A study of allelic series using transcriptomic phenotypes

David Angeles-Albores

Paul W. Sternberg

# Abstract

**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 an allelic series of the *MED12* ortholog in *C. elegans*, *dpy-22*, using transcriptomic phenotypes. 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

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 . 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 made inferences about the role of large deletions in generating null mutants and the modifying effects of placing a loss-of-function mutation in *trans* to a deficiency or large deletion. In multiple senses, 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 . 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 . Although molecular null alleles are important for epistasis measurements, alleles with functional variations are useful to probe the genetic architecture at the single locus level.

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  (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 or for single cells . Although initially expression profiles had a qualitative purpose as descriptive methods to identify genes that are downstream of a perturbation, these profiles are actively being developed as phenotypes for genetic analysis. Transcriptomes have been successful in identifying new cell or organismal states . Finally, genetic pathways can be reconstructed by using single cell sequencing via clustering  and using whole-animal sequencing to measure transcriptome-wide epistasis between the null mutants of two genes . However, to fully characterize a genetic pathway, it is necessary to build allelic series to compare how phenotypes change with varying gene activity.

To explore the relationship between different transcriptomes associated with different alleles, we select an allelic series of the *dpy-22* gene in *C. elegans*. *dpy-22* is the TRAP-230/MED12 orthologue in *C. elegans* . Briefly, Mediator is a complex that globally regulates RNA polymerase II (Pol II) . Mediator is a versatile regulator, a quality often associated with its variable subunit composition , and it can promote transcription as well as inhibit it. The Mediator complex can be associated with four modules: head, middle and tail modules and a kinase module. In *C. elegans*, CDK8-associated kinase module (CKM) consists of *cdk-8*, *mdt-13*, *cic-1* and *dpy-22* (*MED12*) . The CKM is considered a molecular switch, which inhibits Pol II activity by sterically preventing interactions between Mediator and the polymerase . In *C. elegans*, *dpy-22* (*MED12*) has been studied primarily in the context of the male tail , where it was found to interact with the Wnt pathway, as well as vulval formation , where it was found to be an inhibitor of the Ras pathway. The null mutant of *dpy-22* (*MED12*) is lethal , so developmental studies have relied on reduction-of-function alleles to understand the role of this gene in morphogenesis. In particular, studies of the male tail were carried out using an allele, *dpy-22(bx93)*, that generates a truncated protein missing the terminal 900 or so amino acids . In spite of the premature truncation, animals carrying this allele appear phenotypically wild-type. In contrast, the allele used to study the role of *dpy-22* (*MED12*) in the vulva, *dpy-22(sy622)*, is a premature stop codon that removes more than 1,500 amino acids of the protein . 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. We wanted to study how truncations of increasing severity affected transcriptomic phenotypes. These alleles could form a single quantitative series, in which case the *trans*-heterozygote would exhibit a single dosage-dependent phenotype intermediate to the two homozygotes; 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. Alternatively, these alleles could form a mixed series, in which case multiple separable phenotypes would appear that have different behaviors in the *trans*-heterozygote. By using genome-wide transcriptional measurements, we hoped to gain insight into the structure/function relationships within this gene.

Expression profiles are quantitative measurements that hold great potential for dissecting the various molecular functions of a gene, but studying allelic series *de novo* is complicated by the lack of a theoretical framework with which to explore these new phenotypes. To establish a methodology for studying allelic series, we explored a *dpy-22* 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* (*MED12*). 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 over *bx93* 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 intermediate to the expression levels of either homozygote. Thus, genes in this class are responsive to dosage. 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. These findings challenge the way we think about alleles and their phenotypic classes at the transcriptional level and provide an example for how to quantitatively dissect allelic series on a transcriptome-wide level.

