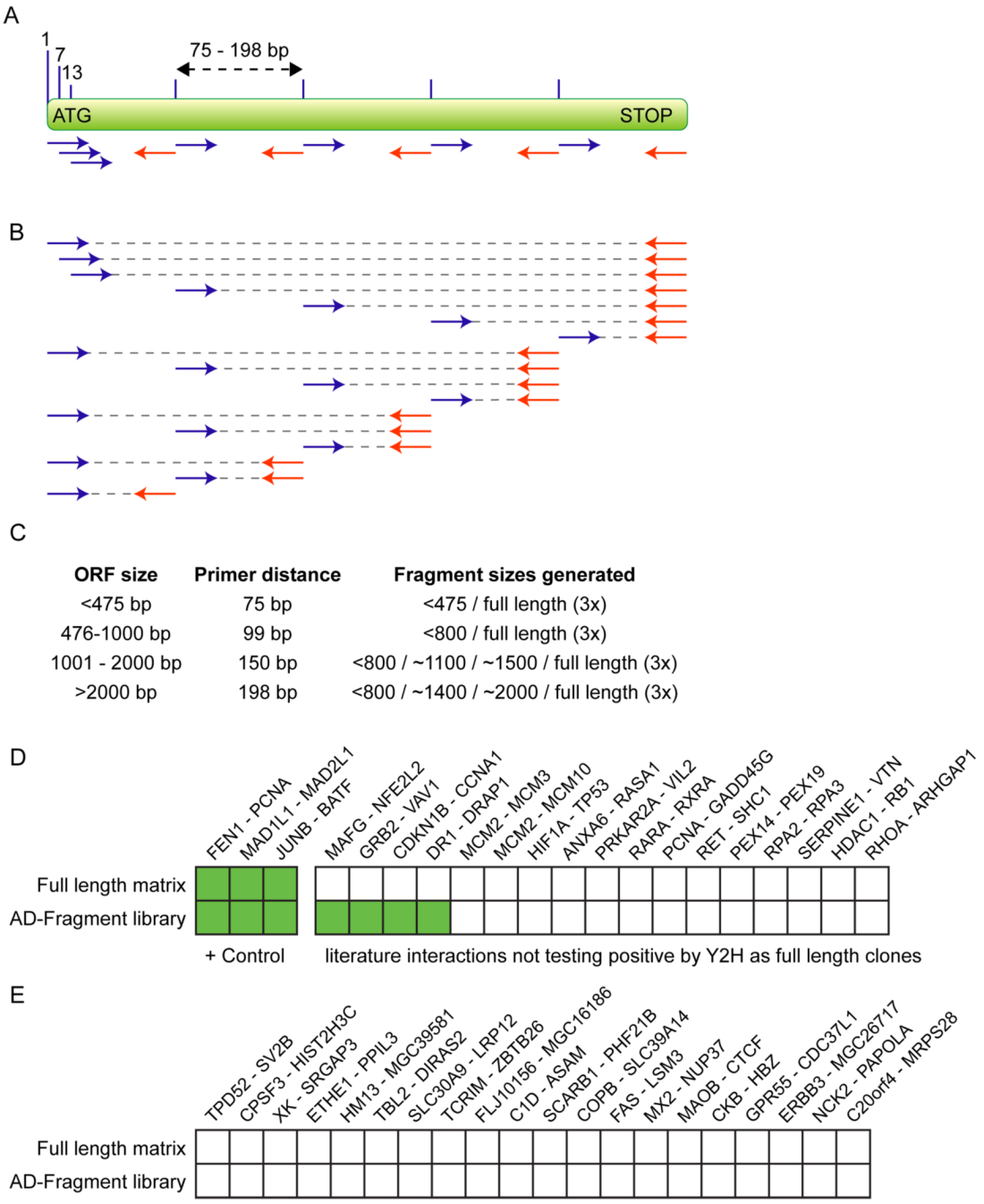

Figure 1. Strategy for generating the AD-Fragment library and effect on Y2H sensitivity and specificity

(A) Primer placement. Primers are designed to start within a 55 bp window surrounding the ideal start positions (lines above ORF).

