Genetic Analysis of a Metazoan Pathway using Transcriptomic Phenotypes

David Angeles-Albores[1, 2,†] Carmie Puckett Robinson[1, 2, 3,†] Brian A. Williams[1]
Barbara J. Wold[1] Paul W. Sternberg[1, 2, *]

May 28, 2017

† These authors contributed equally to this work
1 Division of Biology and Biological Engineering, Caltech, Pasadena, CA, 91125, USA
2 Howard Hughes Medical Institute, Caltech, Pasadena, CA, 91125, USA
3 Department of Neurology, Keck School of Medicine, University of Southern California, Los Angeles, California, 90033, USA
***** Corresponding author. Contact: pws@caltech.edu

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. Interpreting these high-dimensional measurements is unintuitive. We developed a single coefficient to quantify transcriptome-wide epistasis that accurately reflects the underlying interactions. To demonstrate our approach, we sequenced four single and two double *Caenorhabditis elegans* mutants. From these mutants, we reconstructed the known hypoxia pathway. In addition, we uncovered a class of 56 genes that have opposing changes in expression in *egl-9(lf)* and *vhl-1(lf)* but the *egl-9(lf);vhl-1(lf)* mutant has the same phenotype as *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 epistatic 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)[1]. 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[2].

Recently, biological studies have shifted in focus from studying single genes to studying all genes in parallel. In particular, RNA-seq[3] 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[4], better and faster abundance quantification[5,6,7], and improved differential expression analysis methods[8,9]. RNA-seq has been successfully used to identify genetic modules involved in a variety of processes, including T-cell regulation[10,11], the *Caenorhabditis elegans* linker

cell migration[12], and planarian stem cell maintenance[13,14]. For the most part, the role of transcriptional profiling has been restricted to target gene identification.

Transcriptional profiling remains a descriptive tool, and there are only a few examples where transcriptomes have been used to generate genetic models. In population genetics, eQTL studies have established the power of transcriptomes for genetic mapping[15,16,17,18]. However, genetic pathway analysis via epistasis has only been performed once in *Saccharomyces cerevisiae*[19] and once in *Dictyostelium discoideum*[20]. More recently, one group described a protocol for epistasis analysis of a single cell type using single-cell RNA-seq[21]. Epistasis analysis of single cells or single-celled organisms is popular because of the fear that whole-organism sequencing will mix information from multiple cell types and prevent accurate reconstruction of genetic interactions. However, using whole-organism transcriptome profiling, we have recently identified a new developmental state of *C. elegans* caused by loss of a single cell type (sperm cells)[22], which suggests that whole-organism transcriptome profiling contains sufficient information for epistatic analysis. To investigate the ability of whole-organism transcriptomes to serve as quantitative phenotypes for epistatic analysis in metazoans, we sequenced the transcriptomes of of four well-characterized loss-of-function mutants in the *C. elegans* hypoxia pathway[23,24,25,26].

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[27] and are highly conserved in metazoans[28]. A common mechanism for hypoxia-response induction is heterodimerization between a HIFα and a HIFβ subunit; the heterodimer then initiates transcription of target genes[29]. 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*[26], and a single HIFβ gene, *ahr-1*[30]. HIF target genes have been implicated in a wide variety of cellular and extracellular processes including glycolysis, extracellular matrix modification, autophagy and immunity[31,32,33,34,27].

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[35]. 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*[36]. 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, a homolog of sulfhydrylases/cysteine synthases; and CYSL-1 activity is in turn inhibited at by the putative transmembrane O-acyltransferase RHY-1, possibly by post-translational modifications to CYSL-1[37] (see Fig. 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`.

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(sa307)*, *rhy-1(ok1402)*, *vhl-1(ok161)*, *hif-1(ia4)*. We also sequenced the transcriptomes of two double mutants, *egl-9; vhl-1* and *egl-9 hif-1* 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-animal RNA-seq of these mutants at a moderate sequencing depth ($\sim$ 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

