Gene Expression of C. elegans Neurons Carries Significant Information on Their Synaptic Connectivity

Alon Kaufman\textsuperscript{1}, Gideon Dror\textsuperscript{2} & Eytan Ruppin\textsuperscript{3}

\textsuperscript{1}Interdisciplinary Center for Neural Computation, Hebrew University, Jerusalem, 91904, Israel. \textsuperscript{2}The Academic College of Tel-Aviv-Yaffo, Tel Aviv, 64044, Israel. \textsuperscript{3}School of Computer Science and School of Medicine, Tel-Aviv University, Tel-Aviv, 69978, Israel.

Correspondence should be addressed to A.K. (akaufman@alice.nc.huji.ac.il) or to E.R (ruppin@post.tau.ac.il).

The claim that genetic properties of neurons significantly influence their synaptic network structure is a common notion in Neuroscience, which has been substantiated by specifically targeted experimental studies. The nematode C. elegans provides an exciting opportunity to approach this question in a large scale computational manner. Its synaptic connectivity network has been identified and, combined with cellular studies, we now have connectivity and gene expression "signatures" for most of its neurons. This study provides the first large-scale quantitative account of the relation between these two signatures. We show that the expression signature of a neuron carries significant information about its synaptic connectivity, and identify a list of putative genes predicting neural connectivity. Finally, in the amphid neuronal sub-network responsible for chemotaxis, we find that a neuron's expression signature plays a larger role than its connectivity signature in determining its functional role in chemotaxis to different attractants.

It is an accepted common notion that genes play a major role in the formation of the nervous system; they specify neuronal cell types, help destine neurons into defined neural circuits and provide important cues determining their communication\textsuperscript{1}. Many studies have identified specific genes in the nematode C. elegans that disrupt the development of neural circuits. These genes are typically responsible for neuronal
morphology, axon development and synaptogenesis. Such findings include, e.g., axon guidance genes (sax-3, unc-34 and the netrin receptor unc-40\textsuperscript{1,2}), attractive and repulsive interactions (unc-6, unc-40 and unc-5)\textsuperscript{3-5}, pre-synaptic input modulation (unc-4, unc-37)\textsuperscript{6}, pre-synaptic differentiation (sad-1)\textsuperscript{7} and synaptic specificity (syg-1, syg-2)\textsuperscript{8}. These findings have been based on specifically targeted studies, each designed to address a specific pathway, neuron type, receptor or transmitter (for reviews see\textsuperscript{9-13}). Yet, it has been difficult to identify mutations that determine the specific identity of synaptic connections (that is, to whom each neuron is connected), mainly because synaptic specification is one of the last steps in a complex process of neuronal differentiation and axonal migration\textsuperscript{10}. A recent study\textsuperscript{14} presented the first large scale screening for genes involved in the C. elegans neuromuscular junction. The study identified more than 100 novel genes that have specific functions in the transmission of signals across this junction. While latter study was not aimed at identifying synaptic connectivity genes, it has demonstrated the plausibility of addressing such questions in a large scale manner.

The current study presents the first large-scale quantitative investigation of the relation between the genetic properties of neurons and their synaptic connectivity, concomitantly addressing the majority of C. elegans neurons. The existing C. elegans neural wiring diagram provides a “connectivity signature” for each neuron, specifying to whom it is connected (R. M. Durbin (http://elegans.swmed.edu/parts/neurodata.txt), based on the classic work of White \textit{et al}.\textsuperscript{15} and Hall \textit{et al}.\textsuperscript{16}). Each neuron has also an “expression signature” extracted from WormBase (http://wormbase.org), specifying the genes directly associated with it (see Methods). Combined together, this data enables the investigation of the relation between expression and connectivity signatures across most of the C. elegans neurons. We specifically address two attributes of this relation: The first asks whether it is possible to predict a neuron's connectivity signature based solely on its expression signature. The second questions to what extent do neurons with
similar expression signatures have similar connectivity signatures. We show that the expression signatures of neurons carry significant information about their connectivity signatures and further identify specific genes that play a major role in determining this relation. The gene sets we identify do not necessarily have a direct causal influence on synaptic connectivity and specificity; however they provide putative gene targets for further experimental investigation.

