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 8, 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 within 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*, *lin-3* and *lin-12* [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 a 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 devised methods for characterizing allelic series with RNA-seq and we selected three alleles [11,12] of a *C. elegans* Mediator complex subunit gene, *mdt-12*, as a test for these methods. Mediator is a macromolecular complex with ~ 25 subunits [13] and which 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 the 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 the complexity of these phenotypes 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

We adapted the methodology of allelic series, which has been successfully used for scalar phenotypes, to be used in conjunction with expression profiles (see Fig. 1). As a proof of principle, 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. We calculated differential expression with respect to a wild-

Phenotypic Class	Dominance
sy622-specific	1.00 ± 0.00
sy622-associated	0.51 ± 0.01
bx93-associated	0.81 ± 0.01

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

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. Transcripts were considered to have differential expression between wild-type and a mutant if $q \leq 0.1$.

We found 481 genes differentially expressed in *bx93* homozygotes, and 2,863 differentially expressed genes in the *sy622* homozygotes (see Basic Statistics Notebook). We also sequenced *trans*-heterozygotic animals with genotype *dpy-6(e14) bx93/+ 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 allele- or genotype-specific to refer to groups of transcripts that are perturbed in a single genotype. We use the term allele-associated to refer to those groups of transcripts perturbed in at least two genotypes. The ***sy622*-associated** phenotypic class consisted of 720 genes differentially expressed in *sy622* homozygotes and in *trans*-heterozygotes, but which had wild-type expression in *bx93* homozygotes. The ***bx93*-associated** phenotypic class contains 403 genes differentially expressed in all genotypes. 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).

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

Discussion

Our results suggest the existence of various functional units in *mdt-12* (see Fig. 2). The *sy622*-specific phenotypic class is likely controlled by a single functional unit, functional unit 1 (FC1), and the *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 *bx93* allele was largely dominant over the *sy622* allele for the *bx93*-associated class, but gene expression in this class was perturbed in both homozygotes. The perturbations were greater for *sy622* homozygotes than for *bx93* homozygotes. This behavior can be explained if the *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 requires studying 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. This class contains 1226 genes, so is not a statistical artifact, though it could be a strain-specific artifact. If it is not artifactual, the biological meaning of this class is unclear. Phenotypes unique to *trans*-heterozygotes are often the result of physical interactions such as homodimerization, or dosage reduction of a toxic product[19]. In the case of *mdt-12* 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[20]. 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 information, so false positive and false negative events occur frequently enough to create artifactual transcript populations. Moreover, the distribution of false positive and false negative hits may not be uniform across phenotypic classes or their equivalent in other experimental designs. Quantifying signal-to-noise in phenotypic classes prevents 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. Small classes should be viewed with skepticism, particularly if the biological interpretation is implausible Notably, errors of interpretation cannot be avoided by setting a more stringent q-value cut-off. Lowering this cut-off decreases the false positive rate, but increases the false negative rate, leading to artifactual changes in class composition. This highlights the importance of our method, which estimates total error rates in assessing

