METHODOLOGY

Tissue Enrichment Analysis (TEA)

David Angeles-Albores, Raymond Y Lee, Juancarlos Chan and Paul W Sternberg*

*Correspondence: pws@caltech.edu
HHMI and California Institute of Technology, Division of Biology and Biological Engineering, 1200 E California Blvd, 91125, Pasadena, USA
Full list of author information is available at the end of the article

Abstract

Background: Over the last ten years, there has been explosive development in tools for measuring gene expression. These tools can identify thousands of altered genes between conditions, but understanding these datasets and forming hypotheses based on them remains challenging. One way to analyze these datasets is to associate ontologies (hierarchical, descriptive vocabularies with controlled relations between terms) with genes and to look for enrichment of specific terms. Although gene ontology (GO) is available for *Caenorhabditis elegans*, it does not include anatomy information.

Results: We have developed a tool for identifying enrichment of *C. elegans* tissues among gene sets. We applied a hypergeometric model to a pre-existing anatomy ontology for the worm in order to identify enriched tissues and generated a website GUI. To cut down on verbosity, we have come up with three straightforward filtering criteria that slim the ontology by almost tenfold. One filter removes all terms with fewer than an arbitrary number of annotating genes; the second filter removes all sister terms that have identical gene annotations; and the third filter removes all parents whose daughters all survived the previous two filters. We adjusted these filters and validated our tool using a set of 30 gold standards from Expression Cluster data in WormBase.

Conclusions: Our Tissue Enrichment Analysis (TEA) can be found at `www.wormbase.org/tea`, and can be downloaded using Python's standard pip installer. It tests a slimmed-down *C. elegans* tissue ontology for enrichment of specific terms and provides users with a text and graphic representation of the results. Our tool can discriminate between embryonic and larval tissues and can even identify tissues down to the single-cell level.

Keywords: Gene Ontology; Anatomy Ontology; WormBase; RNA-seq; High-throughput biology

Background

RNA-seq and other high-throughput methods in biology have the ability to identify thousands of genes that are altered between conditions. These genes are often correlated in their biological characteristics or functions, but identifying these functions remains challenging. To interpret these long lists of genes, biologists need to abstract genes into fewer terms that are biologically relevant to form hypotheses about what is happening in the data. One such abstraction method relies on Gene Ontology (GO). GO provides a controlled set of hierarchically ordered terms [1, 2] that provide detailed information about the molecular, cellular or biochemical functions of any gene. For a given gene list, certain software programs can query whether a particular term is enriched [3–5]. One area of biological significance that GO does not include is anatomy. One way to address this shortcoming is to test a 'tissue ontology' that provides a complete anatomical description for an organism (e.g 'tissue',

'organ' or 'neuronal cell'), in this case for *C. elegans* [6]. The *C. elegans* database, WormBase [7], maintains a curated list of gene expression data from GFP-reporters. Here we provide a new framework that analyses a user-input list for enrichment of specific tissues. Tissues are physiologically relevant units with broad, relatively well-understood functionalities amenable to hypothesis formation.

Another problem frequently associated with GO enrichment analysis is that it is often difficult to interpret due to the large number of terms associated with a given gene. There exist a number of GO analytic tools for use by the community but a shared complaint for many programs is the very large number of GO terms that are significantly associated with any given gene list. DAVID, a common tool for GO enrichment analysis, clusters enriched terms into broad categories [8], whereas PANTHER [3, 9] attempts to solve this issue by employing a manually reduced ontology, GOslim (H. Yu and P. Thomas, pers. comm.).

To prevent our tool from suffering from verbosity problems, we have filtered our ontology using a small set of well-defined criteria to remove terms that do not contribute additional information. To our knowledge, such filtering has not been performed in an algorithmic fashion for a biological ontology before—indeed, DAVID does not employ term trimming *a priori* of testing, but rather fuzzy clustering *post* testing to reduce the number of ontology terms. Other pruning tools do exist (see for example [10]), but the pruning is query-dependent. We believe our trimming methodology strikes a good balance between detailed tissue calling and conservative testing.

Results

Generating a Useful Dictionary

Reducing term redundancy through a similarity metric

As a first step to generate our tissue enrichment software, we wished to select tissue terms that were reasonably well-annotated, yet specific enough to provide insight and not redundant with other terms. We also wanted to avoid testing tissues at levels where annotation redundancy becomes problematic. For example, several left and right neurons have at least 25 annotated genes and we may want to include them for enrichment testing. However, many left/right neuronal pairs (which are sisters in the ontology) have almost identical annotations, with at most one or two gene differences between them. We reasoned that when two tissues have almost identical annotations, we cannot have statistical confidence in differentiating between them. As a result, testing these sister tissues provides no additional information compared with testing only the parent node to these sisters. We refer to such sisters as 'redundant'. To identify redundancy, we defined a similarity metric (see *Methods* section and Figure 1a). Our similarity metric can be used to identify sisters that have very high similarity between them; alternatively, redundant sisters could be identified if a single sister had a very high similarity score. We referred to these two scoring criteria as 'avg' and 'any' respectively.

Terminal branch terms and parent terms can be safely removed in an algorithmic fashion

Another problem arises from the fact that the tissue ontology is still scarcely populated. Many nodes have 0-10 annotations, which we consider too few to accurately

test. To solve this issue, we implemented a straightforward trimming strategy. For a given terminal node, we test whether the node has more than a threshold number of annotations. If it does not, the node is removed. The next node in the branch is tested and removed recursively until a node which satisfies the condition is found. At that point, no more nodes can be removed from that branch. This is guaranteed by the structure of the ontology: Parent nodes inherit all of the annotations of all of their descendants, so the number of annotated terms monotonically increases with increasing term hierarchy (see Figure 1b). In this way, we ensure that our term dictionary includes only those tissues that are considered sufficiently well annotated for statistical purposes.

