Studying allelic series with transcriptomic phenotypes

David Angeles-Albores

Paul W. Sternberg

# Introduction

Mutations of a gene can yield a series of alleles with different phenotypes that reveal multiple functions encoded by that gene, regardless of the alleles’ molecular nature. Homozygous alleles can be ordered by their phenotypic severity; then, phenotypes of *trans*-heterozygotes carrying two alleles can reveal which alleles are dominant for each phenotype. Together, the severity and dominance hierarchies show intragenic functional regions. In *Caenorhabditis elegans*, these series have helped characterize genes such as , and  1–3.

Biology has moved from expression measurements of single genes towards genome-wide measurements. Expression profiling via RNA-seq 4 enables simultaneous measurement of transcript levels for all genes in a genome, yielding a transcriptome. These measurements can be made on whole organisms, isolated tissues, or single cells 5,6. Transcriptomes have been successfully used to identify new cell or organismal states 7,8. For mutant genes, transcriptomic states can be used for epistasis analysis 9,10, but have not been used to characterize allelic series.

We have devised methods for characterizing allelic series with RNA-seq. To test these methods, we selected three alleles 11,12 of a *C. elegans* Mediator complex subunit gene, . Mediator is a macromolecular complex with subunits 13 that globally regulates RNA polymerase II (Pol II) 14,15. The Mediator complex has at least four biochemically distinct modules: the Head, Middle and Tail modules and a CDK-8-associated Kinase Module (CKM). The CKM associates reversibly with other modules, and appears to inhibit transcription 16,17. In *C. elegans* development, the CKM promotes both male tail formation 11 (through interactions with the Wnt pathway), and vulval formation 18 (through inhibition of the Ras pathway). Homozygotes of allele , encoding a premature stop codon Q2549Amber 11, appear grossly wild-type. In contrast, animals homozygous for a more severe allele, encoding another premature stop codon, Q1698Amber 12, are dumpy (Dpy), have egg-laying defects (Egl), and have multiple vulvae (Muv). Due to pleiotropy, these alleles have not yet been ordered in a series (see Fig. [fig:flowchart]A). In spite of its causative role in a number of neurodevelopmental disorders 19, the structural and functional features of this gene are poorly understood. In humans, is known to have a proline-, glutamine- and leucine-rich domain that interacts with the WNT pathway 20. However, many disease-causing variants fall outside of this domain 21. To study these variants and how they interfere with the functionality of , quantitative and efficient methods are necessary.

RNA-seq phenotypes have the potential to reveal functional regions within genes, but their phenotypic complexity makes this difficult. We developed a method for determining allelic series from transcriptomic phenotypes and used the *C. elegans* gene as a test case. Our analysis revealed functional regions that act to modulate Mediator activity at thousands of genetic loci.

# Results and Discussion

We adapted the allelic series method, previously used for individual phenotypes, for use with expression profiles as multidimensional phenotypes (see Fig. [fig:flowchart]). As a proof of principle, we carried out RNA-seq on biological triplicates of mRNA extracted from homozygotes, homozygotes and wild type controls, along with quadruplicates from *trans*-heterozygotes of both alleles at a depth of 20 million reads per sample. Reads were pseudoaligned using Kallisto 22. We performed a differential expression using a general linear model specified in Sleuth 23 (see ). Differential expression with respect to the wild type control for each transcript in a genotype is measured via a coefficient , which can be loosely interpreted as the natural logarithm of the fold-change. Transcripts were considered to have differential expression between wild-type and a mutant if the false rate, , was less than or equal to 10%.

![image](data:application/pdf;base64,)

By these criteria, we found 481 genes differentially expressed in homozygotes, and 2,863 differentially expressed genes in homozygotes (see [Basic Statistics Notebook](https://wormlabcaltech.github.io/med-cafe/notebook/basic.html)). We also sequenced *trans*-heterozygotes with the genotype , and found 2,214 differentially expressed genes.

We used a false hit analysis to identify four non-overlapping phenotypic classes. We use the term genotype-specific to refer to groups of transcripts that were perturbed in one mutant. We use the term genotype-associated to refer to those groups of transcripts perturbed in two or more mutants. The **-associated** phenotypic class consisted of 720 genes differentially expressed in homozygotes and in *trans*-heterozygotes, but which had wild-type expression in homozygotes. The **-associated** phenotypic class contains 403 genes differentially expressed in all genotypes. We also identified a **-specific** phenotypic class (1,841 genes) and a ***trans*-heterozygote-specific** phenotypic class (1,226 genes; see the  [Phenotypic Classes Notebook](https://wormlabcaltech.github.io/med-cafe/notebook/phenotypic_classes.html)). All genotype-associated phenotypes had Spearman rank correlations , indicating that transcripts within these classes changed in the same direction amongst the genotypes studied.

