Harnessing full-text publications for deep insights into *C. elegans* and *Drosophila* connectomes

Karthick Raja ARULPRAKASAM[1], Janelle Wing Shan TOH[1], Mani R KUMAR[1], Herman FOO[1], Emilia Emmanuelle DAVEY[1], Marek MUTWIL[1], Guillaume THIBAULT[1,2]

[1]School of Biological Sciences, Nanyang Technological University, Singapore, 637551
[2] Mechanobiology Institute, National University of Singapore, Singapore 117411

ORCID
KRA 0000-0002-0403-2910
MM 0000-0002-7848-0126
GT 0000-0002-7926-4812

Correspondence to:
Marek Mutwil, +65 6904 7503, email: mutwil@ntu.edu.sg
Guillaume Thibault, +65 6592 1787, email: thibault@ntu.edu.sg

Running Title: Revolutionizing *C. elegans* and *Drosophila* Connectome Analysis

SUMMARY

In the rapidly expanding domain of scientific research, tracking and synthesizing information from the rapidly increasing volume of publications pose significant challenges. To address this, we introduce a novel high-throughput pipeline that employs ChatGPT to systematically extract and analyze connectivity information from the full-texts and abstracts of 24,237 and 150,538 research publications concerning *Caenorhabditis elegans* and *Drosophila melanogaster*, respectively. This approach has effectively identified 200,219 and 1,194,587 interactions within the *C. elegans* and *Drosophila* connectomes, respectively. Utilizing Cytoscape Web, we have developed comprehensive, searchable online connectomes that link relevant keywords to their corresponding PubMed IDs, thus providing seamless access to an extensive knowledge network encompassing *C. elegans* and *Drosophila*. Our work highlights the transformative potential of integrating artificial intelligence with bioinformatics to deepen our understanding of complex biological systems. By revealing the intricate web of relationships among key entities in *C. elegans* and *Drosophila*, we offer invaluable insights that promise to propel advancements in genetics, developmental biology, neuroscience, longevity, and beyond. We also provide details and discuss significant nodes within both connectomes, including the insulin/IGF-1 signaling (IIS) and the notch pathways. Our innovative methodology sets a robust foundation for future research aimed at unravelling complex biological networks across diverse organisms.

INTRODUCTION

The landscape of biological research has experienced a significant transformation over the past two decades, marked by an exponential surge in the volume of scientific publications. This trend is particularly pronounced in the study of model organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster*, which have long served as pivotal systems for understanding fundamental biological processes. A PubMed search reveals the magnitude of this surge: the number of publications related to *C. elegans* has escalated from 6,166 in 2000 to over 40,000 in 2023, while *Drosophila* research has expanded from 40,793 to more than 120,000 publications within the same timeframe. This proliferation of data, while testament to the fields' dynamism and the research community's productivity, presents a formidable challenge. Researchers are now faced with the Herculean task of staying abreast of emerging insights and effectively synthesizing vast amounts of information. The critical need for sophisticated tools to navigate, manage, and interpret this growing knowledge base has never been more apparent. Without such innovations, the research community's capacity to forge

novel connections and draw meaningful insights from the wealth of available data may be substantially hindered, potentially slowing the trajectory of scientific progress.

Existing connectome and pathway databases, such as BioGRID, the Gene Ontology (GO), and Reactome, offer valuable insights into the complex networks of gene interactions and biological pathways. BioGRID is an extensive repository for interaction datasets, facilitating the exploration of protein and genetic interactions in a variety of organisms (Stark et al., 2006). However, like many databases, it may not capture real-time research dynamics due to the inevitable delay in data curation (Oughtred et al., 2019). The Gene Ontology (GO) provides a comprehensive framework for the representation of gene function across species, yet it can be constrained by its static consensus terminology and may not capture the full spectrum of gene functionality or recent discoveries (Gaudet & Dessimoz, 2017). Reactome, while detailing pathways of numerous biological processes, could also potentially miss species-specific functions and unique cellular conditions (Milacic et al., 2024). Despite the undeniable utility of these tools, they are not without limitations. Their reliance on curated data ensures accuracy but can result in updates lagging behind the latest literature due to the labor-intensive nature of manual curation. Moreover, these databases might not fully encapsulate the multifaceted relationships between genes, such as epistatic interactions, genetic modifiers, and context-dependent effects. Such relationships are essential for a comprehensive understanding of complex phenotypes and diseases. This underscores a crucial gap: the need for an advanced tool capable of dynamically incorporating the latest findings and representing the intricate web of genetic interactions with both depth and breadth.

In the fields of *C. elegans* and *Drosophila* research, gold-standard organism-specific databases such as WormBase and FlyBase have been instrumental in collating vast amounts of genetic information (Gramates et al., 2022; Harris et al., 2010). These platforms offer insights into gene function, protein-protein interactions, and phenotypic data crucial for understanding development, aging, and disease in these model organisms. WormBase, for instance, has been a pivotal tool for *C. elegans* researchers, providing curated genetic, genomic, and biological information. Similarly, FlyBase serves the *Drosophila* community by compiling data on genetic and molecular attributes of *Drosophila* genes and genomes. However, even these comprehensive repositories may not fully capture the dynamic and rapidly evolving insights emerging from current literature. Key information on context-specific gene interactions, the influence of environmental factors on genetic pathways, and the subtleties of temporal and spatial gene expression patterns are often more thoroughly

detailed in individual studies. As such, there exists a gap between curated databases and the nuanced, high-resolution data that can be mined from full-text publications, which often contain rich, yet uncurated, insights into gene function and regulation.

In this work, we unveil a high-throughput text-mining pipeline designed to systematically extract and analyze gene-related information from a vast array of research publications concerning *C. elegans* and *Drosophila*. Utilizing the capabilities of natural language processing technologies, this pipeline transcends the confines of traditional databases to offer a dynamic, enriched view of the genetic interaction landscapes of these model organisms. We detail the development and deployment of our pipeline, showcasing its unparalleled ability to uncover and visualize intricate networks of gene interactions and biological pathways. Through this effort, we aim to equip researchers with a robust tool to navigate the growing wealth of genetic and biological information in *C. elegans* and *Drosophila*, thereby catalyzing significant advances in our systemic understanding of biology. Our study underscores the feasibility and the transformative impact of integrating advanced computational methods with bioinformatics to enhance our grasp of complex biological systems. The *C. elegans* and *Drosophila* Connectomes, developed as a result of this endeavor, are accessible at http://worm.connectome.tools and http://drosophila.connectome.tools, respectively, serving as comprehensive portals to an array of interactions and pathways.

RESULTS

Semantic Analysis of Thousands of Abstracts and Full-Text Papers

In our comprehensive analysis, we processed a substantial amount of literature pertaining to gene function within *C. elegans* and *Drosophila*. To construct the *C. elegans* and *Drosophila* Connectomes, we systematically searched scientific research papers for occurrences of *C. elegans* or *Drosophila* genes in their titles and abstracts. We prioritized full-text papers from open-access publications, those freely available in PubMed Central (PMC), and accessible through Elsevier via the NTU Library. Alternatively, it went back to fetch only titles and abstracts if full-text unavailable. This approach ensured that the selected papers were directly relevant to the genetic makeup and biological processes of these model organisms. The result was a curated selection of articles that form the backbone of our *C. elegans* or *Drosophila* Connectomes databases. Initially, we extracted 24,237 *C. elegans*-related articles, comprising 9,904 full-text articles and 14,332 abstracts. For *Drosophila*, the amount included an even larger set of 150,538 articles, with 71,226 full-text

92 articles and 79,311 abstracts. Articles were initially categorized based on gene nomenclature, which
93 necessitated a subsequent deduplication step due to multiple occurrences of the same articles across different
94 categories. Following this curation process, we filtered the dataset to a final tally of 12,062 articles for the *C.*
95 *elegans* connectome and 36,372 articles for the *Drosophila* connectome. These articles span a wide
96 distribution across 925 journals for *C. elegans* and 1,815 journals for *Drosophila* (Table S1). For a more
97 detailed visualization, we have compiled the top 40 most frequently cited journals in both fields, demonstrating
98 the quantitative distribution of the articles (Figures 1A and 1B). These articles form the foundation of our
99 subsequent analysis using GPT, enabling a refined exploration of the genetic and molecular interplays that
100 define the complex connectomes of these model organisms.
101
102 We exploited the capabilities of OpenAI's GPT API to extract functional relationships between entity pairs (e.g.,
103 gene A interacts with gene B), infer functional annotations of genes (e.g., gene A is implicated in dauer
104 formation), and delineate abbreviations (e.g., FOXO for FORKHEAD BOX O). To optimize our query efficiency,
105 we iteratively tested different prompts with ChatGPT until we refined a distinct prompt (Table S2). The prompt
106 was systematically applied across all abstracts and full-text articles, revealing 200,219 functional relationships
107 for the *C. elegans* connectome and 1,194,587 for the *Drosophila* connectome, along with 112,128 gene
108 function annotations and 73,591 abbreviation identifications. To evaluate the quality of GPT's output, we
109 examined the relationships or 'edges' that it identified between two genes and/or entities within the literature
110 network. The GPT analysis generated statements in the format of “Entity1! Relationship! Entity2”. We classified
111 edges as “good” when at least one of the entities appeared in the GPT analysis, and as “bad” when neither
112 entity was found. For *C. elegans*, out of 334,104 total edges, 268,632 were designated as “good”, while 65,472
113 were “bad” (Figure 1C). In parallel, the *Drosophila* connectome revealed 1,498,156 edges, with 1,056,937
114 categorized as 'good' and 441,319 as “bad” (Figure 1D). Such results lend confidence to the integrity of GPT's
115 output, affirming that most relationships identified align with the expected analytical format. Next to evaluate
116 the precision of our text-mining process, we conducted a manual accuracy assessment on a random sample
117 of 50 abstracts. The outcomes indicated that GPT predominantly identified relationships correctly (Figure 2,
118 Documents S2 and S3). However, we did observe instances where relationships were either missed or
119 inaccurately characterized. Such inaccuracies were notably prevalent in abstracts and full-text articles that did
120 not mention specific gene names, leading to instances of GPT "hallucinating" entities, incorrectly designating
121 them as “gene”.