Finally, we also wanted to remove as many terms as possible from the dictionary with the goals of reducing covariance between terms, decreasing multiple testing and removing as many non-informative terms as possible. Decreasing covariance between terms is important because we employ a frequentist approach that assumes all terms are independent. Large covariation coefficients between some terms means that if one of these tissues tests significant, the other terms are much more likely to pass significance testing as well. This makes adequate correction for false positive rates considerably more difficult. Moreover, from a data analysis perspective, we reasoned that, for any parent node, if all its daughters were selected for testing, there was no additional benefit to test the parent. In other words, if all the daughter nodes are tested, there is little additional information to be gained by including the parent node. To address this issue we removed parent nodes from the analysis if all their daughter nodes passed the annotation threshold (see Figure 1c). We called this a ceiling filter.

Filtering greatly reduces the number of nodes used for analysis

By itself, each of these filters can reduce the number of nodes employed for analysis. These filters are not all commutative: while trimming and redundancy filtering are commutative, applying the ceiling filter is commutative with neither the trimming nor the redundancy filter. If the ceiling filter is applied before any other filter, only terminal nodes will remain, since all the parents have complete daughter sets. Since terminal nodes are the most poorly annotated, after applying the remaining filters very few nodes will be left behind if any. On the other hand, applying the ceiling operator after trimming and redundancy filtering will result in greater numbers of nodes. We always applied the ceiling at the end. For validation (see below) we made a number of different dictionaries. The original ontology has almost 6,000 terms of which 1675 have at least 5 gene annotations. After filtering, dictionary sizes ranged from 21 to a maximum of 460 terms, which shows the number of terms in a scarcely annotated ontology can be reduced by an order of magnitude through application of a few simple filters (see Table 1). These filters were used to compile a static dictionary that we employ for all analyses (see *Validation of the algorithm and parameter selection* section for details).

Tissue enrichment testing via a hypergeometric model

Having built a static dictionary, we generated a Python script that implements a significance testing algorithm based on the hypergeometric model. Briefly, the hypergeometric model assumes the existence of an urn with a pre-determined number

of balls inside it. The balls can be painted one of several colors. The hypergeometric model provides an answer to the question: If an individual removes N balls, what is the probability of observing n_i balls of color i, if the balls are selected without replacement? Mathematically, this is expressed as:

$$\mathrm{P}(n_i|N, m_i, M) = \frac{\binom{m_i}{n_i}\binom{M-m_i}{N-n_i}}{\binom{N}{n_i}}. \quad (1)$$

Here, n_i is the number of balls of type i drawn, N is the total number of draws, m_i is tissue i and M is the total number of balls in the urn. In our specific case, M_i is equal to the total number of annotations in our dictionary. N is found by taking the user-input list and removing any genes that are not in our annotation dictionary. The remaining genes are associated with their annotation profiles—if a tissue is associated with s tissues, it generates s balls of s colors. Our program counts the number of times each tissue appears in the user list, and calculates the probability of having withdrawn as many or more balls for each tissue in the user list. Due to the discrete nature of the hypergeometric distribution, this algorithm can generate artifacts when the list is small. To avoid spurious results, a tissue is never considered significant if there are no annotations for it in the user-provided list.

Once the p-values for each term have been calculated, we apply a standard FDR correction using a Benjamini-Hochberg step-up algorithm [11]. FDR corrected p-values are called q-values. Genes that have a q-value less than a given alpha are considered significant. Our default setting is to set alpha at 0.1.

Users input a gene-list using any valid gene name for *C. elegans*. These names are processed into standard WormBase gene IDs (WBGene IDs). The program returns a text-based table showing the tissues that tested significant, along with their associated q-value, the expected number of hits for a list of that size, the observed number of hits and the enrichment fold change (observed hits / expected hits). Finally, the program can also return a bar chart of the enrichment fold change for the fifteen tissues with the lowest measured q-values. Our software is implemented in an easy to use GUI within WormBase (see Figure 2). Anatomy terms are displayed in human-readable format followed by their unique ontology ID (WBbt ID).

Validation of the algorithm and parameter selection

We wanted to select a dictionary that included enough terms to be specific beyond the largest *C. elegans* tissues, yet would minimize the number of spurious results and which had a good dynamic range in terms of enrichment fold-change. As expected, larger tissues are correlated with better annotation, so increasing term specificity is associated with losses in statistical power. To help us select an appropriate dictionary and validate our tool, we used a set of 30 gold standards based on microarray and RNA-seq literature which are believed to be enriched in specific tissues [12–19]. These data sets are annotated gene lists derived from the corresponding Expression Cluster data in WormBase. Some of these studies have been used to annotate gene expression, and so they did not constitute an independent testing set. To correct this

flaw, we built a clean dictionary that specifically excluded all annotation evidence that came from these studies.

