

Hub connectivity and gene expression in the *C. elegans* connectome



Aurina Arnatkevičiūtė¹*, Ben D. Fulcher¹, Roger Pocock, Taqi Ani, Alex Fornito¹*

1 Brain and Mental Health Laboratory, Monash Institute of Cognitive and Clinical Neurosciences, School of Psychological Sciences, Monash University, 770 Blackburn Rd, Clayton, 3168, VIC, Australia.

* These authors contributed equally to this work.

* alex.fornito@monash.edu

Abstract

WE ♥ CAENORHABDITIS ELEGANS

Author summary

WE ♥ WORMS

Introduction

Neuronal connectivity provides the substrate for integrated brain function. Recent years have seen several systematic and large-scale attempts to generate comprehensive wiring diagrams – connectomes – of nervous systems in species as diverse as the nematode worm *Caenorhabditis elegans* [1,2], the drosophila fruit fly [3,4], zebrafish [5,6], mouse [7,8], rat [[CITE-Swanson]], cat [[CITE-Young]], macaque [?,9] and human [?,10]. One of the most striking findings to emerge from analyses of these diverse data, acquired using different measurement techniques and at resolution scales ranging from nm to mm, is of a strong conservation of certain, non-random topological properties of network organization (reviewed in [?, ?,11,12]; [[CITE-Schroeter et al NRN]]; see also [13]). These properties include a modular organization, such that subsets of functionally related (and usually spatially adjacent) elements are densely interconnected with each other; a hierarchy of modules, such that modules contain nested sub-modules and so on over multiple scales ([14,15]); economical connectivity, such that wiring costs (typically measured in terms of wiring length) are near-minimal given the topological complexity of the system [16,17]; a heterogeneous distribution of connections across network nodes, such that most nodes possess relatively few connections and a small proportion of nodes has a very high degree of connectivity, marking these nodes as network hubs [18,19]; and stronger interconnectivity between hub nodes than expected by chance, leading to the emergence of a so-called rich-club (i.e., the hubs are rich in terms of their connections and they form a club because they are densely interconnected with each other) [4,20–23], features that can only partially be accounted for by spatial constraints [24–26].

Hub connectivity and rich-club organization in particular is thought to play a central role in integrating disparate functional systems in neuronal networks [21,27–30]. Hub nodes typically participate in more than one functional system, or topological

module [21, 22, 31], placing them in a perfect position to integrate diverse neural processes. Experimental data and computational modeling also indicates that hub nodes, and the connections between them, are topologically positioned to mediate a high volume of signal traffic [32–35]. This integrative role comes at a cost however, with connections between hubs extending over longer distances, on average, than other types of connections, a finding reported in the human [32], mouse [36], and nematode [23]. Human positron emission tomography also suggests that hub nodes consume greater metabolic resources and have higher levels of blood flow than other areas [37–39]. This high metabolic cost may underlie the involvement of hub regions in a broad array of human diseases [?, 12, 40].

~~Work linking the large-scale topological organization of brain networks to microscale measure of gene expression has suggested that this high metabolic cost of hub connectivity is associated with a distinct metabolic signature.~~ Fulcher and Fornito [36] combined viral tract tracing data on connectivity between 213 regions of the right hemisphere of the mouse brain [7] with in situ hybridization measures of the expression of 17 642 genes in each of those regions. They found that transcriptional coupling, across the entire genome, was stronger for connected compared to unconnected brain regions and that, in general, this coupling decayed sharply with increasing anatomical distance between brain regions. In stark contrast to this general trend however, was the finding that connected pairs of hubs (i.e., the “rich club” of the brain) showed the highest levels of transcriptional coupling (compared to hub-non-hub and non-hub pairs), despite these regions being separate by longer anatomical distances, on average. Moreover, this coupling was driven predominantly by coupled expression of genes regulating the oxidative synthesis and metabolism of adenosine triphosphate (ATP), the primary energetic currency of neuronal signaling [41, 42]. Vertes et al. [43] later found evidence of a similar transcriptional signature of hub connectivity in humans. Using regional microarray gene expression data of 20 737 genes through 285 cortical areas and applying a partial least squares method to reduce the dimensionality they found that topologically integrative hubs that mediate long distance connections have transcriptional profiles enriched in oxidative metabolism and mitochondria, confirming the hypothesis of their high energy demands.

Together, the analyses of Fulcher and Fornito [36] and Vertes et al. [43], which were performed using measures of mesoscale (μm to mm) and macroscale (mm to cm) connectivity, respectively, suggest that the molecular signature of hub connectivity, which is characterized by elevated coupling of the expression of genes regulating oxidative metabolism, is conserved across species and resolution scales. Here, we sought to test this hypothesis by examining the link between gene expression and microscale connectivity in the nematode worm *C. elegans*. The connectome of this animal, which comprises 302 [[WE HAVE 279 NEURONS AND A CORRESPONDING NUMBER OF CONNECTIONS, BUT THE WHOLE CONNECTOME IS 302 how should we phrase??? WE DON'T HAVE A NUMBER FOR all connections for 302 neurons; 279 neurons 2990 connections]] neurons and 2990 connections, is the only species to have its connectome mapped almost completely at the level of individual neurons and synapses using electron microscopy (EM) [1, 2]. We combined these data on neuronal connectivity with curated information on the binary expression patterns of 948 genes across neurons to examine the relationship between gene expression and the large-scale topological organization of the nematode nervous system. We also took advantage of detailed information on neuron spatial positions, birth times, neuronal lineage, and functional and chemical composition of neurons to understand the mechanisms through which gene expression might influence network topology. We show that [[XXX]].

Materials and methods

We first describe the neural connectivity data used to construct a connectome for *C. elegans* and the methods used to quantify network topology and specific additional properties of neurons (including their neurochemical composition and lineage relationships). We then go on to describe the gene expression data, how we measure expression similarity between pairs of neurons, and our method for scoring genes to perform functional enrichment analysis.

Neuronal connectivity data

The nematode *C. elegans* nervous system is made up of 302 neurons, divided into the pharyngeal nervous system (20 neurons) and the somatic nervous system (282 neurons). The spatial positions of neurons, and their interconnections, are genetically determined and highly reproducible across organisms [44], and do not change through time [?]. Neuronal connectivity data for the adult hermaphrodite *C. elegans* was first mapped by White et al. [1] through detailed electron microscopy, and then revised by Chen et al. [45] and Varshney et al. [2]. Here we analyze the larger somatic nervous system using data from Varshney et al. [2], who mapped connectivity between 279 neurons (282 somatic neurons, excluding CANL/R, and VC6, which do not form synapses with other neurons), downloaded from WormAtlas [46] (www.wormatlas.org/neuronawiring.html#NeuronalconnectivityII).

Connectivity information is available both for electrical gap junctions and chemical synapses. The directionality of electrical gap junctions is unknown (and have previously been represented as bidirectional connections), whereas chemical connectivity information includes presynaptic and postsynaptic neurons (providing information about the connectivity direction) and synapse number (providing information about the projection strength, or connectivity weight). Previous work on *C. elegans* neuronal connectivity has used different processed versions of this data, including: (i) analyzing just synaptic connections [47], (ii) a combination of synaptic connections and gap junctions together [48,49], (iii) a combined, symmetrized connectome that represents unidirectional and reciprocal synaptic connections equivalently [23,50,51], or (iv) comparing a range of connectome representations [52]. In order to account for the influence of both electrical and chemical connectivity, our analysis focuses on the combined directed, binary network that includes both types of connections, with gap junctions represented as bidirectional connections and chemical synapses represented as directed edges from the presynaptic neuron to the postsynaptic neuron. This network of 279 neurons consists of 2287 connected pairs of neurons, spanning a total of 2990 directed connections. Note that the qualitative results presented here are not highly sensitive to the connectome representation, and, e.g., hold when the directed chemical connectivity network is analyzed (excluding gap junctions).

We also assembled a range of data characterizing additional properties of *C. elegans* neurons. Two dimensional spatial coordinates for neurons were obtained as `celegans277.mat` from www.biological-networks.org/?page_id=25 [53]. Coordinates for three neurons (AIBL, AIYL, SMDVL) were missing in this dataset, and were reconstructed by assigning identical coordinates to the corresponding bilateral neurons (AIBR, AIYR, SMDVR) [54]. Neurons were also labeled based on their anatomical location, as: (i) ‘head’, (ii) ‘tail’, or (iii) ‘body’ using data from release WS256 of WormBase [55], (ftp://ftp.wormbase.org/pub/wormbase/releases/WS256/ONTOLOGY/anatomy_association.WS256.wb). These annotations were assigned to individual neurons using the anatomical hierarchy defined in WormBase, which we retrieved using the WormBase python API (WormMine: `intermine.wormbase.org`) [55], propagating each term down the hierarchy to

individual neurons. We manually corrected the assignment of twelve head neurons (ALA, AVFL, AVFR, AVG, RIFL, RIFR, RIGL, RIGR, SABD, SABVL, SABVR, SMDVL), which were assigned as ‘head’ in WormAtlas [46] but not on WormBase. Neurons were labeled by their broad functional type as either a: (i) ‘sensory’ neuron (support receptive function), (ii) ‘motor’ neuron (contain neuromuscular junctions), or (iii) ‘interneuron’ (all other neurons) [1]. A total of 79 neurons are annotated as sensory, 121 annotated as motor, and 97 annotated as interneurons (including eighteen neurons assigned to two categories). The neurotransmitter systems used by each neuron was labeled by matching to data in Table 2 of Pereira et al. [56]. Lineage similarity for all pairs of neurons was obtained from previously published embryonic and post-embryonic lineage trees [57,58]; data were downloaded from WormAtlas (<http://www.wormatlas.org/neuronalwiring.html#Lineageanalysis>) [46]. In this dataset, the closest common ancestor neuron was identified for each pair of neurons, and then the number of cell divisions from the common progenitor was calculated as the lineage distance.

Network analysis

In this section we describe the network methods used to characterize the *C. elegans* connectome.

Degree Neuronal connectivity is most simply quantified by counting the number of neurons that a given neuron projects to, known as its *out-degree*, k_{out} , or by counting the number of neurons that a given neuron receives projections from, known as its *in-degree*, k_{in} . These quantities can be summed to give the total number of connections involving a given neuron, its *degree*, $k_{\text{tot}} = k_{\text{in}} + k_{\text{out}}$, which we refer to as simply the degree, k throughout this work. At a given degree threshold, k , we classified each neuron as either a ‘hub’ (degree $> k$) or a ‘non-hub’ (degree $\leq k$). All connections were subsequently classified in terms of their source and target neurons, as either ‘rich’ (hub \rightarrow hub, or hub \leftrightarrow hub), ‘feed-in’ (non-hub \rightarrow hub), ‘feed-out’ (hub \rightarrow non-hub), or ‘peripheral’ (nonhub \rightarrow nonhub, or nonhub \leftrightarrow nonhub).

Rich-club organization We used the rich-club coefficient, $\phi(k)$, to quantify the interconnectivity of hub neurons:

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k} - 1)}, \quad (1)$$

where $N_{>k}$ is the number of nodes with degree $> k$, and $E_{>k}$ is the number of edges between them [59]. Thus, $\phi(k)$ measures the link density in the subgraph containing nodes with degree $> k$. Because $\phi(k)$ increases with k for general networks (as nodes with higher degree make more connections, yielding a higher expected link density in the subgraph containing nodes with degree $> k$), we compared $\phi(k)$ measured from the *C. elegans* connectome to that of randomized null networks, $\phi_{\text{rand}}(k)$ [59]. An ensemble of 1 000 null networks was generated by shuffling the links in the empirical network while retaining the degree distribution of nodes in the network [60] (rewiring each edge an average of 50 times per null network) using the `randmio.dir` function from the *Brain Connectivity Toolbox* [61]. The normalized rich-club coefficient, $\Phi_{\text{norm}}(k)$, was then computed as the ratio of the rich-club coefficient of the empirical network to the mean rich-club coefficient of the ensemble of randomized networks:

$$\Phi_{\text{norm}}(k) = \frac{\phi(k)}{\langle \phi_{\text{rand}}(k) \rangle}. \quad (2)$$

$\Phi_{\text{norm}} > 1$ indicates rich-club organization of the network, with statistically significant deviations assessed by computing a p -value directly from the empirical null distribution, $\phi_{\text{rand}}(k)$, as a permutation test [20].

Modularity  The modular structure of the connectome was determined applying the Louvain community detection algorithm [62], using the `community.louvain` function implemented in the *Brain Connectivity Toolbox* [63]. To identify the optimal modular assignment, neurons were assigned to modules using consensus clustering  across 1 000 repeats of the Louvain algorithm [64], weighting each partition by its modularity, Q , using the `agreement_weighted` and `consensus_und` functions. We also compared results to a previously reported nine-module partition of neurons derived from an Erdős-Rényi Mixture Model (ERMM) [65].

Gene expression

Neuronal gene expression is measured here as a binary indicator, using data from many individual experiments compiled  to a unified data resource on WormBase [55], indicating which genes are expressed in a given neuron. We use release WS256 of this dataset (ftp://ftp.wormbase.org/pub/wormbase/releases/WS256/ONTOLOGY/anatomy_association.WS256.wb) and analyze annotations made ‘directly’ to individual neurons, excluding ‘uncertain’ annotations (see Supplementary Information for details). We denote the expression of a gene in a neuron using a ‘1’,  but are unable to distinguish: (i) “gene is not expressed” from (ii) “there is no information on whether gene is expressed”, both of which are represented as a ‘0’. Expression data is sparse, in part due to incomplete annotation data—an average of 30 genes are expressed in each neuron (range: 3 to 138 genes), and each gene is expressed in an average of 9 neurons (range: 1 to 148 neurons). A total of 948 genes are expressed in at least one neuron, allowing us to represent the full expression dataset as a binary 279×948 (neuron) \times 948 (genes) matrix, shown in Fig. 1B.

Gene coexpression

Our primary aim in this work is to understand how patterns  of expression across 948 genes  (e.g., rows of Fig. 1B) are related to pairwise patterns of synaptic  connectivity. To  relate our measure of gene expression for individual neurons to measures of coupled expression for pairs of neurons, we required a similarity measure for pairs of neurons that captures the probability of patterns of binary gene expression matching, or *gene coexpression*. We used a binary analogue of the linear Pearson correlation coefficient, the mean square contingency coefficient, r_ϕ [66]:

$$\text{ } \quad r_\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet0}n_{\bullet1}}}, \quad (3)$$

for two vectors x , y , of length $L (= 948)$, where n_{xy} counts the number of observations of each of the four outcomes (e.g., $n_{10} = \sum_i \delta_{x_i,1}\delta_{y_i,0}$ counts the number of times $x = 1$ and $y = 0$ across the length of x and y), while the symbol \bullet sums across a given variable (e.g., $n_{\bullet0} = \sum_i \delta_{y_i,0}$ counts the number of times $y = 0$). A maximum value $r_\phi = 1$ when x and y are identical (such that $n_{11} + n_{00} = L$), and a minimum value $r_\phi = -1$ when x and y are always mismatched (such that $n_{10} + n_{01} = L$). We  found that r_ϕ was not biased by differences in relative number of expressed genes in a neuron, which ranged from 3 (0.3% of 948 genes analyzed here) to 138 (14.6%). Note that other binary correlation measures, including Yule’s Q and Jaccard, exhibited strong bias and were therefore unsuitable (see Supplementary Information).