We measured allelic dominance for each class. The allele is completely recessive to for the -specific phenotypic class. The and alleles are semidominant () to each other for the -associated phenotypic class. The allele is largely dominant over the allele (; see Table [tab:dom]).

Dominance analysis for the allelic series. Dominance values closer to 1 indicate is dominant over , whereas 0 indicates is dominant over .

|  |  |
| --- | --- |
| Phenotypic Class | Dominance |
| -specific |  |
| -associated |  |
| -associated |  |

Our results suggest the existence of various functional regions in (see Fig. [fig:domains]). The -specific phenotypic class is likely controlled by a single functional region, functional region 1 (FC1), and the -associated phenotypic class is likely controlled by a second functional region, functional region 2 (FC2). It is unlikely that these regions are identical because their dominance behaviors are very different. The allele was largely dominant over the allele for the -associated class, but gene expression in this class was perturbed in both homozygotes. The perturbations were greater for homozygotes than for homozygotes. This behavior can be explained if the -associated class is controlled jointly by two distinct effectors, functional regions 3 and 4 (FC3, FR4, see Fig. [fig:domains]). A rigorous examination of this model will require studying alleles that mutate the region between Q1689 and Q2549 using homozygotes and *trans*-heterozygotes.

![ The functional regions associated with each phenotypic class can be mapped intragenically. The number of genes associated with each class is shown. The -associated class may be controlled by two functional regions. FR2 and FR3 could be redundant if FR4 is a modifier of FR2 functionality at -associated loci. ](data:application/pdf;base64,)

The functional regions associated with each phenotypic class can be mapped intragenically. The number of genes associated with each class is shown. The -associated class may be controlled by two functional regions. FR2 and FR3 could be redundant if FR4 is a modifier of FR2 functionality at -associated loci.

We also found a class of transcripts that had perturbed levels in *trans*-heterozygotes only; its biological significance is unclear. Phenotypes unique to *trans*-heterozygotes are often the result of physical interactions such as homodimerization, or dosage reduction of a toxic product 24. In the case of orthologs, how either mechanism could operate is not obvious, since the is expected to assemble in a monomeric manner into the CKM. Massive single-cell RNA-seq of *C. elegans* has recently been reported 25. When this technique becomes cost-efficient, single-cell profiling of these genotypes may provide information that complements the whole-organism expression phenotypes, perhaps explaining the origin of this phenotype.

Intragenic mapping of functional regions associated with phenotypic classes is important, but the biological meaning of functions remains unclear. To assign biological functionality to phenotypic classes, we extracted transcriptomic signatures associated with a Dumpy (Dpy) phenotype using transcriptomes from and mutants (DAA, CPR and PWS *unpublished*), and a -dependent hypoxia response from a previously published analysis 10 and asked whether any phenotypic class was enriched in either response. The *sy622*-specific and -associated classes were enriched in genes that are transcriptionally associated with a Dpy phenotype (fold-change enrichment = 3, , genes observed; fold-change = 1.9, , 82 genes observed). The *bx93*-associated class also showed significant enrichment (fold-change = 2.2, , 68 genes observed). The class that showed the most extreme deviation from random was the *sy622*-specific class. homozygotes are severely Dpy, whereas homozygotes and *trans*-heterozygotes have a slight Dpy phenotype. Plotting the perturbation levels in the *sy622* homozygotes versus the perturbation levels in the *dpy-7* mutants revealed that 75% of the transcripts were strongly correlated in both genotypes. Therefore, the *sy622*-specific phenotypic class contains a transcriptional signature associated with morphological Dpy phenotype (see the  [Enrichment Notebook](https://wormlabcaltech.github.io/med-cafe/notebook/enrichment.html)).

is not known to be upstream of the -dependent hypoxia response in *C. elegans*. Enrichment tests revealed that the hypoxia response was significantly enriched in the *bx93*-associated (fold-change = 2.1, , 63 genes observed), the *sy622*-associated (fold-change = 1.9, , 78 genes observed) and the *sy622*-specific classes (fold-change = 2.4, , 186 genes observed). However, there was no correlation between the expression levels of these genes in genotypes and the expression levels expected from the hypoxia response. Although the hypoxia gene battery can be found in mutants, these genes are not used to deploy a -dependent hypoxia phenotype. Taken together, our results suggest that transcriptomic signatures can be used to understand the biological functionality of phenotypic classes. In *C. elegans*, enrichment analyses of anatomy, phenotype and gene ontologies can also be performed using standardized tools 26,27, but these analyses do not make use of the quantitative information provided by RNA-seq. Additionally, these analyses rely on curated annotations generated from a wide variety of experiments. Transcriptomic signatures are easy to derive, easy to update and are not subject to annotation or experimental bias.