As a first attempt to select a good dictionary, we generated all the possible combinations of dictionaries with minimal annotations of 10, 25, 50 and 100 genes and similarity cutoffs of 0.9, 0.95 and 1, using 'avg' or 'any' similarity thresholding methods (see Table 1). The number of tissues was inversely correlated to the minimum annotation cutoff, as expected, and was largely insensitive to the similarity threshold in the range we explored (0.9-1). Next, we analyzed all 30 datasets using each dictionary. Because of the large number of results, instead of analyzing each set of terms individually, we measured the average q-value for significantly enriched terms in each dataset. When we analyzed the distribution of significant q-values for the dictionaries, we found that the similarity threshold mattered relatively little for any dictionary. We also noticed that the 'any' thresholding method resulted in tighter histograms with a mode closer to 0. For this reason, we chose the 'any' method for dictionary generation. The average q-value increased with decreasing annotation cut-off (see Figure 3), which reflects the decreasing statistical power associated with fewer annotations per term, but we remained agnostic as to how significant is the trade-off between power and term specificity. Based on these observations, we ruled out the dictionary with the 100 gene annotation cut-off: it had the fewest terms and its q-values were not low enough to compensate the trade-off in specificity. To select between dictionaries generated between 50, 33 and 25 annotation cut-offs, and also to ensure the terms that are selected as enriched by our algorithm are reasonable, we looked in detail at the enrichment analysis results. Most results were comparable and in line with expectations. For some sets, all dictionaries performed well. For example, in our 'all neuron enriched sets' [13, 15] the results were an amalgamation of neuron-related terms regardless of the dictionary used (see Table 2). On the other hand, when we looked at a gene set enriched for germline precursor expression in the embryo [13], the dictionary with the 50 cutoff was only able to identify 'oocyte WBbt:006797'; whereas the two smaller dictionaries were able to single out cells germline precursor cells—at the 33-cutoff, our tool identified the larval germline precursor cells 'Z2' and 'Z3' as being five-fold enriched, and at the 25 gene-cutoff the embryonic germline precursor terms 'Psub4','Psub3' and 'Psub2' were identified in addition to 'Z2' and 'Z3'. We also queried an embryonic stage intestine precursor geneset [13]. Notably, this gene set yielded no enrichment when using the 25 cutoff dictionary, nor when using the 50 cutoff dictionary. However, the 33 cutoff dictionary identified the E lineage, which is the intestinal precursor lineage in *C. elegans*, as enriched in this set. Not all queries worked equally well. For example, a number of intestinal enriched genes sets [13, 16] were not enriched in intestine related terms in any dictionary, but they were enriched for pharynx- and hypodermis-related terms. We were somewhat surprised that intestinal gene sets performed poorly, since the intestine is a relatively well-annotated tissue.

We assessed the internal agreement of our tool by using independent gene-sets that we expected to be enriched in the same tissues. We used two pan-neuronal sets [13, 15]; two PVD enriched sets [13, 19]; two GABAergic gene sets [13, 14]; two pharyngeal gene sets [12, 13]; and two intestinal gene sets [13, 16]. Overall, the tool seems to have good internal agreement. On most sets, the same terms were enriched,

although order was somewhat variable (see Table 2). However, most high-scoring terms were preserved between gene sets. All comparisons can be found online in our Github repository (see Availability of data and materials). Overall, the dictionary generated by a 33 gene annotation cutoff with 0.95 redundancy threshold using the 'any' criterion. seemed to perform well, with a good balance between specificity, verbosity and accuracy, so we selected this parameter set to generate our static dictionary.

A brief example

We applied our tool to the RNA-seq datasets developed by Engelmann et al. [20] to gain further understanding of their underlying biology. Engelmann et al. exposed young adult worms to 5 different pathogenic bacteria or fungi for 24 hours, after which mRNA was extracted from the worms for sequencing. We ran TEA on the genes Engelmann *et al* identified as up- or down-regulated. Initially we noticed that genes that are down-regulated tend to be twice better annotated on average than genes that were up-regulated, suggesting that our understanding of the worm immune system is scarce, in spite of important advances made over the last decade.Up-regulated tissues, when detected, almost always included the hypodermis and excretory duct. Three out of the five samples showed enrichment of neuronal tissues or neuronal precursor tissues amongst the down-regulated genes (for an example, see 5). A possible explanation for this might be that the infected worms are sick and the neurons are beginning to shut down; an alternative hypothesis would be that the worm is down-regulating specific neuronal pathways as a behavioral response against the pathogen. Indeed, several studies [21, 22] have provided evidence that *C. elegans* uses chemosensory neurons to identify pathogens. One bacterium did not exhibit the same pattern of down-regulation of neuronal-associated genes. *E. faecalis* showed increased expression of genes associated with neuronal tissues, hinting that *E. faecalis* may have a different pathogenic profile. Our results highlight the involvement of various *C. elegans* neuronal tissues in pathogen defense.

Discussion

We have presented a tissue enrichment analysis tool that employs a standard hypergeometric model to test the *C. elegans* tissue ontology. We have also presented the first, to our knowledge, ontology trimming algorithm for biomedical ontologies. This algorithm, which is very easy to execute, places strong limits on the number of terms selected for testing. Due to the nature of all ontologies as hierarchical, acyclical graphs with term inheritance, term annotations are correlated along any given branch. This correlation reduces the benefits of including all terms for statistical analysis: for any given term along a branch, if that term passes significance, there is a high probability that many other terms along that branch will also pass significant. If the branch is enriched by random chance, error propagation along a branch means that many more false positives will follow. Thus, a researcher might be misled by the number of terms of correlated function and assign importance to this finding; the fact that the branching structure of GO amplifies false positive signals is a powerful argument for either reducing branch length or branch intra-correlation, or both. On the other hand, if a term is actually enriched, we argue

that there is little benefit to presenting the user with additional terms along that branch. Instead, a user will benefit most from testing sparsely along the tree at a suitable specificity for hypothesis formation. Related terms of the same level should only be tested when there is sufficient annotation to differentiate, with statistical confidence, whether one term is enriched above the other. Our algorithm reduces branch length by identifying and removing nodes that are insufficiently annotated and parents that are likely to include sparse information.

Chikina *et al* [23] report a tissue enrichment model based on a Support Vector Machine classifier that has been trained on microarray studies. SVM classifiers are powerful tools, but they require continuous retraining as more tissue expression data becomes available. Moreover, classifiers require that data be rank-ordered by some metric, something which is not possible for certain studies. Our tool relies on an annotation dictionary that is continuously updated, does not require retraining and does not require ranked genes. To our knowledge, there are no other tissue ontology enrichment tools in *C. elegans*, but similar projects exist for humans and zebrafish [24, 25], highlighting the relevance of our tool for high-dimensionality biology.

