LETTER

doi:10.1038/nature11779

Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence

Yuval Tabach[1,2], Allison C. Billi[3,4]*, Gabriel D. Hayes[1,2]*, Martin A. Newman[1,2], Or Zuk[5], Harrison Gabel[1,2], Ravi Kamath[1,2], Keren Yacoby[1], Brad Chapman[1], Susana M. Garcia[1,2], Mark Borowsky[1,2], John K. Kim[3,4] & Gary Ruvkun[1,2]

Genetic and biochemical analyses of RNA interference (RNAi) and microRNA (miRNA) pathways have revealed proteins such as Argonaute and Dicer as essential cofactors that process and present small RNAs to their targets. Well-validated small RNA pathway cofactors such as these show distinctive patterns of conservation or divergence in particular animal, plant, fungal and protist species. We compared 86 divergent eukaryotic genome sequences to discern sets of proteins that show similar phylogenetic profiles with known small RNA cofactors. A large set of additional candidate small RNA cofactors have emerged from functional genomic screens for defects in miRNA- or short interfering RNA (siRNA)-mediated repression in *Caenorhabditis elegans* and *Drosophila melanogaster*[1,2], and from proteomic analyses of proteins co-purifying with validated small RNA pathway proteins[3,4]. The phylogenetic profiles of many of these candidate small RNA pathway proteins are similar to those of known small RNA cofactor proteins. We used a Bayesian approach to integrate the phylogenetic profile analysis with predictions from diverse transcriptional coregulation and proteome interaction data sets to assign a probability for each protein for a role in a small RNA pathway. Testing high-confidence candidates from this analysis for defects in RNAi silencing, we found that about one-half of the predicted small RNA cofactors are required for RNAi silencing. Many of the newly identified small RNA pathway proteins are orthologues of proteins implicated in RNA splicing. In support of a deep connection between the mechanism of RNA splicing and small-RNA-mediated gene silencing, the presence of the Argonaute proteins and other small RNA components in the many species analysed strongly correlates with the number of introns in those species.

Proteins with similar patterns of conservation or divergence across different organisms are more likely to act in the same pathways[5]. To identify proteins that share an evolutionary history with validated small RNA pathway proteins, we determined the phylogenetic profiles of approximately 20,000 *C. elegans* proteins in 85 genomes, representing diverse taxa of the eukaryotic tree of life: 33 animals, 6 land plants, 1 alga, 31 Ascomycota fungi, 3 Basidiomycota fungi and 12 protists. Of the ~20,000 *C. elegans* proteins, 10,054 show homologues in non-nematode eukaryotic genomes (Supplementary Table 1). Following correlation and clustering, this analysis sorts genes into clades of conservation and relative divergence or loss in the various organisms as suites of genes are maintained from common ancestors or diverge in particular lineages[6]. Protein divergence or loss in particular taxonomic clades is not random; entire suites of proteins can diverge or be lost as particular taxa specialize and no longer require ancestral functions. The correlated loss of proteins has been used to assign roles for nuclear-encoded mitochondrial proteins[7] and eukaryotic cilia-associated proteins[8].

We developed a non-binary method of phylogenetic profiling to cluster all protein sequences encoded by *C. elegans* genes. BLAST scores were normalized to the length of the query sequence and for relative phylogenetic distance between *C. elegans* and the queried organism[9]. The matrix of 864,644 conservation scores for the 10,054 *C. elegans* proteins in the 86 genomes was queried either with a single protein to generate a ranking of other *C. elegans* proteins with the most similar pattern of conservation values or using a more global hierarchical clustering method (Fig. 1a). Proteins of the same families exhibit similar patterns of phylogenetic conservation and therefore tend to group together in the hierarchical clustering. However, many phylogenetic clusters include proteins with no sequence similarity; only their conservation or divergence in genomes is correlated. The ability of this non-binary method of phylogenetic profiling to cluster proteins based on function is exemplified by the clustering of proteins known to act as members of complexes. For example, the known protein components of the sensory cilium have highly correlated phylogenetic profiles characterized by loss in particular vertebrates and all fungi and plants and retention in particular protists, whereas the extraordinarily high and universal conservation of ribosomal and translation factor proteins clusters many of these translation components (Supplementary Fig. 1a, b).

