Selecting genes for analysis using historically contingent progress: from RNA changes to protein-protein interactions

Farhaan Lalit[1], Antony M Jose[1]*

Affiliations:

[1]University of Maryland, College Park, MD, USA.

*Corresponding author. Email: amjose@umd.edu

Author Contributions: A. M. J. designed the study; F. L. and A. M. J. performed the analyses (F. L. - all work on data tables and A. M. J. - all work on AlphaFold-based analyses and GO analyses of clusters); and F. L. and A. M. J. wrote the paper.

Competing Interest Statement: The authors declare no competing interests.

Keywords: Mutual Information, AlphaFold, RNA silencing, homeostasis, *C. elegans*.

SUMMARY

Progress in biology has generated numerous lists of genes that share some property. But, advancing from these lists of genes to understanding their roles is slow and unsystematic. Here we use RNA silencing in *C. elegans* to illustrate an approach for prioritizing genes for detailed study given limited resources. The partially subjective relationships between genes forged by both deduced functional relatedness and biased progress in the field was captured as mutual information and used to cluster genes that were frequently identified yet remain understudied. Studied genes in these clusters suggest regulatory links connecting RNA silencing with other processes like the cell cycle. Many proteins encoded by the understudied genes are predicted to physically interact with known regulators of RNA silencing. These predicted influencers of RNA-regulated expression could be used for feedback regulation, which is essential for the homeostasis observed in all living systems. Thus, among the gene products altered when a process is perturbed are regulators of that process, providing a way to use RNA sequencing to identify candidate protein-protein interactions. Together, the analysis of perturbed transcripts and potential interactions of the proteins they encode could help prioritize candidate regulators of any process.

MAIN TEXT

Introduction

Genes and gene products are often collected as lists based on unifying characteristics or based on experiments. For example, genes that show enrichment of a chromatin modification, mRNAs that change abundance in response to a mutation, proteins that interact with another protein, etc. After the initial identification of a set of genes as belonging to a list, multiple approaches [1] are needed to generate an explanatory model. However, many genes do not receive further attention, as evidenced by recent meta-analyses, which highlighted numerous understudied genes in humans [2, 3]. Since single papers often analyze only one or a few genes, a wider view of genes with roles in a process could be gained by comparing lists generated by several studies. Such exploration could identify genes that are present in multiple lists but have not yet been selected for detailed study. Identifying these understudied genes is especially useful during the early stages of a field, when coherent models for most observed phenomena have not yet emerged. While this approach is also extensible to lists of anything that is used to characterize living systems (changes in lipids, metabolites, localizations, etc.), here we focus on lists of mRNAs, proteins, and small RNAs generated by the field of RNA silencing in the nematode *C. elegans*.

A gene present in many lists could be regulated in multiple separable ways and/or be regulated in one or a few ways by connected sets of regulators (Fig. 1*A*). For example, mRNA levels could be regulated through changes in transcription, turnover, localization, small RNA production, etc. or all changes could occur because of turnover regulation by a connected set of regulators. Changes in such genes could alter specific regulatory outputs, making them integrators of inputs from many other regulators. Alternatively, they could have no measurable consequence, but might still be experimentally useful as general sensors of perturbation. One way that organisms could use such general sensing of perturbation in a process could be to return the process to the pre-perturbation state through feedback [4]. Such active resetting would enable restoration of homeostasis faster than through the dissipation of the perturbation alone.

Here we present an approach to identify these regulated but understudied genes in the field of RNA silencing in *C. elegans*. We find that many of these genes encode **p**redicted **i**nfluencers of **R**NA-regulated **e**xpression (PIRE) that can directly interact with key regulators of RNA silencing. The impacts of PIRE proteins on the activity of the regulators need to be evaluated using experiments that selectively perturb the interactions and we expect the other understudied regulated genes also include those with a role in RNA silencing.

Results

Many genes have been repeatedly reported within data tables but remain understudied.

To determine if there are any understudied regulated genes that are relevant for RNA silencing in *C. elegans*, we examined data from past studies in the field. While complete replication of each study might be needed for direct comparisons, this goal is impractical. Even beginning with the 'raw' data deposited to public resources (e.g., fastq files after RNA-seq) and repeating the analyses reported in a paper is not always feasible. Summary tables from previous analyses presented in papers provide a practical intermediate level of data to use for comparisons across studies. Therefore, we collated a total of 432 tables from 112 papers for comparison (see methods and Table S1 for list of studies) and joined the tables together after standardizing gene names to yield genes that can be compared for presence or absence across 432 lists (Fig. 1*B*). To identify a set of genes (g) that receive extensive regulatory input and/or that encode proteins that interact with many other proteins and are yet selectively regulated, we propose a metric r_g (Fig. 1*B*). Since the likelihood of including a gene from the lists increases with g, the metric is specified with a subscript for each analysis (e.g., r_{25} refers to a regulation score when the top 25 genes that are most commonly present in lists are considered) and defined to be:

$$r_g \coloneqq \sum_{i=1}^{n} \frac{S_i}{T_i}$$

where g = size of gene set chosen for analysis, n = total number of lists with altered genes, S_i = number of genes from the ith list that is also present in the gene set g, and T_i = total number of genes in the ith list. The larger the set of genes (g) selected, the greater the chance of a dataset (with T_i genes) having at least one overlapping gene within the selected gene set (probability given by $P(S_i > 0)$ in Fig. 1*C*). The metric r_g is a decision aid that helps with choosing genes for experimental analysis and is not to be taken as an objective measure of the importance of the gene for the biological process under study.

The top 25 genes with the highest r_{25} values included the germline Argonaute proteins CSR-1 [5] and HRDE-1 [6], which have each been the subject of numerous studies (Fig. 1*D*). While most other genes are understudied (fewer than 10 publications on WormBase), among them is W09B7.2/*sdg-1*, which was recently reported to be regulated by the double-stranded RNA (dsRNA) importer SID-1 and encodes a protein with a suggested role in feedback regulation of heritable RNA silencing by colocalizing with perinuclear germ granules [7]. This discovery suggests that the analysis of the additional genes with high r_{25} values could also be fruitful. Of the 16 understudied genes that encode proteins, seven had predicted structures of high confidence (i.e., domains with pLDDT > 90) in the AlphaFold Protein Structure Database [8]. These structures were then used to identify related protein domains using Foldseek [9] (Fig. 1*E*; E-value < 0.05). These include domains with well-defined biochemical activities like de-ubiquitinase (E01G4.5), aspartic protease (K02E2.6), RNAse H1 (RNH-1.3), F-box B (FBXB-97), decapping nuclease (Y47H10A.5), and Histone H1-like (HIL-4) domains. Three more proteins have been proposed to be nucleocapsid-like proteins encoded from genes within retrotransposons ([7, 10]; Fig. 1*E*). These candidates can be experimentally analyzed in the future for possible roles in RNA silencing.

To explore the relationships between these genes with the highest r_{25} values (Fig. 1*F* and 1*G*), we clustered the genes and generated a dendrogram where genes present together in different lists are closer together (see supplementary methods). The dendrogram revealed the gene *hil-4*, which encodes a Histone H1-like protein [11], as the understudied gene clustering closest to *hrde-1* and *csr-1*, making it a strong candidate for a potential role downstream of RNA-mediated gene regulation. Another cluster (brown in Fig. 1*G*) included all four pseudogenes,

suggesting that this method could capture functional relatedness despite the limitations and biases introduced by the available data.

