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 16, 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. We show that genes along a pathway tend to decorrelate as a result of alternative regulatory modes and crosstalk with other pathways; and that this decorrelation accurately reflects functional distance between genes. As a by-product of our analysis, we identified a core set of 400 genes that are involved in the *hif-1*-dependent hypoxia response in the worm.

genetics | RNA-seq | *C. elegans* | hypoxia | transcriptomics

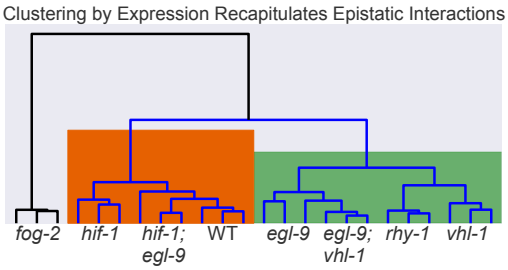

Fig. 1. Blind unsupervised clustering of various *C. elegans* mutants. Genes cluster in a manner that is biologically intuitive. Genes that have visible phenotypes under normoxia cluster away from the wild-type, and when these genes act in the same pathway, they cluster together (i.e, *egl-9*, *vhl-1*,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

T

Results

Clustering visualizes epistatic relationships between genes. 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. *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.

Transcriptomic correlations can predict genetic regulation. 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

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

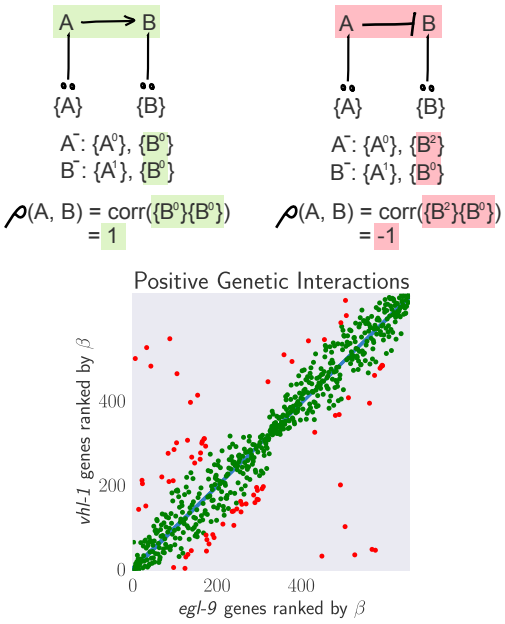

Fig. 2. a. Schematic Diagram showing that genes that interact positively should have a positive transcriptomic correlation, whereas genes that interact negatively should have a negative correlation. Single genes are referred to by their names (A, B), and the transcriptome associated only with gene X 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 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 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

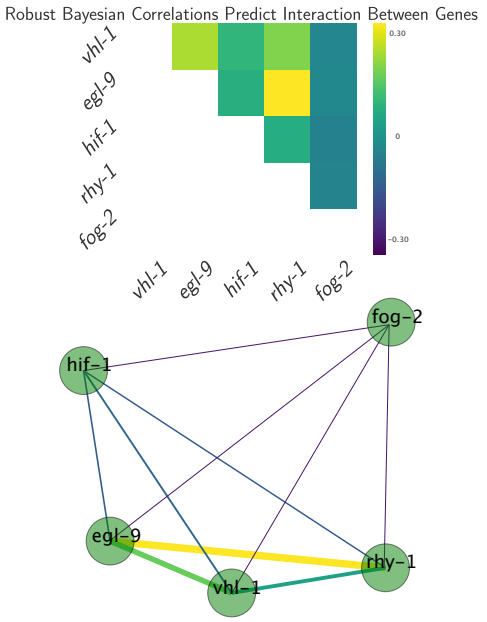

Fig. 3. Top: Heatmap showing pairwise regression values between all single mutants. Bottom: Correlation network drawn from the diagram. Edge width is directly proportional to the regression value.

