Epigenetic context predicts gene expression variation and reproductive traits across genetically identical individuals

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

1. Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA
2. Department of Biochemistry, Oxford University, Oxford, UK

SUMMARY

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 proven to 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 or impossible 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 prior environmental experience and some of which was random. Multiple small sets of genes together were highly predictive of reproductive traits across individuals, explaining on average over half and over a quarter of variation in the two traits. We manipulated mRNA levels of predictive genes using RNA interference 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 individual gene expression differences underlying critical traits may be driven in part by chromatin structure. Together, this work shows that individual differences in gene expression that arise independently of underlying genetic differences are both predictive and causal in shaping reproductive traits at levels that equal or exceed genetic variation.

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 explain only a small fraction of variation in traits and diseases[3], and the clinical utility of such loci is 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, then, the major approaches used to uncover the etiology of complex traits and diseases are limited in their ability to separate genetic and environmental effects and are largely built upon the assumption that genetic differences are required to determine functional causation. 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[11-16]. 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*[17,18]), or they use very few individuals (typically due to difficulty sampling many larger organisms such as mice[19]), precluding knowledge of how

57 individual life trajectories may influence gene expression, epigenetic information, and traits of interest. In
58 recent years, it has become possible to generate individual transcriptomic profiles in *C. elegans* and
59 *Drosophila,* and methods to profile individuals in high throughput have improved[20-26]. This has revealed that
60 genetically identical individuals in a shared environment indeed differ in their gene expression profiles, but
61 the effects that gene expression differences genome-wide may have on complex fitness-related
62 phenotypes -- which also vary substantially across genetically identical individuals – remain largely
63 unknown. To fully separate genetic and environmental causation, we use an isogenic population of *C.*
64 *elegans* to show that individuals in a shared environment at the same developmental stage exhibit
65 substantial differences in two fitness-related reproductive traits. By performing mRNA-seq on each
66 individual worm, we identify hundreds of genes for which variation in expression is strongly associated with
67 reproductive trait variation, identify small sets of genes that are highly predictive of both traits, and show
68 experimentally that reducing mRNA levels of individual genes causally affects progeny production. Finally,
69 we show that predictive gene sets for reproductive traits are enriched for genes with the epigenetic
70 modification H3K27me3, revealing that chromatin environment defines which loci are subject to the
71 individual variation in gene expression that drives variation in complex traits.

72 **Results**

73 To determine if differences in gene expression levels correspond to differences in complex traits,
74 we measured reproductive traits and whole-animal mRNA transcriptomes for each of 180 isogenic *C.*
75 *elegans* in the same environment at the same stage of early adulthood (Fig 1A). Two reproductive traits –
76 egg-laying onset (ELO) and the number of progeny produced in the first 24 hours of egg-laying (early
77 brood) – were measured for each individual. To generate additional phenotypic variation[26] and facilitate
78 dissection of environmental sources of variation, worms experienced controlled environmental
79 perturbations early in development in a two-factor design: they were either born to a day 1 or day 3 adult
80 parent (corresponding to the beginning and middle of egg-laying in *C. elegans*, respectively), and, once
81 hatched, they were subject to either a constant 20°C temperature or an eight-hour shift to 25°C (both
82 corresponding to typical rearing temperatures for *C. elegans*). Isogenic worms exhibited a range of both
83 ELO and early brood trait values (Fig. S1A) and variation in these traits was partially driven by
84 environmental perturbations, consistent with prior work on these perturbations[26-28] (Fig S1B). To determine

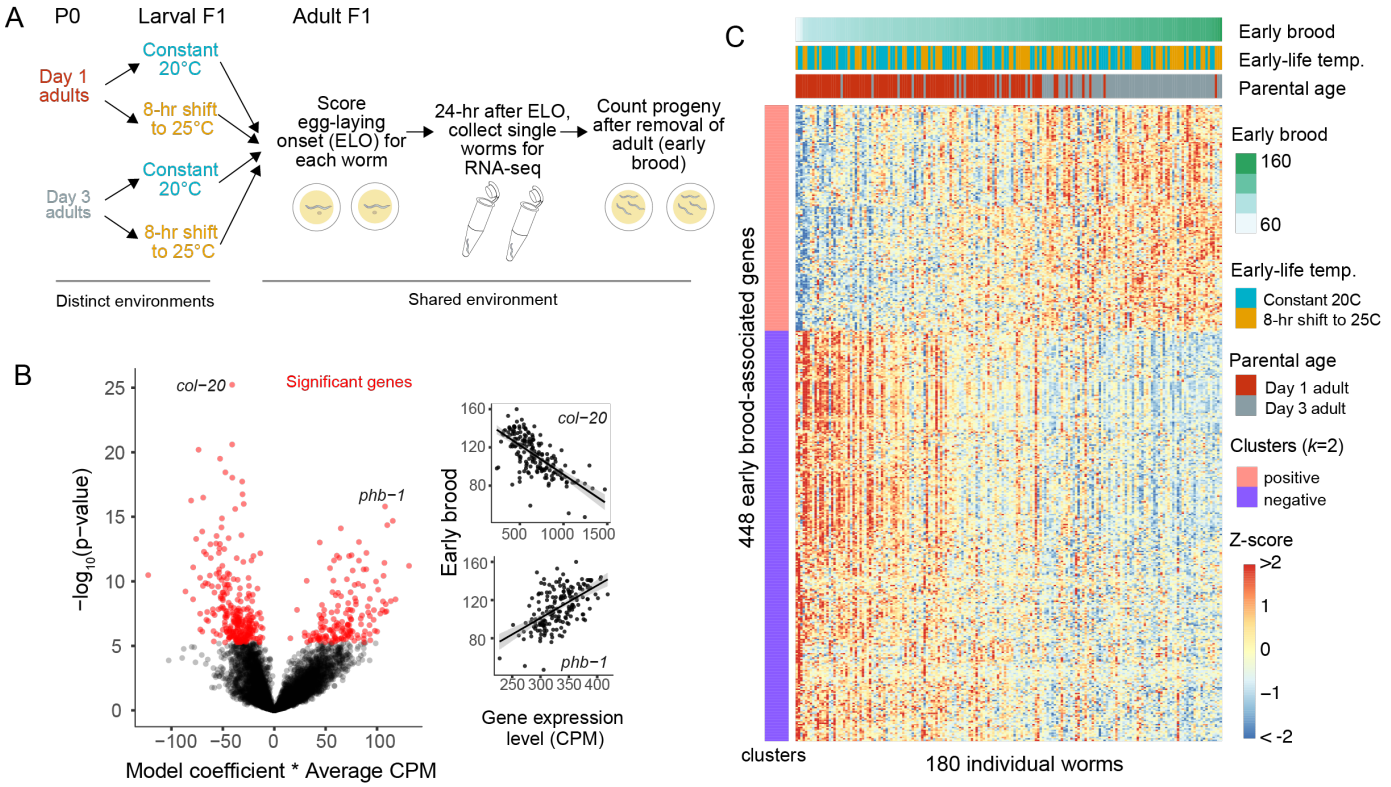

Figure 1. Differences in early brood across isogenic individuals are associated with expression levels of hundreds of genes.
A. Experimental design for single-individual mRNA-seq. P0 worms are allowed to lay progeny as Day 1 or Day 3 adults. F1 progeny are either subject to a constant temperature or an eight-hour shift to 25°C. Beginning in mid-larval stages all F1 worms experience the same environment. For each worm, time to egg-laying onset is scored. 24 hours after egg-laying onset, each worm is collected for single-individual mRNA-seq. Progeny laid on plates in the first 24 hours of egg-laying are counted to determine the early brood of each 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 red, non-significant genes are in black. Significance determined by a Bonferroni-corrected p-value cutoff of 5.67 x 10^{-6}. Phenotypic and expression data for each worm is shown for the top two genes, *col-20* and *phb-1*.
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.

