A study of allelic series using transcriptomic phenotypes in a metazoan

David Angeles-Albores[1,2] and Paul W. Sternberg[1,2,*]

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

October 28, 2017

Expression profiling holds great promise for genetics because of its ability to measure thousands of genes quantitatively in parallel. Although transcriptomes have recently been used to perform epistasis analyses for pathway reconstruction, there has not been a systematic effort to understand how expression profiles will vary among various mutants of the same gene. Here, we study an allelic series in *C. elegans* consisting of one wild type and two mutant alleles of *mdt-12*, a highly pleiotropic gene whose gene product is a subunit of Mediator complex, which is essential for transcriptional initiation in eukaryotes. We developed a false hit analysis to identify which populations of genes commonly differentially expressed with respect to the wild type are likely the result of statistical artifact. We concluded that expression perturbations caused by these alleles split into four distinct modules called phenotypic classes. To understand the dominance relationship between the two mutant alleles, we developed a dominance analysis for transcriptional data. Dominance analysis of these phenotypic classes support a model where *mdt-12* has multiple functional units that function independently to target the Mediator complex to specific genetic loci.

Author Summary

Expression profiling is a way to quickly and quantitatively measure the expression level of every gene in an organism. As a result, these profiles could be used as phenotypes with which to perform genetic analyses (i.e., to figure out what genes interact with each other) as well as to dissect the molecular properties of each gene. Before we can perform these analyses, we have to figure out the rules that apply to these measurements. In this paper, we develop new concepts and methods with which to study an allelic series. Briefly, allelic series are an important aspect of genetics because different alleles encode different versions of a gene. By studying these different versions, we can make statements about how function is encoded within the sequence of a gene. We apply our methods to the *mdt-12* gene, which encodes a subunit of the Mediator complex. Though we know it is essential for all transcriptional activity in eukaryotes, we understand very little about how the Mediator complex functions to generate both general and specific phenotypes. The reason for this is the genes that encode these subunits are associated with general sickness and multiple phenotypes when mutated, which makes them challenging to study genetically. We show that transcriptomic phenotypes renders the study of general factors such as *mdt-12* feasible.

Supplementary Data

The website for the Supplementary Data for this project is still under construction and will be available shortly. All code, data and figures are available upon request.

Introduction

The term 'allelic series' refers to the study of alleles with different phenotypes to understand the molecular properties that this locus controls. Allelic series are historically important for genetics[1]. In early pioneering work, McClintock studied a deficiency of the tail end of chromosome 9 of maize by generating *trans*-heterozygotes with mutants of various genes that she knew existed near the end of chromosome 9. Her work allowed her to infer that the deficiency was modular, effectively generating a double mutant that behaved as a single allele but which could participate phenotypically in two distinct allelic series. From this study, McClintock inferred that deletions could span multiple genes, which behaved as independent modules, and which were identified via complementation assays. This work set the foundations for later observations in yeast that showed two mutant alleles of the same genetic unit, when placed in *trans* to each other, could complement and generate a wild-type phenotype[2]. Allelic series have also been used to study the dose response curve of a phenotype for a particular gene and to infer null phenotypes from hypomorphs. In *C. elegans*, the *let-23*, *lin-3* and *lin-12* allelic series stand out as examples[3,4,5].

Over the last decade, biology has moved from expression measurements of single genes towards genome-wide measurements. Expression profiling via RNA-sequencing[6] (RNA-seq) is a popular method because it enables the simultaneous measurement of transcript levels for all genes in a genome. These measurements can now be made on a whole-organism scale and on single cells[7]. Although initially expression profiles had a qualitative purpose as descriptive methods to identify genes that are downstream of a perturbation, these profiles are now being used as phenotypes for genetic analysis. As a result, transcriptomes have been successfully used to identify new cell or organismal states[8,9]. Genetic pathways have been reconstructed via sequencing cDNA from single cells[10] or by sequencing transcripts from whole-organisms[11]. However, to fully characterize a genetic pathway, it is often necessary to build allelic series to explore whether independent functional units within a gene mediate different aspects of the phenotypes associated with a pathway or gene, or whether the phenotypes are simply the result of gene dosage.

As a proof of principle, we selected a subunit of the Mediator complex in *C. elegans*, *mdt-12* (previously known as *dpy-22*[12]), for genetic analysis. We explored three alleles, including the wild-type allele, of this highly pleiotropic gene because its biological roles are poorly understood. The mutant alleles were generated in previous screens[13,14], where they were associated with specific phenotypes in the male tail and in the vulva. Mediator is a macromolecular complex that contains approximately 25 subunits[15] and which globally regulates RNA polymerase II (Pol II)[16,17]. Mediator is a versatile regulator, a quality often associated with its variable subunit composition[16], and it can promote transcription as well as inhibit it. The Mediator complex consists of four modules: the Head, Middle and Tail modules and a CDK-8-associated Kinase Module (CKM). The CKM can associate reversibly with Mediator. Certain models propose that the CKM functions as a molecular switch, which inhibits Pol II activity by sterically preventing its interaction with the other Mediator modules[18,19]. Other models propose that the CKM negatively modulates interactions between Mediator and enhancers[20]. In *C. elegans*, the CKM consists of CDK-8, MDT-13, CIC-1 and DPY-22[21]. Since *dpy-22* is orthologous to the human Mediator subunits *MED-12* and *MED-12L*[13], we will henceforth refer to this gene as *mdt-12*. *mdt-12* has been studied in the context of the male tail[13], where it was found to interact with the Wnt pathway. It has also been studied in the context of vulval formation[22], where it was found to be an inhibitor of the Ras pathway. Loss of *mdt-12* is lethal in XO animals[23,24], and developmental studies have relied on reduction-of-function alleles to understand the role of this gene in development. Studies of the male tail were carried out using an allele, *dpy-22(bx93)*, that generates a truncated DPY-22 protein missing its C-terminal 949 amino acids as a result of a premature stop codon, Q2549STOP[13]. In spite of the premature truncation, animals carrying this allele grossly appear phenotypically wild-type. In contrast, the allele used to study the role of *mdt-12* in the vulva, *dpy-22(sy622)*, is a premature stop codon, Q1698STOP, that predicted to remove 1,800 amino acids from the C-terminus[14] (see Fig. 1). Animals carrying this mutation are severely dumpy (Dpy), have egg-laying defects (Egl) and have a low penetrance multivulva (Muv) phenotype. These alleles could form a single quantitative series, affecting the same sets of target genes but to different degrees, in which case the *trans*-heterozygote would exhibit a single dosage-dependent phenotype intermediate to the two homozygotes. Alternatively, they could form a single qualitative series, in which case the *trans*-heterozygote should have the same phenotype as the homozygote of the *bx93* allele, since this allele encodes the longer protein. These alleles could also form a mixed series, in which case multiple separable phenotypes would appear that have qualitative or