Finally, in a small set of neurons responsible for chemotaxis in the worm we study the relations between their expression and connectivity signatures and the functional contribution of these neurons. Quantifying these relations, we address a classical question in Neuroscience; what dominates the functionality of a neural circuit – the local, genetic basis of the individual neurons, or the overall network structure determined by their connectivity.

RESULTS

For a majority of the C. elegans neurons we obtained two types of data signatures (Methods): 1. The gene expression signature, describing which genes have expression patterns which are directly associated with a neuron, according to WormBase, and 2. The connectivity signature, describing the outgoing and incoming synaptic connections of each neuron to all other neurons in the network (focusing only on synaptic connections in which the direction is well defined). To avoid a bias caused by the symmetric structure of the data (many of the neurons are situated bilaterally along the nematode body and head), we focused on the right side of the nematode (and including also neurons without a symmetrical companion), retaining 98 such neurons that had both an expression and a connectivity signature (Supplementary Table T1).
The natural starting point for investigating the relation between these two types of signatures revolves around two basic attributes: First, the prediction ability – that is, can the connectivity signatures be predicted based on the expression signatures? Second, a co-variation correlation assay – which essentially measures to what extent are the neighborhood relations between neurons in one space (e.g., expression) similar to their neighborhood relations in the other space (e.g., synaptic connectivity). To study the first, prediction question, we use a standard weighted k-nearest neighbors (KNN) prediction algorithm with multi class targets (see Methods). Based on the expression signature of each neuron, this algorithm predicts its connectivity signature. The resulting prediction accuracy is measured in a conventional manner by the average area under the ROC curve (AUC). The performance obtained was 0.594 and 0.601 in predicting the incoming and outgoing connectivity signatures respectively ($p$-value $=10^{-85}$, and $p$-value $=10^{-75}$ respectively). The predictor’s AUC is moderate, probably reflecting the crude data in hand, but nevertheless, it manifests a markedly statistically significant signal. To study the second question, we applied a co-variation correlation assay (see Methods) to the 98 neurons, finding a Pearson correlation of 0.075 ($p$-value <0.0001) between the gene expression neighborhood relations and the incoming connectivity neighborhood relations, and 0.176 ($p$-value <0.0001) between the expression neighborhood relations and the outgoing synaptic neighborhood relations. These low-magnitude but strongly significant correlations indicate that the neighborhood relations between neurons in the one space bear a moderate similarity to the neighborhood relations in the other space.

Neuron cell type probably plays a major role in determining the connectivity properties of the cell, as distinct cell type-specific properties are determined by the combinatorial functions of multiple transcription factors\textsuperscript{17-19}. Indeed, applying the prediction assay to predict the neuron type based on the expression signatures (where neurons are classified into a number of neuron types according to WormBase (Methods
and Supplementary Table T1)) shows a significant prediction capability (AUC=0.923, 
\( p\text{-value}= 10^{-20} \)). Consequently, the relation between expression and connectivity 
signatures was further examined by the prediction and co-variation correlation assays 
while controlling for specific information about cell types. The relations between the 
two signatures remain marked and significant, both for the incoming and outgoing 
signatures, with AUC=0.599 \( (p\text{-value}=10^{-67}) \) and AUC=0.611 \( (p\text{-value}=10^{-59}) \) in the 
prediction assays and correlations of 0.089 \( (p\text{-value}<0.0001) \) and 0.146 \( (p\text{-value}<0.0001) \) in the co-variation correlation assays (see Supplementary Note for the 
detailed method and results). Interestingly, while applying the co-variation assay to the 
sensory neurons (see Supplementary Note) their expression signatures show a 
significant correlation only with their outgoing synaptic connections (0.432, \( p\text{-value} <0.0001 \)). This may arise because of the absence of data from the sensory receptors, 
their main input sources (the connectivity data includes only connections within the 
neurons).

