A theory of genetic analysis using transcriptomic phenotypes

Thesis by
David Angeles-Albores

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

[2018]
Defended [18 September 2018]

David Angeles-Albores
ORCID: 0000–0001–5497–8264

ACKNOWLEDGEMENTS

This thesis is possible thanks to the unwavering support of a long list of individuals. I would like to thank my advisor, Paul W. Sternberg, for providing a laboratory and an intellectual home for the past few years. I will always remember our conversations over coffee at 9am in the lab. I taught Systems Genetics with Paul as my project came into full maturity; these lectures on genetics gave us a chance to recast the classical interpretations with a genomics perspective. Those lectures are some of my fondest memories at Caltech. I also need to acknowledge my thesis committee, Dianne K. Newman, Elliot Meyerowitz and Matt Thomson. Without their advice, I would be considerably more confused than I am today. Dianne has been a major figure throughout my Ph.D., a great scientist with a heart to match, and I feel lucky to have had an opportunity to learn from her insights. I also need to thank Erich M. Schwarz, who taught me to argue, and taught me to write. Where others were content to say my work was fine, Erich found every possible loophole, every minor detail and every open question and pushed me to be complete without being redundant. I am sincerely grateful for his guidance and his mentorship.

I have benefitted from a fantastic set of collaborators in the Sternberg lab. I have been lucky to work with Hillel Schwartz, a fantastic geneticist and good friend; Carmie Puckett Robinson, with whom I started to work on transcriptome genetics; Daniel Leighton, who taught me about worm pheromones and aging; Raymond Y. Lee and Juancarlos Chan, with whom I learned all the intricacies of WormBase and tool design. Throughout my time here, I have worked with three extremely talented undergraduates who made working in the lab much more exciting: Tiffany Tsou, Kyung Hoi (Joseph) Min and Vladimir Molchanov. Finally, I need to thank all the members of the Sternberg lab for making science come to life: Jon Liu, Han Wang, James Lee, Pei-Yin, Katie Brugman, Cynthia Chai, Wen Chen, Sarah Cohen,

Elizabeth Holman, Sandy Wong, Heenam Park, Daniel Jun Oh, Ravi Nath, Margaret Ho, Srimoyee Ghosh, Sophie Walton, Sarah Torres, Shahla Gharib, Barbara Perry, Animesh Ray and Elizabeth Glater. I cannot name all of the friends I have made here at Caltech; I hope they know how grateful I am for their friendship.

I would like to briefly acknowledge the programs and the people who brought me to Caltech. I would not be here without the EXtraordinary Research Opportunities Program (EXROP) from HHMI, where I met Andrew Quon and Christy Schultz. Through EXROP I met and had a chance to work for Susan Lindquist and her (then) postdoc Georgios Karras, who taught me the beauty of yeast genetics. I wish I could show Sue what I have done with the doors that she opened for me. At Cornell, I was incredibly lucky to be adviced by Laurel Southard, who believed in my potential no matter what grades might say.

Throughout my time at Caltech, I have never been alone. My family has been a source of unconditional support. I thank my parents, Lilia and Josué, and my brother, Andrés, for always believing in me.

Finally, I would like to give my heartfelt thanks to Heather L. Curtis.

Heather, you have been the sun, and the moon, and the stars in my life since I met you. You brought new colors into my world. Every day, I learn to think in new ways, I learn to see new things, thanks to you. I am a better person because I am with you and I am grateful that life brought us together.

This thesis is for you.

ABSTRACT

This thesis deals with the conceptual and computational framework required to use transcriptomes as effective phenotypes for genetic analysis. I demonstrate that there are powerful theoretical reasons why Batesonian epistasis should feature prominently in transcriptional phenotypes. I also show how to compute and interpret the aggregate statistics for transcriptome-wide epistasis and transcriptome-wide dominance using whole-organism transcriptomic profiles of *C. elegans* mutants. Finally, I developed the WormBase Enrichment Suite for enrichment analysis of genomic data.

RNA-seq as a tool has enormous potential because it relies on protocols that are fast, simple and increasingly cheap. In spite of their potential, transcriptomes have seen their use largely limited to single-factor experiments. Even when many transcriptomes are collected, the main analytic approach is to apply clustering algorithms that correlate responses but do not have any power to identify causal mechanisms.

I demonstrate that if a complete genetic experimental design is used (in the form of a full two-factor matrix), transcriptomes can establish genetic interactions between a pair of genes without the need for clustering algorithms. Surprisingly, when we performed epistasis analyses of hypoxia pathway mutants in *C. elegans* we did not simply observe a generalized epistatic interaction between the mutants. In fact, the transcriptomes recapitulated the same Batesonian epistatic relationship that had been observed using classical phenotypes. In other words, we observed that the transcriptomic phenotype of one gene can be masked by the transcriptomic phenotype of a second gene, such that a double mutant of these two genes has exactly the same phenotype as a single mutant of the epistatic gene. Motivated

by this observation, we developed methods to recognize and interpret Batesonian epistasis at the transcriptomic level. This method relies on the calculation of a single aggregate coefficient that we named the transcriptome-wide epistasis coefficient.

The observation that Batesonian epistasis could be reproduced on a transcriptomic level was surprising. To explain how transcriptome-wide epistasis can arise, I studied a simplified model of transcriptional regulation using statistical mechanics. These studies demonstrate that epistatic analysis is equivalent to a perturbative analysis of the partition function of a promoter. Moreover, these studies revealed that a sufficient condition for Batesonian epistasis to occur is if the two genes encode variables that are transformed and multiplied together to form an effective single compound variable. Finally, these studies clearly demonstrate the connection between statistical (or generalized) epistasis and Batesonian epistasis and establish a physical basis for genetic logic.

Genetic analyses of gene functional units can also be carried out using allelic series in tandem with complementation (also known as dominance) tests. I developed a statistical coefficient known as transcriptome-wide dominance to enable analyses of allelic series using expression profiles. A crucial aspect of allelic series is the ability to enumerate the independent phenotypes associated with an arbitrary set of alleles. I developed the concept of phenotypic classes as a transcriptomic analogue of classical phenotypes for this purpose. Briefly, a phenotypic class is a set of transcripts that are differentially expressed in a specific set of genotypes. Thus, an allelic series consisting of two mutant alleles (and a wild-type) can at most result in 7 phenotypic classes. However, some of these phenotypic classes may be artifactual as a result of the significant false positive and false negative rates that are associated with RNA-seq. I developed a simple algorithm that tries to identify phenotypic classes that are artifactual, though often these classes may also be identified through a critical evaluation of their biological implications. I applied

these concepts to a small allelic series of the *dpy-22* gene, which encodes a Mediator subunit in *C. elegans*, and identified 3–4 functional units along with their sequence requirements.

Finally, I developed the WormBase Enrichment Suite by implementing a hypergeometric test on the tissue, gene and phenotype ontology for *C. elegans*. The importance of this tool derives mainly from its integration to WormBase, the repository of all *C. elegans* knowledge, which means that the databases that are tested will undergo continuous improvement and curation, and thus will yield the most accurate results.

PUBLISHED CONTENT AND CONTRIBUTIONS

Angeles-Albores, David, Raymond Y. N. Lee, et al. (2018). "Two new functions in the WormBase Enrichment Suite". In: *Micropublication: biology. Dataset.* DOI: https://doi.org/10.17912/W25Q2N.

Angeles-Albores, David, Carmie Puckett Robinson, et al. (2018). "Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements". In: *Proceedings of the National Academy of Sciences*, p. 201712387. ISSN: 0027-8424. DOI: 10.1073/pnas.1712387115.

Angeles-Albores, David and Paul W Sternberg (2018). "Using Transcriptomes as Mutant Phenotypes Reveals Functional Regions of a Mediator Subunit in *Caenorhabditis elegans*." In: *Genetics*, genetics.301133.2018. ISSN: 1943-2631. DOI: 10.1534/genetics.118.301133.

Angeles-Albores, David, Daniel H. W. Leighton, et al. (2017). "The *Caenorhabditis elegans* Female-Like State: Decoupling the Transcriptomic Effects of Aging and Sperm Status". In: *G3: Genes, Genomes, Genetics* 7.9.

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

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Number *Page*

PREFACE

I have tried to organize this thesis in a way that makes sense. Briefly, the thesis can be viewed in three segments: Epistasis, complementation and software development. In doing so, I have broken the chronological order of my work, but I do not see this as a problem. Science is rarely linear, and it may often turn out that the last concepts to be found are actually those concepts that allow us to make sense of everything else. This has certainly been the case with my work.

In Chapter 1, the reader will find a brief overview of the problem facing transcriptome genetics. This chapter encompasses a review of the relevant literature, but beyond that, I have tried to make arguments I think are important. First, transcriptome genetics has obviated the chasm between statistical epistasis and classical, or Batesonian, epistasis. The confusion between the two (related) terms has been one of the great misfortunes in the field of genetics, since it has hampered a significant amount of work. I am glad to say that in this thesis I have achieved the unification of both concepts, such that no confusion should happen. Second, although we now know how to search signs of epistasis and dominance in transcriptomes, the issue of counting phenotypic classes or modules is becoming increasingly ominous. Unless and until we can confidently identify and purge spurious modules, we will not be able to use these phenotypes to their full extent.

In Chapter 2, I have written a theoretical argument that is the basis for the rest of the chapters dealing with epistasis. In this chapter, I prove that epistasis emerges from statistical mechanics, such that even genes that have enormously complex transcriptional mechanisms can in some cases exhibit Batesonian epistasis. This chapter establishes genetics as a variational method with which to probe an unknown partition function, enabling us to make statements about what values the partition

function is or is not allowed to take.

In Chapter 3, I develop the concept of transcriptome-wide epistasis and use it to reconstruct the well-studied hypoxia pathway. In Chapter 4, I use the concept of transcriptome-wide epistasis to identify a novel stage in the life cycle of the roundworm *C. elegans*.

Chapter 5 deals with the issue of complementation, and its study through expression profiles. In my opinion, this is the most complicated chapter in this thesis. I struggled with every aspect of this project, but the result is, to my mind, pleasing.

Chapter 6 and 7 deal with the creation of the WormBase Enrichment Suite.

Throughout this thesis, I have tried to be pedagogical. If we don't make efforts to explain the computational methods we are developing, biology will pass from a scientific discipline to an astrological pseudo-science, and we will fail to see the true beauty in the stars above and instead imbue them with our human desires and flaws, asking them to help us reach fame instead of helping us to solve the mysteries that abound in our universe.

Chapter 1

INTRODUCTION

Abstract

Transcriptomes are microscopic phenotypes of enormous complexity. In spite of this complexity, it is becoming apparent that transcriptomes follow the same genetic rules as all other mesoscopic and macroscopic phenotypes. Due to their complexity, the genetic rules that bind transcriptomes appear more complicated. There is significant interest in developing statistical and biological methods that can deconvolute transcriptomes to extract the maximum amount of information encoded within them. Here, we review the basic concepts that underlie transcriptome genetics, identify confusions in the field and point towards the emerging challenges and opportunities associated with these intriguing new phenotypes.

Introduction

The recent explosion in genomic technologies has provided us with unparalleled insight into the inner workings of cells. The cost of of sequencing continues to drop, and new technologies are continuously increasing the number of samples that can be sequenced. In turn, these massive datasets have promoted the appearance of increasingly complex algorithms to make sense of them. A common tenet in these methods has been to reduce the dimensionality of these datasets (dimensionality refers to the number of measurements per sample) to look for trends in the data. Though sometimes these methods are rooted in biological principles, more often they come from algebraic methods that have no immediate connection to the underlying biology. This means that although these methods may be quite powerful, the results

may be hard to interpret in biological terms. Moreover, these methods may not utilize the rich structure inherent to biological systems that could place strong constraints on the problem under study to reduce the space of reasonable solutions.

Biological systems can be daunting in their complexity. In general, there is no way to solve these complex systems from first principles, since the identity and activity of each component is in general not known. For this reason, biologists developed a set of methods, collectively referred to as genetic analyses, that do not make many assumptions regarding the underlying molecular details. Genetics is limited to making a limited set of true statements regarding certain kinds of molecular interactions. Due to its limited scope, genetics is robust to biological variation. A major goal over the last century has been to reconstruct the set of all genetic networks that result in a specific phenotype in a specific condition (the genotype-phenotype map). Though spectacular progress has been made in some cases (Costanzo et al., 2016), we are still far from understanding genetic networks. Previously, generating sufficient genotypes in model organisms to analyze any network in detail was a bottleneck to perform thorough genetic reconstructions. However, with the advent of genome engineering, generating specific mutants is rapidly becoming easier. On the other hand, sensitive and fast phenotyping methods have lagged behind. A possible solution to this problem is bulk expression profiling, but the complexity of expression profiles had proved a daunting challenge for genetic analysis. Furthermore, expression profiles have brought to the forefront a major source of confusion in genetics: The definition of genetic interactions.

Biologists identify genetic interactions between genes using a specific method called epistasis analysis. The term ‘epistasis’ was used for the first time over one hundred years ago by William Bateson (Bateson and Mendel, 2009) to refer to the observation that the distribution of offspring phenotypes from a double heterozygote cross did not match the expected distribution prescribed by Mendelian segregation of two loci.

Under Mendelian laws, if two loci are associated with different phenotypes, crossing double heterozygotes of these two loci should generate animals with four phenotypic classes, with each class occurring in a 9:3:3:1 ratio. Bateson realized through segregation analyses that in certain cases, the phenotypic class associated with the double mutant was missing, and instead there was an excess of one phenotypic class typically associated with homozygotes of one mutant allele, an effect similar to Mendel's observations of allelic dominance. He coined the term epistasis to refer to the effect by which an allele at one locus, when present in two copies, can completely mask the phenotypic effect of another allele at a separate locus.

Since he coined the term, Batesonian or classical epistasis has become a popular tool amongst geneticists with which to identify genetic interactions. An important caveat is that in order to perform an epistasis analysis, geneticists must restrict themselves to alleles that are completely devoid of function. When this is the case, the phenotypic transformation of the double mutant is used to construct a genetic pathway (Avery and Wasserman, 1992; Huang and Sternberg, 2006) (see Fig. 11). Classical epistasis has become a cornerstone of biology.

Classical epistasis means that the phenotype of the double mutant is exactly the same as the phenotype of one of the single mutants. However, the problem can also be recast in quantitative terms. Statistical geneticists defined generalized epistasis as a systematic deviation between the observed values and a null model (usually additive or log-additive) that can be corrected by adding a second order interaction term (Fisher, 1919). In the terms of generalized genetics, epistasis in the heterozygote crosses is measured in the systematic excess of one phenotypic class and the systematic depletion of a second class. Notably, generalized epistasis is not constrained in the values it can take, and it is not constrained to measurements of population properties or properties of single individuals.

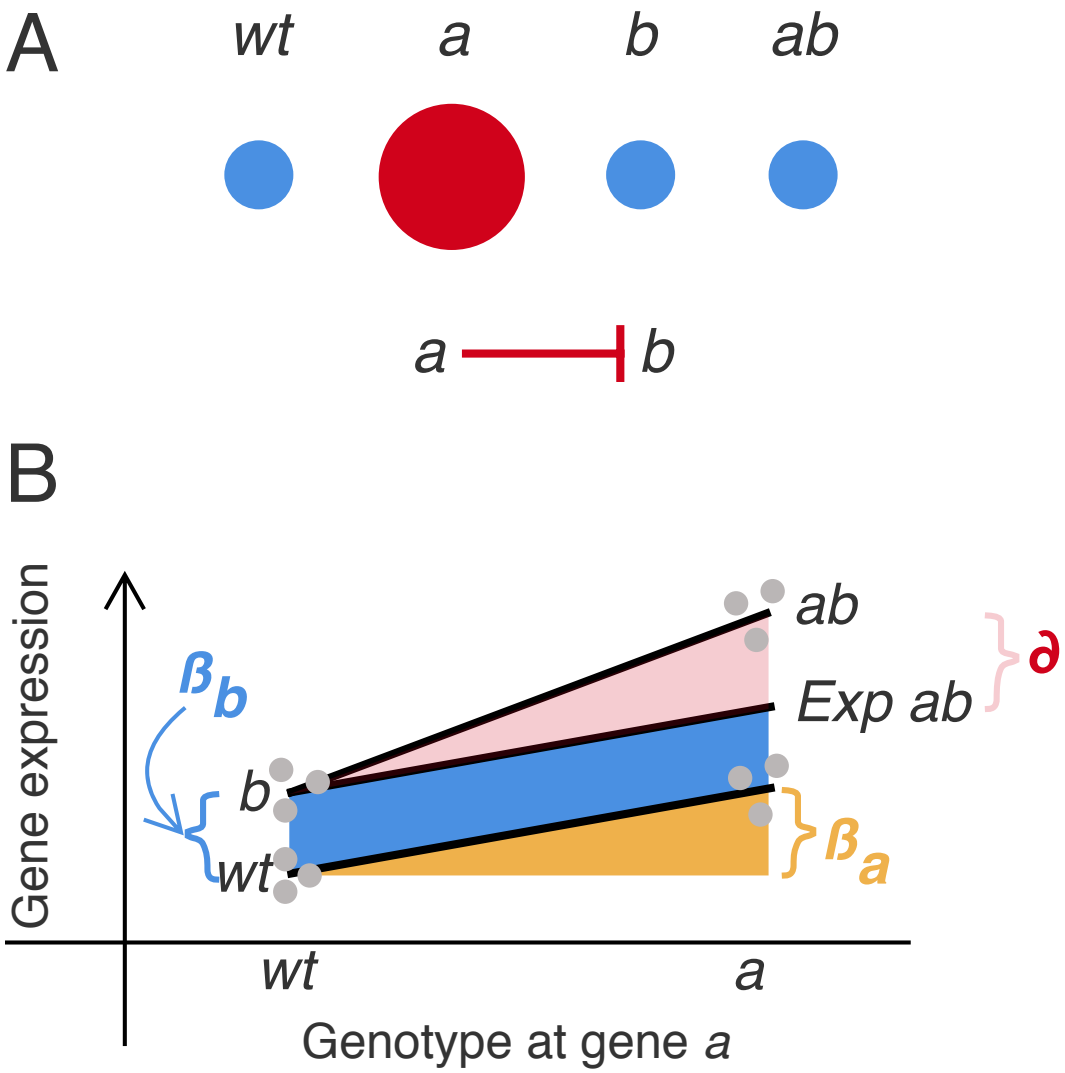


Figure 11 Biologists work with two distinct types of epistasis. **A**. Batesonian, or classical, epistasis refers to those cases where the qualitative phenotype associated with one null mutation is masked completely by the presence of a second mutation at a distinct locus. **B**. Generalized epistasis is used for quantitative phenotypes and measures the systematic deviation in the phenotype of a double mutant relative to a statistical null model. Unlike Batesonian epistasis, generalized epistasis cannot be used to infer genetic pathways, since the choice of null model is arbitrary. The effects associated with allele x are labelled β_x, and the generalized epistasis is given the symbol Δ.

As a result of its definition, the magnitude of generalized epistasis is completely dependent on the null model selected by the researcher. Unlike physical models that can be derived from first principles, statistical models of genetic interactions are heuristic models that may or may not represent the molecular interactions underlying the system accurately. In this sense, second order 'interaction' terms are *ad hoc* corrections, technically useful for machine-learning, but not instructive in terms of understanding the genetic mechanisms at play. The conceptual proof for this is simple: Imagine two different statistical models that describe how two genes interact along a phenotype. Both models perform equally well. One of the models has a statistically significant interaction (generalized epistasis) term whereas the other does not. It is not possible to select one model over the other based on statistical properties. In fact, based on model simplicity, we may even prefer the model with fewer parameters, which could rule out the model that includes an epistasis term.

Like classical epistasis, generalized epistasis has become a useful concept in many areas of biology. Unlike classical epistasis, generalized epistasis measurements have not been restricted to those generated by null alleles; instead, generalized epistasis, particularly in human genetics, is measured between any two molecular variants at different loci measured under a specific null model. As a result of the subtle differences between classical and generalized epistasis, there has been considerable concern about the apparent disagreement between these two concepts (Phillips, 2008; Cordell, 2002; Lehner, 2011). In this review, we will show how generalized and classical epistasis can be successfully unified. Moreover, this unification has important ramifications for our ability to detect genetic interactions between two mutants using in genome-wide studies.

Motivation: A brief introduction to RNA-sequencing

RNA-sequencing (Mortazavi et al., 2008) is a powerful method that can measure all the gene expression levels in an organism simultaneously. These measurements can be made in bulk, from homogenized tissues or even from whole-organisms. Recent technological breakthroughs have made measuring expression levels from single whole organisms (Serra et al., 2018; Chan, Rando, and Conine, 2018; Lott et al., 2011) or even single cells possible (Tang et al., 2009). As a result of its technical advantages, RNA-seq has largely replaced microarrays as the method of choice to monitor gene expression.

Since the advent of genome-wide measurement methods, the idea of a cell- or organismal-state, defined by its gene expression levels, has drawn significant attention. Such states make sense in light of gene regulatory network theory, which posits that the expression of many genes is coordinated by regulatory factors that, when expressed, drive development forward (Britten and Davidson, 1969). A common experimental design used to identify the genes that are controlled by a specific regulatory module is to measure a baseline (typically wild type) sample and a contrast sample where the regulatory module has been perturbed (often through mutation). These experimental designs identify differentially expressed genes between the wild type and the mutant samples. These batteries can then be analyzed through ontological enrichment analyses that attempt to integrate information from all the enriched transcripts and identify the biological processes or signaling pathways contained within this list (see for example Mi et al. (2009) and Angeles-Albores, N. Lee, et al. (2016)). In spite of the enormous amount of quantitative information that RNA-seq can provide about the genes that respond to a downstream perturbation, these single factor experimental designs are generally used to select a small number of novel downstream genes that can be studied to extend a pathway of interest. The problem of how to analyze the rich datasets generated by RNA-seq has proved difficult, and

no one answer will be suitable for all problems. Analyses of these datasets rely on a combination of biological intuition, enrichment analyses or comparisons to other existing datasets.

If we are willing to sacrifice the requirement for interpretability, these datasets are still useful. Their practicality derives significantly from enormous advances in library preparation methods (Picelli et al., 2014) and improved quantification algorithms (Patro, Mount, and Kingsford, 2014; Patro, Duggal, et al., 2016; Bray et al., 2016) that have made RNA-seq an eminently replicable protocol that is fast to execute. As a result, transcriptomes can readily be used to compare the extent to which two perturbations are similar through clustering methods. Thus, transcriptomes could be thought of as extremely long barcodes that are associated with specific, potentially hidden, variables. If two barcodes are similar, then it is plausible to hypothesize that the perturbations applied to generate each barcode were also similar, even though we may not understand what these barcodes mean or how they were generated. However, it is not sufficient to develop algorithms that show two perturbations are similar on average. To use transcriptomes for genetic analysis, we need methods that quantitatively reveal what aspects of two transcriptomes are similar, by how much and that allow us to understand why they are similar.

Genetic interactions detection through sequencing

A brief overview of the problem

Expression profiles are vectors where each entry corresponds to the expression level of a single transcript. Conceptually, each entry could be treated as an independent continuous phenotypes. Since continuous phenotypes can be used to detect statistical epistasis, we could fit a statistical model to explain the expression level of this transcript in each genotype measured (wild type, single and double mutants). This statistical model will fit two parameters, β_a and β_b, that explain the *individual* effects

of each null mutation, and a third parameter, Δ, quantifies the extent to which these individual effects do not add when both null mutations are present at once (see Fig. 11). Each parameter is associated with a *p*-value. These models are generated for every measured transcript. The generated *p*-values should then be adjusted for multiple comparisons (these adjusted values are referred to as *q*-values), and parameters with *q*-values below a pre-specified threshold (often 0.1) are considered statistically different from zero.

As a result, each transcript is associated with six values: the three model parameters and their corresponding *q* values. At this point, the complexity of the problem is obvious: Even if each parameter only acquires one of three values (-, 0, +), there are 27 possible parameter combinations (epistatic classes). Of these 27 classes, only four classes could give rise to the classical epistasis regime (genetic suppression), so only those 4 classes can give rise to a genetic diagram. One approach to visualize and attempt to understand this complex space of epistatic combinations is to use heatmaps to look for patterns and guide interpretation. This approach was used in single-celled organisms (Capaldi et al., 2008; Van Driessche et al., 2005; Sameith et al., 2015; Van De Peppel et al., 2005), and more recently has been used to perform high-throughput analyses of genetic interactions in mammalian cells (Dixit et al., 2016). Regression models with interactions have also been successfully implemented using whole-organism transcriptomic measurements (Angeles-Albores, Leighton, et al., 2017).

The large number of parameter combinations is not the only (or major) drawback to fitting models with interactions for every transcript. Another challenge is the significant false positive and false negative rates for RNA-seq. RNA-seq studies often accept an estimated false discovery rate of 10%, and, although false negative rates are unknown, estimates are as high as 90% for mammalian cells (Pimentel et al., 2017). These rates seriously impair attempts to classify transcripts into any one of the

27 possible classes. If parameters are controlled at a rate of 10%, then the probability that at least one of the parameters in a dense class (classes where all 3 parameters are + or -, not 0) has been falsely accepted is almost 27%. Thus, almost one in three of the transcripts categorized into one of the 8 possible dense epistatic classes (+++, - - -, ++-, etc. . .) is misclassified and instead belongs to one of the twelve doublet epistatic classes (++0, 0- -, etc. . .). The situation becomes considerably worse once we consider false negative rates, which are generally unknown but estimates range up to 90% in mammalian systems (Pimentel et al., 2017). In general, false rates greatly exacerbate the difficulties associated with analyzing transcriptomic datasets. If all transcripts actually belonged to a single epistasis class to begin with, the addition of statistical noise will split this class into many more classes that mimicry complex interactions. The situation is further worsened by the fact that interaction parameters can often be harder to measure than first order parameters. Classifying transcripts into epistatic classes is a major obstacle for successful epistatic analyses, and so far there has been little to no work done to assess which classes are real and which are artifactual (some work has been done in the context of allelic series, see page 14). Equally concerning is the fact that none of these epistatic classes can be translated into genetic diagrams. These epistatic classes do not provide a biological mechanism (genetic, biochemical or cellular) between the genes under study (see Fig. 12).

Occam's razor, information pooling and constrained epistasis

To extract biological mechanisms from transcriptome data, we must apply simplifying constraints. If transcripts are to be classified into 27 possible epistatic classes, we must develop methods to assess which of these classes have sufficient statistical leverage to accept their existence (in other words, we need a statistical test that examines the null hypothesis that such a class could appear purely by chance). However,

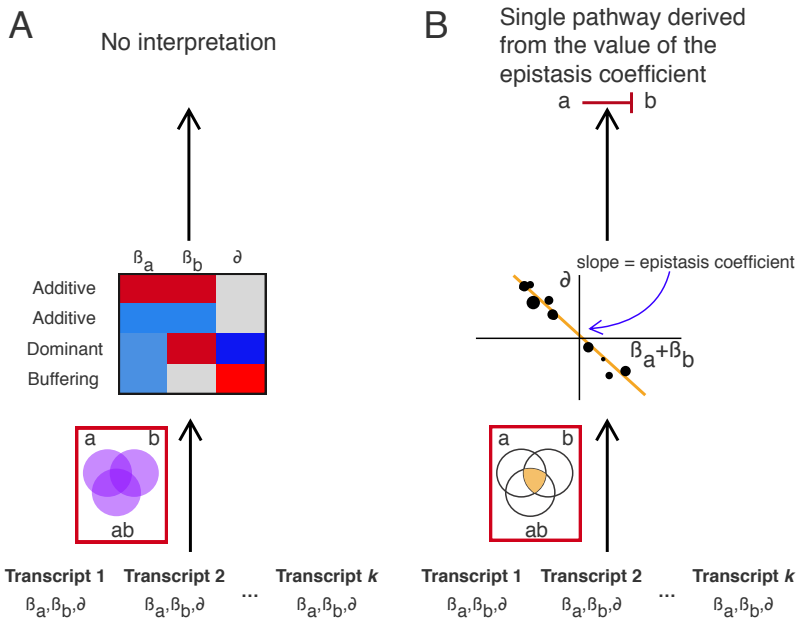


Figure 12 Analysis methodology to infer genetic interactions using transcriptome data. **A**. After fitting all transcripts to a general linear model to calculate the individual and the epistatic components of null mutations in two distinct genes, the resulting parameters can be clustered and visualized in a heatmap. Each observed cluster can be grouped into one of 27 epistatic classes. All clusters are considered biologically relevant regardless of the number of transcripts they contain. A simple conclusion cannot be reached from these heatmaps. This approach was used in Dixit et al. (2016) and Adamson et al. (2016) **B**. Starting from the same statistical model, only transcripts that have all parameters different from zero are considered informative. These transcripts are plotted on a scatterplot, where the x-axis reflects the expected value of the double mutant under an additive or log-additive hypothesis, and the systematic deviation from additivity (generalized epistasis) is plotted on the y-axis. The resulting points form a ray on the plot. The slope of this ray is an aggregate statistic that can be interpreted in terms of a genetic pathway if the two genes exhibit Batesonian epistasis. This approach was used in Angeles-Albores, Leighton, et al. (2017) and Angeles-Albores, Puckett Robinson, et al. (2018)

even if such a test were developed, we still require a method that allows us to summarize the information in these modules, and which lets us build a genetic pathway if the data suggests a pathway exists. A natural way to do this may be to use the natural structure of biological networks to pool the information from all transcripts, and test the interaction of this *structure* between the two mutants, instead of testing the individual transcripts. Information sharing is a powerful concept that allows us to incorporate more data points into a calculation, thus increasing our statistical power for any single test, but it requires the data to be drawn from a structure that permits sharing.

One such information sharing approach was implemented in Angeles-Albores, Puckett Robinson, et al. (2018) and Angeles-Albores, Leighton, et al. (2017). Briefly, these studies obtained whole-organism bulk RNA-seq transcriptome profiles for single and double perturbations and identified differentially expressed transcripts in each condition relative to the wild-type. Next, transcripts that were differentially expressed in all non-control conditions were aggregated and analyzed jointly for systematic deviations from a linear pathway. This systematic deviation was quantified in a single coefficient, called the transcriptome-wide epistasis coefficient. This coefficient can be interpreted in terms of simple genetic pathways because it can be used to test whether the perturbations result in a phenotypic transformation diagnostic of Batesonian epistasis. In this sense, the transcriptome-wide epistasis coefficient represents a unification of generalized epistasis and classical epistasis. This approach is powerful because it avoids multiple hypothesis testing (a single interaction coefficient is tested), and it doesn't rely on any one transcript to draw conclusions. A significant advantage of this method is that these studies were able to test and verify that the generalized epistasis measurements they made were equivalent to Batesonian epistasis (in other words, the double mutant had the same perturbations as one of the single mutants), culminating in a formal genetic

pathway. Both studies assumed that the genetic interaction between two genes is unimodal, in other words, these two genes do not interact along multiple pathways with different strengths and valences. This last assumption may not always hold. This strongly simplifying assumption contrasts with the previously referenced work that assumes unbounded complexity for all genetic interactions. Neither is correct, though it is our opinion that biological interactions tend to be much simpler than is often assumed in genomic studies.

Beyond genetic interactions: Dominance studies to map gene functions

Although transcriptomes have been used as phenotypes for analysis of genetic interactions for many years, their uses need not be restricted for theoretic analysis. In population genetics, transcriptomes have been used as phenotypes with which to identify expression quantitative trait loci in a number of organisms (Brem et al., 2002; DeCook et al., 2006; Kirst et al., 2004; Schadt et al., 2003). Transcriptomes can also be used to compare the genetical properties of different alleles of a single gene (Angeles-Albores and Sternberg, 2018).

Allelic series require considerably more analysis than tests for genetic interactions. To infer functional units from the activity of multiple allelic variants, the phenotypes associated with each variant must be carefully enumerated. Alleles must be ordered according to the phenotypic severity they cause when animals are homozygotes for each variant, with a separate hierarchy drawn for each phenotype. Alleles must also be ordered according to their dominance hierarchy over other alleles along each phenotype by measured the phenotypes of *trans*-heterozygotes. Particular care must be taken to ensure that the phenotypes of the *trans*-heterozygotes are not the result of maternal effects by testing progeny generated from a second, reciprocal, cross. The overall results are examined and the most parsimonious explanation is accepted to draw functional units and establish their sequence requirements. The the number

and resolution of the functional units that can be defined depends on the density of the allelic series that is tested. For a more thorough introduction to dominance and its role in allelic series, see Yook (2005).

As a result of the rigor required to analyze them, allelic series provides an excellent testing ground in which to explore the potential, but also the shortcomings, of transcriptomes as molecular phenotypes. To be successful, the analysis of even the smallest allelic series must order the tested variants. Angeles-Albores and Sternberg (2018) reported the first allelic series, to our knowledge, to be analyzed using expression profiles in any organism. In this analysis, the transcriptomic analogue of distinct phenotypes, phenotypic classes consisting of groups of differentially expressed genes, were identified by labelling each gene with the genotypes where it was differentially expressed. Subsequently, the expression level of these genes in *trans*-heterozygotes was approximated by a linear combination of the expression levels in each homozygote, with the weighting coefficients constrained to add to unity. The weighting coefficients, bounded in this manner, reflect the dominance of one allele over the other. These transcriptome-wide dominance coefficients are analogous to the transcriptome-wide epistasis aggregate statistics derived in previous studies (Angeles-Albores, Puckett Robinson, et al., 2018). The intersections from the Venn diagram (see Fig. 13) are understood to occur as a result of the activity of one or more functional units which may or may not have dosage-saturated activity (this is inferred from the dominance behavior of the given intersection).