We have tried hard to benchmark our tool well. However, our analysis suffers from the drawback that is very hard to to identify spurious term enrichment. Although we were unable to determine false-positive and false-negative rates, we do not believe this should deter scientists from using our tool. Rather, we encourage researchers to use our tool as a guide, integrating evidence from multiple sources to inform the most likely hypotheses. As with any other tool based on statistical sampling, our analysis is most vulnerable to bias in the data set. For example, we know that expression reports are negatively biased against germline expression because of the difficulties associated with expressing DNA in this tissue [26]. As time passes, we are certain the accuracy and power of this tool will improve thanks to the efforts of the combined worm community; indeed, without the community reports of tissue expression in the first place, this tool would not be possible.

Methods

Fetching annotation terms

We used WormBase-curated gene expression data, which includes annotated descriptions of spatial-temporal expression patterns of genes, to build our dictionary. Gene lists per anatomy term were extracted from a Solr document store of gene expression data from the WS252 database provided by WormBase [7]. We used the Solr document store because it provided a convenient access to expression data that included inferred annotations. That is, for each anatomy term, the expression gene list includes genes that were directly annotated to the term, as well as those that were annotated to the term's descendant terms (if there were any). Descendant terms were those connected with the focus term by is_a/part_of relationship chains defined in the anatomy term ontology hierarchy.

Filtering nodes

Defining a Similarity Metric

To identify redundant sisters, we defined the following similarity metric:

$$s_i = \frac{|g_i|}{|\bigcup_{i=0}^{k} g_i|} \tag{2}$$

Where s_i is the similarity for a tissue i with k sisters; g_i refers to the set of tissues associated with tissue i and $|g|$ refers to the cardinality of set g. For a given set of sisters, we called them redundant if they exceeded a given similarity threshold. We envisioned two possible criteria and built different dictionaries using each one. Under a threshold criteron 'any' with parameter S between $(0, 1)$, a given set of sisters j was considered redundant if the condition

$$s_{i,j} > S \tag{3}$$

was true for any sister i in set j. Under a threshold criterion 'avg' with parameter S, a given set of sisters j was considered redundant if the condition

$$\mathrm{E}[s_i]_j > S \tag{4}$$

was true for the set of sisters j (see Figure 1a).

Implementation

All scripts were written in Python. Our software relies on the Pandas, NumPy, Seaborn and SciPy modules to perform all statistical testing and data handling [27–29].

Availability of data and materials

Our web implementation is available at `https://www.wormbase.org/tea`. Our software can also be downloaded using Python's pip installer via the command

```
pip install tissue_enrichment_tool
```

Alternatively, our software is available for download at: `http://dangeles.github.io/TissueEnrichmentAnalysis`

All benchmark gene sets, benchmarking code and Figures can also be found at the same address, under the 'tests' folder.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

DA and PWS conceived of the project; DA developed algorithm; RYL made intellectual contributions to the project; RYL and JC developed the web GUI.

Acknowledgements

We thank Justin Bois for his help and support. We would like to acknowledge all members of the Sternberg lab for helpful discussion.