To identify the genes (features) which highly contribute to the connectivity's 
prediction accuracy and to the expression-connectivity co-variation correlation, an 
extensive feature selection process was performed (see Methods). These feature 
selection assays do not necessarily testify to causal and direct relations but do give rise 
to putative gene candidates for future experimental investigation studies. Supplementary 
Table T2 lists the genes selected in the prediction feature selection assays. Focusing on 
gene sets that provide the highest AUC performance in each of the connectivity 
prediction assays results in 53 genes that yield a predictor with an average AUC of 0.60 
\( (p\text{-value} =10^{-99}) \) for the incoming connections and 30 genes that yield a predictor with 
an average AUC of 0.61 \( (p\text{-value} =10^{-97}) \) for the outgoing connections, as shown in 
Figure 1. Results of the co-variation feature selection assay are shown in Figure 2. As 
the feature selection process used in the correlation co-variation assay is greedy, the 
procedure for feature selection using correlation co-variation is repeated 10 times, each
repetition applying the assay to a random set composed of 90% of the neurons (see Methods). Figure 2 presents the mean and standard deviations of these repetitions. The assay results in statistically significant feature sets, leading to a correlation of 0.252 ($p$-value<0.0001) between expression and incoming connectivity signatures and 0.368 ($p$-value=0.004) with the outgoing connectivity signatures (p-value calculations are described in the Methods). The final gene sets of the co-variation assay (listed in Supplementary Table T3) are produced by focusing only on the genes selected in all 10 repetitions of this assay. The latter results in sets of 12 genes for the incoming connectivity ($p$-value=0.04) and 52 genes for the outgoing connectivity ($p$-value=0.02). Evidently, the correlation obtained with respect to the outgoing signatures remains above that obtained with respect to the incoming signatures, testifying that the expression signatures in hand carry more information about outgoing synaptic patterns than about incoming ones.

The outcome of the feature selection process is a list of genes which bear significant information about the specific targets and sources of neuronal synaptic connectivity. Four such sets, generated for the two connectivity types by using the two assays, are obtained. To compare these gene lists to the contemporary knowledge, we compiled a list of genes known to be involved in neuronal connectivity in C elegans (see Table 1 - these genes are typically involved in axonogenesis and synaptogenesis specificity). All four sets of genes selected in our analysis show a statistically significant overlap with this list of currently known genes: both for the incoming connectivity ($p$-value =0.005 and $p$-value =0.026 in the prediction and co-variation assays respectively, using a hypergeometric significance test), and for the outgoing connectivity ($p$-value =0.021 and $p$-value =0.018 in the prediction and co-variation assays respectively). The end result of this analysis are two connectivity-specific joint gene sets (Supplementary Table T4); genes that appear in both types of outgoing assays (11 genes; ceh-23, che-3,
gpa-3, kin-29, kvs-1, lin-11, osm-3, osm-9, tax-2, tax-4 and unc-5) and genes that appear in both types of incoming assays (5 genes; che-2, mgl-2, mps-1, pef-1 and unc-5).