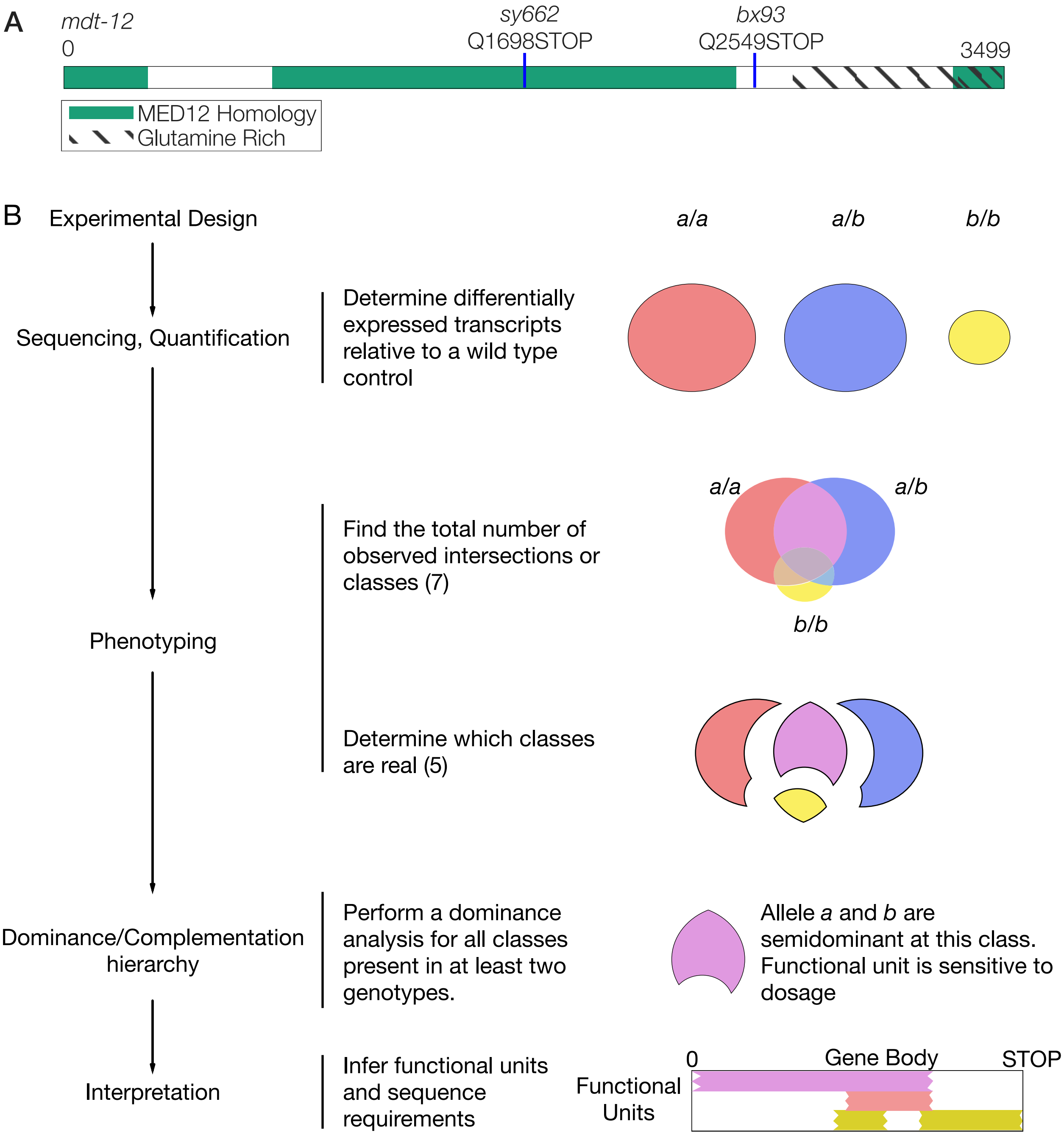
A
mdt-12
0
sy662
Q1698STOP
bx93
Q2549STOP
3499
MED12 Homology
Glutamine Rich
B
Experimental Design
a/a
a/b
b/b
Sequencing, Quantification
Determine differentially expressed transcripts relative to a wild type control
a/a
a/b
Find the total number of observed intersections or classes (7)
b/b
Phenotyping
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.

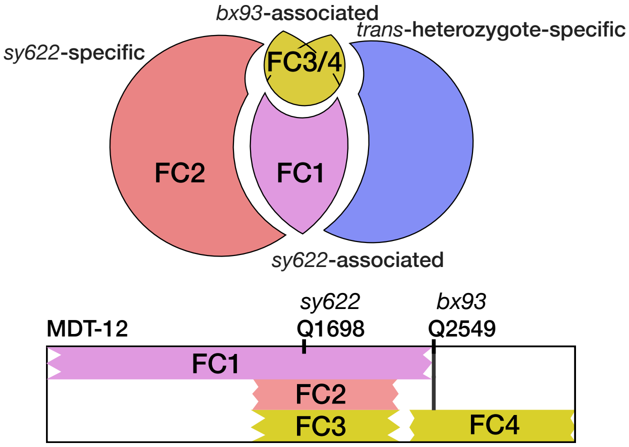

Figure 2. The functional units associated with each phenotypic class can be mapped to intragenic locations. The beginning and end positions of these functional units are unknown, so edges are drawn as ragged lines. Thick horizontal lines show the limit where each function could end, if known. We postulate that the *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.

the plausibility of each class. 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 difficult to identify using 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.

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

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