# Results

## *dpy-22* (*MED12*) conservation, revisited

*dpy-22* was previously identified as an ortholog of the human *MED12* gene . We revisited this orthology using DiOPT , which showed that *dpy-22* (*MED12*) is indeed the *C. elegans* ortholog of *MED12*. However, this search showed that, by sequence, *dpy-22* (*MED12*) has greater sequence conservation with the *MED12L* gene in humans. *MED12L* is a paralog of *MED12*, created after a duplication that occurred in vertebrate lineages . The function of *MED12L* is unknown, although it has been linked to human pathologies , and mass spectrometry reports suggest that can assemble into the Mediator module in a mutually exclusive manner to . We used PredictProtein  to further explore the molecular conservation of this gene. This revealed that *dpy-22* (*MED12*) has sequence conservation matching both *MED12* and *MED12L* at the N-terminal and middle regions of the protein (Eval=0 for both). However, the C-terminal region of *dpy-22* (*MED12*) only shows homology with *MED12L* (Eval=, see Figure [fig:dpy22\_cons]). These conservation results led us to re-examine the regions where the molecular lesions of the *bx93* and the *sy622* alleles fell. We found that the *bx93* allele cleanly deletes the C-terminal region with homology to *MED12L*. On the other hand, the *sy622* allele deletes a significant portion of the gene body that is conserved with both *MED12* and *MED12L* (see Figure [fig:dpy22\_cons]).

 dpy-22 (MED12) molecular conservation A. Diagram of the dpy-22 gene and the bx93 and sy622 alleles. 

dpy-22 (MED12) molecular conservation **A**. Diagram of the dpy-22 gene and the bx93 and sy622 alleles.

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

We studied two alleles of *dpy-22* (a Mediator subunit) that previous studies had suggested could be qualitatively distinct. Allele *bx93* (referred to as weak allele) encodes a premature stop codon that removes the terminal 900 amino acids from the protein. *bx93* homozygotic animals are phenotypically wild-type with a very low incidence of male tail defects . Allele *sy622* (referred to as the strong allele) encodes a premature stop codon that removes the terminal 1700 amino acids from the protein. *sy622* homozygotes grow slowly, are severely dumpy (Dpy), have a low penetrance multivulva (Muv) phenotype and have a prominent egg-laying defective (Egl) phenotype .

We sequenced homozygotes of both alleles and a *trans*-heterozygote of both alleles in triplicate and calculated differential expression with respect to a wild-type control. We found that *bx93* homozygotes expressed 434 differentially expressed genes, and *sy622* homozygotes showed 2,821 differentially expressed genes. Of the 434 differentially expressed genes in the *bx93* mutant, 73% were also differentially expressed in the *sy622* alleles, indicating that the impaired functionalities in the *bx93* homozygote are also impaired in the *sy622* homozygote. Having established that both alleles affect a shared subset of genes, we proceeded to measure whether the *sy622* allele showed greater perturbations in this subset. We observed that the weak allele, *bx93*, had perturbation magnitudes that were on average 39% weaker than the perturbation magnitudes in the strong allele, *sy622* (see Fig. [fig:dpy22]). In summary, the strong allele had more differentially expressed genes than the weak allele, and genes altered commonly in mutants of both alleles were more perturbed in the strong allele than in the weak allele. However, without analysis of the *trans*-heterozygote, it is impossible to conclude whether these differences are the result of greater reduction of function in the *sy622* allele relative to the *bx93* allele or whether the *sy622* allele has qualitative differences from *bx93* as a result of additional deleted functional domains.

 dpy-22 (MED12) is amenable to study by transcriptomic phenotypes. A. Venn diagram of the genotypes we sequenced: A bx93 homozygote, an sy622 homozygote and a bx93/sy622 trans-heterozygote. The size of each box is scaled to the number of differentially expressed genes identified in each genotype. Overlaps between boxes are quantitative, except for the smallest overlaps involving bx93 and sy622, and bx93 and bx93/sy622. B. Genes that are differentially expressed in both homozygotes relative to the wild-type control typically change in the same direction, and they tend to change by 30% less in the bx93 (weak allele) homozygote than in the sy622 (strong allele) homozygote. Inset shows the subset of genes plotted on the diagram. 

dpy-22 (MED12) is amenable to study by transcriptomic phenotypes. **A**. Venn diagram of the genotypes we sequenced: A bx93 homozygote, an sy622 homozygote and a bx93/sy622 trans-heterozygote. The size of each box is scaled to the number of differentially expressed genes identified in each genotype. Overlaps between boxes are quantitative, except for the smallest overlaps involving bx93 and sy622, and bx93 and bx93/sy622. **B**. Genes that are differentially expressed in both homozygotes relative to the wild-type control typically change in the same direction, and they tend to change by 30% less in the bx93 (weak allele) homozygote than in the sy622 (strong allele) homozygote. Inset shows the subset of genes plotted on the diagram.