Multiple proteins encoded by the top r_{25} genes are predicted to interact with known regulators of RNA silencing.

Animals typically recover from silencing initiated by dsRNA within the germline [12] or in somatic cells [13]. This recovery occurs despite the presence of amplification mechanisms, suggesting that silencing ends either when the trigger dsRNA runs out and/or that it is under homeostatic control through feedback inhibition. Evidence for such self-limiting behavior in RNA silencing include the recruitment of an inhibitor of RNA silencing to genes targeted by dsRNA [14] and a feedback loop that limits the production of some endogenous small RNAs [15]. The top understudied regulated genes identified here could encode proteins with such homeostatic roles. Therefore, to test if any of the proteins encoded by genes with the highest r_{25} values could interact with known regulators of RNA silencing, we examined the potential for protein-protein interactions using their predicted structures.

We selected 25 known regulators of RNA silencing (see Fig. 2A) chosen for their roles in different phases of the deduced mechanism(s) of RNA silencing [16-18]. These include proteins with roles in the processing of dsRNA and its regulators; Argonaute proteins and their regulators; proteins with roles in secondary small RNA production and its regulators; components of germ granules; and co-transcriptional regulators (Fig. 2*A*). We then examined their predicted interactions with 16 proteins encoded by understudied regulated genes that have the highest r_{25} values (highlighted in red, Fig. 1G). For 20 RNA regulators, we used AlphaFold 2, which makes extensive use of multiple sequence alignments for computing inter-protein interactions and has a success rate of ~50-60% [19, 20]. Since the computational cost of AlphaFold 2 escalates with the number of amino acids, interactions with the remaining 5 larger regulators (DCR-1, EGO-1, ZNFX-1, NRDE-2, and MET-2) were tested on the recently available but proprietary AlphaFold 3 server [21], which can predict interactions with ligands, and as with AlphaFold 2, uses multiple sequence alignments for its structure predictions. To stratify the predicted interactions, we initially considered the maximal inter-protein predicted aligned error (PAE) and the distance between the interacting residues (distance), which was allowed to be up to twice the length of hydrogen bonds (~3 Å [22]). Examining interactions that satisfy three progressively more stringent criteria (PAE less than 30 and distance less than 2, PAE less than 20 and distance less than 6, and PAE less than 5 and distance less than 6) revealed many interactions with substantial surface area (size of circle in Fig. S1). Even the most stringent criterion (PAE less than 5, which is more stringent than the 8 Å error that has been used successfully [23]) revealed numerous interactions (blue in Fig. S1). Therefore, to constrain the predictions further, we used the ranking scores, which are a combination of interface-predicted template modeling (ipTM) and predicted template modeling scores (pTM): 0.8*ipTM + 0.2*pTM for AlphaFold 2.3 [24] and 0.8*ipTM + 0.2*pTM + 0.5*disorder for AlphaFold 3 [21]. We only considered interactions with a ranking score greater than 0.6, which is relatively high given than ipTM scores as low as ~0.3 can yield true positives [25]. Together, these criteria identified 35 interactions (Fig. 2*B* and Movies S1 to S35). Among the regulators, RDE-3, RDE-8, and SET-25 had the highest numbers of predicted interactors (5 proteins each) and among the proteins encoded by understudied genes, FBXB-97 had the highest number of predicted interactors (7 proteins).

Given the prevalence of predicted interactions and the need to prioritize for follow-up work, we restrict our comments here to interactions that constrain a minimum of 20 residues in the proteins encoded by understudied genes (Fig. S2), which we define as predicted interactions of high confidence. These high-confidence interactors included proteins that were predicted to interact with every phase of the deduced mechanism(s) for RNA silencing (Fig. 2*C*). Since the precise numbers of interacting residues required for a meaningful interaction in vivo is variable

and unknown, interactions that constrain fewer residues could have measurable impacts on function. Nevertheless, we conservatively designate each protein that is predicted to interact with one or more RNA regulators with relatively high confidence as a **P**redicted **I**nfluencer of **R**NA-regulated **E**xpression (PIRE). We name five of these as PIRE-1 through PIRE-5 (C08F11.7, E01G4.5, F15D4.5, K02E2.6, and Y47H10A.5, respectively; Fig. 2*D*) and preserve the names of the three that were already given names based on structural homology (subunit of the **T**ranslocase of the **I**nner **M**itochondrial **M**embrane TIMM-17B.2, the **RN**ase **H** protein RNH-1.3, and the **F**-**b**o**x B** protein FBXB-97) or after detailed study (the **S**ID-1-**d**ependent **g**ene protein W09B7.2/SDG-1). For convenience, these nine putative interactors are collectively referred as PIRE proteins here.

Each of these interactions (Movie S1 to S35) suggest hypotheses for their functional impact based on the known roles of RNA regulators (Table S2) and the domains present in PIRE proteins (Fig. 1*E*). The two PIRE proteins encoded by genes within retrotransposon (PIRE-3 and SDG-1) that also interact with regulators of RNA silencing, supports the idea that retrotransposon-encoded genes influence their own RNA-mediated regulation (e.g., [7]). FBXB-97, which is predicted to be an F-box protein [26], could promote ubiquitin-mediated degradation of its interactors (RDE-4, ERI-1, NRDE-3, DEPS-1, PID-2, RDE-8, and RDE-3) or sequester them, preventing their activity. PIRE-4, which is predicted to be a protease, could cleave its interactors (ERI-1, PID-2, RDE-8, and SET-25) to regulate their activity – a mode of regulation that has been recently elucidated for Argonaute proteins [27] and implicated in RNA silencing within the germline [28]. Additional PIRE proteins with confidently predicted domains (e.g., RNase H in RNH-1.3, SPK domain in PIRE-1, decapping nuclease in PIRE-5, and multiple domains in PIRE-2) potentially implicate new biochemical activities in the process of RNA silencing. In all, two general modes of interaction that are not mutually exclusive could be discerned between PIRE proteins and the tested regulators of RNA silencing (Fig. 3). In one mode exemplified by FBXB-97 (Fig. 3, left), the interactions with most regulators involve nearly the same set of residues. In the other mode exemplified by PIRE-3 (Fig. 3, right), interactions with different regulators involve different sets of residues. Finally, it is possible that protein interactions predicted to have smaller interfaces are nevertheless present in vivo. For example, the predicted interactions of C38D9.2 and PIRE-3 with PRG-1 (Fig. S1) constrain fewer than 20 amino acids (Figs. S2 and S3). Yet, both proteins were pulled down along with PRG-1 in an immunoprecipitation experiment [29].

In summary, predictions using AlphaFold identify numerous interactions that inspire follow-up work to test hypotheses about the roles of PIRE proteins in RNA silencing.

Predictions by AlphaFold 2 and AlphaFold 3 do not always agree.

While AlphaFold 2 predicted all the interactions classified as high-confidence interactions, the one interaction predicted by AlphaFold 3 (EGO-1 and W09B7.1) with a maximal PAE <5 and distance <6 constrained fewer than 20 residues (Fig. 2*B*, Fig. S2, and Fig. S3). The reason for this extreme discrepancy is unclear.

