

Analysis of allelic series with transcriptomic phenotypes

David Angeles-Albores[†] and Paul W. Sternberg^{†,1}

[†]Division of Biology and Biological Engineering, Caltech, Pasadena, CA, 91125, USA

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

KEYWORDS Genetic interactions; Transcriptome; Mediator; Wnt; Ras; Dominance; Cdk8

Mutations of a gene can yield a series of alleles with different phenotypes that reveal multiple functions encoded by that gene, regardless of the alleles' molecular nature. In *Caenorhabditis elegans*, allelic series have characterized genes such as *let-23/EGFR*, *lin-3/EGF* and *lin-12/NOTCH* (Aroian and Sternberg 1991; Ferguson and Horvitz 1985; Greenwald *et al.* 1983). Allelic series provide a way to probe genes where biochemical approaches would be difficult, slow or uninformative with regards to the biological phenomenon of interest. Their power derives from the ability to draw broad conclusions about the gene of interest in terms of gene dosage and functional units, to the extent that these two factors are separable, without regard to the molecular identity of the mutations that created these alleles. Here, gene dosage is defined as the combined effects of transcriptional and translational expression, gene product localization, and biochemical kinetics of the final gene product *in situ*. To study allelic series, we must first enumerate the phenotypes each allele affects, and subsequently order the alleles into severity and dominance hierarchies per phenotype. The resulting hierarchies enable us to better understand how a given gene, which may be highly pleiotropic, can give rise to highly specific mutant phenotypes when mutated in just the right way.

Biology has moved from expression measurements of single

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

We have devised methods for characterizing allelic series using RNA-seq. To test these methods, we selected three alleles (Zhang and Emmons 2000; Moghal and Sternberg 2003) of a *C. elegans* Mediator complex subunit gene, *dpy-22*. Mediator is a macromolecular complex with ~ 25 subunits (Jeronimo and Robert 2017) that globally regulates RNA polymerase II (Pol II) (Allen and Taatjes 2015; Takagi and Kornberg 2006). The Mediator complex has at least four biochemically distinct modules: the Head, Middle and Tail modules and a CDK-8-associated Kinase Module (CKM). The CKM associates reversibly with other modules, and appears to inhibit transcription (Knuesel *et al.* 2009; Elmlund *et al.* 2006). In *C. elegans* development, the CKM promotes the formation of the male tail (Zhang and Emmons 2000) (through interactions with the Wnt pathway), as well as formation of the hermaphrodite vulva (Moghal and Sternberg 2003) (through inhibition of the Ras pathway). Null alleles of *dpy-22* are likely to be lethal, based on embryonic lethal phenotypes observed after RNAi (Wang *et al.* 2004; Lehner *et al.* 2006) and the severe phe-

Copyright © 2018 by the Genetics Society of America

doi: 10.1534/genetics.XXX.XXXXXX

Manuscript compiled: Thursday 19th July, 2018%

¹ Division of Biology and Biological Engineering, Mail Code 156–29, Biology Division, Caltech, 1200 E California Blvd., Pasadena, California 91125. E-mail: pws@caltech.edu

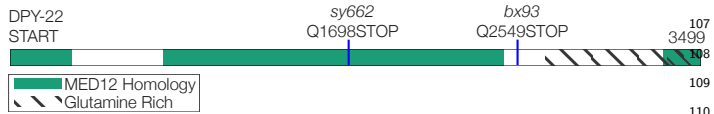


Figure 1. Protein sequence schematic for DPY-22. The positions of the nonsense mutations used are shown.

notypes of a strong *dpy-22* hypomorphic allele, *dpy-22(e652)* (homozygous hermaphrodites are very sick) (Riddle *et al.* 1997). Homozygotes of allele *dpy-22(bx93)*, which encodes a premature stop codon Q2549Amber (Zhang and Emmons 2000), appear grossly wild-type, though this allele does not have complete wild-type functionality, since it fails to fully complement the Muv phenotype of another allele, *sy622*, in a sensitized *let-23* background. In contrast, animals homozygous for a more severe allele, *dpy-22(sy622)* encoding another premature stop codon, Q1698Amber (Moghal and Sternberg 2003), are dumpy (Dpy), have egg-laying defects (Egl), and have multiple vulvae (Muv) (Fig. 1). In humans, MED12 is known to have a proline-, glutamine- and leucine-rich domain that interacts with the WNT pathway (Kim *et al.* 2006). However, many disease-causing variants fall outside of this domain (Yamamoto and Shimojima 2015). In spite of its causative role in a number of neurodevelopmental disorders (Graham and Schwartz 2013), the structural and functional features of this gene are poorly understood, partially because genetic approaches towards studying pleiotropic genes have proved difficult in the past, highlighting the need for new methods.

