nature genetics

A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*

Insuk Lee[1,4], Ben Lehner[2–4], Catriona Crombie[2], Wendy Wong[2], Andrew G Fraser[2] & Edward M Marcotte[1]

The fundamental aim of genetics is to understand how an organism's phenotype is determined by its genotype, and implicit in this is predicting how changes in DNA sequence alter phenotypes. A single network covering all the genes of an organism might guide such predictions down to the level of individual cells and tissues. To validate this approach, we computationally generated a network covering most *C. elegans* genes and tested its predictive capacity. Connectivity within this network predicts essentiality, identifying this relationship as an evolutionarily conserved biological principle. Critically, the network makes tissue-specific predictions—we accurately identify genes for most systematically assayed loss-of-function phenotypes, which span diverse cellular and developmental processes. Using the network, we identify 16 genes whose inactivation suppresses defects in the retinoblastoma tumor suppressor pathway, and we successfully predict that the dystrophin complex modulates EGF signaling. We conclude that an analogous network for human genes might be similarly predictive and thus facilitate identification of disease genes and rational therapeutic targets.

The central goal of genetics is to understand how heritable information encoded in the genome determines the phenotype of an organism. In coming decades, the number of individual human genomes sequenced will grow enormously, and the key emerging problem will be to correlate identified genomic variation to phenotypic variation in health and disease. However, our present ability to predict the outcome of an inherited change in the activity of any single human gene is negligible. We therefore need pragmatic approaches that use current biological knowledge to relate genetic change to phenotypic outcome in multicellular animals.

Probabilistic integrated gene networks, in which linkages between genes indicate their likelihood of being involved in the same biological processes, provide such an approach[1–6]. Developments in functional genomics, proteomics and comparative genomics now allow genome-scale assays of different aspects of gene function. Although datasets from each technique are incomplete and error-prone, they contain much valuable data. Using statistical methods, one can integrate these data to generate a more accurate and comprehensive view of gene function than is contained in any single dataset. Such a network, comprising the majority of genes, might then guide phenotypic predictions in varied tissues and developmental contexts.

Integrated networks have proven successful for the study of unicellular organisms, accurately predicting gene functions[3] and essentiality[7,8]. However, it is unclear whether these methods may be extended to studies of multicellular animals, because the networks do not explicitly reflect multiple cell types, tissues or stages of development. Further, although one might achieve high prediction specificity for small gene networks, retaining this specificity for a network spanning an entire animal proteome (that is, all protein-encoding genes) is challenging and unproven. In our studies, we tested whether it was possible to generate an integrated gene network covering most genes in an animal and also retain the capacity to predict effects of gene perturbation in individual tissues *in vivo*.

We chose to construct a proteome-scale gene network for the model organism *C. elegans*. Not only is the worm complex, with many different cell and tissue types, but the effects of perturbing any gene (or combination of genes) *in vivo* can be rapidly investigated using RNA interference (RNAi). In the worm, one can thus test the ability of a single gene network to predict both gross phenotypes (for example, lethality) and tissue-specific phenotypes. Validated systems approaches for predicting animal gene function might then be applied to human biology and disease.

In this study, we constructed a genetic network comprising ~82% of predicted *C. elegans* genes. We show that this network predicts the effects of perturbing individual genes on the organism's phenotype, identifying genes causing specific phenotypes ranging from cell cycle defects in single embryonic cells to life-span alterations, neuronal defects, and altered patterning of specific tissues. We illustrate how such a network can accelerate research by using network-guided screening to identify genes and interactions for two pathways relevant

[1]Center for Systems and Synthetic Biology, Department of Chemistry and Biochemistry, Institute for Cellular and Molecular Biology, University of Texas, 2500 Speedway, MBB 3.210, Austin, Texas 78712, USA. [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [3]European Molecular Biology Laboratory-Centre for Genomic Regulation (EMBL-CRG) Systems Biology Unit and Institució Catalana de Recerca i Estudis Avançats (ICREA), Center for Genomic Regulation, Universitat Pompeu Fabra, C/ Dr. Aiguader 88, 08003 Barcelona, Spain. [4]These authors contributed equally to this work. Correspondence should be addressed to A.G.F. (agf@sanger.ac.uk) or E.M.M. (marcotte@icmb.utexas.edu).

Received 4 May 2007; accepted 6 November 2007; published online 27 January 2008; doi:10.1038/ng.2007.70

to human disease. We have thus constructed a key resource for gene function prediction in *C. elegans*, which should facilitate characterization of important processes in this model animal. More broadly, this study suggests that similar integrative networks in humans may provide a method for predicting how individual inherited variations in gene sequence and activity relate to individual phenotypes.

RESULTS

Constructing a proteome-scale gene network for *C. elegans*

For a network to predict the phenotypic outcome of perturbing any individual gene, it must include all predicted genes. Previously described gene networks for *C. elegans* include only 3–17% of genes and thus fall short of this goal[9–12]. Maintaining high accuracy of functional predictions while increasing coverage to include most genes is non-trivial, and many methods that work for the strongest linkages (for example, linkages shared between datasets) fail when coverage is increased. For yeast, we and others have shown that integrating high-throughput datasets in a Bayesian framework can produce an accurate probabilistic network that both covers most genes and makes specific predictions of gene function and perturbation phenotypes[2–4,6]. We thus used this approach in the worm.

We assembled diverse datasets, including DNA microarray measurements of the expression of *C. elegans* mRNAs (**Supplementary Table 1** online), assays of physical and/or genetic interactions (**Supplementary Table 2** online) among *C. elegans*[10,13], fly[14], human, and yeast proteins, literature-mined *C. elegans* gene associations, functional associations of yeast orthologs[15] (we term such conserved functional linkages 'associalogs'), estimates of the coinheritance of *C. elegans* genes across bacterial genomes, and the operon structures of bacterial and/or archaeal homologs of *C. elegans* genes. We analyzed over 20 million experimental observations, and where necessary, we mapped observations from other eukaryotes to *C. elegans* via gene orthology[16]. Each dataset links genes either explicitly (for example, through physical and/or genetic interactions) or by inference (as genes sharing expression profiles tend to be functionally linked). Because these data are noisy, a naïve union of these datasets generates a large but error-prone network with poor predictive capacity. Conversely, each dataset in itself is incomplete, and although finding overlaps between multiple datasets identifies high-confidence linkages, it generates a low-coverage network that excludes much high-quality data. More sophisticated methods are thus required to build accurate, high-coverage gene networks.