References

1. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(may):25–29, 2000.
2. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.
3. Huaiyu Mi, Qing Dong, Anushya Muruganujan, Pascale Gaudet, Suzanna Lewis, and Paul D. Thomas. PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research*, 38(SUPPL.1), 2009.
4. Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501, 2010.
5. Da Wei Huang, Richard a Lempicki, and Brad T Sherman. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.
6. R. Y N Lee and Paul W. Sternberg. Building a cell and anatomy ontology of Caenorhabditis elegans, 2003.
7. Kevin L Howe, Bruce J Bolt, Scott Cain, Juancarlos Chan, Wen J Chen, Paul Davis, James Done, Thomas Down, Sibyl Gao, Christian Grove, Todd W Harris, Ranjana Kishore, Raymond Lee, Jane Lomax, Yuling Li, Hans-Michael Muller, Cecilia Nakamura, Paulo Nuin, Michael Paulini, Daniela Raciti, Gary Schindelman, Eleanor Stanley, Mary Ann Tuli, Kimberly Van Auken, Daniel Wang, Xiaodong Wang, Gary Williams, Adam Wright, Karen Yook, Matthew Berriman, Paul Kersey, Tim Schedl, Lincoln Stein, and Paul W Sternberg. WormBase 2016: expanding to enable helminth genomic research. *Nucleic acids research*, 44(November 2015):D774–D780, 2016.
8. Da Wei Huang, Brad T. Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, Robert Stephens, Michael W. Baseler, H. Clifford Lane, and Richard A. Lempicki. DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, 35(SUPPL.2), 2007.
9. Huaiyu Mi, Anushya Muruganujan, and Paul D. Thomas. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41(D1), 2013.
10. Jong Woo Kim, Jordi Conesa Caralt, and Julia K. Hilliard. Pruning bio-ontologies. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 1–10, 2007.
11. Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, 1995.
12. Jeb Gaudet, Srikanth Muttumu, Michael Horner, and Susan E. Mango. Whole-genome analysis of temporal gene expression during foregut development. *PLoS Biology*, 2(11), 2004.
13. W. Clay Spencer, Georg Zeller, Joseph D. Watson, Stefan R. Henz, Kathie L. Watkins, Rebecca D. McWhirter, Sarah Petersen, Vipin T. Sreedharan, Christian Widmer, Jeanyoung Jo, Valerie Reinke, Lisa Petrella, Susan Strome, Stephen E. Von Stetina, Menachem Katz, Shai Shaham, Gunnar Rätsch, and David M. Miller. A spatial and temporal map of C. elegans gene expression. *Genome Research*, 21(2):325–341, 2011.
14. Hulusi Cinar, Sunduz Keles, and Yishi Jin. Expression profiling of GABAergic motor neurons in Caenorhabditis elegans. *Current Biology*, 15(4):340–346, 2005.
15. Joseph D Watson, Shenglong Wang, Stephen E Von Stetina, W Clay Spencer, Shawn Levy, Phillip J Dexheimer, Nurith Kurn, Joe Don Heath, D M Miller 3rd, and David M Miller. Complementary RNA amplification methods enhance microarray identification of transcripts expressed in the C. elegans nervous system. *BMC Genomics*, 9:84, 2008.
16. Florencia Pauli, Yueyi Liu, Yoona a Kim, Pei-Jiun Chen, and Stuart K Kim. Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans. *Development (Cambridge, England)*, 133(2):287–295, 2006.
17. Douglas S. Portman and Scott W. Emmons. Identification of C. elegans sensory ray genes using whole-genome expression profiling. *Developmental Biology*, 270(2):499–512, 2004.
18. Rebecca M Fox, Joseph D Watson, Stephen E Von Stetina, Joan McDermott, Thomas M Brodigan, Tetsunari Fukushige, Michael Krause, D M Miller 3rd, and David M Miller. The embryonic muscle transcriptome of Caenorhabditis elegans. *Genome Biol*, 8(9):R188, 2007.
19. Cody J. Smith, Joseph D. Watson, W. Clay Spencer, Tim O'Brien, Byeong Cha, Adi Albeg, Millet Treinin, and David M. Miller. Time-lapse imaging and cell-specific expression profiling reveal dynamic branching and molecular determinants of a multi-dendritic nociceptor in C. elegans. *Developmental Biology*, 345(1):18–33, 2010.
20. Ilka Engelmann, Aurélien Griffon, Laurent Tichit, Frédéric Montaña-Sanchis, Guilin Wang, Valerie Reinke, Robert H. Waterston, LaDeana W. Hillier, and Jonathan J. Ewbank. A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in C. elegans. *PLoS ONE*, 6(5), 2011.
21. Joshua D. Meisel and Dennis H. Kim. Behavioral avoidance of pathogenic bacteria by Caenorhabditis elegans. *Trends in Immunology*, 35(10):465–470, 2014.
22. Y Zhang, H Lu, and C I Bargmann. Pathogenic bacteria induce aversive olfactory learning in Caenorhabditis elegans. *Nature*, 438(7065):179–184, 2005.
23. Maria D. Chikina, Curtis Huttenhower, Coleen T. Murphy, and Olga G. Troyanskaya. Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. *PLoS Computational Biology*, 5(6), 2009.
24. Young Suk Lee, Arjun Krishnan, Qian Zhu, and Olga G. Troyanskaya. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*, 29(23):3036–3044, 2013.
25. Sergey V Prykhozhij, Annalisa Marsico, and Sebastiaan H Meijsing. Zebrafish Expression Ontology of Gene Sets (ZEOGS): a tool to analyze enrichment of zebrafish anatomical terms in large gene sets. *Zebrafish*, 10(3):303–15, 2013.

26. William G. Kelly, SiQun Xu, Mary K. Montgomery, and Andrew Fire. Distinct requirements for somatic and germline expression of a generally expressed Caenorhabditis elegans gene. *Genetics*, 146(1):227–238, 1997.
27. Wes McKinney. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing*, pages 1–9, 2011.
28. Stéfan Van Der Walt, S. Chris Colbert, and Gael Varoquaux. The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30, 2011.
29. Travis E Oliphant. SciPy: Open source scientific tools for Python. *Computing in Science and Engineering*, 9:10–20, 2007.

Figures

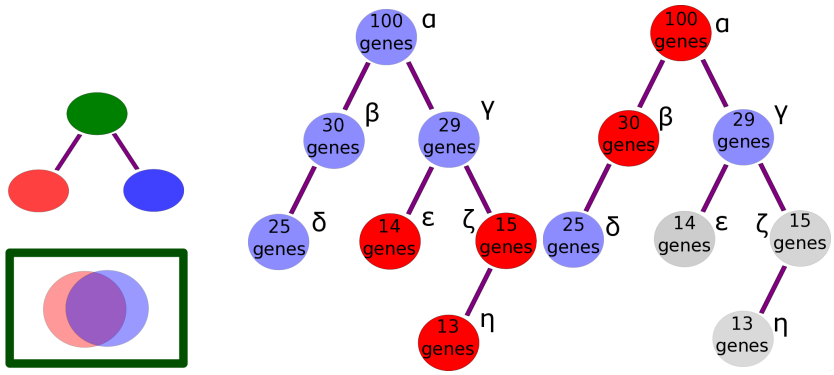


Figure 1: **Schematic representation of trimming filters for an acyclical ontology. a)** The parent node (green) contains at least as many annotations as the union of the two sisters. These two sisters share annotations extensively. Therefore they are too similar and should be removed. **b)** Nodes with less than a threshold number of genes are trimmed (red) and discarded from the dictionary. Here, the example threshold is 25 genes. Nodes ϵ, ζ, η, shown in red will be removed. **c)** We trim parent nodes recursively, starting from the root, if all their daughter nodes have more than the threshold number of annotations. Nodes in grey (ϵ, ζ, η) were removed in the previous step. Nodes α, β shown in red will be trimmed because each one has a complete daughter set. Only nodes γ and δ will be used to generate the static dictionary.

