

# Application of Transcriptional Gene Modules to Analysis of *Caenorhabditis elegans*' Gene Expression Data

Michael Cary,<sup>\*,†,1</sup> Katie Podshivalova,<sup>‡,1</sup> and Cynthia Kenyon<sup>†,‡,2</sup>

<sup>\*</sup>Program in Biomedical Informatics, and <sup>†</sup>Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California 94158, and <sup>‡</sup>Calico Life Sciences LLC, South San Francisco, California 94080

**ABSTRACT** Identification of co-expressed sets of genes (gene modules) is used widely for grouping functionally related genes during transcriptomic data analysis. An organism-wide atlas of high-quality gene modules would provide a powerful tool for unbiased detection of biological signals from gene expression data. Here, using a method based on independent component analysis we call DEXICA, we have defined and optimized 209 modules that broadly represent transcriptional wiring of the key experimental organism *C. elegans*. These modules represent responses to changes in the environment (e.g., starvation, exposure to xenobiotics), genes regulated by transcription factors (e.g., ATFS-1, DAF-16), genes specific to tissues (e.g., neurons, muscle), genes that change during development, and other complex transcriptional responses to genetic, environmental and temporal perturbations. Interrogation of these modules reveals processes that are activated in long-lived mutants in cases where traditional analyses of differentially expressed genes fail to do so. Additionally, we show that modules can inform the strength of the association between a gene and an annotation (e.g., GO term). Analysis of “module-weighted annotations” improves on several aspects of traditional annotation-enrichment tests and can aid in functional interpretation of poorly annotated genes. We provide an online interactive resource with tutorials at <http://genemodules.org/>, in which users can find detailed information on each module, check genes for module-weighted annotations, and use both of these to analyze their own gene expression data (generated using any platform) or gene sets of interest.

## KEYWORDS

gene expression  
data analysis  
microarray  
functional  
annotation of  
genes  
gene  
co-expression  
independent  
component  
analysis  
aging  
mitochondrial  
unfolded  
protein  
response  
respiration  
*hif-1*  
*atfs-1*

Nearly half of the predicted protein-coding genes in *Caenorhabditis elegans* lack a functional annotation based on direct experimental evidence (Lee *et al.* 2018). As a result, querying gene annotation

databases, such as the Gene Ontology (GO) or curated pathways can fail to detect biologically meaningful signals in gene expression data (Ashburner *et al.* 2000; Fabregat *et al.* 2017; Kanehisa *et al.* 2012). An alternative approach to understanding gene function is to use information about gene expression. Gene-expression data can be used to define groups of genes that show similar patterns of expression, or co-variation, across many different conditions, which arise due to changes in genes' transcriptional activity and/or their post-transcriptional regulation. These groups are called transcriptional gene modules, with each module potentially representing a discrete biological phenomenon. Gene modules are routinely constructed when clustering algorithms are applied to gene-expression data and have been used successfully to identify gene regulatory mechanisms in a variety of contexts, from the yeast cell cycle (Wu *et al.* 2006) and sporulation

Copyright © 2020 Cary *et al.*

doi: <https://doi.org/10.1534/g3.120.401270>

Manuscript received June 29, 2020; accepted for publication July 27, 2020; published Early Online August 5, 2020.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supplemental material available at figshare: <https://doi.org/10.25387/g3.12061656>.

<sup>1</sup>These authors contributed equally to this work

<sup>2</sup>Corresponding author: Calico Life Sciences LLC, 1170 Veterans Blvd, South San Francisco, CA 94080, E-mail: [cynthia@calicolabs.com](mailto:cynthia@calicolabs.com)

(Wang *et al.* 2005) to human cancer cells (Teschendorff *et al.* 2007) to cognitive decline in patients with Alzheimer disease (Mostafavi *et al.* 2018). Furthermore, large compendia of data sampling diverse perturbations have been used to define fundamental gene-expression programs of entire organisms (Hughes *et al.* 2000, Engreitz *et al.* 2010, Zhou and Altman 2001).

In *C. elegans*, the most recent effort to generate high-quality fundamental transcriptional modules is now almost two decades old (Kim *et al.* 2001). Co-expressed genes were grouped together into 43 groups (or “mountains”) based on their correlation across 553 microarrays. However, the compendium did not contain all genes (e.g., over 30% of the microarrays only contained ~11,000 of *C. elegans*’ 20,470 protein-coding genes) and each gene was assigned exclusively to one group, although it is well-established that genes can participate in multiple processes (Oeckinghaus *et al.* 2011; Gerstein *et al.* 2012). The number of perturbations for which gene expression data are available has also increased substantially since 2001.

Here, we define 209 transcriptional gene modules in *C. elegans* using a heterogeneous compendium of 1386 microarrays and a method we call DEXICA, for Deep EXtraction Independent Component Analysis. DEXICA builds on prior implementations (Lee and Batzoglou 2003; Gong *et al.* 2007; Engreitz *et al.* 2010) of independent component analysis (ICA) for gene module extraction by maximizing the biological information content of the modules. It does so by varying data pre-processing methods and the number of extracted ICA components (modules) until the number of significantly enriched biological annotations in the module set is maximized. DEXICA also uses an artificial neural network to partition each independent component to determine which genes should be included in each module.

We show that the 209 DEXICA *C. elegans* modules capture gene expression patterns that correspond to biological processes; for example, responses to heat stress, xenobiotics and pathogenic bacteria, and to several individual tissues. Furthermore, data analysis in the module space correctly reveals biological processes that are missed by analyses of differentially expressed genes. We provide a user-friendly web interface, with tutorials, in which users can test which of the 209 gene modules are active in their datasets and find detailed information about each module that helps determine which biological process(es) they represent.

Finally, we explore whether gene modules can be used to improve existing gene annotations. We reason that an annotation shared among co-expressed genes is more likely to be relevant to their function than one that is not, and that annotations of co-expressed genes can provisionally be “transferred” onto their poorly annotated companions. We calculate what we call “module-weighted gene annotations” by weighting the association between a gene and an annotation by the degree to which the annotation appears predictive of the gene’s module membership. We show that matrix-based analysis of module-weighted annotations is more sensitive and specific than common annotation enrichment tests. We provide a framework for using module-weighted annotations to detect significant GO terms and promoter oligonucleotides directly from expression data, and to identify novel GO terms conferred onto genes based on their module membership.

The first part of this paper, which describes the construction and biological validation of the *C. elegans* DEXICA modules, is directed to computational biologists, and the second to *C. elegans* biologists who may or may not have experience in computational biology. Readers simply wishing to apply this analytic method to specific gene lists should feel free to skip directly to the later sections but may also enjoy

reading about the performance of the modules in our various validation test cases.

## MATERIALS AND METHODS

### Compendium construction

To build the compendium of 1386 *C. elegans* Affymetrix arrays, we first downloaded all CEL files with the appropriate platform ID (GPL200) from the GEO database available on March 1, 2014, excluding those for which the organism was not *C. elegans* and the sample type was not RNA. We excluded arrays from experiments for which fewer than 8 hybridizations were performed in order to mitigate the effect that under-sampled conditions might have on predicted modules. We then performed a quality control step using the quality assessment functions provided in the *simpleAffy* (v2.40.0) R package (<http://bioinformatics.picr.man.ac.uk/simpleaffy/>), discarding arrays that did not meet the quality thresholds recommended in the *simpleAffy* documentation.

We generated expression values for probesets separately for each experiment (determined by GEO series IDs) using the RMA pre-processing procedure provided in the *affy* (v1.40.0) R package (Gautier *et al.* 2004), then used the *bias* (v0.0.5) R package (Eklund and Szallasi 2008) to remove intensity-dependent biases in expression levels. We then concatenated the expression matrices for each experiment into a single matrix. Next, we either performed between-experiment quantile normalization on the entire matrix using the *limma* (v3.18.13) R package (Smyth 2005), or omitted this step, depending on the pre-processing method to be tested. Finally, we scaled and centered the arrays, such that the mean of each column was zero and the standard deviation was 1.

### Conducting ICA

To conduct ICA of the gene expression matrix, we used the *fastICA* (v1.2-0) R package (<http://CRAN.R-project.org/package=fastICA>) with default parameters except for the “method” parameter, which we set to “C” to increase computational speed, and the “row.norm” parameter, which we set to “TRUE” in order to balance the total compendium variance between genes with subtle changes in expression values and those with large changes in expression values.

### Partitioning of independent components

To convert independent components to discrete sets of genes, we employed two methods. In the first, for each component, we assigned all genes with a weight  $\geq 3$  to the positive (“a”) hemi-module, and all genes with a weight  $\leq -3$  to the negative (“b”) hemi-module. In the second, we created an artificial neural network using the *neuralnet* (v1.32) R package (<http://CRAN.R-project.org/package=neuralnet>) to predict positive and negative partitioning thresholds for each independent component, based on the component’s skewness and kurtosis (see Supplemental Methods), then assigned genes whose weights exceeded these thresholds to the corresponding hemi-modules.

### Obtaining gene annotations and microarray data

For module optimization, we obtained GO term and REACTOME pathway annotations for *C. elegans* genes using *biomaRt* (v2.18.0) R package (Durinck *et al.* 2009) and *ensembl* mart for data retrieval. To obtain genes’ tissue annotations, we downloaded all available data from the GFP Worm database (<http://gfweb.aecom.yu.edu/>) (Hunt-Newbury *et al.* 2007), which contains annotated expression patterns of promoter::GFP fusion constructs; in total, this dataset provided annotations for

1821 genes across 89 tissue types (*n.b.*, we considered the same tissue in different development stages to be distinct tissue types). For Module Annotation Pages (Tool 3 on <http://genemodules.org/>), known tissues (anatomies) of gene expression were downloaded from the “expression patterns” database from Wormbase (WS275). This dataset included 5376 genes and 2438 anatomical terms. GO term enrichment was calculated using *topGO*(v.2.34.0) and *cellegans.db* (v 3.2.3) R packages. To obtain fold changes for *isp-1* mutants, we used data previously published by our group in which *isp-1(qm150)* mutants were compared to wild type controls (Cristina *et al.* 2009). To obtain fold changes for *hif-1* mutants, we used the *maanova* (v1.33.2) R package (<http://research.jax.org/faculty/churchill>) and data previously published by Shen, *et al.* (Shen *et al.* 2005), to calculate the induced gene fold changes upon mutation of *hif-1*. All other microarray data were obtained directly from the authors of the original publications or from the GEO database (see Supplemental Methods).

### Optimizing gene module prediction

To optimize gene module prediction, we performed ICA with a variety of different data preprocessing options [*e.g.*, the choice of preprocessing algorithm (RMA, GCRMA, PLIER, MAS 5.0), background, perfect match, bias correction, and normalization methods], and with a varied number of extracted components from 5 to 500 by increments of 5. For each parameter combination, we repeated ICA 5 times.

We tested the biological validity of the independent components generated by each ICA run by determining the number of annotations that were enriched in at least one hemi-module. We chose not to optimize based on the number of modules with at least one significant annotation, as tests using simulated data showed that this approach could lead to signals being split into multiple, less accurate representations (data not shown). We first calculated a p-value for the enrichment of genes associated with each annotation term in each hemi-module using the hypergeometric test. We then applied the Simes method (Simes 1986) for multiple hypothesis testing ( $\alpha = 0.05$ ) to the set of p-values for each annotation term. The Simes method is similar to the Benjamini-Hochberg method (Benjamini and Hochberg 1995) for controlling the false discovery rate, but differs in a way that makes it more appropriate here: it aims to answer the question, “Given a set of p-values, what is the likelihood that at least one null hypothesis is false?”, while the Benjamini-Hochberg method asks, “What fraction of rejected null hypotheses are actually true (*i.e.*, falsely discovered)?” Failure of the Simes test indicates that at least one null hypothesis is false at the specified alpha level. To verify the accuracy of our module quality statistics, we repeated all tests using module definition matrices in which gene IDs had been randomly shuffled.