## The *trans*-heterozygote of *dpy-22* strong and weak alleles allows the identification of four phenotypic classes

A standard method to identify whether two alleles differ quantitatively in their activity levels or whether they are qualitatively different because each allele has inactivated protein domains with separable functions is to generate a *trans*-heterozygote. Theoretically, if two alleles are quantitatively different, the *trans*-heterozygote will have a phenotype that is intermediate to the two homozygote phenotypes. On the other hand, if both alleles are inactivating distinct and separable functions of the protein, then the *trans*-heterozygote will exhibit a wild-type phenotype (intragenic complementation). Finally, if one allele is affecting multiple separable activities whereas the other allele is only affecting one, then the *trans*-heterozygote will exhibit the phenotype of the allele that affects the least number of activities (i.e., one allele will exhibit dominance).

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. The *trans*-heterozygote showed 2,930 differentially expressed genes. Using the trans-heterozygote, we were able identify four non-overlapping phenotypic classes by what genotypes caused these genes to become differentially expressed. One phenotypic class consisted of genes that were differentially expressed in the *sy622* homozygote as well as the trans-heterozygote, but not in the *bx93* homozygote (989 differentially expressed genes). We called this the *sy622*-associated phenotype. Another phenotypic class consisted of 1,623 genes that were only dysregulated in the *sy622* homozygote, which we called the *sy622*-specific phenotype because it is entirely suppressed by the presence of a single copy of the *bx93* allele. We also found a trans-heterozygote-specific phenotype consisting of 1,676 genes which is not present in either homozygote. The fourth phenotypic class, called the *bx93*-associated class, was defined as the set of genes dysregulated in both the *bx93* homozygote and the heterozygote, consisting of 310 differentially expressed genes. Although a *bx93*-specific phenotype technically exists, we do not consider it because we believe this class can be mostly explained in terms of false-positives and false-negatives (see discussion: ). Having defined these classes, we set out to describe their properties.

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 phenotypic classes had different perturbation distributions or whether these subsets behaved as if they had been randomly selected from the set of differentially expressed genes in the *sy622* homozygote (see Fig. [fig:classes]). We found that that the coefficients of isoforms within the *bx93*-associated phenotype on average had the largest absolute value (mean ). The *sy622*-associated phenotype had a smaller range of perturbations compared to the *bx93*-associated phenotype (95th percentiles of the two distributions: 3.3 versus 4.2, respectively), and a statistically smaller mean (1.3 vs 0.99, respectively, , non-parametric boostrap). The *sy622*-specific phenotype had the smallest mean of all (0.9, compared with *bx93*-associated phenotype, and compared with the *sy622*-associated phenotype, non-parametric bootstrap). The medians are almost identical between the *sy622*-specific and the *sy622*-associated phenotypes, which indicates that the small difference in the means of these two distributions is primarily driven by the longer tail of the *sy622*-associated phenotype. In conclusion, the *bx93*-associated phenotypic class contains those genes that respond most strongly to loss of function of .

 Within the sy622 homozygote mutant, different phenotypic classes have statistically different perturbation distributions. Genes that are sy622-specific have a different perturbation distribution compared to genes that are bx93-associated or sy622-associated. The lines within the boxes show the 25, 50, and 75 percentiles. Whiskers show the rest of the plot, except for outliers (diamonds). Insets show what genotypes each gene class is expressed in, but the magnitude of the perturbation plotted always corresponds to the sy622 mutant. The means of each distribution were all statistically different from each other, as assessed by a non-parametric bootstrap test. The sy622-specific and the sy622-associated distributions are very similar to each other, and the (small) difference in the means is the result of the heavier tail of the sy622-associated distribution. Notice that the x-axis, |\beta_{sy622}|, is in log-units. 