Expanding the *C. elegans* Interaction Map for a Connectome with Comprehensive Coverage

Leveraging the outputs from GPT analysis, which consist of pairwise relationships between entities, we crafted a network that encapsulates the full spectrum of functional relationships within the *C. elegans* biological system. Despite GPT's instruction to prioritize genes, (Table S2) the analysis yielded interactions that included not only gene-gene interactions but also connections to biological functions, pathways, and phenotypes (Figure 3). During the curation process, we encountered numerous instances where gene functions were discussed generically, using placeholders such as "gene", or the organism names "*Caenorhabditis elegans*" and "worm" were utilized as proxy for specific genetic entities. These non-specific terms were subsequently removed to sharpen the focus on the top 20 most meaningful entities within the *C. elegans* pool. The analysis spotlighted "*daf-16*" and "LIFESPAN" as the most prevalent topics in *C. elegans* literature, underscoring their significance in the field (Figure 3B). Moreover, examining the types of relationships revealed that "regulated", "requires", and "interacts with" emerged as the most common edges, delineating the primary modes of genetic and molecular interactions in *C. elegans* (Figures 3A and 3D). To understand the pattern of mostly investigated genes we pulled out the information of top 5,000 entities, edges between entities, genes as entities, and edges between genes that are appeared in both *C. elegans* and *Drosophila* connectome analysis (Table S3-S6). This analysis showed that "*lin-12*", "*let-23*", and "*par-3*" are mostly investigated genes in *C. elegans* (Figure 3C). The comparative frequency of these edges demonstrates the thematic congruencies within the biological research of these model organisms, reflecting shared foundational processes that are central to understanding their complex biology. The connectome thus constructed offers a comprehensive view of the intricate web of interactions that define *C. elegans* biology, serving as a valuable resource for researchers navigating this model organism's extensive genetic landscape.

Further analysis was conducted to assess the comprehensiveness of the *C. elegans* Connectome in relation to the established BioGRID database, focusing on interacting edges which includes "interact", "bind", and "phosphorylate" relationships from the whole connectome database and the protein-protein interaction (PPI network) only from connectome database (Figures 3E and 3F). We identified 8,565 interacting edges in the connectome. Notably, the Connectome and BioGRID share 311 identical interacting partners. Additionally, there are 158 interactions catalogued in BioGRID that overlap with the Connectome; these, however, are not specifically defined as interacting. In terms of the PPI network, the *C. elegans* Connectome demonstrated an overlap of 298 PPI interacting edges with BioGRID, yet it also identified an additional 4,285 interacting edges

not catalogued by BioGRID (Figure 3F). These findings highlight the Connectome's strength in detecting not only the commonly recognized interactions but also a broader spectrum of biological relationships. Furthermore, the Connectome's capability to capture more nuanced interaction types beyond interacting—such as “colocalize with” and “inhibits”—affirms its utility in providing a more detailed and extensive mapping of molecular interactions (Figure 3).

In our comprehensive mapping of the *Drosophila* connectome, we delineated a network equally rich and intricate as its *C. elegans* counterpart. The extent of interaction types identified among *Drosophila* genes, with “regulates” and “interacts with” emerging as particularly common, indicating frequent protein-protein and gene regulatory interactions (Figure 4A). Broader biological entities and processes that are recurrently discussed, such as “apoptosis” and “lifespan”, highlight their fundamental importance in *Drosophila* studies (Figure 4B). The genes that dominate the *Drosophila* Connectome, such as “*VIA*”, “*cycle*” and “*actin*”, underscoring their prominence and frequent investigation within the species' genetic research (Figure 4C). It is important to note that terms like “VIA” and “cycle” cover both specific gene names and instances where these words do not refer to genes in the prompts/connectomes. Due to this, such terms cannot be distinctly identified as gene-related by ChatGPT without additional contextual analysis. Complementary , the most common gene-to-gene edges include “has”, “regulates” and other informative terms on the type of interactions such as “is required for”, illustrating the extensive interconnectivity within the *Drosophila* genome (Figure 4D). Our Connectome contained 44,382 edges categorized as “interacting” (Figure 4E). Of these, 289 were found to directly correspond with interactions listed in BioGRID. Additionally, the Connectome shared another 132 interactions with BioGRID, though these were not explicitly categorized under the “interacting” descriptor. Regarding the protein-protein interaction (PPI) network, a comparison revealed that 19 PPI edges were common between the *Drosophila* Connectome and BioGRID (Figure 4F). Nevertheless, our Connectome unveiled 9,101 additional PPI edges, absent in BioGRID's catalog. Together, the analysis of both the *C. elegans* and *Drosophila* connectomes illuminates their potential to not only complement but substantially augment existing genetic databases. By unveiling a multitude of previously unavailable interactions, these connectomes serve as unconventional resources that encapsulate the evolving complexity of biological research. They provide researchers with dynamic and current tools essential for exploring into the genetic and molecular fabric of these model organisms.

Interactive Connectome Platforms for *C. elegans* and *Drosophila*

The *C. elegans* and *Drosophila* Connectomes offer an interactive gateway to a vast array of genetic interactions, parallel to a digital atlas for biological functions within these model organisms. Through an intuitive interface, users can query genes, proteins, and other entities, obtaining detailed information pages that include GPT-generated abbreviations, functional annotations, and direct links to scientific articles. The platforms feature KnowledgeNetworks, visual representations of the connectome that allow users to customize the view, isolate nodes, and even download data for advanced analysis. A comprehensive summary accompanies each network, providing immediate insights into the most connected nodes and their respective literature sources. For those requiring programmatic access, an API delivers the connectome's array of data in a structured format. Both Connectomes have been constructed with the same dedication to accessibility and depth of information as the PlantConnectome that we recently reported (Fo et al., 2023).

In the space of model organisms, few genes have garnered as much attention as *daf-16* in *C. elegans*, a gene whose conservation extends to its *Drosophila* counterpart, the fokhead box protein O (*foxo*). These genes pivotal connections within the complex networks of *C. elegans* and *Drosophila* biology, influencing essential processes like development, aging, metabolism, and stress response. DAF-16 and FOXO, transcription factors that interact with genes containing daf-16/FOXO binding elements (DBE), play crucial roles in the Insulin/IGF-1-like signaling (IIS) pathway. A query for "*daf-16*" and "*foxo*" within the *C. elegans* and *Drosophila* Connectomes maps a network sourced from 514 and 213 papers, respectively. Refinement of the search to "regulates" within the "Layout Options" reveals a more focused network from 61 and 71 papers for *C. elegans* and *Drosophila* (Figures 5A and 5B), respectively, featuring *daf-16* and *foxo*'s roles in regulating lifespan, feeding behaviors, dauer development, and stress resistance among other processes.

The *C. elegans* Connectome elucidates *daf-16*'s regulation of genes such as "CYP-35B1/DOD-13" (Iser et al., 2011), "SCL-1" (Ookuma, Fukuda & Nishida, 2003), "PAK-1" (Kennedy, Pham & Grishok, 2013), "COL-19P::GFP" (Wirick et al., 2021), and "SRH-234" (Gruner et al., 2014). Similarly, it highlights how *daf-16* and *foxo* regulate specific phenotypes in *C. elegans* and *Drosophila*, such as "FATTY ACID LIPOLYSIS" (Antebi, 2013) or "LIPOLYSIS IN FAT BODY CELLS" (Roy & Palli, 2018) and "L1 ARREST, DAUER DEVELOPMENT, AND AGING" (Kaplan & Baugh, 2016) or "STEM CELL AGING" (Artoni et al., 2017). These connections and

their respective publications are directly accessible through one-click links, demonstrating the tool's efficacy in providing a rapid, comprehensive overview of protein interactions.

The highly conserved Notch signaling pathway, pivotal in cell fate determinations, was initially discovered in *Drosophila*, marked by its role in wing morphology. Subsequently, the *C. elegans* counterparts, *lin-12* and *glp-1*, were identified, spotlighting their essential functions in developmental processes (Priess, Schnabel & Schnabel, 1987; Greenwald, Sternberg & Horvitz, 1983; Austin & Kimble, 1987). This pathway plays a crucial role in a multitude of biological processes (Kopan & Ilagan, 2009; Suarez Rodriguez, Sanlidag & Sahlgren, 2023), operating through cell-cell interactions initiated by transmembrane ligands that activate Notch receptors on adjacent cells. The activation leads to the cleavage of the Notch receptor's cytosolic domain, which then moves to the nucleus to regulate gene expression.

Leveraging the "notch" query within both the *C. elegans* and *Drosophila* Connectomes yielded networks sourced from 225 and 684 papers, respectively. Further refinement using "regulates" or "interacts with" in the "Layout Options" revealed more focused networks, sourced from 30 and 116 papers for *C. elegans* and *Drosophila*, respectively (Figures 5C and 5D). This analysis underscored the Notch pathway's regulation of "LIN-11 EXPRESSION (Marri & Gupta, 2009), "GERM CELL FATE SPECIFICATION" (Lee et al., 2016), and "C.ELEGANS BEHAVIOR" (Chao et al., 2005) in *C. elegans*. In contrast, the *Drosophila* Connectome illustrated a richer interaction network for the Notch pathway, indicating a more extensive exploration of its interacting partners in *Drosophila* or its suitability as a model organism for studying these interactions. Among the highlighted interactions were "NOTCH" interacts with "MKK4" (Zhou et al., 2021), "DMYC EXPRESSION" (Sun et al., 2008), "MORE THAN 300 GENES" (Ho, Pallavi & Artavanis-Tsakonas, 2015), and "AKAP200" (Bala Tannan et al., 2018), showcasing the pathway's broad influence across *Drosophila*'s genetic landscape.

DISCUSSION

In this study, we unveiled the comprehensive Connectomes for *C. elegans* and *Drosophila*, charting a vast landscape of genetic interactions and biological functions pivotal to these model organisms. Our analysis, underpinned by the innovative application of GPT technology, has facilitated the identification and cataloging of hundreds of thousands of genetic relationships, encompassing both well-documented and previously

unexplored interactions. Notably, genes such as *daf-16*/*foxo* in *C. elegans* and its functional counterpart in *Drosophila*, alongside the Notch signaling pathway, emerged as significant nodes within these networks. These hubs not only underscore the genetic complexity inherent in biological processes like development, aging, and stress response but also highlight the Connectomes' capacity to unearth interactions that span across a broad spectrum of biological research. The creation of these Connectomes marks a significant stride in our ability to navigate the genetic intricacies of *C. elegans* and *Drosophila*, offering an enriched resource that advances our comprehension of their genetic frameworks and sets the stage for future discoveries in genetic regulation and function.