### Quantification of module activity in gene expression data

To project a data vector,  $\mathbf{x}$ , such as a set of gene-expression fold changes, onto a set of gene modules, we used the scalar projection method, in which a mixing vector,  $\mathbf{a}$ , is calculated from the dot product of the data vector and the unit vectors comprising the module definitions,  $\hat{S}$ , as shown in equation 2:

$$\mathbf{a} = \mathbf{x} \cdot \hat{S} \quad (2)$$

The resulting mixing vector,  $\mathbf{a}$ , provides an indication of the weight of each module definition vector in the projected data,  $\mathbf{x}$ . Note that

because  $\mathbf{x}$  is a vector of gene fold changes, this analysis can be applied to data generated using any RNA profiling platform (*e.g.* RNAseq, microarray). Projection of a data matrix,  $\mathbf{X}$ , which generates a mixing matrix,  $\mathbf{A}$ , was carried out using the same procedure.

To calculate signed variance explained (SVE), we calculated the relative variance explained (VE) for each module from  $\mathbf{a}$  as follows, where  $n$  is the total number of modules:

$$VE_i = \frac{\mathbf{a}_i^2}{\sum_{i=1}^n \mathbf{a}_i^2} \quad (3)$$

We then multiplied these values, which are strictly positive, by -1 in each case where  $\mathbf{a}_i < 0$  to obtain SVE.

### Generation of the Enrichment matrix, E

First, ANN-based partitioning of the module definition matrix,  $\mathbf{S}$ , was used to generate matrix  $\mathbf{S}_p$ .  $\mathbf{S}_p$  contains two gene sets per independent component (which we refer to as hemi-modules), for a total of 418. Using a matrix of known Boolean associations between genes and annotations,  $\mathbf{B}$ , we calculated a hypergeometric probability for each annotation in each hemi-module. We used the frequency of genes bearing a particular annotation in the hemi-module, the frequency of such genes in the compendium, the number of genes in the hemi-module, and the number of genes not in the hemi-module as the  $q$ ,  $m$ ,  $k$ , and  $n$  input parameters, respectively, to the *phyper()* function of the *stats* (v3.0.3) R package (<http://www.R-project.org/>).

We used these p-values to populate a matrix,  $\mathbf{E}$ , with a row for each hemi-module and a column for each annotation. For under-represented annotations, we entered the  $\log(p\text{-value})$  in the matrix, and for over-represented annotations we entered the  $-\log(p\text{-value})$ .

### Generation of module-weighted annotations

Our calculation of module-weighted annotations takes advantage of the fact that, in modules generated by ICA, prior to partitioning, each gene has a weight in each module. Given a score, or weight, for each annotation in each module, this allows genes to be associated with annotations via a matrix product calculation. In the  $\mathbf{E}$  matrix (see above) highly positive values correspond to strongly enriched annotations in a hemi-module, highly negative values correspond to strongly depleted annotations. We transform the gene module matrix,  $\mathbf{S}$ , into an unpartitioned hemi-module matrix,  $\mathbf{H}$ , by concatenating it with a negative copy of itself column-wise:

$$\mathbf{H} = [\mathbf{S} \ -\mathbf{S}] \quad (4)$$

The product of this matrix with the  $\mathbf{E}$  matrix produces matrix  $\mathbf{R}$ , which relates genes to annotations:

$$\mathbf{H} \cdot \mathbf{E} = \mathbf{R} \quad (5)$$

As a final step, we normalize the values in this matrix row-wise; *i.e.*, separately for each annotation, by subtracting the mean and dividing by the standard deviation.

### Analysis of GO terms based on gene weights

To test whether the ranking of GO terms based on gene weights (Table 1) was statistically significant, we constructed a list comprising all semantic words used in all GO terms, excluding words shorter than three letters and uninformative words (*e.g.*, “the” and “for”.) We then tested each word for bias toward appearing near the top or

bottom of the ranked GO term list. In agreement with our initial observations, the most significantly top-biased words pertained to macromolecular complexes, such as “nucleosome”, “cilium”, and “ribosomal”, and the most significant bottom-biased words pertained to cell signaling, such as “signal”, “kinase”, and “receptor”. Many of the GO terms containing cell signaling words were generic in nature, e.g., protein kinase regulator activity, thus, our results may partially be explained by a lack of co-regulation among constituents of different signaling pathways. However, some specific cell signaling terms, e.g., *Notch signaling pathway* also appeared near the bottom of the ranked GO term list, suggesting that the genes annotated with such terms are either not strongly co-regulated at the gene expression level or that the biological conditions represented by the compendium did not perturb their expression enough to form modules with our method.

### Analysis of expression data with module-weighted annotations

To test whether a set of gene fold-changes was significantly enriched for specific annotations, given a module-weighted annotation matrix,  $R$ , we calculated the dot product of the data vector,  $x$ , comprising the set of gene fold-changes, and the  $R$  matrix:

$$a = x \cdot R \quad (6)$$

The resulting vector,  $a$ , provides an indication of the degree to which genes with strong weights for each annotation also have strong fold-changes. To generate p-values from these, we permuted the fold-change vector,  $x$ , 1000 times to create a background distribution for each annotation, which we then used to determine z-scores.

### Data Availability

All data supporting the findings of this study are available within the paper and its supplemental materials. Dynamic access to the data, with tutorials, is also provided through a Shiny App graphical user interface at <http://genemodules.org/>. Specifically, 6 separate functionalities are provided:

- Tool 1:** Determine which of the 209 modules defined here are active in the user’s gene expression data (by inputting a list of gene expression fold changes)
- Tool 2:** Using the partitioned 209 modules as gene sets, test whether a list of user-specified genes is enriched in any particular module
- Tool 3:** Detailed description of each of the 209 *C. elegans* gene modules
- Tool 4:** Visualize how genes assigned to a module change under a variety of conditions (conditions can be chosen from the 716 perturbations derived from our microarray compendium or provided by the user)
- Tool 5:** Get module-weighted GO terms associated with a user-specified gene
- Tool 6:** Analyze gene expression data using module-weighted GO terms

Supplemental materials include Supplemental Discussion and Supplemental Methods, Supplemental References, eight Supplemental Figures and four Supplemental Files. File S1, “Partitioned  $S$  matrix”, contains probe sets and genes that belong to each of the 209 DEXICA modules. File S2, “Module annotation summaries”, provides a summary of information about each module - top enriched GO terms and top activating perturbation. File S3, “Finding experiments that activate similar modules” contains a comparison of module activity in all possible pairs of perturbations (contrasts) in

the microarray compendium to facilitate identification of perturbations that activate similar modules. File S4, “Module-weighted GO terms”, contains module-weighted GO annotations of each gene. Supplemental material available at figshare: <https://doi.org/10.25387/g3.12061656>.

Code described in this study is available as an R package: DEXICA R package <https://github.com/MPCary/DEXICA>. Data to reproduce extraction of gene modules can be downloaded from <https://github.com/MPCary/DEXDATA.Celegans>.

To install:

```
install.packages("devtools")
library(devtools)
install_github("MPCary/DEXICA", build_vignettes = TRUE)
install_github("MPCary/DEXDATA.Celegans").
```

## RESULTS

### Development of DEXICA and extraction of optimized gene modules

A large body of gene-expression data is publicly available (Barrett *et al.* 2011; Rustici *et al.* 2013) and has enabled computational prediction of gene modules (Kim *et al.* 2001; Segal *et al.* 2003a; Ihmels *et al.* 2004; Michoel *et al.* 2009; Engreitz *et al.* 2010). We refer to our method for going from a raw compendium of gene expression data to an optimized set of gene modules and a list of genes that belong to each module as DEXICA, for Deep EXtraction Independent Component Analysis (described below).

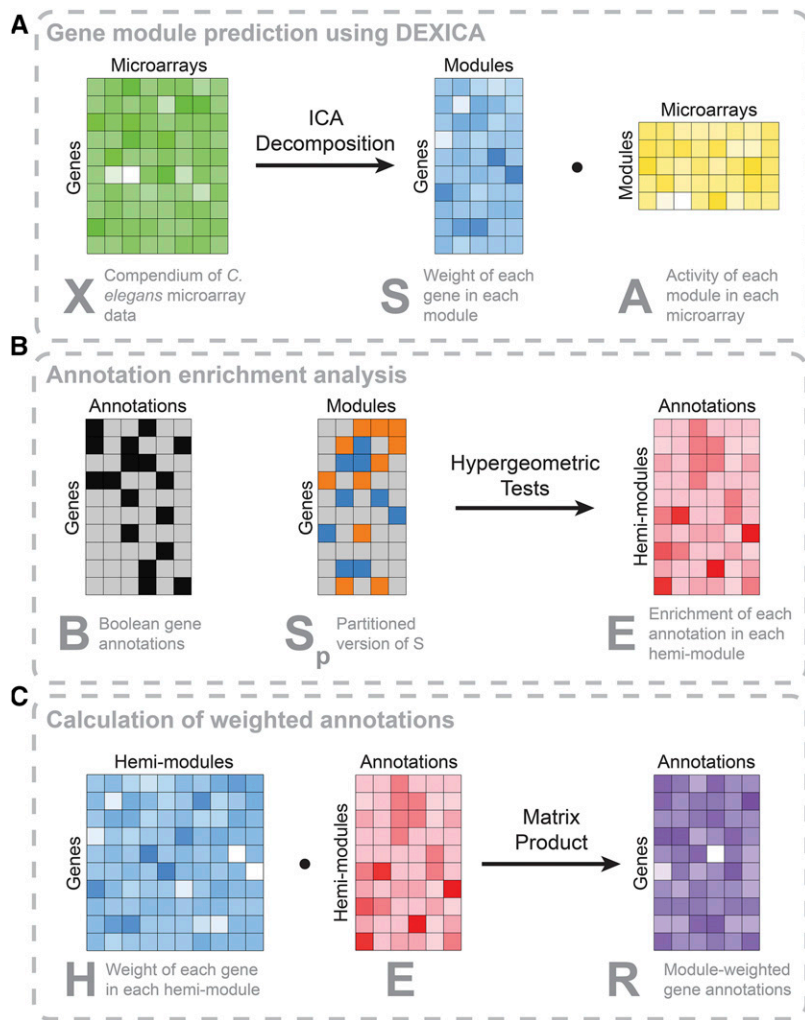
While several methods exist for defining gene modules, independent component analysis (ICA) generally outperforms clustering-based approaches and principle component analysis in extracting biologically relevant signals from large datasets<sup>15,22-24</sup>. Advantages of ICA include its ability to deal well with high dimensional data and to generate modules that can share genes. Furthermore, ICA does not assume that latent signals in the data follow a Gaussian distribution, an important property for gene module prediction, as gene regulation signals appear to be primarily super-Gaussian (Purdom and Holmes 2005). For these reasons, we chose to use ICA for module construction.

Briefly, ICA is a blind source separation method that attempts to “unmix” a signal comprising additive subcomponents by separating it into statistically-independent source signals (Hyvärinen and Oja 2000). In the common notation, an  $n \times m$  data matrix,  $X$ , is decomposed into two new matrices – an  $n \times k$  source matrix,  $S$ , and a  $k \times m$  mixing matrix,  $A$ , where  $k$  is the number of independent components:

$$X = S \cdot A \quad (1)$$

In the context of gene-expression analysis,  $X$  is a matrix of  $m$  measurements (e.g., microarrays) of  $n$  genes, and  $k$  independent components are interpreted as gene modules.  $A$  indicates the weight of each module in each microarray and  $S$  indicates the relative level of inclusion of each gene in each module (Kong *et al.* 2008) (Figure 1a).

