A study of allelic series using transcriptomic phenotypes

David Angeles-Albores[1,2] and Paul W. Sternberg[1,2,*]

[1] *Division of Biology and Biological Engineering, Caltech, Pasadena, CA, 91125, USA*
[2] *Howard Hughes Medical Institute, Caltech, Pasadena, CA, 91125, USA*
[*] *Corresponding author. Contact: pws@caltech.edu*

October 6, 2017

Expression profiling holds great promise for genetics due to its quantitative nature and the large number of genes that are measured. There is increasing interest in using these measurements as phenotypes for classical genetics analysis. Although transcriptomes have recently been used to perform epistasis analyses for pathway reconstruction, there has not been a systematic effort to understand whether different alleles have different transcriptomic qualities. Here, we study two allelic series using transcriptomic phenotypes. We studied two alleles of *dpy-22* that generate prematurely truncated proteins of different lengths. We show that expression perturbations caused by these alleles can be split into three distinct modules, and each module reacts with a different dominance relationship to each allele. Our work formalizes the concept of dominance for transcriptomic phenotypes, and shows the importance of studying allelic series for understanding the molecular qualities of the genes in question.

Author Summary

Expression profiling is a way to quickly and quantitatively measure the expression level of every gene in an organism. As a result, these profiles could be used as phenotypes with which to perform genetic analyses (i.e., to figure out what genes interact with each other) as well as to dissect the molecular properties of each gene. Before we can perform these analyses, we have to figure out the rules that apply to these measurements. In this paper, we develop new concepts and methods with which to study an allelic series. Briefly, allelic series are an important aspect of genetics because different alleles encode different versions of a gene. By studying these different versions, we can make statements about the function of different parts of the gene. By combining allelic series with expression profiling, we can learn much more about the gene under study than we could previously.

Introduction

The term 'allelic series' refers to the study of alleles with different phenotypes to understand the molecular properties that this locus controls. Allelic series are historically important for genetics. The earliest Pubmed-indexed author to use this term was Barbara McClintock[1]. In her work, McClintock studied a deficiency of the tail end of chromosome 9 of maize by generating *trans*-heterozygotes with mutants of various genes that she knew existed near the end of chromosome 9. Her work allowed her to infer that the deficiency was modular, effectively generating a double mutant that behaved as a single allele but which could participate phenotypically in two distinct allelic series. From this study, McClintock inferred that deletions could span multiple genes, which behaved as independent modules, and which were identified via complementation assays. This work set the foundations for later observations in yeast that showed two mutant alleles of the same genetic unit, when placed in *trans* to each other, could complement and generate a wild-type phenotype[2]. Allelic series have also been used to study the dose response curve of a phenotype for a particular gene. In *C. elegans*, the *let-23* allelic series stands out as such an example.

Over the last decade, biology has moved from studies of single genes towards studies of genome-wide measurements. In particular, expression profiling via RNA-sequencing[3] (RNA-seq) is a popular method because it enables the simultaneous measurement of expression levels for all genes in a genome. These measurements can now be made on a whole-organism scale and on single cells[4]. Although initially expression profiles had a qualitative purpose as descriptive methods to identify genes that are downstream of a perturbation, these profiles are now being used as phenotypes for genetic analysis. As a result, transcriptomes have been successfully used to identify new cell or organismal states[5,6]. Genetic pathways have been reconstructed via sequencing single cells[7] or by sequencing whole-organisms. However, to fully characterize a genetic pathway, it is often necessary to build allelic series to compare how phenotypes change with varying gene activity.

As a proof of principle, we selected a subunit of the Mediator complex in *C. elegans*, *dpy-22* (also known as *mdt-12*), for genetic analysis. Mediator is a macromolecular complex that contains 25 or so subunits[8] and which globally regulates RNA polymerase II (Pol II)[9,10]. Mediator is a versatile regulator, a quality often associated with its variable subunit composition[9], and it can promote transcription as well as inhibit it. The Mediator complex consists of four modules: the Head, Middle and Tail modules and a CDK-8-associated Kinase Module (CKM). The CKM can associate reversibly with Mediator. Certain models propose that the CKM functions as a molecular switch, which inhibits Pol II activity by sterically preventing its interaction with the other Mediator modules[11,12]. Other models propose that the CKM negatively modulates interactions between Mediator and enhancers[13]. In *C. elegans*, the CKM consists of CDK-8, MDT-13, CIC-1 and DPY-22[14]. Since *dpy-22* is orthologous to the human Mediator subunits *MED-12* and *MED-12L*[15], we will henceforth refer to this gene as *dpy-22* (*MED-12*).

dpy-22 in gene dosage sentence?

