

WormCat: an online tool for annotation and visualization of *Caenorhabditis elegans* genome-scale data

Amy D. Holdorf^{*}, Daniel P. Higgins[†], Anne C. Hart[‡], Peter R. Boag[§], Gregory J.

Pazour^{**}, Albertha J. M. Walhout^{*,**}, and Amy K. Walker^{**}

^{*}Program in Systems Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA

[†]Department of Computer Science, Georgia Technical University, Atlanta, GA 30332-0765, USA

[‡]Department of Neuroscience, Robert J. and Nancy D. Carney Institute for Brain Science, Brown University, Providence, RI 02912, USA

[§]Department of Biochemistry and Molecular Biology, Monash University, 3800 Clayton Australia

^{**}Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA

Short Title: WormCat enables functional gene set identification

1 **Abstract**

2 The emergence of large gene expression datasets has revealed the need for improved
3 tools to identify enriched gene categories and visualize enrichment patterns. While
4 Gene Ontology (GO) provides a valuable tool for gene set enrichment analysis, it has
5 several limitations. First, it is difficult to graphically compare multiple GO analyses.
6 Second, genes from some model systems are not well represented. For example,
7 around 30% of *Caenorhabditis elegans* genes are missing from analysis in commonly
8 used databases. To allow categorization and visualization of enriched *C. elegans* gene
9 sets in different types of genome-scale data, we developed WormCat, a web-based tool
10 that uses a near-complete annotation of the *C. elegans* genome to identify co-
11 expressed gene sets and scaled heat map for enrichment visualization. We tested the
12 performance of WormCat using a variety of published transcriptomic datasets and show
13 that it reproduces major categories identified by GO. Importantly, we also found
14 previously unidentified categories that are informative for interpreting phenotypes or
15 predicting biological function. For example, we analyzed published RNA-seq data from
16 *C. elegans* treated with combinations of lifespan-extending drugs where one
17 combination paradoxically shortened lifespan. Using WormCat, we identified sterol
18 metabolism as a category that was not enriched in the single or double combinations
19 but emerged in a triple combination along with the lifespan shortening. Thus, WormCat
20 identified a gene set with potential phenotypic relevance that was not uncovered with
21 previous GO analysis. In conclusion, WormCat provides a powerful tool for the analysis
22 and visualization of gene set enrichment in different types of *C. elegans* datasets.

23 Introduction

24 RNA-seq is an indispensable tool for understanding how gene expression changes
25 during development or upon environmental perturbations. As this technology has
26 become less expensive and more robust, it has become more common to generate data
27 from multiple conditions, enabling comparisons of gene expression profiles across
28 biological contexts. The most commonly used method to derive information on the
29 biological function of co-expressed genes is Gene Ontology (GO) (The Gene Ontology
30 2019) (Ashburner *et al.* 2000), where each gene has been annotated by three major
31 classifications: *Biological Process*, *Molecular Function* or *Cellular Component*. For
32 example, the *Biological Process* class is defined as a process that an organism is
33 programmed to execute, and that occurs through specific regulated molecular events.
34 *Molecular Function* refers to protein activities, and *Cellular Component* maps the
35 location of activity. Within each of these classifications, functions are broken down in
36 parent-child relationships with increasing functional specificity (**Fig 1A**). However, child
37 classes can be linked to different parent classes, making statistical analysis not
38 straightforward. For example, the child class *phospholipid biosynthetic process* can be
39 linked to both of the parent groups *metabolic process* and *cellular process*. Thus, GO
40 provides multiple descriptors per gene. Although GO was developed to compare gene
41 function across newly sequenced genomes, it became apparent that it could also be
42 used to identify shared functional classifications within large-scale gene expression data
43 (Eisen *et al.* 1998; Spellman *et al.* 1998). Currently, multiple web-based servers that
44 use different statistical tests can be used to determine enrichment of GO terms for a
45 gene set of interest. For example, PANTHER (www.pantherdb.org) provides enriched

46 GO terms determined by Fisher's Exact test with a Benjamini-Hochberg false discovery
47 rate (FDR) correction for 131 species (Mi *et al.* 2019). Because the multiplicity of GO
48 term parent-child relationships can produce complex data structures, specialized
49 ontologies such as GO-Slim use a restricted set of terms, searching biological
50 processes as default (Mi *et al.* 2019). *P*-values are provided for enriched GO terms.
51 Visualization of gene set enrichment data is important for identifying critical elements
52 and communication of information. PANTHER provides pie or bar charts of individual
53 searches (Mi *et al.* 2019). The GOrilla platform generates tables of *P*-values (Eden *et*
54 *al.* 2009) and links to another service, REVIGO, that use semantic graphs to visualize
55 GO terms data (Supek *et al.* 2011). Thus, the GO databases provide a widely used
56 platform for classifying, comparing, and visualizing functional genomic data. However,
57 as outlined below, GO is of limited use for the analysis of *Caenorhabditis elegans* data
58 and visualization of multiplexed datasets.

59
60 The nematode *C. elegans* has been at the forefront of genomics research. It was the
61 first metazoan organism with a completely sequenced genome (Consortium 1998).
62 After the discovery of RNA interference (RNAi)(Fire *et al.* 1998), multiple RNAi libraries
63 were developed for performing genome-wide knockdown screens (Kamath *et al.* 2003;
64 Rual *et al.* 2004). Gene expression profiling studies using microarrays or RNA-seq
65 have compared gene expression in sex-specific, developmental/aging-related, specific
66 gene deletion, tissue-specific, and dietary or stress-related animal conditions (Reinke *et*
67 *al.* 2000; Hillier *et al.* 2005; Baugh *et al.* 2009; Oliveira *et al.* 2009; Deng *et al.* 2011;
68 Schwarz *et al.* 2012; Bulcha *et al.* 2019). While GO has been used extensively to

69 analyze *C. elegans* gene expression profiling data, it has several limitations. First,
70 around 30% of *C. elegans* genes are not annotated in GO databases (Ding *et al.* 2018),
71 excluding these genes from analysis. Thus, these genes are arbitrarily excluded from
72 enrichment statistics. Second, the visualization of enrichment data from comparative
73 RNA-seq datasets is difficult, and this is true not only for *C. elegans* datasets, but for
74 gene expression profile comparisons in any organism. Most users display the output
75 data as lists with *P*-values (Macneil *et al.* 2013) or as pie or bar charts (Ding *et al.*
76 2015), which are not easily multiplexed for comparison of multiple datasets. Finally, it
77 can be challenging to determine which input genes are associated with a given GO
78 classification, which is critical for interpreting the accuracy and biological importance of
79 enriched gene sets.

80

81 We constructed a web-based gene set enrichment analysis tool we named WormCat
82 (WormCatalog) that works independently from GO to identify potentially co-expressed
83 or co-functioning genes in genome-wide expression studies or functional screens.
84 WormCat (www.wormcat.com), uses a concise list of nested categories where each
85 gene is first assigned to a category based on physiological function, and then to a
86 molecular function or cellular location. WormCat provides a scaled bubble chart that
87 allows the visualization and direct comparison of complex datasets. The tool also
88 provides csv files containing input gene annotations, *P*-values from Fisher's exact tests
89 and Bonferroni multiple hypothesis testing corrections. We used WormCat to identify
90 functional gene sets in published gene expression data and large-scale RNAi screens.
91 WormCat reproducibly identified prior GO classifications and provided an easy way to

92 interpret visualization that enables the facile and intuitive comparison of multiple
93 published datasets. We also identified new groups of enriched categories with
94 potentially important biological significance, showing that WormCat provides enrichment
95 information not revealed by GO. Taken together, WormCat offers an alternative and
96 complementary tool for categorizing and visualizing data for genome-wide *C. elegans*
97 studies and may provide a platform for similar annotations in other model organisms
98 and humans.

99

100 **Materials and Methods**

101 **Annotations:**

102 WormBase version WS270 was used to provide WormBase descriptions and provide
103 phenotype information.

104

105 **Scripts:**

106 The processed data were analyzed using R version 3.4.4 (2018-03-15) and depends on
107 the following R packages: datasets, graphics, grDevices, methods, stats, utils, ggplot2,
108 plot flow, scales, ggthemes, pander, data.table, plyr, gdtools, svglite, FSA.

109

110 **Data Availability:** The code and annotation lists are available under MIT Open Source

111 License and can be downloaded from the GitHub repository

112 <https://github.com/dphiggs01/wormcat> along with version-control information.

113 Alternatively, WormCat can be installed directly as an R package using the devtools

114 library. Supplemental material has been deposited at fig share and includes twelve
115 supplemental figures and fourteen supplemental tables.

116

117 **GO searches:** Genes lists were entered as test sets into GOrilla ([http://cbl-](http://cbl-gorilla.cs.technion.ac.il/)
118 [gorilla.cs.technion.ac.il/](http://cbl-gorilla.cs.technion.ac.il/)) (Eden *et al.* 2009) with the WormCat annotation list used as
119 background so that the same background set was used when comparing WormCat and
120 GOrilla. *All* was selected for ontology choices and the *P*-value thresholds were set to
121 10^{-3} . Output selections were Microsoft Excel and REVIGO (Supek *et al.* 2011).

122

123 **Summary of Supplementary Figures**

124 **Figure S1. GO Analysis of upregulated genes from *sams-1(RNAi)* animals by**
125 **Gene Set Enrichment Analysis.** Bar graph showing GO categories returned from
126 *sams-1(RNAi)* upregulated genes (Ding *et al.* 2015) by the WormBase Gene Set
127 Enrichment Analysis tool (Angeles-Albores *et al.* 2016).

128

129 **Figure S2. WormCat verifies known category enrichments from *sams-1(RNAi)***
130 **downregulated genes. (A-B)** Semantic graphs of GO analysis generated by GOrilla
131 (Eden *et al.* 2009) and visualized by REVIGO (Supek *et al.* 2011) of *sams-1(RNAi)*
132 downregulated genes from untreated (**A**) and choline (Ch) treated (**B**) animals. (**C**)
133 WormCat bubble heat plots comparing *sams-1(RNAi)* with and without choline. Gene
134 expression microarray data for **A-C** were obtained from Ding *et al.*, 2015. Bubble heat
135 plot key is the same as **Fig 1D**. CUB, Complement C1r/C1s, Uegf, Bmp1 Domain; PUF,
136 Pumilio and *fem-3* mRNA Binding Factor; ZF, Zinc Finger. (**D**) Venn diagrams showing

137 overlap between *Stress Response* genes in *sams-1(RNAi)* up (pink) or downregulated
138 genes (blue).

139

140 **Figure S3: WormCat analysis of germline-specific microarray data identifies the**

141 **tau tubulin kinase family as a male-specific category. (A)** Category 1 analysis of

142 Oogenic (Oo) or Spermatogenic (Sp) data sets ordered by most enriched in Oo data.

143 Breakdown of data from the Category 1 level for Cell cycle (**B**), Development (**C**),

144 mRNA Functions (**D**), or Cytoskeleton (**E**). All data is from Reinke *et al.* Gen. Trans.

145 Machinery, General Transcription Machinery; Trans. Chromatin, Transcription:

146 Chromatin; ZF, zinc finger.

147

148 **Figure S4: GO analysis visualized by REVIGO of germline RNA-seq data from**

149 **Ortiz et al.** Semantic graphs of GO analysis generated by GOrilla(Eden *et al.* 2009)and

150 visualized by REVIGO of Gender Neutral (**A**), Oogenic (**B**) and Spermatogenic (**C**)

151 germlines.

152

153 **Figure S5: GO analysis visualized by REVIGO of germline microarray data from**

154 **Reinke et al.** Semantic graphs of GO analysis generated by GOrilla and visualized by

155 REVIGO of Oogenic (**A**) and Spermatogenic (**B**) germlines.

156

157 **Figure S6: GO analysis visualized by REVIGO of larval tissue specific microarray**

158 **data from Spencer et al.** Semantic graphs of GO analysis generated by GOrilla and

159 visualized by REVIGO of microarray from larval Muscle (BWM, body wall muscle) (A)
160 Intestine (Int) (B), Hypodermis (Hyp) (C), Excretory cells (Exc) (D), or Neurons (E).

161

162 **Figure S7: GO analysis visualized by REVIGO of adult tissue specific RNA-seq**
163 **data from Kaletsky, et al.** Semantic graphs of GO analysis generated by GOrilla and
164 visualized by REVIGO of RNA-seq data from adult Muscle (Mus) (A) Intestine (Int) (B),
165 Hypodermis (Hyp) (C), or Neurons (D).

166

167 **Figure S8: GO analysis visualized by REVIGO of larval neuronal subtype**
168 **microarray data from Spencer, et al.** Semantic graphs of GO analysis generated by
169 GOrilla and visualized by REVIGO of microarray from larval dopaminergic (Dopa) (A)
170 GABAergic (GABA) (B), *glr-1* expressing (*glr-1*) (C), or Class A motor neurons (Motor)
171 (D).

172

173 **Figure S9: WormCat analysis of upregulated genes in *C. elegans* treated with**
174 **triple combinations of lifespan-changing drugs.** Category 1, 2, and 3 analysis of
175 upregulated genes found by RNA-seq from triple-drug combinations (Admasu *et al.*
176 2018). Pink box denotes drug combination that causes premature death. Allan,
177 allantoin; CYP, Cytochrome P450; EC Material, Extracellular Material; Maj Sperm
178 Protein, Major Sperm Protein; Neur Function; Neuronal Function; NHR, Nuclear
179 Hormone Receptor; Prot General, Proteolysis General; Psora, Psora-4; Rapa,
180 Rapamycin; Rifa, Rifampicin; Short Chain Dehydr., Short Chain Dehydrogenase; Trans

181 Factor, Transcription Factor; TYR kinase, Tyrosine Kinase; ugt, UDP-
182 glycosyltransferase.

183

184 **Figure S10: WormCat analysis of downregulated genes in *C. elegans* treated with**
185 **triple combinations of lifespan-changing drugs.** Category 1, 2, and 3 analysis of
186 downregulated genes found by RNA-seq from triple-drug combinations (Admasu *et al.*
187 2018). Pink box denotes drug combination that causes premature death. Allan,
188 Allantoin; Psora, Psora-4; Rapa, Rapamycin; Rifa, Rifampicin.