Using simulated data, we found that module-prediction accuracy was highest when the number of extracted components matched the true number of modules (Figure S1). Therefore, we sought to optimize module prediction by evaluating results based upon their biological information content, such as enrichment of Gene Ontology (GO) terms (Ashburner *et al.* 2000), REACTOME pathways (Fabregat *et al.* 2017), and tissue-specific expression (Hunt-Newbury *et al.* 2007). We applied ICA to a diverse compendium of 1386 *C. elegans*

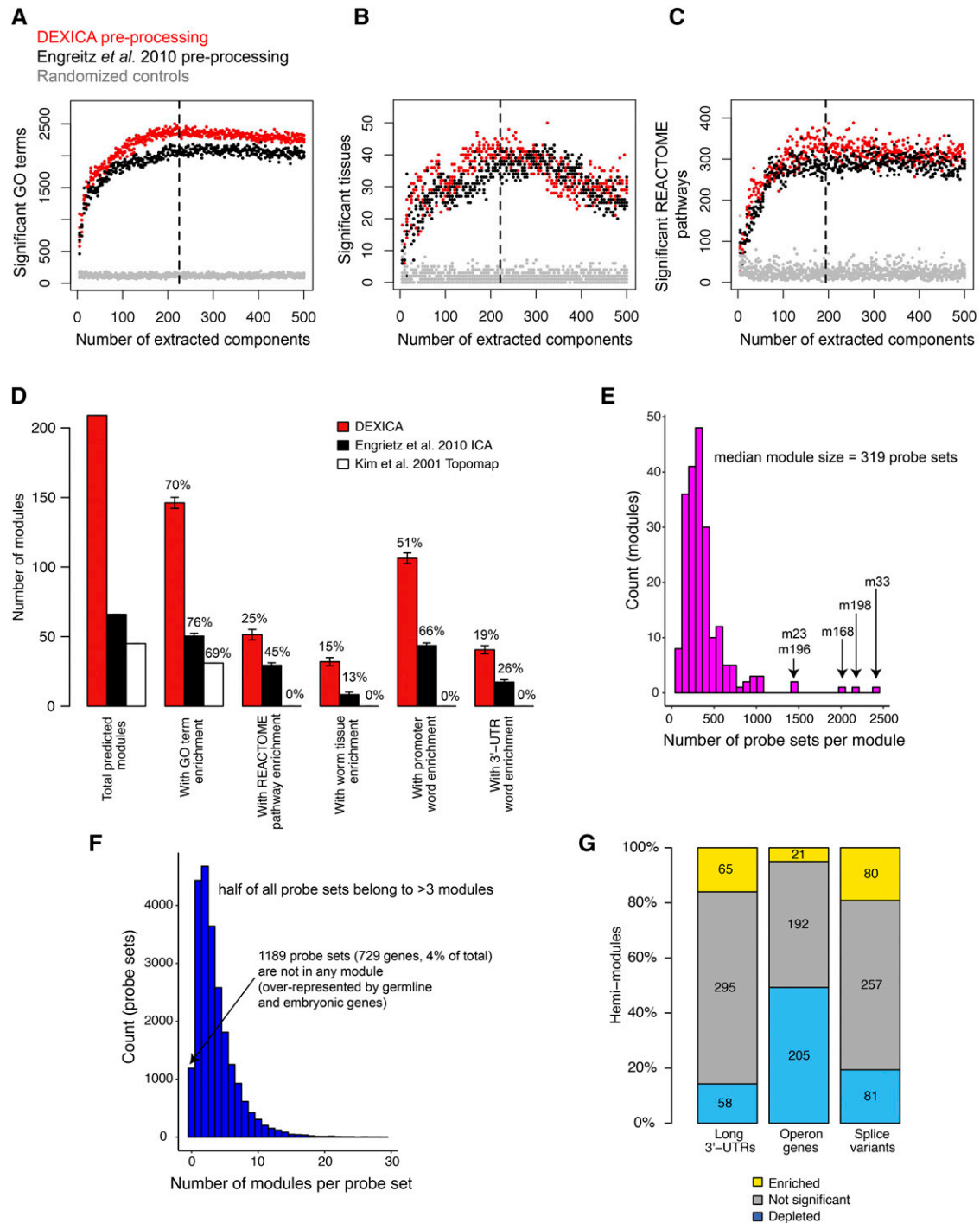


**Figure 1** Schematic of module prediction using DEXICA and derivation of module-weighted annotations. (a) A matrix of gene expression data,  $X$ , is decomposed using independent component analysis (ICA) into a gene module definition matrix,  $S$ , and a matrix containing the weight of each module in each microarray,  $A$ . Rows of the  $S$  matrix indicate the relative degree of inclusion of a given gene in each module. (b) Annotation enrichment within modules is calculated using matrices  $B$  and  $S_p$ . Known associations between genes and annotations are captured in a Boolean matrix  $B$ , where 1 indicates an association between a gene and an annotation and 0 indicates a lack of an association.  $S$  is partitioned into  $S_p$ , where 0 indicates a gene's exclusion from a module, while -1 indicates a gene's assignment to the "a" hemi-module and 1 to the "b" hemi-module (a hemi-module comprises genes that have extreme weights and the same sign in a given column of the  $S$  matrix). Enrichment of genes associated with each annotation in each hemi-module is calculated using hypergeometric tests and  $\log(p\text{-values})$  (for under-represented annotations) or  $-\log(p\text{-values})$  (for over-represented annotations) are recorded in a new matrix,  $E$ . (c) To generate module-weighted annotations, the  $S$  matrix is first transformed into a matrix,  $H$ , which indicates the weight of each gene in each hemi-module.  $H$  contains twice as many columns as  $S$  because there are two hemi-modules per module. The dot product between  $H$  and  $E$  results in the  $R$  matrix, where a row indicates the weighted association between a gene and each annotation. The color intensity in  $R$  indicates annotation weights that would result given the indicated expression values in  $X$  and the Boolean annotations in  $B$ . Color saturation in all matrices except  $B$  (Boolean) and  $S_p$  (trivalued) indicate relative numeric values, with least saturated colors (i.e., white) indicating highly negative values, and most saturated colors indicating highly positive values.

Affymetrix arrays obtained from the Gene Expression Omnibus (GEO) database (Barrett *et al.* 2011). We then compared results obtained from a wide variety of preprocessing methodologies and the total number of components extracted. Omitting between-experiment quantile normalization from the preprocessing procedure produced modules that were more annotation rich than those produced by a published implementation of ICA-based module extraction (Engreitz *et al.* 2010) (Figure 2a-c). Furthermore, annotation content of the modules was maximized when the number of gene modules (i.e., independent components) ranged from 191 to 226.

To enable functional evaluation of the modules, it is useful to summarize them as discrete gene sets. To this end, we partitioned each column of the  $S$  matrix into three sets of genes: one set consisting of genes excluded from the module, and two other sets consisting of genes implied to be regulated in opposite directions. We refer to these latter two sets as "hemi-modules" "a" and "b", one set consisting of genes with highly positive weights and the other consisting of genes with highly negative weights (signs assigned based on skewness) in the independent component. Thus, module genes regulated in one direction (up or down) are part of one hemi-module and genes regulated in the opposite direction (down or up) are part of the other hemi-module. The "a" hemi-module is derived from the more highly skewed side of the independent component, and is usually larger (i.e. contains more genes). While others have used a fixed-threshold

approach to component partitioning (Lee and Batzoglou 2003; Chiappetta *et al.* 2004; Engreitz *et al.* 2010); for example, defining genes with weights exceeding  $\pm 3$  standard deviations from the component mean to be "in" each hemi-module, we found that different individual modules showed maximum annotation enrichments at different thresholds, suggesting that a 'one-size-fits-all' approach to partitioning was sub-optimal. An alternative approach to partitioning that we attempted (described in Frigyesi *et al.* 2006) failed to converge in many cases (data not shown). Therefore, to increase partitioning accuracy, we trained a function to predict partitioning thresholds from the shape of component distributions. Because plots of training data revealed a complex solution surface, we decided to use an artificial neural network (ANN) to predict partitioning thresholds for each component from the skewness and kurtosis of its distribution. The output of the ANN, i.e., predicted partitioning thresholds, is shown in Figure S2a. Although using a fixed-threshold approach to module partitioning produced similar results qualitatively (Figures S2b-d), it resulted in fewer significant annotations across the range of parameters tested than did ANN-based partitioning ( $P < 2.2E-16$ , Figure S2e). Because the mean optimum number of extracted components (dashed vertical lines in Figure 2a-c and Figures S2b-d) was similar for both threshold and ANN partitioning (209, and 209.33, respectively), we chose 209 as the final number of components to extract from the compendium.



**Figure 2** Optimization of gene modules. To determine the optimal preprocessing method and the optimal number of components (gene modules) to extract from a gene expression compendium of *C. elegans* microarray data, we calculated the number of Gene Ontology terms (a), *C. elegans* tissues (b), and REACTOME pathways (c) that were significantly enriched in at least one gene module. Black points show results from a compendium produced using a preprocessing procedure used by Engreitz *et al.* (Engreitz *et al.* 2010); red points show results for the best alternative preprocessing method that we tested. Black dashed lines indicate the point on the x-axis of each graph at which loss regression curves showed the greatest difference between red points and results from randomized controls (gray points). (d) The number of modules produced by DEXICA, by a different ICA-based method used by Engreitz *et al.* (Engreitz *et al.* 2010), and by *C. elegans* gene expression topomap generated by Kim *et al.* (Kim *et al.* 2001). The number and the percentage of total predicted modules that have significant enrichment for various annotations are shown. Error bars indicate s.d. between repeat runs of DEXICA or Engreitz *et al.* method. (e) Distribution of the number of probe sets partitioned to each of the 209 DEXICA modules. (f) Distribution of the number of different DEXICA modules to which a given probe set is partitioned. (g) Each of the 418 DEXICA hemi-modules was tested for enrichment of genes with different structural properties – presence of a long 3'-UTR, transcription as part of an operon, and presence of multiple annotated splice variants.

Gene modules are expected to represent sets of genes that are co-regulated at the level of mRNA expression or stability. Therefore, DEXICA modules should be enriched for DNA and RNA regulatory sequences. To test this, we generated a list of potential regulatory oligonucleotide sequences (called ‘words’) by applying the Mobydick algorithm (Bussemaker *et al.* 2000) to the set of all predicted *C. elegans* 3′-UTRs. We then calculated the statistical significance of the over- or under-representation of genes bearing each word in each gene module. Across multiple runs with 209 components, the mean number of gene modules containing significant promoter words and 3′-UTR words was 106.3 and 40.6, respectively, significantly greater than results produced by other module prediction methods we tested ( $P < 2.2E-16$ , Figure 2d).

Because the ICA algorithm employed by DEXICA, *fastICA*, converges to a final solution from a random starting point (Hyvärinen and Oja 2000), small differences typically exist in the output of different runs; these differences can be seen in the vertical spread of data points in Figures 2a-c, and in the error bars in Figure 2d. While others have reconciled such differences through a clustering approach applied to the output of numerous runs of the algorithm (so called “iterated ICA”) (Frigyesi *et al.* 2006; Engreitz *et al.* 2010), when applied to the *C. elegans* Affymetrix compendium, many of the final components generated by this method were highly correlated to one another, indicating non-independence and potential redundancy among the components (data not shown). We therefore sought to choose a single, high quality, *fastICA* run output to use as predicted gene modules. Because we considered word enrichment the most unbiased measure of module quality (as it relies only on DNA sequence data), we chose as our final module set (File S1) the run from a set of 100 with the best combined rank of significant promoter and 3′-UTR words (it ranked first in promoter words and third in 3′-UTR words, Figure S3). The final median module size is 319 probe sets (Figure 2e) with half of all probe sets belonging to at least 3 modules (Figure 2f). Only 4% of all genes (729 genes, 1189 probe sets) do not belong to any module. These genes show significant enrichment for germline or embryonic expression.

Co-expressed genes often share common structural elements. For example, genes within operons are switched on together during recovery from growth-arrested states in *C. elegans* (Zaslaver *et al.* 2011) and 3′-UTR length is associated with proliferation in cancer cells (Mayr and Bartel 2009). To further explore the information content of DEXICA-extracted modules, we tested each hemi-module for over- and under-enrichment of genes with long 3′-UTRs, for genes appearing in operons and for genes with multiple splice forms. Of the 418 hemi-modules, 65 contained a significant bias toward long 3′-UTR genes and 58 contained a bias toward short 3′-UTR genes ( $q < 0.1$ , threshold chosen based on randomized control trials, see below; Figure 2g). Twenty-one hemi-modules were significantly enriched and 205 hemi-modules were significantly depleted for operon genes, and 81 hemi-modules were enriched and 80 hemi-modules were depleted for genes with multiple splice variants (Figure 2g). Control tests performed on the same module set but with randomly scrambled gene IDs produced no significant modules below  $q = 0.1$  for any of the gene properties we tested (data not shown). Therefore, DEXICA successfully groups genes with common structural features into modules, consistent with the known relationship between gene structure and expression.