dpy-22 (*MED-12*) has been studied in the context of the male tail[15], where it was found to interact with the Wnt pathway. It has also been studied in the context of vulval formation[16], where it was found to be an inhibitor of the Ras pathway. *dpy-22* (*MED-12*) is likely an essential gene, and developmental studies have relied on reduction-of-function alleles to understand the role of this gene in development. Studies of the male tail were carried out using an allele, *dpy-22(bx93)*, that generates a truncated DPY-22 protein missing its C-terminal 900 or so amino acids as a result of a premature stop codon, Q2549STOP[15]. In spite of the premature truncation, animals carrying this allele grossly appear phenotypically wild-type. In contrast, the allele used to study the role of *dpy-22* (*MED-12*) in the vulva, *dpy-22(sy622)*, is a premature stop codon, Q1689STOP, that predicted to remove over 1,500 amino acids from the C-terminus[17]. Animals carrying this mutation are severely dumpy (Dpy), have egg-laying defects (Egl) and have a multivulva (Muv) phenotype that occurs at a very low rate (see Fig. 1). We wanted to study how truncations of increasing severity affected transcriptomic phenotypes. These alleles could form a single quantitative series, affecting the same sets of target genes but to different degrees, in which case the *trans*-heterozygote would exhibit a single dosage-dependent phenotype intermediate to the two homozygotes. Alternatively, they could form a single qualitative series, in which case the *trans*-heterozygote should have the same phenotype as the homozygote of the *bx93* allele, since this allele encodes the longer protein. These alleles could also form a mixed series, in which case multiple separable phenotypes would appear that have qualitative or quantitative behaviors in the *trans*-heterozygote.

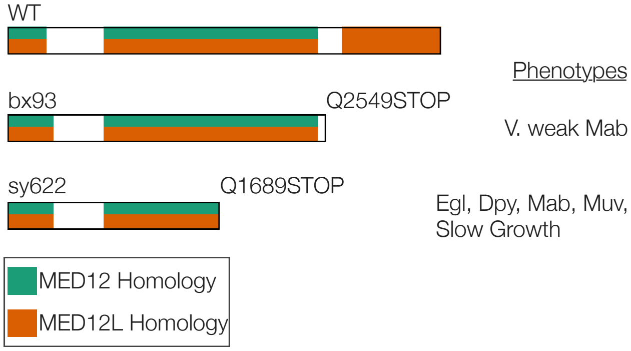

Figure 1. The *dpy-22* (*MED-12*) allelic series, consisting of two amino acid truncations, is amenable to study by transcriptomic phenotypes. Diagram of the *dpy-22* (*MED-12*) gene and the *bx93* and *sy622* alleles. Conservation between *dpy-22* (*MED-12*) and its human orthologs is shown in color.

Expression profiles have the potential to facilitate dissection of molecular structures within genes. To establish a methodology for studying allelic series, we explored three alleles (including the wild-type allele) of the highly pleiotropic gene, *dpy-22* (*MED-12*). For the *dpy-22* (*MED-12*) allelic series, we found that the perturbations caused by the weak loss-of-function allele, *bx93*, are entirely contained within the strong loss-of-function allele, *sy622*. Further, we found that there are three phenotypic classes that are affected by *dpy-22* (*MED-12*). For one class, termed the *sy622*-specific class, the *bx93* homozygote, but not the *sy622* homozygote, shows wild-type functionality. In a *trans*-heterozygote of *sy622/bx93* these genes are suppressed to wild-type levels from the *sy622* levels, which shows that *bx93* is wild-type dominant for this phenotype. A second class, called the *sy622*-associated class, similarly shows wild-type functionality in the *bx93* homozygote but not in the *sy622* homozygote, yet in the *trans*-heterozygote the expression levels of these genes is modulated in a gene-dosage dependent manner. Finally, we identified a third class, called the *bx93*-specific class, which contained genes that were altered in both homozygotes, but which showed an expression level most similar to the *bx93* homozygote, showing that *bx93* has a dominant mutant phenotype for this subset. For each class, we were able to quantitatively measure the dominance level of each allele.

Results

A strong and a weak loss-of-function *dpy-22* allele show different transcriptomic profiles

We sequenced in triplicate mRNA extracted from *sy622* homozygotes, *bx93* homozygote, a *trans*-heterozygote of both alleles and a wild-type control at a depth of 20 million reads. This allowed us to identify 21,954 protein-coding isoforms. We calculated differential expression with respect to a wild-type control using a general linear model (see Methods). Differential expression with respect to the wild-type control for each transcript i in a genotype g is measured via a coefficient $\beta_{g,i}$, which can be loosely interpreted as the natural logarithm of the fold-change. Positive β coefficients indicate up-regulation with respect to the wild-type, whereas negative coefficients indicate down-regulation. Transcripts were considered to have differential expression between wild-type and a mutant if the associated q-value of the β coefficient was less than 0.1.

Using these definitions, we found 434 differentially expressed genes in the *bx93* homozygote transcriptome, and 2,821 differentially expressed genes in the *sy622* homozygote transcriptome. The *trans*-heterozygote transcriptome had 2,930 differentially expressed genes.