189

190 **Figure S11: GO analysis visualized by REVIGO of upregulated genes from RNA-**
191 **seq data from *C. elegans* treated with triple combinations of lifespan extending**
192 **drugs from Admasu *et al.*** Semantic graphs of GO analysis generated by GOrilla and
193 visualized by REVIGO of RNA-seq data from Rifa, Psora, Allan treated (**A**) Rifa, Rapa,
194 Allan treated (**B**), or Rifa, Rapa, Psora treated (**C**).

195

196 **Figure S12: GO analysis visualized by REVIGO of RNA-seq data from a *C. elegans***
197 **RNAi screen for glycogen storage from LaMacchia *et al.*** Semantic graphs of GO
198 analysis generated by GOrilla and visualized by REVIGO of *C. elegans* showing low
199 glycogen storage in an RNAi screen.

200

201 **Summary of Supplementary Tables**

202 **Supplemental Table 1: WormCat annotations.** xlsx file containing *C. elegans* genes
203 arranged alphabetically by Categories with Sequence ID, WormBase ID, and Category

204 1, 2, and 3 annotations along with WormBase descriptions (WormBase version
205 WS270).

206

207 **Supplemental Table 2: WormCat annotation definitions.** xlsx file containing
208 annotation definitions.

209

210 **Supplemental Table 3: Random gene analysis.** xlsx file with tabs containing lists of
211 randomly generated WormBase IDs of 100 (tabs 1-4), 500 (tabs 5-8), 1000 (tabs 9-12
212 and 1500 (tabs 13-16) genes with Category 1, 2 and 3 analysis. NA genes on these
213 tables reflect WormBase IDs that have been merged, marked as dead or updated as
214 not corresponding to a protein-coding gene.

215

216 **Supplemental Table 4: GO analysis of *sams-1(RNAi)* regulated genes.** xlsx file
217 containing GO terms produced by GOrilla (Eden *et al.* 2009) from microarray data for
218 *sams-1(RNAi)* up genes (tab 1), *sams-1(RNAi)* up plus choline (CH) (tab 2), *sams-*
219 *1(RNAi)* down (tab 3), *sams-1(RNAi)* down plus CH (tab 4), and the *sams-1(RNAi)* up
220 genes identified by GOrilla as lipid metabolism with corresponding WormCat
221 annotations (tab 5). Data from Ding *et al.*, 2015.

222

223 **Supplemental Table 5: WormCat analysis of *sams-1(RNAi)* regulated genes.** xlsx
224 file containing Category 1, 2, and 3 analysis from microarray data for *sams-1(RNAi)* up
225 and down genes with or without choline (CH, tabs 1-3) from Ding *et al.*, 2015. Tabs 4-7
226 contain input genes with WormCat annotations for each gene. NA genes on these

227 tables reflect WormBase IDs that have been merged, marked as dead or updated as
228 not corresponding to a protein-coding gene.

229

230 **Supplemental Table 6: WormCat analysis of *sams-1(RNAi)* regulated genes that**
231 **were excluded by GSEA analysis.** xlsx file containing Category 1, 2, and 3 analysis
232 from microarray data for *sams-1(RNAi)* upregulated genes (see **Table S5**, Tab 4) that
233 were excluded by the GSEA tool on WormBase. Data from Ding *et al.*, 2015.

234

235 **Supplemental Table 7: WormCat analysis of germline-expressed genes from Ortiz**
236 ***et al.*** xlsx file containing Category 1, 2, and 3 analysis from RNA-seq data for Germline
237 Neutral (GN), Oogenic (Oo) or Spermatogenic (Sp) datasets (tabs 1-3). Tabs 4-6
238 contain input genes with WormCat annotations for each gene. NA genes on these
239 tables reflect WormBase IDs that have been merged, marked as dead or updated as
240 not corresponding to a protein-coding gene. Tabs 7-10 contain GO analysis by GOrilla
241 for Germline Neutral (GN), Oogenic (Oo) or Spermatogenic (Sp) datasets.

242

243 **Supplemental Table 8: WormCat analysis of germline-expressed genes from**
244 **Reinke *et al.*** xlsx file containing Category 1, 2, and 3 analysis from microarray data of
245 Oogenic (Oo) or Spermatogenic (Sp) datasets (tabs1-3). Tabs 4-5 contain input genes
246 with WormCat annotations for each gene. NA genes on these tables reflect WormBase
247 IDs that have been merged, marked as dead or updated as not corresponding to a
248 protein-coding gene. Tabs 6-7 contain GO analysis by GOrilla for Oogenic (Oo) or
249 Spermatogenic (Sp) datasets.

250

251 **Supplemental Table 9: WormCat analysis of larval tissue-specific genes from**

252 **Spencer *et al.*** xlsx file containing Category 1, 2, and 3 analysis from microarray data

253 from *selective enriched* datasets (tabs 1-3). Tabs 4-12 contain input genes with

254 WormCat annotations for each gene for all tissue and cell types examined. NA genes

255 on these tables reflect WormBase IDs that have been merged, marked as dead or

256 updated as not corresponding to a protein-coding gene. Tabs 13-22 contain GO

257 analysis by GOrilla from selective enriched datasets from Muscle (BWM, body wall

258 muscle), Intestine (Int), Hypodermis (Hyp), Excretory cells (Exc), Neurons (Neuro, pan-

259 neuronal), Dopaminergic (Dopa), GABAergic (GABA), *glr-1* expressing (*glr-1*) or Class

260 A Motor neurons (Motor).

261

262 **Supplemental Table 10: WormCat analysis of adult tissue-specific genes from**

263 **Kaletsky *et al.*** xlsx file containing Category 1, 2, and 3 analysis from RNA-seq data of

264 enriched (en) and unique (un) datasets (tabs 1-3). Tabs 4-11 contain input genes with

265 WormCat annotations for each gene for all tissue and cell types examined. NA genes

266 on these tables reflect WormBase IDs that have been merged, marked as dead or

267 updated as not corresponding to a protein-coding gene. Tabs 12-16 contain GO

268 analysis by GOrilla of enriched genes from Muscle (Mus), Intestine (Int), Hypodermis

269 (Hyp) or Neuron (Neur).

270

271 **Supplemental Table 11: WormCat analysis of upregulated genes from a**

272 **combinatorial RNA-seq study of lifespan enhancing drugs.** xlsx file analysis of data

273 from Admasu *et al.* comparing upregulated genes from single, double, and triple
274 combinations of lifespan inducing drugs. Tabs 1-3: Category 1-3 analysis of genes
275 upregulated by single drugs. Tabs 4-7: Input genes with WormCat annotations for each
276 single drug treatment. Tabs 8-10: Category 1-3 analysis of upregulated genes by double
277 drug combinations. Tabs 11-14: Input genes with WormCat annotations for each double
278 combination drug treatment. Tabs 15-17: Category 1-3 analysis of upregulated genes
279 by triple-drug combinations. Tab 18: Genes from the *Metabolism: lipid: sterol* category
280 from the Rifa/Rapa/Psora set with corresponding lifespan data from Murphy *et al.*
281 (yellow) (Murphy *et al.* 2003). Tabs 19-21: Input genes with WormCat annotations for
282 each triple combination drug treatment. NA genes on these tabs reflect WormBase IDs
283 that have been merged, marked as dead or updated as not corresponding to a protein-
284 coding gene. Tabs 22-24: GO analysis by GOrilla (Eden *et al.* 2009) for each triple
285 combination drug treatment.

286

287 **Supplemental Table 12: WormCat analysis of downregulated genes from a**
288 **combinatorial RNA-seq study of lifespan enhancing drugs.** xlsx file analysis of data
289 from Admasu *et al.* comparing downregulated genes from single, double and triple
290 combinations of lifespan inducing drugs. Tabs 1-3: Category 1-3 analysis of genes
291 downregulated by single drugs. Tabs 4-7: Input genes with WormCat annotations for
292 each single drug treatment. Tabs 8-10: Category 1-3 analysis of downregulated genes
293 by double drugs combinations. Tabs 11-14: Input genes with WormCat annotations for
294 each double combination drug treatment. Tabs 15-17: Category 1-3 analysis of
295 downregulated genes by triple-drug combinations. Tabs 18-20: input genes with

296 WormCat annotations for each triple combination drug treatment. NA genes on these
297 tabs reflect WormBase IDs that have been merged, marked as dead or updated as not
298 corresponding to a protein-coding gene.

299

300 **Supplemental Table 13: WormCat annotations for genes in the Ahringer RNAi**
301 **library.** xlsx file containing gene and WormBase IDs along with Category 1, 2, and 3
302 annotations for the clones represented in the Ahringer RNAi library (Kamath *et al.*
303 2003).

304

305 **Supplemental Table 14: WormCat analysis of a genome scale RNAi screen from**
306 **LaMacchia *et al.*** xlsx file containing Category 1, 2, and 3 analysis of RNAi screen data
307 of all, high and low glycogen stained animals (Tabs 1-3). Tabs 4-6 contain input genes
308 with WormCat annotations for each gene. Tab 7 is Glycogen low genes analyzed by
309 GOrilla. Tabs 8-11 show GO terms with multiple categories containing *cyc-1*, *vha-6*,
310 *pbs-7* and Y71F9AL.17.

311

312 **Results**

313 ***C. elegans* gene annotation**

314 The *C. elegans* genome encodes ~19,800 protein-coding genes, ~260 microRNAs and
315 numerous other non-coding RNAs (WormBase version WS270). We annotated all *C.*
316 *elegans* genes first based on physiological functions, and, when these functions were
317 unknown or pleiotropic, according to molecular function or sub-cellular location (See
318 **Table S1** for annotations, **Table S2** for Category definitions). Our annotations are
319 structured as nested categories, enabling classification into broad (Category 1; Cat1), or
320 more specific categories (Category 2 or 3; Cat2 or Cat3). This annotation has the
321 advantage of including information from multiple sources in addition to GO. For
322 example, we used phenotype information available in WormBase (Lee *et al.* 2018), for
323 Cat1 assignments. Importantly, the phenotypic data present in WormBase (Lee *et al.*
324 2018) was only used if phenotypes were: 1) derived from wild type animals, 2)
325 examined in detail in peer-reviewed publications, and (?) 3) represented in two
326 independent screens. If a gene was ascribed a clear physiological function with these
327 criteria, we assigned it to a physiological category, examples of which include *Stress*
328 *response*, *Development*, and *Neuronal function*. If gene products have multiple
329 functions within the cell, act in multiple cells type or different developmental times, we
330 prioritized assignment to molecular categories. Molecular categories harbor both genes
331 whose products comprise molecular machines, as well as the chaperones or regulatory
332 factors that are necessary for the function of such machines. We used information on
333 molecular function of human orthologs to classify *C. elegans* genes that had not been
334 molecularly defined in nematodes and showed highly similar in BLAST scores. For

335 example, we classified the *C. elegans* gene W03D8.8 in *Metabolism: lipid: beta*
336 *oxidation* based on a BLAST score of $e = 7 \times 10^{-37}$ and similarity over 92% of its length to
337 human ACOT4 (acyl-CoA thioesterase 4). For genes with weaker homology to human
338 genes, we further refined assignments using BLAST (Altschul *et al.* 1990) and the NCBI
339 Conserved Domain server (Marchler-Bauer *et al.* 2017). We used these tools to
340 determine if there was significant homology or shared domains between *C. elegans* and
341 human proteins, then used information in UniProt (www.uniprot.org) for the human
342 proteins to determine molecular classification. For example, we placed the *C. elegans*
343 gene T26E4.3 in *Protein modification: carbohydrate* based on a BLAST core of $e =$
344 4×10^{-7} over 95% of its length to human alpha fucosyltransferase 1 and identification of a
345 Fut1_Fut2-like domain by the NCBI conserved domain server with an e score of
346 6.16×10^{-36} . However, while the gene BE10.3 is referred to in the WormBase description
347 as an ortholog of human FUT9 (fucosyltransferase 9) (**Table S1**), we found no
348 homology to human genes by NCBI BLAST or domain conservation across all
349 organisms with the NCBI Conserved Domain server. Therefore, we classified BE10.3 in
350 *Unknown*. Finally, if no biological or molecular function could be assigned, protein sub-
351 cellular localization was used for annotation. For example, a protein with a predicted
352 membrane-spanning region that lacks characterization as a receptor would be placed in
353 *Transmembrane protein*. Genes with no functional information were classified as
354 *Unknown* (Cat1). There are 8160 genes that lacked sufficient information for
355 classification in physiological, molecular or sub-cellular localization categories and were
356 classified in *Unknown*. Many of these genes are *C. elegans* or nematode-specific,
357 however, some have homology to human genes of unknown function. WormBase also

358 aggregates microarray and RNA-seq information and annotates genes that respond to
359 pharmacological treatments (Lee *et al.* 2018). We also used this information to
360 differentiate genes within *Unknown: regulated by multiple stresses* that respond to at
361 least two commonly used stressors. This classification does not imply these genes have
362 a function in the stress response. It does allow identification of genes with otherwise
363 unknown functions that are common responders to stress. This may be useful to
364 distinguish RNA-seq datasets that respond similarly to pharmacological stressors or can
365 serve as a source to identify specific genes of interest for additional study. We also
366 included pseudogenes and non-coding RNAs in our annotation list. These genes
367 commonly appear in RNA-seq data; including them in the annotation list allows them to
368 be labeled within the user's input dataset. In this way, we were able to leverage
369 multiple data sources to categorize *C. elegans* genes into potentially functional
370 biological groups.

371

372 **WormCat.com allows web-based searches of input genes and generates scaled** 373 **bubble charts and gene lists**

374 WormCat.com maps annotations to input genes, then determines category enrichment
375 for Cat1, Cat2 and Cat3 (**Fig 1B**). Determination of category enrichment in a gene set
376 of interest compared to the entire genome can rely on several commonly used statistics
377 such as the Fisher's exact test and the Mann-Whitney test (Mi *et al.* 2019). We used
378 Fisher's exact test to determine if categories were overrepresented because it is
379 accurate down to small sample sizes, which may occur in high resolution classifications
380 (Mcdonald 2014). In addition, we included the Bonferroni FDR correction (Mcdonald

381 2014). To determine the number of false positives after the Fisher's test or the FDR
382 correction, we tested randomized gene lists of 100, 500, 1000, 1500 genes and found
383 that small numbers of genes were returned using a *P*-value cut-off of 0.05 (for, example
384 5 genes were returned on the 1000 gene random set). Few genes were returned from
385 any of the randomized sets using an FDR cutoff of 0.01 (**Table S3**). Because an FDR <
386 0.01 is relatively stringent, Fisher's exact test *P*-values will also be provided allowing
387 users to make independent evaluations on the statistical cut-offs.