The 92 bilateral pairs of neurons (e.g., AVAL/AVAR, CEPVL/CEPVR, etc.) mostly exhibit highly correlated gene expression patterns, with all bilateral pairs exhibiting $r_\phi > 0.8$, and 96% of pairs exhibiting $r_\phi > 0.95$. To ensure that our results are not driven by the high coexpression between bilateral pairs of neurons, we excluded coexpression values of bilateral pairs of neurons from all coexpression analyses.

Given that r_ϕ treats mutual gene expression (n_{11}) the same as mutual absence of gene expression (n_{00}), we developed a new analytic measure of the probability of mutual gene expression, given its clearer biological interpretation. This measure was useful for validating and interpreting our results, and for motivating our gene enrichment scoring method (described in the following section), however, since it yields similar qualitative outputs to r_ϕ on our data, we quote r_ϕ values (which are easily interpretable as an analogue of the conventional correlation coefficient) throughout this work.

Gene enrichment analysis

We examined the contribution of specific functional sets of genes to patterns of coexpression between neurons. Compared to previous studies of gene functional enrichment [36, 43, 67], performing similar analyses with *C. elegans* data pose additional challenges due to: (i) binary expression data (making robust quantification difficult), (ii) sparse and incomplete data, posing statistical problems for quantifying a signal in genes with limited expression, and (iii) low genome coverage (less than 5% of the protein coding genes in *C. elegans*), limiting our ability to test many relevant Gene Ontology (GO) categories.

Given the challenges posed by the data, we developed new methods to quantify the contribution of individual genes, and functional groups of genes, to patterns of coupled expression (e.g., differences between connected *vs* unconnected, between electrical and chemical synapses, and for rich and feeder connections *versus* peripheral connections), in analogy to previous methods developed for continuous gene expression data [36]. To this end, we first scored each gene for whether it is more likely to be expressed in a given class of neuron pair over another as the probability of obtaining at least as many matches (defined as expression in both neurons of a pair) as observed under a random expression null model using the binomial distribution:

$$p^{(a)} = 1 - \sum_{i=0}^{m-1} \binom{n}{i} p_{\text{class}}^i (1 - p_{\text{class}})^{n-i}, \quad (4)$$

where m is the number of matches on the class of neuron pairs of interest, n is the total number of matches across all neuron pairs considered in the analysis, $p_{\text{class}} = n_{\text{class}}/M$ is the probability of the given class of inter-region pairs as the total number of neuron pairs of that class, n_{class} divided by the maximum number of possible neuron pairs, M . This score can be interpreted as a p -value for the null hypothesis that the number of expression matches of a gene can be explained by random assignment to neuron pairs. For the same reasons as noted above, bilateral pairs of neurons were excluded from all scoring procedures and, to ensure that each gene contributes a meaningful score, we imposed a data quality threshold on the number of possible matches, $n \geq 10$. This provides a score for each gene, a , that is higher for genes that exhibit more matches in the edge class of interest than expected by chance. The p -value scores for each gene, computed using Eq. (4). Details on how this gene scoring method was applied to specific analyses performed here are described in the Supplementary Information.

Given the reduced number of genes remaining for analysis following the quality control, we took the approach of performing an over-representation analysis (ORA) using the genes that contribute most to a given connectivity pattern. ORA assesses whether any gene sets (GO categories) were statistically over-represented in a list of

selected genes that contribute towards increased coexpression for a given link type the most. To obtain the gene list, we used the false discovery rate correction of Benjamini and Hochberg [68] on p -values computed using Eq. (4), which were thresholded at a stringent level of $p = 10^{-4}$. Over-representation for each biological process GO category with between 5 and 100 annotations was quantified as an FDR-corrected p -value (across [[XX]] GO categories) using version 3.0.2 of *ErmineJ* software [69]. Biological process GO annotations [70] were obtained from GEMMA [71] (using *Generic_worm_noParents.an.txt.gz* downloaded on March 31, 2017). Gene Ontology terms and definitions were obtained in RDF XML file format downloaded from *archive.geneontology.org/latest-termdb/go_daily-termdb.rdf-xml.gz* on March 31 2017.

Results

Our main aim in this work is to relate patterns of pairwise connectivity in *C. elegans* to coupled gene expression (gene coexpression), focusing particularly on hub connectivity. A schematic overview of our data is in Fig. 1, including the directed binary connectome (Fig. 1A), additional anatomical data gathered for each neuron (Fig. 1B), and binary gene expression across 948 genes (Fig. 1C).

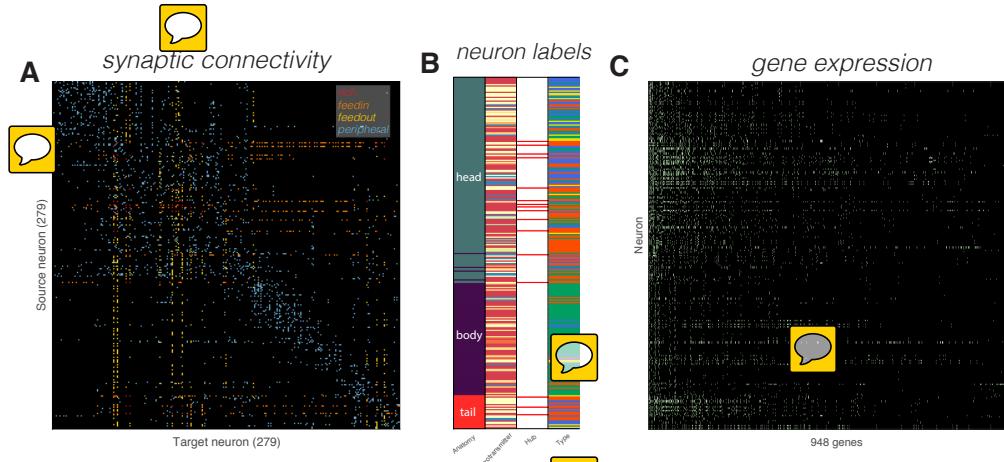


Fig 1. Schematic representation of the data used in this study. All plots show neurons (rows) ordered by their longitudinal position, from the top of the head (upper) to bottom of the tail (lower). Schematic representation of the *C. elegans* nervous system is shown on the left (scale of the neural ring is not preserved). **A** 2990 directed connections between 279 neurons from neuron i (row) to neuron j (column). Connections are colored according to how they connect hubs ($k > 44$) and non-hubs ($k \leq 44$), as ‘rich’ (hub → hub), ‘feed-in’ (nonhub → hub), ‘feed-out’ (nonhub → hub), and ‘peripheral’ (nonhub → nonhub). **B** Anatomical location, neurotransmitter system, functional type and hub assignment for each of 279 neurons. **C** Binary gene expression indicated as a green dot when a gene (column) is expressed in a neuron (row).

We present our analysis in four parts: (i) we characterize the spatial dependency of connection probability and gene coexpression, (ii) we confirm the rich-club organization of the *C. elegans* connectome, (iii) we show that gene coexpression is increased in connected pairs of neurons relative to unconnected pairs, in electrical synapses relative to chemical synapses, and in connected hub neurons relative to other types of connected neuron pairs (mirroring previous results in the mesoscale mouse connectome [36]), (iv) the expression similarity of hubs is not driven by factors like stereotypical interneuron

expression, expression similarity within modules, birth time or lineage similarity, or neuromodulator type; current data suggests that the pattern is driven by genes related to glutamate receptor signaling, and by the high coexpression amongst control interneurons.

284
285
286
287

Spatial dependency

288
289
290
291
292
293
294
295
296
297
298
299
300
301
302

Previous work has demonstrated the importance of spatial effects in driving patterns of gene expression, with more proximal brain areas exhibiting more similar gene expression patterns than more distant brain areas [36, 72–74]. Connection probability also decreases with spatial separation in mammalian brains [24, 26, 36, 75, 76], and in *C. elegans* (cf. Fig. S1 of [48]). Many aspects of connectome organization can be partially explained by geometric effects, such as the decreasing connection probability with increasing separation distance [24, 26, 77]. Unlike network analyses of mammalian brains, where all neurons are confined to a spatially contiguous organ, neurons of the *C. elegans* nervous system are distributed throughout the entire organism, forming a dense cluster of 147 neurons in the head (all within 130 μm), 105 sparser, mostly motor neurons (75%) in the body (spanning 1.02 mm), and a relatively dense cluster of 27 neurons in the tail (all within 90 μm of each other), as plotted in Fig. 2. In order to understand relationships between connectivity and gene expression, we first need to understand the spatial embedding of both measurements.

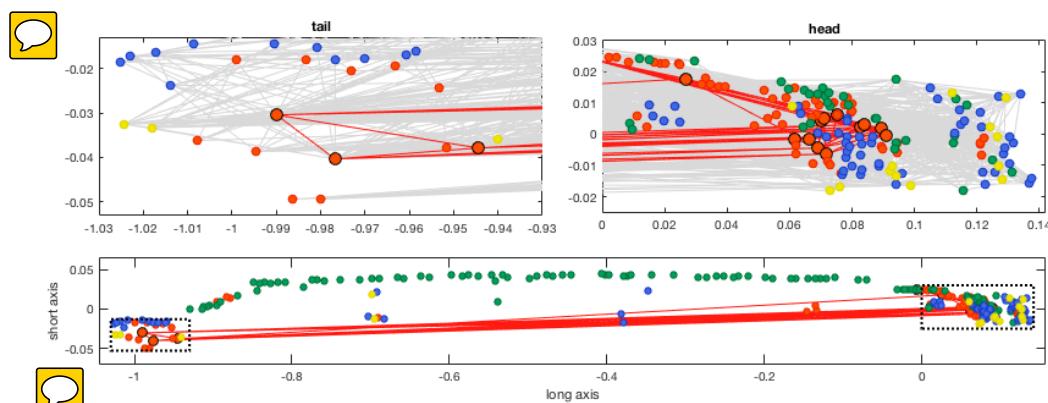


Fig 2. Hub neurons are distributed between the head and tail. Neurons are positioned along the Posterior (P)-Anterior (A) (horizontal), and Ventral (V)-Dorsal (D) (vertical) axes. Neurons are colored by type: (i) interneuron (85 neurons, orange), (ii) sensory (68 neurons, blue), (iii) motor (108 neurons, green), or (iv) multiple assignments (18 neurons, yellow). Hub neurons are shown as larger circles and outlined in black. Rich-club connections (between hub neurons with $k > 44$; red) are shown (in addition to all other connections in the upper plots, shown gray). Axes of each subplot are to scale with each other, and the upper zoomed-in plots of the head and tail are shown as dotted rectangles in the lower plot.

303
304
305
306
307
308
309
310

We first characterize the probability that two neurons will be connected given their source and target types, after labeling each neuron as either: ‘head’, ‘body’, or ‘tail’, according to anatomically distinct regions of the *C. elegans* nervous system. Connection probability is plotted as a function of Euclidean separation distance in Fig. 3 for each combination of source and target neuron labels, across 10 equiprobable distance bins (and with an exponential fit) in each case. Distinguishing connections by source and target neuron types uncovers clear spatial relationships (that are obscured when all connections are grouped, as in [48]), revealing that different classes of connections

exhibit different spatial connectivity relationships. From the very short distance scale of $\lesssim 100 \mu\text{m}$ of head→head and tail→tail connections to the very longest-range head→tail and tail→head connections ($\gtrsim 1 \text{ mm}$), connection probability decreases with separation distance (Fig. 3A). For connections *across the body*, ranging up to $\approx 1 \text{ mm}$, a near-exponential trend is exhibited, mirroring results in other species and across longer length scales [75], including mouse [36, 78], and rodents and primates [26]. Other connections do not exhibit strong spatial connectivity relationships, i.e., connections between the body and head or tail, shown in Fig. 3B.

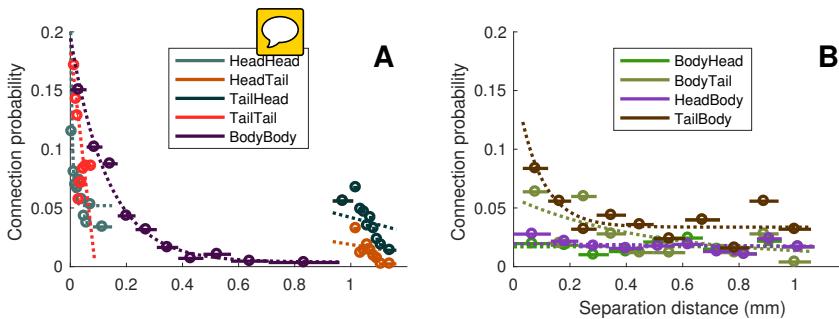


Fig 3. Connection probability decreases with separation distance within and between the head and tail, and within the body. The connection probability for a pair of neurons as a function of their Euclidean distance is estimated in 10 equiprobable distance bins, shown as a circle (bin centers) and a horizontal line (bin extent). There is a decreasing relationship for connections: within the head (aqua), from head→tail (brown) and from tail→head (stone blue), within the tail (red), and within the body (dark purple). Exponential fits of the form $f(x) = A \exp(-\lambda x) + B$, some of which appear approximately linear across the range of the data, are shown as dotted lines. **B** Plots as in **A**, but for connection classes between the body and head/tail: from body→head (forest green), from body→tail (dirt green), connections from head → body (purple), and from tail→body (dark brown). Apart from a small effect at short range for tail→body connections, these connection classes show minimal distance dependence.

We next investigate the spatial dependency of gene coexpression between neuron pairs, r_ϕ , as a function of separation distance, shown in Fig. 4. Gene coexpression decreases slightly with separation distance for the spatially close neurons within the head (Fig. S4A) and within the tail (Fig. S4B), but not for pairs of neurons involving the body (Fig. S4C). Stronger spatial relationships were seen within the head for specific subtypes of neurons, including pairs of interneurons, pairs of sensory neurons, pairs of motor neurons, and motor-interneuron pairs (Figs S4D–H). The weak decreasing trend in r_ϕ with distance in *C. elegans*, is only seen within the head and tail, and is primarily driven by a subset of nearby neurons with high r_ϕ , and may therefore represent a relationship specific to particular functionally related neurons, rather than a general, bulk spatial relationship that is seen in macroscopic mammalian brains [36]. Accordingly, attempting to correct for a bulk trend in the discrete *C. elegans* connectome by taking residuals from an exponential fit produced artifactual reductions in the coexpression of many neuron pairs (shown in Fig. S2). Thus, we found no evidence for bulk spatial relationships of r_ϕ in the specialized and discrete neuronal connectome of *C. elegans*; we analyze raw coexpression data, r_ϕ , throughout this work.