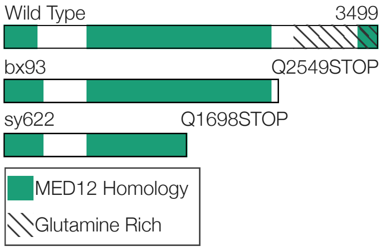

Figure 1. The *mdt-12* allelic series, consisting of two amino acid truncations. Diagram of the MDT-12 wild-type protein and the protein product of *bx93* and *sy622* alleles.

quantitative behaviors in the *trans*-heterozygote.

Expression profiles have the potential to facilitate dissection of molecular structures within genes. For the *mdt-12* allelic series, we found that the perturbations caused by the weak loss-of-function allele, *bx93*, are entirely contained within the perturbations caused by the strong loss-of-function allele, *sy622*. Further, we found three phenotypic classes affected by *mdt-12*. For one class, termed the *sy622*-specific class, the *bx93* homozygote, but not the *sy622* homozygote, shows wild-type functionality. In a *trans*-heterozygote of *sy622/bx93* these perturbations are suppressed to wild-type levels from the *sy622* levels, which shows that *bx93* is wild-type dominant for this phenotype. A second class, called the *sy622*-associated class, similarly shows wild-type functionality in the *bx93* homozygote but not in the *sy622* homozygote, yet in the *trans*-heterozygote these perturbations are modulated in a gene-dosage dependent manner. Finally, we identified a third class, called the *bx93*-specific class, which contained genes that were altered in both homozygotes, but which showed an expression level most similar to the *bx93* homozygote, showing that *bx93* has a dominant mutant phenotype for this subset. For each class, we were able to quantitatively measure the dominance level of each allele.

Results

Strong and weak loss-of-function alleles of *mdt-12* show different transcriptomic profiles

We sequenced in triplicate cDNA synthesized from mRNA extracted from *sy622* homozygotes, *bx93* homozygotes, *trans*-heterozygotes of both alleles and wild-type controls at a depth of 20 million reads per replicate. This allowed us to quantify expression levels of 21,954 protein-coding isoforms. We calculated differential expression with respect to a wild-type control using a general linear model (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. Positive β coefficients indicate up-regulation with respect to the wild-type, whereas negative coefficients indicate down-regulation. Transcripts were tested for differential expression using a Wald test, and the resulting p-values were transformed into q-values that are correcteed for multiple hypothesis testing. Transcripts were considered to have differential expression between wild-type and a mutant if the associated q-value of the β coefficient was less than 0.1. At this threshold, 10% of all differentially expressed genes are expected to be false positive hits.

Using these definitions, we found 481 differentially expressed genes in the *bx93* homozygote transcriptome, and 2,863 differentially expressed genes in the *sy622* homozygote transcriptome (see Fig. 2).

Transcriptome profiling of *mdt-12* *trans*-heterozygotes

We also sequenced *trans*-heterozygotic animals with genotype *dpy-6(e14) bx93/+ sy622*. This *trans*-heterozygote appears phenotypically wild-type, resembling the *bx93* mutant morphologically[14]. The *trans*-heterozygote transcriptome had 2,214 differentially expressed genes.

False hit analysis identifies four phenotypic classes

Overlapping three sets of differentially expressed genes from different genotypes can generate at most seven categories. Each of these seven categories could be interpreted biologically if the population is believed to arise from real effects. If these populations are small, however, there is a real chance that they represent statistical noise, and are not biologically meaningful. If that is the case, these populations may consist largely of genes that are mis-classified and belong to a different cluster, in which case they should be re-classified into the most likely cluster, if it can be determined.

We identified three categories of genes that were most likely to be influenced by statistical noise due to their small size. These populations were those that encompassed genes differentially expressed in *bx93* homozygotes and one other genotype, as well as genes that were differentially expressed specifically in *bx93* homozygotes.

These three categories stand out as candidates for statistical noise not just because of their small size, but also because of the extraordinary biological interpretations required to make sense of them. For example, if there truly is a population of genes that is only perturbed in homozygotes of either allele but not in the *trans*-heterozygote, then this means that the two alleles are somehow intragenically complementing to produce wild-type function. Given the molecular nature of the mutations, this interpretation is unlikely to be correct.

To perform a false hit analysis, we imagined an idealized scenario where the perturbations in *bx93* homozygotes were present in all thre genotypes. We also imagined that in this scenario the *trans*-heterozygote did not exhibit any perturbations not present the *sy622* homozygote. In this simplified scenario, we could model where false positive and false negative hits were most likely to fall (see Fig. 2). Next, we present the results of our hit analysis for eachperturbation category.

We identified 78 genes that are differentially expressed exclusively in *bx93* homozygotes. At a false positive rate of 10% (our defined cut-off) we expect 48 genes to be falsely called as differentially expressed in *bx93* homozygotes. The probability that such a false positive is also differentially expressed in another genotype is 20% (4,392 transcripts identified between the two other genotypes divided by 21,954 the total number of transcripts that were successfully sequenced). Thus, on average we expect 39 false positive hits to be classified into the *bx93*-specific class. On average, half of all genes in the *bx93*-specific class would be expected to be the result of statistical artifacts. Statistical noise is therefore a major contributor towards the existence of this class. Since the biological interpretation of this class is unclear and requiring extraordinary evidence, we find the most parsimonious explanation to be that the *bx93*-specific class does not exist.