Methods

Strains used

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

Strain synchronization, harvesting and RNA sequencing

With the exception of strain MT4866, strains were synchronized by bleaching P₀'s into virgin S. basal (no cholesterol or ethanol added) for 16–18 hours. Arrested L1 larvae were placed in NGM plates seeded with OP50 at 20°C and grown to the young adult stage (assessed by vulval morphology and lack of embryos). We discovered that MT4866 dies upon L1 starvation for this period of time. As a result, we synchronized this strain by double bleaching. Animals were picked if they were young adults, regardless of whether any vulval or morphological phenotypes were present. RNA extraction and sequencing was performed as previously described by Angeles-Albores *et al.* (2018b, 2017). Briefly, young adults were placed in 10 μ L of TE buffer, and digested using Recombinant Proteinase K PCR Grade (Roche Lot 656 No. 03115 838001) incubated with 1% SDS 657 and 1.25 μ L RNA Secure (Ambion AM7005). Total RNA was extracted using the Zymo Research Directzol RNA MicroPrep Kit (Zymo Research, SKU R2061). mRNA was subsequently purified using a NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, NEB, #E7490). Sequencing

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

Read pseudo-alignment and differential expression

Reads were pseudo-aligned to the *C. elegans* genome (WB-cel235) using Kallisto (Bray *et al.* 2016), using 200 bootstraps and with the sequence bias (`-seqBias`) flag. The fragment size for all libraries was set to 200 and the standard deviation to 40. Quality control was performed on a subset of the reads using FastQC, RNAseQC, BowTie and MultiQC (Andrews 2010; Deluca *et al.* 2012; Langmead *et al.* 2009; Ewels *et al.* 2016).

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

False hit analysis

To accurately count phenotypes, we developed a false hit algorithm (Algorithm 1). We implemented this algorithm for comparisons of three genotypes using Python. Such an experiment can result in 128 possible combinations of phenotypic classes (ignoring size). This large number of models necessitates an algorithmic approach that can restrict the number of models. Our algorithm uses a noise function that assumes false hit events are non-overlapping (i.e. the same gene cannot be the result of two false positive events in two or more genotypes) to determine the average noise flux between phenotypic classes. These assumptions break down if false-positive or negative rates are large (>25%).

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

We selected $\lambda = 3$ because classification using this threshold was high across a range of false positive and false negative combinations. A challenge to applying this algorithm to our data is the fact that the false negative rate for our experiment is unknown. Although there has been significant progress in controlling and estimating false positive rates, we know of

no such attempts for false negative rates. It is unlikely that the false negative rate for our study is lower than the false positive rate, because all genotypes except the controls are likely underpowered. We used false negative rates between 10–20% for false hit analysis. All analyses returned the same final model.

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

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

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

Result: $M_{reduced}$, a reduced model that fits the data.

```

begin
  Define a minimal model, K
  Refine the model until convergence or iterations max out
  i ← 0
  Kprev ← ∅
  while (i < imax) | (Kprev ≠ K) do
    Kprev ← K
    Define a noise function to estimate error flows in K
    F ← noise(K, fp, fn)
    for l ∈ L do
      Calculate signal to noise for each labelled class False
      negatives can result in λ < 0 λl ← Mobs,l / Fl if
      (λ > α) | (λ < 0) then
        Kl ← Mobs,l
      end
    end
    end
    i ++
  end
end
Mreduced = K
return Mreduced

```

Dominance analysis

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

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

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

Code

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

Data Availability

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

Results

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

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

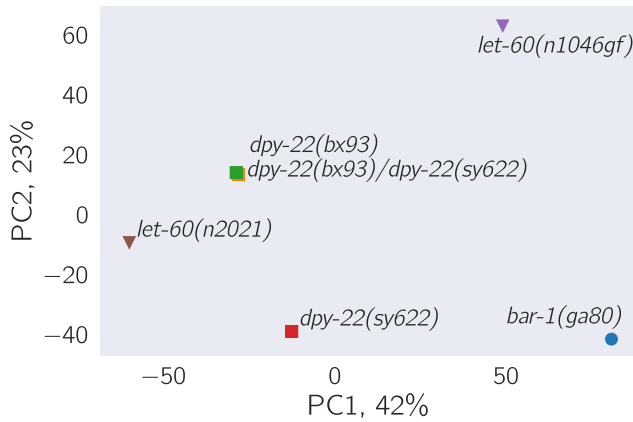


Figure 2. Principal component analysis of the analyzed genotypes. The analysis was performed using only those transcripts that were differentially expressed in at least one genotype. The plot shows that the *trans*-heterozygotes phenocopy the *dpy-22(bx93)* homozygotes along the first two principal dimensions.