The introduction of our *C. elegans* and Drosophila Connectomes represents a significant augmentation to databases like BioGRID, which catalogs nearly 1.6 million interactions across various species through detailed literature annotations (Stark et al., 2006; Oughtred et al., 2019). Our connectomes, by mining full-text publications via computational techniques, offer a complementary approach. This methodology not only enriches the database with the latest research findings but also unravels biological contexts and intricate details of interactions, facilitating a more nuanced understanding of genetic networks. While BioGRID's manual curation process ensures the accuracy and reliability of its data (Oughtred et al., 2019), it may encounter challenges in rapidly integrating new discoveries. Our connectomes aim to mitigate this gap, leveraging natural language processing technologies to capture and incorporate emerging insights directly from the expansive volume of research articles. However, it's essential to recognize the foundational role of databases in the bioinformatics field. Their rigorously vetted information provides a valuable cornerstone that our computational approach seeks to extend, not supplant. The inclusion of CRISPR screen datasets into BioGRID signifies a notable expansion in the types of data curated, reflecting an evolution that our connectomes parallel through the adoption of advanced data mining techniques (Salwinski et al., 2009; Murugesan, Abdulkadhar & Natarajan, 2017). By integrating the strengths of manual curation with the scalability of automated text-mining, we aspire to create a synergistic resource. This combined approach aims to offer researchers a rounded view of the biological landscape, enabling a deeper understanding and facilitating discoveries in the genetics of *C. elegans* and *Drosophila*.

In constructing our *C. elegans* and *Drosophila* Connectomes, we aimed to address some of the challenges inherent in manual curation processes. Our strategy prioritized articles accessible at no cost, including open-

access publications and those available through PubMed Central (PMC) and Elsevier via the NTU Library. While this approach has allowed us to compile a vast and comprehensive database, it inherently limits our ability to immediately incorporate the latest research findings beyond titles and abstracts. This delay in integrating new studies poses a challenge in maintaining the most current view of complex biological networks. Recognizing this limitation, we are exploring innovative strategies to further enhance the timeliness and comprehensiveness of our database. Future directions could include forming collaborations with publishers to secure earlier access to research findings and developing automated text-mining tools that can more rapidly identify and incorporate relevant studies. By augmenting our current resources with these advanced methodologies, we will aim to minimize delays and ensure our databases remain at the forefront of biological research. These steps should not only improve the immediacy of data curation but also reinforce our commitment to providing a dynamic and cutting-edge resource for the scientific community.

In conclusion, the advent of the *C. elegans* and *Drosophila* Connectomes represents an advancement in the field of biological databases, transcending traditional limitations through computational mining and dynamic data incorporation. By leveraging full-text publications, these connectomes offer an enriched, contextually detailed exploration of biological interactions, including both the depth and breadth necessary for decoding complex biological systems. They are not just repositories of information but active platforms for discovery, enabling insights into the intricate interplay of genes and pathways. Moving forward, our focus should remain on expanding accessibility and enhancing data comprehensiveness, ensuring that the *C. elegans* and *Drosophila* Connectomes continue to evolve as indispensable resources in the quest to unravel the complexities of life's fundamental processes.

LIMITATION OF THE STUDY

While our approach significantly advances the scope of interaction data available by leveraging computational techniques to mine full-text publications, it is dependent upon the accessibility of these publications. Reliance on publicly accessible or institutionally available literature means that some recent studies, especially those behind paywalls or subject to embargo periods, may not be immediately integrated into our database. This could introduce delays in reflecting the most current research findings and innovations within the connectomes. Furthermore, while automated data extraction techniques offer scalability, they may not always achieve the accuracy of manual curation, potentially affecting the precision of interaction data. Efforts are ongoing to refine

these methodologies and expand our access to the latest scientific publications, ensuring our databases not only grow in volume but also in the quality and timeliness of the information they provide. Future directions will include exploring collaborations for wider access to publications and enhancing our algorithms for data mining to mitigate these limitations, continually striving to present the most accurate and comprehensive view of the biological landscapes we aim to model.

METHODS

KEY RESOUCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Custom codes	GitHub	https://github.com/mutwil/plant_connectome
Software and Algorithms		
BioPython 1.81	https://biopython.org/	Cock *et al*. (Cock et al., 2009)
ChartJs 4.4.1	https://www.chartjs.org/docs/master	n/a
Cytoscape.js 3.27	https://js.cytoscape.org	Frank *et al*. (Franz et al., 2016)
FileSaver 2.0.5	GitHub	https://github.com/eligrey/FileSaver.js
GPT davinci-002	Open AI	https://platform.openai.com
jQuery 3.7.1	https://jquery.com	n/a
json 2.6.3	PyPI	https://pypi.org
NetworkX 3.2	https://pypi.org/project/networkx/3.1	Hagberg *et al.* (Hagberg, Swart & Schult, 2008)
pickle5 0.0.12	PyPI	https://pypi.org
Python 3.9.15	Python Software Foundation	https://python.org
regex 2023.6.3	PyPI	https://pypi.org

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contacts, Marek Mutwil, at mutwil@ntu.edu.sg and Guillaume Thibault, at thibault@ntu.edu.sg.

Material Availability

This study did not generate new reagents.

Data and Code Availability

The custom codes to generate the connectomes is available at GitHub (https://github.com/mutwil/plant_connectome).

METHOD DETAILS

Retrieval of full-text papers

Using BioPython version 1.81, we downloaded all full-text articles freely available in PubMed. This was followed by the acquisition of institutional token access to Elsevier (NTU Library) and by downloading open-access full-texts. Additionally, we included abstracts from PubMed to ensure a robust dataset. The full-texts included titles, abstracts, introductions, results, and discussions. For analysis, each article was processed using OpenAI's Python API for the davinci 3.5 model, guided by specifically crafted prompts (Table S2). The output underwent further refinement to eliminate single-letter entities (e.g., removing 'Gene !affects! X') and to reframe passive edges into active ones (e.g., converting 'daf-16 ! can restore ! Secretory protein metabolism' to 'daf-16 ! restores ! Secretory protein metabolism'). Furthermore, edges with synonymous meanings were consolidated to enhance clarity and consistency. The model operated under default settings, with the exception of setting the temperature parameter to zero, promoting deterministic outcomes. In our final tally, a total of 24,237 articles for the *C. elegans* connectome and 36,372 articles for the *Drosophila* connectome were processed, encompassing both full-texts and abstracts. This comprehensive collection was assembled and analyzed within a span of two weeks, as outlined in the supplementary abstract.

Construction of *C. elegans* and *Drosophila* Connectome databases

Both connectomes are hosted on a Google Cloud server. The backend was implemented using the Python framework Flask and the Python packages networkx version 3.1, pickle version pickle5 0.0.12, json version 2.6.3, and regex version 2023.6.3. We used JavaScript dependencies jQuery v3.7.1, Cytoscape.js v3.27, ChartJS v4.4.1, and FileSaver v2.0.5 to visualize the KnowledgeNetwork graphs.

API for *C. elegans* and *Drosophila* Connectomes

C. elegans and *Drosophila* Connectomes are equipped with an Application Programming Interface to ease conducting search queries remotely by users. For each successful call to Connectome's API, a JSON object is returned, containing the functional abbreviations, GO terms, other nodes, and text summaries associated

with the search query. To perform searches using the API, users can add "/api/<search type>/<search query>" to the web address, where "<search type>" and "<search query>" are placeholders representing the type of search and user's query, respectively.

ACKNOWLEDGEMENTS

We thank members of Thibault lab for critical reading of the manuscript. This work was supported by funds from the Singapore Ministry of Education Academic Research Fund Tier 1 (RG96/22 to G.T.) and Tier 3 (MOET32022-0002 to M.M.) as well as the Research Scholarship to K.R.A [predoctoral fellowship from Singapore Ministry of Education Academic Research Fund Tier 3 (MOE-MOET32020-0001)].

Author contributions

Conceptualization: M.M. and G.T.; Methodology: M.M. and G.T.; Formal analysis: K.R.A. and J.W.S.T.; Investigation: K.R.A., J.W.S.T., and M.R.K.; Writing - original draft: G.T. and K.R.A.; Writing - review & editing: K.R.A., M.M., and G.T.; Supervision: E.E.D., M.M. and G.T.; Project administration: M.M. and G.T.; Funding acquisition: M.M. and G.T.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

ADDITIONAL FILES

Table S1, Related to Figure 1. List of journals curated for the *C. elegans* and *Drosophila* Connectome. Excel Spreadsheet.

Table S2, Related to Figure 2. An example of an abstract, prompts and outputs from GPT.

Table S3, Related to Figures 3A and 4A. List of the top 5,000 most frequent edges for the *C. elegans* and *Drosophila* Connectome. Excel Spreadsheet.

Table S4, Related to Figures 3B and 4B. List of the top 5,000 most frequent entities for the *C. elegans* and *Drosophila* Connectome. Excel Spreadsheet.

Table S5, Related to Figures 3C and 4C. List of the top 5,000 most frequent genes as entities for the *C. elegans* and *Drosophila* Connectome. Excel Spreadsheet.

Table S6, Related to Figures 3D and 4D. List of the top 5,000 most frequent edges between genes for the *C. elegans* and *Drosophila* Connectome. Excel Spreadsheet.

Document S1. Supplementary File 1.

Document S2. Manual accuracy assessment on a random sample of 50 abstracts for the *C. elegans* Connectome.

Document S3. Manual accuracy assessment on a random sample of 50 abstracts for the *Drosophila* Connectome.

REFERENCES

Antebi, A. (2013) Regulation of longevity by the reproductive system. *Experimental Gerontology*. 48 (7), 596–602. doi:10.1016/j.exger.2012.09.009.

Artoni, F., Kreipke, R.E., Palmeira, O., Dixon, C., Goldberg, Z. & Ruohola-Baker, H. (2017) Loss of foxo rescues stem cell aging in Drosophila germ line. *eLife*. 6, e27842. doi:10.7554/eLife.27842.

Austin, J. & Kimble, J. (1987) glp-1 is required in the germ line for regulation of the decision between mitosis and meiosis in C. elegans. *Cell*. 51 (4), 589–599. doi:10.1016/0092-8674(87)90128-0.

