Analysis of allelic series with transcriptomic phenotypes

David Angeles-Albores[1] and Paul W. Sternberg[1,*]

[1] *Division of Biology and Biological Engineering, Caltech, Pasadena, CA, 91125, USA*
[*] *Corresponding author. Contact: pws@caltech.edu*

May 14, 2018

Although transcriptomes have recently been used 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 is a subunit of the Mediator complex. Our methods identify functional regions within *dpy-22* that modulate Mediator activity upon various genetic 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. Homozygous alleles can be ordered by their phenotypic severity; then, phenotypes of *trans*-heterozygotes carrying two alleles can reveal which alleles are dominant for each phenotype. Together, the severity and dominance hierarchies reveal intragenic functional regions. In *Caenorhabditis elegans*, these series have helped characterize genes such as *let-23/EGFR*, *lin-3/EGF* and *lin-12/NOTCH*[1,2,3]. Allelic series provide a powerful way to probe genes where biochemical approaches would be difficult, slow or uninformative with regards to the biological phenomenon of interest. The power of these allelic series 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*. Thus, allelic series enable geneticists to study alleles with interesting phenotypes, typically found through genetic screens. To study an 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[4] 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[5,6]. Transcriptomes have been successfully used to identify new cell or organismal states[7,8]. For mutant genes, transcriptomic states can be used for epistasis analysis[9,10], but have not been used to characterize allelic series.

We have devised methods for characterizing allelic series with RNA-seq. To test these methods, we selected three alleles[11,12] of a *C. elegans* Mediator complex subunit gene, *dpy-22*. Mediator is a macromolecular complex with $\sim$ 25 subunits[13] that globally regulates RNA polymerase II (Pol II)[14,15]. 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[16,17]. In *C. elegans* development, the CKM promotes both male tail formation[11] (through interactions with the Wnt pathway), and vulval formation[18] (through inhibition of the Ras pathway). Homozygotes of allele *dpy-22(bx93)*, which encodes a premature stop codon Q2549Amber[11], 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*

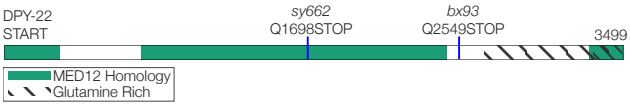

Figure 1. Protein sequence schematic for DPY-22. The positions of the nonsense mutations used are shown.

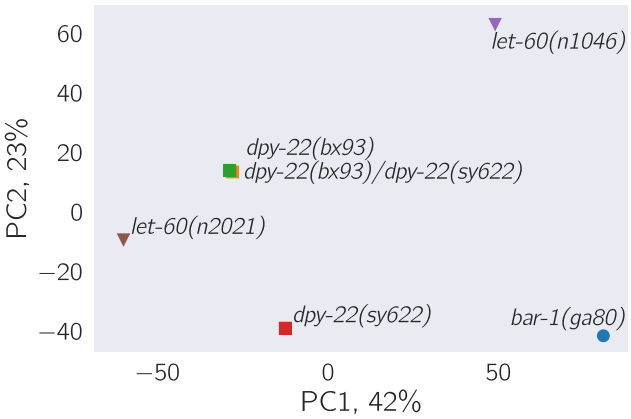

Figure 2. Principal component analysis of the analyzed genotypes. The analysis was performed using only those transcripts that were differentially expressed in at least one genotype. The plot shows that the *trans*-heterozygotes phenocopy the *dpy-22(bx93)* homozygotes along the first two principal dimensions.

background. In contrast, animals homozygous for a more severe allele, *dpy-22(sy622)* encoding another premature stop codon, Q1698Amber[12], are dumpy (Dpy), have egg-laying defects (Egl), and have multiple vulvae (Muv) (see Fig. 1). In spite of its causative role in a number of neurodevelopmental disorders[19], the structural and functional features of this gene are poorly understood. In humans, MED12 is known to have a proline-, glutamine- and leucine-rich domain that interacts with the WNT pathway[20]. However, many disease-causing variants fall outside of this domain[21]. To study these variants and how they interfere with the functionality of *MED12*, quantitative and efficient methods are necessary.

RNA-seq phenotypes have the potential to reveal functional regions within genes, but their phenotypic complexity makes this difficult. We developed a method for determining allelic series from transcriptomic phenotypes and used the *C. elegans dpy-22* gene as a test case. Our analysis revealed functional regions that act to modulate Mediator activity at thousands of genetic loci.