if gene expression differences corresponded to these traits, we used a linear mixed modeling approach and identified significant genes corresponding to each trait (Fig 1B-C, Fig S2). Expression of 448 genes was strongly associated with early brood (Fig 1B-C), while expression of 11 genes was strongly associated with ELO (Fig S3). We confirmed that these effects were not driven by mutations arising in the population over the course of the experiment (see Methods). This reveals many genes, particularly for early brood, for which expression differences across genetically identical individuals strongly associate with the reproductive traits of these individuals.

Differences in early brood were partially driven by parental age but not affected by early-life temperature (Fig 1C and Fig S1B), and previous work has shown that maternal age influences vitellogenin expression to affect progeny phenotypes[26]. We therefore wondered the extent to which the association

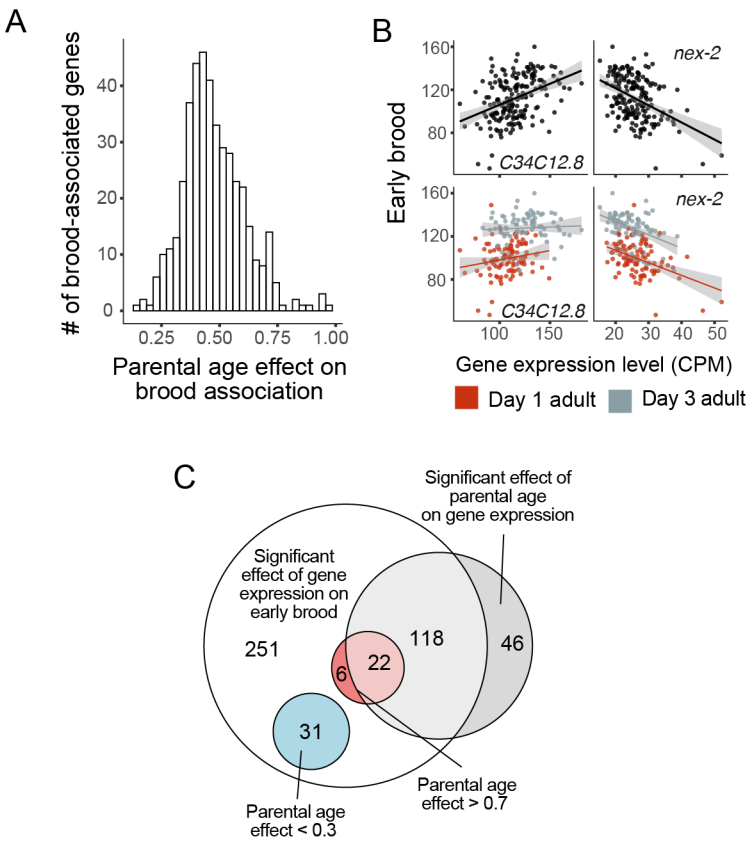

Figure 2. Gene expression correlations with early brood are partially driven by parental age in the previous generation.
A. Histogram of the previously identified brood-associated genes and how much of their association with brood is driven by parental age.
B. Examples of genes that are toward the extremes of the distribution shown in A. C34C12.8 expression is strongly positively associated with brood (black dots) but this positive association is diminished within each environment (red and gray dots). In contrast, *nex-2* expression is strongly negatively associated with early brood but this effect is still robust within each environment.
C. Venn diagram of genes that have a significant association with early brood (white), genes that are affected by parental age (gray), genes that have a significant association with early brood and are strongly affected by parental age (pink), and genes that have a significant association with early brood and are minimally affected by parental age (light blue).

between gene expression and early brood was driven by the parental age for the identified early brood-associated genes. We performed a mediation analysis which revealed that genes varied in the degree to which the original association was driven by parental age (Fig. 2A). For some genes, gene expression variation associated with brood was almost completely driven by parental age (*e.g.,* C34C12.8, Fig 2B) while for others, gene expression variation was largely independent (*e.g., nex-2*, Fig 2B). This was further recapitulated by directly measuring the effect of parental age on gene expression. Brood-associated genes that were the most dependent on parental age exhibited a strong overlap with genes that were differentially expressed in response to parental age. In contrast, genes that exhibited an association with brood independent of parental age exhibited no overlap with genes differentially expressed in response to parental age. As a result, our approach identified gene expression differences that associate with trait variation and revealed that the sources of these gene expression differences can depend on the known life history of individuals or be largely stochastic.

Given that many genes are individually strongly associated with trait variation, we next asked if we could use aggregate gene expression profiles for combinations of genes to predict the reproductive traits of individuals. Principal component analysis (PCA) provides an aggregate profile which simplifies gene