To integrate these datasets, we first estimated how well each dataset links genes known to share biological functions as determined from Gene Ontology (GO) annotations; we expressed these estimates as log likelihood scores (LLS; **Fig. 1a** and **Supplementary Methods** online). These scores are not highly dependent on the use of GO annotations for quality assessment; other schemes for function annotation (for example, KEGG) yield similar results. After generating quality scores for each dataset, we then integrated all datasets into a single network using a modified Bayesian framework (see Methods). This yielded a network (Wormnet v1) comprising 384,700 linkages among 16,113 genes (~82% of the worm proteome). Each linkage is not a binary entity, but instead indicates a probability based on integration of likelihood scores from individual datasets—some linkages are of high confidence, others are only weak. We observed up to five lines of

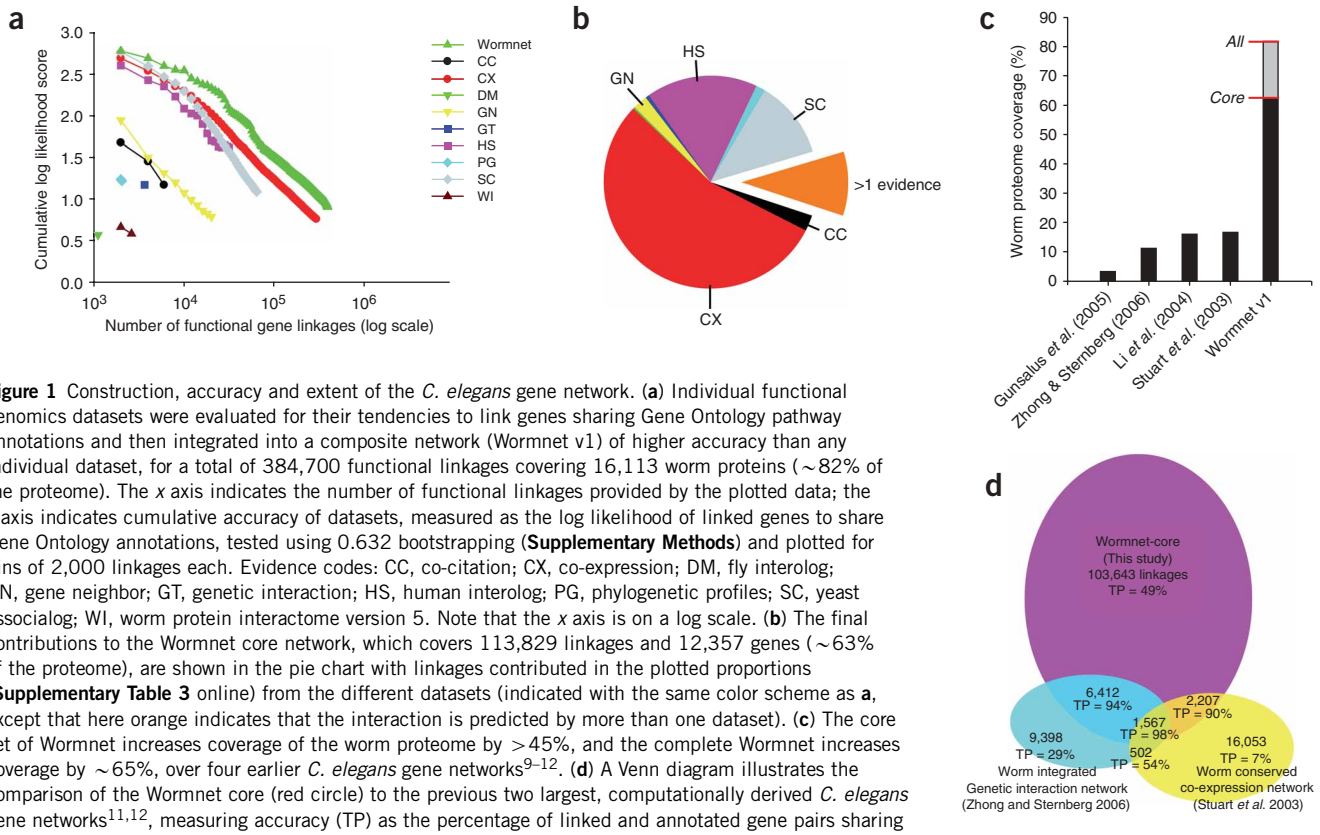

Figure 1 Construction, accuracy and extent of the *C. elegans* gene network. (**a**) Individual functional genomics datasets were evaluated for their tendencies to link genes sharing Gene Ontology pathway annotations and then integrated into a composite network (Wormnet v1) of higher accuracy than any individual dataset, for a total of 384,700 functional linkages covering 16,113 worm proteins (~82% of the proteome). The *x* axis indicates the number of functional linkages provided by the plotted data; the *y* axis indicates cumulative accuracy of datasets, measured as the log likelihood of linked genes to share Gene Ontology annotations, tested using 0.632 bootstrapping (**Supplementary Methods**) and plotted for bins of 2,000 linkages each. Evidence codes: CC, co-citation; CX, co-expression; DM, fly interolog; GN, gene neighbor; GT, genetic interaction; HS, human interolog; PG, phylogenetic profiles; SC, yeast associalog; WI, worm protein interactome version 5. Note that the *x* axis is on a log scale. (**b**) The final contributions to the Wormnet core network, which covers 113,829 linkages and 12,357 genes (~63% of the proteome), are shown in the pie chart with linkages contributed in the plotted proportions (**Supplementary Table 3** online) from the different datasets (indicated with the same color scheme as **a**, except that here orange indicates that the interaction is predicted by more than one dataset). (**c**) The core set of Wormnet increases coverage of the worm proteome by >45%, and the complete Wormnet increases coverage by ~65%, over four earlier *C. elegans* gene networks[9–12]. (**d**) A Venn diagram illustrates the comparison of the Wormnet core (red circle) to the previous two largest, computationally derived *C. elegans* gene networks[11,12], measuring accuracy (TP) as the percentage of linked and annotated gene pairs sharing KEGG pathway annotations. Linkages common to all sets are 98% accurate, those shared between Wormnet and one other set are 90% or 94% accurate, and the ~104,000 linkages unique to Wormnet are still 49% accurate. Similar results are seen in a comparison to the other two previous *C. elegans* gene networks[9,10] (data not shown).