Results and Discussion

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 dpy-22(bx93) / + dpy-22(sy622)*. We also sequenced mRNA extracted from *bar-1(ga80)* (a Wnt ortholog), *let-60(n2021)* and *let-60(n1046gf)* (Ras ortholog) mutants in triplicate because these genes have been previously described to interact with *dpy-22* to form the vulva[12] and the male tail[11]. Sequencing was performed at a depth of 20 million reads per sample. Reads were pseudoaligned using Kallisto[22]. We performed a differential expression using a general linear model specified using Sleuth[23] (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 the false discovery rate, q, was less than or equal to 10%. We used this method to identify the differentially expressed genes associated with each mutant (see Table 1; 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 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 (see Fig. 2). On the PCA

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

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

Although the two dimensional PCA plot suggests strongly that the *dpy-22(bx93)* allele is dominant over the *dpy-22(sy622)* allele, we would like to understand the degree and nature of the dominance that is occurring. 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[10]. We use the term in the plural version, because the shared genes may represent multiple independent modules that formally constitute different phenotypic classes.

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) 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 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 the maximum number of 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. After accepting new groups into the model, we calculate the noise again, and refine the model once more. This process continues until the model converges. We called this method a false hit analysis.

We used false hit analysis to identify four non-overlapping phenotypic classes. We use the term genotype-specific to refer to groups of transcripts that were perturbed in one mutant. We use the

