Gene expression variation across genetically identical individuals predicts reproductive traits

Amy K. Webster[1,2], John H. Willis[1], Erik Johnson[1], Peter Sarkies[3], Patrick C. Phillips[1]

1. Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA
2. Department of Biological Science, Florida State University, Tallahassee, FL, USA
3. Department of Biochemistry, University of Oxford, Oxford, UK

ABSTRACT

In recent decades, genome-wide association studies (GWAS) have been the major approach to understand the biological basis of individual differences in traits and diseases. However, GWAS approaches have limited predictive power to explain individual differences, particularly for complex traits and diseases in which environmental factors play a substantial role in their etiology. Indeed, individual differences persist even in genetically identical individuals, although fully separating genetic and environmental causation is difficult in most organisms. To understand the basis of individual differences in the absence of genetic differences, we measured two quantitative reproductive traits in 180 genetically identical young adult *Caenorhabditis elegans* roundworms in a shared environment and performed single-individual transcriptomics on each worm. We identified hundreds of genes for which expression variation was strongly associated with reproductive traits, some of which depended on individuals' historical environments and some of which was random. Multiple small sets of genes together were highly predictive of reproductive traits, explaining on average over half and over a quarter of variation in the two traits. We manipulated mRNA levels of predictive genes to identify a set of causal genes, demonstrating the utility of this approach for both prediction and understanding underlying biology. Finally, we found that the chromatin environment of predictive genes was enriched for H3K27 trimethylation, suggesting that gene expression variation may be driven in part by chromatin structure. Together, this work shows that individual, non-genetic differences in gene expression are both highly predictive and causal in shaping reproductive traits.

MAIN TEXT

Introduction

Over the past two decades, genome-wide association studies (GWAS) have sought to identify the genetic basis of individual differences driving traits and diseases. While there have been some notable successes[1,2], GWAS analyses often require hundreds of thousands of individuals to identify genetic loci that typically explain a small fraction of variation in traits and diseases[3], and the clinical utility of such loci is often unclear[4,5]. Complex traits and diseases that are driven by a strong environmental component are particularly problematic for GWAS, as these environmental inputs cannot be experimentally controlled in humans[6,7]. This lack of control means that some loci may be spuriously identified and others left undiscovered. More recent approaches, including transcriptome-wide association studies[8] (TWAS) and epigenome-wide association studies[9] (EWAS) recognize the intermediate role of epigenetic regulation and gene expression in driving complex traits and diseases. However, TWAS approaches focus specifically on predicted genetic effects on expression to prioritize GWAS candidates and thus do not account for potentially causal environmentally induced changes in gene expression. EWAS approaches, like GWAS and TWAS, are performed in individuals for which genetic and environmental information are not experimentally controlled, and thus it is unclear if identified epigenomic changes are driven by a genetic difference or an environmental effect[10]. Overall, the major approaches used to uncover the etiology of complex traits and diseases are limited in their ability to separate genetic and environmental effects. Here, we show that even stochastic differences among individuals can be strongly predictive of complex reproductive traits when analyzed through the lens of differences in gene expression.

Complete control of genetic background, especially as layered across different environment inputs, is impossible within human populations. Model organism studies have therefore been invaluable in building our understanding of environmentally induced effects on epigenetic regulation, gene expression, and traits of interest in a controlled genetic background[11–17]. However, such analyses generally focus on bulk populations (typically either to ensure ample material or to average differences across individuals within a replicate for model organisms such as *Drosophila* and *C. elegans*[18,19]), or they use very few individuals (typically due to difficulty sampling many larger organisms such as mice[20]), precluding knowledge of how individual life trajectories may influence gene expression, epigenetic information, and traits of interest. In recent years, it has become possible to generate individual transcriptomic profiles in *C. elegans* and

58 *Drosophila,* and methods to profile individuals in high throughput have improved[21–27]. This has revealed
59 that genetically identical individuals in a shared environment indeed differ in their gene expression profiles.
60 However, the effects that gene expression differences across genetically identical individuals may have on
61 complex fitness-related phenotypes of these same individuals -- which also vary substantially across
62 genetically identical individuals – remain largely unknown. To fully separate genetic and environmental
63 causation, we use an isogenic population of *C. elegans* to show that individuals in a common environment
64 at the same developmental stage exhibit substantial differences in two fitness-related reproductive traits.
65 By performing mRNA-seq on each individual worm, we identify hundreds of genes for which variation in
66 expression is strongly associated with reproductive trait variation, determine the extent to which expression
67 variation is driven by noise versus known environmental history, identify small sets of genes that are highly
68 predictive of both reproductive traits, and show experimentally that reducing mRNA levels of individual
69 genes causally affects progeny production. Finally, we show that predictive gene sets for reproductive traits
70 are enriched for genes with the chromatin modification H3K27me3, revealing that chromatin environment
71 defines which loci are subject to the individual variation in gene expression that drives variation in complex
72 traits.
73
74 **Results**
75 To determine if differences in gene expression levels correspond to differences in complex traits,
76 we measured reproductive traits and whole-animal mRNA transcriptomes for each of 180 isogenic *C.*
77 *elegans* in a shared environment at the same stage of early adulthood (Fig 1A). Two reproductive traits –
78 egg-laying onset (ELO) and the number of progeny produced in the first 24 hours of egg-laying (early
79 brood) – were measured for each individual. To generate additional phenotypic variation and allow us to
80 distinguish between gene expression noise and known environmental sources of variation, worms
81 experienced controlled environmental perturbations early in development in a two-factor design: they were
82 either born to a day 1 or day 3 adult parent (corresponding to the beginning and middle of egg-laying in *C.*
83 *elegans*, respectively), and, once hatched, they were subject to either a constant 20°C temperature or an
84 eight-hour shift to 25°C (both corresponding to typical rearing temperatures for *C. elegans*). This design is
85 ideal for at least two reasons. First, in standard worm rearing, worms that at the same developmental stage
86 in a shared environment have experienced some differences in their environmental histories that are

87 typically unknown to the observer because they appear physiologically identical. Some of these differences
88 in environmental history include factors that we directly controlled in this experiment, including the age of
89 each worm's parent and small fluctuations in rearing temperature. By explicitly incorporating this *historical*
90 environmental heterogeneity, we can distinguish between gene expression variability induced by
91 environmental history and gene expression variability driven by 'noise'. Noise encompasses both unknown
92 extrinsic factors (e.g., microenvironmental differences in food availability) and intrinsic factors (e.g.,
93 stochasticity in biochemical reactions). Second, with this approach we can ask whether sequencing adult
94 worms in a shared environment is a viable approach for phenotyping-by-transcriptomics and whether such
95 an approach depends on knowledge of the historical environment. In other words, if an adult worm is
96 picked off a plate and sequenced, can we determine when that worm began reproducing and how many
97 progeny it produced, without regard to its previous environment or its ancestors' environment? In our
98 experiment, isogenic adult worms in a common environment (with distinct historical environments)
99 exhibited a range of both ELO and early brood trait values (Fig. S1A), and variation in these traits was
100 partially driven by environmental history, consistent with prior work on parental age and rearing
101 temperature in *C. elegans*[27–29](Fig S1B). To determine if gene expression differences identified in single-
102 worm mRNA-seq were directly associated with reproductive traits measured in the same worms, we used a
103 linear mixed modeling approach and, remarkably, identified significant genes corresponding to each trait
104 (Fig 1B-C, Fig S2). Expression of 448 genes was strongly associated with early brood (Fig 1B-C,
105 Bonferroni p-value of 0.05, nominal p-value of 5.7×10^{-6}), while expression of 11 genes was strongly
106 associated with ELO (Fig S2). This model was agnostic to the environmental history of the worm, but
107 accounted for biological replicate, and we confirmed that these effects were not driven by mutations arising
108 in the population over the course of the experiment (see Methods). This reveals many genes, particularly
109 for early brood, for which expression differences across genetically identical individuals strongly associate
110 with the reproductive traits of these individuals. Importantly, this indicates that expression levels of
111 hundreds of genes provide useful information about the reproductive status of individual worms.
112