Tables
Additional Files
Additional file 1 — IPython Notebook
Tutorial for users interested in using our software within a python script

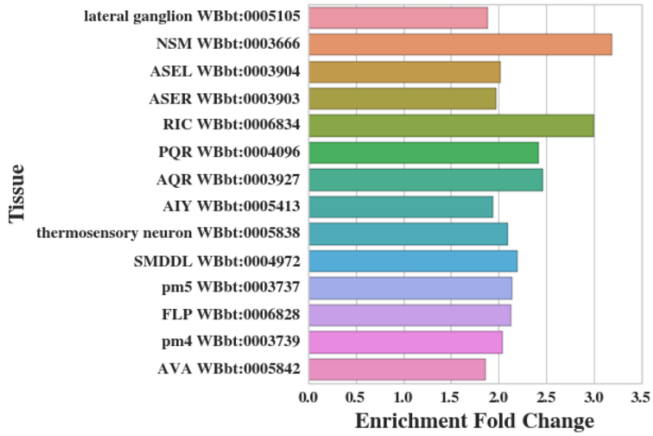
Table 1: Parameter specifications and number of tissues for all dictionaries. The 'Method' column refers to the trimming criterion for the similarity metric. We used two such criteria, 'any' and 'avg'.'any': For a given sister set, if any sister had a similarity exceeding the corresponding threshold, all sisters were removed from the final dictionary. 'avg': For a given sister set, if the average similarity across all the sisters in the set was greater than the corresponding threshold, all sisters were removed from the final dictionary.

Annotation Cutoff	Similarity Threshold	Method	No. Of Terms in Dictionary
25	0.9	any	460
25	0.9	avg	461
25	0.95	any	466
25	0.95	avg	468
25	1.0	any	476
25	1.0	avg	476
33	0.9	any	261
33	0.9	avg	255
33	0.95	any	261
33	0.95	avg	262
33	1.0	any	247
33	1.0	avg	247
50	0.9	any	83
50	0.9	avg	77
50	0.95	any	82
50	0.95	avg	81
50	1.0	any	70
50	1.0	avg	70
100	0.9	any	45
100	0.9	avg	35
100	0.95	any	42
100	0.95	avg	36
100	1.0	any	21
100	1.0	avg	21

Table 2: Comparison of results for a neuronal-enriched geneset from Watson [15]. We ran the same genelist on a dictionary with a minimum annotation cutoff of 50, similarity threshold of 0.95 and similarity method 'any' versus another with a minimum annotation cutoff of 33, similarity threshold of 0.95 and similarity method 'any'. In the table, columns are labeled with their significance value (Q-value) or enrichment fold change followed by a hyphen and a number which indicates which the cutoff for the dictionary that was used for testing. Not all tissues are present in either dictionary. Hyphens denote not-applicable values, which occurs when a particular tissue is not present in both dictionaries. Full table is available on github.

Tissue	Q value-33	Q value-50	Enrichment Fold Change-33	Enrichment Fold Change-50
lumbar ganglion WBbt:0005830	2.9e-19	-	2.2	-
lateral ganglion WBbt:0005105	8.8e-19	8.5e-29	1.9	2.3
nerve ring WBbt:0006749	3.2e-17	8.1e-28	1.7	2.1
retrovesicular ganglion WBbt:0005656	9e-15	1e-19	2.8	3.5
dorsal nerve cord WBbt:0006750	4.2e-12	2.3e-18	2	2.5
pharyngeal nervous system WBbt:0005440	6.7e-10	-	1.9	-
osmosensory neuron WBbt:0008433	2.6e-08	6.2e-12	2.4	3
nociceptor neuron WBbt:0008434	2.6e-08	6.2e-12	2.4	3
ALM WBbt:0005406	2.6e-08	1.1e-11	2.6	3.1
lateral nerve cord WBbt:0006769	2.6e-08	1.2e-11	2.6	3.2

Tissue	Expected	Observed	Enrichment Fold Change	Q value
lateral ganglion WBbt:0005105	75	1.4e+02	1.9	<10^-6
NSM WBbt:0003666	7.2	23	3.2	1.6e-06
ASEL WBbt:0003904	19	39	2	0.0002
ASER WBbt:0003903	19	38	2	0.0004
RIC WBbt:0006834	5	15	3	0.0004
PQR WBbt:0004096	8.2	20	2.4	0.0009
AQR WBbt:0003927	4.9	12	2.5	0.017
AIY WBbt:0005413	11	22	1.9	0.017
thermosensory neuron WBbt:0005838	8.1	17	2.1	0.017
SMDDL WBbt:0004972	5	11	2.2	0.055
pm5 WBbt:0003737	5.6	12	2.1	0.055
FLP WBbt:0006828	5.2	11	2.1	0.065
pm4 WBbt:0003739	5.9	12	2	0.07
AVA WBbt:0005842	7.5	14	1.9	0.096



Drag graph to your desktop to save.
Download output table here

Your list has 2352 valid WormBase genes that have no annotated data or are excluded from testing :

WBGene00000003 - WBGene00000003
WBGene00000005 - WBGene00000005
WBGene00000006 - WBGene00000006
WBGene00000008 - WBGene00000008
WBGene00000017 - WBGene00000017
WBGene00000019 - WBGene00000019

Your list has 745 valid WormBase genes included in statistical testing :

WBGene00000002 - WBGene00000002
WBGene00000023 - WBGene00000023
WBGene00000024 - WBGene00000024
WBGene00000034 - WBGene00000034
WBGene00000037 - WBGene00000037
WBGene00000038 - WBGene00000038

perform another query
if you use this tool, please cite us. If you have ideas, requests or bug-issues, please contact dangeles@caltech.edu

Figure 2: **Screenshot of results from the web GUI.** After inputting a gene-list, the user is provided with the results. An HTML table is output with hyperlinks to the ontology terms. A publication-ready graph is provided below, which can be saved by dragging to the desktop. The graph and the table show anatomy terms in human-readable format, followed by their unique WBbt ID. Finally, lists of the genes used and discarded for the analysis are also presented.