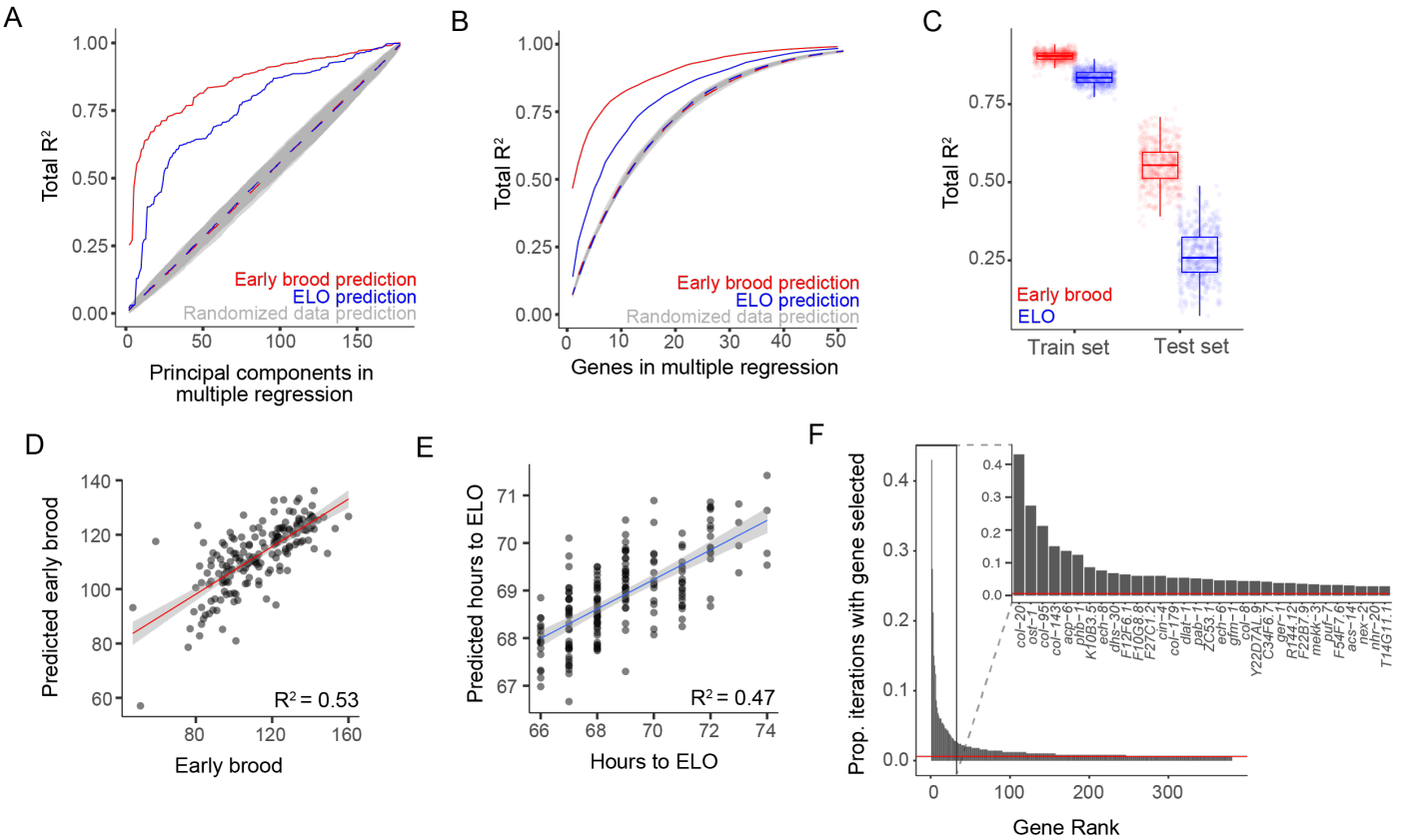

Figure 3. Multi-gene models are highly predictive of early brood and egg-laying onset at an individual level.
A. Total R^2 of gene expression principal components (PCs) in a multiple regression. PCs explain early brood (red), ELO (blue), and randomized data (gray).
B. Total R^2 of top genes in a multiple regression. Gene sets explain early brood (red), ELO (blue), and randomized data (gray).
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.
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.
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.

expression data to linearly independent factors (PCs) that together explain the variation in the gene expression data and are ordered by how much variation they explain (*i.e.*, PC1 explains the most variation, PC2 explains the second-most, etc.). We therefore first asked whether ordered PCs together in a multiple regression explain more variation than expected by chance. Indeed, gene expression PCs explained both ELO and early brood data substantially better than randomized ELO and early brood data, indicating that individual gene expression profiles provide a clear signal of phenotypic information (Fig. 3A).

Because PCA incorporates data from all expressed genes, we next wanted to determine if a smaller subset of genes could be similarly predictive of reproductive traits. We again used a multiple regression

131 approach in which we successively selected each gene by identifying the gene that explained the most
132 additional variance compared to the previous model (Fig. 3B). Genes that exhibit similar expression
133 patterns, even if they are independently associated with traits of interest, would not explain much additional
134 variance in a multiple regression approach. Therefore, this approach prioritizes identifying genes that are
135 most predictive in conjunction with other genes. Relatively small sets of genes predict ELO and early brood
136 phenotypes very well, and do so significantly better than randomized trait data, with diminishing returns for
137 variance explained as the number of genes increases (Fig. 3B). Maximum differences in the proportion of
138 variance explained between trait data and randomized data occur in a model with 4 genes for the early
139 brood analysis (R^2 difference of 0.433) and with 7 genes in the egg-laying onset analysis (R^2 difference of
140 0.195), and substantial differences of at least 0.1 persist to at least a 20-gene model for both traits. This
141 highlights that individual worm gene expression profiles consisting of only a handful of key genes are highly
142 informative for predicting the individual's traits.

143 We next wanted to develop a quantitative estimate of how much trait variation is explained by
144 aggregate gene expression profiles using two distinct approaches that prevent model over-fitting. First, as
145 an extension of the multiple regression analysis identifying predictive genes, we randomly split the full
146 dataset (consisting of trait data and gene expression profiles) into a training set and test set. In the training
147 set, we used the multiple regression approach to identify the set of 10 genes that explained the most
148 variance in the trait data. We then asked how well this set of 10 genes explained variance in the test set
149 and repeated this process 500 times, each time randomly splitting the data into training and test sets (Fig.
150 3C). We found that median proportion of variance explained in the test sets was 0.55 (range 0.36 to 0.73)
151 for early brood and 0.26 (range 0.07 to 0.49) for ELO (Fig. 3C). Thus, sets of only 10 genes explain over
152 half of trait variation in early brood and over a quarter of variation in ELO using test sets of data that were
153 not used to select predictive genes. As a second approach, we used elastic net regression and leave-one-
154 out cross validation on the full set of expressed genes. In other words, we used a machine learning
155 approach designed to prevent overfitting to develop a model for all individuals except one, used the model
156 to predict the trait for the remaining individual, then repeated for all individuals. Consistent with the multiple
157 regression approach, the proportion of variance explained was 0.53 for early brood and 0.47 for ELO (Fig.