Hub connectivity

Next we analyze the topological properties of the *C. elegans* connectome, represented here as a directed, binary connectivity matrix between 279 neurons, combining directed

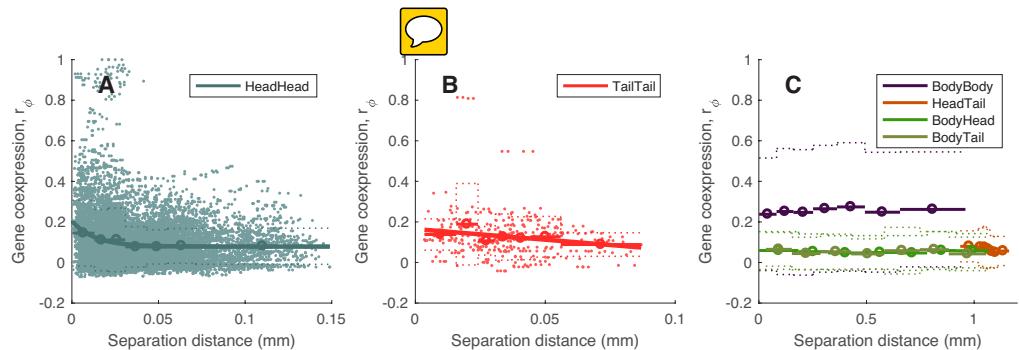


Fig 4. Dependence of gene coexpression, r_ϕ , on spatial separation between pairs of neurons. Gene coexpression, r_ϕ (in all cases excluding left-right homologous pairs of neurons), is shown as a function of the pairwise separation distance between pairs of neurons (shown as the mean (solid) \pm standard deviation (dotted) in seven equiprobable distance bins, with extent shown as horizontal bars), for **A** all pairs of neurons in the head, **B** all pairs of neurons in the tail, and **C** all other pairs (labeled). Scatters for all neuron pairs are added in **A** and **B**. An exponential relationship, $f(x) = A \exp(-\lambda x) + B$, is fitted in **A** and **B**. The weak decreasing trend in r_ϕ with distance, is primarily driven by a small subset of nearby neurons with high r_ϕ , and may therefore represent a more specific relationship between particular neurons, rather than a general, bulk spatial relationship observed in macroscopic mammalian brains [36, 79].

chemical synapses and undirected electrical gap junctions (Fig. 1). The degree distribution is shown in Fig. 5A, where neurons are distinguished by type: 68 sensory neurons, 85 interneurons, 108 motor neurons, and 18 neurons with multiple assignments. Consistent with the results of Towlson et al., who used an undirected version of the connectome (by ignoring the directionality of chemical synapses), we found a positively-skewed degree distribution containing an extended tail of high-degree interneurons that overlap almost completely with the list of Towlson et al. [23] (see Table 1). The hub interneurons of *C. elegans* include the anatomical substrate for behaviors like coordinated locomotion and foraging [80].

We next used the normalized rich-club coefficient, Φ_{norm} , to quantify the extent to which hubs are densely interconnected, with $\Phi_{\text{norm}} > 1$ indicating rich-club organization of the network. Again consistent with the results of Towlson et al. [23], we find rich-club organization in the connectome, as shown in Fig. 5B. The figure plots the variation of Φ_{norm} across degree thresholds, k , at which hubs are defined (as neurons with degree $> k$), with red circles indicating a significant increase in link density among hubs relatively to 1,000 degree-preserving nulls (permutation test, $p < 0.05$). The plot reveals rich-club organization at the upper tail of the degree distribution, particularly across the range $44 < k < 63$, shaded gray in Fig. 5B. Throughout this work, we define a set of hubs as the sixteen neurons with $k > 44$, which corresponds to the lowest degree threshold at which the network displays a contiguous region at high k with significant rich-club organization (permutation test, $p < 0.05$). Similar results were obtained using weighted connectome representations (i.e., using information about the number of synapses in the chemical connectivity network) for two different definitions of the weighted rich-club coefficient [81], as shown in Fig. S6.

The rich-club connectivity of the *C. elegans* connectome is accompanied by an increase in mean hub-hub connection distance [23], with a significant increase through the topological rich-club regime (Welch's t -test, $p < 0.05$), as shown in Fig. 5B. [[TODO-BF: use ranksum test]] This result can be attributed to a relative increase in long-distance hub-hub connections between the head and tail, shown in Fig. S3. The high connection density and long mean anatomical distance of pairs of hub neurons

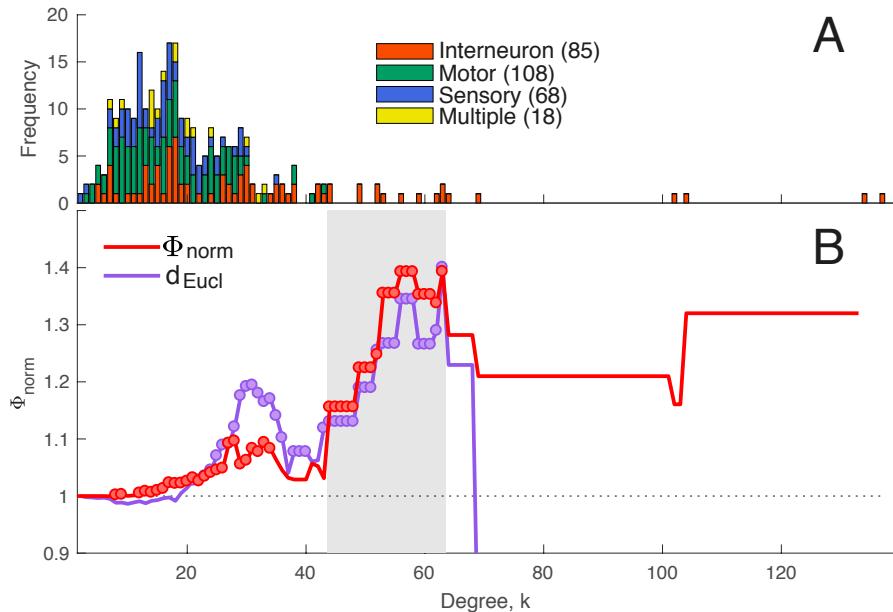


Fig 5. Rich-club organization of the connectome. **A** Degree distribution of the directed *C. elegans* connectome, where neurons are labeled according to four categories: (i) interneuron (85 neurons, orange), (ii) sensory (68 neurons, blue), (iii) motor (108 neurons, green), or (iv) multiple assignments (18 neurons, yellow). The distribution features an extended tail of high-degree interneurons. **B** Normalized rich club coefficient, Φ_{norm} (red), as a function of the degree, k , at which hubs are defined (as neurons with degree $> k$). Also shown is the mean Euclidean separation distance (purple) between connected hub regions (across degree thresholds, k). $\Phi_{\text{norm}} > 1$ indicates that hubs are more densely interconnected among each other than expected by chance, with red circles indicate values of Φ_{norm} that are significantly higher than an ensemble of 1000 degree-matched null networks ($p < 0.05$). Purple circles indicate where the Euclidean distance between connected pairs of hubs is significantly greater than the Euclidean distance for all other pairs of connected regions (Welch's t -test, $p < 0.05$). [[TODO: ranksum?]] [[TODO: Also suggest replacing red with one of these, since red is used in coexpression plot to represent something different]] [[TODO: Add right-side vertical-axis to give units for distance]]

counters the general trend in the *C. elegans* connectome, where the probability of connectivity between two neurons decays with their separation distance (quantified in Fig. 3, cf. spatial visualization Fig. 2). Our results demonstrate that hub-hub connections in *C. elegans* are deep and extend over significantly longer distances than other types of connections in the haptic network of the nematode nervous system, consistent with the rich club forming a central yet costly backbone for neuronal communication [32].

Coexpression and connectivity

We next investigate how the network connectivity properties of *C. elegans* relate to patterns of gene expression. Being a pairwise analysis, we compare pairwise connectivity to pairwise similarity in gene expression, measured by the mean square contingency coefficient, r_ϕ (excluding bilateral pairs of neurons to ensure their high transcriptional similarity does not drive results, see *Methods*).

To investigate whether gene coexpression varies as a function of connectivity, we compared the distribution of r_ϕ for all connected pairs of neurons, and all unconnected pairs of neurons. Connected pairs of neurons have more similar expression profiles than unconnected pairs (Wilcoxon rank-sum test, $p = 1.8 \times 10^{-78}$), mirroring previous results in the mesoscale mouse connectome [36].

We also computed the distribution of r_ϕ for: (i) all pairs of neurons that are connected via a gap junction, (ii) all pairs of neurons that are connected via a chemical synapse, and (iii) all pairs of neurons that have neither connection, shown in Fig. 6A. This analysis involves 474 pairs of neurons connected by a gap junctions, 1940 pairs connected by a synapses, and 36 450 pairs with no connection (after excluding bilateral pairs of neurons). Note that 175 pairs of neurons are connected both synaptically and by a gap junction, and are thus included in both gap junction and synaptic categories. Pairs of neuron that are connected via gap junctions exhibit more similar gene expression profiles than those connected via chemical synapses (Wilcoxon rank-sum test, $p < 5.4 \times 10^{-22}$).

To investigate whether there was a difference in gene coexpression between pairs of neurons connected reciprocally by chemical synapses ($N_1 \leftrightarrow N_2$ for two neurons N_1 and N_2) versus those connected unidirectionally ($N_1 \rightarrow N_2$), we compared the distribution of r_ϕ in both groups. There was no significant difference in gene coexpression between the two groups (Wilcoxon rank-sum test, $p = 0.99$) (distributions shown in Fig. S5), in contrast to significant differences found in the mesoscale mouse connectome [36].

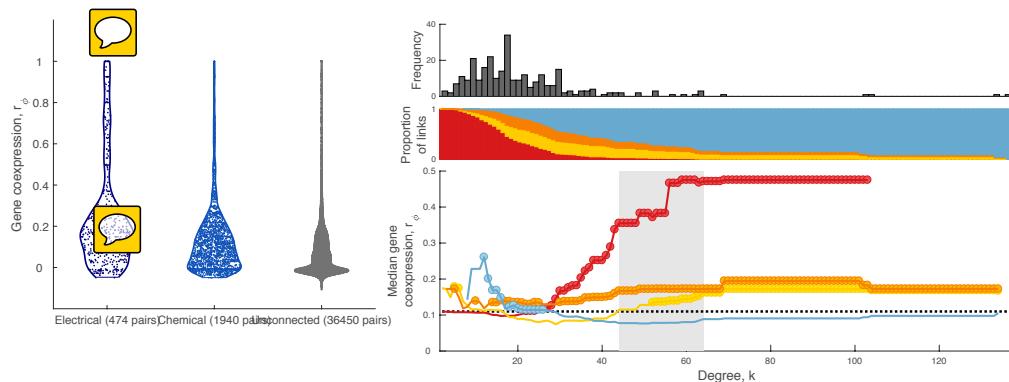


Fig 6. Gene coexpression varies as a function of connectedness and connection type, with the highest coexpression for neurons connected via gap junctions, and for connections involving hubs. **A** Distribution of transcriptional similarity between (i) pairs of neurons connected by gap junctions, (ii) pairs of neurons connected by synapses, and (iii) pairs of neurons that are unconnected, shown as a violin plot, with the median of each distribution represented by a horizontal line. **B** *Top:* Degree distribution, k , of the *C. elegans* connectome. *Middle:* proportion of connections that are: ‘rich’ (hub→hub, red), ‘feed-in’ (nonhub→hub, yellow), ‘feed-out’ (hub→nonhub, orange), or ‘peripheral’ (nonhub→nonhub, blue) as a function of the degree threshold, k , used to define hubs. Note that at high k most neurons are labeled as nonhubs and hence the vast majority of connections are labeled ‘peripheral’. *Bottom:* Median gene coexpression, r_ϕ , for each connection type as a function of k . The median coexpression across all network links is shown as a dotted black line; the topological rich-club regime (determined from the network topology, cf. Fig. 5) is shaded gray. Circles indicate a statistically significant increase in gene coexpression in a given link type relative to the rest of the network (one-sided Wilcoxon rank-sum test; $p < 0.05$). [[TODO: fonts are small, and inconsistent sizes between subplots.]] [[TODO: add labels A, B]] [[TODO: Font size inconsistent in B]]

We next investigated whether coexpression varies across different types of connections defined in terms of their hub connectivity. For a given hub threshold, k , we first labeled each neuron as either a ‘hub’ (nodes with degree $> k$) or a ‘nonhub’ (degree $\leq k$), and then labeled each connection as either ‘rich’ (hub → hub), ‘feed-in’ (nonhub → hub), ‘feed-out’ (hub → nonhub), or ‘peripheral’ (nonhub → nonhub). The median coexpression, \tilde{r}_ϕ for each of these four connection types is plotted in Fig. 6B, with circles indicating statistically significant increases of a given connection type (relative to all other connections, one-sided Wilcoxon rank-sum test; $p < 0.05$). Gene coexpression in rich connections increases with degree, k , particularly in the topological rich-club regime where hubs are densely interconnected (shaded gray in Fig. 6B). Both feed-in and feed-out connections exhibit increased gene coexpression in the topological rich-club regime, while peripheral connections show the lowest levels of coexpression. Full distributions for each edge type for a hub threshold of $k > 44$ are plotted in Fig. 5.

These results, using partial binary annotations of gene expression across just 948 genes in a microscale neuronal connectome, are broadly consistent with a prior analysis of the expression of over 17 000 genes across 213 regions of the mesoscale mouse connectome, including: (i) increased coexpression in connected pairs of nodes; (ii) the highest coexpression in rich connections; and (iii) lowest coexpression in peripheral connections.

Potential drivers of elevated coexpression between hubs

The sixteen hub neurons in *C. elegans* ($k > 44$) are all interneurons in the head and tail that are born prior to hatching, are mostly cholinergic (13/16), and are mostly contained within a single network module. We next investigated whether the similarity of gene expression profiles between hubs was specific to their high connectivity levels, or whether it could rather be explained by similarity in one or more of these other characteristics.

Interneurons All sixteen hubs ($k > 44$) in *C. elegans* are interneurons. To determine whether the increase in gene coexpression in rich connections was specific to interneurons, we plotted the median coexpression for hub-hub connections, \tilde{r}_ϕ , as a function of the degree threshold, k , separately for: (i) connections involving interneurons, (ii) connections involving sensory neurons, and (iii) connections involving motor neurons, as shown in Fig. 7A. For the curve labeled ‘sensory’ neurons, for example, each point is the median \tilde{r}_ϕ across connections involving sensory neurons (i.e., at least one neuron of a connected pair is a sensory neuron), for which both neurons have degree $> k$. We find that the increase in median hub-hub gene coexpression, \tilde{r}_ϕ , with k is strongest for connections involving interneurons, while connections involving motor neurons show a smaller increase with k , and connections involving sensory neurons show no relationship.