Together, these results show that the 209 *C. elegans* gene modules extracted from a large microarray compendium are enriched for GO terms, tissue and pathway annotations as well as for potential

regulatory DNA sequences and gene structural properties. DEXICA is available as an R package (<https://github.com/MPCary/DEXICA>). It provides tools to optimize ICA module extraction and partitioning based on annotation enrichment and can be applied to any gene expression compendium.

### C. elegans DEXICA modules are biologically informative

We observed enrichment of functional gene annotations and oligonucleotide sequences in DEXICA-extracted modules with even the least optimal parameter settings (see Figure 2). To further test the biological significance of the final 209 modules and to begin annotating them, we constructed an alternate microarray compendium comprising not 1386 individual arrays, as in the original compendium, but rather the gene fold changes arising from contrasting experimental and control samples in the same experiment. Projecting the resulting 716-column matrix (one for each contrast) into the space defined by our gene modules allowed us to see which experimental perturbations activate or inhibit which modules.

We compared the most enriched GO terms in each module to the most strongly activating or inhibiting experimental perturbations and observed that in several cases these were in obvious agreement (File S2 and Tool 3 on <http://genemodules.org/>). For example, the strongest activation of m73 (module 73) is achieved by an embryonic development time course experiment. Consistent with this, the top GO terms enriched within genes that comprise m73 describe cell division and the cell cycle. Module 153 is strongly activated by a comparison of L3 lethargus (a larval period prior to molting) to active L3 larvae and, accordingly, this module is enriched for GO terms that describe cuticle remodeling. Similarly, m200 is activated in response to the pathogen *P. aeruginosa* and the top GO terms enriched within m200 genes include “defense response to Gram-negative bacteria” and “innate immune response”.

DEXICA modules can also capture gene expression patterns that define specific tissues. For example, m23 is strongly activated by GFP-based enrichment of neurons (GSE8004). The top GO term associated with m23 is “neuropeptide signaling pathway” and genes in this module have highly significant associations with neuronal anatomy terms. Similarly, m144 is strongly activated by genes specific to embryonic muscle cells (GSE8462) and is enriched for GO terms such as “striated muscle dense body” and anatomies such as “body wall musculature”. Together, these results indicate that DEXICA modules represent biologically meaningful sets of genes.

It is very important to note that a lack of an obvious agreement between the nature of a perturbation that strongly activates a module and the top GO terms enriched in that module does not necessarily indicate that a module is biologically meaningless. On the contrary, this apparent disagreement can be a powerful tool for understanding complex transcriptional responses to perturbations. For example, starvation of L1 larvae induces a developmental arrest and has widespread effects on metabolism and energy allocation (Baugh 2013). Module 51 is strongly activated by L1 starvation and is enriched for GO terms that describe ribosome biogenesis (*e.g.*, “ribosome biogenesis”, “nucleolus”), autophagy (*e.g.*, “autophagosome maturation”) and response to starvation (*e.g.*, “cellular response to nitrogen starvation”). Thus, this module represents several coordinated processes that become activated in response to larval starvation. Obvious agreement between a perturbation and GO terms would also be absent if a process had not previously been associated with a particular condition, making modules useful for hypothesis generation. For example, a wild *C. elegans* isolate, JU1580, shows strong activity of m55 relative to the laboratory *C. elegans* strain, N2,

and genes within m55 are enriched for the “innate immune response” GO term. This leads to a hypothesis that the immune systems of JU1580 and N2 have different levels of basal activity.

### Application of DEXICA modules to improve gene annotations

Because gene annotations can be incomplete and biased, knowledge about which genes tend to be co-expressed in an organism can help inform which annotations are more likely to be biologically relevant and help infer annotations of orphan genes based on annotations of other module members (see Supplemental information for additional discussion). To this end, we devised a calculation of scores, which we call “module-weighted annotations”, based both on how much a given annotation is enriched in each module (matrix *E*, Figure 1b-c) and the degree to which a given gene belongs to that same module (matrix *H*, Figure 1c).

**Module-weighted GO terms:** To test the utility of module-weighted annotations, we first examined whether they generally recapitulate traditional (Boolean) annotations. If so, then genes traditionally associated with each annotation should have larger weights for those annotations (when normalization is omitted from their calculation) than do other genes (two-sample KS test, alpha level = 0.05). For this analysis we used GO terms as annotations and restricted the set to terms with at least 15 annotated genes in order to ensure robust signals; this set comprised 1651 GO terms. In 98.6% (1628/1651) of cases, genes associated with each term had significantly larger module-weighted annotations than did other genes, indicating that module-weighted GO annotations do recapitulate Boolean annotations.

We ranked each GO term by the weight of its most strongly associated gene (Table 1). GO terms with the most highly weighted genes were “ribosome” (CC) and “structural constituent of ribosome” (MF). Consistent with this finding, genes involved in ribosome biogenesis are known to be transcriptionally co-regulated (Wade *et al.* 2006). On the other hand, GO terms associated with signal transduction and kinase activity showed the lowest gene weights. These results are statistically significant (see Supplemental Methods methods) and show that some types of gene groupings (*e.g.*, genes encoding kinases) often used in the analysis of gene expression data may not actually represent functions that are coordinately regulated at the transcript level. Therefore, we suggest that researchers use caution when interpreting a result that certain annotations, like “kinases”, appear to be enriched in a transcriptomics experiment.

A sensitive test will tend to give similar results despite small amounts of random noise being added to the input data. To test the sensitivity of a module-weighted annotation-based enrichment test, we obtained gene fold changes for the 5 most recent (at the time of query) *C. elegans* Affymetrix experiments deposited to the GEO database. These experiments were not included in the data used to construct the modules. We then added varying amounts of Gaussian noise to the fold changes and for each level of noise, we calculated enrichment z-scores for each GO term using three different methods: the Kolmogorov-Smirnov (KS) test, the *t*-test (*n.b.*, one-sample and two-sample *t*-tests produced nearly identical results, data not shown), and scalar projection of the gene fold changes onto a module-weighted GO term matrix. We then compared these results to those obtained by each method when no noise was added (Figure 3a). Dissimilarity to the initial results (zero added noise) increased rapidly with added noise for both the KS test and *t*-test, while the projection-based results were

similar ( $\rho > 0.75$ ) at even the highest noise levels tested (5 standard deviations), suggesting that the projection-based test is more sensitive than both the KS and *t*-tests.

A specific test will tend to show dissimilar results when given dissimilar inputs. To test the specificity of module-weighted annotation analysis, we selected the 100 most dissimilar experiments (correlation of gene fold changes near zero) from a set of 188,805 comparisons. We compared the results of projecting the gene fold changes from each experiment onto the module-weighted annotation matrix to the results obtained from the KS and *t*-tests and found that the experiments showed the weakest similarity in the significance levels of annotations when using the projection method ( $P = 5.2e-13$ , Figure 3b). These results show that annotation enrichment analysis using module-weighted annotations may provide more reliable biological insights than gene set enrichment analyses that rely on the KS (GSEA (Subramanian *et al.* 2005)) or *t*-tests (PAGE (Kim and Volsky 2005), GAGE (Luo *et al.* 2009)). Module-weighted GO terms for each gene (matrix *R*, Figure 1c) are provided in File S4 and their significance in a query gene expression dataset can be tested using Tool 6 on <http://genemodules.org/>.

**Module-weighted promoter words:** While any Boolean gene annotation may be converted into a module-weighted annotation, module-based weights seem particularly well suited for describing regulatory sequences, such as putative transcription factor and microRNA binding sites. To this end, we generated weighted annotations for each of the 5230 words in the promoter word dictionary we constructed using the Mobydick algorithm (described above and in Methods).

To validate predicted regulatory word weights, we searched the literature, the JASPAR database of transcription factor binding profiles (Mathelier *et al.* 2016), and the GEO database for *C. elegans* transcription factors with both an experimentally characterized DNA-binding profile and, separately, a microarray experiment that measured gene expression in a loss-of-function mutant of the transcription factor gene. This search yielded 6 transcription factors: *daf-12*, *daf-16*, *hif-1*, *hlh-30*, *lin-14*, and *nhr-23*. We then projected the loss-of-function microarray experiment data (positively and negatively changing genes separately) onto the module-weighted word matrix, *R* (Figure 1c), and calculated z-scores for each word. Finally, we compared the top-scoring words to the DNA-binding profiles of the respective transcription factors. If the predicted promoter-word weights are accurate, then words that resemble the binding profile should score highly in this analysis.

For *hif-1* and *nhr-23*, the most significantly enriched words in the positively and negatively changing genes, respectively, matched the canonical binding sites (Table 2). A word matching the *hlh-30* binding site scored 6<sup>th</sup> overall among the up-regulated genes, and for *daf-12*, four of the top 20 words for the up-regulated genes contained GAACT or AACTT, which partially matched the reverse complement of a reported *daf-12* binding half-site, AGTTCA (Shostak *et al.* 2004). In the *daf-16* data set, several words matching the so-called “*daf-16* associated element” (DAE) (Zhang *et al.* 2013) scored highly. However, none of the four words matching the canonical *daf-16* binding site, T(G/A)TTTAC were among the words comprising our Mobydick promoter-word dictionary, precluding these from being represented in the analysis. The canonical binding site for the final transcription factor, *lin-14*, is GAAC, but like the canonical *daf-16* binding site, neither this word nor its reverse complement was present in the promoter word dictionary, precluding



■ **Table 1** GO categories with strongest and weakest gene weights. The table shows the top 25 and the bottom 5 GO categories, ranked in decreasing order of the weight of their most strongly associated gene prior to normalization. Similar GO categories are grouped together, and the genes with the highest weights for each group are also shown. *rpl-* (ribosomal protein, large subunit) and *rps-* (ribosomal protein, small subunit) genes encode ribosomal proteins, and *his-* (histone) genes encode nucleosome components

GO Term Group	GO ID	Rank	GO Term	Top Genes
Ribosome	GO:0005840	1	ribosome	<i>rpl-43, rps-15, rps-14, rpl-22, rpl-32</i>
	GO:0003735	2	structural constituent of ribosome	
	GO:0030529	3	ribonucleoprotein complex	
	GO:0006412	4	translation	
	GO:0043228	10	non-membrane-bounded organelle	
	GO:0043232	11	intracellular non-membrane- bounded organelle	
Heme binding	GO:0032991	20	macromolecular complex	
	GO:0020037	5	heme binding	<i>R05D8.9, E02C12.6, T16G1.6, cyp-35A1, C33C12.8</i>
	GO:0046906	6	tetrapyrrole binding	
	GO:0005506	7	iron ion binding	
	GO:0004497	8	monooxygenase activity	
	GO:0009055	9	electron carrier activity	
Nucleosome	GO:0016705	15	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	
	GO:0006334	12	nucleosome assembly	<i>his-4, his-8, his-9, his-45, his-62</i>
	GO:0034728	13	nucleosome organization	
	GO:0031497	14	chromatin assembly	
	GO:0000786	16	nucleosome	
	GO:0032993	17	protein-DNA complex	
	GO:0006333	18	chromatin assembly or disassembly	
	GO:0065004	21	protein-DNA complex assembly	
Lipid glycosylation	GO:0071824	22	protein-DNA complex subunit organization	
	GO:0006323	23	DNA packaging	
Cuticle	GO:0030259	19	lipid glycosylation	<i>C33C12.8, cyp-35A5, cyp-35A1, dhs-23, clec-210</i>
	GO:0042302	24	structural constituent of cuticle	<i>col-2, col-84, col-158, col-44, R07E5.4</i>
Oxidoreductase activity	GO:0016491	25	oxidoreductase activity	<i>B0272.4, Y75B8A.4, acox-1, F58A6.1, dhs-18</i>
...	...	...	...	...
Presynaptic membrane	GO:0042734	1647	presynaptic membrane	<i>K02E11.7, C45G9.6, clec-233, ttr-10, cup-4</i>
Signal transduction	GO:0046579	1648	positive regulation of Ras protein signal transduction	<i>R04B5.6, cyp-33D3, C34D10.1, stdh-2, <sup>19</sup> FF10.1</i>
	GO:0051057	1649	positive regulation of small GTPase mediated signal transduction	
Calcium-dependent kinase	GO:0009931	1650	calcium-dependent protein serine/ threonine kinase activity	<i>tag-83, T21H8.5, cca-1, ZC101.1, ztf-16</i>
	GO:0010857	1651	calcium-dependent protein kinase activity	