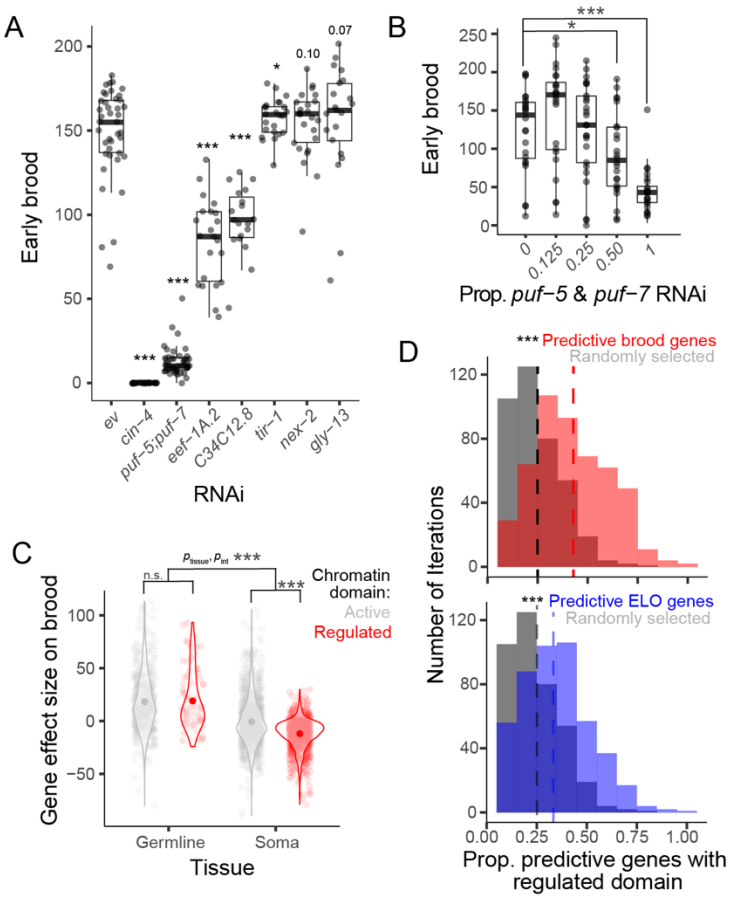

Figure 4. Genes predictive of reproductive traits causally affect early brood and are enriched for H3K27me3
A. RNAi knockdown of selected predictive genes compared to empty vector (ev) and effects on early brood.
B. Dose response of *puf-5* and *puf-7* RNAi together with empty vector and effects on early brood.
A-B. Linear mixed model with RNAi as fixed effect and biological replicate as a random effect. *p<0.05, **p<0.01, ***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

3D-E). Aggregate gene expression profiles identified via multiple approaches are thus highly predictive of reproductive traits among isogenic individuals in a shared environment.

A key question is whether predictive genes are causal to affect phenotypes and whether similar underlying regulation underlies expression variation in predictive genes. To address causality, we focused on the early brood trait because predictive genes explained a higher proportion of variance for this trait compared to ELO. We used RNA interference (RNAi) to reduce the level of gene expression of predictive genes and prioritized genes that were also in the brood-associated set and either highly independent of parental age or dependent on parental age and have an ortholog in humans. Within this curated set, genes causally affected early brood in the majority of cases compared to empty vector (Fig. 4A). We followed up on one pair of genes from this screen, *puf-5* and *puf-7,* to ask whether perturbing the dose of RNAi shows a dose response effect on brood. These genes have previously been shown to act redundantly in fertility in *C. elegans*[29,30] and when knocked down simultaneously in our assay have a large effect on brood without

186 causing sterility (Fig. 4A). Consistent with the idea that differences in the level of mRNA affect the early
187 brood trait, the dose of RNAi corresponds to the effect size on brood (Fig. 4B). Thus, this analysis reveals
188 that the predictive genes also make excellent candidates as causal effectors of complex traits and can
189 affect these traits dependent on the amount of mRNA available.

190 Expression of predictive genes can be causally related to reproductive traits (Fig. 4A-B), but do
191 these genes share underlying regulation? The majority of *C. elegans* genes fall into one of two chromatin
192 domains with distinct levels of gene activity[31-33]. Regulated domains include genes that are expressed at
193 specific stages of development and are marked by the facultative heterochromatin modification H3K27me3.
194 In contrast, active domains include genes that are expressed broadly across cell types, including in the
195 germline, and marked by the euchromatic modification H3K36me3[31,33,34]. These two marks are generally
196 mutually exclusive, and marks are robust across several stages of development in *C. elegans*[33] despite
197 major changes in gene expression that occur during this time. Thus, we can gain insights into the
198 epigenetic status of genes associated with and predictive of reproductive traits. Because active genes
199 marked with H3K36me3 are enriched for being in the germline, we performed a two-factor analysis that
200 takes into account both tissue (soma or germline) and chromatin domain (regulated and active). Among
201 early brood-associated genes, germline genes are strongly positively associated with brood, regardless of
202 chromatin environment, but chromatin environment and tissue also exhibit a significant interaction (Fig.
203 4C). Post-hoc pairwise comparisons reveal that this interaction is driven by differences in the chromatin
204 environment in the soma. In particular, while active somatic genes have on average no effect on brood,
205 regulated somatic genes are strongly negatively associated. Because regulated genes are typically
206 expressed in a temporally controlled or tissue-specific manner in the soma, this may indicate that aberrant
207 expression of these genes has negative consequences for reproduction, or similarly, that appropriate
208 regulation and silencing of these genes promotes reproduction. To determine if epigenetic regulation may
209 underlie variation in predictive genes for both traits overall, without regard to the direction of association,
210 we used all gene sets identified from iterations of predictive sets (Fig. 3C,F) to determine whether these
211 genes are more likely to be found in regulated domains marked with H3K27me3. For both reproductive
212 traits, predictive genes were strongly over-enriched for being in a regulated domain compared to randomly
213 selected gene sets from the same background set (Fig. 4D). Genes in regulated domains are more likely to

214 be variable across developmental stages and cell types[33], and we find in our data that genes with the most
215 expression variation across individuals unexplained by experimental factors are also enriched for
216 H3K27me3 (Fig. S3). Taken together, these results suggest that regulated domains are subject to gene
217 expression variation in genetically identical individuals and that this variation is an important substrate
218 underlying complex trait variation.

219 **Discussion**

220 Individuals differ from each other for a variety of traits over the course of their lives, but being able
221 to reliably predict these differences and understand their causal basis is a challenge. Gene expression is a
222 critical intermediate functional layer between genotype and phenotype, but the role that gene expression
223 plays in shaping individual organismal traits independently of genetic variation and environmental factors
224 has remained elusive. Here we effectively screened for genes that exhibit expression variation between
225 isogenic *C. elegans* individuals and for which this variation corresponds to critical reproductive traits – time
226 to egg-laying onset and early brood. We identified hundreds of associated genes as well as small sets of
227 highly predictive and causal genes. Predictive genes for both traits are enriched for a particular chromatin
228 modification – H3K27me3 – highlighting the possibility that individual differences in chromatin structure
229 may underlie the potential for gene expression to vary between individuals. These findings build upon past
230 research and suggest key areas for future work.

231 Individual gene expression differences underlying reproductive traits suggest that gene expression
232 differences may be important for a variety of other complex traits. Gene expression differences among
233 isogenic individuals in a shared environment have been previously documented as predictive of lifespan[35-
234 37]. In these studies, reporter gene expression for specific candidate genes predicted individuals' later
235 lifespan. Our results extend on such studies by identifying predictive genes in the context of reproductive
236 traits in a high-throughput manner, rather than relying on specific reporters. Our approach can also be used
237 directly in the context of other traits that occur throughout the life cycle, such as those related to
238 development, but must be used indirectly for time-to-death traits like lifespan or stress resistance. Because
239 a worm, once collected for RNA, cannot be subsequently phenotyped for traits like lifespan, using
240 predictive reporters for time-to-death traits in conjunction with transcriptomics will be a useful approach for
241 identifying novel regulators for a broad swath of traits.