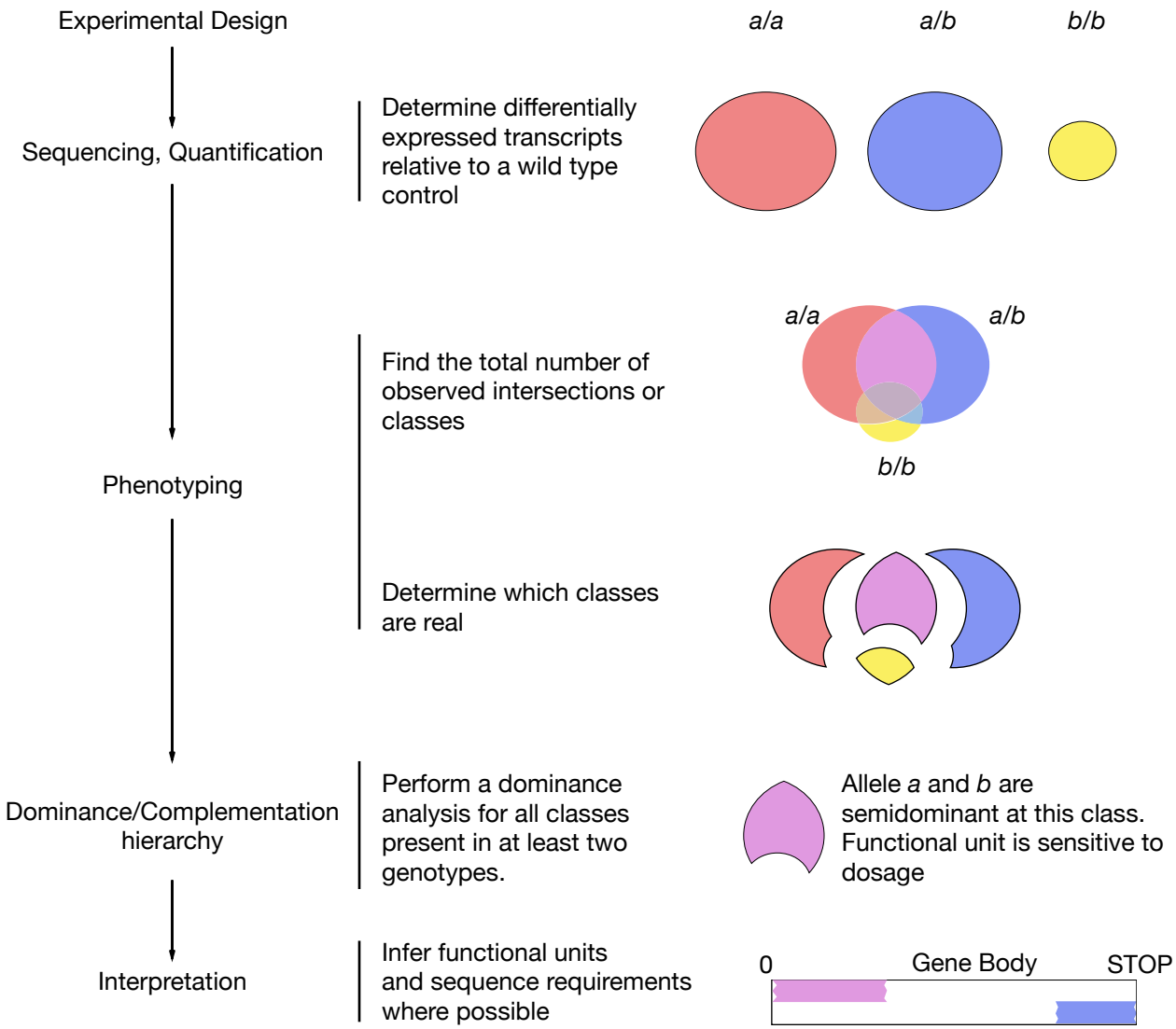

Figure 3. Flowchart for an analysis of arbitrary allelic series. A set of alleles is selected, and the corresponding genotypes are sequenced. Independent phenotypic classes are then identified. For each phenotypic class, the alleles are ordered in a dominance/complementation hierarchy, which can then be used to infer functional regions within the genes in question.

term genotype-associated to refer to those groups of transcripts whose expression was significantly altered in two or more mutants with respect to the wild type control. The ***dpy-22(sy622)*-associated** phenotypic class consisted of 720 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 403 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 (1,841 genes) and a ***trans*-heterozygote-specific** phenotypic class (1,226 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 affects 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 volume**[1]. An important caveat is that volume 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.

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. To assess which allele has more volume, Since the homozygotes of each allele only share a single phenotypic class in common, we need only assess volume along this single phenotype. To calculate a volume 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. 4) 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 volume that is 60% 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. Moreover, within their shared phenotype, the *dpy-22(bx93)* allele encodes functionality that is more similar to wild-type than the functionality encoded in the *dpy-22(sy622)* allele.

[1]We chose the term volume by analogy with a radio. Each allele can affect a different number of phenotypes (channels) with a different perturbation magnitude per channel (volume)

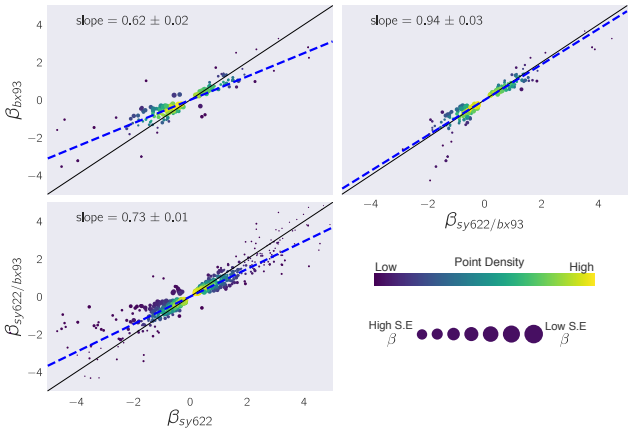

Figure 4. Shared Transcriptomic Phenotypes amongst the *dpy-22* genotypes are regulated in the same direction. For each pairwise comparison, we found those transcripts that were commonly differentially expressed in both genotypes relative to the wild-type control and plotted the β coefficients for each. 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 volume of the *dpy-22(sy622)* allele is greater than the volume of the *dpy-22(bx93)* allele. The other comparisons are shown for completeness.

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*-

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 2. 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)*.

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 other 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 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 2) for the *dpy-22(bx93)*-associated phenotypic class.

Phenotypic classes reflect morphological phenotypes

Having identified a set of phenotypic classes, we wanted to know whether enrichment analysis of anatomical, phenotypic or gene ontology terms revealed relevant physiological aspects underlying our data. To this end, we used the WormBase Enrichment Suite[24] to perform enrichment analysis of each group.

We found that the *dpy-22(bx93)*-associated phenotypic class was enriched in genes involved in 'immune system processes' ($q < 10^{-5}$), and was enriched in genes that are expressed in the 'intestine' ($q < 10^{-4}$). The *dpy-22(sy622)*-associated class on the other hand was enriched in genes expressed in the 'cephalic sheath cell' ($q < 10^{-4}$). Using the Gene Ontology Enrichment Analysis, 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' ($q < 10^{-12}$), namely 'collagen trimers' ($q < 10^{-12}$) 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 involved 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*[12,11] 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)*-specific phenotypic class would exhibit the most significant overlap with *bar-1(ga80)* mutants.