![ sy622 homozygotes show a transcriptional response associated with the Dpy phenotype. A We obtained a set of transcripts associated with the Dpy phenotype from and mutants. We identified the transcripts that were differentially expressed in sy622 homozygotes. Next, we plotted the \beta values of each transcript in sy622 homozygotes against the \beta values in a dpy-7 mutant. A significant portion of the genes are correlated between the two genotypes, showing that the signature is largely intact. 25% of the genes are anti-correlated. B We performed the same analysis using a set of transcripts associated with the -dependent hypoxia response as a negative control. Although sy622 is enriched for the transcripts that make up this response, there is no correlation between the \beta values in sy622 homozygotes and the \beta values in egl-9 homozygotes. In the plots, a colormap is used to represent the density of points. The standard error of the mean is inversely proportional to the standard error of \beta_{mdt-12(sy622)}. ](data:application/pdf;base64,)

*sy622* homozygotes show a transcriptional response associated with the Dpy phenotype. **A** We obtained a set of transcripts associated with the Dpy phenotype from and mutants. We identified the transcripts that were differentially expressed in *sy622* homozygotes. Next, we plotted the values of each transcript in *sy622* homozygotes against the values in a *dpy-7* mutant. A significant portion of the genes are correlated between the two genotypes, showing that the signature is largely intact. 25% of the genes are anti-correlated. **B** We performed the same analysis using a set of transcripts associated with the -dependent hypoxia response as a negative control. Although *sy622* is enriched for the transcripts that make up this response, there is no correlation between the values in *sy622* homozygotes and the values in *egl-9* homozygotes. In the plots, a colormap is used to represent the density of points. The standard error of the mean is inversely proportional to the standard error of .

Transcriptomic phenotypes generate large amounts of differential gene expression data, so false positive and false negative rates can lead to spurious phenotypic classes whose putative biological significance is badly misleading. Such artifacts are particularly likely for small phenotypic classes, which should be viewed with skepticism. Notably, errors of interpretation cannot be avoided by setting a more stringent -value cut-off; doing so will decrease the false positive rate, but increase the false negative rate, which will in turn produce smaller phenotypic classes than expected. Our method avoids this pitfall by using total error rate estimates to assess the plausibility of each class. These conclusions are of broad significance to research where highly multiplexed measurements are compared to identify similarities and differences in the genome-wide behavior of a single variable under multiple conditions.

We have shown that transcriptomes can be used to study allelic series in the context of a large, pleiotropic gene. We identified separable phenotypic classes that would otherwise be obscured by other methods, correlated each class to a functional region, and identified sequence requirements for each region. Given the importance of allelic series for characterizing genetic pathways, we are optimistic that this method will be a useful addition to the geneticist’s arsenal.

# Methods

## Strains used

Strains used were N2 wild-type (Bristol), PS4087 , PS4187 , and PS4176  
. Lines were grown on standard nematode growth media (NGM) Petri plates seeded with OP50 *E. coli* at 20C 28.

## Strain synchronization, harvesting and RNA sequencing

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 grown to the young adult stage (assessed by vulval morphology and lack of embryos). RNA extraction and sequencing was performed as previously described by Angeles-Albores *et al* 7,10.

## Read pseudo-alignment and differential expression

Reads were pseudo-aligned to the *C. elegans* genome (WBcel235) using Kallisto 22, 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 29–32.

Differential expression analysis was performed using Sleuth 23. 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 young adult wild-type replicates from other published 7,10 and unpublished analyses adjusting for batch effects.

