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 22, 2017

Although transcriptomes have recently been used to perform epistasis analyses, they are not yet used to study intragenic function/structure relationships. We developed a theoretical framework to study allelic series using transcriptomic phenotypes. As a proof-of-concept, we apply our methods to an allelic series of *mdt-12*, a highly pleiotropic Mediator subunit gene in *Caenorhabditis elegans*. Our methods identify functional units within *mdt-12* that modulate Mediator activity upon various genetic modules.

Introduction

An 'allelic series' refers to a set of alleles with different phenotypes and can be used to understand the functions encoded within a single gene regardless of the molecular nature of the alleles used. Briefly, in an allelic series a set of alleles are used to generate a set of homozygotes. These genotypes are used to explore the number and severity of phenotypes encoded by the alleles in question. Using the homozygote genotypes, the alleles can be ordered by severity of effect relative to the wild type allele for each measured phenotype. Then, *inter se trans*-heterozygotes are used to establish dominance (complementation) hierarchies within the allele set for each phenotype in question. Together, the severity and dominance hierarchies are used to infer intragenic functional units. Allelic series can be quantitative, in which case all alleles are semidominant to each other, indicating a single functional unit that is dosage-dependent in activity range relevant to the alleles studied; alternatively, an allelic series can be qualitative, in which case some alleles are entirely dominant over the others, indicating that some alleles have wild-type functionality that complements the mutant functionality of the recessive alleles. Allelic series have been used to study the dose response curve of a phenotype for a particular gene and to infer null phenotypes from hypomorphs. In *Caenorhabditis elegans*, the *let-23*, *lin-3* and *lin-12* allelic series stand out as examples[1,2,3].

Biology has moved from expression measurements of single genes towards genome-wide measurements. Expression profiling via RNA-sequencing[4] (RNA-seq) enables simultaneous measurement of transcript levels for all genes in a genome. These measurements can be made on a whole-organism scale or on single cells[5,6]. Transcriptomes have been successfully used to identify new cell or organismal states[7,8] and both methods can be used for genetic analysis[9,10]. 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. Here, we show how to perform an allelic series analysis by counting phenotypic classes (the transcriptomic equivalent of observable phenotypes), and generating a dominance hierarchy between the alleles in question for each class to draw conclusions about the function/structure relationship of the gene under study.

As a proof of principle, we selected three alleles[11,12] of a Mediator complex subunit in *C. elegans*, *mdt-12*. Mediator is a macromolecular complex that contains approximately 25 subunits[13] and which globally regulates RNA polymerase II (Pol II)[14,15]. The Mediator complex consists of four biochemical modules: the Head, Middle and Tail modules and a CDK-8-associated Kinase Module (CKM). The CKM can associate reversibly with the other modules, and it appears to inhibit transcription[16,17]. In *C. elegans*, the CKM consists of CDK-8, MDT-13, CIC-1 and DPY-22[18]. Loss of *mdt-12* is lethal in XO animals[?, ?]. *mdt-12* acts in the formation of the male tail[11], where it interacts with the Wnt pathway, and in vulval formation[19], where it inhibits the Ras pathway. Studies in the male tail were carried out using allele *dpy-22(bx93)*, which generates

../figs/gene_model_dpy22.pdf

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.

a truncated DPY-22 protein as a result of a premature stop codon, Q2549Amber[11]. However, animals homozyguous for this allele grossly appear phenotypically wild-type. In contrast, animals homozyguous for the *dpy-22(sy622)* allele, which encodes a premature stop codon, Q1698Amber[12] (see Fig. 1), are severely dumpy (Dpy), have egg-laying defects (Egl) and a low penetrance multivulva (Muv) phenotype. Due to its pleiotropic effects, a conclusive allelic series analysis has not previously been performed.

Expression profiles have the potential to facilitate dissection of molecular structures within genes, but the high dimensionality of these phenotypes make analysis challenging. We developed a set of conceptual and algorithmic methods to analyze allelic series using transcriptomic phenotypes and applied our methods to an allelic series involving the MDT12 ortholog in *C. elegans*, *mdt-12*. Our analysis revealed a number of functional units that act independently to modulate Mediator activity at thousands of genetic loci.

