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*

August 3, 2017

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 two allelic series using transcriptomic phenotypes. We study gain-of-function and reduction-of-function *let-60 (ras)* alleles and show that they are drastically different and show relationships between themselves that were not predictable. We also study two alleles of *dpy-22* that generate two 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

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 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[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. 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[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 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[4,5]. Finally, genetic pathways can be reconstructed by using single cell sequencing via clustering[6] 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 focused on two allelics series: an allelic series of the *let-60* gene and an allelic series of the *dpy-22* gene in *C. elegans*. *let-60* is the *ras* orthologue in *C. elegans*[7], where it functions to promote the cellular fate of a number of cells during development[8]. *let-60* (*ras*) is a GTPase[7] that cycles between GDP- and GTP-binding states. The GTP-binding state is considered to be the signaling competent state. *dpy-22* is the TRAP-230/MED12 orthologue in *C. elegans*[9,10].

citation

In *C. elegans*, *let-60* (*ras*) always acts downstream of a receptor. Activated *let-60* (*ras*) can activate signaling cascades to transmit information to the eukaryotic nucleus. Often, this signaling cascade consists of *lin-45* (the RAF ortholog)[11], *mek-2*[12] and *mpk-1*[13]. The Ras pathway has been extensively studied in the context of vulval organogenesis, where it acts downstream of *let-23*[14] to promote vulval induction of the P*n*.p cells. In addition to vulval development, *let-60* (*ras*) has been implicated in the migration of the Sex Myoblasts (SM), where it acts both cell autonomously and non-autonomously, the formation of the excretory pore, hypodermal fluid homeostasis, as well as germline development and the formation of the male tail[15]. Thus, the Ras pathway is pleiotropic, but its phenotypes are separable and specific.

The intense scrutiny of the Ras pathway has led to the generation of a large number of loss-of-function alleles for *let-60* (*ras*). Null mutations of *let-60* (*ras*) are known to be lethal, but reduction-of-function alleles have been useful to carefully dissect the molecular properties of this protein. Gain-of-function mutations of *let-60* (*ras*) are also known, although they are much more rare. In particular, a single gain-of-function mutation, *n1046gf*, was famously isolated independently in several vulva formation screens by different laboratories[16,17,18]. These gain-of-function mutations have a privileged history in the Ras pathway, as they have been used in multiple screens to identify genes that have a Suppressor of Ras (Sur) phenotype such as *sur-1* (now *mpk-1*[13]) or *sur-2*[19]. Interestingly, although nulls of these genes are often lethal, strong loss-of-function alleles can be obtained for these suppressors. Moreover many of these reduction-of-function suppressor alleles have no phenotype in a wild-type background. This suggests that many of these genes mediate little Ras signaling under most situations yet have significant dynamic range in the amount of signaling they can accomodate. We selected *let-60* (*ras*) due to the range of phenotypic outcomes associated with this gene and the dependence of these outcomes on the type of lesion present (gain-of-function or reduction-of-function).

Is Ferguson 1985 the right citation here

anything else that I need to cite here?

dpy-22, also known as *mdt-12*, is a subunit of the Mediator complex. Briefly, Mediator is a complex that globally regulates RNA polymerase II (Pol II)[20,21]. Mediator is a versatile regulator, a quality often associated with its variable subunit composition[20], 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* (*med-12*)[22]. The CKM is considered a molecular switch, which inhibits Pol II activity by sterically preventing interactions between Mediator and the polymerase[23,24]. In *C. elegans*, *dpy-22* (*med-12*) has been studied primarily in the context of the male tail[9], where it was found to interact with the Wnt pathway, as well as vulval formation[25], where it was found to be an inhibitor of the Ras pathway. The null mutant of *dpy-22* (*med-12*) 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 DPY-22 protein missing the terminal 900 or so amino acids[9]. 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* (*med-12*) in the vulva, *dpy-22(sy622)*, is a premature stop codon that removes more than 1,500 amino acids of the protein[26]. 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.

who to cite for this?

