Genetic Analysis of a Metazoan Pathway using Transcriptomic Phenotypes

David Angeles-Albores[a,b], **Carmie Puckett Robinson**[a,b], **Brian Williams**[a], **Igor Antoshechkin**[a], **and Paul W Sternberg**[a,b]

[a]Department of Biology and Biological Engineering, Caltech, Pasadena, USA, 91125; [b]Howard Hughes Medical Institute

This manuscript was compiled on September 14, 2016

RNA-seq is a technology that is commonly used to identify genetic modules that are responsive to a perturbation. In theory, global gene expression could also be used as a phenotype, with all the implications that has for genetic analysis. To that end, we sequenced four single mutants and two double mutants of the hypoxia pathway in *C. elegans*. We successfully analyzed the single mutants in a blinded fashion to predict the genetic relationships between the genes, and used the double mutants as a test of our predictions and to infer the directionality of the relationship. As a result of our analysis, we identified a core set of 400 genes that are involved in the *hif-1*-dependent hypoxia response in the worm.

Genetic Analysis | RNA-seq | *C. elegans* | hypoxia | transcriptomics

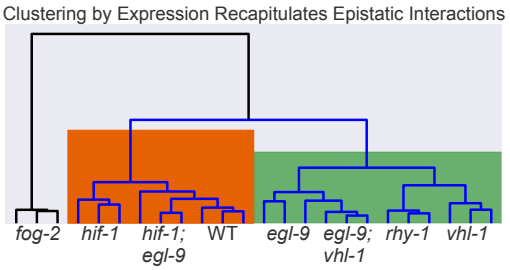

Fig. 1. Blind unsupervised clustering of various *C. elegans* mutants. Genes cluster in a manner that is biologically intuitive. Genes that have visible phenotypes under normoxia cluster away from the wild-type, and when these genes act in the same pathway, they cluster together (i.e, *egl-9*, *vhl-1*, and *rhy-1*). Genes that look wild-type under normoxia cluster near the wild-type..

T

Results

Clustering is a well-known technique in bioinformatics to identify relationships between data. As a first step in our analysis, we wanted to make sure that clustering by Transcripts Per Million (TPM) yielded genetically relevant information. Indeed, when blind, unsupervised clustering was performed on the data, three clusters emerged naturally (see Fig.). *hif-1* and *egl-9*;*hif-1* clustered along with the wild-type; whereas *egl-9*, *egl-9*;*vhl-1*, *vhl-1* and *rhy-1* all clustered away from the wild-type. Finally, our negative control *fog-2* was in its own cluster (see Fig. 1). These clusters make intuitive biological sense: *hif-1* does not have a large role in normoxic circumstances, and is continuously degraded in a normal environment []. As a result, *hif-1* exists only at low levels in a normoxic worm. This strong control on the protein levels of *hif-1* is known to be mediated by a pathway involving *egl-9*, *vhl-1* and *rhy-1*. Whereas the *hif-1* is largely wildtype in normoxic environments, genes that control *hif-1* expression have visible phenotypes. The expectation that *hif-1* should therefore cluster near the wild-type and the control genes should cluster away from the wild-type are therefore realized. Moreover, unsupervised clustering correctly identified epistatic relationships in double mutants: the *hif-1*;*egl-9* double mutant clustered with the wild-type (this double mutant no longer has an *egl* phenotype), and the *egl-9*;*vhl-1* mutant clusters with the *egl-9* and *vhl-1* single mutants. Thus, we conclude that expression data contains enough signal to cluster genes in a meaningful manner.