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

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

As a first step in our analysis, we performed dimensionality reduction on the transcriptomes we sequenced using Principal Component Analysis (PCA). Briefly, PCA identifies the vectors along which there is most variation in the data. These vectors can be used to project the data into lower dimensions to assess whether samples cluster, though interpreting the biological reasons for this clustering can be challenging. To perform PCA, we selected only those transcripts that were differentially expressed in at least one genotype, and used the β coefficients associated with these genes to perform PCA. Projecting the data into two dimensions maintains 65% of the variation. The first dimension separates the gain and loss of function *let-60* mutants. The second dimension separates the *dpy-22* mutants (Fig. 2). On the PCA plot, the *trans*-heterozygote mutants appear to phenocopy the *dpy-22(bx93)* mutants, recapitulating previous experiments that showed the *dpy-22(bx93)* allele to be dominant over the *dpy-22(sy622)* allele.

Three *dpy-22* genotypes have shared transcriptomic phenotypes

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

may represent multiple independent modules that formally constitute different phenotypic classes.

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

False hit analysis identifies four non-overlapping phenotypic classes

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

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

We used false hit analysis to identify four non-overlapping phenotypic classes (Fig. 3). We use the term genotype-specific to refer to groups of transcripts that were perturbed in one mutant genotype. We use the term genotype-associated to refer to those groups of transcripts whose expression was significantly altered in two or more mutants genotypes with respect to the wild type control. The *dpy-22(sy622)*-associated phenotypic class consisted of 665 genes differentially expressed in *dpy-22(sy622)* homozygotes and in *trans*-heterozygotes, but which had wild-type expression in *dpy-22(bx93)* homozygotes. The *dpy-22(bx93)*-associated phenotypic class contains 229 genes differentially expressed in all genotypes. The *dpy-22(bx93)*-associated class included re-classified transcripts that had been found to be differentially expressed in the *dpy-22(bx93)* homozygote and one other genotype, because these were very likely to be the result of false negative