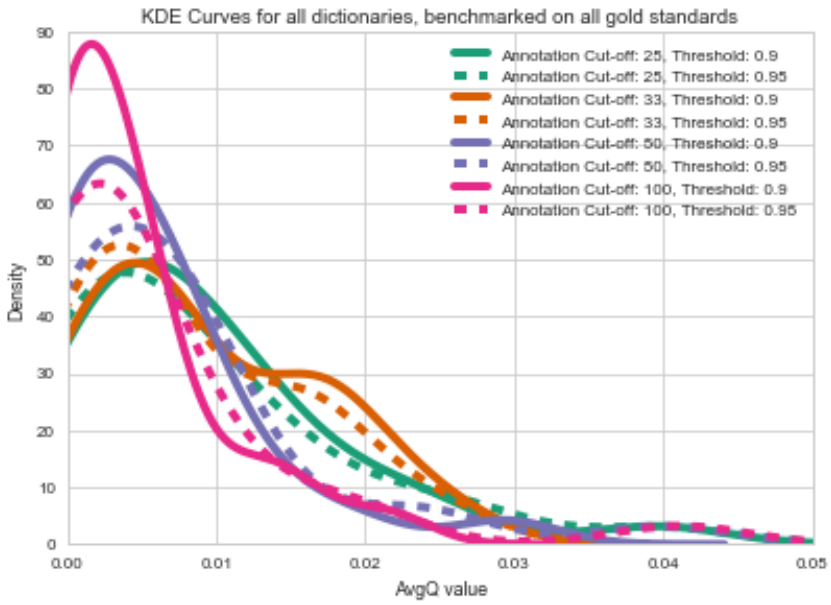
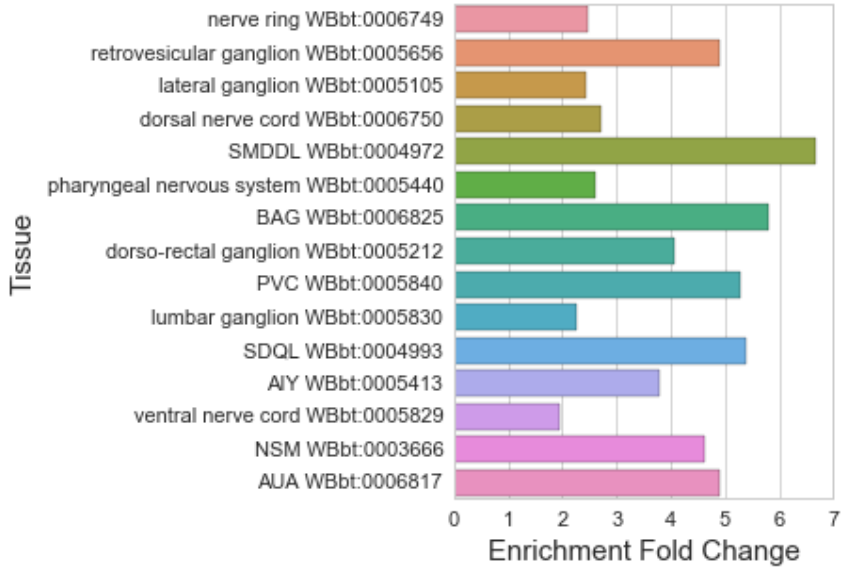
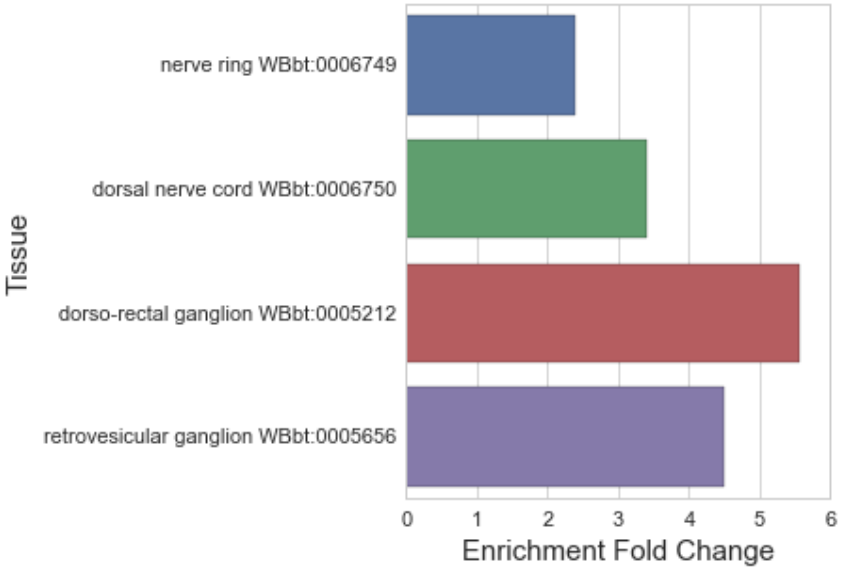


Figure 3: **Kernel density estimates for 30 gold standard datasets.** We ran TEA on 30 datasets we believed to be enriched in particulae tissues and pooled all the results to observe the distribution of q-values. The mode of the distribution for dictionaries with annotation cut-offs of 100 and 50 genes are very similar; however, when the cut-off is lowered to 25 genes, the mode of the distribution shifts to the left, potentially signalling a decrease in measurement power.



(a)



(b)

Figure 4: **Independently derived genesets show similar results within a given dictionary** Bar labels are anatomy terms in human-readable format, followed by their unique WBbt ID. Dictionary with cutoff: 33; threshold: 0.95; method: 'any'. **a)** GABAergic gene set from Watson [15]. **b)** GABAergic gene set from Spencer [13].