There was no significant overlap between any of the mutants we tested and the genes in the *dpy-22(bx93)*-specific class, consistent with our interpretation of this class as the result of random false positives (see Fig. 5). All other classes showed significant enrichment with genes perturbed in *bar-1(ga80)*. Similarly, *let-60(n2021)* showed en-

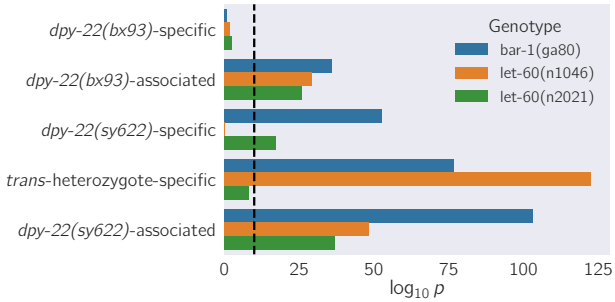

Figure 5. *dpy-22* phenotypic classes are statistically significantly enriched for signatures of *let-60* (ras) and *bar-1* (wnt) signaling. We tested whether the overlap between the differentially expressed genes in *bar-1(ga80)*, *let-60(n1046gf)* or *let-60(n20201)* and the *dpy-22* phenotypic classes 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).

richment 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

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* (see Fig. 6). 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. 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 0 and 2,549. 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. 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. 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. 6). Such a model would propose that the sequences necessary for FR3 functionality are within the interval 0 and 1,698, 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 FR2, and FR4 modifies its activity 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[25]. In the case of *dpy-22/MDT12* orthologs, how either mechanism could operate is not obvious, since DPY-22 is expected to assemble in a monomeric manner into the CKM. Massive single-cell RNA-seq of *C. elegans* has recently been reported[26]. When this technique becomes cost-efficient, single-cell profiling of these genotypes may provide information that complements the whole-organism expression phenotypes, perhaps explaining the origin of this phenotype.

Phenotypes, not molecular pathways

In an attempt to better dissect interactions between *dpy-22* and *let-60* (ras) and *bar-1* (wnt), we obtained expression profiles of these mutants and looked for enrichment within each phenotypic class. Prior research suggested that the *bx93*-associated class should show STPs with *bar-1* mutant transcriptomes but not with

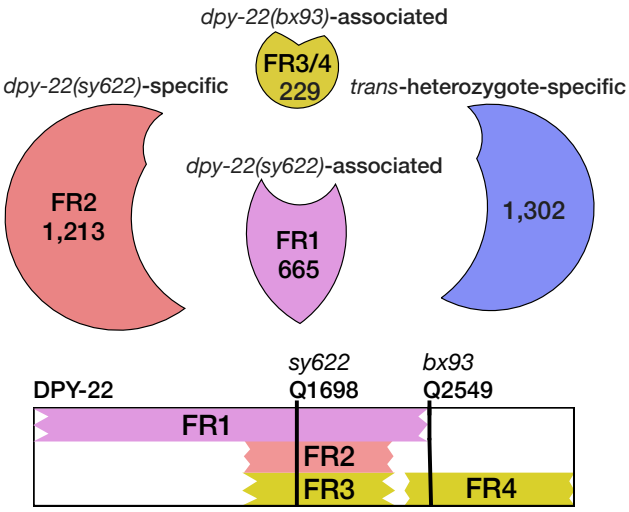

Figure 6. The functional regions associated with each phenotypic class can be mapped intragenically. The number of genes associated with each class is shown. The *dpy-22(bx93)*-associated class may be controlled by two functional regions. FR2 and FR3 could be redundant if FR4 is a modifier of FR2 functionality 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).

let-60 mutant transcriptomes, and the *sy622*-specific class should show STPs with gain-of-function *let-60* mutants but not with *bar-1* null mutants. Instead, we found that *let-60* and *bar-1* loss-of-function mutants had STPs with all the phenotypic classes; and *let-60* gain-of-function mutants did not show STPs with the *sy622*-associated class. Thus, although we could predict that both of these genes are interactors with *dpy-22*, our experiments did not enable us to predict what functional regions mediate this interaction.