We estimated that statistical noise could account for > 80% of the genes that were differentially expressed in both *bx93* and *sy622* homozygotes and not differentially expressed in the *trans*-heterozygote. Further, we estimated that statistical artifacts could explain > 80% of the transcripts that were differentially expressed in the *trans*-heterozygote and *bx93* homozygotes but not in *sy622* homozygotes. For both of these populations, we estimate that the majority of the false hits emerge from false negative results. In other words, most of the noise in these populations is the result of mis-classification. Finally, the biological interpretation of either population is implausible given the molecular nature of the alle-

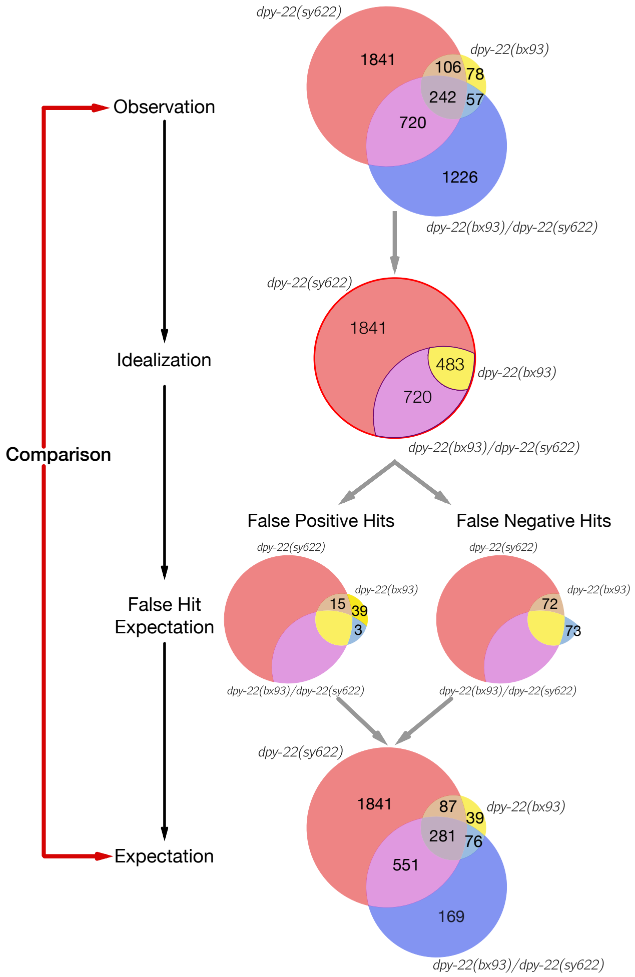

Figure 2. False hit analysis. To assess the extent to which statistical artifacts could affect the interpretation of certain intersections, we first idealized the Venn diagram and asked whether false positive and false negative results could distort the diagram back to its original shape. We estimated the false negative rate at 15% and used a false positive rate of 10%. For simplicity, only false hit analysis for *bx93* groups is shown. False hits can explain the existence of a groups of genes that are differentially expressed in *bx93* homozygotes only, in *bx93* homozygotes and *trans*-heterozygotes, and in *bx93* homozygotes and *sy622* homozygotes. Genes that are solely expressed in *bx93* homozygotes are unlikely to exist, whereas genes that are differentially expressed in *bx93* homozygotes and one other genotype are probably misclassified and should be differentially expressed in all genotypes. The *trans*-heterozygote specific class cannot be explained by statistical artifacts.

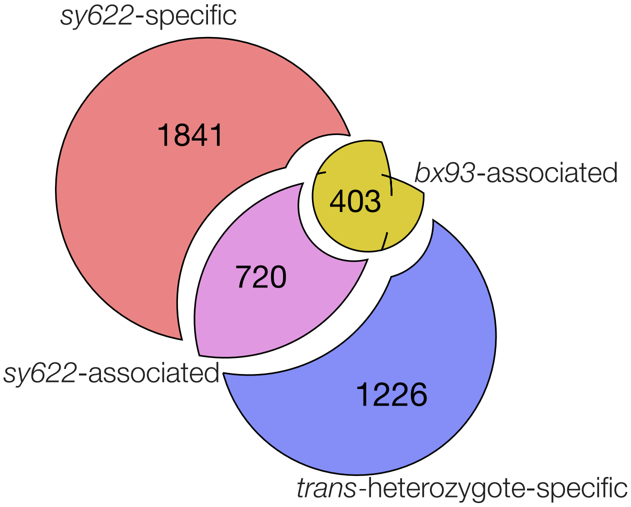

Figure 3. Transcripts under the control of *mdt-12* belong to distinct phenotypic classes. Exploded Venn diagram highlighting the four identified phenotypic classes.

les. Taken together, a false hit analysis of these two categories strongly suggests that they contain genes that have been mis-classified and which most likely are differentially expressed in all three genotypes.

A false hit analysis identified four non-overlapping phenotypic classes (see Fig. 3). We use the term allele- or genotype-specific to refer to groups of transcripts that are solely perturbed in a single genotype. On the other hand, we use the term allele-associated to refer to those groups of transcripts that are perturbed in at least two genotypes. We identified a *sy622*-associated phenotypic class, which consisted of 720 genes differentially expressed in *sy622* homozygotes and in *trans*-heterozygotes, but which were not differentially expressed in *bx93* homozygotes. We also identified a *bx93*-associated phenotypic class. Following the argument of the previous paragraph, this class included all genes that were differentially expressed in *bx93* homozygotes and at least one other genotype, since it is likely that of these genes should actually be differentially expressed in all genotypes. As a result, this class contains 403 genes. We also identified a *sy622*-specific phenotypic class (1,841 genes) and a *trans*-heterozygote-specific phenotypic class (1,226 genes). Having identified these phenotypic classes, we set out to confirm whether each class actually behaved as an independent phenotypic module in an allelic series and whether each class could be interpreted biologically to shed light on the functions of *mdt-12*.

Different phenotypic classes behave differently in an *sy622* homozygote

We asked whether these classes had perturbation distributions distinct from each other within a single homozygote. Specifically, we wanted to test whether these sets behaved as randomly selected sets. If this were the case, then within a single genotype, each class would be expected to have the same distribution of perturbations (see Fig. 4). We found that that the β coefficients of isoforms within the *bx93*-associated phenotype on average had the largest absolute value (mean: 1.2). The *sy622*-associated phenotype had a smaller range of perturbations compared to the *bx93*-associated phenotype (95th percentiles of the two distributions: 2.9 versus 3.2, respectively), and a statistically smaller median (0.91 vs 1.2, respectively, $p < 10^{-6}$, non-parametric boostrap). The medians of the *sy622*-specific and -associated classes were the same ($p = 0.15$). There are systematic differences between the behaviors of each class. This rejects the null hypothesis that the transcripts in each class were randomly selected.