We next investigated whether hub neurons show a unique transcriptional signature amongst interneurons. Pairs of hub interneurons indeed exhibit significantly higher coexpression compared to pairs of all other interneurons (Wilcoxon rank sum test, $p = 5 \times 10^{-31}$). However, we tested whether this observed increase could be driven by the anatomical properties of hub interneurons [[TODO AA: this still needs some unpacking to distinguish from the rank sum test just performed]]. We selected a subset of interneurons that most closely resemble the anatomical properties of hub interneurons, in terms of their origination and termination locations, arriving at the following set of eight nonhub head interneurons: AVFL, AVFR, AVHL, AVHR, AVKL, AVKR, AVJL, and AVJR [[any reproducible criteria through which these were selected??]]. As shown in Fig. 7B, pairs of hub interneurons exhibit significantly higher gene coexpression than pairs of anatomically-matched nonhub interneurons (Wilcoxon

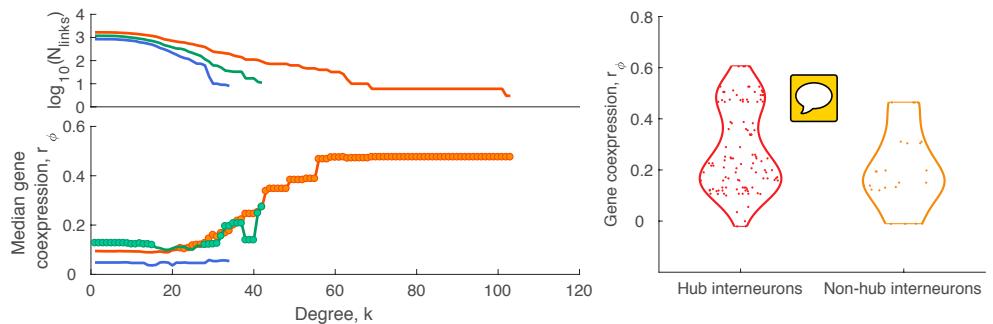


Fig 7. The increase in coexpression with degree is specific to high-degree interneurons, and hub interneurons have higher coexpression than nonhub interneurons. **A** Top: The number of connections involving interneurons (orange), sensory neurons (blue), and motor neurons (green) across degree threshold, k represented as $\log_{10}(\text{number of links})$. Bottom: Median gene coexpression as a function of degree for connections involving different types of neurons. Circles indicate a statistically significant increase in gene coexpression in a given link type relative to the rest of the network (one-sided Wilcoxon rank-sum test; $p < 0.05$). **B** Coexpression distributions for all pairs of hub interneurons (red) and all pairs of non-hub interneurons (yellow) (Wilcoxon rank sum test, $p = 0.047$). [[TODO-AA: Is the top-left plot actually counting pairs?? Or is it actual connectome edges?]] [[TODO-AA: Add labels A, B, C using the LabelCurrentAxes function]] [[TODO-AA: Clarify axis tick labels in B: at the moment appears like a node-level analysis on horizontal axis but pairwise analysis on vertical]] [[TODO-AA: Recolor, or somehow update caption for B that reflects the new command interneuron story – will do when plots are finalized]]

rank sum test, $p = 0.047$). Thus, the increase in coexpression amongst hub interneurons cannot be explained simply by the fact that hubs are interneurons. 451
452

Modular organization The *C. elegans* connectome has a modular organization, and can be decomposed into groups of neurons or modules with dense inter-module connectivity (with relatively sparse connectivity between modules) [15, 52, 82], or into groups of neurons with more similar patterns of connectivity within groups than between groups [65, 83]. Given that connectome modules may correspond to functionally specialized systems, it is therefore plausible that neurons belonging to the same module might exhibit similar gene expression patterns, as has been shown for macroscopic functional networks in the human brain [84]. Brain network hubs generally participate in multiple modules [31, 85], and their dense rich-club connectivity implies that they may form their own module that overlaps with other, more specialized systems [21, 22, 27]. We therefore investigated whether increased gene coexpression between pairs of hub neurons in the *C. elegans* connectome can be explained by their distribution across connectome modules. 453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474

We used the Louvain community detection algorithm [62] to extract modules from the *C. elegans* synaptic connectome using consensus clustering (see *Methods*). Four modules were extracted, with eleven hubs in module one (which contains 111 neurons), four hubs in module two (96 neurons), one hub in module three (40 neurons), and no hubs in module four (32 neurons). We compared the results of this modular partition of neurons to a previously reported nine-module partition derived from an Erdős-Rényi Mixture Model (ERMM) [65]. For the Louvain consensus modules, gene coexpression, r_ϕ , was significantly increased for connected neurons in the same module (1552 pairs) relative to connected pairs in different modules (687 pairs) (Wilcoxon rank sum test,

475
476
477
478
479
480
481
482
483
484
485
486
487
488
489

$p = 6.6 \times 10^{-4}$ but there was no significant difference between intra-modular connected neurons and inter-modular connected neurons for the nine-module ERMM partition (Wilcoxon rank sum test, $p = 0.46$). The resolution and type of modular decomposition thus affects the relationship between coexpression, connectivity, and modular network structure in *C. elegans*. Given that modular structure of the network can influence coexpression (for the four-module Louvain decomposition), we tested whether connected hubs exhibit more similar coexpression within and between modules analyzing both the consensus Louvain and ERMM decompositions. For pairs of connected neurons within the same module, coexpression is higher for pairs of hubs than pairs of nonhubs (Wilcoxon rank sum test, $p = 6.9 \times 10^{-17}$ for consensus Louvain modules, shown in Fig. 8A; 9.3×10^{-7} for ERMM partition). We found a similar result for connected neurons in different modules: pairs of connected hubs exhibit increased coexpression than other types of connected pairs of neurons (Wilcoxon rank sum test, $p = 1.6 \times 10^{-5}$ for consensus Louvain modules, shown in Fig. 8B; $p = 1.6 \times 10^{-16}$ for ERMM). Thus, for both types of modular decompositions, connected pairs of hub neurons exhibit increased coexpression relative to other intra-modular and inter-modular connections.

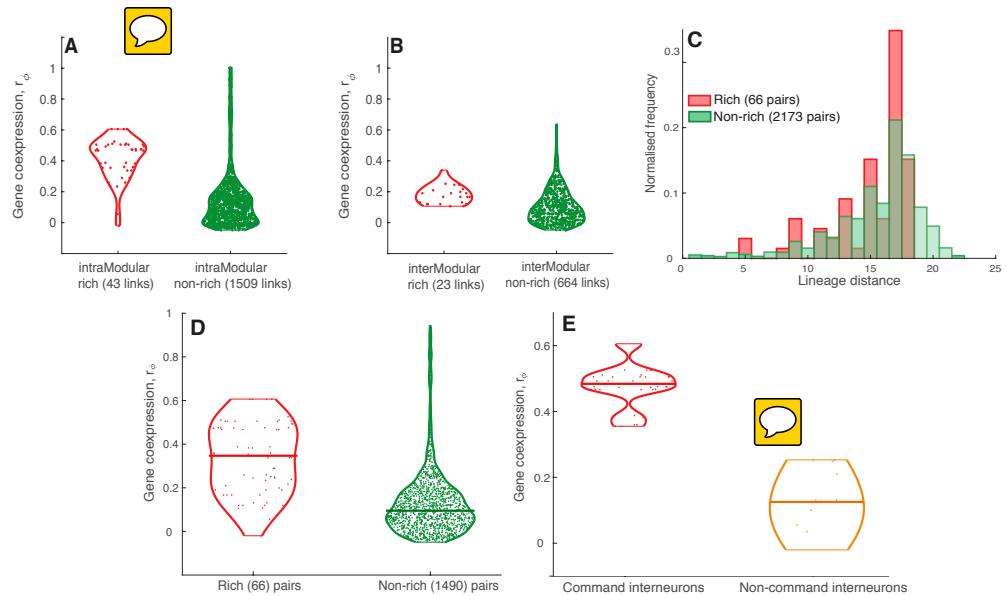


Fig 8. Increased coexpression of hub neurons is not driven by their membership of similar topological modules, similar birth times, or lineage distance. **A** Distributions of gene coexpression, r_ϕ , for intra-modular rich (red) non-rich (green) connections, shown as violin plots with the median shown as a horizontal bar (Wilcoxon rank sum test, $p = 6.9 \times 10^{-17}$). **B** Distributions of gene coexpression, r_ϕ , for inter-modular rich (red) and non-rich (green) connections, shown as violin plots with the median shown as a horizontal bar (Wilcoxon rank sum test, $p = 1.6 \times 10^{-5}$). **C** Distributions of gene coexpression, r_ϕ , between early born hubs (rich links, red) and non-hubs (non-rich links, green) shown as violin plots with the median shown as a horizontal bar. (Wilcoxon rank sum test, $p = 3.9 \times 10^{-22}$). **D** Distributions of lineage distance between rich links (red) and non-rich links (green), plotted as histograms due to a discrete nature of this measure. (Wilcoxon rank sum test, $p = 0.079$). **E** Distributions of gene coexpression between hub command interneurons (red) and hub non-commans interneurons (yellow) shown as violin plots (Wilcoxon rank sum test, $p = 3.3 \times 10^{-8}$). [[TODO: change colors for distributions, fill distributions, e.g., intra-modular and inter-modular shouldn't use red (rich) use green for non-rich]]

Birth time The genesis of neurons in the *C. elegans* nervous system is separated into two distinct time periods: before hatching (birth time < 550 min – ‘early’) and after hatching (birth time > 1200 min – ‘late’), with no neurons formed during intermediate times [54]. Distinguishing neurons into these two groups is appropriate as birth time differences between pairs of neurons are dominated by the difference between these two clusters of neuron birth periods. As a broad group, connected pairs of early-born neurons do not exhibit significantly different gene coexpression compared to connected pairs of late-born neurons (Wilcoxon rank sum test, $p = 0.64$), but connected pairs of early-born neurons do exhibit significantly higher gene coexpression relative to pairs of connected neurons for which one neuron is born prior to hatching and the other is born post-hatching (Wilcoxon rank sum test, $p = 4.2 \times 10^{-3}$). Since all *C. elegans* hub neurons are born prior to hatching [55], neurons born at similar times may share similar connectivity properties [56], we investigated whether the increase in coexpression between hub neurons of *C. elegans* may be driven by their similar birth times. Focusing on the 201 neurons (16 hub and 185 nonhub neurons) born prior to hatching, connected pairs of hub neurons (rich) exhibit significantly increased coexpression relative to other pairs of connected early-born neurons (i.e., feed-in, feed-out, and peripheral connections; Wilcoxon rank sum test, $p < 10^{-22}$), as shown in Fig. 8D.

Lineage distance The lineage distance between a pair of neurons is defined as the sum of total divisions that have taken place since the most recent common ancestor cell [51,57,58]. In the mammalian brain, neuronal lineage has been associated with both functional [?, ?] and connectivity-related properties [?]. Moreover, tissue distance (resembling lineage distance on a cellular scale) was found to correlate with gene expression divergence, meaning that tissues from the same branch on the fate map share more similar gene expression pattern in both human and mouse mesoderm as well as ectoderm tissues [87]. Given the ectoderm afterwards differentiates to form the nervous system, this finding suggests a possible relationship between lineage distance and gene coexpression in a microscale neuronal system such as of *C. elegans*. While there is no significant correlation between lineage distance and gene coexpression in *C. elegans* (Spearman $\rho = -0.027$, $p = 0.2$), all hub neurons are formed early in development as characterized above, and hence it is possible that they originate from common ancestry cells and share intrinsic similarities in gene expression that influence the observed increase in coexpression between hub neurons. However, as shown in Fig. 8E, there is no significant difference in lineage distance between hub (rich connections) and non-hub neurons (non-rich connections) (Wilcoxon rank sum test, $p = 0.079$). [[TODO-AA: make clear whether connected or pairs]] Thus, we can not attribute the transcriptional similarity of connected hub neurons to their neuronal lineage.

Neurotransmitter system Hub neurons ($k > 44$) consist of thirteen cholinergic neurons, two glutamatergic neurons, and one neuron of unknown type [56]. Neuron pairs show different coexpression relationships as a function of their neurotransmitter type, e.g., pairs of GABAergic neurons have high median coexpression $\tilde{r}_\phi = 0.59$, while pairs of glutamatergic neurons exhibit a relatively low median coexpression $\tilde{r}_\phi = 0.18$. To determine whether the similarity in gene coexpression between pairs of hub neurons could be explained by their neurotransmitter types, we used a permutation test to compute the mean coexpression between all pairs of random groups of sixteen neurons of the same neurotransmitter types (e.g., thirteen random cholinergic neurons, two random glutamatergic neurons, and one random neuron of unknown neurotransmitter type) across 1×10^6 permutations. The increased coexpression between hub neurons cannot be explained by their neurotransmitter identities, with hub neurons displaying a significant increase in median coexpression relative to random sets of neurons with the

same neurotransmitter types as hub neurons (permutation test, $p = 2 \times 10^{-6}$). 541

Broad anatomy. Of the sixteen hub neurons, thirteen are in the head and three are 542 in the tail; none are located in the body. Given that coexpression varies as a function of 543 anatomical position, with pairs of neurons in the same broad anatomical class exhibiting 544 the highest mean coexpression (e.g., pairs of neurons within the body have a median 545 $\tilde{r}_\phi = 0.14$, pairs of neurons within the tail have $\tilde{r}_\phi = 0.12$, and pairs of neurons within 546 the head have $\tilde{r}_\phi = 0.07$), with lower mean coexpression in mixed classes (e.g., pairs of 547 neurons between head and body have $\tilde{r}_\phi = 0.01$). Given this variation, we tested 548 whether the increased gene expression in hub neurons could be explained by their 549 anatomical distribution using the permutation testing procedure described above for 550 neurotransmitter systems. That is, by comparing the distribution of coexpression 551 between hubs to a null distribution formed from 1×10^6 random permutations of 552 thirteen head neurons and three tail neurons. The median coexpression between hub 553 neurons is significantly increased relative to random sets of thirteen head neurons and 554 three tail neurons (permutation test, $p = 2 \times 10^{-6}$). Within the head, gene coexpression 555 is significantly increased amongst hub neurons relatively to other pairs of head neurons 556 (Wilcoxon rank sum test, $p = 4.1 \times 10^{-11}$ for 46 hub-hub pairs and 1186 others), and 557 also amongst head/tail pairs of neurons (Wilcoxon rank sum test, $p = 1.6 \times 10^{-14}$ for 23 558 hub-hub pairs, 174 others), but not for the three hub neurons in the tail (Wilcoxon rank 559 sum test, $p = 0.15$ for three hub-hub pairs, 53 others). 560

Command interneurons. *C. elegans* neurons can be divided into distinct groups 561 that each perform a specialized behavioral function [?]. The ‘command interneurons’ 562 are a functional group of ten neurons that govern forward (AVBL, AVBR, PVCL, 563 PVCR) and backward (AVAL, AVAR, AVDL, AVDR, AVEL, AVER) locomotion [?]; all 564 of these neurons are hubs. Given the overlap between hub neurons and the command 565 interneurons, we investigated whether command interneurons exhibit more similar 566 expression than other hub interneurons, and may therefore drive the increase in 567 coexpression amongst hubs as a whole. We compared gene coexpression, r_ϕ , between all 568 pairs of command interneurons (ten neurons), and between all pairs of hubs that are not 569 command interneurons (six neurons), shown in Fig. 8E. Gene coexpression between 570 command interneurons is significantly greater than between other hub neurons 571 (Wilcoxon rank sum test, $p = 3.3 \times 10^{-8}$), indicating that command interneurons play a 572 major role in driving the increased coexpression amongst hub neurons. Relative to 573 macroscopic brains in which hubs are often distributed across diverse neural 574 systems [36], it is not possible to disentangle the role of specialized functional circuits 575 from connectivity in *C. elegans*, with highly connected hubs containing the 576 transcriptionally uniform set of command interneurons. 577

Functional enrichment

Having characterized a robust relationship between hub connectivity and gene 579 expression, we next investigate whether specific functional groups of genes contribute to 580 the observed differences in gene coexpression. We developed a method to score genes 581 according to their contribution to a given coexpression pattern, which was used to 582 perform an overrepresentation enrichment analysis using gene ontology (GO) categories 583 for biological processes [69, 70] (see Methods). Relative to previous results in mouse, for 584 example [36, 88–90], the current data are more challenging due to incomplete binary 585 expression measurements in a small proportion of the genome. 586

We first determined which functional groups of genes drive increased coexpression in 587 connected pairs of neurons relative to unconnected pairs. The top GO categories are 588

related to glutamate receptor signaling: ‘ionotropic glutamate receptor signaling pathway’ and ‘glutamate receptor signaling pathway’, as well as cell surface receptor signaling, ‘ion transport’, and ‘ion transmembrane transport’, but none are significant at a false discovery rate of 0.05. Other categories are involved in the regulation of growth rate and several catabolic processes (see Table 2). While these incomplete binary expression data do not allow us to deduce functional categories with statistical significance, our findings point to the possible role of some relevant neuronal processes in driving the increased coexpression of connected pairs of neurons relative to unconnected pairs.