Theoretically, two genes that have a linear positive interaction should be positively correlated in their overlapping transcriptomes, whereas two genes that have a linear negative interaction should be negatively correlated in their transcriptomes. Formally, if we consider that a gene A has a transcriptome $\{A\}$ associated with it, and if we consider a second gene B with an associated transcriptome $\{B\}$ that is activated by A (that is, $B \in \{A\}$, such that $\{B\} \subset \{A\}$), then it follows that genetic knockout of A or B should both lead to the same perturbation of the transcriptome $\{B\}$. Therefore, it follows that two genes should be strongly positively correlated in the overlap between their transcriptomes if they have positive regulatory associations. Conversely, it follows that if two mutants have overlapping transcriptomes, and if these transcriptomes have a strong positive association, it is likely

Significance Statement

Measurements of global gene expression are often used as descriptive tools that identify genes that are downstream a perturbation. In theory, there is no reason why measurements of global transcriptomes could not be used as a quantitative phenotype for genetic analysis. Here, we show that transcriptomes can be used for epistasis analysis in a metazoan, and that transcriptomes afford far more information per experiment than classic genetic analysis. By using transcriptomes as quantitative phenotypes, we can accurately predict interactions between genes, while at the same time identifying genes common to a pathway. When pathways branch, it is also possible to identify gene batteries that are associated with each end of the branch point. Finally, genes that would result in invisible visible phenotypes in an animal are not likely to be invisible at the transcriptome phenotype due to the exquisite granularity present in these structures, which represents an important advance towards studying small effect genes that make up the majority of animals' genetic repertoire.

DA and PWS wrote the manuscript. CPR performed all experiments. BW performed library preparation. IA performed sequencing. DA performed all analysis and blinded genetic reconstruction.

The authors declare no conflict of interest.

[2]To whom correspondence should be addressed. E-mail: pws@caltech.edu

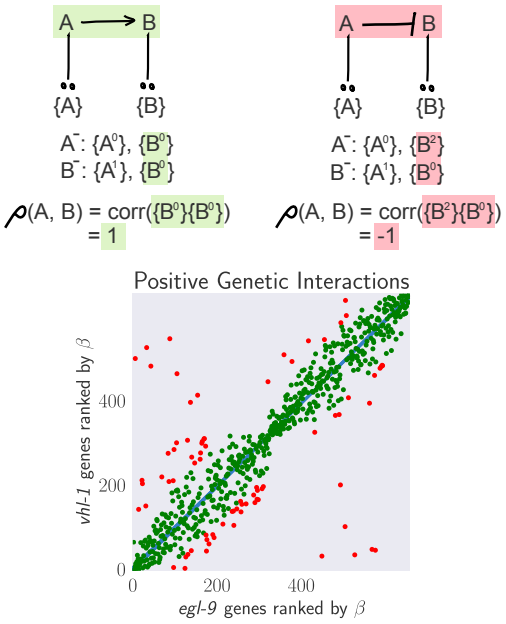

Fig. 2. a. Schematic Diagram showing that genes that interact positively should have a positive transcriptomic correlation, whereas genes that interact negatively should have a negative correlation. Single genes are referred to by their names (A, B), and the transcriptome associated only with gene X is referred to as {X}. We use superscripts to denote expression level. In this case, 0 = no expression (knockout); 1 = WT level; 2 = Greater than WT level. b. Empirical demonstration that transcriptomes between two interacting genes can be extremely well correlated when genes are ranked by expression changes relative to a wild-type.

that these two genes share a positive regulatory association. In other words, transcriptomic correlation is a good predictor of genetic regulation.

Although transcriptomic correlations could theoretically be used for the purposes of identifying genetic regulation, noise from measuring 20,000 genes in multiple different genotypes could cause serious interference with any inferences. Additionally, genes sometimes experience multiple modes of regulation, including positive and negative regulation, from the same gene or pathway. If a positive and a negative signal are both present in a transcriptome, running a naive regression will result in a value close to zero. Therefore, we took steps to mitigate noise in the form of outliers. As a first mitigation attempt, we rank-tranformed the regression coefficients output by Kallisto. This has the effect of mitigating outliers by resetting the difference between adjacent coefficients to unity. Secondly, we performed robust Bayesian regressions using a Student T distribution as a prior. A Student T distribution decays less quickly than a normal distribution, which causes the model to consider outliers to be less informative than traditional frequentist regressions which effectively use a normal prior.