With a simple query of one of the central proteins in RNAi, the Argonaute RDE-1, we generated a rank-ordered list of proteins with phylogenetic profiles most similar to that of RDE-1 (Fig. 1b). The 26 other *C. elegans* Argonautes represent the top correlated proteins, a trivial consequence of protein sequence similarity within the Argonaute family. The signature phylogenetic profile of the Argonaute proteins is that they are absent in 9 out of 31 Ascomycota species, 1 out of 3 Basidiomycota species, and 6 out of 14 protist species, but have not been lost in any of the 33 animal or 6 land plant species compared. The retention of Argonaute proteins correlates with the ability to inactivate genes by RNAi[10], and the loss of RNAi in about one-half of the sequenced Ascomycota fungi is correlated with the 'killer' RNA virus[11]. Additional *C. elegans* proteins that cluster with the Argonautes but show no sequence similarity include an asparaginase encoded by *K01G5.9*, the CAND-1 elongation factor and another elongation factor, the THO complex protein THOC-1. THO complex members have emerged from genetic screens for defective transgene and RNAi silencing in *Arabidopsis thaliana*[12].

Another validated *C. elegans* RNAi protein is MUT-2, a polyA polymerase implicated in a step downstream of the production of primary siRNAs by Dicer[13]. Out of the 50 *C. elegans* proteins with phylogenetic profiles most closely correlated with MUT-2 (Supplementary Fig. 1c), 10 are Argonautes, which bear no sequence similarity to MUT-2, demonstrating the efficacy of this approach to detect validated small

[1]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Life Sciences Institute, University of Michigan, Ann Arbor, Michigan 48109, USA. [4]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA. [5]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA.
*These authors contributed equally to this work.