This study highlighted the importance of recognizing and characterizing the statistical artifacts that can occur in genomic datasets (see Fig. 14). The analyzed dataset had sufficiently large false positive and false negative rates to generate artificial phenotypic classes that nevertheless could be identified and removed from the analysis. Unlike epistatic classes, for which we do not have a sense of what classes can most easily arise as a result of statistical artifacts, all the phenotypic classes

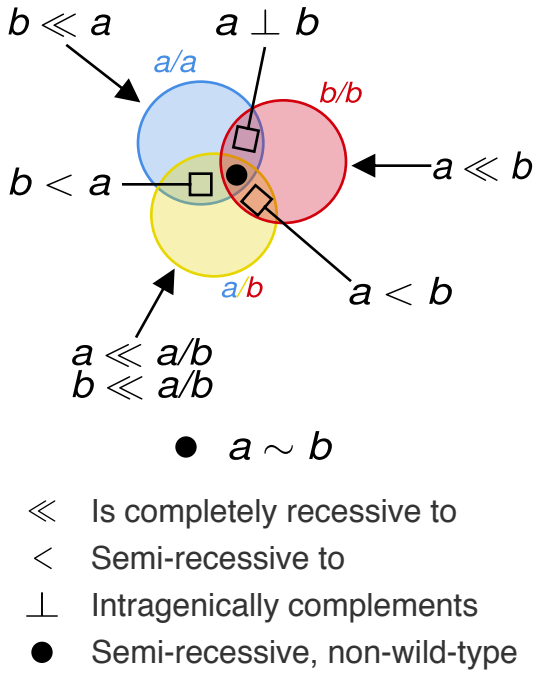


Figure 13 Genes that are differentially expressed in genotypes containing mutant (a, b) alleles relative to a wild type homozygote can be categorized into phenotypic classes. Each phenotypic class can in turn be associated with a dominance behavior. The Venn diagram represents differentially expressed transcripts in each genotype relative to the wild-type control. Each of the possible 7 intersections is labelled with its dominance interpretation if the intersection is real. In this context, semi-recessiveness means that one allele is partially or completely dominant to the other along a continuous spectrum between 0 and 1. The dominance sign between an allele and the heterozygote genotype indicates heterosis or overdominance.

arising from allelic series analyses can be readily interpreted in terms of inter-allelic complementation, a phenomenon that is extremely well characterized in genetics. Allelic series provide an excellent testing ground in which to explore algorithms to partition transcriptomes into gene batteries that have sufficient statistical support, since it is possible to have an intuition for artifactual classes.

Open problems and opportunities

RNA-sequencing is becoming increasingly easier and cheaper. RNA-seq offers a powerful, unbiased approach to genetics that can be multiplexed in many systems relatively easily. We expect that genetics using expression profiles will be an

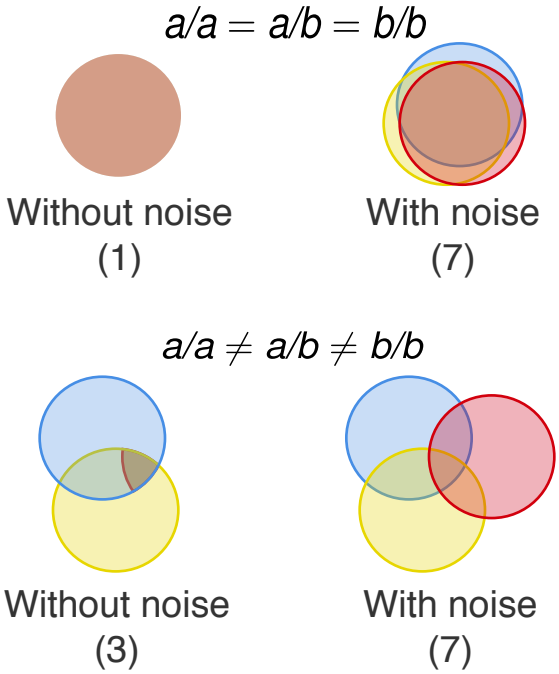


Figure 14 RNA-seq artifacts can greatly exaggerate apparent biological complexity. We considered the case where we have two phenotypically identical alleles that can be used to generate the genotypes *a/a*, *b/b* and *a/b*. In the absence of artifacts, the set of differentially expressed transcripts relative to a wild-type control should be the same amongst all three genotypes. However, if measurement error occurs, then instead of observing a single Venn intersection, we will observe seven intersections. If these intersections are not identified as false, we would wrongly conclude that allele *a* and *b* are not phenotypically equivalent, incurring in an error rate of 600%. Even in the case where the three genotypes are not equivalent, statistical noise will tend to significantly increase the apparent biological complexity present in the system (from 3 to 7 in this example). In general, statistical artifacts are so common in genomic assays that they will tend to generate all the possible intersections in a comparison. This highlights the need to apply simplifying constraints on transcriptome data before interpreting the results.

excellent first-pass assay because of the speed and sheer amount of information associated with the generation of expression profiles. These properties make RNA-seq particularly advantageous for groups that are studying relatively unknown genes or genes with subtle phenotypes. RNA-seq may also be a powerful method to complement genetics in emerging model organisms where conventional genetics may be laborious and where researchers may wish to minimize the number of experiments performed while maximizing the amount they can learn.

A major challenge moving forward will be mixed epistasis analyses with allelic series. Such mixed analyses try to identify the sequence requirements of one gene to participate in an epistatic interaction, and to test whether the observed epistatic interaction between two genes reflects a single biochemical function or the joint activity of distinct molecular properties. For example, in *C. elegans* the inhibition of *hif-1* by *egl-9* is mediated partially by the hydroxylation of HIF-1 by EGL-9, and partially through a hydroxylation-independent mechanism that is not well understood (Shao, Zhang, and Powell-Coffman, 2009). The high false positive and false negative rates inherent to RNA-seq means that all interactions amongst all genes will appear to be the compounded result of many independent activities. The solution to this problem will require methods that can incorporate information not just between single and double mutants, or homozygotes and heterozygotes, but amongst epistatic modules and dominance modules while searching for the most parsimonious structure that can explain all the expression profiles.

A second challenge will be the association of gene batteries with other observable phenotypes to develop signatures that allow us to read and interpret a transcriptome in terms of biological covariates. In other words, we would like signatures that allowed us to infer what the organism was doing when the RNA was extracted, what pathways had been disrupted or activated, what cellular or morphological phenotypes it exhibited. Such signatures could be derived by allowing organisms

to undergo a specific life history, then extracting the transcriptome and associating the differentially expressed genes in response to this life history relative to a control history to derive a signature. Alternatively, single-cell or single-organism methods may be able to track organisms, recording their behavior, before extracting their RNA (Lane et al., 2017). These signatures, although useful, should not be treated as causal, because the derivation of these signatures is through correlation. Deriving causal signatures would be very interesting and potentially useful as well, since this would make the discovery and association of novel pathways considerably easier. A significant weakness of expression signatures is that they only make sense relative to a baseline control, and therefore signatures can only be associated with events that have a sufficient dynamic range relative to the baseline. Another problem with signatures is the arbitrary definition, since they will inevitably be defined according to a q-value cut-off. It seems reasonable to postulate that eventually we must abandon the concept of differential expression: It is too brittle, too relativistic and prevents us from thinking about the transcriptome as a complete object.

Without transcriptional signatures of some sort, understanding modules will be all but impossible. Even with signatures, modules will be explained only phenomenologically: We know this signature is correlated to this phenotype, therefore this module is correlated to the same phenotype. With time, we may be able to understand mechanistically why specific phenotypes are correlated with the expression of specific genes. For the moment, such understanding seems far from our reach.

In the end, the major challenge for transcriptome genetics is likely to be our own creativity. New phenotypes always have their difficulties and drawbacks, and expression profiles are no exception. Expression profiles will not, on their own, reconstruct every network or solve all of biology. However, expression profiles are an object of a new kind, with behaviors that we do not fully understand hiding novel biological phenomena. It has become evident that genetics is applicable at an enormous

range of phenotypes, from population phenotypes to organismal to macroscopic and mesoscopic phenotypes. Transcriptomes represent a new phenotype at the microscopic and genomic level. Perhaps surprisingly, these microscopic phenotypes, in spite of all their complexity, seem to obey the genetic properties that bind all other phenotypes. The challenge, then, is how to use transcriptomes to discover biological principles that help us understand how the hierarchy of cells, organs, organisms and populations emerges from the collective actions of a string of atoms.

References

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

Angeles-Albores, David, Daniel H. W. Leighton, et al. (2017). "The *Caenorhabditis elegans* Female-Like State: Decoupling the Transcriptomic Effects of Aging and Sperm Status". In: *G3: Genes, Genomes, Genetics* 7.9.

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

Angeles-Albores, David, Carmie Puckett Robinson, et al. (2018). "Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements". In: *Proceedings of the National Academy of Sciences*, p. 201712387. ISSN: 0027-8424. DOI: 10.1073/pnas.1712387115.

Angeles-Albores, David and Paul W Sternberg (2018). "Using Transcriptomes as Mutant Phenotypes Reveals Functional Regions of a Mediator Subunit in *Caenorhabditis elegans*." In: *Genetics*, genetics.301133.2018. ISSN: 1943-2631. DOI: 10.1534/genetics.118.301133.

Avery, Leon and Steven Wasserman (1992). *Ordering gene function: the interpretation of epistasis in regulatory hierarchies*. DOI: 10.1016/0168-9525(92)90263-4. arXiv: NIHMS150003.

Bateson, William and Gregor Mendel (2009). *Mendel's principles of heredity: A defence, with a translation of mendel's original papers on hybridisation*, pp. 1–212. ISBN: 9780511694462. DOI: 10.1017/CBO9780511694462.

Bray, Nicolas L et al. (2016). "Near-optimal probabilistic RNA-seq quantification." In: *Nature biotechnology* 34.5, pp. 525–7. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. arXiv: 1505.02710.

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

Britten, Roy J. and Eric H. Davidson (1969). "Gene regulation for higher cells: A theory". In: *Science* 165.3891, pp. 349–357. ISSN: 00368075. DOI: 10.1126/science.165.3891.349.

Capaldi, Andrew P et al. (Nov. 2008). "Structure and function of a transcriptional network activated by the MAPK Hog1". In: *Nature Genetics* 40.11, pp. 1300–1306. ISSN: 1061-4036. DOI: 10.1038/ng.235.

Chan, Io Long, Oliver J Rando, and Colin C Conine (2018). "Effects of Larval Density on Gene Regulation in *Caenorhabditis elegans* During Routine L1 Synchronization". In: *G3: Genes|Genomes|Genetics* 8.5, 1787 LP –1793. ISSN: 2160-1836. DOI: 10.1534/g3.118.200056.

Cordell, Heather J (2002). "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". In: *Human Molecular Genetics* 11.20, pp. 2463–2468. DOI: 10.1093/hmg/11.20.2463.

Costanzo, Michael et al. (2016). "A global genetic interaction network maps a wiring diagram of cellular function". In: *Science* 353.6306. ISSN: 10959203. DOI: 10.1126/science.aaf1420. arXiv: arXiv:1011.1669v3.

DeCook, Rhonda et al. (2006). "Genetic regulation of gene expression during shoot development in Arabidopsis". In: *Genetics* 172.2, pp. 1155–1164. ISSN: 00166731. DOI: 10.1534/genetics.105.042275.

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

Fisher, R. A. (1919). "XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance". In: *Transactions of the Royal Society of Edinburgh* 52.2, pp. 399–433. ISSN: 00804568. DOI: 10.1017/S0080456800012163.

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

Kirst, Matias et al. (2004). "Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus." In: *Plant physiology* 135.4, pp. 2368–78. ISSN: 0032-0889. DOI: 10.1104/pp.103.037960.

Lane, Keara et al. (Apr. 2017). "Measuring Signaling and RNA-Seq in the Same Cell Links Gene Expression to Dynamic Patterns of NF-κB Activation". In: *Cell Systems* 4.4, 458–469.e5. ISSN: 24054712. DOI: 10.1016/j.cels.2017.03.010.

Lehner, Ben (Aug. 2011). "Molecular mechanisms of epistasis within and between genes". In: *Trends in Genetics* 27.8, pp. 323–331. ISSN: 01689525. DOI: 10.1016/j.tig.2011.05.007.

Lott, Susan E. et al. (2011). "Noncanonical compensation of zygotic X transcription in early *Drosophila melanogaster* development revealed through single-embryo RNA-Seq". In: *PLoS Biology* 9.2. ISSN: 15449173. DOI: 10.1371/journal.pbio.1000590.

Mi, Huaiyu et al. (2009). "PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium". In: *Nucleic Acids Research* 38.SUPPL.1. ISSN: 03051048. DOI: 10.1093/nar/gkp1019.

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226. arXiv: 1111.6189v1.

Patro, Rob, Geet Duggal, et al. (2016). "Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference". In: *bioRxiv*, p. 021592. DOI: 10.1101/021592. arXiv: 1505.02710.

Patro, Rob, Stephen M. Mount, and Carl Kingsford (2014). "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms". In: *Nature biotechnology* 32.5, pp. 462–464. ISSN: 1546-1696. DOI: 10.1038/nbt.2862. arXiv: 1308.3700.

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

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

Pimentel, Harold et al. (2017). "Differential analysis of RNA-seq incorporating quantification uncertainty". In: *Nature Methods* 14.7, pp. 687–690. ISSN: 15487105. DOI: 10.1038/nmeth.4324.

Sameith, Katrin et al. (2015). "A high-resolution gene expression atlas of epistasis between gene-specific transcription factors exposes potential mechanisms for genetic interactions". In: *BMC Biology* 13.1. ISSN: 17417007. DOI: 10.1186/s12915-015-0222-5.

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

Serra, Lorrayne et al. (2018). "Adapting the Smart-seq2 Protocol for Robust Single Worm RNA-seq". In: *BIO-PROTOCOL* 8.4. ISSN: 2331-8325. DOI: 10.21769/BioProtoc.2729.

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

Tang, Fuchou et al. (May 2009). "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature Methods* 6.5, pp. 377–382. ISSN: 1548-7091. DOI: 10.1038/nmeth.1315.

Van De Peppel, Jeroen et al. (2005). "Mediator expression profiling epistasis reveals a signal transduction pathway with antagonistic submodules and highly specific downstream targets". In: *Molecular Cell* 19.4, pp. 511–522. ISSN: 10972765. DOI: 10.1016/j.molcel.2005.06.033.

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

Yook, Karen (2005). "Complementation". In: *WormBook*. ISSN: 15518507. DOI: 10.1895/wormbook.1.24.1.

Chapter 2

A STATISTICAL MECHANICAL THEORY OF GENETICS USING GENE EXPRESSION PHENOTYPES

Abstract

Genetics is a powerful method that can be used to probe the molecular function of individual genes and to reconstruct genetic interaction networks. A cornerstone of genetics is the concept of epistasis: The ability of the phenotype associated with a null allele of one gene to block the phenotype of a null allele associated with a second gene. Epistasis is a widespread phenomenon in biology, and it has been observed using phenotypes that span many orders of magnitude in length-scales. In spite of its importance, a theoretical derivation for why epistasis occurs at so many scales is lacking. Here, we use statistical mechanics to derive epistasis from first principles.

Introduction

Imagine, if you will, a partition function of unbounded complexity. Imagine that there are thousands of potential variables that could, but do not have to, participate in the system. Further imagine that this system is not immediately experimentally tractable: The number of particles in the system cannot be easily controlled, the various energies, enthalpies and entropies of the system cannot be measured, and an analytical function cannot be satisfactorily written from theoretical principles. Such a problem might at first sight appear to be intractable through the methods of statistical mechanics, or may require highly complex numerical methods to estimate the properties of the partition function.

This thermodynamic regime is the regime occupied by many biological systems.

Bacteria, archaea, fungi or animals consist of a large and unknown number of particles (proteins) encoded within genes in a genome. Any one characteristic (phenotype) in an organism is the result of a large and unknown number of particles interacting in a large and unknown number of configurations.

For systems where little or nothing is known about the factors controlling a phenotype, biology relies on a powerful empirical method: genetics. Briefly, in classical genetics, mutants of random genes are generated until a gene is found that, when removed, causes a mutant phenotype. Mutants that exhibit the desired mutant phenotypes can be combined to generate double mutants and the phenotype of the resulting double mutant is inspected. If the two genes under investigation exhibit a phenomenon known as classical (or Batesonian) epistasis, then these two genes are said to have a genetic interaction, and they can be ordered into a genetic pathway where one gene activates or inhibits the second. Classical epistasis is not equivalent to the concept of epistasis used in population genetics or human genetics (Cordell, 2002). In these fields, epistasis is represented by second order interaction terms between arbitrary mutations in a linear or log-linear statistical model. There has been significant work on the effect these second order terms have on these statistical models (Crow, 2010; Mackay, 2014). However, these statistical models are not grounded in a principled theory of genetic interaction and the presence or absence of epistasis is dependent on the choice of statistical model (Cordell, 2002). In this text, we restrict our studies to classical epistasis.

Genetics has been of major importance for finding the genes that control phenotypes of interest and for ordering them into networks that are amenable for biochemical or biophysical characterization. The choice of phenotype is important, and the introduction of new phenotypes has led to significant breakthroughs. The phenotypes used for genetic analysis include animal morphology (Sulston and Brenner, 1974) and development (Jürgens et al., 1984), behavior (Benzer, 1967), cellular differen-

tiation, metabolism (Beadle and Tatum, 1941), and most recently gene expression levels (Angeles-Albores et al., 2018; Hughes et al., 2000; Capaldi et al., 2008). The ability of genetics to establish interactions between particles using phenotypes that vary by 6–8 orders of magnitude of length scales has been extremely useful and is deeply intriguing from a theoretical perspective. In spite of its strong logical foundation and enormous empirical evidence of its usefulness, there is no general theoretical description for why genetics is so effective.

Statistical Mechanics of Genetic Interactions

We will derive epistasis in gene expression phenotypes using a toy model for gene expression (Garcia et al., 2007; Bintu et al., 2005). This model can be derived from the assumption that the level at which a gene is expressed is directly proportional to the probability that RNA Polymerase II is bound at that gene's promoter. This probability depends upon the RNA Polymerase levels, ρ, and on other factors, $\{A, B, \ldots\}$ that can bind the promoter and RNA polymerase:

$$p_{bound}(A, B, \ldots) = \frac{1}{1 + \frac{1}{\rho F_{reg}(A,B,\ldots)}}. \tag{2.1}$$

$p_{bound}(A, B, \ldots)$ is the probability that RNA Polymerase is bound at the locus of interest. F_{reg} is a rational function,

$$F_{reg}(A, B) = P(A, B)/Q(A, B).$$

It represents the effective number of RNA polymerases at the promoter of interest and in general cannot be analytically determined for all but the simplest systems. To ensure that Eq. 2.1 is a probability, the range of F_{reg} is restricted to positive real numbers. This factor is used to model a variety of transcription factors, such as activators, or inhibitors, and the physical interactions between them. A and B

represent the activities of the gene products of genes a, b. The variables X are related to the physical number of proteins of X, X_{protein} through the equation:

$$X = \frac{X_{\text{protein}} e^{-\varepsilon_{xd}}}{N_{NS}}.$$

Here, ε_{xd} refers to the energy of binding of protein X to DNA at the promoter (in units of $k_B T$), and N_{NS} refers to the total number of non-specific sites on the genome. A major assumption in deriving this equation is that all proteins are bound to either the promoter or alternatively to non-specific sites in the genome. A further assumption is that $N_{NS} \gg X$.

We are interested in what occurs when either variable A, B or both A and B are set to zero instead of the levels found in a non-mutant, or wild-type, reference organism (X_{wt}). Specifically, we will explore the constraints on the functional form between A and B such that the distribution of gene expression levels in a mutant lacking protein A is completely independent of the levels of protein B, even though in the general case of all non-zero levels of protein A, the probability of RNAP binding is conditional on the levels of both A and B.

Experimentally, this is tested by generating single and double mutants and measuring the expression level of a reporter gene. We search for gene pairs, a and b, where the expression level of a reporter in a mutant lacking protein A is equal to the expression level of the reporter in a mutant lacking both proteins,

$$p_{bound}(A = 0, B = 0) = p_{bound}(A = 0, B_{wt}). \tag{2.2}$$

This condition is trivially satisfied if B does not play a role in controlling the gene expression level of the reporter, so we will only consider cases where perturbing the value of B away from B_{wt} changes the expression levels of the reporter gene.

The identity in Eq. 2.2 is called classical (or Batesonian) epistasis, and demonstrates that a null allele of one gene (a) can mask the phenotype associated with a null allele of a second gene (b). The gene that is masked is said to be *hypostatic*, while the masking gene is *epistatic*. When a pair of genes shows epistasis, there is a genetic interaction between gene a and gene b (note: the genetic interaction is not said to occur between the two proteins). Since p_{bound} depends on proteins A and B only through $F_{reg}(\cdot,\cdot)$, Eq. 2.2 can be re-written in terms of the regulatory function.

$$F_{reg}(A=0,B=0) = F_{reg}(A=0,B_{wt}).$$

We can approximate the right hand side of this equation as a Taylor function of B around 0, letting $F_{reg}(A=0,B=0)=\phi$:

$$\phi = \phi + \sum_i \frac{\partial^i}{\partial B^i} F(A=0,B=0)\frac{B_{wt}^i}{i!}.$$

This equation can be satisfied for arbitrary values of B if and only if:

$$\frac{\partial^i}{\partial B^i} F(A=0,B=0) = 0. \tag{2.3}$$

Thus, enforcing epistasis (Eq. 2.2) is equivalent to constraining all partial derivates with respect to the *hypostatic* variable, B, of F_{reg} at $(0,0)$ to sum to zero. Since F_{reg} is a rational function we can re-write Eq. 2.3 using the chain rule,

$$\sum_{i=1}^{\infty}\left[\frac{\partial^i F_{reg}}{\partial P^i}\frac{\partial^i P}{\partial B^i} + \frac{\partial^i F_{reg}}{\partial Q^i}\frac{\partial^i Q}{\partial B^i}\right]_{A=0,B=0} = 0,$$

which reveals that all partial derivatives of P and Q with respect to B must be zero at the point $(A=0,B=0)$.

We cast the polynomials P and Q into the general form

$$X = \sum_{j,k=0}^{\infty} \lambda_{jk}^{X} A^j B^k.$$

Using this general form, we will now find the constraints on λ_{jk}^{X} such that all the partial derivatives of this polynomial family vanish when $A = 0$. From inspection, all terms of order $j \geq 1$ will be zero when $A = 0$. Then, it follows that if $\lambda_{0,k}^{X} = 0, \forall k > 0$, all the partial derivates of P and Q with respect to B are 0 when $A = 0$. Thus, if two genes, a and b satisfy Eq. 2.2 when they are mutated, then P and Q can be written as:

$$X = \lambda_{00}^{X} + \sum_{j=1}^{\infty} \lambda_{j0}^{X} A^j + \sum_{j,k\geq 1}^{\infty} \lambda_{jk}^{X} A^j B^k. \tag{2.4}$$

Where X is either P or Q. We refer to each term in Eq. 2.4 as a *mesostate*. Briefly, a mesostate is a combination of microstates containing a defined set of species with an unknown stoichiometric distribution. In Eq. 2.4 there are three mesostates. If a mesostate is compatible with a set of epistatic relationships, then it must be non-empty in either P or Q.

Epistasis is transitive

Suppose there are three genes, a, b and c encoding proteins A, B and C respectively. Suppose further that epistasis analyses performed using a specific reporter as an expression phenotype show that a is epistatic over b and b is epistatic over c. Is a epistatic over c?

Once again, we let F_{reg} be a rational function of the polynomials P and Q. The general form of these polynomials is:

$$X = \sum_{j,k,l} \lambda_{jkl}^{X} A^j B^k C^l$$

Since a is epistatic over b, it follows that $\lambda^X_{0,k,l} = 0 \forall k > 0$. Since b is epistatic over c, it follows that $\lambda^X_{j,0,l} = 0 \forall l > 0$. Therefore, these polynomials can be written as

$$X = \lambda^X_{000} + \sum_{j=1}^{\infty} \lambda^X_{j00} A^j + \sum_{j,k\geq 1}^{\infty} \lambda^X_{jk0} A^j B^k + \sum_{j,k,l\geq 1}^{\infty} \lambda^X_{jkl} A^j B^k C^l. \quad (2.5)$$

From this functional form, it is clear that a is epistatic over c. Therefore, epistasis is transitive.

Epistasis is hierarchical

Suppose there are three genes, a, b and d encoding proteins A, B and D respectively. Suppose further that epistasis analyses performed using a specific reporter as an expression phenotype show that a is epistatic over b and d is epistatic over b. Must it be the case that either a is epistatic over d, or d is epistatic over a?

Once again, we let F_{reg} be a rational function of the polynomials P and Q. The general form of these polynomials is:

$$X = \sum_{j,k,l} \lambda^X_{jkl} A^j B^k D^l$$

Since a is epistatic over b, it follows that $\lambda^X_{0,k,l} = 0, \forall k > 0$. Since d is epistatic over b, it follows that $\lambda^X_{j,k,0} = 0, \forall l > 0$. Therefore, these polynomials can be written as

$$X = \lambda^X_{000} + \sum_{j=1}^{\infty} \lambda^X_{j00} A^j + \sum_{j,k\geq 1}^{\infty} \lambda^X_{jk0} A^j B^k + \sum_{l=1}^{\infty} \lambda^X_{00l} D^l + \sum_{k,l\geq 1}^{\infty} \lambda^X_{0kl} B^k D^l + \sum_{j,k,l\geq 1}^{\infty} \lambda^X_{jkl} A^j B^k D^l. \quad (2.6)$$

This functional form means that we cannot conclude anything about the epistatic relationship between a and d without generating a double mutant of a and d.

The results from the preceding two sections mean that epistasis is similar to an inequality statement. Therefore, we propose the notation:

$$a > b \quad (2.7)$$

to represent the genetic epistasis of gene a over b as defined by Eq. 2.2.

Epistasis enables qualitative functional inferences

Suppose that two genes, a and b, $a > b$. Further suppose that the phenotypes can be arranged in the following order:

$$p_{wt}(A_{wt}, B = 0) < p_{wt}(A = 0, B_{wt}) = p_{wt}(A_{wt}, B_{wt}).$$

This order can be immediately rephrased in terms of F_{reg},

$$F_{reg}(A_{wt}, B = 0) < F_{reg}(A = 0, B_{wt}) = F_{reg}(A_{wt}, B_{wt}).$$

Since F_{reg} is a function of polynomials P and Q, both of which have the functional form,

$$X = \lambda^X_{00} + \sum_{j=0}^{\infty} \lambda^X_{j0} A^j + \sum_{j,k\geq 1}^{\infty} \lambda^X_{jk} A^j B^k$$

Since we know that at least one term in each mesostate of P and Q must be non-zero, we conclude that the effective activity of A must be 0. With this information, the family of functions P and Q that will satisfy this hierarchy is:

$$Q(A_{wt}, B_{wt}) = \lambda^Q_{00} + \sum_j \lambda^Q_{j0} A^j + \sum_{j,k\geq 1} \lambda^Q_{jk} A^j_{wt} B^k_{wt} \tag{2.8}$$

and

$$P(A_{wt}, B_{wt}) = \lambda^P_{00} + \sum_{j,k\geq 1} \lambda^P_{jk} A^j_{wt} B^k_{wt}. \tag{2.9}$$

For these arguments to be true, it must also be the case that $B_{wt} \gg A_{wt}$ (equality is only achieved in the case when either B_{wt} becomes infinite or A_{wt} is zero). Genetically, gene b is a net inhibitor of gene a, and gene a is a net genetic inhibitor of our reporter phenotype.

We consider a different epistatic relationship between two different genes, c and d, such that $c > d$ and the phenotypes can be ordered:

$$F_{reg}(C_{wt}, D_{wt}) < F_{reg}(C_{wt}, D = 0) < F_{reg}(C = 0, D_{wt}).$$

We recall that:

$$F_{reg}(C = 0, D_{wt}) = \frac{\lambda^P_{00}}{\lambda^Q_{00}}$$

A suitable family of functions for this hierarchy is:

$$Q(C_{wt}, D_{wt}) = \lambda^Q_{00} + \sum_j \lambda^Q_{j0} C^j + \sum_{j,k\geq 1} \lambda^Q_{jk} C^j_{wt} D^k_{wt} \tag{2.10}$$

and

$$P(A_{wt}, B_{wt}) = \lambda_{00}^{P} + \sum_{j,k\geq 1} \lambda_{jk}^{P} C_{wt}^{j} D_{wt}^{k}, \tag{2.11}$$

subject to the constraint:

$$0 < \sum_{j,k\geq 1} (\lambda_{jk}^{Q} - \lambda_{jk}^{P}) C_{wt}^{j} D_{wt}^{k}. \tag{2.12}$$

This family of functions allows us to conclude that gene c is a net genetic inhibitor of our reporter phenotype, and d is a net promoter of the genetic activity of gene c.

Interpretation of epistasis for statistical mechanical systems

We have shown that classical epistasis is an identity between single and double mutants (Eq. 2.2) that is the result of nested polynomials. To understand epistasis, we introduce the concept of *mesostates*. We define mesostates as combinations of microstates involving a defined set of species with an unknown stoichiometric distribution. In Eq. 2.6, for example, is a sum of six mesostates. The first mesostate represents promoter leakiness; the second consists of all microstates that depend on the presence of protein A to form; the third consists of all microstates that depend on the presence of proteins A and B to form; the fourth consists of all microstates that depend on the presence of protein D to form; the fifth consists of all microstates that depend on the presence of protein B and D to form; and the sixth consists of all microstates that depend on the presence of proteins A, B and D to form. Epistasis on its own cannot tell us how many microstates correspond to a single mesostate, but it can tell us what mesostates are not accessible to the system, thus ruling out families of microstates.

Throughout this text, we have assumed that our imaginary proteins participated directly in binding to the promoter of interest. However, we can dispense with

this requirement, and instead imagine that these proteins function as switches that permit the existence of specific mesostates accessible to the promoter. This is the reason why we refer to epistatic interactions as genetic interactions: Epistasis does not provide any guarantee that the gene products ever interact physically, chemically or even that they coexist in the same space at the same time.

Here, we have shown that classical genetics, which has had a rich history over the past century and is a cornerstone of modern biology, is equivalent to a perturbative, parameter-free study of the partition function of a thermodynamic system. Though we have limited ourselves to applying this method to gene expression phenotypes, this approach is generalizable, and in fact, is not even limited to biological systems. We believe that the statistical mechanical basis for genetics explains its ability to explain phenotypes that span many orders of magnitude in time, space and molecular complexity.

References

Angeles-Albores, David et al. (2018). "Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements". In: *Proceedings of the National Academy of Sciences*, p. 201712387. ISSN: 0027-8424. DOI: 10.1073/pnas.1712387115.

Beadle, G. W. and E. L. Tatum (1941). "Genetic Control of Biochemical Reactions in Neurospora". In: *Proceedings of the National Academy of Sciences* 27.11, pp. 499–506. ISSN: 0027-8424. DOI: 10.1073/pnas.27.11.499.

Benzer, S. (1967). "BEHAVIORAL MUTANTS OF Drosophila ISOLATED BY COUNTERCURRENT DISTRIBUTION". In: *Proceedings of the National Academy of Sciences* 58.3, pp. 1112–1119. ISSN: 0027-8424. DOI: 10.1073/pnas.58.3.1112.

Bintu, Lacramioara et al. (Apr. 2005). "Transcriptional regulation by the numbers: Models". In: *Current Opinion in Genetics and Development* 15.2, pp. 116–124. ISSN: 0959437X. DOI: 10.1016/j.gde.2005.02.007. arXiv: 0412011 [q-bio].

Capaldi, Andrew P et al. (Nov. 2008). "Structure and function of a transcriptional network activated by the MAPK Hog1". In: *Nature Genetics* 40.11, pp. 1300–1306. ISSN: 1061-4036. DOI: 10.1038/ng.235.

Cordell, Heather J (2002). "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". In: *Human Molecular Genetics* 11.20, pp. 2463–2468. DOI: 10.1093/hmg/11.20.2463.

Crow, James F. (2010). "On epistasis: why it is unimportant in polygenic directional selection". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365.1544, pp. 1241–1244. ISSN: 0962-8436. DOI: 10.1098/rstb.2009.0275.

Garcia, Hernan G. et al. (2007). "A First Exposure to Statistical Mechanics for Life Scientists". In: p. 27. ISSN: 0036-8075. arXiv: 0708.1899.

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

Jürgens, G. et al. (1984). "Mutations affecting the pattern of the larval cuticle in Drosophila melanogaster - II. Zygotic loci on the third chromosome". In: *Wilhelm Roux's Archives of Developmental Biology* 193.5, pp. 283–295. ISSN: 03400794. DOI: 10.1007/BF00848157.

Mackay, Trudy F.C. (2014). *Epistasis and quantitative traits: Using model organisms to study gene-gene interactions*. DOI: 10.1038/nrg3627.

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

Chapter 3

RECONSTRUCTING A METAZOAN GENETIC PATHWAY WITH TRANSCRIPTOME-WIDE EPISTASIS MEASUREMENTS

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. Also, quantifying and interpreting epistatic effects using expression profiles remains a challenge. We developed a single coefficient to quantify transcriptome-wide epistasis that reflects the underlying interactions and which can be interpreted intuitively. To demonstrate our approach, we sequenced four single and two double mutants of *Caenorhabditis elegans*. From these mutants, we reconstructed the known hypoxia pathway. In addition, we uncovered a class of 56 genes with *hif-1*-dependent expression that have opposite changes in expression in mutants of two genes which cooperate to negatively regulate HIF-1 abundance; however, the double mutant of these genes exhibits suppression epistasis. This class violates 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. If the mutants of two distinct genes have a quantifiable phenotype, and the double mutant has a phenotype that is not the sum of the phenotypes of the single mutants, this non-additivity is referred to as generalized epistasis, and

indicates that these genes interact functionally. Such interactions can occur at the biochemical level between their products or as a consequence of their functions (L. S. Huang and Paul W Sternberg, 2006). Epistasis analysis remains a cornerstone of genetics today (Phillips, 2008).

Recently, biological studies have shifted in focus from studying single genes to studying all genes in parallel. In particular, RNA-seq (Mortazavi et al., 2008) enables biologists to identify genes that change expression in response to a perturbation. RNA-seq has been used to identify genetic modules involved in a variety of processes, such as in the *Caenorhabditis elegans* linker cell migration (Schwarz, Kato, and Paul W. Sternberg, 2012), planarian stem cell maintenance (Van Wolfswinkel, Wagner, and Reddien, 2014; Scimone et al., 2014). The role of transcriptional profiling has been restricted to target gene identification, and so far there are only a few examples where transcriptomes have been used to generate quantitative genetic models of any kind. In quantitative genetics, eQTL studies have established the power of transcriptomes for genetic mapping (Brem et al., 2002; Schadt et al., 2003; Li et al., 2006; King et al., 2014). Genetic pathway analysis via epistasis has been performed in *Saccharomyces cerevisiae* (Hughes et al., 2000; Capaldi et al., 2008) and in *Dictyostelium discoideum* (Van Driessche et al., 2005). Recently, Dixit *et al* described a protocol for epistasis analysis in dendritic and K562 cells using single-cell RNA-seq (Dixit et al., 2016). Epistasis analysis of single cells or single-celled organisms is popular because of the concern that whole-organism sequencing will mix information from multiple cell types, preventing the accurate reconstruction of genetic interactions. 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) (Angeles-Albores, Leighton, et al., 2017), 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 four well-characterized loss-of-function mutants in the *C. elegans* hypoxia pathway (Epstein et al., 2001; Shen, Shao, and Powell-Coffman, 2006; Shao, Zhang, and Powell-Coffman, 2009; H. Jiang, Guo, and Powell-Coffman, 2001).