To directly compare both approaches for predicting protein-protein interactions, we examined some of the interactions predicted by each approach using the other. We first examined interactions predicted with a high ranking score according to AlphaFold 2 (> 0.8, Fig. 4*A*). Of these, only the interaction between RNH-1.3 and RDE-3 was confidently predicted by AlphaFold 3, albeit with a lower score (0.85 for AF2 vs 0.68 for AF3). Aligning both predicted complexes using the RDE-3 protein reveals that both predictions are in good agreement (Fig. 4*B*). We next considered two proteins, PIRE-3 and FBXB-97, for which multiple interactors were predicted by AlphaFold 2 with varying confidence. While AlphaFold 2 predicted interactions between PIRE-3 and 3 RNA regulators (ranking = 0.73, 0.70, and 0.61), and between FBXB-97 and 7 RNA regulators (ranking = 0.66, 0.67, 0.75, 0.78, 0.75, 0.77, and 0.76), AlphaFold 3 only predicted an

interaction with RDE-3 for both proteins (Fig. 4*C,* ranking = 0.78 and 0.79). The RDE-3-interacting residues of FBXB-97 predicted by both approaches overlapped but those of PIRE-3 did not (Fig. 4*C*). Furthermore, aligning the predicted protein-protein complexes using RDE-3 showed a large discrepancy in the positions of the interacting partners in both cases (Fig. 4*C*, FBXB-97, *left*; PIRE-3, *right*). Similarly, comparing the predictions for interactions between EGO-1 and W09B7.1 also revealed large discrepancies (Fig. 4*C*). While a region of interaction was predicted using AlphaFold 3 with PAE <5 and distance <6 (Fig. 4*E*, right), regions of interaction were only detectable using AlphaFold 2 when the maximal PAE allowed was increased to 10 (Fig. 4*E*, left). Even at this lower threshold for error, the predicted interacting regions differed between the two approaches (black ovals in Fig. 4*D*).

The reasons for the differences between predictions by AlphaFold 2 and AlphaFold 3 could be varied. For example, differences in sampling of predictions, which is expected to correlate with success rate [30] (25 models used in AlphaFold 2 here versus five on the AlphaFold 3 server) and/or differences in handling intrinsically disordered regions, for which structures can be identified by AlphaFold 2 if they conditionally fold [31]. Modifications to these algorithms that extend capabilities continue to be developed (e.g., modeling of interacting interfaces within intrinsically disordered regions [32], predicting multiple conformations [33], and predicting large protein assemblies [34]). Therefore, further comparisons of multiple algorithms for predicting protein-protein interactions and customized exploration of criteria for interactions [35] may be useful for determining when each algorithm can aid the generation of hypotheses. However, determining if, when, and where any predicted interactions occur in vivo will require many future experiments.

AlphaFold-predicted structures could reveal interactions that are challenging to demonstrate experimentally.

Obtaining experimental support for direct interactions between proteins can be difficult. For example, an interaction between the most abundant Gα protein in the brain (Gα_o [36], GOA-1 in *C. elegans*) and the diacylglycerol kinase DGK-1 is strongly predicted by genetic analysis [37, 38]. Both AlphaFold 2 and AlphaFold 3 predict the same extensive binding between GOA-1 and DGK-1 (Fig. S4). Furthermore, the interaction interface is largely preserved and reliably predicted by AlphaFold 3 when GOA-1 is by itself or bound to either GTP or GDP (Fig. S4). Yet, early attempts using purified proteins failed to reveal a detectable interaction between DGK-1 and GOA-1 in vitro [39], and this interaction has remained a conjecture for more than two decades.

While biochemical approaches rely on preserving or recreating in vitro the unknown conditions in vivo to coax a detectable interaction between proteins, prediction algorithms that incorporate extensive multiple sequence alignments (e.g., AlphaFold 2 and to an unknown extent AlphaFold 3) can use the co-evolution of residues to deduce the interaction. Given these complementary strengths, systematic analyses using both multiple experimental approaches [1] and multiple prediction algorithms are needed to find the edge of predictability for protein-protein interactions.

Clustering based on Historical Mutual Information among genes with the highest r_{100} values reveals a group of genes that link RNA silencing to other processes.

To examine if the observations on understudied yet regulated genes using r_{25} hold when analyzing a larger set of genes, we examined the top 100 genes with the highest r_{100} values. To quantify the correlated presence or absence of genes in different lists we used a measure of mutual information [40] named here as historical mutual information (HMI) to emphasize the subjective nature of this measure because it depends on both functional relatedness of the genes and biased availability of data (see supplementary methods). Using HMI to cluster these genes revealed three major clusters (43, 42, and 11 genes), another cluster with two genes and two

other unconnected genes (Fig. 5*A*, Table S3). Only one cluster (cluster 1 in Fig. 5*A*) had significant numbers of genes associated with gene ontology terms. Many of these genes encode proteins that bind and/or hydrolyze RNA (Fig. 5*B*, *top*), localize to cytoplasmic ribonucleoprotein granules (Fig. 5*B*, *middle*), and/or play roles in other processes such as cell division (Fig. 5*B*, *bottom*). Consistently, this cluster also had the greatest number of genes that have been described in multiple publications (Fig. 5*C*), including all the genes that have been featured in abstracts on RNA silencing (Fig. 5*D*). Therefore, the analysis of additional genes in this cluster could be relevant for RNA silencing and connect such regulation to other processes (e.g., the cell cycle). Since four of the five pseudogenes are in a small cluster (Fig. 5*E*, 4 of 11 genes in cluster 2), the other genes in this cluster could potentially be targets of regulation without specific downstream regulation or be co-regulated sensors of pseudogene RNA levels. Intriguingly, there is a large overlap between a set of genes that require HRDE-1 for downregulation (67 genes in both replicates from worms grown at 15ºC [41]) and genes in a single cluster (Fig. 5*F*, 17 of 42 genes in cluster 3). One possible explanation for this abundance and clustering could be that *hrde-1*-dependent gene lists are among the most numerous generated by the field and/or included in our analysis (44 of 298 lists with fewer than 2000 genes). Alternatively, genes that are subject to HRDE-1-dependent silencing could be extensively regulated by many other regulators and require this additional downregulation for fitness – i.e., overexpression of these genes is detrimental. Consistent with this possibility, loss of HRDE-1 results in progressive sterility that can be reversed by restoring HRDE-1 activity [41]. Also, as expected for the use of HRDE-1 downstream of SID-1, genes upregulated using *sid-1* (18 genes in animals with a deletion in *sid-1* [7]) overlap with genes in the same cluster (Fig. 5*F*, 4 of 42 in cluster 3). Future studies by labs working on multiple aspects of RNA silencing in *C. elegans* have the potential to test and enrich the classification of the regulated yet understudied genes revealed here, including the identification of many more PIRE proteins.

Discussion

Our analysis has identified selectively regulated yet understudied genes in the field of RNA silencing in *C. elegans*, some of which encode predicted influencers of RNA-regulated expression that act through protein-protein interactions.

The inevitable bias of progress. Bias during progress in a field is unavoidable and its causes are complex, including availability of technology, researcher pre-disposition, perceived importance of a direction, current societal need, etc. Therefore, the comprehensive appraisal of a field through equal representation of all important aspects is impractical. Indeed, our analysis involved the manual collation of many datasets for comparison which could have resulted in omissions and inclusions that spark disagreements. While future extensions of this work could automate the process of aggregating and comparing data, flexible inclusion of different lists in the analysis would be needed to enable customization based on the expertise, interests, and risk tolerance of individual labs. Furthermore, earlier studies using older technologies could have led to conclusions that need revision. For example, when analyzed using multi-copy transgenes, the dsRNA-binding protein RDE-4 showed a cell non-autonomous effect [42, 43], but when analyzed using single-copy transgenes, RDE-4 showed a cell autonomous effect [44]. Since different researchers could interpret such conflicting data differently (e.g., differences in levels of tissue-restricted expression versus differences in extent of misexpression in other tissues), it is useful to preserve the ability to customize lists. With the expanding number of lists generated by large-scale experimental approaches in different fields, identifying selectively regulated yet understudied genes could aid the prioritization of genes for detailed mechanistic studies using the limited resources and time available for any lab.

