A study of allelic series using 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*

December 2, 2017

Expression profiling holds great promise for genetics because of its ability to measure thousands of genes quantitatively. 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 distinct 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 regulation 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 epistasis analyses or with which to dissect the functional units contained within a single gene. To perform these analyses, we must extend genetic methods, developed for scalar phenotypes, to multiple dimensions. We developed new concepts and methods to study allelic series using expression profiling techniques. Briefly, allelic series are an important aspect of genetics because different alleles encode different versions of a gene. By studying phenotypic similarities and differences among 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 Mediator generates both general and specific phenotypes. We show that transcriptomic phenotypes renders the study of general factors such as *mdt-12* feasible.

Introduction

An 'allelic series' refers to a set of alleles with different phenotypes and can be used to understand the functions of a single locus. 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 in terms of its Mendelian inheritance pattern, 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 (see for example[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, and to identify what aspects of the pathway are sensitive to gene dosage. To address this problem, we developed a conceptual framework that enables us to count the number of phenotypes exhibited by an ensemble of alleles and their genotypes, perform dominance and complementation tests on these phenotypes to establish a dominance and complementation hierarchy, and which ultimately allows us to make inferences about the functional units contained within a gene and the sequence requirements for each unit.

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 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

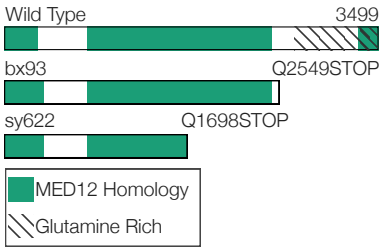

Figure 1. The *mdt-12* allelic series, consisting of two amino acid truncations. Diagram of the MDT-12 wild-type protein and the protein product of *bx93* and *sy622* alleles.

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

A conceptual framework for analyzing allelic series

Allelic series offer a way to study the functional units within a gene without requiring prior knowledge about the molecular structure of the mutations involved. In an allelic series, a set of alleles are selected. Then, a set of homozygotes of each allele is created, the phenotypes of each homozygote are enumerated and their severity scored to order the alleles by loss (or gain) of function relative to the wild-type phenotype. Next, alleles are placed in *trans* to each other to check whether one allele is dominant or semidominant over the other for each phenotype. In this way, alleles are ordered into a dominance hierarchy. This dominance hierarchy can be used to identify functional units and their sequence requirements within a gene (see Fig. 2).

Counting phenotypes in RNA-seq data

In theory, a gene could act along distinct pathways that alter the expression of distinct modules or clusters in different ways. These functions could be independent of one another, and could be encoded in different sites along a gene. To parse the functions contained within a gene from gene expression data, it is therefore crucial to enumerate the phenotypes a given set of alleles can generate. For two different alleles, we reasoned that the number of independent phenotypes should be equal to the number of intersection classes occurring in a Venn diagram of the two homozygotes and the *trans*-heterozygote in question. Thus, to enumerate the independent phenotypes associated with a set of (two) alleles, in a noiseless experiment it is sufficient to sequence a wild-type control, the homozygotes of the mutant alleles and the corresponding *trans*-heterozygote, label which genotypes perturb the expression of each transcript in the genome, and count the number of combinations of labels, which we call phenotypic classes, that occur. For a comparison involving 3 genotypes, the maximum number of phenotypes that can be identified is 7. In general, for a comparison involving N genotypes, the maximum number of phenotypes that can be identified is $2^N - 1$.

The presence of noise in an experiment complicates matters substantially. We envision a case where homozygotes of two different null alleles of a gene are sequenced and their transcriptomes are compared to each other. In the absence of noise, we expect the two transcriptomes to match exactly. When false positive and false negative rates are non-zero, however, the two transcriptomes will no longer contain exactly the same genes. In fact, if the false positive and the false negative rate are 10%, we should expect the two transcriptomes to overlap approximately at 60% of the genes that are differentially expressed in either genotype relative to a wild-type control. To appropriately count the number of phenotypes detected in this thought experiment, it would be necessary to ask whether each of the three classes contains a number of genes significantly greater than the number expected from statistical artifacts. Failure to account for this would determine the number of phenotypes to be three, an error rate of 200%. Importantly, the two null alleles would be determined to differ in their molecular functions, which constitutes a major qualitative error. In general, noise will make genetic interactions and genetic functions appear much more complex by increasing the number of seemingly dif-