## False hit analysis

To accurately count phenotypes, we developed a false hit algorithm (Algorithm [alg:false]). We implemented this algorithm for three-way comparisons in Python. Although experimentally restricted, a three-way comparison can result in possible sets (ignoring size). This large number of models necessitates an algorithmic approach that can at least restrict the possible number of models. Our algorithm uses a noise function that assumes false hit events are non-overlapping (i.e. the same gene cannot be the result of two false positive events in two or more genotypes) to determine the average noise flux between phenotypic classes. These assumptions break down rapidly if false-positive or negative rates exceed 20%.

To benchmark our algorithm, we generated one thousand Venn diagrams at random. For each Venn diagram, we calculated the average false positive and false negative flux matrices. Then, we added noise to each phenotypic class in the Venn diagram, assuming that fluxes were normally distributed with mean and standard deviation equal to the flux coefficient calculated. We input the noised Venn diagram into our false hit analysis and collected classification statistics. For a given signal-to-noise cutoff, , classification accuracy varied significantly with changes in the total error rate. In the absence of false negative hits, false hit analysis can accurately identify non-empty genotype-associated phenotypic classes, but identifying genotype-specific classes becomes difficult if the experimental false positive rate is high. On the other hand, even moderate false negative rates () rapidly degrade signal from genotype-associated classes. For classes that are associated with three genotypes, an experimental false negative rate of 30% is enough on average to prevents this class from being observed.

We selected because classification using this threshold was high across a range of false positive and false negative combinations. A challenge to applying this algorithm to our data is the fact that the false negative rate for our experiment is unknown. Although there has been significant progress in controlling and estimating false positive rates, we know of no such attempts for false negative rates. It is unlikely that the false negative rate for our study is lower than the false positive rate, because all genotypes except the controls are likely underpowered. We used false negative rates between 10–20% for false hit analysis. When the false negative rate was set at 15% or higher, the algorithm converged on the same five classes shown above. For false negative rates between 10–15%, the algorithm output the same five classes, but also accepted the (,)-associated class. We selected the model corresponding to false negative rates of 15–20% because this model had lower values than the model selected with a false negative rate of 10–15% (4,212 versus 100,650).

We asked whether re-classification of some classes into others could improve model fit. We manually re-classified the (,)-associated and the (, *trans-heterozygote*)-associated classes into the *bx93*-associated class (which is associated with all genotypes), and we compared statistics between a re-classified reduced model and a reduced model. The re-classified model had a lower (181). Thus, we concluded that the re-classified reduced model is the most likely model to give rise to our data.

[alg:false]

*Return the reduced model*

## Dominance analysis

We modeled allelic dominance as a weighted average of allelic activity:

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

Code was written in Jupyter notebooks 33 using the Python programming language. The Numpy, pandas and scipy libraries were used for computation 34–36 and the matplotlib and seaborn libraries were used for data visualization 37,38. Enrichment analyses were performed using the WormBase Enrichment Suite 26. For all enrichment analyses, a -value of less than was considered statistically significant. For gene ontology enrichment analysis, terms were considered statistically significant only if they also showed an enrichment fold-change greater than 2.

## Data Availability

Raw and processed reads were deposited in the Gene Expression Omnibus. Scripts for the entire analysis can be found with version control in our Github repository, <https://github.com/WormLabCaltech/med-cafe>. A user-friendly, commented website containing the complete analyses can be found at <https://wormlabcaltech.github.io/med-cafe/>. Raw reads and quantified abundances for each sample were deposited at the NCBI Gene Expression Omnibus (GEO) 39 under the accession code GSE107523 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107523>).

# Acknowledgements

This work was supported by HHMI with whom PWS was an investigator, by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology, and by the NIH grant U41 HG002223. This article would not be possible without help from Dr. Igor Antoshechkin and Dr. Vijaya Kumar who performed the library preparation and sequencing. We would like to thank Carmie Puckett Robinson for the unpublished Dpy transcriptional signature. Han Wang, Hillel Schwartz, Erich Schwarz, Porfirio Quintero and Carmie Puckett Robinson provided valuable input throughout the project.

1. Aroian, R. V. & Sternberg, P. W. Multiple functions of let-23, a *Caenorhabditis elegans* receptor tyrosine kinase gene required for vulval induction. *Genetics* **128,** 251–67 (1991).

2. Ferguson, E. & Horvitz, H. R. Identification and characterization of 22 genes that affect the vulval cell lineages of *Caenorhabditis elegans*. *Genetics* **110,** 17–72 (1985).

3. Greenwald, I. S., Sternberg, P. W. & Robert Horvitz, H. The lin-12 locus specifies cell fates in *Caenorhabditis elegans*. *Cell* **34,** 435–444 (1983).

4. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).

5. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6,** 377–382 (2009).

6. Schwarz, E. M., Kato, M. & Sternberg, P. W. Functional transcriptomics of a migrating cell in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 16246–51 (2012).

7. Angeles-Albores, D. *et al.* The *Caenorhabditis elegans* Female State: Decoupling the Transcriptomic Effects of Aging and Sperm-Status. *G3: Genes, Genomes, Genetics* (2017).

8. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356,** (2017).

9. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167,** 1853–1866.e17 (2016).

10. Angeles Albores, D., Puckett Robinson, C., Williams, B. A., Wold, B. J. & Sternberg, P. W. Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements. *bioRxiv* (2017). doi:[10.1101/112920](https://doi.org/10.1101/112920)

11. Zhang, H. & Emmons, S. W. A *C. elegans* mediator protein confers regulatory selectivity on lineage-specific expression of a transcription factor gene. *Genes and Development* **14,** 2161–2172 (2000).

12. Moghal, N. A component of the transcriptional mediator complex inhibits RAS-dependent vulval fate specification in *C. elegans*. *Development* **130,** 57–69 (2003).

13. Jeronimo, C. & Robert, F. The Mediator Complex: At the Nexus of RNA Polymerase II Transcription. **27,** 765–783 (2017).

14. Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nature reviews. Molecular cell biology* **16,** 155–166 (2015).

15. Takagi, Y. & Kornberg, R. D. Mediator as a general transcription factor. *The Journal of biological chemistry* **281,** 80–9 (2006).

16. Knuesel, M. T., Meyer, K. D., Bernecky, C. & Taatjes, D. J. The human CDK8 subcomplex is a molecular switch that controls Mediator coactivator function. *Genes & development* **23,** 439–51 (2009).

17. Elmlund, H. *et al.* The cyclin-dependent kinase 8 module sterically blocks Mediator interactions with RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 15788–93 (2006).

18. Moghal, N. & Sternberg, P. W. A component of the transcriptional mediator complex inhibits RAS-dependent vulval fate specification in *C. elegans*. *Development* **130,** 57–69 (2003).

19. Graham, J. M. & Schwartz, C. E. MED12 related disorders. *American Journal of Medical Genetics, Part A* **161,** 2734–2740 (2013).

20. Kim, S., Xu, X., Hecht, A. & Boyer, T. G. Mediator is a transducer of Wnt/-catenin signaling. *Journal of Biological Chemistry* **281,** 14066–14075 (2006).

21. Yamamoto, T. & Shimojima, K. A novel MED12 mutation associated with non-specific X-linked intellectual disability. *Human Genome Variation* **2,** 15018 (2015).

22. Bray, N. L., Pimentel, H. J., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34,** 525–7 (2016).

23. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *brief communications nature methods* **14,** (2017).

24. Yook, K. Complementation. *WormBook* (2005). doi:[10.1895/wormbook.1.24.1](https://doi.org/10.1895/wormbook.1.24.1)

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

26. Angeles-Albores, D., N. Lee, R. Y., Chan, J. & Sternberg, P. W. Tissue enrichment analysis for *C. elegans* genomics. *BMC Bioinformatics* **17,** 366 (2016).

27. 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* (2017). doi:[10.1101/106369](https://doi.org/10.1101/106369)

28. Brenner, S. The Genetics of *Caenorhabditis elegans*. *Genetics* **77,** 71–94 (1974).

29. Andrews, S. FastQC: A quality control tool for high throughput sequence data. 1 (2010). doi:[citeulike-article-id:11583827](https://doi.org/citeulike-article-id:11583827)

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

31. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Bowtie: An ultrafast memory-efficient short read aligner. *Genome biology* R25 (2009). doi:[10.1186/gb-2009-10-3-r25](https://doi.org/10.1186/gb-2009-10-3-r25)

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

33. Pérez, F. & Granger, B. IPython: A System for Interactive Scientific Computing Python: An Open and General- Purpose Environment. *Computing in Science and Engineering* **9,** 21–29 (2007).

34. 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).

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

36. Oliphant, T. E. SciPy: Open source scientific tools for Python. *Computing in Science and Engineering* **9,** 10–20 (2007).

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

38. Waskom, M. *et al.* seaborn: v0.7.0 (January 2016). (2016). doi:[10.5281/zenodo.45133](https://doi.org/10.5281/zenodo.45133)

39. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30,** 207–10 (2002).