We next investigated whether particular functional groups of genes drive differences in coexpression between connections involving hub neurons (i.e., in rich, feed-in, and feed-out connections) relative to peripheral connections between pairs of nonhub neurons. While previous work implicated genes regulating oxidative metabolism for connections involving hubs in mouse [36] (and for hub regions in human [43]), the *C. elegans* gene expression dataset was insufficient to analyze these same gene categories here (e.g., only one of the 948 genes annotated to the GO categories related to hub connectivity in mouse is present in our gene expression dataset: *unc-32* is annotated to ‘ATP hydrolysis coupled proton transport’). Thus, although a direct test of the metabolic hypothesis is not possible from current data, we investigated whether relevant GO categories were overrepresented in pairs of connected hubs. Top GO categories are listed in Table 3; none are statistically significant at a false discovery rate of 0.05. The same two glutamate receptor signaling categories implicated for connectivity above: ‘ionotropic glutamate receptor signaling pathway’ and ‘glutamate receptor signaling pathway’, were again at the top of the list of categories related to hub connectivity. Other top categories include a more general receptor signaling class (‘cell surface receptor signaling pathway’), ‘regulation of locomotion’, and various metabolism and biosynthesis related processes including ‘positive regulation of biosynthetic process’, two groups responsible for positive regulation of nitrogen and nucleobase-containing compound metabolic process, ‘positive regulation of cellular metabolic process’, and ‘positive regulation of macromolecule biosynthetic process’.

The implication of glutamate receptor signaling for synaptic connectivity, and hub connectivity in particular, is somewhat surprising given that only 26% of neurons with known neurotransmitter type are glutamatergic (whereas 55% are cholinergic) and just two of the sixteen hubs are glutamatergic (thirteen are cholinergic, and one is of unknown type). Furthermore, just 32% of neurons that project to hubs are glutamatergic. Our inability to get a comprehensive understanding of neurotransmitter signaling for the gene expression data used here is related to its incompleteness, e.g., only one acetylcholine-related GO category—‘synaptic transmission, cholinergic’ (GO:0007271)—contained sufficiently many genes (fourteen) to be represented in our enrichment analysis. Glutamate release categories were also not represented here (each were annotated with fewer than five genes in our dataset and hence excluded). Future work leveraging more comprehensive expression data could use the methods developed here for scoring genes according to their contribution to coexpression signatures could be used to better understand the relationship between network topology, neurotransmitter signaling, and other gene-regulated functions.

Discussion

- First coexpression analysis in worm. Despite noisy gene expression data we get some insights into the genetic basis of connectivity on a neuronal level

Our results on the spatial dependence of connection probability and gene expression,

using sparsely-annotated binary expression data in a highly specialized neuronal nervous system, mirror many trends observed in bulk regions of macroscopic mammalian brains of mouse [36] and human [79]. It is therefore likely that the spatial effects in this specialized neuronal connectome are not simply due to bulk spatial trends in coexpression, as in macroscopic brains containing millions of neurons, but may reflect true organizational principles. Spacey spacey.

Compared to our previous work in the mesoscale mouse connectome [36], the present findings in a near-fully mapped neuronal chemical connectome show similarities and differences. Despite the incredible difference in spatial scale (from each node containing 10^5 neurons in mouse, to just a single neuron in worm), we found a similar basic qualitative signal of increased coexpression in connected pairs of neurons, and amongst connected pairs, the most similar expression signatures in pairs of connected hub neurons. These overall findings were specific to connections involving interneurons, i.e., were not found in motor or sensory neurons, being driven by the relatively expression similarity of high-degree interneuron hubs, were not driven by similarity in lineage, neurotransmitter type, or modular membership, and were robust to a range of data processing choices (including connectome type, and coexpression measure). These findings come despite sparsely annotated, binary gene expression data, with less than 5% coverage of the full genome, with preliminary enrichment analysis indicating a role for glutamate in both connectivity and hub-connectivity. Our results indicate that even in a highly specialized neuronal nervous system, costly hub-hub connections display a distinctive transcriptional signature that may reflect their unique functional role in the network. Further work across different species and scales may shed light on deeper principles of how connectomes organize to facilitate efficient biological functioning. The hubs of the mouse connectome are broadly distributed across anatomical divisions [36], whereas *C. elegans* hubs mostly correspond to a specific functional circuit; we attempted to address by looking at potential confounds, to characterize whether high connectivity was related to aspects like interneuron type, neurotransmitter system, etc. Results may be more a statement about similarity of control interneurons, which are more densely annotated with similar genes due to their functional specialization, rather than being specific to topology. Teasing apart specific effects of topology is much harder in a highly specialized neural system like *C. elegans*, but we did our best.

differ significantly (being a neuronal connectome), and with so few neurons, each highly specialized, it may be hard to pick up broad expression patterns that are network topology-specific. Quite different to the case in mouse, where regions were spread across the brain, and so effects of regional specialization of function, cytoarchitecture, gene expression were averaged, and the remaining hub-related signal could be isolated.

It is worth noting that given sparse gene annotations in the current binary gene expression data we were required to perform additional quality control steps to ensure that all top genes supplied for the ORA analysis have a sufficiently high number of both possible and actual matches and therefore our enrichment results are not driven by the noise in the data. [[include the whole list of genes for both a analysis as excel data with numbers of possible and actual matches as well as corrected and uncorrected p values]] Taking into consideration limitations of the current dataset for such an analysis, we still find enrichment in meaningful GO categories for both connected neurons as well as connections involving hubs that, while are not statistically significant, allow us to provide some interpretation about the underlying mechanisms of hub connectivity in the neuronal connectome of the *C. elegans* and are in line with findings in both meso and macro level connectomes.

Being a EM connectome, avoids statistical estimation of connectomes from tract-tracing experiments [91], inference of axonal connectivity from diffusion MRI, or the need to form a discrete parcellation of a continuous brain.

[[delete?:]] [[It has been shown that cells with identical fates can be formed by different gene regulatory pathways [92] [[relevance?]], however to the best of knowledge the relationship between neuron lineage distance and gene expression patterns has not been investigated yet [[move to discussion?]].]]

The recent availability of gene expression data in the brain has allowed structural connectome properties to be related to molecular information [93]. Initial statistical work done using binary gene expression annotations in *C. elegans* showed that the expression of a small number of genes can be used to predict connectivity between pairs of neurons [94–96]. Later work, utilizing detailed quantitative gene expression data in mouse from the Allen Mouse Brain Atlas [97] revealed a relationship between gene expression and connectivity in mouse [88,89], with pairs of brain regions with similar gene expression patterns also displaying similar connectivity profiles (using connectivity data from rat) [98]. Comparing different types of connections in mouse, we have demonstrated that pairs of connected hubs have the most similar gene expression profiles, and that hub connectivity is associated with the correlated expression of genes related to oxidative energy metabolism [36]. Similar findings have been reported at the regional level in the human cortex, with integrative hub regions (with high inter-modular degree and long connection distance) being associated with expression of genes regulating oxidative metabolism and mitochondria [43]. Both results are consistent with the high metabolic cost of hubs [37,38], which may be due to the high signal traffic they’re thought to mediate [32,34], painting a picture of hubs as a highly conserved and metabolically costly feature of structural brain networks.

[[TODO: add details on lineage]] While the relationship between cell lineage and gene expression was studied in other species [87,99].

LIMITATIONS: Compared to continuous expression measurements of > 17 000 genes in the mouse brain [97] or > 20 000 genes in the human brain [100], which permit more detailed analysis [36,43,67,88,89,98], the key challenges of working with *C. elegans* expression data are: (i) incomplete (< 5%) coverage of the full worm genome (which contains > 20 000 protein-coding genes [55]), (ii) binary rather than continuous expression indicators, and (iii) an inability to distinguish missing data from lack of expression.

- Did the best we had with sparsely-annotated binary gene expression data
- No way of discriminating between missing data and the absence of expression
- only around 5 percent of genes in the genome available
- annotation problems: different qualifiers, loosing sensitivity/specificity if including too much or too little - need to balance
- Future work may have actual high-throughput RNAseq gene data for all neurons, could do this analysis properly!
- Would surely allow more subtle understanding of how gene expression varies with connectivity roles

Compared to continuous expression information provided from techniques like RNA sequencing, *in situ* hybridization, or microarray with whole-brain coverage using a systematic experimental procedure [97,101,102], the data analyzed here presents numerous challenges. With access to only a relatively limited coverage of the genome (948 out of >19 000 genes [?] annotated in the current dataset) [[TODO: how many worm genes in total? Add proportion of worm genes represented “19,735 protein-coding genes—with >90% directly supported by experimental evidence—and >1300 noncoding RNA genes [?]”]], the enrichment results provided here are severely constrained in scope.

Further, using the current data, we are unable to distinguish expression levels between neurons that express a given gene, and we are unable to distinguish missing data, from negative data (both are represented as '0' in our data). Future work systematically mapping gene expression in a unified experimental paradigm would allow a more subtle and comprehensive analysis of gene expression and function.

Thus, despite minimal gene coverage (5% of the total genome), we developed a new method that was able to find interpretable patterns in the existing binary expression data for *C. elegans*, providing clues as to function of genes that may drive molecular differences between topologically different pairs of neurons.

Our results are broadly consistent with expectations. For example, as the majority of hubs are command interneurons, which control forward and backward locomotion, enrichment of genes related to locomotory behavior is sensible. Furthermore, although we couldn't test the exact gene categories from mouse or human, enrichment of related metabolic processes associated with connectivity and hub connectivity hints at an increase in energy demand as a function of connectivity.

Conclusion

Conclusion text.

Acknowledgments

Thanks to Alex Fornito for being a big deal. We thank WormBase.

References

- White JG, Southgate E, Thomson JN. The structure of the nervous system of the nematode *Caenorhabditis elegans*. Trans R Soc Lond B Biol 1986;;
- Varshney LR, Chen BL, Paniagua E, Hall DH, Chklovskii DB. Structural properties of the *Caenorhabditis elegans* neuronal network. PLoS computational biology. 2011;7(2):e1001066. doi:10.1371/journal.pcbi.1001066.
- Chiang AS, Lin CY, Chuang CC, Chang HM, Hsieh CH, Yeh CW, et al. Three-dimensional reconstruction of brain-wide wiring networks in drosophila at single-cell resolution. Curr Biol. 2011;21(1):1–11.
- Shih CT, Sporns O, Yuan SL, Su TS, Lin YJ, Chuang CC, et al. Connectomics-Based Analysis of Information Flow in the Drosophila Brain. Curr Biol. 2015;.
- Wanner AA, Genoud C, Masudi T, Siksou L, Friedrich RW. Dense EM-based reconstruction of the interglomerular projectome in the zebrafish olfactory bulb. Nat Neurosci. 2016;19(6):816–825.
- Hildebrand DGC, Cicconet M, Torres RM, Choi W, Quan TM, Moon J, et al. Whole-brain serial-section electron microscopy in larval zebrafish. Nature. 2017;545(7654):345–349.
- Oh SW, Harris JA, Ng L, Winslow B, Cain N, Mihalas S, et al. A mesoscale connectome of the mouse brain. Nature. 2014;508(7495):207–214. doi:10.1038/nature13186.

8. Zingg B, Hintiryan H, Gou L, Song MY, Bay M, Bienkowski MS, et al. Neural Networks of the Mouse Neocortex. *Cell.* 2014;156(5):1096–1111.
9. Markov NT, Ercsey-Ravasz MM, Gomes ARR, Lamy C, Magrou L, Vezoli J, et al. A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex. *Cereb Cortex.* 2012;24(1):bhs270–36.
10. Hagmann P, Cammoun L, Gigandet X, Meuli R, Honey CJ, Wedeen VJ, et al. Mapping the Structural Core of Human Cerebral Cortex. *PLoS Biol.* 2008;6(7):e159.
11. Bullmore ET, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci.* 2009;10(3):186–198.
12. Bullmore ET, Sporns O. The economy of brain network organization. *Nat Rev Neurosci.* 2012;13:337–349.
13. Fornito A, Zalesky A, Bullmore E. Fundamentals of brain network analysis. Academic Press; 2016.
14. Meunier D, Lambiotte R, Bullmore ET. Modular and Hierarchically Modular Organization of Brain Networks. *Front Psychiat.* 2010;4:1–11.
15. Bassett DS, Greenfield DL, Meyer-Lindenberg A, Weinberger DR, Moore SW, Bullmore ET. Efficient Physical Embedding of Topologically Complex Information Processing Networks in Brains and Computer Circuits. *PLoS computational biology.* 2010;6(4):e1000748. doi:10.1371/journal.pcbi.1000748.
16. Betzel RF, Avena-Koenigsberger A, Goñi J, He Y, de Reus MA, Griffa A, et al. Generative models of the human connectome. *NeuroImage.* 2016;124:1054–1064.
17. Bassett DS, Greenfield DL, Meyer-Lindenberg A, Weinberger DR, Moore SW, Bullmore ET. Efficient Physical Embedding of Topologically Complex Information Processing Networks in Brains and Computer Circuits. *PLoS Comp Biol.* 2010;6(4):e1000748.
18. van den Heuvel MP, Sporns O. Network hubs in the human brain. *Trends in Cognitive Sciences.* 2013;17(12):683–696.
19. Varshney LR, Chen BL, Paniagua E, Hall DH, Chklovskii DB. Structural Properties of the Caenorhabditis elegans Neuronal Network. *PLoS Comp Biol.* 2011;7(2):e1001066.
20. van den Heuvel MP, Sporns O. Rich-Club Organization of the Human Connectome. *J Neurosci.* 2011;31(44):15775–15786.
21. Zamora-López G, Zhou C, Kurths J. Cortical Hubs Form a Module for Multisensory Integration on Top of the Hierarchy of Cortical Networks. *Front Neuroinf.* 2010;4:1.
22. de Reus MA, van den Heuvel MP. Rich Club Organization and Intermodule Communication in the Cat Connectome. *J Neurosci.* 2013;33(32):12929–12939.
23. Towlson EK, Vértes PE, Ahnert SE. The rich club of the *C. elegans* neuronal connectome. *J Neurosci.* 2013;33(15):6380–6387.
24. Henderson JA, Robinson PA. Relations Between the Geometry of Cortical Gyrification and White-Matter Network Architecture. *Brain Conn.* 2014;4(2):112–130.