Function(s) of the x-dependent gene. Different properties of a single protein or RNA could be important for different biological roles [45, 46], or the same properties could be important

for different processes. Despite such variety, a gene found in many lists could become associated with a single label because of the historical sequence of discovery (e.g., HRDE-1-dependent genes; many in cluster 3, Fig. 2*F*), thereby obscuring additional roles of that gene. All nine PIRE proteins are predicted to interact with more than one tested regulator of RNA silencing (Figs. 2*B*, S1, S2, and S3). If these interactions are validated through experimental analyses, it will not be possible to classify these PIRE proteins into single pathways. Indeed it can be challenging to delineate pathways when an intersecting network of regulators make quantitative contributions to an observed effect [13]. The well-recognized difficulty in defining the function of a gene [47] is exacerbated in these cases, making it more appropriate to consider these proteins as entities within a system whose roles depend on context.

From transcript changes to protein-protein interactions. Positive feedback loops that drive growth and development are a ubiquitous feature of life [48]. Yet, living systems are also characterized by homeostasis [4], which needs negative feedback to suppress runaway processes. For example, in a chain of biochemical reactions, product inhibition [49] can be used to regulate production to match need. While this organization enables compensation in response to change, complete compensation for all processes is clearly not possible as evidenced by the fact that many mutations have measurable consequences. A specific case of this general principle is transcriptional adaptation, where the mutation-induced degradation of a transcript results in compensatory changes in the levels of other transcripts [50]. The existence of PIRE proteins suggests that another way for organisms to compensate for the perturbation of a protein that regulates a process is to change the levels of other proteins that can regulate the same process through protein-protein interactions. Thus, we speculate that perturbing a protein could sometimes alter the mRNA levels of its interactors because of the prevalence of feedback regulation in living systems. If true, this feature of life provides a strategy for combining RNA sequencing and protein structure predictions to identify protein-protein interactions of regulatory importance by analyzing changes in the transcriptome.

Methods Summary

Data tables from 112 studies on RNA silencing in *C. elegans* that were published between 2007 and 2023 were downloaded (Table S1), reformatted manually and/or using custom scripts, and filtered to generate lists that only include entries with reported p-values or adjusted p-values < 0.05, when such values were available. Gene names were standardized across datasets using tools from Wormbase [51]. The top 'g' genes that occur in the greatest numbers of tables were culled as the most frequently identified genes. A measure for the extent of regulation of each gene (r_g) was used to aid prioritization for detailed study. Co-occurrence patterns of genes in different tables were captured using the Jaccard distance (d_J) [52] or as a symmetric measure of normalized mutual information [40], defined here as Historical Mutual Information (HMI). The d_J values were used to generate a dendrogram using the average linkage method (Fig. 1*G*). HMI was used to group genes into clusters according to the Girvan-Newman algorithm [53] and different sets of genes were highlighted (Fig. 2). Gene ontology (GO) analyses were performed using Gene Ontology Resource (https://geneontology.org/; [54, 55]) and significant GO terms were collected for visualization using REVIGO [56].

Prediction of dimer formation between the 16 proteins encoded by understudied genes among the top r_{25} genes and 25 key regulators of RNA silencing were obtained using AlphaFold 2 [8, 24] run on a high-performance cluster (Zaratan at UMD) and/or using the AlphaFold 3 [21] server online (https://golgi.sandbox.google.com/). Large regulators (DCR-1, EGO-1, ZNFX-1, NRDE-2, and MET-2) were tested on the AlphaFold 3 server initially and positive hits, if any, were examined again using AlphaFold 2 (e.g., EGO-1 interaction with W09B7.1). The computed models were processed using custom shell scripts, python programs, and ChimeraX [57]. Briefly, the highest ranked model for each pair of proteins were depicted with the predicted aligned error used to highlight inter-protein interactions as pseudobonds colored according to the alphafold pae palette on ChimeraX (Movie S1 to S35). For criteria of maxPae (5, 20, or 30) and distance (6 or 2), an approximation of the interaction area was calculated by isolating the mutually constrained residues and using the 'buriedarea' command (ChimeraX). This area was divided by the product of the number of amino acids in each protein to get a normalized value and scaled uniformly before plotting (e.g., Fig. 2*B* for maxPae <5 and distance <6). Finally, the ranking scores (0.8*ipTM + 0.2pTM for AlphaFold 2.3 and 0.8*ipTM + 0.2pTM + 0.5*disorder for AlphaFold 3) were used to shade the circle representing each interaction (Fig. 2*B*).

All scripts used in this study are available at GitHub (AntonyJose-Lab/Lalit_Jose_2024).

Acknowledgements

We thank Tom Kocher, Brian Pierce, and members of the Jose lab for comments on the manuscript. We acknowledge the University of Maryland supercomputing resources (http://hpcc.umd.edu) made available for conducting the research reported in this paper. We thank Brian Pierce for discussions and Rui Yin for installing AlphaFold 2 on the UMD HPCC, and Carlos Retamal and José Feijo for getting us started with AlphaFold. This work is supported in part by National Institutes of Health Grant R01GM124356 and National Science Foundation Grant 2120895 to A.M.J.

References

1. A. M. Jose, The analysis of living systems can generate both knowledge and illusions. *Elife* **9**, e56354 (2020).
2. R. A. K. Richardson, H. T. Navarro, L. A. Nunes Amaral, T. Stoeger, Meta-Research: understudied genes are lost in a leaky pipeline between genome-wide assays and reporting of results. *eLife* **12**, RP93429 (2023).
3. J. J. Rocha *et al.*, Functional unknomics: Systematic screening of conserved genes of unknown function. *PLoS Biol* **21**, e3002222 (2023).
4. G. E. Billman, Homeostasis: The Underappreciated and Far Too Often Ignored Central Organizing Principle of Physiology. *Front Physiol* **11**, 200 (2020).
5. J. M. Claycomb *et al.*, The Argonaute CSR-1 and its 22G-RNA cofactors are required for holocentric chromosome segregation. *Cell* **139**, 123-134 (2009).
6. B. A. Buckley *et al.*, A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature* **489**, 447-451 (2012).
7. N. Shugarts *et al.*, Intergenerational transport of double-stranded RNA limits heritable epigenetic changes. *bioRxiv* 10.1101/2021.10.05.463267, 2021.2010.2005.463267 (2024).
8. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
9. M. van Kempen *et al.*, Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243-246 (2024).
10. S. E. J. Fischer, G. Ruvkun, Caenorhabditis elegans ADAR editing and the ERI-6/7/MOV10 RNAi pathway silence endogenous viral elements and LTR retrotransposons. *Proc Natl Acad Sci U S A* **117**, 5987-5996 (2020).
11. M. A. Jedrusik, E. Schulze, A single histone H1 isoform (H1.1) is essential for chromatin silencing and germline development in Caenorhabditis elegans. *Development* **128**, 1069-1080 (2001).
12. S. Devanapally *et al.*, Mating can initiate stable RNA silencing that overcomes epigenetic recovery. *Nat Commun* **12**, 4239 (2021).
13. D. R. Knudsen-Palmer, P. Raman, F. Ettefa, L. Ravin, A. M. Jose, Target-specific requirements for RNA interference can arise through restricted RNA amplification despite the lack of specialized pathways. *bioRxiv* 10.1101/2023.02.07.527351 (2024).
14. R. Perales *et al.*, Transgenerational Epigenetic Inheritance Is Negatively Regulated by the HERI-1 Chromodomain Protein. *Genetics* **210**, 1287-1299 (2018).
15. A. K. Rogers, C. M. Phillips, A Small-RNA-Mediated Feedback Loop Maintains Proper Levels of 22G-RNAs in C. elegans. *Cell Rep* **33**, 108279 (2020).
16. N. Frolows, A. Ashe, Small RNAs and chromatin in the multigenerational epigenetic landscape of Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci* **376**, 20200112 (2021).
17. A. C. Billi, S. E. Fischer, J. K. Kim, Endogenous RNAi pathways in C. elegans. *WormBook* 10.1895/wormbook.1.170.1, 1-49 (2014).
18. A. E. Sundby, R. I. Molnar, J. M. Claycomb, Connecting the Dots: Linking Caenorhabditis elegans Small RNA Pathways and Germ Granules. *Trends Cell Biol* **31**, 387-401 (2021).