©2013 Macmillan Publishers Limited. All rights reserved

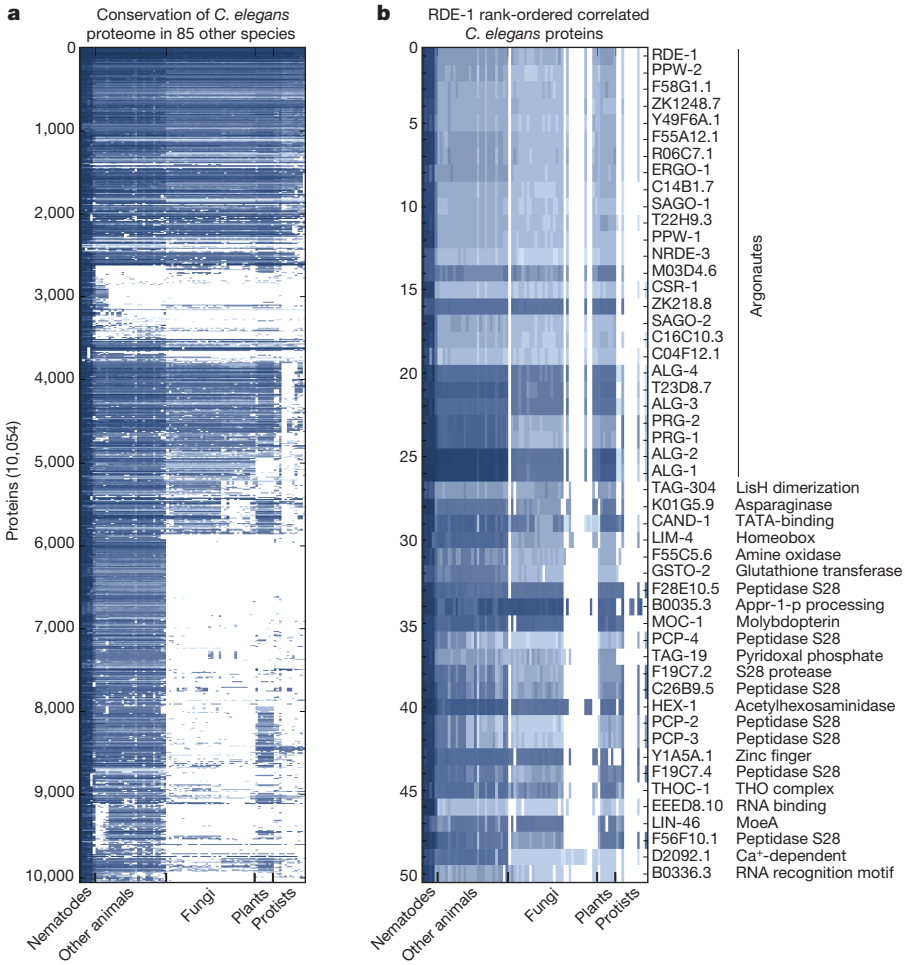

Figure 1 | Phylogenetic profiling analysis shows correlated conservation patterns of *C. elegans* proteins. a, Phylogenetic profiles of 10,054 conserved *C. elegans* proteins across 85 other eukaryotic genomes. For each *C. elegans* query protein, the normalized ratio of the BLAST score for the top-scoring protein sequence similarity is indicated in the column corresponding to each genome. Values range from 0 (white, no similarity) to 1 (blue, 100% similarity). **b**, Phylogenetic profiles of validated RNAi factor RDE-1 and the 49 most correlated proteins in rank order.

RNA pathway proteins. The splicing components MAG-1, RSP-8, RNP-4, RSP-5 and DDB-1 and the translation factors EIF-3.D and EIF-3.E, many of which score in the validation tests below, also have similar phylogenetic profiles. In addition, out of the proteins most correlated with the *C. elegans* orthologue of Dicer (DCR-1), a nuclease that processes siRNAs and miRNAs, 3 Argonaute proteins emerge among the top 50 correlated phylogenetic profiles (Supplementary Fig. 1d and Supplementary Table 2).

The RNA-dependent RNA polymerases[14], siRNA-amplifying cofactors, are present in only 5 out of 27 animals (all the nematode species and, surprisingly, the tick), in all of the land plants, in 2 out of 4 Basidiomycota fungi, in 18 out of 27 Ascomycota fungi and in 4 out of 14 protists, but are not present in green algae. A query of the RNA-dependent RNA polymerase RRF-3 (Supplementary Fig. 1e) revealed the cofactor-independent phosphoglycerate mutase F57B10.3 as a dramatically correlated non-homologous protein ($R = 0.93$). Inactivation of this phosphoglycerate mutase gene causes defects in the endogenous siRNA response as well as transgene silencing, validating its role in RNA silencing (Supplementary Table 2). It is possible that either the biochemical substrate or product of this glycolysis pathway protein, or its enzymatic activity as a phosphatase, couples it to small RNA pathways.

To identify candidate small RNA pathway proteins more comprehensively, we globally ranked proteins based on phylogenetic-profile correlation with multiple validated siRNA and miRNA cofactors. After assigning all conserved *C. elegans* proteins to hierarchical clusters, we gave each protein a score to reflect its phylogenetic clustering with the validated set of small RNA proteins (Supplementary Fig. 2). This analysis identified 60 proteins not previously implicated in small RNA pathways whose phylogenetic profiles correlate highly with those of validated siRNA and miRNA pathway proteins (Fig. 2).

The validated siRNA and miRNA protein cofactors identified so far probably constitute a small fraction of the total number of proteins that mediate small RNA function. Full-genome RNAi screens for defects in siRNA or miRNA pathway function have identified hundreds of additional candidate small RNA pathway proteins. We integrated ten genome-scale studies into the phylogenetic cluster analysis: five *C. elegans* gene-inactivation screens for defects in RNAi or miRNA function[1,15,16], *C. elegans* orthologues of *Drosophila* genes identified in two full-genome RNAi screens for impaired siRNA or miRNA response[2] and three proteomic studies of complexes containing the known RNAi proteins DCR-1 (ref. 4), ERI-1 (ref. 17) and AIN-2 (ref. 18). Candidate genes identified in these studies show little overlap (Supplementary Table 3 and Supplementary Fig. 3a, b). However, the candidates from the different studies have similar phylogenetic profiles to each other and to validated small RNA cofactors (Fig. 3, Supplementary Fig. 3c, d and Supplementary Table 4).