Dominance can be quantified in transcriptomic phenotypes

Dominance relationships between alleles are phenotype-specific. In other words, an allele can be dominant over another for one phenotype, yet not for others. An example is the *let-23* allelic series—nulls of *let-23* are recessive lethal (Let) and presumably also recessive vulvaless (Vul) relative to the wild-type allele. The *sy1* allele of *let-23* is dominant viable relative to null alleles, but is recessive Vul[3] to the wild-type allele. Above, we postulated that there are four phenotypic classes, three of which are composed of genes whose expression is significantly perturbed in the *sy622* homozygote. If these classes are indeed modular phenotypes, then the dominance relationships within each class should be the same from gene to gene. In other words, a single dominance coefficient should be sufficient to explain the gene expression in the *trans*-heterozygote for every gene within a class.

To quantify this dominance, we implemented and maximized a Bayesian model (see Methods). Briefly, we asked what the linear combination of β coefficients from each homozygote would best predict the observed β values of the heterozygote, subject to the constraint that the coefficients added up to 1 (see Dominance analysis). We reasoned that if this was a modular phenotype controlled by a single functional unit encoded within the gene of interest, then a plot

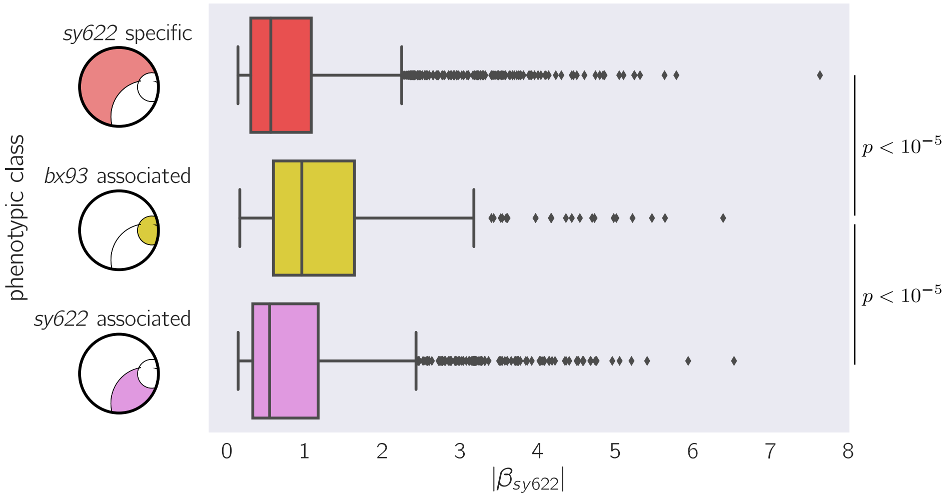

Figure 4. Within the *sy622* homozygote mutant, transcripts whose differential expression pattern places them in different phenotypic classes have statistically different distributions. The lines within the boxes show the 25th, 50th, and 75th percentiles. Whiskers show the 0th and 100th percentiles, with the exception of outliers (diamonds). Diagrams show what genotypes each gene class is expressed in, but the magnitude of the perturbation plotted always corresponds to the *sy622* mutant. The x-axis shows the absolute magnitude of the perturbation for each transcript in *sy622* homozygotes, $|\beta_{sy622}|$. The medians of the *sy622*-specific and the *sy622*-associated classes were statistically significantly different from the median of the *bx93*-specific class, as assessed by a non-parametric bootstrap test.

of the predicted β values from the optimized model against the observed β values of the heterozygote for each transcript should show the data falling along a line with slope equal to unity. Systematic deviations from linear behavior would indicate that the transcripts plotted are not part of a modular phenotypic class controlled by a functional unit.

The *sy622*-specific class expression phenotype of the *sy622* homozygote is complemented to wild-type levels by the presence of a *bx93* allele

Since our previous testing showed that the transcript expression of genes in this class was dysregulated in *sy622* homozygotes, and wild-type in both *bx93* homozygotes and *trans*-heterozygotes we can conclude that these transcripts are complemented to their wild-type levels by the presence of the *bx93* allele. Applying the Bayesian model yields identical results ($d_{bx93} = 1$). Thus, there is a module that has wild-type functionality in the *bx93* allele but is partially or completely deleted in the *sy622* allele. This functionality must require protein encoded between the amino acid position 1,698 where the *sy622* protein product truncates prematurely, and the position 2,549 where the *bx93* protein product ends.

The *bx93* allele is dominant over the *sy622* for the *bx93*-associated phenotype

We explored how expression levels of transcripts within the *bx93*-associated phenotypic class were controlled by these two alleles. Transcripts in this class are differentially expressed in homozygotes of either allele. Moreover, transcripts in this class are more perturbed in *sy622* homozygotes than in *bx93* homozygotes. This is consistent with a single functional unit that is impaired in the *bx93* allele, and even more impaired in the *sy622* allele (see Fig. 5).

If a single functional unit is being impaired, then we would expect these alleles to form a quantitative allelic series for this phenotypic class. In a quantitative series, alleles exhibit semidominance. We quantified the dominance coefficient for this class and found that the *bx93* allele is largely but not completely dominant over the *sy622* allele ($d_{bx93} = 0.81$; see Fig. 5). Dominance in the context of an allelic series indicates a qualitative allelic series, which is evidence that MDT-12 protein produced from the *bx93* allele has an intact functional unit that is deleted in protein product from the *sy622* allele. Mixed evidence for quantitative and qualitative allelic series at this phenotypic class precludes a definitive conclusion.