113

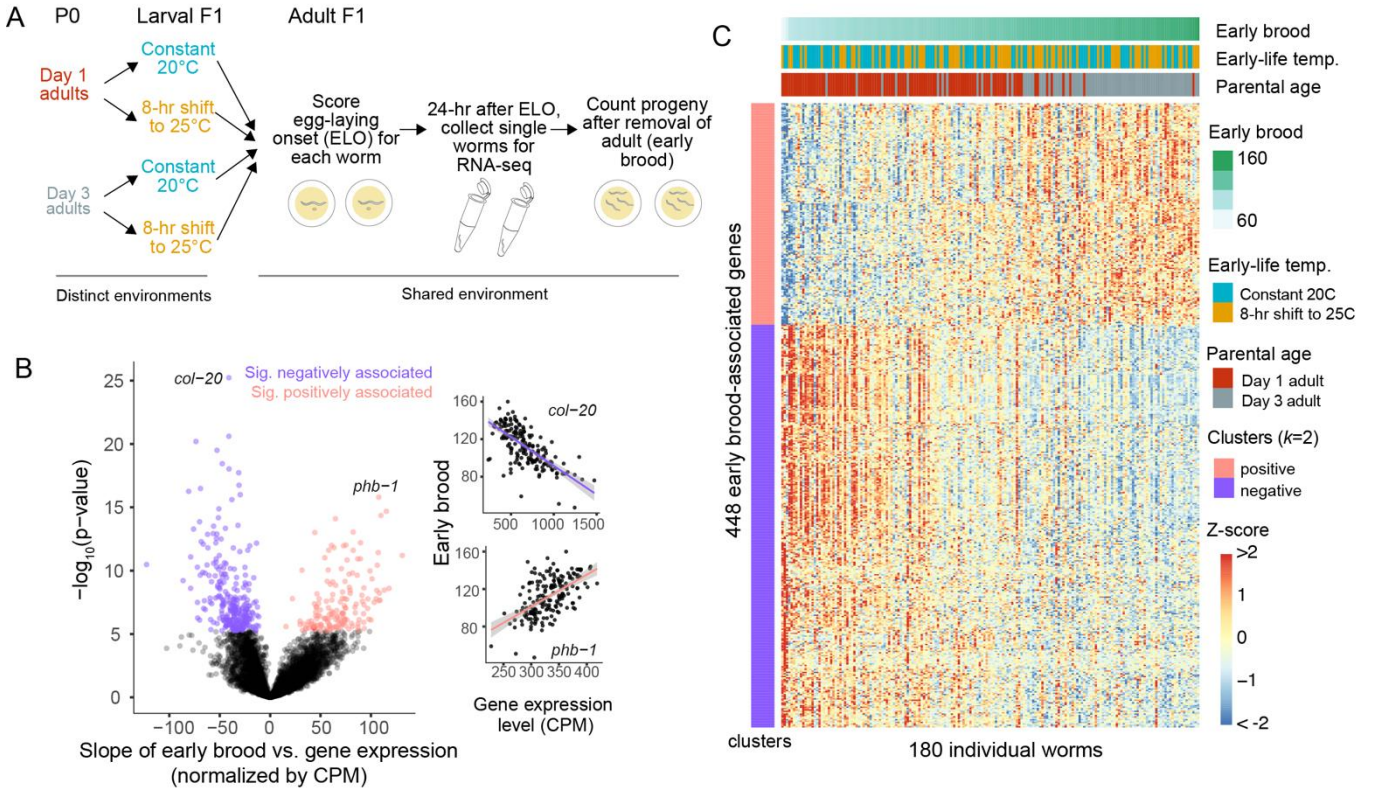


Figure 1. Differences in early brood across isogenic individuals are associated with expression levels of hundreds of genes.
A. Experimental design for single-individual mRNA-seq. P0 worms are allowed to lay progeny as Day 1 or Day 3 adults. F1 progeny are either subject to a constant temperature or an eight-hour shift to 25°C. Beginning in mid-larval stages all F1 worms experience the same environment. For each F1 worm, time to egg-laying onset is scored. 24 hours after egg-laying onset, each F1 worm is collected for single-individual mRNA-seq. F2 progeny laid on plates in the first 24 hours of egg-laying are counted to determine the early brood of each F1 worm.
B. Each of 8824 expressed genes was analyzed in a mixed model to assess the strength of its association with early brood. Significant genes are in pink (positively associated) or purple (negatively associated), non-significant genes are in black. Significance determined by a Bonferroni-corrected p-value cutoff of 0.05 (nominal p-value of 5.67×10^{-6}). Phenotypic and expression data for each worm is shown for the top two genes, *col-20* and *phb-1* with linear model fits in black and 95% confidence intervals in gray.
C. Heatmap of 448 brood-associated genes for all 180 individuals. Worms are sorted from left to right from lowest to highest early brood. Early brood data and environmental information is shown in the bars at the top. Genes are sorted into two clusters based on whether the association was positive or negative.

114 We next quantified the effect sizes on reproductive traits explained by the individual genes that
115 were significant in the previous analysis. Further, because of our experimental design, we also determined
116 the effect size and significance of the association after accounting for environmental history in our model,
117 allowing us to distinguish how much of the effect of gene expression on reproductive traits is gene
118 expression 'noise' and not driven by known environmental history. At the phenotypic level, differences in
119 early brood were partially driven by parental age but not affected by early-life temperature (Fig 1C and Fig
120 S1B). Differences in the timing of egg-laying onset were partially driven by both parental age and

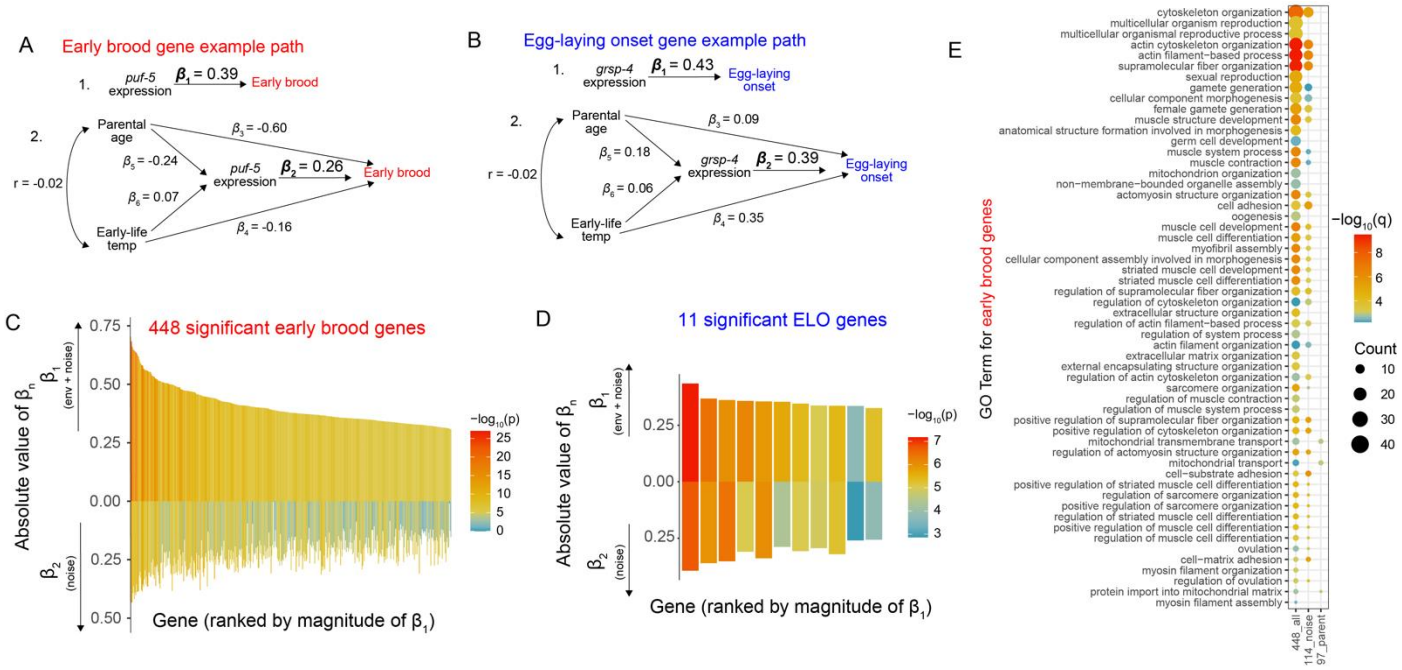


Figure 2: Associations between gene expression and reproductive traits are partially driven by both noise and historical environment.
A. Two example path analyses showing the relationship between gene expression and early brood. In Model 1, gene expression of a particular gene, *puf-5*, drives differences in early brood and β_1 is the coefficient of the linear mixed model. β_1 represents the total effect of gene expression on early brood. In Model 2, β_2 represents the effect of *puf-5* on early brood that is independent of the historical environment.
B. Two example path analyses showing the relationship between gene expression and egg-laying onset, analogous to A. In Model 1, β_1 is the total effect of *grsp-4* on egg-laying onset, while in Model 2, β_2 represents the effect of *grsp-4* that is independent of historical environment.
C. The 448 genes with for which expression variation is significantly associated with early brood, ordered by the magnitude of their β_1 effect sizes. The absolute values of β_1 and β_2 are plotted upward and downward, respectively, for each gene.
D. The 11 genes for which expression variation is significantly associated with egg-laying onset across individuals, ordered by the magnitude of their β_1 effect sizes. The absolute values of β_1 and β_2 are plotted upward and downward, respectively, for each gene.
C-D. Legend indicates the $-\log_{10}$(p-value) of β_1 and β_2 for each gene.
E. GO Terms for the 448 genes associated with early brood with a q-value < 0.01 are shown. GO Terms for the two subsets of the 448 genes with overlapping GO Terms are also shown: 114 genes that have significant values of β_2 but parental age does not have a significant effect on expression ('noise' genes), and 97 genes that do not have significant values of β_2 but parental age does have a significant effect on expression (parental age genes).

temperature. Because temperature and parental age have effects on progeny phenotypes, we wanted to test the extent to which the association between gene expression and reproductive traits was driven by environmental history (parental age or early-life temperature) for the identified genes. Previous work on parental age found that maternal age affects vitellogenin gene expression to impact progeny phenotypes[27], but how other genes may be affected is not yet known. To determine effect sizes, we calculated the standardized beta coefficient for each gene using a linear mixed model without environmental history, in line with the analysis done in Figure 1 (β_1 in Fig. 2A-B). We then generated a linear mixed model that incorporates environmental history and calculated the standardized beta for the expression component of

129 the model, effectively isolating gene expression 'noise' that is not explained by environmental history (β_2 in
130 Fig. 2A-B). These two approaches to calculating standardized beta values for expression are illustrated as
131 path analyses in Fig. 2A-B. Model 1 considers expression alone and its association with brood, while Model
132 2 is a causal model that incorporates environmental history and its effects on gene expression. Equations
133 for these models, including the coefficients for $\beta_3 - \beta_6$, can be found in the methods. Values for all
134 coefficients can be found in Supplementary File 1. Here we primarily focus on standardized beta values β_1
135 and β_2, because these respectively represent 1) the total effect of expression on each trait and 2) the effect
136 of expression on each trait after controlling for historical environment. In Figure 2A, *puf-5* is shown as an
137 example gene that is part of the group of 448 genes significantly associated with brood. The β_1 value of
138 0.39 for *puf-5* is significant and indicates that a 1 standard deviation change in gene expression is
139 associated with a 0.39 standard deviation change in early brood. The β_2 value of 0.26 for *puf-5* is also
140 significant, meaning that after accounting for historical environment, a 1 standard deviation change in gene
141 expression is associated with a 0.26 standard deviation change in early brood. The standard deviation in
142 early brood is 20.3, so this indicates that these gene expression changes associated with *puf-5* account for
143 approximately 5-8 progeny produced in the first day of egg-laying, which is a sizable effect on fitness.
144 Similarly, in Figure 2B, *grsp-4* is shown as an example gene with a significant association with egg-laying
145 onset both before and after accounting for environmental history. The absolute values of β_1 and β_2 for all
146 significant genes are shown in Figure 2C-D. Early brood genes had a median absolute value of β_1 of 0.38
147 and a median absolute value of β_2 of 0.23, indicating that for a 1 standard deviation change in gene
148 expression for a typical significant gene, there was a corresponding 0.38 or 0.23 standard deviation
149 change, respectively, in early brood. Egg-laying onset genes had a median β_1 value of 0.35 and β_2 value of
150 0.31. Thus, while we are able to identify many more genes for early brood compared to egg-laying onset,
151 genes for both traits have largely comparable effect sizes. Among significant genes for both traits, β_2 values
152 were consistently lower than β_1 (Fig. 2C-D), suggesting some of the total effect size was driven by
153 environmental history rather than pure noise. Nonetheless, β_2 values were often still highly significant with
154 substantial explanatory power, indicating that gene expression noise does indeed explain differences in
155 reproductive traits. For egg-laying onset, 5 of 11 genes have highly significant values of β_2 (Bonferroni
156 corrected p-value of less than 0.05), though all are nominally significant. Of the 448 genes associated with

157 early brood, 157 genes have highly significant values of β_2, indicating that gene expression noise
158 independent of historical environment is driving associations for these genes.

159 In our previous analysis, we identified genes for which variation is associated with early brood or
160 egg-laying onset and determined which of these genes had expression variation that was not explained by
161 the historical environment (i.e., the 157 noise genes). Because of the unique structure of our dataset, we
162 can also directly ask whether the historical environment has effects on gene expression, regardless of
163 whether these expression changes lead to changes in our reproductive traits. To address this, we used a
164 negative binomial mixed model to identify genes that are differentially expressed due to parental age or
165 early-life temperature. We identified 186 genes with expression changes significantly affected by parental
166 age, and just 2 genes significantly affected by early-life temperature. 140 of the 186 genes differentially
167 expressed due to parental age in the previous generation overlapped with the original 448 genes for which
168 expression variation is associated with early brood, suggesting that the historical environment, i.e. parental
169 age, at least partially accounts for the gene expression changes that are associated with early brood. 43 of
170 the 186 genes also overlap with our 'noise' genes, meaning they exhibit gene expression variability
171 significantly explained by *both* historical environment and by noise; that is, these two sources of variation
172 operate together for these genes. Thus, among genes for which expression variation is associated with
173 early brood, we can gain insight on whether this variability in primarily driven by expression 'noise',
174 historical environmental perturbations, or both.