(B) Fragments generated by combining primers.

(C) Distances in between primers and fragment sizes produced for ORFs of the indicated lengths.

(D,E) Literature derived interactions and random protein pairs tested as full-length fusions (results from Venkatesan *et al.* personal communication) and using an AD-Fragment library. Green boxes indicate detection of an interaction. Protein names correspond to Entrez names.

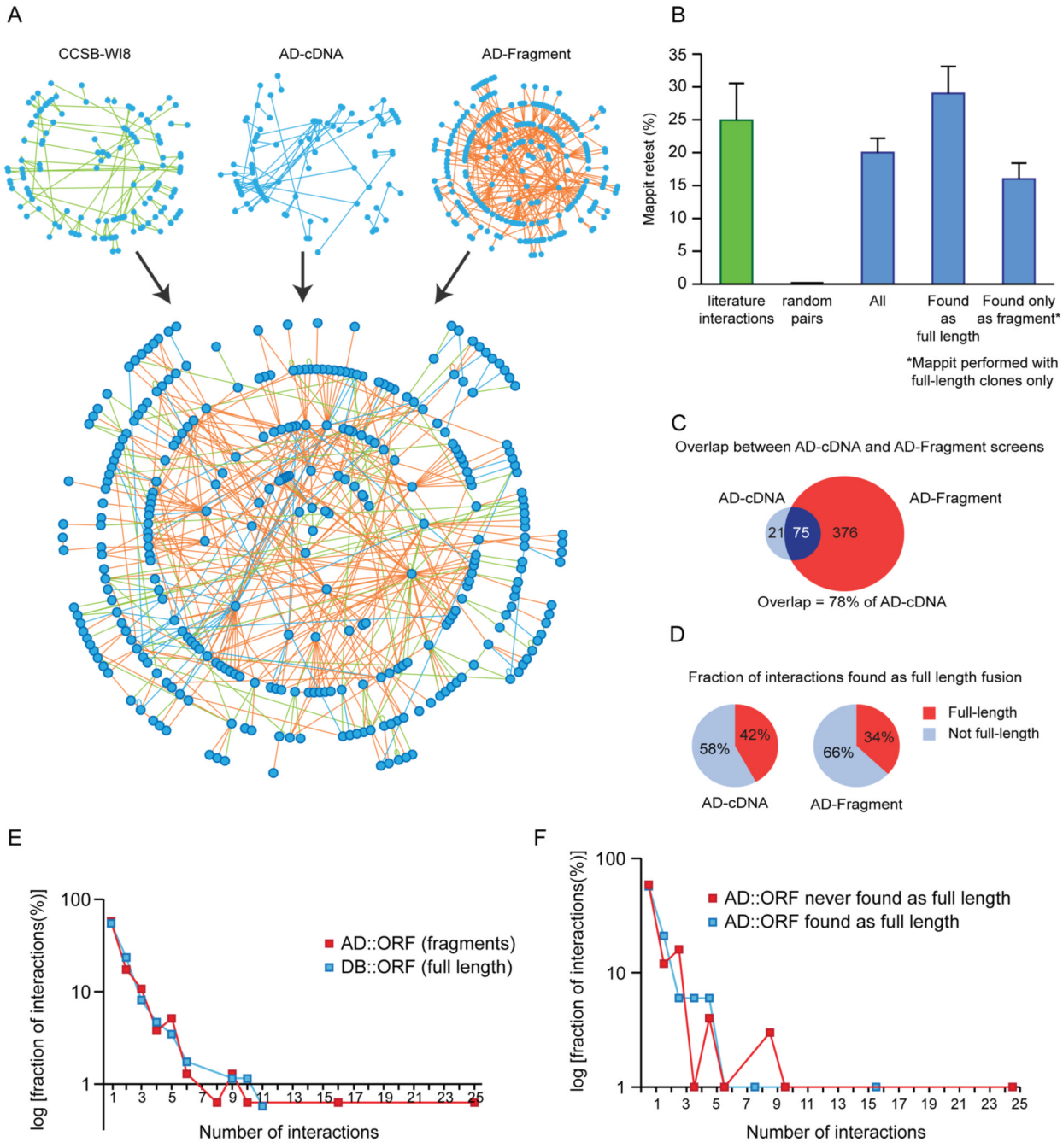

Figure 2. Properties of the Y2H protein-protein interaction network

(A) Network graph of the protein-protein interactions between early embryogenesis proteins, compiled from data in the most recent release of the worm interactome (CCSB-WI8), and from the AD-cDNA and AD-Fragment screens described here.

(B) Retest rate of interactions in MAPPIT. Green bar: interactions derived from literature (results from Simonis *et. al.* personal communication). Random protein pairs did not interact. Blue bars: retest of 355 interactions described here, split into: (1) all 355 interactions, (2) those found as full-length fusions (124 interactions), and (3) those found as truncated fusions only (225 interactions). Error bars correspond to binomial standard error.

(C) Overlap between AD-cDNA and AD-Fragment library derived interactions within the early embryogenesis protein space.
(D) Fraction of interactions found as full-length fusions in AD-cDNA and AD-Fragment library screens.
(E) Comparison of connectivity of bait and prey proteins.
(F) Comparison of connectivity of prey proteins that were found as full-length at least once, with those that were never found as full-length.