388

389 The WormCat website (www.wormcat.com) provides gene enrichment outputs in
390 multiple formats (**Fig 1C**). First, all input genes are listed with mapped annotations
391 (*rgs_and_categories.csv*). Genes that matched at least one Cat1, Cat2 and Cat3
392 classification are returned with Fisher's exact test *P*-values (*Cat1.csv*, *Cat2.csv* or
393 *Cat3.csv*). Next, Cat1, Cat2, and Cat3 matches with an FDR correction of < 0.01 are
394 returned as CSV files named *Cat1.apv*, *Cat2.apv* and *Cat3.apv* (appropriate *P*-value).
395 Finally, the *Cat.apv* files are used to generate two types of graphical output. First, it
396 constructs scaled heat map bubble charts (*Cat1.*, *Cat2.*, *Cat3.sgv*) where color signifies
397 *P*-values and size specifies the number of genes in the category (**Fig 1D**). The scaling
398 for these graphs is fixed so that multiple datasets can be compared and graphed
399 together. Second, a sunburst graph is built with concentric rings of Cat1, Cat2, and
400 Cat3 values (**Fig 1E**). In these graphs, sections of each ring correspond to categories,
401 with the size of the section proportional to the number of genes in the category. On the
402 website, each ring section is clickable to generate a sub-graph-based division within a
403 section. For example, clicking a single Cat1 section would generate a subgraph with all

404 the Cat2 and Cat3 subdivisions located within. This graphical output is likely to be most
405 useful for visualization of a single RNA-seq dataset, or genetic screening data. Thus,
406 WormCat provides multiple outputs to allow inspection of individual input genes,
407 generation of gene tables and *P*-values, and graphical visualization of enrichments.

408

409 **Comparison of GO and WormCat analysis of *sams-1(RNAi)* enrichment data**

410 To determine the utility of the WormCat annotations, we first analyzed microarray data
411 we previously generated to compare gene expression changes after knockdown of
412 *sams-1*, with and without dietary supplementation of choline (Ding *et al.* 2015). *sams-1*
413 encodes an S-adenosylmethionine (SAM) synthase, which is an enzyme that produces
414 nearly all of the methyl groups used in methylation of histones and nucleic acids, in
415 addition to the production of the membrane phospholipid phosphatidylcholine (PC)
416 (Mato and Lu 2007). *sams-1* RNAi or loss-of-function (*lof*) animals have extended
417 lifespan (Hansen 2005), increased lipid stores (Walker *et al.* 2011), and activated innate
418 immune signatures (Ding *et al.* 2015). *sams-1* animals have low PC (Walker *et al.*
419 2011), but those levels are restored with supplementation of choline (Ding *et al.* 2015),
420 which supports SAM-independent phosphatidylcholine synthesis (Vance 2014) (**Fig**
421 **2A**). Gene expression changes in *sams-1(RNAi)* animals could result from perturbation
422 in different SAM-dependent pathways. To determine which transcriptional changes
423 occurred downstream of alterations in PC synthesis, we performed microarrays with
424 RNA from *sams-1(RNAi)* and *sams-1(RNAi)* animals supplemented with choline. 90%
425 of genes that changed in expression in *sams-1(RNAi)* animals returned to wild type
426 levels after choline supplementation. Therefore, the expression of the remaining 10% of

427 genes was altered by *sams-1* RNAi independently of phosphatidylcholine levels (Ding *et*
428 *al.* 2015).

429

430 In order to identify GO terms enrichment with WormCat, we submitted genes up or
431 down regulated 2-fold or more in *sams-1(RNAi)* animals to both WormCat and GOrilla
432 (Eden *et al.* 2009). We used REVIGO (Supek *et al.* 2011) to visualize GO output. Both
433 GOrilla/REVIGO (**Fig 2B; FigS2 A, B; Table S4**) and WormCat (**Fig 2C; Table S5**)
434 identified categories of stress-response and metabolism linked to lipid accumulation in
435 the genes that are upregulated upon *sams-1* RNAi, which is in agreement with our
436 previous analysis (Ding *et al.* 2015). Interestingly, the relative importance of lipid
437 metabolism is different in the two analyses. In the WormCat analysis, *Metabolism: lipid*
438 was the third most enriched Cat2 category with a *P*-value of 1.2×10^{-9} (**Table S5**). In the
439 GO analysis, however, *lipid metabolic process* was found with a modest enrichment of
440 FDR corrected *P*-value = 5×10^{-2} (**Table S4**). WormCat identified 41 genes in the
441 *Metabolism: lipid* category, whereas GOrilla's GO term search identified 33 genes in
442 *lipid metabolic process* (**Fig 2E; Table S4**). Further inspection showed that six of the
443 genes identified by solely by GOrilla were phospholipid lipases or phosphatases, one
444 was an undefined hydrolase with no homology or domain similarity to genes with known
445 lipid functions, and one was a transmembrane protein that may be better classified in
446 other categories (see **Table S4** for GO lipid genes annotated by WormCat, tab 5
447 "GO_lipid_sams_up"). For example, lipases that hydrolyze phospholipids are the end
448 points of metabolic pathways but produce second messengers acting in signaling
449 pathways. One of these genes, Y69A2AL.2 has significant similarity to the human

450 phospholipase A2 gene, PLA2G1B (BLAST e score of 2×10^{-11}). This class of
451 phospholipases cleave 3-sn-phosphoglycerides to produce the signaling molecule
452 arachidonic acid (Xu *et al.* 2009); therefore a classification of *Signaling* is likely more
453 reflective of its biological function than *Metabolism: lipid*. Taken together, WormCat
454 identifies more genes that are directly relevant to the increased lipid storage phenotype
455 observed with *sams-1(RNAi)* or (*lof*) animals (Walker *et al.* 2011; Smulan *et al.* 2016).

456
457 Next, we compared WormCat analysis of *sams-1(RNAi)* upregulated genes to the Gene
458 Set Enrichment Analysis (GSEA) tool located in the WormBase suite (Angeles-Albores
459 *et al.* 2016). GSEA, a GO based tool, identified similar categories as GOrilla with a
460 concurrently high score for *lipid catabolic process* (**Fig S1**). Our test set included 773
461 genes (**Table S5, tab4**); however, 286 of these genes were excluded from the GSEA
462 analysis (**Table S6**), similar to the percentage excluded in a GOrilla analysis (Ding, et
463 al. 2018). Unlike GOrilla, GSEA provides the user with gene IDs of excluded genes
464 (**Table S6**). Therefore, we asked if these genes were excluded because their functions
465 were undefined or if they were instead capable of classification. We found that 118 of
466 the 286 excluded genes were classified as *Unknown* by WormCat (**Table S6**).

467 However, 92 of the 476 genes GSEA included were also *Unknown* in WormCat analysis
468 (**Table S5, tab 4**). Thus, the genes within this set that are classified as *Unknown* by
469 WormCat only partially overlap with those that are excluded from GO analysis.

470 Furthermore, WormCat classified 117 genes within the 286 genes excluded from GSEA,
471 with 16 in non-coding categories and the remaining 101 in protein coding categories
472 such as *Cytoskeleton*, *Metabolism* and *Proteolysis: proteasome* (**Table S6**). Thus,

473 analysis of genes excluded from GO shows that an important fraction can be annotated
474 and that *Unknown WormCat* categories are represented in both genes included and
475 excluded from GO analysis.

476
477 Next, we used WormCat to analyze genes downregulated in *sams-1(RNAi)* animals.
478 We noted enrichment in *Development: germline and mRNA function* categories in
479 *sams-1(RNAi)* animals and that this enrichment is lost with choline treatment (**Fig S2D**,
480 **Table S5**). This is consistent with the reduction in embryo production after *sams-*
481 *1(RNAi)* and the rescue of fertility when PC levels are restored by choline
482 supplementation (Walker *et al.* 2011; Ding *et al.* 2015). *Stress response* categories,
483 however, are enriched in downregulated genes from both *sams-1(RNAi)* and *sams-*
484 *1(RNAi)* choline treated animals (**Fig S2C; Table S5**). This appears to contrast with the
485 complete loss of enrichment after choline treatment in the upregulated stress-response
486 genes (**Fig 2C; Table S5**). However, inspection of the annotated gene lists returned by
487 WormCat shows that the individual genes within the down-regulated *Stress response*
488 category are different (**Fig S2E; Table S5**). Thus, on a gene by gene level, this data
489 shows that the effects of choline supplementation are distinct for the up and
490 downregulated genes in the *Stress response* category. In addition, this demonstrates
491 that by providing both gene set enrichment and annotation of individual genes,
492 WormCat provides a level of analysis that is difficult to achieve by traditional GO
493 methods.

494

495 **Tau-tubulin kinases family are enriched in spermatogenic germlines**

496 *C. elegans* is a robust model system for studying development and differentiation.
497 Study of hermaphrodite germline development has been of particular interest, as it first
498 produces sperm, after which it switches to oocyte production (Hubbard and Greenstein
499 2005). This concurs with distinct gene expression programs for both processes
500 (Greenstein 2005; L'hernault 2006). Recently, the Kimble lab performed RNA-seq on
501 dissected germlines from genetically female (*fog-2(q71)*) and genetically male (*fem-*
502 *3(q96)*) animals (Ortiz *et al.* 2014) (**Fig 3A**). Genes that were expressed in both
503 germlines were called gender-neutral (GN), in contrast to genes that are specific to
504 female (Oo, oogenic) or male (Sp, spermatogenic) germlines (Ortiz *et al.* 2014). We
505 used WormCat to analyze the categories that were enriched in each dataset. We found
506 that GN genes are strongly enriched for growth, DNA, transcription, and mRNA
507 functions (**Fig 3B; Table S7**), which is expected because the germline is undergoing
508 extensive mitotic and meiotic divisions. We further found that *Chromosome dynamics*
509 and *Meiotic functions* were enriched in the GN dataset (**Fig 3C; Table S7**), as were
510 *mRNA functions of Processing and Binding* (**Fig 3D; Table S7**). Oo genes were
511 enriched for mRNA binding proteins, especially the zinc finger (ZF) class (**Fig 3D; Table**
512 **S7**). These include such as maternally deposited *oma-1*, *pie-1*, *pos-1*, and *mex-1*, *mex-*
513 *5* and *mex-6* mRNAs, which are known to function in oocytes (Lee and Schedl 2006)
514 (**Table S7**). ZF proteins with unknown nucleic acid binding specificity were also
515 enriched in the Oo dataset (**Fig 3D; Table S7**), suggesting that many of these may also
516 be produced in the maternal germline. In an independent data set comparing RNA from
517 germline-less (*glp-4(bn2)*), oocyte (*fem-3(gof)*) and sperm-producing (*fem-1(lof)*)
518 animals by microarray analysis (Reinke *et al.* 2000), we also observed that categories in

519 mRNA functions, transcription, development and cell cycle control were enriched (**Fig**
520 **S3A-D, Table S8**).

521

522 As expected, Sp genes are enriched for *Major Sperm Proteins* (MSPs), which are
523 necessary for sperm crawling (**Fig 3B; Table S7**). Interestingly, a class of potential
524 cytoskeletal regulators, *tau-tubulin kinases* (TTKs), were also enriched in Sp genes (64
525 of 71, P -value of 8.8×10^{-34}) (**Fig 3E; Table S7**). One TTK, *spe-6*, was previously
526 isolated in a screen for spermatogenesis defects and is thought to be involved in
527 phosphorylation of MSPs to allow the sperm to crawl (Varkey *et al.* 1993).

528 Underscoring the potential importance of the TTKs in the male germline, WormCat also
529 produced an enrichment in *tau tubulin kinases* in the Reinke, et al. spermatogenic gene
530 sets (**Fig S3E, Table S8**). Thus, WormCat has identified a class of kinases that may be
531 important for sperm-specific functions (**Fig 3F**).

532

533 To directly compare gene set enrichment from WormCat and GO, we analyzed each of
534 these germline-enriched datasets with GOrilla and used REVIGO (Supek *et al.* 2011) for
535 visualization (**Fig S4A-C, Fig S5A-B; Table S7, S8**). For the GN genes, the top 5 of
536 the 544 significantly enriched categories were nucleic acid metabolic process
537 (GO:0090304), nucleobase-containing compound metabolic process (GO:0006139),
538 heterocycle metabolic process (GO:0046483), cellular aromatic compound metabolic
539 process (GO:0006725), and organic cyclic compound metabolic process (GO:1901360)
540 (**FigS4A, Table S7, see tabs 7, 8**). These GO categories are highly overlapping and
541 are linked to multiple general processes involving nucleic acids. One gene

542 GO:0006139, *gut-2*, an LSM RNA binding protein, was present in 23 different GO
543 categories (**Table S7**). Comparison of these GO categories found that each contains
544 genes placed in distinct WormCat categories. For example, *gut-2* was placed in *mRNA*
545 *Functions* in WormCat, *ama-1*, the RNA Pol II large subunit, placed in *Transcription:*
546 *General Machinery*, *brc-1*, the BRCA1 ortholog, placed in *DNA* and *nsun-5*, a
547 mitochondrial RNA methyltransferase placed in *Metabolism: mitochondria*. These
548 WormCat categories are the top five identified in the GN dataset (**Fig 3B, Table S7**).
549 Thus, while WormCat and GO are both identify nucleic acid-related processes as
550 among the most highly enriched in the GN dataset, the WormCat data is more concise
551 and easily aligned with the molecular processes.