175 We next assessed Gene Ontology (GO) enrichments for genes associated with reproductive trait
176 variation. We focused on early brood genes because hundreds of genes were identified, giving us power to
177 detect enriched terms. We evaluated GO terms for all 448 genes as well as two subsets of the 448 genes:
178 1) genes that have a significant value of β_2 and a non-significant effect of parental age on gene expression
179 (114 genes, the 'noise' group) and 2) genes that have a non-significant value of β_2 and a significant value
180 of parental age on gene expression (97 genes, the 'parental age' group). Enrichments for all three gene
181 groups are shown in Figure 2E. We find that the majority of enriched terms for the 448 genes are also
182 enriched in either the noise group or parental age group, in a mutually exclusive manner (i.e. no term is
183 enriched in both subsets), suggesting that these two sources of gene expression variation may, to some
184 extent, be regulated by distinct processes. Indeed, the enriched parental age terms all impinge on

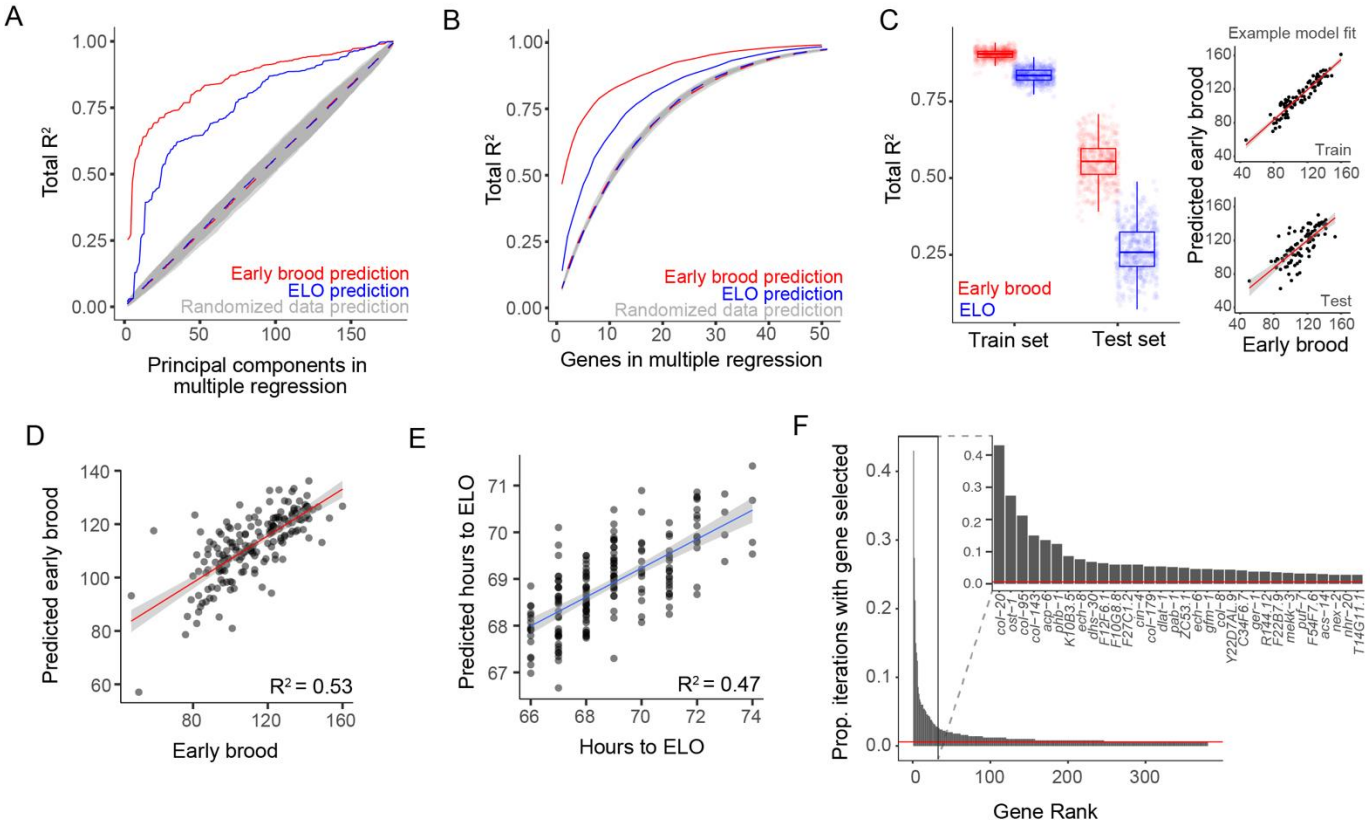


Figure 3. Multi-gene models are highly predictive of early brood and egg-laying onset at an individual level.
A. Total R^2 of gene expression principal components (PCs) in a multiple regression. PCs explain early brood (red), ELO (blue), and randomized data (dotted lines and gray shading).
B. Total R^2 of top genes in a multiple regression. Gene sets explain early brood (red), ELO (blue), and randomized data (dotted lines and gray shading).
A-B. Dotted lines represent average of randomized data iterations and gray shading indicates standard deviation.
C. Total R^2 of top 10 genes in a multiple regression identified in a train set and total R^2 of the same 10 genes used in a test set. Train and test sets randomly selected 500 times and top 10 genes identified in each iteration. Box plot center lines indicate median and box limits indicate upper and lower quartiles. An example of one iteration of train and test sets and corresponding model fits are shown for early brood.
D-E. Machine learning model elastic net regression and leave-one-out cross validation used to identify a predicted trait for each worm given transcriptomic profile and compared to true trait data. Linear fit used to determine R^2. Lines indicate linear fit, and gray indicates 95% confidence intervals.
D. Early brood model, E. ELO model
F. Proportion of 500 iterations from C in which a given gene is selected as one of the top 10 predictive genes for early brood, ordered by those selected from most to least often, and an inset showing the genes that are most frequently selected as predictive.

185 mitochondrial regulation, consistent with parental age affecting mitochondria-related gene expression in
186 progeny. Notable categories among enriched noise terms include cytoskeleton organization and female
187 gamete generation, with the latter category particularly relevant for affecting progeny production.

188 The analysis shown in Figure 2 provides a framework to understand individual gene expression as
189 useful both for prediction and understanding causal mechanisms. Toward the goal of using gene
190 expression profiles to predict phenotypes, we next asked if we could use *aggregate* gene expression
191 profiles for combinations of genes to predict the reproductive traits of individuals. For example, the β_1 value

192 of a single gene provides an indication of to what extent the phenotype is associated by that single gene. If
193 instead of generating individual regressions for each gene, we combined genes into a single model, would
194 we be able to better predict phenotypes? Because we are focused on prediction rather than the distinction
195 between expression noise and the historical environment in this analysis, we focus on an analysis that is
196 agnostic to environmental history, but we emphasize that all worms are genetically identical and collected
197 while in a shared environment at the same developmental stage. Principal component analysis (PCA)
198 provides one such aggregate profile which simplifies gene expression data to linearly independent factors
199 (PCs) that together explain the variation in the gene expression data and are ordered by how much
200 variation they explain (*i.e.*, PC1 explains the most variation, PC2 explains the second-most, etc.). We
201 therefore first asked whether ordered PCs together in a multiple regression explain more variation than
202 expected by chance. Indeed, gene expression PCs explained both ELO and early brood data substantially
203 better than randomized ELO and early brood data, indicating that individual gene expression profiles
204 provide a clear signal of phenotypic information (Fig. 3A). Because we have phenotypic and transcriptomic
205 data for 180 worms, and therefore up to 180 linearly independent PCs that can be added a multiple
206 regression model, the total R^2 eventually rises to 1 even for randomized data, as shown in Fig. 3A.
207 Therefore, the optimal number of PCs occurs at the inflection points of the graph, which is after only 7 PCs
208 for early brood (R^2 of 0.55) but 28 PCs for ELO (R^2 of 0.56). Thus, for early brood in particular, very few
209 PCs explain over half of the variation in the data and explain substantially more variation than for shuffled
210 phenotypic data. Therefore, whole transcriptomes are a viable approach to trait prediction in isogenic *C.*
211 *elegans*.
212 Because PCA incorporates data from all expressed genes, we next wanted to determine if a smaller
213 subset of genes could be similarly predictive of reproductive traits. We again used a multiple regression
214 approach in which we successively selected each gene by identifying the gene that explained the most
215 additional variance compared to the previous model (Fig. 3B). Genes that exhibit highly correlated
216 expression patterns, even if they are each individually associated with traits of interest, would not explain
217 much additional variance in a multiple regression approach. Therefore, this approach prioritizes identifying
218 genes that are most predictive in conjunction with other genes. We found relatively small sets of genes
219 predict ELO and early brood phenotypes very well, and do so significantly better than randomized trait
220 data, with diminishing returns for variance explained as the number of genes increases (Fig. 3B). Maximum

221 differences in the proportion of variance explained between trait data and randomized data occur in a
222 model with only 4 genes for the early brood analysis (R^2 difference of 0.433) and with 7 genes in the egg-
223 laying onset analysis (R^2 difference of 0.195). While less powerful than utilizing the entire transcriptome,
224 this result highlights that individual worm gene expression profiles consisting of only a handful of key
225 selected genes are highly informative for predicting the individual's traits.

226 We next wanted to develop a quantitative estimate of how much trait variation is explained by
227 aggregate gene expression profiles using two distinct approaches that prevent model over-fitting and would
228 therefore be more feasible to use for trait prediction in a new dataset. First, as an extension of the multiple
229 regression analysis identifying predictive genes, we randomly split the full dataset (consisting of trait data
230 and gene expression profiles) into a training set and test set. In the training set, we used the multiple
231 regression approach to identify the set of 10 genes that explained the most variance in the trait data. We
232 then asked how well this set of 10 genes explained variance in the test set and repeated this process 500
233 times, each time randomly splitting the data into training and test sets (Fig. 3C). We found that median
234 proportion of variance explained in the test sets was 0.55 (range 0.36 to 0.73) for early brood and 0.26
235 (range 0.07 to 0.49) for ELO, and an example of the fit of one of these iterations is also shown (Fig. 3C).
236 Thus, sets of only 10 genes explain over half of trait variation in early brood and over a quarter of variation
237 in ELO using test sets of data that were not used to select predictive genes. All 500 sets of 10 predictive
238 genes are available as part of Supplementary File 1. As a second approach, we used elastic net regression
239 and leave-one-out cross validation on the full set of expressed genes. In other words, we used a machine
240 learning approach designed to prevent overfitting to develop a model for all individuals except one, used
241 the model to predict the trait for the remaining individual, then repeated for all individuals. Consistent with
242 the multiple regression approach, the proportion of variance explained was 0.53 for early brood and 0.47
243 for ELO (Fig. 3D-E). Aggregate gene expression profiles identified via multiple approaches are thus highly
244 predictive of reproductive traits among isogenic individuals in a shared environment, reliably explaining
245 about half of reproductive trait variation among genetically identical individuals.