Bala Tannan, N., Collu, G., Humphries, A.C., Serysheva, E., Weber, U. & Mlodzik, M. (2018) AKAP200 promotes Notch stability by protecting it from Cbl/lysosome-mediated degradation in Drosophila melanogaster. *PLoS genetics*. 14 (1), e1007153. doi:10.1371/journal.pgen.1007153.

Chao, M.Y., Larkins-Ford, J., Tucey, T.M. & Hart, A.C. (2005) lin-12 Notch functions in the adult nervous system of C. elegans. *BMC neuroscience*. 6, 45. doi:10.1186/1471-2202-6-45.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M.J.L. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 25 (11), 1422–1423. doi:10.1093/bioinformatics/btp163.

Fo, K., Chuah, Y.S., Foo, H., Davey, E.E., Fullwood, M., Thibault, G. & Mutwil, M. (2023) *PlantConnectome: knowledge networks encompassing >100,000 plant article abstracts*.p.2023.07.11.548541. doi:10.1101/2023.07.11.548541.

Franz, M., Lopes, C.T., Huck, G., Dong, Y., Sumer, O. & Bader, G.D. (2016) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*. 32 (2), 309–311. doi:10.1093/bioinformatics/btv557.

Gaudet, P. & Dessimoz, C. (2017) Gene Ontology: Pitfalls, Biases, and Remedies. In: C. Dessimoz & N. Škunca (eds.). *The Gene Ontology Handbook*. Methods in Molecular Biology. New York, NY, Springer. pp. 189–205. doi:10.1007/978-1-4939-3743-1_14.

Gramates, L.S., Agapite, J., Attrill, H., Calvi, B.R., Crosby, M.A., Dos Santos, G., Goodman, J.L., Goutte-Gattat, D., Jenkins, V.K., Kaufman, T., Larkin, A., Matthews, B.B., Millburn, G., Strelets, V.B., & the FlyBase Consortium (2022) FlyBase: a guided tour of highlighted features. *Genetics*. 220 (4), iyac035. doi:10.1093/genetics/iyac035.

Greenwald, I.S., Sternberg, P.W. & Horvitz, H.R. (1983) The lin-12 locus specifies cell fates in Caenorhabditis elegans. *Cell*. 34 (2), 435–444. doi:10.1016/0092-8674(83)90377-x.

Gruner, M., Nelson, D., Winbush, A., Hintz, R., Ryu, L., Chung, S.H., Kim, K., Gabel, C.V. & van der Linden, A.M. (2014) Feeding state, insulin and NPR-1 modulate chemoreceptor gene expression via integration of sensory and circuit inputs. *PLoS genetics*. 10 (10), e1004707. doi:10.1371/journal.pgen.1004707.

Hagberg, A., Swart, P.J. & Schult, D.A. (2008) *Exploring network structure, dynamics, and function using NetworkX*. https://www.osti.gov/biblio/960616.

Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., et al. (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research*. 38 (Database issue), D463-467. doi:10.1093/nar/gkp952.

Ho, D.M., Pallavi, S.K. & Artavanis-Tsakonas, S. (2015) The Notch-mediated hyperplasia circuitry in Drosophila reveals a Src-JNK signaling axis. *eLife*. 4, e05996. doi:10.7554/eLife.05996.

Iser, W.B., Wilson, M.A., Wood, W.H., Becker, K. & Wolkow, C.A. (2011) Co-regulation of the DAF-16 target gene, cyp-35B1/dod-13, by HSF-1 in C. elegans dauer larvae and daf-2 insulin pathway mutants. *PloS One*. 6 (3), e17369. doi:10.1371/journal.pone.0017369.

Kaplan, R.E.W. & Baugh, L.R. (2016) L1 arrest, daf-16/FoxO and nonautonomous control of post-embryonic development. *Worm*. 5 (2), e1175196. doi:10.1080/21624054.2016.1175196.

Kennedy, L.M., Pham, S.C.D.L. & Grishok, A. (2013) Nonautonomous regulation of neuronal migration by insulin signaling, DAF-16/FOXO, and PAK-1. *Cell Reports*. 4 (5), 996–1009. doi:10.1016/j.celrep.2013.07.045.

Kopan, R. & Ilagan, M.X.G. (2009) The Canonical Notch Signaling Pathway: Unfolding the Activation Mechanism. *Cell*. 137 (2), 216–233. doi:10.1016/j.cell.2009.03.045.

Lee, M.-H., Mamillapalli, S.S., Keiper, B.D. & Cha, D.S. (2016) A systematic mRNA control mechanism for germline stem cell homeostasis and cell fate specification. *BMB reports*. 49 (2), 93–98. doi:10.5483/bmbrep.2016.49.2.135.

Marri, S. & Gupta, B.P. (2009) Dissection of lin-11 enhancer regions in Caenorhabditis elegans and other nematodes. *Developmental Biology*. 325 (2), 402–411. doi:10.1016/j.ydbio.2008.09.026.

Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., et al. (2024) The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*. 52 (D1), D672–D678. doi:10.1093/nar/gkad1025.

Murugesan, G., Abdulkadhar, S. & Natarajan, J. (2017) Distributed smoothed tree kernel for protein-protein interaction extraction from the biomedical literature. *PloS One*. 12 (11), e0187379. doi:10.1371/journal.pone.0187379.

Ookuma, S., Fukuda, M. & Nishida, E. (2003) Identification of a DAF-16 transcriptional target gene, scl-1, that regulates longevity and stress resistance in Caenorhabditis elegans. *Current biology: CB*. 13 (5), 427–431. doi:10.1016/s0960-9822(03)00108-8.

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K. & Tyers, M. (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Research*. 47 (D1), D529–D541. doi:10.1093/nar/gky1079.

Priess, J.R., Schnabel, H. & Schnabel, R. (1987) The glp-1 locus and cellular interactions in early C. elegans embryos. *Cell*. 51 (4), 601–611. doi:10.1016/0092-8674(87)90129-2.

Roy, A. & Palli, S.R. (2018) Epigenetic modifications acetylation and deacetylation play important roles in juvenile hormone action. *BMC genomics*. 19 (1), 934. doi:10.1186/s12864-018-5323-4.

Salwinski, L., Licata, L., Winter, A., Thorneycroft, D., Khadake, J., Ceol, A., Aryamontri, A.C., Oughtred, R., Livstone, M., Boucher, L., Botstein, D., Dolinski, K., Berardini, T., Huala, E., Tyers, M., Eisenberg, D., Cesareni, G. & Hermjakob, H. (2009) Recurated protein interaction datasets. *Nature Methods*. 6 (12), 860–861. doi:10.1038/nmeth1209-860.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*. 34 (Database issue), D535-539. doi:10.1093/nar/gkj109.

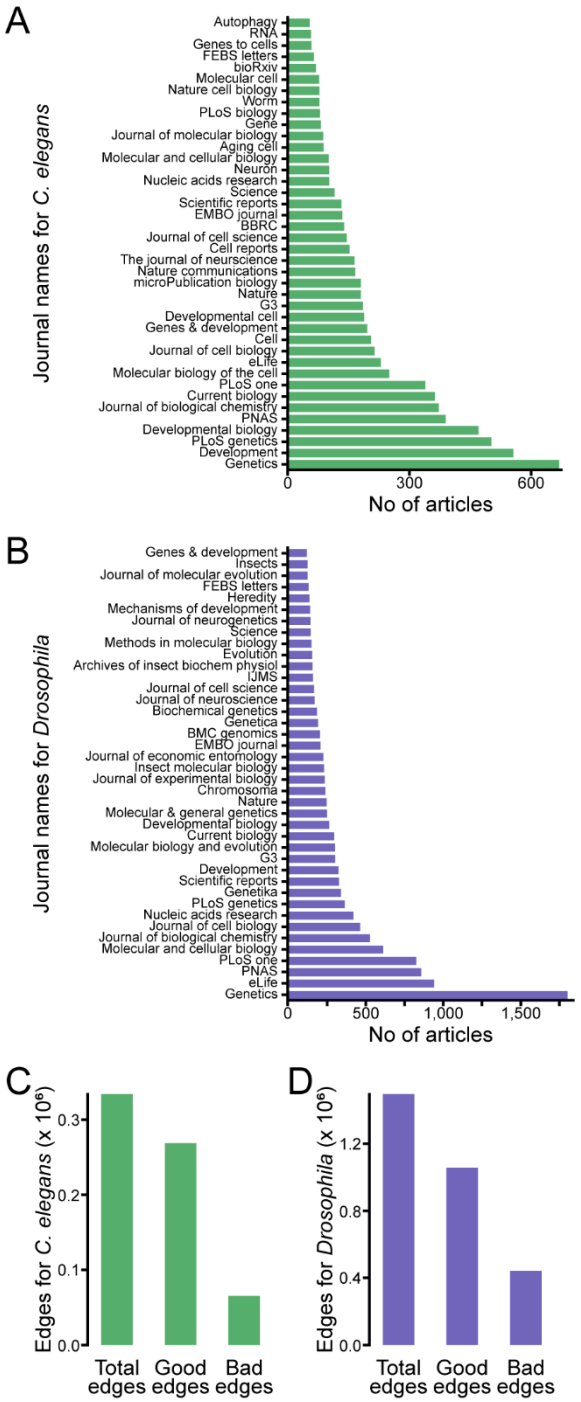
Suarez Rodriguez, F., Sanlidag, S. & Sahlgren, C. (2023) Mechanical regulation of the Notch signaling pathway. *Current Opinion in Cell Biology*. 85, 102244. doi:10.1016/j.ceb.2023.102244.

Sun, J., Smith, L., Armento, A. & Deng, W.-M. (2008) Regulation of the endocycle/gene amplification switch by Notch and ecdysone signaling. *The Journal of Cell Biology*. 182 (5), 885–896. doi:10.1083/jcb.200802084.

Wirick, M.J., Cale, A.R., Smith, I.T., Alessi, A.F., Starostik, M.R., Cuko, L., Lalk, K., Schmidt, M.N., Olson, B.S., Salomon, P.M., Santos, A., Schmitter-Sánchez, A., Galagali, H., Ranke, K.J., Wolbert, P.A., Knoblock, M.L., Kim, J.K. & Karp, X. (2021) daf-16/FOXO blocks adult cell fate in Caenorhabditis elegans dauer larvae via lin-41/TRIM71. *PLoS genetics*. 17 (11), e1009881. doi:10.1371/journal.pgen.1009881.

