A theory of genetic analysis using transcriptomic phenotypes

Thesis by

David Angeles-Albores

In Partial Fulfillment of the Requirements for the
degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

[Year Degree Conferred]
Defended [Exact Date]

Choose one from the choices in the source code!! And delete this `todo` when you're done that

ACKNOWLEDGEMENTS

[Add acknowledgements here. If you do not wish to add any to your thesis, you may simply add a blank titled Acknowledgements page.]

ABSTRACT

[This abstract must provide a succinct and informative condensation of your work. Candidates are welcome to prepare a lengthier abstract for inclusion in the dissertation, and provide a shorter one in the CaltechTHESIS record.]

PUBLISHED CONTENT AND CONTRIBUTIONS

[Include a bibliography of published articles or other material that are included as part of the thesis. Describe your role with the each article and its contents. Citations must include DOIs or publisher URLs if available electronically.

If you are incorporating any third-party material in the thesis, including works that you have authored/co-authored but for which you have transferred copyright, you must indicate that permission has been secured to use the material. For example: “Fig. 2 reprinted with permission from the copyright holder, holder name”

Add the option `iknowwhattodo` to this environment to dismiss this message.]

Cahn, J. K. B., A. Baumschlager, et al. (2016). “Mutations in adenine-binding pockets enhance catalytic properties of NAD (P) H-dependent enzymes”. In: *Protein Engineering Design and Selection* 19.1, pp. 31–38. DOI: 10.1093/protein/gzv057.
J.K.B.C participated in the conception of the project, solved and analyzed the crystal structures, prepared the data, and participated in the writing of the manuscript.

Cahn, J. K. B., S. Brinkmann-Chen, et al. (2015). “Cofactor specificity motifs and the induced fit mechanism in class I ketol-acid reductoisomerases”. In: *Biochemical Journal* 468.3, pp. 475–484. DOI: 10.1042/BJ20150183.
J.K.B.C participated in the conception of the project, solved and analyzed the crystal structures, prepared the data, and participated in the writing of the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

NOMENCLATURE

Asteroid. A very small planet ranging from 1,000 km to less than one km in diameter. Asteroids are found commonly around other larger planets.

Galaxy. A system of stars independent from all other systems.

Chapter 1

INTRODUCTION

Chapter 2

GENETIC ANALYSIS OF A METAZOAN PATHWAY USING TRANSCRIPTOMIC PHENOTYPES

Abstract

RNA-seq is commonly used to identify genetic modules that respond to perturbations. In single cells, transcriptomes have been used as phenotypes, but this concept has not been applied to whole-organism RNA-seq. Linear models can quantify expression effects of individual mutants and identify epistatic effects in double mutants. However, interpreting these high-dimensional measurements is unintuitive. We developed a single coefficient to quantify transcriptome-wide epistasis which accurately reflects the underlying interactions. To demonstrate the power of our approach, we sequenced four single and two double C. elegans mutants. From these mutants, we successfully reconstructed the known hypoxia pathway. Using this approach, we uncovered a class of 31 genes that have opposing changes in expression in *egl-9(lf)* and *vhl-1(lf)* but the *egl-9(lf);vhl-1(lf)* mutant phenocopies *egl-9(lf)*. These changes violate the classical model of HIF-1 regulation, but can be explained by postulating a role of hydroxylated HIF-1 in transcriptional control.

Introduction

Genetic analysis of molecular pathways has traditionally been performed through epistatis analysis. Generalized epistasis indicates that two genes interact functionally; such interaction can involve the direct interaction of their products or the interaction of any consequence of their function (small molecules, physiological or behavioral effects) (L. S. Huang and Paul W Sternberg, 2006). If two genes interact, and the mutants of these genes have a quantifiable phenotype, the double mutant of interacting genes will have a phenotype that is not the sum of the phenotypes of the single mutants that make up its genotype. Epistasis analysis remains a cornerstone of genetics today (Phillips, 2008).

Recently, biological studies have shifted in focus from studying single genes to studying all genes in parallel. In particular, RNA-seq (Mortazavi et al., 2008) enables biologists to identify genes that change expression in response to a perturbation. Gene expression profiling using RNA-seq has become much more sensitive thanks to deeper and more frequent sequencing due to lower sequencing costs (Metzker, 2010), better and faster abundance quantification (Patro, Mount, and Kingsford, 2014; Bray et al., 2016; Patro, Duggal, et al., 2016), and improved differential expression analysis methods (Pimentel et al., 2016; Trapnell et al., 2013). RNA-seq has been successfully used to identify genetic modules involved in a variety of processes, including T-cell regulation (Singer et al., 2016; Shalek et al., 2013), the *Caenorhabditis elegans* (*C. elegans*) linker cell migration (Schwarz, Kato, and Paul W. Sternberg, 2012), and planarian stem cell maintenance (Van Wolfswinkel, Wagner, and Reddien, 2014; Scimone et al., 2014). For the most part, the role of transcriptional profiling has been restricted to target gene identification.

Although transcriptional profiling has been primarily used for descriptive purposes, transcriptomic phenotypes have previously been used to make genetic inferences. Microarray analyses in *S. cerevisiae* and *D. discoideum* were used to show that transcriptomes can be interpreted to infer genetic relationships in simple eukaryotes (Hughes et al., 2000; Van Driessche et al., 2005). eQTL studies in many organisms, from yeast to humans, have established the usefulness of transcriptomic phenotypes for population genetics studies (Brem et al., 2002; Schadt et al., 2003; Li et al., 2006; King et al., 2014). In cell culture, single-cell RNA-seq has seen significant progress towards using transcriptomes as phenotypes with which to test genetic interactions (Adamson et al., 2016; Dixit et al., 2016). More recently, we have identified a new developmental state of *C. elegans* using whole-organism transcriptome profiling (Angeles-Albores, Leighton, et al., 2016). To investigate the ability of whole-organism transcriptomes to serve as quantitative phenotypes for epistasis analysis in metazoans, we sequenced the transcriptomes of of four well-characterized loss of function mutants in the *C. elegans* hypoxia pathway (Epstein et al., 2001; Shen, Shao, and Powell-Coffman, 2006; Shao, Zhang, and Powell-Coffman, 2009; H. Jiang, Guo, and Powell-Coffman, 2001).

Metazoans depend on the presence of oxygen in sufficient concentrations to support aerobic metabolism. Genetic pathways evolved to rapidly respond to any acute or chronic changes in oxygen levels at the cellular or organismal level. Biochemical and genetic approaches identified the Hypoxia Inducible Factors (HIFs) as an important group of oxygen-responsive genes that are involved in a broad range of human pathologies (Gregg L. Semenza, 2012).

Hypoxia Inducible Factors are highly conserved in metazoans (Loenarz et al., 2011). A common mechanism for hypoxia-response induction is heterodimerization between a HIFα and a HIFβ subunit; the heterodimer then initiates transcription of target genes (B. H. Jiang et al., 1996). The number and complexity of HIFs varies throughout metazoans, with humans having three HIFα subunits and two HIFβ subunits, whereas in the roundworm *C. elegans* there is a single HIFα gene, *hif-1* (H. Jiang, Guo, and Powell-Coffman, 2001) and a single HIFβ gene, *ahr-1* (Powell-Coffman, Bradfield, and Wood, 1998). HIF target genes have been implicated in a wide variety of cellular and extracellular processes including glycolysis, extracellular matrix modification, autophagy and immunity (G L Semenza et al., 1994; Bishop et al., 2004; Shen, Nettleton, et al., 2005; Bellier et al., 2009; Gregg L. Semenza, 2012).

Levels of HIFα proteins tend to be tightly regulated. Under conditions of normoxia, HIF-1α exists in the cytoplasm and partakes in a futile cycle of continuous protein production and rapid degradation (L. E. Huang et al., 1996). HIF-1α is hydroxylated by three proline hydroxylases in humans (PHD1, PHD2 and PHD3) but is only hydroxylated by one proline hydroxylase (EGL-9) in *C. elegans* (Kaelin and Ratcliffe, 2008). HIF-1 hydroxylation increases its binding affinity to Von Hippel Lindau Tumor Suppressor 1 (VHL-1), which allows ubiquitination of HIF-1 leading to its subsequent degradation. In *C. elegans*, EGL-9 activity is inhibited by binding of CYSL-1, and CYSL-1 activity is in turn inhibited at the protein level by RHY-1, possibly by post-translational modifications to CYSL-1 (Ma et al., 2012) (see Fig. 2.1).

Here, we show that transcriptomes contain robust signals that can be used to infer relationships between genes in complex metazoans by reconstructing the hypoxia pathway in *C. elegans* using RNA-seq. Furthermore, we show that the phenomenon of phenotypic epistasis, a hallmark of genetic interaction, holds at the molecular systems level. We also demonstrate that transcriptomes contain sufficient information, under certain circumstances, to order genes in a pathway using only single mutants. Finally, we were able to identify genes that appear to be downstream of *egl-9* and *vhl-1*, but do not appear to be targets of *hif-1*. Using a single set of genome-wide measurements, we were able to observe and quantitatively assess significant fraction of the known transcriptional effects of *hif-1* in *C. elegans*. A complete version of the analysis, with ample documentation, is available at `https://wormlabcaltech.github.io/mprsq`.

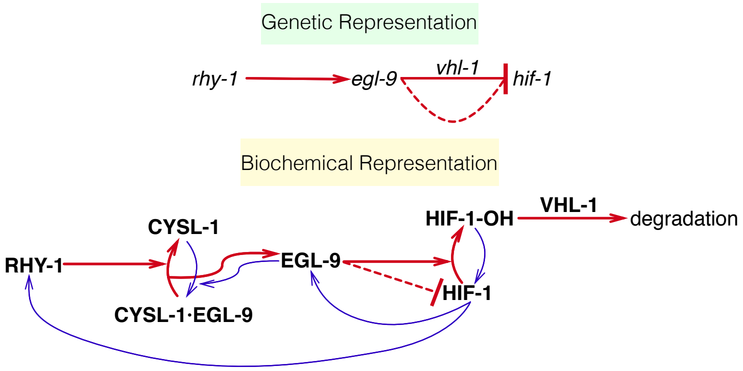


Figure 2.1: Genetic and biochemical representation of the hypoxia pathway in *C. elegans*. Red arrows are arrows that lead to inhibition of HIF-1, and blue arrows are arrows that increase HIF-1 activity or are the result of HIF-1 activity. EGL-9 is known to exert *vhl-1*-dependent and independent repression on HIF-1 as shown in the genetic diagram. The *vhl-1*-independent repression of HIF-1 by EGL-9 is denoted by a dashed line and is not dependent on the hydroxylating activity of EGL-9. Technically, RHY-1 inhibits CYSL-1, which in turn inhibits EGL-9, but this interaction was abbreviated in the genetic diagram for clarity.

Results

The hypoxia pathway controls thousands of genes in *C. elegans*

We selected four single mutants within the hypoxia pathway for expression profiling: *egl-9(lf)* (*sa307*), *rhy-1(lf)* (*ok1402*), *vhl-1(lf)* (*ok161*), *hif-1(lf)* (*ia4*). We also sequenced the transcriptomes of two double mutants, *egl-9(lf);vhl-1(lf)* (*sa307*, *ok161*) and *egl-9(lf) hif-1(lf)* (*sa307*, *ia4*) as well as wild-type N2 as a control sample. Each genotype was sequenced in triplicate at a depth of 15 million reads. We performed whole-organism RNA-seq of these mutants at a moderate sequencing depth (∼ 7 million mapped reads for each individual replicate) under normoxic conditions. For single samples, we identified around 22,000 different isoforms per sample, which allowed us to measure differential expression of 18,344 isoforms across all replicates and genotypes (this constitutes ∼70% of the protein coding isoforms in *C. elegans*). We also included in our analysis a *fog-2(lf)* (*q71*) mutant which we have previously studied (Angeles-Albores, Leighton, et al., 2016), because *fog-2* is not reported to interact with the hypoxia pathway. We analyzed our data using a general linear model on logarithm-transformed counts. Changes in gene expression are reflected in the regression coefficient, β which is specific to each isoform within a genotype. Statistical significance is achieved when the q-values for each β (p-values adjusted for multiple testing) are less than 0.1. Genes that are significantly altered between wild-type and a given mutant have β values that are statistically significantly different from 0. These coefficients are not equal to the average log-fold change per gene, although they are loosely related to this quantity. Larger magnitudes of β correspond to larger perturbations. These coefficients can be used to study the RNA-seq data in question.

In spite of the moderate sequencing depth, transcriptome profiling of the hypoxia pathway revealed that this pathway controls thousands of genes in *C. elegans*. The *egl-9(lf)* transcriptome showed differential expression of 1,806 genes. Similarly, 2,103 genes were differentially expressed in *rhy-1(lf)* mutants. The *vhl-1(lf)* transcriptome showed considerably fewer differentially expressed genes (689), possibly because it is a weaker controller of *hif-1(lf)* than *egl-9(lf)* (Shao, Zhang, and Powell-Coffman, 2009). The *egl-9(lf)*;*vhl-1(lf)* double mutant transcriptome showed 2,376 differentially expressed genes. The *hif-1(lf)* mutant also showed a transcriptomic phenotype involving 546 genes. The *egl-9(lf) hif-1(lf)* double mutant showed a similar number of genes with altered expression (404 genes, see Table 2.1).

Genotype	Differentially Expressed Genes
egl-9(lf)	1,806
rhy-1(lf)	2,103
vhl-1(lf)	689
egl-9(lf);vhl-1(lf)	2,376
egl-9(lf) hif-1(lf)	404
fog-2(lf)	2090

Table 2.1: Number of differentially expressed genes in each mutant.

Principal Component Analysis visualizes epistatic relationships between genotypes

Principal Component Analysis (PCA) is a well-known technique in bioinformatics that is used to identify relationships between high dimensional data points (Yeung and Ruzzo, 2001) We performed PCA on our data to examine whether each genotype clustered in a biologically relevant manner. PCA identifies the vector that can explain most of the variation in the data;this is called the first PCA dimension. Using PCA, one can identify the first n dimensions that can explain more than 95% of the variation in the data. Sample clustering in these n dimensions often indicates biological relationships between the data, although interpreting PCA dimensions can be difficult.

After applying PCA, we expected *hif-1(lf)* to cluster near *egl-9(lf) hif-1(lf)*, because *hif-1(lf)* exhibits no phenotypic defects under normoxic conditions, in contrast to *egl-9(lf)*, which exhibits an egg-laying (Egl) phenotype in the same environment. In *egl-9(lf) hif-1(lf)* mutants the Egl phenotype of *egl-9(lf)* mutants is suppressed and instead the grossly wild-type phenotype of *hif-1(lf)* is observed. On the other hand, we expected *egl-9(lf)*, *rhy-1(lf)*, *vhl-1(lf)* and *egl-9(lf);vhl-1(lf)* to form a separate cluster since each of these genotypes is Egl and has a constitutive hypoxic response. Finally, we included as a negative control a *fog-2(lf)* mutant we have analyzed previously (Angeles-Albores, Leighton, et al., 2016). This data was obtained at a different time from the other genotypes, so we included a batch-normalization term in our equations to account for this. Since *fog-2* has not been described to interact with the hypoxia pathway, we expected that it should appear far away from either cluster.

The first dimension of the PCA analysis was able to discriminate between mutants that have constitutive high levels of HIF-1 and mutants that have no HIF-1, whereas the second dimension was able to discriminate between mutants within the hypoxia pathway and outside the hypoxia pathway (see Fig. 2.2). Therefore expression profiling measures enough signal to cluster genes in a meaningful manner in complex metazoans.

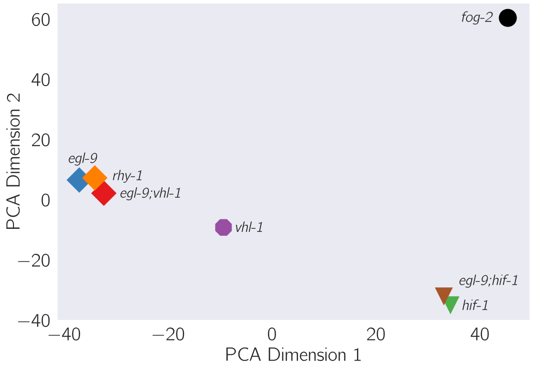


Figure 2.2: Principal component analysis of various *C. elegans* mutants. Genotypes that have an activated hypoxia response (*i.e*, *egl-9(lf)*, *vhl-1(lf)*, and *rhy-1(lf)*) cluster far from *hif-1(lf)*. *hif-1(lf)* clusters with the suppressed *egl-9(lf) hif-1(lf)* double mutant. The *fog-2(lf)* transcriptome, used as an outgroup, is far away from either cluster.

Reconstruction of the hypoxia pathway from first genetic principles

Having shown that the signal in the mutants we selected was sufficient to cluster mutants using the values of the regression coefficients β, we set out to reconstruct the hypoxia pathway from genetic first principles. In general, to reconstruct a pathway, we must first assess whether two genes act on the same phenotype. If they do not act on the same phenotype (the set of commonly differentially regulated genes between two mutants is empty), these mutants are independent. If they are not independent, then two mutants have a shared transcriptomic phenotype (STP)—a set of genes or isoforms that are differentially expressed in both mutants, without taking into account what direction they change in. In this case, we must measure whether these genes act additively or epistatically on the measured phenotype; if there is epistasis we must measure whether it is positive or negative, in order to assess whether the epistatic relationship is a genetic suppression or a synthetic interaction.

Genes in the hypoxia mutant act on the same transcriptional phenotype

We observed that all the hypoxia mutants had significant shared transcriptomic phenotypes (fraction of the transcriptomes that was shared between mutants ranged from a minimum of 6.8% shared between *hif-1(lf)* and *egl-9(lf);vhl-1(lf)* to a maximum of 31% shared genes between *egl-9(lf)* and *egl-9(lf);vhl-1(lf)*). For comparison, we also analyzed a previously published *fog-2(lf)* transcriptome (Angeles-Albores, Leighton, et al., 2016). The *fog-2* gene is involved in masculinization of the *C. elegans* germline, which enables sperm formation, and is not known to be involved in the hypoxia pathway. The hypoxia pathway mutants and the *fog-2(lf)* mutant also showed shared transcriptomic phenotypes (3.6%–12% genes), but correlations between expression level changes were considerably weaker (see below), suggesting that there is minor cross-talk between these pathways.