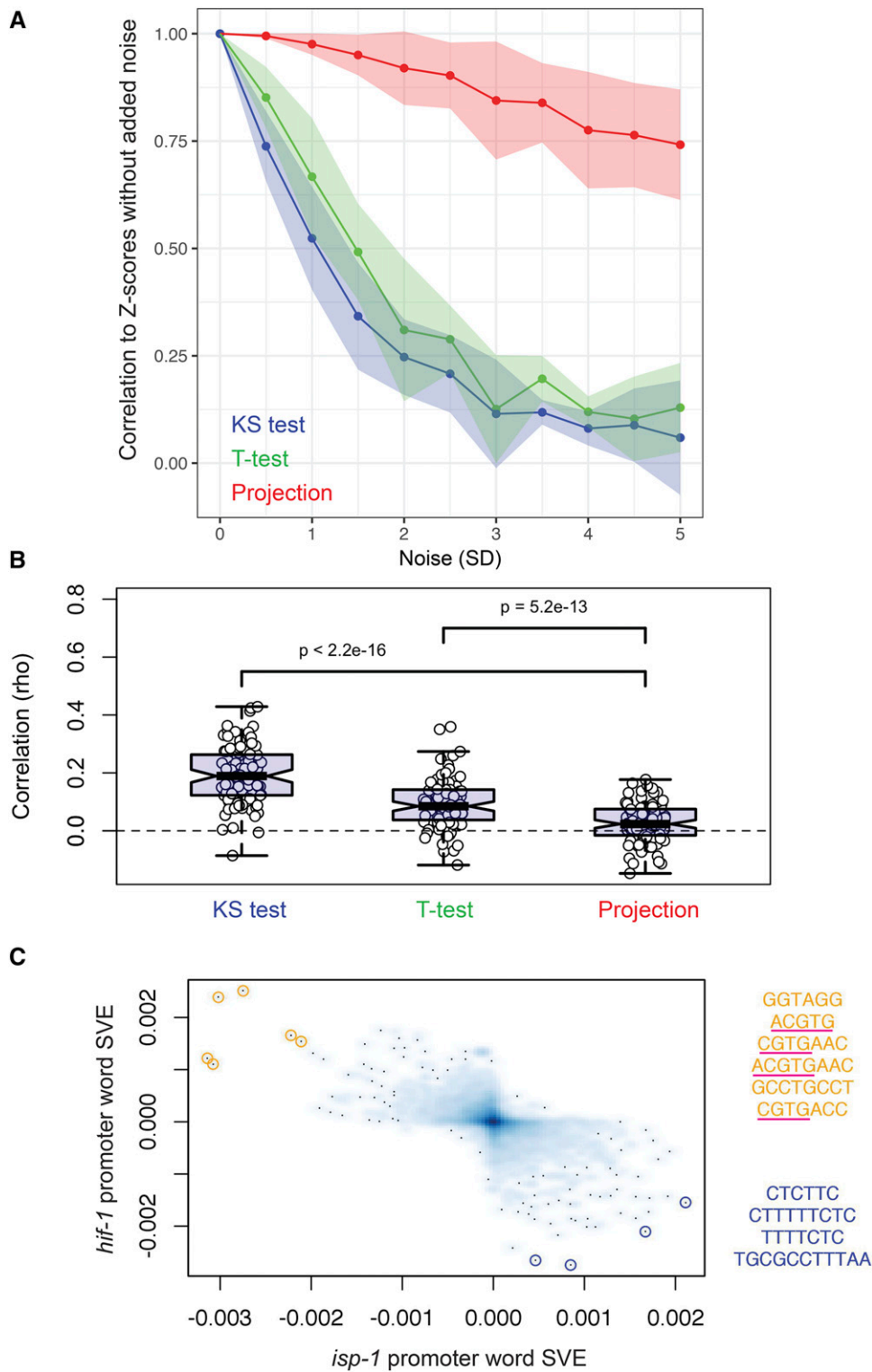
it from representation. Taken together, these results suggest that module weighted regulatory sequences can be used to determine important regulatory sites in a gene expression experiment, and further validate our method for calculating weighted annotations.

Finally, we tested whether module-weighted words could reveal activity of the transcription factor HIF-1, previously shown to be required for longevity of the *isp-1* mitochondrial respiration mutants (Lee *et al.* 2010). We calculated promoter word z-scores for the *isp-1* microarray data set and compared the results to those for *hif-1*. As predicted, we observed a very strong anti-correlation between the promoter word z-scores for *isp-1* mutants and those for *hif-1* mutants (Figure 3c,  $R = -0.581$ ,  $P < 2.2E-16$ ) and four of the six most active words matched the canonical HIF-1 binding site [(A/G)CGTG; underlined in Figure 3c]. These results further support the utility of module-weighted promoter oligonucleotides

for identifying biologically relevant regulatory sequences directly from gene expression data.

### Informative descriptions and annotations of DEXICA modules

Gene co-expression modules are useful for interpreting gene expression data because they often represent transcriptional signatures of biological processes (Segal *et al.* 2003b). Therefore, detailed information about modules that are active in a gene-expression experiment would be of primary interest to an investigator. We showed above that *C. elegans* gene modules constructed using DEXICA are biologically informative. As a comprehensive resource, we created a Module Annotation Page for each of the 209 modules (see Data availability, Tool 3). Each page shows the module's significantly enriched GO terms (excerpted in File S2), tissues of known



**Figure 3** Performance of module-weighted annotation-based tests. (a) To assess sensitivity, Gaussian noise was added to gene expression data from 5 recent *C. elegans* Affymetrix experiments not included in the compendium used to train the modules. [The standard deviation of the noise distribution ( $\mu = 0$ ) was varied from 0.5x to 5x the standard deviation of the gene fold change distribution.] At each noise level, z-scores for each GO term (based on 100 random permutations of the gene IDs in the input data) were calculated using three different methods: KS – Kolmogorov-Smirnov test; t-test – two-sample t-test (one-sample test gave highly similar results, data not shown); projection – projection of the gene fold changes into the space defined by the module-weighted annotation matrix,  $R$ . When conducting KS and t-test, fold changes of genes assigned to each GO term were compared to fold changes of genes not assigned to that term. Spearman correlation coefficients were calculated between Z-scores at each noise level and Z-scores without any noise added. (b) To assess specificity, 100 highly dissimilar pairs of experiments (Spearman correlation,  $\rho$ , of gene fold changes near 0) were selected from a set of 188,805 pairs generated using published microarray data. For each contrast belonging to a pair, we determined Z-scores for GO annotations using the three methods and calculated the rank correlation ( $\rho$ ) of the absolute value of these Z-scores between pair members. The center of the box represents the median value and whiskers extend to the most extreme data point that is not further than 1.5 times the IQR from the box. (c) Gene fold changes in *isp-1* and *hif-1* mutants were projected into promoter word space using module-weighted promoter words (see Methods). The 10 points furthest from the origin are highlighted with colored circles in the figure (orange = positive SVE for *hif-1*, blue = negative SVE for *hif-1*), and their words corresponding to these points are shown to the right of the figure. Four of the six words highlighted in orange contain full or partial matches to the canonical *hif-1* binding site, (A/G) CGTG (underlined).

expression, enriched promoter and 3'-UTR oligonucleotide words, transcription-factor binding sites predicted using HOMER (Heinz et al. 2010) and component genes (each linked to a WormBase description). The component genes can be interrogated further for enrichment of additional gene annotations or regulatory motifs [e.g.

using RSAT (Nguyen et al. 2018)] to gain a deeper understanding of a module's function.

A limitation of using GO terms and other pre-existing annotations to interpret gene modules is that some cellular activities are poorly annotated (see UPR<sup>mt</sup> example in section 1 below). For this reason, an

■ **Table 2** Top scoring promoter words for transcription factor perturbation experiments. The table shows the top three most significant words (those with the most extreme z-scores), calculated using module-weighted promoter words, for each transcription factor loss-of-function microarray experiment. If a full or partial match (4 or more bases) to the canonical binding site of the factor does not occur within the top 3 ranking words, the most significant such match is also shown. Bold letters indicate matching positions to canonical binding sites

Factor	Canonical site	Top words	Projection set	Z score	Z score rank
DAF-16	T(G/A)TTTAC, CTTATCA*	<b>TTCTTATCA</b>	Down-regulated	-3.95	1
		GGAAG	Down-regulated	-3.69	2
		TTTTCTG	Down-regulated	3.47	3
HIF-1	ACGTG	<b>ACGTGAAC</b>	Up-regulated	4.39	1
		<b>CGTGAAC</b>	Up-regulated	4.38	2
		<b>ACGTG</b>	Up-regulated	3.71	3
NHR-23	AGGTCA	<b>AGGTCA</b>	Down-regulated	-5.29	1
		<b>TGACCTA</b>	Down-regulated	-4.72	2
		CCTCCCCC	Down-regulated	-4.39	3
HLH-30	TCACGTGA(C/T)	CTTACTATT	Up-regulated	-4.3	1
		CGTAATCC	Up-regulated	4.14	2
		CTTTTTTCT	Down-regulated	4.07	3
		<b>CACGTG</b>	Up-regulated	-3.09	20
LIN-14	GAAC	CCTACCTACCTA	Down-regulated	4.35	1
		GCGCGTCAAATA	Up-regulated	-3.95	2
		GCCGCGCACCCC	Down-regulated	-3.78	3
		<b>GTTCTGG</b>	Down-regulated	-2.53	109
DAF-12	AGTTCA	CCCCAC	Down-regulated	-4.41	1
		GCTC	Up-regulated	4.28	2
		CCCCGCC	Down-regulated	-4	3
		<b>AACTTTT</b>	Up-regulated	3.51	11

additional strategy to understanding what a given module represents is to examine its activity under a variety of conditions. To facilitate this, in the Module Annotation Pages we have also provided a ranked list of perturbations (from the 716 we generated) that activate each module significantly (also see File S2). While it is unlikely that a typical gene module can be described completely using a single semantic annotation due to the complexity of gene regulatory networks, we suggest processes that each module likely captures based on our comparison of module activity under several conditions and our examination of enriched GO terms (File S2). In a typical workflow, once the active modules in a gene expression experiment are identified by a researcher, we recommend that the researcher consult Module Annotation Pages and File S2 to infer the biological process(es) described by the modules of interest.

### Guide to using DEXICA modules

**Gene modules can uncover processes that are poorly annotated and therefore missed by conventional analyses:** Mutations in components of the electron transport chain reduce respiration, slow development and reproduction and increase lifespan in *C. elegans* (Dancy *et al.* 2014). The mitochondrial unfolded protein response (UPR<sup>mt</sup>) is induced in response to a stoichiometric imbalance between nuclear and mitochondrial proteins within mitochondria, and this process is known to be active in long-lived respiration mutants (Baker *et al.* 2012; Durieux *et al.* 2011). However, analysis of significantly differentially expressed genes in the *isp-1* respiration mutant, either using conventional GO term- or KEGG pathway- overrepresentation tests or using a more recent tool designed to query more comprehensive annotations of *C. elegans* genes, WormCat (Holdorf *et al.* 2020), failed to identify UPR<sup>mt</sup> or, for that matter, any process related to the mitochondria (Figure 4a). We asked whether application of DEXICA modules could successfully identify a UPR<sup>mt</sup> gene expression signature in *isp-1* mutants.