Table 1: List of genes involved in the analysis* which have previously been reported in the literature as acting in axonogenesis and synaptogenesis. The incoming and outgoing synaptic connectivity columns indicate (with +) if a gene reported in the literature was indeed identified by one of our corresponding assays of gene selection.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Incoming connectivity</th>
<th>Outgoing connectivity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>unc-4</td>
<td></td>
<td></td>
<td>Specifies synaptic choice and axonal morphology.</td>
</tr>
<tr>
<td>unc-5</td>
<td>+</td>
<td>+</td>
<td>Effects axon guidance and outgrowth.</td>
</tr>
<tr>
<td>unc-6</td>
<td>+</td>
<td></td>
<td>Effects axon guidance and outgrowth.</td>
</tr>
<tr>
<td>unc-37</td>
<td></td>
<td></td>
<td>Specifies synaptic choice.</td>
</tr>
<tr>
<td>unc-30</td>
<td>+</td>
<td>+</td>
<td>Defects in axonal pathfinding and synaptic connections.</td>
</tr>
<tr>
<td>unc-40</td>
<td>+</td>
<td></td>
<td>Effects axon guidance and outgrowth.</td>
</tr>
<tr>
<td>unc-53</td>
<td></td>
<td>+</td>
<td>Acts in the migration and outgrowth of axons.</td>
</tr>
<tr>
<td>unc-73</td>
<td></td>
<td>+</td>
<td>Required for cell migrations and axon guidance.</td>
</tr>
<tr>
<td>unc-76</td>
<td></td>
<td></td>
<td>Mutants show axon outgrowth defects.</td>
</tr>
<tr>
<td>slt-1</td>
<td>+</td>
<td></td>
<td>Directs ventral axon guidance and guidance at the midline.</td>
</tr>
<tr>
<td>sax-3</td>
<td>+</td>
<td></td>
<td>Defects in axon patterning at the ventral midline, maintenance of nerve ring placement.</td>
</tr>
<tr>
<td>tax-2</td>
<td>+</td>
<td>+</td>
<td>Mutations display axon outgrowth defects.</td>
</tr>
<tr>
<td>Gene</td>
<td>Symbol</td>
<td>Expression</td>
<td>Functional Contribution</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>tax-4</td>
<td>+</td>
<td>+</td>
<td>Mutations display axon outgrowth defects^{27}.</td>
</tr>
<tr>
<td>vab-8</td>
<td></td>
<td></td>
<td>Guides directed axon outgrowth and cell migration^{28}.</td>
</tr>
<tr>
<td>cam-1</td>
<td>+</td>
<td>+</td>
<td>Guides cell migration and axon outgrowth^{29}.</td>
</tr>
<tr>
<td>lin-11</td>
<td>+</td>
<td>+</td>
<td>Effects axon guidance and outgrowth^{30,31}.</td>
</tr>
<tr>
<td>syg-1</td>
<td></td>
<td></td>
<td>Effects synaptic specificity^{8}.</td>
</tr>
</tbody>
</table>

*The table lists only genes that are included in the expression signatures defined in WormBase and hence can potentially be discovered by our gene selection procedures. Some genes, like syg-2, vab-7, sad-1, unc-34 and others are known from the literature to play a major role in axonogenesis but do not appear in the pertaining WormBase gene lists.

Lastly, we turn to examine the relation between the neuronal expression and connectivity signatures and the functional contribution of each neuron to the animal's behavior, on a small subset of the neurons where the data required is available. The latter functional “neuron contribution signature” is obtained via Multi-Perturbation Analysis (MPA)^{32} of neural laser ablation data published by Bargmann and Horvitz^{33}, focusing on 8 amphid neuron pairs. The Methods section and Supplementary Methods describe the computation of the contribution signature (contribution to chemotaxis behavior) for each of the 8 amphid neuron pairs in a set of chemotaxis assays (see supplementary table T5). Applying the co-variation correlation assay to the full data results in insignificant correlations between the expression-contribution and the connectivity-contribution signatures. However, using feature selection successfully identifies subsets of genes and neuronal connections with borderline significance levels. The expression signature shows the highest correlation with the neuronal functional contribution signature (0.880, \( p\text{-value} = 0.04 \)). The incoming connectivity signature shows no significant correlation and the outgoing synaptic connectivity does manifest a correlation of 0.603 (\( p\text{-value} = 0.05 \)). Evidently, the outgoing synaptic connections of the amphid neurons are more important for successful chemotaxis than their input connections from other neurons (recall that synaptic inputs from sensory receptors are
not included in the connectivity data). Combining both the expression and connectivity signatures to a single joint features signature for each neuron and repeating the assay does not yield any improvement over the correlation obtained using genetic features solely (0.881, \( p\)-value = 0.06). These results testify to the dominant role of the expression properties in determining the behavioral roles of amphid neurons during chemotaxis (at least given the current data, which lacks information on incoming sensory receptors' connectivity). The latter result may also be explained by the high correlation between the expression and connectivity signatures manifested earlier.