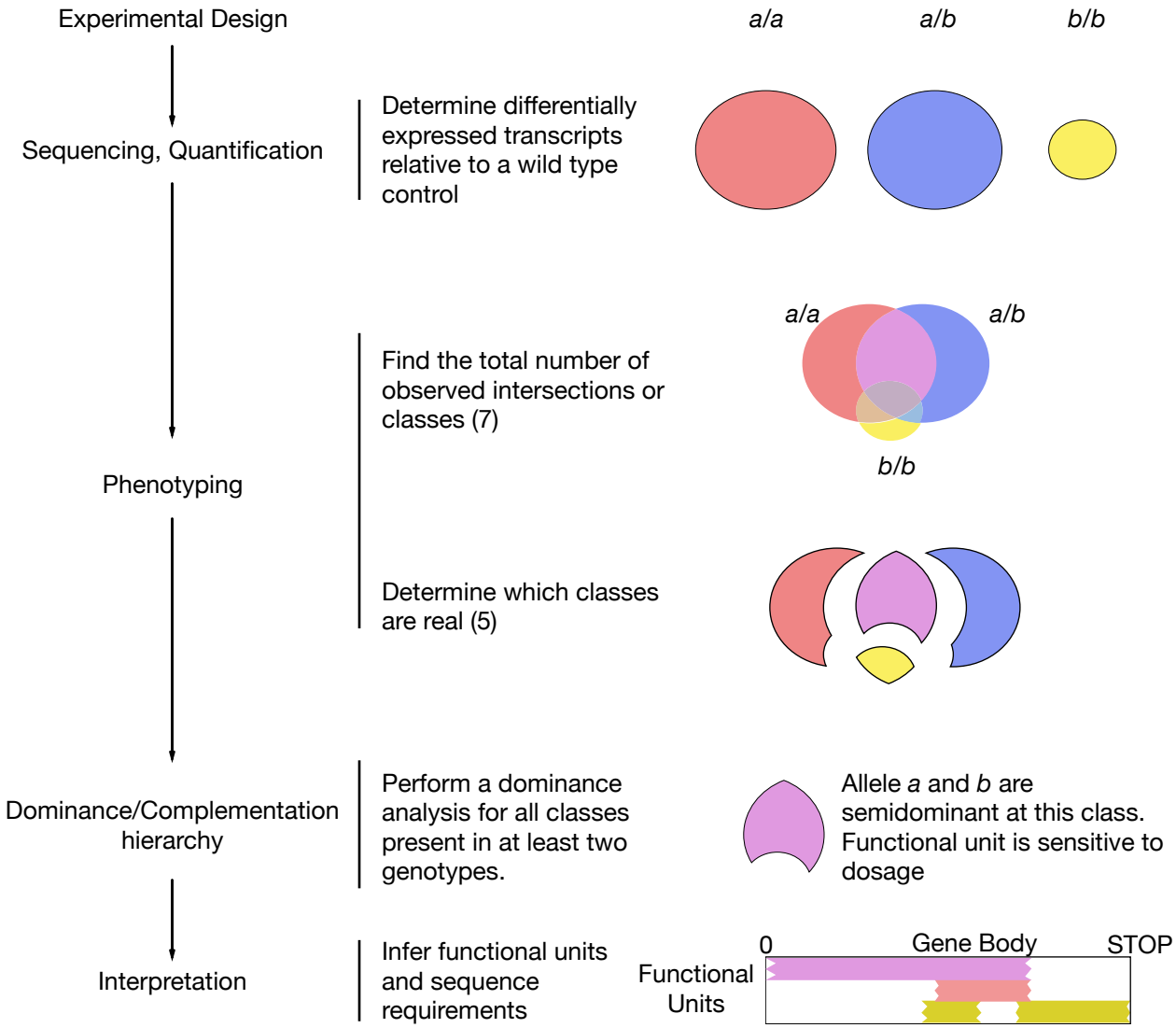

Figure 2. Flowchart for an analysis of arbitrary allelic series. A set of 2 or more alleles is selected, and the corresponding genotypes are generated sequenced to generate observable data. The number of phenotypes is estimated, correcting for statistical artifacts. For each phenotype, the alleles are ordered in a dominance/complementation hierarchy. The resulting dominance hierarchy can then be interpreted in terms of functional units within the genes in question.

ferent phenotypes that are detected.