246 A key remaining question is whether predictive genes are causal to affect phenotypes and whether
247 similar underlying regulation underlies expression variation in predictive genes. To address causality, we
248 focused on the early brood trait because predictive genes explained a higher proportion of variance for this
249 trait compared to ELO. We used RNA interference (RNAi) to reduce the level of gene expression of

250 predictive genes and prioritized genes that were also in the 448 brood-associated genes, either highly
251 independent of parental age or dependent on parental age, and have an ortholog in humans. Within this
252 curated set, genes causally affected early brood in 5 of 7 cases compared to empty vector (Fig. 4A). We
253 followed up on one pair of genes from this screen, *puf-5* and *puf-7,* to ask whether perturbing the dose of
254 RNAi shows a dose response effect on brood. These genes have previously been shown to act
255 redundantly in fertility in *C. elegans*[30,31] and when knocked down simultaneously in our assay have a large
256 effect on brood without causing sterility (Fig. 4A). Consistent with the idea that differences in the level of
257 mRNA affect the early brood trait, the dose of RNAi corresponds to the effect size on brood (Fig. 4B). Thus,
258 this analysis reveals that the predictive genes also make excellent candidates as causal effectors of
259 complex traits and can affect these traits dependent on the amount of mRNA available.

260 Expression of predictive genes can be causally related to reproductive traits (Fig. 4A-B), but do
261 these genes share underlying regulation? The majority of *C. elegans* genes fall into one of two chromatin
262 domains with distinct levels of gene activity[32–34]. Regulated domains include genes that are expressed at
263 specific stages of development and are marked by the facultative heterochromatin modification H3K27me3.
264 In contrast, active domains include genes that are expressed broadly across cell types, including in the
265 germline, and marked by the euchromatic modification H3K36me3[32,34,35]. These two marks are generally
266 mutually exclusive, and marks are robust across several stages of development in *C. elegans*[34] despite
267 major changes in gene expression that occur during this time. Importantly, the robustness of these
268 chromatin marks throughout development means that we can use existing data to determine whether
269 genes in active or regulated domains are differentially associated with gene expression variation linked to
270 reproductive traits in our study. We therefore used genes previously classified as regulated or active to
271 gain insights into the epigenetic status of genes associated with and predictive of reproductive traits.
272 Because active genes are enriched for being expressed in the germline, we performed a two-factor
273 analysis that takes into account both tissue (soma or germline) and chromatin domain (regulated and
274 active). Among early brood-associated genes, germline genes are strongly positively associated with
275 brood, regardless of chromatin environment, but chromatin environment and tissue also exhibit a significant

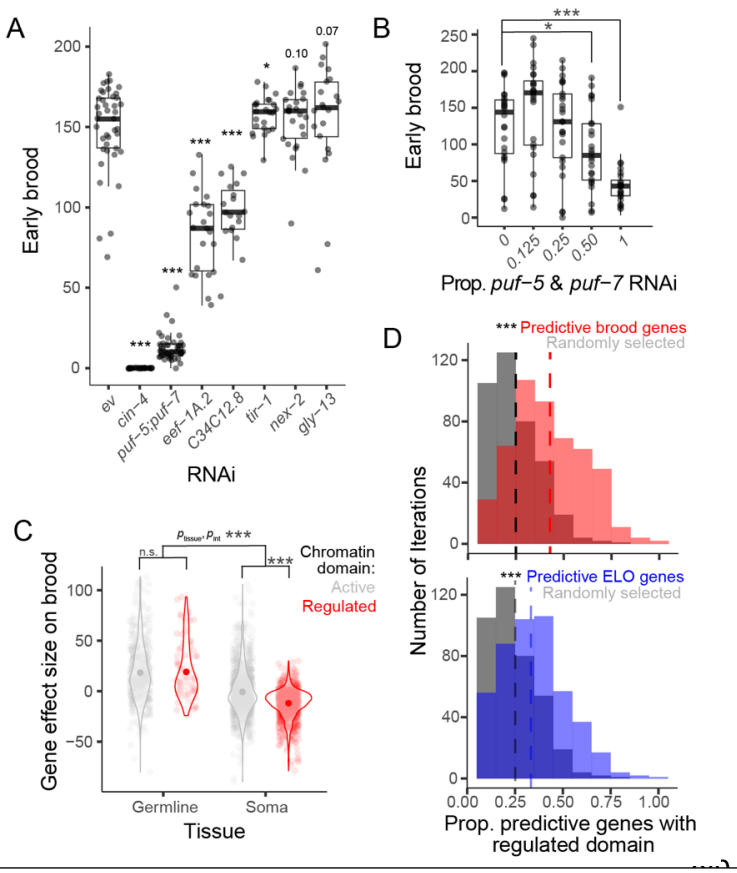


Figure 4. Genes predictive of reproductive traits causally affect early brood and are enriched for H3K27me3
A. RNAi knockdown of selected predictive genes compared to empty vector (ev) and effects on early brood.
B. Dose response of *puf-5* and *puf-7* RNAi together with empty vector and effects on early brood.
A-B. Linear mixed model with RNAi as fixed effect and biological replicate as a random effect. Center lines, median; box limits, upper and lower quartiles. *p<0.05, ***p<0.001
C. Effect sizes on brood for genes categorized by tissue (soma or germline) and chromatin domain (active or regulated). Two-way ANOVA showed significant interaction between tissue and chromatin domain ($p = 1.58 \times 10^{-6}$) and a significant main effect of tissue ($p < 2 \times 10^{-16}$). Post-hoc Tukey tests corrected for multiple testing showed a significant difference between chromatin domains within somatic tissues (adjusted p = 0) but not germline tissues (adjusted p = 0.98). Within somatic genes, active somatic genes effects on average do not differ from 0 (p = 0.3, one-sample t-test with $\mu = 0$), while regulated somatic genes on average have a significant negative association with brood ($p < 2.2 \times 10^{-16}$, one-sample t-test with $\mu = 0$). ***p<0.001
D. Histogram of all iterations identifying predictive gene sets and the proportion of each set in a regulated chromatin domain marked with H3K27me3 compared to randomly selected sets of genes from the same background set. Early brood iterations are shown in pink, ELO in blue, and the same gray control is shown for both histograms. The vertical lines are color-coded to indicate the median proportion of genes in regulated domains in the corresponding condition. Kolmogorov-Smirnov test used to determine statistical significance. ***p<0.001

interaction (Fig. 4C). Post-hoc pairwise comparisons reveal that this interaction is driven by differences in the chromatin environment in the soma. In particular, while active somatic genes have on average no effect on brood, regulated somatic genes are strongly negatively associated. Because regulated genes are typically expressed in a temporally controlled or tissue-specific manner in the soma, this may indicate that aberrant expression of these genes has negative consequences for reproduction, or similarly, that appropriate regulation and silencing of these genes promotes reproduction. To determine if epigenetic regulation may underlie variation in predictive genes for both traits overall, without regard to the direction of association, we used all gene sets identified from iterations of predictive sets (Fig. 3C,F) to determine whether these genes are more likely to be found in regulated domains. For both reproductive traits, predictive genes were strongly over-enriched for being in a regulated domain compared to randomly selected gene sets from the same background set (Fig. 4D). Genes in regulated domains are more likely to be variable across developmental stages and cell types[34], and we find in our data that genes with the most expression variation across individuals unexplained by experimental

305 factors are also more likely to be found in regulated domains (Fig. S3). Taken together, these results
306 suggest that regulated domains are subject to gene expression variation in genetically identical individuals
307 and that this variation is an important substrate underlying complex trait variation.
308
309 **Discussion**
310 Individuals differ from each other for a variety of traits over the course of their lives, but being able
311 to reliably predict these differences and understand their causal basis is a challenge. Gene expression is a
312 critical intermediate functional layer between genotype and phenotype, but the role that gene expression
313 plays in shaping individual organismal traits independently of genetic variation and environmental factors
314 has remained elusive. Here we effectively screened for genes that exhibit expression variation between
315 isogenic *C. elegans* individuals and for which this variation corresponds to two critical reproductive traits –
316 time to egg-laying onset and early brood. We identified hundreds of genes associated with these traits,
317 differentiated between expression variation that is dependent and independent of historical environmental
318 factors, and defined sets of highly predictive and causal genes. Predictive genes for both traits are
319 enriched for a particular chromatin modification – H3K27me3 – highlighting the possibility that chromatin
320 structure may underlie the potential for gene expression to vary between individuals. These findings build
321 upon past research and suggest key areas for future work.
322 With our framework, we identify individual differences in gene expression that are associated with
323 reproductive traits, and we parse whether these effects are primarily driven by noise or by known historical
324 environmental factors. For example, *puf-5* and *puf-7* expression is highly associated with early brood after
325 accounting for the known historical environment, and knockdown of these genes causally affects early
326 brood in a dose-dependent manner. On the other hand, genes like C34C12.8 exhibit an overall association
327 with early brood that is primarily driven by the effect of parental age on gene expression; nonetheless,
328 C34C12.8 is also causally linked to early brood. Thus, by experimentally manipulating and measuring
329 *individuals*, our approach is highly efficient in dissecting multiple sources of environmental variation --
330 including noise and known environmental perturbations – underlying multiple traits that would otherwise be
331 obscured by bulk experiments. In both cases, the mechanism(s) generating gene expression differences
332 across individuals remains unknown. For 'noise' genes in particular, whether differences in gene
333 expression are driven by extrinsic or intrinsic factors is unclear. While every effort was made to control the

334 environment for our experiment, isogenic individuals nonetheless exhibit some extrinsic differences in their
335 lives, such as how they choose to wander on a plate of food, that may to some extent impact gene
336 expression and reproduction. We further expect that some of the variation observed is intrinsic, based on
337 literature exploring the stochastic nature of transcription[36], as well as examples that suggest the amount of
338 phenotypic variation present among individuals of a given genetic background is itself genetically
339 controlled[37]. Continued experimentation to determine the extent to which intrinsic or extrinsic factors
340 regulate gene expression variability in this context -- where such variability is directly related to complex
341 phenotypes -- is of high interest for future work.

342 When genes expression differences are identified across individuals by single-worm mRNA-seq,
343 such differences could occur because individuals have different numbers of cells of a given type (e.g.,
344 germ cells) but each cell express genes at consistent levels, or alternatively, that individuals differ because
345 within each cell, additional copies of a gene are being transcribed while cell numbers themselves remain
346 constant. Of course, a combination of these scenarios is possible as well. *C. elegans* uniquely possess an
347 invariant somatic cell lineage[38], but can exhibit differences in cell number in the adult germline, including in
348 the primordial zone[39] or in the uterus [40]. Our results show that, in general, genes expressed in the germline
349 tend to be positively associated with early brood, while somatic genes tend to be negatively associated with
350 brood. Of somatic genes, genes marked with H3K27me3 are particularly predictive of early brood. Somatic
351 genes marked with H3K27me3 are also more likely to exhibit gene expression variability in general. Given
352 that the somatic lineage is invariant and somatic genes marked with H3K27me3 are the most variable in
353 expression and most predictive of early brood, it is unlikely that cell number differences are the major driver
354 of gene expression differences. While single-cell RNAseq has been performed in *C. elegans*, such analysis
355 requires thousands of individuals to be pooled to obtain single cells and is not technically possible at the
356 level of single individuals. Thus, to fully understand cell-type specific variation in single *C. elegans*, and
357 how these relate to individuals' traits, further innovation on this front is required.