We wanted to know whether it was informative to look at quantitative agreement within STPs. For each mutant pair, we rank-transformed the regression coefficients β of each isoform within the STP, and calculated lines of best fit using Bayesian regression with a Student-T distribution to mitigate noise from outliers and plotted the results in a rank plot (see Fig 2.3). For transcriptomes associated with the hypoxia pathway, we found that these correlations tended to have values higher than 0.9 with a tight distribution around the line of best fit. The correlations for mutants from the hypoxia pathway with the *fog-2(lf)* mutant were considerably weaker, with magnitudes between 0.6–0.85 and greater variance around the line of

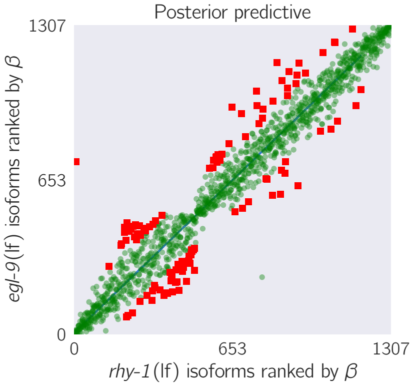


Figure 2.3: Strong transcriptional correlations can be identified between genes that share a positive regulatory connection. We took the *egl-9(lf)* and the *rhy-1(lf)* transcriptomes, identified differentially expressed genes common to both transcriptomes and ranked each gene according to its differential expression coefficient β. We plotted the rank of each gene in *rhy-1(lf)* versus the rank of the same gene in the *egl-9(lf)* transcriptome. The result is an almost perfect correlation. Green, transparent large points mark inliers to the primary regressions (blue lines); red squares mark outliers to the primary regressions.

best fit. Although *hif-1* is known to be genetically repressed by *egl-9*, *rhy-1* and *vhl-1* (Epstein et al., 2001; Shen, Shao, and Powell-Coffman, 2006), all the correlations between mutants of these genes and *hif-1(lf)* were positive.

After we calculated the pairwise correlation within each STP, we weighted the result of each regression by the number of isoforms within the STP and divided by the total number of differentially expressed isoforms present in the two mutant transcriptomes that contributed to that specific STP, $N_{\text{overlap}}/N_{\text{g}_1\cup\text{g}2}$. The weighted regressions recapitulated a module network (see Fig. 2.4). We identified a strong positive interaction between *egl-9(lf)* and *rhy-1(lf)*. The magnitude of this weighted correlation derives from the magnitude of the transcriptomes for these mutants (1,806 and 2,103 differentially expressed genes respectively) and the overlap between both genes was extensive, which makes the weighting factor considerably larger than other pairs. The weak correlation between *hif-1(lf)* and *egl-9(lf)* results from the small size of the *hif-1(lf)* transcriptome and the small overlap between the transcriptomes.

The fine-grained nature of transcriptional phenotypes means that these weighted correlations between transcriptomes of single mutants are predictive of genetic interaction.

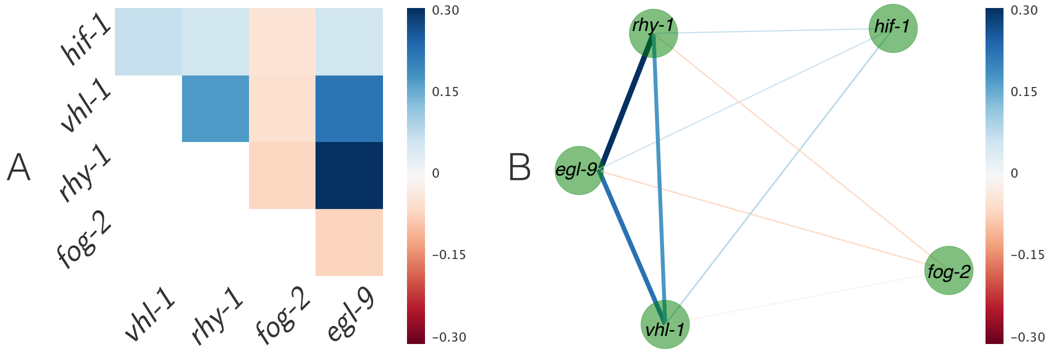


Figure 2.4: **A**. Heatmap showing pairwise regression values between all single mutants. **B**. Correlation network drawn from **A**. Edge width is proportional to the logarithm of the magnitude of the weighted correlation between two nodes divided by absolute value of the weighted correlation value of smallest magnitude. Edges are also colored according to the heatmap in **A**. Inhibitors of *hif-1* are tightly correlated and form a control module; *hif-1* is positively correlated to its inhibitors, albeit weakly; and *fog-2*, a gene that is not reported to interact with the hypoxia pathway, has the smallest, negative correlation to any gene.

A quality check of the transcriptomic data reveals excellent agreement with the literature

One way to establish whether genes are acting additively or epistatically to each other is to perform qPCR of a reporter gene in the single and double mutants. This approach was used to successfully map the relationships within the hypoxia pathway (see, for example (Shao, Zhang, and Powell-Coffman, 2009; Shen, Shao, and Powell-Coffman, 2006)). A commonly used hypoxia reporter gene is *nhr-57*, which is known to exhibit a several-fold increase in mRNA expression when HIF-1 accumulates (Shen, Shao, and Powell-Coffman, 2006; Shen, Nettleton, et al., 2005; Park et al., 2012). Likewise, increased HIF-1 fucntion is known to cause increased of *rhy-1* and *egl-9* (Powell-Coffman, 2010).

We can selectively look at the expression of a few genes at a time. Therefore, we queried the changes in expression of *rhy-1*, *egl-9*, and *nhr-57*. We included the nuclear laminin gene *lam-3* as a representative negative control not believed to be responsive to alterations in the hypoxia pathway. *nhr-57* was upregulated in *egl-9(lf)*, *rhy-1(lf)* and *vhl-1(lf)*, but remains unchanged in *hif-1(lf)*. *egl-9(lf);vhl-1(lf)* had an expression level similar to *egl-9(lf)*; whereas the *egl-9(lf) hif-1(lf)* mutant showed wild-type levels of the reporter expression, as reported previously (Shen, Shao, and Powell-Coffman, 2006) (see Fig. 2.5).

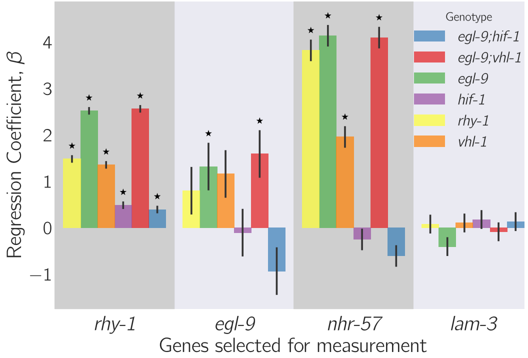


Figure 2.5: **Top**: Observed β values of select genes. We selected four genes (*rhy-1*, *egl-9*, *nhr-57* and *lam-3*, shown on the x-axis) and plotted their regression coefficients, β, as measured for every genotype (represented by one of six colors) to study the epistatic relationships between each gene. Asterisks above a bar represent a regression coefficient statistically significantly different from 0 ($q < 10^{-1}$) relative to a wild-type control. Error bars show standard error of the mean value of β. *nhr-57* is an expression reporter that has been used previously to identify *hif-1* regulators (Shen, Shao, and Powell-Coffman, 2006; Shao, Zhang, and Powell-Coffman, 2009). *lam-3* is shown here as a negative control that should not be altered by mutations in this pathway. We measured modest increases in the levels of *rhy-1* mRNA when *hif-1(lf)* is knocked out.

We observed changes in *rhy-1(lf)* expression consistent with previous literature (Shen, Shao, and Powell-Coffman, 2006) when HIF-1 accumulates. We also observed increases in *egl-9* expression in *egl-9(lf)*. *egl-9* is known as a hypoxia responsive gene (Powell-Coffman, 2010). Although changes in *egl-9* expression were not statistically significantly different from the wild-type in *rhy-1(lf)* and *vhl-1(lf)* mutants, the mRNA levels of *egl-9* still trended towards increased expression in these genotypes. As with *nhr-57*, *egl-9* and *rhy-1* expression were wild-type in *egl-9(lf) hif-1(lf)* and *egl-9(lf);vhl-1(lf)* mutant showed expression phenotypes identical to *egl-9(lf)*. This dataset also showed that knockout of *hif-1* resulted in a modest increase in the levels of *rhy-1*. This suggests that *hif-1*, in addition to being a positive regulator of *rhy-1*, also inhibits it, which constitutes a novel observation. Using a single reporter we would have been able to reconstruct an important fraction of the genetic relationships between the genes in the hypoxia pathway—-but would likely fail to observe yet other genetic interactions, such as the evidence for *hif-1* negatively regulating *rhy-1* transcript levels.

Transcriptome-wide epistasis

Ideally, any measurement of transcriptome-wide epistasis should conform to certain expectations. First, it should make use of the regression coefficients of as many genes as possible. Second, it should be summarizable in a single, well-defined number. Third, it should have an intuitive behavior, such that special values of the statistic should each have an unambiguous interpretation.

One way of displaying transcriptome-wide epistasis is to plot transcriptome data onto an epistasis plot (see Fig 2.6). In an epistasis plot, the X-axis represents the expected expression of a double mutant a^-b^- if a and b interact additively. In other words, each individual isoform's x-coordinate is the sum of the regression coefficients from the single mutants a^- and b^-. The Y-axis represents the deviations from the additive (null) model, and can be calculated as the difference between the observed regression coefficient and the predicted regression coefficient. Only genes that are differentially expressed in all three genotypes are plotted. Assuming that the two genes interact via a simple phenotype (for example, if both genes affect a transcription factor that generates the entire transcriptome), these plots will generate specific patterns that can be described through linear regressions. The slope of these lines, $s_{a,b}$, is the transcriptome-wide epistasis coefficient.

Epistasis plots can be understood intuitively for simple cases of genetic interactions. If two genes act additively on the same set of differentially expressed isoforms then all the plotted points will fall along the line $y = 0$. If two genes interact in an unbranched pathway, then a^- and b^- should have identical phenotypes for a^-, b^- and a^-b^-, if all the genotypes are homozygous for genetic null alleles (L. S. Huang and Paul W Sternberg, 2006). It follows that the data points should fall along a line with slope equal to $-\frac{1}{2}$. On the other hand, in the limit of complete inhibition of a by b, the plots should show a line of best fit with slope equal to -1[1]. Genes that interact synthetically (*i.e.*, through an OR-gate) will fall along lines with slopes > 0. When there is epistasis of one gene over another, the points will fall along a line of best fit with slope $s_{ab=b}$ or $s_{ab=a}$. This slope must be determined from the single-mutant data. From this information, we can use the single mutant data to predict the distribution of slopes that results for each case stated above, as well as for each epistatic combination ($a^-b^- = a^-$ or $a^-b^- = b^-$). The transcriptome-wide epistasis coefficient ($s_{a,b}$), emerges as a powerful way to quantify epistasis because

[1] Specifically, this follows from assuming that b^- is wild-type under the conditions assayed; and $a^-b^- = b^-$ = wild-type

it integrates information from many different genes or isoforms into a single number (see Fig. 2.6).

In our experiment, we studied two double mutants, *egl-9(lf) hif-1(lf)* and *egl-9(lf);vhl-1(lf)*. We wanted to understand how well an epistasis analysis based on transcriptome-wide coefficients agreed with the epistasis results reported in the literature, which were based on qPCR of single genes. Therefore, we performed orthogonal distance regression on the two gene combinations we studied (*egl-9* and *vhl-1*; and *egl-9* and *hif-1*) to determine the epistasis coefficient for each gene pair. We also generated models for the special cases mentioned above (additivity, $a^-b^- = a^-$, strong suppression, etc. . .) using the single mutant data. For every simulation, as well as for the observed data, we used bootstraps to generate probability distributions of the epistasis coefficients.

When we compared the predictions for the transcriptome-wide epistasis coefficient, $s_{egl-9,vhl-1}$ under different assumptions with the observed slope (-0.42). We observed that the predicted slope matched the simulated slope for the case where *egl-9* is epistatic over *vhl-1* (*egl-9(lf) = egl-9(lf);vhl-1(lf)*, see Fig. 2.6) and did not overlap with any other prediction. Next, we predicted the distribution of $s_{egl-9,hif-1}$ for different pathways and contrasted with the observed slope. In this case, we saw that the uncertainty in the observed coefficient overlapped significantly with the strong suppression model, where EGL-9 strongly suppresses HIF-1, and also with the model where *hif-1(lf) = egl-9(lf) hif-1(lf)*. In this case, both models are reasonable—HIF-1 is strongly suppressed by EGL-9, and we know from previous literature that the epistatic relationship, *hif-1(lf) = egl-9(lf) hif-1(lf)*, is true for these mutants. In fact, as the repression of HIF-1 by EGL-9 becomes stronger, the epistatic model should converge on the limit of strong repression (see Epistasis).

Another way to test which model best explains the epistatic relationship between *egl-9* and *vhl-1* is to use Bayesian model selection to calculate an odds ratio between two models to explain the observed data. Models can be placed into two categories: parameter-free and fit. Parameter free models are ‘simpler’ because their parameter space is smaller (0 parameters) than the fit models (n parameters). By Occam’s razor, simpler models should be preferred to more complicated models. However, simple models suffer from the drawback that systematic deviations from them cannot be explained or accomodated, whereas more complicated models can alter the fit values to maximize their explanatory power. In this sense, more complicated models should be preferred when the data shows systematic deviations from the simple model.

Odds-ratio selection gives us a way to quantify the trade-off between simplicity and explanatory power.

We reasoned that comparing a fit model ($y = \alpha \cdot x$, where α is the slope of best fit) against a parameter-free model ($y = \gamma \cdot x$, where γ is a single number) constituted a conservative approach towards selecting which theoretical model (if any) best explained the data. In particular, this approach will tend to strongly favor the line of best fit over simpler model for all but very small, non-systematic deviations. We decided that we would reject the theoretical models only if the line of best-fit was 10^3 times more likely than the theoretical models (odds ratio, OR $> 10^3$). Comparing the odds-ratio between the line of best fit and the different pathway models for *egl-9* and *vhl-1* showed similar results to the simulation. Only the theoretical model *egl-9(lf)* = *egl-9(lf);vhl-1(lf)* could not be rejected (OR = 0.46), whereas all other models were significantly less likely than the line of best fit (OR $> 10^{44}$). Therefore, *egl-9* is epistatic to *vhl-1*. Moreover, since $s_{egl-9,vhl-1}$ is strictly between and not equal to 0 and -0.5, we conclude that *egl-9* acts on its transcriptomic phenotype in *vhl-1*-dependent and independent manners. A branched pathway that can lead to epistasis coefficients in this range is a pathway where *egl-9* interacts with its transcriptomic phenotype via branches that have the same valence (both positive or both negative) (Shao, Zhang, and Powell-Coffman, 2009). When we performed a similar analysis to establish the epistatic relationship between *egl-9* and *hif-1*, we observed that the best alternative to a free-fit model was a model where *hif-1* is epistatic over *egl-9* (OR= 2551), but the free-fit model was still preferred. All other models were strongly rejected (OR $> 10^{25}$).

Epistasis can be predicted

Given our success in measuring epistasis coefficients, we wanted to know whether we could predict the epistasis coefficient between *egl-9* and *vhl-1* in the absence of the *egl-9(lf)* genotype. Since RHY-1 indirectly activates EGL-9, the *rhy-1(lf)* transcriptome should contain more or less equivalent information to the *egl-9(lf)* transcriptome. Therefore, we generated predictions of the epistasis coefficient between *egl-9* and *vhl-1* by substituting in the *rhy-1(lf)* data. We predicted $s_{rhy-1,vhl-1} = -0.45$. Similarly, we used the *egl-9(lf);vhl-1(lf)* double mutant to measure the epistasis coefficient while replacing the *egl-9(lf)* dataset with the *rhy-1(lf)* dataset. We found that the epistasis coefficient using this substitution was -0.40. This coefficient was different from -0.50 (OR $> 10^{62}$), reflecting the same qualitative conclusion that the

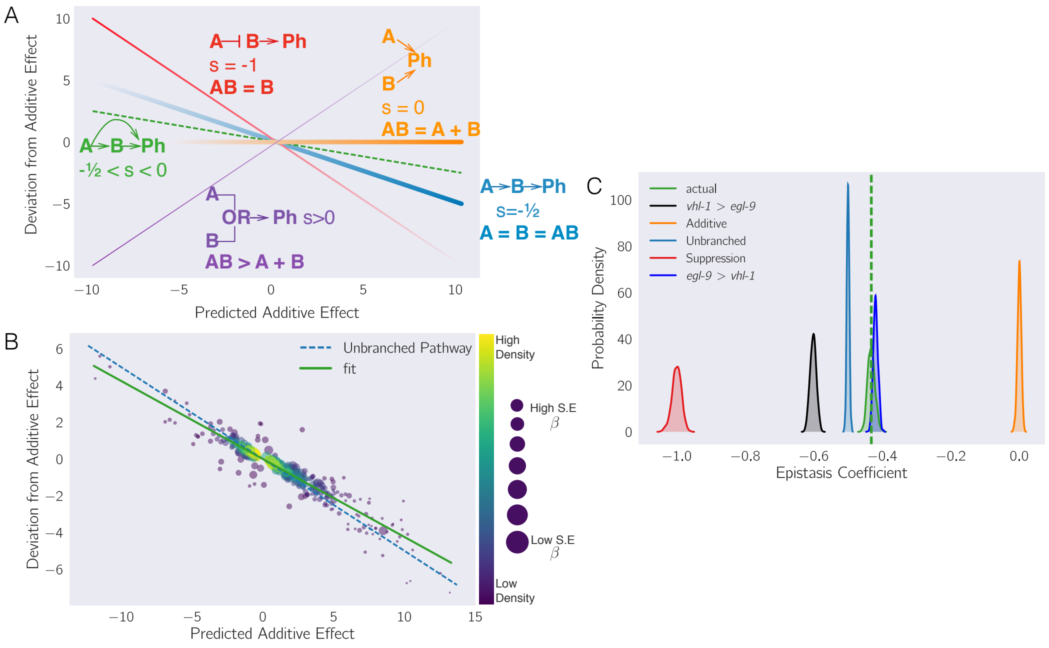


Figure 2.6: (**A**) Schematic diagram of an epistasis plot. The X-axis on an epistasis plot is the expected coefficient for a double mutant under an additive model (null model). The Y-axis plots deviations from this model. Double mutants that deviate in a systematic manner from the null model exhibit transcriptome-wide epistasis (s). To measure s, we perform a linear regression on the data. The slope of the line of best fit is s. This coefficient is related to genetic architectures. Genes that act additively on a phenotype **(Ph)** will have $s = 0$ (orange line); whereas genes that act along an unbranched pathway will have $s = -1/2$ (blue line). Strong repression is reflected by $s = -1$ (red line). Cases where $s > 0$ correspond to synthetic interactions (purple line), and in the limit as $s \longrightarrow \infty$, the synthetic interaction must be an OR-gate. Cases where $0 < s < -1/2$ correspond to circuits that have multiple positive branches; whereas cases where $-1/2 < s < -1$ correspond to cases where the branches have different valence. Cases where $s < -1$ represent inhibitory branches. (**B**) Epistasis plot showing that the *egl-9(lf);vhl-1(lf)* transcriptome deviates significantly from a null additive. Points are colored qualitatively according to density (purple—low, yellow—high) and size is inversely proportional to the standard error (S.E.) of the y-axis (larger points, higher accuracy). The purple line is the line of best fit from an orthogonal distance regression. (**C**) Comparison of simulated epistatic coefficients against the observed coefficient. Green curve shows the bootstrapped observed transcriptome-wide epistasis coefficient for *egl-9* and *vhl-1*. Dashed green line shows the mean value of the data. Using the single mutants, we simulated coefficient distributions for a linear model (light blue, centered at −0.5); an additive model (orange, centered at 0); a model where either *egl-9* or *vhl-1* masks the other phenotype (dark blue and black, respectively) and a complete suppression model (red, centered at −1). The observed coefficient overlaps the predicted epistasis curve for *egl-9(lf);vhl-1(lf)* = *egl-9(lf)* (green and dark blue).

hypoxia pathway is branched. In conclusion, we were able to obtain a quantitatively close prediction of the epistasis coefficient for two mutants using the transcriptome of a related, upstream mutant. Finally, we showed that in the absence of a single mutant, an upstream locus can under some circumstances be used to estimate epistasis between two genes.