242 Our results suggest that differences in epigenetic regulation may underlie differences in gene
243 expression and traits across individuals. Regulated genes are robustly marked with H3K27me3, defining
244 regulated domains across developmental stages, and in our data these genes are enriched both among
245 genes with the most unexplained variation and among genes for which variation is predictive of
246 reproductive traits. In contrast, germline genes that predominantly consist of genes in active domains are
247 specifically positively associated with brood. Taken together, one possible model is that regulated genes
248 marked with H3K27me3 are more subject to inter-individual variation which may lead to random differences
249 in gene expression and, in turn, reproductive traits. This is consistent with observations that regulated
250 genes exhibit a higher coefficient of variation across conditions and that increased expression of genes
251 marked with H3K27me3 is not associated with a change in histone post-translational modifications (*e.g.,* to
252 an active mark like H3K36me3)[38]. In contrast, active genes expressed in the germline may be less subject
253 to variability in general but variability that does occur may be particularly important for reproductive traits.
254 While it is clear that regulated genes are more subject to inter-individual variability in expression, at this
255 stage it is unclear whether this variability is driven by differences in the level of H3K27me3. Performing
256 analysis on histone modifications genome-wide in individual worms, which is still technically challenging,
257 will be a key next step to determine whether differences in chromatin structure correlate with gene
258 expression changes at an individual level.
259 Genetically identical individuals can have large differences in phenotype if gene expression varies
260 across those individuals. As genetics has evolved as a field, the prevailing paradigm holds that only DNA
261 sequence differences can be used to understand biological differences among individuals, with all other
262 variation being counted as “noise.” Yet, for most traits, this apparent noise represents the majority of
263 variation, including the preponderance of risk leading to disease in human populations. This work illustrates
264 that differences in gene expression can be causally linked to important reproductive phenotypes given
265 sufficient control of genetic background. This brings us to a point in which we can begin to understand the
266 causal biology of environmental variation underlying individual differences. If even a small fraction of
267 variation of this type can be harnessed to intervene in a personalized way to improve human health, then it
268 holds the potential to have a much larger impact than nearly any variant identified to date via genetic
269 association studies.

SUPPLEMENTARY FIGURES

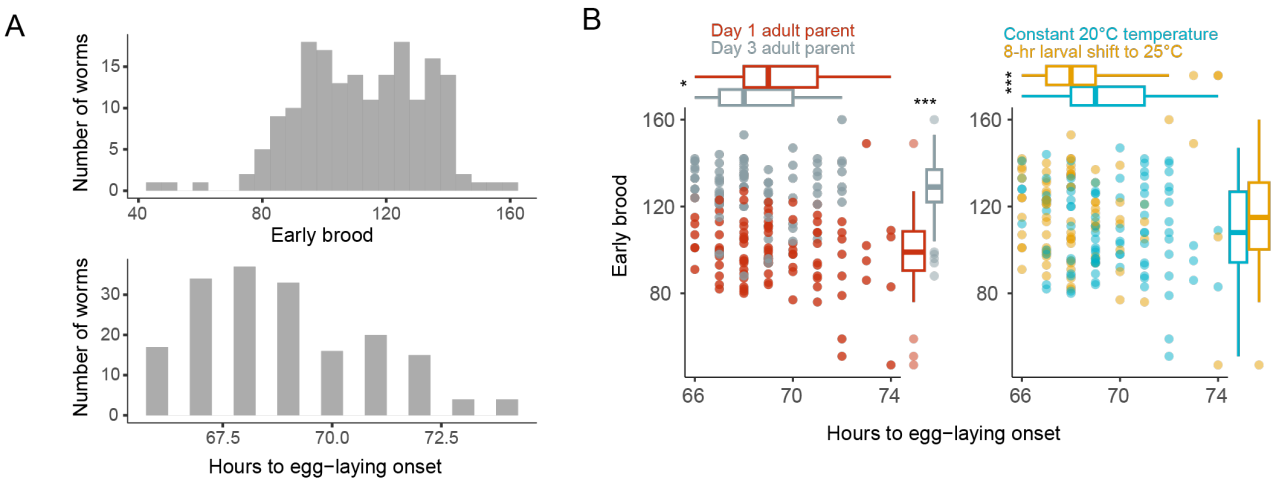

Figure S1. Isogenic worms at the same developmental stage exhibit variability in egg-laying onset and early brood that partially depends on previous environmental conditions

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

B. Bivariate scatterplots of early brood and egg-laying onset trait for all worms. The same data is plotted twice but color-coded according to two previous environmental perturbations (age of parent and early-life temperature). Significance determined using a linear mixed model with environmental perturbation as a fixed effect and biological replicate as a random effect, ***p<0.001, *p<0.05

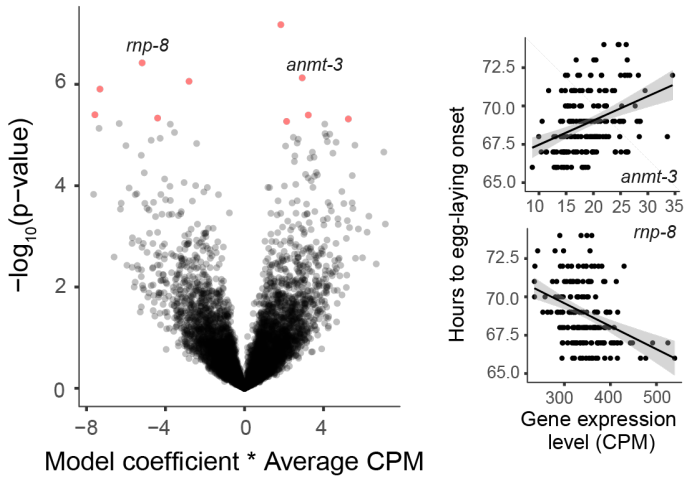

Figure S2. Variation in 11 genes is significantly associated with egg-laying onset.

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 5.67 x 10^{-6}. Phenotypic and expression data for each worm is shown for two exemplar genes, *anmt-3* and *rnp-8*.