358 We observed that differences in chromatin domains are associated with differences in gene
359 expression and traits across individuals. Regulated genes are defined by the histone modification
360 H3K27me3, a mark that is robust across developmental stages despite differences in gene expression,
361 while active genes are marked with H3K36me3 and are typically broadly expressed, including in the
362 germline. In our data, regulated genes are enriched both among genes with the most unexplained variation

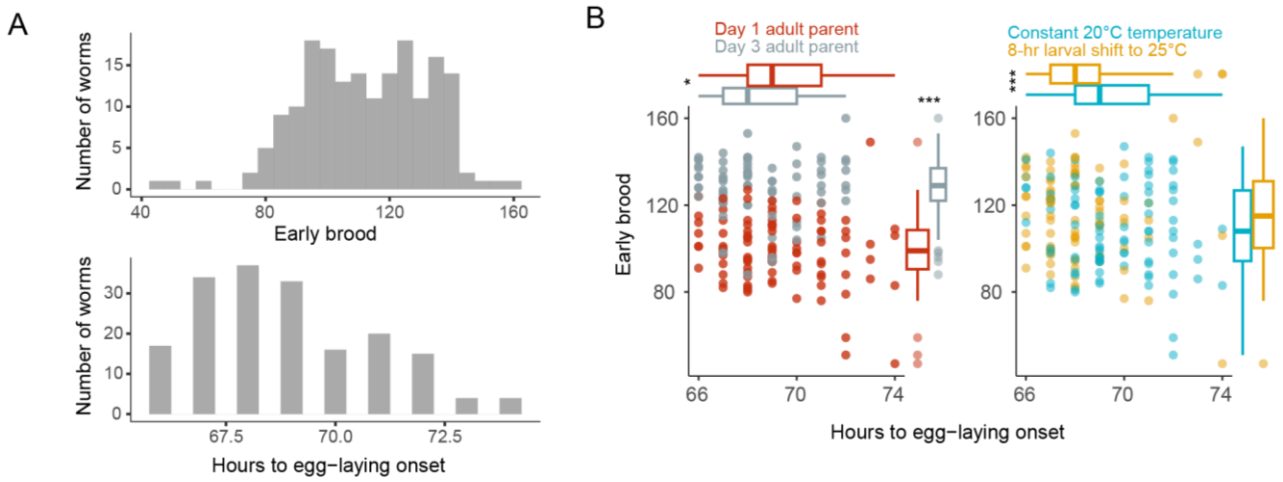
363 and among genes for which variation is predictive of reproductive traits. In contrast, germline genes that
364 predominantly consist of genes in active domains are specifically positively associated with brood. Taken
365 together, one possible model is that regulated genes marked with H3K27me3 are more subject to inter-
366 individual variation than genes marked with H3K36me3, which may lead to differences in gene expression
367 without changing the levels of these marks. This is consistent with observations that regulated genes
368 exhibit a higher coefficient of variation across conditions and that increased expression of genes marked
369 with H3K27me3 is not associated with a change in histone post-translational modifications (*e.g.,* to an
370 active mark like H3K36me3)[41]. An alternative hypothesis is that the levels of the marks themselves vary
371 and that this leads to changes in gene expression across individuals. While it is clear that regulated genes
372 are more subject to inter-individual variability in expression, at this stage it is unknown whether this
373 variability is driven by differences in the level of chromatin marks. Performing analysis on histone
374 modifications genome-wide in individual worms, which is still technically challenging, will be a key next step
375 to determine whether differences in chromatin structure correlate with gene expression changes at an
376 individual level.

377 Individual gene expression differences underlying reproductive traits suggest that gene expression
378 differences may be important for a variety of other complex traits. Gene expression differences among
379 isogenic individuals in a shared environment have been previously documented as predictive of lifespan[42–
380 44]. In these studies, reporter gene expression for specific candidate genes predicted individuals' later
381 lifespan. Our results extend on such studies by identifying predictive genes in the context of reproductive
382 traits in a high-throughput manner, rather than relying on specific reporters. Our approach can in theory be
383 used directly in the context of other traits that occur throughout the life cycle, such as those related to
384 development, but must be used indirectly for time-to-death traits like lifespan or stress resistance. Because
385 a worm, once collected for RNA, cannot be subsequently phenotyped for traits like lifespan, using
386 predictive reporters for time-to-death traits in conjunction with transcriptomics will likely be a useful
387 approach for identifying novel regulators for a broad swath of traits.

388 Genetically identical individuals can have large differences in phenotype if gene expression varies
389 across those individuals. Our work illustrates that differences in gene expression can be causally linked to,
390 and are highly predictive of, important reproductive phenotypes among genetically identical individuals.
391 Assaying isogenic individuals for both organismal and molecular phenotypes enables us to uniquely

392 understand the basis of non-genetic variation underlying individual differences and provides an avenue for
393 high-throughput phenotyping by transcriptomics[17]. Given the major role of non-genetic factors in regulating
394 complex traits and diseases in humans, identification of these molecular drivers of variation may be a
395 powerful next step for advancing precision medicine.

396 **SUPPLEMENTARY FIGURES**



397

398 Figure S1. Isogenic worms at the same developmental stage exhibit variability in egg-laying onset and

399 early brood that partially depends on previous environmental conditions

400 A. Univariate histograms of early brood and egg-laying onset trait values for all worms

401 B. Bivariate scatterplots of early brood and egg-laying onset trait for all worms. The same data is plotted

402 twice but color-coded according to two previous environmental perturbations (age of parent and early-life

403 temperature). Center lines for box plots indicate median; box limits indicate upper and lower quartiles.

404 Significance determined using a linear mixed model with environmental perturbation as a fixed effect and

405 biological replicate as a random effect, ***$p<0.001$, *$p<0.05$

406

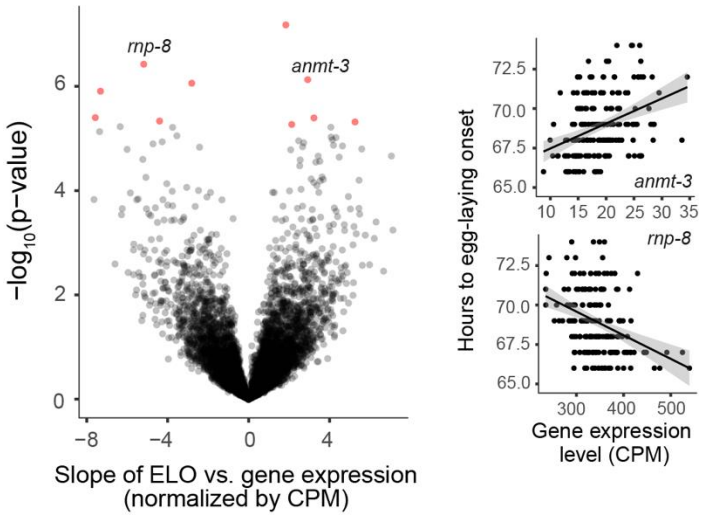


Figure S2. Variation in 11 genes is significantly associated with egg-laying onset.

Each of 8824 expressed genes was analyzed in a mixed model to assess the strength of its association with egg-laying onset. Significant genes are in red, non-significant genes are in black. Significance determined by a Bonferroni-corrected p-value cutoff of 0.05 (nominal p-value of 5.67×10^{-6}). Phenotypic and expression data for each worm is shown for two exemplar genes, *anmt-3* and *rnp-8*. Black lines from linear model fit and gray shading indicates 95% confidence intervals.

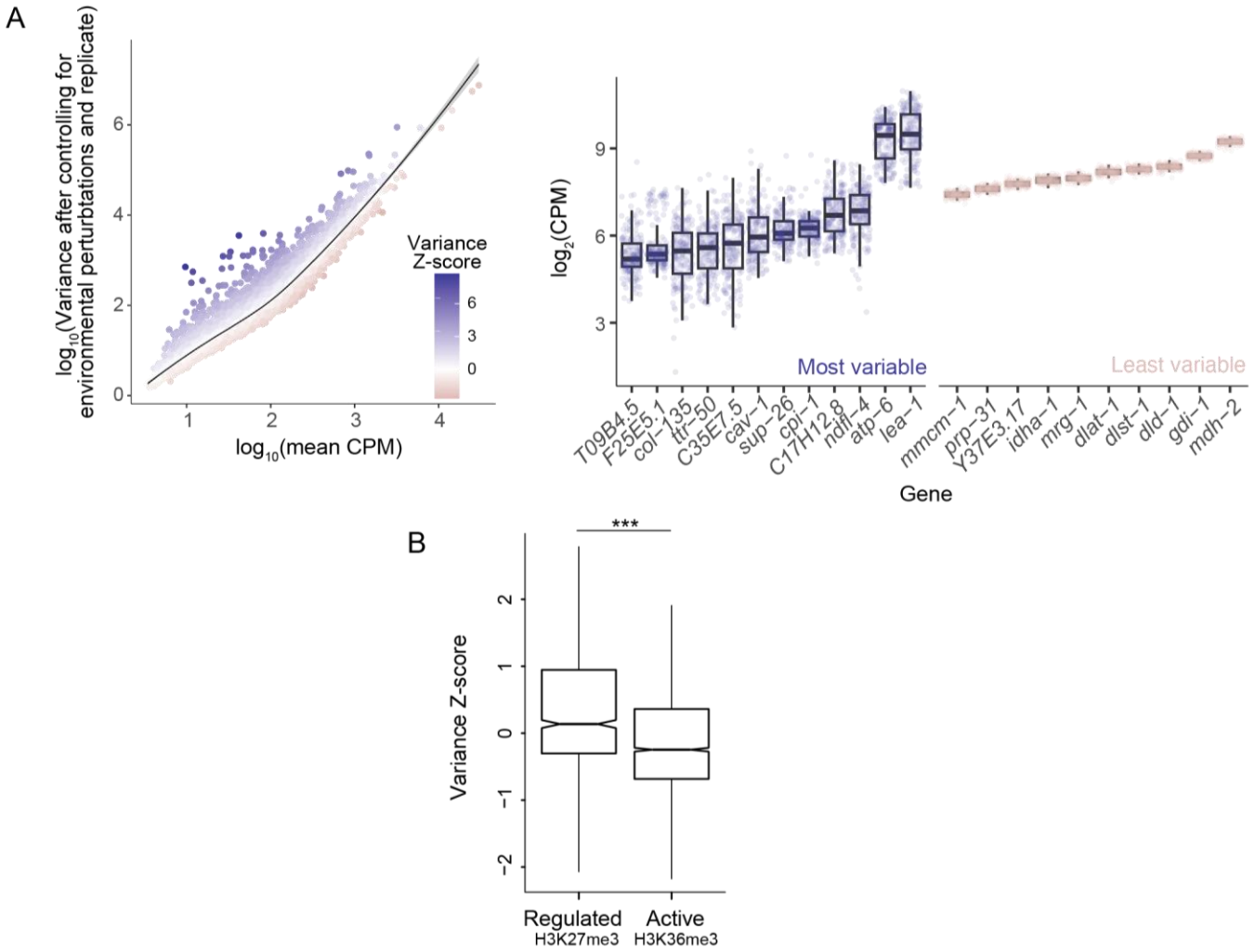


Figure S3. Single-individual mRNA-seq reveals most and least variable genes after controlling for environmental perturbations and biological replicate.