Transcriptomic decorrelation can be used to infer functional distance

So far, we have shown that RNA-seq can accurately measure genetic interactions. However, genetic interactions are far removed from biochemical interactions: Genetic interactions do not require two gene products to interact physically, nor even to be physically close to each other. RNA-seq cannot measure physical interactions between genes, but we wondered whether expression profiling contains sufficient information to order genes along a pathway.

Single genes are often regulated by multiple independent sources. The connection between two nodes can in theory be characterized by the strength of the edges connecting them (the thickness of the edge); the sources that regulate both nodes (the fraction of inputs common to both nodes); and the genes that are regulated by both nodes (the fraction of outputs that are common to both nodes). In other words, we expected that expression profiles associated with a pathway would respond quantitatively to quantitative changes in activity of the pathway. Targeting a pathway at multiple points would lead to expression profile divergence as we compare nodes that are separated by more degrees of freedom, reflecting the flux in information between them.

We investigated the possibility that transcriptomic signals do in fact contain relevant information about the degrees of separation by weighting the robust Bayesian regression between each pair of genotypes by the size of the shared transcriptomic phenotype of each pair divided by the total number of isoforms differentially expressed in either mutant ($N_{\text{Intersection}}/N_{\text{Union}}$). We plotted the weighted correlation of each gene pair, ordered by increasing functional distance (see Fig. 2.7). In every case, we see that the weighted correlation decreases monotonically due mainly, but not exclusively, to a smaller STP. We believe that this result is not due to random noise or insufficiently deep sequencing. Instead, we propose a framework in which every gene is regulated by multiple different molecular species, which induces progressive decorrelation. This decorrelation in turn has two consequences. First, decorrelation within a pathway implies that two nodes may be almost independent

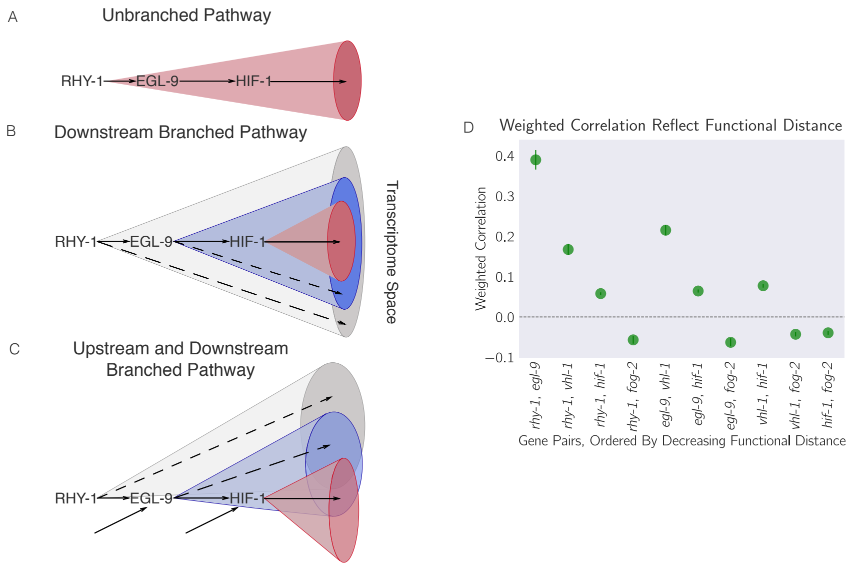


Figure 2.7: Theoretically, transcriptomes can be used to order genes in a pathway under certain assumptions. Arrows in the diagrams above are intended to show the direction of flow, and do not indicate valence. **A**. A linear pathway in which *rhy-1* is the only gene controlling *egl-9*, which in turn controls *hif-1* does not contain information to infer the order between genes. **B**. If *rhy-1* and *egl-9* have transcriptomic effects that are separable from *hif-1*, then the *rhy-1* transcriptome should contain contributions from *egl-9*, *hif-1* and *egl-9*- and *hif-1*-independent pathways. This pathway contains enough information to infer order. **C**. If a pathway is branched both upstream and downstream, transcriptomes will show even faster decorrelation. Nodes that are separated by many edges may begin to behave almost independently of each other with marginal transcriptomic overlap or correlation. **D**. The hypoxia pathway can be ordered. We hypothesize the rapid decay in correlation is due to a mixture of upstream and downstream branching that happens along this pathway. Bars show the standard error of the weighted coefficient from the Monte Carlo Markov Chain computations.

of each other if the functional distance between them is large. Second, it may be possible to use decorrelation dynamics to infer gene order in a branching pathway, as we have done with the hypoxia pathway.

The circuit topology of the hypoxia pathway explains patterns in the data

We noticed that while some of the rank plots contained a clear positive correlation (see Fig. 2.3), other rank plots showed a discernible cross-pattern (see Fig. 2.8). In particular, this cross-pattern emerged between *vhl-1(lf)* and *rhy-1(lf)* or between *vhl-1(lf)* and *egl-9(lf)*, even though genetically *vhl-1*, *rhy-1* and *egl-9* are all inhibitors of *hif-1(lf)*. Such cross-patterns could be indicative of feedback loops or other complex interaction patterns.

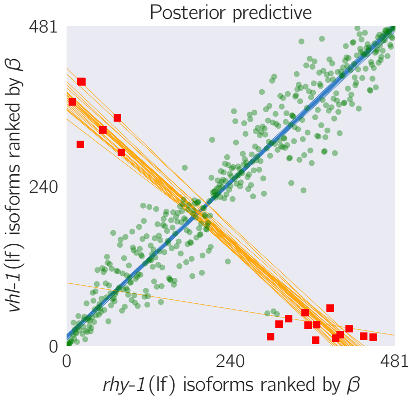


Figure 2.8: A feedback loop can generate transcriptomes that are both correlated and anti-correlated. The *vhl-1(lf)*/*rhy-1(lf)* STP shows a cross-pattern. Green large points are inliers to the first regression. Red squares are outliers to the first regression. Only the red small points were used for the secondary regression. Blue lines are representative samples of the primary bootstrapped regression lines. Orange lines are representative samples of the secondary bootstrapped regression lines.

If the above is correct, then it should be possible to identify *egl-9*-independent, *rhy-1(lf)*-dependent target genes in a logically consistent way. One erroneous way to identify these targets is via subtractive logic. Using subtractive logic, we would identify genes that are differentially expressed in *rhy-1(lf)* mutants but not in *egl-9(lf)* mutants. Such a gene set would consist of almost 700 genes. One major drawback of subtractive logic is that it cannot be applied when feedback loops exist between the genes in question. Another problem is that the set of identified genes are statistically indistinguishable from false positive and false negative hits because they have no distinguishing property beyond the condition that they should be differentially expressed in one mutant but not the other. In fact, this is exactly the behavior expected of false-positive or false-negative hits—presence in one, but not multiple, mutants. We need to consider the relationship between two genes before we can begin to identify targets which expression is dependent on one gene and independent of the other.

rhy-1 and *egl-9* share a well-defined relationship. RHY-1 inhibits CYSL-1, which in turn inhibits EGL-9 (Ma et al., 2012). Therefore, loss of RHY-1 leads to inactivation of EGL-9, which leads to increase in the cellular levels of HIF-1. HIF-1 in turn causes the mRNA levels of *rhy-1* and *egl-9* to increase, as they are involved in the *hif-1*-dependent hypoxia response. However, since *rhy-1* has been mutated, the observed transcriptome is RHY-1 ‘null’; EGL-9 ‘null’; HIF-1 ‘on’. The situation is similar for *egl-9(lf)*, except that RHY-1 is not inactive, and therefore the observed transcriptome is the result of RHY-1 ‘up’; EGL-9 ‘null’; and HIF-1 ‘on’. From this pattern, we conclude that the *egl-9(lf)* and *rhy-1(lf)* transcriptomes should exhibit a cross-pattern when plotted against each other: The positive arm of the cross is the result of the EGL-9 ‘null’; HIF-1 ‘on’ dynamics; and the negative arm reflects the different direction of RHY-1 activity between transcriptomes. No negative arm is visible (with the exception of two outliers, which are annotated as pseudogenes in WormBase). Therefore, in this dataset we do not find genes that have *egl-9* independent, *rhy-1*-dependent expression patterns.

We also identified a main hypoxia response induced by disinhibiting *hif-1* (355 genes) by identifying genes that were commonly up-regulated amongst *egl-9(lf)*, *rhy-1(lf)* and *vhl-1(lf)* mutants. Although the hypoxic response is likely to involve between three and seven times more genes (assuming the *rhy-1(lf)* transcriptome reflects the maximal hypoxic response), this is a conservative estimate that minimizes false positive results, since these changes were identified in four different genotypes with three replicates each. This response included five transcription factors (*W02D7.6*, *nhr-57*, *ztf-18*, *nhr-135* and *dmd-9*). The full list of genes associated with the hypoxia response can be found in the Supplementary Table 1.

hif-1-independent effects of *egl-9* have been reported previously (Park et al., 2012), which led us to question whether we could identify similar effects in our dataset. We have observed that *hif-1(lf)* displays a modest increase in the transcription of *rhy-1*, from which we speculated that EGL-9 would have increased activity in the *hif-1(lf)* mutant compared to the wild-type. Therefore, we searched for genes that were regulated in an opposite manner between *hif-1(lf)* and *egl-9(lf) hif-1(lf)*, and that were regulated in the same direction between all *egl-9(lf)* genotypes. We did not find any genes that met these conditions.

We also searched for genes with *hif-1* independent, *vhl-1*-dependent gene expression and found 45 genes, which can be found in the Supplementary Table 2. Finally, we searched for candidates directly regulated by *hif-1*. Initially, we searched for genes that had were significantly altered in *hif-1(lf)* genotypes in one direction, but altered in the opposite direction in mutants that activate the HIF-1 response. Only two genes (*R08E5.3*, and *nit-1*) met these conditions. This could reflect the fact that HIF-1 exists at very low levels in *C. elegans*, so loss of function mutations in *hif-1* might only have mild effects on its transcriptional targets. We reasoned that genes that are overexpressed in mutants that induce the HIF-1 response would be enriched for genes that are direct candidates. We found 195 genes which have consistently increased expression in mutants with a constitutive hypoxic response. These genes can be found in the Supplementary Table 3.

Enrichment analysis of the hypoxia response

To validate that our transcriptomes were correct, and to understand how functionalities may vary between them, we subjected each decoupled response to enrichment analysis using the WormBase Enrichment Suite (Angeles-Albores, N. Lee, et al., 2016; Angeles-Albores, R. Y. Lee, et al., 2017).

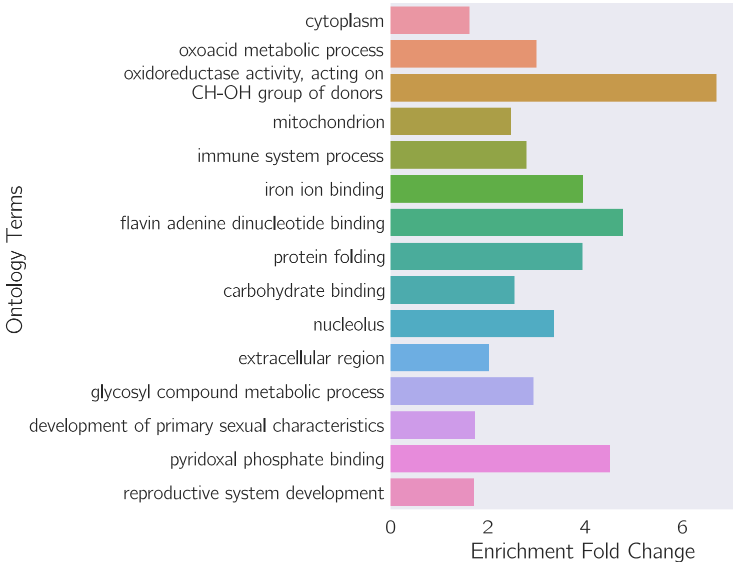


Figure 2.9: Gene ontology enrichment analysis of genes associated with the main hypoxia response. A number of terms reflecting catabolism and bioenergetics are enriched.

We used gene ontology enrichment analysis (GEA) on the main hypoxia response program. This showed that the terms 'oxoacid metabolic process' ($q < 10^{-4}$, 3.0 fold-change, 24 genes), 'iron ion binding' ($q < 10^{-2}$, 3.8 fold-change, 10 genes), and 'immune system process' ($q < 10^{-3}$, 2.9 fold-change, 20 genes) were significantly enriched. GEA also showed enrichment of the term 'mitochondrion' ($q < 10^{-3}$, 2.5 fold-change, 29 genes) (see Fig. 2.9). Indeed, *hif-1(lf)* has been implicated in all of these biological and molecular functions (Luhachack et al., 2012; Ackerman and Gems, 2012; Romney et al., 2011; Gregg L. Semenza, 2011). As benchmark on the quality of our data, we selected a set of 22 genes known to be responsive to HIF-1 levels from the literature and asked whether these genes were present in our hypoxia response list. We found 8/22 known genes, which constitutes a statistically significant result ($p < 10^{10}$). The small number of reporters found in this list probably reflects the conservative nature of our estimates. We studied the *hif-1*-independent, *vhl-1*-dependent gene set using enrichment analysis but no terms were significantly enriched.

Identification of non-classical epistatic interactions

hif-1(lf) has traditionally been viewed as existing in a genetic OFF state under normoxic conditions. However, our dataset indicates that 546 genes show altered expression when *hif-1* function is removed in normoxic conditions. Moreover, we observed positive correlations between *hif-1(lf)* β coefficients and *egl-9(lf)*, *vhl-1(lf)* and *rhy-1(lf)* β coefficients in spite of the negative regulatory relationships between these genes and *hif-1*. Such positive correlations could indicate a different relationship between these genes than has previously been reported, so we attempted to substantiate them through epistasis analyses.

To perform epistasis analyses, we first identified genes that exhibited violations of the canonical genetic model of the hypoxia pathway. To this end, we searched for genes that exhibited different behaviors between *egl-9(lf)* and *vhl-1(lf)*, or between *rhy-1(lf)* and *vhl-1(lf)* (we assume that all results from the *rhy-1(lf)* transcriptome reflect a complete loss of *egl-9* activity). We found 31 that satisfied this condition (see Fig. 2.10, Supplemental Table 4). Additionally, many of these genes exhibited a new kind of epistasis. Namely, *egl-9* was epistatic over *vhl-1*. Identification of a set of genes that have a consistent set of relationships between themselves suggests that we have identified a new aspect of the hypoxia pathway.

To illustrate this, we focused on three genes, *nlp-31*, *ftn-1* and *ftn-2*, which epistasis patterns that we felt reflected the population well. *ftn-1* and *ftn-2* are both described in the literature as genes that are responsive to mutations in the hypoxia pathway. Moreover, these genes have been previously described to have aberrant behaviors (Ackerman and Gems, 2012; Romney et al., 2011), specifically the opposite effects of *egl-9(lf)* and *vhl-1(lf)*. These studies showed that loss of *vhl-1(lf)* decreases expression of *ftn-1* and *ftn-2* using both RNAi and alleles, which allays concerns of strain-specific interference. Moreover, Ackerman and Gems (2012) showed that *vhl-1* is epistatic to *hif-1* for the *ftn-1* expression phenotype, and that loss of HIF-1 is associated with increased expression of *ftn-1* and *ftn-2*. We observed that *hif-1* was epistatic to *egl-9*, and that *egl-9* and *hif-1* both promoted *ftn-1* and *ftn-2* expression.

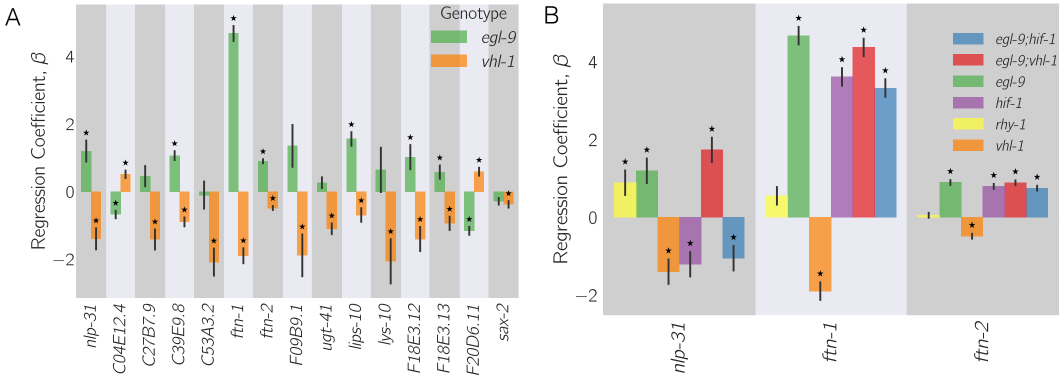


Figure 2.10: **A**. 27 genes in *C. elegans* exhibit non-classical epistasis in the hypoxia pathway, characterized by opposite effects on gene expression, relative to the wild-type, of of the *vhl-1(lf)* compared to *egl-9(lf)* (or *rhy-1(lf)*) mutants. Shown are a random selection of 15 the 27 genes for illustrative purposes. **B**. Representative genes showing that non-canonical epistasis shows a consistent pattern. *vhl-1(lf)* mutants have an opposite effect to *egl-9(lf)*, but *egl-9* remains epistatic to *vhl-1* and loss-of-function mutations in *hif-1* suppress the *egl-9(lf)* phenotype. Asterisks show β values significantly different from 0 relative to wild-type ($q < 10^{-1}$).

Epistasis analysis of *ftn-1* and *ftn-2* expression reveals that *egl-9* is epistatic to *hif-1*; that *vhl-1* has opposite effects to *egl-9*, and that *vhl-1* is epistatic to *egl-9*. Analysis of *nlp-31* reveals similar relationships. *nlp-31* expression is decreased in *hif-1(lf)*, and increased in *egl-9(lf)*. However, *egl-9* is epistatic to *hif-1*. Like *ftn-1* and *ftn-2*, *vhl-1* has the opposite effect to *egl-9*, yet is epistatic to *egl-9*. We propose in the Discussion a model for how HIF-1 might regulate these targets.

HIF-1 in the cellular context

We identified the transcriptional changes associated with bioenergetic pathways in *C. elegans* by extracting from WormBase all genes associated with the tricarboxylic acid (TCA) cycle, the electron transport chain (ETC) and with the *C. elegans* GO term energy reserve. Previous research has described the effects of mitochondrial dysfunction in eliciting the hypoxia response (S. J. Lee, Hwang, and Kenyon, 2010), but transcriptional feedback from HIF-1 into bioenergetic pathways has not been as extensively in *C. elegans*, as in vertebrates (see, for example (G L Semenza et al., 1994; Gregg L. Semenza, 2012)). We also searched for the changes in ribosomal components and the proteasome, as well as for terms relating to immune response (see Fig 2.11).