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 two well-characterized sets of alleles. We sequenced a weak gain-of-function (*n1046gf*) and a strong loss-of-function (*n2021*) allele of Ras, as well as a weak loss-of-function (*bx93*) and a strong-loss-of-function (*sy622*) allele of *dpy-22* (*med-12*). We expected that the Ras alleles would show perturbations in the same set of genes in opposite directions. Instead, we find that the two sets of altered genes are largely distinct; of those genes that are common to both alleles, a majority change in the same direction, contrary to expectations. 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 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 DPY-22 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

Comparing gain-of-function and loss-of-function alleles

We sequenced in triplicate two alleles of *let-60*, *let-60(n1046gf)* and *let-60(n2021)* at the young adult stage and compared them with wild-type samples at the same stage. We sequenced each sample at a depth of 20M reads. These reads were pseudo-aligned using Kallisto[27], which allowed us to quantify 21,800 protein-coding isoforms. After quantification, we performed differential expression analysis using Sleuth[28]. Briefly, Sleuth uses a General Linear Model to identify genes that are differentially expressed by log-transforming the estimated counts of a given isoform in all the samples and performing a linear regression between a given mutant and wild-type. The slope of the linear regression, β, is a measurement of the magnitude of the perturbation, loosely analogous to the natural logarithm of the fold-change (see Angeles-Albores *et al*,

Figure 1), for each isoform. These slopes are tested against the null hypothesis that they are equal to zero. An isoform is considered to be differentially expressed when the q-value (p-value corrected for false discovery rate) is less than 0.1. β values can be positive or negative—positive values of β indicate an isoform that is up-regulated in the mutant relative to the wild-type control, whereas negative values of β indicate an isoform that is down-regulated in the mutant relative to the wild-type control. When we refer to β and q-values, it will always be in reference to isoforms. However, when speaking about the size of a gene set, we will always quantify it as the number of individual genes contained in the set. For *C. elegans*, the difference between the number of isoforms and genes is negligible because most protein-coding genes have a single isoform.

We predicted that the transcriptomes of these alleles would show the same set of differentially expressed genes (henceforth referred to as a shared transcriptomic phenotype, STP) because these alleles are both perturbing the GTP-binding potential of the protein product. However, we predicted that whereas these genes would be up-regulated (down-regulated) in the gain-of-function homozygote, they would be down-regulated (up-regulated) in the loss-of-function. In other words, the gain-of-function allele would have opposite effects relative to the loss-of-function allele for genes contained in the STP. We identified 3,021 differentially expressed genes in the *let-60(gf)* mutant relative to the wild-type control, and 857 differentially expressed genes in the *let-60(lf)* mutant relative to the wild-type control (see Fig. 1). Contrary to our expectations, the STP between these two alleles consisted of 31% of the differentially expressed genes in the *let-60(lf)* transcriptome, totalling 269 genes. Moreover, we found that this STP showed a strong positive correlation between the two alleles. To assess whether one allele was significantly stronger than another we split the STP into its correlated and anti-correlated components and found the regression line in each component. We measured a correlation coefficient of 1.03 and -1.00 for each component, indicating that the two alleles have effects that are on average exactly opposite. Our results suggest the existence of four phenotype classes: A *let-60(gf)*-specific class, a *let-60(lf)*-specific class, a positively correlated shared transcriptomic phenotype class and a negatively correlated shared transcriptomic phenotype class.

We used the WormBase Enrichment Suite [29] to perform Gene, Tissue and Phenotype Ontology enrichment analyses of each class. Analysis of the *let-60(gf)*-specific phenotype revealed enrichment of genes expressed in the early embryo (AB lineage), in the male, the hypodermis and the reproductive system whereas the *let-60(lf)*-specific phenotype showed enrichment of the intestine. Phenotype term enrichment associated the *let-60(gf)*-specific transcriptomic phenotype with the linker cell migration, lipid metabolism and neuropil development, and the *let-60(lf)*-specific transcriptomic phenotype was enriched in mitochondrial alignment. Gene ontology enrichment analysis also showed that the *let-60(gf)*-specific transcriptomic phenotype was enriched in regulation of cell-shape, immune system response and side of membrane. The *let-60(lf)*-specific phenotype showed enrichment in terms related to striated muscle, collagen trimer and cell death. Our results suggest that gain- and loss-of-function alleles have separable functions that do not conform to the gain/loss of structure phenotypes typically associated with them.