Metazoans depend on the presence of oxygen in sufficient concentrations to support aerobic metabolism. Hypoxia inducible factors (HIFs) are an important group of oxygen-responsive genes that are highly conserved in metazoans (Loenarz et al., 2011). A common mechanism for hypoxia-response induction is heterodimerization between a HIFα and a HIFβ subunit; the heterodimer then initiates transcription of target genes (B. H. Jiang et al., 1996). The number and complexity of HIFs varies throughout metazoans. In the roundworm *C. elegans* there is a single HIFα gene, *hif-1* (H. Jiang, Guo, and Powell-Coffman, 2001), and a single HIFβ gene, *ahr-1* (Powell-Coffman, Bradfield, and Wood, 1998).

Levels of HIFα proteins are tightly regulated. Under conditions of normoxia, HIF-1α exists in the cytoplasm and partakes in a futile cycle of protein production and rapid degradation (L. E. Huang et al., 1996). In *C. elegans*, HIF-1α is hydroxylated by a proline hydroxylase (EGL-9) (Kaelin and Ratcliffe, 2008). HIF-1 hydroxylation increases its binding affinity to Von Hippel-Lindau tumor suppressor 1 (VHL-1), which in turn allows ubiquitination of HIF-1 leading to its 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 by the putative transmembrane O-acyltransferase RHY-1, possibly by post-translational modifications to CYSL-1 (Ma et al., 2012) (see Fig. 31).

Our reconstruction of the hypoxia pathway in *C. elegans* shows that whole-animal transcriptome profiles can be used as phenotypes for genetic analysis and that epis-

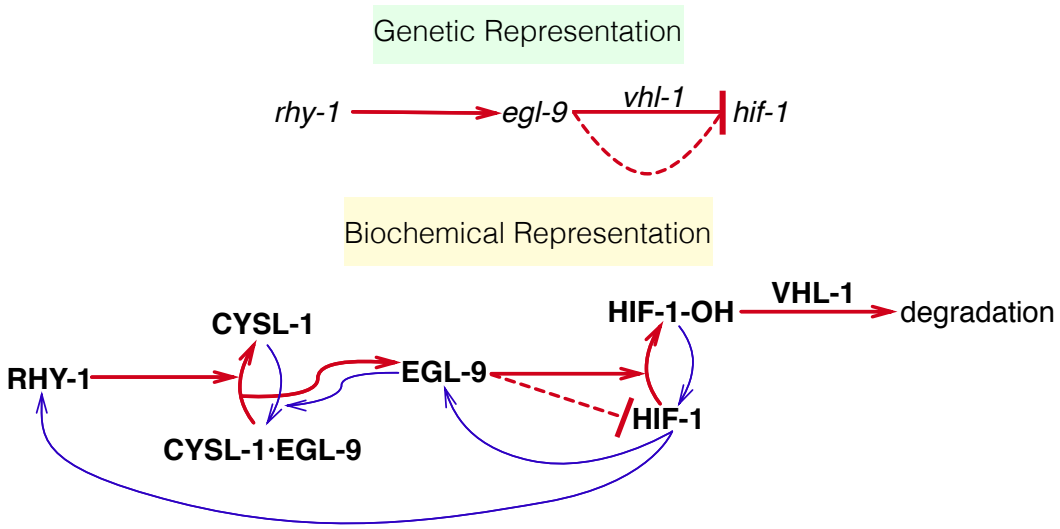


Figure 31 Genetic and biochemical representation of the hypoxia pathway in *C. elegans*. Red arrows are arrows that lead to inhibition of HIF-1, and blue arrows are arrows that increase HIF-1 activity or are the result of HIF-1 activity. EGL-9 is known to exert VHL-1-dependent and independent repression 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. RHY-1 inhibits CYSL-1, which in turn inhibits EGL-9, but this interaction was abbreviated in the genetic diagram for clarity.

tasis, a hallmark of genetic interaction observed in double mutants, holds at the molecular systems level. We demonstrate that transcriptomes can aid in ordering genes in a pathway using only single mutants. We were able to identify genes that appear to be downstream of *vhl-1*, but not downstream of *hif-1*. Using a single set of transcriptome-wide measurements, we observed most of the known transcriptional effects of *hif-1* as well as novel effects not described before in *C. elegans*. Taken together, this analysis demonstrates that whole-animal RNA-seq is a fast and powerful method for genetic analyses in an area where phenotypic measurements are now the rate-limiting step.

Results

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

We selected four null 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). Each genotype was sequenced in triplicate using mRNA extracted from 30 worms at a depth of 15 million reads per sample. Of these 15 million reads, 50% of the reads mapped to the *C. elegans* genome on average. All samples were analyzed under normoxic conditions. We measured differential expression of 19,676 isoforms across all replicates and genotypes (~70% of the protein coding isoforms in *C. elegans*; see Basic Statistics Notebook). We included in our analysis a *fog-2(q71)* mutant we have previously studied (Angeles-Albores, Leighton, et al., 2017), 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-value of a β coefficient (p-values adjusted for multiple testing) are less than 0.1. Transcripts that are differentially expressed between the 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 to the logarithm of the fold-change between the mutant and the wild type. Larger magnitudes of β correspond to larger perturbations (see Fig. 32). 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 by the number of differentially expressed genes (DEGs), not isoforms, they contain. For the case of *C. elegans*, this difference is negligible since the great majority of protein-coding genes have a single isoform. We have opted for this method of

referring to gene sets because it simplifies the language considerably. A complete version of the code used for this analysis with ample documentation, is available at https://wormlabcaltech.github.io/mprsq.

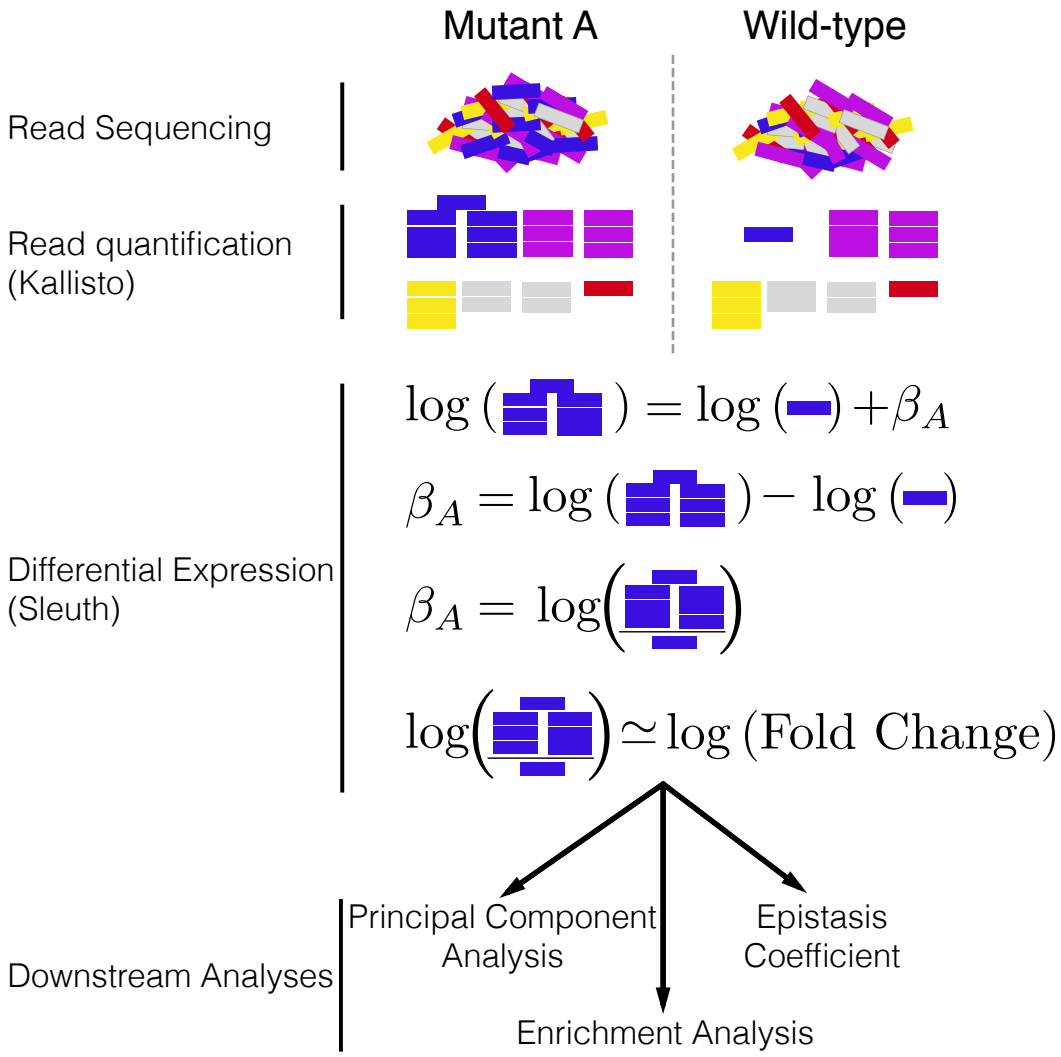


Figure 32 Analysis workflow. After sequencing, reads are quantified using Kallisto. Bars show estimated counts for each isoform. Differential expression is calculated using Sleuth, which outputs one β coefficient per isoform per genotype. β coefficients are analogous to the natural logarithm of the fold-change relative to a wild type control. Downstream analyses are performed with β coefficients that are statistically significantly different from 0. q-values less than 0.1 are considered statistically different from 0.

Transcriptome profiling of the hypoxia pathway revealed that this pathway controls thousands of genes in *C. elegans* (see Table 31, see SI File 1 for a complete list

of differentially expressed genes). The *egl-9(lf)* transcriptome showed differential expression of 2,549 genes. 3,005 genes were differentially expressed in *rhy-1(lf)* mutants. The *vhl-1(lf)* transcriptome showed considerably fewer DEGs (1,275), possibly because *vhl-1* is a weaker inhibitor of *hif-1* than *egl-9* (Shao, Zhang, and Powell-Coffman, 2009). The *egl-9(lf)*;*vhl-1(lf)* double mutant transcriptome showed 3,654 DEGs. The *hif-1(lf)* mutant 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). We do not think that this transcriptional response is the due to transiently induced hypoxia during harvesting. If the wild type strain had become hypoxic, then the *hif-1(lf)* genotype should show significantly lower levels of *nhr-57*, a marker that increases during hypoxia. We do not observe altered levels of *nhr-57* when comparing the wild type and *hif-1(lf)* mutant, nor between the wild type and *egl-9(lf) hif-1(lf)* double mutant. Finally, the *egl-9(lf)*, *vhl-1(lf)*, *rhy-1(lf)* and *egl-9(lf); vhl-1(lf)* mutants did show altered *nhr-57* transcript levels (see Quality Control Notebook, SI Figure 1). Of the differentially expressed genes in *hif-1(lf)* mutants, 161/1,075 were also differentially expressed in *egl-9(lf) hif-1(lf)* mutants, which suggests these transcripts are *hif-1*-dependent under normoxia. For the remaining genes, we cannot rule out cumulative effects from loss of *hif-1*, strain-specific eQTLs present in the strain background or that loss of *egl-9* suppresses the mutant phenotype. We designed our experiments to probe the constitutive hypoxia response, and not the effects of *hif-1* under normoxia, which we did not foresee. As a result, we have limited resolving power to explain the transcriptome of *hif-1(lf)* mutants.

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 31 Number of differentially expressed genes in each mutant strain with respect to the wild type (N2).

Principal Component Analysis visualizes epistatic relationships between genotypes

Principal component analysis (PCA) is used to identify relationships between high-dimensional data points (Yeung and Ruzzo, 2001). We used PCA examine whether each genotype clustered in a biologically relevant manner. PCA identifies the vector that explains most of the variation in the data; this is called the first principal component. PCA can identify the first *n* components that explain more than 95% of the data variance. Clustering in these *n* dimensions can indicate biological relationships, although interpreting principal components can be difficult. In our analysis, the first principal component discriminated mutants that have constitutive high levels of HIF-1 from mutants that have no HIF-1, whereas the second component was able to discriminate between mutants within the hypoxia pathway and outside the hypoxia pathway (see Fig. 33; *fog-2* is not reported to act in the hypoxia pathway and acts as a negative control; see Genetic Interactions Notebook).

Reconstruction of the hypoxia pathway from first genetic principles

To reconstruct a genetic pathway, we must assess whether two genes act on the same phenotype. If they do not act on the same phenotype (two mutations do not cause the same genes to become differentially expressed relative to wild type), these mutants

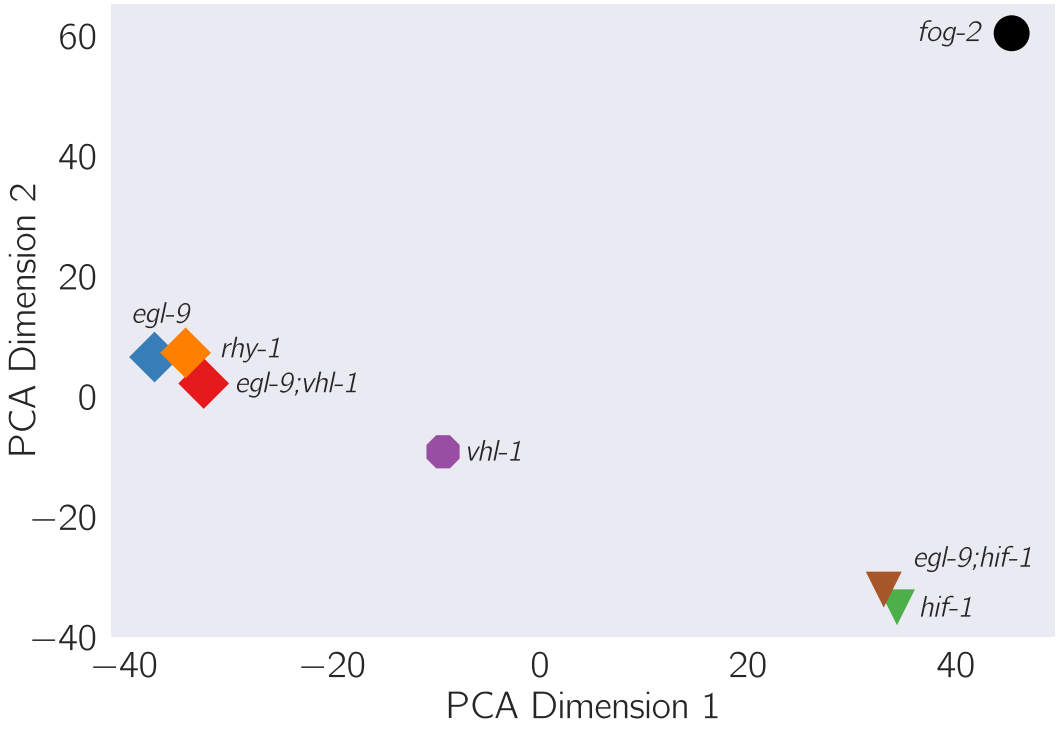


Figure 33 Principal component analysis of various *C. elegans* mutants. Genotypes that have an constitutive hypoxia response (i.e. *egl-9(lf)*) cluster far from genotypes that do not have a hypoxic response (i.e. *hif-1(lf)*) along the first principal component. The second principal component separates genotypes that do not participate hypoxic response pathway.

are independent. Otherwise, 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) between two mutants as the shared set of genes or isoforms whose expression in both mutants are different from wild-type, regardless of the direction of change.

Genes in the hypoxia mutant act on the same transcriptional phenotype

All the hypoxia mutants had a significant STP: the fraction of differentially expressed genes that was shared between mutants ranged from a minimum of 10% between *hif-1(lf)* and *egl-9(lf); vhl-1(lf)* to a maximum of 32% between *egl-9(lf)* and *egl-9(lf); vhl-1(lf)* (see SI Table 1). For comparison, we also analyzed a previously published *fog-2(lf)* transcriptome (Angeles-Albores, Leighton, et al., 2017). 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 had STPs (8.8%–14%).

Next, we analyzed pairwise correlations between all mutant pairs. We rank-transformed the β coefficients of each isoform between the STP of two mutants, and plotted the transcript ranks between genotypes (see Fig 34). Although *hif-1* is known to be genetically repressed by *egl-9*, *rhy-1* and *vhl-1* (Epstein et al., 2001; Shen, Shao, and Powell-Coffman, 2006), all the correlations between mutants of these genes and *hif-1(lf)* were positive (see Genetic Interactions Notebook). We reasoned that this apparent contradiction could be due to either strain-specific effects in our N2 background (an artifactual signal) or that it could reflect a previously unrecognized aspect of HIF-1 biology. This motivated us to look for genes that exhibited verifiable extreme patterns of anomalous behavior and led us to propose a new model of the hypoxia pathway (see Identification of non-classical epistatic interactions).

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

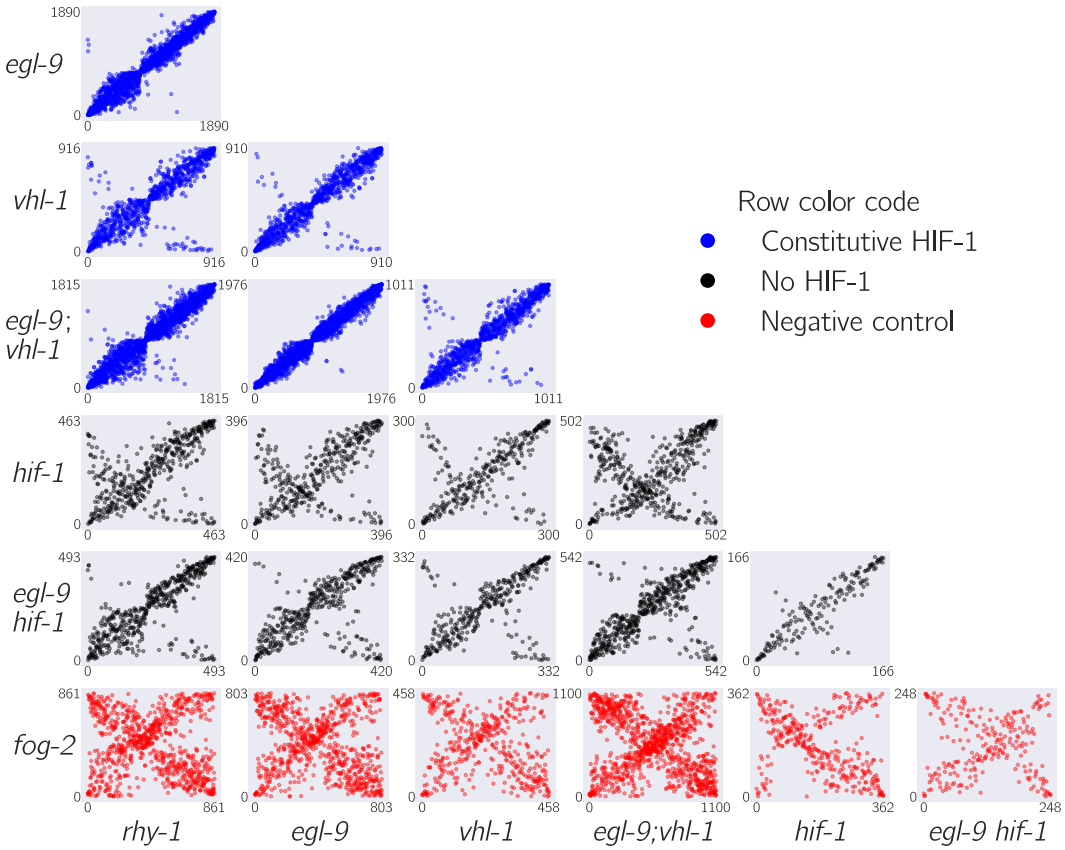


Figure 34 Interacting genes have correlated transcriptional signatures. The rank order of transcripts contained in the shared transcriptional phenotype is plotted for each pairwise combination of genotypes.Correlations between in-pathway genotypes are strong whereas comparisons with a *fog-2(lf)* genotype are dominated by noise. Comparisons between some genotypes show populations of transcripts that are anticorrelated, possibly as a result of feedback loops. Plots are color-coded by row. Comparisons with genotypes with a constitutive hypoxia response are in blue; comparisons with genotypes negative for *hif-1(lf)* are black; and comparisons involving *fog-2(lf)* are red. X- and y-axes show the rank of each transcript within each genotype.

statistic have an unambiguous interpretation.

We found an approach that satisfies all of the above conditions and which can be graphed in an epistasis plot (see Fig 35) In an epistasis plot, the X-axis represents the expected β coefficient for given gene in a double mutant $a^- b^-$ if a and b interact log-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 log-additive (null) model, and can be calculated as the difference between the predicted and the observed β coefficients. Only isoforms that are differentially expressed in all three genotypes are plotted. This attempts to ensure that the isoforms to be examined are regulated by both genes. These plots will generate specific patterns that can be described through linear regressions. The slope of these lines, to which we assign the mathematical notation $s(a, b)$, is the transcriptome-wide epistasis coefficient. Importantly, the transcriptome-wide epistasis coefficient is fundamentally distinct from Pearson or Spearman correlation coefficients and need not have a simple linear mapping. In other words, negative correlation coefficients do not imply a specific sign of the epistasis coefficient, and *vice versa*. For suppression to occur, for example, the only requirement is that the phenotype of the double mutant should match one, and only one, of the two single mutants. The value of the correlation coefficient is not relevant.

Transcriptome-wide epistasis coefficients can be understood intuitively for simple cases of genetic interactions if complete genetic nulls are used. 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 all the mutants should have the same phenotype. It follows that data from this pathway will form line with slope equal to $-\frac{1}{2}$. On the other hand, in the limit of complete genetic inhibition of b by a in an unbranched pathway (i.e., a is in great excess over b, such that under the conditions measured b has no activity),

the plots should show a line of best fit with slope equal to -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 one of two possible slopes that must be determined empirically from the single mutant data. We can use both single mutant data to predict the distribution of slopes that results for the cases stated above. Thus, the transcriptome-wide epistasis coefficient integrates information from many different isoforms into a single number (see Fig. 35).

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 determined the epistasis coefficient of the two gene combinations we studied (*egl-9* and *vhl-1*, and *egl-9* and *hif-1*). In addition to computing an epistasis coefficient from these factors, we would like to know which gene is suppressed in the double mutant. Suppression means that the double mutant should have exactly the phenotype of one and only one mutant, we can simulate the double mutant by replacing the double mutant data with either of the two single mutants and matching the simulated result to the observed result. The result that most closely matches the real data will reveal which gene is being suppressed, which in turn allows us to order the genes along a pathway.

We measured the epistasis coefficient between *egl-9* and *vhl-1*, s(*egl-9 vhl-1*) $= -0.41 \pm 0.01$ (see Epistasis Notebook). Simulations using just the single mutant data showed that the double mutant exhibited the *egl-9(lf)* phenotype (see Fig. 35). We used Bayesian model selection to reject a linear pathway (odds ratio (OR) $> 10^{92}$), which leads us to conclude *egl-9* is upstream of *vhl-1* acting on a phenotype in a branched manner. We also measured epistasis between *egl-9* and *hif-1*, s(*egl-9*, *hif-1*) $= -0.80 \pm 0.01$ (see SI Figure 2), and we found that this behavior could be predicted by modeling *hif-1* downstream of *egl-9*. We also rejected the null

hypothesis that these two genes act in a positive linear pathway (OR> 10^{93}). Taken together, this leads us to conclude that *egl-9* strongly inhibits *hif-1*.

Epistasis between two genes can be predicted using an upstream component

Given our success in measuring epistasis coefficients, we wanted to know whether it would be possible to 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, we reasoned that the *rhy-1(lf)* transcriptome should contain almost 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, predicting $s(rhy-1, vhl-1) = -0.45$. 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 -0.38 ± 0.01. This coefficient was different from -0.50 (OR $> 10^{102}$), reflecting the same qualitative conclusion that *vhl-1* represents a branch in the hypoxia pathway. We were able to obtain a close prediction of the epistasis coefficient for two mutants using the transcriptome of a related, upstream mutant.

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 do not require two gene products to interact biochemically, 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

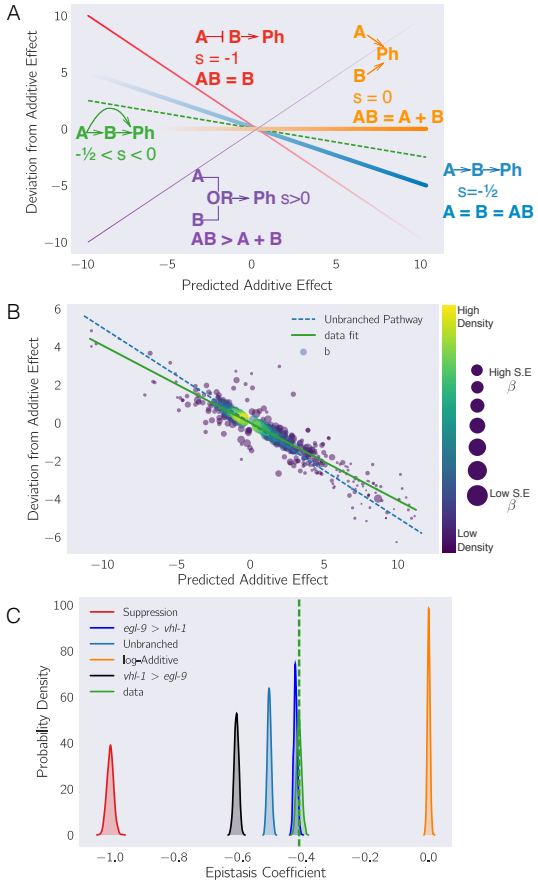


Figure 35 (**A**) Schematic diagram of an epistasis plot. The X-axis on an epistasis plot is the expected coefficient for a double mutant under an log-additive model (null model). The Y-axis plots deviations from this model. Double mutants that deviate in a systematic manner from the null model exhibit transcriptome-wide epistasis (s). To measure s, we find the line of best fit and determine its slope. Genes that act log-additively on a phenotype **(Ph)** will have $s = 0$ (null hypothesis, 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), whereas $s > 0$ correspond to synthetic interactions (purple line). (**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. The green 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. Simulations use only the single mutant data to idealize what expression of the double mutant should look like. $a > b$ means that the phenotype of a is observed in a double mutant $a^{-}b^{-}$.

(the fraction of inputs common to both nodes); and the genes that are regulated by both nodes (the fraction of outputs that are common to both nodes). In other words, we expected that expression profiles associated with a pathway would respond quantitatively to quantitative changes in activity of the pathway. Targeting a pathway at multiple points would lead to expression profile divergence as we compare nodes that are separated by more degrees of freedom, reflecting the flux in information between them.

We investigated this possibility 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. 36). In every case, we see that the weighted correlation decreases monotonically due mainly, but not exclusively, to a smaller STP (see Decorrelation Notebook).

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.

Classical epistasis identifies a core hypoxic response

We searched for genes whose expression obeyed the two epistatic equality relationships, *hif-1(lf)=egl-9(lf) hif-1(lf)* and *egl-9(lf)=egl-9(lf); vhl-1(lf)*, since these equalities define the hypoxia pathway. We excluded genes whose expression deviated from this relationship by more than 2 standard deviations or that had opposite changes in

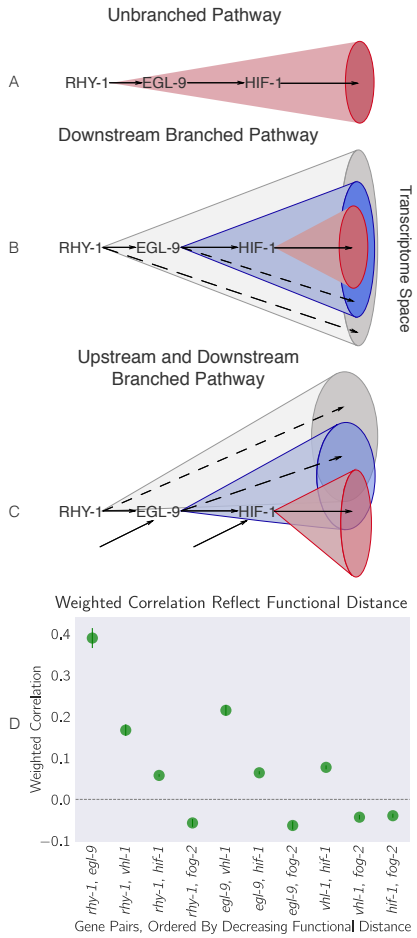


Figure 36 Transcriptomes can be used to order genes in a pathway under certain assumptions. Arrows in the diagrams above are intended to show the direction of flow, and do not indicate valence. **A**. A linear pathway in which *rhy-1* is the only gene controlling *egl-9*, which in turn controls *hif-1* 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.

direction. Using these criteria, we identified 1,258 genes in the hypoxia response. Tissue Enrichment Analysis showed that the intestine and epithelial system were enriched in this response ($q < 10^{-10}$ for both terms), consistent with previous reports (Budde and Roth, 2010). Gene Enrichment Analysis (Angeles-Albores, N. Lee, et al., 2018) showed enrichment in the mitochondrion and in collagen trimers ($q < 10^{-10}$) (see Enrichment Analysis Notebook and SI Figures 3 and 4). This response included 15 transcription factors. Even though HIF-1 is an activator, not all of these genes were up-regulated. We reasoned that only genes that are up-regulated in HIF-1-inhibitor mutants are candidates for direct regulation by HIF-1. We found 264 such genes.

Feedback can be inferred

While some of the rank plots contained a clear positive correlation, others showed a discernible cross-pattern (see Fig. 34). 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 *vhl-1*, *rhy-1* and *egl-9* are all inhibitors of *hif-1(lf)*. Such cross-patterns could be indicative of feedback loops or other complex interaction patterns. If the above is correct, then it should be possible to identify genes that are regulated by *rhy-1* in a logically consistent way: Since loss of *egl-9* causes *rhy-1* mRNA levels to increase, if this increase leads to a significant change in RHY-1 activity, then it follows that the *egl-9(lf)* and *rhy-1(lf)* should show anti-correlation in a subset of genes. Since we do not observe many genes that are anti-correlated, we conclude that is unlikely that the change in *rhy-1* mRNA expression causes a significant change in RHY-1 activity under normoxic conditions. We also searched for genes with *hif-1*-independent, *vhl-1*-dependent gene expression and found 71 genes (SI File 1).

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 relationship between these genes that has not been reported previously.

We identified genes that exhibited violations of the canonical genetic model of the hypoxia pathway (see Fig. 37; also Non-canonical epistasis notebook). We searched for genes that changed in different directions between *egl-9(lf)* and *vhl-1(lf)*, or, equivalently, 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) without specifying any further conditions. We found 56 that satisfied this condition (see Fig. 37, SI File 1). When we checked expression of these genes in the double mutant, we found that *egl-9* remained epistatic over *vhl-1* for this class of genes. This class of genes may in fact be larger because it overlooks genes that have wild-type expression in an *egl-9(lf)* background, altered expression in a *vhl-1(lf)* background, and suppressed (wild-type) expression in an *egl-9(lf); vhl-1(lf)* background. As a result, it could help explain why the *hif-1(lf)* mutant transcriptome is positively correlated with its inhibitors.

Although this entire class had similar behavior, we focused on two genes, *nlp-31* and *ftn-1* which have representative expression patterns. *ftn-1* is described to be responsive to mutations in the hypoxia pathway and has been reported to have aberrant behaviors; specifically, loss of function of *egl-9* and *vhl-1* have opposing effects on *ftn-1* expression (Ackerman and Gems, 2012; Romney et al., 2011). These studies showed the same *ftn-1* expression phenotypes using RNAi and alleles,

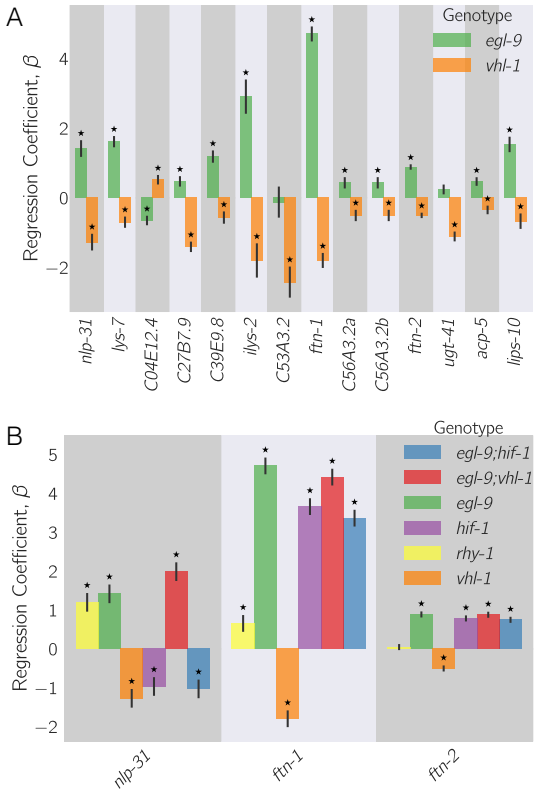


Figure 37 **A**. 56 genes in *C. elegans* exhibit non-classical epistasis in the hypoxia pathway, characterized by opposite effects on gene expression, relative to the wild type, of the *vhl-1(lf)* compared to *egl-9(lf)* (or *rhy-1(lf)*) mutants. Shown are a random selection of 15 out of 56 genes for 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 ($q < 10^{-1}$).

allaying concerns of strain-specific interference. We observed that *hif-1* was epistatic to *egl-9*, and that *egl-9* and *hif-1* both promoted *ftn-1* expression.

Analysis of *ftn-1* 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*, *vhl-1* has the opposite effect to *egl-9*, yet is epistatic to *egl-9*. We propose in the Discussion a novel model for how HIF-1 might regulate these targets.