We used a naive Bayesian classifier to assign predictive values to six genome-scale studies of RNAi cofactors and five miRNA cofactors (see Supplementary Methods)[19,20]. To the phylogenetic profiles, we added a score for each *C. elegans* gene that is co-expressed on microarrays[21] or whose encoded gene product interacts with validated small RNA pathway proteins[22]. The top 105 genes identified by this analysis are

©2013 Macmillan Publishers Limited. All rights reserved

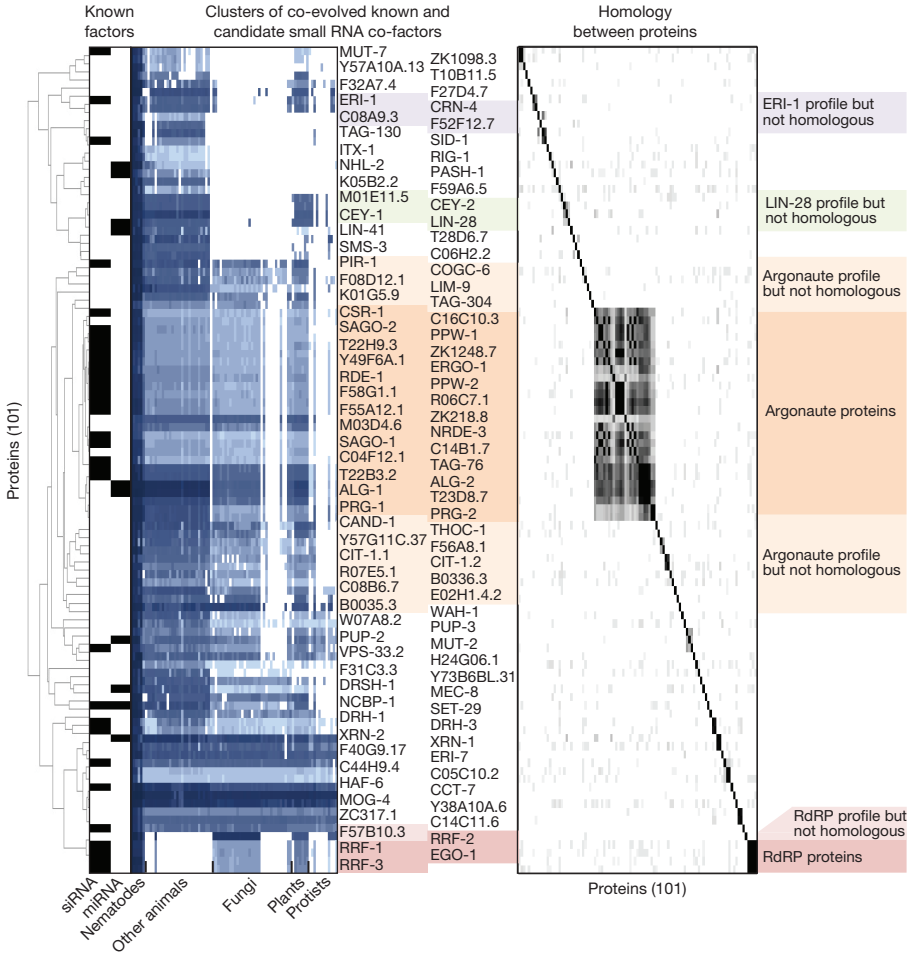

Figure 2 | Phylogenetic clusters of candidate small RNA pathway proteins. Validated miRNA and siRNA pathway proteins map non-randomly on the phylogenetic profile; proteins that map to the same clusters are likely to function in small RNA pathways. Left panel, clusters enriched for validated miRNA and siRNA pathway proteins (black boxes). Darker blue, higher protein-sequence similarity. Right panel, pairwise local protein-sequence alignment of all pairs of proteins in the cluster. White, no similarity; black, significant similarity.

enriched with 41 well-validated siRNA pathway genes (Supplementary Fig. 7 and Supplementary Table 2). The other genes on this list are excellent candidates to mediate siRNA or related small RNA functions. More than 20 of these genes encode RNA recognition motifs including RNP ($P < 0.00001$) and helicase ($P < 0.00001$), an approximately 20-fold enrichment relative to the entire data set. Nine proteins from this list constitute components of the spliceosome (Supplementary Fig. 3).