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 [26].

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

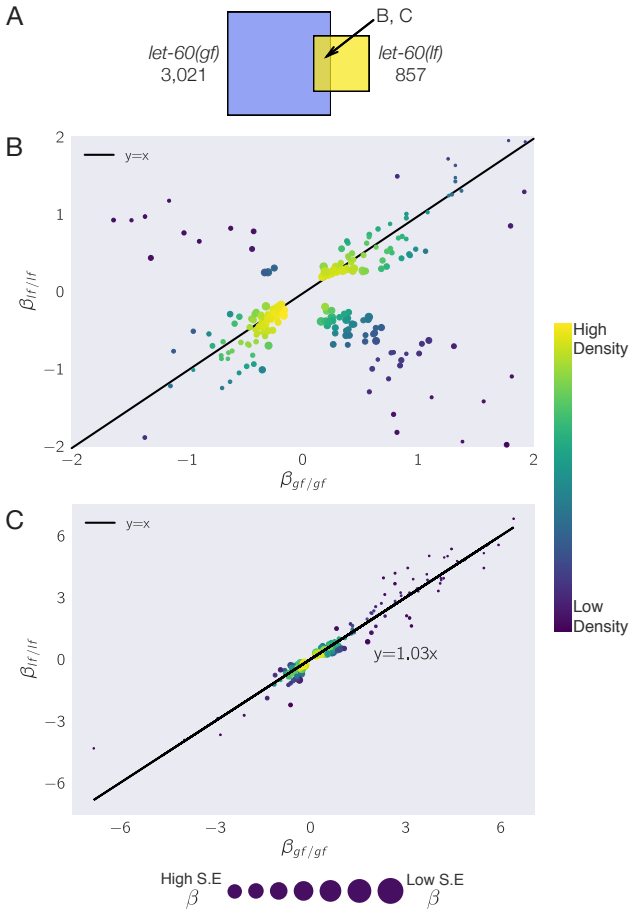

Figure 1. **A**. Venn diagram of the differentially expressed genes in a *let-60(gf)* and a *let-60(lf)* mutant. Diagram is to scale. **B**. β coefficients for isoforms that are differentially expressed in both mutants. **C** When two isoforms change in the same direction in both mutants, the magnitude of the change is also the same.