Discussion

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

We have shown that whole-organism transcriptomic phenotypes can be used to reconstruct genetic pathways and to discern previously uncharacterized genetic interactions. We successfully reconstructed the hypoxia pathway including the order of action of the genetic components and its branching pattern. These results highlight the potential of whole-animal expression profiles for dissecting molecular pathways that are expressed in a large number of cells within an organism. While our results are promising, it remains to be seen whether our approach will also work for pathways that act in a few cells. We selected a previously characterized pathway because *C. elegans* is less amenable to high-throughput screens compared to cultured cells. That said, the striking nature of our results makes us optimistic that this technique could be successfully used to reconstruct unknown pathways.

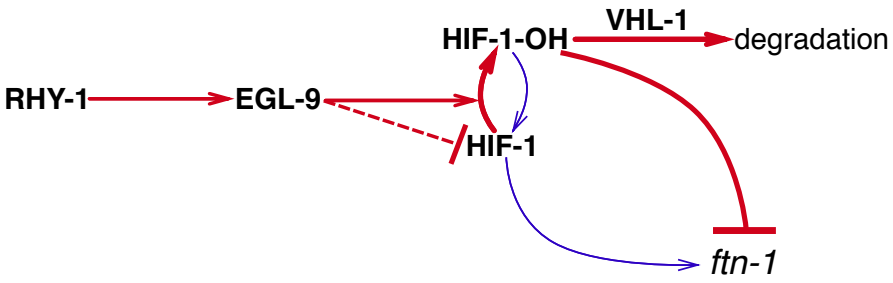
Interpretation of the non-classical epistasis in the hypoxia pathway

The 56 genes that exhibit a striking pattern of non-classical epistasis suggest the existence of previously undescribed aspects of the hypoxia pathway. Some of these non-classical behaviors had been observed previously (Ackerman and Gems, 2012; Romney et al., 2011; Luhachack et al., 2012), but no satisfactory mechanism has been proposed to explain them. Previous studies (Romney et al., 2011; Ackerman and Gems, 2012) suggested that HIF-1 integrates information on iron concentration in the cell to determine its binding affinity 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 (Luhachack et al., 2012) have previously reported that certain genes important for the *C. elegans* immune response against

pathogens reflect similar non-canonical expression patterns. Their interpretation was that *swan-1*, which encodes a binding partner to EGL-9 (Shao, Zhang, Ye, et al., 2010), is important for modulating HIF-1 activity in some manner. The lack of a conclusive double mutant analysis in this work means the role of SWAN-1 in modulation of HIF-1 activity remains to be demonstrated. Other mechanisms, such as tissue-specific differences in the pathway (Budde and Roth, 2010) could also modulate expression, though it is worth pointing out that *ftn-1* expression appears restricted to a single tissue, the intestine (Kim et al., 2004). Another possibility is that *egl-9* controls *hif-1* mRNA stability via other *vhl-1*-independent pathways, but we did not see a decreases in *hif-1* level in *egl-9(lf)*, *rhy-1(lf)* or *vhl-1(lf)* mutants. Another possibility, such as control of protein stability via *egl-9* independently of *vhl-1* (Chintala et al., 2012) will not lead to splitting unless it happens in a tissue-specific manner.

One parsimonious solution 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, the effects of HIF-1 for certain genes like *ftn-1* or *nlp-31* are antagonized by HIF-1-hydroxyl, which is present at only a low level in the cell in normoxia because it is degraded in a *vhl-1*-dependent fashion. This means that loss of *vhl-1* stabilizes HIF-1-hydroxyl. If *vhl-1* is inactivated, genes that are sensitive to HIF-1-hydroxyl will be inhibited as a result of the increase in HIF-1-hydroxyl, despite the increased levels of non-hydroxylated HIF-1. On the other hand, *egl-9(lf)* abrogates the generation of HIF-1-hydroxyl, stimulating accumulation of non-hydroxylated HIF-1 and promoting gene expression. 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. 38). HIF-1-hydroxyl is challenging to study genetically, and if it does have the activity suggested by our genetic evidence this may have prevented such a role from being detected.

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 38 A hypothetical model showing a mechanism where HIF-1-hydroxyl antagonizes HIF-1 in normoxia. **A**. Diagram showing that RHY-1 activates EGL-9. EGL-9 hydroxylates HIF-1 in an oxygen-dependent manner. HIF-1 is rapidly hydroxylated and the product, HIF-1-OH is rapidly degraded in a VHL-1-dependent fashion. EGL-9 can also inhibit HIF-1 in an oxygen-independent fashion. In our model, HIF-1 and HIF-1-OH have opposing effects on transcription. The width of the arrows represents rates in normoxic conditions. **B**. Table showing the effects of loss-of-function mutations on HIF-1 and HIF-1-OH activity, showing how this can potentially explain the *ftn-1* expression levels in each case. S.S = Steady-state.

No mimetic mutations are known with which to study the pure hydroxylated HIF-1 species, and mutations in the Von Hippel-Lindau gene that stabilize the hydroxyl species also increase the quantity of non-hydroxylated HIF-1 by mass action.

Because HIF-1 is detected at low levels in cells under normoxic conditions (Wang and Semenza, 1993), total HIF-1 protein levels are assumed to be so low as to be biologically inactive. However, our data show 1,075 genes change expression in response to loss of *hif-1* under normoxic conditions, which 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.

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

For yet other sets of genes that must change expression in response to the hypoxia pathway, it may not be sufficient 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 robustly to transient fluctuations in the levels of their components (Hart, Antebi, et al., 2012; Hart and Alon, 2013).

Our RNA-seq data suggests that one of these atypical targets of HIF-1 may be RHY-1. Although *rhy-1* does not exhibit non-classical epistasis, all genotypes containing a *hif-1(lf)* mutation 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 auto-regulates both positively and negatively.

Strengths and weaknesses of the methodology

We have described a set of methods that can in principle be applied to any multi-dimensional phenotype. Although we have not applied these methods to *de novo* pathway discovery, we believe that they will be broadly applicable to a wide variety of genetic problems. One aspect of our methodology is the use of whole-organism expression data. Data collection from whole-organisms can be rapid with low technical barriers. On the other hand, a concern is that whole-organism data will average signals across tissues, which would limit the scope of this technology to the study of genetic pathways that are systemic or expressed in large tissues. In reality, our method may be applicable for pathways that are expressed even in a small number of cells in an organism. If a pathway is active in a single cell, this does

not mean that it does not have cell-non-autonomous effects that could be detected on an organism-wide level. Thus, pathways that act in single cells could still be characterized via whole-organism transcriptome profiling. If the non-autonomous effects are long-lasting, then the profiling could take place after the time-of-action of this pathway. In fact, this is how the female-like state in *C. elegans* was recently identified (Angeles-Albores, Leighton, et al., 2017): *fog-2* is involved in translation repression of *tra-2* in the somatic gonad, thereby promoting sperm formation in late larvae (Clifford et al., 2000). Loss of this gene causes non-cell-autonomous effects that can be detected well after the time-of-action of *fog-2* in the somatic gonad has ended. Therefore, we believe that our methodology will be applicable to many genetic cases, with the exception of pathways that acts in complex, antagonistic manners depending on the cell type, or if the pathway minimally affects gene expression.

Genetic analysis of transcriptomic data has proved challenging as a result of its complexity. Although dimensionality reduction techniques such as PCA have emerged as powerful methods with which to understand these data, these methods generate reduced coordinates which are difficult or impossible to interpret. As an example, the first principal component in this paper (see Fig. 33) could be interpreted as HIF-1 pseudo-abundance (Lönnberg et al., 2017). However, another equally reasonable, yet potentially completely different interpretation, is as a pseudo-HIF-1/HIF-1-OH ratio. Another way to analyze genetic interactions is via general linear models (GLMs) that include interaction terms between two or more genes. GLMs can quantify the genetic interactions on single transcripts. We and others (Dixit et al., 2016; Angeles-Albores, Leighton, et al., 2017) have used GLMs to perform epistasis analyses of pathways using transcriptomic phenotypes. GLMs are powerful, but they generate a different interaction coefficient for each gene measured. The large number of coefficients makes interpretation of the genetic interaction between

two mutants difficult. Previous approaches (Dixit et al., 2016) visualize these coefficients via clustered heatmaps. However, two clusters cannot be assumed to be evidence that two genes interact via entirely distinct pathways. Indeed, the non-classical epistasis examples we described here might cluster separately even though a reasonable model can be invoked that does not require any new molecular players.

The epistasis plots shown here are a useful way to visualize epistasis in vectorial phenotypes. We have shown how an epistasis plot can be used to identify interactions between two genes by examining the transcriptional phenotypes of single and double mutants. 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. Epistasis plots and GLMs are not mutually exclusive. A GLM could be used to quantify epistasis interactions at single-transcript resolution, and the results then analyzed using an epistasis plot (for a non-genetic example, see Angeles-Albores, Leighton, et al. (2017)). A benefit of epistasis plots is that they enable the computation of a single, aggregate statistic that describes the ensemble behavior of a set of genes. This aggregate statistic is not enough to describe all possible behaviors in a system, but it can be used to establish whether the genes under study are part of a single pathway. In the case of the hypoxia pathway, phenotypes that are downstream of the hypoxia pathway should conform to the genetic equalities, *egl-9(lf) hif-1(lf)* = *hif-1(lf)* **AND** *egl-9(lf); vhl-1(lf)* = *egl-9(lf)*. Genes whose expression levels behave strangely, yet satisfy these equalities are downstream of the hypoxia pathway. These anomalous genes cannot be identified via the epistasis coefficient but the epistasis coefficient does provide a unifying framework with which to analyze them by constraining the space of plausible hypotheses.

Until relatively recently, the rapid generation and molecular characterization of null mutants was a major bottleneck for genetic analyses. Advances in genomic

engineering mean that, for a number of organisms, production of mutants is now rapid and efficient. As mutants become easier to produce, biologists are realizing that phenotyping and characterizing the biological functions of individual genes is challenging. This is particularly true for whole organisms, where subtle phenotypes can go undetected for long periods of time. We have shown that whole-animal RNA-sequencing is a sensitive method that can be seamlessly incorporated with genetic analyses of epistasis.

Methods

Nematode strains and culture

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

RNA isolation

Lines were synchronized by harvesting eggs via sodium hypochlorite treatment and subsequently plating eggs on food. Worms were staged and based on the time after plating, vulva morphology and the absence of eggs. 30–50 non-gravid young adults were picked and placed in 100 µL of TE pH 8.0 (Ambion AM9849) in 0.2 mL PCR tubes on ice. Worms were allowed to settle or spun down by centrifugation and $\sim$ 80 µL of supernatant removed before flash-freezing in liquid N_2. These samples were digested with Recombinant Proteinase K PCR Grade (Roche Lot No. 03115 838001) for 15 min at 60° in the presence of 1% SDS and 1.25 µL RNA Secure (Ambion AM7005). 5 volumes of Trizol (Tri-Reagent Zymo Research) were added to the RNA samples and treated with DNase I using Zymo Research Quick-RNA MicroPrep R1050. Samples were analyzed run on an Agilent 2100 BioAnalyzer (Agilent Technologies). Replicates were selected that had RNA integrity numbers

equal to or greater than 9.0 and without bacterial ribosomal bands, except for the ZG31 mutant where one of three replicates had a RIN of 8.3.

Library preparation and sequencing

10 ng of total RNA from each sample was reverse-transcribed into cDNA using the Clontech SMARTer Ultra Low Input RNA for Sequencing v3 kit (catalog #634848) in the SMARTSeq2 protocol (Picelli et al., 2014). RNA was denatured at 70°C for 3 min 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 a LNA TSO primer (Picelli et al., 2014), and run at 42°C for 90 minutes, followed by denaturation at 70°C for 10 min. The first strand reaction was used as template for 13 cycles of PCR using the Clontech v3 kit. Reactions were purified with Ampure XP SPRI beads (catalog #A63880). After quantification using the Qubit High Sensitivity DNA assay, a 3 ng aliquot of the 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 1,076 bp. Tagmentation of the full length cDNA 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 on an Illumina HiSeq2500 machine in single read mode with a read length of 50 nt to a depth of 15 million reads per sample. Base calls were performed with RTA 1.13.48.0 followed by conversion to FASTQ with bcl2fastq 1.8.4.

Read alignment and differential expression analysis

We used Kallisto (Bray et al., 2016) to perform read pseudo-alignment and performed differential analysis using Sleuth (Pimentel et al., 2016). We fit a general

linear model for an isoform t in sample i:

$$y_{t,i} = \beta_{t,0} + \beta_{t,genotype} \cdot X_{t,i} + \beta_{t,batch} \cdot Y_{t,i} + \epsilon_{t,i} \tag{3.1}$$

where $y_{t,i}$ was the logarithm transformed counts of isoform t in sample i; $\beta_{t,genotype}$ and $\beta_{t,batch}$ were parameters of the model for the isoform t, and which could be interpreted as biased estimators of the log-fold change; $X_{t,i}, Y_{t,i}$ were indicator variables describing the experimental conditions of the isoform t in sample i; and $\epsilon_{t,i}$ was the noise associated with a particular measurement. After fitting the general linear model, we tested isoforms for differential expression using the built-in Wald-test in Sleuth (Pimentel et al., 2016), which outputs a q-value that has been corrected for multiple hypothesis testing.

Genetic Analysis, Overview

The processed data were analyzed using Python 3.5. We used the Pandas, Matplotlib, Scipy, Seaborn, Sklearn, Networkx, PyMC3, and TEA libraries (McKinney, 2011; Oliphant, 2007; Pedregosa et al., 2012; Salvatier, Wiecki, and Fonnesbeck, 2015; Van Der Walt, Colbert, and Varoquaux, 2011; Hunter, 2007; Angeles-Albores, N. Lee, et al., 2016; Waskom et al., 2016). Our analysis is available in Jupyter Notebooks (Pérez and Granger, 2007). All code and processed data 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/` in html format. Raw reads were deposited in the Short Read Archive under the study accession number SRP100886 and in the GEO under the accession number GSE97355.

Weighted correlations

Correlations between mutants were calculated by identifying their STP. Transcripts were rank-ordered according to their regression coefficient, β. Regressions were

performed using a Student-T distribution with the PyMC3 library (Salvatier, Wiecki, and Fonnesbeck, 2015) (`pm.glm.families.StudenT` in Python). If the correlations had an average value > 1, the average correlation coefficient was set to 1. Weights were calculated as the number of genes that were inliers divided by the number of DEGs present in either mutant.

Epistatic analysis

The epistasis coefficient between two null mutants a and b was calculated as:

$$s(a, b) = \frac{\beta_{a,b} - \beta_a - \beta_b}{\beta_a + \beta_b} \tag{3.2}$$

Null models for various epistatic relationships were generated by sampling the single mutants in an appropriate fashion. For example, to generate the distribution for two mutants that obey the epistatic relationship $a^- = a^- b^-$, we substituted $\beta_{a,b}$ with β_a and bootstrapped the result.

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:

$$P(\alpha \mid M_1, D) \propto \prod_{(x_i, y_i, \sigma_i) \in D} \exp\left[\frac{(y_i - \alpha \cdot x_i)^2}{2\sigma_i^2}\right] \cdot (1 + \alpha^2)^{-3/2}, \tag{3.3}$$

where α was the slope to be determined, x_i, y_i are the of each point, and σ_i was the standard error associated with the y-value. We used equation 3.3 to obtain the most likely slope given the data, D, via minimization (`scipy.optimize.minimize` in Python). Finally, we approximated the odds ratio as:

$$OR = \frac{P(D \mid \alpha^*, M_1) \cdot (2\pi)^{1/2} \sigma_{\alpha^*}}{P(D \mid M_i)}, \tag{3.4}$$

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

Enrichment analysis

Tissue, Phenotype and Gene Ontology Enrichment Analysis were carried out using the WormBase Enrichment Suite for Python (Angeles-Albores, N. Lee, et al., 2018; Angeles-Albores, N. Lee, et al., 2016).

References

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

Angeles-Albores, David, Daniel H. W. Leighton, et al. (2017). "The *Caenorhabditis elegans* Female-Like State: Decoupling the Transcriptomic Effects of Aging and Sperm Status". In: *G3: Genes, Genomes, Genetics* 7.9.

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

– (2018). "Two new functions in the WormBase Enrichment Suite". In: *Micropublication: biology. Dataset.* DOI: https://doi.org/10.17912/W25Q2N.

Bray, Nicolas L et al. (2016). "Near-optimal probabilistic RNA-seq quantification." In: *Nature biotechnology* 34.5, pp. 525–7. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. arXiv: 1505.02710.

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

Budde, Mark W. and Mark B. Roth (2010). "Hydrogen Sulfide Increases Hypoxia-inducible Factor-1 Activity Independently of von Hippel–Lindau Tumor Suppressor-1 in *C. elegans*". In: *Molecular biology of the cell* 21, pp. 212–217. ISSN: 1939-4586. DOI: 10.1091/mbc.E09.

Capaldi, Andrew P et al. (Nov. 2008). "Structure and function of a transcriptional network activated by the MAPK Hog1". In: *Nature Genetics* 40.11, pp. 1300–1306. ISSN: 1061-4036. DOI: 10.1038/ng.235.

Chintala, Sreenivasulu et al. (2012). "Prolyl hydroxylase 2 dependent and Von-Hippel-Lindau independent degradation of Hypoxia-inducible factor 1 and 2 alpha by selenium in clear cell renal cell carcinoma leads to tumor growth inhibition". In: *BMC Cancer* 12.1, p. 293. ISSN: 1471-2407. DOI: 10.1186/1471-2407-12-293.

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

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

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

Goentoro, Lea et al. (2009). "The Incoherent Feedforward Loop Can Provide Fold-Change Detection in Gene Regulation". In: *Molecular Cell* 36.5, pp. 894–899. ISSN: 10972765. DOI: 10.1016/j.molcel.2009.11.018. arXiv: NIHMS150003.

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

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

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

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

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

Hunter, John D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science and Engineering* 9.3, pp. 99–104. ISSN: 15219615. DOI: 10.1109/MCSE.2007.55. arXiv: 0402594v3 [arXiv:cond-mat].

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

Jiang, Huaqi, Rong Guo, and Jo Anne Powell-Coffman (2001). "The *Caenorhabditis eleganshif-1* gene encodes a bHLH-PAS protein that is required for adaptation to hypoxia." In: *Proceedings of the National Academy of Sciences of the United States of America* 98.14, pp. 7916–7921. ISSN: 0027-8424. DOI: 10.1073/pnas.141234698.

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

Kim, Young-Il et al. (2004). "Transcriptional Regulation and Life-span Modulation of Cytosolic Aconitase and Ferritin Genes in *C.elegans*". In: *Journal of Molecular Biology* 342.2, pp. 421–433. ISSN: 00222836. DOI: 10.1016/j.jmb.2004.07.036.

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

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

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

Lönnberg, Tapio et al. (2017). "Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria". In: *Science Immunology* 2.9.

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

Ma, Dengke K. et al. (2012). "CYSL-1 Interacts with the O 2-Sensing Hydroxylase EGL-9 to Promote H 2S-Modulated Hypoxia-Induced Behavioral Plasticity in *C. elegans*". In: *Neuron* 73.5, pp. 925–940. ISSN: 08966273. DOI: 10.1016/j.neuron.2011.12.037. arXiv: NIHMS150003.

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

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226. arXiv: 1111.6189v1.

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

Pedregosa, Fabian et al. (2012). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830. ISSN: 15324435. DOI: 10.1007/s13398-014-0173-7.2. arXiv: 1201.0490.

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

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

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

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

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

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

Salvatier, John, Thomas Wiecki, and Christopher Fonnesbeck (2015). "Probabilistic Programming in Python using PyMC". In: *PeerJ Computer Science* 2.e55, pp. 1–24. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.55. arXiv: 1507.08050.

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

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

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

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

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

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

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

Van Der Walt, Stéfan, S. Chris Colbert, and Gaël Varoquaux (2011). "The NumPy array: A structure for efficient numerical computation". In: *Computing in Science and Engineering* 13.2, pp. 22–30. ISSN: 15219615. DOI: 10.1109/MCSE.2011.37. arXiv: 1102.1523.

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

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

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

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

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

Chapter 4

THE *CAENORHABDITIS ELEGANS* FEMALE-LIKE STATE: DECOUPLING THE TRANSCRIPTOMIC EFFECTS OF AGING AND SPERM STATUS

Abstract

Understanding genome and gene function in a whole organism requires us to fully comprehend the life cycle and the physiology of the organism in question. *Caenorhabditis elegans* XX animals are hermaphrodites that exhaust their sperm after 3 days of egg-laying. Even though *C. elegans* can live for many days after cessation of egg-laying, the molecular physiology of this state has not been as intensely studied as other parts of the life cycle, despite documented changes in behavior and metabolism. To study the effects of sperm depletion and aging of *C. elegans* during the first 6 days of adulthood, we measured the transcriptomes of 1st day adult hermaphrodites; 6th day sperm-depleted adults; and at the same time points, mutant *fog-2(lf)* worms that have a feminized germline phenotype. We found that we could separate the effects of biological aging from sperm depletion. For a large subset of genes, young adult *fog-2(lf)* animals had the same gene expression changes as sperm-depleted 6th day wild-type hermaphrodites, and these genes did not change expression when *fog-2(lf)* females reached the 6th day of adulthood. Taken together, this indicates that changing sperm status causes a change in the internal state of the worm, which we call the female-like state. Our data provide a high-quality picture of the changes that happen in global gene expression throughout the period of early aging in the worm.

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

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

Here, we investigate what we argue is a distinct state in the *C. elegans* life cycle. Although *C. elegans* hermaphrodites emerge into adulthood replete with sperm, after about 3 days of egg-laying the animals become sperm-depleted and can only reproduce by mating. This marks a transition into what we define as the endogenous

female-like state. This state is behaviorally distinguished by increased male-mating success (Garcia et al., 2007), which may be due to an increased attractiveness to males (Morsci, Haas, and Barr, 2011). This increased attractiveness acts at least partially through production of volatile chemical cues (Leighton et al., 2014). These behavioral changes are also coincident with functional deterioration of the germline (Andux and Ellis, 2008), muscle (Herndon et al., 2002), intestine (McGee et al., 2011) and nervous system (J. Liu et al., 2013), changes traditionally attributed to the aging process (T. R. Golden and Melov, 2007).

To decouple the effects of aging and sperm-loss, we devised a two factor experiment. We examined wild-type XX animals at the beginning of adulthood (before worms contained embryos, referred to as 1st day adults) and after sperm depletion (6 days after the last molt, which we term 6th day adults). Second, we examined feminized XX animals that fail to produce sperm but are fully fertile if supplied sperm by mating with males (see Fig. 41). We used *fog-2* null mutants to obtain feminized animals. *fog-2* is involved in germ-cell sex determination in the hermaphrodite worm and is required for sperm production (Schedl and Kimble, 1988; Clifford et al., 2000). *C. elegans* defective in sperm formation will emerge from the larval stage as female adults. As time moves forward, these spermless worms only exhibit changes related to biological aging. As a result, *fog-2(lf)* mutants should show fewer gene changes during the first 6 days of adulthood compared to their egg-laying counterparts that age and also transition from egg-laying into a sperm depleted stage.

Here, we show that we can detect a transcriptional signature associated with loss of hermaphroditic sperm marking entrance into the endogenous female-like state. We can also detect changes associated specifically with biological aging. Biological aging causes transcriptomic changes consisting of 5,592 genes in *C. elegans*. 4,552 of these changes occur in both genotypes we studied, indicating they do not depend on sperm status. To facilitate exploration of the data, we have generated a website where

we have deposited additional graphics, as well as all of the code used to generate these analyses: https://wormlabcaltech.github.io/Angeles_Leighton_2016/

.

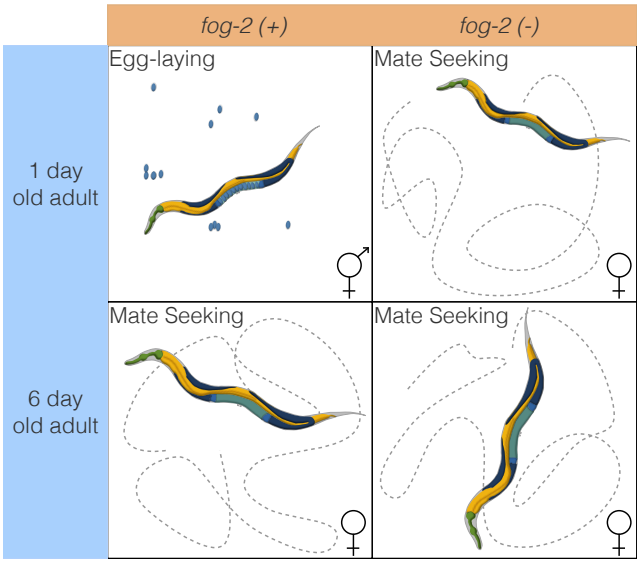


Figure 41 Experimental design to identify genes associated with sperm loss and with aging. Studying the wild-type worm alone would measure time- and sperm-related changes at the same time, without allowing us to separate these changes. Studying the wild-type worm and a *fog-2(lf)* mutant would enable us to measure sperm-related changes but not time-related changes. By mixing both designs, we can measure and separate both modules.

Materials and Methods

Strains

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

RNA extraction

Synchronized worms were grown to either young adulthood or the 6th day of adulthood prior to RNA extraction. Synchronization and aging were carried out

according to protocols described previously (Leighton et al., 2014). 1,000–5,000 worms from each replicate were rinsed into a microcentrifuge tube in S basal (5.85 g/L NaCl, 1 g/L K_2HPO_4, 6 g/L KH_2PO_4), and then spun down at 14,000 rpm for 30 s. The supernatant was removed and 1mL of TRIzol was added. Worms were lysed by vortexing for 30 s at room temperature and then 20 min at 4°. The TRIzol lysate was then spun down at 14,000 rpm for 10 min at 4°C to allow removal of insoluble materials. Thereafter the Ambion TRIzol protocol was followed to finish the RNA extraction (MAN0001271 Rev. Date: 13 Dec 2012). 3 biological replicates were obtained for each genotype and each time point.

RNA-Seq

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

for Bioanalyzer (Agilent Technologies #5067–4626). Libraries were sequenced on Illumina HiSeq2500 in single read mode with the read length of 50nt following manufacturer's instructions. Base calls were performed with RTA 1.13.48.0 followed by conversion to FASTQ with bcl2fastq 1.8.4.

Statistical Analysis

RNA-Seq Analysis.

RNA-Seq alignment was performed using Kallisto (Bray et al., 2016) with 200 bootstraps. Differential expression analysis was performed using Sleuth (Pimentel et al., 2016). The following General Linear Model (GLM) was fit:

$$\log(y_i) = \beta_{0,i} + \beta_{G,i} \cdot G + \beta_{A,i} \cdot A + \beta_{A::G,i} \cdot A \cdot G,$$

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

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

Tissue, Phenotype and Gene Ontology Enrichment Analyses (TEA, PEA and

GEA, respectively) were performed using the WormBase Enrichment Suite for Python (Angeles-Albores, N. Lee, et al., 2016; Angeles-Albores, Lee, et al., 2018). Briefly, the WormBase Enrichment Suite accepts a list of genes and identifies the terms to which these genes are annotated. Terms are annotated by frequency of occurrence, and the probability that a term appears at this frequency under random sampling is calculated using a hypergeometric probability distribution. The hypergeometric probability distribution is extremely sensitive to deviations from the null distribution, which allows it to identify even small deviations from the null.

Data Availability

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

Results and Discussion

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

In order to decouple time-dependent effects from changes associated with loss of hermaphroditic sperm, we measured wild-type and *fog-2(lf)* adults at the 1st day adult stage (before visible embryos were present) and 6th day adult stage, when all wild-type hermaphrodites have laid all their eggs (see Fig 41), but mortality is still low (< 10%) (Stroustrup et al., 2013). We obtained 16–19 million reads mappable to the *C. elegans* genome per biological replicate, which enabled us to identify 14,702 individual genes totalling 21,143 isoforms (see Figure 42a).

One way to analyze the data from this two-factor design is by pairwise comparison

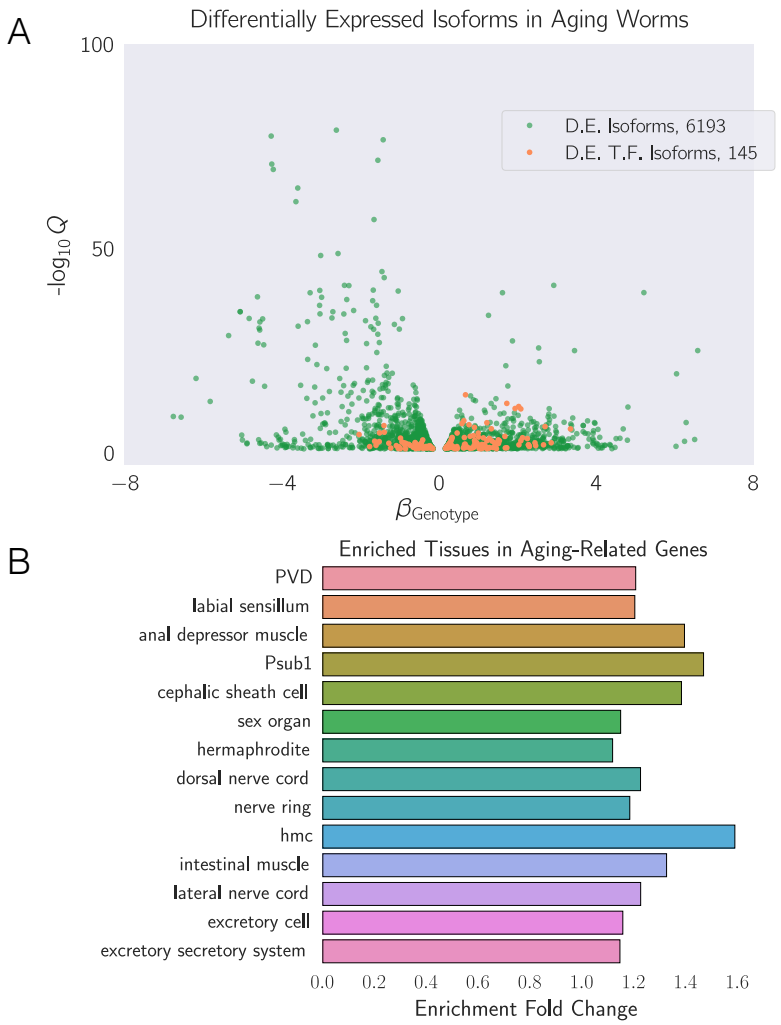


Figure 42 A. Differentially expressed isoforms in the aging category. We identified a common aging expression signature between N2 and *fog-2(lf)* animals, consisting of 6,193 differentially expressed isoforms totaling 5,592 genes. The volcano plot is randomly down-sampled 30% for ease of viewing. Each point represents an individual isoform. β_{Aging} is the regression coefficient. Larger magnitudes of β indicate a larger log-fold change. The y-axis shows the negative logarithm of the q-values for each point. Green points are differentially expressed isoforms; orange points are differentially expressed isoforms of predicted transcription factor genes (Reece-Hoyes et al., 2005). An interactive version of this graph can be found on our website. **B**. Enriched tissues in aging-associated genes. Tissue Enrichment Analysis (Angeles-Albores, N. Lee, et al., 2016) showed that genes associated with muscle tissues and the nervous system are enriched in aging-related genes. Only statistically significantly enriched tissues are shown. Enrichment Fold Change is defined as $Observed/Expected$. hmc stands for head mesodermal cell.

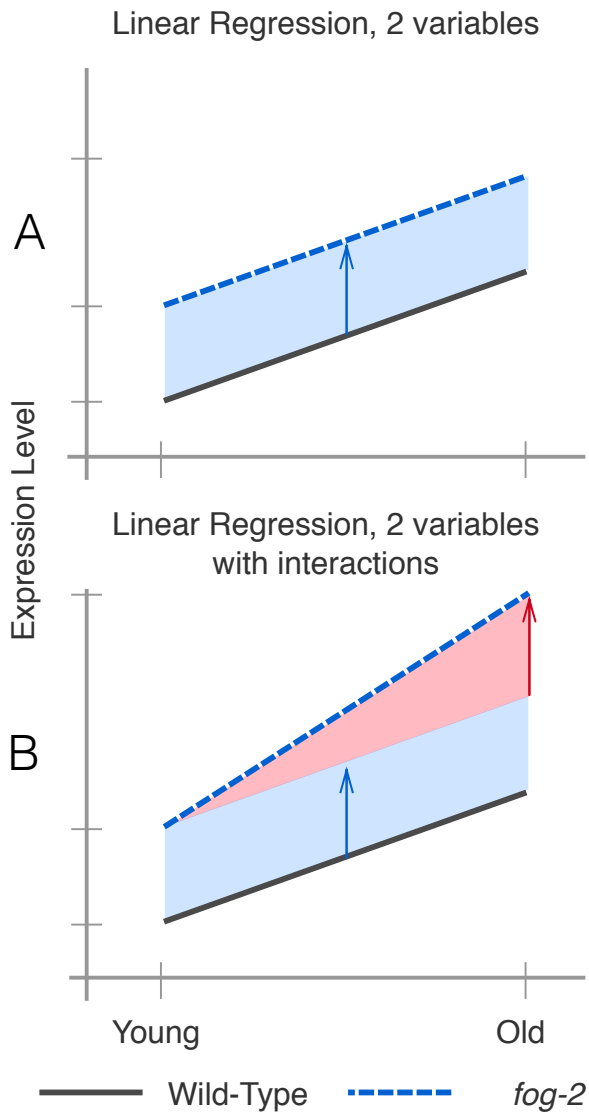


Figure 43 Explanation of linear regressions with and without interactions. **A**. A linear regression with two variables, age and genotype. The expression level of a hypothetical gene increases by the same amount as worms age regardless of genotype. However, *fog-2(lf)* has higher expression of this gene than the wild-type at all stages (blue arrow). **B**. A linear regression with two variables and an interaction term. In this example, the expression level of this hypothetical gene is different between wild-type worms and *fog-2(lf)* (blue arrow). Although the expression level of this gene increases with age, the slope is different between wild-type and *fog-2(lf)*. The difference in the slope can be accounted for through an interaction coefficient (red arrow).

of the distinct states. However, such an analysis would not make full use of all the statistical power afforded by this experiment. Another method that makes full use of the information in our experiment is to perform a linear regression in 3 dimensions (2 independent variables, age and genotype, and 1 output). A linear regression with 1 parameter (age, for example) would fit a line between expression data for young and old animals. When a second parameter is added to the linear regression, said parameter can be visualized as altering the y-intercept, but not the slope, of the first line in question (see Fig. 43a).