Within the sy622 homozygote mutant, different phenotypic classes have statistically different perturbation distributions. Genes that are sy622-specific have a different perturbation distribution compared to genes that are bx93-associated or sy622-associated. The lines within the boxes show the 25, 50, and 75 percentiles. Whiskers show the rest of the plot, except for outliers (diamonds). Insets show what genotypes each gene class is expressed in, but the magnitude of the perturbation plotted always corresponds to the sy622 mutant. The means of each distribution were all statistically different from each other, as assessed by a non-parametric bootstrap test. The sy622-specific and the sy622-associated distributions are very similar to each other, and the (small) difference in the means is the result of the heavier tail of the sy622-associated distribution. Notice that the x-axis, , is in log-units.

## Dominance can be quantified in transcriptomic phenotypes

We reasoned that if one allele was dominant over the other in the heterozygote, then plotting the 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. A classical 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.

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

We explored how expression levels changed within the *bx93*-associated phenotypic class between the homozygotes and the heterozygote. We selected the genes within the *bx93*-associated phenotypic class, and plotted the coefficients of these genes in the *bx93* allele against the coefficients in the heterozygote. The coefficients fell along a line with slope of 1.1, indicating that the *trans*-heterozygote has a strong resemblance to the *bx93* homozygote, although on average its phenotype is 10% worse (see SIXXX).

The close resemblance of the *bx93* levels to the trans-heterozygote levels suggested that the *bx93* is dominant over the *sy622* allele. To quantify this dominance, we implemented and maximized a Bayesian model. Briefly, we asked whether there was a linear combination of the coefficients of each homozygote that would predict the observed values of the heterozygote, subject to the constraint that the coefficients added up to 1. Our results suggested that the *bx93* allele was responsible for of the gene expression phenotypes of the trans-heterozygote. We wanted to explore how well this model explained the data. We reasoned that if this was a modular phenotype, then it should be possible to plot the predicted values against the observed values of the heterozygote using this coefficient. If the model fit well, we expected to observe a clearly linear relationship between both axes. In particular, we should not observe systematic deviations from this model. The plot revealed that the results fit remarkably well, furthering the case that the *bx93*-associated class indeed constitutes a modular phenotype (see Figure [fig:transhet]).

 For each phenotype, alleles have a single dominance behavior. A. Schematic explaining codominance. The closer the trans-heterozygote is to one of the homozygotes, the more dominant the allele corresponding to that homozygote is considered to be. Codominance is only valid when the heterozygote has a phenotype between the two alleles being studied. Dominance is phenotype-specific, so two alleles can share different dominance relationships for different phenotypes. B. The bx93 allele is dominant over sy622 for the expression level of genes that fall into the bx93-associated class. The bx93 is 80% dominant over the sy622 allele. C. The sy622 is co-dominant with the bx93 allele for the expression level of genes that fall into the sy622-associated class. Genes within this class show attenuated perturbations in the trans-heterozygote relative to the sy622 homozygote but do not show full complementation by the bx93 allele, even though the bx93 allele shows wild-type expression at these loci. X-axes are the predicted \beta values for each subset in heterozygote using the data from both homozygotes. The dominance coefficient was estimated via maximum likelihood estimates on the datasets. Y-axes show the observed \beta values for each subset in the heterozygote. Insets show the subset of genes plotted in each graph. 

For each phenotype, alleles have a single dominance behavior. **A**. Schematic explaining codominance. The closer the trans-heterozygote is to one of the homozygotes, the more dominant the allele corresponding to that homozygote is considered to be. Codominance is only valid when the heterozygote has a phenotype between the two alleles being studied. Dominance is phenotype-specific, so two alleles can share different dominance relationships for different phenotypes. **B**. The bx93 allele is dominant over sy622 for the expression level of genes that fall into the bx93-associated class. The bx93 is 80% dominant over the sy622 allele. **C**. The sy622 is co-dominant with the bx93 allele for the expression level of genes that fall into the sy622-associated class. Genes within this class show attenuated perturbations in the trans-heterozygote relative to the sy622 homozygote but do not show full complementation by the bx93 allele, even though the bx93 allele shows wild-type expression at these loci. X-axes are the predicted values for each subset in heterozygote using the data from both homozygotes. The dominance coefficient was estimated via maximum likelihood estimates on the datasets. Y-axes show the observed values for each subset in the heterozygote. Insets show the subset of genes plotted in each graph.