472 Zhou, D., Stobdan, T., Visk, D., Xue, J. & Haddad, G.G. (2021) Genetic interactions regulate hypoxia tolerance
473 conferred by activating Notch in excitatory amino acid transporter 1-positive glial cells in Drosophila
474 melanogaster. *G3 (Bethesda, Md.)*. 11 (2), jkab038. doi:10.1093/g3journal/jkab038.

475 **FIGURES AND LEGENDS**



476
477
478 **Figure 1. Comprehensive meta-analysis of full-text articles in *C. elegans* and *Drosophila***
479 (**A**-**B**) Quantitative distribution of articles from the top 40 journals featuring *C. elegans*- (**A**) and *Drosophila*-
480 related (**B**) research.
481 (**C**-**D**) Profile of *C. elegans* (**C**) and *Drosophila* (**D**) interaction data showing total, validated (good), and
482 erroneous (bad) edges (in millions).

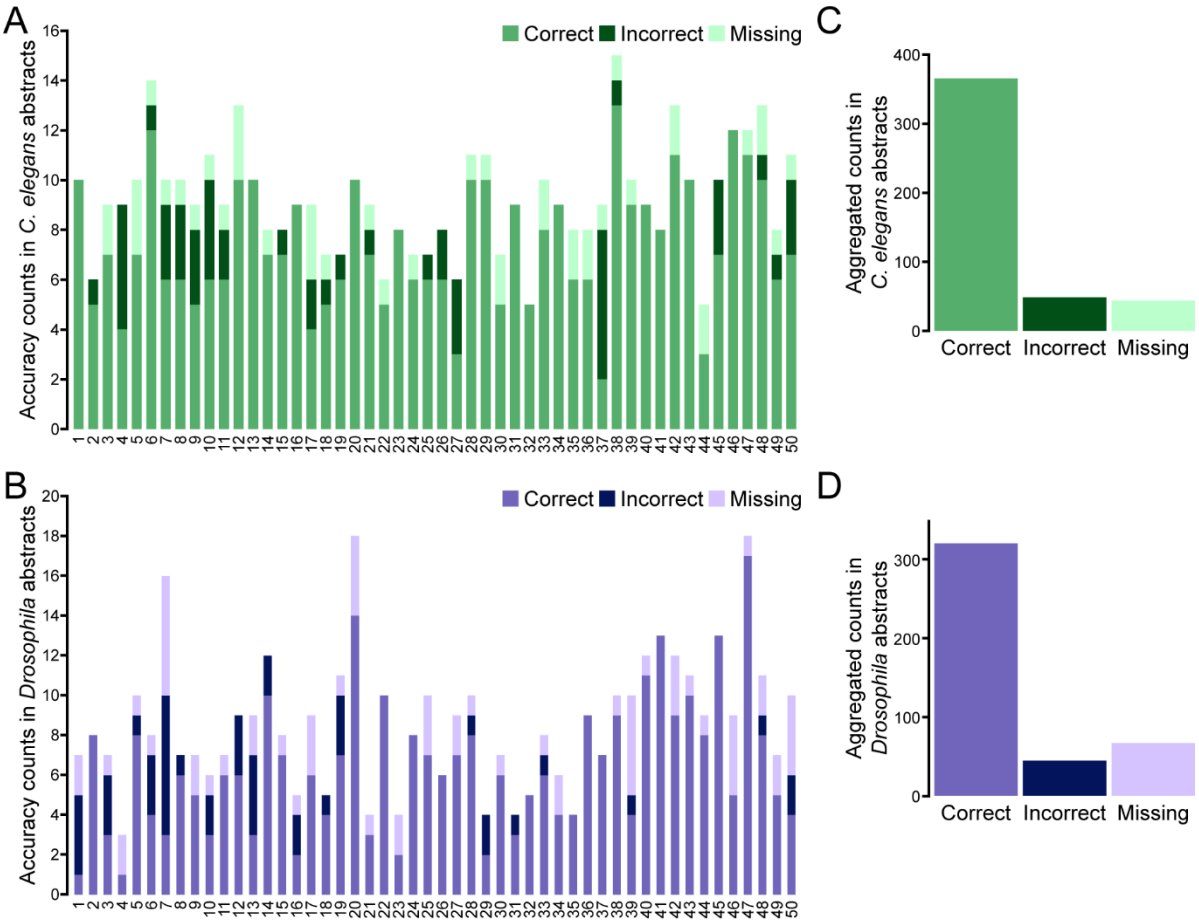


Figure 2. Accuracy assessment of GPT-processed abstracts for model organisms
(**A**-**B**) Analysis accuracy for 50 *C. elegans* (**A**) and 50 *Drosophila* (**B**) abstracts, categorized by the incidence of correct, incorrect, and missing information.
(**C**-**D**) Aggregated counts of accurate (correct), erroneous (incorrect), and overlooked (missing) statements from *C. elegans* (**C**) and *Drosophila* (**D**) abstracts shown in (A) and (B), respectively.

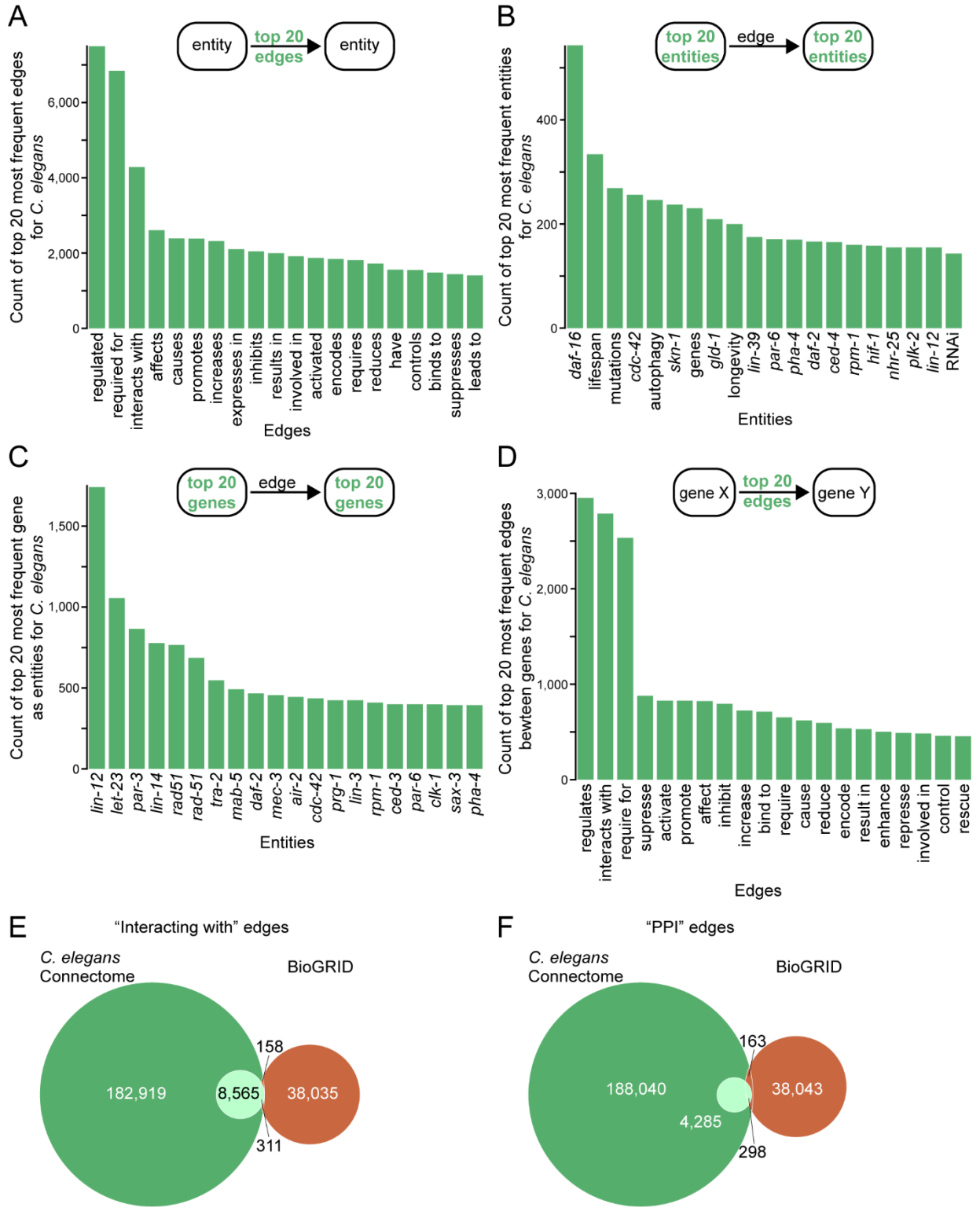


Figure 3. Network analysis highlights in *C. elegans* Connectome

(**A**) Frequency distribution of the top 20 interaction edges in the *C. elegans* connectome.
(**B**) Frequency distribution of the top 20 interaction entities in the *C. elegans* connectome.
(**C**) Frequency distribution of the top 20 interaction entities as genes in the *C. elegans* connectome.
(**D**) Frequency distribution of the top 20 interaction edges between genes in the *C. elegans* connectome.
(**E-F**) Venn diagram illustrating the commonalities "interacting with" (**E**) and protein-protein interaction (PPI) (**F**) edges in the *C. elegans* Connectome and the interacting proteins listed in BioGRID.