As a first step of gene expression analysis using DEXICA modules, we determine a fold change of every detected gene between *isp-1*

mutants and wild-type animals. This calculation is done without imposing any cut offs on the data (e.g., p-values or effect sizes). We then project the resulting vector onto the module definition matrix (*S*, see Figure 1), which can be done by using Tool 1 of the online user interface (<http://genemodules.org/>). Three modules - m47, m66 and m169 - displayed the highest activity (Figure 4b). Module activity is represented as Signed Variance Explained (SVE), where + or - indicates direction of activity (useful when comparing two or more experiments), and the absolute value of SVE represents the relative proportion of variance in a gene expression experiment explained by a particular module. The sum of absolute values of SVE for all modules is 1.

By consulting Module Annotation Pages (see section “Annotation of DEXICA modules” and Tool 3 on <http://genemodules.org/>), we determine that the same three modules are strongly active in other respiratory-chain mutants, but also that the most strongly activating non-respiration perturbation of m47 is a mutation in *spg-7*, a mitochondrial protein quality-control protease, disruption of which is known to induce UPR<sup>mt</sup> (Pellegrino *et al.* 2014) (Figure 4b - a plot like this can be generated by calculating module activity in each experiment using Tool 1 and plotting the resulting SVE values against each other). This finding suggests that m47 may encompass UPR<sup>mt</sup> genes. If true, then m47 should be active in animals with constitutive activity of the transcriptional UPR<sup>mt</sup> regulator, ATFS-1. To test this prediction, we calculated module activity in *atfs-1* gain-of-function mutants (this perturbation was not part of the original compendium, GEO Accession number GSE73669). Indeed, induction of the UPR<sup>mt</sup> transcriptional response in otherwise healthy animals using this mutation strongly and specifically induced activity of m47 (Figure 4c). Furthermore, genes that belong to m47 showed concordant expression in the *isp-1* respiration mutant, in response to UPR<sup>mt</sup> induction by disruption of *spg-7*, and in response to constitutive activity of ATFS-1 in normal animals (Figure 4c, d). As expected, mutation of *atfs-1* prevented these changes in animals with an

**A** 633 genes significantly differentially expressed in the *isp-1(-)* respiration mutant:

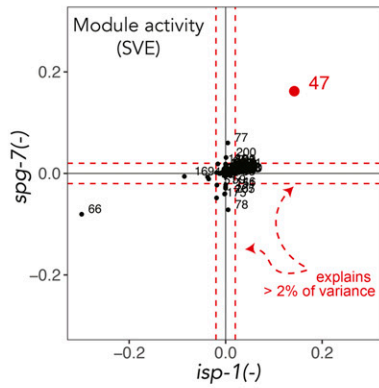
GO Term	Description	Adj. p-value
BP GO:0008360	regulation of cell shape	0.00732
BP GO:0018209	peptidyl-serine modification	0.0117
BP GO:0018105	peptidyl-serine phosphorylation	0.0175
BP GO:0035335	peptidyl-tyrosine dephosphorylation	0.0459
MF GO:0046982	protein heterodimerization activity	0.00615
MF GO:0004725	protein tyrosine phosphatase activity	0.0118
MF GO:0003824	catalytic activity	0.0472

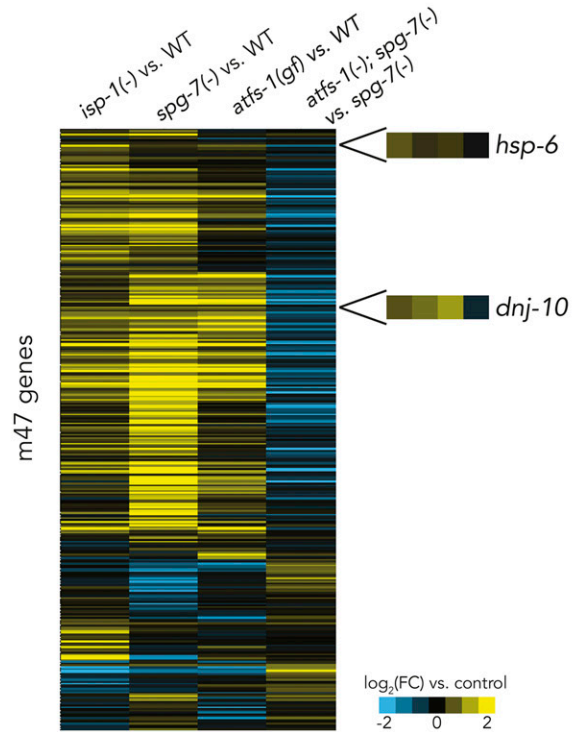
KEGG Pathway	Description	Adj. p-value
cel01130	Biosynthesis of antibiotics	0.025
cel01100	Metabolic pathways	0.016

WormCat Category	Description	Adj. p-value
1	Unknown	9.32E-09
	Metabolism	1.08E-05
	Stress response	3.45E-05
	Transcription: chromatin	3.44E-04
	Cytoskeleton	1.53E-03
2	Transcription: chromatin structure	5.94E-06
	Cytoskeleton: microtubule	9.51E-06
	Signaling: phosphatase	1.35E-03
	Stress response: detoxification	6.53E-03
	Stress response: pathogen	6.99E-03
3	Unknown: regulated by multiple stresses	1.57E-26
	Cytoskeleton: microtubule: tau tubulin kinase	1.32E-06
	Transcription: chromatin structure: histone	5.74E-06
	Signaling: phosphatase: Y	7.98E-05

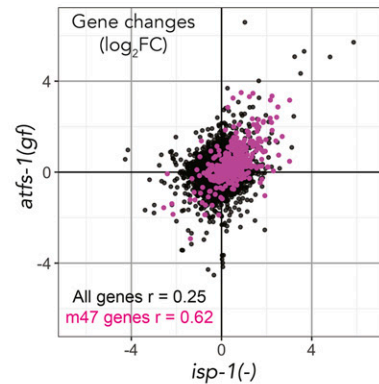
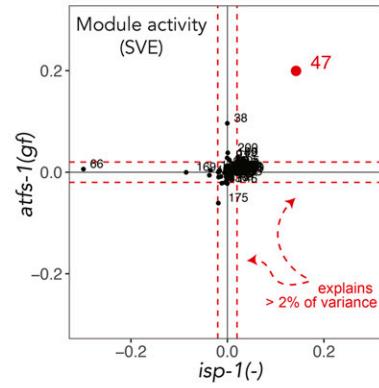
**B**



**D**

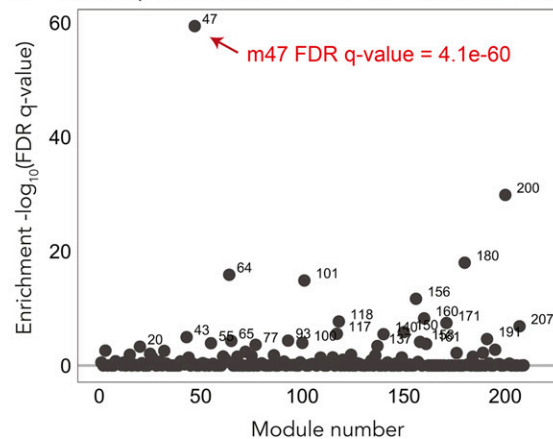


**C**



**E**

UPR<sup>mt</sup> genes up-regulated by ATFS-1 (Nargund et al. 2012)



**Figure 4** Analysis of module activity in *isp-1* respiration mutants reveals activity of the key mitochondrial unfolded protein response (UPR<sup>mt</sup>) transcription factor, ATFS-1. (a) Conventional annotation analyses of genes that are differentially expressed in *isp-1(qm150)* respiration mutants. GO term-overrepresentation was determined using GOrilla(Eden et al. 2009) and KEGG pathway-overrepresentation was determined using

induced UPR<sup>mt</sup> (*spg-7* RNAi) (column 4 Figure 4d). Moreover, two chaperone genes known to be induced during UPR<sup>mt</sup> (*hsp-6* and *dnj-10*) are part of m47. Taken together, these data strongly suggest that m47 represents genes induced by ATFS-1 during the UPR<sup>mt</sup>. This is an example of how identifying active gene modules in a mutant of interest and finding other conditions that activate the same modules can reveal biologically informative gene signatures.

Analysis of module activity, but not GO term or KEGG pathway enrichment analyses, was able to identify correctly a key biological process using gene expression data from *isp-1* mutants. Why did GO term enrichment analysis fail to recover the “mitochondrial unfolded protein response” term (GO:0034514)? To our surprise, there are only eight *C. elegans* genes annotated with this GO term (*haf-1*, *ubl-5*, *gcn-2*, *atfs-1*, *clpp-1*, *dve-1*, *hsp-6* and *hsp-60*). Among these, only *hsp-6* and *hsp-60* are induced during UPR<sup>mt</sup>, whereas the others are genes needed for activation of UPR<sup>mt</sup>. In the KEGG pathway database, UPR<sup>mt</sup> is not included as a discrete entry, but is part of the “Longevity regulating pathway – worm” (map04212) entry, and similar to GO, primarily includes inducers rather than mediators of the UPR<sup>mt</sup>. These examples illustrate how incomplete annotations and/or overly general groupings (e.g., containing both activators and mediators) can result in a failure of standard annotation enrichment methods to detect biological signals in gene expression data.

Nargund *et al.* have defined a set of ATFS-1-targeted genes based on up-regulation of these genes with or without *atfs-1* and in the absence or presence of mitochondrial stress (Nargund *et al.* 2012). While module analysis was able to reveal UPR<sup>mt</sup> without reliance on any pre-existing gene annotations or pre-defined gene sets, we wondered whether GSEA (Subramanian *et al.* 2005) analysis of *isp-1* gene expression data using this gene set would have been able to detect ATFS-1 activity. The Nargund *et al.* ATFS-1 gene set showed enrichment within *isp-1* genes, but this enrichment was not statistically significant (Figure S4). In contrast, when we tested enrichment of the ATFS-1 gene set within modules (enabled by Tool 2 on <http://genemodules.org/>), we found a highly significant enrichment within m47 (Figure 4e). These results show that given a set of functionally related genes, testing enrichment of that set within modules can be more informative than testing enrichment within a ranked list of gene changes, likely because modules comprise groups of genes that are functionally related.

**Gene modules can uncover gene signatures despite a lack of significant differential expression:** Reduction-of-function *isp-1* mutations extend lifespan in a manner dependent on the hypoxia inducible factor 1 (HIF-1) (Lee *et al.* 2010), but, unexpectedly, we found that the overlap between the significant genes obtained by comparing *isp-1* or *hif-1* mutants to wild type was not statistically significant ( $\chi^2$  test p-value = 0.17; Figure S5a). The other highly active modules in *isp-1* mutants besides m47 are m66 and m169. We

wondered if these modules represent genes regulated by HIF-1. Indeed, when we compared gene module activity in *hif-1* and *isp-1* mutants, we found a strong anti-correlation (Pearson correlation between SVE = -0.730,  $P = 4.7E-36$ ; Figure S5b) driven by m66 and m169. The negative correlation is consistent with the finding that the life extension observed when *isp-1* activity is reduced requires activity of HIF-1. The idea that HIF-1 directly regulates transcription of genes in m66 and m169 is further supported by the fact that the canonical HIF-1 binding site [(A/G)CGTG] is the top oligonucleotide sequence enriched in the promoters of the genes comprising these modules (q-values 8.13e-105 and 1.47e-28, respectively). The similarity between module activity in *isp-1* and *hif-1* datasets, despite a lack of similarity among their most differentially expressed genes, suggests that the role of HIF-1 in regulating the lifespan of *isp-1* mutants may be to instigate small but coordinated expression changes in many genes, most of which fail significance tests for differential expression in one or both datasets. These results demonstrate that gene modules are sensitive toward identification of transcriptional signatures and therefore can be useful for analysis of datasets in which few gene changes reach statistical significance.