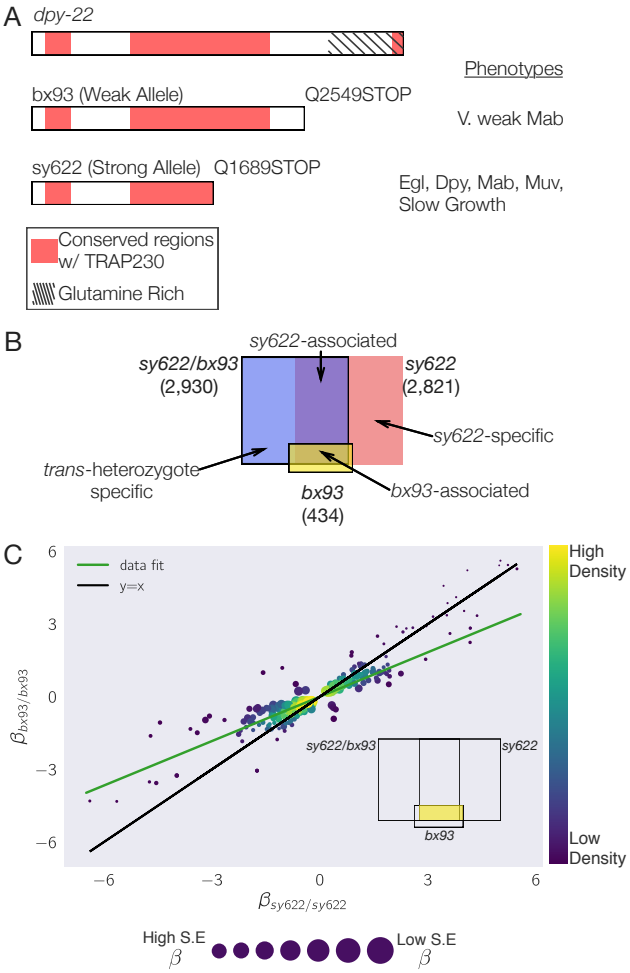

Figure 2. The *dpy-22* allelic series, consisting of two amino acid truncations, is amenable to study by transcriptomic phenotypes. **A**. Diagram of the *dpy-22* gene and the *bx93* and *sy622* alleles. **B**. Venn diagram of the genotypes we sequenced: A *bx93* homozygote, an *sy622* homozygote and a *bx93/sy622* trans-heterozygote. **C**. Genes that are commonly differentially expressed in both homozygotes 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.

magnitudes in the strong allele, *sy622* (see Fig. 2). 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.

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: Loss-of-function allelic series reveal unknown functionality). 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. 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: 3.3 versus 4.2, respectively), and a statistically smaller mean (1.3 vs 0.99, respectively, $p < 10^{-5}$, non-parametric boostrap). The *sy622*-specific phenotype had the smallest mean of all (0.9, $p < 10^{-5}$ compared with *bx93*-associated phenotype, and $p = 0.01$ 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 DPY-22.

Dominance can be quantified in transcriptomic phenotypes

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. 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 β coefficients of these genes in the *bx93* allele against the coefficients in the heterozygote. The

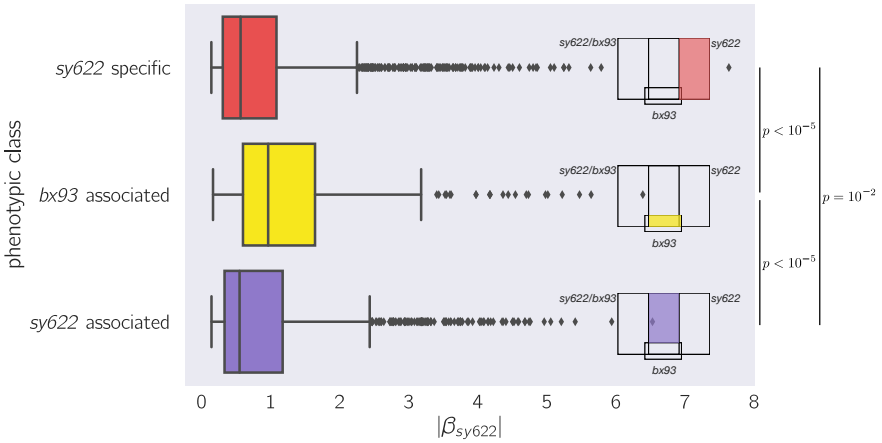

Figure 3. Within the *sy622* homozygote mutant, different phenotypic classes have statistically different perturbation distributions. Genes that are *sy622*-specific have a different perturbation distribution compared to genes that are *bx93*-associated or *sy622*-associated. The lines within the boxes 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.

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 β coefficients of each homozygote that would predict the observed β 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 $80\% \pm 1\%$ 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 β values against the observed β 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 4).

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 $48\% \pm 1\%$. 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.