Although a simple linear model is oftentimes useful, sometimes it is not appropriate to assume that the two variables under study are entirely independent. For example, in our case, three out of the four timepoint-and-genotype combinations we studied did not have sperm, and sperm-status is associated with both the *fog-2(lf)* self-sterile phenotype and with biological age of the wild-type animal. One way to statistically model such correlation between variables is to add an interaction term to the linear regression. This interaction term allows extra flexibility in describing how changes occur between conditions. For example, suppose a given theoretical gene *X* has expression levels that increase in a *fog-2*-dependent manner, but also increases in an age-dependent manner. However, aged *fog-2(lf)* animals do not have the expression levels of *X* that would be expected from adding the effect of the two perturbations; instead, the expression levels of *X* in this animal are considerably above what is expected. In this case, we could add a positive interaction coefficient to the model to explain the effect of genotype on the y-intercept as well as the slope (see Fig. 43b). When the two perturbations affect a single genetic pathway, these interactions can be interpreted as epistatic interactions.

For these reasons, we used a general linear model with interactions to identify a transcriptomic profile associated with the *fog-2(lf)* genotype independently of age, as well as a transcriptomic profile of *C. elegans* aging common to both genotypes.

The change associated with each variable is referred as β; this number, although related to the natural logarithm of the fold change, is not equal to it. However, it is true that larger magnitudes of β indicate greater change. Thus, for each gene we performed a linear regression, and we evaluated the whether the β values associated with each coefficient were significantly different from 0 via a Wald test corrected for multiple hypothesis testing. A coefficient was considered to be significantly different from 0 if the q-value associated with it was less than 0.1.

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

We identified a transcriptomic signature consisting of 5,592 genes that were differentially expressed in 6th day adult animals of either genotype relative to 1st day adult animals (see S1). This constitutes more than one quarter of the genes in *C. elegans*. Tissue Enrichment Analysis (TEA) (Angeles-Albores, N. Lee, et al., 2016) showed that nervous tissues including the 'nerve ring', 'dorsal nerve cord', 'PVD' and 'labial sensillum' were enriched in genes that become differentially expressed through aging. Likewise, certain muscle groups ('anal depressor muscle', 'intestinal muscle') were enriched. (see Figure 42b). Gene Enrichment Analysis (GEA) (Angeles-Albores, Lee, et al., 2018) revealed that genes that were differentially expressed during the course of aging were enriched in terms involving respiration ('respiratory chain', 'oxoacid metabolic process'); translation ('cytosolic large ribosomal subunit'); and nucleotide metabolism ('purine nucleotide', 'nucleoside phosphate' and 'ribose phosphate' metabolic process). Phenotype Enrichment Analysis (PEA) (Angeles-Albores, Lee, et al., 2018) showed this gene list was associated with phenotypes that affect the *C. elegans* gonad, including 'gonad vesiculated', 'gonad small', 'oocytes lack nucleus' and 'rachis narrow'.

To verify the quality of our dataset, we generated a list of 1,056 golden standard

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

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

The whole-organism *fog-2(lf)* differential expression signature

We identified 1,881 genes associated with the *fog-2(lf)* genotype, including 60 transcription factors (see S2). TEA showed that the terms 'AB', 'somatic gonad', 'uterine muscle', 'cephalic sheath cell', 'spermathecal-uterine junction', and 'PVD' were enriched in this gene set. The 'somatic gonad' and 'spermathecal-uterine junction' are both near the site of action of *fog-2(lf)* (the germline) and possibly reflect physiological changes from a lack of sperm. Phenotype ontology enrichment analysis showed that only a single phenotype term, 'spindle orientation variant' was enriched in the *fog-2(lf)* transcriptional signature ($q < 10^{-1}$, 38 genes, 2-fold enrichment). Most genes annotated as 'spindle orientation variant' were slightly upregulated, and therefore are unlikely to uniquely reflect reduced germline proliferation. GO

term enrichment was very similar to the aging gene set and reflected enrichment in annotations pertaining to translation and respiration. Unlike the aging gene set, the *fog-2(lf)* signature was significantly enriched in 'myofibril' and 'G-protein coupled receptor binding' ($q < 10^{-1}$). Enrichment of the term 'G-protein coupled receptor binding' was due to 14 genes: *cam-1*, *mom-2*, *dsh-1*, *spp-10*, *flp-6*, *flp-7*, *flp-9*, *flp-13*, *flp-14*, *flp-18*, *K02A11.4*, *nlp-12*, *nlp-13*, and *nlp-40*. *dsh-1*, *mom-2* and *cam-1* are members of the Wnt signaling pathway. Most of these genes' expression levels were up-regulated, suggesting increased G-protein binding activity in *fog-2(lf)* mutants.

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

Of the 1,881 genes that we identified in the *fog-2(lf)* signature, 1,040 genes were also identified in our aging set. Moreover, of these 1,040 genes, 905 genes changed in the same direction in response to either aging or germline feminization. The overlap between these signatures suggests an interplay between sperm-status and age. The nature of the interplay should be captured by the interaction coefficients in our model. There are four possibilities. First, the *fog-2(lf)* worms may have a fast-aging phenotype, in which case the interaction coefficients should match the sign of the aging coefficient. Second, the *fog-2(lf)* worms may have a slow-aging phenotype, in which case the interaction coefficients should have an interaction coefficient that is of opposite sign, but not greater in magnitude than the aging coefficient (if a gene increases in aging in a wild-type worm, it should still increase in a *fog-2(lf)* worm, albeit less). Third, the *fog-2(lf)* worms exhibit a rejuvenation phenotype. If this is the case, then these genes should have an interaction coefficient that is of opposite sign and greater magnitude than their aging coefficient, such that the change of these genes in *fog-2(lf)* mutant worms is reversed relative to the wild-type. Finally, if these genes are indicative of a female-like state, then these genes should not change with

age in *fog-2(lf)* animals, since these animals do not exit this state during the course of the experiment. Moreover, because wild-type worms become female as they age, a further requirement for a transcriptomic signature of the female-like state is that aging coefficients for genes in this signature should have genotype coefficients of equal sign and magnitude. In other words, entrance into the female-like state should be not be path-dependent.

To evaluate which of these possibilities was most likely, we selected the 1,040 genes that had aging, genotype and interaction coefficients significantly different from zero and we plotted their temporal coefficients against their genotype coefficients (see Fig. 44a). We observed that the aging coefficients were strongly predictive of the genotype coefficients. Most of these genes fell near the line $y = x$, suggesting that these genes define a female-like state.

We considered how loss-of-function of *fog-2* and aging could both interact to cause entry into this state. We reasoned that a plausible mechanism is that *fog-2* promotes sperm-production, and aging promotes sperm-depletion. This simple pathway model suggests that a double perturbation consisting of aging and loss of function of *fog-2* should show non-additivity of phenotypes (generalized epistasis). To test whether these two perturbations deviate from additivity, we generated an epistasis plot using this gene set. We have previously used epistasis plots to measure transcriptome-wide epistasis between genes in a pathway (Angeles-Albores, Puckett Robinson, et al., 2018). Briefly, an epistasis plot shows the expected expression of a double perturbation under an additive model (null model) on the x-axis, and the deviation from this null model in the y-axis. In other words, we calculated the x-coordinates for each point by adding $\beta_{\text{Genotype}} + \beta_{\text{Aging}}$, and the y-coordinates are equal to $\beta_{Interaction}$ for each isoform. Previously we have shown that if two genes or perturbations act within a linear pathway, an epistasis plot will generate a line with slope equal to -0.5. When we generated an epistasis plot and found the line of best

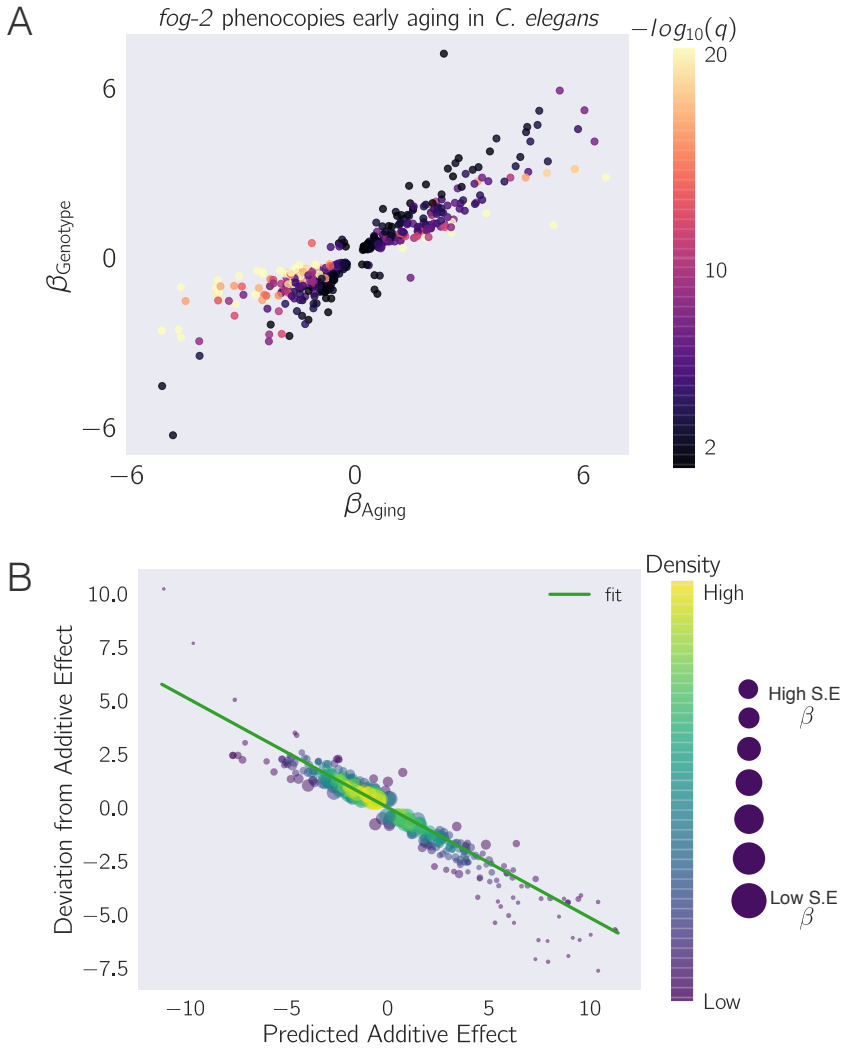


Figure 44 *fog-2(lf)* partially phenocopies early aging in *C. elegans*. The β in each axes is the regression coefficient from the GLM, and can be loosely interpreted as an estimator of the log-fold change. Loss of *fog-2* is associated with a transcriptomic phenotype involving 1,881 genes. 1,040/1,881 of these genes are also altered in wild-type worms as they progress from young adulthood to old adulthood, and 905 change in the same direction. However, progression from young to old adulthood in a *fog-2(lf)* background results in no change in the expression level of these genes. **A**. We identified genes that change similarly during feminization and aging. The correlation between feminization and aging is almost 1:1. **B**. Epistasis plot of aging versus feminization. Epistasis plots indicate whether two genes (or perturbations) act on the same pathway. When two effects act on the same pathway, this is reflected by a slope of -0.5. The measured slope was -0.51 ± 0.01.

fit, we observed a slope of -0.51 ± 0.01, which suggests that the *fog-2* gene and time are acting to generate a single transcriptomic phenotype along a single pathway. Overall, we identified 405 genes that changed in the same direction through age or mutation of the *fog-2(lf)* gene and that had an interaction coefficient of opposite sign to the aging or genotype coefficient (see S3). Taken together, these observations suggests that these 405 genes define a female-like state in *C. elegans*.

Analysis of the female-like state expression signature

To better understand the changes that happen after sperm loss, we performed tissue enrichment, phenotype enrichment and gene ontology enrichment analyses on the set of 405 genes that we associated with the female-like state (see Fig. 45). TEA showed no tissue enrichment using this gene-set. GEA showed that this gene list was enriched in constituents of the ribosomal subunits almost four times above background ($q < 10^{-5}$, 17 genes). The enrichment of ribosomal constituents in this gene set in turn drives the enriched phenotypes: 'avoids bacterial lawn', 'diplotene absent during oogenesis', 'gonad vesiculated', 'pachytene progression during oogenesis variant', and 'rachis narrow'. The expression of most of these ribosomal subunits is down-regulated in aged animals or in *fog-2(lf)* mutants.

Discussion

Defining an Early Aging Phenotype

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

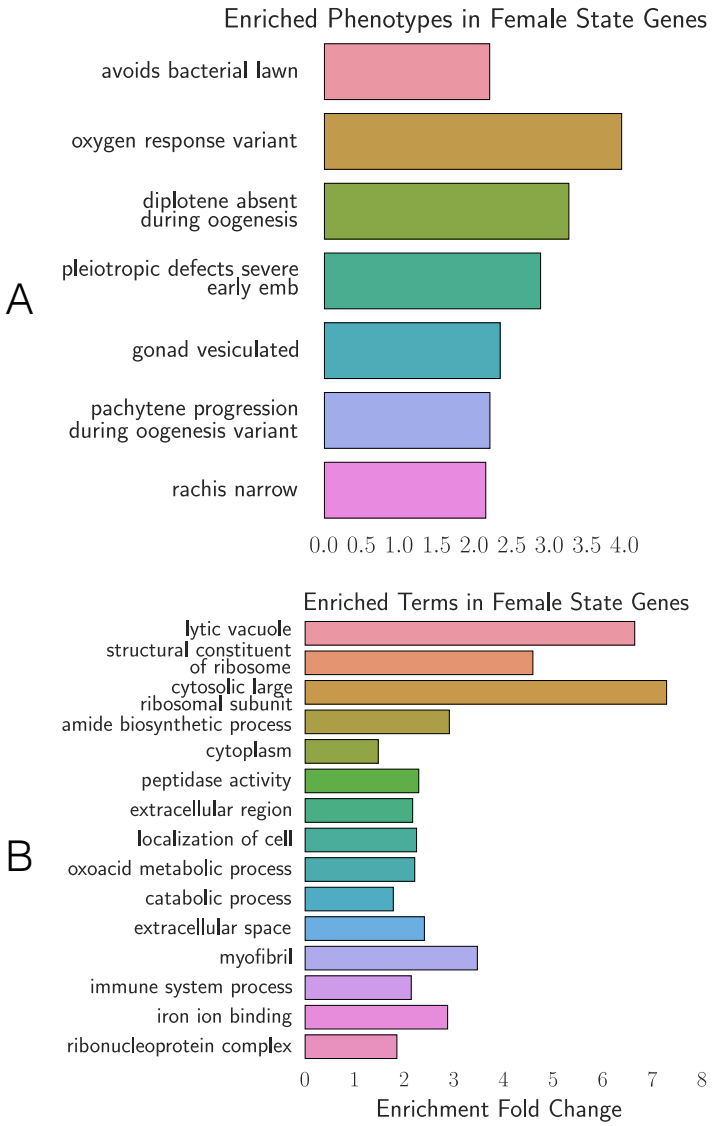


Figure 45 Phenotype and GO enrichment of genes involved in the female-like state. **A**. Phenotype Enrichment Analysis. **B**. Gene Ontology Enrichment Analysis. Most of the terms enriched in PEA reflect the abundance of ribosomal subunits present in this gene set.

aging transcriptome literature, this disagreement can be explained by a difference in methodology; earlier publications typically addressed the aging of fertile wild-type hermaphrodites only indirectly, or queried aging animals at a much later stage of their life cycle.

General linear models enable epistasis measurements

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

The *C. elegans* female-like state

Our explorations have shown that the loss of *fog-2(lf)* partially phenocopies the transcriptional events that occur naturally as *C. elegans* ages from the 1st day of adulthood to the 6th day of adulthood. Moreover, epistasis analysis of these perturbations suggests that they act on the same pathway, namely sperm generation and depletion (see Fig. 46). Self-sperm generation promotes the hermaphrodite

state, whereas sperm depletion marks entry into the female-like state. Given the enrichment of neuronal transcription factors that are associated with sperm loss in our dataset, we believe this dataset should contain some of the transcriptomic modules that are involved in these pheromone production and behavioral pathways, although we have been unable to find these genes.

Behavioral and physiological changes upon mating are not unknown in other species. In particular, in the fruit fly *Drosophila melanogaster*, sex peptide present in the male seminal fluid is known to drive changes in gene expression (H. Liu and Kubli, 2003; Xue and Noll, 2000; Avila et al., 2011; Heifetz et al., 2014; Rezával et al., 2014; Mack et al., 2006) as well as behavior. More recently, sperm was found to be necessary to drive changes in aggression in the fruit fly (Bath et al., 2017). These changes are often reversible upon the disappearance of seminal fluid or sperm. In the case of *C. elegans*, we have observed that sperm loss is associated with gene expression changes that probably reflect physiological changes in the worm. Our experimental design did not include a test for reversibility of these changes. The possibility of a rescue experiment with males raises interesting possibilities: What fraction of the changes observed upon loss of self-sperm are reversible? Do male seminal fluid or male sperm cause changes beyond rescue?

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

C. elegans has a complicated life cycle, with two alternative developmental pathways that have multiple stages (larval development and dauer development), followed by reproductive adulthood. In addition to its developmental stages, researchers have recognized that *C. elegans* has numerous life states that it can enter into when given instructive environmental cues. One such state is the L1 arrest state, where development ceases entirely upon starvation (Johnson et al., 1984; Baugh and Sternberg, 2006). More recently, researchers have described additional diapause

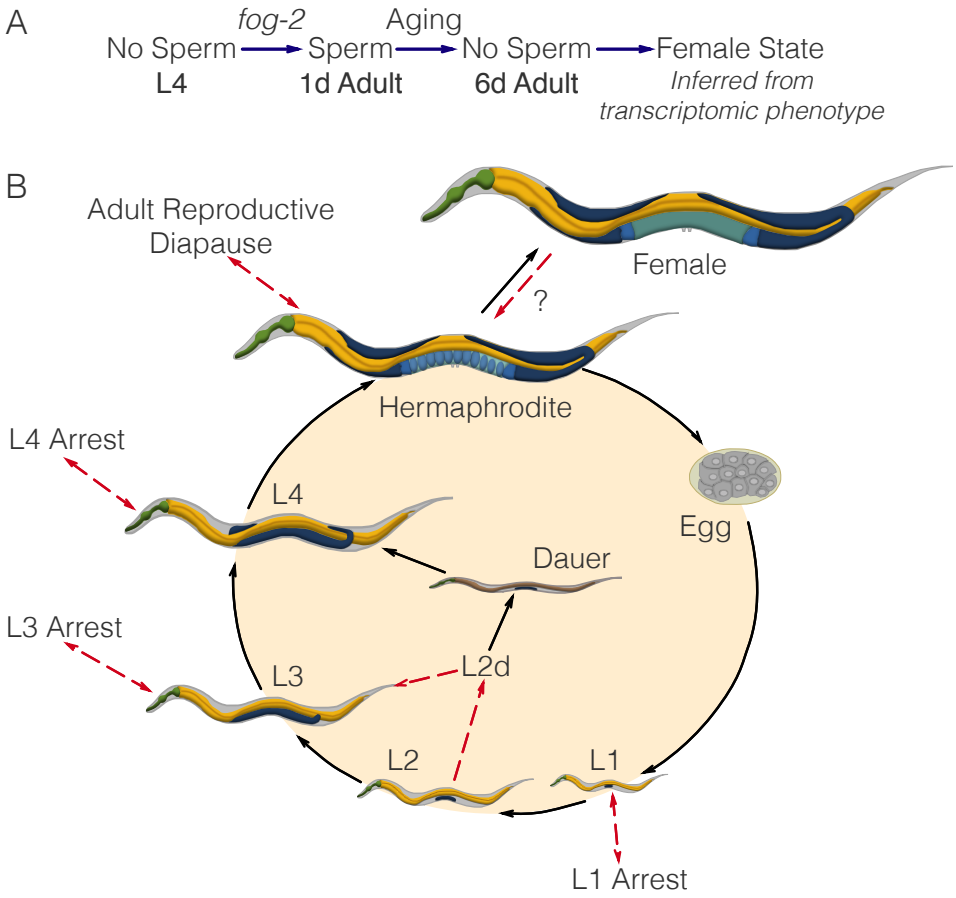


Figure 46 A. A substrate-dependent model showing how *fog-2* promotes sperm generation, whereas aging promotes sperm depletion, leading to entry to the female-like state. Such a model can explain why *fog-2* and aging appear epistatic to each other. **B**. The complete *C. elegans* life cycle. Recognized stages of *C. elegans* are marked by black arrows. States are marked by red arrows to emphasize that at the end of a state, the worm returns to the developmental timepoint it was at before entering the state. The L2d state is an exception. It is the only stage that does not return to the same developmental timepoint; rather, the L2d state is a permissive state that allows entry into either dauer or the L3 stage. We have presented evidence of a female-like state in *C. elegans*. At this point, it is unclear whether the difference between hermaphrodites and females is reversible by males. Therefore, it remains unclear whether it is a stage or a true state.

states that the worm can access at the L3, L4 and young adult stages under conditions of low food (Angelo and Gilst, 2009; Seidel and Kimble, 2011; Schindler, Baugh, and Sherwood, 2014). Not all states of *C. elegans* are arrested, however (see Fig. 46). For example, the L2d state is induced by crowded and nutrient poor conditions (J. W. Golden and Riddle, 1984). While within this state, the worm is capable of entry into either dauer or the L3 larval stage, depending on environmental conditions. Thus, the L2d state is a permissive state, and marks the point at which the nematode development is committed to a single developmental pathway.

Identification of the *C. elegans* life states has often been performed by morphological studies (as in the course of L4 arrest or L2d) or via timecourses (L1 arrest). However, not all states may be visually identifiable, or even if they are, the morphological changes may be very subtle, making positive identification difficult. However, the detailed information afforded by a transcriptome should in theory provide sufficient information to definitively identify a state, since transcriptomic information underlies morphology. Moreover, transcriptomics can provide an insight into the physiology of complex metazoan life states. By identifying differentially expressed genes and using ontology enrichment analyses to identify gene functions, sites of expression or phenotypes that are enriched in a given gene set, we can obtain a clear picture of the changes that occur in the worm analogous to identifying gross morphological changes.

RNA-seq is a powerful technology that has been used successfully in the past as a qualitative tool for target acquisition, though recent work has successfully used RNA-seq to measure genetic interactions via epistasis (Dixit et al., 2016; Angeles-Albores, Puckett Robinson, et al., 2018). Here, we have shown that whole-organism RNA-seq data can also be used to successfully identify internal states in a multicellular organism.

References

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

Angeles-Albores, David, Raymond YN Lee, et al. (2018). "Two new functions in the WormBase Enrichment Suite". In: *Micropublication: biology. Dataset.* DOI: https://doi.org/10.17912/W25Q2N.

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

Angeles-Albores, David, Carmie Puckett Robinson, et al. (Mar. 2018). "Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements." In: *Proceedings of the National Academy of Sciences of the United States of America* 115.13, E2930–E2939. ISSN: 1091-6490. DOI: 10.1073/pnas.1712387115.

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

Avila, Frank W. et al. (Jan. 2011). "Insect Seminal Fluid Proteins: Identification and Function". In: *Annual Review of Entomology* 56.1, pp. 21–40. ISSN: 0066-4170. DOI: 10.1146/annurev-ento-120709-144823.

Bath, Eleanor et al. (May 2017). "Sperm and sex peptide stimulate aggression in female *Drosophila*". In: *Nature Ecology & Evolution* 1.6, p. 0154. ISSN: 2397-334X. DOI: 10.1038/s41559-017-0154.

Baugh, L. Ryan and Paul W. Sternberg (2006). "DAF-16/FOXO Regulates Transcription of *cki-1*/Cip/Kip and Repression of *lin-4* during *C. elegans* L1 Arrest".

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

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

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

Bray, Nicolas L et al. (2016). "Near-optimal probabilistic RNA-seq quantification." In: *Nature biotechnology* 34.5, pp. 525–7. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. arXiv: 1505.02710.

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

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

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

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

Garcia, Hernan G. et al. (2007). "A First Exposure to Statistical Mechanics for Life Scientists". In: p. 27. ISSN: 0036-8075. arXiv: 0708.1899.

Gerstein, Mark B. et al. (2014). "Comparative analysis of the transcriptome across distant species". In: *Nature* 512, pp. 445–448. ISSN: 0028-0836. DOI: 10.1038/nature13424. arXiv: NIHMS150003.

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

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

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

Heifetz, Yael et al. (Mar. 2014). "Mating Regulates Neuromodulator Ensembles at Nerve Termini Innervating the *Drosophila* Reproductive Tract". In: *Current Biology* 24.7, pp. 731–737. ISSN: 09609822. DOI: 10.1016/j.cub.2014.02.042.

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

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

Hunter, John D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science and Engineering* 9.3, pp. 99–104. ISSN: 15219615. DOI: 10.1109/MCSE.2007.55. arXiv: 0402594v3 [arXiv:cond-mat].

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

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

Liu, Huanfa and Eric Kubli (Aug. 2003). "Sex-peptide is the molecular basis of the sperm effect in *Drosophila* melanogaster." In: *Proceedings of the National Academy of Sciences of the United States of America* 100.17, pp. 9929–33. ISSN: 0027-8424. DOI: 10.1073/pnas.1631700100.

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

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

Mack, Paul D et al. (July 2006). "Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.27, pp. 10358–63. ISSN: 0027-8424. DOI: 10.1073/pnas.0604046103.

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

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

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

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

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

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226. arXiv: 1111.6189v1.

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

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

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

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

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

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

Reece-Hoyes, John S. et al. (2005). "A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks." In: *Genome biology* 6.13, R110. ISSN: 1474-760X. DOI: 10.1186/gb-2005-6-13-r110.

Rezával, Carolina et al. (Mar. 2014). "Sexually Dimorphic Octopaminergic Neurons Modulate Female Postmating Behaviors in *Drosophila*". In: *Current Biology* 24.7, pp. 725–730. ISSN: 09609822. DOI: 10.1016/j.cub.2013.12.051.

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

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

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

Stroustrup, Nicholas et al. (2013). "The *Caenorhabditis elegans* Lifespan Machine." In: *Nature methods* 10.7, pp. 665–70. ISSN: 1548-7105. DOI: 10.1038/nmeth.2475. arXiv: NIHMS150003.

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

Sulston, J. E. and H. R. Horvitz (Mar. 1977). "Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*". In: *Developmental Biology* 56.1, pp. 110–156. ISSN: 00121606. DOI: 10.1016/0012-1606(77)90158-0.

Sulston, J. E., E. Schierenberg, et al. (Nov. 1983). "The embryonic cell lineage of the nematode *Caenorhabditis elegans*". In: *Developmental Biology* 100.1, pp. 64–119. ISSN: 00121606. DOI: 10.1016/0012-1606(83)90201-4.

Van Der Walt, Stéfan, S. Chris Colbert, and Gaël Varoquaux (2011). "The NumPy array: A structure for efficient numerical computation". In: *Computing in Science and Engineering* 13.2, pp. 22–30. ISSN: 15219615. DOI: 10.1109/MCSE.2011.37. arXiv: 1102.1523.

Wang, Charles et al. (2014). "The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance." In: *Nature biotechnology* 32.9, pp. 926–32. ISSN: 1546-1696. DOI: 10.1038/nbt.3001. arXiv: NIHMS150003.

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

Xue, L. and M. Noll (Mar. 2000). "*Drosophila* female sexual behavior induced by sterile males showing copulation complementation". In: *Proceedings of the National Academy of Sciences* 97.7, pp. 3272–3275. ISSN: 0027-8424. DOI: 10.1073/pnas.97.7.3272.

Chapter 5

USING TRANSCRIPTOMES AS MUTANT PHENOTYPES REVEALS FUNCTIONAL REGIONS OF A MEDIATOR SUBUNIT IN *C. ELEGANS*

Abstract

Although transcriptomes have recently been used as phenotypes with which to perform epistasis analyses, they are not yet used to study intragenic function/structure relationships. We developed a theoretical framework to study allelic series using transcriptomic phenotypes. As a proof-of-concept, we apply our methods to an allelic series of *dpy-22*, a highly pleiotropic *Caenorhabditis elegans* gene orthologous to the human gene *MED12*, which encodes a subunit of the Mediator complex. Our methods identify functional units within *dpy-22* that modulate Mediator activity upon various genetic programs, including the Wnt and Ras modules.

Introduction

Mutations of a gene can yield a series of alleles with different phenotypes that reveal multiple functions encoded by that gene, regardless of the alleles' molecular nature. In *Caenorhabditis elegans*, allelic series have characterized genes such as *let-23/EGFR*, *lin-3/EGF* and *lin-12/NOTCH* (Aroian and Paul W Sternberg, 1991; Ferguson and Horvitz, 1985; Greenwald, Paul W. Sternberg, and Robert Horvitz, 1983). Allelic series provide a way to probe genes where biochemical approaches would be difficult, slow or uninformative with regards to the biological phenomenon of interest. Their power derives from the ability to draw broad conclusions about the gene of interest in terms of gene dosage and functional units, to the extent that these two factors are separable, without regard to the molecular identity of the

mutations that created these alleles. Here, gene dosage is defined as the combined effects of transcriptional and translational expression, gene product localization, and biochemical kinetics of the final gene product *in situ*. To study allelic series, we must first enumerate the phenotypes each allele affects, and subsequently order the alleles into severity and dominance hierarchies per phenotype. The resulting hierarchies enable us to better understand how a given gene, which may be highly pleiotropic, can give rise to highly specific mutant phenotypes when mutated in just the right way.

Biology has moved from expression measurements of single genes towards genome-wide measurements. Expression profiling via RNA-seq (Mortazavi et al., 2008) enables simultaneous measurement of transcript levels for all genes in a genome, yielding a transcriptome. These measurements can be made on whole organisms, isolated tissues, or single cells (Tang et al., 2009; Schwarz, Kato, and Paul W. Sternberg, 2012). Transcriptomes have been successfully used to identify new cell or organismal states (Angeles-Albores, Leighton, et al., 2017; Villani et al., 2017). Transcriptomic states can be used to perform epistatic analyses (Dixit et al., 2016; Angeles-Albores, Puckett Robinson, Brian A Williams, et al., 2018), but have not been used to characterize allelic series.

We have devised methods for characterizing allelic series using RNA-seq. To test these methods, we selected three alleles (Zhang and Emmons, 2000; Moghal and Paul W. Sternberg, 2003) of a *C. elegans* Mediator complex subunit gene, *dpy-22*. Mediator is a macromolecular complex with $\sim$ 25 subunits (Jeronimo and Robert, 2017) that globally regulates RNA polymerase II (Pol II) (Allen and Taatjes, 2015; Takagi and Kornberg, 2006). The Mediator complex has at least four biochemically distinct modules: the Head, Middle and Tail modules and a CDK-8-associated Kinase Module (CKM). The CKM associates reversibly with other modules, and appears to inhibit transcription (Knuesel et al., 2009; Elmlund et al., 2006). In *C. el-*

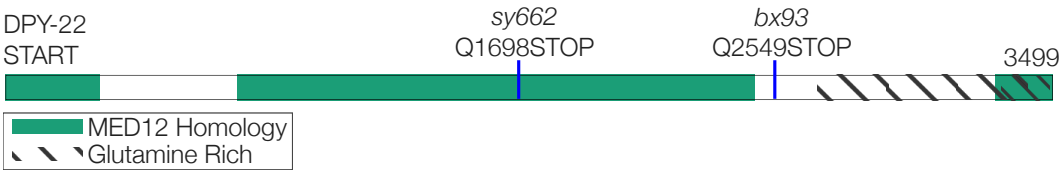


Figure 51 Protein sequence schematic for DPY-22. The positions of the nonsense mutations used are shown.

egans development, the CKM promotes the formation of the male tail (Zhang and Emmons, 2000) (through interactions with the Wnt pathway), as well as formation of the hermaphrodite vulva (Moghal and Paul W. Sternberg, 2003) (through inhibition of the Ras pathway). Null alleles of *dpy-22* are likely to be lethal, based on embryonic lethal phenotypes observed after RNAi (Wang et al., 2004; Lehner et al., 2006) and the severe phenotypes of a strong *dpy-22* hypomorphic allele, *dpy-22(e652)* (homozygous hermaphrodites are very sick) (Riddle et al., 1997). Homozygotes of allele *dpy-22(bx93)*, which encodes a premature stop codon Q2549Amber (Zhang and Emmons, 2000), appear grossly wild-type, though this allele does not have complete wild-type functionality, since it fails to fully complement the Muv phenotype of another allele, *sy622*, in a sensitized *let-23* background. In contrast, animals homozygous for a more severe allele, *dpy-22(sy622)* encoding another premature stop codon, Q1698Amber (Moghal and Paul W. Sternberg, 2003), are dumpy (Dpy), have egg-laying defects (Egl), and have multiple vulvae (Muv) (Fig. 51). In humans, MED12 is known to have a proline-, glutamine- and leucine-rich domain that interacts with the WNT pathway (Kim et al., 2006). However, many disease-causing variants fall outside of this domain (Yamamoto and Shimojima, 2015). In spite of its causative role in a number of neurodevelopmental disorders (Graham and Schwartz, 2013), the structural and functional features of this gene are poorly understood, partially because genetic approaches towards studying pleiotropic genes have proved difficult in the past, highlighting the need for new methods.

Methods

Strains used

Strains used were N2 wild-type (Bristol) (Sulston and Brenner, 1974), PS4087 *dpy-22(sy622)* (Moghal and Paul W. Sternberg, 2003), PS4187 *dpy-22(bx93)* (Zhang and Emmons, 2000), PS4176 *dpy-6(e14) dpy-22(bx93)/+ dpy-22(sy622)* (Moghal and Paul W. Sternberg, 2003), MT4866 *let-60(n2021)* (Beitel, Clark, and Horvitz, 1990), MT2124 *let-60(n1046gf)* (Beitel, Clark, and Horvitz, 1990) and EW15 *bar-1(ga80)* (Eisenmann et al., 1998). Lines were grown on standard nematode growth media (NGM) Petri plates seeded with OP50 *E. coli* at 20°C (Sulston and Brenner, 1974).