The transcriptome of a *trans*-heterozygote of *dpy-22* identifies four phenotypic classes

We sequenced a *trans*-heterozygote of the *bx93* and *sy622* alleles with genotype *dpy-6(e14) bx93/+ sy622*. This *trans*-heterozygote appears phenotypically wild-type, resembling the *bx93* mutant morphologically[17]. Using the *trans*-heterozygote, we identified five non-overlapping phenotypic classes by what genotypes caused these genes to become differentially expressed (see Fig. 2). We called the set of genes that were differentially expressed only in the *bx93* homozygote relative to the wild type the *bx93*-specific phenotypic class. We do not analyze this class further due to its small size (43 genes; see Discussion). Next, we defined the set of 296 genes that were differentially expressed in the *bx93* homozygote and at least one other genotype as the *bx93*-associated phenotypic class. The *sy622*-associated phenotypic class, which consisted of 825 genes, was defined as the set of genes that were differentially expressed in the *sy622* homozygote and in the *trans*-heterozygote, but which did not already belong to the *bx93*-associated phenotypic class. The *sy622*-specific phenotypic class (1,477 genes) and the *trans*-heterozygote-specific phenotypic classes (1,536 genes) were defined as the sets of genes that were only differentially expressed in each genotype. Having defined these phenotypic classes, we set out to confirm whether each class actually behaved as an independent phenotypic module in an allelic series and whether each class could be interpreted biologically to shed light on the structure of *dpy-22* (*MED-12*).

Different phenotypic classes behave differently in an *sy622* homozygote

We asked whether these classes had perturbation distributions distinct from each other within a single homozygote. Specifically, in the context of the *sy622* homozygote, we wanted to know whether the *sy622*-specific, the *sy622*-associated and the *bx93*-associated phenotypes were different in the magnitude of their perturbations or whether these subsets behaved as if they had been randomly selected from the set of differentially expressed genes in the *sy622* homozygote, in which case the distributions of effects would be the same for all classes (see Fig. 3). We found that that the β coefficients of isoforms within the *bx93*-associated phenotype on average had the largest absolute value (mean 1.3). The *sy622*-associated phenotype had a smaller range of perturbations compared to the *bx93*-associated phenotype (95th percentiles of the two distributions: 3.6 versus 4.1, respectively), and a statistically smaller mean (1.1 vs 1.3, respectively, $p = 1.7 \cdot 10^{-5}$, non-parametric boostrap). The *sy622*-specific phenotype had the smallest mean of all (0.95, $p < 10^{-6}$ compared with *bx93*-associated phenotype).

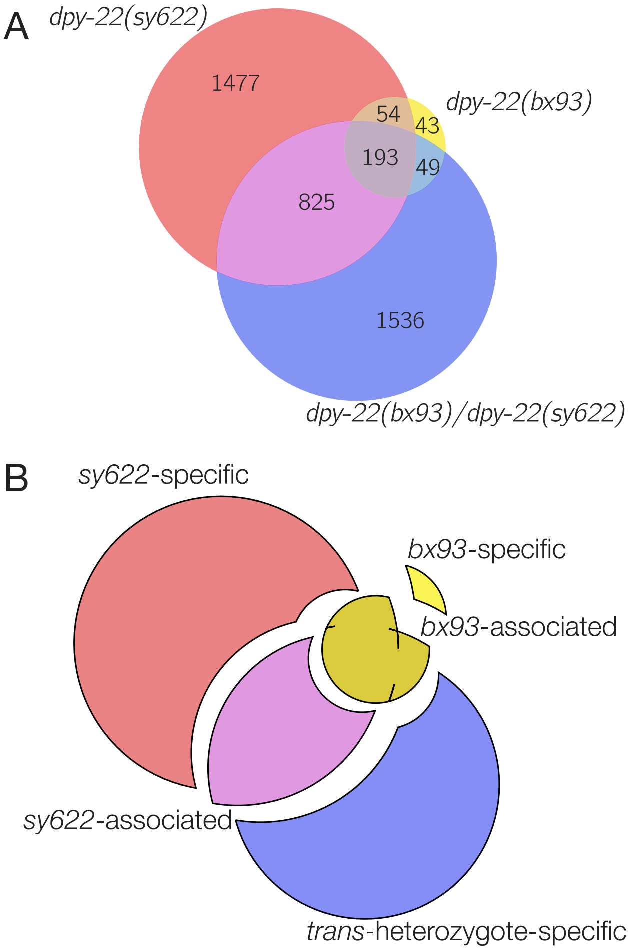

Figure 2. Transcripts under the control of *dpy-22* (*MED-12*) belong to distinct phenotypic classes. **A** Venn diagram showing number of genes in each subset. **B** Exploded Venn diagram highlighting the five identified phenotypic classes.

Dominance can be quantified in transcriptomic phenotypes

quantitative genetics citation?

We reasoned that if one allele was dominant over the other in the heterozygote, then plotting the β coefficients in the homozygote of the dominant allele versus the heterozygote should lead to a slope of 1. Deviations from a slope with magnitude equal to unity should therefore be interpreted as deviations from a standard dominant-recessive model. When expression in a *trans*-heterozygote is intermediate between the two homozygotes, this suggests a co-dominance regime where both alleles are contributing to the phenotype in a weighted fashion.