**DISCUSSION**

The gene expression data gathered from the public databases is obviously not optimal for comprehensively addressing the challenges raised in this study, as it is quite crude and noisy. Ideally, single-neuron genomic expression data collected at different time points during axonal growth and synaptic development should be examined. Indeed, first steps in obtaining more refined data are now being performed\(^{34,35}\). Yet, within the limitations of the currently existing large-scale data we successfully identify statistically significant information characterizing the fundamental relation between the expression and synaptic properties of neurons. Our estimations of the predictive information that resides in the neuronal expression data about their connectivity should hence be regarded as rough lower bounds on the true values of this information, given the noisy quality of the data.

The gene sets identified in the current study are putative candidates for playing a key role in determining and maintaining the synaptic connectivity structure of C. elegans, carrying the highest level of information about the connectivity signatures. The list of 15 genes described in Supplementary Table T4 (results of the intersection between the prediction and co-variation correlation assays) compose our most plausible gene targets for further investigation. Indeed, some of the genes in this list were already
identified in previous axonogenesis studies, such as unc-5, tax-2, tax-4 \textsuperscript{27} and lin-11 \textsuperscript{30,31}. Yet, some interesting clues indirectly point to the additional involvement of genes from our list which have not been previously known to be directly involved in synaptogenesis: Though there is currently no evidence that che-23 plays a role in axon guidance and formation in the C. elegans, its Drosophila melanogaster homolog, dhlb-9 is known to be involved in neural development, axonal pathfinding and target recognition\textsuperscript{36,37}. Even though mps-1 and kvs-1 have not been directly associated with axon guidance and development, they have been previously reported as causes of neuronal defects and dysfunction after their inactivation in RNA interference experiments\textsuperscript{38}. Some genes on the list are also likely to play part in the axon guidance, targeting and development due to the processes to which they are annotated; such are genes encoding for G proteins involved in signal transduction (GPA-3) or proteins expressed in the cilia of ciliated neurons (CHE-2, CHE-3). Interestingly, some genes on the list are known to act as specific neuron type identifiers (osm-9, osm-3) - hence the information they are carrying regarding the connectivity signatures is probably mediated via their effects on determining cellular fate. Finally, mgl-2, identified as specific to the incoming synaptic signature, has a human homolog, grm-1, which is known to function as a post-synaptic metabotropic G protein-coupled receptor\textsuperscript{39} - this is in line with its appearance solely in the incoming assay list, and supports its role in axon development and regulation.

The relations between the expression and connectivity signatures of neurons and their functional contributions are obviously highly complex and transcend across many levels. Previous studies have shown a correlation between the neuronal transcriptome and the electrophysiological phenotypes of neurons, and have shown that one can build a predictor from the former to the latter\textsuperscript{40}. But the link from these electrophysiological properties to the neuron's actual functional contribution has been missing. Our results point to the existence of a correlation between behavioral measures of a neuron's
function to its "local", specific properties, such as its expression and connectivity signatures. Such a correlation would not have been observed if the neuron's functional contribution would have been solely a global, emergent property of the network in which it is embedded. Furthermore, we find that, at least with the currently available data, the expression signatures are more important than the neuronal connectivity patterns in determining neuronal functional contributions. Indeed, one should note that the C. elegans has a relatively simple, hard-wired network and that in more complex organisms the pertaining findings may be different.

