Selecting genes for analysis using historically contingent progress

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; and F. L. and A. M. J. wrote the paper.

Competing Interest Statement: The authors declare no competing interests.

Keywords: Mutual Information, RNA silencing, gene regulation, *C. elegans*.

This file includes:

Main Text

Figures 1 and 2

SUMMARY

Progress in biology has generated numerous lists of genes that share some property. But, advancing from the initial implication of a set of genes in a process to understanding their roles in the process is slow and unsystematic. Here we use RNA silencing in *C. elegans* to illustrate a general approach for comparing lists of data accumulated by a field to prioritize genes for detailed study given limited resources. The partially subjective relationships between genes forged by both functional relatedness of the genes and biased progress in the field was captured as historical mutual information (HMI) and used as a quantitative measure for clustering genes. These clusters suggest regulatory links connecting RNA silencing with other processes like the cell cycle and identify understudied regulated genes that could be used to sense perturbation or mediate feedback inhibition.

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 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. 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 a group of studies or even all studies in a relatively young field. 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 that provide explanations 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, making them experimentally useful as general sensors of perturbation. Here we present an approach to identify these regulated but understudied genes in the field of RNA silencing in *C. elegans*.

Results

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 ideal 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 [2] and HRDE-1 [3], 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 importer SID-1 and encodes a protein with a suggested role in feedback regulation of heritable RNA silencing by colocalizing with perinuclear germ granules [4]. 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, 7 had high-confidence AlphaFold structures [5], which were then used to identify related protein domains using Foldseek [6] (Fig. 1*E*). Three more proteins have been proposed to be nucleocapsids encoded from genes within retrotransposons ([4, 7]; Fig. 1*E*). These candidates can be experimentally analyzed in the future for roles in RNA silencing, if any.

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 [8], as the understudied gene clustering closest to *hrde-1* and *csr-1*, making it a strong candidate for a role potentially 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.

To examine if the observations 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 [9] 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. 2*A*). Only one cluster (cluster 1 in Fig. 2*A*) had significant numbers of genes associated with gene ontology terms. Many of these genes encode proteins that bind and/or hydrolyze RNA (Fig. 2*B*, *top*), localize to cytoplasmic ribonucleoprotein granules (Fig. 2*B*, *middle*), and/or play roles in other processes such as cell division (Fig. 2*B*, *bottom*). Consistently, this cluster also had the greatest number of genes that have been described in multiple publications (Fig. 2*C*), including all the genes that have been featured in abstracts on RNA silencing (Fig. 2*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. 2*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 [10]) and genes in a single cluster (Fig. 2*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 [10]. 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* [4]) overlap with genes in the same cluster (Fig. 2*F*, 4 of 42 in cluster 3). Future studies by labs working on RNA silencing in *C. elegans* have the potential to test and enrich the classification of regulated yet understudied genes revealed here.

Discussion

Our analysis has identified selectively regulated yet understudied genes in the field of RNA silencing in *C. elegans*. Clustering these genes reveal that better studied genes are together in one cluster. Many of these genes have known roles, providing regulatory links between RNA silencing and other processes. The other two larger clusters include genes and pseudogenes that have been described as targets of RNA regulation. These regulated genes, which are present in many lists, could be under the independent control of multiple regulators and/or be jointly regulated by a connected set of regulators, making them experimentally useful general sensors of perturbation in RNA silencing and/or regulators that mediate feedback inhibition of RNA silencing.

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 that had to use older technologies with limitations 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 [11, 12], but when analyzed using single-

copy transgenes, RDE-4 showed a cell autonomous effect [13]. 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 customization for the lists included.

Different properties of a single protein or RNA could be important for different biological roles [14, 15], or the same properties could be important for different processes. Despite such plurality, 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. With the expanding number of lists generated through 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.

Methods

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. 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) [16] or as a symmetric measure of normalized mutual information [9], 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 [17] and different sets of genes were highlighted (Fig. 2). Gene ontology (GO) analyses were performed

using Gene Ontology Resource (https://geneontology.org/; [18, 19]) and significant GO terms were collected for visualization using REVIGO [20]. All programs used in this study are available at GitHub (AntonyJose-Lab/Lalit_Jose_2024).