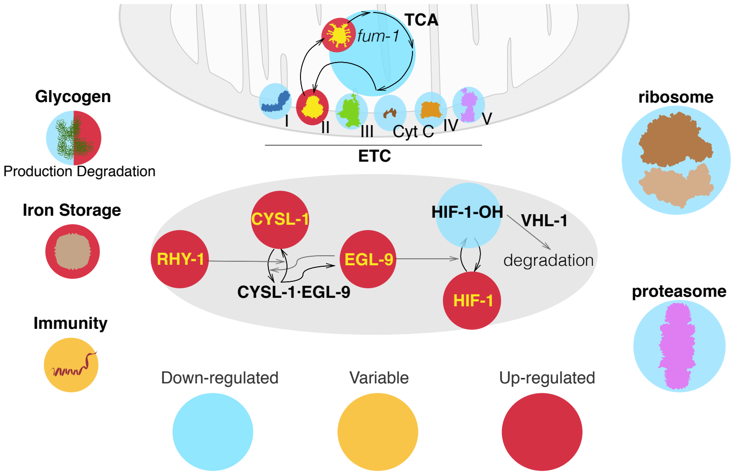


Figure 2.11: A graphic summary of the genome-wide effects of HIF-1 from our RNA-seq data.

Bioenergetic pathways

Our data shows that most of the enzymes involved in the TCA cycle and in the ETC are down-regulated when HIF-1 is induced in agreement with the previous literature (Gregg L. Semenza, 2012). However, the fumarase gene *fum-1* and the mitochondrial complex II stood out as notable exceptions to the trend, as they were up-regulated in every single genotype that causes deployment of the hypoxia response. FUM-1 catalyzes the reaction of fumarate into malate, and complex II catalyzes the reaction of succinate into fumarate. Complex II has been identified as a source of reserve respiratory capacity in neonatal rat cardiomyocytes previously (Pfleger, He, and Abdellatif, 2015). We found two energy reserve genes that were down-regulated by HIF-1. *aagr-1* and *aagr-2*, which are predicted to function in glycogen catabolism (Sikora et al., 2010). Three distinct genes involved in energy reserve were up-regulated. These genes were *ogt-1*, which encodes O-linked GlcNac Transferase gene; *T04A8.7*, encoding an ortholog of human glucosidase, acid beta (GBA); and *T22F3.3*, encoding ortholog of human glycogen phosphorylase isozyme in the muscle (PYGM).

Protein synthesis and degradation

hif-1(lf) is also known to inhibit protein synthesis and translation in varied ways. (Brugarolas et al., 2004). Most reported effects of HIF-1 on the translation machinery are posttranslational, and no reports to date show transcriptional control of the ribosomal machinery in *C. elegans* by HIF-1. We used the WormBase Enrichment

Suite Gene Ontology dictionary (Angeles-Albores, R. Y. Lee, et al., 2017) to extract 143 protein-coding genes annotated as 'structural constituents of the ribosome' and we queried whether they were differentially expressed in our mutants. *egl-9(lf)*, *vhl-1(lf)*, *rhy-1(lf)* and *egl-9(lf)*;*vhl-1(lf)* showed differential expression of 91 distinct ribosomal constituents (not all constituents were detected in all genotypes). For every one of these genotypes, these genes were always down-regulated. In contrast, *hif-1(lf)* showed up-regulation of a single ribosomal constituent.

Next, we asked whether HIF-1 has any transcriptional effects on the proteasomal constituents; no such effects of HIF-1 on the proteasome have been reported in *C. elegans*. Out of 40 WormBase-annotated proteasomal constituents, we found 31 constituents that were differentially expressed in at least one of the four genotypes that induce a hypoxic response. Every gene we found was down-regulated in at least two out of the four genotypes we studied.

Discussion

The *C. elegans* hypoxia pathway can be reconstructed entirely from RNA-seq data

In this paper, we have shown that whole-organism transcriptomic phenotypes can be used to reconstruct genetic pathways and to discern previously overlooked or uncharacterized genetic interactions. We successfully reconstructed the hypoxia pathway, and inferred order of action (*rhy-1* activates *egl-9*, *egl-9* and *vhl-1* inhibit *hif-1*), and we were able to infer from transcriptome-wide epistasis measurements that *egl-9* exerts *vhl-1*-dependent and independent inhibition on *hif-1*.

HIF-1 and the cellular environment

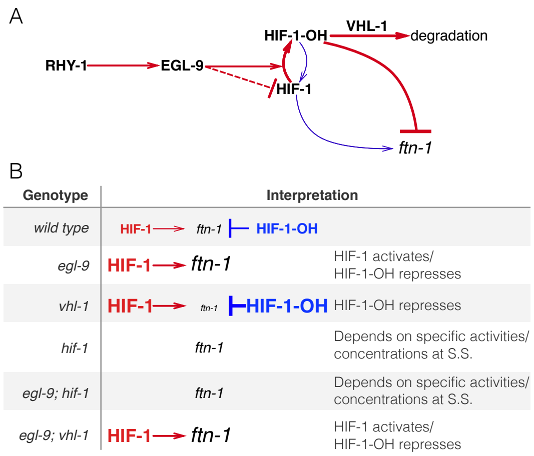
In addition to reconstructing the pathway, our dataset allowed us to observe a wide variety of physiologic changes that occur as a result of the HIF-1-dependent hypoxia response. In particular, we observed down-regulation of most components of the TCA cycle and the mitochondrial electron transport chain with the exceptions of *fum-1* and the mitochondrial complex II. The mitochondrial complex II catalyzes the reaction of succinate into fumarate. In mouse embryonic fibroblasts, fumarate has been shown to antagonize HIF-1 prolyl hydroxylase domain (PHD) enzymes, which are orthologs of EGL-9 (Sudarshan et al., 2009). If the inhibitory role of fumarate on PHD enzymes is conserved in *C. elegans*, upregulation of complex II by HIF-1 during hypoxia may increase intracellular levels of fumarate, which in turn could lead to artificially high levels of HIF-1 even after normoxia resumes. The increase in fumarate produced by the complex could be compensated by increasing expression of *fum-1*. Increased fumarate degradation allows *C. elegans* to maintain plasticity in the hypoxia pathway, keeping the pathway sensitive to oxygen levels.

Interpretation of the non-classical epistasis in the hypoxia pathway

The observation of almost 30 genes that exhibit a specific pattern of non-classical epistasis suggests the existence of previously undescribed aspects of the hypoxia pathway. Some of these non-classical epistases had been observed previously (Ackerman and Gems, 2012; Romney et al., 2011; Luhachack et al., 2012), but no satisfactory mechanism has been proposed to explain this biology. Romney et al. (2011) and Ackerman and Gems (2012) suggest that HIF-1 integrates information on iron concentration in the cell to bind to the *ftn-1* promoter, but could not definitively establish a mechanism. It is unclear why deletion of *hif-1* induces *ftn-1* expression, deletion of *egl-9* also causes induction of *ftn-1* expression, but deletion of *vhl-1* removes this inhibition. Moreover, Luhachack et al. (2012) have previously reported that certain genes important for the *C. elegans* immune response against pathogens reflect similar expression patterns. Their interpretation was that *swan-1*, which encodes a binding partner to EGL-9 (Shao, Zhang, Ye, et al., 2010), is important for modulating HIF-1 activity in some manner. The lack of a conclusive double mutant analysis in this work means the role of SWAN-1 in modulation of HIF-1 activity remains to be demonstrated. Nevertheless, mechanisms that call for additional transcriptional modulators become less likely given the number of genes with different biological functions that exhibit the same pattern.

One way to resolve this problem without invoking additional genes is to consider HIF-1 as a protein with both activating and inhibiting states. In fact, HIF-1 already exists in two states in *C. elegans*: unmodified HIF-1 and HIF-1-hydroxyl (HIF-1-OH). Under this model, HIF-1-hydroxyl antagonizes the effects of HIF-1 for certain genes like *ftn-1* or *nlp-31*. Loss of *vhl-1* stabilizes HIF-1-hydroxyl. A subset of genes that are sensitive to HIF-1-hydroxyl will be inhibited as a result of the increase in the amount of this species, in spite of loss of *vhl-1* function also increasing the level of non-hydroxylated HIF-1. On the other hand, *egl-9(lf)* selectively removes all HIF-1-hydroxyl, stimulating accumulation of HIF-1 and promoting gene activity. Whether deletion of *hif-1(lf)* is overall activating or inhibiting will depend on the relative activity of each protein state under normoxia (see Fig. 2.12).

Multiple lines of circumstantial evidence that HIF-1-hydroxyl plays a role in the



Genotype	Interpretation	
wild type	HIF-1 → *ftn-1* ⊣ HIF-1-OH	
egl-9	HIF-1 → *ftn-1*	HIF-1 activates/ HIF-1-OH represses
vhl-1	HIF-1 → *ftn-1* ⊣ HIF-1-OH	HIF-1-OH represses
hif-1	*ftn-1*	Depends on specific activities/ concentrations at S.S.
egl-9; hif-1	*ftn-1*	Depends on specific activities/ concentrations at S.S.
egl-9; vhl-1	HIF-1 → *ftn-1*	HIF-1 activates/ HIF-1-OH represses

Figure 2.12: A hypothetical model showing a mechanism where HIF-1-hydroxyl antagonises HIF-1. **A**. Diagram showing that RHY-1 activates EGL-9. EGL-9 hydroxylates HIF-1 in an oxygen dependent fashion. Under normoxia, HIF-1 is rapidly hydroxylated and only slowly does hydroxylated HIF-1 return to its original state. EGL-9 can also inhibit HIF-1 in an oxygen-independent fashion. HIF-1 hydroxyl is rapidly degraded in a VHL-1 dependent fashion. In our model, HIF-1 and HIF-1 hydroxyl have opposing effects on transcription. The width of the arrows represents the rates under normoxic conditions. **B**. Table showing the effects of loss-of-function mutations on HIF-1 and HIF-1 hydroxyl activity, showing how this can potentially explain the behavior of *ftn-1* in each case. S.S = Steady-state.

functionality of the hypoxia pathway. First, HIF-1-hydroxyl is challenging to study genetically because no mimetic mutations are available with which to study the pure hydroxylated HIF-1 species. Second, although mutations in the Von-Hippel Landau gene stabilize the hydroxyl species, they also increase the quantity of non-hydroxylated HIF-1 by mass action. Finally, since HIF-1 is detected low levels in cells under normoxic conditions (Wang and G L Semenza, 1993), total HIF-1 protein (unmodified HIF-1 plus HIF-1-hydroxyl) is often tacitly assumed to be vanishingly rare and therefore biologically inactive.

Our data show hundreds of genes that change expression in response to loss of *hif-1* under normoxic conditions. This establishes that there is sufficient total HIF-1 protein to be biologically active. Our analyses also revealed that *hif-1(lf)* shares positive correlations with *egl-9(lf)*, *rhy-1(lf)* and *vhl-1(lf)*, and that each of these genotypes also shows a secondary negative rank-ordered expression correlation with each other. These cross-patterns between all loss of function of inhibitors of HIF-1 and *hif-1(lf)* can be most easily explained if HIF-1-hydroxyl is biologically active.

A homeostatic argument can be made in favor of the activity of HIF-1-hydroxyl. At any point in time, the cell must measure the levels of multiple metabolites at once. The *hif-1*-dependent hypoxia response integrates information from O_2, α-ketoglutarate (2-oxoglutarate) and iron concentrations in the cell. One way to integrate this information is by encoding it only in the effective hydroxylation rate of HIF-1 by EGL-9. Then the dynamics in this system will evolve exclusively as a result of the total amount of HIF-1 in the cell. Such a system can be sensitive to fluctuations in the absolute concentration of HIF-1 (Goentoro et al., 2009). Since the absolute levels of HIF-1 are low in normoxic conditions, small fluctuations in protein copy-number represent can represent a large fold-change in HIF-1 levels. These fluctuations would not be problematic for genes that must be turned on only under conditions of severe hypoxia—presumably, these genes would be associated with low affinity sites for HIF-1, so that they are only activated when HIF-1 levels are far above random fluctuations.

For yet other sets of genes that must change expression in response to the hypoxia pathway, it may not make as much sense to integrate metabolite information exclusively via EGL-9-dependent hydroxylation of HIF-1. In particular, genes that may function to increase survival in mild hypoxia may benefit from regulatory mechanisms that can sense minor changes in environmental conditions and which therefore benefit from robustness to transient changes in protein copy number. Likewise, genes that are involved in iron or α-ketoglutarate metabolism (such as *ftn-1*) may benefit from being able to sense, accurately, small and consistent deviations from basal concentrations of these metabolites. For these genes, the information may be better encoded by using HIF-1 and HIF-1-hydroxyl as an activator/repressor pair. Such circuits are known to possess distinct advantages for controlling output in a manner that is robust to transient fluctuations in the levels of their components (Hart, Antebi, et al., 2012; Hart and Alon, 2013).

Our RNA-seq data suggests that one of these atypical targets of HIF-1 may be RHY-1. Although *rhy-1* does not exhibit non-classical epistasis, *hif-1(lf)* and *egl-9(lf) hif-1(lf)* both had increased expression levels of *rhy-1*. We speculate that if *rhy-1* is controlled by both HIF-1 and HIF-1-hydroxyl, then this might imply that HIF-1 regulates the expression of its pathway (and therefore itself) in a manner that is robust to total HIF-1 levels.

Insights into genetic interactions from vectorial phenotypes

Here, we have described a set of straightforward methods that can be in theory applied to any vectorial phenotype. Genome-wide methods afford a lot of information, but genome-wide interpretation of the results is often extremely challenging. Each method has its own advantages and disadvantages. We briefly discuss these methods, their uses and their drawbacks.

Principal component analysis is computationally tractable and clusters can often be visually detected with ease. However, PCA can be misleading, especially when the dimensions represented do not explain a very large fraction of the variance present in the data. In addition, principal dimensions are the product of a linear combination of vectors, and therefore must be interpreted with extreme care. In this case, the first principal dimension separated genotypes that increase HIF-1 protein levels from those that decrease it, but this dimension is a mix of vectors of change in gene expression. Although PCA showed that there is information hidden in these genotypes, it was not enough by itself to provide biological insight.

Whereas PCA operates on all genotypes simultaneously, correlation analysis is a pairwise procedure that measures how predictable the gene expression changes are in a mutant given the vector of expression changes in another. Like PCA, correlation analysis is easy and fast to perform. Unlike PCA, the product of a correlation analysis is a single number with a straightforward interpretation. However, correlation analysis is particularly sensitive to outliers. Although a common strategy is to rank-transform expression data to mitigate outliers, rank-transformations do not remove the cross-patterns that appear when feedback loops or other complex interactions are present between two genes. Such cross-patterns can still lead to vanishing correlations if both patterns are equally strong. Therefore, correlation analyses must take into account the possible existence of systematic outliers. Moreover, correlation values must be measured for both interactions in cross-patterned rank plots. Weighted correlations could be informative for ordering genes along pathways. A drawback of correlation analysis is that the number of pairwise comparisons that must be made increases combinatorially, though strategies could be used to decrease the total number of effective comparisons.

Epistasis plots are a novel way to visualize epistasis in vectorial phenotypes. Here, we have shown how an epistasis plot can be used to identify interactions between two single mutants and a double mutant. In reality, epistasis plots can be generated for any set of measurements involving a set of N mutants and an N-mutant genotype.

Epistasis plots can accumulate an arbitrary number of points within them, possess a rich structure that can be visualized and have straightforward interpretations for special slope values.

Another way to analyze epistasis is via general linear models (GLMs) that include interaction terms between two or more genes. In this way, GLMs can quantify the epistatic effect of an interaction on single genes. We and others (Dixit et al., 2016; Angeles-Albores, Leighton, et al., 2016) have previously used GLMs to identify gene sets that are epistatically regulated by two or more inputs. While powerful, GLMs suffer from the multiple comparison problem. Correcting for false positives using well-known multiple comparison corrections such as FDR (Storey and Tibshirani, 2003) tends to increase false negative rates. Moreover, since GLMs attempt to estimate effect magnitudes for individual gene or isoform expression levels, they effectively treat each gene as an independent quantity, which prevents better estimation of the magnitude and direction of the epistasis between two genes.

Epistasis plots do not suffer from the multiple comparison problem because the number of tests performed is orders of magnitudes smaller than the number of tests performed by GLMs. Ideally, in an epistasis plot we need only perform 3 tests—rejection of additive, unbranched and suppressive null models—compared with the tens of thousands of tests that are performed in GLMs. Moreover, the magnitude of epistasis between two genes can be estimated using hundreds of genes, which greatly improves the statistical resolution of the epistatic coefficient. This increased resolution is important because the size and magnitude of the epistasis has specific consequences for the type of pathway that is expected.

Any quantitative use of genome-wide datasets requires a good experimental setup. Here, we have demonstrated that whole-organism RNA-seq can be used to dissect molecular pathways in exquisite detail when paired with experimental designs that are motivated by classical genetics. Much more research will be necessary to understand whether epistasis has different consequences in the microscopic realm of transcriptional phenotypes than in the macroscopic world that geneticists have explored previously. Our hope is that these tools, coupled with the classic genetics experimental designs, will reveal hitherto unknown aspects of genetics theory.

Methods

Nematode strains and culture

Strains used were N2 wild-type Bristol, CB5602 *vhl-1*(*ok161*), CB6088 *egl-9*(*sa307*) *hif-1*(*ia4*), CB6116 *egl-9*(*sa307*);*vhl-1*(*ok161*), JT307 *egl-9*(*sa307*), ZG31 *hif-1*(*ia4*), RB1297 *rhy-1*(*ok1402*). All lines were grown on standard nematode growth media (NGM) plates seeded with OP50 *E. coli* at 20°C (Brenner 1974).

RNA Isolation

Unsynchronized lines were grown on NGM plates at 20C and eggs harvested by sodium hypochlorite treatment. Eggs were plated on 6 to 9 6cm NGM plates with ample OP50 *E. coli* to avoid starvation and grown at 20°C. Worms were staged and harvested based on the time after plating, vulva morphology and the absence of eggs. Approximately 30–50 non-gravid young adults were picked and placed in 100μL of TE pH 8.0 at 4°C in 0.2mL PCR tubes. After settling and a brief spin in microcentrifuge approximately 80μL of TE (Ambion AM 9849) was removed from the top of the sample and individual replicates were snap frozen in liquid N2. These replicate samples were then digested with Proteinase K (Roche Lot No. 03115 838001 Recombinant Proteinase K PCR Grade) for 15min at 60° in the presence of 1% SDS and 1.25μL RNA Secure (Ambion AM 7005). RNA samples were then taken up in 5 Volumes of Trizol (Tri Reagent Zymo Research) and processed and treated with DNase I using Zymo MicroPrep RNA Kit (Zymo Research Quick-RNA MicroPrep R1050). RNA was eluted in RNase-free water and divided into aliquots and stored at -80°C. One aliquot of each replicate was analyzed using a NanoDrop (Thermo Fisher) for impurities, Qubit for concentration and then analyzed on an Agilent 2100 BioAnalyzer (Agilent Technologies). Replicates were selected that had RNA integrity numbers (RIN) equal or greater than 9.0 and showed no evidence of bacterial ribosomal bands, except for the ZG31 mutant where one of three replicates had a RIN of 8.3.