Having mitigated the effect of outliers, we saw that for certain gene pairs, their transcriptomes correlated very well when genes were ranked by their expression changes (see Fig. 2). Having confirmed that we could extract strong signals from these transcriptomes, we proceeded to generate all pairwise correlations between interactomes and we weighted the correlations by the number of genes that participated in the correlation (that were not outliers) divided by the total number of genes detected in all samples. The regression slopes recapitulated a network with three 'modules': A control module, a responder module and an uncorrelated module (see Fig. 3). We were able to identify a strong positive interaction between *egl-9* and *rhy-1*. Part of the reason for this lies in the fact that the transcriptomes for these genes consisted of 1,813 and 2,457 significantly altered genes respectively and the overlap between both genes was quite extensive. On the other hand, none of the correlations between *hif-1* and its controlling genes are negative. On the one hand, we expect that the *hif-1* transcriptome is most susceptible to noise because the protein is expressed at low levels in normoxic environments. However, the *hif-1* transcriptome consists of 937 differentially expressed genes, and the overlap between *hif-1* and all its controlling genes was always greater than 200 genes. Moreover, the unweighted correlation between all the pairwise genes was >0.7 for all comparisons. This means it is unlikely that the positive correlations are purely a result of noise.

In order to rule out noise, we calculated the probability that *hif-1* and its regulatory genes are drawing their transcriptomes from a common pool; in other words, the probability that *hif-1* and the regulatory genes share an isotranscriptome (we use the word isotranscriptomes to refer to two transcriptomes that have the same set of genes, and where these genes change in the same way relative to a control). One way to do this is to take, for example, *egl-9* or *hif-1*, and select whichever transcriptome has a greater number of differentially expressed genes and paint these genes as red balls in an urn, whereas any genes that are not differentially expressed are painted as white balls. Then, we can ask what the probability of selecting N balls in total of which K are red balls is using a hypergeometric function, taking into account that a gene will be red only if it changes in the same way in both transcriptomes. We find that the probability that *hif-1* is interacting positively with *egl-9*, *rhy-1* and *vhl-1* is essentially unity. We conclude that under a normoxic environment, *hif-1* has a positive genetic association with *egl-9*, *rhy-1* and *vhl-1*.

Previous work in the hypoxia pathway suggests that this pathway may have feedback loops. Using the same genetic formalism as above, we realized that interactomes due to the fine-grained nature of the data can identify two regulatory interactions if they are of opposite sign. Consider a system in which an arbitrary gene A activates a gene B, which in turn blocks a gene C. Each gene X has a specific transcriptome *{X}*.Under this system, B and C should have transcriptomes that are negatively correlated. If C activates A, however, then knocking out B should augment expression of C, which should in turn increase expression of A. However, knocking out C should lead to less A, which in turn will lead to less B. Under this thought experiment, suppose that we know the specific transcriptomes associated with A, B and C:*{A}*, *{B}*, *{C}*. Then it must be the case that the genetic knockout of B must have a perturbed transcriptomes $\{A^2\}, \{B^0\}, \{C^2\}$-in other words, knocking out B increases the levels of A, which leads to an overexpression perturbation of the specific transcriptome associated with A, and so forth. On the other hand, knocking out C must lead to the perturbed transcriptomes $\{A^0\}, \{B^0\}, \{C^0\}$. Now, if we were able to correlate each specific transcriptome between correlations, we would find that the specific transcriptomes associated with A and C are anti-correlated; whereas the specific transcriptome associated with B is correlated between both genotypes. This should lead to

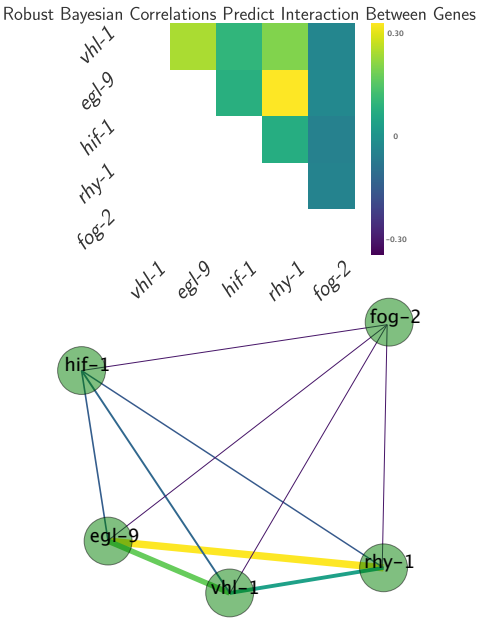

Fig. 3. Top: Heatmap showing pairwise regression values between all single mutants. Bottom: Correlation network drawn from the diagram. Edge width is directly proportional to the regression value.

