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 13, 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

Mutations of a gene can yield a series of alleles with different phenotypes that reveal multiple functions encoded by that gene, regardless of the alleles' molecular nature. Homozygous alleles can be ordered by their phenotypic severity; tehn, phenotypes of *trans*-heterozygotes carrying two alleles can reveal which alleles are dominant for each phenotype. Together, the severity and dominance hierarchies show intragenic functional units. In *Caenorhabditis elegans*, these series have helped characterize genes such as *let-23/EGFR*, *lin-3/EGF* and *lin-12/NOTCH* [1,2,3].

Biology has moved from expression measurements of single genes towards genome-wide measurements. Expression profiling via RNA-seq [4] enables simultaneous measurement of transcript levels for all genes in a genome, yielding a transcriptome. These measurements can be made on whole organisms, isolated tissues, or on single cells [5,6]. Transcriptomes have been successfully used to identify new cell or organismal states [7,8]. For mutant genes, transcriptomic states can be used for epistasis analysis [9,10], but have not been used to characterize allelic series.

We have devised methods for characterizing allelic series with RNA-seq. To test these methods, we selected three alleles [11,12] of a *C. elegans* Mediator complex subunit gene, *mdt-12*. Mediator is a macromolecular complex with ~ 25 subunits [13] that globally regulates RNA polymerase II (Pol II) [14,15]. The Mediator complex has at least four biochemically distinct modules: the Head, Middle and Tail modules and a CDK-8-associated Kinase Module (CKM). The CKM associates reversibly with other modules, and appears to inhibit transcription [16,17]. In *C. elegans* development, the CKM promotes both male tail formation [11], (through interactions with the Wnt pathway), and vulval formation [18], (through inhibition of the Ras pathway). Homozygotes of allele *dpy-22(bx93)*, encoding a premature stop codon Q2549Amber [11], appear grossly wild-type. In contrast, animals homozyguous for a more severe allele, *dpy-22(sy622)* encoding another premature stop codon, Q1698Amber [12], are dumpy (Dpy), have egg-laying defects (Egl), and have multiple vulvae (Muv). Due to its pleiotropy, these alleles have not yet been ordered in a series (see Fig. 1A).

RNA-seq phenotypes have the potential to reveal functional units within genes, but their phenotypic complexity makes this difficult. We developed a method for determining allelic series from transcriptomic phenotypes and we used the *C. elegans mdt-12* gene as a test case. Our analysis revealed functional units that act to modulate Mediator activity at thousands of genetic loci.

Results and discussion

We adapted the allelic series method, previously used for individual phenotypes, for use with expression profiles as multidimensional phenotypes (see Fig. 1). As a proof of principle, we carried out RNA-seq on biological triplicates of mRNA extracted from *mdt-12(sy622)* homozygotes, *mdt-12(bx93)* homozygotes and wild type controls, and quadruplicates from *trans*-heterozygotes of both alleles at a depth of 20 million reads per sample. Reads were pseudoaligned using Kallisto[19]. We performed a differential expression using a general linear model specified in Sleuth[20] (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 $q \leq 0.1$.

By these criteria, we found 481 genes differentially expressed in *mdt-12(bx93)* homozygotes, and 2,863 differentially expressed genes in *mdt-12(sy622)* homozygotes (see Basic Statistics Notebook). We also sequenced *trans*-heterozygotes with the genotype *dpy-6(e14) mdt-12(bx93)/+ mdt-12(sy622)*, and found 2,214 differentially expressed genes.

We used a false hit analysis to identify four non-overlapping phenotypic classes. We use the term genotype-specific to refer to groups of transcripts that perturbed in one mutant. We use the term genotype-associated to refer to those groups of transcripts perturbed in two or more mutants. The ***mdt-12(sy622)*-associated** phenotypic class consisted of 720 genes differentially expressed in *mdt-12(sy622)* homozygotes and in *trans*-heterozygotes, but which had wild-type expression in *mdt-12(bx93)* homozygotes. The ***mdt-12(bx93)*-associated** phenotypic class contains 403 genes differentially expressed in all genotypes. We also identified a ***mdt-12(sy622)*-specific** phenotypic class (1,841 genes) and a ***trans*-heterozygote-specific** phenotypic class (1,226 genes; see the Phenotypic Classes Notebook).