19. R. Yin, B. Y. Feng, A. Varshney, B. G. Pierce, Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci* **31**, e4379 (2022).
20. P. Bryant, G. Pozzati, A. Elofsson, Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* **13**, 1265 (2022).
21. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 10.1038/s41586-024-07487-w (2024).
22. D. Herschlag, M. M. Pinney, Hydrogen Bonds: Simple after All? *Biochemistry* **57**, 3338-3352 (2018).
23. I. R. Humphreys *et al.*, Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).
24. R. Evans *et al.*, Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 10.1101/2021.10.04.463034, 2021.2010.2004.463034 (2022).
25. S. Weeratunga *et al.*, Interrogation and validation of the interactome of neuronal Munc18-interacting Mint proteins with AlphaFold2. *J Biol Chem* **300**, 105541 (2024).
26. J. W. Harper, B. A. Schulman, Cullin-RING Ubiquitin Ligase Regulatory Circuits: A Quarter Century Beyond the F-Box Hypothesis. *Annu Rev Biochem* **90**, 403-429 (2021).
27. R. K. Gudipati *et al.*, Protease-mediated processing of Argonaute proteins controls small RNA association. *Mol Cell* **81**, 2388-2402 (2021).
28. M. Placentino *et al.*, Intrinsically disordered protein PID-2 modulates Z granules and is required for heritable piRNA-induced silencing in the Caenorhabditis elegans embryo. *EMBO J* **40**, e105280 (2021).
29. P. J. Batista *et al.*, PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. *Mol Cell* **31**, 67-78 (2008).
30. B. Wallner, AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics* **39**, btad573 (2023).
31. T. R. Alderson, I. Pritisanac, D. Kolaric, A. M. Moses, J. D. Forman-Kay, Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc Natl Acad Sci U S A* **120**, e2304302120 (2023).
32. H. Bret, J. Gao, D. J. Zea, J. Andreani, R. Guerois, From interaction networks to interfaces, scanning intrinsically disordered regions using AlphaFold2. *Nat Commun* **15**, 597 (2024).
33. H. K. Wayment-Steele *et al.*, Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832-839 (2024).
34. B. Shor, D. Schneidman-Duhovny, CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2. *Nat Methods* **21**, 477-487 (2024).
35. D. Yu, G. Chojnowski, M. Rosenthal, J. Kosinski, AlphaPulldown-a python package for protein-protein interaction screens using AlphaFold-Multimer. *Bioinformatics* **39**, btac749 (2023).
36. S. M. Strittmatter, D. Valenzuela, T. E. Kennedy, E. J. Neer, M. C. Fishman, G0 is a major growth cone protein subject to regulation by GAP-43. *Nature* **344**, 836-841 (1990).
37. Y. M. Hajdu-Cronin, W. J. Chen, G. Patikoglou, M. R. Koelle, P. W. Sternberg, Antagonism between G(o)alpha and G(q)alpha in Caenorhabditis elegans: the RGS protein EAT-16 is necessary for G(o)alpha signaling and regulates G(q)alpha activity. *Genes Dev* **13**, 1780-1793 (1999).

38. S. Nurrish, L. Ségalat, J. M. Kaplan, Serotonin inhibition of synaptic transmission: Gαo decreases the abundance of UNC-13 at release sites. *Neuron* **24**, 231-242 (1999).
39. A. M. Jose, M. R. Koelle, Domains, amino acid residues, and new isoforms of Caenorhabditis elegans diacylglycerol kinase 1 (DGK-1) important for terminating diacylglycerol signaling in vivo. *J Biol Chem* **280**, 2730-2736 (2005).
40. I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, San Francisco, ed. 2nd, 2005).
41. J. Z. Ni *et al.*, A transgenerational role of the germline nuclear RNAi pathway in repressing heat stress-induced transcriptional activation in *C. elegans*. *Epigenetics Chromatin* **9**, 3 (2016).
42. A. M. Jose, G. A. Garcia, C. P. Hunter, Two classes of silencing RNAs move between Caenorhabditis elegans tissues. *Nat Struct Mol Biol* **18**, 1184-1188 (2011).
43. D. Blanchard *et al.*, On the nature of in vivo requirements for rde-4 in RNAi and developmental pathways in C. elegans. *RNA Biol* **8**, 458-467 (2011).
44. P. Raman, S. M. Zaghab, E. C. Traver, A. M. Jose, The double-stranded RNA binding protein RDE-4 can act cell autonomously during feeding RNAi in C. elegans. *Nucleic Acids Res* **45**, 8463-8473 (2017).
45. C. E. Chapple *et al.*, Extreme multifunctional proteins identified from a human protein interaction network. *Nat Commun* **6**, 7412 (2015).
46. D. M. Ribeiro, G. Briere, B. Bely, L. Spinelli, C. Brun, MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins. *Nucleic Acids Res* **47**, D398-D402 (2019).
47. D. M. Keeling, P. Garza, C. M. Nartey, A. R. Carvunis, The meanings of 'function' in biology and the problematic case of de novo gene emergence. *Elife* **8**, e47014 (2019).
48. A. M. Jose, Heritable epigenetic changes are constrained by the dynamics of regulatory architectures. *eLife* **12**, RP92093 (2024).
49. E. Frieden, C. Walter, Prevalence and Significance of the Product Inhibition of Enzymes. *Nature* **198**, 834-837 (1963).
50. T. E. Sztal, D. Y. R. Stainier, Transcriptional adaptation: a mechanism underlying genetic robustness. *Development* **147**, dev186452 (2020).
51. P. Davis *et al.*, WormBase in 2022-data, processes, and tools for analyzing Caenorhabditis elegans. *Genetics* **220**, iyac003 (2022).
52. P. Jaccard, Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société vaudoise des sciences naturelles* **37**, 241-272 (1901).
53. M. Girvan, M. E. Newman, Community structure in social and biological networks. *Proc Natl Acad Sci U S A* **99**, 7821-7826 (2002).
54. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
55. G. O. Consortium *et al.*, The Gene Ontology knowledgebase in 2023. *Genetics* **224** (2023).
56. F. Supek, M. Bosnjak, N. Skunca, T. Smuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
57. E. C. Meng *et al.*, UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci* **32**, e4792 (2023).