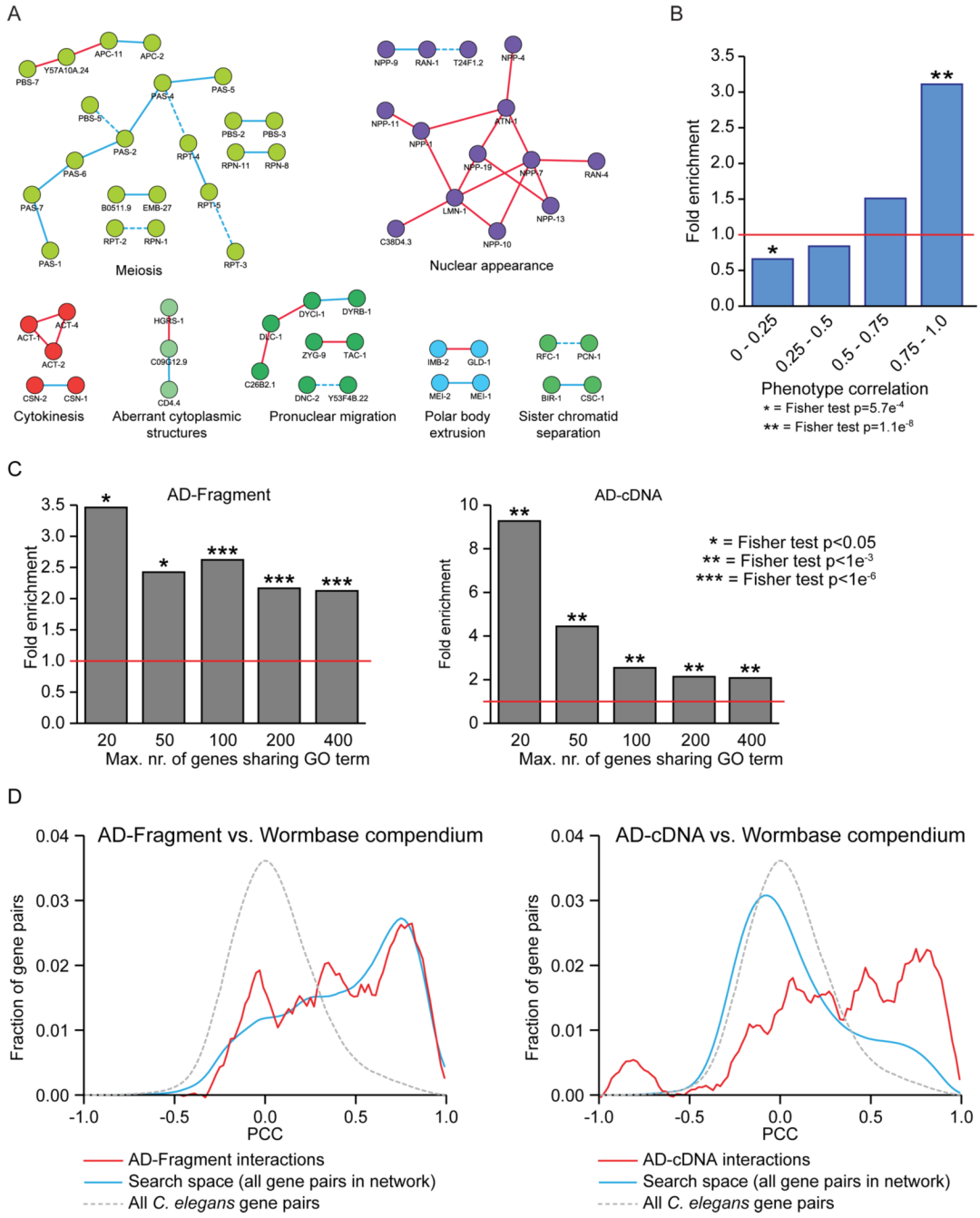

Figure 3. Enrichment in similar phenotypes, GO terms, and mRNA expression profiles for interacting protein pairs

(A) Examples of interactions between proteins assigned to the same functional class based on their RNAi phenotypes. Red lines: new Y2H interactions. Blue lines: known Y2H interactions re-identified. Blue dotted lines: known Y2H interactions not found.

(B) Enrichment in phenotypic correlation for interacting protein pairs relative to average value of all possible protein pairs in the interaction network.

(C) Enrichment in shared GO terms at different levels of specificity.

(D) Pearson correlation coefficients (PCCs) for the mRNAs corresponding to each pair of proteins in the interaction data sets (red lines), the protein space searched (blue lines), and the

entire worm genome (dotted grey lines). Early embryogenesis genes already have highly similar expression profiles compared to the entire worm genome, hence no further enrichment can be observed for interactions derived from the AD-Fragment library (left panel).