Library Preparation and Sequencing

10ng of quality checked total RNA from each sample was reverse-transcribed into cDNA using the Clontech SMARTer Ultra Low Input RNA for Sequencing v3 kit (catalog #634848) in the SMARTSeq2 protocol (Picelli et al., 2014). RNA was denatured at 70°C for 3 minutes in the presence of dNTPs, oligo dT primer and spiked-in quantitation standards (NIST/ERCC from Ambion, catalog #4456740). After chilling to 4°C, the first-strand reaction was assembled using the LNA TSO primer described in Picelli et al. (2014), and run at 42°C for 90 minutes, followed by denaturation at 70°C for 10 minutes. The entire first strand reaction was then used as template for 13 cycles of PCR using the Clontech v3 kit. Reactions were cleaned up with 1.8X volume of Ampure XP SPRI beads (catalog #A63880) according to the manufacturer's protocol. After quantification using the Qubit High Sensitivity DNA assay, a 3ng aliquot of the amplified cDNA was run on the Agilent HS DNA chip to confirm the length distribution of the amplified fragments. The median value for the average cDNA lengths from all length distributions was 1076bp. Tagmentation of the full length cDNA for sequencing was performed using the Illumina/Nextera DNA library prep kit (catalog #FC-121–1030). Following Qubit quantitation and Agilent BioAnalyzer profiling, the tagmented libraries were sequenced. Libraries were sequenced on Illumina HiSeq2500 in single read mode with the read length of 50nt to an average depth of 15 million reads per sample following manufacturer's instructions. Base calls were performed with RTA 1.13.48.0 followed by conversion to FASTQ with bcl2fastq 1.8.4. Spearman correlation of the transcripts per million (TPM) for each genotype showed that every pairwise correlation within genotype was > 0.9.

Read Alignment and Differential Expression Analysis

We used Kallisto to perform read pseudo-alignment and performed differential analysis using Sleuth. We fit a general linear model for a transcript t in sample i:

$$y_{t,i} = \beta_{t,0} + \beta_{t,genotype} \cdot X_{t,i} + \beta_{t,batch} \cdot Y_{t,i} + \epsilon_{t,i} \tag{2.1}$$

where $y_{t,i}$ are the logarithm transformed counts; $\beta_{t,genotype}$ and $\beta_{t,batch}$ are parameters of the model, and which can be interpreted as biased estimators of the log-fold change; $X_{t,i}, Y_{t,i}$ are indicator variables describing the conditions of the sample; and $\epsilon_{t,i}$ is the noise associated with a particular measurement.

Genetic Analysis, Overview

Genetic analysis of the processed data was performed in Python 3.5. Our scripts made extensive use of the Pandas, Matplotlib, Scipy, Seaborn, Sklearn, Networkx, Bokeh, PyMC3, and TEA libraries (Bokeh Development Team, 2014; McKinney, 2011; Oliphant, 2007; Pedregosa et al., 2012; Salvatier, Wiecki, and Fonnesbeck, 2015; Van Der Walt, Colbert, and Varoquaux, 2011; Hunter, 2007; Angeles-Albores, N. Lee, et al., 2016; Waskom et al., 2016). Our analysis is available in a Jupyter

Notebook (Pérez and Granger, 2007). All code and required data (except the raw reads) are available at `https://github.com/WormLabCaltech/mprsq` along with version-control information. Our Jupyter Notebook and interactive graphs for this project can be found at `https://wormlabcaltech.github.io/mprsq/`. Raw reads were deposited in the Short Read Archive under the study accession number SRP100886.

Weighted Correlations

Pairwise correlations between transcriptomes where calculated by first identifying the set of differentially expressed genes (DEGs) common to both transcriptomes under analysis. DEGs were then rank-ordered according to their regression coefficient, β. Bayesian robust regressions were performed using a Student-T distribution. Bayesian analysis was performed using the PyMC3 library (Salvatier, Wiecki, and Fonnesbeck, 2015) (`pm.glm.families.StudenT` in Python). If the correlation has an average value > 1, the correlation coefficient was set to 1.

Weights were calculated as the proportion of genes that were < 1.5 standard deviations away from the primary regression out of the entire set of shared DEGs for each transcriptome.

Epistasis Analysis

For a double mutant X^-Y^-, we used the single mutants X^- and Y^- to find expected value of the coefficient for a double mutant under an additive model for each isoform i. Specifically,

$$\beta_{\text{Add},i} = \beta_{X,i} + \beta_{Y,i}. \tag{2.2}$$

Next, we find the difference, Δ_i, between the observed double mutant expression coefficient, $\beta_{XY,\text{Obs},i}$, and the predicted expression coefficient under an additive model for each isoform i.

To calculate the transcriptome-wide epistasis coefficient, we plotted $(\beta_{\text{Add},i}, \Delta_i)$ and found the line of best fit using orthogonal distance regression using the `scipy.odr` package in Python. We performed non-parametric bootstrap sampling of the ordered tuples with replacement using 5,000 iterations to generate a probability distribution of slopes of best fit.

There are as many models as epistatic relationships. For quantitative phenotypes, epistatic relationships (except synthetic interactions) can be generally expressed as:

$$\beta_{XY} = \sum_{g \in G} \lambda_g \beta_g, \tag{2.3}$$

where P_i is the quantitative phenotype belonging to the genotype i; G is the set of single mutants $\{X, Y\}$ that make up the double mutant, XY; and λ_g is the contribution of the phenotype P_g to P_{XY}. Additive interactions between genes are the result of setting $\lambda_g = 1$. All other relationships correspond to setting $\lambda_X = 0$, $\lambda_Y = 1$ or $\lambda_X = 1$, $\lambda_Y = 0$.

A given epistatic interaction can be simulated by predicting the double mutant phenotype under that interaction and re-calculating the y-coordinates. The recalculated y-coordinates can then be used to predict the possible epistasis coefficients for the cases where X is epistatic over Y, and Y is epistatic over X.

To select between theoretical models, we implemented an approximate Bayesian Odds Ratio. We defined a free-fit model, M_1, that found the line of best fit for the data:

$$P(\alpha \mid M_1, D) \propto \prod_{(x_i, y_i, \sigma_i) \in D} \exp \frac{(y_i - \alpha \cdot x_i)^2}{2\sigma_i} \cdot (1 + \alpha^2)^{-3/2}, \tag{2.4}$$

where α is the slope of the model to be determined, x_i, y_i were the x- and y-coordinates of each point respectively, and σ_i was the standard error associated with the y-value. We minimized the negative logarithm of equation 2.4 to obtain the most likely slope given the data, D (`scipy.optimize.minimize` in Python). Finally, we approximated the odds ratio as:

$$OR = \frac{P(D \mid \alpha^*, M_1) \cdot (2\pi)^{1/2} \sigma_{\alpha^*}}{P(D \mid M_i)}, \tag{2.5}$$

where α^* is the slope found after minimization, σ_α^* is the standard deviation of the parameter at the point α^* and $P(D \mid M_i)$ is the probability of the data given the parameter-free model, M_i.

Enrichment Analysis

Tissue, Phenotype and Gene Ontology Enrichment Analysis were carried out using the WormBase Enrichment Suite for Python (Angeles-Albores, R. Y. Lee, et al., 2017; Angeles-Albores, N. Lee, et al., 2016).

Author Contributions:

This work was supported by HHMI with whom PWS is an investigator and by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology. All strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). This article was written with support of the Howard Hughes Medical Institute. This article wouldn't be possible without help from Dr._ Igor Antoshechkin who performed all sequencing. We thank Hillel Schwartz for all of his careful advice. We would like to thank Jonathan Liu, Han Wang, and Porfirio Quintero for helpful discussion.

References

Ackerman, Daniel and David Gems (2012). "Insulin/IGF-1 and hypoxia signaling act in concert to regulate iron homeostasis in *Caenorhabditis elegans*". In: *PLoS Genetics* 8.3. ISSN: 15537390. DOI: 10.1371/journal.pgen.1002498.

Adamson, Britt et al. (2016). "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response". In: *Cell* 167.7, 1867–1882.e21. ISSN: 00928674. DOI: 10.1016/j.cell.2016.11.048.

Angeles-Albores, David, Raymond Y Lee, et al. (2017). "Phenotype and gene ontology enrichment as guides for disease modeling in C. elegans". In: *bioRxiv*. DOI: 10.1101/106369.

Angeles-Albores, David, Daniel H W Leighton, et al. (2016). "Transcriptomic Description of an Endogenous Female State in *C. elegans*". In: *bioRxiv*.

Angeles-Albores, David, Raymond Y. N. Lee, et al. (2016). "Tissue enrichment analysis for *C. elegans* genomics". In: *BMC Bioinformatics* 17.1, p. 366. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1229-9.

Bellier, Audrey et al. (2009). "Hypoxia and the hypoxic response pathway protect against pore-forming toxins in *C. elegans*". In: *PLoS Pathogens* 5.12. ISSN: 15537366. DOI: 10.1371/journal.ppat.1000689.

Bishop, Tammie et al. (2004). "Genetic Analysis of Pathways Regulated by the von Hippel-Lindau Tumor Suppressor in *Caenorhabditis elegans*". In: *PLoS Biology* 2.10. ISSN: 15449173. DOI: 10.1371/journal.pbio.0020289.

Bokeh Development Team (2014). "Bokeh: Python library for interactive visualization". In:

Bray, Nicolas L et al. (2016). "Near-optimal probabilistic RNA-seq quantification." In: *Nature biotechnology* 34.5, pp. 525–7. ISSN: 1546-1696. DOI: 10.1038/nbt.3519.

Brem, Rachel B. et al. (2002). "Genetic Dissection of Transcriptional Regulation in Budding Yeast". In: *Science* 296.5568.

Brugarolas, James et al. (2004). “Regulation of mTOR function in response to hypoxia by REDD1 and the TSC1/TSC2 tumor suppressor complex”. In: *Genes and Development* 18.23, pp. 1–12. ISSN: 08909369. DOI: 10.1101/gad.1256804.(mTOR).

Dixit, Atray et al. (2016). “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens”. In: *Cell* 167.7, 1853–1866.e17. ISSN: 00928674. DOI: 10.1016/j.cell.2016.11.038.

Epstein, Andrew C. R. et al. (2001). “*C. elegans* EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation”. In: *Cell* 107.1, pp. 43–54. ISSN: 00928674. DOI: 10.1016/S0092-8674(01)00507-4.

Goentoro, Lea et al. (2009). “The Incoherent Feedforward Loop Can Provide Fold-Change Detection in Gene Regulation”. In: *Molecular Cell* 36.5, pp. 894–899. ISSN: 10972765. DOI: 10.1016/j.molcel.2009.11.018.

Hart, Yuval and Uri Alon (2013). “The Utility of Paradoxical Components in Biological Circuits”. In: *Molecular Cell* 49.2, pp. 213–221. ISSN: 10972765. DOI: 10.1016/j.molcel.2013.01.004.

Hart, Yuval, Yaron E Antebi, et al. (2012). “Design principles of cell circuits with paradoxical components”. In: *Proceedings of the National Academy of Sciences of the United States of America* 109.21, pp. 8346–8351. ISSN: 0027-8424. DOI: 10.1073/pnas.1117475109.

Huang, L. Eric et al. (1996). “Activation of hypoxia-inducible transcription factor depends primarily upon redox-sensitive stabilization of its α subunit”. In: *Journal of Biological Chemistry* 271.50, pp. 32253–32259. ISSN: 00219258. DOI: 10.1074/jbc.271.50.32253.

Huang, Linda S and Paul W Sternberg (2006). “Genetic dissection of developmental pathways.” In: *WormBook : the online review of* C. elegans *biology* 1995, pp. 1–19. ISSN: 1551-8507. DOI: 10.1895/wormbook.1.88.2.

Hughes, Timothy R. et al. (2000). “Functional Discovery via a Compendium of Expression Profiles”. In: *Cell* 102.1, pp. 109–126. ISSN: 00928674. DOI: 10.1016/S0092-8674(00)00015-5.

Hunter, John D. (2007). “Matplotlib: A 2D graphics environment”. In: *Computing in Science and Engineering* 9.3, pp. 99–104. ISSN: 15219615. DOI: 10.1109/MCSE.2007.55.

Jiang, B H et al. (1996). “Dimerization, DNA binding, and transactivation properties of hypoxia-inducible factor 1.” In: *The Journal of biological chemistry* 271.30, pp. 17771–17778. ISSN: 00219258. DOI: 10.1074/jbc.271.30.17771.

Jiang, Huaqi, Rong Guo, and Jo Anne Powell-Coffman (2001). “The *Caenorhabditis elegans* hif-1 gene encodes a bHLH-PAS protein that is required for adaptation to hypoxia.” In: *Proceedings of the National Academy of Sciences of the United*

States of America 98.14, pp. 7916–7921. ISSN: 0027-8424. DOI: 10.1073/pnas.141234698.

Kaelin, William G. and Peter J. Ratcliffe (2008). "Oxygen Sensing by Metazoans: The Central Role of the HIF Hydroxylase Pathway". In: *Molecular Cell* 30.4, pp. 393–402. ISSN: 10972765. DOI: 10.1016/j.molcel.2008.04.009.

King, Elizabeth G. et al. (2014). "Genetic Dissection of the *Drosophila melanogaster* Female Head Transcriptome Reveals Widespread Allelic Heterogeneity". In: *PLoS Genetics* 10.5. Ed. by Greg Gibson, e1004322. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004322.

Lee, Seung Jae, Ara B. Hwang, and Cynthia Kenyon (2010). "Inhibition of respiration extends *C. elegans* life span via reactive oxygen species that increase HIF-1 activity". In: *Current Biology* 20.23, pp. 2131–2136. ISSN: 09609822. DOI: 10.1016/j.cub.2010.10.057.

Li, Yang et al. (2006). "Mapping Determinants of Gene Expression Plasticity by Genetical Genomics in *C. elegans*". In: *PLoS Genetics* 2.12, e222. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.0020222.

Loenarz, Christoph et al. (2011). "The hypoxia-inducible transcription factor pathway regulates oxygen sensing in the simplest animal, *Trichoplax adhaerens*". In: *EMBO reports* 12.1, pp. 63–70. ISSN: 1469-221X. DOI: 10.1038/embor.2010.170.

Luhachack, Lyly G. et al. (2012). "EGL-9 Controls *C. elegans* Host Defense Specificity through Prolyl Hydroxylation-Dependent and -Independent HIF-1 Pathways". In: *PLoS Pathogens* 8.7, p. 48. ISSN: 15537366. DOI: 10.1371/journal.ppat.1002798.

Ma, Dengke K. et al. (2012). "CYSL-1 Interacts with the O_2-Sensing Hydroxylase EGL-9 to Promote H_2S-Modulated Hypoxia-Induced Behavioral Plasticity in *C. elegans*". In: *Neuron* 73.5, pp. 925–940. ISSN: 08966273. DOI: 10.1016/j.neuron.2011.12.037.

McKinney, Wes (2011). "pandas: a Foundational Python Library for Data Analysis and Statistics". In: *Python for High Performance and Scientific Computing*, pp. 1–9.

Metzker, Michael L (2010). "Sequencing technologies - the next generation." In: *Nature reviews. Genetics* 11.1, pp. 31–46. ISSN: 1471-0056. DOI: 10.1038/nrg2626.

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226.

Oliphant, Travis E (2007). "SciPy: Open source scientific tools for Python". In: *Computing in Science and Engineering* 9, pp. 10–20. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.58.

Park, Eun Chan et al. (2012). “Hypoxia regulates glutamate receptor trafficking through an HIF-independent mechanism”. In: *The EMBO Journal* 31.6, pp. 1618–1619. ISSN: 0261-4189. DOI: 10.1038/emboj.2012.44.

Patro, Rob, Geet Duggal, et al. (2016). “Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference”. In: *bioRxiv*, p. 021592. DOI: 10.1101/021592.

Patro, Rob, Stephen M. Mount, and Carl Kingsford (2014). “Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms”. In: *Nature biotechnology* 32.5, pp. 462–464. ISSN: 1546-1696. DOI: 10.1038/nbt.2862.

Pedregosa, Fabian et al. (2012). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830. ISSN: 15324435. DOI: 10.1007/s13398-014-0173-7.2.

Pérez, F. and B.E. Granger (2007). “IPython: A System for Interactive Scientific Computing Python: An Open and General- Purpose Environment”. In: *Computing in Science and Engineering* 9.3, pp. 21–29. ISSN: 15219615. DOI: doi:10.1109/MCSE.2007.53..

Pfleger, J, M He, and M Abdellatif (2015). “Mitochondrial complex II is a source of the reserve respiratory capacity that is regulated by metabolic sensors and promotes cell survival”. In: *Cell Death and Disease* 6.7, pp. 1–14. ISSN: 2041-4889. DOI: 10.1038/cddis.2015.202.

Phillips, Patrick C (2008). “Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems”. In: *Nat Rev Genet* 9.11, pp. 855–867. ISSN: 1471-0056. DOI: 10.1038/nrg2452.

Picelli, Simone et al. (2014). “Full-length RNA-seq from single cells using Smart-seq2.” In: *Nature protocols* 9.1, pp. 171–81. ISSN: 1750-2799. DOI: 10.1038/nprot.2014.006.

Pimentel, Harold J et al. (2016). “Differential analysis of RNA-Seq incorporating quantification uncertainty”. In: *bioRxiv*, p. 058164. DOI: 10.1101/058164.

Powell-Coffman, Jo Anne (2010). “Hypoxia signaling and resistance in *C. elegans*”. In: *Trends in Endocrinology and Metabolism* 21.7, pp. 435–440. ISSN: 10432760. DOI: 10.1016/j.tem.2010.02.006.

Powell-Coffman, Jo Anne, Christopher A. Bradfield, and William B. Wood (1998). “*Caenorhabditis elegans* Orthologs of the Aryl Hydrocarbon Receptor and Its Heterodimerization Partner the Aryl Hydrocarbon Receptor Nuclear Translocator”. In: *Proceedings of the National Academy of Sciences* 95.6, pp. 2844–2849. ISSN: 0027-8424. DOI: 10.1073/pnas.95.6.2844.

Romney, Steven Joshua et al. (2011). “HIF-1 regulates iron homeostasis in *Caenorhabditis elegans* by activation and inhibition of genes involved in iron uptake and storage”. In: *PLoS Genetics* 7.12. ISSN: 15537390. DOI: 10.1371/journal.pgen.1002394.

Salvatier, John, Thomas Wiecki, and Christopher Fonnesbeck (2015). “Probabilistic Programming in Python using PyMC”. In: *PeerJ Computer Science* 2.e55, pp. 1–24. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.55.

Schadt, Eric E. et al. (2003). “Genetics of gene expression surveyed in maize, mouse and man”. In: *Nature* 422.6929, pp. 297–302. ISSN: 00280836. DOI: 10.1038/nature01434.