25. Roberts JA, Perry A, Lord AR, Roberts G, Mitchell PB, Smith RE, et al. The contribution of geometry to the human connectome. *NeuroImage*. 2016;124:379–393.
26. Horvát S, Gămănuț R, Ercsey-Ravasz M, Magrou L, Gămănuț B, Van Essen DC, et al. Spatial Embedding and Wiring Cost Constrain the Functional Layout of the Cortical Network of Rodents and Primates. *PLoS Biol*. 2016;14(7):e1002512.
27. Fornito A, Zalesky A, Breakspear MJ. The connectomics of brain disorders. *Nat Rev Neurosci*. 2015;16(3):159–172.
28. van den Heuvel MP, Sporns O. An Anatomical Substrate for Integration among Functional Networks in Human Cortex. *Journal of Neuroscience*. 2013;33(36):14489–14500. doi:10.1523/JNEUROSCI.2128-13.2013.
29. Crossley NA, Mechelli A, Scott J, Carletti F, Fox PT, McGuire P, et al. The hubs of the human connectome are generally implicated in the anatomy of brain disorders. *Brain: a journal of neurology*. 2014;137(Pt 8):2382–95. doi:10.1093/brain/awu132.
30. Crossley NA, Mechelli A, Vértes PE, Winton-Brown TT, Patel AX, Ginestet CE, et al. Cognitive relevance of the community structure of the human brain functional coactivation network. *Proc Natl Acad Sci USA*. 2013;110(28):11583–11588.
31. de Reus MA, Saenger VM, Kahn RS, van den Heuvel MP. An edge-centric perspective on the human connectome: link communities in the brain. *Phil Trans Roy Soc B*. 2014;369(1653):20130527–20130527.
32. van den Heuvel MP, Kahn RS, Goñi J, Sporns O. High-cost, high-capacity backbone for global brain communication. *Proc Natl Acad Sci USA*. 2012;109(28):11372–11377.
33. Harriger L, van den Heuvel MP, Sporns O. Rich Club Organization of Macaque Cerebral Cortex and Its Role in Network Communication. *PLoS ONE*. 2012;7(9):e46497.
34. Mišić B, Sporns O, McIntosh AR. Communication Efficiency and Congestion of Signal Traffic in Large-Scale Brain Networks. *PLoS Comp Biol*. 2014;10(1):e1003427.
35. Mišić B, Betzel RF, Nematzadeh A, Goñi J, Griffa A, Hagmann P, et al. Cooperative and Competitive Spreading Dynamics on the Human Connectome. *Neuron*. 2015;86(6):1518–1529.
36. Fulcher BD, Fornito A. A transcriptional signature of hub connectivity in the mouse connectome. *Proc Natl Acad Sci USA*. 2016;113(5):1435–1440.
37. Tomasi D, Wang GJ, Volkow ND. Energetic cost of brain functional connectivity. *Proc Natl Acad Sci USA*. 2013;110(33):13642–13647.
38. Collin G, Sporns O, Mandl RCW, van den Heuvel MP. Structural and Functional Aspects Relating to Cost and Benefit of Rich Club Organization in the Human Cerebral Cortex. *Cereb Cortex*. 2014;24(9):2258–2267.
39. Liang X, Zou Q, He Y, Yang Y. Coupling of functional connectivity and regional cerebral blood flow reveals a physiological basis for network hubs of the human brain. *Proceedings of the National Academy of Sciences*. 2013;110(5):1929–34. doi:10.1073/pnas.1214900110.

40. Fornito A, Zalesky A, Breakspear M. The connectomics of brain disorders. *Nature Reviews Neuroscience*. 2015;16(3):159–172. doi:10.1038/nrn3901.
41. Lennie P. The Cost of Cortical Computation. *Curr Biol*. 2003;13(6):493–497.
42. Laughlin SB, Sejnowski TJ. Communication in neuronal networks. *Science*. 2003;.
43. Vértes PE, Rittman T, Whitaker KJ, Romero-Garcia R, Váša F, Kitzbichler MG, et al. Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. *Philosophical Transactions of the Royal Society B*. 2016;371(1705):735–769. doi:10.1098/rstb.2015.0362.
44. Riddle DL, Blumenthal T, Meyer BJ, Priess JR. *C. elegans* II; 1997. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK19997/>.
45. Chen BL, Hall DH, Chklovskii DB. Wiring optimization can relate neuronal structure and function. *Proc Natl Acad Sci USA*. 2006;103(12):4723–4728.
46. Altun ZF, Herndon LA, Wolkow CA, Crocker C, Lints R, Hall DH. WormAtlas; 2002–2017. Available from: <http://www.wormatlas.org>.
47. Kashtan N, Itzkovitz S, Milo R, Alon U. Topological generalizations of network motifs. *Phys Rev E*. 2004;70(3):031909.
48. Azulay A, Itskovits E, Zaslaver A. The *C. elegans* Connectome Consists of Homogenous Circuits with Defined Functional Roles. *PLoS Comp Biol*. 2016;12(9):e1005021.
49. Kim S, Kim H, Kralik JD, Jeong J. Vulnerability-Based Critical Neurons, Synapses, and Pathways in the *Caenorhabditis elegans* Connectome. *PLoS Comp Biol*. 2016;12(8):e1005084.
50. Kim JS, Kaiser M. From *Caenorhabditis elegans* to the human connectome: a specific modular organization increases metabolic, functional and developmental efficiency. *Phil Trans Roy Soc B*. 2014;369(1653):20130529–20130529.
51. Pavlovic DM, Vértes PE, Bullmore ET, Schafer WR, Nichols TE. Stochastic Blockmodeling of the Modules and Core of the *Caenorhabditis elegans* Connectome. *PLoS ONE*. 2014;9(7):e97584.
52. Pan RK, Chatterjee N, Sinha S. Mesoscopic Organization Reveals the Constraints Governing *Caenorhabditis elegans* Nervous System. *PLoS ONE*. 2010;5(2):e9240.
53. Choe Y, McCormick B, Koh W. Network connectivity analysis on the temporally augmented *C. elegans* web: A pilot study. In: *Soc Neurosci Abstr*. vol. 30; 2004.
54. Varier S, Kaiser M, Koh W, Meuli R, Honey C. Neural Development Features: Spatio-Temporal Development of the *Caenorhabditis elegans* Neuronal Network. *PLoS Computational Biology*. 2011;7(1):e1001044. doi:10.1371/journal.pcbi.1001044.
55. Harris TW, Antoshechkin I, Bieri T, Blasius D, Chan J, Chen WJ, et al. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research*. 2009;38:D463–D467.

56. Pereira L, Kratsios P, Serrano-Saiz E, Sheftel H, Mayo AE, Hall DH, et al. A cellular and regulatory map of the cholinergic nervous system of *C. elegans*. *eLife*. 2015;4:e12432.
57. Sulston JE, Horvitz HR. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Developmental biology*. 1977;56(1):110–56.
58. Sulston JE, Schierenberg E, White JG, Thomson JN. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental biology*. 1983;100(1):64–119.
59. Colizza V, Flammini A, Serrano M, Vespignani A. Detecting rich-club ordering in complex networks. *Nat Phys*. 2006;2(2):110–115.
60. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science*. 2002;296:910–913.
61. Rubinov M, Sporns O. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*. 2010;52(3):1059–1069.
62. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech*. 2008;2008(10):P10008.
63. Rubinov M, Sporns O. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*. 2010;52(3):1059–1069. doi:10.1016/j.neuroimage.2009.10.003.
64. Lancichinetti A, Fortunato S, Thomson JN, Brenner S, Moon S. Consensus clustering in complex networks. *Scientific Reports*. 2012;2:547–579. doi:10.1038/srep00336.
65. Pavlovic DM, Vértes PE, Bullmore ET, Schafer WR, Nichols TE. Stochastic blockmodeling of the modules and core of the *Caenorhabditis elegans* connectome. *PLoS ONE*. 2014;9(7):e97584. doi:10.1371/journal.pone.0097584.
66. Warrens MJ. On Association Coefficients for 2x2 Tables and Properties That Do Not Depend on the Marginal Distributions. *Psychometrika*. 2008;73(4):777–789. doi:10.1007/s11336-008-9070-3.
67. Parkes L, Fulcher BD, Yücel M, Fornito A. Transcriptional signatures of connectomic subregions of the human striatum. *Genes, Brain and Behavior*. 2017;25:1176.
68. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B*. 1995;57(1):289–300.
69. Gillis J, Mistry M, Pavlidis P. Gene function analysis in complex data sets using ErmineJ. *Nature protocols*. 2010;5(6):1148–59. doi:10.1038/nprot.2010.78.
70. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for The Unification of Biology. *NATURE GENETICS*. 2000;25(1):25–29. doi:10.1038/75556.
71. Zoubarev A, Hamer KM, Keshav KD, Luke McCarthy E, Santos JRC, Van rossum T, et al. Gemma: A resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*. 2012;28(17):2272–2273. doi:10.1093/bioinformatics/bts430.

72. Krienen FM, Yeo BTT, Ge T, Buckner RL, Sherwood CC. Transcriptional profiles of supragranular-enriched genes associate with corticocortical network architecture in the human brain. *Proceedings of the National Academy of Sciences*. 2016;113(4):E469–78. doi:10.1073/pnas.1510903113.
73. Pantazatos SP, Li X. Commentary: BRAIN NETWORKS. Correlated gene expression supports synchronous activity in brain networks. *Science* 348, 1241-4. bioRxiv. 2016; p. 079202.
74. Richiardi J, Altmann A, Greicius M. Distance Is Not Everything In Imaging Genomics Of Functional Networks: Reply To A Commentary On Correlated Gene Expression Supports Synchronous Activity In Brain Networks. bioRxiv. 2017; p. 132746.
75. Wang XJ, Kennedy H. Brain structure and dynamics across scales: in search of rules. *Current Opinion in Neurobiology*. 2016;37:92–98.
76. Markov NT, Ercsey-Ravasz M, Van Essen DC, Knoblauch K, Toroczkai Z, Kennedy H. Cortical High-Density Counterstream Architectures. *Science*. 2013;342(6158):1238406–1238406.
77. Roberts JA, Perry A, Lord AR, Roberts G, Mitchell PB, Smith RE, et al. The contribution of geometry to the human connectome. *NeuroImage*. 2016;124(Pt A):379–93. doi:10.1016/j.neuroimage.2015.09.009.
78. Goulas A, Uylings HBM, Hilgetag CC. Principles of ipsilateral and contralateral cortico-cortical connectivity in the mouse. *Brain Struct Funct*. 2016;252:1–15.
79. Krienen FM, Yeo BTT, Ge T, Buckner RL, Sherwood CC. Transcriptional profiles of supragranular-enriched genes associate with corticocortical network architecture in the human brain. *Proc Natl Acad Sci USA*. 2016;113(4):E469–78.
80. Tsali EL, Hobert O. Functional mapping of neurons that control locomotory behavior in *Caenorhabditis elegans*. *Journal of neurobiology*. 2003;56(2):178–197.
81. Opsahl T, Colizza V, Panzarasa P, Ramasco JJ. Prominence and control: the weighted rich-club effect. *Physical review letters*. 2008;101(16):168702. doi:10.1103/PhysRevLett.101.168702.
82. Kim JS, Kaiser M. From *Caenorhabditis elegans* to the human connectome: a specific modular organization increases metabolic, functional and developmental efficiency. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2014;369(1653). doi:10.1098/rstb.2013.0529.
83. Achacoso T, Yamamoto W. AY's Neuroanatomy of *C. elegans* for Computation. CRC Press, Boca Raton, FL; 1992.
84. Richiardi J, Altmann A, Milazzo AC, Chang C, Chakravarty MM, Banaschewski T, et al. Correlated gene expression supports synchronous activity in brain networks. *Science*. 2015;348(6240):1241–1244. doi:10.1126/science.1255905.
85. van den Heuvel MP, Sporns O. An Anatomical Substrate for Integration among Functional Networks in Human Cortex. *J Neurosci*. 2013;33(36):14489–14500.
86. Schröter M, Paulsen O, Bullmore E. Micro-connectomics: probing the organization of neuronal networks at the cellular scale. *Nat Rev Neurosci*. 2017; p. 1–16.

87. Cui Q, Yu Z, Purisima EO, Wang E, Li X, Schwarz P, et al. MicroRNA regulation and interspecific variation of gene expression. *Trends in Genetics*. 2007;23(8):372–375. doi:10.1016/j.tig.2007.04.003.
88. Ji S, Fakhry A, Deng H. Integrative analysis of the connectivity and gene expression atlases in the mouse brain. *NeuroImage*. 2014;84:245–253.
89. Fakhry A, Ji S. High-resolution prediction of mouse brain connectivity using gene expression patterns. *Methods*. 2015;73:71–78.
90. French L, Pavlidis P. Relationships between gene expression and brain wiring in the adult rodent brain. *PLoS Comp Biol*. 2011;7(1):e1001049.
91. Ypma RJF, Bullmore E. Statistical Analysis of Tract-Tracing Experiments Demonstrates a Dense, Complex Cortical Network in the Mouse. *PLoS Comp Biol*. 2016;12(9):e1005104–22.
92. Liu X, Long F, Peng H, Aerni SJ, Jiang M, Sánchez-Blanco A, et al. Analysis of Cell Fate from Single-Cell Gene Expression Profiles in *C. elegans*. *Cell*. 2009;139(3):623–633. doi:10.1016/j.cell.2009.08.044.
93. van den Heuvel MP, Yeo BTT. A Spotlight on Bridging Microscale and Macroscale Human Brain Architecture. *Neuron*. 2017;93(6):1248–1251.
94. Varadan V, Miller III DM, Anastassiou D. Computational inference of the molecular logic for synaptic connectivity in *C. elegans*. *Bioinformatics*. 2006;22(14):e497–e506.
95. Kaufman A, Dror G, Meilijson I, Ruppin E. Gene expression of *Caenorhabditis elegans* neurons carries information on their synaptic connectivity. *PLoS computational biology*. 2006;2(12):e167. doi:10.1371/journal.pcbi.0020167.
96. Baruch L, Itzkovitz S, Golan-Mashiach M, Shapiro E, Segal E. Using Expression Profiles of *Caenorhabditis elegans* Neurons To Identify Genes That Mediate Synaptic Connectivity. *PLoS computational biology*. 2008;4(7):e1000120. doi:10.1371/journal.pcbi.1000120.
97. Lein E, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. 2007;445(7124):168–176.
98. French L, Pavlidis P. Relationships between gene expression and brain wiring in the adult rodent brain. *PLoS computational biology*. 2011;7(1):e1001049. doi:10.1371/journal.pcbi.1001049.
99. Kluger Y, Tuck DP, Chang JT, Nakayama Y, Poddar R, Kohya N, et al. Lineage specificity of gene expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101(17):6508–6513. doi:10.1073/pnas.0401136101.
100. Hawrylycz MJ, Lein E, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*. 2012;489(7416):391–399.
101. Shen EH, Overly CC, Jones AR. The Allen Human Brain Atlas: comprehensive gene expression mapping of the human brain. *Trends Neurosci*. 2012;32(12):711–714.

102. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci*. 2016;19(2):335–346.

Supporting information

Expression annotations from WormBase

Neuronal gene expression is measured as a binary indicator on WormBase [55], based on curated data collated from many individual experiments.

Expression annotations are made either ‘directly’ to individual neurons (when an experiment indicates expression in an individual neuron), or ‘indirectly’ to broader classes of neurons like ‘interneuron’ or ‘head’ (meaning that some members of that class exhibit expression of that gene). In order to maintain specificity of annotations, we only analyzed ‘direct’ annotations here.

Annotations of gene G to neuron N are also made with varying levels of certainty: ‘certain’ (G was observed to be expressed in N), ‘enriched’ (G has been found to be enriched in a certain dataset through microarray, RNAseq or proteomic analysis), ‘partial’ (G was observed to be expressed in some cells of a group of neurons that include N), ‘blank’ (data from studies before 2005), or ‘uncertain’ (G was sometimes observed to be expressed in N , or G was observed to be expressed in a cell that could be N). Our analysis excluded annotations labeled as ‘uncertain’ to avoid including false positives.

Gene coexpression sensitivity

Measures of correlation between binary vectors can be biased by the relative proportion of ones between two vectors. Individual neurons range from between 3 to 138 expressed genes (0.3% to 14.6% of all 948 genes considered). To ensure that our measure of coexpression is not biased by differences in the relative proportions of expression annotations, we conducted a sensitivity analysis in which we compared the r_ϕ metric, Eq. (3), with alternative methods for quantifying correlations between binary vectors: the Jaccard index, $n_{11}/(n_{10} + n_{01} + n_{11})$, Yule’s Q coefficient, $(n_{00}n_{11} - n_{01}n_{10})/(n_{00}n_{11} + n_{01}n_{10})$, and the χ^2 index, $N(n_{00}n_{11} - n_{01}n_{10})/(n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1})$ [95], where n_{xy} counts the number of observations of each of the four binary pairwise possibilities: n_{00} , n_{01} , n_{10} , and n_{11} (as outlined in the main text), while the symbol \bullet sums across a given variable (e.g., $n_{\bullet 0} = n_{00} + n_{10}$).