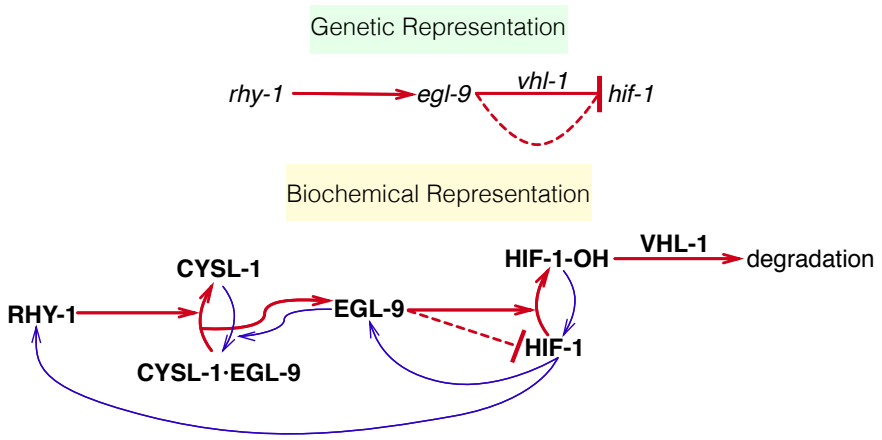


Figure 1. Genetic and biochemical representation of the hypoxia pathway in *C. elegans*. Red arrows are arrows that lead to inhibition of HIF-1, and blue arrows are arrows that increase HIF-1 activity or are the result of HIF-1 activity. EGL-9 is known to exert *vhl-1*-dependent and independent repression 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.

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[22], 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 (excluding wild-type, which is used as baseline). 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 (i.e. greater than 0 or less than 0). β coefficients are analogous, thought not equal to, the logarithm of the fold-change between mutants and wild-type. Larger magnitudes of β correspond to larger perturbations (see Fig. 2). These coefficients can be used to study the RNA-seq data in question. When we refer to β coefficients and q-values, it will always be in reference to isoforms. However, we report the sizes of each gene set in by the number of genes they contain, not isoforms. For the case of *C. elegans*, this difference is negligible since the great majority of protein-coding genes have a single isoform, and most of the remaining protein-coding genes have two, with relatively few genes having more. We have opted for this method of referring to gene sets because it simplifies the language considerably.

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 2,549 genes. Similarly, 3,005 genes were differentially expressed in *rhy-1(lf)* mutants. The *vhl-1(lf)* transcriptome showed considerably fewer differentially expressed genes (1,275), possibly because it is a weaker controller of *hif-1(lf)* than *egl-9(lf)*[25]. The *egl-9(lf);vhl-1(lf)* double mutant transcriptome showed 3,654 differentially expressed genes. The *hif-1(lf)* mutant also showed a transcriptomic phenotype involving 1,075 genes. The *egl-9(lf) hif-1(lf)* double mutant showed a similar number of genes with altered expression (744 genes, see Table 1).

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[38]. 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

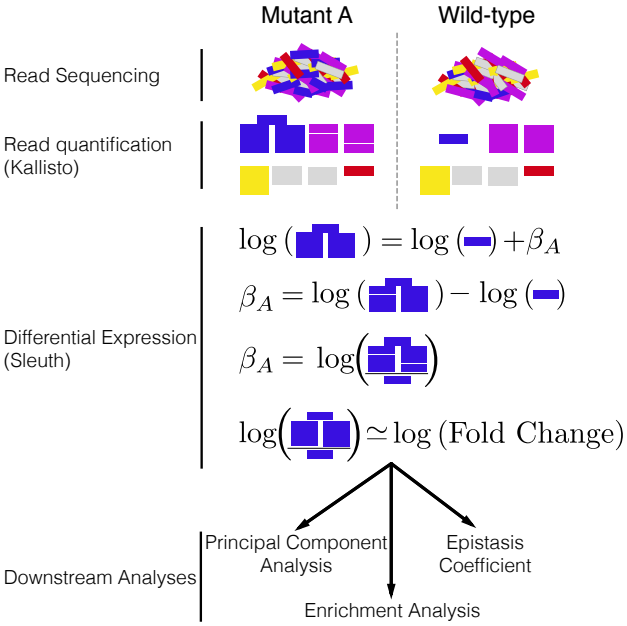


Figure 2. Workflow used to quantify changes in isoform expression. After sequencing, reads are quantified using Kallisto. Differential expression is calculated using Sleuth, which outputs one β coefficient per detected isoform per perturbation or genotype except wild-type. β coefficients are analogous to, though not exactly the same as, the natural logarithm of the fold-change. Downstream analyses can be performed after identifying β coefficients that are statistically significantly different from 0. This can be done because each β coefficient is associated with a q-value. Q-values less than 0.1 are considered statistically different from 0.

Genotype	Differentially Expressed Genes
egl-9(lf)	2,549
rhy-1(lf)	3,005
vhl-1(lf)	1,275
hif-1(lf)	1,075
egl-9(lf);vhl-1(lf)	3,654
egl-9(lf) hif-1(lf)	744
fog-2(lf)	2,840

Table 1. Number of differentially expressed genes in each mutant with respect to wildtype (N2).

dimensions often indicates biological relationships between the data, although interpreting PCA dimensions can be difficult.

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. 3; *fog-2* is not reported to act in the hypoxia pathway and acts as a negative control). Therefore expression profiling measures enough signal to cluster genes in a meaningful manner in complex metazoans.