552

553 Within the spermatogenic datasets from Ortiz *et al.* and Reinke *et al.*, WormCat
554 identified a class of kinases, tau tubulin kinases (TTKs), that have the potential to
555 function in sperm motility. General categories of phosphorus metabolic process
556 (GO:0006793), phosphate-containing compound metabolic process (GO:0006796) and
557 peptidyl-threonine phosphorylation (GO:0018107) were among the top five most
558 enriched categories by GO from the Spermatogenic dataset, however, the TTKs as a
559 group were not selectively identified from these very broad signaling categories in either
560 spermatogenic data set (**Table S7, Table S8**). Thus, WormCat provided advantages
561 over GO in the germline data sets by providing less redundant and more easily
562 interpreted data and, most importantly, by identifying novel categories with potential
563 links to biological function.

564

565 **Identification of post-embryonic tissue-specific gene expression categories**

566 Improved technologies for cell-type-specific marker expression, nematode disruption,
567 and deep sequencing of small RNA quantities have allowed construction of gene
568 expression datasets from larval (Spencer *et al.* 2011) and adult somatic tissues
569 (Kaletsky *et al.* 2018). To generate data from larval cell types, the Miller lab used cell-
570 type specific tagged green fluorescent proteins to label a wide variety of larval tissues
571 and examined mRNA expression in tiling microarrays (Spencer *et al.* 2011). RNA from
572 each cell type would include tissue-specific, broadly expressed and ubiquitously
573 expressed genes. To define cell-type specific transcripts, Spencer *et al.* designated
574 *selectively enriched genes* as expressed more than 2-fold vs. the whole animal and as
575 present in few cell types (Spencer *et al.* 2011). First, we performed WormCat analysis
576 on the *selectively enriched* gene sets and found distinct gene set enrichments for each
577 tissue type (**Fig 4A, Table S9**). For instance, body wall muscle (BWM) was enriched for
578 *Muscle Function* and *Cytoskeleton* (**Fig 4B; Table S9**). The category *Metabolism* was
579 enriched in both intestine (Int) and hypodermis (Hyp), whereas *Stress responses*
580 appeared more specific for the intestine, and *Extracellular material* for the hypodermis
581 (**Fig 4B, C; Table S9**). This likely reflects the role of the intestine in mediating contact
582 with the bacterial diet after ingestion and the importance of the hypodermis for cuticle
583 formation in larval development. While metabolic genes are expected to be required
584 across multiple cell types, some cell types have specialized metabolic requirements.
585 Both intestine and hypodermis are enriched for lipid metabolism genes at the Cat2 level.
586 However, Cat3 analysis shows that sterol and sphingolipid genes drive this enrichment
587 in the intestine while hypodermal lipid enrichment involves more broad categories with

588 minor enrichments in *Metabolism: lipid: binding* and *Metabolism: lipid: lipase* (*P*-values
589 of 4.51×10^{-04} and 2.86×10^{-04} , which did not satisfy the FDR cutoff)(**Fig 4D; Table S9**).
590 The Cat1 level analysis showed strong enrichment of transmembrane (TM) transporters
591 in all tissues including the intestine, excretory cells and in neurons, however the Cat2
592 level shows enrichment of distinct classes of transporters (**Fig 4B; Table S9**) aligning
593 with functions such as nutrient uptake, waste processing, and channel activity in each of
594 these cell types.

595
596 Next we examined the data from Kaletsky *et al.*, who performed RNA-seq from adult *C.*
597 *elegans* sorted for muscle (Mus), intestinal (Int), hypodermal (Hyp) and neurons
598 (Kaletsky *et al.* 2018) (**Fig 4E; Table S10**). They computationally separated genes to
599 distinguish expression specificity, demarking "enriched", "unique" and "ubiquitously"
600 expressed categories. We used the "enriched" gene sets in WormCat analysis and
601 found that WormCat correctly mapped muscle or neuronal genes to those cell types
602 (**Fig 4F; Table S10**). At the Cat1 level, *Extracellular material* was enriched in muscle,
603 hypodermis and intestine (**Fig 4F; Table S10**). At the Cat2 levels, *Extracellular material*
604 diverged with *matrix* showing enrichment in muscle and *collagen* showing enrichment in
605 intestine and hypodermis (**Fig 4G; Table S10**). However, the collagen genes enriched
606 in intestine and hypodermis were distinct (**Fig 4G; Table S10**), perhaps reflecting
607 differing roles for these collagens in the cuticle vs. in basement membranes.

608 Distinguishing individual genes for this comparison is very cumbersome in commonly
609 used GO servers and therefore represents an advantage of using WormCat. Previous
610 studies found that two intestinal basement membrane collagens were produced in non-

611 hypodermal tissues (Graham *et al.* 1997); however, this data suggests that others could
612 be locally produced by the intestine. Kaletsky *et al.* also noted enrichment of metabolic
613 function in adult hypodermis with GO analysis. Metabolic gene enrichment was also
614 detected by WormCat analysis of their data (**Fig 4H; Table S10**), as well as in the larval
615 data from Spencer *et al.* (**Fig 4D; Table S9**).

616

617 In our annotation strategy, we chose to restrict genes in categories such as *Neuronal*
618 *function* to those that are specific to that tissue, and that have a described physiological
619 function. Genes which functioned in neurons as well as other tissues were placed in
620 more general molecular function-based categories. With this approach, we hoped to
621 reduce false-positive identification of neuronal categories outside the nervous system,
622 yet permit the identification of related, yet functionally less-specific groups. For
623 example, while the WormCat analysis of the neuronal tissues in the Spencer *et al.* and
624 Kaletsky *et al.* datasets showed strong enrichment of neuronal-specific categories, it
625 also included categories of genes likely to function in both neurons and other tissues, or
626 that contained genes that had not yet been classified *in vivo*. These categories include
627 *Metabolism: insulin* (**Fig 4D, H; Table S10**), *Transmembrane (TM) transport*, *Signaling*
628 (**Fig 4B, F; Table S10**) and *Transmembrane protein* (**Fig 4B; Table S10**). This is in line
629 with the analysis by both Kaletsky *et al.* and Ritter *et al.* (Ritter *et al.* 2013) which also
630 noted insulin expression across tissues and noted that more insulin genes were
631 expressed at higher levels in adult neurons.

632

633 In order to distinguish the utility of WormCat from GO for the tissue-specific Spencer *et*
634 *al.* and Kaletsky *et al.* datasets, we used GOrilla (Eden *et al.* 2009) to generate GO
635 analysis and visualized the data with REVIGO (Supek *et al.* 2011) (**Figure S6-S8;**
636 **Table S9, S10**). There were many similarities among the categories. For example,
637 categories linked to the *Cytoskeleton* are highly enriched in the muscle datasets from
638 Kaletsky *et al.* by GOrilla and WormCat (**Fig 4F, Fig S7A, Table S10**). In another
639 example, *Stress response* categories were highly enriched by both WormCat and GO in
640 the larval (Spencer *et al.* 2011) and adult (Murphy *et al.* 2003) intestine (**Fig 4F, Fig**
641 **S6B, S7B, Table S10**). However, as shown above, WormCat identified the insulin gene
642 family as strongly enriched in both the larval (**Fig 4D**) and adult (**Fig 4H**) neuronal
643 tissue. Insulins were not identified as a class by our GO analysis. Instead, they were
644 distributed among less specific categories such as biological regulation (GO:0065007),
645 regulation of biological process (GO:0050789) and regulation of cellular process
646 (GO:0050794) (**Fig S5, S6; Table S9, S10**). Thus, WormCat finds the major categories
647 shown by GOrilla in the tissue-specific data and also identifies additional enriched
648 groups.

649

650 The seven transmembrane protein family in *C. elegans* presented an annotation
651 challenge. This class comprises around 8% of all protein-coding genes that seem likely
652 to function in neurons, yet whose functions are undescribed (Robertson and Thomas
653 2006). Some have significant homology to mammalian G protein-coupled receptors
654 (GPCRs), while others are nematode or *C. elegans* specific (Robertson and Thomas
655 2006). In order to identify and classify these proteins as accurately as possible, GPCRs

656 with strong evidence for neuron-specific activity were placed in *Neuronal function*, while
657 all other potential GPCRs were classified by protein domain and homology. For
658 developing a list of potential GPCRs, we selected genes identified in WormBase as
659 containing a transmembrane domain as well as those we initially annotated as GPCRs
660 in the *Signaling* category. To recover any genes missed by these approaches, we
661 added all *Unknown* proteins from our annotation list. We submitted the protein
662 sequences for these genes to the NCBI Conserved Domain search tool (Marchler-Bauer
663 *et al.* 2017) and selected all the genes in these groups that contained a seven-
664 transmembrane (7TM) domain (**Fig 5A**). Next, we used BLASTP to determine the
665 degree of homology to human GPCRs, which would reflect the conservation of function.
666 Genes that had BLASTP scores of $e < 0.05$ on the NCBI server were classified in
667 *Signaling: heteromeric G protein: receptor*. Those with e scores > 0.05 were classified
668 as *TM protein: 7TM*, with class designated by WormBase in Cat3. Thus, genes with
669 classified within *Neuronal function* or *Signaling* have a strong likelihood of GPCR
670 function, whereas those in *TM protein: 7TM* have not been sufficiently defined.
671 *Signaling: G protein* categories are enriched in neuronal genes sets from both Kaletsky
672 *et al.* and Spencer *et al.* (**Fig 5B, C; Table S9, S10**) and 7TM proteins show enrichment
673 in the larval pan-neuronal, *glr-1*-expressing neurons and motor neurons (**Fig 5C; Table**
674 **S9, S10**). Thus, our annotation strategy allows separation of GPCRs with a highly
675 likelihood of neuronal function, yet still permits enrichment of the larger class of 7TM
676 proteins in neuronal tissues.
677

678 In order to directly compare WormCat and GO on the larval neuronal data sets, we
679 examined category enrichment of Spencer *et al.* pan-neuronal and motor neuron genes
680 in GO by GOrilla (Eden *et al.* 2009), using REVIGO (Supek *et al.* 2011) for visualization
681 (**Fig S6, S8; Table S9**). The most enriched category in the pan-neuronal or motor
682 neuron datasets was G protein-coupled receptor signaling (GO:0007186). Next, we
683 used WormCat to determine how we had annotated genes within GO:0007186 and
684 found that this GO category included genes we had classified in *Signaling: Heteromeric*
685 *G protein (G-alpha subunits and receptors)*, *Neuronal Function: Synaptic function*
686 (neuropeptides and neurotransmitter receptors) and *TM protein: 7TM receptor* (**Fig 5C,**
687 **Table S9**). While inclusion of the G protein signaling apparatus and neuropeptide
688 ligands is appropriate for the broad category of G protein signaling, the GO categories
689 do not differentiate between GPCRs with a high likelihood of function from the 7TM
690 proteins that have not been functionally characterized. In addition, many of the *nlp*
691 genes listed in GO:0007186 have not been functionally characterized and thus, it is not
692 clear if they are bona fide GPCR ligands or could interact with other receptors outside of
693 GPCR signaling (Li and Kim 2008). Therefore, WormCat improves on GO analysis for
694 these datasets by providing more nuanced information on the function of these genes in
695 GPCR pathways.

696
697 Neuronal genes from adult (Kaletsky *et al.* 2018) and larval gene sets (Spencer *et al.*
698 2011) also showed strong enrichment in Cat2 and Cat3 classifications within *Neuronal*
699 *function*, such as *Synaptic function*, *neuropeptide*, and *neurotransmitter (nt) receptor*
700 (**Fig 5D, E; Tables S9-S10**). *Cilia* genes were also enriched in the pan-neuronal and

701 dopaminergic larval gene sets (**Fig 5D; Table S9**). Neurons are the only ciliated cells in
702 *C. elegans* and cilia occur on multiple neuronal subtypes (Inglis *et al.* 2007). However,
703 all dopaminergic neurons are ciliated (Inglis *et al.* 2007), and are therefore more likely to
704 show enrichment. Taken together, our WormCat analysis of these large tissue-specific
705 gene sets provides a detailed view of gene classes specific to muscle, hypodermis,
706 intestine, and neurons in larvae and adults. We have identified differential enrichment
707 in lipid metabolism genes and collagens from intestine and hypodermis, defined a
708 classification system for GPCRs and 7TMs and identified *Cilia* as a major enriched
709 category in dopaminergic neurons. Much of this information goes beyond what is
710 revealed in GO analysis and provides predictions that can be useful to design future
711 studies. Identification of these types of nuanced tissue-specific patterns is an important
712 step to understanding how specific cell types function.

713

714 **Drug interactions limiting lifespan induce changes in sterol metabolism**

715 *C. elegans* is particularly suited to studies determining gene expression changes in
716 response to a panel of treatments in a whole animal, and to correlate these changes to
717 physiological function. For example, Admasu *et al.* generated a complex gene
718 expression dataset by performing parallel RNA-seq on animals treated with five
719 lifespan-increasing drugs that affect distinct pathways (Allantoin, Rapamycin, Metformin,
720 Psora-5, and Rifampicin). They used five pairwise combinations and three triple drug
721 combinations to determine if any combination lead to further lifespan extension, and to
722 identify gene expression profiles associated with increased longevity (Admasu *et al.*
723 2018). They found that one triple drug combination (Rifa/Psora/Allan) activated

724 lipogenic metabolism through the transcription factor SBP-1/SREBP-1 and determined
725 that the drug-induced longevity was dependent on SBP-1 function (Admasu *et al.* 2018).
726 The authors also made the striking observation that a distinct triple drug combination
727 (Rifa/Rapa/Psora) reduced lifespan, even though each single drug or drug pairs
728 increased longevity (Admasu *et al.* 2018). To determine if any gene expression
729 categories might explain this effect, we used WormCat to analyze category enrichment
730 for the up and downregulated genes for each single drug, pairwise or triple drug
731 combination (**Fig 6A, Figs S9, S10; Tables S11, S12**). Similar to the author's KEGG
732 analysis (Admasu *et al.* 2018), we observed *Metabolism: lipid* enrichment in long-lived
733 Rifa/Rapa/Psora-treated animals (**Fig 6A, Table S11**), however, we also noted that
734 *Metabolism: lipid* was enriched in all three combinations with WormCat. Next, we
735 examined the up and downregulated genes to determine if any categories correlated
736 with the failure to survive in the Rifa/Rapa/Psora treated animals. We did not find
737 category signatures in the downregulated genes that appeared to correlate with the
738 decrease in longevity (**Fig S10; Table S12**). However, upregulated genes from the
739 short-lived Rifa/Rapa/Psora treated animals were enriched in another specific class of
740 lipid metabolic genes: sterol metabolism (**Fig 6A, Fig S9**). Closer examination of the
741 single and pairwise combinations showed that the enrichment of sterol metabolic genes
742 only appeared in the triple combination with poor survival (**Fig 6B**). *C. elegans* do not
743 use cholesterol as a membrane component (Ashrafi 2007). Thus, this category does not
744 include cholesterol synthesis genes, but does include genes involved in modification of
745 sterols, for example, in steroid hormone production (Watts and Ristow 2017).
746 Examination of individual genes (**Table S11, Tab 18 Sterol Genes**) showed that five of

747 the 19 had lifespan phenotypes and four had lethality related phenotypes in WormBase,
748 consistent with their effects on survival in Admasu *et al.* Furthermore, Murphy *et al.*
749 showed that three of the 19 sterol genes are upregulated in another long-lived model,
750 *daf-2(mu150)*, and two of these, *stdh-1* and *stdh-3* are required for lifespan extension in
751 *daf-2(mu150)* animals (Murphy *et al.* 2003). Thus, the category enrichments captured
752 by WormCat for this drug study have identified sterol metabolism genes as potential
753 players in the paradoxical lifespan shortening effects of the Rifa/Rapa/Psora
754 combination.