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, homozygotes of each allele are generated (if possible), 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 allele. Finally, alleles are placed in *trans* to each other to check whether one allele is dominant or semidominant over the other for each phenotype, resulting in an ordered 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 encode multiple independent functions along its sequence. To parse the separate functions contained within a gene from gene expression data, this data must be broken up into phenotypic classes, and each class associated with the genotypes that perturb them. 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 differentially expressed genes (relative to a wild type control) of the mutant homozygotes and the *inter se trans*-heterozygotes. 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. 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, the list of differentially expressed genes and its perturbation are equal between the two homozygotes. When false positive and false negative rates are non-zero, however, the lists differentially expressed genes will diverge. If the false positive and the false negative rate are 10%, we should expect the two lists to overlap at 60% of the genes in either list. To accept the hypothesis that these two alleles are different in nature, it is necessary to test whether each of the three classes contains a number of genes significantly different from the number expected from statistical artifacts. Failure to account for this would lead to the conclusion that both alleles are not identical and at least one of them is not a null allele, a major qualitative error. Noise will make genetic interactions and genetic functions appear more complex by increasing the number of seemingly different phenotypes detected.

We addressed the problem of phenotype counting by developing a false hit analysis (see Fig. **??**). Briefly, the purpose of a false hit analysis is to identify gene classes that are likely to be statistical artifacts and remove or re-classify them into the correct classes. To identify these classes, we defined a minimal model as the gene class altered in all genotypes studied relative to the wild-type control. Subsequently, the noise generated by that minimal model

../figs/flowchart.pdf

Figure 2. Flowchart for an analysis of arbitrary allelic series. A set of alleles is selected, and the corresponding genotypes are sequenced. 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.

is simulated from known or estimated false positive and negative rates. The noising process will generate additional classes not contained in the minimal model. 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 them. The algorithm stops once the classes contained in the model have converged. The simulation of 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 classes—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 affecting these classes). Because certain classes can become artificially enriched in false positives, whereas others will become depleted of false positive hits (and *vice versa* for false negative hits), false hit analysis can significantly reduce the effective false positive rate within a study. This is important, because although setting more stringent q-value thresholds leads to fewer false positives, it will inevitably increase the false negative rate, which in turn will create certain classes at the expense of others. Additionally, smalller classes will suffer from decreased statistical power, limiting their ability to detect small effects.

./figs/false_hit_analysis.pdf

Figure 3. Schematic of a false hit analysis. 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.

Establishing a dominance hierarchy between alleles

After enumerating the phenotypic classes, which presumably reflect independent regulatory mechanisms, a dominance or complementation hierarchy between the set of alleles must be derived for each phenotypic class. This hierarchy may then be used to identify a minimum and maximum number of functional units affected within the allele set. 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.

To quantify dominance for a phenotypic class, we implemented and maximized a Bayesian model (see Methods; see also the Dominance Notebook). 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.

Proof-of-principle 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. **??**; 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[12]. The *trans*-heterozygote transcriptome had 2,214 differentially expressed genes (see the Basic Statistics Notebook).

False hit analysis identifies four phenotypic classes

We used a false hit analysis to identify 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. 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. The ***bx93*-specific** class of transcripts is expected to contain on average 39 false positive hits. Therefore, this class is most likely artifactual, and we do not consider it for the rest of our analysis. 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).

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, the *sy622* allele is recessive to the *bx93* allele, which has wild-type functionality for this phenotypic class. Therefore, there is a functional unit that is impaired in the *sy622* allele but not in the *bx93* allele. This unit requires protein encoded between amino acid position 1,698 where the *sy622* protein product truncates prematurely, and 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*

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.

and *bx93* alleles are semidominant ($d_{bx93} = 0.51$) to each other within this phenotypic class. This suggests that there is a functional unit requiring amino acids 1–2,549.

Transcripts in the *bx93*-associated phenotypic class do not have wild-type expression in any genotype except the wild type. 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. 4). If a single functional unit was impaired, then we would expect these alleles to form a quantitative allelic series within 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. 4), which indicates a qualitative allelic series. These two alleles form a mixed allelic series at this phenotype, precluding a definitive conclusion regarding functional units.

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 *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[12], 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 enriched 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[10], 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.

../figs/dpy_phenotype.pdf

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 classes can inform structure/function predictions, and enrichment analysis of each class can be used to associate transcriptional classes 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 pathways, 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.