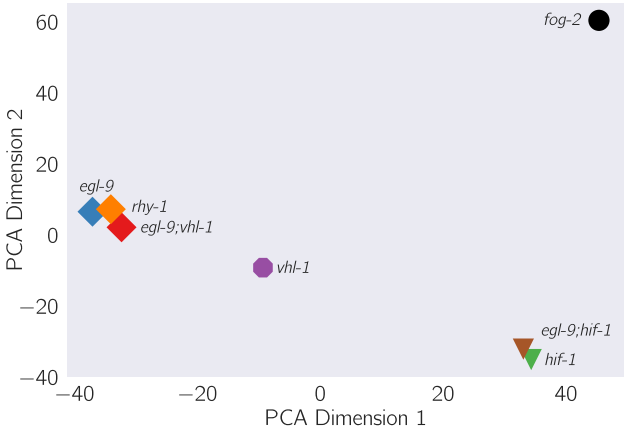


Figure 3. Principal component analysis of various *C. elegans* mutants. Genotypes that have an activated hypoxia response (*i.e*, *egl-9(lf)*, *vhl-1(lf)*, and *rhy-1(lf)*) cluster far from *hif-1(lf)*. *hif-1(lf)* clusters with the suppressed *egl-9(lf) hif-1(lf)* double mutant. The *fog-2(lf)* transcriptome, 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 do not act on the same phenotype, these mutants are independent. If they are not independent, then we must measure whether these genes act additively or epistatically on the phenotype of interest; 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. To allow coherent comparisons of different mutant transcriptomes (the phenotype we are studying here), we define the shared transcriptomic phenotype (STP) as the set of genes or isoforms that are differentially expressed in two mutants being compared, regardless of the direction of change.

Genes in the hypoxia mutant act on the same transcriptional phenotype

We observed that all the hypoxia mutants had a significant STP: the 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[22]. 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

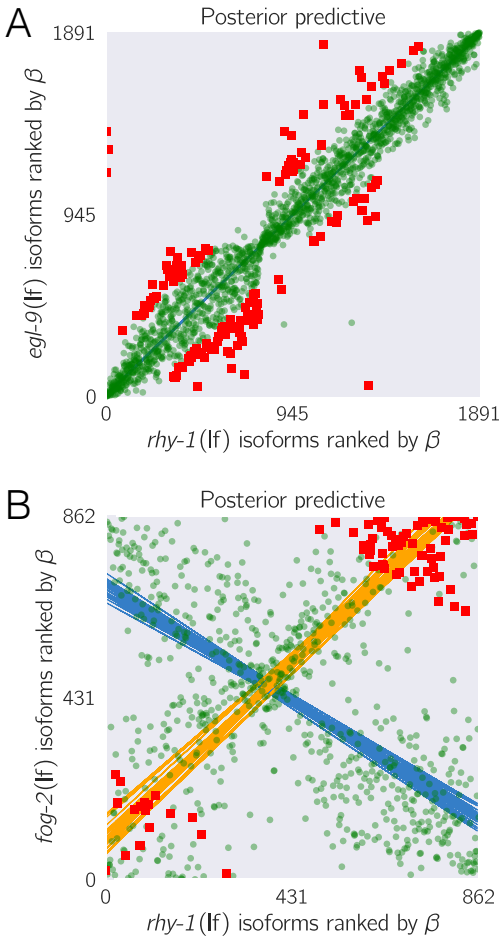


Figure 4. Strong transcriptional correlations can be identified between genes that share a positive regulatory connection. **A**. We obtained the *egl-9(lf)* and the *rhy-1(lf)* transcriptomes, identified differentially expressed genes common to both transcriptomes and ranked each gene according to its 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. **B**. For comparison, we followed the same procedure with the *fog-2(lf)* and *rhy-1(lf)* transcriptomes. *fog-2(lf)* is not known to interact with the primary hypoxia pathway. Notice that although there are correlations between both transcriptomes, the points have considerably more scattered than in **A**. Green, transparent large points mark inliers to the primary regressions (blue lines); red squares mark outliers to the primary regressions and orange lines represent the secondary correlations involving the outliers.

outliers and plotted the results in a rank plot (see Fig 4). 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 best fit. Although *hif-1* is known to be genetically repressed by *egl-9*, *rhy-1* and *vhl-1* [23,24], all the correlations between mutants of these genes and *hif-1(lf)* were positive.

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 5). 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 act positively 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[1]. 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 in an unbranched pathway, 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 it integrates information from many different genes or isoforms into a single number (see Fig. 5).