We addressed the problem of phenotype counting by developing a false hit analysis algorithm (see Fig. 3). Briefly, the purpose of a false hit analysis is to identify those gene classes that are likely to be statistical artifacts and remove or re-classify them into the correct classes. To identify these classes, we first defined a minimal model as the gene class that is altered in all genotypes studied relative to the wild-type control. Subsequently, the noise generated by that minimal model is simulated from known or estimated false positive and negative rates. The noising process will generate additional classes not contained in the minimal model. Then, a signal-to-noise ratio can be estimated by comparing the observed size of each class with the size of of the class generated by the noising process. Classes that have a signal-to-noise greater than an arbitrary threshold are accepted and the minimal model is expanded to include those classes that have a high signal-to-noise ratio. The algorithm stops once the classes contained in the model have converged. A beneficial aspect of this algorithm is that the noising process can be split up into false positive and false negative components. False negative hits will tend to break up a single class into multiple—by estimating which classes are the result of false negative hits, and by modeling which class the false negative hits are most likely from, we can re-classify spurious classes, expanding the size and identity of biologically relevant classes. Once the number of phenotypic classes have been accounted for, these classes can be interpreted in terms of functional units depending on the genotypes that label them (the genotypes that affect these classes).

Establishing a dominance hierarchy between alleles

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. For gene expression data, after enumerating the phenotypic classes, which presumably reflect gene modules that are regulated independently from one another, a dominance or complementation hierarchy between the set of alleles must be derived for each phenotypic class. For some classes, this can be done immediately. Here, we briefly explain the logic for an allelic series