../figs/inferred_domains.pdf

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 units shown may be redundant.

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. 5). 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. 5). 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 functional units 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.

Controlling statistical artifacts

Transcriptomic phenotypes generate large amounts of information that can be used to determine functional units. However, false positive and false negative events occur frequently enough to create artifactual populations of transcripts. Moreover, the distribution of false positive and false negative hits may not be uniform when comparing differentially expressed transcripts relative to a wild-type control between multiple genotypes. Identifying what classes are most at risk of these events and to what extent these classes may be polluted by false signals will prevent over-interpretation and may significantly decrease the apparent complexity of a gene or a genetic interaction, because artifactual classes can often exhibit 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 implausible. These conclusions are of broad significance to chromatin research where highly multiplexed measurements are 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 is not 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 strain-specific artifacts. 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.

If the *trans*-heterozygote specific class is not artifactual, its biological interpretation 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[20]. 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 explains the *trans*-heterozygote specific class if the response 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 apparent. Single-cell sequencing of *C. elegans* has recently been reported[21]. 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 origin of this phenotype.

Analysis of allelic series using transcriptome-wide measurements

The potential of transcriptomes to perform epistasis analyses has been amply demonstrated[9,7], 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[22].

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)*. Lines were grown on standard nematode growth media (NGM) Petri plates seeded with OP50 *E. coli* at 20°C[23].

Strain synchronization, harvesting and RNA sequencing

Strains were synchronized by bleaching P_0's into virgin S. basal (no cholesterol or ethanol added) for 8–12 hours. Arrested L1 larvae were placed in NGM plates seeded with OP50 at 20°C and grown to the young adult stage (assessed by vulval morphology and lack of embryos). RNA extraction was performed as described in[10] and sequenced using a previously described protocol[7].

Read pseudo-alignment and differential expression

Reads were pseudo-aligned to the *C. elegans* genome (WBcel235) using Kallisto[24], 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[25,26,27,28]. All libraries had good quality scores.

Differential expression analysis was performed using Sleuth[29]. We used a general linear model to identify genes that were differentially expressed between wild-type and mutant libraries. To increase our statistical power, we pooled young adult wild-type replicates from other published[10,7] and unpublished analysis adjusting for batch effects.

Non-parametric bootstrap

We performed non-parametric bootstrap testing to identify whether two distributions had the same test statistic using 10^6 bootstraps. If no statistics equal to or greater than the observed statistic was 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:

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

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

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

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

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

Code

Code was written in Jupyter notebooks[30] using the Python programming language. The Numpy, pandas and scipy libraries were used for computation[31,32,33] and the matplotlib and seaborn libraries were used for data visualization[34,35]. Enrichment analyses were performed using the WormBase Enrichment Suite[36]. For all enrichment analyses, a q-value of less than 10^{-3} was considered statistically significant. For gene ontology enrichment analysis, terms were considered statistically significant only if they also showed an enrichment fold-change greater than 2.

Data Availability

Raw and processed reads were deposited in the Gene Expression Omnibus. Scripts for the entire analysis can be found with version control in our Github repository, https://github.com/WormLabCaltech/med-cafe. A user-friendly, commented website containing the complete analyses can be found at https://wormlabcaltech.github.io/med-cafe/. Raw reads and quantified abundances for each sample were deposited at the NCBI Gene Expression Omnibus (GEO)[37] under the accession code GSE107523 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107523).

Acknowledgements

This work was supported by HHMI with whom PWS was an investigator, by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology, and by the NIH grant U41 HG002223. This article would not be possible without help from Dr. Igor Antoshechkin and Dr. Vijaya Kumar who performed the library preparation and sequencing. 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. Aroian, R. V. & Sternberg, P. W. Multiple functions of let-23, a *Caenorhabditis elegans* receptor tyrosine kinase gene required for vulval induction. *Genetics* **128**, 251–67 (1991).

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

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

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

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

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

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

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

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

10. Angeles Albores, D., Puckett Robinson, C., Williams, B. A., Wold, B. J. & Sternberg, P. W. Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements. *bioRxiv* (2017).

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

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

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

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

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

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

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

18. 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).

19. 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).

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

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

22. 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).

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

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

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

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

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

28. 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).

29. 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).

30. 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).

31. 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).

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

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

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

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

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

37. 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).