To evaluate each measure, we generated random binary vectors of length 948 containing different proportions of 1s seen in our data, ranging from the minimum, 1, to a maximum, 148. For all pairwise combinations of proportions, we computed the coexpression measure, taking an average across 1 000 permutations, and then recorded the resulting mean correlation value, as plotted in Fig. S1. Because all vectors are independent random binary strings, the mean coexpression value should not show any systematic variation, which would indicate bias. The mean square contingency coefficient, r_ϕ (Fig. S1A) and our developed coexpression matching index (Fig. S1D) show no systematic bias to the proportion of ones in the vector (varying within $\approx 10^{-3}$ and ≈ 0.5 respectively), whereas Yule’s Q shows a negative bias for small annotation proportions (Fig. S1B), while the Jaccard index shows a strong positive bias across the full range (Fig. S1C). Based on these numerical experiments, we selected Pearson’s mean square contingency coefficient, r_ϕ , here, to ensure that changes in coexpression were due to matching structure and not simply the number of gene annotations.

Coexpression matching index

Existing measures of binary correlation (described below) are symmetric between ‘0’ and ‘1’ and thus do not directly distinguish between the biologically relevant case of genes being expressed together from the case in which neither gene is expressed. We thus developed a measure of the probability, $P(m)$, that two binary strings will contain

m ‘positive matches’ (i.e., m genes are expressed in both neurons), which can be computed as:

$$P(m) = \binom{n_2}{m} \binom{N - n_2}{n_1 - m} / \binom{N}{n_1}, \quad (5)$$

for two binary expression vectors, x_i, y_i , of length N , containing n_1 and n_2 1s, respectively ($n_2 \leq n_1$), with m matches (n_{11}). Our index, p_{match} , is approximately equal to the probability of obtaining as many or fewer matches than observed (weighting the probability of the observed number of matches at 0.5 for symmetry), computed as:

$$p_{\text{match}} = \sum_{x=0}^{m-1} P(x) + P(m)/2, \quad (6)$$

for m positive matches. We verified that this index yields qualitatively similar results to the mean square contingency coefficient, r_ϕ , used throughout this work, validating our related positive match method for scoring individual genes (as an approximately single-gene contribution to the probability score computed in Eq. 6 above). Given the qualitative similarity of this measure to r_ϕ , we chose to focus on r_ϕ throughout this work due to its ease of interpretability as a correlation coefficient ranging from -1 to 1.

Binary enrichment scoring

In this section we provide more information about how genes were scored in the connectivity and hub connectivity analyses performed in this work. When comparing pairs of neurons that contain a structural connection to pairs of neurons that do not share a connection, $p_{\text{class}} = 0.059$ is the proportion of neuron pairs that are connected, n is the total number of neuron pairs that both exhibit expression of gene a , and m is the number of neuron pairs that are structurally connected for which both neurons express gene a . Then, we compare pairs of connected neurons for which at least one is a hub, to pairs in which both neurons are nonhubs. In this case, $p_{\text{class}} = 0.35$ is the proportion of connected pairs of neurons that involve hubs, n is the number of connected neuron pairs for which gene a is expressed in both, and m counts the number of connected neuron pairs involving hubs for which gene a is expressed. [[TODO-AA: Aurina – can you add the electrical/chemical stuff to this section?]] The data quality criterion of $n \geq 10$ matches was satisfied by 414 (/948) genes for the connected/unconnected analysis and 168 genes for hub connectivity analysis.

S1 Table Hub neurons of the *C. elegans* connectome.

S2 Table Enrichment results for connected vs unconnected neurons.

S3 Table. Enrichment results for links involving hubs.

Table 1. Hub neurons of the *C. elegans* connectome. Hubs are defined as neurons with $k > 41$. For each hub, we list the neuron name, degree, k , location, and function. Entries have been sorted (descending) by degree.

Neuron	Degree, k	Description
AVAR	137	Head interneuron, role in locomotor decisions
AVAL	134	Head interneuron, role in locomotor decisions
AVBR	104	Head interneuron, role in locomotor decisions
AVBL	102	Head interneuron, role in locomotor decisions
PVCR	69	Tail interneuron, role in locomotor decisions
PVCL	64	Tail interneuron, role in locomotor decisions
AVDR	63	Head interneuron, role in locomotor decisions
AVER	63	Head interneuron, role in locomotor decisions
AVEL	62	Head interneuron, role in locomotor decisions
DVA	59	Head interneuron, mechanosensory integration
RIBL	56	Head interneuron
AVKL	53	Head interneuron
AVDL	52	Head interneuron, role in locomotor decisions
RIBR	52	Head interneuron
AIBR	49	Head interneuron
RIGL	49	Head interneuron

Table 2. Biological processes GO categories showing increased gene coexpression for connected neurons compared to unconnected neurons. Entries have been sorted by p -value (uncorrected).

GOcategory	Description	No. of genes	p (uncorr)	p (corr)
GO:0035235	ionotropic glutamate receptor signaling pathway	7	0.0005	0.3245
GO:0007215	glutamate receptor signaling pathway	9	0.0041	0.9367
GO:0007166	cell surface receptor signaling pathway	37	0.0048	0.9367
GO:0006811	ion transport	71	0.0058	0.9367
GO:0034220	ion transmembrane transport	57	0.0079	1
GO:0055085	transmembrane transport	67	0.0129	1
GO:1901575	organic substance catabolic process	33	0.031	1
GO:0040009	regulation of growth rate	7	0.0417	1
GO:0040010	positive regulation of growth rate	7	0.0417	1
GO:0030163	protein catabolic process	27	0.0428	1
GO:0009056	catabolic process	35	0.0493	1
GO:0009057	macromolecule catabolic process	28	0.0548	1
GO:0071495	cellular response to endogenous stimulus	14	0.0575	1
GO:0045927	positive regulation of growth	29	0.0689	1
GO:0050830	defense response to Gram-positive bacterium	5	0.0695	1

Table 3. Biological processes GO categories showing increased gene coexpression for links involving hubs (rich, feed-in and feed-out) compared to peripheral links. Categories have been sorted by *p*-value (uncorrected).

GOcategory	Description	No. of genes	p (uncorr)	p (corr)
GO:0007215	glutamate receptor signaling pathway	6	0.0009	0.1741
GO:0035235	ionotropic glutamate receptor signaling pathway	6	0.0009	0.1741
GO:0007166	cell surface receptor signaling pathway	20	0.0035	0.3473
GO:0009891	positive regulation of biosynthetic process	11	0.0063	0.3473
GO:0031328	positive regulation of cellular biosynthetic process	11	0.0063	0.3473
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	11	0.0063	0.3473
GO:0051173	positive regulation of nitrogen compound metabolic process	11	0.0063	0.3473
GO:0031325	positive regulation of cellular metabolic process	12	0.0109	0.5258
GO:0040012	regulation of locomotion	24	0.0156	0.5467
GO:0010557	positive regulation of macromolecule biosynthetic process	10	0.0212	0.5467
GO:0045893	positive regulation of transcription, DNA-templated	10	0.0212	0.5467
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	10	0.0212	0.5467
GO:0051254	positive regulation of RNA metabolic process	10	0.0212	0.5467
GO:1902680	positive regulation of RNA biosynthetic process	10	0.0212	0.5467
GO:1903508	positive regulation of nucleic acid-templated transcription	10	0.0212	0.5467

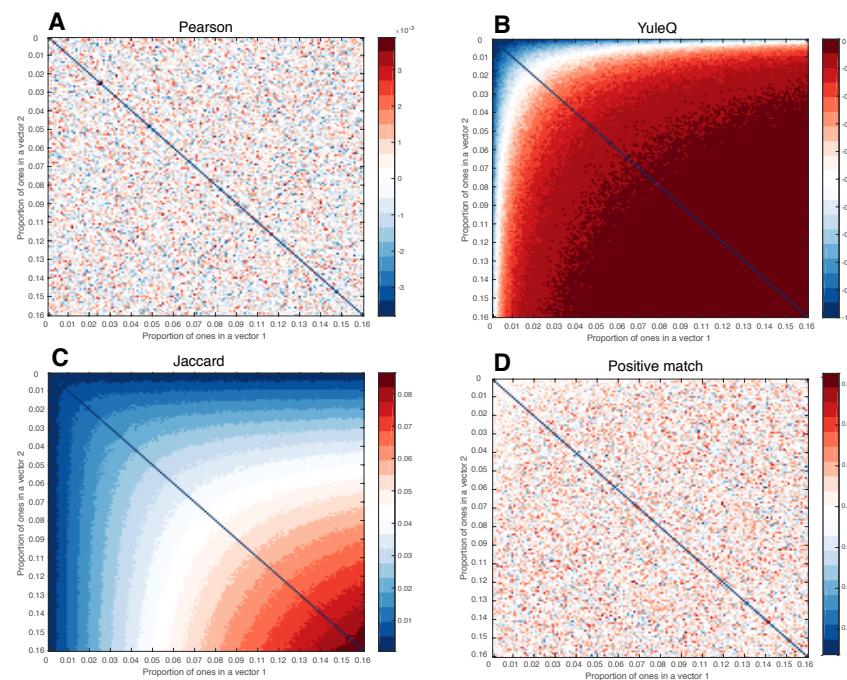


Fig S1. Dependence of binary coexpression metrics on the proportion of positive annotations. Shown here is the mean correlation between pairs of random, binary vectors of length 948, with different proportions of 1s (between 0–0.16; matrix size is 150×150 , therefore the number of ones ranges from 1 to 150). Each point is averaged over 1 000 different pairs of random vectors. Bias is indicated by a systematic trend in correlation values, driven simply by the proportion of positive annotations for a pair of vectors. **A** Pearson, r_ϕ , the measure used throughout this work as well as our developed positive match measure **D** shows no systematic bias (note the color axis scale, 10^{-3} for Pearson, r_ϕ and 0.5 for positive match measure). By contrast, the remaining measures to show a dependence on the proportion of 1s in each vector: **B** Jaccard, **C** Yule's Q .

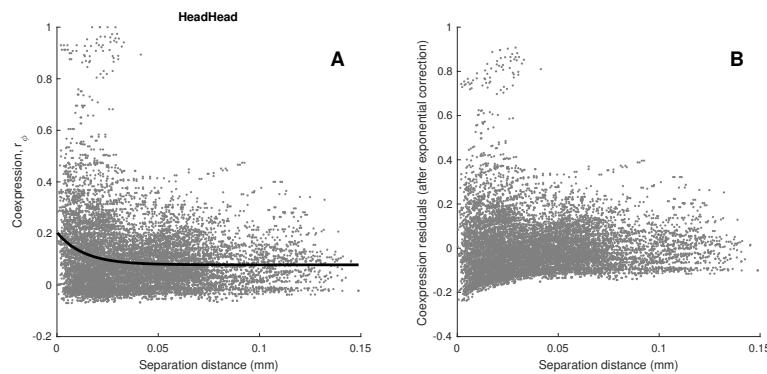


Fig S2. Applying an exponential trend correction to skewed coexpression data produces artifacts in the resulting corrected data. **A**, Coexpression values, r_ϕ , are plotted as a function of Euclidean separation distance for all pairs of neurons within the head (gray dots). An exponential trend fitted to the data is shown in black, $f(x) = A \exp(-\lambda x) + B$. The trend is not a bulk effect that holds in general across the brain, but is driven primarily by a small number of neuron pairs with high r_ϕ at short distances ($\lesssim 25\mu\text{m}$). Thus, taking residuals from this trend, shown in **B**, does not adequately correct the spatial trend.

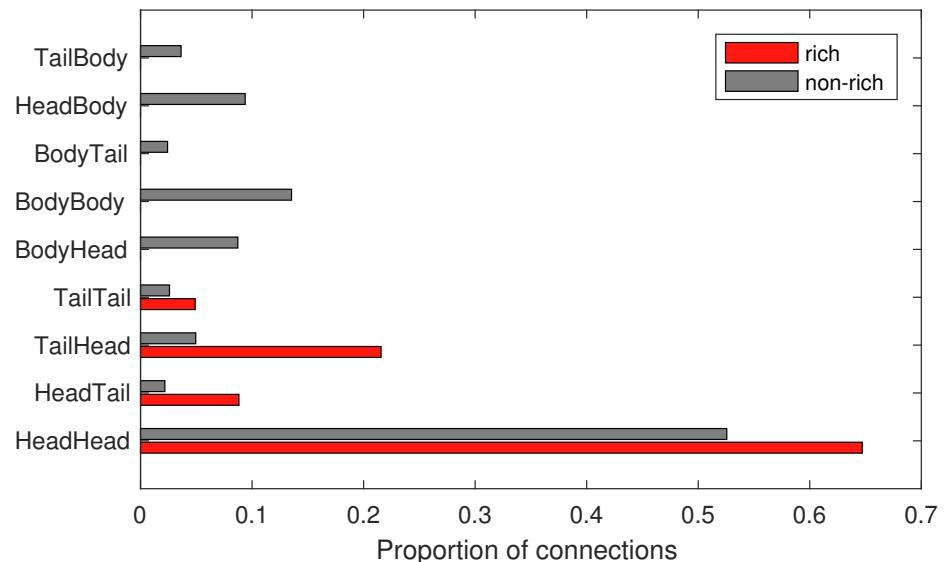


Fig S3. Distribution of hub-hub connections ('rich', red), and all other connections ('non-rich', i.e., feeder and peripheral, gray) across anatomical divisions (head, body, and tail), defining hubs as those with degree, $k > 44$, corresponding to the lowest threshold at which the *C. elegans* connectome exhibits rich-club organization. The increased separation distance between hubs is driven by a relative increase in connection density between the head and tail [[TODO-BF: is the plor updated? because the caption was not]].