58. M. Varadi *et al.*, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, D439-D444 (2022).

Figures and Figure Legends

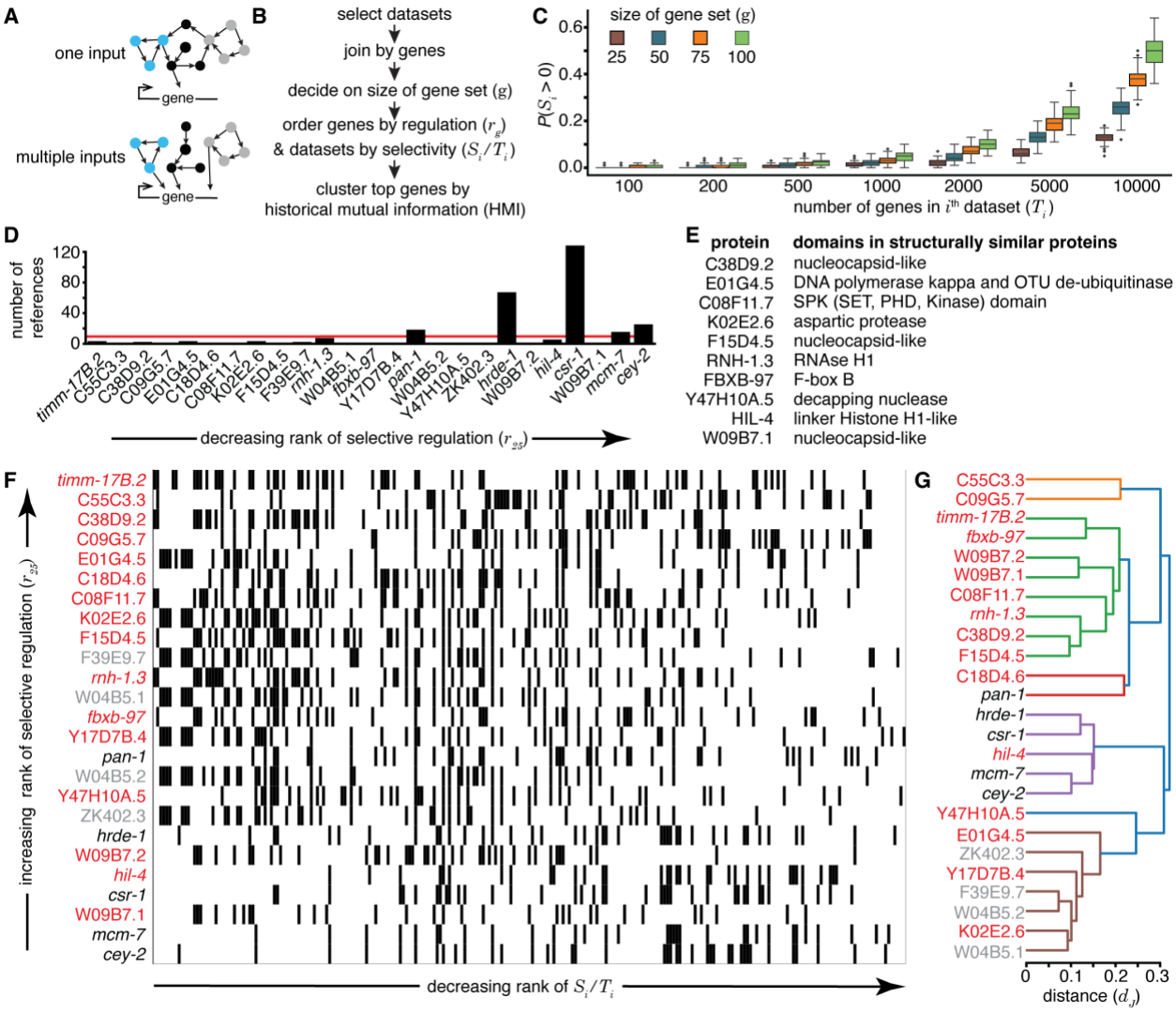

Figure 1. Some genes are selectively regulated, reported as part of many lists, and yet are understudied. (*A*) Schematics of possible regulatory architectures for genes found on multiple lists. (*top*) Gene receiving one input form a large network. (*bottom*) Gene receiving multiple inputs from separable networks. (*B*) Strategy for the identification of regulated genes. See Methods for details. (*C*) Relationship between S_i, T_i, and g obtained using simulated data for an organism with 20,000 genes. Distributions of 100 runs for each parameter combination are presented as box and whisker plots. (*D*) Numbers of publications listed on WormBase for the top 25 regulated genes as measured using r_{25} in the field of RNA silencing in *C. elegans*. Red line marks 10 publications. (*E*) Domains present in proteins encoded by understudied genes among the top 25 genes that are suggestive of function. Proteins with high-confidence AlphaFold structures [8] were used to identify related proteins using Foldseek [9] or based on the literature ([7, 10]; C38D9.2, F15D4.5, and W09B7.2). (*F*) Heat map showing the top 25 regulated genes. Presence (black) or absence (white) of each gene in each dataset is indicated. Relatively understudied (<10 references on WormBase) genes (red) or pseudogenes (grey) identified in (*D*) are indicated. (*G*) Hierarchical clustering of the top 25 genes based on co-occurrence in studies, where gene names colored as in (*D*) and 'distance (d_J)' indicates Jaccard distance.