755

756 In order to compare gene set enrichment of the triple drug combinations from WormCat
757 with GO, we analyzed upregulated genes from the Rifa/Psora/Allan, Rifa/Rapa/Allan
758 and Rifa/Rapa/Psora treated animals in GOrilla (Eden, 2009) and visualized the data
759 with REVIGO (Supek, 2011) (**Fig S11; Table S11**). WormCat and GO showed multiple
760 similarities. For example, WormCat and GO identified extracellular matrix-linked
761 categories in all three triple combinations (WormCat: EC MATERIAL; GOrilla:
762 GO:0030198: extracellular matrix organization) (**Fig S9; Table S11**). However,
763 WormCat identified *Metabolism: lipid* in all three combinations, whereas GO analysis by
764 GOrilla only identified categories linked to lipid metabolism (GO:0006629: lipid
765 metabolic process ($q = 5.63 \times 10^{-03}$), GO:0044255 cellular lipid metabolic process ($q =$
766 1.49×10^{-02}) and GO:0006631 fatty acid metabolic process ($q = 2.16 \times 10^{-02}$)) in the
767 Rifa/Rapa/Psora dataset (**Table S11**). WormCat also showed a much higher
768 enrichment score for *Metabolism: lipid*, $p = 2.00 \times 10^{-14}$) (**Table S11**). Thus, as in the

769 *sams-1* microarray data discussed previously, WormCat provides an improved tool for
770 determining enrichment of metabolic genes.

771

772 WormCat also found an enrichment of transcription factors in each of the triple
773 combinations, with specific enrichments in nuclear hormone receptors and
774 homeodomain genes in the Rifa/Psora/Allan upregulated set (**Fig S9**) Enrichments of
775 nuclear hormone receptors in *C. elegans* is potentially of interest as they may regulate
776 multiple metabolic regulatory networks (Arda *et al.* 2010). However, GOrilla only
777 identified categories linked to transcription factors (GO:0006355: regulation of
778 transcription, DNA-templated, GO:0051252: regulation of RNA metabolic process,
779 GO:2001141: regulation of RNA biosynthetic process, GO:1903506 regulation of nucleic
780 acid-templated transcription and GO:0019219 regulation of nucleobase-containing
781 compound metabolic process) in the Rifa/Psora/Allan dataset. No individual class of
782 transcription factors were identified in any of the triple combinations by GO (**Table S11**),
783 thus WormCat offers a clear advantage over GO by providing increased coverage
784 across diverse categories of gene function.

785

786 **Identification of gene set enrichments in RNAi screening data**

787 In order to use WormCat to analyze genome-scale RNAi screening data, we mapped
788 WormCat annotations to the list of genes in the Ahringer library (Kamath *et al.* 2003)
789 (**Table S13**). To test this approach, we used data from the Roth lab who screened the
790 Ahringer library for changes in glycogen storage in *C. elegans* and identified more than
791 600 genes, scored as glycogen high, glycogen low and abnormal localization

792 (Lamacchia *et al.* 2015) (**Fig 7A, Table S14**). The authors functionally classified all hits
793 from the screen with an in-house annotation list, graphed the percentage within each
794 group, and noted high percentages of genes with roles in metabolism (electron transport
795 chain), signaling, protein synthesis or stability, and trafficking (Lamacchia *et al.* 2015),
796 however, they were unable to assign statistical significance to any of the groups.
797 WormCat identified similar groups as the LaMacchia *et al.* functional classification for
798 the glycogen low candidates. For example, we identified *Metabolism: mitochondria*,
799 complex I, III, IV, and V and found that these categories were statistically enriched (**Fig**
800 **7B; Table S14**). However, signaling was not enriched (**Table S14**). Thus, WormCat is
801 able to identify statistically relevant pathways in genome-scale RNAi screen data.

802

803 To provide a direct comparison between WormCat and GO with this data set, we
804 determined the GO term associated with the glycogen low data by GOrilla (Eden *et al.*
805 2009) and visualized the data with REVIGO (Supek *et al.* 2011) (**Figure S12; Table**
806 **S14**). 185 separate GO terms were identified in this data set compared to the four Cat1
807 level terms identified by WormCat (*Metabolism, Lysosome, Proteolysis Proteasome* and
808 *Trafficking*) (**Fig 7B, Table S14**). WormCat also finds a limited number of Cat2
809 groupings within these sets including *Metabolism: mitochondria, Lysosome: vacuolar*
810 *ATPase, Proteolysis Proteasome:19S, 20S, and Trafficking:ER/Golgi*) (**Fig 7B, Table**
811 **S14**). This large difference in number of significantly enriched categories stems from
812 the multiple, overlapping categories present in the GO analysis. For example, the
813 mitochondrial gene *cyc-1* (Cytochrome C oxidase) is represented in 87 of the GO terms,
814 whereas the annotation in WormCat is *METABOLISM: mitochondria* (**Table S14, tab 8**).

815 Similarly, the vacuolar ATPase *vha-6* is represented in 39 of GO terms returned, the
816 proteasomal component *psb-7* is present in 23, and the ER/Golgi COP I component
817 Y71F9AL.17 is in 21 (see **Table S14**, tabs 9-11). This GO term redundancy provides
818 the user with a complex, hard to interpret list. In addition, GO terms that are repeated
819 fewer times (such as those containing the trafficking gene Y71F9AL.17) become
820 marginalized in a complex list. Thus, with this dataset WormCat provides easily
821 distinguished categories with clear links to biological or molecular function. The GO
822 terms show the same genes repeated in a large fraction of the categories and obscure
823 categories with less gene redundancy.

824

825

826

827 **Discussion**

828 **WormCat provides new insights into comparative RNA-seq data**

829 Current technology allows for the routine use of genome-scale experiments for the
830 generation of gene expression data. The goal of these experiments is often to identify
831 classes of genes that add insight to biological functions, as well as to highlight selected
832 genes for individual analysis. GO analysis, while widely used, is difficult to apply to
833 datasets with multiple combinations of treatments or genetic perturbations. Further, for
834 *C. elegans*, current GO analysis is often inaccurate and misses useful physiological and
835 molecular information. Here we have shown that WormCat can annotate gene
836 categories, provide enrichment statistics, and display user-friendly graphics for gene
837 sets identified from *C. elegans* gene expression studies. Furthermore, our visualization
838 strategy allows comparison across multiple datasets, facilitating identification of
839 categories that can be linked to shared biological functions.

840

841 Our initial, script-based, smaller-scale version of WormCat highlighted changes in
842 metabolic gene expression in *C. elegans* with changes in levels of the methyl donor S-
843 adenosylmethionine (SAM) or methyltransferases modifying H3K4me3 (Ding *et al.*
844 2018). In this study, we have expanded the annotation list, developed a web-based
845 server, and added an additional graphical output. We used WormCat to successfully
846 analyze data from metabolic, tissue-specific, and drug-induced expression changes.
847 This analysis provides not only validation and use-case examples but also additional
848 insights into the known gene expression patterns. For example, our examination of
849 germline gene expression datasets from the Kimble and Kim labs (Reinke *et al.* 2000;

850 Ortiz *et al.* 2014) identified a large class of microtubule kinases (tau tubulin kinases,
851 TTK) as enriched in spermatogenic gene sets and as a co-enriched gene set with major
852 sperm proteins (MSPs). One TTK, *spe-6*, has been previously identified in a screen for
853 mutants with defects in sperm development (Varkey *et al.* 1993). Our results suggest
854 that many genes in this family could have important functions in spermatogenesis and
855 that the appearance of MSPs and TTKs in a dataset could also serve as a marker for
856 maleness. Finally, we used WormCat to analyze a dataset consisting of RNA-seq from
857 *C. elegans* treated with multiple lifespan changing drugs alone or in combination, plus
858 one mutation animal strain that extends lifespan (Admasu *et al.* 2018). The
859 classification and graphical output allowed us to identify upregulation of sterol
860 metabolism genes in a triple-drug combination that was not present in the single or
861 double drug treatments. Thus, WormCat identified a gene set that may be important for
862 the effects of the lifespan-altering drugs in this assay.

863

864 **Strengths and weaknesses of WormCat**

865 We developed WormCat to overcome some of the limitations of GO analysis when
866 analyzing *C. elegans* gene expression data and to utilize specific phenotype data
867 available in WormBase. In addition, we specifically engineered WormCat to classify
868 data for identification of co-expressed or co-functioning gene sets. Finally, we
869 developed two graphical outputs, a scaled heat map/bubble plot and a sunburst plot.
870 The modular nature of the bubble plot allows multiple datasets to be grouped and
871 compared, while the sunburst plot gives a concise view of single datasets, as may be
872 obtained with screening data. Our validation with random gene testing and analysis of

873 *C. elegans* gene expression data from metabolic, tissue-specific, and drug-treated
874 animals shows that WormCat is a robust tool that provides biologically relevant gene
875 enrichment information. There are three main areas that WormCat provides an
876 advantage over using GO that are apparent in our case studies. First, as discussed
877 above, we found that in some of our test cases, WormCat identified broader sets of
878 genes within categories or categories that were not identified by GO. Second, the
879 WormCat output is much easier to interpret; the bubble charts provide intuitive
880 visualization and the tables provide clear access to the enrichment statistics and
881 annotation of the input genes. Third, the availability of the annotations for each input
882 gene enables comparisons between genes in categories. For example, we found that
883 while *Extracellular material: collagen* was enriched in both intestine and hypoderm in
884 the Kaletsky *et al.* data set, the genes were non-overlapping, suggesting tissue-specific
885 expression of collagen genes. This comparison would be difficult to make with GO, as
886 many common GO servers do not supply the genes with each category in an easily
887 accessible manner. Directly comparing the genes within WormCat and GO categories
888 from our previously published dataset of gene expression after *sams-1* knockdown, we
889 found that WormCat identified a broader set of lipid metabolic genes than GO analysis
890 from GOrilla and that the genes identified only by GO analysis might be better classified
891 in different categories to reflect their biological functions. Thus, WormCat provides an
892 alternative to GO with advantages in output that improve data interpretation and access
893 to gene annotations that allow deeper comparisons among categories. In some cases,
894 WormCat also identifies categories that are not found by GO.
895

896 However, there are several limitations to WormCat. First, while multiple researchers
897 with varied expertise curated our annotation list, some genes may be mis-annotated, or
898 some Cat2 or Cat3 groups may fit better in other Cat1 classifications. We will update
899 the WormCat annotation list at periodic intervals while providing access to the previous
900 annotation lists. Second, each *C. elegans* gene was given a single, nested annotation,
901 rather than a group of annotations as in GO. We chose to prioritize the visualization of
902 enriched gene sets in this instance, using a single annotation per gene to permit
903 graphing in scaled heat maps. Access to the program and annotation lists for the local
904 application also allows users to customize the annotation lists according to their
905 preferences.

906

907 Annotation lists of genome-scale data are likely to contain errors. We have defined
908 several sources of error and have taken corrective steps. In some cases, a gene may
909 be simply mis-annotated. For example, a component of the *General transcription*
910 *machinery* was placed in *Signaling* by the annotator. In others, the classification system
911 may be incorrect. An example of this would be classifying enzymes that modify small
912 molecules as protein modification. To estimate the mis-classification error rate, we
913 generated a list of 3000 random WormBase IDs. We mapped each ID to our annotation
914 list and rechecked the annotations. We found 29/2294 genes (1.3%) whose
915 annotations were incorrect by our criteria (13 of these were Unknown genes which
916 could be classified in other categories). This suggests around 300 genes in the entire
917 dataset may be mis-annotated by our criteria, many representing Unknown genes which

918 could acquire classification. We will periodically update the WormCat annotation lists to
919 accommodate new gene information and correct errors.

920

921 It is important to note that some gene classifications depend on criteria that are open for
922 interpretation. For example, transcription factors regulating genes within a pathway are
923 grouped within a linked category to allow identification of co-functioning genes. For
924 instance, *efl-1*, a master regulator of cell cycle genes is annotated as *Cell cycle:*
925 *transcriptional regulator* instead of with the more broadly acting trans-regulatory factors
926 in *Transcription factor: E2F*. To allow for different interpretations of the annotation
927 strategy, we have set up a GitHub site (<https://github.com/dphiggs01/wormcat>) where
928 the annotation list and scripts for executing WormCat can be downloaded and
929 customized by the user to accommodate differences in annotation preference.