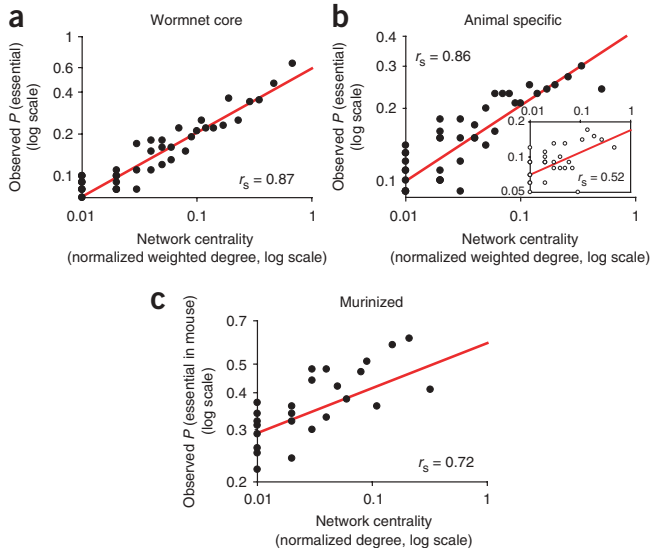

Figure 2 Network topology-based prediction of gene essentiality. (**a**) The connectivity of a gene, measured for each gene as the normalized sum of log likelihood scores of its linkages, is highly correlated with the tendency for the gene to be essential. All genes in Wormnet were sorted by connectivity and binned into successive bins of 200 genes; the mean connectivity and frequency of tested genes conferring nonviable RNAi phenotypes[19] were then calculated for each bin (bins are plotted as filled circles; same bin size for plots **a**–**c**). (**b**) The correlation is not a simple consequence of including yeast-derived linkages, as a *C. elegans* network with all yeast-derived linkages removed shows a similar trend. Moreover, the trend is still evident and significant ($P < 0.01$) even after removal of all genes with yeast orthologs (**b**, inset). (**c**) This trend is evolutionarily conserved across animals—connectivity in a 'murinized' *C. elegans* gene network (the subset of 6,924 genes with mouse orthologs[16]) correlates with the tendency for the orthologous mouse genes to be essential (measured as the frequency of tested mouse genes whose knockouts and/or gene disruptions result in prenatal or perinatal mortality[49]). Here, connectivity is measured as the normalized number of linkages per gene.

evidence per linkage; the number of lines of evidence correlates well with increasing LLS (**Supplementary Methods**). Applying an empirical cut-off to the complete network yields a higher confidence core network comprising 113,829 linkages for 12,357 worm genes (~63% of the proteome). **Figure 1b** indicates the relative contributions of each dataset to the core network.

The core network is substantially larger than earlier *C. elegans* functional networks[9–12], contributing 102,088 previously unidentified linkages and covering an additional ~45% of the *C. elegans* proteome (**Fig. 1c**). This increased coverage does not come at the expense of accuracy: Wormnet has a comparable or better false-positive rate than the previous smaller networks, as assessed by the recapitulation of known functional relationships in the GO and KEGG databases (**Fig. 1d** and **Supplementary Methods**). The network extends considerably beyond previously described associations: 83,946 links in the core network (74%) neither derive from literature-mined relationships nor overlap with known Gene Ontology pathway relationships. Notably, the core network does not have a scale-free node degree distribution, but is better fit by a power-law for genes with lower connectivity and an exponential decay for genes above a characteristic threshold (**Supplementary Methods**).

A key goal of this network is that it should predict the phenotypic consequences of perturbing genes. We thus tested the capacity of this network to predict loss-of-function phenotypes of different levels of specificity. We first tested whether the network could predict gene essentiality. We next tested whether we could accurately identify genes with diverse functions ascertained through previous genome-wide RNAi screens. Finally, we tested whether we could identify previously unknown genes that modulate pathways relevant to human disease and then experimentally validate these predictions. In this way, we validate the ability of this single network to make accurate, specific predictions about the outcome of gene perturbations on the phenotype of the entire organism.

Network connectivity predicts essentiality in animals

We first examined whether the core network could predict gene essentiality. Although gene essentiality is a general prediction, it is nonetheless important—for example, accurately predicting mouse gene essentiality would influence the targeting strategy for major knockout programs. Similarly, there are thousands of potential drug targets in the 'druggable genome'—knowing which targets are essential could refine the search space.

The number of interaction partners a yeast protein has correlates with the tendency of the gene to be essential[7]. It was recently suggested that this relationship might not hold for animal gene networks[17], although a similar trend was observed for *C. elegans* protein-DNA interactions[18]. We observed a robust correlation between gene connectivity in the functional network and the frequency of nonviable RNAi phenotypes[19] (Spearman rank correlation $r_s = 0.87$; **Fig. 2a**). This trend is not simply a consequence of yeast-derived linkages: a worm network with all yeast-derived linkages omitted also shows a strong correlation ($r_s = 0.86$; **Fig. 2b**), and the relationship is statistically significant ($P < 0.01$) even after removing all genes with yeast orthologs from the network ($r_s = 0.52$; **Fig. 2b** inset). Thus, we can predict the probability that a gene has a non-viable RNAi phenotype using only network connectivity.

We further tested whether this relationship between connectivity and essentiality is likely to hold for mammalian gene networks. We constructed a mouse gene network using linkages from Wormnet and the orthology relationships between mouse and worm genes. Although this 'murinized' network did not contain any data from the mouse in forming the gene linkages, we found that there was again a strong relationship between the connectivity of a gene and the probability of embryonic or perinatal lethality in knockout and/or gene-disrupted mice ($r_s = 0.72$, **Fig. 2c**). This finding demonstrates that network connectivity predicts essentiality from yeast to animals and suggests that this likely represents a fundamental principle of biology.

A single network can predict diverse phenotypes

Genes that are tightly linked in probabilistic networks of the kind we have constructed here are very likely to be involved in the same biological process (**Fig. 1**); thus, genes linked in Wormnet are likely to share similar loss-of-function phenotypes. The key to the success of this 'guilt by association' approach[20,21] is in the combined high quality and high coverage of the linkages.

To test the capability of Wormnet to identify new genes sharing loss-of-function phenotypes with previously studied genes, we first used data deriving from genome-wide RNAi screens. We did not use linkages deriving from any of these screens to construct Wormnet. Forty-three RNAi phenotypes have been screened on a genome-wide scale. These phenotypes range from gross (for example, sterility, lethality and growth defects) to specific, such as those affecting single

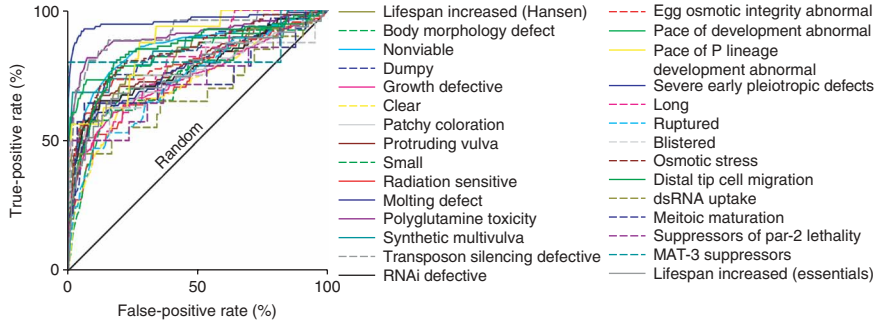

Figure 3 Network-based prediction of loss-of-function phenotypes. Genes conferring a specific RNAi phenotype can be predicted on the basis of connectivity to other genes conferring the phenotype (the 'seed' set), as shown here by leave-one-out prediction of genes giving rise to each of the 29 indicated RNAi phenotypes (also see **Supplementary Figure 1**). For a given phenotype, each gene in the worm proteome was rank-ordered by the sum of its linkage log likelihood scores to the seed set of genes already known to show that phenotype (omitting each seed gene from the seed set for purposes of evaluation). High-scoring genes are therefore most tightly connected to the seed set and are the most likely candidates to share that phenotype. This trend is evident in a ROC plot in which we measure recovery of genes with the given phenotype as a function of rank, calculating the true-positive prediction rate (sensitivity; TP/(TP+FN)) versus the false-positive prediction rate (1–specificity; FP/(FP+TN)).