From the proteins best correlated with validated small RNA pathway cofactors by phylogenetic profile or in the naive Bayesian analysis (Figs 1–3), we tested 87 representative candidates using two different tests for defects in RNAi. Transgene silencing in the somatic cells of the enhanced RNAi mutant *eri-1*(*mg366*) is mediated by an RNAi mechanism[1]. We tested a set of 87 predicted small RNA pathway genes using this strain, and 43 scored as significantly RNAi-defective (Supplementary Table 2, and Fig. 4a). We also tested candidates using a green fluorescent protein (GFP)-based sensor for the abundant *C. elegans* endogenous siRNA 22G siR-1 (ref. 23) to monitor whether any of the gene inactivations affect the production or response to this endogenous siRNA. Thirty-three out of 87 genes tested scored in this assay (Supplementary Table 2 and Fig. 4b). Eight of the nine predicted splicing components scored strongly in these validation screens.

The enrichment for RNA splicing components (Supplementary Fig. 4) points to a close mechanistic connection between splicing and small RNA regulation. Among the Ascomycota and protist species that have lost the Argonaute proteins, most show an extreme loss of introns, from 10^4–10^5 introns in species with Argonautes to 10^2 or fewer introns in most species without Argonautes (Supplementary Fig. 5). We screened for defects in RNAi a cherry-picked gene inactivation sublibrary of *C. elegans* orthologues of known splicing factors that have emerged from biochemical and genetic screens for splicing components from other systems. From a set of 46 *C. elegans* genes annotated in KEGG (Kyoto Encyclopedia of Genes and Genomes) to encode the orthologues of known splicing proteins that could be tested for roles in RNAi in our assays, 16 and 22 of these splicing-factor genes scored strongly in the *eri-1* transgene desilencing assay and the endogenous 22G siR-1 sensor assay. Many of the splicing components that scored strongly in these screens show a phylogenetic profile similar to the Argonaute proteins (Supplementary Fig. 6 and Supplementary Table 6). However, a subset of splicing factors that are well conserved across phylogeny also scored strongly in these assays.

We used the *eri-1* transgene desilencing system to conduct a full-genome screen for gene inactivations that disable transgene silencing and identified 855 genes required for transgene silencing, with more than 200 scoring above 3 on a scale of 0 to 4 for desilencing (Supplementary Table 7). Among gene inactivations that caused the greatest desilencing, 11% correspond to the highest ranked predictions from the siRNA naive Bayesian analysis, a 30-fold enrichment ($P = 4.7 \times 10^{-13}$ using a hypergeometric test) for positives. Out of the 84 splicing factors that have been assigned to specific splicing steps, 49 scored in the full genome screen as required for transgene silencing, and 32 showed phylogenetic profiles clustering with known small RNA factors. The splicing factors that couple to small RNA pathways were not isolated to any particular step of RNA splicing. Splicing factor mutations in *Schizosaccharomyces pombe* disrupt the RNAi-based