Note: the bx93-associated class is actually 3 classes merged together. 2 of these classes aren't DE in all 3 genotypes (only 2 each), but my analyses strongly suggest this is the result of false negatives

We measured allelic dominance for each class. The *mdt-12(sy622)* allele is completely recessive to the *mdt-12(bx93)* for the *mdt-12(sy622)*-specific phenotypic class. The *mdt-12(sy622)* and *mdt-12(bx93)* alleles are semidominant ($d_{bx93} = 0.51$) to each other for the *mdt-12(sy622)*-associated phenotypic class. The *mdt-12(bx93)* allele is largely dominant over the *mdt-12(sy622)* allele ($d_{bx93} = 0.81$; see Table 1).

Our results suggest the existence of various functional units in *mdt-12/MDT12* (see Fig. 2). The *mdt-12(sy622)*-specific phenotypic class is likely controlled by a single functional unit, functional unit 1 (FC1), and the *mdt-12(sy622)*-associated phenotypic class is likely controlled by a second functional unit, functional unit 2 (FC2). It is unlikely that these units are identical because their dominance behaviors are very different. The *mdt-12(bx93)* allele was largely dominant over the *mdt-12(sy622)* allele for the *mdt-12(bx93)*-associated class, but gene expression in this class was perturbed in both homozygotes. The perturbations were greater for *mdt-12(sy622)* homozygotes than for *mdt-12(bx93)* homozygotes. This behavior can be explained if the *mdt-12(bx93)*-associated class is controlled jointly by two distinct effectors, functional units 3 and 4 (FC3, FC4, see Fig. 2). A rigorous examination of this model will require studying alleles that mutate the region between Q1689 and Q2549 using homozygotes and *trans*-heterozygotes.

Phenotypic Class	Dominance
mdt-12(sy622)-specific	1.00 ± 0.00
mdt-12(sy622)-associated	0.51 ± 0.01
mdt-12(bx93)-associated	0.81 ± 0.01

Table 1. Dominance analysis for the *mdt-12/MDT12* allelic series. Dominance values closer to 1 indicate *mdt-12(bx93)* is dominant over *mdt-12(sy622)*, whereas 0 indicates *mdt-12(sy622)* is dominant over *mdt-12(bx93)*.

We also found a class of transcripts that had perturbed levels in *trans*-heterozygotes only; its biologicla 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[21]. In the case of *mdt-12/MDT12* orthologs, how either mechanism could operate is not obvious, since the MDT-12 is expected to assemble in a monomeric manner into the CKM. Massive single-cell sequencing of *C. elegans* has recently been reported[22]. When this technique becomes cost-efficient, single-cell profiling of these genotypes may provide information that complements the whole-organism expression phenotypes, perhaps explaining the origin of this phenotype.

Transcriptomic phenotypes generate large amounts of differential gene expression data, so false positive and false negative rates can lead to spurious phenotypic classes whose putative biological significance is badly misleading. Such artifacts are particularly likely for small phenotypic classes, which should be viewed with skepticism. Notably, errors of interpretation cannot be avoided by setting a more stringent q-value cut-off; doing so will decrease the false positive rate, but increase the false negative rate, which will in turn produce smaller phenotypic classes than expected. Our method avoids this pitfall by using total error rate estimates to assess the plausibility of each class. These conclusions are of broad significance to