cellular processes (for example, mismatch repair defects, apoptosis) or single tissues (for example, vulva development). For each genome-wide RNAi screen, we wanted to assess how tightly linked the hits were to each other relative to other genes. If genes that share any given loss-of-function phenotype associated tightly together, this would indicate that Wormnet has the capability to identify additional genes sharing loss-of-function phenotypes with previously studied genes.

To examine how tightly genes with a given loss-of-function phenotype were associated, we calculated how closely every gene in the complete network was linked to the set of previously identified genes (the 'seed' set) conferring that phenotype (see Methods). We found that the network strongly predicts genes with 29 of the 43 phenotypes reported from genome-wide screens (**Fig. 3** and **Supplementary Figure 1**); a further 10 screens can be reasonably predicted (**Supplementary Methods**). Previously described networks are not as capable at identifying the associations between the majority of these genes (**Supplementary Methods**), primarily because of lack of proteome coverage (3–17%, compared to 82% in Wormnet).

In the cases where we failed to make strong predictions, we anticipate three causes of failure. First, the screen itself could have been noisy, thus impacting predictability. For example, of two identical RNAi screens for longevity, one is highly predictable[22] and the other is only moderately so[23], suggesting differences in screens rather than inherent phenotype predictability. Second, the phenotype may be too broad; for example, viable post-embryonic phenotype is a weakly predicted, broad category covering several phenotypes, many of which are strongly predictable when taken in isolation (clear, small, patchy, body morphology defect, protruding vulva). Finally, the network may be still incomplete and lacking informative data about certain biological pathways—this might be the case for poor predictions of genes that regulate fat content, which may require data more directly relevant to metabolism. As this informatic approach is easily extendable, such datasets can be incorporated in future versions.

Wormnet's predictive power is evident in the case of genes regulating lifespan. Three independent RNAi screens have been done for genes affecting lifespan, allowing us to assess validation rates of predictions made using data from an individual screen. A previous study reported 29 genes that extend lifespan when inhibited[22]. Of the 50 and 200 new genes most highly connected to these in Wormnet (**Supplementary Methods**), 10 (20%) and 21 (10.5%) are validated in the independent RNAi screens[23,24]. These hit rates represent an 8.3- to 100-fold and a 4.4- to 52.5-fold enrichment compared to the hit rates of the three genomic screens. Wormnet-predicted genes include some of the most potent effectors of longevity identified, such as *egl-45*, shown to increase lifespan by 55% following RNAi[24]. Thus, Wormnet facilitates the identification of genes required even for complex phenotypes, such as the regulation of lifespan.

Taken together, the accurately predicted phenotypes in **Figure 3** and **Supplementary Figure 1** represent diverse cellular, developmental and physiological processes. In total, we have shown that we can predict phenotypes for genes specifically affecting the following individual tissues: oocyte, single cell embryo, single lineage in embryo (P lineage), cuticle and hypodermal cells, vulva, neurons, gut, germline and excretory cell. These results demonstrate that a single gene network can predict effects of gene perturbation for diverse aspects of animal biology, and that it is not essential to construct a specialized subnetwork for each particular process.

Network guided screening: the retinoblastoma pathway

We show above that genes sharing loss-of-function phenotypes are tightly linked in the network. Thus, given previously studied genes known to share loss-of-function phenotypes, one can query the network and identify other genes that are tightly linked to these known genes and that are thus likely to have the same loss-of-function phenotype. This approach, which we term 'network guided screening', should be particularly useful for complex systems such as humans, where *in vivo* genome-wide screens are not possible. To illustrate this approach, we focused on two disease relevant pathways that are highly conserved between worm and human and for which there is limited prior knowledge.

We first used Wormnet to identify genes that function in the retinoblastoma tumor suppressor pathway. In *C. elegans*, the retinoblastoma–synMuv B pathway acts genetically redundantly with the synMuv A pathway to repress development of the hermaphrodite vulva[25]. The *C. elegans* vulva consists of 22 cells that derive entirely from three vulval precursor cells (VPCs). However, in worms defective in both the synMuv A and synMuv B pathways, an additional three cells can be induced to form ectopic vulvae (the synthetic multivulva or 'synMuv' phenotype; **Fig. 4a**). Vulva development is therefore controlled by gene interactions in a small number of cells and presents a good example to test the capability of a proteome-scale animal network to predict precise developmental functions for genes.

Recently, six genes were identified that encode suppressors of the synMuv pathway[26,27]; thus, we used these six genes to predict additional synMuv suppressor genes. The six genes interacted with 62 and 142 genes in the core and non-core of Wormnet, respectively, and we used RNAi to test whether targeting these candidate genes suppressed

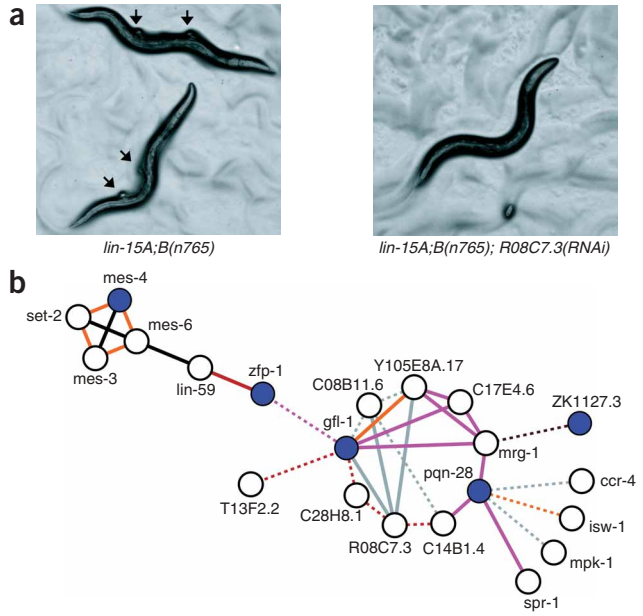

Figure 4 Identification of genes that antagonize the retinoblastoma-synMuv pathway. (**a**) Inactivation of both the synMuv A and synMuv B pathways in the strain *lin-15A;B(n765)* results in the induction of ectopic vulvae (synMuv phenotype, arrows). RNAi against genes that antagonize this pathway (such as the gene *R08C7.3*) can suppress the induction of these ectopic vulvae. (**b**) RNAi against 10 of 50 core and 6 of 124 non-core interaction partners (clear nodes) of the five known suppressors of the synMuv pathway (blue nodes) could also suppress the synMuv phenotype of *lin-15A;B(n765)*. Interactions between these genes in the core of Wormnet are shown as solid edges, and non-core interactions are shown as dashed edges. The dataset used to predict each interaction is indicated by the color of each edge, using the same color-scheme as in **Figure 1a**.