©2013 Macmillan Publishers Limited. All rights reserved

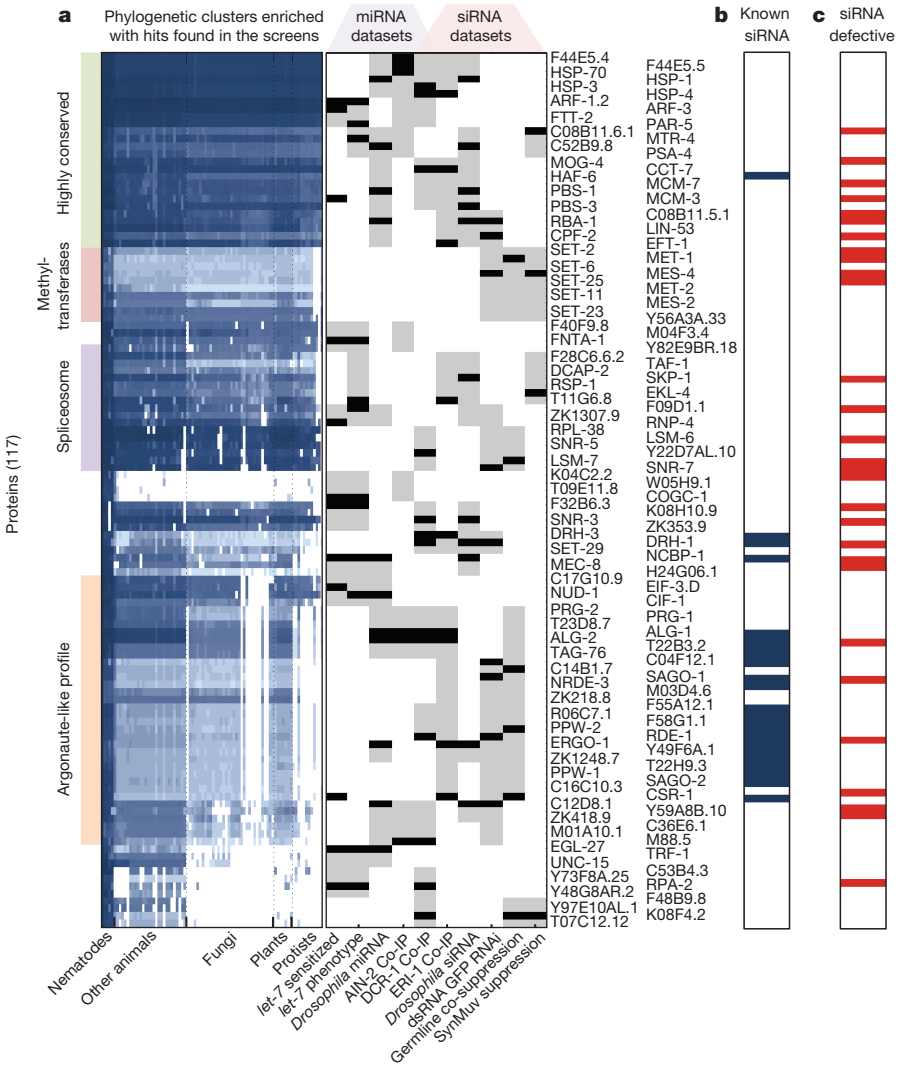

Figure 3 | Select phylogenetic clusters enriched with hits from proteomic and functional genomic small RNA screens. **a**, The phylogenetic profile matrix was clustered and a Max Ratio score (MRS) was calculated for every protein in each screen; 117 proteins scored significantly in miRNA (56 proteins) or siRNA (75 proteins) functional genomic screens, or both (14 proteins). Middle panel, black tick, hit in screens; grey tick, significant MRS. **b**, Blue boxes, the 23 known small RNA pathway proteins identified. **c**, From the 117 proteins predicted by the phylogenetic profile, 28 proteins (red boxes) show defects in siRNA silencing ($P < 3 \times 10^{15}$).