Schwarz, Erich M., Mihoko Kato, and Paul W. Sternberg (2012). “Functional transcriptomics of a migrating cell in *Caenorhabditis elegans*.” In: *Proceedings of the National Academy of Sciences of the United States of America* 109.40, pp. 16246–51. ISSN: 1091-6490. DOI: 10.1073/pnas.1203045109.

Scimone, M. Lucila et al. (2014). “Neoblast specialization in regeneration of the planarian *Schmidtea mediterranea*”. In: *Stem Cell Reports* 3.2, pp. 339–352. ISSN: 22136711. DOI: 10.1016/j.stemcr.2014.06.001.

Semenza, G L et al. (1994). “Transcriptional regulation of genes encoding glycolytic enzymes by hypoxia-inducible factor 1.” In: *The Journal of Biological Chemistry* 269.38, pp. 23757–63. ISSN: 0021-9258.

Semenza, Gregg L. (2011). “Hypoxia-inducible factor 1: Regulator of mitochondrial metabolism and mediator of ischemic preconditioning”. In: *Biochimica et Biophysica Acta - Molecular Cell Research* 1813.7, pp. 1263–1268. ISSN: 01674889. DOI: 10.1016/j.bbamcr.2010.08.006.

– (2012). “Hypoxia-inducible factors in physiology and medicine”. In: *Cell* 148.3, pp. 399–408. ISSN: 00928674. DOI: 10.1016/j.cell.2012.01.021.

Shalek, Alex K. et al. (2013). “Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells”. In: *Nature* 498.7453, pp. 236–40. ISSN: 1476-4687. DOI: 10.1038/nature12172.

Shao, Zhiyong, Yi Zhang, and Jo Anne Powell-Coffman (2009). “Two Distinct Roles for EGL-9 in the Regulation of HIF-1-mediated gene expression in *Caenorhabditis elegans*”. In: *Genetics* 183.3, pp. 821–829. ISSN: 00166731. DOI: 10.1534/genetics.109.107284.

Shao, Zhiyong, Yi Zhang, Qi Ye, et al. (2010). “*C. elegans* SWAN-1 binds to EGL-9 and regulates HIF-1- mediated resistance to the bacterial pathogen *Pseudomonas aeruginosa* PAO1”. In: *PLoS Pathogens* 6.8, pp. 91–92. ISSN: 15537366. DOI: 10.1371/journal.ppat.1001075.

Shen, Chuan, Daniel Nettleton, et al. (2005). “Roles of the HIF-1 hypoxia-inducible factor during hypoxia response in *Caenorhabditis elegans*”. In: *Journal of Biological Chemistry* 280.21, pp. 20580–20588. ISSN: 00219258. DOI: 10.1074/jbc.M501894200.

Shen, Chuan, Zhiyong Shao, and Jo Anne Powell-Coffman (2006). “The *Caenorhabditis elegans rhy-1* Gene Inhibits HIF-1 Hypoxia-Inducible Factor Activity in a Negative Feedback Loop That Does Not Include *vhl-1*”. In: *Genetics* 174.3, pp. 1205–1214. ISSN: 00166731. DOI: 10.1534/genetics.106.063594.

Sikora, Jakub et al. (2010). “Bioinformatic and biochemical studies point to AAGR-1 as the ortholog of human acid α-glucosidase in *Caenorhabditis elegans*”. In: *Molecular and Cellular Biochemistry* 341.1-2, pp. 51–63. ISSN: 03008177. DOI: 10.1007/s11010-010-0436-3.

Singer, Meromit et al. (2016). “A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells”. In: *Cell* 166.6, 1500–1511.e9. ISSN: 00928674. DOI: 10.1016/J.CELL.2016.08.052.

Storey, John D and Robert Tibshirani (2003). “Statistical significance for genomewide studies.” In: *Proceedings of the National Academy of Sciences of the United States of America* 100.16, pp. 9440–5. ISSN: 0027-8424. DOI: 10.1073/pnas.1530509100.

Sudarshan, S. et al. (2009). “Fumarate hydratase deficiency in renal cancer induces glycolytic addiction and hypoxia-inducible transcription factor 1α stabilization by glucose-dependent generation of reactive oxygen species”. In: *Mol Cell Biol* 29.15, pp. 4080–4090. ISSN: 0270-7306. DOI: 10.1128/MCB.00483-09.

Trapnell, Cole et al. (2013). “Differential analysis of gene regulation at transcript resolution with RNA-seq.” In: *Nature biotechnology* 31.1, pp. 46–53. ISSN: 1546-1696. DOI: 10.1038/nbt.2450.

Van Der Walt, Stéfan, S. Chris Colbert, and Gaël Varoquaux (2011). “The NumPy array: A structure for efficient numerical computation”. In: *Computing in Science and Engineering* 13.2, pp. 22–30. ISSN: 15219615. DOI: 10.1109/MCSE.2011.37.

Van Driessche, Nancy et al. (2005). “Epistasis analysis with global transcriptional phenotypes”. In: *Nature Genetics* 37.5, pp. 471–477. ISSN: 1061-4036. DOI: 10.1038/ng1545.

Van Wolfswinkel, Josien C., Daniel E. Wagner, and Peter W. Reddien (2014). “Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment”. In: *Cell Stem Cell* 15.3, pp. 326–339. ISSN: 18759777. DOI: 10.1016/j.stem.2014.06.007.

Wang, G L and G L Semenza (1993). “Characterization of hypoxia-inducible factor 1 and regulation of DNA binding activity by hypoxia.” In: *The Journal of Biological Chemistry* 268.29, pp. 21513–8. ISSN: 0021-9258.

Waskom, Michael et al. (2016). “seaborn: v0.7.0 (January 2016)”. In: doi: 10.5281/zenodo.45133.

Yeung, K. Y. and W. L. Ruzzo (2001). “Principal component analysis for clustering gene expression data.” In: *Bioinformatics (Oxford, England)* 17.9, pp. 763–774. issn: 1367-4803. doi: 10.1093/bioinformatics/17.9.763.

Chapter 3

TRANSCRIPTOMIC DESCRIPTION OF AN ENDOGENOUS FEMALE STATE IN *C. ELEGANS*

Abstract

Understanding genome and gene function in a whole organism requires us to fully comprehend the life cycle and the physiology of the organism in question. Although *C. elegans* is traditionally thought of as a hermaphrodite, XX animals exhaust their sperm and become endogenous females after 3 days of egg-laying. The molecular physiology of this state has not been as intensely studied as other parts of the life cycle, despite documented changes in behavior and metabolism that occur at this stage. To study the female state of *C. elegans*, we measured the transcriptomes of 1st day adult hermaphrodites; endogenous, 6th day adult females; and at the same time points, mutant *fog-2(lf)* worms that have a feminized germline phenotype. At these time points, we could separate the effects of biological aging from the transition into the female state. *fog-2(lf)* mutants partially phenocopy 6 day adult wild-type animals and exhibit fewer differentially expressed genes as they age throughout these 6 days. Therefore, *fog-2* is epistatic to age as assessed by this transcriptomic phenotype, which indicates that both factors act on sperm status to mediate entry into the female state. These changes are enriched in transcription factors canonically associated with neuronal development and differentiation. Our data provide a high-quality picture of the changes that happen in global gene expression throughout the period of early aging in the worm.

Introduction

Transcriptome analysis by RNA-seq (Mortazavi et al., 2008) has allowed for in-depth analysis of gene expression changes between life stages and environmental conditions in many species (Gerstein et al., 2014; Blaxter et al., 2012). *Caenorhabditis elegans*, a genetic model nematode with extremely well defined and largely invariant development (Sulston and Horvitz, 1977; Sulston, Schierenberg, et al., 1983), has been subjected to extensive transcriptomic analysis across all stages of larval development (Hillier et al., 2009; Boeck et al., 2016; Murray et al., 2012) and many stages of embryonic development (Boeck et al., 2016). Although RNA-seq was used to develop transcriptional profiles of the mammalian aging process soon after its invention (Magalhães, Finch, and Janssens, 2010), few such studies have been conducted in *C. elegans* past the entrance into adulthood.

A distinct challenge to the study of aging transcriptomes in *C. elegans* is the hermaphroditic lifestyle of wild-type individuals of this species. Young adult hermaphrodites are capable of self-fertilization (Sulston and Brenner, 1974; Corsi, Wightman, and Chalfie, 2015), and the resulting embryos will contribute RNA to whole-organism RNA extractions. Most previous attempts to study the *C. elegans* aging transcriptome have addressed the aging process only indirectly, or relied on the use of genetically or chemically sterilized animals to avoid this problem (Murphy et al., 2003; Halaschek-Wiener et al., 2005; Lund et al., 2002; McCormick et al., 2012; Eckley et al., 2013; Boeck et al., 2016; Rangaraju et al., 2015). In addition, most of these studies obtained transcriptomes using microarrays, which are less accurate than RNA-seq, especially for low-expressed genes (Wang et al., 2014).

Here, we investigate what we argue is a distinct state in the *C. elegans* life cycle, the endogenous female state. Although *C. elegans* hermaphrodites emerge into adulthood already replete with sperm, after about 3 days of egg-laying the animals become sperm-depleted and can only reproduce by mating, This marks a transition into what we define as the endogenous female state. This state is behaviorally distinguished by increased male-mating success (Garcia, LeBoeuf, and Koo, 2007), which may be due to an increased attractiveness to males (Morsci, Haas, and Barr, 2011). This increased attractiveness acts at least partially through production of volatile chemical cues (Leighton et al., 2014). These behavioral changes are also coincident with functional deterioration of the germline (Andux and Ellis, 2008), muscle (Herndon et al., 2002), intestine (McGee et al., 2011) and nervous system (Liu et al., 2013), changes traditionally attributed to the aging process (T. R. Golden and Melov, 2007).

To decouple the effects of aging and sperm-loss, we devised a two factor experiment. We examined wild-type XX animals at the beginning of adulthood (before worms contained embryos, referred to as 1st day adults) and after sperm depletion (6 days after the last molt, which we term 6th day adults). Second, we examined feminized XX animals that fail to produce sperm but are fully fertile if supplied sperm by mating with males (see Fig. 3.1). We used *fog-2(lf)* mutants to obtain feminized animals. *fog-2* is involved in germ-cell sex determination in the hermaphrodite worm and is required for sperm production (Schedl and Kimble, 1988; Clifford et al., 2000).

C. elegans defective in sperm formation will never transition into or out of a hermaphroditic stage. As time moves forward, these spermless worms only exhibit changes related to biological aging. We also reasoned that we might be able to identify gene expression changes due to different life histories: whereas hermaphrodites lay almost 300 eggs over three days, spermless females do not lay a single one. The different life histories could affect gene expression.

Here, we show that we can detect a transcriptional signature associated both with loss of hermaphroditic sperm and entrance into the endogenous female state. We can also detect changes associated specifically with biological aging. Loss of sperm leads to increases in the expression levels of transcription factors that are canonically associated with development and cellular differentiation and enriched in neuronal functions. Biological aging causes transcriptomic changes consisting of 5,592 genes in *C. elegans*. 4,552 of these changes occur in both genotypes we studied, indicating they do not depend on life history or genotype. To facilitate exploration of the data, we have generated a website where we have deposited additional graphics, as well as all of the code used to generate these analyses: `https://wormlabcaltech.github.io/Angeles_Leighton_2016/` .

Materials and Methods

Strains

Strains were grown at 20°C on NGM plates containing *E. coli* OP50. We used the laboratory *C. elegans* strain N2 as our wild-type strain (Sulston and Brenner, 1974). We also used the N2 mutant strain JK574, which contains the *fog-2(q71)* allele, for our experiments.

RNA extraction

Synchronized worms were grown to either young adulthood or the 6th day of adulthood prior to RNA extraction. Synchronization and aging were carried out according to protocols described previously (Leighton et al., 2014). 1,000–5,000 worms from each replicate were rinsed into a microcentrifuge tube in S basal (5.85g/L NaCl, 1g/L K_2HPO_4, 6g/L KH_2PO_4), and then spun down at 14,000rpm for 30s. The supernatant was removed and 1mL of TRIzol was added. Worms were lysed by vortexing for 30 s at room temperature and then 20 min at 4°. The

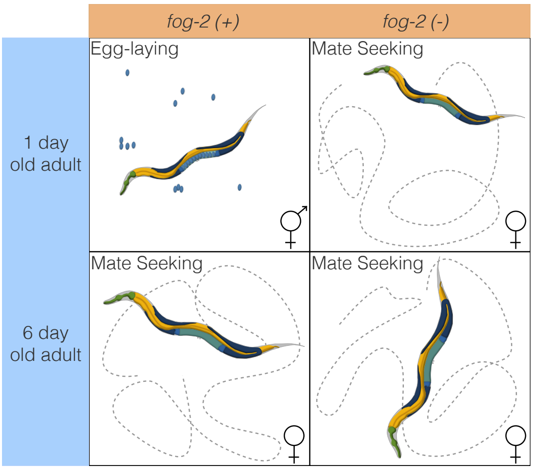


Figure 3.1: Experimental design to identify genes associated with sperm loss and with aging. Studying the wild-type worm alone would measure time- and sperm-related changes at the same time, without allowing us to separate these changes. Studying the wild-type worm and a *fog-2(lf)* mutant would enable us to measure sperm-related changes but not time-related changes. By mixing both designs, we can measure and separate both modules.

TRIzol lysate was then spun down at 14,000rpm for 10 min at 4°C to allow removal of insoluble materials. Thereafter the Ambion TRIzol protocol was followed to finish the RNA extraction (MAN0001271 Rev. Date: 13 Dec 2012). 3 biological replicates were obtained for each genotype and each time point.

RNA-Seq

RNA integrity was assessed using RNA 6000 Pico Kit for Bioanalyzer (Agilent Technologies #5067–1513) and mRNA was isolated using NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, NEB, #E7490). RNA-Seq libraries were constructed using NEBNext Ultra RNA Library Prep Kit for Illumina (NEB #E7530) following manufacturer's instructions. Briefly, mRNA isolated from $\sim 1\mu$g of total RNA was fragmented to the average size of 200nt by incubating at 94°C for 15 min in first strand buffer, cDNA was synthesized using random primers and ProtoScript II Reverse Transcriptase followed by second strand synthesis using Second Strand Synthesis Enzyme Mix (NEB). Resulting DNA fragments were end-repaired, dA tailed and ligated to NEBNext hairpin adaptors (NEB #E7335). After ligation, adaptors were converted to the 'Y' shape by treating with USER enzyme and DNA fragments were size selected using Agencourt AMPure XP beads (Beckman Coulter #A63880) to generate fragment sizes between 250 and 350 bp. Adaptor-ligated DNA was PCR amplified followed by AMPure XP bead clean up. Libraries were quantified with Qubit dsDNA HS Kit (ThermoFisher Scientific #Q32854) and the size distribution was confirmed with High Sensitivity DNA Kit for Bioanalyzer (Agilent Technologies #5067–4626). Libraries were sequenced on Illumina HiSeq2500 in single read mode with the read length of 50nt following manufacturer's instructions. Base calls were performed with RTA 1.13.48.0 followed by conversion to FASTQ with bcl2fastq 1.8.4.

Statistical Analysis

RNA-Seq Analysis

RNA-Seq alignment was performed using Kallisto (Bray et al., 2016) with 200 bootstraps. The commands used for read-alignment are in the S.I. file 1. Differential expression analysis was performed using Sleuth (Pimentel et al., 2016). The following General Linear Model (GLM) was fit:

$$\begin{aligned}\log(y_i) = &\beta_{0,i} + \beta_{G,i}{}^{\cdot} G + \\ &\beta_{A,i}{}^{\cdot} A + \beta_{A::G,i}{}^{\cdot} A\ G,\end{aligned}$$

where y_i are the TPM counts for the ith gene; $\beta_{0,i}$ is the intercept for the ith gene, and $\beta_{X,i}$ is the regression coefficient for variable X for the ith gene; A is a binary age variable indicating 1st day adult (0) or 6th day adult (1) and G is the genotype variable indicating wild-type (0) or *fog-2(lf)* (1); $\beta_{A::G,i}$ refers to the regression coefficient accounting for the interaction between the age and genotype variables in the ith gene. Genes were called significant if the FDR-adjusted q-value for any regression coefficient was less than 0.1. Our script for differential analysis is available on GitHub.

Regression coefficients and TPM counts were processed using Python 3.5 in a Jupyter Notebook (Pérez and Granger, 2007). Data analysis was performed using the Pandas, NumPy and SciPy libraries (McKinney, 2011; Van Der Walt, Colbert, and Varoquaux, 2011; Oliphant, 2007). Graphics were created using the Matplotlib and Seaborn libraries (Waskom et al., 2016; Hunter, 2007). Interactive graphics were generated using Bokeh (Bokeh Development Team, 2014).

Tissue, Phenotype and Gene Ontology Enrichment Analyses (TEA, PEA and GEA, respectively) were performed using the WormBase Enrichment Suite for Python (Angeles-Albores, N. Lee, et al., 2016; Angeles-Albores, Lee, et al., 2017).

Data Availability

Strains are available from the *Caenorhabditis* Genetics Center. All of the data and scripts pertinent for this project except the raw reads can be found on our Github repository `https://github.com/WormLabCaltech/Angeles_Leighton_2016`. File S1 contains the list of genes that were altered in aging regardless of genotype. File S2 contains the list of genes and their associations with the *fog-2(lf)* phenotype. File S3 contains genes associated with the female state. Raw reads were deposited to the Sequence Read Archive under the accession code SUB2457229.

Results and Discussion

Decoupling time-dependent effects from sperm-status via general linear models

In order to decouple time-dependent effects from changes associated with loss of hermaphroditic sperm, we measured wild-type and *fog-2(lf)* adults at the 1st day adult stage (before visible embryos were present) and 6th day adult stage, when all wild-type hermaphrodites had laid all their eggs (see Fig 3.1), but mortality was still low ($< 10\%$) (Stroustrup et al., 2013). We obtained 16–19 million reads mappable to the *C. elegans* genome per biological replicate, which enabled us to identify 14,702 individual genes totalling 21,143 isoforms (see Figure 3.2a).

One way to analyze the data from this two-factor design is by pairwise comparison of the distinct states. However, such an analysis would not make full use of all the statistical power afforded by this experiment. Another method that makes full use of the information in our experiment is to perform a linear regression in 3 dimensions (2 independent variables, age and genotype, and 1 output). A linear regression with 1 parameter (age, for example) would fit a line between expression data for young and old animals. When a second parameter is added to the linear regression, said parameter can be visualized as altering the y-intercept, but not the slope, of the first line in question (see Fig. 3.3a).