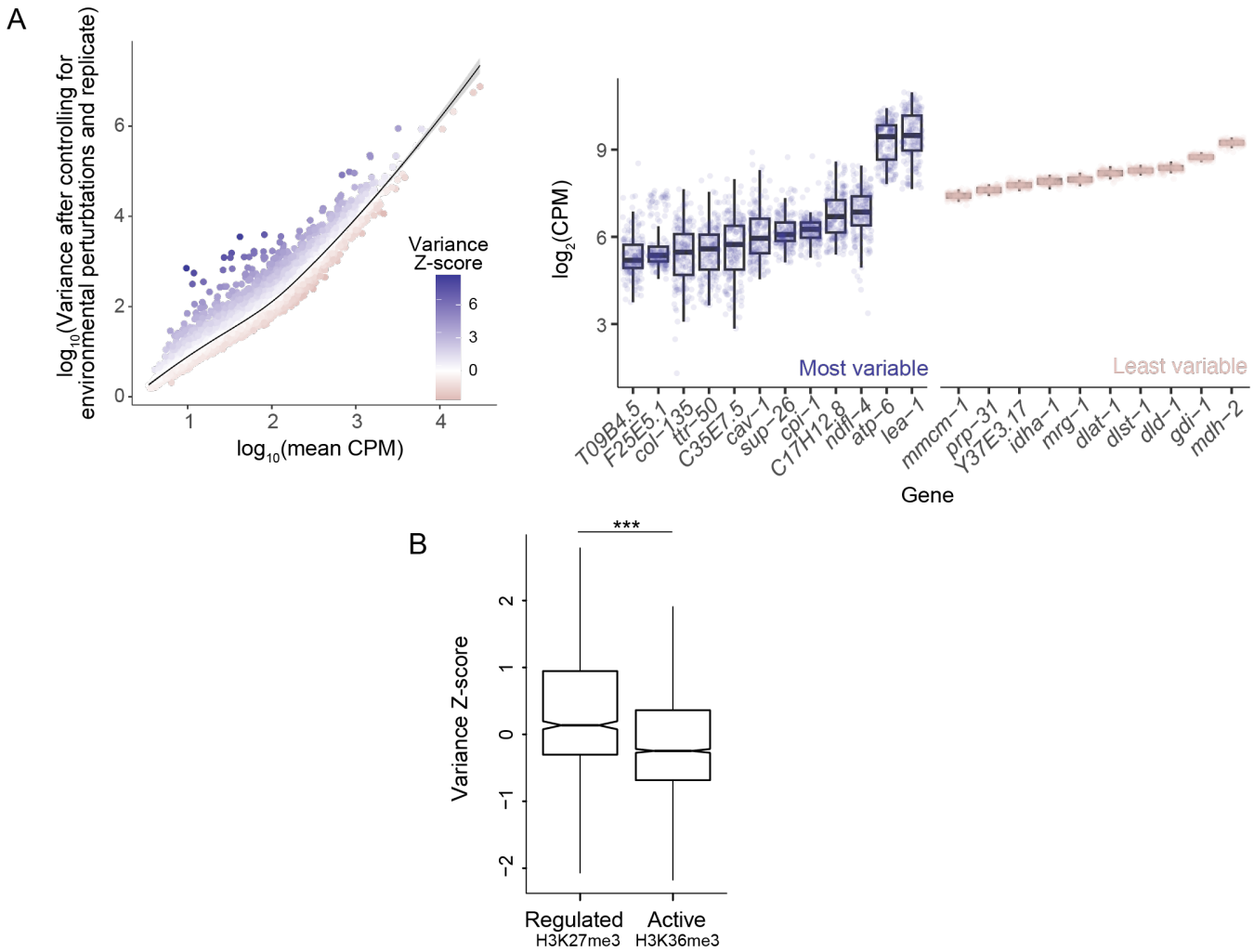

Figure S3. Single-individual mRNA-seq reveals most and least variable genes after controlling for environmental perturbations and biological replicate.

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 with 95% confidence interval is plotted in black. 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.

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

297 **METHODS**

298 *Strains used*

299 Strain N2-PD1073 was used for all experiments. Worms were passaged on *E. coli* OP50 except
300 during RNAi experiments, when worms were fed HT115 bacteria expressing dsRNA for genes of interest
301 (see 'RNAi' section).

302

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

304 Worms were well-fed at 20°C for several generations prior to beginning experiments. The
305 experiment takes place over the course of 10 days. On day 1, for each of 5 biological replicates, P0
306 embryos and L4s were picked to separate 5 cm NGM plates with OP50 and kept at 20°C. Approximately
307 10-20 individuals were placed on each plate and several plates were picked for each stage, ensuring that
308 worms do not experience dietary restriction. On day 4, P0 worms were either in the first day of adulthood
309 ('young adult') or third day of adulthood ('older adult'). P0 young adults and older adults were singled to
310 new plates, allowed to lay progeny for two hours at 20°C, then removed from the plate. Hour 0 of
311 development for F1 worms is considered the end of this egg-laying window. F1 worms were then kept at
312 20°C until day 5 of the experiment (approximately 18 hours of development as L1/L2 larvae). On day 5,
313 approximately half of these larvae were continually left at 20°C and the other half were shifted to 25°C for 8
314 hours then shifted back to 20°C. On day 6 all worms remained at 20°C and were singled to individual NGM
315 plates with OP50 to ensure they can be individually phenotyped for reproductive traits. On day 7, from 66-
316 74 hours of development, egg-laying onset was assessed every hour by checking for embryos on each
317 plate. Worms that died, bagged, or otherwise did not reach egg-laying onset in this time interval were not
318 included in subsequent analysis, and all together this affected 6% of worms across all biological replicates.
319 Time to egg-laying onset was noted for each worm. 24 hours after egg-laying onset, on day 8, each worm
320 was collected for cell lysate generation for mRNA-seq (see 'Cell lysate generation' section of Materials and
321 Methods). On day 10, the number of progeny laid on each plate during the 24 hours between egg-laying
322 onset and collection were counted as late larvae/young adults to ensure all progeny were counted. All
323 together, each F1 worm collected experienced two environmental perturbations (young adult or older adult

324 parent, and early-life constant temperature or temperature shift) and was measured for two traits (egg-
325 laying onset time and number of progeny produced within 24 hours) prior to mRNA-seq.
326
327 *Cell lysate generation*
328 Individual worms were collected for chemical lysis 24 hours after egg-laying onset (that is, worms
329 that reached egg-laying onset at 72 hours were collected at 96 hours, worms reaching egg-laying onset at
330 73 hours were collected at 97 hours, etc. for each hour). 200 μL of lysis mix (187.2 μL Elution buffer and
331 12.8 μL Proteinase K) was made fresh every 2-3 hours on ice. 6 μL of lysis mix was pipetted into each
332 PCR tube according to how many worms needed to be collected in a given hour. Each worm was picked
333 into a 10 μL drop of dH2O for a quick wash to remove any excess bacteria, then transferred to a second
334 drop of dH2O. After this, each worm was picked into an individual PCR tube with lysis mix. Immediately
335 after worms are placed into tubes with lysis mix, we proceeded with the lysis protocol (typically no more
336 than a dozen worms were prepared at once). The lysis protocol is 65°C for 10 min, 85°C for 1 min, and 4°C
337 hold and is based on a previously published protocol[24]. After the lysis protocol, worms were moved to ice
338 and then stored at -80°C until mRNA-seq library preparation.
339
340 *Preparation and sequencing of mRNA-seq libraries*
341 The KAPA mRNA-seq preparation kit (KK8580) was used to prepare libraries. 5 μL of each cell
342 lysate preparation described above, which contains total RNA, was used for each sample. Reagents were
343 used at 0.25X and kit instructions were followed to generate libraries. Library concentrations were
344 quantified using Qubit, and samples were pooled together in equimolar ratios in two batches of 96 for
345 sequencing. Custom dual-indexed adapters were designed by the University of Oregon Genomics and Cell
346 Characterization Core Facility and sequencing data was generated with the Illumina NovaSeq 6000 at the
347 same facility.
348
349 *Analysis of single-worm mRNA-seq data*
350 FASTQ files were aligned to WBcel235 version of the *C. elegans* genome using subread-align[39]
351 with flag -t 0. Bam files were counted using featureCounts with flag -T 5. 12 samples were filtered out at