Strain synchronization, harvesting and RNA sequencing

With the exception of strain MT4866, strains were synchronized by bleaching P_0's into virgin S. basal (no cholesterol or ethanol added) for 16–18 hours. Arrested L1 larvae were placed in NGM plates seeded with OP50 at 20°C and grown to the young adult stage (assessed by vulval morphology and lack of embryos). We discovered that MT4866 dies upon L1 starvation for this period of time. As a result, we synchronized this strain by double bleaching. Animals were picked if they were young adults, regardless of whether any vulval or morphological phenotypes were present. RNA extraction and sequencing was performed as previously described by Angeles-Albores, Puckett Robinson, Brian A Williams, et al. (2018) and Angeles-Albores, Leighton, et al. (2017). Briefly, young adults were placed in 10 μL of TE buffer, and digested using Recombinant Proteinase K PCR Grade (Roche Lot 656 No. 03115 838001) incubated with 1% SDS 657 and 1.25 μL RNA Secure (Ambion AM7005). Total RNA was extracted using the Zymo Research Directzol RNA MicroPrep Kit (Zymo Research, SKU R2061). mRNA was subsequently purified using a NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, NEB, #E7490).

Sequencing libraries were generated using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB #E7530). These libraries were sequenced using an Illumina HiSeq2500 machine in single-read mode with a read length of 50 nucleotides.

Read pseudo-alignment and differential expression

Reads were pseudo-aligned to the *C. elegans* genome (WBcel235) using Kallisto (Bray et al., 2016), using 200 bootstraps and with the sequence bias (`-seqBias`) flag. The fragment size for all libraries was set to 200 and the standard deviation to 40. Quality control was performed on a subset of the reads using FastQC, RNAseQC, BowTie and MultiQC (Andrews, 2010; Deluca et al., 2012; Langmead et al., 2009; Ewels et al., 2016).

Differential expression analysis was performed using Sleuth (Pimentel et al., 2017). We used a general linear model to identify genes that were differentially expressed between wild-type and mutant libraries. To increase our statistical power, we pooled young adult wild-type replicates from other published (Angeles-Albores, Puckett Robinson, Brian A Williams, et al., 2018; Angeles-Albores, Leighton, et al., 2017) and unpublished analyses adjusting for batch effects. Briefly, batch effects were controlled by including the identity of the person who collected the worms and the method by which the libraries were generated as covariates.å

False hit analysis

To accurately count phenotypes, we developed a false hit algorithm (Algorithm 1). We implemented this algorithm for comparisons of three genotypes using Python. Such an experiment can result in 128 possible combinations of phenotypic classes (ignoring size). This large number of models necessitates an algorithmic approach that can restrict the number of models. Our algorithm uses a noise function that assumes false hit events are non-overlapping (i.e. the same gene cannot be the result of two false positive events in two or more genotypes) to determine the average noise

flux between phenotypic classes. These assumptions break down if false-positive or negative rates are large (>25%).

To benchmark our algorithm, we generated one thousand Venn diagrams at random. For each Venn diagram, we calculated the average false positive and false negative flux matrices. Then, we added noise to each phenotypic class in the Venn diagram, assuming that fluxes were normally distributed with mean and standard deviation equal to the flux coefficient calculated. We input the noised Venn diagram into our false hit analysis and collected classification statistics. For a given signal-to-noise cutoff, λ, classification accuracy varied significantly with changes in the total error rate. In the absence of false negative hits, false hit analysis can accurately identify non-empty genotype-associated phenotypic classes, but identifying genotype-specific classes becomes difficult if the experimental false positive rate is high. On the other hand, even moderate false negative rates ($>$ 10%) rapidly degrade signal from genotype-associated classes. For classes that are associated with three genotypes, an experimental false negative rate of 30% is enough on average to prevents this class from being observed.

We selected $\lambda = 3$ because classification using this threshold was high across a range of false positive and false negative combinations. A challenge to applying this algorithm to our data is the fact that the false negative rate for our experiment is unknown. Although there has been significant progress in controlling and estimating false positive rates, we know of no such attempts for false negative rates. It is unlikely that the false negative rate for our study is lower than the false positive rate, because all genotypes except the controls are likely underpowered. We used false negative rates between 10–20% for false hit analysis. All analyses returned the same final model.

We asked whether re-classification of some classes into others could improve model

fit. We manually re-classified the (*dpy-22(sy622)*,*dpy-22(bx93)*)-associated and the (*dpy-22(bx93)*, *trans-heterozygote*)-associated classes into the *bx93*-associated class (which is associated with all genotypes), and compared χ^2 statistics between a re-classified reduced model ($\chi^2 = 72$) and a reduced model ($\chi^2 = 130$). Based on the lower χ^2 of the re-classified reduced model, we concluded that it is the most likely model given our data.

Algorithm 1 False Hit Algorithm. Briefly, the algorithm initializes a reduced model with the phenotypic class or classes labelled by the largest number of genotypes. This reduced model is used to estimate noise fluxes, which in turn can be used to estimate a signal-to-noise metric between observed and modelled classes. Classes that exhibit a high signal-to-noise are incorporated into the reduced model.

Data: $\mathbf{M}_{obs} = \{N_l\}$, an observed set of classes, where each class is labelled by $l \in L$ and is of size N_l. f_p, f_n, the false positive and negative rates respectively. α, the signal-to-noise threshold for acceptance of a class.
Result: $\mathbf{M}_{reduced}$, a reduced model that fits the data.

begin
  Define a minimal model, **K**
  Refine the model until convergence or iterations max out
  $i \leftarrow 0$
  $\mathbf{K_{prev}} \leftarrow \emptyset$
  while $(i < i_{\max})$ | $(\mathbf{K_{prev}} \neq \mathbf{K})$ **do**
    $\mathbf{K_{prev}} \leftarrow \mathbf{K}$
    Define a noise function to estimate error flows in **K** $\mathbf{F} \leftarrow \text{noise}(\mathbf{K}, f_p, f_n)$
    for $l \in L$ **do**
      Calculate signal to noise for each labelled class False negatives can result in $\lambda < 0$ $\lambda_l \leftarrow \mathbf{M}_{obs,l}/F_l$ **if** $(\lambda > \alpha)$ | $(\lambda < 0)$ **then**
        $\mathbf{K}_l \leftarrow \mathbf{M}_{obs,l}$
      end
    end
    $i++$
  end
end
$\mathbf{M}_{reduced} = \mathbf{K}$
return $\mathbf{M}_{reduced}$

Dominance analysis

We modeled allelic dominance as a weighted average of allelic activity:

$$\beta_{a/b,i,\text{Pred}}(d_a) = d_a \cdot \beta_{a/a,i} + (1 - d_a) \cdot \beta_{b/b,i}, \tag{5.1}$$

where $\beta_{k/k,i}$ refers to the β value of the ith isoform in a genotype k/k, and d_a is the dominance coefficient for allele a.

To find the parameters d_a that maximized the probability of observing the data, we found the parameter, d_a, that maximized the equation:

$$P(d_a|D, H, I) \propto \prod_{i \in S} \exp - \frac{(\beta_{a/b,i,\text{Obs}} - \beta_{a/b,i,\text{Pred}}(d_a))^2}{2\sigma_i^2} \tag{5.2}$$

where $\beta_{a/b,i,\text{Obs}}$ was the coefficient associated with the ith isoform in the *trans*-het a/b and σ_i was the standard error of the ith isoform in the *trans*-heterozygote samples as output by Kallisto. S is the set of isoforms that participate in the regression (see main text). This equation describes a linear regression which was solved numerically.

Code

Code was written in Jupyter notebooks (Pérez and Granger, 2007) using the Python programming language. The Numpy, pandas and scipy libraries were used for computation (Van Der Walt, Colbert, and Varoquaux, 2011; McKinney, 2011; Oliphant, 2007) and the matplotlib and seaborn libraries were used for data visualization (Hunter, 2007; Waskom et al., 2016). Enrichment analyses were performed using the WormBase Enrichment Suite (Angeles-Albores, N. Lee, et al., 2016; Angeles-Albores, Puckett Robinson, Brian A. Williams, et al., 2018). For all enrichment analyses, a q-value of less than 10^{-3} was considered statistically significant. For gene ontology enrichment analysis, terms were considered statistically significant only if they also showed an enrichment fold-change greater than 2.

Data Availability

Raw and processed reads were deposited in the Gene Expression Omnibus. Scripts for the entire analysis can be found with version control in our Github repository, https://github.com/WormLabCaltech/med-cafe. A user-friendly, commented website containing the complete analyses can be found at https://wormlabcaltech.github.io/med-cafe/. Raw reads and quantified abundances for each sample were deposited at the NCBI Gene Expression Omnibus (GEO) (Edgar, Domrachev, and Lash, 2002) under the accession code GSE107523 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107523).

Results

RNA-sequencing of three *dpy-22* alleles and two known interactor genes

We carried out RNA-seq on biological triplicates of mRNA extracted from *dpy-22(sy622)* homozygotes, *dpy-22(bx93)* homozygotes, and wild type controls, along with quadruplicates from *trans*-heterozygotes of both alleles with the genotype *dpy-6(e14) dpy-22(bx93)/+ dpy-22(sy622)*. We also sequenced mRNA extracted from *bar-1(ga80)* (the β-catenin ortholog in *C. elegans*), *let-60(n2021)* and *let-60(n1046gf)* (the Ras ortholog in *C. elegans*) mutants in triplicate because these genes have been previously described to interact with *dpy-22* to form the vulva (Moghal and Paul W. Sternberg, 2003) and the male tail (Zhang and Emmons, 2000). Sequencing was performed at a depth of 20 million reads per sample. Reads were pseudoaligned using Kallisto (Bray et al., 2016). We performed a differential expression using a general linear model specified using Sleuth (Pimentel et al., 2017) (see Methods). Differential expression with respect to the wild type control for each transcript i in a genotype g is measured via a coefficient $\beta_{g,i}$, which can be loosely interpreted as the natural logarithm of the fold-change. Transcripts were considered to have differential expression between wild-type and a mutant if their false discovery rate, q, was less

Genotype	Differentially Expressed Genes
dpy-22(bx93)	266
dpy-6(e14) dpy-22(bx93) / + dpy-22(sy622)	2,128
dpy-22(sy622)	2,036
bar-1(ga80)	4613
let-60(n2021)	509
let-60(n1046gf)	2526

Table 51 The number of differentially expressed genes relative to the wild-type control for each genotype with a significance threshold of 0.1.

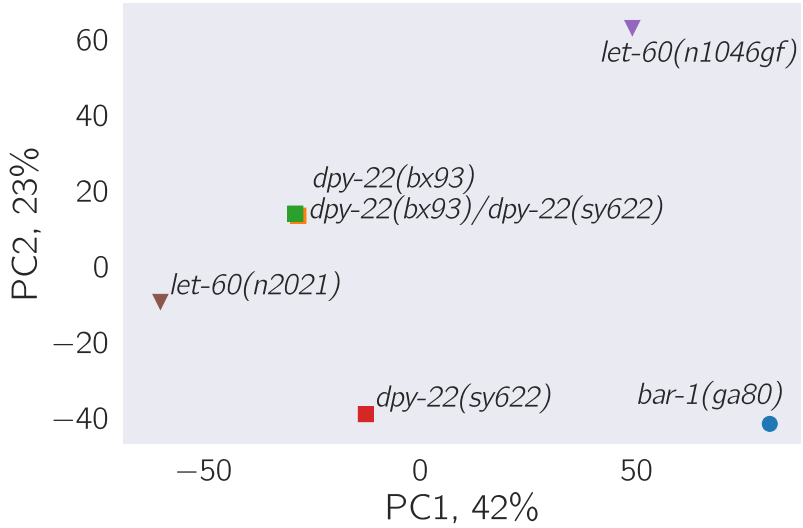


Figure 52 Principal component analysis of the analyzed genotypes. The analysis was performed using only those transcripts that were differentially expressed in at least one genotype. The plot shows that the *trans*-heterozygotes phenocopy the *dpy-22(bx93)* homozygotes along the first two principal dimensions.

than or equal to 10%. We used this method to identify the differentially expressed genes associated with each mutant (Table 51; Basic Statistics Notebook) Supplementary File 1 contains all the beta values associated with this project. We have also generated a website containing complete details of all the analyses available at the following URL: https://wormlabcaltech.github.io/med-cafe/analysis.

Principal component analysis visualizes the allelic dominance of the *dpy-22(bx93)* allele over *dpy-22(sy622)*

As a first step in our analysis, we performed dimensionality reduction on the transcriptomes we sequenced using Principal Component Analysis (PCA). Briefly, PCA identifies the vectors along which there is most variation in the data. These vectors can be used to project the data into lower dimensions to assess whether samples cluster, though interpreting the biological reasons for this clustering can be challenging. To perform PCA, we selected only those transcripts that were differentially expressed in at least one genotype, and used the β coefficients associated with these genes to perform PCA. Projecting the data into two dimensions maintains 65% of the variation. The first dimension separates the gain and loss of function *let-60* mutants. The second dimension separates the *dpy-22* mutants (Fig. 52). On the PCA plot, the *trans*-heterozygote mutants appear to phenocopy the *dpy-22(bx93)* mutants, recapitulating previous experiments that showed the *dpy-22(bx93)* allele to be dominant over the *dpy-22(sy622)* allele.

Three *dpy-22* genotypes have shared transcriptomic phenotypes

We would like to understand the degree and nature of the dominance between these *dpy-22* alleles. To construct a severity and dominance hierarchy, we must establish how many transcriptomic phenotypes are represented among the three *dpy-22* genotypes, and of those phenotypes, how many of them are shared transcriptomic phenotypes (STPs). Shared transcriptomic phenotypes are defined as the set of genes that are commonly differentially expressed in two mutant genotypes relative to a wild-type control, regardless of the direction of change, as defined previously (Angeles-Albores, Puckett Robinson, Brian A Williams, et al., 2018). We use the term in the plural version, because the shared genes may represent multiple independent modules that formally constitute different phenotypic classes.

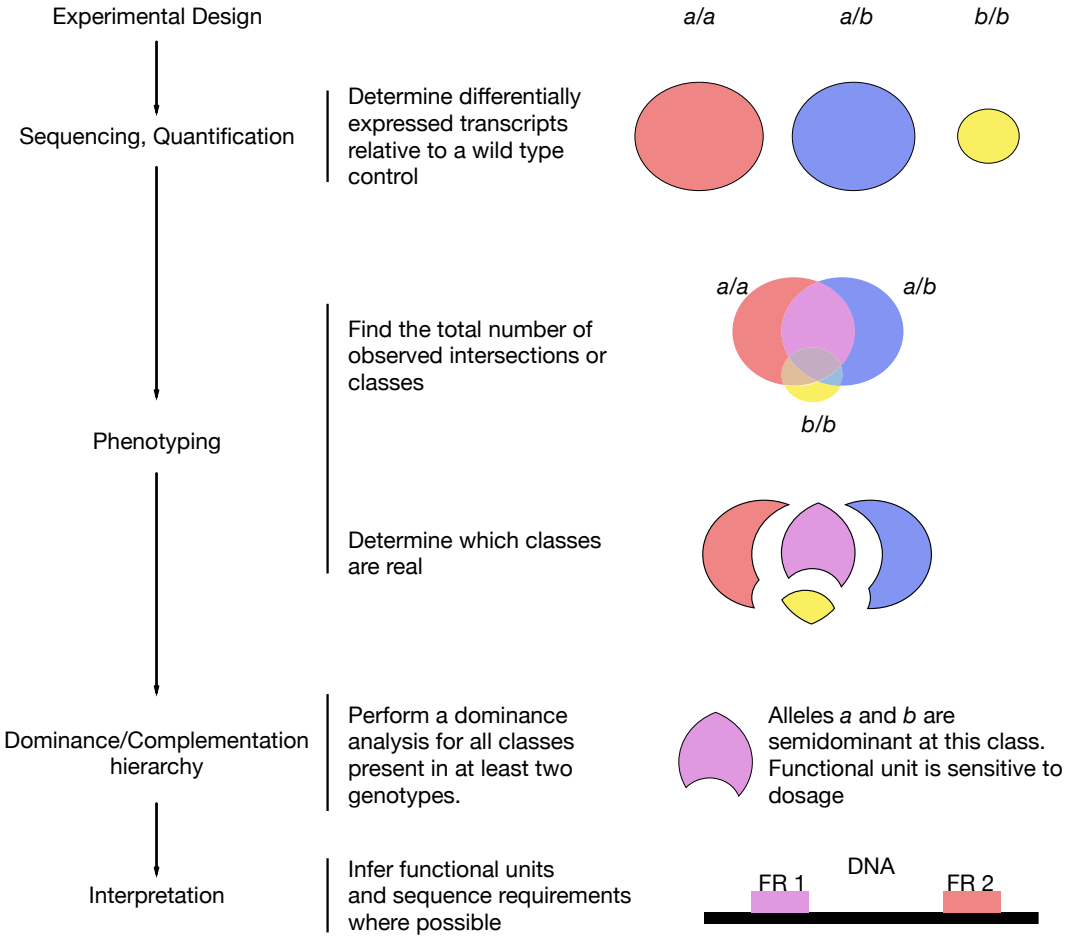


Figure 53 Flowchart for an analysis of arbitrary allelic series. A set of alleles is selected, and the corresponding genotypes are sequenced. Independent phenotypic classes are identified, and classes that are the result of noise are discarded via a false hit analysis. For each phenotypic class, the alleles are ordered in a dominance/complementation hierarchy, which can then be used to infer functional regions (FR) within the genes in question.

We identified significant pairwise STPs between all *dpy-22* mutants. The transcripts that were differentially expressed in *dpy-22(bx93)* homozygotes were almost all differentially expressed in *dpy-22(sy622)* homozygotes (189/266) and in *trans*-heterozygotes (192/266). On the other hand, although *dpy-22(sy622)* homozygotes and *trans*-heterozygotes exhibited a similar number of differentially expressed genes, less than half of these were shared between the two genotypes.

False hit analysis identifies four non-overlapping phenotypic classes

Severity and dominance hierarchies must be calculated with respect to each independent phenotype associated with the alleles under study. A challenge with expression profiles is to identify these independent phenotypes. We reasoned that comparing the expression profiles of the two *dpy-22* homozygotes and the *trans*-heterozygote would naturally partition the expression profiles into groups that would constitute phenotypic classes. However, a three-way comparison can give rise to 7 ($2^3 - 1$) possible groupings: transcripts perturbed in only a single genotype (3), transcripts perturbed in two genotypes (3) and transcripts perturbed in all three genotypes (1). A shortcoming of RNA-seq is that it is prone to false positive and false negative artifacts, and these artifacts could be numerous enough to cause the appearance of certain groups that would not be there otherwise. In other words, we might find a subset of genes that are differentially expressed in a single genotype, but if this subset is small enough, we ought to be concerned that this subset is caused by false positive hits within this genotype or false negative hits in the other genotypes. This thought experiment highlights the need to assess which groups have sufficient statistical support to consider as phenotypic classes.

We developed a method to assess whether groups in a Venn diagram are likely to be the result of statistical artifacts. Briefly, the algorithm works by first assuming all of the data is the result of false positive and false negative hits except for the group of transcripts that is differentially expressed in most genotypes. Then, using estimates for the false positive and negative response, we calculate the expected sizes of all the groups after adding noise under this model. If an observed group is much larger than expected by noise, we refine the data model to accept the group. This process is iterated until the data model converges. We called this method a false hit analysis.

We used false hit analysis to identify four non-overlapping phenotypic classes (Fig. 53). We use the term genotype-specific to refer to groups of transcripts

that were perturbed in one mutant genotype. We use the term genotype-associated to refer to those groups of transcripts whose expression was significantly altered in two or more mutants genotypes with respect to the wild type control. The ***dpy-22(sy622)*-associated** phenotypic class consisted of 665 genes differentially expressed in *dpy-22(sy622)* homozygotes and in *trans*-heterozygotes, but which had wild-type expression in *dpy-22(bx93)* homozygotes. The ***dpy-22(bx93)*-associated** phenotypic class contains 229 genes differentially expressed in all genotypes. The *dpy-22(bx93)*-associated class included re-classified transcripts that had been found to be differentially expressed in the *dpy-22(bx93)* homozygote and one other genotype, because these were very likely to be the result of false negative hits in the missing genotype, and re-classifying these transcripts improved our model substantially. We also identified a ***dpy-22(sy622)*-specific** phenotypic class (1,213 genes) and a ***trans*-heterozygote-specific** phenotypic class (1,302 genes; see the Phenotypic Classes Notebook).

Severity hierarchy of a *dpy-22* allelic series

Having separated the expression profiles into phenotypic classes, we can ask what the severity hierarchy is between the *dpy-22(bx93)* allele and the *dpy-22(sy622)* allele. Broadly speaking, there are two ways to assess severity. First, we can ask which allele causes more mutant phenotypes or phenotypic groups as a homozygote (**allelic pleiotropy**). Alternatively, we can identify the allele which causes the greatest change in expression in a homozygote at each shared phenotype among the homozygotes of both alleles, which we refer to as **allelic magnitude**. An important caveat is that magnitude only makes sense if the homozygotes of each allele are well correlated (i.e., they have a linear relationship with small spread). If the phenotypes have zero or negative correlation between two homozygotes, then the two alleles under inspection are not of the same kind, i.e., they cannot both be loss-of-function

alleles or gain-of-function alleles for this phenotype, though the converse is not necessarily true.

The *dpy-22(sy622)* homozygote shows more differentially expressed genes that participate in a greater number of phenotypic classes relative to the *dpy-22(bx93)* homozygote. Thus, the *dpy-22(sy622)* allele is a more pleiotropic mutation than the *dpy-22(bx93)* allele. Since the homozygotes of each allele only share a single phenotypic class in common, we need only assess magnitude along this single phenotype. To calculate a magnitude coefficient, for genes in the *dpy-22(bx93)*-associated phenotypic class, we plotted the β coefficients from the *dpy-22(sy622)* homozygote against the β coefficients from the *dpy-22(bx93)* homozygote (see Fig. 54) and performed a linear regression to find the slope of this line. Using this method, we found that the *dpy-22(bx93)* homozygote has a magnitude that is 62% ±2% of the *dpy-22(sy622)* homozygote. Taken together, these results suggest that the *dpy-22(sy622)* allele represents a more severe alteration-of-function mutation than the mutation within the *dpy-22(bx93)* allele.

Dominance hierarchy of a *dpy-22* allelic series

We measured allelic dominance for each class using a dominance coefficient (see Methods). The dominance coefficient is a measure of the contribution of each allele to the total expression level in *trans*-heterozygotes. By definition, the *dpy-22(sy622)* allele is completely recessive to *dpy-22(bx93)* for the *dpy-22(sy622)*-specific phenotypic class. To determine the dominance coefficient for the remaining phenotypic classes, we first selected the transcripts within those classes, and asked what linear combination of the homozygotic β coefficients best approximated the β coefficients of the *trans*-heterozygote, subject to the constraint that the sum of the weights for the two homozygotes should be equal to unity. We solved this problem by finding the maximum likelihood estimate for these weights. Using this method, we found

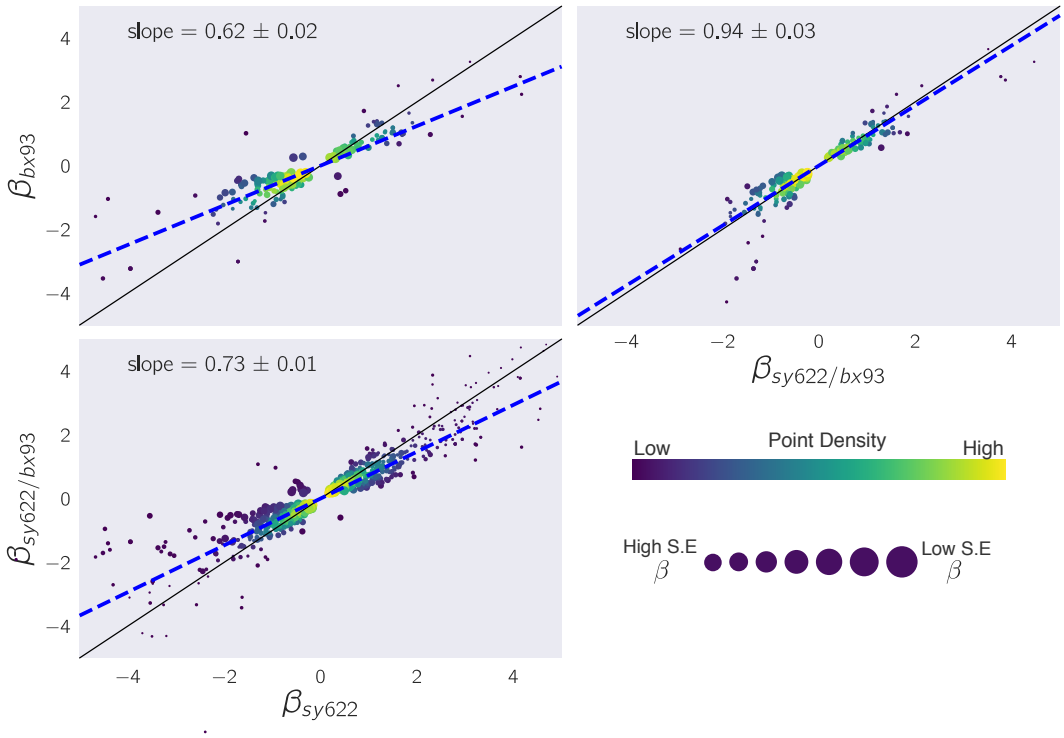


Figure 54 Shared Transcriptomic Phenotypes amongst the *dpy-22* genotypes are regulated in the same direction. For each pairwise comparison, we found those transcripts that were commonly differentially expressed in both genotypes relative to the wild-type control and plotted the β coefficients for each. We performed a linear regression on each plot to find the line of best fit (broken blue line). Only the comparison between *dpy-22(sy622)* and *dpy-22(bx93)* homozygotes was used to establish that the magnitude of the *dpy-22(sy622)* allele is greater than the magnitude of the *dpy-22(bx93)* allele. The other comparisons are shown for completeness.

that the *dpy-22(sy622)* and *dpy-22(bx93)* alleles are semidominant (d_{bx93} = 0.48) to each other for the *dpy-22(sy622)*-associated phenotypic class. The *dpy-22(bx93)* allele is largely dominant over the *dpy-22(sy622)* allele (d_{bx93} = 0.82; see Table 52) for the *dpy-22(bx93)*-associated phenotypic class.

Phenotypic classes reflect morphological phenotypes

We performed enrichment analysis of anatomical, phenotypic and gene ontology terms using the WormBase Enrichment Suite (Angeles-Albores, N. Lee, et al., 2016; Angeles-Albores, Puckett Robinson, Brian A. Williams, et al., 2018). The

Phenotypic Class	Dominance
dpy-22(sy622)-specific	1.00 ± 0.00
dpy-22(sy622)-associated	0.48 ± 0.01
dpy-22(bx93)-associated	0.82 ± 0.01

Table 52 Dominance analysis for the *dpy-22/MDT12* allelic series. Dominance values closer to 1 indicate *dpy-22(bx93)* is dominant over *dpy-22(sy622)*, whereas 0 indicates *dpy-22(sy622)* is dominant over *dpy-22(bx93)*.

dpy-22(bx93)-associated phenotypic class was enriched in genes involved in 'immune system processes' ($q < 10^{-5}$), and was enriched in genes expressed in the 'intestine' ($q < 10^{-4}$). The *dpy-22(sy622)*-associated class was enriched in genes expressed in the 'cephalic sheath cell' ($q < 10^{-4}$). Using ontology enrichment analysis from the WormBase Enrichment Suite, we found that the *dpy-22(sy622)*-associated class is enriched in histones and histone-like proteins ('DNA packaging complex' $q < 10^{-3}$) as well as genes involed in 'immune system processes' ($q < 10^{-5}$). The *dpy-22(sy622)*-specific class was enriched in genes that have expression in the 'intestine' ($q < 10^{-7}$), 'muscular system' ($q < 10^{-3}$) and 'epithelial system' ($q < 10^{-2}$). The genes in this class are known to cause bacterial lawn avoidance when knocked down or knocked out ($q < 10^{-2}$). Finally, GO enrichment showed that the *dpy-22(sy622)*-specific class is specifically enriched in 'structural constituents of cuticle' ($q < 10^{-12}$), and in genes involved in respiration ($q < 10^{-6}$). This last result recapitulates the fact that *dpy-22(sy622)* homozygotes show a severe Dumpy phenotype. The *trans*-heterozygote specific class was enriched in genes expressed in 'male' animals ($q < 10^{-63}$) and genes expressed in the 'reproductive system' ($q < 10^{-21}$). GO enrichment of genes in the *trans*-heterozygote specific class showed enrichment of the genes involved in the 'regulation of cell shape' ($q < 10^{-6}$) and in a variety of terms involving phosphate metabolism, such as 'nucleoside phosphate binding' ($q < 10^{-5}$), 'dephosphorylation' ($q < 10^{-3}$) or 'phosphorylation' ($q < 10^{-2}$), suggesting that this class may be enriched in genes in-

volved in signal transduction though the reason for this enrichment remains unclear. The *dpy-22(bx93)*-specific class did not show enrichment on any test, consistent with our interpretation that this class is the result of random false positive hits.

Predicted interactions of Mediator with Wnt and Ras pathways in *C. elegans*

Previous work in *C. elegans* (Moghal and Paul W. Sternberg, 2003; Zhang and Emmons, 2000) has implicated *dpy-22* as an inhibitor of the Wnt and Ras pathways during the formation of the vulva and the male tail. We obtained expression profiles for *bar-1(ga80)* mutants as well as loss-of-function and gain-of-function Ras mutants, *let-60(n2021)* and *let-60(n1046gf)* respectively. We predicted that the *dpy-22(sy622)*-specific phenotypic class would exhibit the most significant overlap (assessed by a hypergeometric enrichment test) with differentially expressed genes in *let-60(n1046gf)* mutants, whereas the *dpy-22(bx93)*-associated phenotypic class would exhibit the most significant overlap with *bar-1(ga80)* mutants.

The *dpy-22(bx93)*-specific class did not show a transcriptomic signature associated with either the Wnt or the Ras pathway, consistent with our interpretation of this class as false positive (Fig. 55). All other classes showed significant enrichment with genes perturbed in *bar-1(ga80)*. Similarly, *let-60(n2021)* showed enrichment in all real phenotypic classes, with the exception of the *trans*-heterozygote specific class. Contrary to our hypotheses, differentially expressed genes in *let-60(n1046gf)* did not show significant overlap with the *dpy-22(sy622)*-specific phenotype, but they did show significant overlap with all remaining real phenotypic classes.

Discussion

A conceptual framework for analyses of allelic series using transcriptomic phenotypes

Although transcriptomic phenotypes have been used for epistatic analyses (Dixit et al., 2016; Angeles-Albores, Puckett Robinson, Brian A Williams, et al., 2018;

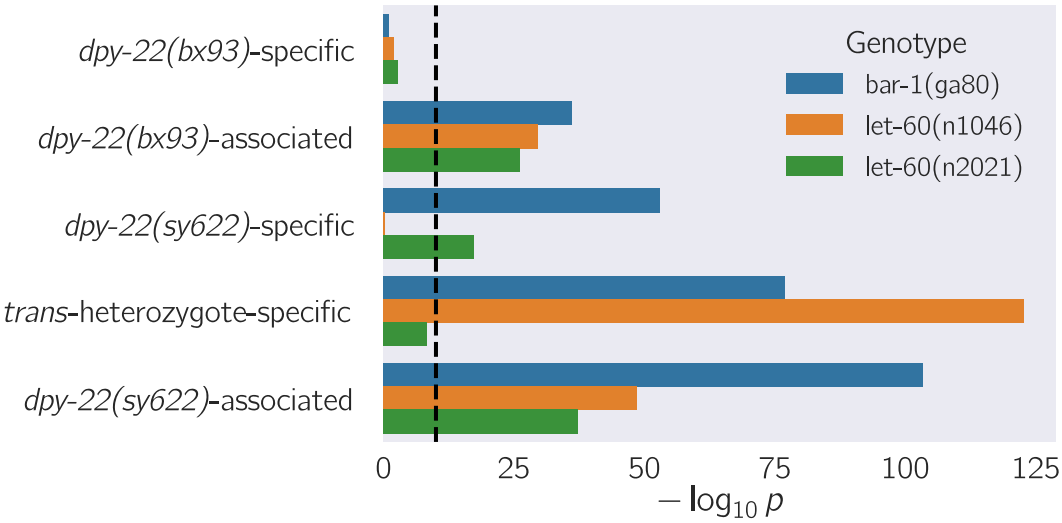


Figure 55 *dpy-22* phenotypic classes are statistically significantly enriched for signatures of *let-60* (ras) and *bar-1* (wnt) signaling. We tested whether the overlap between the differentially expressed genes in *bar-1(ga80)*, *let-60(n1046gf)* or *let-60(n2021)* and the *dpy-22* phenotypic classes was statistically significant using a hypergeometric enrichment test. Since the hypergeometric enrichment test is very sensitive to deviations from random, and since we suspect that there may be a broad genotoxic response to all mutants, we used a statistical significance threshold of $p < 10^{-10}$ (dashed black line).

Angeles-Albores, Leighton, et al., 2017), they have not been used to study gene function in the context of an allelic series. Outstanding challenges for transcriptomes in allelic series were how to count or identify distinct phenotypes within the different transcriptomes, how to order alleles in a severity hierarchy and how to order alleles in a dominance hierarchy. In this work, we present solutions to these problems, and propose a set of unifying concepts that we believe will be useful for future analyses. We re-analyzed an allelic series of the Mediator subunit gene *dpy-22* that had been studied previously (Moghal and Paul W. Sternberg, 2003), recapitulating and extending previous results as a proof of principle for our methodology. In our results, we derived a set of methods that do not rely on the nature of the mutations. In the subsequent discussion, we use the fact that the mutations we used were truncations to derive further insights into the functional units present in this gene.

To interpret our phenotypic classes in a biological context, we investigated whether

these phenotypic classes contained Ras and Wnt expression signatures. Our attempts were partially successful, but a more rigorous analysis awaits the availability of a larger mutant set to establish empirically the overlap that is biologically significant. In part, we reason that some genes may form part of a broad stress response. If that were the case, many mutants may share similar transcriptomic signatures.