the multivulva phenotype of a *lin-15A;B* strain. We found that RNAi against 10 of 50 tested core interactors (20%) and 6 of 124 non-core interactors (5%) appreciably suppressed this phenotype (**Fig. 4a,b**). In contrast, in a large-scale screen, inactivation of 0.9% (17 of 1,748) of genes suppressed the *lin-15A;B* synMuv phenotype (data not shown). Predictions arising from network connections are thus 21-fold and fivefold better than those expected by chance in the core and non-core networks, respectively. Given the false-negative rate of RNAi, this validation rate is consistent with the true-positive rate calculated in **Figure 1** using functional annotations. The genes identified in the large-scale screen that were not predicted by Wormnet using the six previously known genes are still highly clustered in Wormnet (area under ROC curve, AUC = 0.845), but are not connected to the seed genes. Most of these genes encode proteins involved in a separate pathway that can affect the retinoblastoma-synMuv phenotype (the EGF pathway).

The interactions that predicted the newly identified retinoblastoma-synMuv pathway suppressors derived from diverse datasets from multiple organisms (**Fig. 4b**): only 3 of these 16 interactions were supported by known physical and/or genetic interactions in *C. elegans* (**Table 1**), highlighting the power of integrating data across species. Further, only two of these interactions were predicted in previously published *C. elegans* networks[9–12], underlining the strength of this more comprehensive network. Finally, we compared our predicted results to those of a recent genome-wide RNAi screen for synMuv suppressors[28]. This study reported the identity of 26 suppressors, and we correctly predicted 27% of these. Again, the genes that were not predicted by Wormnet are clustered in the network (AUC = 0.737), but they are not directly attached to the six previously known suppressors. Moreover, 56% of the suppressors that we newly identified via prediction were not reported in the genome-wide screen. Thus, although both network-based prediction and direct genome-wide screening show false negatives, the signal from prediction (16 positives from 174 genes screened) is far stronger than that from unbiased screening, dramatically reducing candidates for experimental validation. This is likely to be a key benefit in animals where genome-wide screens *in vivo* are not currently feasible.

We note that most of the genes that we newly identified as suppressors of the retinoblastoma-synMuv pathway have human orthologs (**Table 1**), and it is possible that inactivation of these orthologs may revert the effects of losing this key tumor suppressor, thereby defining potential targets for pharmaceutical intervention or prevention of human retinoblastoma-linked tumors.

The dystrophin complex modulates EGF-Ras-MAPK signaling

To further illustrate the predictive power of Wormnet, we tested the functional significance of a predicted connection between the dystrophin associated protein complex (DAPC), components of which

Table 1 Suppression of the synMuv phenotype by RNAi against each of the newly identified genes antagonizing the retinoblastoma-synMuv pathway

Gene	Human ortholog(s)	Suppression of *lin-15(n765)* (%)	*n*	Evidence
Y105E8A.17	*DMAP1*	88	81	SC, HS, DM
mrg-1	*MORF4L1/L2*	97	128	HS, WI
mes-3	-	22	150	CC
mes-6	*?EED*	59	69	CC, GT
set-2	*Q9UPS6, SET1A*	5	83	CC, GT
spr-1	*RCOR1/2/3*	13	40	HS
C14B1.4	*WDR5/5B*	39	193	HS
C17E4.6	*VPS72*	13	246	HS
lin-59	*?ASH1L*	48	87	CX
R08C7.3	*ENSG00000177907, SUB1*	96	107	SC
mpk-1	*MAPK1/3*	29	21	SC
T13F2.2	*H2AFZ/H2AFV, ENSG00000187025*	16	80	CX
C28H8.1	*BCL7A/B/C*	3	192	CX
isw-1	*SMARCA1/5*	38	130	CX, SC
C08B11.6	*ACTR6*	69	159	SC
ccr-4	*CNOT6/6L*	4	204	SC
Control		0	531	

Core interactions are shaded. Human orthologs were identified using INPARANOID[16], except for the genes *mes-6* and *lin-59*, which do not have clear human orthologs and where the best BLAST hits as defined in WormBase are indicated with a question mark (?). Evidence codes are as described in **Figure 1**.

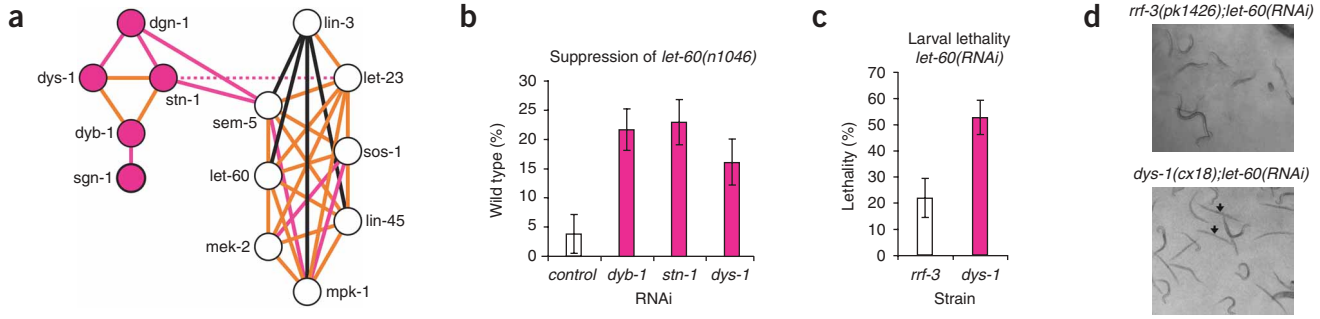

Figure 5 Verifying a functional interaction between the dystrophin associated protein complex (DAPC) and EGF-Ras-MAPK signaling. (**a**) Network representation of the predicted interactions in Wormnet between the DAPC (pink nodes) and the EGF-Ras-MAPK pathway (clear nodes). Edges are colored according to evidence codes as in **Figure 4**. (**b**) Inactivation of DAPC components by RNAi can suppress the induction of ectopic vulvae by a gain-of-function *Ras/let-60* gene. The percentage of wild-type animals (those without a multivulva phenotype) is shown for each treatment. Control RNAi experiments used a bacterial feeding clone that does not target an expressed *C. elegans* gene (Ahringer library clone Y95B8A_84.g). (**c**) Mutations in the *dys-1* gene enhance the larval lethal phenotype of *let-60(RNAi)*. RNAi against *let-60* results in larval lethality in 22% of progeny in the RNAi-hypersensitive strain *rrf-3(pk1426)*, but in 53% of progeny of the *dys-1(cx18)* strain. Larvae die with a characteristic rod-like phenotype due to absence of a functional excretory cell[31]. (**d**) Representative photos of the data presented in **c** illustrating the strongly enhanced penetrance of the 'rod-like' larval lethality (arrows) after RNAi against *let-60* in the *dys-1* strain. Error bars in **b** and **c**, s.d.

are defective in muscular dystrophies, and the EGF signaling pathway, which is frequently defective in cancers.