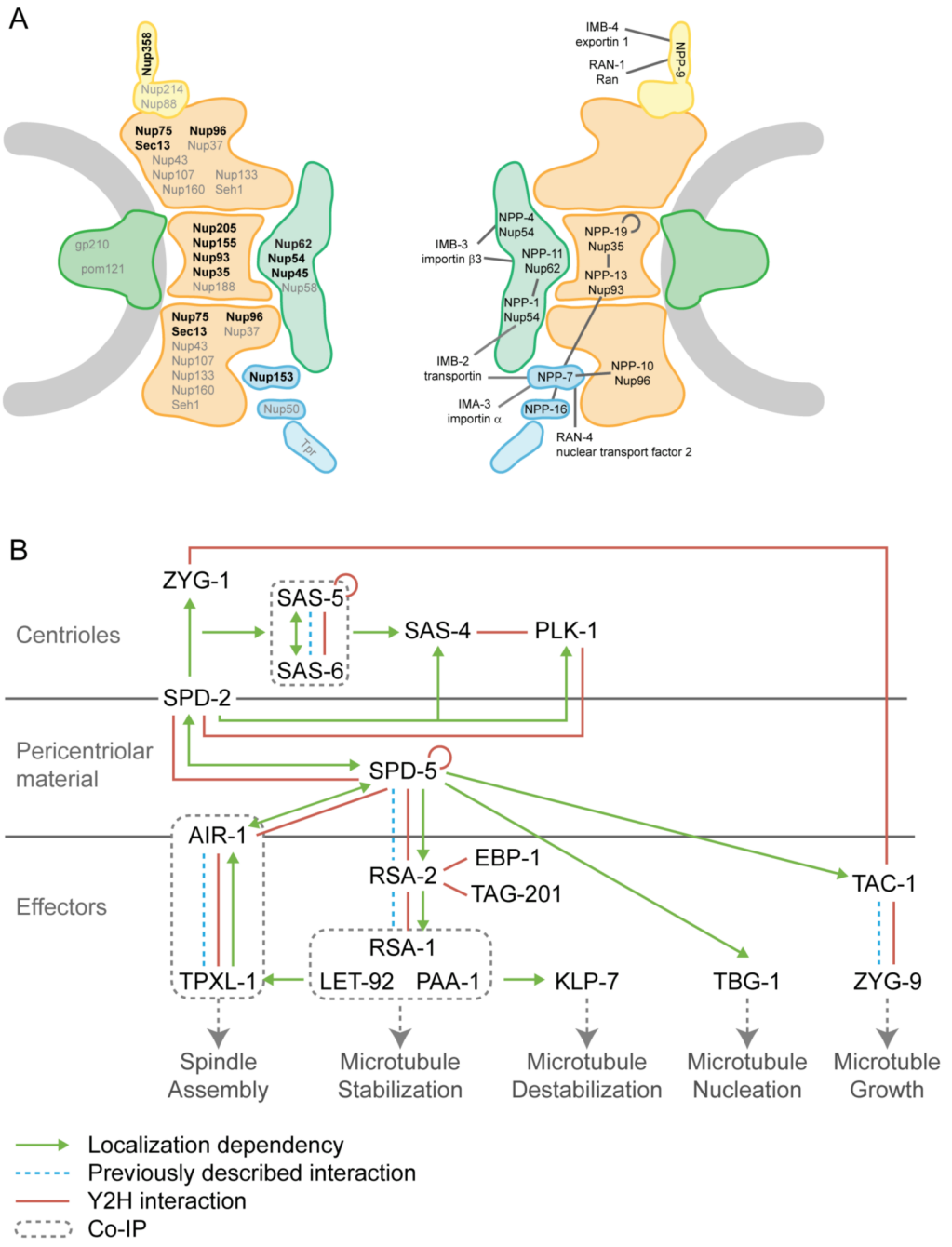

Figure 4. Y2H results of nuclear pore complex (NPC) and centrosome screens

(A) Schematic drawing of NPC. Shown are nuclear membrane (grey) with membrane rings (green), inner and outer scaffold rings (orange), FG nucleoporins (green), cytoplasmic tendrils (yellow), and nuclear basket (blue). Left: approximate localization of mammalian proteins within the NPC. *C. elegans* homologs of proteins in black were used as baits in our screens. Right: Interactions found between *C. elegans* NPC and import/export machinery proteins. (B) Diagram of centrosome assembly pathway. Green arrows represent localization dependencies, dotted blue lines previously described binary interactions, red lines Y2H interactions discovered here, and dotted boxes co-IP complexes.