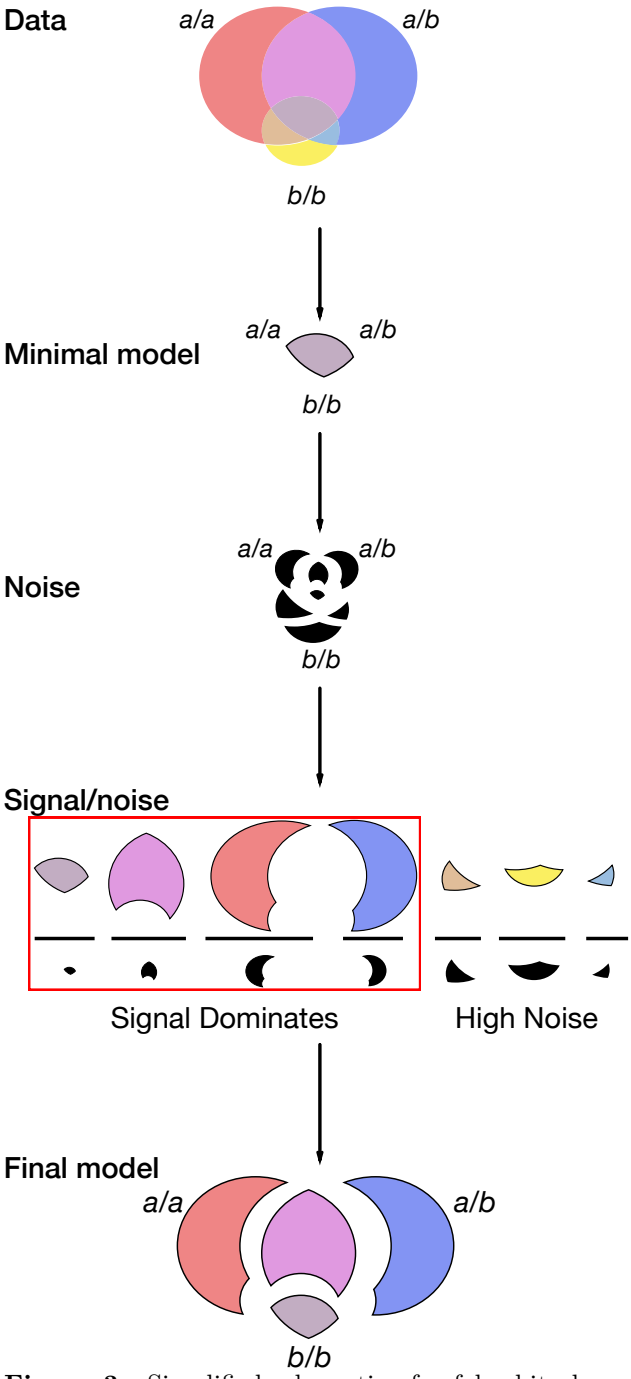

Figure 3. Simplified schematic of a false hit algorithm. For data from an arbitrary experimental design, a minimal model is inferred. Typically, the minimal model should be the intersection of all the groups measured. A noising process simulates the sizes of all the other classes expected if only the minimal model was true. A signal-to-noise statistic is calculated for each class. Classes that exceed a user-defined threshold are accepted and the model is refined. The algorithm converges when the classes no longer change from iteration to iteration.

containing three alleles (+, a, and b). Importantly, in the following text we assume that each phenotypic class reflects the same internal mechanism, such that a single dominance coefficient can accurately explain the behavior of each transcript within the class.

For example, a phenotypic class that contains genes that are perturbed only in one homozygotic genotype (a/a) relative to the wild-type control ($+/+$) but not in the other homozygotic genotype (b/b) or the *trans*-heterozgyotes (a/b) reveals regulation that is impaired in one allele but not in the other in a way such that the allele with wild-type functionality can complement the impaired allele to wild-type levels. This suggests that the gene contains at least one functional unit, impaired in one allele (a) but not in the other (b), and that the function of this unit is not dosage-dependent within the activity levels relevant to this allelic pair. Interpretation of this class of genes ideally does not require a further dominance analysis.

On the other hand, a phenotypic class where two genotypes, a/a and a/b, exhibit perturbed expression levels requires a dominance analysis for biological interpretation of this class. If the two alleles are semidominant, meaning the perturbation levels in the *trans*-heterozygote (a/b) are intermediate between the levels of the two homozygotes, then this means that one functional unit is impaired in one allele but not the other, and the function of this unit is dosage-sensitive within the activity levels relevant to these alleles. On the other hand, if the perturbations associated with the *trans*-heterozygotes are equal to those in the homozygotes a/a, then a is dominant negative over b. Dominant negative alleles are important biologically because they generate poisonous products that can often be highly informative of gene function.

Systematic analysis of each phenotypic class for two pairs of alleles will reveal the maximum and minimum number of functional units that may be differentially impaired between the two alleles. For example, if one phenotypic class shows the two alleles to be semidominant, whereas the other class shows one allele to be dominant over the other, the most parsimonious explanation is that there are at least two functional units encoded within the gene and which are affected by this allelic set. On the other hand, another hypothesis may posit that there is a single functional unit which is dominant in some specific context but semidominant in another. Thus, there may be 1–2 functional units in the gene under study. This number may be more exactly determined by studying many more alleles that affect the sequences required for these functions.

To quantify this dominance, we implemented and maximized a Bayesian model (see Methods; see also the Dominance Notebook). Briefly, for each transcript within a given class, we asked how the logarithm-transformed fold-change (which we refer to as a β coefficient) measured for each homozygote should be weighted to best predict the observed β values observed for the *trans*-heterozygote, subject to the constraint that the weights added up to 1 (see Dominance analysis). We reasoned that for modular phenotypes controlled by a single functional unit encoded within the gene of interest, then a plot 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.

RNA-seq analysis of an allelic series of a Mediator subunit

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. 3; see also the Basic Statistics Notebook).