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

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

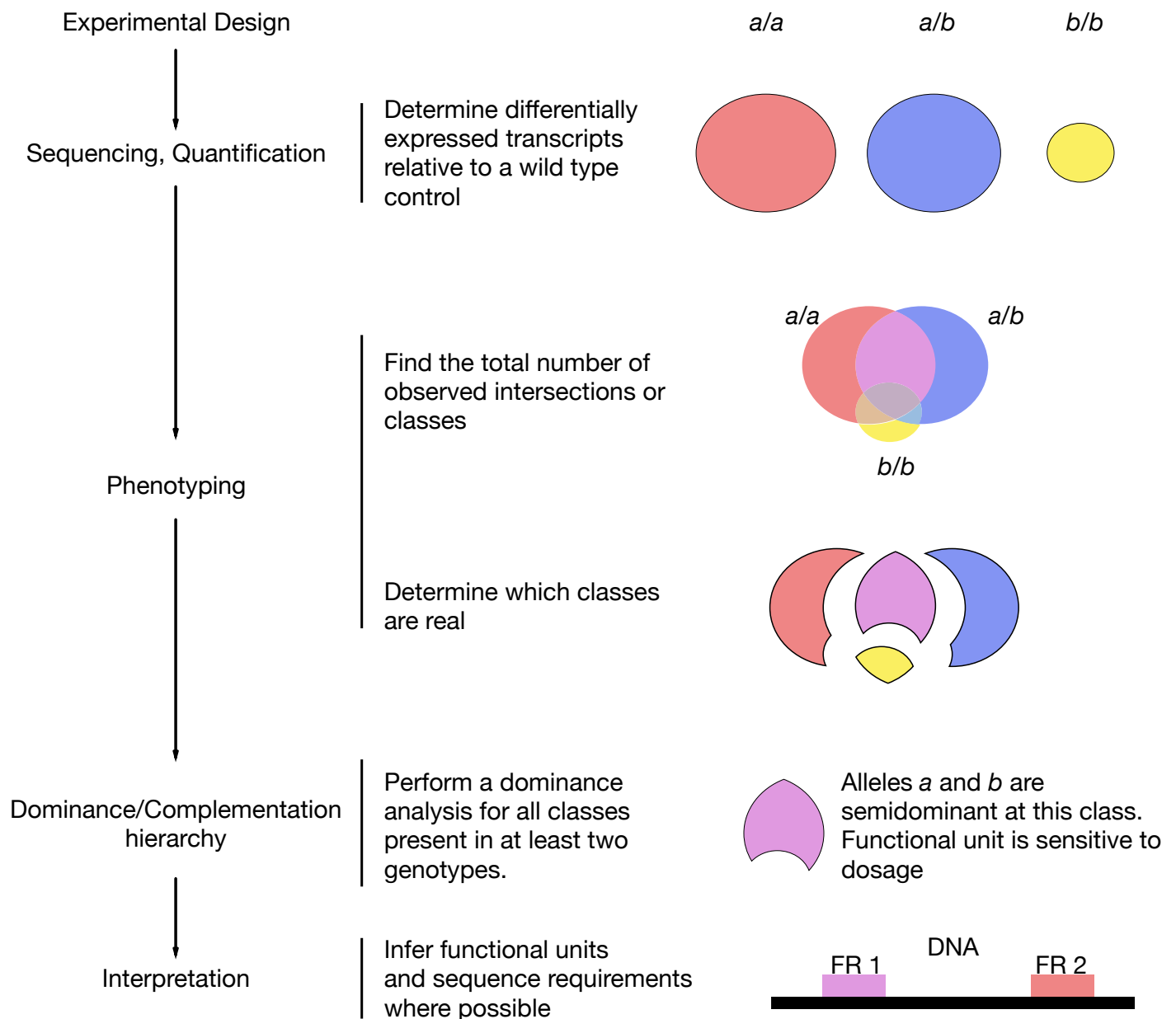


Figure 3. Flowchart for an analysis of arbitrary allelic series. A set of alleles is selected, and the corresponding genotypes are sequenced. Independent phenotypic classes are identified, and classes that are the result of noise are discarded via a false hit analysis. For each phenotypic class, the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

hits in the missing genotype, and re-classifying these transcripts improved our model substantially. We also identified a *dpy-22(sy622)*-specific phenotypic class (1,213 genes) and a *trans*-heterozygote-specific phenotypic class (1,302 genes; see the [Phenotypic Classes Notebook](#)).

Severity hierarchy of a *dpy-22* allelic series

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

The *dpy-22(sy622)* homozygote shows more differentially expressed genes that participate in a greater number of phenotypic classes relative to the *dpy-22(bx93)* homozygote. Thus, the *dpy-22(sy622)* allele is a more pleiotropic mutation than the *dpy-22(bx93)* allele. To assess which allele has more magnitude, Since the homozygotes of each allele only share a single phenotypic class in common, we need only assess magnitude along this single phenotype. To calculate a magnitude coefficient, for genes in the *dpy-22(bx93)*-associated phenotypic class, we plotted the β coefficients from the *dpy-22(sy622)* homozygote against the β coefficients from the *dpy-22(bx93)* homozygote (see Fig. 4) and performed a linear regression to find the slope of this line. Using this method, we found that the *dpy-22(bx93)* homozygote has a magnitude that is 62% \pm 2% of the *dpy-22(sy622)* homozygote. Taken together, these results suggest that the *dpy-22(sy622)* allele represents a more severe alteration-of-function mutation than the mutation within the *dpy-22(bx93)* allele.

Dominance hierarchy of a *dpy-22* allelic series

We measured allelic dominance for each class using a dominance coefficient (see [Methods](#)). The dominance coefficient is a measure of the contribution of each allele to the total expression level in *trans*-heterozygotes. By definition, the *dpy-22(sy622)* allele is completely recessive to *dpy-22(bx93)* for the *dpy-22(sy622)*-specific phenotypic class. To determine the dominance coefficient for the remaining phenotypic classes, we first selected the transcripts within those classes, and asked what linear combination of the homozygotic β coefficients best approximated the β coefficients of the *trans*-heterozygote, subject to the constraint that the sum of the weights for the two homozygotes should be equal to unity. We solved this problem by finding the maximum likelihood estimate for these weights. Using this method, we found that the *dpy-22(sy622)* and *dpy-22(bx93)* alleles are semidominant ($d_{bx93} = 0.48$) to each other for the *dpy-22(sy622)*-associated phenotypic class. The *dpy-22(bx93)* allele is largely dominant over the *dpy-22(sy622)* allele ($d_{bx93} = 0.82$; see Table 2) for the *dpy-22(bx93)*-associated phenotypic class.