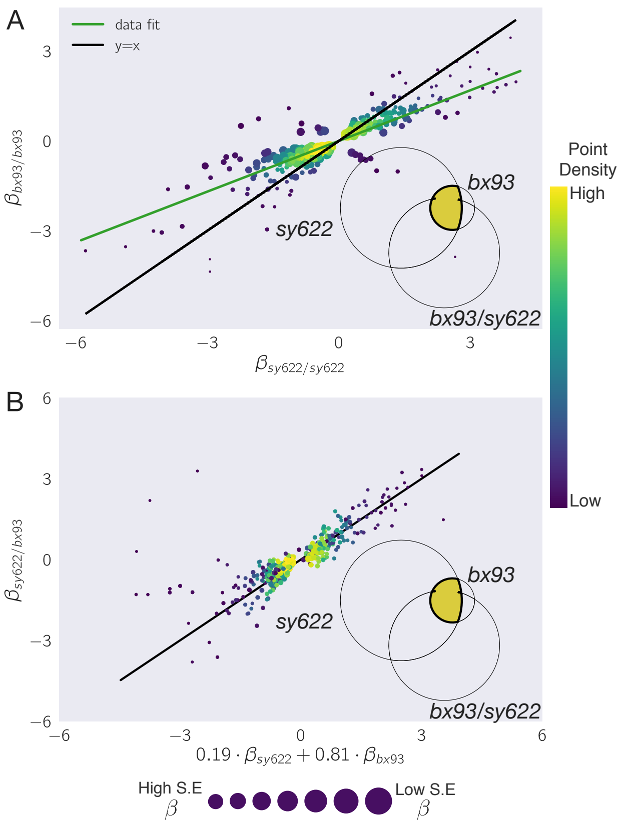

Figure 5. The *bx93*-associated class has properties of both quantitative and qualitative allelic series. **A** In *bx93* homozygotes, transcripts within the *bx93*-associated class are less perturbed than in *sy622* homozygotes. The line of best fit (green) is $\beta_{bx93/bx93} = 0.56 \cdot \beta_{sy622/sy622}$. **B** In a *trans*-heterozygote, the *bx93* allele is largely dominant over the *sy622* allele for the expression levels of transcripts in the *bx93*-associated class. In the graphs above, densely packed points are colored yellow as a visual aid. The size of the point is inversely proportional to the standard error of the β coefficients.

The *bx93* allele is semidominant with *sy622* for the *sy622*-associated phenotypic class

We quantified the relative dominance of *bx93* and *sy622* on the expression level of transcripts that belonged to the *sy622*-associated class. We found that both alleles are semidominant ($d_{bx93} = 0.51$). This suggests that there is a structure distributed evenly throughout the gene body starting the first amino acid position and ending before position 2,549. Since the two alleles are semidominant for transcript expression in this class, the functionality encoded in this gene must be dosage-dependent for this model to hold.

The *sy622*-specific class is strongly enriched for a Dpy transcriptional signature

bx93 homozygotic animals are almost wild-type, but careful measurements show that they have a slight body length defect causing them to be slightly Dpy, and *sy622* homozygotic animals are known to be severely Dpy[14], but this phenotype is complemented almost to *bx93* levels when this allele is placed in *trans* to the *sy622* allele. The only class that is fully complemented to wild-type levels is the *sy622*-specific class. Therefore, we hypothesized that the *sy622*-specific class should show a strong transcriptional Dpy signature.

To test this hypothesis, we derived a Dpy signature from two Dpy mutants (*dpy-7* and *dpy-10*, DAA, CPR and PWS *unpublished*) consisting of 628 genes. We used this gene set to look for a transcriptional Dpy signature in each phenotypic class using a hypergeometric probabilistic model (see Methods). We found that the *sy622*-specific and -associated classes were enriched in genes that are transcriptionally associated with a Dpy phenotype. The *bx93*-associated class also showed significant enrichment (fold-change $= 2.2$, $p = 4 \cdot 10^{-10}$, 68 genes observed). The enrichment was of considerably greater magnitude in the *sy622*-specific class (fold-change enrichment $= 3$, $p = 2 \cdot 10^{-40}$, 167 genes observed) than the enrichment in the *sy622*-associated class (fold-change $= 1.9$, $p = 9 \cdot 10^{-9}$, 82 genes observed) or in the *bx93*-associated class. Correlation analysis showed that a majority of the genes in the *sy622*-specific class were strongly correlated between the expression levels in the Dpy signature and the expression levels in *sy622* homozygotes, while 25% of the genes were anti-correlated (Spearman R $= 0.42$, $p = 6 \cdot 10^{-15}$, see Fig. 6). If the anti-correlated values are excluded from the Spearman regression, the statistical value of the regression improves significantly (Spearman R = 0.94, $p = 2 \cdot 10^{-108}$). Taken together, this suggests that the *sy622*-specific phenotypic class contains a transcriptional signature that can be associated with the morphological Dpy phenotype.

We also tested a hypoxia dataset[11], since *mdt-12* is not known to be upstream of the *hif-1*-dependent hypoxia response in *C. elegans*. Enrichment tests revealed that the hypoxia response was significantly enriched in the *bx93*-associated (fold-change = 2.1, $p = 10^{-8}$, 63 genes observed), the *sy622*-associated (fold-change = 1.9, $p = 4 \cdot 10^{-8}$, 78 genes observed) and the *sy622*-specific classes (fold-change = 2.4, $p = 9 \cdot 10^{-55}$, 186 genes observed). However, there was no correlation between the expression levels of these genes in *mdt-12* genotypes and the expression levels expected from the hypoxia response. Although the hypoxia gene battery can be found in *mdt-12* mutants, these genes are not used to deploy a hypoxia response, and the animals do not have a hypoxic-response phenotype.