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 epistatic 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), 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. 5) 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)*

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

= *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 both *vhl-1*-dependent and *vhl-1*-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)[25]. 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 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

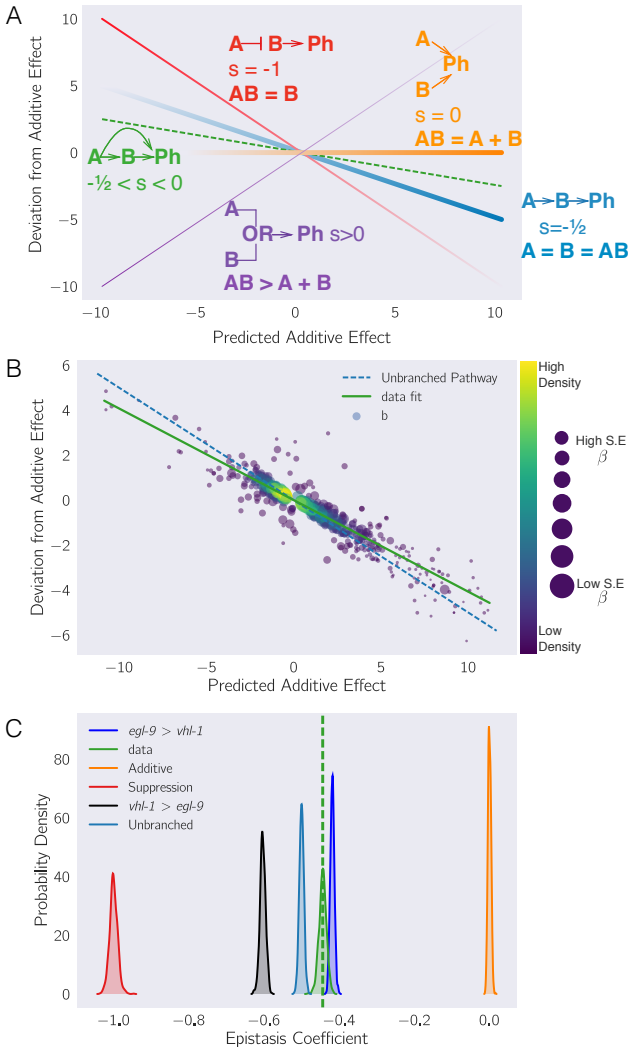


Figure 5. (**A**) Schematic diagram of an epistasis plot. The X-axis on an epistasis plot is the expected coefficient for a double mutant under an additive model (null model). The Y-axis plots deviations from this model. Double mutants that deviate in a systematic manner from the null model exhibit transcriptome-wide 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 $s \to \infty$, 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 epistasis 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).

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. 6). 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 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. 4), other rank plots showed a discernible cross-pattern (see Fig. 7). 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 suppressors of *hif-1(lf)*. Such cross-patterns could be indicative of feedback loops or other complex interaction patterns.

If the above is correct, then it should be possible to identify *egl-9*-independent, *rhy-1*-dependent target genes in a logically consistent way. One misleadingly simple way to identify these targets is via subtractive logic. Using subtractive logic, we would identify genes that are differentially expressed in in *rhy-1(lf)* mutants but not in *egl-9(lf)* mutants. Such a gene set would consist of almost 700 genes. On the one hand, given the q-value of < 0.1 for our RNA-seq analysis, it certainly is reasonable to assume that many of these genes are genuinely affected by *rhy-1(lf)*. However, a major pitfall of subtractive logic is that it fails to take into account the possibility that feedback loops exist between the genes in question. One way to obviate the feedback loops between *rhy-1* and *egl-9* would be to generate a double *rhy-1(lf)*;*egl-9(lf)* mutant and identify differentially expressed genes common to both *rhy-1(lf)* and *rhy-1(lf)*;*egl-9(lf)* but not differentially expressed in *egl-9(lf)* mutants. This would reveal RHY-1 target genes that likely belong to biological pathways entirely distinct from the hypoxia pathway, which is the subject of our analysis. We need to dissect the transcriptomic relationship between two genes both falling in a single pathway, before we can begin to identify target genes which expression is dependent on one gene and independent of the other within that pathway.

rhy-1 and *egl-9* share a well-defined relationship. RHY-1 inhibits CYSL-1, which in turn inhibits EGL-9[37]. 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 'off'; 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* (655 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