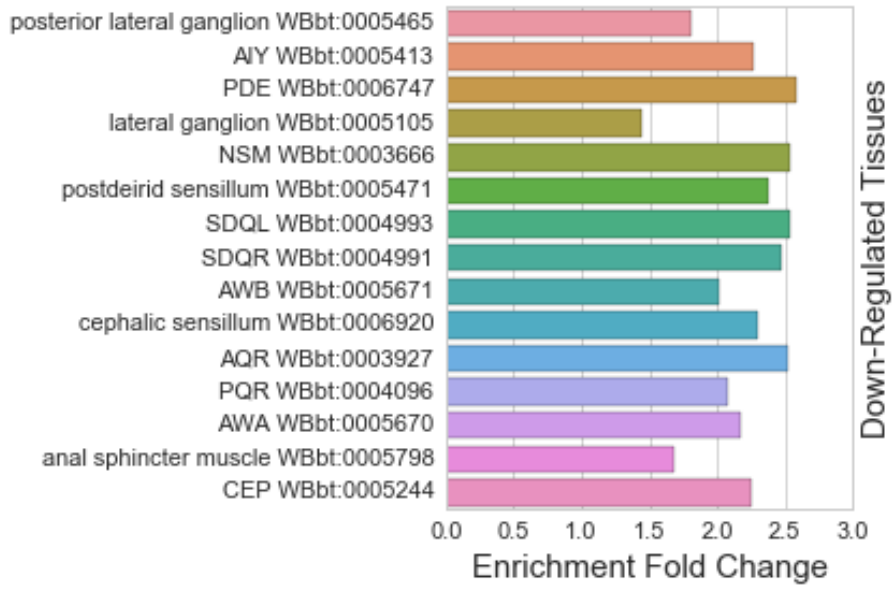


Figure 5: ***D. coniospora* enrichment results** Genes altered in *C. elegans* after 24hr exposure to *D. coniospora*, a fungus, show enrichment in many neuronal tissues.

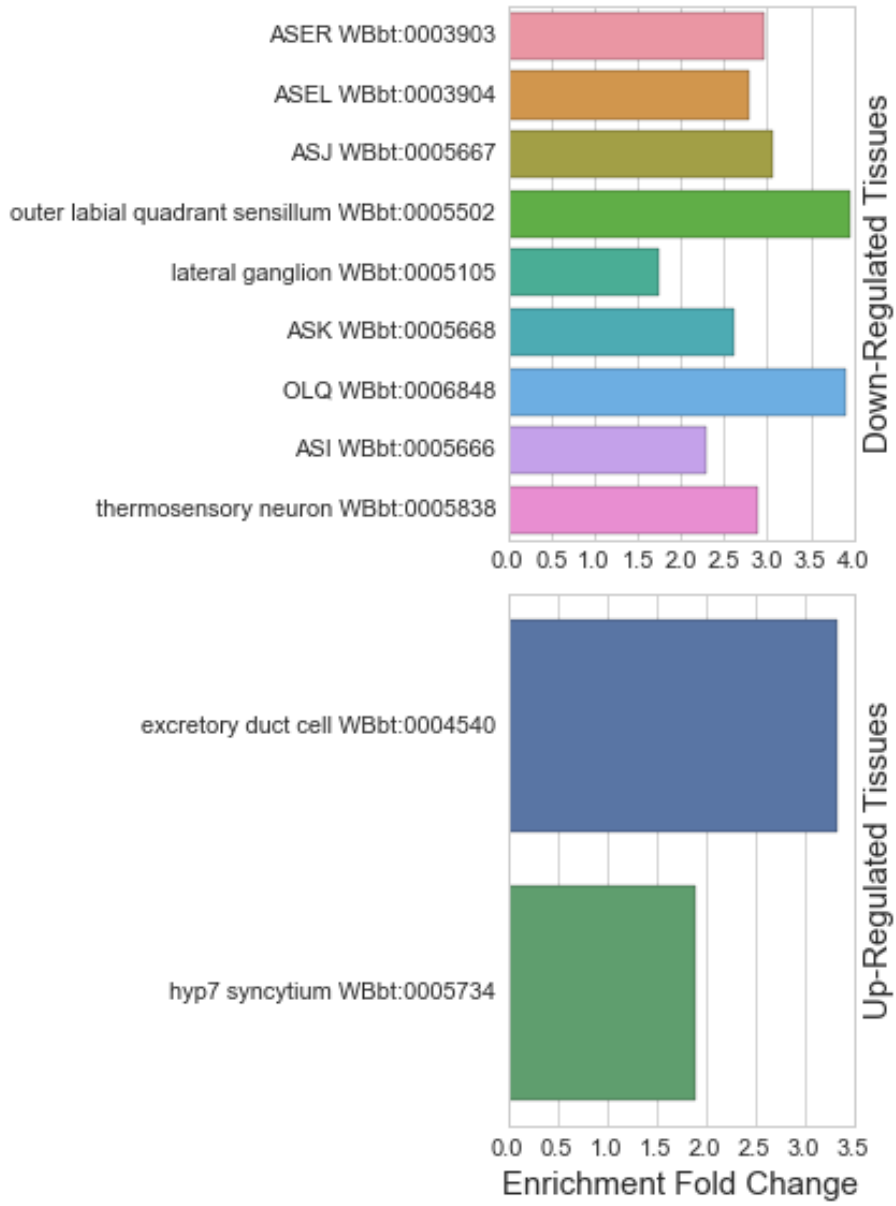


Figure 6: ***Serratia marcescens* enrichment results**. Genes altered in *C. elegans* after 24hr exposure to *Serratia marcescens*, a bacterium, shows down-regulation of chemosensory neurons notably ASER, ASEL, ASJ and ASK. The lateral ganglions are also affected. Up-regulated genes are overrepresented for terms involving the excretory duct cell and hypodermis.