### 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. The *sy622*-associated genes are genes that are differentially expressed in the *sy622* homozygote or the trans-heterozygote, but not the *bx93* homozygote. Genes in this group showed a 23% reduction in the magnitude of their perturbations in the heterozygote compared to the homozygote (see SIXXX). Therefore, these genes are attenuated by the presence of a single copy of the *bx93* allele. To determine the relative dominance of *bx93* and *sy622*, we implemented the same model as above and found the coefficient that maximized the probability of observing the data. We found that *bx93* and *sy622* are almost perfectly codominant. *sy622* has a dominance coefficient of . This behavior is qualitatively different from genes in the *bx93*-associated phenotypic class, where *bx93* was 80% dominant. Finally, the behaviors of both of these classes are distinct from the behavior of genes in the *sy622*-specific class, which show differential expression in a *sy622* homozygote, but this dysregulation is complemented by the *bx93* allele (by definition, the dominance coefficient associated with *bx93* must be 1 for this class). This establishes that alleles can have differences in dominance for different phenotypic classes at the gene expression level.

## Insights into the gy of *sy622* homozygotes

Whereas the *sy622* homozygote is strongly phenotypic (see Fig. [fig:dpy22]), 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. To better understand this phenotypic class, we used the Wormbase Enrichment Suite  to query what anatomical, phenotypic or gene ontological terms were enriched in this gene set (see Table [tab:enrich]).

lcc Term & , *sy622*-specific & , *bx93*-associated  
Intestine & &   
Intestinal muscle & &   
PVD & &   
Muscular system & &   
pm3/5 & &   
Severe pleiotropic defects early embryo & &   
Rachis absent & &   
Meiosis defective early embryo & &   
Dauer constitutive & &  
Dauer metabolism & &  
Collagen trimers & &   
Muscle cell development & &   
Contractile fibers & &   
Oviposition & &   
Glucuronosyltransferase activity & &   
Monocarboxylic acid catabolic process & &   
Lytic vacuole & &

The *sy622*-specific phenotypic class was enriched for genes expressed in the intestine and intestinal muscle. We also found enrichment in cell-types that could reasonably be associated with egg-laying defects, namely the PVD neuron, and the muscular system. Phenotype ontology enrichment revealed that the *sy622*-specific phenotypic class was enriched for terms associated with embryonic lethality and small brood size, such as severe pleiotropic defects in the early embryo, oocytes lack nucleus, rachis absent and meiosis defective in the early embryo. Gene ontology enrichment showed that the *sy622*-specific phenotype was enriched in collagen trimers, muscle cell development, contractile fibers and oviposition. In contrast, the *bx93*-specific phenotypic class does not enrich identical terms. Rather, the *bx93*-specific class shows enrichment for genes expressed in the intestine and pharyngeal muscle cells, pm3 and pm5. It shows enrichment of genes associated with dauer constitutive and dauer metabolism phenotypes, and the gene ontology enrichment primarily reflects terms associated with metabolism, such as glucuronosyltransferase activity, monocarboxylic acid catabolic process, and lytic vacuole.