Phenotypic classes and their sequence requirements

Because the mutations we used are truncations, our results suggest the existence of various functional regions in *dpy-22/MDT12* (Fig. 56). These functional regions could encode protein domains with biochemical activity, or they could encode biochemically active amino acid motifs, such as nuclear localization sequences or protein binding sites. These functional regions could confer stability to the protein, thereby regulating its levels. As a caveat, we note that we have interpreted the effects these mutations have in terms of their putative effects at the protein level. In the case of our alleles, the relevant homozygotes had wild-type *dpy-22* mRNA levels, suggesting that these mutations do not affect the stability of the mRNA.

The *dpy-22(sy622)*-specific phenotypic class is likely controlled by a single functional region, functional region 1 (FR1). Sequence necessary for wild-type FR1 functionality is encoded between amino acid positions 1 and 2,549, since this is the sequence that is intact in the *bx93* allele. We speculate that this functional region may be the reason that *bx93* is unable to complement the Muv phenotype of *sy622* in a sensitized *let-23* background, since *trans*-heterozygotes in this background exhibit a semidominant Muv phenotype. The *dpy-22(sy622)*-associated phenotypic class is likely controlled by a second functional region, functional region 2 (FR2), and some necessary sequences for wild-type function are encoded between amino acid positions 1,698 and 2,549, but additional sequence could lie between amino acids 1 and 1,698. It is unlikely that FR1 and FR2 are identical because their dominance

behaviors are very different. The *dpy-22(bx93)* allele was largely dominant over the *dpy-22(sy622)* allele for the *dpy-22(bx93)*-associated class, but gene expression in this class was perturbed in both homozygotes. The perturbations were greater for *dpy-22(sy622)* homozygotes than for *dpy-22(bx93)* homozygotes. This behavior can be explained if the *dpy-22(bx93)*-associated class is controlled jointly by two distinct effectors, functional regions 3 and 4 (FR3, FR4, see Fig. 56). Such a model would propose that the sequences necessary for FR3 functionality are within the interval 1 and 2,549, and some sequences necessary for FR4 functionality are encoded between positions 2549 and 3499. This model explains how expression levels of the *bx93*-associated phenotypic class in the *trans*-heterozygote are complemented to the levels of the *bx93* homozygote, because FR3 is complemented in *trans*, but FR4 is defective. Thus, FR3 encodes a functionality that is not dosage-dependent. One possibility is that FR3 is equivalent to FR1 or FR2, and FR4 modifies activity of either of these regions at a subset of loci. A rigorous examination of this model will require studying many alleles that mutate the region between Q1689 and Q2549 using homozygotes and *trans*-heterozygotes.

We also found a class of transcripts that had perturbed levels in *trans*-heterozygotes only; its biological significance is unclear. Phenotypes unique to *trans*-heterozygotes are often the result of physical interactions such as homodimerization, or dosage reduction of a toxic product (Yook, 2005). In the case of *dpy-22/MDT12* orthologs, these explanations seem unlikely since DPY-22 is a monomeric subunit of the CKM. Another possibility is that the *trans*-heterozygote-specific class is the result of complex tissue cross-talk. Massive single-cell RNA-seq of *C. elegans* has recently been reported (Cao et al., 2017), and this tool could provide valuable information regarding this hypothesis. Another possibility is that the *cis*-marker we used for the *bx93* allele, *dpy-6(e14)*, which we assumed to be recessive in all phentoypes, actually has dominant transcriptomic phenotype.

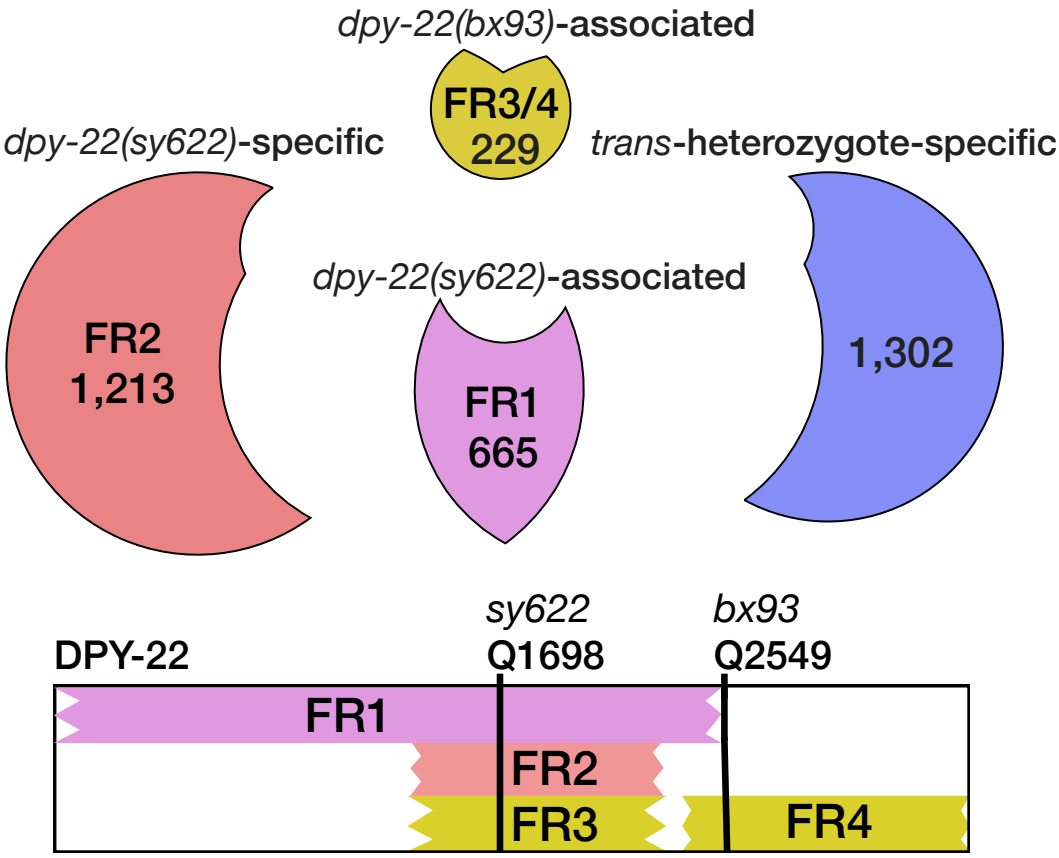


Figure 56 The functional regions associated with each phenotypic class can be mapped intragenically. The number of genes associated with each class is shown. The *dpy-22(bx93)*-associated class may be controlled by two functional regions. FR1 is a dosage-sensitive unit. FR2 and FR3 could be redundant if FR4 is a modifier of FR2 functionality at *dpy-22(bx93)*-associated loci. Note that the *dpy-22(bx93)*-associated phenotypic class is actually three classes merged together. Two of these classes are DE in *dpy-22(bx93)* homozygotes and one other genotype. Our analyses suggested that these two classes are likely the result of false negative hits and genes in these classes should be differentially expressed in all three genotypes, so we merged these three classes together (see Methods).

Occam's razor

Transcriptomic phenotypes generate large amounts of differential gene expression data, so false positive and false negative rates can lead to spurious phenotypic classes whose putative biological significance is misleading. Such artifacts are particularly likely when a phenotypic class is small. Notably, errors of interpretation cannot be avoided by setting a more stringent q-value cut-off: doing so will decrease the false positive rate, but increase the false negative rate, which will in turn produce smaller phenotypic classes than expected. Our method tries to avoid this pitfall by using total error rate estimates to assess the plausibility of each class, though a major drawback is that it relies on a subjective estimation of the false negative rate. These conclusions are of broad significance to research where highly multiplexed measurements are compared to identify similarities and differences in the genome-wide behavior of a single variable under multiple conditions.

We have shown that transcriptomes can be used to study allelic series in the context of a large, pleiotropic gene. We identified separable phenotypic classes that would otherwise be obscured by other methods, correlated each class to a functional region, and identified sequence requirements for each region. Given the importance of allelic series for characterizing gene function and their roles in specific genetic pathways, we are optimistic that this method will be a useful addition to the geneticist's arsenal.

References

Allen, Benjamin L and Dylan J Taatjes (2015). "The Mediator complex: a central integrator of transcription." In: *Nature reviews. Molecular cell biology* 16.3, pp. 155–166. ISSN: 1471-0080. DOI: 10.1038/nrm3951.

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. DOI: citeulike-article-id:11583827.

Angeles-Albores, David, Daniel H. W. Leighton, et al. (2017). "The *Caenorhabditis elegans* Female-Like State: Decoupling the Transcriptomic Effects of Aging and Sperm Status". In: *G3: Genes, Genomes, Genetics* 7.9.

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

Angeles-Albores, David, Carmie Puckett Robinson, Brian A. Williams, et al. (2018). "Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements". In: *Proceedings of the National Academy of Sciences*, p. 201712387. ISSN: 0027-8424. DOI: 10.1073/pnas.1712387115.

Angeles-Albores, David, Carmie Puckett Robinson, Brian A Williams, et al. (Mar. 2018). "Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements." In: *Proceedings of the National Academy of Sciences of the United States of America* 115.13, E2930–E2939. ISSN: 1091-6490. DOI: 10.1073/pnas.1712387115.

Aroian, Raffi V and Paul W Sternberg (1991). "Multiple functions of let-23, a *Caenorhabditis elegans* receptor tyrosine kinase gene required for vulval induction." In: *Genetics* 128.2, pp. 251–67. ISSN: 0016-6731.

Beitel, Greg J., Scott G. Clark, and H. Robert Horvitz (Dec. 1990). "*Caenorhabditis elegans* ras gene *let-60* acts as a switch in the pathway of vulval induction". In: *Nature* 348.6301, pp. 503–509. ISSN: 0028-0836. DOI: 10.1038/348503a0.

Bray, Nicolas L et al. (2016). "Near-optimal probabilistic RNA-seq quantification." In: *Nature biotechnology* 34.5, pp. 525–7. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. arXiv: 1505.02710.

Cao, Junyue et al. (Aug. 2017). "Comprehensive single-cell transcriptional profiling of a multicellular organism." In: *Science (New York, N.Y.)* 357.6352, pp. 661–667. ISSN: 1095-9203. DOI: 10.1126/science.aam8940.

Deluca, David S. et al. (2012). "RNA-SeQC: RNA-seq metrics for quality control and process optimization". In: *Bioinformatics* 28.11, pp. 1530–1532. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts196.

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

Edgar, Ron, Michael Domrachev, and Alex E Lash (Jan. 2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." In: *Nucleic acids research* 30.1, pp. 207–10. ISSN: 1362-4962.

Eisenmann, D M et al. (1998). "The beta-catenin homolog BAR-1 and LET-60 Ras coordinately regulate the Hox gene *lin-39* during *Caenorhabditis elegans* vulval development." In: *Development (Cambridge, England)* 125, pp. 3667–3680. ISSN: 0950-1991.

Elmlund, Hans et al. (Oct. 2006). "The cyclin-dependent kinase 8 module sterically blocks Mediator interactions with RNA polymerase II." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.43, pp. 15788–93. ISSN: 0027-8424. DOI: 10.1073/pnas.0607483103.

Ewels, Philip et al. (2016). "MultiQC: Summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19, pp. 3047–3048. ISSN: 14602059. DOI: 10.1093/bioinformatics/btw354.

Ferguson, E and H. Robert Horvitz (1985). "Identification and characterization of 22 genes that affect the vulval cell lineages of *Caenorhabditis elegans*". In: *Genetics* 110.1, pp. 17–72.

Graham, John M. and Charles E. Schwartz (Nov. 2013). "MED12 related disorders". In: *American Journal of Medical Genetics, Part A* 161.11, pp. 2734–2740. ISSN: 15524825. DOI: 10.1002/ajmg.a.36183.

Greenwald, Iva S., Paul W. Sternberg, and H. Robert Horvitz (Sept. 1983). "The *lin-12* locus specifies cell fates in *Caenorhabditis elegans*". In: *Cell* 34.2, pp. 435–444. ISSN: 00928674. DOI: 10.1016/0092-8674(83)90377-X.

Hunter, John D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science and Engineering* 9.3, pp. 99–104. ISSN: 15219615. DOI: 10.1109/MCSE.2007.55. arXiv: 0402594v3 [arXiv:cond-mat].

Jeronimo, Célia and François Robert (Oct. 2017). *The Mediator Complex: At the Nexus of RNA Polymerase II Transcription*. DOI: 10.1016/j.tcb.2017.07.001.

Kim, Seokjoong et al. (May 2006). "Mediator is a transducer of Wnt/β-catenin signaling". In: *Journal of Biological Chemistry* 281.20, pp. 14066–14075. ISSN: 00219258. DOI: 10.1074/jbc.M602696200.

Knuesel, Matthew T et al. (Feb. 2009). "The human CDK8 subcomplex is a molecular switch that controls Mediator coactivator function." In: *Genes & development* 23.4, pp. 439–51. ISSN: 1549-5477. DOI: 10.1101/gad.1767009.

Langmead, Ben et al. (2009). "Bowtie: An ultrafast memory-efficient short read aligner." In: *Genome biology* 10, R25. DOI: 10.1186/gb-2009-10-3-r25.

Lehner, Ben et al. (Jan. 2006). "Loss of LIN-35, the Caenorhabditis elegans ortholog of the tumor suppressor p105Rb, results in enhanced RNA interference". In: *Genome Biology* 7.1, R4. ISSN: 14656906. DOI: 10.1186/gb-2006-7-1-r4.

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

Moghal, N. and Paul W. Sternberg (2003). "A component of the transcriptional mediator complex inhibits RAS-dependent vulval fate specification in *C. elegans*." In: *Development* 130.1, pp. 57–69. ISSN: 09501991. DOI: 10.1242/dev.00189.

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226. arXiv: 1111.6189v1.

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

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

Pimentel, Harold et al. (2017). "Differential analysis of RNA-seq incorporating quantification uncertainty". In: *Nature Methods* 14.7, pp. 687–690. ISSN: 15487105. DOI: 10.1038/nmeth.4324.

Riddle, Donald L et al. (1997). *C. elegans II*. ISBN: 0879695323. DOI: NBK20183.

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

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

Takagi, Yuichiro and Roger D Kornberg (Jan. 2006). "Mediator as a general transcription factor." In: *The Journal of biological chemistry* 281.1, pp. 80–9. ISSN: 0021-9258. DOI: 10.1074/jbc.M508253200.

Tang, Fuchou et al. (May 2009). "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature Methods* 6.5, pp. 377–382. ISSN: 1548-7091. DOI: 10.1038/nmeth.1315.

Van Der Walt, Stéfan, S. Chris Colbert, and Gaël Varoquaux (2011). "The NumPy array: A structure for efficient numerical computation". In: *Computing in Science and Engineering* 13.2, pp. 22–30. ISSN: 15219615. DOI: 10.1109/MCSE.2011.37. arXiv: 1102.1523.

Villani, Alexandra-Chloé et al. (2017). "Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors". In: *Science* 356.6335. ISSN: 0036-8075. DOI: 10.1126/science.aah4573.

Wang, Jen-Chywan et al. (July 2004). "The Caenorhabditis elegans ortholog of TRAP240, CeTRAP240/let-19, selectively modulates gene expression and is essential for embryogenesis." In: *The Journal of biological chemistry* 279.28, pp. 29270–7. ISSN: 0021-9258. DOI: 10.1074/jbc.M401242200.

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

Yamamoto, Toshiyuki and Keiko Shimojima (June 2015). “A novel MED12 mutation associated with non-specific X-linked intellectual disability”. In: *Human Genome Variation* 2, p. 15018. ISSN: 2054-345X. DOI: 10.1038/hgv.2015.18.

Yook, Karen (2005). “Complementation”. In: *WormBook*. ISSN: 15518507. DOI: 10.1895/wormbook.1.24.1.

Zhang, H. and S. W. Emmons (2000). “A *C. elegans* mediator protein confers regulatory selectivity on lineage-specific expression of a transcription factor gene”. In: *Genes and Development* 14.17, pp. 2161–2172. ISSN: 08909369. DOI: 10.1101/gad.814700.

Chapter 6

TISSUE ENRICHMENT ANALYSIS FOR *C. ELEGANS* GENOMICS

Abstract

Background Over the last ten years, there has been explosive development in methods for measuring gene expression. These methods can identify thousands of genes altered between conditions, but understanding these datasets and forming hypotheses based on them remains challenging. One way to analyze these datasets is to associate ontologies (hierarchical, descriptive vocabularies with controlled relations between terms) with genes and to look for enrichment of specific terms. Although Gene Ontology (GO) is available for *Caenorhabditis elegans*, it does not include anatomical information.

Results We have developed a tool for identifying enrichment of *C. elegans* tissues among gene sets and generated a website GUI where users can access this tool. Since a common drawback to ontology enrichment analyses is its verbosity, we developed a very simple filtering algorithm to reduce the ontology size by an order of magnitude. We adjusted these filters and validated our tool using a set of 30 gold standards from Expression Cluster data in WormBase. We show our tool can even discriminate between embryonic and larval tissues and can even identify tissues down to the single-cell level. We used our tool to identify multiple neuronal tissues that are down-regulated due to pathogen infection in *C. elegans*.

Conclusions Our Tissue Enrichment Analysis (TEA) can be found within WormBase, and can be downloaded using Python's standard pip installer. It tests a slimmed-down *C. elegans* tissue ontology for enrichment of specific terms and

provides users with a text and graphic representation of the results.

Background

RNA-seq and other high-throughput methods in biology have the ability to identify thousands of genes that are altered between conditions. These genes are often correlated in their biological characteristics or functions, but identifying these functions remains challenging. To interpret these long lists of genes, biologists need to abstract genes into concepts that are biologically relevant to form hypotheses about what is happening in the system. One such abstraction method relies on Gene Ontology (GO). GO provides a controlled set of hierarchically ordered terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2015) that provide detailed descriptions about the molecular, cellular or biochemical functions of any gene. For a given gene list, existing software programs can query whether a particular term is enriched (Mi, Dong, et al., 2009; McLean et al., 2010; Huang, Lempicki, and Brad T Sherman, 2009; Pathan et al., 2015). One area of biological significance that GO does not include is anatomy. One way to address this shortcoming is to use a 'tissue ontology' that provides a complete anatomical description for an organism (e.g.'tissue', 'organ' or 'specific cell'), in this case for *C. elegans*. Such an ontology has been described previously for this organism (R. Y. N. Lee and Sternberg, 2003). Cells and tissues are physiologically relevant units with broad, relatively well-understood functionalities amenable to hypothesis formation. The *C. elegans* database, WormBase (Howe et al., 2016), maintains a curated list of gene expression data from the literature. Here we provide a new framework that analyzes a user-input list for enrichment of specific cells and tissues.

Another problem frequently associated with GO enrichment analysis is that it is often difficult to interpret due to the large number of terms associated with a given gene (which we refer to as 'result verbosity'). DAVID, a common tool for GO enrichment

analysis, clusters enriched terms into broad categories (Huang, Brad T. Sherman, et al., 2007), whereas PANTHER (Mi, Dong, et al., 2009; Mi, Muruganujan, and Thomas, 2013) attempts to solve this issue by employing a manually reduced ontology, GOslim (pers._ comm., H. Yu and P. Thomas). To reduce verbosity, we have filtered our ontology using a small set of well-defined criteria to remove terms that do not contribute additional information. To our knowledge, such filtering has not been performed in an algorithmic fashion for a biological ontology before; indeed, DAVID does not employ term trimming *a priori* of testing, but rather fuzzy clustering *post* testing to reduce the number of ontology terms. Other pruning methods do exist (see for example (J. W. Kim, Caralt, and Hilliard, 2007; Garrido and Requena, 2012)), but the pruning is query-dependent or generates a brand new 'brief ontology' which satisfies a set of logic relationships and has certain connectivity requirements. We do not propose to regenerate a new 'brief ontology', but instead we use our approach to select those nodes that have sufficient annotated evidence for statistical testing. We believe our trimming methodology strikes a good balance between detailed tissue calling and conservative testing.

We have developed a tool that tests a user-provided list of genes for term enrichment using a nematode-specific tissue ontology. This ontology, which is not a module of Gene Ontology, is verbose. We select nodes from the ontology for statistical testing using an algorithmic approach, outlined below, that reduces multiple hypothesis testing issues by limiting testing to terms that are well-annotated. The results are provided to the user in a GUI that includes a table of results and an automatically generated bar-chart. This software addresses a previously unmet need in the *C. elegans* community for a tool that reliably and specifically links gene expression with changes in specific cells, organs or tissues in the worm.

Results

Generating a Gene-Tissue Dictionary by Specific Node Selection

Reducing term redundancy through a similarity metric

For our tool, we employ a previously generated cell and tissue ontology for *C. elegans* (R. Y. N. Lee and Sternberg, 2003), which is maintained and curated by WormBase. This ontology contains thousands of anatomiy terms, but not every term is equally well-annotated. As a first step to generate our tissue enrichment software, we wished to select tissue terms that were reasonably well-annotated, yet specific enough to provide insight and not redundant with other terms. For example, nematodes have a number of neurons that are placed symmetrically along the left/right body axis, and are functionally similar. These left/right neuronal pairs (which are sisters in the ontology) have almost identical annotations, with at most one or two gene differences between them, and therefore we cannot have statistical confidence in differentiating between them. As a result, testing these sister terms provides no additional information compared with testing only the parent node to these sisters. To identify redundancy, we defined two possible similarity metrics (see *Methods* section and Figure 61) that can be used to identify ontology sisters that have very high similarity between them. Intuitively, a set of sisters can be considered very similar if they share most gene annotations. Within a given set of sisters, we can calculate a similarity score for a single node by counting the number of unique annotations it contains and dividing by the total number of unique annotations in the sister set. Having assigned to each sister a similarity score, we can identify the **average** similarity score for this set of sisters, and if this average value exceeds a threshold, these sisters are not considered testable candidates. An alternative method is check whether **any** of the scores exceeds a predetermined threshold, and if so remove this sister set from the ontology. We referred to these two scoring criteria as '**avg**' and '**any**' respectively.

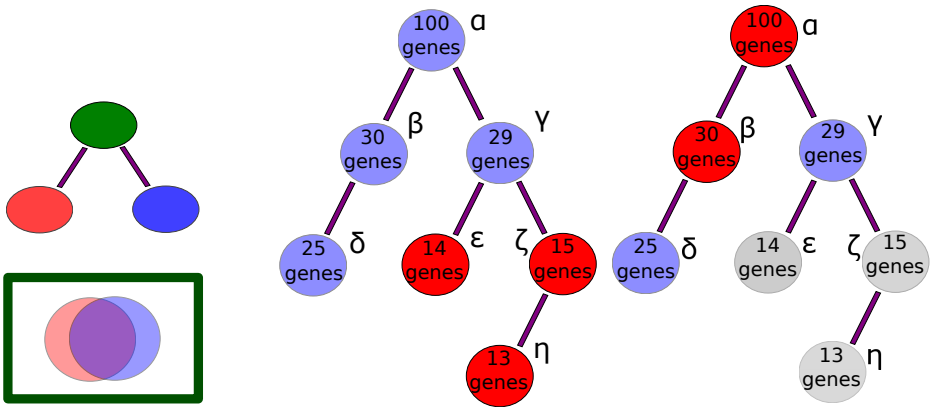


Figure 61 Schematic representation of trimming filters for an acyclical ontology. **a.** The parent node (green) contains at least as many annotations as the union of the two sisters. These two sisters share annotations extensively, as expressed by the overlap in the Venn diagram, so they qualify for removal. **b.** Nodes with less than a threshold number of genes are trimmed (red) and discarded from the dictionary. Here, the example threshold is 25 genes. Nodes ϵ, ζ, η, shown in red are removed. **c.** Parent nodes are removed recursively, starting from the root, if all their daughter nodes have more than the threshold number of annotations. Nodes in grey (ϵ, ζ, η) were removed in the previous step. Nodes α, β shown in red are trimmed because each one has a complete daughter set. Only nodes γ and δ will be used to generate the static dictionary.

Terminal branch terms and parent terms can be safely removed in an algorithmic fashion

Another problem arises from the ontology being scarcely populated. Many nodes have 0–10 annotations, which we consider too few to accurately test. To solve this issue, we implemented another straightforward node selection strategy. For a given terminal node, we test whether the node has more than a threshold number of annotations. If it does not, the node is not used for statistical testing. The next higher node in the branch is tested and removed recursively until a node that satisfies the condition is found. At that point, no more nodes can be removed from that branch. This completion is guaranteed by the structure of the ontology: parent nodes inherit all of the annotations of all of their descendants, so the number of annotated terms monotonically increases with increasing term hierarchy (see Figure 61). In this way,

we ensure that our term dictionary includes only those tissues that are considered sufficiently well annotated for statistical purposes.

Additionally, we reasoned that for any parent node if all its daughters were selected for testing, there was no additional benefit to test the parent. We removed parent nodes from the analysis if all their daughter nodes passed the annotation threshold (see Figure 61). We called this a ceiling filter. Applying these three filters reduced the number of ontology terms by an order of magnitude.

Filtering greatly reduces the number of nodes used for analysis

By itself, each of these filters can reduce the number of nodes employed for analysis, but applying the filters in different orders removes different numbers of nodes (not all the filters are commutative). We chose to always execute annotation and similarity thresholding first, followed by the ceiling filter. For validation (see below) we made a number of different dictionaries. The original ontology has almost 6,000 terms of which 1675 have at least 5 gene annotations. After filtering, dictionary sizes ranged from 21 to a maximum of 460 terms, which shows the number of terms in a scarcely annotated ontology can be reduced by an order of magnitude through the application of a few simple filters (see Table 61). These filters were used to compile a static dictionary that we employ for all analyses (see *Validation of the algorithm and parameter selection* section for details). Our trimming pipeline is applied as part of each new WormBase release. This ensures that the ontology database we are using remains up-to-date with regards to both addition or removal of specific terms as well as with regard to gene expression annotations.

Tissue enrichment testing via a hypergeometric model

Having built a static dictionary, we generated a Python script that implements a significance testing algorithm based on the hypergeometric model. Briefly, the

Table 61 Parameter specifications and number of tissues for all dictionaries. The 'Method' column refers to the trimming criterion for the similarity metric. We used two such criteria, 'any' and 'avg'.'any': For a given sister set, if any sister had a similarity exceeding the corresponding threshold, all sisters were removed from the final dictionary. 'avg': For a given sister set, if the average similarity across all the sisters in the set was greater than the corresponding threshold, all sisters were removed from the final dictionary.

Annotation Cutoff	Similarity Threshold	Method	No. Of Terms in Dictionary
25	0.9	any	460
25	0.9	avg	461
25	0.95	any	466
25	0.95	avg	468
25	1.0	any	476
25	1.0	avg	476
33	0.9	any	261
33	0.9	avg	255
33	0.95	any	261
33	0.95	avg	262
33	1.0	any	247
33	1.0	avg	247
50	0.9	any	83
50	0.9	avg	77
50	0.95	any	82
50	0.95	avg	81
50	1.0	any	70
50	1.0	avg	70
100	0.9	any	45
100	0.9	avg	35
100	0.95	any	42
100	0.95	avg	36
100	1.0	any	21
100	1.0	avg	21

hypergeometric model tests the probability of observing n_i occurences of a tissue i in a list of size M if there are m_i labels for that tissue in a dictionary of total size N that are drawn without replacement. Mathematically, this is expressed as:

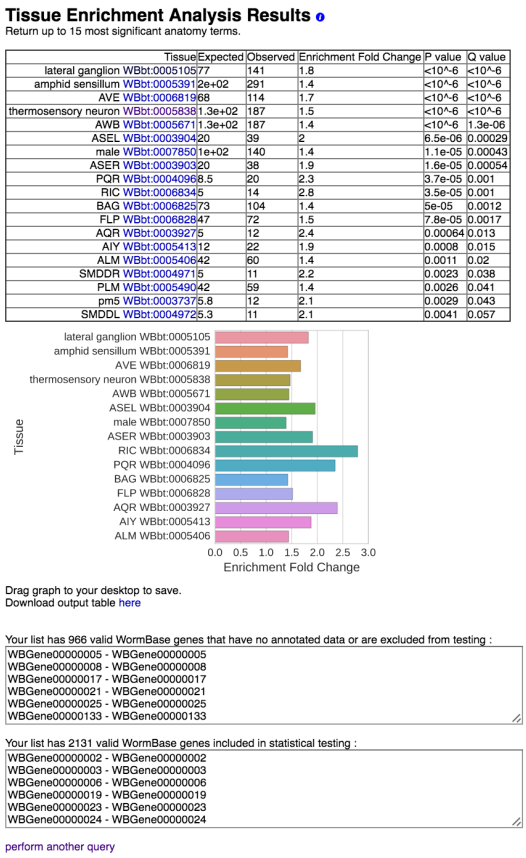
$$\mathrm{P}(n_i|N, m_i, M) = \frac{\binom{m_i}{n_i}\binom{M-m_i}{N-n_i}}{\binom{N}{n_i}}. \tag{6.1}$$

Although a user will input gene IDs, we test the number of ocurrences of a term within the gene list, so a single gene can contribute to multiple terms. Due to the discrete nature of the hypergeometric distribution, this algorithm can generate artifacts when the list is small. To avoid spurious results, a tissue is never considered significant if there are no annotations for it in the user-provided list.

Once the p-values for each term have been calculated, we apply a standard FDR correction using a Benjamini-Hochberg step-up algorithm (Benjamini and Hochberg, 1995). FDR corrected p-values are called q-values. Genes that have a q-value less than a given alpha are considered significant. Our default setting is an alpha of 0.1, which is a standard threshold broadly agreed upon by the scientific community (see for example (Love, Huber, and Anders, 2014; Pawitan et al., 2005; Storey and Tibshirani, 2003)). This threshold cannot be altered in the web GUI, but is user tunable through our command-line implementation.

Users input a gene list using any valid gene name for *C. elegans*. These names are processed into standard WormBase gene IDs (WBGene IDs). The program returns a table containing all the enriched terms and associated information such as number of terms in gene list and expected number of terms. Finally, the program can also return a bar chart of the enrichment fold change for the fifteen tissues with the lowest measured q-values. The bars in the graph are sorted in ascending order of

q-value and then in descending order of fold-change. Bars are colored for ease of viewing, and color does not convey information. Our software is implemented in an easy to use GUI (see Figure 62). Anatomy terms are displayed in human-readable format followed by their unique ontology ID (WBbt ID). In summary, each time the ontology annotations are updated, a new trimmed ontology is generated using our filters; in parallel, users can submit their gene lists through WormBase for testing, with results output in a number of formats (see Figure 63).



Tissue Enrichment Analysis Results

Return up to 15 most significant anatomy terms.

Tissue	Expected	Observed	Enrichment Fold Change	P value	Q value
lateral ganglion WBbt:0005105	77	141	1.8	<10^-6	<10^-6
amphid sensillum WBbt:0005391	2e+02	291	1.4	<10^-6	<10^-6
AVE WBbt:0006819	68	114	1.7	<10^-6	<10^-6
thermosensory neuron WBbt:0005838	1.3e+02	187	1.5	<10^-6	<10^-6
AWB WBbt:0005671	1.3e+02	187	1.4	<10^-6	1.3e-06
ASEL WBbt:0003904	20	39	2	6.5e-06	0.00029
male WBbt:0007850	1e+02	140	1.4	1.1e-05	0.00043
ASER WBbt:0003903	20	38	1.9	1.6e-05	0.00054
PQR WBbt:0004096	8.5	20	2.3	3.7e-05	0.001
RIC WBbt:0006834	5	14	2.8	3.5e-05	0.001
BAG WBbt:0006825	73	104	1.4	5e-05	0.0012
FLP WBbt:0006828	47	72	1.5	7.8e-05	0.0017
AQR WBbt:0003927	5	12	2.4	0.00064	0.013
AIY WBbt:0005413	12	22	1.9	0.0008	0.015
ALM WBbt:0005406	42	60	1.4	0.0011	0.02
SMDDR WBbt:0004971	5	11	2.2	0.0023	0.038
PLM WBbt:0005490	42	59	1.4	0.0026	0.041
pm5 WBbt:0003737	5.8	12	2.1	0.0029	0.043
SMDDL WBbt:0004972	5.3	11	2.1	0.0041	0.057

Drag graph to your desktop to save.
Download output table here

Your list has 966 valid WormBase genes that have no annotated data or are excluded from testing :

WBGene00000005 - WBGene00000005
WBGene00000008 - WBGene00000008
WBGene00000017 - WBGene00000017
WBGene00000021 - WBGene00000021
WBGene00000025 - WBGene00000025
WBGene00000133 - WBGene00000133

Your list has 2131 valid WormBase genes included in statistical testing :

WBGene00000002 - WBGene00000002
WBGene00000003 - WBGene00000003
WBGene00000006 - WBGene00000006
WBGene00000019 - WBGene00000019
WBGene00000023 - WBGene00000023
WBGene00000024 - WBGene00000024

perform another query

Figure 62 Screenshot of results from the web GUI. After inputting a gene-list, the user is provided with the results. An HTML table is output with hyperlinks to the ontology terms. A publication-ready graph is provided below, which can be saved by dragging to the desktop. The graph is colored for better visualization; color is not intended to convey information. The graph and the table show anatomy terms in human-readable format, followed by their unique WBbt ID. Finally, lists of the genes used and discarded for the analysis are also presented.

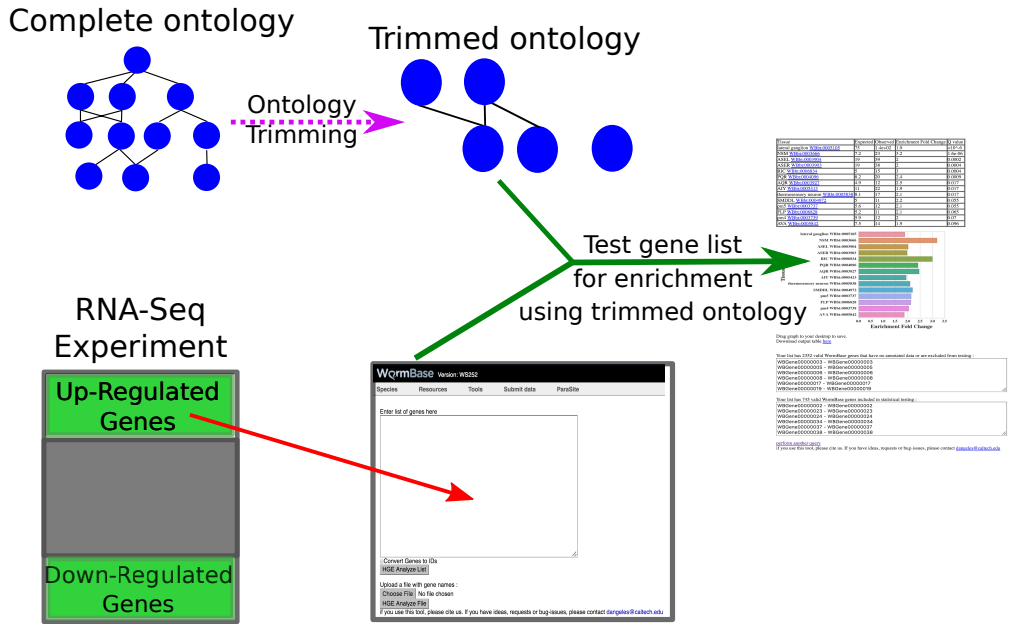


Figure 63 TEA Workflow. The complete ontology is annotated continuously by WormBase curators. After each update, the ontology is processed to remove uninformative terms, and the remaining terms are used for statistical testing. Users can select a gene list and input it into our tool using our WormBase portal. The gene list is tested for enrichment using the trimmed ontology, and results are output in tabular and graphic formats for analysis.