Dominance relationships between alleles are phenotype-specific. In other words, an allele can be dominant over another for one phenotype, yet not for others. An example is the *let-23* allelic series—nulls of *let-23* are recessive lethal (Let) and presumably also recessive vulvaless (Vul) relative to the wild-type allele. The *sy1* allele of *let-23* is viable dominant relative to null alleles, but is recessive Vul to the wild-type allele. Above, we postulated that there are four phenotypic classes, three of which are perturbed in the *sy622* homozygote. If these classes are indeed modular phenotypes, then the dominance relationships within each class should be the same from gene to gene. In other words, a single dominance coefficient should be sufficient to explain the gene expression in the *trans*-heterozygote for every gene within a class.

To quantify this dominance, we implemented and maximized a Bayesian model. Briefly, we asked what the linear combination of β coefficients from each homozygote would best predict the observed β values of the heterozygote, subject to the constraint that the coefficients added up to 1 (see Dominance analysis). We reasoned that if this was a modular phenotype controlled by a single structure encoded within the gene of interest, then a plot of the predicted β values from the optimized model against the observed β values of the heterozygote for each transcript should show the data falling along a line with slope equal to unity. Systematic deviations from linear behavior

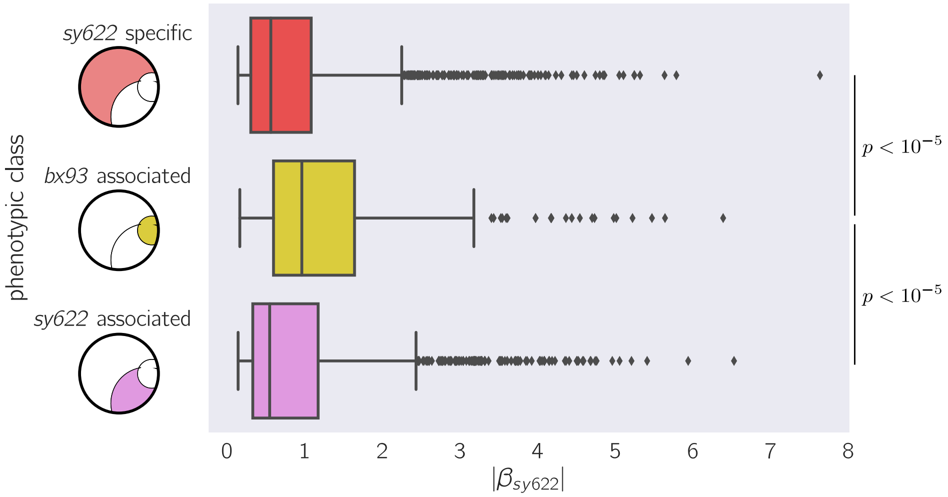

Figure 3. Within the *sy622* homozygote mutant, different phenotypic classes have statistically different distributions. The lines within the boxes show the 25, 50, and 75 percentiles. Whiskers show the rest of the plot, except for outliers (diamonds). Diagrams show what genotypes each gene class is expressed in, but the magnitude of the perturbation plotted always corresponds to the *sy622* mutant. The medians of the *sy622*-specific and the *sy622*-associated classes were statistically significantly different from the mean of the *bx93*-specific class, as assessed by a non-parametric bootstrap test.

would indicate that the transcripts plotted are not part of a modular phenotypic class controlled by a single structure.

The *sy622*-specific class expression phenotype of the *sy622* homozygote is complemented to wild-type levels by the presence of a *bx93* allele

Since our previous testing showed that the transcript expression of genes in this class was dysregulated in *sy622* homozygotes, and wild-type in both *bx93* homozygotes and *trans*-heterozygotes we can conclude that these transcripts are complemented to their wild-type levels by the presence of the *bx93* allele. Applying the Bayesian model yields identical results. Thus, there is a module that has wild-type functionality in the *bx93* allele but is partially or completely deleted in the *sy622* allele. This functionality must be encoded between amino acid position 1,689, where the *sy622* allele truncates prematurely, and the position 2,549 where the *bx93* allele stops.

The *bx93* allele is dominant over the *sy622* for the *bx93*-associated phenotype

We explored how expression levels of transcripts within the *bx93*-associated phenotypic class were controlled by these two alleles. We first applied our dominance analysis to transcripts in this class. We found that the *bx93* allele is largely dominant ($d_{bx93} = 0.82$) over the *sy622* allele (see Fig. 4). A large dominance coefficient might indicate that, for this phenotypic class, the *bx93* has a functional structure that is not present in the *sy622* allele. However, a significant portion of the transcripts within this class are differentially expressed in both homozygotes studied. Therefore, if this class is controlled by a single structure, then the functionality of this structure cannot be intact in the *bx93* homozygote. Moreover, when we compared the expression levels of transcripts in *sy622* and *bx93* homozygotes, we found that the *bx93* homozygotes had β coefficients that were on average 39% weaker than in the *sy622* homozygote. This implies that the two alleles should be codominant to each other, which is at odds with the dominance coefficient we observed. The mixed evidence precludes a conclusion about the structure/function relationship underlying this phenotypic class.

