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 5, 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 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 epistasis analysis[9,10].

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]. *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 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], are dumpy (Dpy), have egg-laying defects (Egl) and a 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 framework to analyze allelic series using transcriptomic phenotypes and applied our methods to a series involving the MDT12 ortholog in *C. elegans*, *mdt-12*. Our analysis revealed a number of functional units that act to modulate Mediator activity at thousands of genetic loci.

Results

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. We adapted this methodology, 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-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 $q \leq 0.1$.

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 Basic Statistics Notebook). We also sequenced *trans*-heterozygotic animals with genotype *dpy-6(e14) bx93/+ sy622*. The *trans*-heterozygote transcriptome had 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 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, which contains 403 genes. We also identified a ***sy622*-specific** phenotypic class (1,841 genes) and a ***trans*-heterozygote-specific** phenotypic class (1,226 genes; see the Phenotypic Classes Notebook).

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*.

To dissect these alleles, establishment of a dominance hierarchy is required for each allele at each phenotypic class. 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. The *sy622* and *bx93* alleles are semidominant ($d_{bx93} = 0.51$) to each other within this phenotypic class. The *bx93* allele is largely but not completely 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* that control distinct phenotypic classes (see Fig. 2). 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 unlikely that these functional units are the same because the dominance behaviors are different among the two phenotypic classes.

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. One interpretation 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. 2). 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.

In our study, we found a class of transcripts that were exclusively differentially expressed in *trans*-

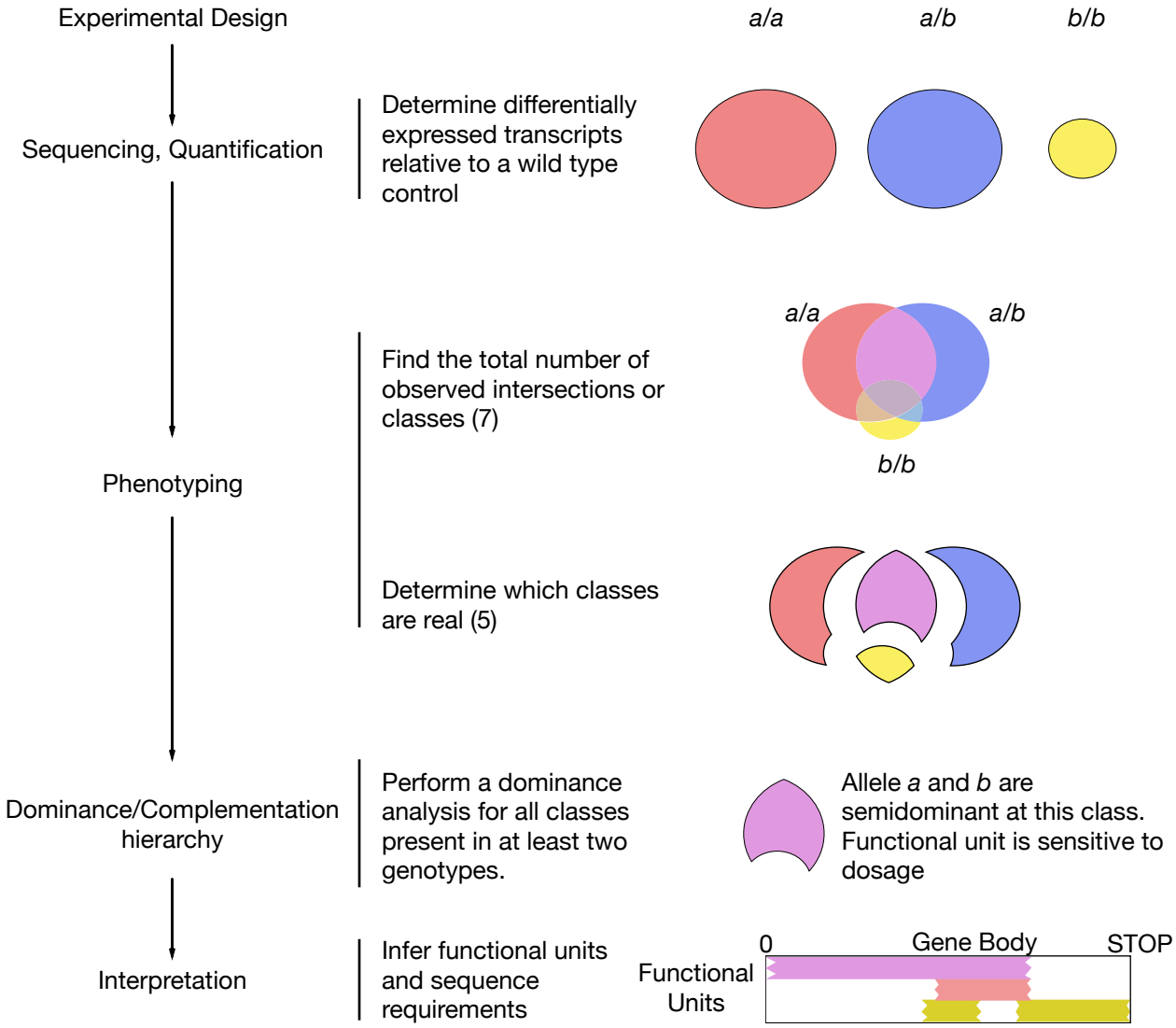

Figure 1. Flowchart for an analysis of arbitrary allelic series. A set of alleles is selected, and the corresponding genotypes are sequenced. Independent phenotypic classes are then identified. For each phenotypic class, the alleles are ordered in a dominance/complementation hierarchy, which can then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

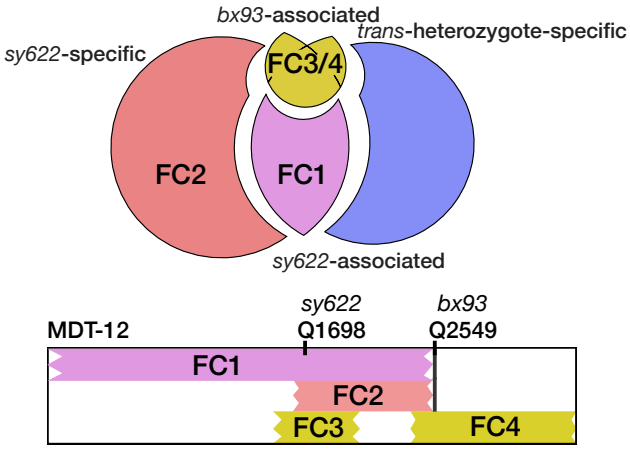

Figure 2. The functional units associated with each phenotypic class can be mapped to intragenic locations. The beginning and end positions of these functional units are unknown, so edges are drawn as ragged lines. Thick horizontal lines show the limit where each function could end, if known. We postulate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

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. If the *trans*-heterozygote specific class is not artifactual, its biological interpretation is still unclear. Phenotypes unique to *trans*-heterozygotes are often the result of physical interactions such as homodimerization, or dosage reduction of a toxic product[20]. 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. 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.

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 across phenotypic classes or their equivalent in other experimental designs. 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.

We have shown that transcriptomes can be used to study allelic series to partition the transcriptomic effects of a large, pleiotropic gene into separable phenotypic classes that would otherwise be difficult to identify using other methods. 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.

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. 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).