A. Log_{10}-transformed unexplained variance in mRNA-seq after controlling for environmental factors and biological replicate is plotted against log_{10}-transformed average CPM across all worms and a loess fit line is shown in black with a 95% confidence interval in gray. Each point is a different gene and genes are color-coded according to their distance from the fit line. Genes in purple are most variable given their level of expression, while genes in pink are the least variable. Exemplar genes with comparable expression levels are shown on the right. Center lines for box plots indicate median; box limits indicate upper and lower quartiles.

B. Variance Z-scores for genes with regulated domains (H3K27me3) compared to variance Z-scores for genes with active domains (H3K36me3). Wilcoxon test, ***$p < 2 \times 10^{-16}$

428 **METHODS**

429 *Strains used*

430 Strain N2-PD1073 was used for all experiments. N2-PD1073 is a subclone of the N2-derivative
431 strain VC2010 that was generated in the process of assembling a new N2 reference. Worms were
432 passaged on *E. coli* OP50 except during RNAi experiments, when worms were fed HT115 bacteria
433 expressing dsRNA for genes of interest (see 'RNAi' section).

434

435 *Experimental set-up for single-worm RNA-seq*

436 Worms were well-fed at 20°C for several generations prior to beginning experiments. The
437 experiment takes place over the course of 10 days. On day 1, for each of 5 biological replicates, P0
438 embryos and L4s were picked to separate 5 cm NGM plates with OP50 and kept at 20°C. Approximately
439 10-20 individuals were placed on each plate and several plates were picked for each stage, ensuring that
440 worms do not experience dietary restriction. On day 4, P0 worms were either in the first day of adulthood
441 ('young adult') or third day of adulthood ('older adult'). P0 young adults and older adults were singled to
442 new plates, allowed to lay progeny for two hours at 20°C, then removed from the plate. Hour 0 of
443 development for F1 worms is considered the end of this egg-laying window. F1 worms were then incubated
444 at 20°C until day 5 of the experiment (undergoing approximately 18 hours of development as L1/L2 larvae).
445 On day 5, approximately half of these larvae remained at 20°C while the other half were shifted to 25°C for
446 8 hours then shifted back to 20°C. On day 6 all worms remained at 20°C and were singled to individual
447 NGM plates with OP50 to ensure they can be individually phenotyped for reproductive traits. On day 7,
448 from 66-74 hours of development, egg-laying onset was assessed every hour by checking for embryos on
449 each plate. Worms that died, bagged, or otherwise did not reach egg-laying onset in this time interval were
450 not included in subsequent analysis, and all together this affected 6% of worms across all biological
451 replicates. Time to egg-laying onset was noted for each worm. 24 hours after egg-laying onset, on day 8,
452 each worm was collected for cell lysate generation for mRNA-seq (see 'Cell lysate generation' section of
453 Materials and Methods). On day 10, the number of progeny laid on each plate during the 24 hours between
454 egg-laying onset and collection were counted as late larvae/young adults to ensure all progeny were
455 counted. All together, each F1 worm collected experienced two environmental perturbations (young adult

456 or older adult parent, and early-life constant temperature or temperature shift) and was measured for two
457 traits (egg-laying onset time and number of progeny produced within 24 hours) prior to mRNA-seq.
458
459 *Cell lysate generation*
460 Individual worms were collected for chemical lysis 24 hours after egg-laying onset (that is, worms
461 that reached egg-laying onset at 72 hours were collected at 96 hours, worms reaching egg-laying onset at
462 73 hours were collected at 97 hours, etc. for each hour). 200 μL of lysis mix, consisting of 187.2 μL Elution
463 buffer (10 mM Tris, 0.1 mM EDTA) and 12.8 μL Proteinase K, was made fresh every 2-3 hours on ice. 6 μL
464 of lysis mix was pipetted into each PCR tube according to how many worms needed to be collected in a
465 given hour. Each worm was picked into a 10 μL drop of dH2O for a quick wash to remove any excess
466 bacteria, then transferred to a second drop of dH2O. After this, each worm was picked into an individual
467 PCR tube with lysis mix. Immediately after worms are placed into tubes with lysis mix, we proceeded with
468 the lysis protocol (typically no more than a dozen worms were prepared at once). The lysis protocol is 65°C
469 for 10 min, 85°C for 1 min, and 4°C hold and is based on a previously published protocol[22]. After the lysis
470 protocol, worms were moved to ice and then stored at -80°C until mRNA-seq library preparation.
471
472 *Preparation and sequencing of mRNA-seq libraries*
473 The KAPA mRNA-seq preparation kit (KK8580, Roche Sequencing Solutions) was used to prepare
474 libraries. 5 μL of each cell lysate preparation described above, which contains total RNA, was used for
475 each sample. Reagents were used at 0.25X and kit instructions were followed to generate libraries. Library
476 concentrations were quantified using Qubit, and samples were pooled together in equimolar ratios in two
477 batches of 96 for sequencing. Custom dual-indexed adapters were designed by the University of Oregon
478 Genomics and Cell Characterization Core Facility and sequencing data was generated with the Illumina
479 NovaSeq 6000 at the same facility.
480
481 *Analysis of single-worm mRNA-seq data*
482 FASTQ files were aligned to WBcel235 version of the *C. elegans* genome using subread-align[45]
483 with flag -t 0. Bam files were counted using featureCounts with flag -T 5. 12 of 192 samples were filtered

484 out at this stage due to particularly low read counts (less than 1 million reads) and/or poor mapping quality
485 (less than 70% of reads both mapped to the genome and assigned to a genomic feature). The remaining
486 180 samples had an average of 5.8 million reads (SD 3.4 million), and an average of 85.8% (SD 3.1%) of
487 reads both mapping to the genome and assigned to a genomic feature. Resulting count data from these
488 180 samples was CPM-normalized in R using the package edgeR[46] and only protein-coding genes were
489 included. Depending on the downstream analysis, genes were filtered in one of two ways for further
490 analysis: 1) genes expressed at a level of CPM > 10 in at least one library (8824 genes) or 2) genes
491 expressed at a level of CPM > 1 across all libraries (7938 genes). While these two sets are largely
492 overlapping, the first way of filtering allows identification of genes that are not expressed at all in some
493 samples but moderately expressed in others. However, as models become more complex, having some
494 samples with 0 expression causes problems for model fitting and thus having a more stringently-defined
495 set is more appropriate. To identify genes associated with early brood and egg-laying onset, we looped
496 through the set of 8824 genes and tested two linear mixed models using the 'nlme' package[47] in R to test
497 the effect of each gene on each of two reproductive phenotypes. In both cases the fixed effect was the
498 normalized gene expression level in each worm and the random effect was biological replicate. For the first
499 model, the dependent variable was early brood and for the second model the dependent variable was
500 hours to egg-laying onset. To determine the effect of environmental perturbations on gene expression, we
501 looped through each gene and fit a negative binomial mixed model to each gene. In this case, for each
502 gene, fixed effects included environmental perturbations (parental age and early-life temperature), the
503 random effect included biological replicate, and the dependent variable was gene expression level. For
504 both sets of models, the summary() function was used to generate p-values, and a Bonferroni correction
505 was used to determine genes with significant effects. Gene expression variation that was unexplained by
506 the negative binomial (residual variance compared to the model) was used to determine the most and least
507 variable genes in the data after controlling for environmental effects and biological replicate. Because
508 variance increases with higher expression, we identified most and least variable genes by fitting a loess
509 model to $\log_{10}$(unexplained variance) vs. $\log_{10}$(mean CPM) for the 7938 genes. Z-scores were calculated
510 for each gene relative to the loess fit to determine genes that were most and least variable given their level
511 of gene expression.
512

513 *Path analysis and effect size calculations*

514 Using the R package 'lme4'[48], for each individual *i* and biological replicate *j*, the following linear
515 mixed models were fit for each gene *k*:

516 $(\text{Early brood})_{ij} = \beta_0 + \beta_1(\text{Expression level of gene } k) + u_j + e_{ij}$

517 $(\text{Early brood})_{ij} = \beta_0 + \beta_2(\text{Expression level of gene } k) + \beta_3(\text{Parental age}) + \beta_4(\text{Early-life temperature}) + u_j + e_{ij}$

518 $(\text{Egg-laying onset})_{ij} = \beta_0 + \beta_1(\text{Expression level of gene } k) + u_j + e_{ij}$

519 $(\text{Egg-laying onset})_{ij} = \beta_0 + \beta_2(\text{Expression level of gene } k) + \beta_3(\text{Parental age}) + \beta_4(\text{Early-life temperature}) + u_j$
520 $+ e_{ij}$

521 $(\text{Expression level of gene } k) = \beta_0 + \beta_5(\text{Parental age}) + \beta_6(\text{Early-life temperature}) + u_j + e_{ij}$

522 Parental age was binarized such that progeny of day 1 adults were given the value of 1 and progeny of day
523 3 adults were given the value of 0. Early-life temperature was binarized such that individuals experiencing
524 a constant 20°C temperature throughout their lives were given the value of 1 and individuals that
525 experienced a 25°C temperature shift were given the value of 0. All values of β_n were standardized such
526 that a 1 standard deviation change in the independent variable represents a shift in the dependent variable
527 corresponding to the value of β_n. u_j represents the random intercept for each biological replicate, and e_{ij} is
528 the error term. Coefficients β_1 through β_6 are shown in a path analysis format in Figure 2A-B, and
529 coefficients β_1 and β_2 are shown graphically in Figures 2C-D for all significant genes from Figure 1 and
530 Figure S2. The values of all coefficients for all significant genes are in Supplementary File 1.

531

532 *Gene Ontology (GO) analysis*

533 GO Term analysis was performed using the R package 'clusterProfiler'[49]. The enrichGO function
534 was used for each gene list with the following parameters: OrgDb = org.Ce.eg.db, keyType = "SYMBOL",
535 ont = "BP", pAdjustMethod = "BH", qvalueCutoff = 0.05. Output for all three gene lists is included in
536 Supplementary File 1. Figure 2E shows all enriched terms with a q-value < 0.01 for the list of 448 genes
537 and any enriched terms that overlap with this set from that are enriched from the subsets of 97 parental
538 age genes or 114 noise genes. The set of 97 genes was generated by identifying genes in the set of 448
539 that were significantly affected by parental age and did not have a significant value of β_2. The set of 114
540 genes was generated by identifying genes in the set of 448 that were not significantly affected by parental
541 age and had a significant value of β_2.