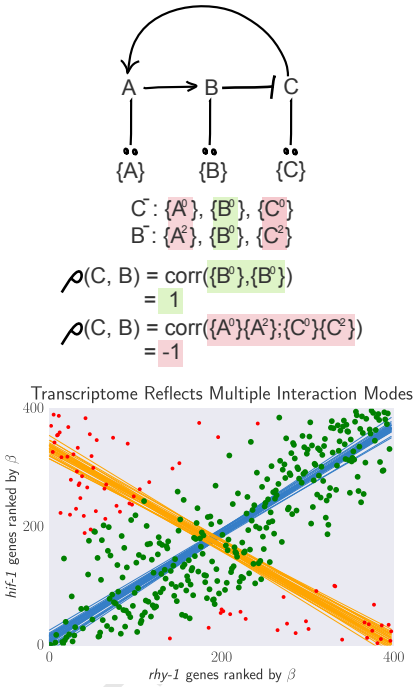

Fig. 4. Top: A feedback loop can generate transcriptomes that are both correlated and anti-correlated. Bottom: *hif-1* transcriptome correlated to the *rhy-1* transcriptome. Green large points are inliers to the first regression. Red small points are outliers to the first regression. Only the red small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

$\{A^0\}, \{B^0\}, \{C^0\}$. 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 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 unweighted correlation values close to 1. 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%.

***in silico* qPCR reveals extensive feedback in the 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 be unchanged in *hif-1*. The epistasis test, using the double *egl-9*;*hif-1* double mutant should result in no change; whereas the *egl-9*;*vhl-1* double mutant should have a similar change to the *vhl-1* and the *egl-9* mutants. In fact, our datasets reflected these known interactions, showing that the RNA-seq measurements can be used in a semi-quantitative fashion to perform inferences on genetic regulation. Next, we decided to perform *in silico* qPCR of every gene under scrutiny in order to get a clearer idea of the relationships between them (see Fig. **??**). We found that *rhy-1*, and to a lesser extent *egl-9* were activated by mutations in *egl-9*, *rhy-1* and *vhl-1*. This suggests that *hif-1* is a positive regulator of *rhy-1*. Given that *rhy-1* post-translationally controls *egl-9* [], it is unlikely that the increase in *egl-9* is driven by the increase in *rhy-1* levels. Therefore, our experiment also suggests that *hif-1* is a positive regulator of *egl-9*. On the other hand, we also discovered that mutation of *hif-1* increased levels of *rhy-1*. This suggests that *hif-1* is also a negative regulator of *rhy-1*. One potential mechanism through which *hif-1* could be both a positive and a negative regulator would be for hydroxylation of *hif-1* to change its activity. Under this mechanism, loss of *hif-1* hydroxylation leads to activation of *rhy-1* and *egl-9* as a homeostatic mechanism; whereas excessive hydroxylation causes inhibition of these genes.

Whereas loss of hydroxylation seems to lead to overexpression of *rhy-1* and *egl-9*, there is no change in *hif-1* levels. The only change in expression level of this gene occurs in the *hif-1* mutant. Therefore, we postulate that *hif-1* positively autoregulates itself only in the hydroxylated state.

Performing the *in silico* experiment with the *egl-9*;*vhl-1* double mutant shows a similar increase in activity of the two genes in question. This provides confirmatory evidence that *hif-1* up-regulates *egl-9*, but also suggests that *egl-9* and *vhl-1* are epistatic to one another. Such epistasis can only occur in one of two ways: Either the genes are acting linearly, or they are acting in AND gated fashion, with both genes required to mediate an effect. Similarly, the *egl-9*;*hif-1* double mutant exhibits the same expression profile as *hif-1*, which means *egl-9* is an inhibitor of *hif-1*.