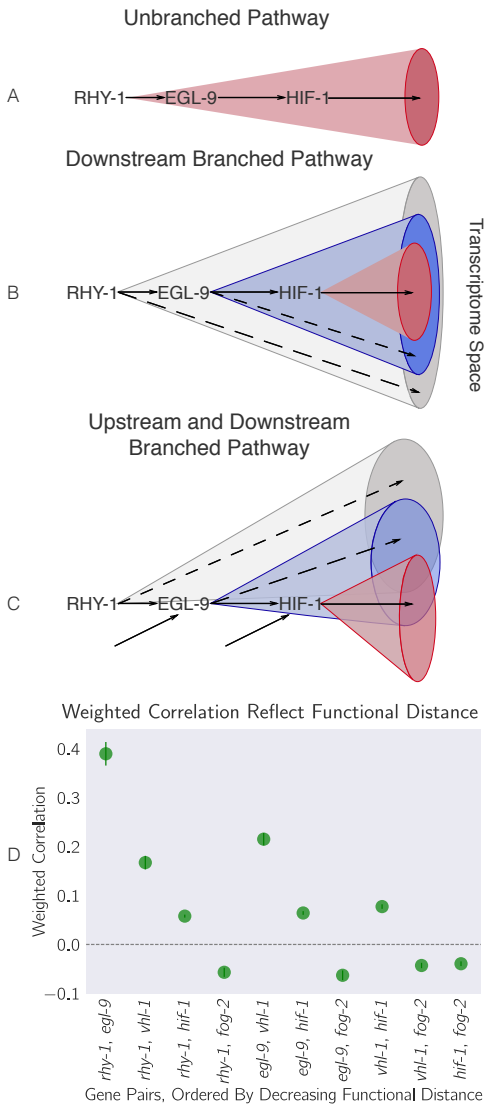


Figure 6. Theoretically, transcriptomes can be used to order genes in a pathway under certain assumptions. Arrows in the diagrams above are intended to show the direction of flow, and do not indicate valence. **A**. A linear pathway in which *rhy-1* is the only gene controlling *egl-9*, which in turn 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.

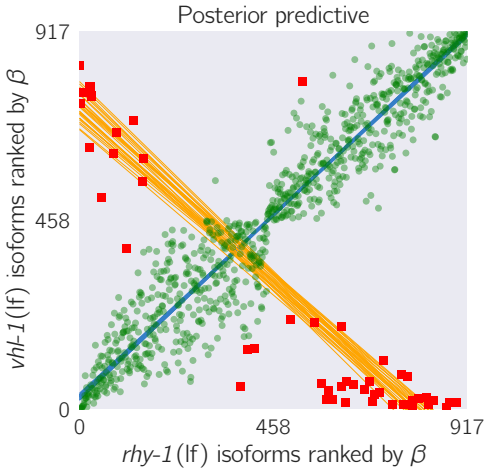


Figure 7. A feedback loop can generate transcriptomes that are both correlated and anti-correlated. The *vhl-1(lf)/rhy-1(lf)* STP shows a cross-pattern. Green large points are inliers to the first regression. Red squares are outliers to the first regression. Only the red small points were used for the secondary regression. Blue lines are representative samples of the primary bootstrapped regression lines. Orange lines are representative samples of the secondary bootstrapped regression lines.

included five transcription factors (*W02D7.6*, *nhr-57*, *ztf-18*, *nhr-135* and *dmd-9*; Supplementary Table 1)

hif-1-independent effects of *egl-9* have been reported previously[39], 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 71 genes (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 312 genes which have consistently increased expression in mutants with a constitutive hypoxic response (Supplementary Table 3).

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 1,075 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 56 that satisfied this condition (see Fig. 8, Supplemental Table 4). Additionally, many of these genes exhibited a new kind of epistasis: *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.