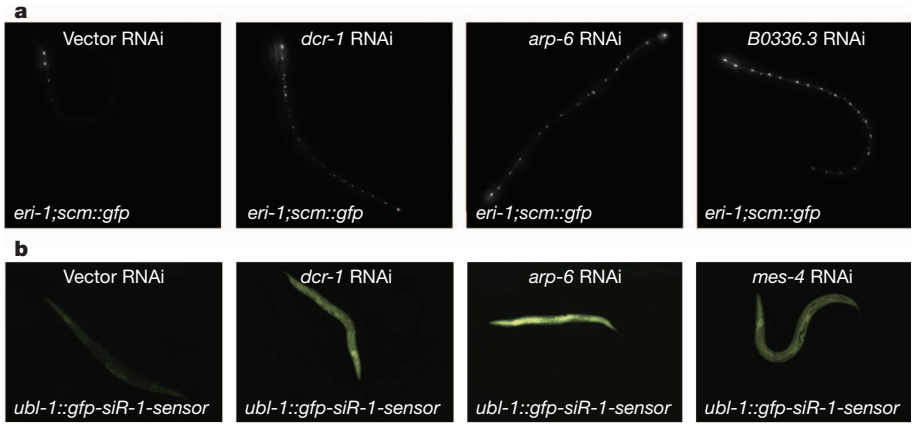

Figure 4 | Inactivation of genes implicated in RNAi pathways re-animates transgenes that are silenced by RNAi. **a**, Expression of *scm::gfp* in the seam cells of an *eri-1(mg366)* mutant, where it is normally silenced by RNAi. Animals shown were treated with control, *dcr-1*, *arp-6* or *B0336.3* RNAi. **b**, GFP expression from the *ubl-1::gfp-siR-1-sensor* transgene, which is normally silenced by the siR-1 endogenous siRNA. Animals shown were treated with control, *dcr-1*, *arp-6* or *mes-4* RNAi.

©2013 Macmillan Publishers Limited. All rights reserved

centromeric silencing[24]. Both splicing proteins and siRNA and miRNA pathway proteins co-localize to cytoplasmic processing bodies (P-bodies) and nuclear Cajal bodies[25], further supporting the possibility of functional crosstalk between splicing and RNAi.

Early genome sequence comparisons of *S. pombe*, *Saccharomyces cerevisiae* and a small set of eukaryotes suggested that loss of introns and splicing components is highly correlated with loss of Argonaute proteins[26]. One interpretation was that the loss of RNAi in *S. cerevisiae* enabled viral invasion and a subsequent loss of introns through reverse transcription of genes by the invading viral replication enzymes. However, such a scenario would not predict that inactivation of splicing components in a species bearing the RNAi apparatus would cause an RNAi-defective phenotype. One model is that splicing could regulate RNAi indirectly by modulating spliced isoforms of key RNAi factors. However, the observations that only a subset of splicing cofactors are required for RNAi and the co-immunoprecipitation of splicing factors and DCR-1, ERI-1 and AIN-2 disfavours this indirect model. A mechanistic coupling between RNAi and RNA splicing explains these new data better. RNAi factors also affect splicing: Dicer is required for efficient spliceosomal RNA maturation in *Candida albicans*[27]. If RNAi engages introns intimately by, for example, engaging nascent transcripts through the Argonaute NRDE-3 before splicing[28], then the selective advantage of introns may fade once the RNAi pathway is lost.

Our data suggest that a large subset of the proteins that mediate steps in the maturation of mRNAs bearing introns are also required for RNAi, and that those genomes that have lost most of their introns no longer require the RNAi pathway. Superimposed on the mRNA splicing pathway is an RNA surveillance system that eliminates aberrantly processed or mutant pre-mRNAs and mRNAs. It is possible that RNAi constitutes another level of mRNA surveillance that acts in parallel to—and using many of the same components as—the splicing quality control surveillance pathways.

METHODS SUMMARY

Informatics. The Normalized Phylogenetic Profile (NPP) data matrix was clustered through MATLAB statistical toolbox using the average linkage method and Pearson correlation coefficient as a similarity measure. Clustering was performed on the rows of the matrix. To identify *C. elegans* proteins with phylogenetic profiles similar to published small RNA co-factors (Supplementary Table 9), the fraction of the validated proteins in each phylogenetic cluster was calculated and optimized to define a Max Ratio Score (MRS) (Supplementary Fig. 2).

Received 16 April; accepted 8 November 2012.
Published online 23 December 2012.