This study is the first to address the relation between neuronal expression and connectivity properties in a large scale quantitative manner. Despite the rough and low precision data available the results presented and the genes identified provide a promising starting point for further, more detailed computational and experimental investigations. The use of DNA microarrays with hundreds or thousands of simultaneously-measured mRNAs, along with extensive behavioral assays spanning many tasks, should further facilitate our understanding of the relationship between gene expression and behavior. Further studies would be required to determine whether the findings presented here appear across a variety of other functions and in other species, using genetic and connectivity information that is gradually being identified, e.g., in cats\(^41\) and humans\(^42\).

**METHODS**

**The data.** The Neurons expression signatures were obtained from the public WormBase database (http://wormbase.org version WS140), which lists for each neuron the genes with expression patterns directly associated with it (a gene is directly associated with a neuron according to its appearance in the pertaining expression pattern assay). The data includes 181 neurons (out of the 302 C. elegans neurons), each having
at least one gene associated with it. To avoid a bias caused by the symmetric structure of the data we focused on 98 neurons (only neurons on the right side of the nematode and neurons without a symmetrical companion - Supplementary Table T1). The resulting expression signature of each neuron is a binary vector of 289 genes (see gene list in Supplementary Table T6), coded as one if the corresponding gene appears to be associated to the neuron and zero otherwise. Neurons' connectivity signatures were obtained from the updated version of C. elegans synaptic wiring diagrams, formed by serial-sections electron microscopic reconstructions; R. M. Durbin (http://elegans.swmed.edu-parts/neurodata.txt) based on the classic work by White et al.\textsuperscript{15} and Hall et al.\textsuperscript{16} We focus on chemical synapses, in which the identities of the pre-synaptic and post-synaptic neurons are well-defined. Any two neurons may be connected or not with a direction assigned to their connection. Each neuron is thus described by two binary vectors characterizing its connectivity, one for the outgoing synaptic connections (the synapses sent out by the axon of the respective neuron) and one for the incoming synaptic connections (the synapses impinging on it).

The neuron type classification. Each of the 98 neurons analyzed were assigned to one or more neuron types according to the WormBase database. The neuron types are: sensory, amphid (including amphid interneurons), cord, motor, ring, labial and interneurons. See supplementary Table T1 for the classification of the neurons.

The neurons' contribution signature. The signature describes the causal roles (contributions) that a neuron plays during behavior (chemotaxis). The contributions are determined via a Multi-Perturbation Analysis (MPA)\textsuperscript{32,43} of neuronal laser ablation experiments probing the chemotaxis behavior of the nematode to various chemical attractants\textsuperscript{33}. The contribution of each neuron represents its relative importance to successful chemotaxis. It is computed by measuring (or estimating) the decrease in the chemotaxis ability of the nematode when the pertaining neuron is ablated, averaging
over all possible multi-knockout configurations (see Supplementary Methods for a detailed description of the MPA and its application to the worm chemotaxis data). This contribution computation is repeated for 4 different chemotaxis tasks to 4 different attractants, resulting in a contribution signature for each neuron as a vector of 4 continuous values. For each task, the contributions of all neurons are normalized such that they all sum up to 1 and reflect relative contributions (see Supplementary Table T5).