Inactivation of DAPC genes in *C. elegans* (that is, the genes *dys-1* (dystrophin), *dyb-1* (dystrobrevin), *dgn-1* (dystroglycan), *stn-1* (syntrophin) and *sgn-1* (sarcoglycan)) results in a mild hyperactivity, probably from abnormal neuromuscular junction function[29]. We found that two DAPC genes were directly connected by three interactions to components of the EGF-Ras-MAPK pathway (**Fig. 5a**); these interactions predict a possible functional interaction between the DAPC and the EGF-Ras-MAPK pathway. None of the DAPC genes have any predicted interactions in previous *C. elegans* networks[9–12].

In *C. elegans*, the best characterized function of EGF signaling is as an inductive signal during vulval development[30]; therefore, we tested whether inactivating DAPC components modulated EGF signaling in the vulva. Inactivation of all three tested DAPC genes (*dys-1, dyb-1* and *stn-1*) strongly suppressed induction of ectopic vulvae by an activated *Ras/let-60* gene (**Fig. 5b**), suggesting that the DAPC positively regulates EGF signaling during vulva induction. Consistent with this result, *dys-1* mutations enhanced the 'rod-like' larval lethal phenotype resulting from inhibition of the *Ras/let-60* signal required to specify development of the excretory cell[31] (**Fig. 5c**,**d**). Hence we conclude that in *C. elegans* the DAPC positively regulates EGF-Ras-MAPK signaling both during vulva induction in larvae and excretory cell development in embryogenesis.

DAPC regulation of EGF signaling in *C. elegans* suggests that misregulation of EGF-Ras-MAPK signaling may be important for muscular dystrophy pathology. In support of this, Ras activity in rat muscle preparations is reduced after depleting β-dystroglycan, suggesting the DAPC-Ras functional coupling also exists in vertebrate muscle[32]. Therefore, the EGF-Ras-MAPK pathway may represent a therapeutic target in muscular dystrophy.

We have thus shown that we can use a single network covering most predicted *C. elegans* genes to predict genes affecting differentiation of three vulval precursor cells (for the synMuv suppressor genes), or of the single excretory cell (for the DAPC-EGF connection). Furthermore, we required only six query genes for the SynMuv suppressors and five for the DAPC complex. This is similar in scale to the amount of knowledge of many human congenital diseases.

DISCUSSION

Although construction of single, organism-wide, probabilistic gene networks from existing data has been successful for unicellular organisms, it was unclear that a single network could accurately predict gene function in an animal. To test this, we used existing data to generate a single network including most *C. elegans* genes. We demonstrated that the network predicts diverse cellular, developmental and physiological processes with great specificity. We then applied this network to newly identify genes interacting with key pathways related to human disease and to successfully predict functional connections between pathways. Thus a single gene network covering the majority of genes can make accurate, specific predictions about effects of perturbing individual genes in varied developmental contexts, tissues and genetic backgrounds.

We note that the current network is still incomplete: it covers >80% of genes, but it is certainly missing many functional linkages between them. Additional datasets will help, such as higher coverage physical interaction maps in the worm[10]. Transcriptional analyses of individual tissues and mutant strains should also increase predictive accuracy for specific tissues. We intend to release new versions of Wormnet as such data become available.

Several considerations should be noted regarding the possibility that a similar approach may be successful in humans. First, although few of the network datasets derive from measurements made in specific tissues, we can nonetheless predict effects of gene perturbation on individual tissues. Second, the types and scales of *C. elegans* data are already available for humans. The only major worm datasets not available for humans are genome-scale RNAi screens, which directly measure effects of gene perturbation. However, we omitted these from Wormnet, reserving these data for validating the network's predictive capacity.

Finally, gene networks such as Wormnet identify genes likely to share functions, but they do not identify these functions directly nor indicate positive or negative regulation of a process, as activators and repressors frequently cluster together. We cannot predict a priori the phenotypic effect of gene perturbation without knowing results for other genes. We therefore need a query set of genes with known functions or loss-of-function phenotypes to identify other candidate genes. For example, for any given human genetic disease, we would need to know some genes associated with that disease before

predicting new candidates. For both the retinoblastoma and the DAPC genes, we validated predictions made using less than six query genes, similar to the number known to be mutated in many diseases suggesting that this is a realistic prospect.

This resource is accessible to the community through a website that allows any researcher with genes of interest to identify interacting genes, rank-ordered by association strength to the query list (see URLs section in Methods). Just as BLAST[33] allows researchers to identify genes related by sequence to genes of interest, each scored for confidence, so probabilistic networks allow researchers to identify genes related functionally to genes of interest. Neither BLAST nor gene networks offer guarantees—the top BLAST hits are not necessarily functional orthologs of a gene, and the closest genes in a network may not function in the same process. They are, however, the most likely candidates based on extensive functional data.

In summary, we have demonstrated that it is possible to construct a single network that both covers the majority of genes in *C. elegans* and that predicts which genes are important for the majority of systematically tested phenotypes in this animal. A similar network for humans might therefore facilitate the identification of genes important for diverse processes and diseases.

METHODS

Network construction and evaluation. Computational methods are described in full in the **Supplementary Methods**. We analyzed several datasets. In July 2005, we downloaded expression data from the Stanford Microarray Database[34], identified sets of DNA microarray experiments that showed significant correlation between the extent of mRNA co-expression and functional associations between genes, and selected 6 sets encompassing 220 DNA microarray experiments[35–39] and 635 additional array experiments previously published[11]. Within each microarray set, we filtered out genes out that failed to show significant (typically >1.2-fold) expression changes in some minimal number of experiments, optimizing this value for each expression dataset by recall-precision analysis. Gene pairs were then analyzed for co-expression as previously described[2].