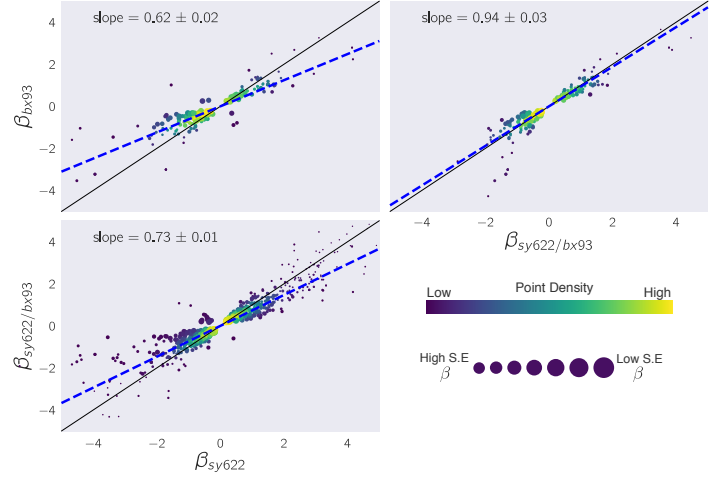


Figure 4. Shared Transcriptomic Phenotypes amongst the *dpy-22* genotypes are regulated in the same direction. For each pairwise comparison, we found those transcripts that were commonly differentially expressed in both genotypes relative to the wild-type control and plotted the β coefficients for each. We performed a linear regression on each plot to find the line of best fit (broken blue line). Only the comparison between *dpy-22(sy622)* and *dpy-22(bx93)* homozygotes was used to establish that the magnitude of the *dpy-22(sy622)* allele is greater than the magnitude of the *dpy-22(bx93)* allele. The other comparisons are shown for completeness.

Phenotypic Class	Dominance
<i>dpy-22(sy622)</i> -specific	1.00 \pm 0.00
<i>dpy-22(sy622)</i> -associated	0.48 \pm 0.01
<i>dpy-22(bx93)</i> -associated	0.82 \pm 0.01

Table 2. Dominance analysis for the *dpy-22/MDT12* allelic series. Dominance values closer to 1 indicate *dpy-22(bx93)* is dominant over *dpy-22(sy622)*, whereas 0 indicates *dpy-22(sy622)* is dominant over *dpy-22(bx93)*.

Phenotypic classes reflect morphological phenotypes

We performed enrichment analysis of anatomical, phenotypic and gene ontology terms using the WormBase Enrichment Suite (Angeles-Albores *et al.* 2016, 2018a). The *dpy-22(bx93)*-associated phenotypic class was enriched in genes involved in ‘immune system processes’ ($q < 10^{-5}$), and was enriched in genes expressed in the ‘intestine’ ($q < 10^{-4}$). The *dpy-22(sy622)*-associated class was enriched in genes expressed in the ‘cephalic sheath cell’ ($q < 10^{-4}$). Using ontology enrichment analysis from the WormBase Enrichment Suite, we found that the *dpy-22(sy622)*-associated class is enriched in histones and histone-like proteins (‘DNA packaging complex’ $q < 10^{-3}$) as well as genes involved in ‘immune system processes’ ($q < 10^{-5}$). The *dpy-22(sy622)*-specific class was enriched in genes that have expression in the ‘intestine’ ($q < 10^{-7}$), ‘muscular system’ ($q < 10^{-3}$) and ‘epithelial system’ ($q < 10^{-2}$). The genes in this class are known to cause bacterial lawn avoidance when knocked down or knocked out ($q < 10^{-2}$). Finally, GO enrichment showed that the *dpy-22(sy622)*-specific class is specifically enriched in ‘structural constituents of cuticle’ ($q < 10^{-12}$), and in genes involved in respiration ($q < 10^{-6}$). This last result recapitulates the fact that *dpy-22(sy622)* homozygotes show a severe Dumpy phenotype. The *trans*-heterozygote specific class was enriched in genes expressed in ‘male’ animals ($q < 10^{-63}$) and genes expressed in the ‘reproductive system’ ($q < 10^{-21}$). GO enrichment of genes in the *trans*-heterozygote specific class showed enrichment of the genes involved in the ‘regulation of cell shape’ ($q < 10^{-6}$) and in a variety of terms involving phosphate metabolism, such as ‘nucleoside phosphate binding’ ($q < 10^{-5}$), ‘dephosphorylation’ ($q < 10^{-3}$) or ‘phosphorylation’ ($q < 10^{-2}$), suggesting that this class may be enriched in genes involved in signal transduction though the reason for this enrichment remains unclear. The *dpy-22(bx93)*-specific class did not show enrichment on any test, consistent with our interpretation that this class is the result of random false positive hits.