**Gene modules as a hypothesis generation tool:** As shown above, the correlation between the module activities of *hif-1* and *isp-1* was substantial ( $r = -0.730$ ), and therefore served as a strong validation of the module approach to uncovering functionally significant processes. To determine how often two sets of module activities generated from different experiments could be expected to show this degree of similarity, we determined module activity correlations for all possible pairs of the 716 experimental contrasts, excluding pairs in which both contrasts originated from the same experiment (*i.e.*, from the same GEO series). This produced 188,805 contrast pairs, 13,376 (7.08%) of which showed a statistically significant correlation (Holm corrected p-value < 0.05). As expected, the highest-ranking pairs were the same experiment performed by different labs or at different times. The strength of the *hif-1* and *isp-1* correlation would have fallen within the top 1%, had those experiments been part of the contrast set. Interestingly, had we observed the similarity between the *hif-1* and *isp-1* projections without any prior genetic data, we might have hypothesized a role for *hif-1* in *isp-1* mutants before such a role was discovered from genetic screening (Lee *et al.* 2010). Therefore, to give the reader a chance to find other unexpected correlations between perturbations (*i.e.*, gene expression changes that activate similar modules but that are generated by different experiments), we include the full set of contrast comparisons in File S3. We hope these comparisons will prove useful for hypothesis generation.

### Guide to using module-weighted annotations

To aid in functional interpretation of gene expression data despite sparse annotations of many *C. elegans* genes, we have developed a

---

DAVID (Huang *et al.* 2009). WormCat (Holdorf *et al.* 2020) is an online analysis tool that queries the GO database but works independently of GO and addresses some of its limitations. Significant gene expression changes were determined using the *limma* R package and FDR-adjusted p-value of < 0.05. (b) Knockdown of a protein quality-control protease *spg-7* is the top non-respiration microarray experiment in the compendium that induces activity of the *isp-1* module m47. Module activity is expressed as signed variance explained (SVE). (c) Comparison of *isp-1* (*qm150*) mutants and animals that express a constitutively active form of *atfs-1*, a key transcription factor required for the UPR<sup>mt</sup>. The gene expression scatter plot shows all detected genes (22,625; black) overlaid with genes that belong to m47 (367; pink).  $r$ , Pearson correlation coefficients. (d) Normalized changes in expression of genes that belong to m47 (367 genes) in response to inhibition of respiration (*isp-1* mutation, column 1), activation of UPR<sup>mt</sup> by disrupting mitochondrial protein quality control (column 2), constitutive activation of *atfs-1* under normal conditions (column 3), and activation of UPR<sup>mt</sup> in the absence of *atfs-1* (column 4). (e) Genes up-regulated by ATFS-1 during UPR<sup>mt</sup>, as determined in Nargund *et al.* 2012 (Nargund *et al.* 2012), are strongly enriched in m47. Enrichment was calculated using hypergeometric statistics.

process of transferring gene annotations based on the degree of gene co-expression to create “module-weighted gene annotations” (see description above). This approach provisionally transfers annotations, such as GO terms, that are enriched in a module to all gene members of that module. This can be useful in at least two ways: for studying individual genes and for identifying functional categories within gene expression data.

For studying individual genes: if a query gene has a strong module-weighted association to a GO term with which it is not traditionally associated, it means that this gene is co-expressed with other genes that *are* traditionally associated with the GO term. For example, by applying Tool 5 (<http://genemodules.org/>) to the small ribosomal subunit S16 (*rps-16*), we find that most of the module-weighted GO terms associated with this gene have something to do with the ribosome or translation, indicating that *rps-16* is tightly co-expressed with other genes involved in ribosome biogenesis and not much else. On the other hand, while *sod-3*, a superoxide dismutase, does have significant weights for GO terms with which it is traditionally associated (e.g., oxidoreductase activity, superoxide metabolic process and response to superoxide), “catalase activity” ranks more highly for *sod-3*. This indicates that *sod-3* is co-expressed with genes annotated as catalases. While it is unlikely that *sod-3* has a novel catalase activity, it is more likely that expression of *sod-3*, a superoxide dismutase, is coordinated with expression of catalases, since hydrogen peroxide produced by *sod-3* is further degraded by catalases to avoid damage to the cell. Some of the other top-ranking module-weighted GO terms for *sod-3* describe dioxygenase activity. Similarly, this may indicate that dioxygenases generate superoxide radicals, and therefore increased expression of these enzymes is typically correlated with increased expression of a superoxide dismutase. Therefore, analysis of module-weighted annotations of individual genes can help form hypotheses about novel gene functions and transcriptional co-regulation of distinct biological processes.

The second application of module-weighted annotations is similar to a common analysis scheme whereby over-represented annotations (e.g., GO terms) are identified within a list of significantly differentially expressed genes. The key difference is that traditional annotations have been extended using information about gene co-expression (i.e., gene modules). Furthermore, this approach leverages all expression data, even genes that do not change in a significant way. Instead of calculating enrichment p-values, Z-scores are calculated based on the strength of the module-weighted association between each gene and each annotation and the magnitude of that gene’s fold-change in the list of gene fold-changes. This analysis scheme is more selective and sensitive toward identifying functional categories of genes that change in response to a treatment than traditional over-representation analyses (see section “Application of DEXICA modules to improve gene annotations” above). Tool 6 on <http://genemodules.org/> accepts a list of gene fold-changes and returns a table of module-weighted GO terms ranked by their predicted activity.

## DISCUSSION

We have captured much of the transcriptional wiring of *C. elegans* by extracting gene co-expression modules from a large compendium of gene expression data. These 209 modules represent RNA-level signatures of diverse biological processes that can occur in *C. elegans* and, along with the tools we provide, can be used as a resource for analyzing gene expression data, or in fact to interrogate gene lists of any sort. Because genes are grouped into modules based purely on how they behave in experimental assays, and because modules encompass both previously-annotated and unannotated genes alike,

modules can reveal signals within transcriptomic data that otherwise would be missed due to incomplete knowledge of gene function, subtle gene expression changes or noisy data.

Experimentally, gene-expression modules can be used to deconvolve complex phenomena into subsets of co-regulated genes, genes that likely act together to mediate a specific process. The modules can be annotated extensively, as we have done, and these annotations can be applied provisionally to all genes in the module. Regulatory factors associated with specific annotations (like 5’ or 3’ oligonucleotide words) can be implicated as well, and functional links can be revealed between dissimilar conditions that activate the same modules (e.g. see File S3).

ICA has been applied to the prediction of gene modules before, but we could find no examples in the literature of optimizing the process using biological metrics in the manner that we describe. Combined with the improved ability to partition independent components provided by our artificial neural network approach, we expect that DEXICA will be useful for constructing gene modules for other organisms. The DEXICA software and our *C. elegans* modules and data are freely available as R packages online (see Data availability).

While we have taken steps to maximize module prediction accuracy for the microarray compendium we assembled, many additional gene modules may exist in *C. elegans* that were not perturbed sufficiently in the samples comprising the compendium to be detected. These gene modules would remain hidden. As new areas of research are explored and new experiments are published, however, new gene modules will be discovered. For example, most of the gene expression data in our compendium was collected using whole animals. Data collected from isolated cells or tissues could help produce modules that are active in a relatively small number of cells. The DEXICA package can be used to create new and improved gene modules using compendia with expanded information content.

The impetus for constructing gene modules in *C. elegans* was to create gene groupings that do not rely on existing annotations, because annotations of many genes are missing or incomplete. We then wondered whether numeric scores based on gene co-expression could actually improve the existing gene annotations, leading us to develop the concept of module-weighted annotations. By weighting an association between each gene and each annotation by the degree to which that annotation appears to predict gene modularity, annotations that are shared by co-expressed genes are “boosted” and those that are not are diminished. It is important to note, however, that relevance scores between genes and annotations could be calculated using other metrics of gene behavior as well. For example, a gene might be weakly associated with a term in the context of gene expression but strongly associated with it in the context of physical protein interactions.

We have found that analysis of module-weighted GO terms is less sensitive to noise than are typical statistical over-representation tests. Furthermore, because module-weighted annotations incorporate information about gene expression, they effectively model the gene-gene correlation structure of the system. This is useful because typical over-representation tests do not perform well with gene sets that have a high level of gene-gene correlation [i.e., annotations assigned to genes that are strongly co-expressed are more likely to be significant (Goeman and Buhlmann 2007; Gatti *et al.* 2010; Tamayo *et al.* 2016)]. GSEA, for example, deals with gene-gene correlation issue using permutation. We show that module-weighted GO terms produce significantly fewer false positives than do over-representation tests (see Figure 3b). Therefore, we think that module-weighted

annotations are a promising new way to address the gene-gene correlation problem and are working to develop it further for annotation enrichment analysis.

Module-weighted annotations may be especially useful for promoter/3'-UTR word analysis. Projection of gene expression changes onto the module-weighted word matrix allows identification of potential regulatory sequences directly from a list of gene fold changes, bypassing many steps required by traditional regulatory sequence detection (sequence retrieval, repeat masking, over-representation analysis, and background correction). More importantly, this analysis is more likely to yield true positive results because it makes use of a large amount of historical data (via the modules) to filter out sequences with no apparent regulatory function.

We used module-weighted word analysis to analyze data from six transcription factor perturbation microarray experiments. Four of the six produced high scoring promoter words that closely matched the known DNA binding sites of the corresponding transcription factors. For the two that did not, *daf-16* and *lin-14*, words that exactly matched the factors' canonical binding sites were not present in the promoter word dictionary generated by the Mobydick algorithm, and results for these factors could be poor for this reason. Interestingly, the second highest scoring word for the genes down-regulated upon *daf-16* perturbation was GGAAG, and this word occurs twice more as a substring among the top 20 scoring words. This sequence is a partial match to an alternative *daf-16* binding site reported in hookworm (Gao *et al.* 2010), G(A/G)(C/G)A(A/T)G, suggesting that this site may be functional in *C. elegans* as well.

Finally, because annotations shared among a significant fraction of the genes in a module are "transferred" to all genes in the module, module-weighted annotations can be used provisionally to infer novel functions for genes, which is especially useful for studying poorly annotated genes.

## ACKNOWLEDGMENTS

We would like to thank H. Li, H. El-Samad, H. Madhani and F. Li for critical review and discussion of this work. This work was supported by the NIH R36/R01 grant AG011816 to C. K. and by Calico Life Sciences. This work was initiated in C.K.'s lab at UCSF and completed at Calico Life Sciences.

## LITERATURE CITED

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25: 25–29. <https://doi.org/10.1038/75556>

Baker, B. M., A. M. Nargund, and T. Sun, 2012 Protective Coupling of Mitochondrial Function and Protein Synthesis via the eIF2 $\alpha$  Kinase GCN-2. *PLoS Genet.* 8: e1002760. <https://doi.org/10.1371/journal.pgen.1002760>

Barrett, T., D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista *et al.*, 2011 NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.* 39: D1005–D1010. <https://doi.org/10.1093/nar/gkq1184>

Baugh, L. R., 2013 To grow or not to grow: nutritional control of development during *Caenorhabditis elegans* L1 arrest. *Genetics* 194: 539–555. <https://doi.org/10.1534/genetics.113.150847>

Benjamini, Y., Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JSTOR* 57: 289–300.