542

543 *Prediction analysis*

544 Principal component analysis was performed on 7938 CPM-normalized genes (without regard to
545 reproductive phenotype data). CPM values for each gene were mean-normalized and log2-transformed
546 prior to use in the prcomp function in R. To determine whether principal components explained variance in
547 the reproductive traits, each principal component was successively added to a multiple linear regression
548 model, beginning with the principal component explaining the most variance in the gene expression data
549 (*i.e.,* PC1), to determine the total variance explained by cumulative principal components. As a control
550 comparison, phenotypic data was shuffled repeatedly (100x) and the same multiple regression analysis
551 was performed. This analysis was performed for both the early brood trait and egg-laying onset trait.

552 To identify a set of genes that explains variance in each trait, all 7938 genes were looped through in
553 a linear regression model to identify the single gene explaining the most variance. Then the remaining
554 7937 genes were looped to identify which gene explained the most additional variance when the first gene
555 is already included in the model. This was repeated until the total variance explained was close to 1. This
556 analysis was repeated for shuffled phenotypic data as a control comparison.

557 To determine how well sets of genes would be likely to work in a new dataset, we used two
558 approaches. First, we randomly split the data (including matched phenotypic data and gene expression
559 data) into two equal groups, the train set and test set, 500 times. Each time, for the train set, the protocol
560 described above was used to identify the top 10 genes that together explain the most variance in the data
561 using a multiple regression. This set of 10 genes was then used on the test set to determine how much
562 variation is explained when the genes were not selected because of how well they work in this set. The
563 typical variance explained in the test set provides an estimate of how well the genes should explain
564 variance in a new dataset. In a second machine learning approach, we used elastic net regression using
565 the glmnet package[50] in R with alpha = 0.5 and leave-one-out cross validation. That is, we trained the
566 model for all data points except for one, then used the model to predict that remaining data point, and
567 repeated for all data points.

568

569 *Enrichment analysis*

570 To determine whether brood-associated genes and variable gene sets were enriched for particular
571 chromatin environments and tissue specificities we used our previously generated annotation for *C.*
572 *elegans* transcriptional units[51]. Briefly, in this approach we used published data[32,35] to obtain tissue
573 specificity and chromatin environment for *C. elegans* regulatory elements genome wide and then used
574 bedtools2[52] to annotate regulatory elements to the nearest gene using the coordinates of transcriptional
575 units extracted from the Wormbase gff3 (version 279).
576 To determine if there was an interaction between tissue and chromatin environment for effects on
577 early brood (Fig. 4C), we restricted genes to those expressed in the soma or germline, and in either an
578 active or regulated chromatin domain. We then performed a two-way ANOVA to determine if these factors
579 affected the effect sizes for each gene on brood (the same effect sizes plotted in Fig. 1B), followed by post-
580 doc Tukey's tests for pairwise comparisons and one-sample t-tests. To determine if gene expression
581 variation differed by chromatin domain (Fig. S3B), we performed a Wilcoxon test comparing the variance z-
582 scores for regulated genes to the variance z-scores for active genes.
583 To determine whether predictive genes were enriched for regulated or active genes as shown in
584 Figure 4D, we used the 500 sets of 10 genes for each trait described in the 'Prediction analysis' section of
585 the methods that were used to predict reproductive phenotypes (Figure 3C). We filtered the chromatin
586 dataset described above to include only genes that were classified as either active or regulated. We then
587 merged this with the background of 7938 expressed genes from our gene expression analysis. Of 10
588 predictive genes, a median of 7 genes were either active or regulated for each of the 500 iterations and
589 included in subsequent analysis. For each iteration, the proportion of regulated genes was calculated. As a
590 control, 500 random sets of 10 genes were selected from our background set, merged with the filtered
591 chromatin data, and the proportion of regulated and active genes was determined. The distribution of the
592 proportion of predictive regulated genes for each trait was compared to the proportion of regulated genes
593 from random selected sets of genes using a Kolmogorov-Smirnov statistical test.
594 *RNAi*
595 RNAi was performed by feeding. Colonies of RNAi bacteria (*E. coli* HT115 containing a specific
596 RNAi construct) were individually grown in 5 mL LB with carbenicillin overnight (~16-18 hours) at 37°C with
597 shaking. RNAi bacteria was obtained from the Ahringer library[53] for empty vector (ev) and for genes of
598 interest: ZK1127.7 (*cin-4*), F54C9.8 (*puf-5*), B0273.2 (*puf-6/7),* R03G5.1 (*eef-1A.2*), C34C12.8, F13B10.1

599 (*tir-1*), T07C4.9 (*nex-2*), and B0416.6 (*gly-13*). Plates used for RNAi were made with NGM, carbenicillin,
600 and IPTG as described previously[54]. These plates were seeded with RNAi bacteria and allowed to dry and
601 grow for ~30 hours. L4 larvae were plated in groups of ~10 on these plates, then singled the next day (~16-
602 18 hours later) onto new RNAi plates prepared at the same time as the original batch of plates. Adult
603 worms were allowed to lay progeny for 24 hours, then adults were removed from the plates. If a worm died
604 during this time it was censored. The number of progeny laid by each adult was counted 2 days later,
605 resulting in an early brood measurement for these experiments. For the *puf-5* and *puf-7* dose response,
606 *puf-5* and *puf-7* RNAi constructs were grown separately as described then pooled together 1:1. This mixed
607 RNAi was considered the full dose of *puf-5* and *puf-7* RNAi and was pooled with ev in the ratios shown in
608 Fig. 4C for the varying doses. Statistical analysis on early brood data was performed by comparing RNAi
609 treatment targeting a particular gene to empty vector for biological replicates within a batch. Linear mixed
610 effects models were used in the R package 'nlme'[55] with RNAi treatment as a fixed effect and biological
611 replicate as a random effect. The summary() function was used to generate p-values.
612
613 *Validation of isogenic worms*
614 We used the mRNA-seq data to validate that the frequency of genetic variants among the
615 essentially isogenic individual worms was very low. First, we used samtools[56] to generate coverage files
616 from the sorted bam files resulting from mRNA-seq data processing. We filtered coverage files for each
617 worm to include only nucleotides that had at least 20 reads mapping to that nucleotide. For each worm and
618 nucleotide, we determined whether the base call of that nucleotide was unambiguous (all 20+ reads have
619 the same call), ambiguous with two genotypes (exactly two different bases were called), or ambiguous with
620 more than two genotypes called. All worms with sufficient depth at a nucleotide were used to determine
621 putative homozygous or heterozygous variants. A particular nucleotide was included in subsequent
622 analysis if at least 10 worms had sufficient coverage at the site. This resulted in 8,361,705 unique sites and
623 a total of 768,738,255 sites across all worms. Candidate homozygous variants included nucleotides that
624 were called as one base unambiguously in at least one worm and called as a different base unambiguously
625 in other worms. Candidate heterozygous variants were identified by filtering ambiguous nucleotides with
626 two genotypes in which the frequency of each genotype was present at a frequency of between 0.3 and 0.7
627 and for which at least 10 other worms were unambiguously called a single genotype. These criteria reduce

628 the possibility that a heterozygous mutant is erroneously called if a site was merely subject to rare
629 sequencing errors. This resulted in a homozygous variant rate of 2.21 x 10^{-8} and a heterozygous variant
630 rate of 9.86 x 10^{-7}. While already quite low, these rates likely represent over-estimates because visual
631 inspection of bam files in some cases revealed ambiguous base calls. Over three quarters of sites (78.5%)
632 with a putative heterozygous variant were found only in a single worm, meaning that these rare variants
633 cannot account for an association with expression across 180 worms.
634
635 *Statistical analysis*
636 Statistical tests are described in detail in figure legends and in the corresponding section of the
637 methods. All individual data points are plotted in figures whenever possible.
638
639 *Data availability statement*
640 All raw mRNA-seq data is available at NCBI GEO accession GSE244875. All data used to generate
641 figures are available as part of Supplementary File 1.
642
643 *Code availability statement*
644 Scripts are available at https://github.com/amykwebster/SingleWormRNAseq and have been
645 preserved on Zenodo at time of publication at https://doi.org/10.5281/zenodo.15489998.
646
647 **ACKNOWLEDGEMENTS**
648 We thank Chien-Hui Chuang and the Bruce Bowerman lab for sharing RNAi clones. Funding was
649 provided by grants F32GM146402 to AKW and R35GM131838 to PCP. We thank members of the Phillips
650 lab for critical feedback. We would also like to thank WormBase.

651 **REFERENCES**
652

653 1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017).
654 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.
655 10.1016/j.ajhg.2017.06.005.

656 2. Edwards, A.O., Ritter, R., Abel, K.J., Manning, A., Panhuysen, C., and Farrer, L.A. (2005).
657 Complement factor H polymorphism and age-related macular degeneration. Science *308*, 421–424.
658 10.1126/science.1110189.

659 3. Nolte, I.M., van der Most, P.J., Alizadeh, B.Z., de Bakker, P.I., Boezen, H.M., Bruinenberg, M.,
660 Franke, L., van der Harst, P., Navis, G., Postma, D.S., et al. (2017). Missing heritability: is the gap
661 closing? An analysis of 32 complex traits in the Lifelines Cohort Study. Eur. J. Hum. Genet. *25*, 877–
662 885. 10.1038/ejhg.2017.50.

663 4. Alsheikh, A.J., Wollenhaupt, S., King, E.A., Reeb, J., Ghosh, S., Stolzenburg, L.R., Tamim, S.,
664 Lazar, J., Davis, J.W., and Jacob, H.J. (2022). The landscape of GWAS validation; systematic
665 review identifying 309 validated non-coding variants across 130 human diseases. BMC Med.
666 Genomics *15*, 74. 10.1186/s12920-022-01216-w.

667 5. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The Post-GWAS Era: From Association to
668 Function. Am. J. Hum. Genet. *102*, 717–730. 10.1016/j.ajhg.2018.04.002.

669 6. Barban, N., Jansen, R., de Vlaming, R., Vaez, A., Mandemakers, J.J., Tropf, F.C., Shen, X., Wilson,
670 J.F., Chasman, D.I., Nolte, I.M., et al. (2016). Genome-wide analysis identifies 12 loci influencing
671 human reproductive behavior. Nat. Genet. *48*, 1462–1472. 10.1038/ng.3698.

672 7. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S.,
673 Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for
674 obesity biology. Nature *518*, 197–206. 10.1038/nature14177.

675 8. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel,
676 R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for
677 transcriptome-wide association studies. Nat. Genet. *51*, 592–599. 10.1038/s41588-019-0385-z.

678 9. Rakyan, V.K., Down, T.A., Balding, D.J., and Beck, S. (2011). Epigenome-wide association studies
679 for common human diseases. Nat. Rev. Genet. *12*, 529–541. 10.1038/nrg3000.