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

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

The *dpy-22(bx93)*-specific class did not show a transcriptomic signature associated with either the Wnt or the Ras pathway, consistent with our interpretation of this class as false positive (Fig. 5). All other classes showed significant enrichment with genes perturbed in *bar-1(ga80)*. Similarly, *let-60(n2021)* showed enrichment in all real phenotypic classes, with the exception of the *trans*-heterozygote specific class. Contrary to our hypotheses, differentially expressed genes in *let-60(n1046gf)* did not show significant overlap with the *dpy-22(sy622)*-specific phenotype, but they did show significant overlap with all remaining real phenotypic classes.

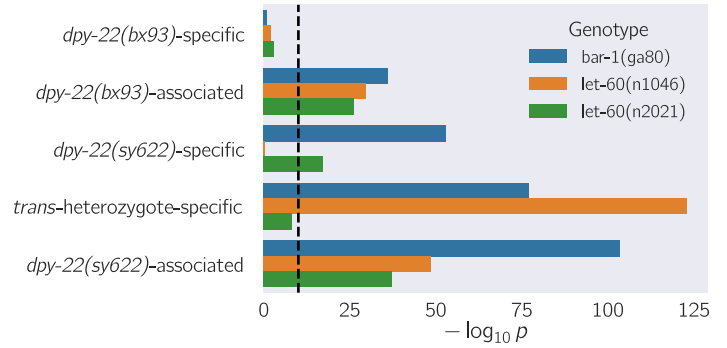


Figure 5. *dpy-22* phenotypic classes are statistically significantly enriched for signatures of *let-60* (ras) and *bar-1* (wnt) signaling. We tested whether the overlap between the differentially expressed genes in *bar-1(ga80)*, *let-60(n1046gf)* or *let-60(n2021)* and the *dpy-22* phenotypic classes was statistically significant using a hypergeometric enrichment test. Since the hypergeometric enrichment test is very sensitive to deviations from random, and since we suspect that there may be a broad genotoxic response to all mutants, we used a statistical significance threshold of $p < 10^{-10}$ (dashed black line).

Discussion

A conceptual framework for analyses of allelic series using transcriptomic phenotypes

Although transcriptomic phenotypes have been used for epistatic analyses (Dixit *et al.* 2016; Angeles-Albores *et al.* 2018b, 2017), they have not been used to study gene function in the context of an allelic series. Outstanding challenges for transcriptomes in allelic series were how to count or identify distinct phenotypes within the different transcriptomes, how to order alleles in a severity hierarchy and how to order alleles in a dominance hierarchy. In this work, we present solutions to these problems, and propose a set of unifying concepts that we believe will be useful for future analyses. We re-analyzed an allelic series of the Mediator subunit gene *dpy-22* that had been studied previously (Moghal and Sternberg 2003), recapitulating and extending previous results as a proof of principle for our methodology. In our results, we derived a set of methods that do not rely on the nature of the mutations. In the subsequent discussion, we use the fact that the mutations we used were truncations to derive further insights into the functional units present in this gene.

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

Phenotypic classes and their sequence requirements

Because the mutations we used are truncations, our results suggest the existence of various functional regions in *dpy-22/MDT12* (Fig. 6). These functional regions could encode protein domains with biochemical activity, or they could encode biochemically active amino acid motifs, such as nuclear localization sequences or protein binding sites. These functional

regions could confer stability to the protein, thereby regulating its levels. As a caveat, we note that we have interpreted the effects these mutations have in terms of their putative effects at the protein level. In the case of our alleles, the relevant homozygotes had wild-type *dpy-22* mRNA levels, suggesting that these mutations do not affect the stability of the mRNA.

The *dpy-22(sy622)*-specific phenotypic class is likely controlled by a single functional region, functional region 1 (FR1). Sequence necessary for wild-type FR1 functionality is encoded between amino acid positions 1 and 2,549, since this is the sequence that is intact in the *bx93* allele. We speculate that this functional region may be the reason that *bx93* is unable to complement the Muv phenotype of *sy622* in a sensitized *let-23* background, since *trans*-heterozygotes in this background exhibit a semidominant Muv phenotype. The *dpy-22(sy622)*-associated phenotypic class is likely controlled by a second functional region, functional region 2 (FR2), and some necessary sequences for wild-type function are encoded between amino acid positions 1,698 and 2,549, but additional sequence could lie between amino acids 1 and 1,698. It is unlikely that FR1 and FR2 are identical because their dominance behaviors are very different. The *dpy-22(bx93)* allele was largely dominant over the *dpy-22(sy622)* allele for the *dpy-22(bx93)*-associated class, but gene expression in this class was perturbed in both homozygotes. The perturbations were greater for *dpy-22(sy622)* homozygotes than for *dpy-22(bx93)* homozygotes. This behavior can be explained if the *dpy-22(bx93)*-associated class is controlled jointly by two distinct effectors, functional regions 3 and 4 (FR3, FR4, see Fig. 6). Such a model would propose that the sequences necessary for FR3 functionality are within the interval 1 and 2,549, and some sequences necessary for FR4 functionality are encoded between positions 2549 and 3499. This model explains how expression levels of the *bx93*-associated phenotypic class in the *trans*-heterozygote are complemented to the levels of the *bx93* homozygote, because FR3 is complemented in *trans*, but FR4 is defective. Thus, FR3 encodes a functionality that is not dosage-dependent. One possibility is that FR3 is equivalent to FR1 or FR2, and FR4 modifies activity of either of these regions at a subset of loci. A rigorous examination of this model will require studying many alleles that mutate the region between Q1689 and Q2549 using homozygotes and *trans*-heterozygotes.