Discussion

Allelic series using transcriptomic phenotypes can dissect the functional units of a gene

We have shown that whole-organism transcriptomic phenotypes can be analyzed in the context of an allelic series to partition the transcriptomic effects of a large, pleiotropic gene into separable classes. Analysis of these modules can inform structure/function predictions at the molecular level, and enrichment analysis of each class can be subsequently correlated with other morphologic or behavioral phenotypes. This method shows promise for analysing pathways that have major effects on gene expression in an organism, and which do not have complex, antagonistic tissue-specific effects on expression. Given the importance of allelic series for fully characterizing genetic pathways, we are optimistic that this method will be a useful addition towards making full use of the potential of these molecular phenotypes. Specifically, allelic series coupled with false hit analyses show great promise to identify distinct phenotypic classes that would be difficult or impossible to measure using standard methods. The sensitivity and quantitative nature of transcriptomic phenotypes makes identification of these phenotypes considerably more feasible. Once the phenotypic classes have been identified, dominance and enrichment analyses can be performed easily with significant statis-

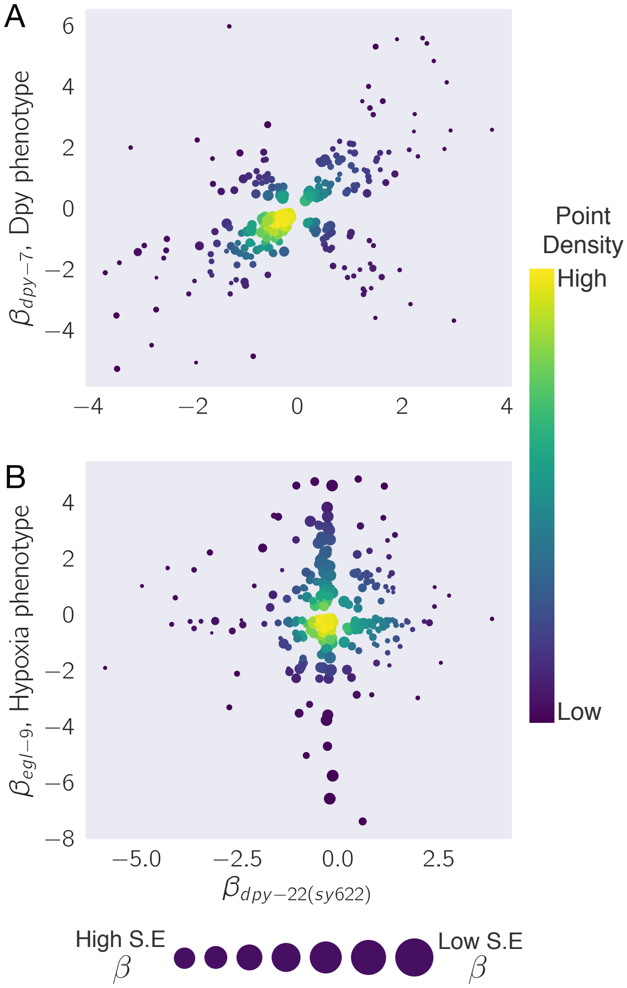

Figure 6. *sy622* homozygotes show a transcriptional response associated with the Dpy phenotype. **A** We obtained a set of transcripts associated with the Dpy phenotype from *dpy-7* and *dpy-10* mutants. We identified the transcripts that were differentially expressed in *sy622* homozygotes. Next, we plotted the β values of each transcript in *sy622* homozygotes against the β values in a *dpy-7* mutant. A significant portion of the genes are correlated between the two genotypes, showing that the signature is largely intact. 25% of the genes are anti-correlated. **B** We performed the same analysis using a set of transcripts associated with the *hif-1*-dependent hypoxia response as a negative control. Although *sy622* is enriched for the transcripts that make up this response, there is no correlation between the β values in *sy622* homozygotes and the β values in *egl-9* homozygotes. In the plots, a colormap is used to represent the density of points. The standard error of the mean is inversely proportional to the standard error of $\beta_{mdt-12(sy622)}$.

tical power. These properties highlight the power of coupling the genetical properties of *C. elegans* with next-generation sequencing methods.

A structure/function diagram of *mdt-12*

Our results strongly suggest the existence of various functional units in *mdt-12* that control distinct phenotypic classes (see Fig. 7). The *sy622*-specific class of transcripts is regulated normally in the presence of the *bx93* allele, indicating that the mutated protein product retains wild-type functionality for regulating these genes. This functionality is decreased or absent in MDT-12 produced from the *sy622* allele. Therefore, the functional unit that controls this class, functional unit 1 (FC1), must require sequence between amino-acid position 1,689 and position 2,549.

A similar argument can be made for a functional unit that controls *sy622*-associated transcripts, functional unit 2 (FC2). These genes are strongly perturbed in *sy622* homozygotes and they are also perturbed in *bx93/sy622 trans*-heterozygotes, albeit to a lesser degree. For this argument to hold, however, the functional unit must work in a dosage-dependent manner, since the *bx93* allele is semidominant with the *sy622* allele, and this unit is likely intact in the protein product made by the *bx93* allele. This is in contrast to FC1, which is not dosage-dependent.

Evidence in favor of a *bx93*-associated functional unit was mixed. Although dominance analysis suggested that the *bx93* allele was largely dominant over the *sy622* allele for expression levels of genes in this class, the expression of these genes deviated from wild-type levels in both alleles. The latter suggests that the *bx93*-associated module is perturbed quantitatively in both alleles, whereas dominance analyses favor an interpretation where the module is present in one allele but not in the other. One possibility is that the *bx93*-associated function we observed is the joint activity of two distinct effectors, functional units 3 and 4 (FC3, FC4, see Fig. 7). In this model, FC4 loses partial function in the *bx93* allele, whereas the FC3 retains its complete activity. This leads to non-wild-type expression levels of the *bx93*-associated class of transcripts. In the *sy622* allele, FC4 is further impaired, causing an increase in the severity of the observable phenotype. A rigorous examination of this model requires studying alleles that mutate the region between Q1689 and Q2549 using homozygotes and *trans*-heterozygotes. Future work should be able to establish how many modules exist in total, and how they may interact to drive gene expression. The phenotypic classes identified here could be compared against transcriptomic signatures from other transcription factors to identify candidate cofactors.

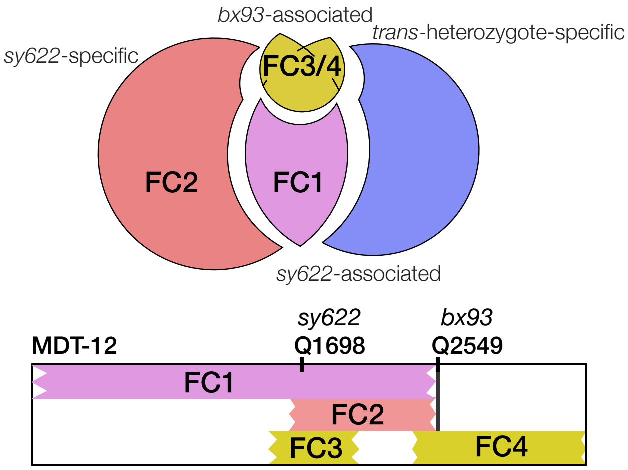

Figure 7. The functional units associated with each phenotypic class can be mapped to intragenic locations. The beginning and end positions of these functional units are unknown, so edges are drawn as ragged lines. Thick horizontal lines show the limit where each function could end, if known. We postulate that the *bx93*-associated class is controlled by two functional units, FC3 and FC4, in the tail region of this gene. Some of the modules shown may represent the same structures. Future experiments are required to make a more complete determination of the number and nature of these modules.