# Conclusions

## Loss-of-function allelic series reveal unknown functionality

Our sequencing results demonstrate that *sy622*, an allele that truncates 1700 amino acids from , has a more severe phenotype than *bx93* when assayed transcriptomically. This worsening manifests as an increase in the number of differentially expressed genes in *sy622* relative to wild-type compared to the number of differentially expressed genes in *bx93*. Moreover, the genes that are commonly dysregulated in both alleles show greater perturbations on average in the *sy622* homozygote relative to the *bx93* allele (see Fig. [fig:dpy22]). Notably, the set of genes differentially expressed in the *bx93* is contained within the *sy622* with few exceptions. We can account for most of these exceptions by invoking a 10% false positive rate (the cutoff to declare differential expression in our study) and a similar false negative rate in all samples. Thus, it seems reasonable to state that the set of genes affected by *bx93* is a subset of the set of genes affected by *sy622*. It follows that this subset is biologically equivalent to the *bx93*-specific phenotypic class. Ignoring for the moment the gene regulatory changes unique to the *trans*-heterozygote (which we will return to momentarily, see ), we can idealize the Venn diagram of gene expression changes in each genotype, which in turn will allow us to infer structure/function relationships within this complex gene (see Fig. [fig:inferred\_domains]).

 The loss-of-function allelic series of dpy-22 (MED12) allows us to infer the existence of functional domains within this complex gene. A. Idealized Venn diagram of the genotypes studied here. Notice that the trans-heterozygote specific phenotypic class is missing, since there is not a clear biological explanation for it (see  for a detailed discussion). B. Conservation of dpy-22 (MED12) amino acid sequence and the functional domains identified in this study. The bx93-specific domain corresponds with a region that has homology to but not to . 

The loss-of-function allelic series of dpy-22 (MED12) allows us to infer the existence of functional domains within this complex gene. **A**. Idealized Venn diagram of the genotypes studied here. Notice that the trans-heterozygote specific phenotypic class is missing, since there is not a clear biological explanation for it (see  for a detailed discussion). **B**. Conservation of dpy-22 (MED12) amino acid sequence and the functional domains identified in this study. The bx93-specific domain corresponds with a region that has homology to but not to .

We used the overlap between genotypes to define phenotypic classes—these classes in turn could be compared within a genotype or between them to establish the extent to which one allele was dominant over another at each class. This combined analysis showed that the *bx93* allele is almost completely dominant for the *bx93*-specific phenotypic class, which suggests that the terminal 900 or so amino acids encode a functional domain that helps to target *dpy-22* (*MED12*) to the genes contained within this class. Similarly, we found a second class of genes, termed the *sy622*-specific phenotypic class for which the *bx93* homozygote exhibited dominant wild-type expression, the *sy622* homozygote exhibited dysregulated expression, which was recessive to the *bx93* allele. This allows us to conclude that between Q1689 and Q2549 there is a functional domain that is intact in the *bx93* allele and deleted (partially or entirely) in the *sy622* allele. Finally, we found that there is a third phenotypic class, the *sy622*-associated phenotypic class, for which both alleles are codominant. This suggests that, for genes within this class, the deletions in both alleles are quantitatively affecting a single functionality within *dpy-22* (*MED12*), either by lowering the stability of the protein or by removing it from its site of action (the *sy622* allele is reported to fail to translocate to the nucleus) .

## The *trans*-heterozygote specific phenotypic class

Intriguingly, we found a phenotypic class specifically altered in the trans-heterozygote that was not present in either homozygote. A weakness in our study is that we have limited power to study the gene expression changes associated with the trans-heterozygote. This phenotypic class is puzzling because *dpy-22* (*MED12*) is not known to have homotypic interactions, which are a classical explanation for trans-heterozygote-specific phenotypes. Moreover, since this class is specific to a single strain we cannot rule out that this class is actually a result of a strain-specific mutation or set of mutations. In particular, the genotype of the heterozygote includes a mutation at the *dpy-6* locus to balance 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 a carrier mutation that fixed in the balanced strain. Finally, it is also possible that the *bx93* and the *sy622* strains had background mutations that did not have effects on their own, but when combined generate a synthetic phenotype with themselves or with one or both of the *dpy-22* alleles. Although we cannot definitively pinpoint the origin of the *trans*-heterozygote-specific phenotypic class, the other phenotypic classes are unlikely to be the result of mutational background since both alleles came from different screens carried out in different laboratories at different times.