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

Occam's razor

Transcriptomic phenotypes generate large amounts of differential gene expression data, so false positive and false negative rates can lead to spurious phenotypic classes whose putative biological significance is misleading. Such artifacts are particularly likely when a phenotypic class is small. Notably, errors of

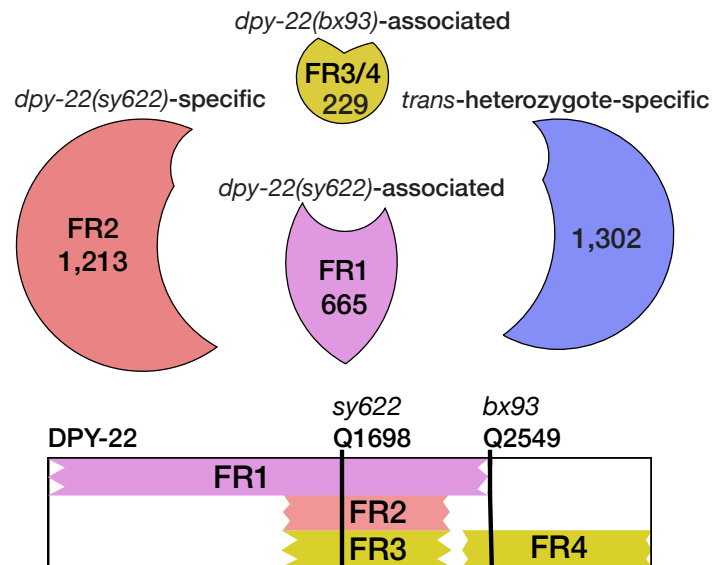


Figure 6. The functional regions associated with each phenotypic class can be mapped intragenically. The number of genes associated with each class is shown. The *dpy-22(bx93)*-associated class may be controlled by two functional regions. FR1 is a dosage-sensitive unit. FR2 and FR3 could be redundant if FR4 is a modifier of FR2 functionality at *dpy-22(bx93)*-associated loci. Note that the *dpy-22(bx93)*-associated phenotypic class is actually three classes merged together. Two of these classes are DE in *dpy-22(bx93)* homozygotes and one other genotype. Our analyses suggested that these two classes are likely the result of false negative hits and genes in these classes should be differentially expressed in all three genotypes, so we merged these three classes together (see Methods).

interpretation cannot be avoided by setting a more stringent q -value cut-off: doing so will decrease the false positive rate, but increase the false negative rate, which will in turn produce smaller phenotypic classes than expected. Our method tries to avoid this pitfall by using total error rate estimates to assess the plausibility of each class, though a major drawback is that it relies on a subjective estimation of the false negative rate. These conclusions are of broad significance to research where highly multiplexed measurements are compared to identify similarities and differences in the genome-wide behavior of a single variable under multiple conditions.

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

Acknowledgements

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

Literature Cited

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

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

Angeles-Albores, D., R. Y. Lee, J. Chan, and P. W. Sternberg, 2018a Two new functions in the WormBase Enrichment Suite. Micropublication: biology. Dataset. .

Angeles-Albores, D., D. H. W. Leighton, T. Tsou, T. H. Khaw, I. Antoshechkin, and P. W. Sternberg, 2017 The *Caenorhabditis elegans* Female State: Decoupling the Transcriptomic Effects of Aging and Sperm-Status. *G3: Genes, Genomes, Genetics* .

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

Angeles-Albores, D., C. Puckett Robinson, B. A. Williams, B. J. Wold, and P. W. Sternberg, 2018b Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements. *Proceedings of the National Academy of Sciences of the United States of America* **115**: E2930–E2939.

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

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

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

Brenner, S., 1974 The Genetics of *Caenorhabditis elegans*. *Genetics* **77**: 71–94.