Controlling statistical artifacts

Transcriptomic phenotypes generate large amounts of information that can be used to determine functional units. However, due to the large number of tests performed, false positive and false negative events occur frequently enough to create populations of transcripts that have anomalous behaviors. It is necessary to identify what modules or populations are most at risk of these events and to what extent these modules may be polluted by false signals to prevent overinterpretation. In our experiment, we can estimate statistical noise in each population. There is a rich literature in genomics devoted to controlling and estimating false positive rates[25,26], but false negative rates have largely been ignored because they do not create spurious signal in simple experimental designs and because there is ample signal in most RNA-seq experiments. For allelic series experiments to be successful, systematic algorithms to estimate and control false negative rates, and to identify the populations most at risk for enrichment of false hits, must be developed, because false negative hits can create populations of genes that have fantastical biological behaviors (such as contrived examples of intragenic complementation or dosage models).

We performed a false hit analysis, estimating the false negative rate at 15%, to identify the clusters or classes of genes most at risk for statistical noise. As a general rule, small clusters or classes should be viewed with skepticism, particularly if the biological interpretation is complex. To perform a false hit analysis, we found it crucial is to appropriately idealize the shape of the Venn diagram. This idealized Venn diagram can then be "squeezed" with false negative and false positive rates to observe how it deforms. The deformed diagram can then be compared with reality to estimate the contribution of false hits to the existence of each class.

The *trans*-heterozygote specific phenotypic class is not a statistical artifact

In our study, we found a class of transcripts that were exclusively differentially expressed in *trans*-heterozygotes. The size of this class, 1226 genes, means it cannot be a statistical artifact. As a result, this class must be interpreted either as a legitimate aspect of *mdt-12* biology, possibly reflecting dosage- or tissue-specific effects, or as a strain-specific artifact. The genotype of the heterozygote includes a mutation at the *dpy-6* locus which acts as a cis-marker for the *bx93* mutation. One possibility is that the *dpy-6* loss-of-function mutation is not recessive for transcriptomic phenotypes and is responsible for the dysregulation of the new genes observed in the heterozygote. Another possibility is that the *dpy-6* strain had background mutations that affect gene expression levels in a complex manner. These issues could be addressed by re-generating the alleles used in this study using genome engineering tools like CRISPR Cas9, which have few off-target effects in *C. elegans*[27]. However, even if these issues were addressed, the biological interpretation of this class is not straightforward.

Phenotypes that are exacerbated or are unique to *trans*-heterozygotes often indicate that the protein products of the two alleles are somehow interfering with each other. This interference can often be the result of physical interactions such as homodimerization, or through a dosage reduction of a toxic product[28]. In the case of *mdt-12* orthologs, the protein products are not known to form oligomers. Instead, MDT-12 and its orthologs are expected to assemble in a monomeric manner into the CDK-8 Kinase Module.

A dosage model could explain the *trans*-heterozygote specific class if the dosage curve is bell-shaped. In this model, a switch is only activated at a very specific *mdt-12* gene dosage. Beyond this dosage, the switch remains off. Although such a model explains the data, mechanisms that could generate such a dosage curve are not immediately obvious. One possibility is that this switch is enacted at the level of cell specification or cell division, and that at the appropriate dosage of *mdt-12*, two cells that would typically collaborate to form a phenotype now act antagonistically, pushing *trans*-heterozygotes into a different state from the homozygotes. If this is the case, whole-organism RNA-seq may have limited resolution to identify what tissues or cells are being perturbed. Single-cell sequencing of *C. elegans* has recently been reported. As this technique becomes more widely adopted, and with decreasing cost, single-cell profiling of these genotypes may provide information that complements the whole-organism expression phenotypes, perhaps explaining the mysterious origin of this phenotype.

Analysis of allelic series using transcriptome-wide measurements

The potential of transcriptomes to perform epistasis analyses has been amply demonstrated[10,8], but their potential to perform allelic series analyses has been less studied. Though similar in some respects, epistasis analyses and allelic series studies call for different methods to solve different problems. To successfully perform an allelic series analysis, we must be able to identify the number and identity of the phenotypic classes, and a dominance analysis must be performed for each class to determine whether the alleles interact qualitatively or quantitatively with each other. Additionally, if an allelic series includes more than two alleles, the number of experimental outcomes and the number of possible outcomes rapidly become large.

The general problem of partitioning a set of genes into phenotypic classes is a common problem in bioinformatics. This problem has been tackled through clustering, matrix-based methods such as PCA or non-negative matrix factorization, or through q-value-based classification (as we have done here). Although these methods can classify genes or transcripts into clusters, by themselves they cannot ascertain the probability that any one cluster is real. For allelic series studies, this represents a major problem, since each cluster can in theory represent a new, independent functional unit within the molecular structure of the gene under study. Failure to identify clusters that are the result of statistical artifacts in general will cause researchers to identify inflated numbers of functional units within a molecular structure that appear to behave in a biologically spectacular fashion. We attempted to solve this problem for our series by estimating contributions of statistical noise to each class, although a challenge is that we do not know the false negative rate in our experiment. For our analysis, we exploited the molecular structure of our alleles (nested truncations) to create an idealized version of how gene clusters should behave. We then used our false positive rate and an estimated false negative rate to estimate the signal/noise ratio for each class. This method allows us to identify false classes, and in so doing it also reduces the apparent complexity of the molecular structure of the gene under study.

A challenge for allelic series studies will be the biological interpretation of unexpected classes, such as the *trans*-heterozygote specific class in our analysis. This class is too large to be explained by statistical anomalies. If this class is not an artifact of background or strain construction, the biological interpretation of this class is still not clear. Moreover, even if the biological interpretation of this class were clear, it is not immediately apparent what experimental design could establish the veracity of our interpretation. This problem could perhaps be ameliorated by correlating transcriptomic signatures with more morphologic, behavioral or cellular phenotypes, as has been done in single-cell studies[29].