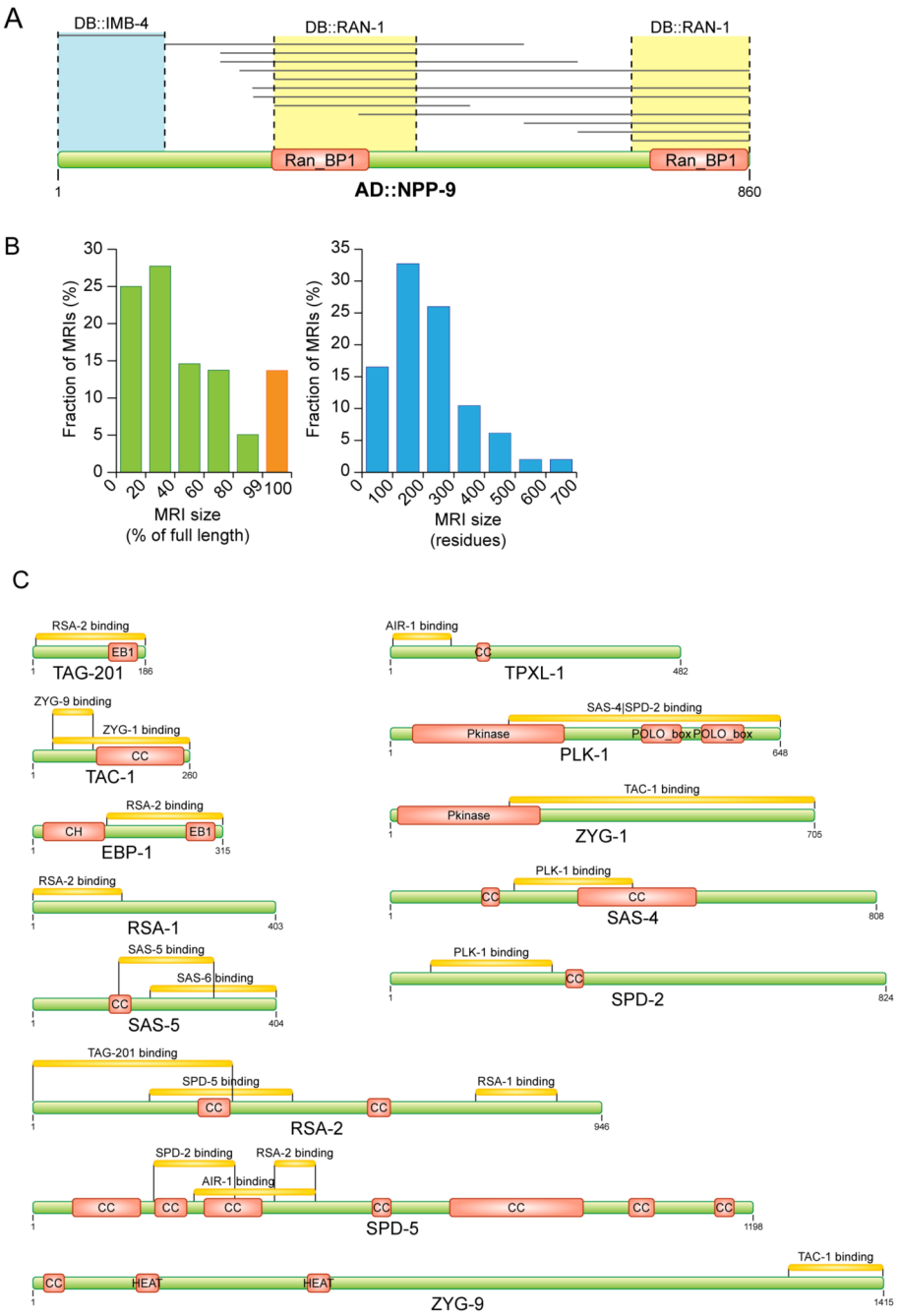

Figure 5. Identification and validation of minimal regions required for interaction (MRIs)

(A) Example of identification of an MRI. The AD-Fragment library was screened with full length DB::RAN- 1 and DB::IMB-4. Grey lines indicate protein fragments of NPP-9 that interacted with RAN-1 or IMB-4.

(B) Sizes of MRIs identified in the AD-Fragment library screens expressed as percentage of corresponding full-length protein and absolute amino acids.

(C) MRIs identified in proteins involved in centrosome assembly. Green bars represent full-length proteins. Yellow bars represent regions of the full-length protein required for interaction with the indicated binding partner (e.g. the N-terminal region of TPXL-1 is required for binding

to AIR-1). Pfam-A domain signatures are drawn as red boxes. CC = coiled-coil prediction. The region of RSA-2 that mediates binding to SPD-5 was further refined manually (not shown).