Our failure to predict what functional regions mediate interactions with other pathways or to predict the valence of the interactions reflects the fact that we are treating expression profiles as phenotypes, and not as a method to read out the activity of molecular pathways. Expression profiles as phenotypes have advantages, namely that they can be used for genetic analyses, but they also come with disadvantages. In our case, we sequenced animals after development had occurred, so the expression levels we observe are the average expression levels of all tissues after they have finished forming. Moreover, as with any other phenotype, we cannot discriminate between proximal or direct effects on gene expression from second order or ripple effects. As with morphological phenotypes, it may turn out to be the case that expression profiles, like any other phenotype, will often reflect a limited number of phenotypic classes. In some cases, these phenotypic classes may be associated with a molecular pathway by the use of index sets generated from mutants in these pathways. In other cases, they may reflect stresses and insults that the animals are suffering as a result of suboptimal activity. The difference between these two classes matters, since the conclusions that can be drawn from studying each one vary in their scopes; however, it seems reasonable likely that both will prove useful.

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 badly misleading. Such artifacts are particularly likely for small phenotypic classes, which should be viewed with skepticism. 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 plausibil-

ity of each class. 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.

Methods

Strains used

Strains used were N2 wild-type (Bristol)[27], PS4087 *dpy-22(sy622)*[12], PS4187 *dpy-22(bx93)*[11], PS4176 *dpy-6(e14) dpy-22(bx93)/ + dpy-22(sy622)*[12], MT4866 *let-60(n2021)*[28], MT2124 *let-60(n1046gf)*[28] and EW15 *bar-1(ga80)*[29]. Lines were grown on standard nematode growth media (NGM) Petri plates seeded with OP50 *E. coli* at 20°C[27].

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 8–12 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 *et al*[10,7]. Briefly, young adults were placed in 10μL 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[22], 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[30,31,32,33].

Differential expression analysis was performed using Sleuth[23]. 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[10,7] and unpublished analyses adjusting for batch effects. Briefly, batches were assigned based on the a covariate that represented the combination of the person who collected the worms, the person who extracted the RNA, the month in which the samples were sequenced and the library preparation method.

False hit analysis

To accurately count phenotypes, we developed a false hit algorithm (Algorithm 1). We implemented this algorithm for three-way comparisons in Python. Although experimentally restricted, a three-way comparison can result in 128 possible sets (ignoring size). This large number of models necessitates an algorithmic approach that can at least restrict the possible 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 rapidly if false-positive or negative rates exceed 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 cut-

off, 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. When the false negative rate was set at 15% or higher, the algorithm converged on the same five classes shown above. For false negative rates between 10–15%, the algorithm output the same five classes, but also accepted the (*dpy-22(sy622),dpy-22(bx93)*)-associated class. We selected the model corresponding to false negative rates of 15–20% because this model had lower χ^2 values than the model selected with a false negative rate of 10–15% (4,212 versus 100,650).

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 we compared χ^2 statistics between a re-classified reduced model and a reduced model. The re-classified model had a lower χ^2 (181). Thus, we concluded that the re-classified reduced model is the most likely model to give rise to our data.

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 set to initialize the reduced model
  $\mathbf{K} = \{\min_{l \in L} N_l\}$
  Refine the model until the model converges or iterations max out
  $i \leftarrow 0$
  $\mathbf{K_{prev}} \leftarrow \emptyset$
  while $(i < i_{\max}) \mid (\mathbf{K_{prev}} \neq \mathbf{K})$ **do**
    $\mathbf{K_{prev}} \leftarrow \mathbf{K}$
    Define a noise function to estimate error flows in $\mathbf{K}$
    $\mathbf{F} \leftarrow \text{noise}(\mathbf{K}, f_p, f_n)$
    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 $(\lambda > \alpha) \mid (\lambda < 0)$ **then**
        $\mathbf{K}_l \leftarrow \mathbf{M}_{obs,l}$
      end
    end
    $i++$
  end
end
Return the reduced model
$\mathbf{M}_{reduced} = \mathbf{K}$
return $\mathbf{M}_{reduced}$

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.

Dominance analysis

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

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

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:

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

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[34] using the Python programming language. The Numpy, pandas and scipy libraries were used for computation[35,36,37] and the matplotlib and seaborn libraries were used for data visualization[38,39]. Enrichment analyses were performed using the WormBase Enrichment Suite[24,40]. 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)[41] under the accession code GSE107523 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107523).

Acknowledgements

This work was supported by HHMI with whom PWS was an investigator, by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology, and by the NIH grant U41 HG002223. This article would not be possible without help from Dr. Igor Antoshechkin and Dr. Vijaya Kumar who performed the library preparation and sequencing. Han Wang, Hillel Schwartz, Erich Schwarz, Porfirio Quintero and Carmie Puckett Robinson provided valuable input throughout the project.