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 (see the Basic Statistics Notebook).

False hit analysis identifies four phenotypic classes

We applied our false hit analysis to identify four non-overlapping phenotypic classes (see Fig. 4). 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. The false hit analysis suggested that this class should include all genes that were differentially expressed in *bx93* homozygotes and at least one other genotype, since it is likely that these genes represented false negative hits in the missing genotype. After re-classification, 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; see the Phenotypic Classes Notebook).

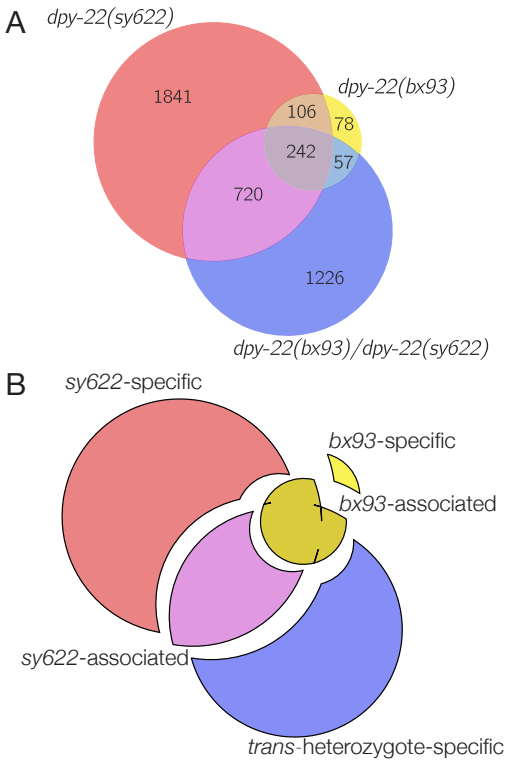

Figure 4. Transcripts under the control of *mdt-12* belong to distinct phenotypic classes. **A** Venn diagram of genes differentially expressed in each sequenced genotype relative to wild type before false hit analysis. **B** Exploded Venn diagram highlighting the four identified phenotypic classes after a false hit analysis.

Measurement of a dominance hierarchy

To dissect these alleles, establishment of a dominance hierarchy is required for each allele at each phenotypic class. Since the *sy622*-specific class is perturbed only in the *sy622* homozygotes, we conclude that the *sy622* allele is recessive to the *bx93* allele, which has wild-type functionality for this phenotypic class. Therefore, there is a module that is partially or completely deleted in the *sy622* allele but retains wild-type functionality in the *bx93* allele. This functionality requires 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 *sy622*-associated class contains genes with perturbed expression levels in both *sy622* homozygotes and in *trans*-heterozygotes. Interpretation of

this class requires a dominance analysis. The *sy622* and *bx93* alleles are semidominant ($d_{bx93} = 0.51$) to each other within this phenotypic class. This suggests that there is a structure requiring amino acids 1–2,549. Semidominance further indicates that the functionality encoded in this gene is dosage-dependent within the relevant activity levels of these alleles.

We also 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 was 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.

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). The *sy622*-specific and -associated classes were en-