a characteristic X pattern in the ranked data. Although in this particular example the cross is due to feedback loops, it is important to point out that there are other patterns that could generate these patterns. We investigated whether any pairwise comparisons between our single mutants generated this cross pattern. Indeed, we found that comparing *hif-1* with *rhy-1*, and *hif-1* with *egl-9* yielded negative correlations. In fact, 8/12 possible comparisons showed a cross pattern with correlation values close to 0. However, using a hypergeometric test to examine the probability that these pairs have negative regulatory patterns, we find that the probability of negative regulation for any pair is between 4% and 20%. At this moment, it is unclear whether there are complex regulatory interactions in this set. Strictly speaking, we can only say that the data does not reject the possibility of feedback, or other complex regulatory interactions.

***in silico* qPCR.** Given the uncertain results from our transcriptome-wide analysis of secondary interactions, we wanted additional evidence for feedback loops in hypoxia pathway. We realized that our dataset enabled us to perform a sort of *in silico* qPCR. In order to verify the quality of our data and the veracity of *in silico* qPCR, we first queried the changes in expression of *nhr-57*. This particular reporter has been shown to be under direct control of *hif-1*. Thus, we expected that this gene should go up in *egl-9*, *rhy-1* and *vhl-1*, and it should go down in *hif-1*. Moreover, the behaviour in the double mutants should be interpreted as an epistasis test.

Supporting Information (SI). Because PNAS edits SI and composes it into a single PDF, authors must provide the following file formats only.

SI Text. Supply Word, RTF, or LaTeX files (LaTeX files must be accompanied by a PDF with the same file name for visual

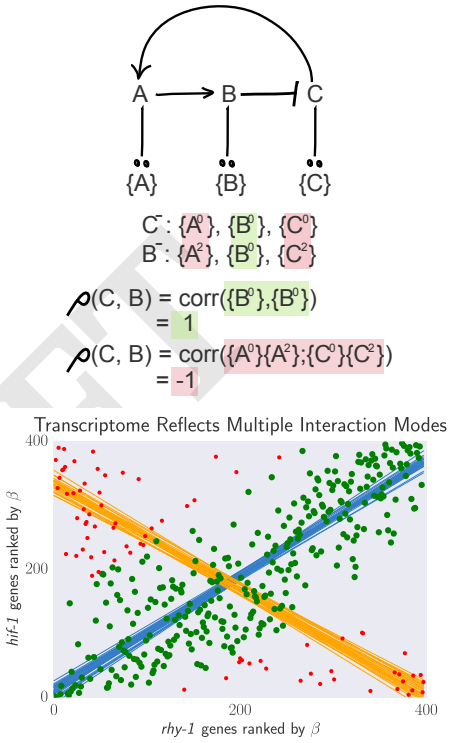

Fig. 4. Top: A feedback loop can generate transcriptomes that are both correlated and anti-correlated. Bottom: *hif-1* transcriptome correlated to the *rhy-1* transcriptome. Green large points are inliers to the first regression. Red small points are outliers to the first regression. Only the red small 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.

reference).

SI Figures. Provide a brief legend for each supporting figure after the supporting text. Provide figure images in TIFF, EPS, high-resolution PDF, JPEG, or GIF format; figures may not be embedded in manuscript text. When saving TIFF files, use only LZW compression; do not use JPEG compression. Do not save figure numbers, legends, or author names as part of the image. Composite figures must be pre-assembled.

SI Tables. Supply Word, RTF, or LaTeX files (LaTeX files must be accompanied by a PDF with the same file name for visual reference); include only one table per file. Do not use tabs or spaces to separate columns in Word tables.

SI Datasets. Supply Excel (.xls), RTF, or PDF files. This file type will be published in raw format and will not be edited or composed.

Materials and Methods

ACKNOWLEDGMENTS. The authors would like to acknowledge the rest of the Sternberg lab for their comments and support.