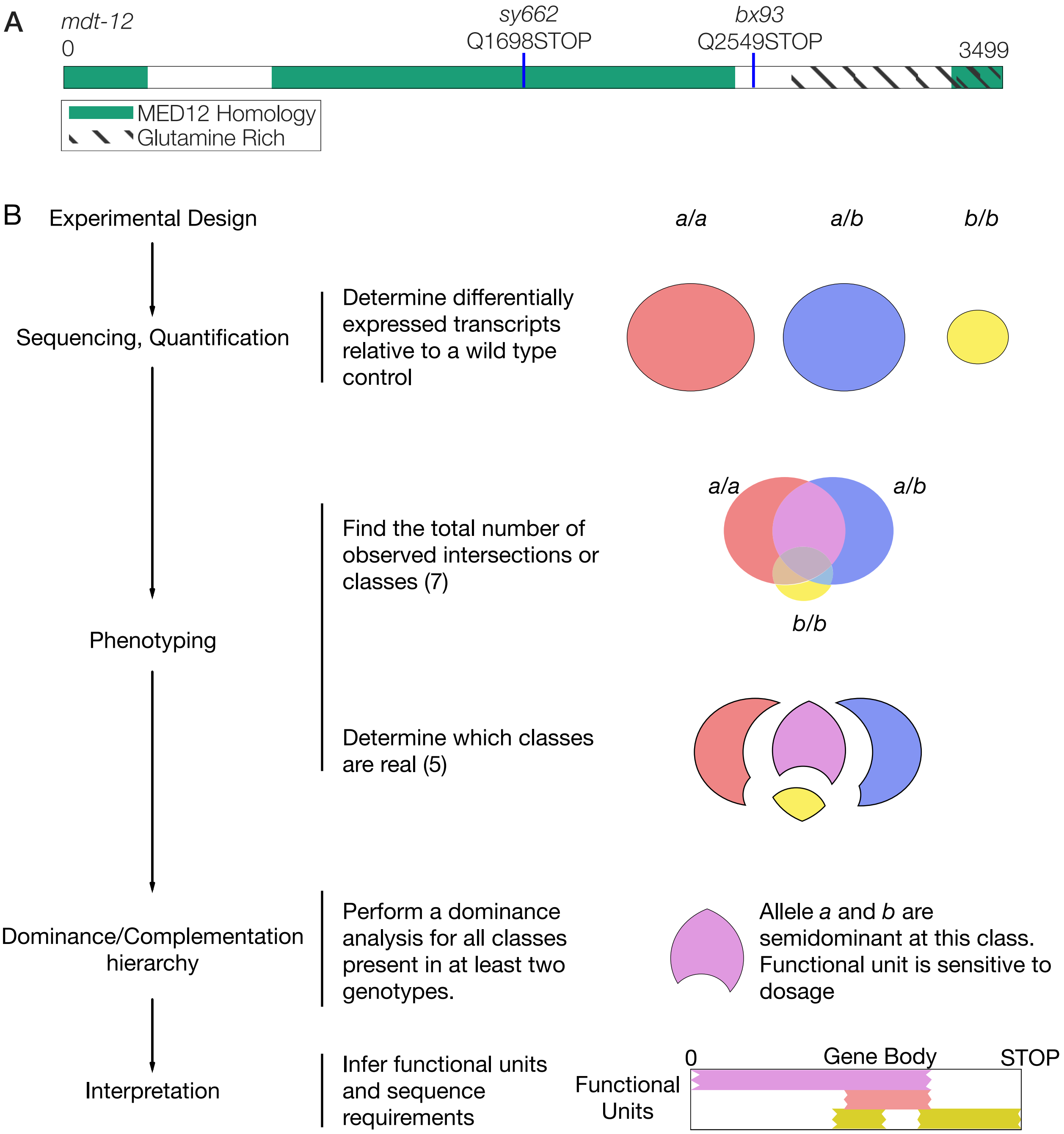
A
mdt-12
0
sy662
Q1698STOP
bx93
Q2549STOP
3499
MED12 Homology
Glutamine Rich
B
Experimental Design
a/a
a/b
b/b
Sequencing, Quantification
Determine differentially expressed transcripts relative to a wild type control
Find the total number of observed intersections or classes (7)
a/a
a/b
b/b
Phenotyping
Determine which classes are real (5)
Dominance/Complementation hierarchy
Perform a dominance analysis for all classes present in at least two genotypes.
Allele a and b are semidominant at this class. Functional unit is sensitive to dosage
Interpretation
Infer functional units and sequence requirements
0
Gene Body
STOP
Functional Units

Figure 1. **A** Protein sequence of *mdt-12*. The positions of the nonsense mutations used are shown. **B** Flowchart for an analysis of arbitrary allelic series. A set of alleles is selected, and the corresponding genotypes are sequenced. Independent phenotypic classes are then identified. For each phenotypic class, the alleles are ordered in a dominance/complementation hierarchy, which can then be used to infer functional units within the genes in question.