Expression profiling as a method for phenotypic profiling

The possibility of identifying distinct phenotypes using expression profiling is an exciting prospect. With the advent of facile genome editing technologies, the allele generation has become routine. As a result, phenotypification is now the rate-limiting step for genetic analyses. We believe that RNA-seq can be used in conjunction with allelic series to exhaustively enumerate independent phenotypes with minor effort. We should push to sequence allelic diversity to more fully understand genotype-genotype variation.

Methods

Strains used

Strains used were N2 wild-type (Bristol), PS4087 *mdt-12(sy622)*, PS4187 *mdt-12(bx93)*, and PS4176 *dpy-6(e14) mdt-12(bx93)/ + mdt-12(sy622)*. All lines were grown on standard nematode growth media (NGM) Petri plates seeded with OP50 *E. coli* at 20°C[30].

Strain synchronization, harvesting and RNA sequencing

All 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 allowed to grow to the young adult stage (as assessed by vulval morphology and lack of embryos). RNA extraction was performed as described in[11] and sequenced using a previously described protocol[8].

Read pseudo-alignment and differential expression

Reads were pseudo-aligned to the *C. elegans* genome (WBcel235) using Kallisto[31], 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[32,33,34,35]. All libraries had good quality scores.

Differential expression analysis was performed using Sleuth[36]. Briefly, 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 wild-type replicates from other published and unpublished analysis. All wild-type replicates were collected at the same stage (young adult). In total, we had 10 wild-type replicates from 4 different batches, which heightened our statistical power. Batch effects were smaller than between-genotype effects, as assessed by principal component analysis (PCA), except when switching between samples constructed by different library methods. Wild-type samples constructed using the same library method clustered together and away from all other mutant samples. However, clustering wild-type samples by themselves revealed that the samples clusters correlated with the person who collected them. Therefore, we added batch correction terms to our model to account for batch effects from library construction as well as from the person who collected the samples.

Non-parametric bootstrap

We performed non-parametric bootstrap testing to identify whether two distributions had the same mean. Briefly, the two datasets were mixed, and samples were selected at random with replacement from the mixed population into two new datasets. We calculated the difference in the means of these new datasets. We iterated this process 10^6 times. To calculate a p-value that the null hypothesis is true, we identified the number of times a difference in the means of the simulated populations was greater than or equal to the observed difference in the means of the real population. We divided this result by 10^6 to complete the calculation for a p-value. If an event where the difference in the simulated means was greater than the observed difference in the means was not observed, we reported the p-value as $p < 10^{-6}$. Otherwise, we reported the exact p-value. We chose to reject the null hypothesis that the means of the two datasets are equal to each other if $p < 0.05$.

Dominance analysis

We modeled allelic dominance as a weighted average of allelic activity. Briefly, our model proposed that β coefficients of the heterozygote, $\beta_{a/b,i,\text{Pred}}$, could be modeled as a linear combination of the coefficients of each homozygote:

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

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:

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

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

All code was written in Jupyter notebooks[37] using the Python programming language. The Numpy, pandas and scipy libraries were used for computation[38,39,40] and the matplotlib and seaborn libraries were used for data visualization[41,42]. Enrichment analyses were performed using the WormBase Enrichment Suite[43]. 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.

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. We would like to thank Carmie Puckett Robinson for access to the unpublished Dpy transcriptional signature. Han Wang, Hillel Schwartz, Erich Schwarz and Porfirio Quintero provided valuable input throughout the project.

References

1. McClintock, B. THE RELATION OF HOMOZYGOUS DEFICIENCIES TO MUTATIONS AND ALLELIC SERIES IN MAIZE. *Genetics* **29**, 478–502 (1944).

2. FINCHAM, J. R. S. & PATEMAN, J. A. Formation of an Enzyme through Complementary Action of Mutant 'Alleles' in Separate Nuclei in a Heterocaryon. *Nature* **179**, 741–742 (1957).

3. 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).

4. 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).

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

6. 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).

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

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

9. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).

10. 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).

11. 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. *bioRxiv* (2017).

12. Bourbon, H.-M. *et al.* A Unified Nomenclature for Protein Subunits of Mediator Complexes Linking Transcriptional Regulators to RNA Polymerase II. *Molecular Cell* **14**, 553–557 (2004).

13. 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).

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

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

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

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

18. 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).

19. 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).

20. van de Peppel, J. *et al.* Mediator Expression Profiling Epistasis Reveals a Signal Transduction Pathway with Antagonistic Submodules and Highly Specific Downstream Targets. *Molecular Cell* **19**, 511–522 (2005).

21. Grants, J. M., Goh, G. Y. S. & Taubert, S. The Mediator complex of *Caenorhabditis elegans*: insights into the developmental and physiological roles of a conserved transcriptional coregulator. *Nucleic acids research* **43**, 2442–53 (2015).

22. 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).

23. Hodgkin, J., Horvitz, H. R. & Brenner, S. NONDISJUNCTION MUTANTS OF THE NEMATODE CAENORHABDITIS ELEGANS. *Genetics* **91** (1979).

24. Meneely, P. M. & Wood, W. B. Genetic Analysis of X-Chromosome Dosage Compensation in Caenorhabditis elegans. *Genetics* **117** (1987).

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

26. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing (1995).

27. Chiu, H., Schwartz, H. T., Antoshechkin, I. & Sternberg, P. W. Transgene-free genome editing in Caenorhabditis elegans using CRISPR-Cas.

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

29. Lane, K. *et al.* Measuring Signaling and RNA-Seq in the Same Cell Links Gene Expression to Dynamic Patterns of NF-κB Activation. *Cell Systems* **4**, 458–469.e5 (2017).

30. Brenner, S. The Genetics of CAENORHABDITIS ELEGANS. *Genetics* **77**, 71–94 (1974).

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

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

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

34. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Bowtie: An ultrafast memory-efficient short read aligner. [http://bowtie.cbcb.umd.edu/]. *Genome biology* R25 (2009).

35. 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).

36. 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).

37. 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).

38. 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).

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

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

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

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

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