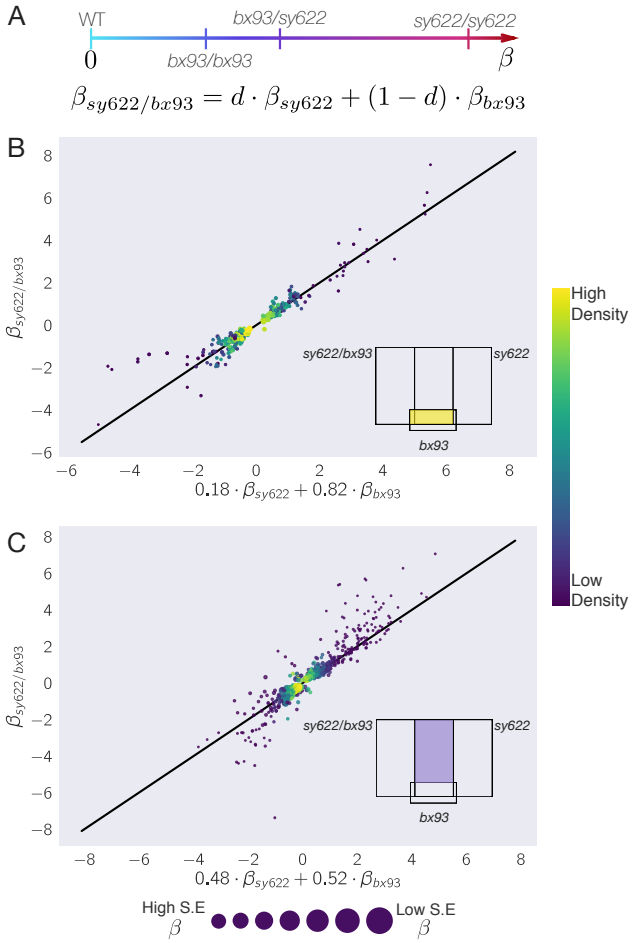

Figure 4. For each phenotype, alleles have a single dominance behavior. **A**. Schematic explaining codominance. The closer the trans-heterozygote is to one of the homozygotes, the more dominant the allele corresponding to that homozygote is considered to be. Codominance is only valid when the heterozygote 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 β 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 β values for each subset in the heterozygote. Insets show the subset of genes plotted in each graph.

Insights into the physiology of *sy622* homozygotes

Whereas the *sy622* homozygote is strongly phenotypic (see Fig. 2), 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[29,30] to query what anatomical, phenotypic or gene ontological terms were enriched in this gene set (see Table 1).

Term	$-\log_{10} q$, *sy622*-specific	$-\log_{10} q$, *bx93*-associated
Intestine	12	–
Intestinal muscle	4	< 1
PVD	3	< 1
Muscular system	7	< 1
pm3/5	< 1	2
Severe pleiotropic defects early embryo	4	< 1
Rachis absent	4	< 1
Meiosis defective early embryo	3	< 1
Dauer constitutive	< 1	2
Dauer metabolism	< 1	2
Collagen trimers	25	< 1
Muscle cell development	13	< 1
Contractile fibers	14	< 1
Oviposition	12	< 1
Glucuronosyltransferase activity	< 1	9
Monocarboxylic acid catabolic process	< 1	4
Lytic vacuole	< 1	4

Table 1. List of enriched terms for the *sy622*-specific and *bx93*-specific phenotypic classes. The *sy622*-specific and the *bx93*-associated columns show $-\log_{10} q$ for each term. q-values are calculated using a hypergeometric model and adjusted via a Benjamini-Hochberg algorithm. q-values shown are rounded to the nearest power of 10 for simplicity. < 1 implies that the q-value did not pass the significance threshold. Enrichment analysis was carried out using the WormBase Enrichment Suite[29].

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

Gain-of-function and loss-of-function alleles can lead to different transcriptomic states.