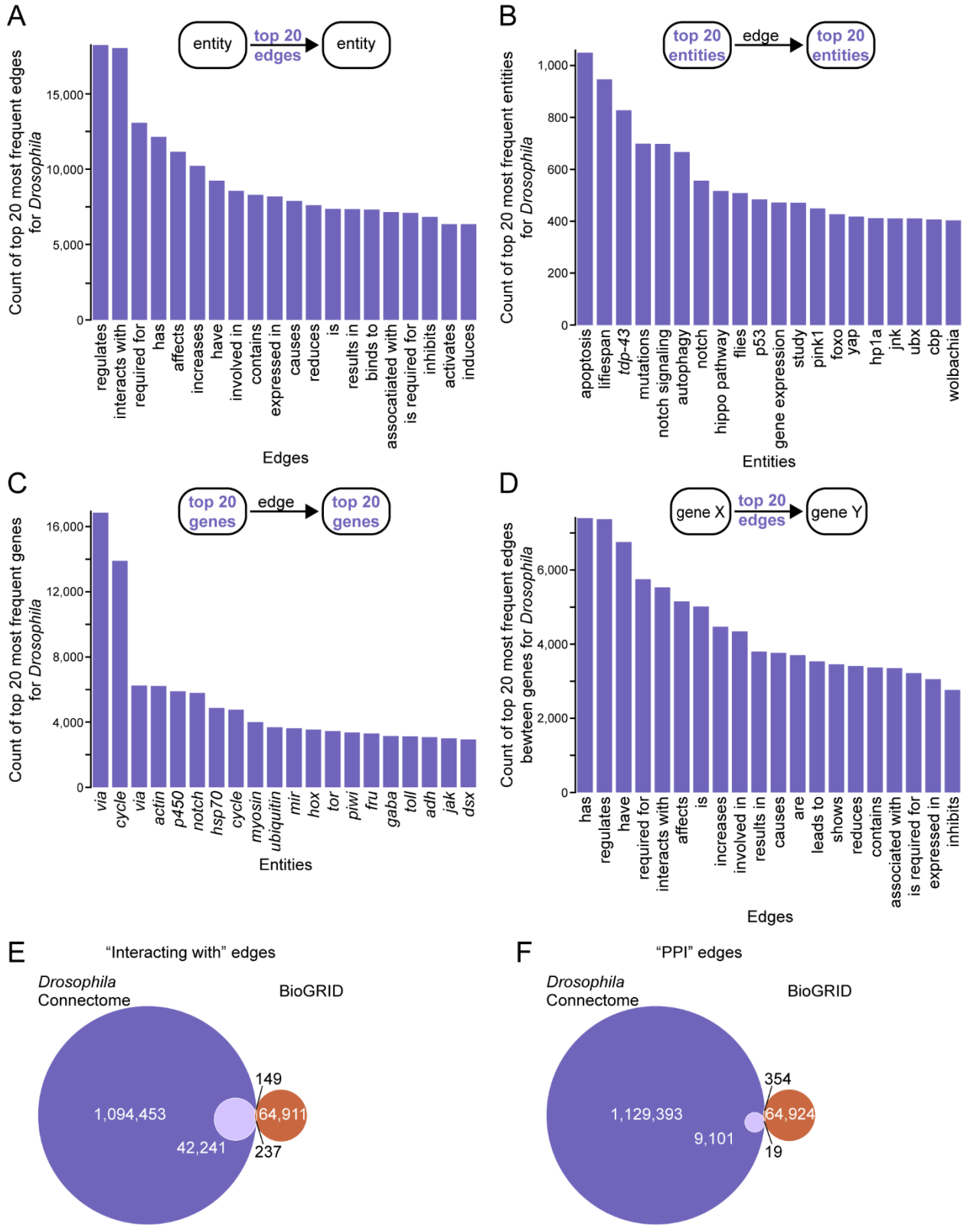


Figure 4. Network analysis highlights in *Drosophila* connectome

(**A**) Frequency distribution of the top 20 interaction edges in the *Drosophila* connectome.
(**B**) Frequency distribution of the top 20 interaction entities in the *Drosophila* connectome.
(**C**) Frequency distribution of the top 20 interaction entities as genes in the *Drosophila* connectome.
(**D**) Frequency distribution of the top 20 interaction edges between genes in the *Drosophila* connectome.
(**E-F**) Venn diagram illustrating the commonalities "interacting with" (**E**) and protein-protein interaction (PPI)
(**F**) edges in the *Drosophila* Connectome and the interacting proteins listed in BioGRID.