We also downloaded genome-wide yeast two-hybrid interactions between *C. elegans* proteins[10] and the associated literature-derived protein-protein interactions from the Worm Interactome database[10]. Interactions of different confidence (literature, scaffold, core1, core2 and noncore) were benchmarked separately. Genetic interactions (for $\sim$800 genes and $\sim$4,000 interactions) were included from WormBase[13], derived from $>1{,}000$ primary publications. Human protein interactions were collected from existing literature-derived databases[40–44] as well as large-scale yeast two-hybrid analysis[45], then transferred by orthology to *C. elegans* via orthologs defined using INPARANOID[16]. Additional protein interactions were derived in this manner from fly yeast two-hybrid interactions[14], as well as by transferring functional interactions from the yeast functional gene network[15]. Interactions were assigned confidence scores before integration (**Supplementary Methods**).

We derived comparative genomics linkages from the analysis of 133 genomes (117 bacteria and 16 archaea) using the methods of phylogenetic profiling (with modifications in the **Supplementary Methods**) and gene neighbors, calculated as previously described[46,1].

We derived linkages from co-citation of *C. elegans* gene names (perfect matches to either the systematic names or common names) in a set of 7,732 abstracts, downloaded on December 2004, containing the word 'elegans'. Gene pairs were scored as previously described[2].

We used two primary reference pathway sets to evaluate and integrate datasets. The *C. elegans* Gene Ontology (GO) annotation from WormBase[13] (downloaded March 2005) served as the principal reference set for training and benchmarking the network. As gold-standard positive functional linkages, we selected genes sharing GO "biological process" annotation terms from levels 2 through 10 of the GO hierarchy. We excluded five terms annotating an excessive number of genes in order to reduce functional bias in the benchmark set (**Supplementary Methods**): embryonic development, positive regulation of growth rate, growth, locomotory behavior and regulation of transcription (DNA dependent). As gold-standard negative linkages, we selected pairs of genes from this set that did not share annotation terms. A second set of functional annotation was derived from KEGG database[47] annotations (downloaded November 2005; we excluded three annotations accounting for $>40\%$ of the linkages: oxidative phosphorylation, purine metabolism and ribosome). In all, there were 786,056 gene pairs sharing annotation from the GO reference set (available in the **Supplementary Data File**), and 9,406 from KEGG, with 5,069 pairs shared between the two reference sets. An additional test set (KEGG minus GO) was created by removing all GO pairs from the KEGG set. Other tests of network accuracy are described in the **Supplementary Methods**.

We carried out integration and benchmarking of datasets essentially as previously described[2], measuring the linkages derived from each dataset for their enrichment for gold standard positive gene pairs relative to negative gene pairs using a log likelihood criterion, evaluated as a function of the dataset-specific confidence scores associated with the linkages (for example, correlation coefficient between mRNA expression vectors or mutual information between phylogenetic profiles). To obtain accurate estimates of dataset accuracy, we employed 0.632 bootstrapping[48] for all log likelihood evaluations. Linkages from each dataset were then combined using a modified naïve Bayes integration, accounting for dataset correlation using an empirically optimized weighting parameter[15]. A summary of all linkages in the full Wormnet v1.0 network is available in the **Supplementary Data File** online. Previous work on analogous networks in yeast[2] suggests that a cut-off of 1.5 as a likelihood score, corresponding to a minimum $\sim$60% accuracy for individual linkages, provides a high quality network that makes accurate predictions about gene function while maintaining as high coverage as possible. We therefore used this empirically determined cut-off to define the 'core' network of higher confidence interactions. Additional tests of network clustering, topology and functional coherence are described in the **Supplementary Methods**.

Predicting gene perturbation phenotypes. For the prediction of essential genes, we obtained RNAi phenotype profiles from WormBase (April 2005). These included profiles for 15,824 worm genes ($\sim$80% of worm proteome). Of these, 2,349 genes ($\sim$15%) show some phenotypic change during worm development following RNAi knockdown. We used the RNAi phenotype profiles and measures of network connectivity for the 15,824 worm genes for analysis of essential genes. Genes conferring Emb (embryonic lethal), Ste (sterile), Lvl (larval lethal), Stp (sterile progeny), or Adl (adult lethal) phenotypes were considered essential. For analysis of essential mouse genes, we transferred linkages between *C. elegans* genes to pairs of mouse genes according to orthology derived using INPARANOID[16]. We considered only the 6,924 mouse proteins that exist in the murinized Wormnet. Here, 'essential' is defined as genes whose disruption causes embryonic or perinatal lethality in mice[49].

For the prediction of phenotypes from large-scale RNAi screens, we collected results of 43 genome-wide RNAi phenotypic screens from 24 publications, listed in the **Supplementary Methods**. For each phenotype, positive examples are defined as genes known to give rise to the phenotype (the 'seed set'), and negative examples are defined as all other genes screened in that assay. We evaluated the prediction of each phenotype using a leave-one-out approach (that is, each gene in the seed set was omitted from the set for the purposes of evaluating it): each gene in the proteome was rank ordered by the sum of its log likelihood scores to the seed set of genes known to show that phenotype, with high-scoring genes most likely to share the phenotype. We then measured the recovery of genes with the given phenotype, calculating true-positive rate (sensitivity, TP/(TP+FN)) and false-positive rate (1–specificity, FP/(FP+TN)) as a function of rank, evaluating performance using receiver operating characteristic (ROC) analysis. For these analyses, we employed the full Wormnet v1 network, spanning 82% of the proteome, so that we might maximally recover genes with each phenotype. The probabilistic scoring of the linkages simplifies using the entire network—lower confidence linkages still often provide legitimate functional information but are appropriately weighted by confidence when calculating the sum of log likelihood scores to the seed set.

RNAi feeding experiments. *Suppression of synMuv phenotype.* RNAi feeding experiments were done in 12-well agar plates using a previously described protocol[19]. Each selected bacterial clone from the Ahringer RNAi feeding library[19] was grown overnight in 2TY with 100 μg/ml ampicillin and spotted onto every well of a 12-well plate (one complete plate per RNAi feeding clone).

Mixed stage 3 and 4 (L3/4) larval MT8189 *lin-15A;B(n765)* worms were obtained by bleaching adults and incubating at 20 °C for 48 h on OP50 seeded plates. We added approximately ten L3/4 worms to each well of the first column ('well 0') of each 12-well RNAi feeding plate, and the plates incubated at 16 °C. After three days a single adult worm from each well 0 was transferred into each of the other wells of the RNAi feeding plates and allowed to lay eggs for 24 h at 20 °C before being removed. Progeny were scored for the presence of multiple vulvae after a further 3 to 4 days at 20 °C. The **Supplementary Methods** list tested genes.