The *sy622*-associated phenotype is attenuated by the presence of *bx93* in the *trans*-heterozygote

We also wanted to know whether the *sy622*-associated phenotype showed differences depending on genotypic context. We quantified the relative dominance of *bx93* and *sy622* on the expression level of transcripts of this class. We found that both alleles are codominant ($d_{bx93} = 0.51$). This suggests that there is a structure distributed evenly throughout the gene body starting the first amino acid position and ending before position 2,549 (the site of the *bx93* truncation, otherwise these transcripts would be differentially expressed in *bx93* homozygotes). Since the two alleles are co-dominant for transcript expression in this class, the functionality encoded in this gene must be dosage-dependent for this model to hold.

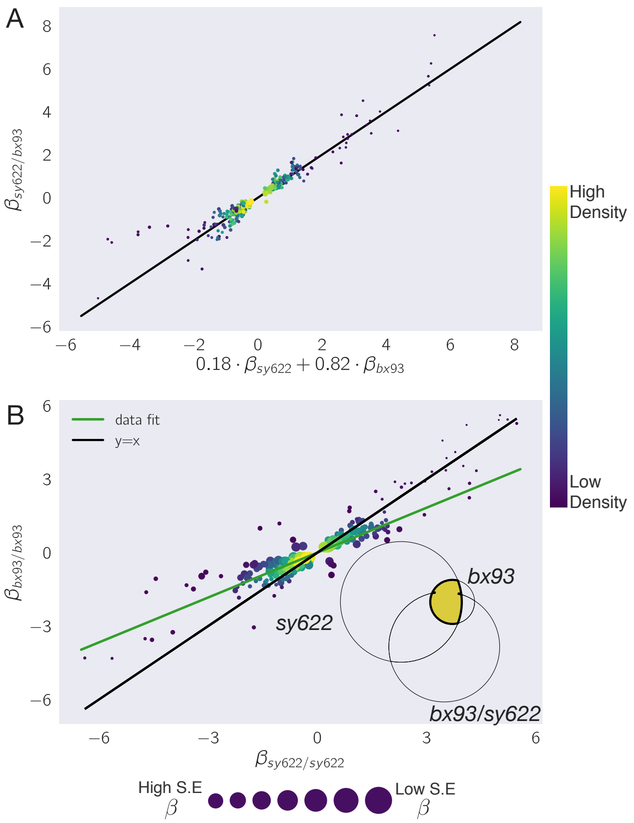

Figure 4. The *bx93*-associated class has properties of both quantitative and qualitative allelic series. **A** In a *trans*-heterozygote, the *bx93* allele is largely dominant over the *sy622* allele for the expression levels of transcripts in the *bx93*-associated class. **B** A majority of the transcripts in the *bx93*-associated class are differentially expressed in homozygotes of both alleles. In *bx93* homozygotes, these transcripts are less perturbed than in *sy622* homozygotes.

Ontology enrichment correlates transcriptomic phenotypic classes with morphological phenotypes

Whereas the *sy622* homozygote is visibly different from the wild type (see Fig. 1), the *bx93* is almost entirely wild-type. Since the *trans*-heterozygote also appears grossly wild-type, we hypothesized that the *sy622*-specific phenotypic class was associated with the macroscopic phenotypes visible in the *sy622* allele. We used the Wormbase Enrichment Suite[18,19] to query what anatomical, phenotypic or gene ontological terms were enriched in each phenotypic class.

The *bx93*-specific gene class returned no enriched anatomy, phenotype or gene ontology terms, consistent with our interpretation that this class plays a minor role in the biology of *dpy-22* (*MED-12*). The *trans*-heterozygote specific class was enriched in the anatomy terms ‘male’ (Enrichment Fold Change, $FC = 2$, $q < 10^{-40}$) and ‘reproductive system’ ($FC = 1.3$, $q < 10^{-19}$), but showed no phenotype enrichment. Gene enrichment for this category showed enrichment of many terms. Some of the top terms included ‘cell death’ ($FC = 3.9$, $q < 10^{-28}$) and ‘gene silencing by RNA’ ($FC = 5$, $q < 10^{-13}$).

The *sy622*-associated class showed enrichment in the intestine ($FC = 1.3$, $q < 10^{-3}$) . Phenotype enrichment analysis of this class showed no enrichment Gene ontology enrichment analysis identified terms such as ‘oviposition’ ($FC = 4.1$, $q < 10^{-6}$), ‘tube development’ ($FC = 9.1$, $q < 10^{-8}$) and ‘collagen trimer’ ($FC = 4.6$, $q < 10^{-4}$).

The *sy622*-specific class showed enrichment in the intestine ($FC = 1.4$, $q < 10^{-15}$), the muscular system ($FC = 1.3$, $q < 10^{-7}$), the epithelial system ($FC = 1.2$, $q < 10^{-3}$) and sex organs ($FC = 1.3$, $q < 10^{-3}$). This class had enrichment of terms associated with general sickness and pleiotropy, namely ‘avoids bacterial lawn’ ($FC = 1.9$, $q < 10^{-6}$), ‘gonad vesiculated’ ($FC = 2.0$, $q < 10^{-3}$) and ‘severe pleiotropic defects in the early embryo’ ($FC = 2.3$, $q < 10^{-3}$). Gene ontology enrichment analysis showed enrichment of many terms, including ‘respiratory chain’ ($FC = 7.3$, $q < 10^{-13}$), ‘muscle cell development’ ($FC = 6.6$, $q < 10^{-15}$), ‘oviposition ($FC = 2.6$, $q < 10^{-10}$) and ‘collagen trimer’ ($FC = 7.7$, $q < 10^{-21}$).