References

1. Aroian, R. V. & Sternberg, P. W. Multiple functions of let-23, a *Caenorhabditis elegans* receptor tyrosine kinase gene required for vulval induction. *Genetics* **128**, 251–67 (1991).

2. Ferguson, E. & Horvitz, H. R. Identification and characterization of 22 genes that affect the vulval cell lineages of *Caenorhabditis elegans*. *Genetics* **110**, 17–72 (1985).

3. Greenwald, I. S., Sternberg, P. W. & Robert Horvitz, H. The lin-12 locus specifies cell fates in *Caenorhabditis elegans*. *Cell* **34**, 435–444 (1983).

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

5. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).

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

7. Angeles-Albores, D. *et al.* The *Caenorhabditis elegans* Female State: Decoupling the Transcriptomic Effects of Aging and Sperm-Status. *G3: Genes, Genomes, Genetics* (2017).

8. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356** (2017).

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

10. Angeles-Albores, D., Puckett Robinson, C., Williams, B. A., Wold, B. J. & Sternberg, P. W. Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E2930–E2939 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29531064http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5879656.

11. Zhang, H. & Emmons, S. W. A *C. elegans* mediator protein confers regulatory selectivity on lineage-specific expression of a transcription factor gene. *Genes and Development* **14**, 2161–2172 (2000).

12. Moghal, N. A component of the transcriptional mediator complex inhibits RAS-dependent vulval fate specification in *C. elegans* . *Development* **130**, 57–69 (2003).

13. Jeronimo, C. & Robert, F. The Mediator Complex: At the Nexus of RNA Polymerase II Transcription (2017).

14. Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nature reviews. Molecular cell biology* **16**, 155–166 (2015).

15. Takagi, Y. & Kornberg, R. D. Mediator as a general transcription factor. *The Journal of biological chemistry* **281**, 80–9 (2006).

16. Knuesel, M. T., Meyer, K. D., Bernecky, C. & Taatjes, D. J. The human CDK8 subcomplex is a molecular switch that controls Mediator coactivator function. *Genes & development* **23**, 439–51 (2009).

17. Elmlund, H. *et al.* The cyclin-dependent kinase 8 module sterically blocks Mediator interactions with RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 15788–93 (2006).

18. Moghal, N. & Sternberg, P. W. A component of the transcriptional mediator complex inhibits RAS-dependent vulval fate specification in *C. elegans*. *Development* **130**, 57–69 (2003).

19. Graham, J. M. & Schwartz, C. E. MED12 related disorders. *American Journal of Medical Genetics, Part A* **161**, 2734–2740 (2013).

20. Kim, S., Xu, X., Hecht, A. & Boyer, T. G. Mediator is a transducer of Wnt/β-catenin signaling. *Journal of Biological Chemistry* **281**, 14066–14075 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16565090.

21. Yamamoto, T. & Shimojima, K. A novel MED12 mutation associated with non-specific X-linked intellectual disability. *Human Genome Variation* **2**, 15018 (2015).

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

23. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *brief communications nature methods* **14** (2017).

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

25. Yook, K. Complementation. *WormBook* (2005).

26. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (New York, N.Y.)* **357**, 661–667 (2017).

27. Brenner, S. The Genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).

28. Beitel, G. J., Clark, S. G. & Horvitz, H. R. *Caenorhabditis elegans* ras gene *let-60* acts as a switch in the pathway of vulval induction. *Nature* **348**, 503–509 (1990).

29. Eisenmann, D. M., Maloof, J. N., Simske, J. S., Kenyon, C. & Kim, S. K. The beta-catenin homolog BAR-1 and LET-60 Ras coordinately regulate the Hox gene lin-39 during Caenorhabditis elegans vulval development. *Development (Cambridge, England)* **125**, 3667–3680 (1998).

30. Andrews, S. FastQC: A quality control tool for high throughput sequence data (2010).

31. Deluca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).

32. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Bowtie: An ultrafast memory-efficient short read aligner. *Genome biology* R25 (2009).

33. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

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

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

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

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

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

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

40. Angeles-Albores, D., Lee, R. Y., Chan, J. & Sternberg, P. W. Two new functions in the WormBase Enrichment Suite. *Micropublication: biology. Dataset.* (2018).

41. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207–10 (2002).