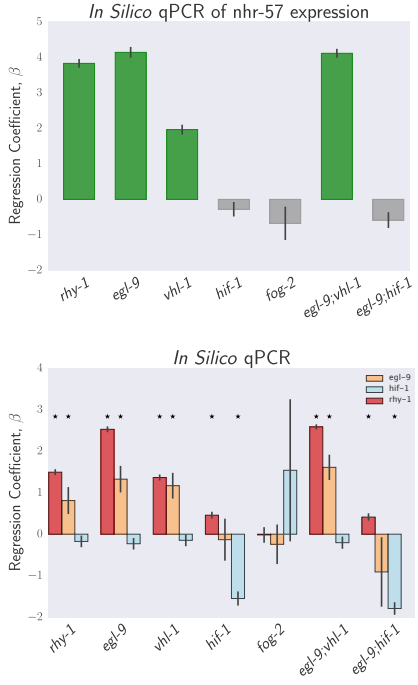

Fig. 5. Top: *In silico* qPCR results using *nhr-57* as an expression reporter. The results from the single mutants suggest that loss of *egl-9* activates *nhr-57* transcription, and the double mutant results show that *egl-9* inhibits *hif-1*; and that *egl-9* and *vhl-1* act in concert to inhibit *hif-1*. Green bars show statistically significant increases in expression relative to wild-type. Bottom: *In silico* qPCR of genes in the hypoxia pathway and a *fog-2* control. These results suggest that *hif-1* activates *rhy-1*, and possibly *egl-9*, when it is not hydroxylated. *hif-1* also appears to autoactivate in a hydroxylation-dependent manner. Like with *nhr-57*, these results support the hypothesis that *egl-9* and *vhl-1* together inhibit the *hif-1* response.

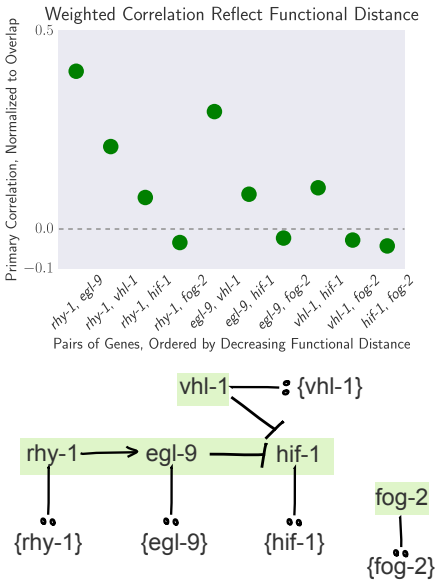

Fig. 6. Top: Pairwise weighted correlations between transcriptomes can be used to infer functional distance between interacting genetic partners. Pairwise correlations are ordered by increasing network distance between genes. Notice that correlations involving the *fog-2* negative control are very near zero. Bottom: Simplified schematic of the hypoxia pathway shown to illustrate functional distance between genes in the pathway.

In summary, the *in silico* qPCR results suggest that *egl-9* and *vhl-1* act in concert to inhibit *hif-1*. Likewise, these results taken together with the transcriptome-wide cross-patterns that emerge from pairwise comparisons between genes in the hypoxia pathway suggest that there are positive and negative feedback loops feeding into *rhy-1* and possibly *egl-9*. These feedack loops explain why *hif-1* is positively transcriptomically correlated with *egl-9*–since we are observing

Transcriptomic decorrelation can be used to infer functional distance. We were interested in figuring out whether RNA-seq could be used to identify functional interactions within a genetic pathway. Although there is no *a priori* reason why global gene expression should reflect functional interactions, we were encouraged by the strength of the unweighted correlations between genes in the hypoxia pathway, and conversely by the weak correlation of these genes with the *fog-2* mutant.

We investigated the possibility that transcriptomic signals might contain relevant information about the degrees of separation by weighting the robust bayesian regression of each pairwise analysis by $N_{\mathrm{Overlap}}/N_{\mathrm{detected}}$. We then 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 decreasing N_{Overlap}. 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. Even in unbranched pathways, this would induce progressive decorrelation between genes proportionally to the distance between them.

Discussion

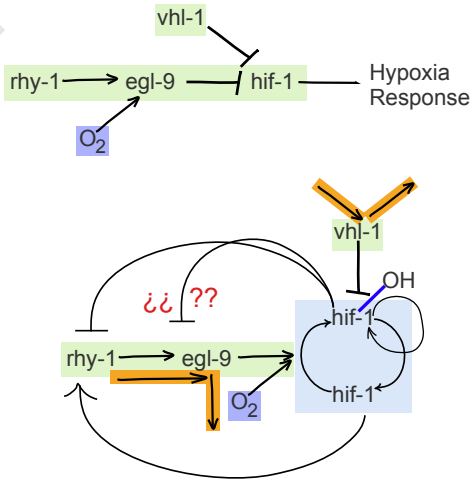

Fig. 7. Top: Bottom:

Materials and Methods

hi. Hello

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