352 this stage due to particularly low read counts (less than 1 million reads) and/or poor mapping quality (less
353 than 70% of reads both mapped to the genome and assigned to a genomic feature). The remaining 180
354 samples had an average of 5.8 million reads (SD 3.4 million), and an average of 85.8% (SD 3.1%) of reads
355 both mapping to the genome and assigned to a genomic feature. Resulting count data from these 180
356 samples was CPM-normalized in R using the package edgeR[40] and only protein-coding genes were
357 included. Depending on the downstream analysis, genes were filtered in one of two ways for further
358 analysis: 1) genes expressed at a level of CPM > 10 in at least one library (8824 genes) or 2) genes
359 expressed at a level of CPM > 1 across all libraries (7938 genes). While these two sets are largely
360 overlapping, the first way of filtering allows identification of genes that are not expressed at all in some
361 samples but moderately expressed in others. However, as models become more complex, having some
362 samples with 0 expression causes problems for model fitting and thus having a more stringently-defined
363 set is more appropriate. To identify genes associated with early brood and egg-laying onset, we looped
364 through the set of 8824 genes and tested two linear mixed models using the 'nlme' package[41] in R to test
365 the effect of each gene on each of two reproductive phenotypes. In both cases the fixed effect was the
366 normalized gene expression level in each worm and the random effect was biological replicate. For the first
367 model, the dependent variable was early brood and for the second model the dependent variable was
368 hours to egg-laying onset. To determine the effect of environmental perturbations on gene expression, we
369 looped through the set of 7938 genes and fitted a negative binomial mixed model to each gene. In this
370 case, for each gene, fixed effects included environmental perturbations (parental age and early-life
371 temperature), the random effect included biological replicate, and the dependent variable was gene
372 expression level. For both sets of models, the summary() function was used to generate p-values, and a
373 Bonferroni correction was used to determine genes with significant effects. Gene expression variation that
374 was unexplained by the negative binomial (residual variance compared to the model) was used to
375 determine the most and least variable genes in the data after controlling for environmental effects and
376 biological replicate. Because variance increases with higher expression, we identified most and least
377 variable genes by fitting a loess model to $\log_{10}$(unexplained variance) vs. $\log_{10}$(mean CPM) for the 7938
378 genes. Z-scores were calculated for each gene relative to the loess fit to determine genes that were most
379 and least variable given their level of gene expression.

380

381 *Mediation analysis*

382 To determine the degree to which the association between early brood and gene expression
383 variation was driven by parental age, we looped through each of 7938 genes and used the 'mediate'
384 function in the psych package[42] with n.iter = 1000 and parental age as the mediating variable. To determine
385 the proportion of the gene expression effect driven by gene expression independent of parental age for the
386 significant genes identified in the original association analysis, the model coefficient for gene expression
387 when both gene expression and parental age were included in the additive model was divided by the
388 coefficient when only gene expression was included in the model.

389

390 *Prediction analysis*

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

400 To identify a set of genes that explains variance in each trait, all 7938 genes were looped through in
401 a linear regression model to identify the single gene explaining the most variance. Then the remaining
402 7937 genes were looped to identify which gene explained the most additional variance when the first gene
403 is already included in the model. This was repeated until the total variance explained was close to 1. This
404 analysis was repeated for shuffled phenotypic data as a control comparison.

405 To determine how well sets of genes would be likely to work in a new dataset, we used two
406 approaches. First, we randomly split the data (including matched phenotypic data and gene expression
407 data) into two equal groups, the train set and test set, 500 times. Each time, for the train set, the protocol

408 described above was used to identify the top 10 genes that together explain the most variance in the data
409 using a multiple regression. This set of 10 genes was then used on the test set to determine how much
410 variation is explained when the genes were not selected because of how well they work in this set. The
411 typical variance explained in the test set provides an estimate of how well the genes should explain
412 variance in a new dataset. In a second machine learning approach, we used elastic net regression using
413 the glmnet package[43] in R with alpha = 0.5 and leave-one-out cross validation. That is, we trained the
414 model for all data points except for one, then used the model to predict that remaining data point, and
415 repeated for all data points.
416
417 *Enrichment analysis*
418 To determine whether brood-associated genes and variable gene sets were enriched for particular
419 chromatin environments and tissue specificities we used our previously generated annotation for *C.*
420 *elegans* transcriptional units[44]. Briefly, in this approach we used published data[31,34] to obtain tissue
421 specificity and chromatin environment for *C. elegans* regulatory elements genome wide and then used
422 bedtools2[45] to annotate regulatory elements to the nearest gene using the coordinates of transcriptional
423 units extracted from the Wormbase gff3 (version 279).
424 To determine if there was an interaction between tissue and chromatin environment for effects on
425 early brood (Fig. 4C), we restricted genes to those expressed in the soma or germline, and in either an
426 active or regulated chromatin domain. We then performed a two-way ANOVA to determine if these factors
427 affected the effect sizes for each gene on brood (the same effect sizes plotted in Fig. 1B), followed by post-
428 doc Tukey's tests for pairwise comparisons and one-sample t-tests. To determine if gene expression
429 variation differed by chromatin domain (Fig. S3B), we performed a Wilcoxon test comparing the variance z-
430 scores for regulated genes to the variance z-scores for active genes.
431 To determine whether predictive genes were enriched for regulated or active genes as shown in
432 Figure 4D, we used the 500 sets of 10 genes for each trait described in the 'Prediction analysis' section of
433 the methods that were used to predict reproductive phenotypes (Figure 3C). We filtered the chromatin
434 dataset described above to include only genes that were classified as either active or regulated. We then
435 merged this with the background of 7938 expressed genes from our gene expression analysis. Of 10