Contrary to our expectations, the gain-of-function and loss-of-function *let-60* (*ras*) alleles had significant differences in the genes they perturbed, and those genes that were commonly regulated tended to be per-

turbed in the same direction. To explain these findings, we considered a scenario where the gain-of-function allele signals constitutively through pathways that are only sparingly used in wild-type animals. This model would predict that the gain-of-function allele mutant would show changes in these pathways as they are flooded with new information through channels that are rarely used. In contrast, the loss-of-function allele mutant should show fewer changes in these pathways, because in this model most of the changes are rarely in use. Thus, genes that are differentially expressed in the gain-of-function allele homozygote would be interpreted as genes where *let-60* (*ras*) is sufficient to trigger changes in expression, and the changes in the loss-of-function allele represent genes where *let-60* (*ras*) is necessary for appropriate expression. Although this scenario begins to explain the difference in size between the number of differentially expressed genes in the gain-of-function allele homozygote and the loss-of-function allele homozygote, it does not explain why there are genes that are differentially expressed in the loss-of-function allele homozygote that are not present in the gain-of-function allele. Moreover, this model does not explain why the STP between the loss-of-function allele and gain-of-function allele homozygotes is predominantly correlated, instead of being entirely anti-correlated. Therefore, it seems unlikely that the transcriptomes of these alleles result from a model where *let-60* (*ras*) is capable of signaling through a large number of pathways but predominantly uses only a subset.

A mechanism that could feasibly generate *let-60(lf)*- and *let-60(gf)*-specific effects as well as correlated and anti-correlated effects would be to postulate that there are four signaling states available to LET-60. In this model, a protein bound to GTP constitutively can signal through a set of proteins, whereas a protein bound to GDP signals through a different, non-overlapping pathway. A third pathway requires GTP-to-GDP cycling at a specific average rate, such that interfering with that cycling by stabilizing either the GTP- or GDP-bound states has the same effect. A similar effect has been described previously for the Sec4 GTPase.

cite Peter Novick here

Another way to understand the effects we observed in *let-60(lf)* and *let-60(gf)* mutants is by thinking of the transcriptomes of each mutant as their respective states. Recently, transcriptome profiling has been used to identify novel states in both single cells and whole-organisms[4,5]. If these transcriptomes are states, then these states are the result of continous *let-60* (*ras*) activity (or inactivity) throughout the lifespan of the animal, and reflect the proximal or immediate effects of altered LET-60 activity a well as compensatory changes due to altered development or life history. This framework has the advantage that it does not *a priori* suggest that the gain-of-function allele homozygote and loss-of-function allele homozygote should have similar, or even overlapping, effects.

Without many more experiments, we cannot definitively point at a mechanism. It is possible that background mutations are contributing to the *let-60(lf)*- and *let-60(gf)*-specific changes, since these mutants were identified in different screens carried out in different labs. A rigorous methodology to exclude background mutations would call for sequencing multiple independent lines containing each allele. As library generation and sequencing costs fall, these experiments will become more feasible. Even assuming that background effects are responsible for the lack of overlap between the two mutants, the positive and negative correlations between these two alleles raise important questions about Ras biology. The fact that either correlation has a value with magnitude exactly of 1 suggests the existence of circuits that monitor Ras signaling levels with quantitative accuracy in *C. elegans*.

Loss-of-function allelic series reveal unknown functionality

Our sequencing results demonstrate that *sy622*, an allele that truncates 1700 amino acids from DPY-22, 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. 2). Unlike the *let-60* (*ras*) alleles we studied, 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 (which was the cutoff for 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.

Although a comparison between these two alleles proves fruitful in establishing differences in phenotypic severity, this comparison alone does not allow us to answer whether or not *bx93* and *sy622* act in different

ways on subsets of genes. To this end, we sequenced a trans-heterozygote of both alleles, which allowed us to identify four phenotypic classes, which in turn are informative about the biological effects of each allele. We found a *sy622*-specific phenotypic class for which the *bx93* has a dominant wild-type phenotype; we found a *bx93*-associated phenotypic class for which *bx93* also has a dominant phenotype, but this phenotype is definitely not wild-type. We also found a *sy622*-associated class for which *sy622* and *bx93* appear almost perfectly co-dominant. Intriguingly, we also found a phenotype specific to 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* (*med-12*) 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.