Although a simple linear model is oftentimes useful, sometimes it is not appropriate to assume that the two variables under study are entirely independent. For example, in our case, three out of the four timepoint-and-genotype combinations we studied did not have sperm, and sperm-status is associated with both the *fog-2(lf)* self-sterile phenotype and with biological age of the wild-type animal. One way to statistically model such correlation between variables is to add an interaction term to the linear regression. This interaction term allows extra flexibility in describing how changes occur between conditions. For example, suppose a given theoretical gene *X* has

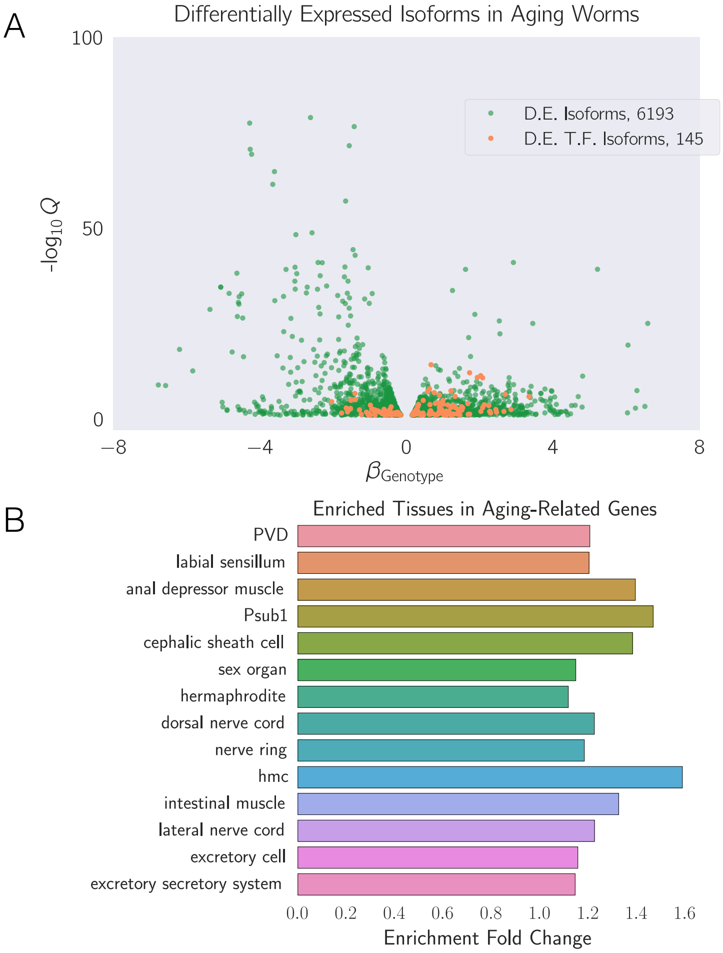


Figure 3.2: **A** We identified a common aging transcriptome between N2 and *fog-2(lf)* animals, consisting of 6,193 differentially expressed isoforms totaling 5,592 genes. The volcano plot is randomly down-sampled 30% for ease of viewing. Each point represents an individual isoform. β_{Aging} is the regression coefficient. Larger magnitudes of β indicate a larger log-fold change. The y-axis shows the negative logarithm of the q-values for each point. Green points are differentially expressed isoforms; orange points are differentially expressed isoforms of predicted transcription factor genes (Reece-Hoyes et al., 2005). An interactive version of this graph can be found on our website. **B** Tissue Enrichment Analysis (Angeles-Albores, N. Lee, et al., 2016) showed that genes associated with muscle tissues and the nervous system are enriched in aging-related genes. Only statistically significantly enriched tissues are shown. Enrichment Fold Change is defined as $Observed/Expected$. hmc stands for head mesodermal cell.

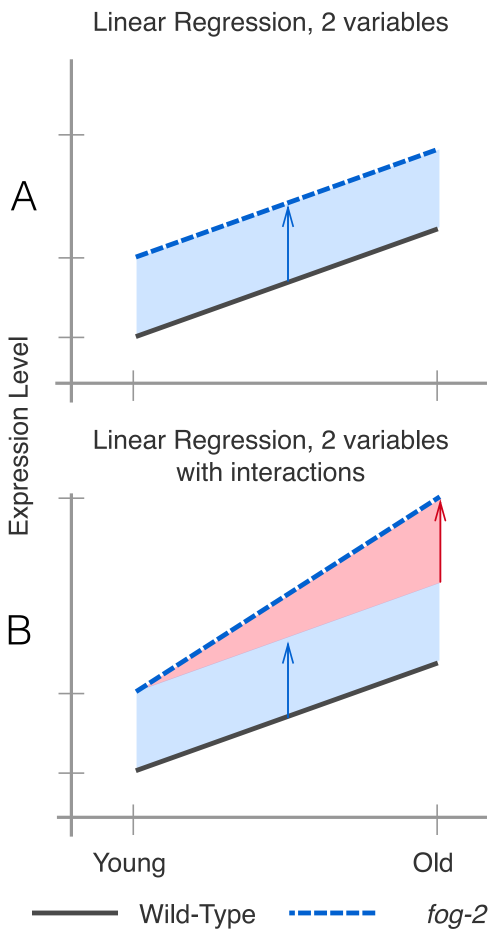

Linear Regression, 2 variables
A
Expression Level
Linear Regression, 2 variables
with interactions
B
Young
Old
Wild-Type
fog-2

Figure 3.3: **A**. A linear regression with two variables, age and genotype. The expression level of a gene increases by the same amount as worms age regardless of genotype. However, *fog-2(lf)* has more mRNA than the wild-type at all stages (blue arrow). **B**. A linear regression with two variables and an interaction term. In this example, the expression level of this hypothetical gene is different between wild-type worms and *fog-2(lf)* (blue arrow). Although the expression level of this gene increases with age, the slope is different between wild-type and *fog-2(lf)*. The difference in the slope can be accounted for through an interaction coefficient (red arrow).

expression levels that increase in a *fog-2*-dependent manner, but also increases in an age-dependent manner. However, aged *fog-2(lf)* animals do not have expression levels of X that would be expected from adding the effect of the two perturbations; instead, the expression levels of X in this animal are considerably above what is expected. In this case, we could add a positive interaction coefficient to the model to explain the effect of genotype on the y-intercept as well as the slope (see Fig. 3.3b). When the two perturbations are loss-of-function mutations, such interactions are epistatic interactions.

For these reasons, we used a linear generalized model (see Statistical Analysis) with interactions to identify a transcriptomic profile associated with the *fog-2(lf)* genotype independently of age, as well as a transcriptomic profile of *C. elegans* aging common to both genotypes. The change associated with each variable is referred as β; this number, although related to the natural logarithm of the fold change, is not equal to it. However, it is true that larger magnitudes of β indicate greater change. Thus, for each gene we performed a linear regression, and we evaluated the whether the β values associated with each coefficient were significantly different from 0 via a Wald test corrected for multiple hypothesis testing. A coefficient was considered to be significantly different from 0 if the q-value associated with it was less than 0.1.

A quarter of all genes change expression between the 1st day of adulthood and the 6th day of adulthood in *C. elegans*.

We identified a transcriptomic signature consisting of 5,592 genes that were differentially expressed in 6th day adult animals of either genotype relative to 1st day adult animals (see SI file 2). This constitutes more than one quarter of the genes in *C. elegans*. Tissue Enrichment Analysis (TEA) (Angeles-Albores, N. Lee, et al., 2016) showed that nervous tissues including the ‘nerve ring’, ‘dorsal nerve cord’, ‘PVD’ and ‘labial sensillum’ were enriched in genes that become differentially expressed through aging. Likewise, certain muscle groups (‘anal depressor muscle’, ‘intestinal muscle’) were enriched. (see Figure 3.2b). Gene Enrichment Analysis (GEA) (Angeles-Albores, Lee, et al., 2017) revealed that genes that were differentially expressed during the course of aging were enriched in terms involving respiration (‘respiratory chain’, ‘oxoacid metabolic process’); translation (‘cytosolic large ribosomal subunit’); and nucleotide metabolism (‘purine nucleotide’, ‘nucleoside phosphate’ and ‘ribose phosphate’ metabolic process). Phenotype Enrichment Analysis (PEA) (Angeles-Albores, Lee, et al., 2017) showed enrichment of phenotypes that affect the *C. elegans* gonad, including ‘gonad vesiculated’, ‘gonad small’, 'oocytes lack nucleus' and 'rachis narrow'.

To verify the quality of our dataset, we generated a list of 1,056 golden standard genes expected to be altered in 6th day adult worms using previous literature reports including downstream genes of *daf-12*, *daf-16*, and aging and lifespan extension datasets (Murphy et al., 2003; Halaschek-Wiener et al., 2005; Lund et al., 2002; McCormick et al., 2012; Eckley et al., 2013). Out of 1,056 standard genes, we found 506 genes in our time-responsive dataset. This result was statistically significant with a p-value $< 10^{-38}$.

Next, we used a published compendium (Reece-Hoyes et al., 2005) to search for known or predicted transcription factors. We found 145 transcription factors in the set of genes with differential expression in aging nematodes. We subjected this list of transcription factors to TEA to understand their expression patterns. 6 of these transcription factors were expressed in the 'hermaphrodite specific neuron' (HSN), a neuron physiologically relevant for egg-laying (*hlh-14*, *sem-4*, *ceh-20*, *egl-46*, *ceh-13*, *hlh-3*), which represented a statistically significant 2-fold enrichment of this tissue ($q < 10^{-1}$). The term 'head muscle' was also overrepresented at twice the expected level ($q < 10^{-1}$, 13 genes). Many of these transcription factors have been associated with developmental processess, and it is unclear why they would change expression in adult animals.

The whole-organism *fog-2(lf)* transcriptome in *C. elegans*.

We identified 1,881 genes associated with the *fog-2(lf)* genotype, including 60 transcription factors (see SI file 3). TEA showed that the terms 'AB', 'midbody', 'uterine muscle', 'cephalic sheath cell', 'anal depressor muscle' and 'PVD' were enriched in this gene set. The terms 'AB' and 'midbody' likely reflect the impact of *fog-2(lf)* on the germline. Phenotype enrichment showed that only a single phenotype, 'spindle orientation variant' was enriched in the *fog-2(lf)* transcriptome ($q < 10^{-1}$, 38 genes, 2-fold enrichment). Most genes annotated as 'spindle orientation variant' were slightly upregulated, and therefore are unlikely to uniquely reflect reduced germline proliferation. GO term enrichment was very similar to the aging gene set and reflected enrichment in annotations pertaining to translation and respiration. Unlike the aging gene set, the *fog-2(lf)* transcriptome was significantly enriched in 'myofibril' and 'G-protein coupled receptor binding' ($q < 10^{-1}$). Enrichment of the term 'G-protein coupled receptor binding' was due to 14 genes: *cam-1*, *mom-2*, *dsh-1*, *spp-10*, *flp-6*, *flp-7*, *flp-9*, *flp-13*, *flp-14*, *flp-18*, *K02A11.4*, *nlp-12*, *nlp-13*, and *nlp-40*. *dsh-1*, *mom-2* and *cam-1* are members of the Wnt signaling pathway. Most of these genes' expression levels were up-regulated, suggesting increased G-protein binding activity in *fog-2(lf)* mutants.

The *fog-2(lf)* transcriptome overlaps significantly with the aging transcriptome

Of the 1,881 genes that we identified in the *fog-2(lf)* transcriptome, 1,040 genes were also identified in our aging set. Moreover, of these 1,040 genes, 905 genes changed in the same direction in response either aging or germline feminization. The overlap between these transcriptomes suggests an interplay between sperm-status and age. The nature of the interplay should be captured by the interaction coefficients in our model. There are four possibilities. First, the *fog-2(lf)* worms may have a fast-aging phenotype, in which case the interaction coefficients should match the sign of the aging coefficient. Second, the *fog-2(lf)* worms may have a slow-aging phenotype, in which case the interaction coefficients should have an interaction coefficient that is of opposite sign, but not greater in magnitude than the aging coefficient (if a gene increases in aging in a wild-type worm, it should still increase in a *fog-2(lf)* worm, albeit less). Third, the *fog-2(lf)* worms exhibit a rejuvenation phenotype. If this is the case, then these genes should have an interaction coefficient that is of opposite sign and greater magnitude than their aging coefficient, such that the change of these genes in *fog-2(lf)* mutant worms is reversed relative to the wild-type. Finally, if these genes are indicative of a female state, then these genes should not change with age in *fog-2(lf)* animals, since these animals do not exit this state during the course of the experiment. Moreover, because wild-type worms become female as they age, a further requirement for a transcriptomic signature of the female state is that aging coefficients for genes in this signature should have genotype coefficients of equal sign and magnitude. In other words, entrance into the female state should be not be path-dependent.

To evaluate which of these possibilities was most likely, we selected the 1,040 genes that had aging, genotype and interaction coefficients significantly different from zero and we plotted their temporal coefficients against their genotype coefficients (see Fig. 3.4a). We observed that the aging coefficients were strongly predictive of the genotype coefficients. Most of these genes fell near the line $y = x$, suggesting that these genes define a female state. As a further test that these genes actually define a female state, we generated an epistasis plot using this gene set. We have previously used epistasis plots to measure transcriptome-wide epistasis between genes in a pathway (). Briefly, an epistasis plot plots the expected expression of

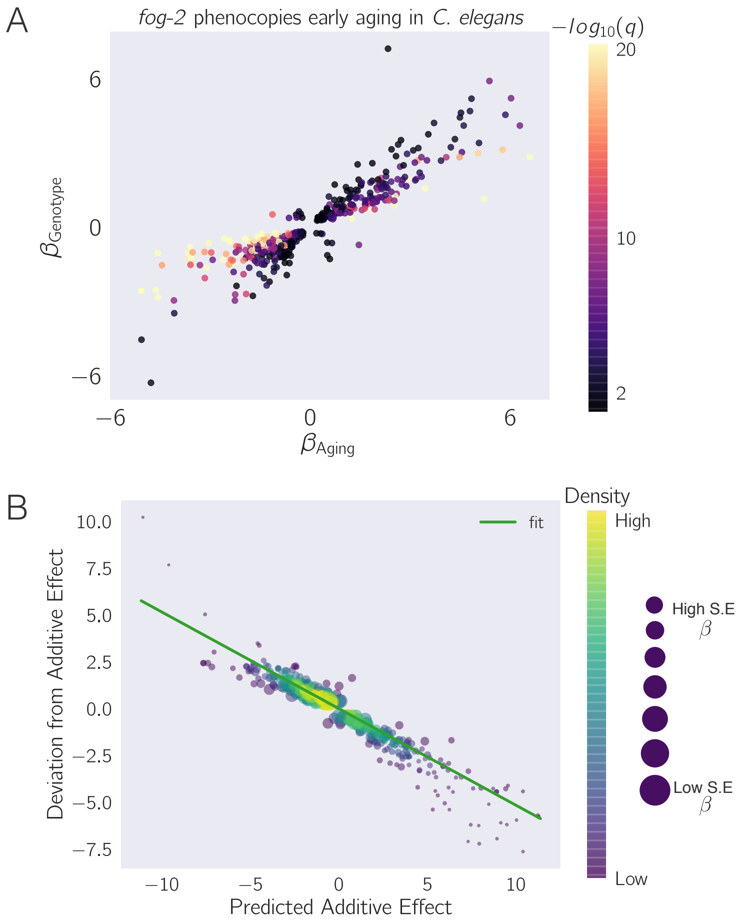


Figure 3.4: *fog-2(lf)* partially phenocopies early aging in *C. elegans*. The β in each axes is the regression coefficient from the GLM, and can be loosely interpreted as an estimator of the log-fold change. Feminization by loss of *fog-2(lf)* is associated with a transcriptomic phenotype involving 1,881 genes. 1,040/1,881 of these genes are also altered in wild-type worms as they progress from young adulthood to old adulthood, and 905 change in the same direction. However, progression from young to old adulthood in a *fog-2(lf)* background results in no change in the expression level of these genes. **A** We identified genes that change similarly during feminization and aging. The correlation between feminization and aging is almost 1:1. **B** Epistasis plot of aging versus feminization. Epistasis plots indicate whether two genes (or perturbations) act on the same pathway. When two effects act on the same pathway, this is reflected by a slope of -0.5. The measured slope was -0.51 ± 0.01.

a double perturbation under an additive model (null model) on the x-axis, and the deviation from this null model in the y-axis. In other words, we calculated the x-coordinates for each point by adding $\beta_{\text{Genotype}} + \beta_{\text{Aging}}$, and the y-coordinates are equal to $\beta_{Interaction}$ for each isoform. Previously we have shown that if two genes act in a linear pathway, an epistasis plot will generate a line with slope equal to -0.5. When we generated an epistasis plot and found the line of best fit, we observed a slope of -0.51 ± 0.01, which suggests that the *fog-2* gene and time are acting to generate a single transcriptomic phenotype along a single pathway. Overall, we identified 405 genes that increased in the same direction through age or mutation of the *fog-2(lf)* gene and that had an interaction coefficient of opposite sign to the aging or genotype coefficient (see SI file 4). Taken together, this information suggests that these 405 genes define a female state in *C. elegans*.

Analysis of the Female State Transcriptome

To better understand the changes that happen after sperm loss, we performed tissue enrichment, phenotype enrichment and gene ontology enrichment analyses on the set of 405 genes that we associated with the female state. TEA showed no tissue enrichment using this gene-set. GEA showed that this gene list was enriched in constituents of the ribosomal subunits almost four times above background ($q < 10^{-5}$, 17 genes). The enrichment of ribosomal constituents in this gene set in turn drives the enriched phenotypes: 'avoids bacterial lawn', 'diplotene absent during oogenesis', 'gonad vesiculated', 'pachytene progression during oogenesis variant', and 'rachis narrow'. The expression of most of these ribosomal subunits is down-regulated in aged animals or in *fog-2(lf)* mutants.

Discussion

Defining an Early Aging Phenotype

Our experimental design enables us to decouple the effects of egg-laying from aging. As a result, we identified a set of almost 4,000 genes that are altered similarly between wild-type and *fog-2(lf)* mutants. Due to the read depth of our transcriptomic data (20 million reads) and the number of samples measured (3 biological replicates for 4 different life stages/genotypes), this dataset constitutes a high-quality description of the transcriptomic changes that occur in aging populations of *C. elegans*. Although our data only capture $\sim 50\%$ of the expression changes reported in earlier aging transcriptome literature, this disagreement can be explained by a difference in methodology; earlier publications typically addressed the aging of fertile wild-type

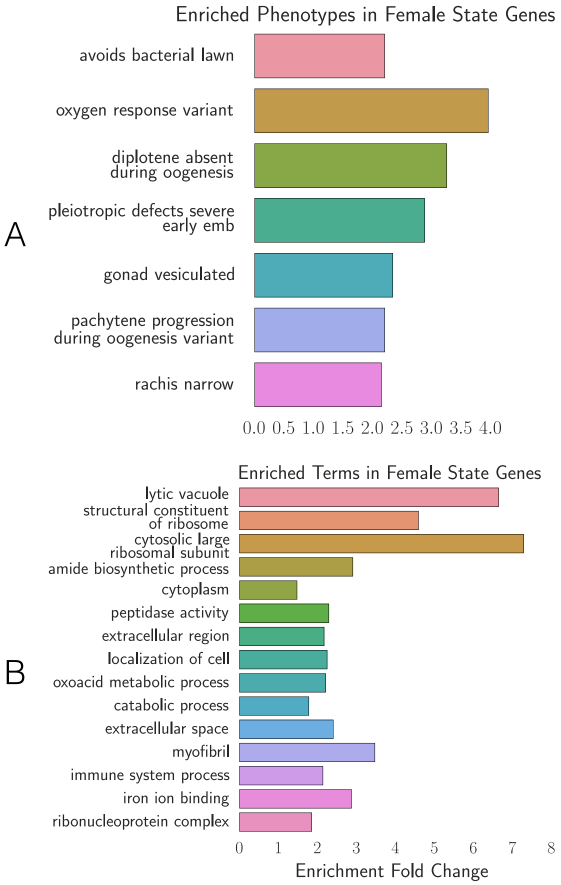


Figure 3.5: Phenotype and GO enrichment of genes involved in the female state. **A**. Phenotype Enrichment Analysis. **B**. Gene Ontology Enrichment Analysis. Most of the terms enriched in PEA reflect the abundance of ribosomal subunits present in this gene set.

hermaphrodites only indirectly, or queried aging animals at a much later stage of their life cycle.