680 10. Battram, T., Gaunt, T.R., Relton, C.L., Timpson, N.J., and Hemani, G. (2022). A comparison of the
681 genes and genesets identified by GWAS and EWAS of fifteen complex traits. Nat. Commun. *13*,
682 7816. 10.1038/s41467-022-35037-3.

683 11. Torres-Garcia, S., Yaseen, I., Shukla, M., Audergon, P.N.C.B., White, S.A., Pidoux, A.L., and
684 Allshire, R.C. (2020). Epigenetic gene silencing by heterochromatin primes fungal resistance. Nature
685 *585*, 453–458. 10.1038/s41586-020-2706-x.

686 12. Lismer, A., Dumeaux, V., Lafleur, C., Lambrot, R., Brind'Amour, J., Lorincz, M.C., and Kimmins, S.
687 (2021). Histone H3 lysine 4 trimethylation in sperm is transmitted to the embryo and associated with
688 diet-induced phenotypes in the offspring. Dev. Cell *56*, 671-686.e6. 10.1016/j.devcel.2021.01.014.

689 13. Kaletsky, R., Moore, R.S., Vrla, G.D., Parsons, L.R., Gitai, Z., and Murphy, C.T. (2020). C. elegans
690 interprets bacterial non-coding RNAs to learn pathogenic avoidance. Nature *586*, 445–451.
691 10.1038/s41586-020-2699-5.

692 14. DiVito Evans, A., Fairbanks, R.A., Schmidt, P., and Levine, M.T. (2023). Histone methylation
693 regulates reproductive diapause in Drosophila melanogaster. PLoS Genet. *19*, e1010906.
694 10.1371/journal.pgen.1010906.

695 15. Werner, M.S., Loschko, T., King, T., Reich, S., Theska, T., Franz-Wachtel, M., Macek, B., and
696 Sommer, R.J. (2023). Histone 4 lysine 5/12 acetylation enables developmental plasticity of
697 Pristionchus mouth form. Nat. Commun. *14*, 2095. 10.1038/s41467-023-37734-z.

698 16. Öst, A., Lempradl, A., Casas, E., Weigert, M., Tiko, T., Deniz, M., Pantano, L., Boenisch, U., Itskov,
699 P.M., Stoeckius, M., et al. (2014). Paternal diet defines offspring chromatin state and
700 intergenerational obesity. Cell *159*, 1352–1364. 10.1016/j.cell.2014.11.005.

701 17. Webster, A.K., and Phillips, P.C. (2025). Epigenetics and individuality: from concepts to causality
702 across timescales. Nat. Rev. Genet. 10.1038/s41576-024-00804-z.

703 18. Meyer, D.H., and Schumacher, B. (2021). BiT age: A transcriptome-based aging clock near the
704 theoretical limit of accuracy. Aging Cell *20*, e13320. 10.1111/acel.13320.

705 19. Daines, B., Wang, H., Wang, L., Li, Y., Han, Y., Emmert, D., Gelbart, W., Wang, X., Li, W., Gibbs,
706 R., et al. (2011). The Drosophila melanogaster transcriptome by paired-end RNA sequencing.
707 Genome Res. *21*, 315–324. 10.1101/gr.107854.110.

708 20. Li, B., Qing, T., Zhu, J., Wen, Z., Yu, Y., Fukumura, R., Zheng, Y., Gondo, Y., and Shi, L. (2017). A
709 Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. Sci. Rep. *7*, 4200.
710 10.1038/s41598-017-04520-z.

711 21. Chang, D., Serra, L., Lu, D., Mortazavi, A., and Dillman, A. (2021). A Revised Adaptation of the
712 Smart-Seq2 Protocol for Single-Nematode RNA-Seq. Methods Mol. Biol. *2170*, 79–99. 10.1007/978-
713 1-0716-0743-5_6.

714 22. Serra, L., Chang, D.Z., Macchietto, M., Williams, K., Murad, R., Lu, D., Dillman, A.R., and
715 Mortazavi, A. (2018). Adapting the Smart-seq2 Protocol for Robust Single Worm RNA-seq. Bio
716 Protoc *8*. 10.21769/BioProtoc.2729.

717 23. Pallares, L.F., Picard, S., and Ayroles, J.F. (2020). TM3'seq: A Tagmentation-Mediated 3'
718 Sequencing Approach for Improving Scalability of RNAseq Experiments. G3 (Bethesda) *10*, 143–
719 150. 10.1534/g3.119.400821.

720 24. Lin, Y., Chen, Z.-X., Oliver, B., and Harbison, S.T. (2016). Microenvironmental Gene Expression
721 Plasticity Among Individual Drosophila melanogaster. G3 (Bethesda) *6*, 4197–4210.
722 10.1534/g3.116.035444.

723 25. Werkhoven, Z., Bravin, A., Skutt-Kakaria, K., Reimers, P., Pallares, L.F., Ayroles, J., and de Bivort,
724 B.L. (2021). The structure of behavioral variation within a genotype. eLife *10*. 10.7554/eLife.64988.

725 26. Wang, A.J., Wibisono, P., Geppert, B.M., and Liu, Y. (2022). Using single-worm RNA sequencing to
726 study C. elegans responses to pathogen infection. BMC Genomics *23*, 653. 10.1186/s12864-022-
727 08878-x.

728 27. Perez, M.F., Francesconi, M., Hidalgo-Carcedo, C., and Lehner, B. (2017). Maternal age generates
729 phenotypic variation in Caenorhabditis elegans. Nature *552*, 106–109. 10.1038/nature25012.

730 28. Gouvêa, D.Y., Aprison, E.Z., and Ruvinsky, I. (2015). Experience Modulates the Reproductive
731 Response to Heat Stress in C. elegans via Multiple Physiological Processes. PLoS ONE *10*,
732 e0145925. 10.1371/journal.pone.0145925.

733 29. Hirsh, D., Oppenheim, D., and Klass, M. (1976). Development of the reproductive system of
734 Caenorhabditis elegans. Dev. Biol. *49*, 200–219. 10.1016/0012-1606(76)90267-0.

735 30. Stumpf, C.R., Kimble, J., and Wickens, M. (2008). A Caenorhabditis elegans PUF protein family
736 with distinct RNA binding specificity. RNA *14*, 1550–1557. 10.1261/rna.1095908.

31. Lublin, A.L., and Evans, T.C. (2007). The RNA-binding proteins PUF-5, PUF-6, and PUF-7 reveal multiple systems for maternal mRNA regulation during C. elegans oogenesis. Dev. Biol. *303*, 635–649. 10.1016/j.ydbio.2006.12.004.

32. Jänes, J., Dong, Y., Schoof, M., Serizay, J., Appert, A., Cerrato, C., Woodbury, C., Chen, R., Gemma, C., Huang, N., et al. (2018). Chromatin accessibility dynamics across C. elegans development and ageing. eLife *7*. 10.7554/eLife.37344.

33. Ahringer, J., and Gasser, S.M. (2018). Repressive Chromatin in Caenorhabditis elegans: Establishment, Composition, and Function. Genetics *208*, 491–511. 10.1534/genetics.117.300386.

34. Evans, K.J., Huang, N., Stempor, P., Chesney, M.A., Down, T.A., and Ahringer, J. (2016). Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. Proc Natl Acad Sci USA *113*, E7020–E7029. 10.1073/pnas.1608162113.

35. Serizay, J., Dong, Y., Jänes, J., Chesney, M., Cerrato, C., and Ahringer, J. (2020). Distinctive regulatory architectures of germline-active and somatic genes in C. elegans. Genome Res. *30*, 1752–1765. 10.1101/gr.265934.120.

36. Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci USA *99*, 12795–12800. 10.1073/pnas.162041399.

37. Yang, C.-H., Fagnocchi, L., Apostle, S., Wegert, V., Casaní-Galdón, S., Landgraf, K., Panzeri, I., Dror, E., Heyne, S., Wörpel, T., et al. (2022). Independent phenotypic plasticity axes define distinct obesity sub-types. Nat. Metab. *4*, 1150–1165. 10.1038/s42255-022-00629-2.

38. Sulston, J.E., and Horvitz, H.R. (1977). Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev. Biol. *56*, 110–156. 10.1016/0012-1606(77)90158-0.

39. Tolkin, T., and Hubbard, E.J.A. (2021). Germline Stem and Progenitor Cell Aging in C. elegans. Front. Cell Dev. Biol. *9*, 699671. 10.3389/fcell.2021.699671.

40. Mignerot, L., Gimond, C., Bolelli, L., Bouleau, C., Sandjak, A., Boulin, T., and Braendle, C. (2024). Natural variation in the Caenorhabditis elegans egg-laying circuit modulates an intergenerational fitness trade-off. eLife *12*. 10.7554/eLife.88253.

41. Pérez-Lluch, S., Blanco, E., Tilgner, H., Curado, J., Ruiz-Romero, M., Corominas, M., and Guigó, R. (2015). Absence of canonical marks of active chromatin in developmentally regulated genes. Nat. Genet. *47*, 1158–1167. 10.1038/ng.3381.

42. Rea, S.L., Wu, D., Cypser, J.R., Vaupel, J.W., and Johnson, T.E. (2005). A stress-sensitive reporter predicts longevity in isogenic populations of Caenorhabditis elegans. Nat. Genet. *37*, 894–898. 10.1038/ng1608.

43. Kinser, H.E., Mosley, M.C., Plutzer, I.B., and Pincus, Z. (2021). Global, cell non-autonomous gene regulation drives individual lifespan among isogenic C. elegans. eLife *10*. 10.7554/eLife.65026.

44. Mosley, M.C., Kinser, H.E., Martin, O.M.F., Stroustrup, N., Schedl, T., Kornfeld, K., and Pincus, Z. (2024). Similarities and differences in the gene expression signatures of physiological age versus future lifespan. Aging Cell, e14428. 10.1111/acel.14428.

45. Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. *41*, e108. 10.1093/nar/gkt214.

46. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. 10.1093/bioinformatics/btp616.

47. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and Team, R.C. (2007). Linear and nonlinear mixed effects models. R package version.

48. Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Package “lme4.” ….

49. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. 10.1089/omi.2011.0118.

50. Friedman, J., Hastie, T., and Tibshirani, R. (2021). Package “glmnet.” CRAN R ….

51. Wilson, R., Le Bourgeois, M., Perez, M., and Sarkies, P. (2023). Fluctuations in chromatin state at regulatory loci occur spontaneously under relaxed selection and are associated with epigenetically inherited variation in C. elegans gene expression. PLoS Genet. *19*, e1010647. 10.1371/journal.pgen.1010647.

52. Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinformatics *47*, 11.12.1-34. 10.1002/0471250953.bi1112s47.

53. Kamath, R.S., and Ahringer, J. (2003). Genome-wide RNAi screening in Caenorhabditis elegans. Methods *30*, 313–321. 10.1016/s1046-2023(03)00050-1.

54. Ahringer, J. (2006). Reverse genetics. WormBook. 10.1895/wormbook.1.47.1.

55. Van Willigen, B., and Maintainer, R. (2017). Package “nlme.” Linear and ….

56. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.