The prediction assay. Prediction is performed using a standard weighted multi-class K-Nearest Neighbor (KNN) algorithm, using Euclidean distance between the neurons in the input expression signatures' space. As a preprocessing stage we eliminated features (genes) that are shared by no more than one neuron. The prediction targets, given an input neuron, are its synaptic connections (incoming and outgoing, separately) to all other neurons (each represented as a class in the multi-class prediction). The prediction model’s performance score is based on 5-fold cross validation (training on 80% of the neurons and testing on the resulting 20%). Prediction accuracy is measured by the average area under ROC curve (AUC), where averaging is performed over all output classes. The cross validation was further used for finding the optimal value of K, the single hyper-parameter of KNN. Statistical significance of the prediction performance was calculated against an empiric null hypothesis, constructed from repeating the prediction procedure with shuffling: On each such repetition the neuron signatures were shuffled amongst all neurons (that is, shuffling the neuron labels - thus eliminating any functional relation between a neuron and its corresponding signatures while preserving the actual distribution of signatures). To calculate significance levels a one sided t-test was applied, comparing the mean result achieved in the 5-fold cross validation of the actual data to the empirical distribution achieved with the shuffling. (The t-test requires data generated from a normal distribution; this assumption was verified by analyzing the Q-Q plots of the empirical distribution).
The co-variation correlation assay. To examine the correlation between two signatures across all neurons under investigation we use an assay similar to the one used by Toledo-Rodriguez et al.\textsuperscript{40} Given a set of $N$ neurons, where each has two signatures, $s_1$ and $s_2$, we construct two $N \times N$ similarity matrices $S_1$ and $S_2$, where $S_1$ ($S_2$) represents the pairwise Pearson correlation between the $s_1$ ($s_2$) signatures of the neurons. The $(N^2-N/2-N)$ entries forming the lower triangle of $S_1$ ($S_2$) are concatenated to form a co-variation vector $v_1$ ($v_2$). The Pearson correlation between the two co-variation vectors $v_1$ and $v_2$ describes the extent of which the neighborhood relations of the neurons in the two signature spaces $s_1$ and $s_2$ are similar. The statistical significance of the resulting correlation is computed using an empiric null hypothesis constructed from repeating the procedure with shuffling. On each repetition the neuron signatures were shuffled amongst all neurons (shuffling the neuron labels as described above in the prediction assay). To calculate p-values we repeated the shuffling 1000 times and computed the probability to achieve a score equal or higher than the score of the original (non-shuffled) data. We methodology was used since in contrast with the prediction assay, the scores obtained by the shuffling procedure were not normally distributed.

Feature selection – prediction assay. Feature selection was used to find a small subset of genes that yield high accuracy prediction (at least as equally good as that obtained with all the features). A filtering feature selection method was used, ranking the features according to their average mutual information with respect to the multi-class targets\textsuperscript{44}. For various sizes of feature sets (2, 3, 5, 10, 17, 30, 53, 93, 164, 289) the average AUC performance achieved by KNN (similar procedure as in the general case) was calculated (note that the ranking of the features and selecting the optimal K were performed only on part of the data, and measuring the actual performance measure was done on a validation set, not available to the training stages). Statistical significance was computed for the feature set resulting in the highest AUC. The significance level is calculated against an empirical null hypothesis constructed from repeating the feature selection
procedure with the same shuffling procedure described above in the prediction-without-selection case, and applying the same t-test. (Q-Q plots verify a normal distribution which has permitted us to perform the t-test)

**Feature selection – correlation co-variation assay.** Here we used a greedy backward elimination algorithm\(^\text{44}\), starting from the complete gene set and iterating while eliminating genes via a greedy algorithm which maximizes the correlation co-variation measure. In each iteration 25% of the features (genes) are eliminated according to their marginal influence on the correlation co-variation measure when excluded. The feature selection process was repeated 10 times, each utilizing 90% of the data, to avoid overfitting and local minima. The selected set of genes used throughout this paper includes only genes that were selected in all 10 repetitions of the feature selection process. The statistical significance of the outcome is calculated against an empiric null hypothesis constructed from repeating the identical feature selection procedure with a shuffling procedure as described above, for a 1000 times. The p-values for testing the statistical significance of the optimal feature set were computed as follows: For each number of features \( j \) we calculate the mean correlation achieved when applying the feature selection to the shuffled data \( C_j^{\text{null}} \) and its standard deviation \( S_j^{\text{null}} \) (this forms a null hypothesis empirical distribution). For the correlation achieved by the maximum chosen set, \( C_j^{\text{true}} \) (on the actual non-shuffled data), with \( j^* \) being the number of features in the chosen set, we calculated its variation from the null model \( \Delta=(C_j^{\text{true}}-C_j^{\text{null}})/S_j^{\text{null}} \). The p-value is the probability of achieving such a variation, \( \Delta \) (or larger) in any of the 1000 shuffled repetitions with any number of features (for each shuffle we consider the optimal number of features maximizing the variation from the null model). Hence, if significant it testifies that the probability of achieving such a variation by chance (no matter with how many features) is low. The significance level of the size of the gene set chosen (those selected in all 10 repetitions) is computed versus the probability of
achieving similar sizes or larger when applying an identical procedure to randomly shuffled data.