Suppression of let-60(n1046) multivulva phenotype. We carried out RNAi feeding experiments exactly as described for the SynMuv suppressor assay, except that we grew MT2124 *let-60(n1046)* worms continuously at 20 °C and transferred adults after two days. All experiments were done in triplicate.

let-60(n1046) synthetic larval lethality. To assay the larval lethal phenotype of RNAi targeting *let-60*, we carried out RNAi feeding experiments in liquid culture as previously described[50] adding ~10 L1 stage worms to 40 µl of bacterial RNAi feeding clones and counting the number of viable and 'rod-like' larval lethal progeny after 4 d incubation at 20 °C. We used the RNAi hypersensitive strain NL2099 *rrf-3(pk1426)* and LS292 *dys-1(cx18)*. Four repeats were done per strain.

URLs. Wormnet, http://www.functionalnet.org/wormnet.

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

ACKNOWLEDGMENTS

Worm strains used in this work were provided by the *Caenorhabditis* Genetics Center, which is funded by the US National Institutes of Health National Center for Research Resources (NCRR). SAGE data were obtained from the Genome BC *C. elegans* Gene Expression Consortium and were produced at the Michael Smith Genome Sciences Centre with funding from Genome Canada. This work was supported by grants from the N.S.F. (IIS-0325116, EIA-0219061), N.I.H. (GM06779-01), Welch (F1515), and a Packard Fellowship (E.M.M.). B.L. was supported by a Sanger Institute Postdoctoral Fellowship and the EMBL-CRG Systems Biology Program, and the Institució Catalana de Recerca i Estudis Avançats (ICREA). A.G.F., W.W. and C.C. are supported by the Wellcome Trust. We thank A. Enright for help with constructing the web interface for analyzing this data.

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

1. Bowers, P.M. *et al.* Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* **5**, R35 (2004).
2. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
3. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. USA* **100**, 8348–8353 (2003).
4. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
5. Rhodes, D.R. *et al.* Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.* **23**, 951–959 (2005).
6. Chen, Y. & Xu, D. Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **32**, 6414–6424 (2004).
7. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
8. Famili, I., Forster, J., Nielsen, J. & Palsson, B.O. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. USA* **100**, 13134–13139 (2003).
9. Gunsalus, K.C. *et al.* Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 861–865 (2005).
10. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
11. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
12. Zhong, W. & Sternberg, P.W. Genome-wide prediction of *C. elegans* genetic interactions. *Science* **311**, 1481–1484 (2006).
13. Chen, N. *et al.* WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res.* **33**, D383–D389 (2005).
14. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
15. Lee, I., Li, Z. & Marcotte, E.M. An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS ONE* **2**, e988 (2007).
16. Remm, M., Storm, C.E. & Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
17. Gandhi, T.K. *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* **38**, 285–293 (2006).
18. Deplancke, B. *et al.* A gene-centered *C. elegans* protein-DNA interaction network. *Cell* **125**, 1193–1205 (2006).
19. Kamath, R.S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
20. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
21. Walker, M.G., Volkmuth, W. & Klingler, T.M. Pharmaceutical target discovery using Guilt-by-Association: schizophrenia and Parkinson's disease genes. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1999**, 282–286 (1999).
22. Hansen, M., Hsu, A.L., Dillin, A. & Kenyon, C. New genes tied to endocrine, metabolic, and dietary regulation of lifespan from a *Caenorhabditis elegans* genomic RNAi screen. *PLoS Genet.* **1**, 119–128 (2005).
23. Hamilton, B. *et al.* A systematic RNAi screen for longevity genes in *C. elegans*. *Genes Dev.* **19**, 1544–1555 (2005).
24. Curran, S.P. & Ruvkun, G. Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS Genet.* **3**, e56 (2007).
25. Lu, X. & Horvitz, H.R. lin-35 and lin-53, two genes that antagonize a *C. elegans* Ras pathway, encode proteins similar to Rb and its binding protein RbAp48. *Cell* **95**, 981–991 (1998).
26. Lehner, B. *et al.* Loss of LIN-35, the *Caenorhabditis elegans* ortholog of the tumor suppressor p105Rb, results in enhanced RNA interference. *Genome Biol.* **7**, R4 (2006).
27. Wang, D. *et al.* Somatic misexpression of germline P granules and enhanced RNA interference in retinoblastoma pathway mutants. *Nature* **436**, 593–597 (2005).
28. Cui, M., Kim, E.B. & Han, M. Diverse chromatin remodeling genes antagonize the Rb-involved SynMuv pathways in *C. elegans*. *PLoS Genet.* **2**, e74 (2006).
29. Segalat, L. Dystrophin and functionally related proteins in the nematode *Caenorhabditis elegans*. *Neuromuscul. Disord.* **12**Suppl 1, S105–S109 (2002).
30. Sundaram, M.V. RTK/Ras/MAP kinase signaling. *WormBook* Vol. 1.80.1 <http://www.wormbook.org>(2005).
31. Yochem, J., Sundaram, M. & Han, M. Ras is required for a limited number of cell fates and not for general proliferation in *Caenorhabditis elegans*. *Mol. Cell. Biol.* **17**, 2716–2722 (1997).
32. Chockalingam, P.S. *et al.* Dystrophin-glycoprotein complex and Ras and Rho GTPase signaling are altered in muscle atrophy. *Am. J. Physiol. Cell Physiol.* **283**, C500–C511 (2002).
33. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
34. Gollub, J. *et al.* The Stanford microarray database: data access and quality assessment tools. *Nucleic Acids Res.* **31**, 94–96 (2003).
35. Jiang, M. *et al.* Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **98**, 218–223 (2001).
36. Lund, J. *et al.* Transcriptional profile of aging in *C. elegans*. *Curr. Biol.* **12**, 1566–1573 (2002).
37. Reinke, V. *et al.* A global profile of germline gene expression in *C. elegans*. *Mol. Cell* **6**, 605–616 (2000).
38. Romagnolo, B. *et al.* Downstream targets of let-60 Ras in *Caenorhabditis elegans*. *Dev. Biol.* **247**, 127–136 (2002).
39. Wang, J. & Kim, S.K. Global analysis of dauer gene expression in *Caenorhabditis elegans*. *Development* **130**, 1621–1634 (2003).
40. Alfarano, C. *et al.* The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.* **33**, D418–D424 (2005).
41. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432 (2005).
42. Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371 (2003).
43. Ramani, A.K., Bunescu, R.C., Mooney, R.J. & Marcotte, E.M. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* **6**, R40 (2005).
44. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006).
45. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
46. Date, S.V. & Marcotte, E.M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21**, 1055–1062 (2003).
47. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
48. Efron, B. & Tibshirani, R. *An introduction to the bootstrap,* **439** (Chapman & Hall, New York, 1993).
49. Eppig, J.T. *et al.* The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res.* **33**, D471–D475 (2005).
50. Lehner, B., Crombie, C., Tischler, J., Fortunato, A. & Fraser, A.G. Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat. Genet.* **38**, 896–903 (2006).