930

931 The value of gene set enrichment is also highly dependent on the criteria used to
932 specify the regulated genes. In the present study, we used the same criteria as the
933 respective authors, except that we separated up and downregulated genes where
934 necessary. For example, in the Kaletsky *et al.* tissue-specific data, the authors provided
935 data for all genes expressed in each tissue, enriched genes (expressed at FDR great
936 than 0.05, and \log_2 fold change greater than 2 relative to other tissues), or unique genes
937 (\log_2 RPKM greater than 5) significantly differentially expressed in comparison to the
938 expression of each of the three other tissues (FDR greater than 0.05, \log_2 fold change
939 greater than 2 for each comparison) (Kaletsky *et al.* 2018). We found the best
940 resolution of WormCat categories between the tissues occurred with the enriched

941 datasets, rather than with all genes or unique gene sets. This suggests that gene lists
942 with all expressed genes may require more stringent statistical cutoffs, but also that
943 WormCat may not be as suited to highly filtered data.

944

945 **Application to other organisms**

946 By developing WormCat specifically for analyzing *C. elegans* gene sets, we were able
947 to take advantage of available data on WormBase but limited the applicability of our
948 annotation list with other organisms. Although researchers in mammalian fields can
949 access pathway analysis pipelines such as Ingenuity Pathway Analysis (Qiagen,
950 (Kramer *et al.* 2014)) that are focused on identifying functionally linked genes, these
951 programs do not necessarily provide a simple graphical output for comparative analysis.
952 WormCat analysis generating the scaled heat/bubble charts can be adapted for use
953 with other organisms by running the program locally with altered annotation lists.
954 Replacing gene IDs and the Cat1, Cat2 and Cat3 values with any annotation allows
955 customization of the pipeline to any other organism. Thus, the modular nature of
956 WormCat allows adaptation to multiple annotation strategies within *C. elegans* or to
957 other organisms, allowing a streamlined visualization for examining genome-scale
958 expression or screen data.

959

960 **Acknowledgments**

961 We wish to thank members of the Walker and Walhout labs for helpful discussion.
962 Funding to A.K.W NIH NIA 1R01AG053355. A.J.M.W. grants NIH grants DK068429 and
963 GM122502.

964

965 **Figure Legends**

966 **Figure 1: WormCat annotates and visualizes *C. elegans* gene enrichment from**

967 **genome-scale data. (A)** Diagram comparing the parent-child methods for linking GO

968 terms with the nested tree strategy used for annotating *C. elegans* genes in WormCat.

969 **(B)** Screenshot of the WormCat web page showing the data entry form. **(C)** Flow chart

970 diagramming steps and outputs from the WormCat program. Data outputs are in tabular

971 comma-separated values (CSV) and scalable vector graphics (SVG) formats. **(D)**

972 Legend for scaled bubble charts showing number of genes referenced to size and *P*-

973 value referenced to color. In graphs, Category 1, 2 and 3 are differentiated by

974 capitalization, size and italics. **(E)** Legend for sunburst plots showing concentric rings

975 visualizing Category 1, 2 and 3 data.

976

977 **Figure 2: WormCat verifies known category enrichments *sams-1(RNAi)***

978 **upregulated genes. (A)** Schematic showing metabolic pathways linking methionine,

979 SAM, choline, and phosphatidylcholine. Gene expression microarray data for **B-D** were

980 obtained from (Ding *et al.* 2015). **(B)** Semantic plot of GO enriched classifications

981 generated by REVIGO (Supek *et al.* 2011) from *sams-1(RNAi)* Up genes. **(C)** WormCat

982 visualization of categories enriched in genes upregulated in *sams-1(RNAi)* animals with

983 and without choline supplementation in order of Cat1 strongest enrichment. Categories

984 2 and 3 are listed under each Category 1, with Category 2 or 3 sets that appeared

985 independently of a Category 1 listed last. Bubble heat plot key is the same as Fig 1D.

986 **(D)** *sams-1(RNAi)* Up plus choline (Ch) genes visualized by REVIGO. **(E)** Venn diagram

987 showing overlap between WormCat *Metabolism: lipid* and GO *Lipid process* gene
988 annotations. ABC, ATP-Binding Cassette; Ch, Choline; CUB, Complement C1r/C1s,
989 Uegf, Bmp1 domain; EC Material, Extracellular Material; NHR, Nuclear Hormone
990 Receptor; Prot General, Proteolysis General; Prot Proteasome, Proteolysis
991 Proteasome; SAM, S-adenosylmethionine; TM Transport, Transmembrane Transport;
992 ugt, UDP-glycosyltransferase

993

994 **Figure 3: Analysis of germline-specific RNA-seq data identifies the tau tubulin**
995 **kinase family as a male-specific category. (A)** Schematic showing germlines used
996 for female (top) or male (bottom)-specific RNA-seq analysis from Ortiz *et al.*, and the
997 mutant alleles to cause these phenotypes. **(B)** WormCat Category 1 analysis of
998 Germline neutral (GN), Oogenic (Oo) or Spermatogenic (Sp) datasets ordered by most
999 enriched in GN data. **(C-E)** Breakdown of WormCat enrichment from the Category 1
1000 level for Cell Cycle **(C)**, mRNA Functions and Nucleic Acid **(D)**, and Cytoskeleton **(E)**.
1001 Bubble heat plot key is the same as Fig 1D. **(F)** Schematic showing predicted
1002 phosphorylation and organization of MSPs during *C. elegans* sperm maturation based
1003 on WormCat findings. APC, Anaphase Promoting Complex; Chr Dynamics,
1004 Chromosome Dynamics; mRNA Func., mRNA Function; MSP, Major Sperm Protein;
1005 Phos, Phosphorylation; Protein Mod, Protein Modification; Prot Proteasome, Proteolysis
1006 Proteasome; RBM, RNA Binding Motif; TTK, Tau Tubulin Kinase; TM Transport,
1007 Transmembrane Transport; Trans: Gen Mach, Trans: Chromatin, Transcription:
1008 Chromatin; Transcription: General Machinery; Trans Factor, Transcription Factor; ZF,
1009 Zinc Finger

1010

1011 **Figure 4: WormCat analysis of tissue-specific gene sets reveals the importance of**

1012 **the intestine in stress-responsive categories. (A)** Diagram showing larval tissues

1013 isolated in tiling array data used in figures **B-D** from Spencer *et al.* **(B)** WormCat

1014 Category 1 enrichment for larval tissue-specific *selective enriched* gene sets shows

1015 differentiation of Body wall muscle (BWM), Intestine (Int), Hypodermis (Hyp), Excretory

1016 cells (Exe) and Neurons (Neuro). **(C-D)** Category 2 and 3 breakdown of Stress

1017 Response **(C)** and Metabolism **(D)**. **(E)** Schematic showing adult tissues isolated for

1018 RNA-seq used in figures **F-I** from Kaletsky *et al.* **(F)** Category 1 analysis of enriched

1019 genes shows the differentiation of muscle and neuronal functions. **(G-H)** Category 2 and

1020 3 breakdown of Extracellular Material gene enrichment including a Venn Diagram

1021 showing relationships between collagen genes in intestine and hypodermis **(G)**, and

1022 Metabolism **(H)**. Bubble heat plot key is the same as Fig 1D. 1CC, 1-Carbon Cycle; EC

1023 Material, Extracellular Material; GST, Glutathione-S-transferase; Maj Sperm Protein,

1024 Major Sperm Protein; Neur Function, Neuronal Function; Prot General, Proteolysis

1025 General; Short Chain Dehyd, Short Chain Dehydrogenase; TM Transport,

1026 Transmembrane Transport

1027

1028 **Figure 5: Detailed analysis of neuronal tissue-specific gene sets reveals specific**

1029 **enrichment for cilia gene expression on dopaminergic neurons. (A)** Flow chart

1030 showing the process for annotating seven transmembrane (7 TM) proteins. *e* value is

1031 the statistical score provided by the NCBI BLAST server. Asterisk on Signaling notes

1032 that only predicted GPCRs within this category were submitted to the NCBI conserved

1033 domain server. **(B-E)** Breakdown of Neuronal Function to Category 2 and 3 from larval
1034 data in Kaletsky *et al.* **(B, D)** or adult data in Spencer *et al.* **(C, E)**. 7TM receptor, Seven
1035 Transmembrane Receptor; BWM, Body Wall Muscle; dmsr, DroMyoSuppressin
1036 Receptor Related; Dopa, Dopaminergic Neurons; Exe, Excretory Cells; GABA, Gamma-
1037 Aminobutyric Acid-Specific Neurons; *glr-1*, Glutamate Receptor-Specific Neurons;
1038 Hetero G protein, Heterotrimeric G Protein; Hyp, Hypodermis; IFT, Intraflagellar
1039 Transport; Int, Intestine; mks module, Meckel-Gruber syndrome Module; Motor, Motor
1040 Neurons; nt Receptor, Neurotransmitter Receptor; Neuro, Neurons; Pan-N, Pan-
1041 Neuronal

1042

1043 **Figure 6: WormCat analysis of RNA-seq data from *C. elegans* treated with**
1044 **combinations of lifespan-lengthening drugs reveals the emergence of sterol**
1045 **metabolism in drug combinations limiting survival. (A)** Comparison of *Metabolism:*
1046 *lipid: sterol* enrichment in single, double and triple-drug combinations shows sterol
1047 emergence in the Rifa/Rapa/Psora gene set (Admasu *et al.* 2018). **(B)** Diagram
1048 showing a summary of data from lifespan changes after triple drug treatment from
1049 Admasu *et al.* Pink box denotes drug combination that causes premature death. Bubble
1050 heat plot key is the same as Fig 1D. Allan, Allantoin; Psora, Psora-4; Rapa, Rapamycin;
1051 Rifa, Rifampicin

1052

1053 **Figure 7: WormCat analysis of a genome-scale RNAi screen quantitates**
1054 **categories of candidate genes. (A)** Schematic of the RNAi screen from LaMacchia *et*

1055 *a/*. identifying candidate genes that altered glycogen staining. **(B)** Sunburst diagram

1056 from low glycogen candidates showing significantly enriched categories.

1057

1058

1059

1060 Literature cited

- 1061 Admasu, T. D., K. Chaithanya Batchu, D. Barardo, L. F. Ng, V. Y. M. Lam *et al.*, 2018
1062 Drug Synergy Slows Aging and Improves Healthspan through IGF and SREBP
1063 Lipid Signaling. *Dev Cell* 47: 67-79 e65.
1064
- 1065 Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local
1066 alignment search tool. *J Mol Biol* 215: 403-410.
1067
- 1068 Angeles-Albores, D., N. L. RY, J. Chan and P. W. Sternberg, 2016 Tissue enrichment
1069 analysis for *C. elegans* genomics. *BMC Bioinformatics* 17: 366.
1070
- 1071 Arda, H. E., S. Taubert, L. T. MacNeil, C. C. Conine, B. Tsuda *et al.*, 2010 Functional
1072 modularity of nuclear hormone receptors in a *Caenorhabditis elegans* metabolic
1073 gene regulatory network. *Mol Syst Biol* 6: 367.
1074
- 1075 Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.*, 2000 Gene ontology:
1076 tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:
1077 25-29.
1078
- 1079 Ashrafi, K., 2007 Obesity and the regulation of fat metabolism. *Wormbook*: 1-20.
1080
- 1081 Baugh, L. R., J. Demodena and P. W. Sternberg, 2009 RNA Pol II accumulates at
1082 promoters of growth genes during developmental arrest. *Science* 324: 92-94.
1083
- 1084 Bulcha, J. T., G. E. Giese, M. Z. Ali, Y. U. Lee, M. D. Walker *et al.*, 2019 A Persistence
1085 Detector for Metabolic Network Rewiring in an Animal. *Cell Rep* 26: 460-468
1086 e464.
1087
- 1088 Consortium, C. e. S., 1998 Genome sequence of the nematode *C. elegans*: a platform
1089 for investigating biology. *Science* 282: 2012-2018.
1090
- 1091 Deng, X., J. B. Hiatt, D. K. Nguyen, S. Ercan, D. Sturgill *et al.*, 2011 Evidence for
1092 compensatory upregulation of expressed X-linked genes in mammals,
1093 *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat Genet* 43: 1179-
1094 1185.
1095
- 1096 Ding, W., D. P. Higgins, D. K. Yadav, A. A. Godbole, R. Pukkila-Worley *et al.*, 2018
1097 Stress-responsive and metabolic gene regulation are altered in low S-
1098 adenosylmethionine. *PLoS Genet* 14: e1007812.
1099
- 1100 Ding, W., L. J. Smulan, N. S. Hou, S. Taubert, J. L. Watts *et al.*, 2015 s-
1101 Adenosylmethionine Levels Govern Innate Immunity through Distinct
1102 Methylation-Dependent Pathways. *Cell Metab*.
1103

- 1104
1105 Eden, E., R. Navon, I. Steinfeld, D. Lipson and Z. Yakhini, 2009 GOrilla: a tool for
1106 discovery and visualization of enriched GO terms in ranked gene lists. BMC
1107 Bioinformatics 10: 48.
1108
1109 Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein, 1998 Cluster analysis and
1110 display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:
1111 14863-14868.
1112
1113 Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver *et al.*, 1998 Potent and
1114 specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*.
1115 Nature 391: 806-811.
1116
1117 Graham, P. L., J. J. Johnson, S. Wang, M. H. Sibley, M. C. Gupta *et al.*, 1997 Type IV
1118 collagen is detectable in most, but not all, basement membranes of
1119 *Caenorhabditis elegans* and assembles on tissues that do not express it. J Cell
1120 Biol 137: 1171-1183.
1121
1122 Greenstein, D., 2005 Control of oocyte meiotic maturation and fertilization. WormBook:
1123 1-12.
1124
1125 Hansen, M., Hsu, A.L., Dillin, A., Kenyon, C., 2005 New Genes tied to Endocrine,
1126 Metabolic and Dietary Regulation of Lifespan from a *Caenorhabditis elegans*
1127 RNAi Screen. PLoS Genetics 1: 119-128.
1128
1129 Hillier, L. W., A. Coulson, J. I. Murray, Z. Bao, J. E. Sulston *et al.*, 2005 Genomics in *C.*
1130 *elegans*: so many genes, such a little worm. Genome Res 15: 1651-1660.
1131
1132 Hubbard, E. J., and D. Greenstein, 2005 Introduction to the germ line. WormBook: 1-4.
1133
1134 Inglis, P. N., G. Ou, M. R. Leroux and J. M. Scholey, 2007 The sensory cilia of
1135 *Caenorhabditis elegans*. WormBook: 1-22.
1136
1137 Kaletsky, R., V. Yao, A. Williams, A. M. Runnels, A. Tadych *et al.*, 2018 Transcriptome
1138 analysis of adult *Caenorhabditis elegans* cells reveals tissue-specific gene and
1139 isoform expression. PLoS Genet 14: e1007559.
1140
1141 Kamath, R. S., A. G. Fraser, Y. Dong, G. Poulin, R. Durbin *et al.*, 2003 Systematic
1142 functional analysis of the *Caenorhabditis elegans* genome using RNAi.[see
1143 comment]. Nature 421: 231-237.
1144
1145 Kramer, A., J. Green, J. Pollard, Jr. and S. Tugendreich, 2014 Causal analysis
1146 approaches in Ingenuity Pathway Analysis. Bioinformatics 30: 523-530.
1147
1148 L'Hernault, S. W., 2006 Spermatogenesis. WormBook: 1-14.
1149