Validation of the algorithm and optimizing parameter selection

We wanted to select a dictionary that included enough terms to be specific beyond the most basic *C. elegans* tissues, yet would minimize the number of spurious results and which had a good dynamic range in terms of enrichment fold-change. Larger tissues are correlated with better annotation, so increasing term specificity is associated with losses in statistical power. To help us select an appropriate dictionary and validate our tool, we used a set of 30 gold standards based on microarray and RNA-seq literature which are believed to be enriched in specific tissues (Gaudet et al., 2004; Spencer et al., 2011; Cinar, Keles, and Jin, 2005; Watson et al., 2008; Pauli et al., 2006; Portman and Emmons, 2004; Fox et al., 2007; Smith et al., 2010). These data sets are annotated gene lists derived from the corresponding Expression Cluster data in WormBase. Some of these studies have been used to annotate gene

expression, and so they did not constitute an independent testing set. To correct this flaw, we built a clean dictionary that specifically excluded all annotation evidence that came from these studies.

As a first attempt to select a dictionary, we generated all possible combinations of dictionaries with minimal annotations of 10, 25, 33, 50 and 100 genes and similarity cutoffs of 0.9, 0.95 and 1, using 'avg' or 'any' similarity thresholding methods (see Table 61). The number of remaining ontology terms was inversely correlated to the minimum annotation cutoff, and was largely insensitive to the similarity threshold in the range we explored. Next, we analyzed all 30 datasets using each dictionary. Because of the large number of results, instead of analyzing each set of terms individually, we measured the average q-value for significantly enriched terms in each dataset without regard for the perceived accuracy of the terms that tested significant. We found that the similarity threshold mattered relatively little for any dictionary. We also noticed that the 'any' thresholding method resulted in tighter histograms with a mode closer to 0. For this reason, we chose the 'any' method for dictionary generation. The average q-value increased with decreasing annotation cut-off (see Figure 64), which reflects the decreasing statistical power associated with fewer annotations per term, but we remained agnostic as to how significant is the trade-off between power and term specificity. Based on these observations, we ruled out the dictionary with the 100 gene annotation cut-off: it had the fewest terms and its q-values were not low enough in our opinion to compensate for the trade-off in specificity.

To select between dictionaries generated between 50, 33 and 25 annotation cut-offs, and also to ensure the terms that are selected as enriched by our algorithm are reasonable, we looked in detail at the enrichment analysis results. Most results were comparable and expected. For some sets, all dictionaries performed well. For example, in our 'all neuron enriched sets' (Spencer et al., 2011; Watson et al.,

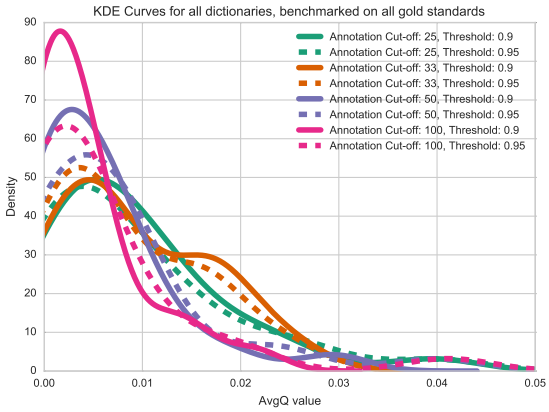


Figure 64 Kernel density estimates (KDE) for 30 gold standard datasets. We ran TEA on 30 datasets we believed to be enriched in particular tissues and pooled all the results to observe the distribution of q-values. The mode of the distribution for dictionaries with annotation cut-offs of 100 and 50 genes are very similar; however, when the cut-off is lowered to 25 genes, the mode of the distribution shifts to the left, potentially signalling a decrease in measurement power.

2008) all terms were neuron-related regardless of the dictionary used (see Table 62). On the other hand, for a set enriched for germline precursor expression in the embryo (Spencer et al., 2011), the 50 cutoff dictionary was only able to identify 'oocyte WBbt:006797', which is not a germline precursor although it is germline related; whereas the two smaller dictionaries singled out actual germline precursor cells—at the 33 cutoff, our tool identified the larval germline precursor cells 'Z2' and 'Z3' as enriched, and at the 25 gene cutoff the embryonic germline precursor terms 'P_4','P_3' and 'P_2' were identified in addition to 'Z2' and 'Z3'. We also queried an intestine precursor set (Spencer et al., 2011). Notably, this gene set yielded no enrichment when using the 25 cutoff dictionary, nor when using the 50 cutoff dictionary. However, the 33 cutoff dictionary identified the E lineage, which is the intestinal precursor lineage in *C. elegans*, as enriched. Both of these results capture specific aspects of *C. elegans* that are well known to developmental biologists.

Not all queries worked equally well. For example, a number of intestinal sets (Spencer et al., 2011; Pauli et al., 2006) were not enriched in intestine-related terms in any

dictionary, but were enriched for pharynx and hypodermis. We were surprised that intestinal gene sets performed poorly, since the intestine is a relatively well-annotated tissue.

We assessed the internal agreement of our tool by using independent gene sets that we expected to be enriched in the same tissues. We used two pan-neuronal sets (Spencer et al., 2011; Watson et al., 2008); two PVD sets (Spencer et al., 2011; Smith et al., 2010); and two GABAergic sets (Spencer et al., 2011; Cinar, Keles, and Jin, 2005). Overall, the tool has good internal agreement. On most sets, the same terms were enriched, although order was somewhat variable (see Table 65), and most high-scoring terms were preserved between sets. All comparisons can be found online in our Github repository (see Availability of data and materials). Overall, the dictionary generated by a 33 gene annotation cutoff with 0.95 redundancy threshold using the 'any' criterion performed best, with a good balance between specificity, verbosity and accuracy, so we selected this parameter set to generate our static dictionary. As of this publication, the testable dictionary contains 261 terms.

Applying the tool

We applied our tool to the RNA-seq datasets developed by Engelmann et al. (Engelmann et al., 2011) to gain further understanding of their underlying biology. Engelmann et al._ exposed young adult worms to 5 different pathogenic bacteria or fungi for 24 hours, after which mRNA was extracted from the worms for sequencing. We ran TEA on the genes Engelmann *et al* identified as up- or down-regulated. Initially we noticed that genes that are down-regulated tend to be twice as better annotated on average than genes that were up-regulated, suggesting that our understanding of the worm immune system is scarce, in spite of important advances made over the last decade. Up-regulated tissues, when detected, almost always included the hypodermis and excretory duct. Three of the five samples showed enrichment of

Table 62 Comparison of results for a GABAergic neuronal-enriched gene set from Watson (Watson et al., 2008) showing that results are similar regardless of annotation cutoff. We ran the same gene list on a dictionary with a minimum annotation cutoff of 50, similarity threshold of 0.95 and similarity method 'any' versus another with a minimum annotation cutoff of 33, similarity threshold of 0.95 and similarity method 'any'. In the table, columns are abeled with their significance value (Q-value) or enrichment fold change followed by a hyphen and a number which indicates which the cutoff for the dictionary that was used for testing. Not all tissues are present in either dictionary. Hyphens denote not-applicable values, which occurs when a particular tissue is not present in both dictionaries.

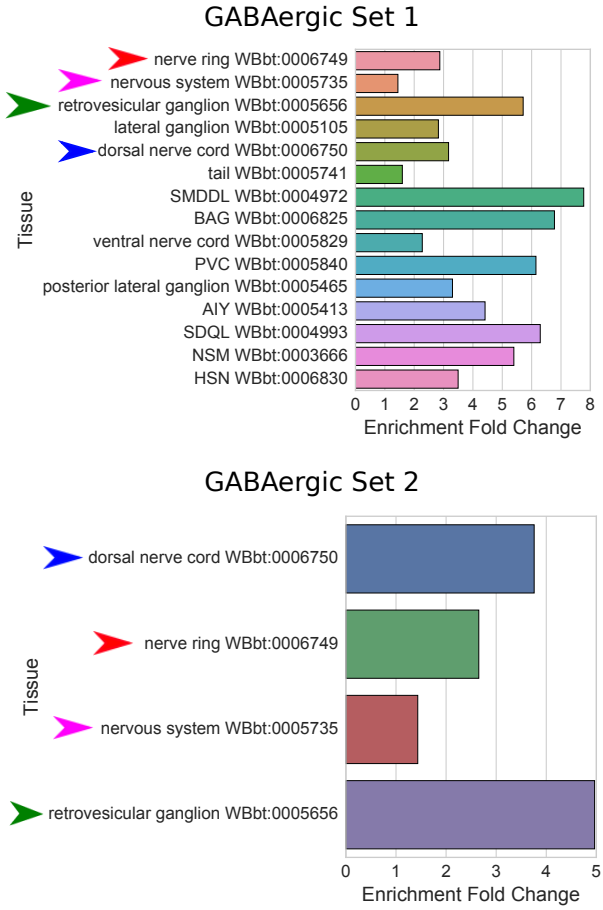


Figure 65 Independently derived gene sets show similar results when tested with the same dictionary. **Set 1.** GABAergic gene set from Watson (Watson et al., 2008). **Set 2.** GABAergic gene set from Spencer (Spencer et al., 2011). Arrowheads highlight identical terms between both analyses. All terms refer to neurons or neuronal tissues and are GABA-associated. Dictionary with cutoff: 33; threshold: 0.95; method: 'any'.

neuronal tissues or neuronal precursor tissues among the down-regulated genes. As an independent verification, we also performed GO analysis using PANTHER on the down-regulated genes for *D. coniospora*. These results also showed enrichment in terms associated with neurons (see Figure 66). A possible explanation for this neuronal association might be that the infected worms are sick and the neurons are beginning to shut down; an alternative hypothesis would be that the worm is down-regulating specific neuronal pathways as a behavioral response against the pathogen. Indeed, several studies (Meisel and D. H. Kim, 2014; Zhang, Lu, and Bargmann, 2005) have provided evidence that *C. elegans* uses chemosensory neurons to identify pathogens. Our results highlight the involvement of various *C. elegans* neuronal tissues in pathogen defense.

Discussion

We have presented a tissue enrichment analysis tool that employs a standard hypergeometric model to test the *C. elegans* tissue ontology. We use a hypergeometric function to test a user-provided gene list for enrichment of anatomical terms in *C. elegans*. Our hope is that the physiological relevance of anatomical terms will enable researchers to make hypotheses about high-dimensionality data. Specifically, we believe an enriched term may broadly suggest one of two hypotheses: if a list is enriched in a particular anatomical region, that anatomical region is affected by the experimental treatment; alternatively, the anatomical regions that are enriched reflect biologically relevant interactions between tissues. We believe the first hypothesis is a reasonable one to make in the case of whole-worm RNA-seq data for example, whereas the second hypothesis may be more plausible in cases where a researcher already knows what tissues a particular gene list came from, as may be the case in single-cell RNA-seq.

Our tool relies on an annotation dictionary that is continuously updated primarily

a) GO Enrichment Analysis

Displaying only results with P<0.05; click here to display all results

	Caenorhabditis elegans (REF)	Client Text Box Input (Δ Hierarchy NEW!)				
PANTHER GO-Slim Biological Process	#	#	expected	Fold Enrichment	+/-	P value
Unclassified	12877	745	925.08	.81	-	0.00E00
translation	362	46	26.01	1.77	+	4.19E-02
└protein metabolic process	1879	246	134.99	1.82	+	1.16E-17
└primary metabolic process	4498	603	323.14	1.87	+	6.74E-58
└metabolic process	5383	697	386.71	1.80	+	6.88E-65
sensory perception	454	59	32.62	1.81	+	3.14E-03
└neurological system process	631	116	45.33	2.56	+	4.50E-17
└system process	887	183	63.72	2.87	+	4.45E-34
└single-multicellular organism process	956	196	68.68	2.85	+	2.39E-36
└multicellular organismal process	956	196	68.68	2.85	+	2.39E-36
cellular protein modification process	904	123	64.94	1.89	+	5.36E-09
regulation of transcription from RNA polymerase II promoter	695	95	49.93	1.90	+	8.54E-07
└transcription from RNA polymerase II promoter	893	115	64.15	1.79	+	5.34E-07
└transcription, DNA-dependent	928	123	66.67	1.84	+	2.64E-08
└RNA metabolic process	1258	160	90.37	1.77	+	8.08E-10
└nucleobase-containing compound metabolic process	1936	247	139.08	1.78	+	2.25E-16
└regulation of nucleobase-containing compound metabolic process	826	116	59.34	1.95	+	3.13E-09
└regulation of biological process	1619	213	116.31	1.83	+	3.53E-15
└biological regulation	2130	326	153.02	2.13	+	6.25E-37
response to stress	370	53	26.58	1.99	+	6.24E-04

b) TEA

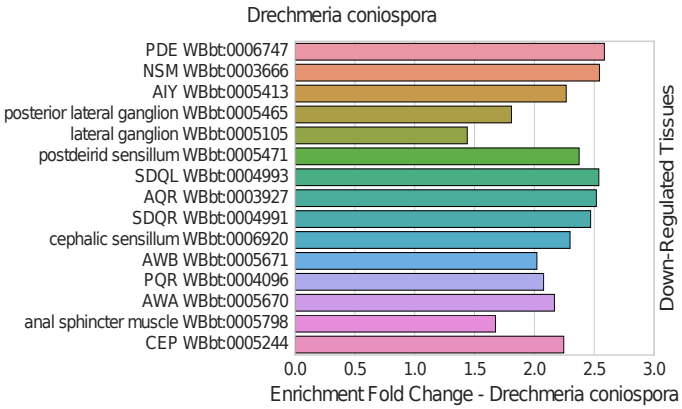


Figure 66 *D. coniospora* Gene Enrichment Analysis and Tissue Enrichment Analysis results. We compared and contrasted the results from a gene enrichment analysis program, pantherDB, with TEA by analyzing genes that were significantly down-regulated when *C. elegans* was exposed to *D. coniospora* in a previously published dataset by Engelmann *et al* (Engelmann et al., 2011) with both tools. **a.** pantherDB screenshot of results, sorted by p-value. Only top hits shown. **b.** TEA results, sorted by q-value (lowest on top) and fold-change. Both pantherDB and TEA identify terms associated with neurons (red square). The two analyses provide complementary, not redundant, information.

with data from single gene qualitative analyses, does not require retraining and does not require ranked genes. To our knowledge, this is the first tool that tests tissue enrichment in *C. elegans* via the hypergeometric method, but similar projects exist for humans and zebrafish (Y. S. Lee et al., 2013; Prykhozhij, Marsico, and Meijsing, 2013), highlighting the relevance of our tool for high-dimensionality biology. Chikina *et al* (Chikina et al., 2009) have previously reported a tissue enrichment model for *C. elegans* based on a Support Vector Machine classifier that has been trained on microarray studies. SVMs are powerful tools, but they require continuous retraining as more tissue expression data becomes available. Moreover, classifiers require that data be rank-ordered by some metric, something which is not possible for certain studies. Furthermore, this tissue enrichment tool provides users with enrichment results for only 6 large tissues. In contrast, our tool routinely tests a much larger number of terms, and we have shown it can even accurately identify enrichment of embryonic precursor lineages for select data sets.

We have also presented the first, to our knowledge, ontology term filtering algorithm applied to biomedical ontologies. This algorithm, which is very easy to execute, identifies terms that have specificity and statistical power for hypothesis testing. Due to the nature of all ontologies as hierarchical, acyclical graphs with term inheritance, term annotations are correlated along any given branch. This correlation reduces the benefits of including all terms for statistical analysis: for any given term along a branch, if that term passes significance, there is a high probability that many other terms along that branch will also pass significance. If the branch is enriched by random chance, error propagation along a branch means that many more false positives will follow. Thus, a researcher might be misled by the number of terms of correlated function and assign importance to this finding; the fact that the branching structure of GO amplifies false positive signals is a powerful argument for either reducing branch length or branch intracorrelation, or both. On the other hand, if a

term is actually enriched, we argue that there is little benefit to presenting the user with additional terms along that branch. Instead, a user will benefit most from testing sparsely along the tree at a suitable specificity for hypothesis formation. Related terms of the same level should only be tested when there is sufficient annotation to differentiate, with statistical confidence, whether one term is enriched above the other. Our algorithm reduces branch length by identifying and removing nodes that are insufficiently annotated and parents that are likely to include sparse information.

We endeavoured to benchmark our tool well, but our analysis cannot address problems related to spurious term enrichment. Although we were unable to determine false-positive and false-negative rates, we do not believe this should deter scientists from using our tool. Rather, we encourage researchers to use our tool as a guide, integrating evidence from multiple sources to inform the most likely hypotheses. As with any other tool based on statistical sampling, our analysis is most vulnerable to bias in the data set. For example, expression reports are negatively biased against germline expression because of the difficulties associated with expressing transgenes in this tissue (Kelly et al., 1997). As time passes, we are certain the accuracy and power of this tool will improve thanks to the continuing efforts of the worm research community; indeed, without the community reports of tissue expression in the first place, this tool would not be possible.

Conclusions

We have built a tissue enrichment tool that employs a tissue ontology previously developed by WormBase. We use a simple algorithm to identify the best ontology terms for statistical testing and in this way minimize multiple testing problems. Our tool is available within WormBase or can be downloaded for offline use via ‘pip install’.

Methods

Fetching annotation terms

We used WormBase-curated gene expression data, which includes annotated descriptions of spatial-temporal expression patterns of genes, to build our dictionary. Gene lists per anatomy term were extracted from a Solr document store of gene expression data from the WS252 database provided by WormBase (Howe et al., 2016). We used the Solr document store because it provided a convenient access to expression data that included inferred annotations. That is, for each anatomy term, the expression gene list includes genes that were directly annotated to the term, as well as those that were annotated to the term's descendant terms (if there were any). Descendant terms were those connected with the focus term by is_a/part_of relationship chains defined in the anatomy term ontology hierarchy.

Filtering nodes

Defining a Similarity Metric

To identify redundant sisters, we defined the following similarity metric:

$$s_i = \frac{|g_i|}{|\bigcup_{i=0}^{k} g_i|} \tag{6.2}$$

Where s_i is the similarity for a tissue i with k sisters; g_i refers to the set of tissues associated with tissue i and $|g|$ refers to the cardinality of set g. For a given set of sisters, we called them redundant if they exceeded a given similarity threshold. We envisioned two possible criteria and built different dictionaries using each one. Under a threshold criteron 'any' with parameter S between $(0, 1)$, a given set of sisters j was considered redundant if the condition

$$s_{i,j} > S \tag{6.3}$$

was true for any sister i in set j. Under a threshold criterion 'avg' with parameter S, a given set of sisters j was considered redundant if the condition

$$\mathrm{E}[s_i]_j > S \tag{6.4}$$

was true for the set of sisters j (see Figure 61).

Since nodes can have multiple parents (and therefore multiple sister sets), a complete set of similarity scores was calculated before trimming the ontology, and nodes were removed from the ontology if they exceeded the similarity threshold at least once in any comparison.

Implementation

All scripts were written in Python 3.5. Our software relies on the pandas, NumPy, Seaborn and SciPy modules to perform all statistical testing and data handling (McKinney, 2011; Van Der Walt, Colbert, and Varoquaux, 2011; Oliphant, 2007).

Availability of data and materials

Our web implementation is available at http://www.wormbase.org/tools/enrichment/tea/tea.cgi. Our software can also be downloaded using Python's pip installer via the command

```
pip install tissue_enrichment_analysis
```

Alternatively, our software is available for download at: http://dangeles.github.io/TissueEnrichmentAnalysis

All benchmark gene sets, benchmarking code and Figures can also be found at the same address, under the 'tests' folder.

Abbreviations

- TEA—Tissue Enrichment Analysis
- GO—Gene Ontology
- WBbt ID—A unique ID assigned to reference ontology terms
- WBgene ID—A unique ID assigned to reference nematode genes

Additional Files

Additional file 1 — TEA Tutorial

Tutorial for users interested in using our software within a python script

Additional file 2 — Folder Structure for SI files 3 and 4

A file detailing the folder structure of the zipped folders 3 and 4.

Additional file 3 — Golden Gene Sets

A list of all the genes used for our benchmarking process.

Additional file 4 — Results

A folder containing a complete version of the results we generated for this paper.

References

Ashburner, M et al. (2000). "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1, pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. arXiv: 10614036.

Benjamini, Yoav and Yosef Hochberg (1995). *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. DOI: 10.2307/2346101.

Chikina, Maria D. et al. (2009). "Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans". In: *PLoS Computational Biology* 5.6. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000417.

Cinar, Hulusi, Sunduz Keles, and Yishi Jin (2005). "Expression profiling of GABAergic motor neurons in Caenorhabditis elegans". In: *Current Biology* 15.4, pp. 340–346. ISSN: 09609822. DOI: 10.1016/j.cub.2005.02.025.

Engelmann, Ilka et al. (2011). "A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in C. elegans". In: *PLoS ONE* 6.5. ISSN: 19326203. DOI: 10.1371/journal.pone.0019055.

Fox, Rebecca M et al. (2007). "The embryonic muscle transcriptome of Caenorhabditis elegans". In: *Genome Biol* 8.9, R188. ISSN: 14656906. DOI: 10.1186/gb-2007-8-9-r188.

Garrido, Julián and Ignacio Requena (2012). "Towards summarizing knowledge: Brief ontologies". In: *Expert Systems with Applications* 39.3, pp. 3213–3222. ISSN: 09574174. DOI: 10.1016/j.eswa.2011.09.008.

Gaudet, Jeb et al. (2004). "Whole-genome analysis of temporal gene expression during foregut development". In: *PLoS Biology* 2.11. ISSN: 15449173. DOI: 10.1371/journal.pbio.0020352.

Howe, Kevin L et al. (2016). "WormBase 2016: expanding to enable helminth genomic research." In: *Nucleic acids research* 44.November 2015, pp. D774–D780. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1217.

Huang, Da Wei, Richard a Lempicki, and Brad T Sherman (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." In: *Nature Protocols* 4.1, pp. 44–57. ISSN: 1750-2799. DOI: 10.1038/nprot.2008.211.

Huang, Da Wei, Brad T. Sherman, et al. (2007). "DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists". In: *Nucleic Acids Research* 35.SUPPL.2. ISSN: 03051048. DOI: 10.1093/nar/gkm415.

Kelly, William G. et al. (1997). "Distinct requirements for somatic and germline expression of a generally expressed Caenorhabditis elegans gene". In: *Genetics* 146.1, pp. 227–238. ISSN: 00166731.

Kim, Jong Woo, Jordi Conesa Caralt, and Julia K. Hilliard (2007). "Pruning bio-ontologies". In: *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 1–10. ISSN: 15301605. DOI: 10.1109/HICSS.2007.455.

Lee, R. Y N and Paul W. Sternberg (2003). *Building a cell and anatomy ontology of Caenorhabditis elegans*. DOI: 10.1002/cfg.248.

Lee, Young Suk et al. (2013). "Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies". In: *Bioinformatics* 29.23, pp. 3036–3044. ISSN: 13674803. DOI: 10.1093/bioinformatics/btt529.

Love, Michael I., Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." In: *Genome biology* 15.12, p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

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

McLean, Cory Y et al. (2010). "GREAT improves functional interpretation of cis-regulatory regions." In: *Nature biotechnology* 28.5, pp. 495–501. ISSN: 1087-0156. DOI: 10.1038/nbt.1630.

Meisel, Joshua D. and Dennis H. Kim (2014). "Behavioral avoidance of pathogenic bacteria by Caenorhabditis elegans". In: *Trends in Immunology* 35.10, pp. 465–470. ISSN: 14714981. DOI: 10.1016/j.it.2014.08.008.

Mi, Huaiyu, Qing Dong, et al. (2009). "PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium". In: *Nucleic Acids Research* 38.SUPPL.1. ISSN: 03051048. DOI: 10.1093/nar/gkp1019.

Mi, Huaiyu, Anushya Muruganujan, and Paul D. Thomas (2013). "PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees". In: *Nucleic Acids Research* 41.D1. ISSN: 03051048. DOI: 10.1093/nar/gks1118.

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

Pathan, Mohashin et al. (2015). "FunRich: An open access standalone functional enrichment and interaction network analysis tool". In: *Proteomics* 15.15, pp. 2597–2601. ISSN: 16159861. DOI: 10.1002/pmic.201400515.

Pauli, Florencia et al. (2006). "Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans." In: *Development (Cambridge, England)* 133.2, pp. 287–295. ISSN: 0950-1991. DOI: 10.1242/dev.02185.

Pawitan, Yudi et al. (2005). "False discovery rate, sensitivity and sample size for microarray studies". In: *Bioinformatics* 21.13, pp. 3017–3024. ISSN: 13674803. DOI: 10.1093/bioinformatics/bti448.

Portman, Douglas S. and Scott W. Emmons (2004). "Identification of C. elegans sensory ray genes using whole-genome expression profiling". In: *Developmental Biology* 270.2, pp. 499–512. ISSN: 00121606. DOI: 10.1016/j.ydbio.2004.02.020.

Prykhozhij, Sergey V, Annalisa Marsico, and Sebastiaan H Meijsing (2013). "Zebrafish Expression Ontology of Gene Sets (ZEOGS): a tool to analyze enrichment of zebrafish anatomical terms in large gene sets." In: *Zebrafish* 10.3, pp. 303–15. ISSN: 1557-8542. DOI: 10.1089/zeb.2012.0865.

Smith, Cody J. et al. (2010). "Time-lapse imaging and cell-specific expression profiling reveal dynamic branching and molecular determinants of a multi-dendritic nociceptor in C. elegans". In: *Developmental Biology* 345.1, pp. 18–33. ISSN: 00121606. DOI: 10.1016/j.ydbio.2010.05.502.

Spencer, W. Clay et al. (2011). "A spatial and temporal map of C. elegans gene expression". In: *Genome Research* 21.2, pp. 325–341. ISSN: 10889051. DOI: 10.1101/gr.114595.110.

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

The Gene Ontology Consortium (2015). "Gene Ontology Consortium: going forward". In: *Nucleic Acids Research* 43.D1, pp. D1049–D1056. ISSN: 0305-1048. DOI: 10.1093/nar/gku1179.

Van Der Walt, Stéfan, S. Chris Colbert, and Gaël Varoquaux (2011). "The NumPy array: A structure for efficient numerical computation". In: *Computing in Science and Engineering* 13.2, pp. 22–30. ISSN: 15219615. DOI: 10.1109/MCSE.2011.37. arXiv: 1102.1523.

Watson, Joseph D et al. (2008). "Complementary RNA amplification methods enhance microarray identification of transcripts expressed in the C. elegans nervous system". In: *BMC Genomics* 9, p. 84. ISSN: 1471-2164. DOI: 10.1186/1471-2164-9-84.

Zhang, Y, H Lu, and C I Bargmann (2005). "Pathogenic bacteria induce aversive olfactory learning in Caenorhabditis elegans". In: *Nature* 438.7065, pp. 179–184. ISSN: 0028-0836. DOI: 10.1038/nature04216.

Chapter 7

TWO NEW FUNCTIONS IN THE WORMBASE ENRICHMENT SUITE

Genome-wide experiments routinely generate large amounts of data that can be hard to interpret biologically. A common approach to interpreting these results is to employ enrichment analyses of controlled languages, known as ontologies, that describe various biological parameters such as gene molecular or biological function. In *C. elegans*, three distinct ontologies, the Gene Ontology (GO), Anatomy Ontology (AO), and the Worm Phenotype Ontology (WPO) are used to annotate gene function, expression and phenotype, respectively (Ashburner et al., 2000; Lee and Sternberg, 2003; Schindelman et al., 2011). Previously, we developed software to test datasets for enrichment of anatomical terms, called the Tissue Enrichment Analysis (TEA) tool (Angeles-Albores et al., 2016). Using the same hypergeometric statistical method, we extend enrichment testing to include WPO and GO, offering a unified approach to enrichment testing in *C. elegans*. The WormBase Enrichment Suite can be accessed via a user-friendly interface at `http://www.wormbase.org/tools/enrichment/tea/tea.cgi`.

To validate the tools, we analyzed a previously published extracellular vesicle (EV)-releasing neuron (EVN) signature gene set derived from dissociated ciliated EV neurons(Wang et al., 2015) using the WormBase Enrichment Suite based on the WS262 WormBase release. TEA correctly identified the CEM, hook sensillum and IL2 neuron as enriched tissues. The top phenotype associated with the EVN signature was chemosensory behavior. Gene Ontology enrichment analysis showed that cell projection and cell body were the most enriched cellular components in this gene set, followed by the biological processes neuropeptide sig-

naling pathway and vesicle localization further down. The tutorial script used to generate the figure above can be viewed at: https://github.com/dangeles/TissueEnrichmentAnalysis/blob/master/tutorial/Tutorial.ipynb

The addition of Gene Enrichment Analysis (GEA) and Phenotype Enrichment Analysis (PEA) to WormBase marks an important step towards a unified set of analyses that can help researchers to understand genomic datasets. These enrichment analyses will allow the community to fully benefit from the data curation ongoing at WormBase.

Methods

Using the methods described in Angeles-Albores et al. (2016), we generated ontology dictionaries using the Anatomy, Phenotype and Gene Ontology annotations for *C. elegans*. The dictionary similarity parameter was set to 95for all ontologies. The annotation per term minimum was set to 33 annotations for the AO, a 50 annotations for the WPO, and 33 annotations for GO. Terms within the dictionary are tested using a hypergeometric probability test and corrected using the Benjamini-Hochberg step-up algorithm. In WS262, there are 1320 anatomy terms, 1117 phenotypes, and 3025 GO terms that have at least 11 genes annotated to them. The dictionaries are freely accessible using the Python version of the Suite, which can be installed using the pip tool for Python libraries: `pip install tissue_enrichment_analysis`. The dictionary can then be automatically downloaded by importing the enrichment analysis library in a Python script by writing `import tissue_enrichment_analysis as ea`. The dictionaries can then be downloaded by typing `ea.fetch_dictionary(dict)` into Python, where 'dict ' is one of the strings 'tissue', 'phenotype' or 'go' to specify which dictionary to download. If the function does not receive an argument, the dictionary corresponding to the AO is downloaded by default. See the tutorial above for an example

implementation.

References

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

Ashburner, M et al. (2000). "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1, pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. arXiv: 10614036.

Lee, R. Y N and Paul W. Sternberg (2003). *Building a cell and anatomy ontology of Caenorhabditis elegans*. DOI: 10.1002/cfg.248.

Schindelman, Gary et al. (2011). "Worm Phenotype Ontology: integrating phenotype data within and beyond the C. elegans community." In: *BMC bioinformatics* 12, p. 32. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-32.

Wang, Juan et al. (2015). "Cell-Specific Transcriptional Profiling of Ciliated Sensory Neurons Reveals Regulators of Behavior and Extracellular Vesicle Biogenesis". In: *Current Biology* 25.24, pp. 3232–3238. ISSN: 09609822. DOI: 10.1016/j.cub.2015.10.057.

CONCLUSION

As the phenotypes that we analyze grow more and more complex, it may be tempting to deploy methods to deal with large datasets developed in other disciplines to attempt to extract and understand the mechanics underlying our systems. These methods can be extremely effective, but only if they are fed the appropriate experimental designs and analyzed with biological principles in mind.

Biology is replete with systems that have an enormous number of variables (genes) with highly non-linear relationships between them. As a result, modeling with differential equations will always be difficult. Differential equations are brittle, with great dependence on the model specifications and parameter values, particularly when the equations have non-linearities susceptible to bifurcations. Therefore, we need methods that are parameter-free and linear. In genetics, we find both properties.

I have tried to demonstrate that transcriptomes are phenotypes, and I have tried to show examples of how these phenotypes can be rigorously and robustly manipulated to draw biologically meaningful conclusions. The work is not without flaws, and some of it may even be wrong. However, the principles are sound. Moreover, it seems apparent that these principles can be applied to many experimental settings, not just transcriptomes. It would be particularly interesting to apply these methods to two experimental tools: Metagenomics and Transposon sequencing. Metagenomics is used to survey communities and like RNA-seq can measure abundances with quantitative accuracy, but, like with RNA-seq, a major challenge has been the interpretation of the resulting datasets. I suspect that if metagenomes are inspected through single and double perturbations, the underlying community complexity will begin to unravel. Specifically, I **strongly** expect that Batesonian epistasis will play a major role in microbiomic communities, just as it does in genetic networks. I

believe that if these complex communities are inspected for examples of epistasis, we will be able to identify groups of bacteria that are functionally interconnected, and we will begin to understand the mechanisms underlying these interconnections. A second area that is of increasing interest to me is Transposon sequencing (Tn-seq). In Tn-seq experiments, a mutant library is generated through transposon insertion. The library is subjected to a defined selection process and the library is sequenced to obtain the mutant proportions after selection. Thus, Tn-seq can be used to identify genetic elements linked causally to the phenotype of interest. Tn-seq can be performed on mutant backgrounds, not just wild-type strains. So far, Tn-seq experiments have performed single-factor analyses of wild-type and mutant enrichments to identify interactors in the desired environment. It would be incredibly interesting to perform Tn-seq epistasis experiments because the phenotype used to reconstruct the interaction be the *genetic regulatory network associated with the genes under study*. At the same time, Tn-seq of multiple mutants would immediately yield the next layer of interactors or synthetic interactors in the network, naturally providing the next set of experiments to perform. I suspect this avenue could be particularly fruitful.

It is my hope that these concepts prove useful to the greater scientific community.