Supplementary Information accompanies the paper on …

REFERENCES

2. Yu, T.W., Hao, J.C., Lim, W., Tessier-Lavigne, M. & Bargmann, C.I. Shared 
   receptors in axon guidance: SAX-3/Robo signals via UNC-34/Enabled and a 
3. Lim, Y.-s. & Wadsworth, W.G. Identification of Domains of Netrin UNC-6 that 
   Mediate Attractive and Repulsive Guidance and Responses from Cells and 
   guide circumferential migrations of pioneer axons and mesodermal cells on the 
5. Adler, C.E., Fetter, R.D. & Bargmann, C.I. UNC-6/Netrin induces neuronal 
   asymmetry and defines the site of axon formation. Nat Neurosci 9, 511-518 
   Presynaptic Vesicle Clustering and Axon Termination. Neuron 29, 115-129 
   the Synaptic Guidepost Protein SYG-2 and Its Receptor, SYG-1. Cell 116, 869-
   THE GROWTH CONE: Ligand-Receptor Complexes and the Control of Axon 


**ACKNOWLEDGMENT**

We thank Ya'acov Ritov, Issaq Meilijson, Danny Yekutieli, Nir Yosef, Tomer Shlomi, Roy Varshavsky, Millet Treinin and Cori Bargmann for their valuable comments and suggestions. A. Kaufman is supported by the Yeshaya Horowitz Association through the Center of Complexity Science. E. Ruppin's research is supported by the Tauber Fund, the Yeshaya Horowitz Association through the Center of Complexity Science and the Israeli Science Fund (ISF).

**AUTHOR CONTRIBUTIONS**

A.K. gathered the data from the public available sources. A.K., G.D., and E.R. designed the analysis methods. A.K. and G.D applied them. A.K. and E.R. wrote the manuscript.

**COMPETING INTEREST STATEMENTS**

The authors declare they have no competing financial interests.
**Figure 1:** Prediction of synaptic connectivity signatures as a function of the most informative genes. The accuracy of the predictor as a function of the number of genes selected for the predictor is described by the blue line. Prediction accuracy is measured by the average area under ROC curve (AUC). The right most point (289 genes) denotes the prediction outcome before any feature selection is applied to the data. The blue line represents 5-fold cross validation repetitions of the selection-prediction scheme (mean and standard deviations are displayed). The red dashed lines represent the empirical null hypothesis distribution of performing the selection-prediction scheme on random data (constructed by shuffling the identities of the neurons, see Methods). Maximum AUC measurements are achieved with 53 and 30 features in the incoming and outgoing assays respectively with corresponding p-values of $p=10^{-99}$ and $p=10^{-97}$, calculated by applying a one-sided t-test between the original and shuffled data (see Methods).
Figure 2: Co-variation correlation feature selection assay. The mean and standard deviation of the Pearson correlation (blue line) between the neurons’ neighborhood relations in the expression and connectivity spaces is displayed as a function of the number of genes used to determine the expression signature (results of 10 repetitions of the assay each with 90% of the neurons). The right most point (289 genes) denotes the correlation before any feature selection is applied to the data. The dashed red line represents the empirical null hypothesis distribution of the co-variation correlation on random data (constructed by shuffling 1000 times the identities of the neurons and reapplying the analysis to the shuffled data). Maximum correlation measurements are achieved with 39 and 92 features in the incoming and outgoing assays respectively with corresponding p-values of p<0.0001, p=0.004 respectively (see Methods).