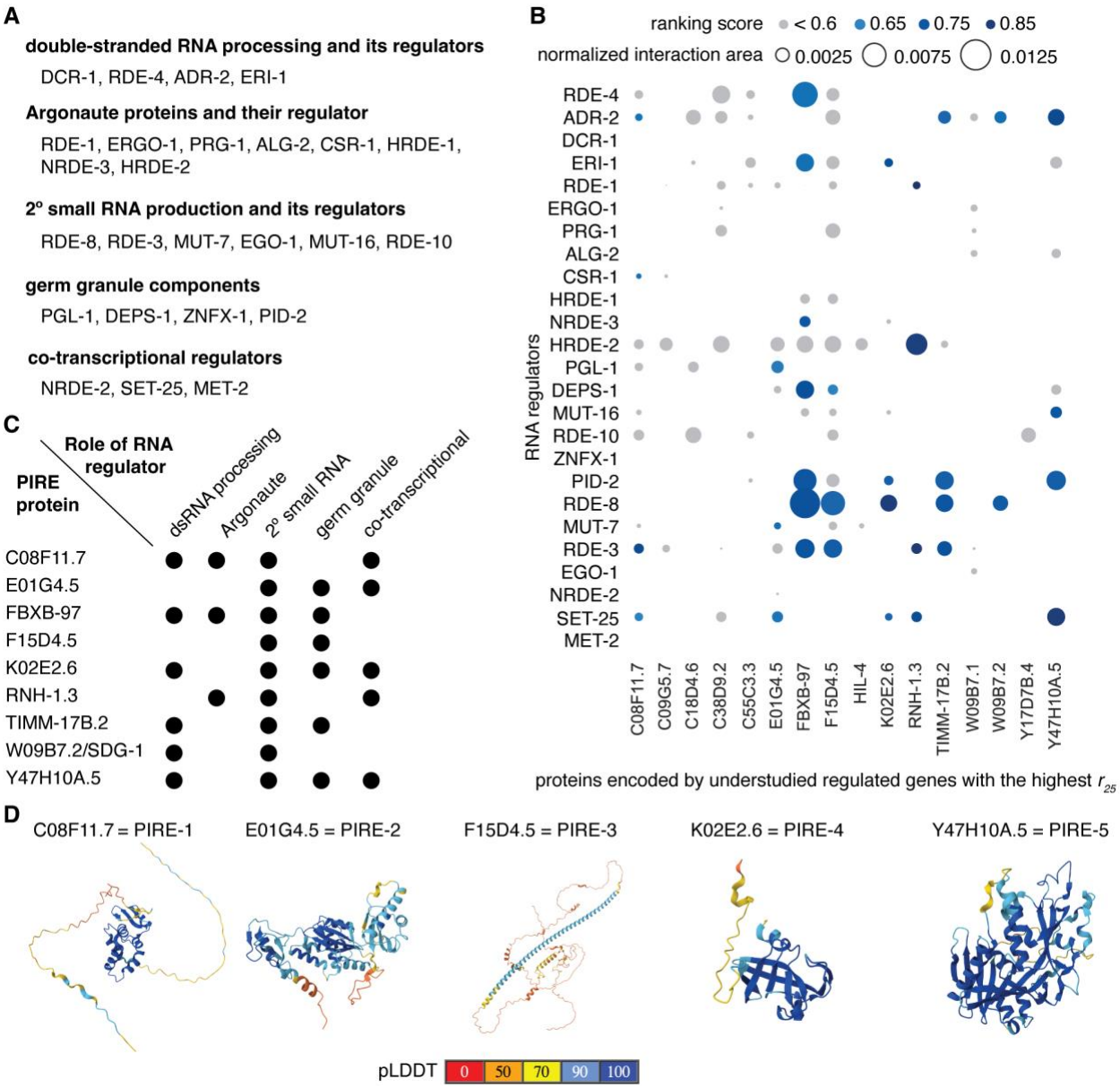

Figure 2. Understudied regulated genes encode proteins predicted to interact with key regulators of RNA silencing. (A) Regulators of RNA silencing in different categories examined for predicted interactions with proteins encoded by understudied genes identified in this study. See text for details. (B) Predicted interactions between proteins encoded by genes with the highest r_{25} scores and known regulators of RNA silencing in *C. elegans*. The area of the interaction surface between partners normalized by the product of the sizes of the interactors is shown as a bubble plot (inter-protein predicted aligned error <5 and inter-residue distance <6). Also see Fig. S1 and Movies S1 to S35. (C) Proteins encoded by understudied genes with significant interactions (top), including five named as predicted influencers of RNA-regulated expression (PIRE) (bottom). (D) Predicted structures for the five newly named PIRE proteins are shown with the per-residue confidence (pLDDT) as present in the AlphaFold protein database [58].

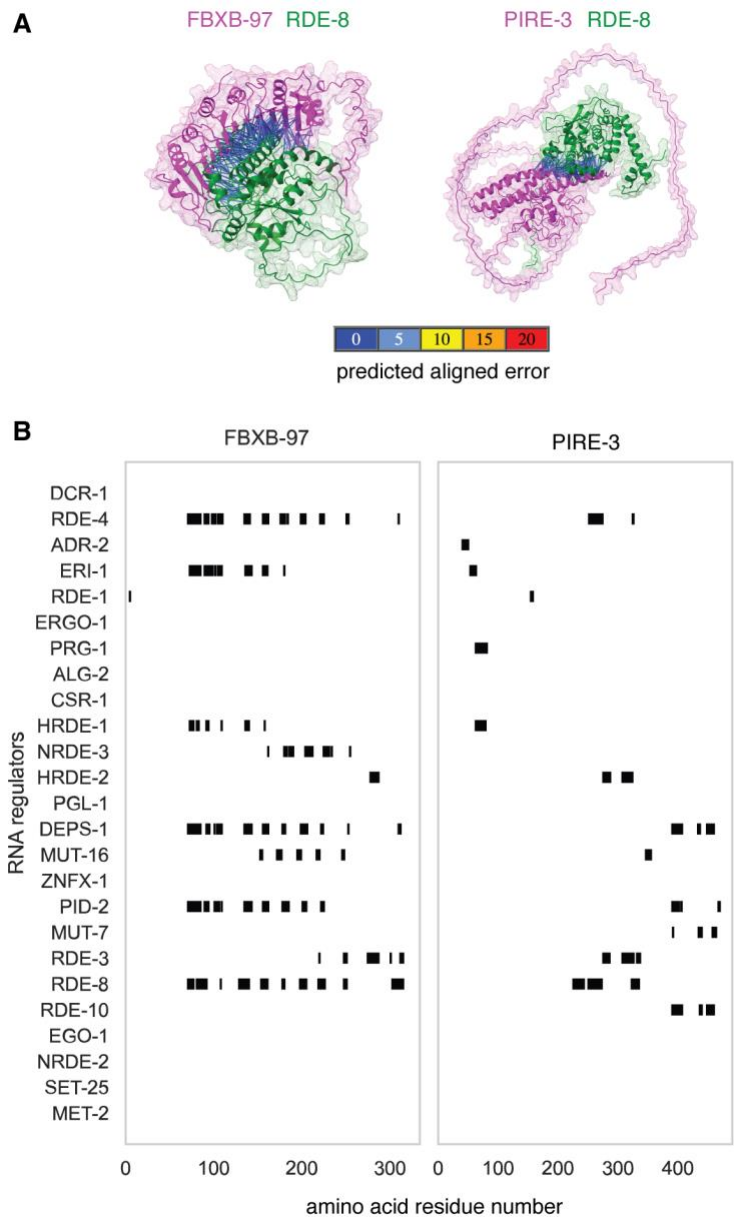

Figure 3. Predicted Influencer of RNA-regulated Expression (PIRE) proteins interact with regulators of RNA silencing in two general modes. (A) Predicted interactions between the PIRE proteins (magenta) FBXB-97 (left) and PIRE-3 (right) with the known regulator RDE-8 (green) that are of high confidence (constraining more than 20 amino acid residues with an inter-Cα distance less than 6 and PAE less than 5) are indicated with pseudo bonds. (B) Regions of the PIRE protein sequence constrained by the interacting regulator. Markers (black) are enlarged with respect to the X-axis for visibility (e.g., the marker denoting the interaction between RDE-1 and FBXB-97 only indicates one residue).