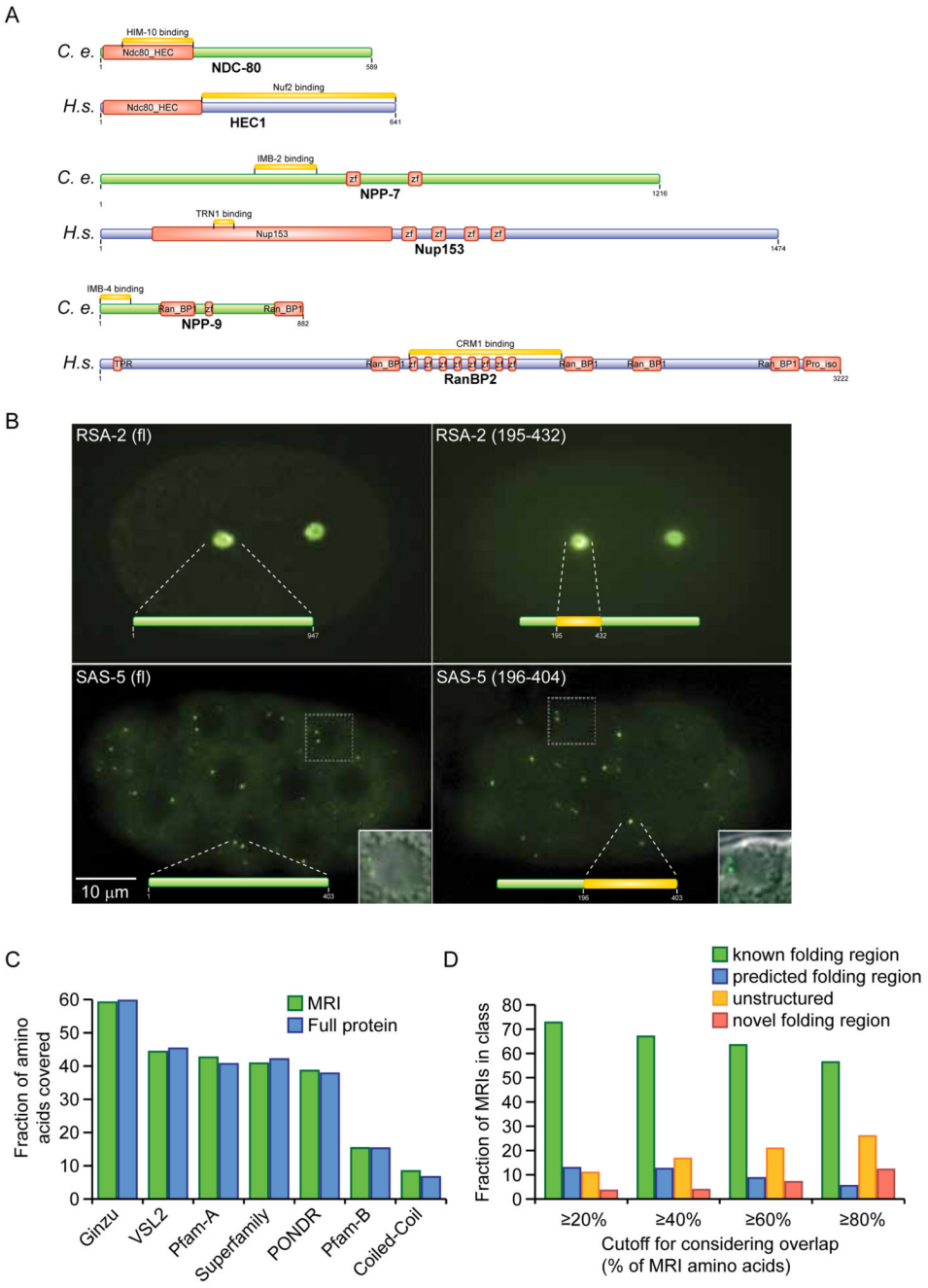

Figure 6. Comparison of MRIs with computational domain predictions

(A) Three cases where interacting regions differ between *C. elegans* and the orthologous proteins in human.

(B) Localization of GFP fusions of full-length RSA-2 and SAS-5 and their MRIs required for binding to SPD-5 and SAS-6, respectively.

(C) Fraction of amino acids of MRIs and the corresponding full proteins that are covered by computationally predicted domains of the indicated types.

(D) Fraction of MRIs classified as ‘known folding region,’ ‘predicted folding region,’ ‘unstructured,’ or ‘novel folding region,’ based on overlap with computational predictions.