Cao, J., J. S. Packer, V. Ramani, D. A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S. N. Furlan, F. J. Steemers, A. Adey, R. H. Waterston, C. Trapnell, and J. Shendure, 2017 Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (New York, N.Y.)* **357**: 661–667.

Deluca, D. S., J. Z. Levin, A. Sivachenko, T. Fennell, M. D. Nazaire, C. Williams, M. Reich, W. Winckler, and G. Getz, 2012 RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**: 1530–1532.

Dixit, A., O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev, 2016 Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**: 1853–1866.e17.

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

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

Elmlund, H., V. Baraznenok, M. Lindahl, C. O. Samuelsen, P. J. B. Koeck, S. Holmberg, H. Hebert, and C. M. Gustafsson, 2006 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.

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

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

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

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

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

Jeronimo, C. and F. Robert, 2017 The Mediator Complex: At the Nexus of RNA Polymerase II Transcription.

Kim, S., X. Xu, A. Hecht, and T. G. Boyer, 2006 Mediator is a transducer of Wnt/ β -catenin signaling. *Journal of Biological Chemistry* **281**: 14066–14075.

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

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

Lehner, B., A. Calixto, C. Crombie, J. Tischler, A. Fortunato, M. Chalfie, and A. Fraser, 2006 Loss of LIN-35, the

- Caenorhabditis elegans ortholog of the tumor suppressor p105Rb, results in enhanced RNA interference. *Genome Biology* **7**: R4.
- McKinney, W., 2011 pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing* pp. 1–9.
- Moghal, N. and P. W. Sternberg, 2003 A component of the transcriptional mediator complex inhibits RAS-dependent vulval fate specification in *C. elegans*. *Development* **130**: 57–69.
- Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621–628.
- Oliphant, T. E., 2007 SciPy: Open source scientific tools for Python. *Computing in Science and Engineering* **9**: 10–20.
- Pérez, F. and B. Granger, 2007 IPython: A System for Interactive Scientific Computing Python: An Open and General-Purpose Environment. *Computing in Science and Engineering* **9**: 21–29.
- Pimentel, H., N. L. Bray, S. Puente, P. Melsted, and L. Pachter, 2017 Differential analysis of RNA-seq incorporating quantification uncertainty. *brief communications nature methods* **14**.
- Riddle, D. L., T. Blumenthal, B. J. Meyer, and J. R. Priess, 1997 *C. elegans II*.
- Schwarz, E. M., M. Kato, and P. W. Sternberg, 2012 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.
- Takagi, Y. and R. D. Kornberg, 2006 Mediator as a general transcription factor. *The Journal of biological chemistry* **281**: 80–9.
- Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani, 2009 mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**: 377–382.
- Van Der Walt, S., S. C. Colbert, and G. Varoquaux, 2011 The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering* **13**: 22–30.
- Villani, A.-C., R. Satija, G. Reynolds, S. Sarkizova, K. Shekhar, J. Fletcher, M. Griesbeck, A. Butler, S. Zheng, S. Lazo, L. Jardine, D. Dixon, E. Stephenson, E. Nilsson, I. Grundberg, D. McDonald, A. Filby, W. Li, P. L. De Jager, O. Rozenblatt-Rosen, A. A. Lane, M. Haniffa, A. Regev, and N. Hacohen, 2017 Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**.
- Wang, J.-C., A. Walker, T. K. Blackwell, and K. R. Yamamoto, 2004 The *Caenorhabditis elegans* ortholog of TRAP240, CeTRAP240/let-19, selectively modulates gene expression and is essential for embryogenesis. *The Journal of biological chemistry* **279**: 29270–7.
- Waskom, M., S. St-Jean, C. Evans, J. Warmenhoven, K. Meyer, M. Martin, L. Rocher, P. Hobson, P. Bachant, T. Nagy, D. Wehner, O. Botvinnik, T. Megies, S. Lukauskas, Drewokane, E. Ziegler, T. Yarkoni, A. Miles, A. Lee, L. P. Coelho, Y. Halchenko, T. Augspurger, G. Hitz, J. Vanderplas, C. Fitzgerald, J. B. Cole, Gkunter, S. Villalba, S. Hoyer, and E. Quintero, 2016 seaborn: v0.7.0 (January 2016) .
- Yamamoto, T. and K. Shimojima, 2015 A novel MED12 mutation associated with non-specific X-linked intellectual disability. *Human Genome Variation* **2**: 15018.
- Yook, K., 2005 Complementation. *WormBook* .
- Zhang, H. and S. W. Emmons, 2000 A *C. elegans* mediator protein confers regulatory selectivity on lineage-specific expression of a transcription factor gene. *Genes and Development* **14**: 2161–2172.