Our analyses indicate that the *sy622*-specific phenotypic class regulates a large number (39) of collagen genes that is not expected to be the result of random sampling. The expression of these genes is not homogeneous, and 19/40 isoforms are down-regulated and the rest are up-regulated. On the other hand, the *sy622*-associated class regulates 12 collagen genes, most of which increase in expression. A similar situation occurs with genes annotated as 'oviposition' genes. The *sy622*-specific class contains 38 oviposition genes, compared to 23 in the *sy622*-associated class. It is plausible that changes in these genes are behind the Dpy and Egl phenotypes of the *sy622* homozygote. In particular, the enrichments in the *sy622*-specific are the most extreme deviations from random sampling, as we hypothesized.

Discussion

Allelic series using transcriptomic phenotypes can dissect the molecular structure of a gene

We have shown that whole-organism transcriptomic phenotypes can be analyzed in the context of an allelic series to partition the transcriptomic effects of a large, pleiotropic gene into separable classes. Analysis of these modules can inform structure/function predictions at the molecular level, and enrichment analysis of each class can be subsequently correlated with observable phenotypes. This method shows promise for analysing pathways that have major effects on gene expression in an organism, and which do not have complex, antagonistic tissue-specific effects on expression. Given the importance of allelic series for fully characterizing genetic pathways, we are optimistic that this method will be a useful addition towards making full use of the potential of these molecular phenotypes.

A structure/function diagram of *dpy-22* (*MED-12*)

Our results strongly suggest the existence of two structures in *dpy-22* (*MED-12*) that control distinct phenotypic classes. The *sy622*-specific class retains wild-type functionality in the *bx93* allele, but this functionality is decreased in the *sy622* allele. Therefore, the function that controls this class must exist between amino-acid position 0 and position 2,549. A similar argument can be made for a structure that controls *sy622*-associated genes. For this argument to hold, however, the functionality associated with this structure must be dosage-dependent, since the *bx93* allele is codominant with the *sy622* allele, and this structure is likely intact in the *bx93* allele.

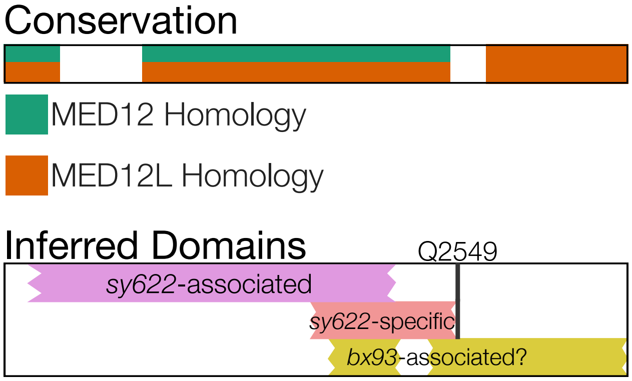

Figure 5. The modules associated with each phenotypic class can be mapped to intragenic locations. The beginning and end positions of these functions are unknown, so edges are drawn as ragged lines. Thick horizontal lines show the limit where each function could end, if known. We postulate that the *bx93*-associated function exists as two distinct modules in the tail region of this gene. Some of the modules shown may represent the same structures. Future experiments are required to make a complete determination of the number and nature of these modules.

Evidence in favor of a *bx93*-associated functionality was mixed. Although dominance analysis suggested that the *bx93* allele was dominant over the *sy622* allele for expression levels of genes in this class, the expression of these genes deviated from wild-type levels in both alleles. The latter suggests that the *bx93*-associated module is perturbed quantitatively in both genes, whereas dominance analyses favor an interpretation where the module is present in one allele but not in the other. One possibility is that the *bx93*-associated function we observed is the joint activity of two distinct effectors. In this model, one effector loses partial function in the *bx93* allele, whereas the second effector retains its complete activity. This leads to non-wild-type expression levels of the *bx93*-associated class of transcripts. In the *sy622* allele, both effectors are completely deleted, causing an increase in the severity of the observable phenotype. 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 whether how many modules exist in total, and how they may interact to drive gene expression.

Statistical artifacts associated with this analysis

Transcriptomic phenotypes generate large amounts of information that can be used to accurately determine molecular structures. However, due to the large number of tests performed, false positive and false negative events occur frequently enough to create populations of transcripts that have anomalous behaviors. It is necessary to identify what modules or populations are most at risk of these events and to what extent these modules may be polluted by false signals to prevent over-interpretation. In our experiment, we can identify two populations that are most at risk for statistical artifacts.