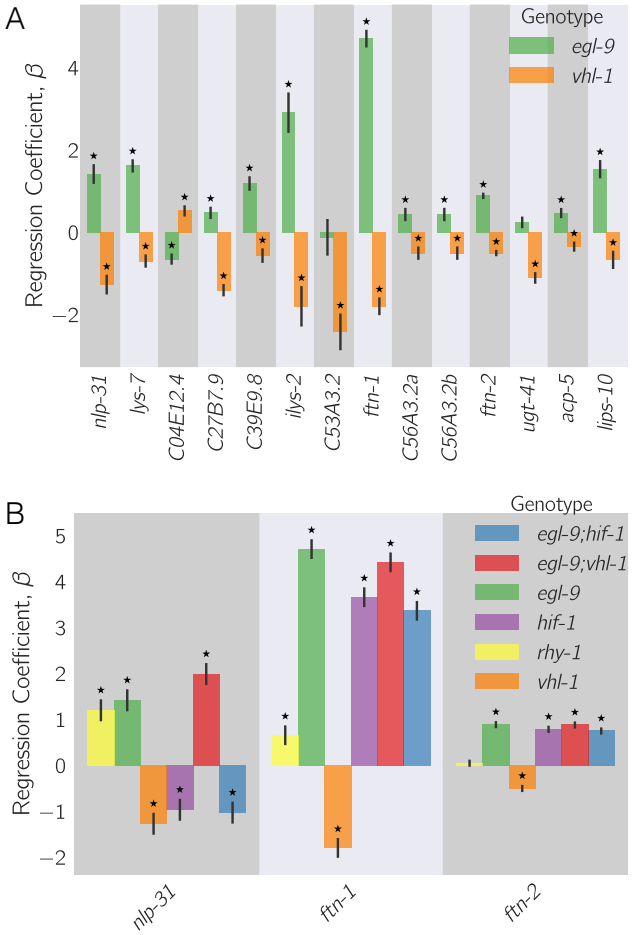


Figure 8. **A**. 56 genes in *C. elegans* exhibit non-classical epistasis in the hypoxiapathway, characterized by opposite effects on gene expression, relative to the wild-type, of of the *vhl-1(lf)* compared to *egl-9(lf)* (or *rhy-1(lf)*) mutants. Shown are a random selection of 15 the 56 genes for illustrative purposes. **B**. Genes that behave non-canonically have 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 ($q < 10^{-1}$).

To illustrate this, we focused on three genes, *nlp-31*, *ftn-1* and *ftn-2*, which had 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[40,41], 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[40] 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.

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.

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*.

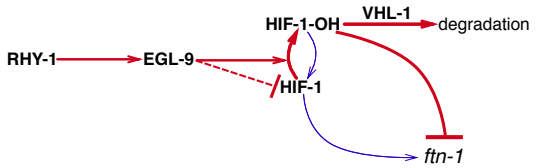
Interpretation of the non-classical epistasis in the hypoxia pathway

The observation of 56 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[40,41,42], but no satisfactory mechanism has been proposed to explain this biology. Previous studies[41,40] 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* and deletion of *egl-9* both cause induction of *ftn-1* expression, but deletion of *vhl-1* abolishes this induction. Moreover, Luchachack et al[42] 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[43], 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. 9).

HIF-1-hydroxyl is challenging to study genetically, and this may have prevented its possible role in the hypoxia pathway from being detected. No mimetic mutations are available with which to study the pure hydroxylated HIF-1 species, and mutations in the Von-Hippel Landau gene that stabilize the hydroxyl species also increase the quantity of non-hydroxylated HIF-1 by mass action. Because HIF-1 is detected low levels

A



B

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 9. 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.

in cells under normoxic conditions[44], 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 1,075 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[45]. Since the absolute levels of HIF-1 are low in normoxic conditions, small fluctuations in protein copy-number 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[46,47].

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. Transcriptome profiling methods afford a lot of information, but transcriptome-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. Although PCA showed that there is information hidden in these genotypes, it is 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 sensitive to outliers. Although outliers can be mitigated via rank-transformations or other methods, these approaches cannot remove outliers resulting from systematic variation caused,for example, by feedback loops. Such interactions can lead to vanishing correlations if both are equally strong and should be accounted for. 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.

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. GLMs can quantify the epistatic effect of an interaction on single genes. We and others[21,22] 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[48] 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 transcriptome-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 20°C and eggs harvested by sodium hypochlorite treatment. Eggs were plated on 6–9 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[49]. 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[49], 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 estimated counts 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:

$$y_{t,i} = \beta_{t,0} + \beta_{t,genotype} \cdot X_{t,i} + \beta_{t,batch} \cdot Y_{t,i} + \epsilon_{t,i} \quad (1)$$

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 [50,51,52,53,54,55,56,57,58]. Our analysis is available in a Jupyter Notebook [59]. 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 [54] (`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.

Epistatic 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,

$$\beta_{\text{Add},i} = \beta_{X,i} + \beta_{Y,i}. \quad (2)$$

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:

$$\beta_{XY} = \sum_{g \in G} \lambda_g \beta_g, \quad (3)$$

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:

$$P(\alpha \mid M_1, D) \propto \prod_{(x_i,y_i,\sigma_i)\in D} \exp\left[\frac{(y_i - \alpha \cdot x_i)^2}{2\sigma_i^2}\right] \cdot (1+\alpha^2)^{-3/2}, \tag{4}$$

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 4 to obtain the most likely slope given the data, D (`scipy.optimize.minimize` in Python). Finally, we approximated the odds ratio as:

$$OR = \frac{P(D \mid \alpha^*, M_1) \cdot (2\pi)^{1/2} \sigma_{\alpha^*}}{P(D \mid M_i)}, \tag{5}$$

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[60,57].

Acknowledgements

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 would not 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. We also thank Andrés Ortiz and Erich Schwarz for invaluable conversations regarding the extension of phenotypes in multidimensional spaces.

Competing interests:

The authors declare they have no conflicts of interest.