1. Kim, J. K. *et al.* Functional genomic analysis of RNA interference in *C. elegans*. *Science* **308,** 1164–1167 (2005).
2. Zhou, R. *et al.* Comparative analysis of argonaute-dependent small RNA pathways in *Drosophila*. *Mol. Cell* **32,** 592–599 (2008).
3. Meister, G. *et al.* Identification of novel argonaute-associated proteins. *Curr. Biol.* **15,** 2149–2155 (2005).
4. Duchaine, T. F. *et al.* Functional proteomics reveals the biochemical niche of *C. elegans* DCR-1 in multiple small-RNA-mediated pathways. *Cell* **124,** 343–354 (2006).
5. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96,** 4285–4288 (1999).
6. Gabaldón, T. Evolution of proteins and proteomes: a phylogenetics approach. *Evol. Bioinform. Online* **1,** 51–61 (2005).
7. Pagliarini, D. J. *et al.* A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134,** 112–123 (2008).
8. Avidor-Reiss, T. *et al.* Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell* **117,** 527–539 (2004).
9. Enault, F., Suhre, K., Abergel, C., Poirot, O. & Claverie, J. M. Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics* **19** (Suppl. 1), i105–i107 (2003).
10. Drinnenberg, I. A. *et al.* RNAi in budding yeast. *Science* **326,** 544–550 (2009).
11. Drinnenberg, I. A., Fink, G. R. & Bartel, D. P. Compatibility with killer explains the rise of RNAi-deficient fungi. *Science* **333,** 1592 (2011).
12. Yelina, N. E. *et al.* Putative *Arabidopsis* THO/TREX mRNA export complex is involved in transgene and endogenous siRNA biosynthesis. *Proc. Natl Acad. Sci. USA* **107,** 13948–13953 (2010).
13. Ketting, R. F. & Plasterk, R. H. A genetic link between co-suppression and RNA interference in *C. elegans*. *Nature* **404,** 296–298 (2000).
14. Simmer, F. *et al.* Loss of the putative RNA-directed RNA polymerase RRF-3 makes *C. elegans* hypersensitive to RNAi. *Curr. Biol.* **12,** 1317–1319 (2002).
15. 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).
16. Parry, D. H., Xu, J. & Ruvkun, G. A whole-genome RNAi Screen for *C. elegans* miRNA pathway genes. *Curr. Biol.* **17,** 2013–2022 (2007).
17. Thivierge, C. *et al.* Tudor domain ERI-5 tethers an RNA-dependent RNA polymerase to DCR-1 to potentiate endo-RNAi. *Nature Struct. Mol. Biol.* **19,** 90–97 (2012).
18. Zhang, L. *et al.* Systematic identification of *C. elegans* miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol. Cell* **28,** 598–613 (2007).
19. Calvo, S. *et al.* Systematic identification of human mitochondrial disease genes through integrative genomics. *Nature Genet.* **38,** 576–582 (2006).
20. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302,** 449–453 (2003).
21. Hibbs, M. A. *et al.* Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* **23,** 2692–2699 (2007).
22. Simonis, N. *et al.* Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nature Methods* **6,** 47–54 (2009).
23. Montgomery, T. A. *et al.* PIWI associated siRNAs and piRNAs specifically require the *Caenorhabditis elegans* HEN1 ortholog henn-1. *PLoS Genet.* **8,** e1002616 (2012).
24. Bayne, E. H. *et al.* Splicing factors facilitate RNAi-directed silencing in fission yeast. *Science* **322,** 602–606 (2008).
25. Pontes, O. & Pikaard, C. S. siRNA and miRNA processing: new functions for Cajal bodies. *Curr. Opin. Genet. Dev.* **18,** 197–203 (2008).
26. Aravind, L., Watanabe, H., Lipman, D. J. & Koonin, E. V. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl Acad. Sci. USA* **97,** 11319–11324 (2000).
27. Bernstein, D. A. *et al. Candida albicans* Dicer (CaDcr1) is required for efficient ribosomal and spliceosomal RNA maturation. *Proc. Natl Acad. Sci. USA* **109,** 523–528 (2012).
28. Guang, S. *et al.* An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science* **321,** 537–541 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Duchaine for access to his ERI-1 proteomic data before it was published and to S. Fischer, C. Zhang and T. Montgomery for helpful discussions. The work was supported by NIH GM088565 and the Pew Charitable Trusts (J.K.K.) and NIH GM44619 and GM098647 (G.R.).

Author Contributions Y.T., J.K.K. and G.R. designed experiments; Y.T. developed analytical tools and analysed data; and Y.T., A.C.B., G.D.H., M.A.N., S.M.G., H.G., R.K. and J.K.K. designed and carried out experiments. O.Z. gave statistical support and conceptual advice. Y.T., K.Y, B.C. and M.B. wrote code. Y.T., A.C.B., J.K.K. and G.R. wrote the paper. G.R. and J.K.K. supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.R. (ruvkun@molbio.mgh.harvard.edu).

©2013 Macmillan Publishers Limited. All rights reserved