The *sy622*-associated class and the *bx93*-associated class are presented in our analysis as independent modules. A transcript that is differentially expressed in both *sy622* homozygotes and *trans*-heterozygotes is assigned to the *sy622*-associated class if and only if it is not differentially expressed in *bx93* homozygotes. If a transcript is falsely found to have wild-type expression in *bx93* homozygotes, then this transcript will be misclassified, and it will contribute signal to the *sy622*-associated class. Assuming a false negative rate of 10%, the number of transcripts that are mis-classified in this manner is approximately 34 (10% of 339 genes differentially expressed in *bx93* homozygotes). This constitutes almost 5% of the signal in the *sy622*-associated class. On the other hand, a transcript could be misclassified in the *bx93*-associated class in several ways. We enumerate the most likely events next. First, an *sy622*-associated transcript could be called as differentially expressed in *bx93* homozygotes. This event would contribute $\sim$ 19 genes (10% of 193 genes differentially expressed in all genotypes) to the *bx93*-associated class. Second, a transcript in the *sy622*-associated class could be falsely identified as differentially expressed in *bx93* homozygotes. This would be expected to contribute 5 transcripts. A similar event would also contribute 5 transcripts if the misclassification occurred in *trans*-heterozygotes. Therefore, we might expect that 29 transcripts are falsely contributing to the *bx93*-associated class, which constitutes $\sim$ 10% of the total signal. Therefore, the *bx93*-associated class is twice as vulnerable to statistical artifacts as the *sy622*-associated class. Moreover, most of the signal comes from transcripts that should have been classified in the *sy622*-associated class. Therefore, statistical noise will tend to make these two classes appear more similar than they really are. Fortunately, since both classes contained hundreds of genes and statistical contamination was less than 20%, our signal/noise ratio is able to resolve differences in the behaviors of these populations without trouble.

The phenotypic class that is most likely to be artifactual is the *bx93*-specific class. This class contains 43 genes. The expected number of transcripts that are falsely assigned as differentially expressed in *bx93* homozygotes is 33. The probability that such a false positive appears in any other genotype is approximately 20% (3,000 genes identified between the two other genotypes divided by 21,0000 the total number of genes that were successfully sequenced). Thus, the *bx93*-specific class on average contains 26 genes (80% of 33) that are false positives. False negative rates will also contribute to the *bx93*-specific class by moving genes from the *bx93*-associated class into the *bx93*-specific class. Such misassignments are expected to contribute $\sim$ 11 transcripts assuming a 10% false negative rate. In total, we estimate 37/43 genes in the *bx93*-specific class can be explained by sources of statistical artifacts, leading us to conclude that this phenotypic class does not exist and is simply the result of statistical noise.

The *trans*-heterozygote specific phenotypic class is not a statistical artifact

In our study, we found a large class of transcripts that were exclusively differentially expressed in *trans*-heterozygotes. The size of this class makes a statistical artifact unlikely. As a result, this class must be understood as either a legitimate aspect of *dpy-22* (*MED-12*) biology, reflecting antagonistic dosage-responsive tissue-specific effects, or as a strain-specific artifact. The genotype of the heterozygote includes a mutation at the *dpy-6* locus which acts as a balancer for the *bx93* mutation. One possibility is that the *dpy-6* loss-of-function mutation is not recessive for transcriptomic phenotypes and is responsible for the dysregulation of the new genes observed in the heterozygote. Another possibility is that the *dpy-6* strain had eQTLs that are affecting gene expression levels in a complex manner. As the cost of sequencing becomes lower, and with improved genetic engineering tools that allow the creation of background-free mutations, it will become increasingly important to rule out these hypotheses by sequencing additional independently derived identical alleles.

Transcriptomic genetics

Allelic series are a cornerstone of genetic analyses. Classically, these series have been important to understand multiple aspects of a gene by comparing and contrasting the properties of different alleles in homozygotes as well as heterozygotes. Due to their sensitivity and quantitative nature, transcriptomic phenotypes represent an exciting new phenotype with which to study these series. Here, we have shown that transcriptomic phenotypes can quickly and easily partition gene sets into phenotypic classes that have different statistical and physiological properties with minimal bioinformatic complexity. Expression profiles can be used for genetic pathway analysis[7] as well as for the identification of novel cellular or animal states[5,6]. In addition to sequencing various cell types to understand cellular diversity, we should sequence diverse alleles to understand genotype-genotype variation.

Methods

Strains used

Strains used were N2 wild-type (Bristol), PS4087 *dpy-22(sy622)*, PS4187 *dpy-22(bx93)*, and PS4176 *dpy-6(e14) dpy-22(bx93)/ + dpy-22(sy622)*. All lines were grown on standard nematode growth media (NGM) Petri plates seeded with OP50 *E. coli* at 20°C[20].

Strain synchronization, harvesting and RNA sequencing

All 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 allowed to grow to the young adult stage (as assessed by vulval morphology and lack of embryos). RNA extraction was performed as described in and sequenced using a previously described protocol[5].

Read pseudo-alignment and differential expression

Reads were pseudo-aligned to the *C. elegans* genome (WBcel235) using Kallisto[21], 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[22,23,24,25]. All libraries had good quality scores.

Differential expression analysis was performed using Sleuth[26]. Briefly, 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 wild-type replicates from other published and unpublished analysis. All wild-type replicates were collected at the same stage (young adult). In total, we had 10 wild-type replicates from 4 different batches, which heightened our statistical power. To account for batch effects, we added a batch correction term to our general linear model.