Bussemaker, H. J., H. Li, and E. D. Siggia, 2000 Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* 97: 10096–10100. <https://doi.org/10.1073/pnas.180265397>

Chiappetta, P., M. C. Roubaud, and B. Torresani, 2004 Blind source separation and the analysis of microarray data. *J. Comput. Biol.* 11: 1090–1109. <https://doi.org/10.1089/cmb.2004.11.1090>

Cristina, D., M. Cary, A. Lunceford, C. Clarke, and C. Kenyon, 2009 A regulated response to impaired respiration slows behavioral rates and increases lifespan in *Caenorhabditis elegans*. *PLoS Genet.* 5: e1000450. <https://doi.org/10.1371/journal.pgen.1000450>

Dancy, B. M., M. M. Sedensky, and P. G. Morgan, 2014 Effects of the mitochondrial respiratory chain on longevity in *C. elegans*. *Exp. Gerontol.* 56: 245–255. <https://doi.org/10.1016/j.exger.2014.03.028>

Durieux, J., S. Wolff, and A. Dillin, 2011 The cell-non-autonomous nature of electron transport chain-mediated longevity. *Cell* 144: 79–91. <https://doi.org/10.1016/j.cell.2010.12.016>

Durinck, S., P. T. Spellman, E. Birney, and W. Huber, 2009 Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4: 1184–1191. <https://doi.org/10.1038/nprot.2009.97>

Eden, E., R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, 2009 GOrrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10: 48. <https://doi.org/10.1186/1471-2105-10-48>

Eklund, A. C., and Z. Szallasi, 2008 Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol.* 9: R26. <https://doi.org/10.1186/gb-2008-9-2-r26>

Engreitz, J. M., B. J. Daigle, J. J. Marshall, and R. B. Altman, 2010 Independent component analysis: mining microarray data for fundamental human gene expression modules. *J. Biomed. Inform.* 43: 932–944. <https://doi.org/10.1016/j.jbi.2010.07.001>

Fabregat, A., S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie *et al.*, 2017 The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 46: D649–D655. <https://doi.org/10.1093/nar/gkx1132>

Frigyesi, A., S. Veerla, D. Lindgren, and M. Hoglund, 2006 Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinformatics* 7: 290. <https://doi.org/10.1186/1471-2105-7-290>

Gao, X., Z. Wang, J. Martin, S. Abubucker, X. Zhang *et al.*, 2010 Identification of hookworm DAF-16/FOXO response elements and direct gene targets. *PLoS One* 5: e12289. <https://doi.org/10.1371/journal.pone.0012289>

Gatti, D. M., W. T. Barry, A. B. Nobel, I. Rusyn, and F. A. Wright, 2010 Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics* 11: 574. <https://doi.org/10.1186/1471-2164-11-574>

Gautier, L., L. Cope, B. M. Bolstad, and R. A. Irizarry, 2004 affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20: 307–315. <https://doi.org/10.1093/bioinformatics/btg405>

Gerstein, M. B., A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan *et al.*, 2012 Architecture of the human regulatory network derived from ENCODE data. *Nature* 489: 91–100. <https://doi.org/10.1038/nature11245>

Goeman, J. J., and P. Buhlmann, 2007 Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23: 980–987. <https://doi.org/10.1093/bioinformatics/btm051>

Gong, T., J. Xuan, C. Wang, H. Li, E. Hoffman *et al.*, 2007 Gene module identification from microarray data using nonnegative independent component analysis. *Gene Regul. Syst. Bio.* 1: 349–363.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin *et al.*, 2010 Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell.* 28: 576–589. <https://doi.org/10.1016/j.molcel.2010.05.004>

Holdorf, A. D., D. P. Higgins, A. C. Hart, P. R. Boag, G. J. Pazour *et al.*, 2020 WormCat: An Online Tool for Annotation and Visualization of *Caenorhabditis elegans* Genome-Scale Data. *Genetics* 214: 279–294. <https://doi.org/10.1534/genetics.119.302919>

Huang, D. W., B. T. Sherman, and R. A. Lempicki, 2009 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4: 44–57. <https://doi.org/10.1038/nprot.2008.211>

Hughes, T. R., M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton *et al.*, 2000 Functional discovery via a compendium of expression profiles. *Cell* 7: 109–126. <https://pubmed.ncbi.nlm.nih.gov/10929718>

Hunt-Newbury, R., R. Viveiros, R. Johnsen, A. Mah, D. Anastas *et al.*, 2007 High-throughput in vivo analysis of gene expression in

- Caenorhabditis elegans. *PLoS Biol.* 5: e237. <https://doi.org/10.1371/journal.pbio.0050237>
- Hyvärinen, A., and E. Oja, 2000 Independent Component Analysis: Algorithms and Application. *Neural Netw.* 13: 411–430. [https://doi.org/10.1016/S0893-6080\(00\)00026-5](https://doi.org/10.1016/S0893-6080(00)00026-5)
- Ihmels, J., S. Bergmann, and N. Barkai, 2004 Defining transcription modules using large-scale gene expression data. *Bioinformatics* 20: 1993–2003. <https://doi.org/10.1093/bioinformatics/bth166>
- Kanehisa, M., S. Goto, Y. Sato, and M. Furumichi, 2012 KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40: D109–D114. <https://doi.org/10.1093/nar/gkr988>
- Kim, S. K., J. Lund, M. Kiraly, K. Duke, M. Jiang *et al.*, 2001 A Gene Expression Map for *Caenorhabditis elegans*. *Science* 293: 2087–2092. <https://doi.org/10.1126/science.1061603>
- Kim, S. Y., and D. J. Volsky, 2005 PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 6: 144. <https://doi.org/10.1186/1471-2105-6-144>
- Kong, W., C. R. Vanderburg, H. Gunshin, J. T. Rogers, and X. Huang, 2008 A review of independent component analysis application to microarray gene expression data. *Biotechniques* 45: 501–520. <https://doi.org/10.2144/000112950>
- Lee, R. Y. N., K. L. Howe, T. W. Harris, V. Arnaboldi, S. Cain *et al.* 2018 WormBase 2017: molting into a new stage. *Nucleic acids res.* 46: D869–D874. <https://doi.org/10.1093/nar/gkx998>
- Lee, S.-I., and S. Batzoglou, 2003 Application of independent component analysis to microarrays. *Genome Biol.* 4: R76. <https://doi.org/10.1186/gb-2003-4-11-r76>
- Lee, S. J., A. B. Hwang, and C. Kenyon, 2010 Inhibition of respiration extends *C. elegans* life span via reactive oxygen species that increase HIF-1 activity. *Curr. Biol.* 20: 2131–2136. <https://doi.org/10.1016/j.cub.2010.10.057>
- Luo, W., M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf, 2009 GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 10: 161. <https://doi.org/10.1186/1471-2105-10-161>
- Mathelier, A., O. Fornes, D. J. Arenillas, C. Y. Chen, G. Denay *et al.* 2016 JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids res* 44: D110–D115. <https://doi.org/10.1093/nar/gkv1176>
- Mayr, C., and D. P. Bartel, 2009 Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138: 673–684. <https://doi.org/10.1016/j.cell.2009.06.016>
- Michoel, T., R. De Smet, A. Joshi, K. Marchal, and Y. Van de Peer, 2009 Reverse-engineering transcriptional modules from gene expression data. *Ann. N. Y. Acad. Sci.* 1158: 36–43. <https://doi.org/10.1111/j.1749-6632.2008.03943.x>
- Mostafavi, S., C. Gaiteri, S. E. Sullivan, C. C. White, S. Tasaki *et al.*, 2018 A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat. Neurosci.* 21: 811–819. <https://doi.org/10.1038/s41593-018-0154-9>
- Nargund, A. M., M. W. Pellegrino, C. J. Fiorese, B. M. Baker, and C. M. Haynes, 2012 Mitochondrial import efficiency of ATRF-1 regulates mitochondrial UPR activation. *Science* 337: 587–590. <https://doi.org/10.1126/science.1223560>
- Nguyen, N. T. T., B. Contreras-Moreira, J. A. Castro-Mondragon, W. Santana-Garcia, and R. Ossio 2018 RSAT 2018: regulatory sequence analysis tools 20th anniversary *Nucleic Acids Res.* 2: W209–W214. <https://pubmed.ncbi.nlm.nih.gov/29722874/>
- Oeckinghaus, A., M. S. Hayden, and S. Ghosh, 2011 Crosstalk in NF- $\kappa$ B signaling pathways. *Nat. Immunol.* 12: 695–708. <https://doi.org/10.1038/ni.2065>
- Pellegrino, M. W., A. M. Nargund, N. V. Kirienco, R. Gillis, C. J. Fiorese *et al.*, 2014 Mitochondrial UPR-regulated innate immunity provides resistance to pathogen infection. *Nature* 516: 414–417. <https://doi.org/10.1038/nature13818>
- Purdom, E., and S. P. Holmes, 2005 Error distribution for gene expression data. *Stat Appl Genet Mol Biol* 4: Article16. <https://doi.org/10.2202/1544-6115.1070>
- Rustici, G., N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag *et al.*, 2013 ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res.* 41: D987–D990. <https://doi.org/10.1093/nar/gks1174>
- Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein *et al.*, 2003b Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34: 166–176. <https://doi.org/10.1038/ng1165>
- Segal, E., R. Yelensky, and D. Koller, 2003a Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* 19: i273–i282. <https://doi.org/10.1093/bioinformatics/btg1038>
- Shen, C., D. Nettleton, M. Jiang, S. K. Kim, and J. A. Powell-Coffman, 2005 Roles of the HIF-1 hypoxia-inducible factor during hypoxia response in *Caenorhabditis elegans*. *J. Biol. Chem.* 280: 20580–20588. <https://doi.org/10.1074/jbc.M501894200>
- Shostak, Y., M. R. Van Gilst, A. Antebi, and K. R. Yamamoto, 2004 Identification of *C. elegans* DAF-12-binding sites, response elements, and target genes. *Genes Dev.* 18: 2529–2544. <https://doi.org/10.1101/gad.1218504>
- Simes, R. J., 1986 An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* 73: 751–754. <https://doi.org/10.1093/biomet/73.3.751>
- Smyth, G. K., 2005 Limma: linear models for microarray data, pp. 397–420 in *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, edited by Gentleman, R., V. Carey, S. Dudoit, R. Irizarry, and W. Huber, Springer, New York. [https://doi.org/10.1007/0-387-29362-0\\_23](https://doi.org/10.1007/0-387-29362-0_23)
- Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert *et al.*, 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102: 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Tamayo, P., G. Steinhardt, A. Liberzon, and J. P. Mesirov, 2016 The limitations of simple gene set enrichment analysis assuming gene independence. *Stat. Methods Med. Res.* 25: 472–487. <https://doi.org/10.1177/0962280212460441>
- Teschendorff, A. E., M. Journée, P. A. Absil, R. Sepulchre, and C. Caldas, 2007 Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLOS Comput. Biol.* 3: e161. <https://doi.org/10.1371/journal.pcbi.0030161>
- Wade, C. H., M. A. Umbarger, and M. A. McAlear, 2006 The budding yeast rRNA and ribosome biosynthesis (RRB) regulon contains over 200 genes. *Yeast* 23: 293–306. <https://doi.org/10.1002/yea.1353>
- Wang, W., J. M. Cherry, Y. Nochomovitz, E. Jolly, D. Botstein *et al.*, 2005 Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl. Acad. Sci. USA* 102: 1998–2003. <https://doi.org/10.1073/pnas.0405537102>
- Wu, W.-S., W.-H. Li, and B.-S. Chen, 2006 Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. *BMC Bioinformatics* 7: 421. <https://doi.org/10.1186/1471-2105-7-421>
- W. Zhou, and R. B. Altman 2018 Data-driven human transcriptomic modules determined by independent component analysis. *BMC Bioinform.* 17: 327. <https://pubmed.ncbi.nlm.nih.gov/30223787/>
- Zaslavner, A., L. R. Baugh, and P. W. Sternberg, 2011 Metazoan operons accelerate recovery from growth-arrested states. *Cell* 145: 981–992. <https://doi.org/10.1016/j.cell.2011.05.013>
- Zhang, P., M. Judy, S. J. Lee, and C. Kenyon, 2013 Direct and indirect gene regulation by a life-extending FOXO protein in *C. elegans*: roles for GATA factors and lipid gene regulators. *Cell Metab.* 17: 85–100. <https://doi.org/10.1016/j.cmet.2012.12.013>
- Zhou, W. and R. B. Altman, 2018 Data-driven human transcriptomic modules determined by independent component analysis. *BMC Bioinform.* 19: 327. <https://doi.org/10.1186/s12859-018-2338-4>

Communicating editor: S. Lee