Measurement of a female state is enabled by linear models

We set out to study the self-fertilizing (hermaphroditic) to self-sterile (female) transition by comparing wild-type animals with *fog-2(lf)* mutants as they aged. Our computational approach enabled us to separate between two biological processes that are correlated within samples. Because of this intra-sample correlation, identifying this state via pairwise comparisons would not have been straightforward. Although it is a favored method amongst biologists, such pairwise comparisons suffer from a number of drawbacks. First, pairwise comparisons are unable to draw on the full statistical power available to an experiment because they discard almost all information except the samples being compared. Second, pairwise comparisons require a researcher to define *a priori* which comparisons are informative. For experiments with many variables, the number of pairwise combinations is explosively large. Indeed, even for this two-factor experiment, there are 6 possible pairwise comparisons. On the other hand, by specifying a linear regression model, each gene can be summarized with three variables, each of which can be analyzed and understood without the need to resort to further pairwise combinations.

Our explorations have shown that the loss of *fog-2(lf)* partially phenocopies the transcriptional events that occur naturally as *C. elegans* ages from the 1st day of adulthood to the 6th day of adulthood. Moreover, epistasis analysis of these perturbations suggest that they act on the same pathway, namely sperm generation and depletion (see Fig. 3.6). Sperm generation promotes a non-female states, whereas sperm depletion causes entry into the female state. Given the enrichment of neuronal transcription factors that are associated with sperm loss in our dataset, we believe this dataset should contain some of the transcriptomic modules that are involved in these pheromone production and behavioral pathways, although we have been unable to find these genes. Currently, we cannot judge how many of the changes induced by loss of hermaphroditic sperm are developmental (i.e., irreversible), and how many can be rescued by mating to a male. While an entertaining thought experiment, establishing whether these transcriptomic changes can be rescued by males is a daunting experimental task, given that the timescales for physiologic changes could reasonably be the same as the timescale of onset of embryonic transcription. All in all, our research supports the idea that wide-ranging transcriptomic effects of aging in various tissues can be observed well before onset of mortality, and that *C. elegans* continues to develop as it enters a new state of its life cycle.

The *C. elegans* life cycle, life stages and life states

C. elegans has a complicated life cycle, with two alternative developmental pathways that have multiple stages (larval development and dauer development), followed by reproductive adulthood. In addition to its developmental stages, researchers have recognized that *C. elegans* has numerous life states that it can enter into when given instructive environmental cues. One such state is the L1 arrest state, where development ceases entirely upon starvation (Johnson et al., 1984). More recently, researchers have described additional diapause states that the worm can access at the L3, L4 and young adult stages under conditions of low food (Angelo and Gilst, 2009; Seidel and Kimble, 2011; Schindler, Baugh, and Sherwood, 2014). Not all

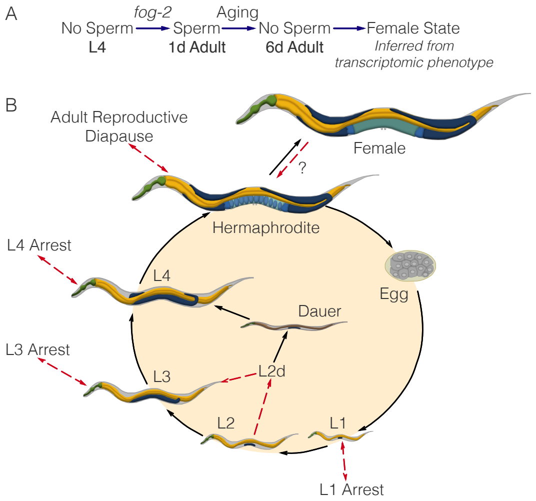


Figure 3.6: **A**. A substrate model showing how *fog-2* promotes sperm generation, whereas aging promotes sperm depletion, leading to entry to the female state. Such a model can explain why *fog-2* and aging appear epistatic to each other. **B**. The complete *C. elegans* life cycle. Recognized stages of *C. elegans* are marked by black arrows. States are marked by red arrows to emphasize that at the end of a state, the worm returns to the developmental timepoint it was at before entering the state. The L2d state is an exception. It is the only stage that does not return to the same developmental timepoint; rather, the L2d state is a permissive state that allows entry into either dauer or the L3 stage. We have presented evidence of a female state in *C. elegans*. At this point, it is unclear whether the difference between hermaphrodites and females is reversible by males. Therefore, it remains unclear whether it is a stage or a true state.

states of *C. elegans* are arrested, however (see Fig. 3.6). For example, the L2d state is induced by crowded and nutrient poor conditions (J. W. Golden and Riddle, 1984). While within this state, the worm is capable of entry into either dauer or the L3 larval stage, depending on environmental conditions. Thus, the L2d state is a permissive state, and marks the point at which the nematode development is committed to a single developmental pathway.

Identification of the *C. elegans* life states has often been performed by morphological studies (as in the course of L4 arrest or L2d) or via timecourses (L1 arrest). However, not all states may be visually identifiable, or even if they are, the morphological changes may be very subtle, making positive identification difficult. However, the detailed information afforded by a transcriptome should in theory provide sufficient information to definitively identify a state, since transcriptomic information underlies morphology. Moreover, transcriptomics can provide an informative description into the physiology of complex metazoan life state's via measurements of global gene expression. By identifying differentially expressed genes and using ontology enrichment analyses to identify gene functions, sites of expression or phenotypes that are enriched in a given gene set, researchers can obtain a clearer picture of the changes that occur in the worm in a less biased manner than by identifying gross morphological changes. RNA-seq is emerging as a powerful technology that has been used successfully in the past as a qualitative tool for target acquisition. More recent work has successfully used RNA-seq to establish genetic interactions between genes (Dixit et al., 2016; Adamson et al., 2016). In this work, we have shown that whole-organism RNA-seq data can also be analyzed via a similar formalism to successfully identify internal states in a multi-cellular organism.

References

Adamson, Britt et al. (2016). "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response". In: *Cell* 167.7, 1867–1882.e21. ISSN: 00928674. DOI: 10.1016/j.cell.2016.11.048.

Andux, Sara and Ronald E. Ellis (2008). "Apoptosis maintains oocyte quality in aging Caenorhabditis elegans females". In: *PLoS Genetics* 4.12. ISSN: 15537390. DOI: 10.1371/journal.pgen.1000295.

Angeles-Albores, David, Raymond Y Lee, et al. (2017). "Phenotype and gene ontology enrichment as guides for disease modeling in C. elegans". In: *bioRxiv*. DOI: 10.1101/106369. URL: http://biorxiv.org/content/early/2017/02/07/106369.

Angeles-Albores, David, Raymond Y. N. Lee, et al. (2016). "Tissue enrichment analysis for *C. elegans* genomics". In: *BMC Bioinformatics* 17.1, p. 366. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1229-9.

Angelo, Giana and Marc R Van Gilst (2009). "Cells and Extends Reproductive". In: *Science* 326.November, pp. 954–958.

Blaxter, M. et al. (2012). "Genomics and transcriptomics across the diversity of the Nematoda". In: *Parasite Immunology* 34.2-3, pp. 108–120. ISSN: 01419838. DOI: 10.1111/j.1365-3024.2011.01342.x.

Boeck, Max E et al. (2016). "The time-resolved transcriptome of C . elegans". In: *Genome Research*, pp. 1–10. ISSN: 15495469. DOI: 10.1101/gr.202663.115.Freely.

Bokeh Development Team (2014). "Bokeh: Python library for interactive visualization". In:

Bray, Nicolas L et al. (2016). "Near-optimal probabilistic RNA-seq quantification." In: *Nature biotechnology* 34.5, pp. 525–7. ISSN: 1546-1696. DOI: 10.1038/nbt.3519.

Clifford, Robert et al. (2000). “FOG-2, a novel F-box containing protein, associates with the GLD-1 RNA binding protein and directs male sex determination in the C. elegans hermaphrodite germline.” In: *Development (Cambridge, England)* 127.24, pp. 5265–5276. ISSN: 0950-1991.

Corsi, Ann K., Bruce Wightman, and Martin Chalfie (2015). “A transparent window into biology: A primer on Caenorhabditis elegans”. In: *Genetics* 200.2, pp. 387–407. ISSN: 19432631. DOI: 10.1534/genetics.115.176099.

Dixit, Atray et al. (2016). “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens”. In: *Cell* 167.7, 1853–1866.e17. ISSN: 00928674. DOI: 10.1016/j.cell.2016.11.038.

Eckley, D. Mark et al. (2013). “Molecular characterization of the transition to midlife in Caenorhabditis elegans”. In: *Age* 35.3, pp. 689–703. ISSN: 01619152. DOI: 10.1007/s11357-012-9401-2.

Garcia, L. Rene, Brigitte LeBoeuf, and Pamela Koo (2007). “Diversity in mating behavior of hermaphroditic and male-female Caenorhabditis nematodes”. In: *Genetics* 175.4, pp. 1761–1771. ISSN: 00166731. DOI: 10.1534/genetics.106.068304.

Gerstein, Mark B. et al. (2014). “Comparative analysis of the transcriptome across distant species”. In: *Nature* 512, pp. 445–448. ISSN: 0028-0836. DOI: 10.1038/nature13424. arXiv: NIHMS150003. URL: http://www.nature.com/doifinder/10.1038/nature13424%7B%5C%%7D5Cnhttp://dx.doi.org/10.1038/nature13424.

Golden, James W. and Donald L. Riddle (1984). “The Caenorhabditis elegans dauer larva: Developmental effects of pheromone, food, and temperature”. In: *Developmental Biology* 102.2, pp. 368–378. ISSN: 00121606. DOI: 10.1016/0012-1606(84)90201-X.

Golden, Tamara R and Simon Melov (2007). “Gene expression changes associated with aging in C. elegans.” In: *WormBook : the online review of C. elegans biology*, pp. 1–12. ISSN: 1551-8507. DOI: 10.1895/wormbook.1.127.2.

Halaschek-Wiener, Julius et al. (2005). “Analysis of long-lived C. elegans daf-2 mutants using serial analysis of gene expression”. In: *Genome Research*, pp. 603–615. DOI: 10.1101/gr.3274805..

Herndon, Laura a et al. (2002). “Stochastic and genetic factors influence tissue-specific decline in ageing C. elegans.” In: *Nature* 419.6909, pp. 808–814. ISSN: 0028-0836. DOI: 10.1038/nature01135.

Hillier, Ladeana W. et al. (2009). “Massively parallel sequencing of the polyadenylated transcriptome of C. elegans”. In: *Genome Research* 19.4, pp. 657–666. ISSN: 10889051. DOI: 10.1101/gr.088112.108.

Hunter, John D. (2007). “Matplotlib: A 2D graphics environment”. In: *Computing in Science and Engineering* 9.3, pp. 99–104. ISSN: 15219615. DOI: 10.1109/MCSE.2007.55.

Johnson, Thomas E. et al. (1984). “Arresting development arrests aging in the nematode Caenorhabditis elegans”. In: *Mechanisms of Ageing and Development* 28.1, pp. 23–40. ISSN: 00476374. DOI: 10.1016/0047-6374(84)90150-7.

Leighton, Daniel H. W. et al. (2014). “Communication between oocytes and somatic cells regulates volatile pheromone production in Caenorhabditis elegans”. In: *Proceedings of the National Academy of Sciences* 111.50, pp. 17905–17910. ISSN: 1091-6490. DOI: 10.1073/pnas.1420439111. URL: http://www.pnas.org/content/111/50/17905.abstract.

Liu, Jie et al. (2013). “Functional aging in the nervous system contributes to age-dependent motor activity decline in C. elegans”. In: *Cell Metabolism* 18.3, pp. 392–402. ISSN: 15504131. DOI: 10.1016/j.cmet.2013.08.007. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1016/j.cmet.2013.08.007.

Lund, James et al. (2002). “Transcriptional profile of aging in C. elegans”. In: *Current Biology* 12.18, pp. 1566–1573. ISSN: 09609822. DOI: 10.1016/S0960-9822(02)01146-6.

Magalhães, Jp De, Ce Finch, and G Janssens (2010). “Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions”. In: *Ageing research reviews* 9.3, pp. 315–323. ISSN: 1872-9649. DOI: 10.1016/j.arr.2009.10.006.Next-generation.

McCormick, Mark et al. (2012). “New genes that extend Caenorhabditis elegans’ lifespan in response to reproductive signals”. In: *Aging Cell* 11.2, pp. 192–202. ISSN: 14749718. DOI: 10.1111/j.1474-9726.2011.00768.x. arXiv: NIHMS150003.

McGee, Matthew D. et al. (2011). “Loss of intestinal nuclei and intestinal integrity in aging C. elegans”. In: *Aging Cell* 10.4, pp. 699–710. ISSN: 14749718. DOI: 10.1111/j.1474-9726.2011.00713.x.

McKinney, Wes (2011). “pandas: a Foundational Python Library for Data Analysis and Statistics”. In: *Python for High Performance and Scientific Computing*, pp. 1–9.

Morsci, Natalia S., Leonard A. Haas, and Maureen M. Barr (2011). “Sperm status regulates sexual attraction in Caenorhabditis elegans”. In: *Genetics* 189.4, pp. 1341–1346. ISSN: 00166731. DOI: 10.1534/genetics.111.133603.

Mortazavi, Ali et al. (2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226.

Murphy, Coleen T. et al. (2003). “Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans.” In: *Nature* 424.6946, pp. 277–283. ISSN: 00280836. DOI: 10.1038/nature01789.

Murray, John Isaac et al. (2012). “Multidimensional regulation of gene expression in the C . elegans embryo Multidimensional regulation of gene expression in the C . elegans embryo”. In: pp. 1282–1294. ISSN: 1088-9051. DOI: 10.1101/gr.131920.111.

Oliphant, Travis E (2007). “SciPy: Open source scientific tools for Python”. In: *Computing in Science and Engineering* 9, pp. 10–20. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.58.

Pérez, F. and B.E. Granger (2007). “IPython: A System for Interactive Scientific Computing Python: An Open and General- Purpose Environment”. In: *Computing in Science and Engineering* 9.3, pp. 21–29. ISSN: 15219615. DOI: doi:10.1109/MCSE.2007.53..

Pimentel, Harold J et al. (2016). “Differential analysis of RNA-Seq incorporating quantification uncertainty”. In: *bioRxiv*, p. 058164. DOI: 10.1101/058164.

Rangaraju, Sunitha et al. (2015). “Suppression of transcriptional drift extends C. elegans lifespan by postponing the onset of mortality”. In: *eLife* 4.December2015, pp. 1–39. ISSN: 2050084X. DOI: 10.7554/eLife.08833.

Reece-Hoyes, John S. et al. (2005). “A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks.” In: *Genome biology* 6.13, R110. ISSN: 1474-760X. DOI: 10.1186/gb-2005-6-13-r110. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1414109%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Schedl, Tim and Judith Kimble (1988). “fog-2, a germ-line-specific sex determination gene required for hermaphrodite spermatogenesis in Caenorhabditis elegans.” In: *Genetics* 119.1, pp. 43–61. ISSN: 00166731.

Schindler, Adam J., L. Ryan Baugh, and David R. Sherwood (2014). “Identification of Late Larval Stage Developmental Checkpoints in Caenorhabditis elegans Regulated by Insulin/IGF and Steroid Hormone Signaling Pathways”. In: *PLoS Genetics* 10.6, pp. 13–16. ISSN: 15537404. DOI: 10.1371/journal.pgen.1004426.

Seidel, Hannah S. and Judith Kimble (2011). “The oogenic germline starvation response in c. elegans”. In: *PLoS ONE* 6.12. ISSN: 19326203. DOI: 10.1371/journal.pone.0028074.

Stroustrup, Nicholas et al. (2013). “The Caenorhabditis elegans Lifespan Machine.” In: *Nature methods* 10.7, pp. 665–70. ISSN: 1548-7105. DOI: 10.1038/nmeth.2475. arXiv: NIHMS150003. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3865717%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Sulston, J. E. and S. Brenner (1974). “The DNA of *Caenorhabditis elegans*.” In: *Genetics* 77.1, pp. 95–104. ISSN: 00166731.

Sulston, J. E. and H. R. Horvitz (1977). “Post-embryonic cell lineages of the nematode, Caenorhabditis elegans”. In: *Developmental Biology* 56.1, pp. 110–156. ISSN: 00121606. DOI: 10.1016/0012-1606(77)90158-0. URL: http://www.sciencedirect.com/science/article/pii/0012160677901580.

Sulston, J. E., E. Schierenberg, et al. (1983). “The embryonic cell lineage of the nematode Caenorhabditis elegans”. In: *Developmental Biology* 100.1, pp. 64–119. ISSN: 00121606. DOI: 10.1016/0012-1606(83)90201-4. URL: http://www.ncbi.nlm.nih.gov/pubmed/6684600.

Van Der Walt, Stéfan, S. Chris Colbert, and Gaël Varoquaux (2011). “The NumPy array: A structure for efficient numerical computation”. In: *Computing in Science and Engineering* 13.2, pp. 22–30. ISSN: 15219615. DOI: 10.1109/MCSE.2011.37.

Wang, Charles et al. (2014). “The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance.” In: *Nature biotechnology* 32.9, pp. 926–32. ISSN: 1546-1696. DOI: 10.1038/nbt.3001. arXiv: NIHMS150003. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4243706%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Waskom, Michael et al. (2016). “seaborn: v0.7.0 (January 2016)”. In: DOI: 10.5281/zenodo.45133.

Chapter 4

UNDERSTANDING THE SYNMUV PHENOTYPE FROM A TRANSCRIPTOMIC PERSPECTIVE

Chapter 5

A STUDY OF DOSAGE RESPONSE IN *C. ELEGANS* USING TRANSCRIPTOME PROFILING

Chapter 6

STRUCTURAL GENES HAVE TRANSCRIPTIONAL CONSEQUENCES

Chapter 7

GENETIC ANALYSIS OF A PATHWAY USING TRANSCRIPTOMIC TRANS-PHENOTYPES

Chapter 8

TISSUE ENRICHMENT ANALYSIS FOR *C. ELEGANS* GENOMICS

Chapter 9

PHENOTYPE AND GENE ONTOLOGY ENRICHMENT AS GUIDES FOR DISEASE MODELING IN *C. ELEGANS*

Chapter 10

AN AUTOMATED FRAMEWORK FOR TRANSCRIPTIONAL PROFILING

Appendix A

QUESTIONNAIRE

Appendix B

CONSENT FORM

[1] Endnotes are notes that you can use to explain text in a document.

POCKET MATERIAL: MAP OF CASE STUDY SOLAR SYSTEMS