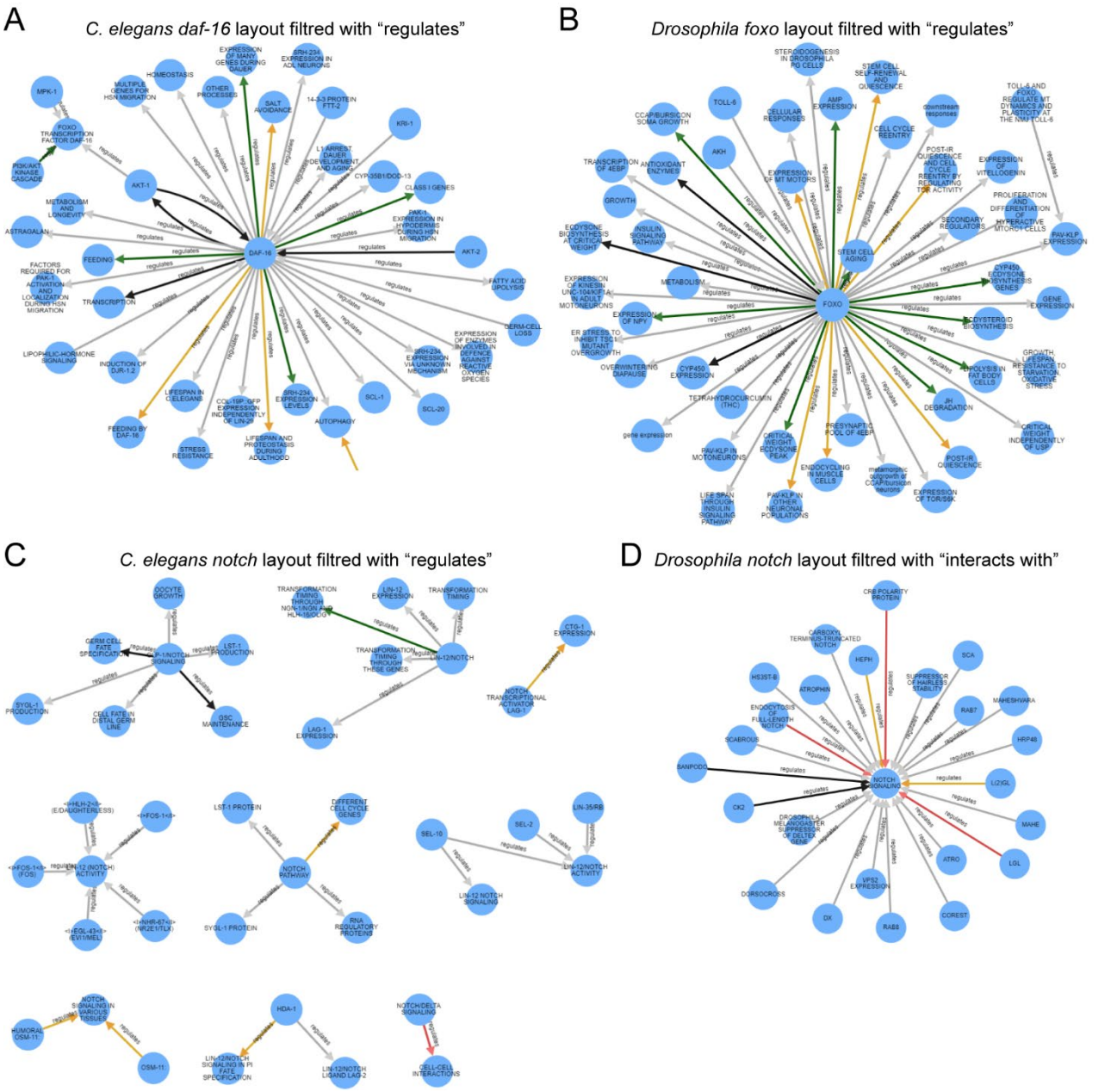
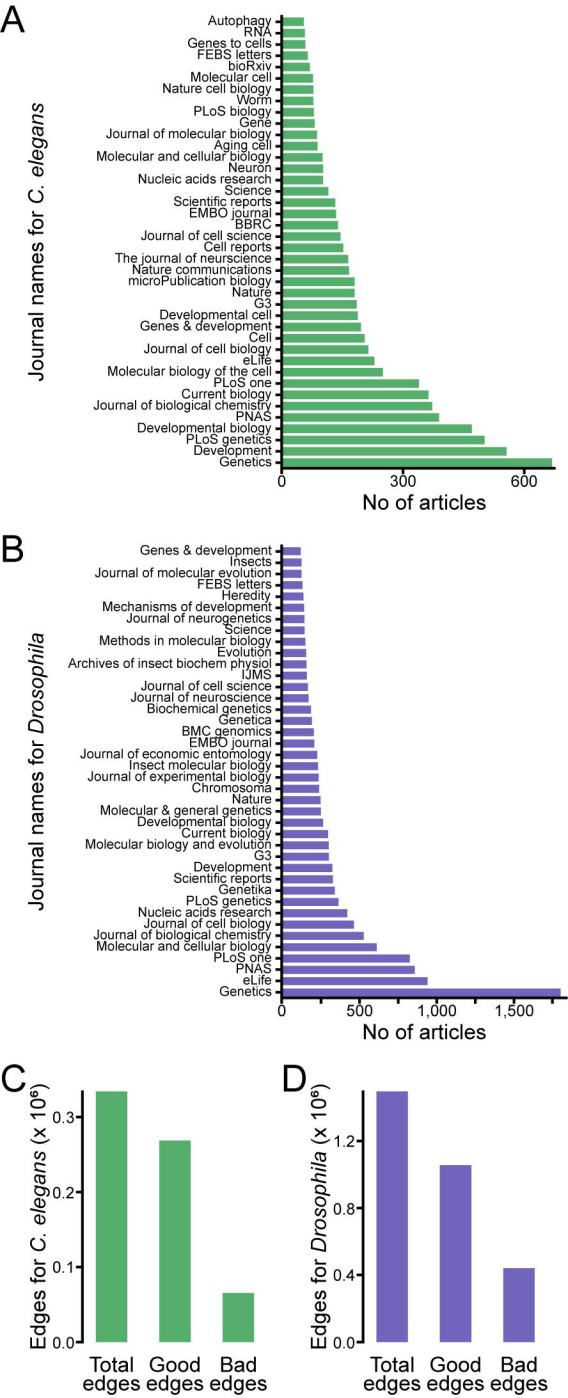
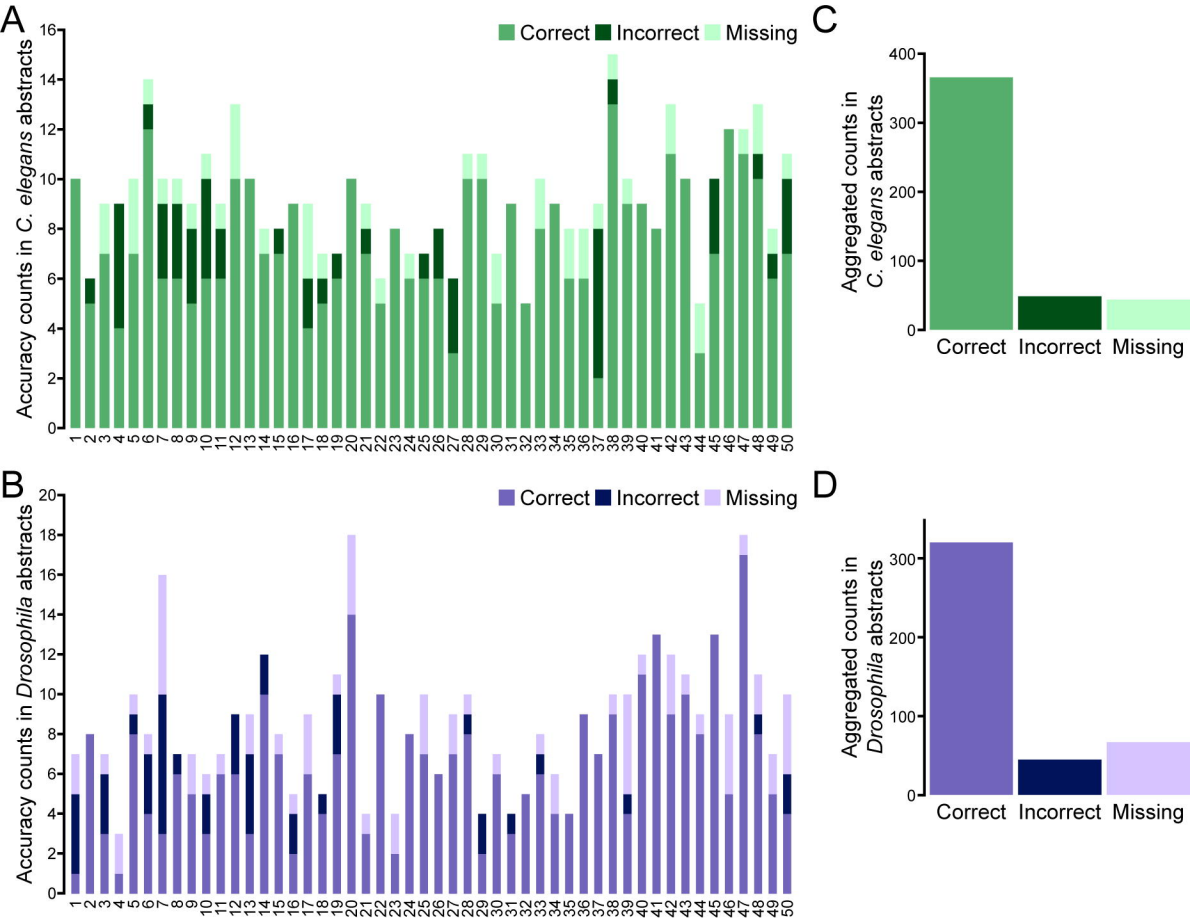
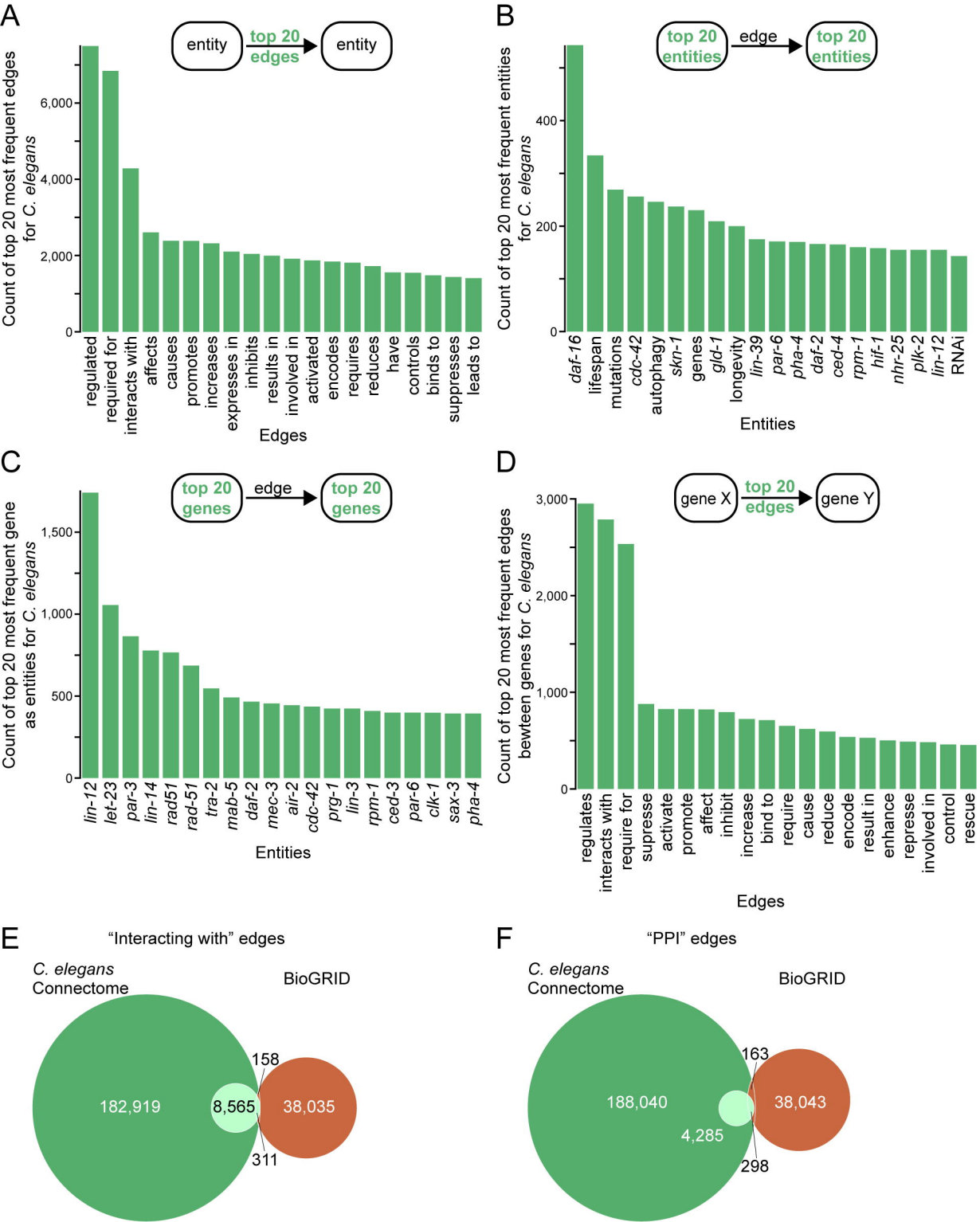
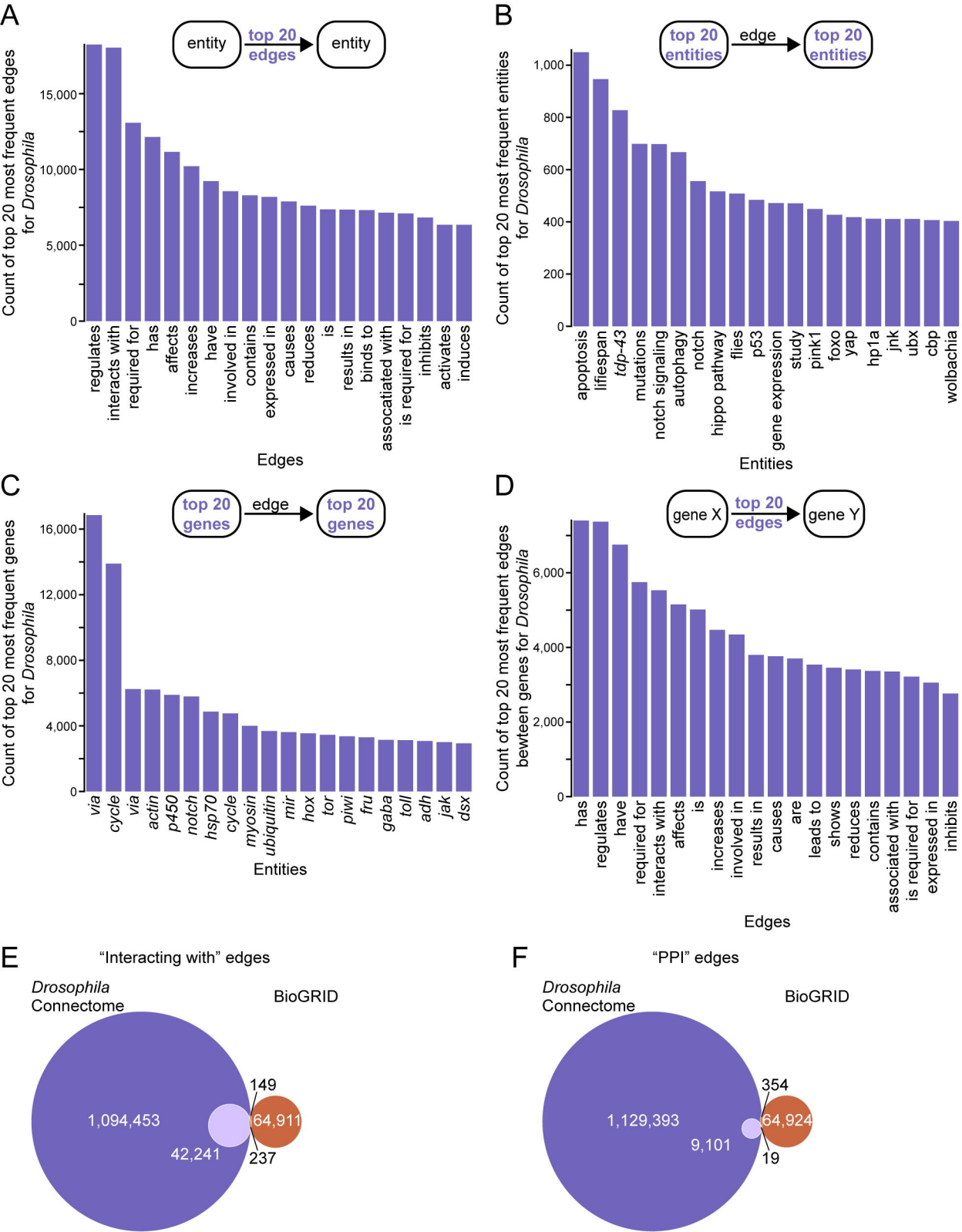