References

1. Huang, L. S. & Sternberg, P. W. Genetic dissection of developmental pathways. *WormBook : the online review of* C. elegans *biology* 1–19 (2006).

2. Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**, 855–867 (2008).

3. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).

4. Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31–46 (2010).

5. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology* **32**, 462–464 (2014).

6. Bray, N. L., Pimentel, H. J., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525–7 (2016).

7. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. *bioRxiv* 021592 (2016).

8. Pimentel, H. J., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv* 058164 (2016).

9. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology* **31**, 46–53 (2013).

10. Singer, M. *et al.* A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. *Cell* **166**, 1500–1511.e9 (2016).

11. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–40 (2013).

12. Schwarz, E. M., Kato, M. & Sternberg, P. W. Functional transcriptomics of a migrating cell in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 16246–51 (2012).

13. Van Wolfswinkel, J. C., Wagner, D. E. & Reddien, P. W. Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment. *Cell Stem Cell* **15**, 326–339 (2014).

14. Scimone, M. L., Kravarik, K. M., Lapan, S. W. & Reddien, P. W. Neoblast specialization in regeneration of the planarian *Schmidtea mediterranea*. *Stem Cell Reports* **3**, 339–352 (2014).

15. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296** (2002).

16. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).

17. Li, Y. *et al.* Mapping Determinants of Gene Expression Plasticity by Genetical Genomics in *C. elegans*. *PLoS Genetics* **2**, e222 (2006).

18. King, E. G., Sanderson, B. J., McNeil, C. L., Long, A. D. & Macdonald, S. J. Genetic Dissection of the *Drosophila melanogaster* Female Head Transcriptome Reveals Widespread Allelic Heterogeneity. *PLoS Genetics* **10**, e1004322 (2014).

19. Hughes, T. R. *et al.* Functional Discovery via a Compendium of Expression Profiles. *Cell* **102**, 109–126 (2000).

20. Van Driessche, N. *et al.* Epistasis analysis with global transcriptional phenotypes. *Nature Genetics* **37**, 471–477 (2005).

21. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).

22. Angeles-Albores, D. *et al.* Transcriptomic Description of an Endogenous Female State in *C. elegans*. *bioRxiv* (2016).

23. Epstein, A. C. R. *et al.* *C. elegans* EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation. *Cell* **107**, 43–54 (2001).

24. Shen, C., Shao, Z. & Powell-Coffman, J. A. The *Caenorhabditis elegans rhy-1* Gene Inhibits HIF-1 Hypoxia-Inducible Factor Activity in a Negative Feedback Loop That Does Not Include *vhl-1*. *Genetics* **174**, 1205–1214 (2006).

25. Shao, Z., Zhang, Y. & Powell-Coffman, J. A. Two Distinct Roles for EGL-9 in the Regulation of HIF-1-mediated gene expression in *Caenorhabditis elegans*. *Genetics* **183**, 821–829 (2009).

26. Jiang, H., Guo, R. & Powell-Coffman, J. A. The *Caenorhabditis elegans* hif-1 gene encodes a bHLH-PAS protein that is required for adaptation to hypoxia. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 7916–7921 (2001).

27. Semenza, G. L. Hypoxia-inducible factors in physiology and medicine. *Cell* **148**, 399–408 (2012).

28. Loenarz, C. *et al.* The hypoxia-inducible transcription factor pathway regulates oxygen sensing in the simplest animal, *Trichoplax adhaerens*. *EMBO reports* **12**, 63–70 (2011).

29. Jiang, B. H., Rue, E., Wang, G. L., Roe, R. & Semenza, G. L. Dimerization, DNA binding, and transactivation properties of hypoxia-inducible factor 1. *The Journal of biological chemistry* **271**, 17771–17778 (1996).

30. Powell-Coffman, J. A., Bradfield, C. A. & Wood, W. B. *Caenorhabditis elegans* Orthologs of the Aryl Hydrocarbon Receptor and Its Heterodimerization Partner the Aryl Hydrocarbon Receptor Nuclear Translocator. *Proceedings of the National Academy of Sciences* **95**, 2844–2849 (1998).

31. Semenza, G. L., Roth, P. H., Fang, H. M. & Wang, G. L. Transcriptional regulation of genes encoding glycolytic enzymes by hypoxia-inducible factor 1. *The Journal of Biological Chemistry* **269**, 23757–63 (1994).

32. Bishop, T. *et al.* Genetic Analysis of Pathways Regulated by the von Hippel-Lindau Tumor Suppressor in *Caenorhabditis elegans*. *PLoS Biology* **2** (2004).

33. Shen, C., Nettleton, D., Jiang, M., Kim, S. K. & Powell-Coffman, J. A. Roles of the HIF-1 hypoxia-inducible factor during hypoxia response in *Caenorhabditis elegans*. *Journal of Biological Chemistry* **280**, 20580–20588 (2005).