Non-parametric bootstrap

We performed non-parametric bootstrap testing to identify whether two distributions had the same mean. Briefly, the two datasets were mixed, and samples were selected at random with replacement from the mixed population into two new datasets. We calculated the difference in the means of these new datasets. We iterated this process 10^6 times. To calculate a p-value that the null hypothesis is true, we identified the number of times a difference in the means of the simulated populations was greater than or equal to the observed difference in the means of the real population. We divided this result by 10^6 to complete the calculation for a p-value. If an event where the difference in the simulated means was greater than the observed difference in the means was not observed, we reported the p-value as $p < 10^{-5}$. Otherwise, we reported the exact p-value. We chose to reject the null hypothesis that the means of the two datasets are equal to each other if $p < 0.05$.

Dominance analysis

We modeled allelic dominance as a weighted average of allelic activity. Briefly, our model proposed that β coefficients of the heterozygote, $\beta_{a/b,i,\text{Pred}}$, could be modeled as a linear combination of the coefficients of each homozygote:

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

where $\beta_{k/k,i}$ refers to the β value of the 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:

$$P(d_a|D,H,I) = \prod_{i\in S} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \frac{(\beta_{a/b,i,\mathrm{Obs}} - \beta_{a/b,i,\mathrm{Pred}}(d_a))^2}{2\sigma_i^2} \tag{2}$$

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

All code was written in Jupyter notebooks[27] using the Python programming language. The Numpy, pandas and scipy libraries were used for computation[28,29,30] and the matplotlib and seaborn libraries were used for data visualization[31,32]. Enrichment analyses were performed using the WormBase Enrichment Suite[18]. 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.

Acknowledgements

This work was supported by HHMI with whom PWS is an investigator and by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology. All strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). This article would not be possible without help from Dr. Igor Antoshechkin and Dr. Vijaya Kumar who performed the library preparation and sequencing.

References

1. McClintock, B. THE RELATION OF HOMOZYGOUS DEFICIENCIES TO MUTATIONS AND ALLELIC SERIES IN MAIZE. *Genetics* 478–502.

2. FINCHAM, J. R. S. & PATEMAN, J. A. Formation of an Enzyme through Complementary Action of Mutant 'Alleles' in Separate Nuclei in a Heterocaryon. *Nature* 741–742.

3. 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). URL http://dx.doi.org/10.1038/nmeth.1226{%}5Cnhttp://www.nature.com/nmeth/journal/v5/n7/suppinfo/nmeth.1226{_}S1.html{%}5Cnhttp://www.nature.com/doifinder/10.1038/nmeth.1226{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/18516045. 1111.6189v1.

4. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009). URL http://www.nature.com/doifinder/10.1038/nmeth.1315.

5. Angeles-Albores, D. *et al.* The Caenorhabditis elegans Female State: Decoupling the Transcriptomic Effects of Aging and Sperm-Status. *G3: Genes, Genomes, Genetics* (2017). URL http://www.g3journal.org/content/early/2017/07/26/g3.117.300080.

6. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* eaah4573.

7. 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). URL http://linkinghub.elsevier.com/retrieve/pii/S0092867416316105.

8. Jeronimo, C. & Robert, F. The Mediator Complex: At the Nexus of RNA Polymerase II Transcription (2017). URL http://www.sciencedirect.com/science/article/pii/S0962892417301162?via{%}3Dihub{#}bib0075.

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

10. Takagi, Y. & Kornberg, R. D. Mediator as a general transcription factor. *The Journal of biological chemistry* 80–9.

11. 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* 439–51.

12. 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* 15788–93.

13. van de Peppel, J. *et al.* Mediator Expression Profiling Epistasis Reveals a Signal Transduction Pathway with Antagonistic Submodules and Highly Specific Downstream Targets. *Molecular Cell* **19**, 511–522 (2005). URL http://linkinghub.elsevier.com/retrieve/pii/S1097276505014371.

14. 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* 2442–53.

15. 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).

16. 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).

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

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

19. Angeles-Albores, D., Lee, R. Y., Chan, J. & Sternberg, P. W. Phenotype and gene ontology enrichment as guides for disease modeling in *C. elegans* . *bioRxiv* .

20. Sulston, J. E. & Brenner, S. The DNA of Caenorhabditis elegans. *Genetics* **77**, 95–104 (1974).

21. Bray, N. L., Pimentel, H. J., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* 525–7. 1505.02710.

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

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

24. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Bowtie: An ultrafast memory-efficient short read aligner. [http://bowtie.cbcb.umd.edu/]. *Genome biology* R25.

25. 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).

26. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *brief communications nature methods* .

27. Pérez, F. & Granger, B. IPython: A System for Interactive Scientific Computing Python: An Open and General- Purpose Environment. *Computing in Science and Engineering* 21–29.

28. 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). 1102.1523.

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

30. Oliphant, T. E. SciPy: Open source scientific tools for Python. *Computing in Science and Engineering* 10–20.

31. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* **9**, 99–104 (2007). 0402594v3.

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