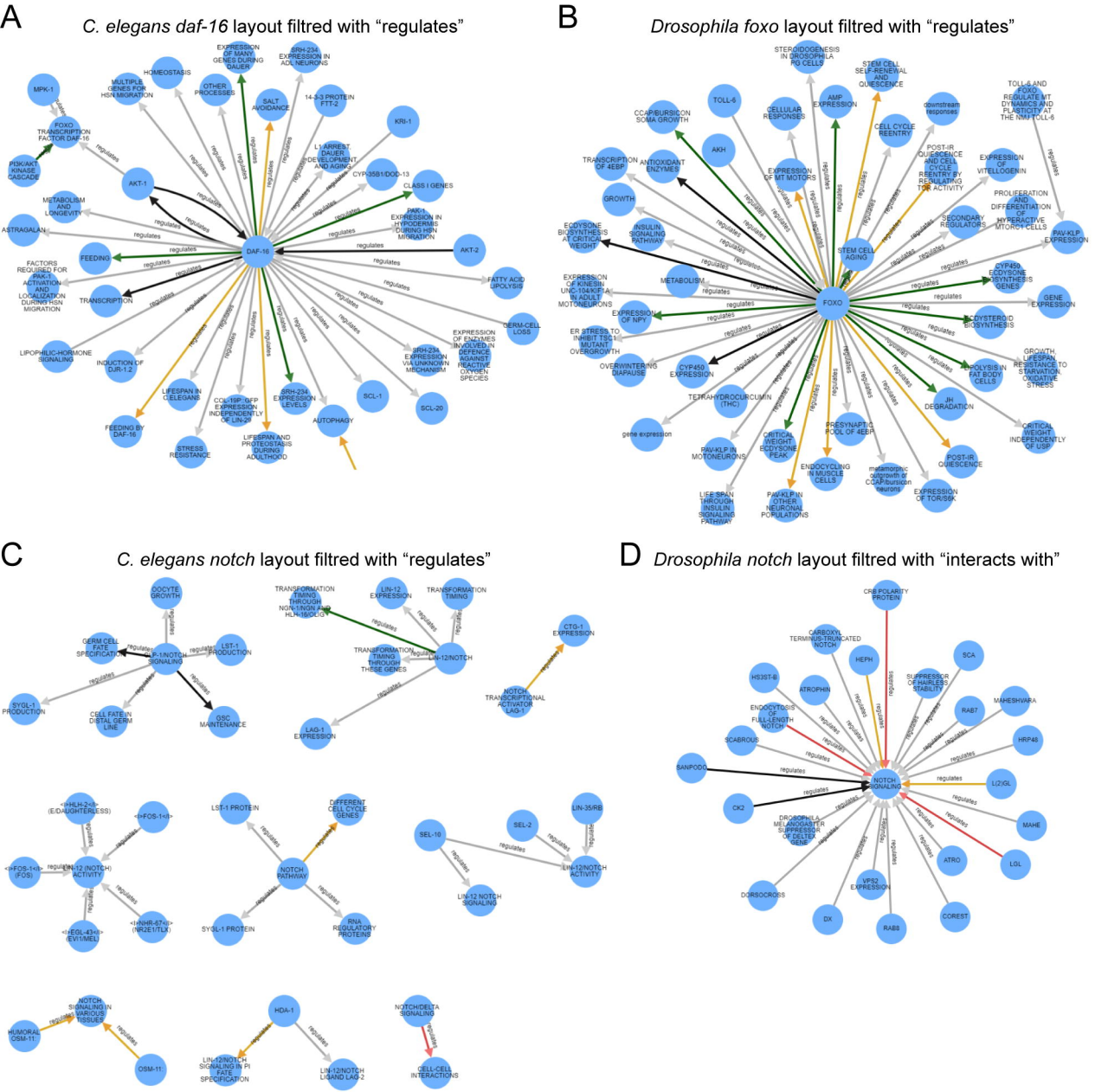
Figure 5. Connectome visualization for key genes in *C. elegans* and *Drosophila*. (**A**) Knowledge network of *C. elegans* gene *daf-16* with Layout Options "regulates". (**B**) knowledge network of *Drosophila* gene *foxo* Layout Options "regulates". (**C**) Knowledge network of *C. elegans* "notch" genes with Layout Options "regulates". (**D**) Knowledge network of *Drosophila* gene *notch* with Layout Options "interacts with".


A
Journal names for C. elegans
Autophagy
RNA
Genes to cells
FEBS letters
bioRxiv
Molecular cell
Nature cell biology
Worm
PLoS biology
Gene
Journal of molecular biology
Aging cell
Molecular and cellular biology
Neuron
Nucleic acids research
Science
Scientific reports
EMBO journal
BBRC
Journal of cell science
Cell reports
The journal of neurscience
Nature communications
microPublication biology
Nature
G3
Developmental cell
Genes & development
Cell
Journal of cell biology
eLife
Molecular biology of the cell
PLoS one
Current biology
Journal of biological chemistry
PNAS
Developmental biology
PLoS genetics
Development
Genetics
0
300
600
No of articles
B
Journal names for Drosophila
Genes & development
Insects
Journal of molecular evolution
FEBS letters
Heredity
Mechanisms of development
Journal of neurogenetics
Science
Methods in molecular biology
Evolution
Archives of insect biochem physiol
IJMS
Journal of cell science
Journal of neuroscience
Biochemical genetics
Genetica
BMC genomics
EMBO journal
Journal of economic entomology
Insect molecular biology
Journal of experimental biology
Chromosoma
Nature
Molecular & general genetics
Developmental biology
Current biology
Molecular biology and evolution
G3
Development
Scientific reports
Genetika
PLoS genetics
Nucleic acids research
Journal of cell biology
Journal of biological chemistry
Molecular and cellular biology
PLoS one
PNAS
eLife
Genetics
0
500
1,000
1,500
No of articles
C
Edges for C. elegans (x 10^6)
0.0
0.1
0.2
0.3
Total edges
Good edges
Bad edges
D
Edges for Drosophila (x 10^6)
0.0
0.4
0.8
1.2
Total edges
Good edges
Bad edges


A
Correct
Incorrect
Missing
Accuracy counts in C. elegans abstracts
0
2
4
6
8
10
12
14
16
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
C
Aggregated counts in C. elegans abstracts
0
100
200
300
400
Correct
Incorrect
Missing
B
Correct
Incorrect
Missing
Accuracy counts in Drosophila abstracts
0
2
4
6
8
10
12
14
16
18
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
D
Aggregated counts in Drosophila abstracts
0
100
200
300
Correct
Incorrect
Missing


A
entity
top 20 edges
entity
Count of top 20 most frequent edges for C. elegans
6,000
4,000
2,000
0
regulated
required for
interacts with
affects
causes
promotes
increases
expresses in
inhibits
results in
involved in
activated
encodes
requires
reduces
have
controls
binds to
suppresses
leads to
Edges
B
top 20 entities
edge
top 20 entities
Count of top 20 most frequent entities for C. elegans
400
200
0
daf-16
lifespan
mutations
cdc-42
autophagy
skn-1
genes
gld-1
longevity
lin-39
par-6
pha-4
daf-2
ced-4
rpm-1
hif-1
nhr-25
plk-2
lin-12
RNAi
Entities
C
top 20 genes
edge
top 20 genes
Count of top 20 most frequent gene as entities for C. elegans
1,500
1,000
500
0
lin-12
let-23
par-3
lin-14
rad51
rad-51
tra-2
mab-5
daf-2
mec-3
air-2
cdc-42
prg-1
lin-3
rpm-1
ced-3
par-6
clk-1
sax-3
pha-4
Entities
D
gene X
top 20 edges
gene Y
Count of top 20 most frequent edges bewteen genes for C. elegans
3,000
2,000
1,000
0
regulates
interacts with
require for
supresse
activate
promote
affect
inhibit
increase
bind to
require
cause
reduce
encode
result in
enhance
represse
involved in
control
rescue
Edges
E
"Interacting with" edges
C. elegans Connectome
BioGRID
182,919
8,565
158
311
38,035
F
"PPI" edges
C. elegans Connectome
BioGRID
188,040
4,285
163
298
38,043


A
entity
top 20 edges
entity
Count of top 20 most frequent edges for Drosophila
15,000
10,000
5,000
0
regulates
interacts with
required for
has
affects
increases
have
involved in
contains
expressed in
causes
reduces
is
results in
binds to
assocatiated with
is required for
inhibits
activates
induces
Edges
B
top 20 entities
edge
top 20 entities
Count of top 20 most frequent entities for Drosophila
1,000
800
600
400
200
0
apoptosis
lifiespan
tdp-43
mutations
notch signaling
autophagy
notch
hippo pathway
flies
p53
gene expression
study
pink1
foxo
yap
hp1a
jnk
ubx
cbp
wolbachia
Entities
C
top 20 genes
edge
top 20 genes
Count of top 20 most frequent genes for Drosophila
16,000
12,000
8,000
4,000
0
via
cycle
via
actin
p450
notch
hsp70
cycle
myosin
ubiquitin
mir
hox
tor
piwi
fru
gaba
toll
adh
jak
dsx
Entities
D
gene X
top 20 edges
gene Y
Count of top 20 most frequent edges bewteen genes for Drosophila
6,000
4,000
2,000
0
has
regulates
have
required for
interacts with
affects
is
increases
involved in
results in
causes
are
leads to
shows
reduces
contains
associated with
is required for
expressed in
inhibits
Edges
E
"Interacting with" edges
Drosophila Connectome
BioGRID
1,094,453
42,241
149
64,911
237
F
"PPI" edges
Drosophila Connectome
BioGRID
1,129,393
9,101
354
64,924
19


A
C. elegans daf-16 layout filtred with “regulates”
B
Drosophila foxo layout filtred with “regulates”
C
C. elegans notch layout filtred with “regulates”
D
Drosophila notch layout filtred with “interacts with”