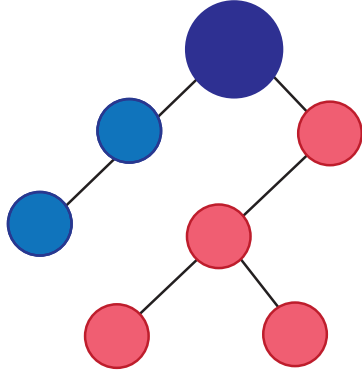
- 1150 LaMacchia, J. C., H. N. Frazier, 3rd and M. B. Roth, 2015 Glycogen Fuels Survival
1151 During Hyposmotic-Anoxic Stress in *Caenorhabditis elegans*. *Genetics* 201: 65-
1152 74.
1153
- 1154 Lee, M. H., and T. Schedl, 2006 RNA-binding proteins. *WormBook*: 1-13.
1155
- 1156 Lee, R. Y. N., K. L. Howe, T. W. Harris, V. Arnaboldi, S. Cain *et al.*, 2018 *WormBase*
1157 2017: molting into a new stage. *Nucleic Acids Res* 46: D869-D874.
1158
- 1159 Li, C., and K. Kim, 2008 Neuropeptides. *WormBook*: 1-36.
1160
- 1161 Macneil, L. T., E. Watson, H. E. Arda, L. J. Zhu and A. J. M. Walhout, 2013 Diet-
1162 induced developmental acceleration independent of TOR and insulin in *C.*
1163 *elegans*. *Cell* 153: 240-252.
1164
- 1165 Marchler-Bauer, A., Y. Bo, L. Han, J. He, C. J. Lanczycki *et al.*, 2017 CDD/SPARCLE:
1166 functional classification of proteins via subfamily domain architectures. *Nucleic*
1167 *Acids Res* 45: D200-D203.
1168
- 1169 Mato, J. M., and S. C. Lu, 2007 Role of S-adenosyl-L-methionine in liver health and
1170 injury. *Hepatology* 45: 1306-1312.
1171
- 1172 McDonald, J. H., 2014 *Handbook of Biological Statistics*. Sparky House Publishing,
1173 Baltimore, MD.
1174
- 1175 Mi, H., A. Muruganujan, X. Huang, D. Ebert, C. Mills *et al.*, 2019 Protocol Update for
1176 large-scale genome and gene function analysis with the PANTHER classification
1177 system (v.14.0). *Nat Protoc* 14: 703-721.
1178
- 1179 Murphy, C. T., S. A. McCarroll, C. I. Bargmann, A. Fraser, R. S. Kamath *et al.*, 2003
1180 Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis*
1181 *elegans*. *Nature* 424: 277-283.
1182
- 1183 Oliveira, R. P., J. Porter Abate, K. Dilks, J. Landis, J. Ashraf *et al.*, 2009 Condition-
1184 adapted stress and longevity gene regulation by *Caenorhabditis elegans* SKN-
1185 1/Nrf. *Aging Cell* 8: 524-541.
1186
- 1187 Ortiz, M. A., D. Noble, E. P. Sorokin and J. Kimble, 2014 A new dataset of
1188 spermatogenic vs. oogenic transcriptomes in the nematode *Caenorhabditis*
1189 *elegans*. *G3 (Bethesda)* 4: 1765-1772.
1190
- 1191 Reinke, V., H. E. Smith, J. Nance, J. Wang, C. Van Doren *et al.*, 2000 A global profile of
1192 germline gene expression in *C. elegans*. *Mol Cell* 6: 605-616.
1193

- 1194 Ritter, A. D., Y. Shen, J. Fuxman Bass, S. Jeyaraj, B. Deplancke *et al.*, 2013 Complex
1195 expression dynamics and robustness in *C. elegans* insulin networks. *Genome*
1196 *Res* 23: 954-965.
1197
- 1198 Robertson, H. M., and J. H. Thomas, 2006 The putative chemoreceptor families of *C.*
1199 *elegans*. *WormBook*: 1-12.
1200
- 1201 Rual, J. F., J. Ceron, J. Koreth, T. Hao, A. S. Nicot *et al.*, 2004 Toward improving
1202 *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi
1203 library. *Genome Res* 14: 2162-2168.
1204
- 1205 Schwarz, E. M., M. Kato and P. W. Sternberg, 2012 Functional transcriptomics of a
1206 migrating cell in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* 109: 16246-
1207 16251.
1208
- 1209 Smulan, L. J., W. Ding, E. Freinkman, S. Gujja, Y. J. Edwards *et al.*, 2016 Cholesterol-
1210 Independent SREBP-1 Maturation Is Linked to ARF1 Inactivation. *Cell Rep* 16: 9-
1211 18.
1212
- 1213 Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders *et al.*, 1998
1214 Comprehensive identification of cell cycle-regulated genes of the yeast
1215 *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9: 3273-
1216 3297.
1217
- 1218 Spencer, W. C., G. Zeller, J. D. Watson, S. R. Henz, K. L. Watkins *et al.*, 2011 A spatial
1219 and temporal map of *C. elegans* gene expression. *Genome Res* 21: 325-341.
1220
- 1221 Supek, F., M. Bosnjak, N. Skunca and T. Smuc, 2011 REVIGO summarizes and
1222 visualizes long lists of gene ontology terms. *PLoS One* 6: e21800.
1223
- 1224 The Gene Ontology, C., 2019 The Gene Ontology Resource: 20 years and still GOing
1225 strong. *Nucleic Acids Res* 47: D330-D338.
1226
- 1227 Vance, D. E., 2014 Phospholipid methylation in mammals: from biochemistry to
1228 physiological function. *Biochimica et biophysica acta* 1838: 1477-1487.
1229
- 1230 Varkey, J. P., P. L. Jansma, A. N. Minniti and S. Ward, 1993 The *Caenorhabditis*
1231 *elegans spe-6* gene is required for major sperm protein assembly and shows
1232 second site non-complementation with an unlinked deficiency. *Genetics* 133: 79-
1233 86.
1234
- 1235 Walker, A. K., R. L. Jacobs, J. L. Watts, V. Rottiers, K. Jiang *et al.*, 2011 A conserved
1236 SREBP-1/phosphatidylcholine feedback circuit regulates lipogenesis in
1237 metazoans. *Cell* 147: 840-852.
1238

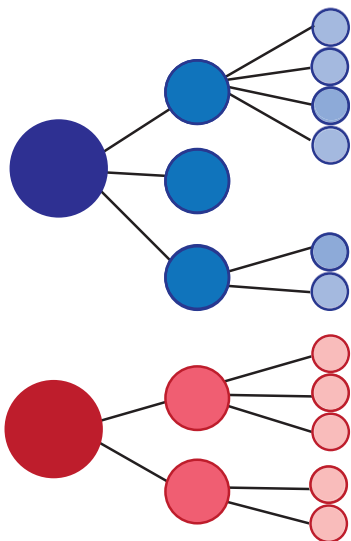
- 1239 Watts, J. L., and M. Ristow, 2017 Lipid and Carbohydrate Metabolism in *Caenorhabditis*
1240 *elegans*. *Genetics* 207: 413-446.
1241
1242 Xu, W., L. Yi, Y. Feng, L. Chen and J. Liu, 2009 Structural insight into the activation
1243 mechanism of human pancreatic prolympholipase A2. *J Biol Chem* 284: 16659-
1244 16666.
1245

A.

GO: Parent-Child



WormCat: Nested Tree



CATEGORY 1 Category 2 category 3

B. WormCat

Process your Regulated Gene Set with WormCat

Name: Email:

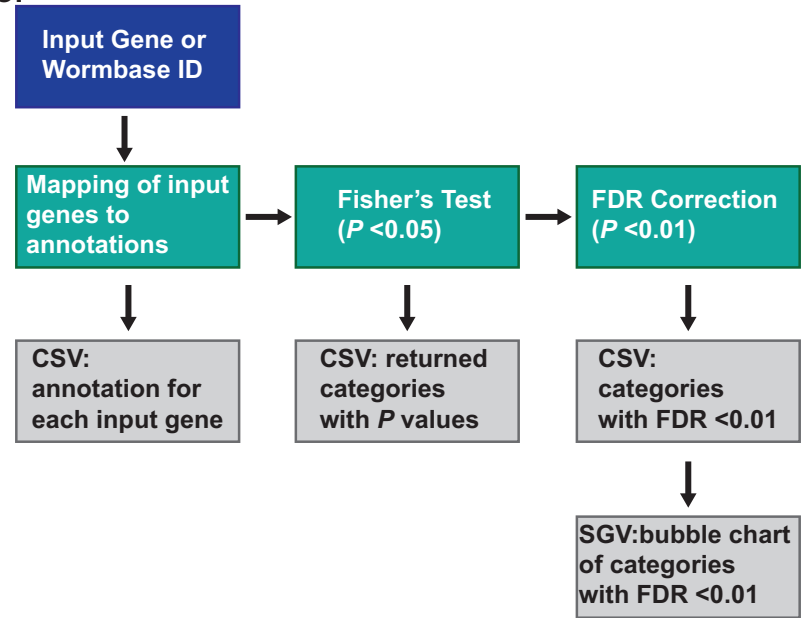
Annotation Type: Input Type:

Title:

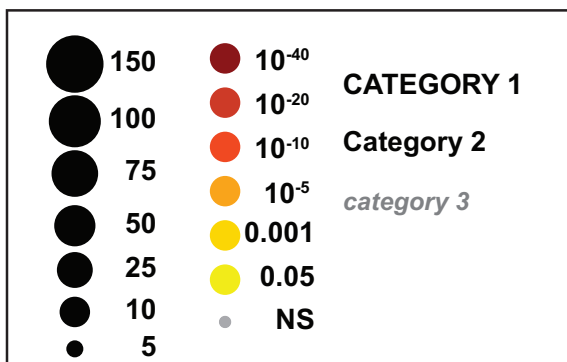
Regulated Gene Set:

Drop File Here

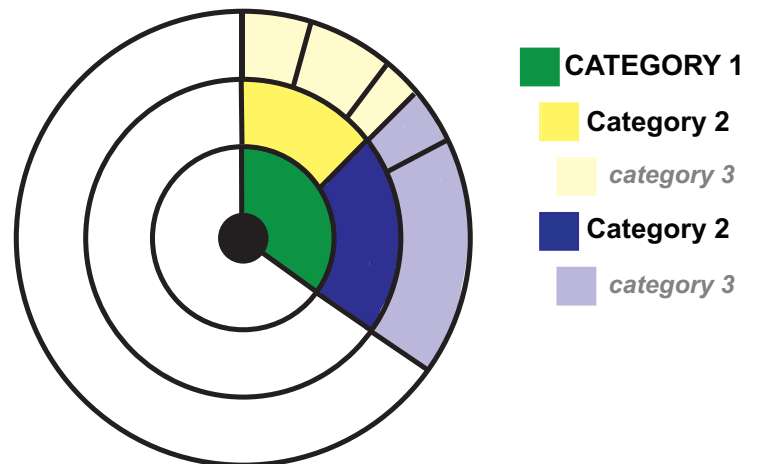
C.