Acknowledgements

We thank Tom Kocher and members of the Jose lab for comments on the manuscript. 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** (2020).
2. 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).
3. B. A. Buckley *et al.*, A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature* **489**, 447-451 (2012).
4. N. Shugarts *et al.*, SID-1 regulates a retrotransposon-encoded gene to tune heritable RNA silencing. *bioRxiv* 10.1101/2021.10.05.463267 (2023).
5. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
6. M. van Kempen *et al.*, Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* 10.1038/s41587-023-01773-0 (2023).
7. 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).
8. 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).
9. I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, San Francisco, ed. 2nd, 2005).
10. 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).
11. 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).
12. 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).
13. 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).
14. C. E. Chapple *et al.*, Extreme multifunctional proteins identified from a human protein interaction network. *Nat Commun* **6**, 7412 (2015).
15. 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).
16. 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).
17. M. Girvan, M. E. Newman, Community structure in social and biological networks. *Proc Natl Acad Sci U S A* **99**, 7821-7826 (2002).
18. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

19. G. O. Consortium *et al.*, The Gene Ontology knowledgebase in 2023. *Genetics* **224** (2023).
20. F. Supek, M. Bosnjak, N. Skunca, T. Smuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).

Figures and Figure Legends

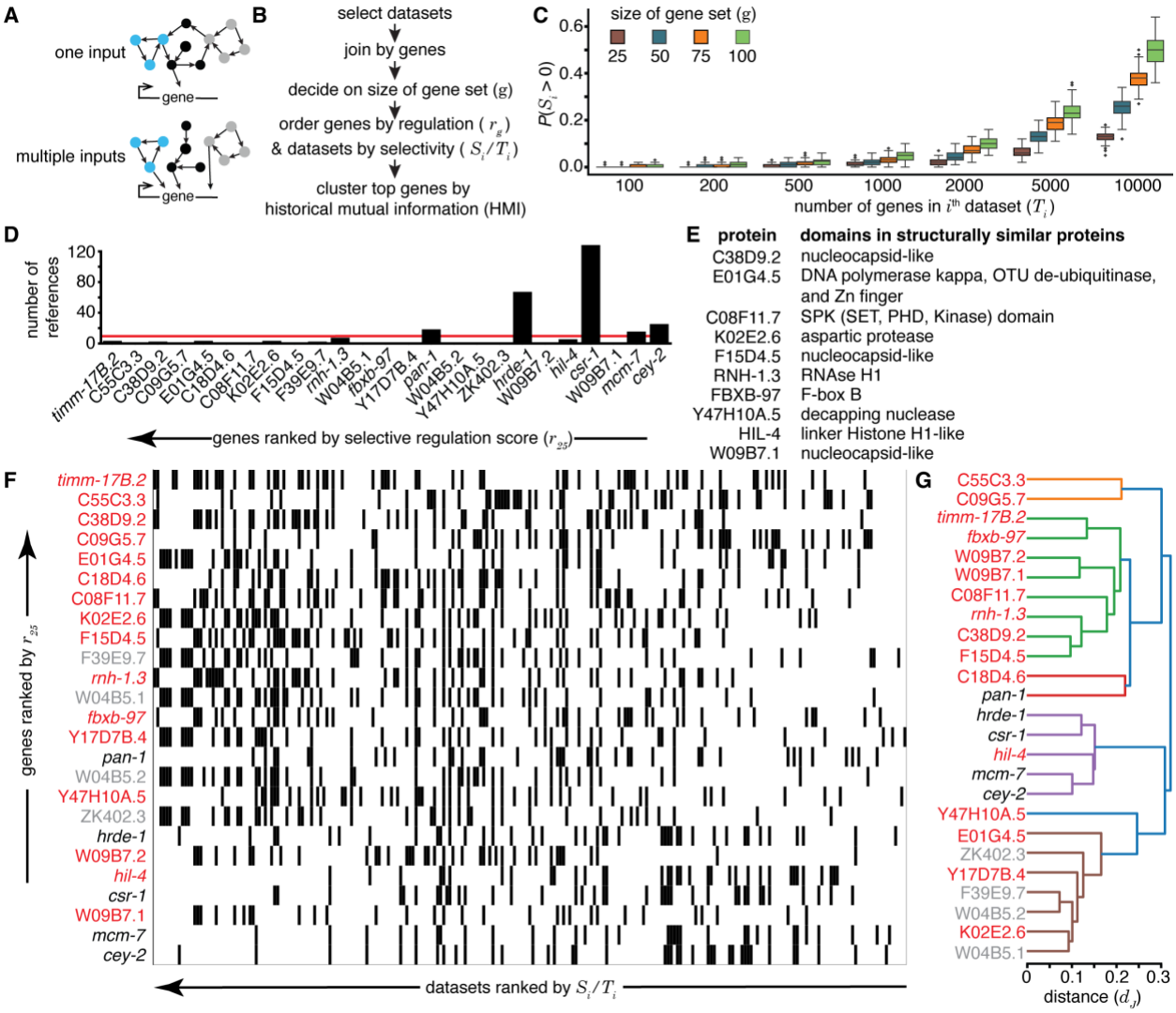

Figure 1. Some genes are selectively regulated, reported as part of many lists, and yet are understudied. (*A*) Schematics of possible regulatory architectures for genes found on multiple lists. (*top*) Gene receiving one input form a large network. (*bottom*) Gene receiving multiple inputs from separable 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 from 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 [5] were used to identify related proteins using Foldseek [6] or based on the literature ([4, 7]; 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.

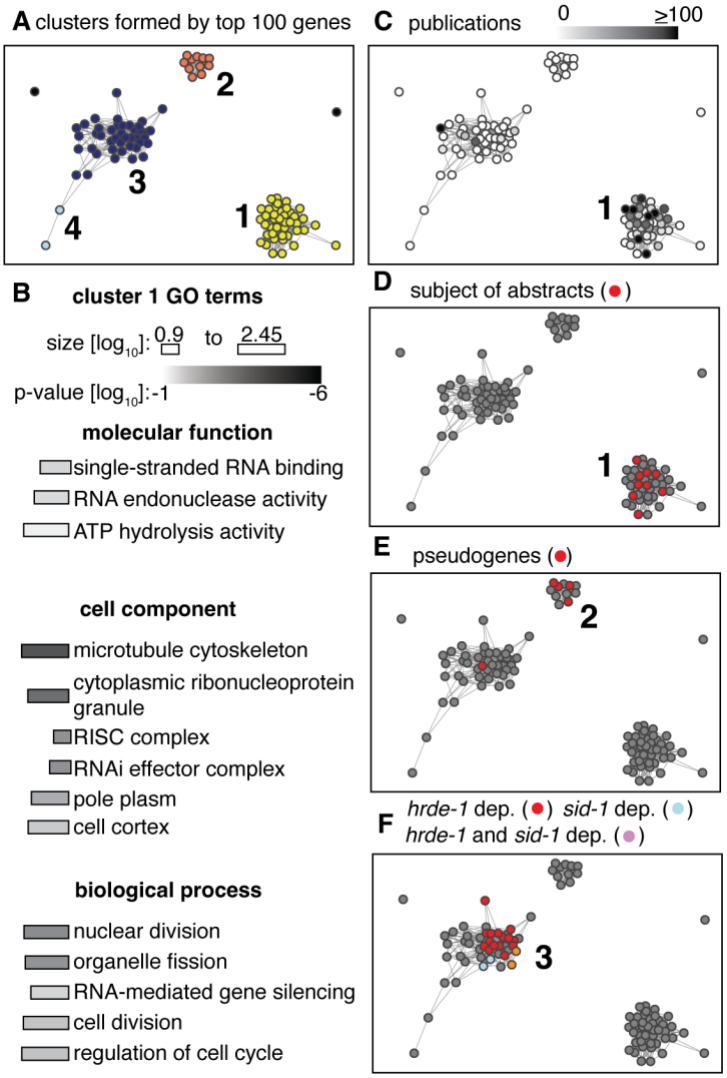

Figure 2. Clusters formed by regulated genes suggest priorities for detailed study. (*A* and *B*) Properties of the top 100 regulated genes as measured using r_{100} in the field of RNA silencing in *C. elegans.* (*A*) Clusters of genes based on their historical mutual information. Threshold for link: HMI > 0.9. Also see Table S2. (*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 [10] (red), a *sid-1* mutant [4] (blue), or both (orange) (*F*).