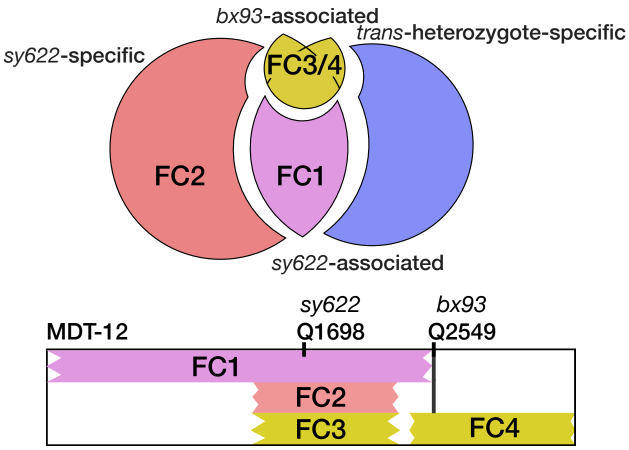

Figure 2. The functional units associated with each phenotypic class can be mapped to intragenic locations. The beginning and end positions of these functional units are unknown, so edges are drawn as ragged lines. Thick horizontal lines show the limit where each function could end, if known. We postulate that the *mdt-12(bx93)*-associated class is controlled by two functional units, FC3 and FC4, in the tail region of this gene. FC2 and FC3 may be redundant.

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 unit, and identified sequence requirements for each unit. Given the importance of allelic series for characterizing genetic pathways, we are optimistic that this method will be a useful addition to the geneticists arsenal.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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[19], 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[24,25,26,27]. All libraries had good quality scores.

Differential expression analysis was performed using Sleuth[20]. 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.

False hit analysis

To accurately count phenotypes, we developed a false hit algorithm (see Algorithm 1). We implemented this algorithm for three-way comparisons. We set our signal-to-noise threshold, α, to 4 after random benchmarking showed that this threshold performs well when false positive and negative rates are close to 10%. Using this threshold, the algorithm can correctly identify genotype-specific classes with 75% accuracy if the classes contain > 800 transcripts. Our method can identify when genotype-specific classes are empty with extremely high precision. Determining whether a genotype-specific class is real is most difficult when the class contains 50–500 transcripts. False negative rates for genotype-associated classes are extremely low using this method for classes with any number of transcripts. However, false positive rates are on the order of 30%.

```
Data: $\mathbf{M}_{obs} = \{N_l\}$, an observed set of classes,
      where each class is labelled by $l \in L$ and
      is of size $N_l$. $f_p, f_n$, the false positive and
      negative rates respectively. $\alpha$, the
      signal-to-noise threshold for acceptance
      of a class.
Result: $\mathbf{M}_{reduced}$, a reduced model that fits the
        data.
begin
    Define a minimal set to initialize the reduced model
    $\mathbf{K} = \{\min_{l \in L} N_l\}$
    Refine the model until the model converges or iterations max out
    $i \leftarrow 0$
    $\mathbf{K_{prev}} \leftarrow \emptyset$
    while $(i < i_{\max}) \mid (\mathbf{K_{prev}} \neq \mathbf{K})$ do
        $\mathbf{K_{prev}} \leftarrow \mathbf{K}$
        Define a noise function to estimate error flows in $\mathbf{K}$
        $\mathbf{F} \leftarrow \mathrm{noise}(\mathbf{K}, f_p, f_n)$
        for $l \in L$ do
            Calculate signal to noise for each labelled class
            False negatives can result in $\lambda < 0$
            $\lambda_l \leftarrow \mathbf{M}_{obs,l}/F_l$
            if $(\lambda > \alpha) \mid (\lambda < 0)$ then
                $\mathbf{K}_l \leftarrow \mathbf{M}_{obs,l}$
            end
        end
        $i++$
    end
end
Return the reduced model
return $\mathbf{K}$
```

Algorithm 1: False Hit Algorithm

Dominance analysis

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

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

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:

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

where $\beta_{a/b,i,\mathrm{Obs}}$ 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[28] using the Python programming language. The Numpy, pandas and scipy libraries were used for computation[29,30,31] and the matplotlib and seaborn libraries were used for data visualization[32,33]. Enrichment analyses were performed using the WormBase Enrichment Suite[34]. 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)[35] under the accession code GSE107523 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107523).

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

10. Angeles Albores, D., Puckett Robinson, C., Williams, B. A., Wold, B. J. & Sternberg, P. W. Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements. *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. Moghal, N. & Sternberg, P. W. A component of the transcriptional mediator complex inhibits RAS-dependent vulval fate specification in *C. elegans*. *Development* **130**, 57–69 (2003).

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

20. 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).

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

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

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

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

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

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

27. 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).

28. 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).

29. 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).

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

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

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

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

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

35. 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).