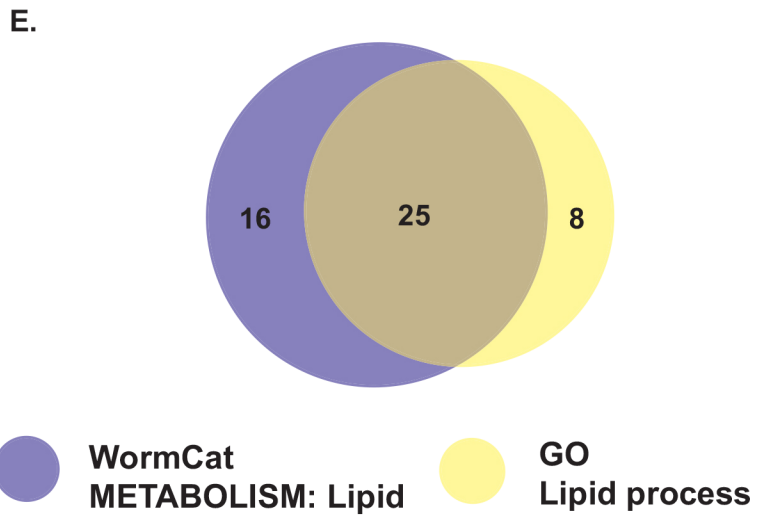
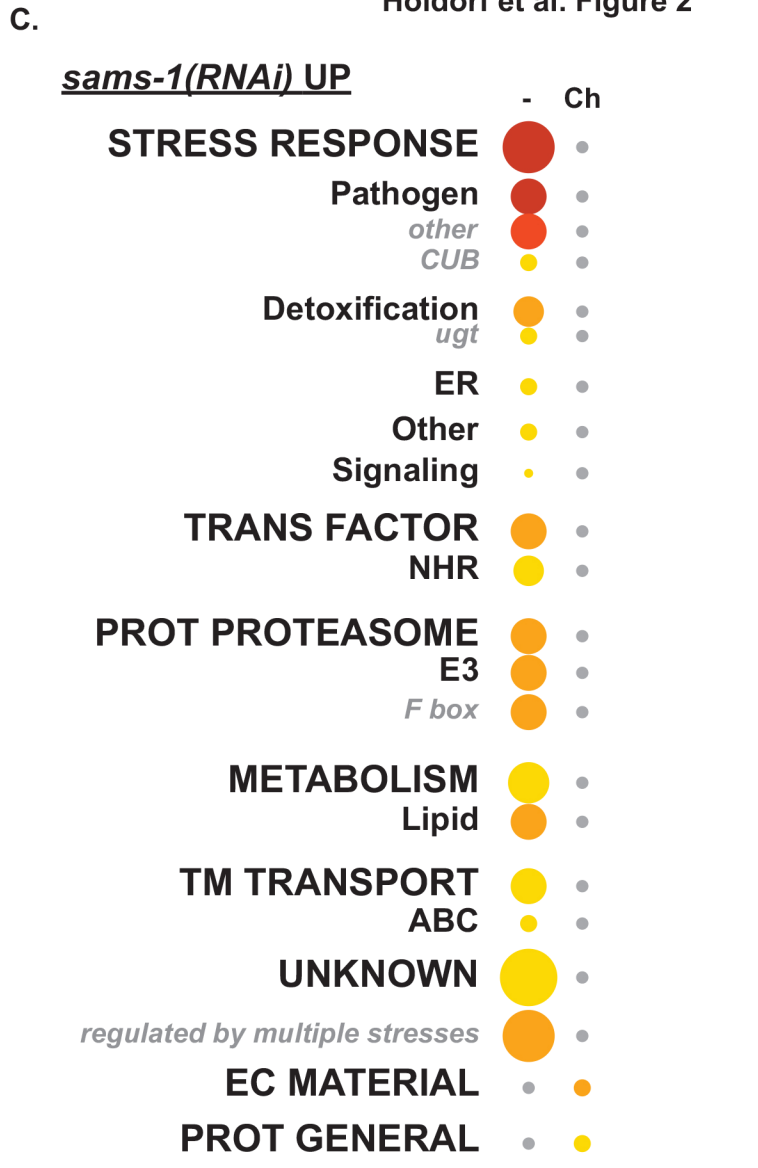
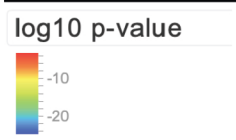
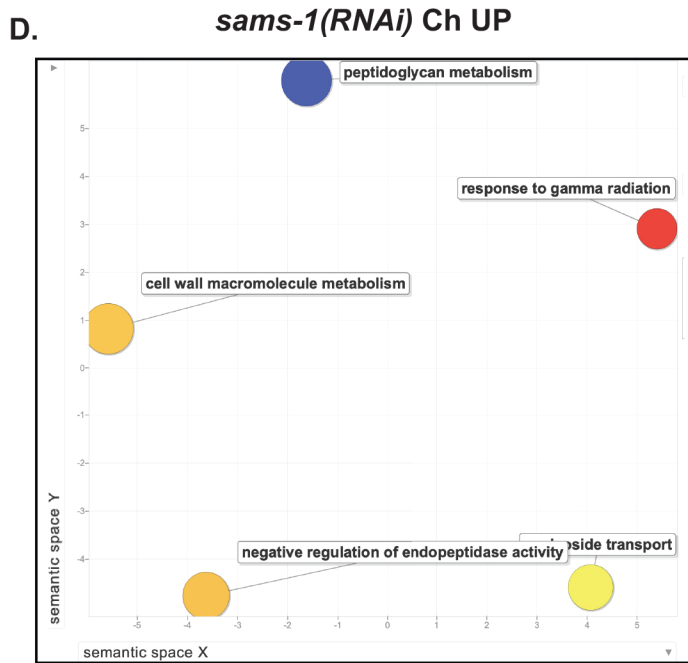
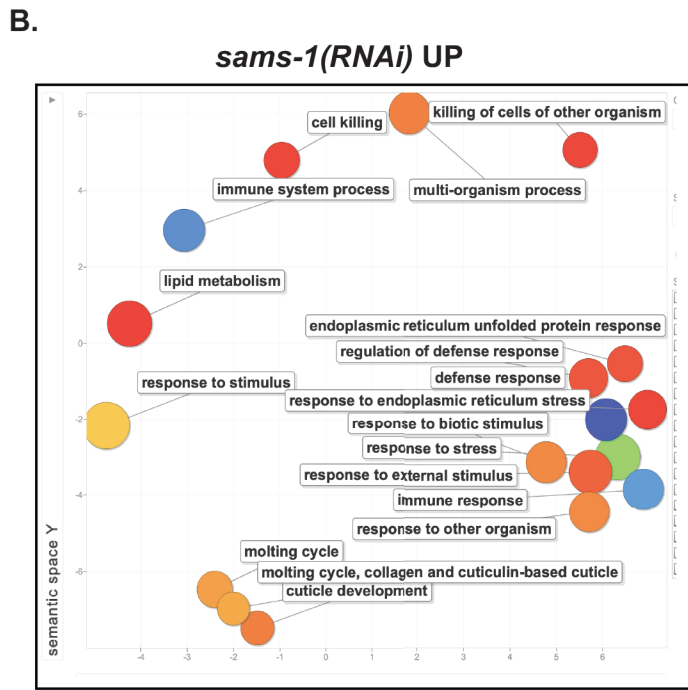
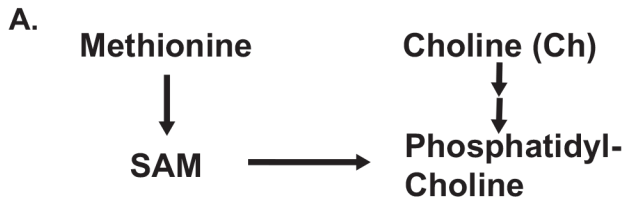


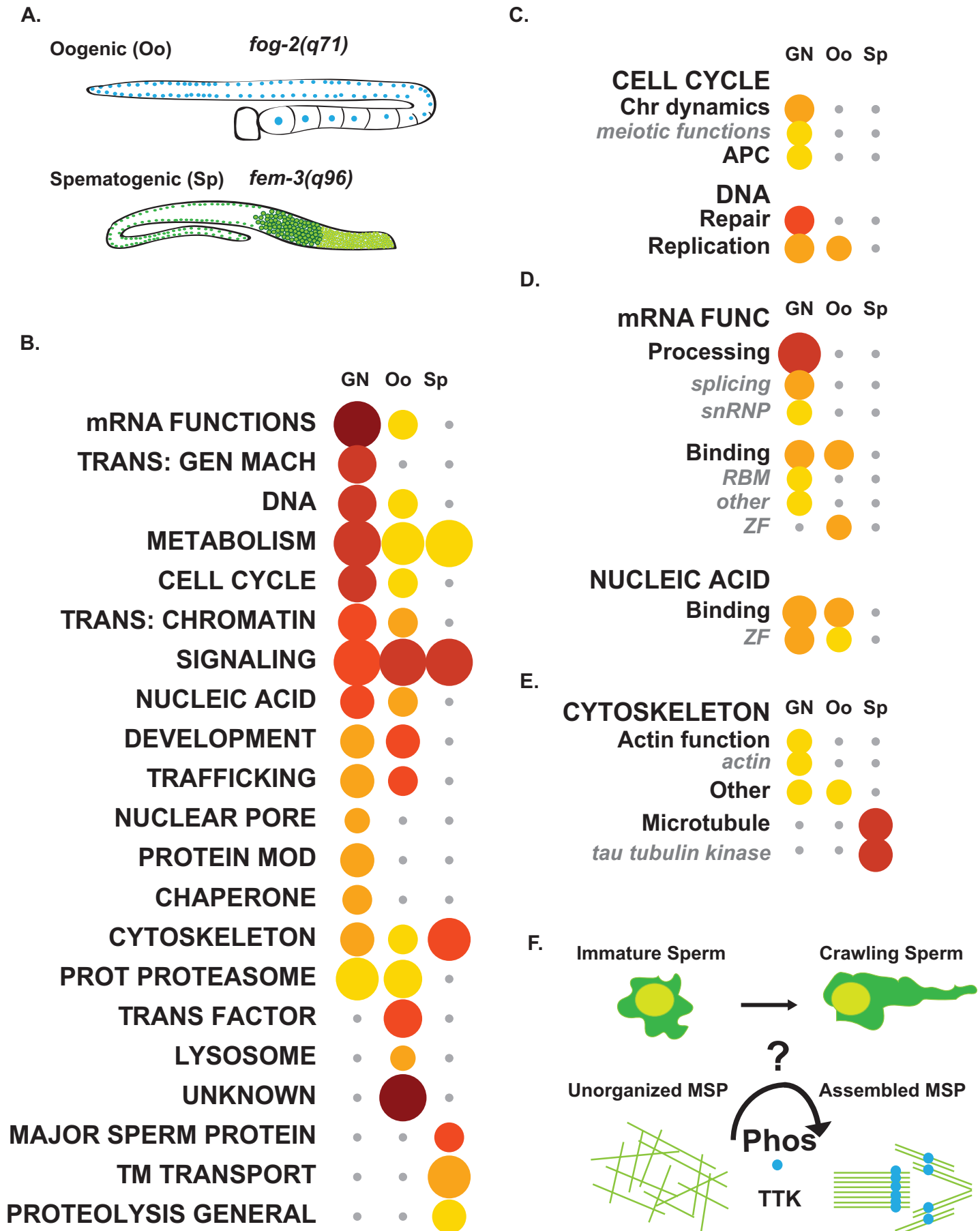
D.

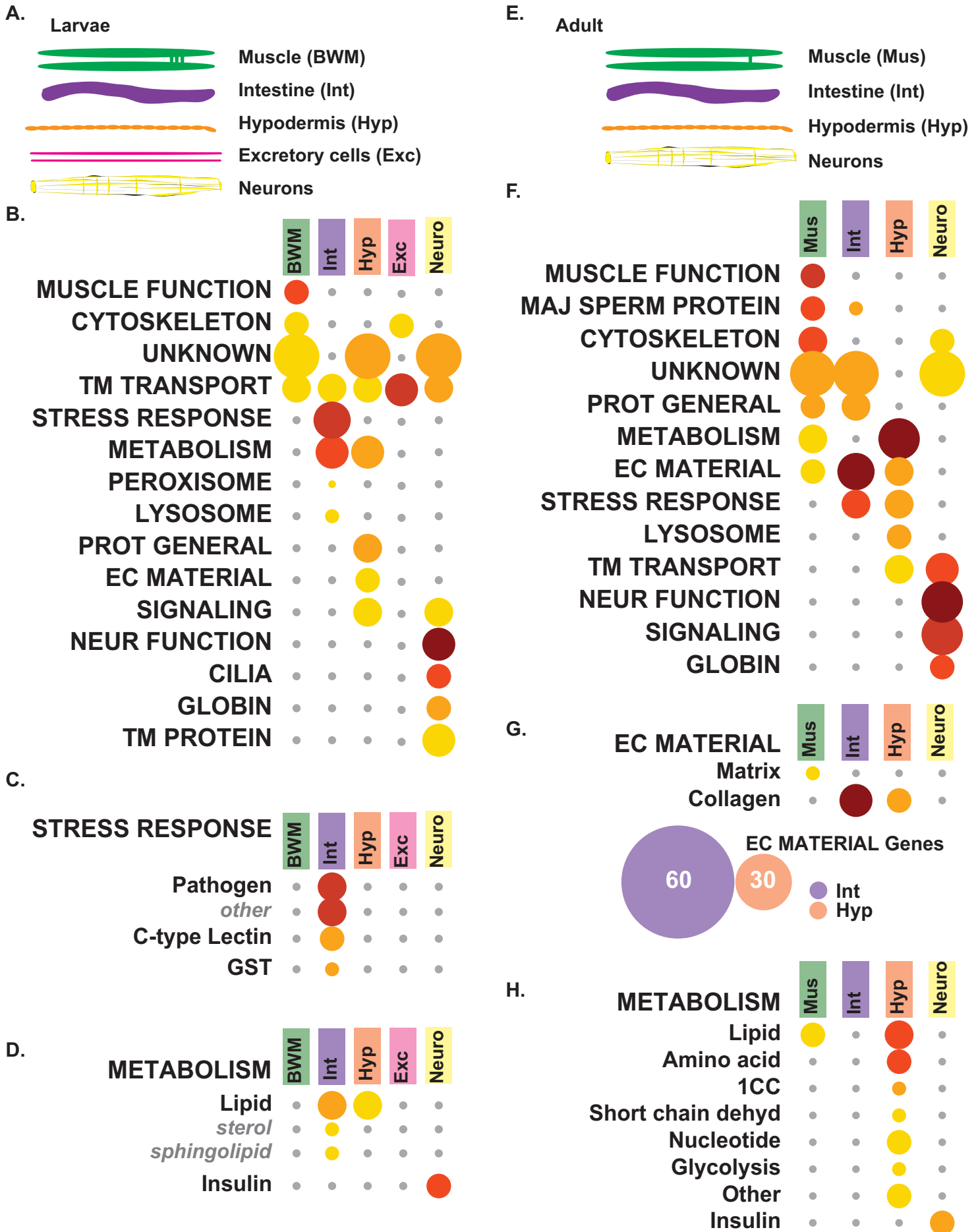


E.









A.

SINGLE

| | Psora | Rapa | Rifa | Allan |
|-------------------|-------|------|------|-------|
| METABOLISM | • | • | • | • |
| Lipid | • | • | • | • |
| <i>sterol</i> | • | • | • | • |

DOUBLE

| | Rapa | Rapa | Rifa | Rifa |
|-------------------|-------|------|-------|-------|
| | Psora | Rifa | Allan | Psora |
| METABOLISM | ● | ● | ● | ● |
| Lipid | ● | • | • | ● |
| <i>sterol</i> | • | • | • | • |

TRIPLE

| | Rifa | Rifa | Rifa |
|-------------------|-------|-------|-------|
| | Psora | Rapa | Rapa |
| | Allan | Allan | Psora |
| METABOLISM | ● | ● | ● |
| Lipid | ● | ● | ● |
| <i>sterol</i> | • | • | ● |

B.