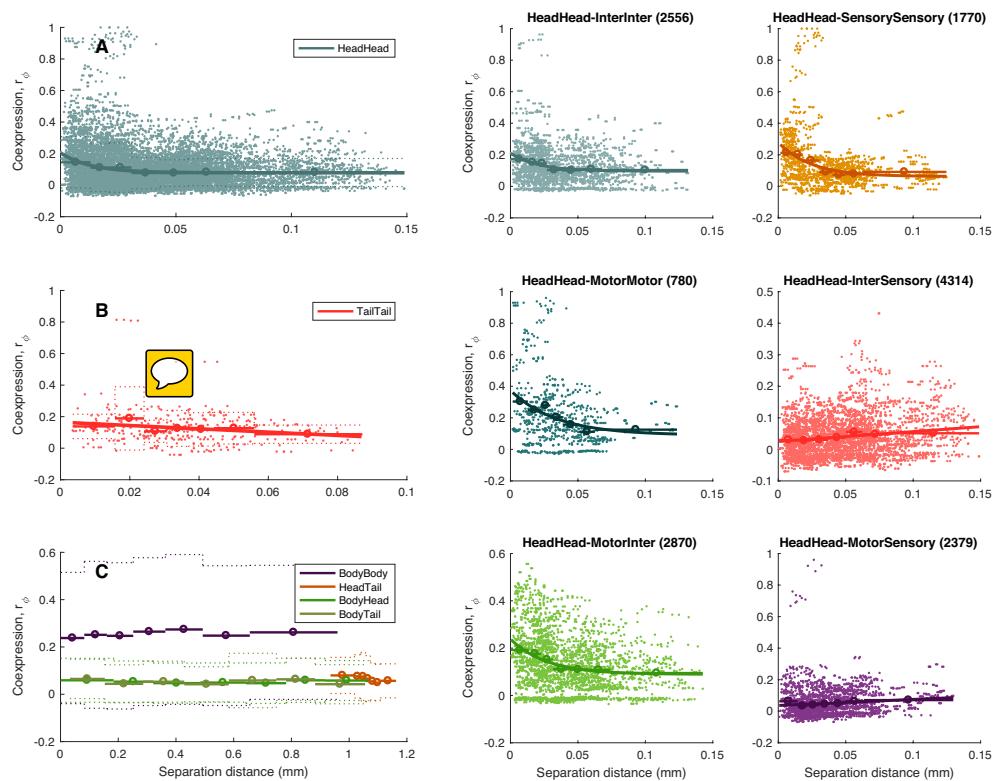


Fig S4. Gene coexpression decreases with separation distance within the head and tail. Gene coexpression, r_ϕ (in all cases excluding left-right homologous pairs of neurons), is shown as a function of the pairwise separation distance between pairs of neurons (shown as the mean (solid) \pm standard deviation (dotted) in seven equiprobable distance bins, with extent shown as horizontal bars), for **A** all pairs of neurons in the head, **B** all pairs of neurons in the tail, and **C** all other pairs (labeled). Scatters for all neuron pairs are added in **A** and **B**. An exponential relationship, $f(x) = a \exp(-bx) + c$, is fitted in **A** and **B**, revealing a slight decreasing trend in r_ϕ with distance, as per macroscopic mammalian brains [36, 79]. Looking within the head, we find different spatial relationships as a function of neuron type, shown for **C** pairs of interneurons, **D** pairs of motor neurons, **E** motor-interneuron pairs, **F** pairs of sensory neurons, **G** interneuron-sensory neuron pairs, and **H** motor-sensory pairs. [[TODO: add labels for subplots C–H, label all axes? Change axis label to Gene coexpression, r_ϕ]]

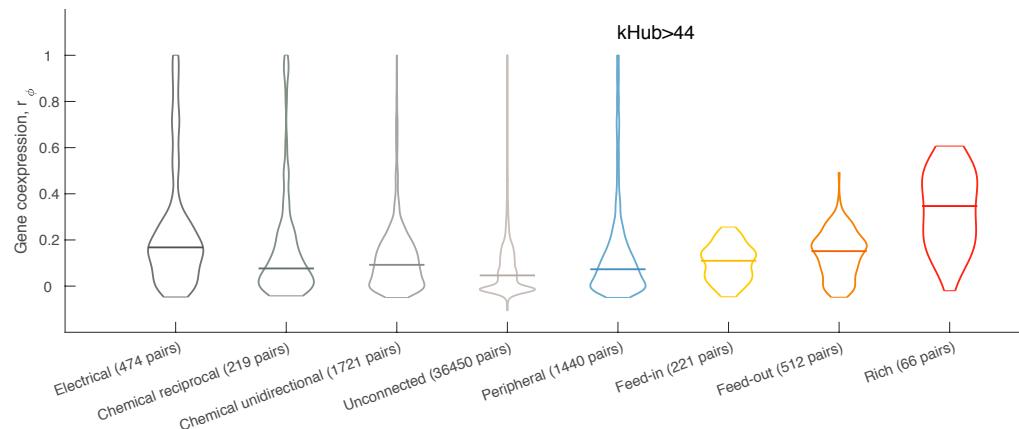


Fig S5. Coexpression for different types of connectome edges. *Left:* Distribution of transcriptional similarity between reciprocally connected (219 pairs), unidirectionally connected (1721) and unconnected (36 749 pairs) pairs of neurons as a violin plot, with the median of each distribution shown as a horizontal line. Connections between symmetrical (left/right) pairs of neurons are excluded from the analysis. **Coexpression values for Coexpression**, r_ϕ is increased in connected (reciprocally and unidirectionally) pairs of neurons ($p = 1.810^{-78}$, Wilcoxon rank sum test). Coexpression between neurons connected via gap junctions is significantly higher than between neurons connected via chemical synapses ($p = 5.410^{-22}$, Wilcoxon rank sum test) *Right:* Distribution of transcriptional similarity for rich, feed-in, feed-out and peripheral links, where hubs are neurons with degree $k > 44$. [[TODO: color inside of the distributions Add $p < 0.05$ values on the plot where significant.]]

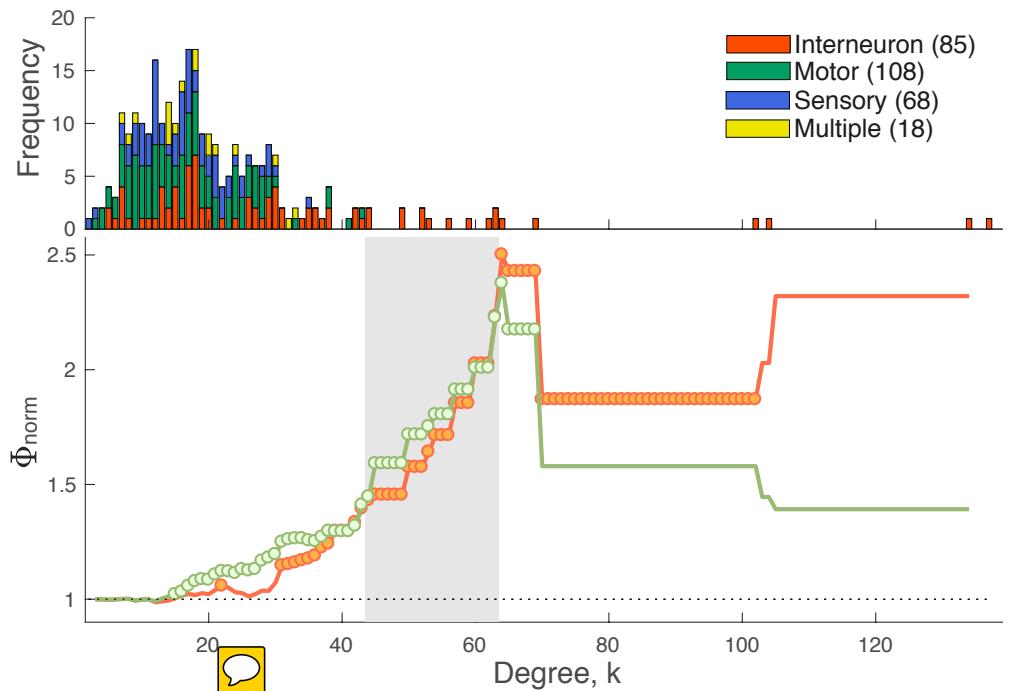


Fig S6. **A** Degree distribution of the binary synaptic connectome, where neurons are labeled according to four categories: (i) interneuron (85 neurons, orange), (ii) sensory (68 neurons, blue), (iii) motor (108 neurons, green), or (iv) multimodal (18 neurons, yellow). An extended tail of high-degree neurons can be seen, which are mostly interneurons. **B** Normalized weighted (topology fixed and weights randomized in the null model, shown orange) $\Phi_{\text{norm}}^{\text{weighted}}$ and mixed (both topology and weights mixed in the null model, shown green) $\Phi_{\text{norm}}^{\text{mixed}}$ rich club coefficient as a function of the degree, k , at which hubs are defined (as neurons with degree $> k$). Circles indicate values of Φ_{norm} that are significantly higher than an ensemble of 1 000 degree-matched null networks (Welch's t -test, $p < 0.05$). [[TODO: Change the labels within the plot for mixed and weighted as superscripts]] [[TODO: define/explain the two definitions and give references]]

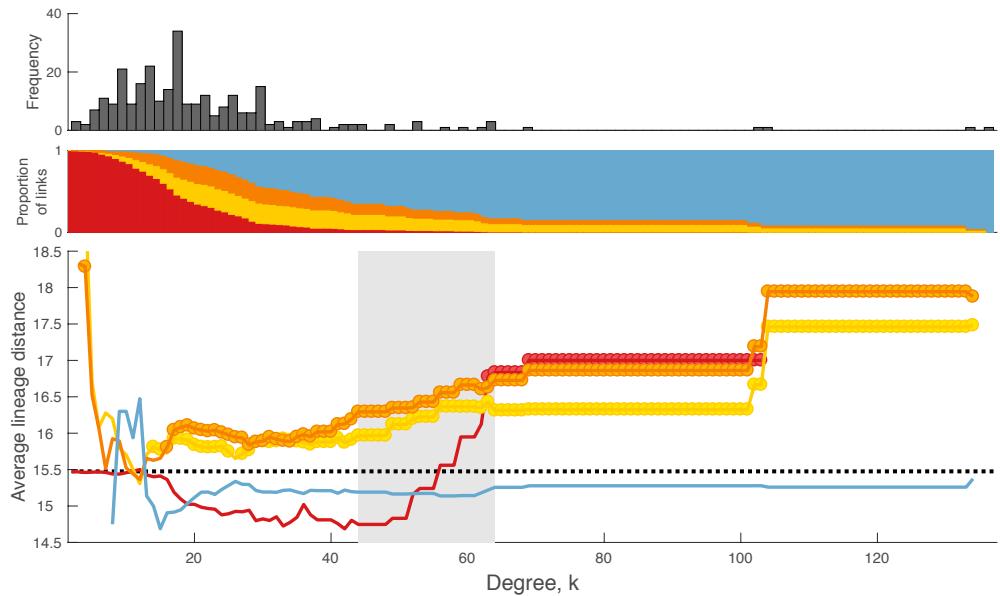


Fig S7. Lineage distance between pairs of neurons for different connection types as a function of the degree at which hubs are defined, k . *Top:* Degree distribution. *Middle:* proportion of connections that are ‘rich’ (hub→hub, red), ‘feed-in’ (nonhub→hub, yellow), ‘feed-out’ (hub→nonhub, orange), and ‘peripheral’ (nonhub→nonhub, blue) as a function of the degree threshold, k , used to define hubs. Note that at high k , most neurons are labeled as nonhubs, and hence the vast majority of connections are ‘peripheral’. *Lower:* Mean lineage distance for each connection type as a function of k . Lineage distance across all network links shown as a dotted black line; the topological rich-club regime (determined from the network topology, cf. Fig. 5) is shaded gray. Circles indicate a statistically significant increase or decrease in birth time difference in a given link type relative to the rest of the network (two-sided Welch’s t test; $p < 0.05$).

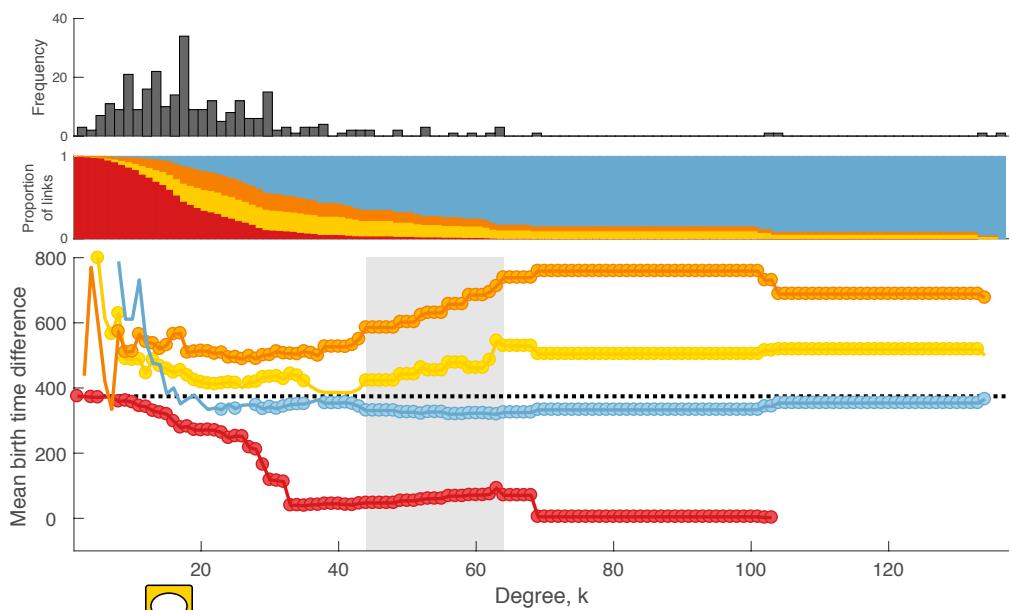


Fig S8. *Top:* Degree distribution. *Middle:* proportion of connections that are ‘rich’ (hub→hub, red), ‘feed-in’ (nonhub→hub, yellow), ‘feed-out’ (hub→nonhub, orange), and ‘peripheral’ (nonhub→nonhub, blue) as a function of the degree threshold, k , used to define hubs. Note that at high k , most neurons are labeled as nonhubs, and hence the vast majority of connections are ‘peripheral’. *Bottom:* Mean birth time difference for each connection type as a function of k . The mean birth time difference across all network links shown as a dotted black line; the topological rich-club regime (determined from the network topology, cf. Fig. 5) is shaded gray. Circles indicate a statistically significant increase or decrease in birth time difference in a given link type relative to the rest of the network (two-sided Welch’s t test; $p < 0.05$). 

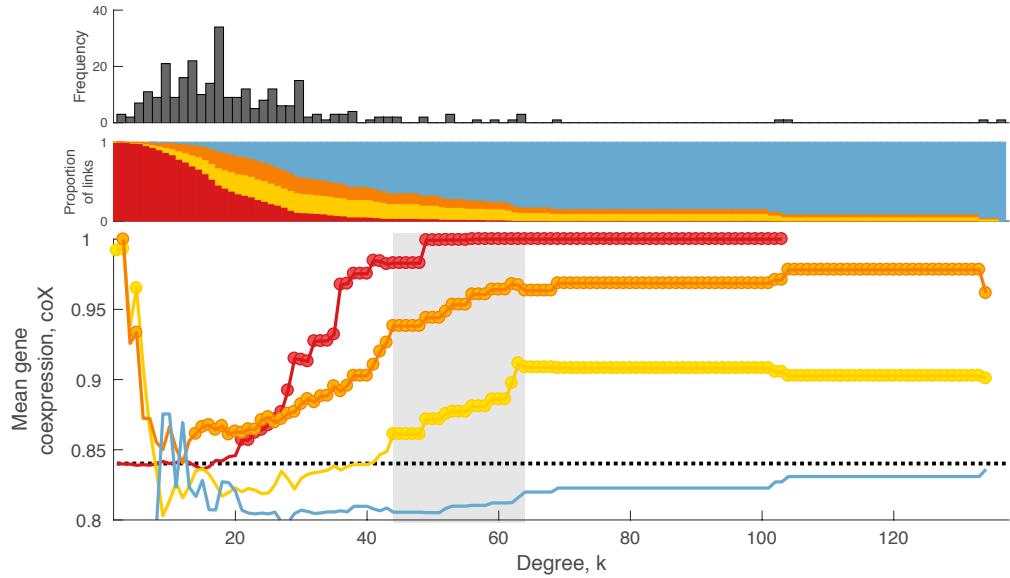


Fig S9. Gene coexpression, as measured by a positive matching probability index, Eq. (5). *Top:* Degree distribution. *Middle:* proportion of connections that are ‘rich’ (hub→hub, red), ‘feed-in’ (nonhub→hub, yellow), ‘feed-out’ (hub→nonhub, orange), and ‘peripheral’ (nonhub→nonhub, blue) as a function of the degree threshold, k , used to define hubs. Note that at high k , most neurons are labeled as nonhubs, and hence the vast majority of connections are ‘peripheral’. *Bottom:* Mean gene coexpression calculated using similarity index from only positive matches, r_ξ , for each connection type as a function of k . The mean coexpression across all network links shown as a dotted black line; the topological rich-club regime (determined from the network topology, cf. Fig. 5) is shaded gray. Circles indicate a statistically significant increase in gene coexpression in a given link type relative to the rest of the network (one-sided Welch’s t test; $p < 0.05$).