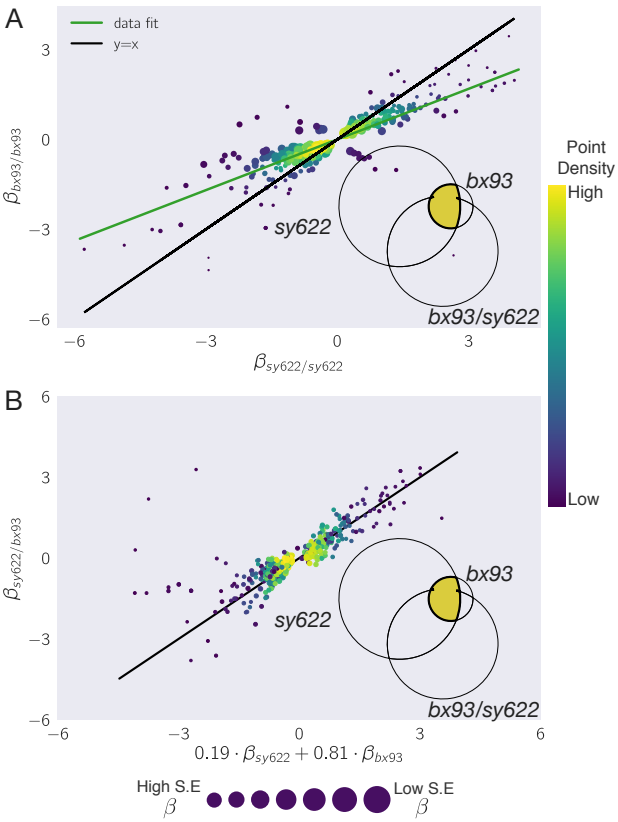

Figure 5. The *bx93*-associated class has properties of both quantitative and qualitative allelic series. **A** In *bx93* homozygotes, transcripts within the *bx93*-associated class are less perturbed than in *sy622* homozygotes. The line of best fit (green) is $\beta_{bx93/bx93} = 0.56 \cdot \beta_{sy622/sy622}$. **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.

riched in genes that are transcriptionally associated with a Dpy phenotype (fold-change enrichment = 3, $p = 2 \cdot 10^{-40}$, 167 genes observed; fold-change = 1.9, $p = 9 \cdot 10^{-9}$, 82 genes observed). The *bx93*-associated class also showed significant enrichment (fold-change = 2.2, $p = 4 \cdot 10^{-10}$, 68 genes observed). The class that showed the most extreme deviation from random was the *sy622*-specific class, consistent with our hypothesis. Plotting the perturbation levels in the *sy622* homozygotes versus the perturbation levels in the *dpy-7* mutants revealed that 75% of the transcripts were strongly correlated in both genotypes. Therefore, the *sy622*-specific phenotypic class contains a transcriptional signature associated with morphological Dpy phenotype (see the Enrichment Notebook).

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 *hif-1*-dependent hypoxia phenotype.

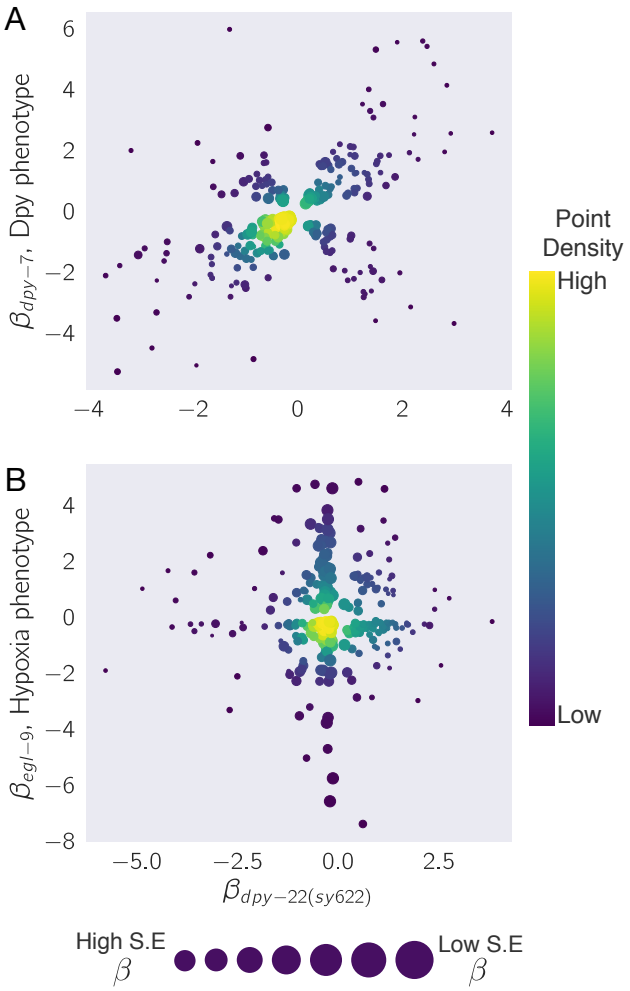

Figure 6. *sy622* homozygotes show a transcriptional response associated with the Dpy phenotype. **A** We obtained a set of transcripts associated with the Dpy phenotype from *dpy-7* and *dpy-10* mutants. We identified the transcripts that were differentially expressed in *sy622* homozygotes. Next, we 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 $\beta_{mdt-12(sy622)}$.