Our enrichment analysis of the *sy622*-specific and the *bx93*-associated phenotypic classes revealed that they reflect functionally distinct aspects of *dpy-22* (*med-12*) 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* (*med-12*) 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. Specifically, for the *sy622* and the *bx93* alleles, we can rule out that these alleles form a strict qualitative allelic series, since such a series should not exhibit a dosage-dependent phenotype (the *sy622*-associated class). However, it is possible that *sy622* and *bx93* represent a mixed quantitative-qualitative series, because we found phenotypic classes where *bx93* complemented *sy622*, *bx93* was dominant over *sy622* and where *sy622* was co-dominant with or partially dominant over *bx93*.

Genetics in multi-dimensional phenotypes

Add citations below.

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 (Perturb-Seq). 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 20°C[31].

Strain synchronization, harvesting and RNA sequencing

With the exception of *let-60(lf)*, 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). 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[27], 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[32,33,34,35]. All libraries had good quality scores.

Differential expression analysis was performed using Sleuth[28]. 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 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 10^5 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:

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

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:

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

where $\beta_{a/b,i,\text{Obs}}$ 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[36] using the Python programming language. The Numpy, pandas and scipy libraries were used for computation[37,38,39] and the matplotlib and seaborn libraries were used for data visualization[40,41]. Enrichment analyses were performed using the WormBase Enrichment Suite[29].

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. 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.

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

6. 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.

7. Han, M. & Sternberg, P. W. let-60, a gene that specifies cell fates during *C. elegans* vulval induction, encodes a ras protein. *Cell* 921–931.

8. Yochem, J., Sundaram, M. & Han, M. Ras is required for a limited number of cell fates and not for general proliferation in *Caenorhabditis elegans* . *Molecular and cellular biology* 2716–22.

9. 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).

10. Bourbon, H.-M. *et al.* A Unified Nomenclature for Protein Subunits of Mediator Complexes Linking Transcriptional Regulators to RNA Polymerase II.

11. Han, M., Golden, A., Han, Y. & Sternberg, P. W. *C. elegans* lin-45 raf gene participates in let-60 ras-stimulated vulval differentiation. *Nature* 133–140.

12. Wu, Y., Han, M. & Guan, K. L. MEK-2, a *Caenorhabditis elegans* MAP kinase kinase, functions in Ras-mediated vulval induction and other developmental events. *Genes and Development* **9**, 742–755 (1995).

13. Lackner, M. R., Kornfeld, K., Miller, L. M., Robert Horvitz, H. & Kim, S. K. A MAP kinase homolog, mpk-1, is involved in ras-mediated induction of vulval cell fates in *Caenorhabditis elegans* . *Genes and Development* **8**, 160–173 (1994).

14. Sternberg, P. W. *et al.* LET-23-mediated signal transduction during *Caenorhabditis elegans* development. *Molecular Reproduction and Development* 523–528.

15. Sundaram, M. RTK/Ras/MAPK signaling. *WormBook* .

16. Han, M., Aorian, R. V. & Sternberg, P. W. The let-60 locus controls the switch between vulval and nonvulval cell fates in *Caenorhabditis elegans* . *Genetics* 899–913.

17. Beitel, G. J., Clark, S. G. & Horvitz, H. R. Caenorhabditis elegans ras gene let-60 acts as a switch in the pathway of vulval induction. *Nature* 503–509.

18. Ferguson, E. L. & Horvitz, H. R. IDENTIFICATION AND CHARACTERIZATION OF 22 GENES THAT AFFECT THE VULVAL CELL LINEAGES OF THE NEMATODE *Caenorhabditis elegans* . *Genetics* .

19. Singh, N. & Han, M. sur-2, a novel gene, functions late in the let-60 ras-mediated signaling pathway during *Caenorhabditis elegans* vulval induction. *Genes & development* 2251–65.

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

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

22. 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.

23. 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.

24. 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.

25. 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).

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

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

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

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

30. 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* .

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

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

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

34. 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.

35. Ewels, P., Magnusson, M., Lundin, S. & K\"{a}ller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

36. 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.

37. 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.

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

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

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

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