34. Bellier, A., Chen, C. S., Kao, C. Y., Cinar, H. N. & Aroian, R. V. Hypoxia and the hypoxic response pathway protect against pore-forming toxins in *C. elegans*. *PLoS Pathogens* **5** (2009).

35. Huang, L. E., Arany, Z., Livingston, D. M. & Franklin Bunn, H. Activation of hypoxia-inducible transcription factor depends primarily upon redox-sensitive stabilization of its α subunit. *Journal of Biological Chemistry* **271**, 32253–32259 (1996).

36. Kaelin, W. G. & Ratcliffe, P. J. Oxygen Sensing by Metazoans: The Central Role of the HIF Hydroxylase Pathway. *Molecular Cell* **30**, 393–402 (2008).

37. Ma, D. K., Vozdek, R., Bhatla, N. & Horvitz, H. R. CYSL-1 Interacts with the O_2-Sensing Hydroxylase EGL-9 to Promote H_2S-Modulated Hypoxia-Induced Behavioral Plasticity in *C. elegans*. *Neuron* **73**, 925–940 (2012).

38. Yeung, K. Y. & Ruzzo, W. L. Principal component analysis for clustering gene expression data. *Bioinformatics (Oxford, England)* **17**, 763–774 (2001).

39. Park, E. C. *et al.* Hypoxia regulates glutamate receptor trafficking through an HIF-independent mechanism. *The EMBO Journal* **31**, 1618–1619 (2012).

40. Ackerman, D. & Gems, D. Insulin/IGF-1 and hypoxia signaling act in concert to regulate iron homeostasis in *Caenorhabditis elegans*. *PLoS Genetics* **8** (2012).

41. Romney, S. J., Newman, B. S., Thacker, C. & Leibold, E. A. HIF-1 regulates iron homeostasis in *Caenorhabditis elegans* by activation and inhibition of genes involved in iron uptake and storage. *PLoS Genetics* **7** (2011).

42. Luhachack, L. G. *et al.* EGL-9 Controls *C. elegans* Host Defense Specificity through Prolyl Hydroxylation-Dependent and -Independent HIF-1 Pathways. *PLoS Pathogens* **8**, 48 (2012).

43. Shao, Z., Zhang, Y., Ye, Q., Saldanha, J. N. & Powell-Coffman, J. A. *C. elegans* SWAN-1 binds to EGL-9 and regulates HIF-1- mediated resistance to the bacterial pathogen *Pseudomonas aeruginosa* PAO1. *PLoS Pathogens* **6**, 91–92 (2010).

44. Wang, G. L. & Semenza, G. L. Characterization of hypoxia-inducible factor 1 and regulation of DNA binding activity by hypoxia. *The Journal of Biological Chemistry* **268**, 21513–8 (1993).

45. Goentoro, L., Shoval, O., Kirschner, M. W. & Alon, U. The Incoherent Feedforward Loop Can Provide Fold-Change Detection in Gene Regulation. *Molecular Cell* **36**, 894–899 (2009).

46. Hart, Y., Antebi, Y. E., Mayo, A. E., Friedman, N. & Alon, U. Design principles of cell circuits with paradoxical components. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 8346–8351 (2012).

47. Hart, Y. & Alon, U. The Utility of Paradoxical Components in Biological Circuits. *Molecular Cell* **49**, 213–221 (2013).

48. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–5 (2003).

49. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols* **9**, 171–81 (2014).

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

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

52. Oliphant, T. E. SciPy: Open source scientific tools for Python. *Computing in Science and Engineering* **9**, 10–20 (2007).

53. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2012).

54. Salvatier, J., Wiecki, T. & Fonnesbeck, C. Probabilistic Programming in Python using PyMC. *PeerJ Computer Science* **2**, 1–24 (2015).

55. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering* **13**, 22–30 (2011).

56. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* **9**, 99–104 (2007).

57. Angeles-Albores, D., N. Lee, R. Y., Chan, J. & Sternberg, P. W. Tissue enrichment analysis for *C. elegans* genomics. *BMC Bioinformatics* **17**, 366 (2016).

58. Waskom, M. *et al.* seaborn: v0.7.0 (January 2016) (2016).

59. Pérez, F. & Granger, B. IPython: A System for Interactive Scientific Computing Python: An Open and General- Purpose Environment. *Computing in Science and Engineering* **9**, 21–29 (2007).

60. Angeles-Albores, D., Lee, R. Y., Chan, J. & Sternberg, P. W. Phenotype and gene ontology enrichment as guides for disease modeling in C. elegans. *bioRxiv* (2017).