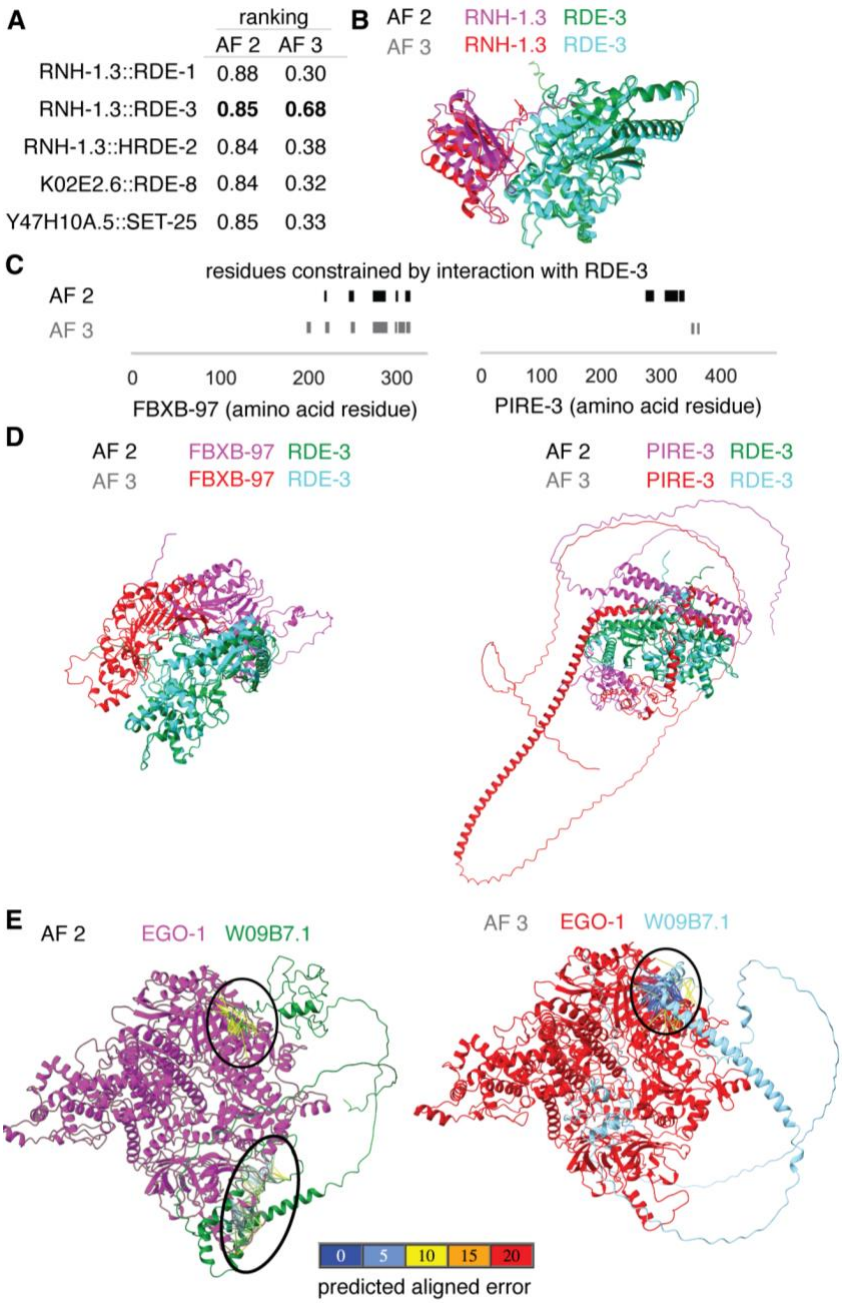

Figure 4. Interactions predicted by AlphaFold 2 and by the AlphaFold 3 server can differ. (A) Comparison of the top ranking interactions between known regulators of RNA silencing and the PIRE proteins predicted by AlphaFold 2 (AF 2 [24]; 0.8*ipTM + 0.2*pTM) compared with the score generated by AlphaFold 3 (AF3 [21]; 0.8*ipTM + 0.2*pTM + 0.5*disorder). A high-confidence prediction by both approaches is highlighted in bold. (B) Models for the interaction of RNH-1.3 with RDE-3 generated by AF2 and AF3 overlayed using RDE-3. Also see Movie S36. (C) Comparison of residues of PIRE proteins constrained through interactions as predicted by AF2 (black) or by AF3 (grey). (D) Comparison of interactions between FBXB-97 and RDE-3 (left), and between PIRE-3 and RDE-3 (right) as predicted by AF2 (black) and the AF3 server (grey), respectively. Structures are shown with differential coloring of each protein and overlayed using the RDE-3 structures in both cases. Also see Movies S37 and S38. (E) Interactions between

EGO-1 (magenta or red) and W09B7.1 (green or cyan) predicted by AF2 or AF3. Black ovals indicate interacting regions with inter-protein PAE <10 (left) or <5 (right). Also see Movie S39.

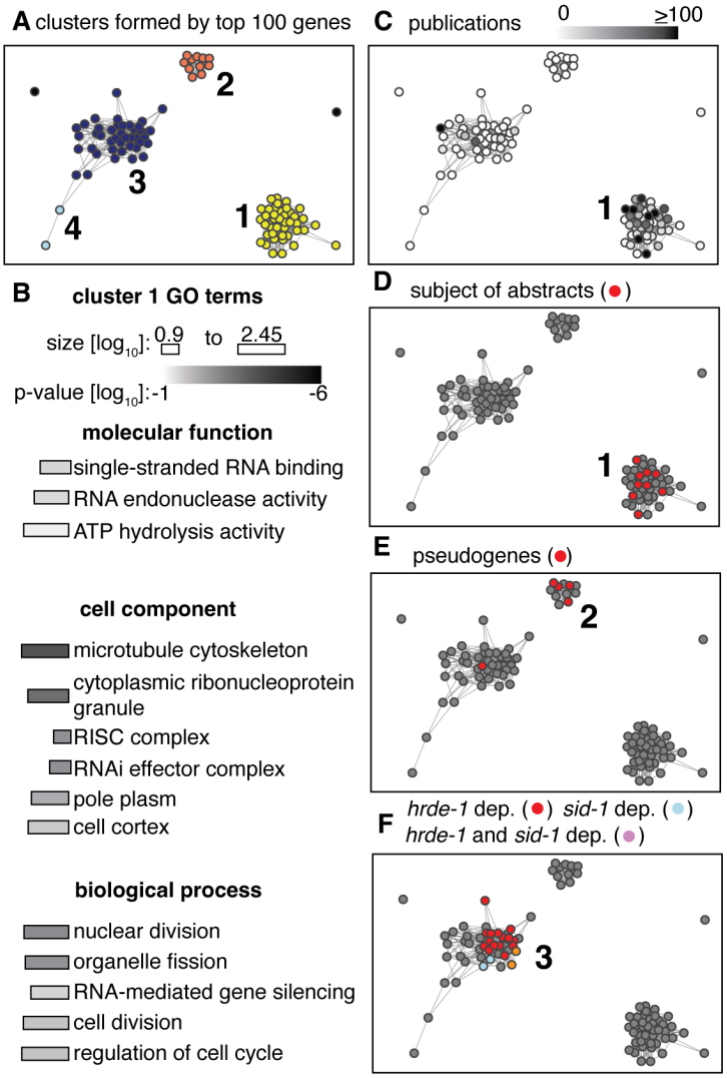

Figure 5. Clusters formed by understudied regulated genes suggest priorities for detailed study. (*A* and *B*) Properties of the top 100 regulated genes as measured using r_{100} in the field of RNA silencing in *C. elegans.* (*A*) Clusters of genes based on their historical mutual information. Threshold for link: HMI > 0.9. Also see Table S3. (*B*) Molecular functions (*top*), cell components (*middle*), and biological processes (*bottom*) of genes in cluster 1 as in (*A*). Length of boxes near each term in (*B*) indicates $\log_{10}$(annotations for GO term in *C. elegans*), with largest and smallest bars indicating ~285 and ~8 annotations, and shading indicates -$\log_{10}$(Bonferroni-corrected p-value), with black and white indicating a p-value of ~10^{-6} and ~10^{-1}, respectively. (*C-F*) Network in (*A*) with nodes colored to show number of publications per gene (white, 0; black, ≥100) (*C*), genes that have been the main subject of abstracts on RNA silencing in *C. elegans* (*D*), pseudogenes (red) (*E*), and genes changed in *hrde-1* mutants [41] (red), a *sid-1* mutant [7] (blue), or both (orange) (*F*).