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 phenotypic classes that would otherwise be difficult if not impossible to identify using other methods. Analysis of these modules can inform structure/function predictions, and enrichment analysis of each class can be used to associate transcriptional modules with morphologic or behavioral phenotypes. This method shows promise for analyzing 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 path-

ways, we are optimistic that this method will be a useful addition towards making full use of the potential of these molecular phenotypes. Once the phenotypic classes have been identified, dominance and enrichment analyses can be performed easily with significant statistical power.

A structure/function diagram of *mdt-12*

Our results suggest the existence of various functional units in *mdt-12* that control distinct phenotypic classes (see Fig. 7). It seems likely that the *sy622*-specific phenotypic class is controlled by a single functional unit, functional unit 1 (FC1), whereas the *sy622*-associated phenotypic class is controlled by a second functional unit, functional unit 2 (FC2). Although possible, it is highly unlikely that these functional units are the same because the dominance behaviors are different among the two phenotypic classes.

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 a 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. FC3 could be the same unit as FC2, since neither unit behaves in a dosage sensitive manner. 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.

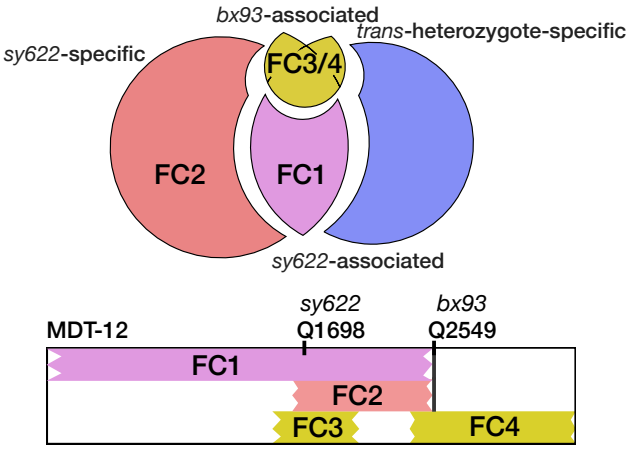

Figure 7. The functional units associated with each phenotypic class can be mapped to intragenic locations. The beginning and end positions of these functional units are unknown, so edges are drawn as ragged lines. Thick horizontal lines show the limit where each function could end, if known. We postulate 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 over-interpretation. In our experiment, we can estimate statistical noise in each population. There is a rich literature in genomics devoted to estimating and controlling 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). As a general rule, small clusters or classes should be viewed with

skepticism, particularly if the biological interpretation is complex. We expect these conclusions will be of broad significance to genomics research where highly multiplexed measurements are often compared to identify similarities and differences in the genome-wide behavior of a single variable under multiple conditions.

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 that 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* activity level. Beyond this threshold, 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[29]. 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.

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[30].

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[31].

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[32], 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[33,34,35,36]. All libraries had good quality scores.

Differential expression analysis was performed using Sleuth[37]. 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[38] using the Python programming language. The Numpy, pandas and scipy libraries were used for computation[39,40,41] and the matplotlib and seaborn libraries were used for data visualization[42,43]. Enrichment analyses were performed using the WormBase Enrichment Suite[44]. 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 will be 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)[45] under the accession code GSE107523 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSExxx).

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 the unpublished Dpy transcriptional signature. Han Wang, Hillel Schwartz, Erich Schwarz, Porfirio Quintero and Carmie Puckett Robinson 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** (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 (2013).

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

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

30. 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).

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

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

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

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

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

36. 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).

37. 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).

38. 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).

39. 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).

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

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

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

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

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

45. 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). URL http://www.ncbi.nlm.nih.gov/pubmed/11752295http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC99122.