In a complete genetic analysis, which is beyond the scope of this paper, the above possibilities should be rigorously tested. To rule out background as the cause of the *trans*-heterozygote phenotypic class, the alleles should be regenerated using a genome engineering tool such as Cas9  and the trans-heterozygote re-sequenced. Alternatively, more alleles coding for similar molecular lesions should be sequenced along with the respective heterozygote. To rule out dominant effects from the *dpy-6* locus, the locus could be restored to a wild-type status using standard co-conversion Cas9 techniques . Taken together, these experiments would help establish whether the *trans*-heterozygote phenotypic class is a result of strain background or not. 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.

## Physiological effects of deleting *dpy-22* (*MED12*)

Our enrichment analysis of the *sy622*-specific and the *bx93*-associated phenotypic classes revealed that they reflect functionally distinct aspects of *dpy-22* (*MED12*) biology. The *sy622*-specific class contains genes that are associated with severe pleiotropic effects, embryonic lethality and sterility. It also contains genes that are associated with muscle development and function, and there is enrichment of genes expressed in the PVD neuron. Collagen trimers are also overrepresented in this gene class. Taken together, these terms suggest that perturbing this gene class away from the wild-type should lead to an animal that is sickly, has a small brood size and has altered locomotion as well as altered collagen production. Indeed, the *sy622* homozygote is Dpy and Egl and RNAi against *dpy-22* (*MED12*) is known to cause embryonic and larval lethality . The *bx93*-associated class is enriched in a different set of terms, which suggests that transcriptome profiling can be used in conjunction with allelic series to separate genes into distinct phenotypic classes that are biologically relevant.

## Genetics in multi-dimensional phenotypes

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. Recents developments in the fields of transcriptomics have shown that expression profiles can be used for genetic pathway analysis  as well as for the identification of novel cellular or animal states . In particular, single-cell sequencing has shown great potential as a tool because it can help understand transcriptional heterogeneity at the cellular level, but also because random screens can be used to simultaneously knock out random combinations of genes and infer genetic interactions . Our work shows the importance of understanding allelic diversity towards understanding distinct biological properties of the genes in question. In addition to sequencing great numbers of cells to understand cell-cell heterogeneity and diversity, we should also sequence diverse alleles to better understand genotype-genotype heterogeneity.

# Methods

## Strains used

Strains used were N2 wild-type (Bristol), PS4087 *dpy-22(sy622)*, PS4187 *dpy-22(bx93)*, PS4176  
*dpy-6(e14) dpy-22(bx93)/ + dpy-22(sy622)*, MT4866 *let-60(n2021)*, and MT2124 *let-60(n1046gf)*. All lines were grown on standard nematode growth media (NGM) Petri plates seeded with OP50 *E. coli* at 20C .

## Strain synchronization, harvesting and RNA sequencing

With the exception of *let-60(lf)*, all strains were synchronized by bleaching P’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 20C and allowed to grow to the young adult stage (as assessed by vulval morphology and lack of embryos). The *let-60(lf)* strain was discovered to have a severe arrest phenotype that caused a 96% lethality rate after 12 hours of starvation. *let-60(lf)* worms were bleached once, then adults were selected at the young adult stage (assessed by timing and visual inspection of adult alae). RNA extraction and sequencing was performed as previously described .

## Read pseudo-alignment and differential expression

Reads were pseudo-aligned using Kallisto , 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 . All libraries had good quality scores.

Differential expression analysis was performed using Sleuth . 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 greatly 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 times. To calculate a -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 to complete the calculation for a -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 -value as . Otherwise, we reported the exact -value. We chose to reject the null hypothesis that the means of the two datasets are equal to each other if .

## Dominance analysis

We modeled allelic dominance as a weighted average of allelic activity. Briefly, our model proposed that coefficients of the heterozygote, , could be modeled as a linear combination of the coefficients of each homozygote:

where refers to the value of the th isoform in a genotype , and is the dominance coefficient for allele .

To find the parameters that maximized the probability of observing the data, we found the parameter, , that maximized the equation:

where was the coefficient associated with the th isoform in the trans-het and was the standard error of the th isoform in the trans-heterozygote samples as output by Kallisto. 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  using the Python programming language. The Numpy, pandas and scipy libraries were used for computation  and the matplotlib and seaborn libraries were used for data visualization . Enrichment analyses were performed using the WormBase Enrichment Suite .

# 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.