436 predictive genes, a median of 7 genes were either active or regulated for each of the 500 iterations and
437 included in subsequent analysis. For each iteration, the proportion of regulated genes was calculated. As a
438 control, 500 random sets of 10 genes were selected from our background set, merged with the filtered
439 chromatin data, and the proportion of regulated and active genes was determined. The distribution of the
440 proportion of predictive regulated genes for each trait was compared to the proportion of regulated genes
441 from random selected sets of genes using a Kolmogorov-Smirnov statistical test.
442
443 *RNAi*
444 RNAi was performed by feeding. Colonies of RNAi bacteria (*E. coli* HT115 containing a specific
445 RNAi construct) were individually grown in 5 mL LB with carbenicillin overnight (~16-18 hours) at 37°C with
446 shaking. RNAi bacteria was obtained from the Ahringer library[46] for empty vector (ev) and for genes of
447 interest: ZK1127.7 (*cin-4*), F54C9.8 (*puf-5*), B0273.2 (*puf-6/7),* R03G5.1 (*eef-1A.2*), C34C12.8, F13B10.1
448 (*tir-1*), T07C4.9 (*nex-2*), and B0416.6 (*gly-13*). Plates used for RNAi were made with NGM, carbenicillin,
449 and IPTG as described previously[47]. These plates were seeded with RNAi bacteria and allowed to dry and
450 grow for ~30 hours. L4 larvae were plated in groups of ~10 on these plates, then singled the next day (~16-
451 18 hours later) onto new RNAi plates prepared at the same time as the original batch of plates. Adult
452 worms were allowed to lay progeny for 24 hours, then adults were removed from the plates. If a worm died
453 during this time it was censored. The number of progeny laid by each adult was counted 2 days later,
454 resulting in an early brood measurement for these experiments. For the *puf-5* and *puf-7* dose response,
455 *puf-5* and *puf-7* RNAi constructs were grown separately as described then pooled together 1:1. This mixed
456 RNAi was considered the full dose of *puf-5* and *puf-7* RNAi and was pooled with ev in the ratios shown in
457 Fig. 4C for the varying doses. Statistical analysis on early brood data was performed by comparing RNAi
458 treatment targeting a particular gene to empty vector for biological replicates within a batch. Linear mixed
459 effects models were used in the R package nlme with RNAi treatment as a fixed effect and biological
460 replicate as a random effect. The summary() function was used to generate p-values.
461
462 *Validation of isogenic worms*

We used the mRNA-seq data to validate that the frequency of genetic variants among the essentially isogenic individual worms was very low. First, we used samtools[48] to generate coverage files from the sorted bam files resulting from mRNA-seq data processing. We filtered coverage files for each worm to include only nucleotides that had at least 20 reads mapping to that nucleotide. For each worm and nucleotide, we determined whether the base call of that nucleotide was unambiguous (all 20+ reads have the same call), ambiguous with two genotypes (exactly two different bases were called), or ambiguous with more than two genotypes called. All worms with sufficient depth at a nucleotide were used to determine putative homozygous or heterozygous variants. A particular nucleotide was included in subsequent analysis if at least 10 worms had sufficient coverage at the site. This resulted in 8,361,705 unique sites and a total of 768,738,255 sites across all worms. Candidate homozygous variants included nucleotides that were called as one base unambiguously in at least one worm and called as a different base unambiguously in other worms. Candidate heterozygous variants were identified by filtering ambiguous nucleotides with two genotypes in which the frequency of each genotype was present at a frequency of between 0.3 and 0.7 and for which at least 10 other worms were unambiguously called a single genotype. These criteria reduce the possibility that a heterozygous mutant is erroneously called if a site was merely subject to rare sequencing errors. This resulted in a homozygous variant rate of 2.21×10^{-8} and a heterozygous variant rate of 9.86×10^{-7}. While already quite low, these rates likely represent over-estimates because visual inspection of bam files in some cases revealed ambiguous base calls. Over three quarters of sites (78.5%) with a putative heterozygous variant were found only in a single worm, meaning that these rare variants cannot account for an association with expression across 180 worms.

Data and code availability

All raw mRNA-seq data is available at NCBI GEO accession GSE244875. All data used to generate figures are available as part of Supplementary File 1. Scripts are available at https://github.com/amykwebster/SingleWormRNAseq and will be preserved with a DOI using Zenodo upon publication.

ACKNOWLEDGEMENTS

We thank Chien-Hui Chuang and the Bruce Bowerman lab for sharing RNAi clones. Funding was provided by grants F32GM146402 to AKW and R35GM131838 to PCP. We thank members of the Phillips lab for critical feedback. We would also like to thank WormBase.

REFERENCES

1 Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22, doi:10.1016/j.ajhg.2017.06.005 (2017).

2 Edwards, A. O. *et al.* Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421-424, doi:10.1126/science.1110189 (2005).

3 Nolte, I. M. *et al.* Missing heritability: is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study. *Eur J Hum Genet* **25**, 877-885, doi:10.1038/ejhg.2017.50 (2017).

4 Alsheikh, A. J. *et al.* The landscape of GWAS validation; systematic review identifying 309 validated non-coding variants across 130 human diseases. *BMC Med Genomics* **15**, 74, doi:10.1186/s12920-022-01216-w (2022).

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

6 Barban, N. *et al.* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat Genet* **48**, 1462-1472, doi:10.1038/ng.3698 (2016).

7 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206, doi:10.1038/nature14177 (2015).

8 Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51**, 592-599, doi:10.1038/s41588-019-0385-z (2019).

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

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

11 Torres-Garcia, S. *et al.* Epigenetic gene silencing by heterochromatin primes fungal resistance. *Nature* **585**, 453-458, doi:10.1038/s41586-020-2706-x (2020).

12 Lismer, A. *et al.* Histone H3 lysine 4 trimethylation in sperm is transmitted to the embryo and associated with diet-induced phenotypes in the offspring. *Dev Cell* **56**, 671-686 e676, doi:10.1016/j.devcel.2021.01.014 (2021).

13 Kaletsky, R. *et al.* C. elegans interprets bacterial non-coding RNAs to learn pathogenic avoidance. *Nature* **586**, 445-451, doi:10.1038/s41586-020-2699-5 (2020).

14 Evans, A. D., Fairbanks, R. A., Schmidt, P. & Levine, M. T. An epigenetic basis of adaptive plasticity in Drosophila melanogaster. *bioRxiv*, 2022.2010.2011.511590, doi:10.1101/2022.10.11.511590 (2023).

15 Werner, M. S. *et al.* Histone 4 lysine 5/12 acetylation enables developmental plasticity of Pristionchus mouth form. *Nat Commun* **14**, 2095, doi:10.1038/s41467-023-37734-z (2023).

16 Öst, A. *et al.* Paternal diet defines offspring chromatin state and intergenerational obesity. *Cell* **159**, 1352-1364 (2014).

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

18 Daines, B. *et al.* The Drosophila melanogaster transcriptome by paired-end RNA sequencing. *Genome Res* **21**, 315-324, doi:10.1101/gr.107854.110 (2011).

19 Li, B. *et al.* A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. *Sci Rep* **7**, 4200, doi:10.1038/s41598-017-04520-z (2017).

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

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

22 Lin, Y., Chen, Z. X., Oliver, B. & Harbison, S. T. Microenvironmental Gene Expression Plasticity Among Individual Drosophila melanogaster. *G3 (Bethesda)* **6**, 4197-4210, doi:10.1534/g3.116.035444 (2016).

23 Werkhoven, Z. *et al.* The structure of behavioral variation within a genotype. *Elife* **10**, doi:10.7554/eLife.64988 (2021).

24 Serra, L. *et al.* Adapting the Smart-seq2 Protocol for Robust Single Worm RNA-seq. *Bio Protoc* **8**, doi:10.21769/BioProtoc.2729 (2018).
25 Wang, A. J., Wibisono, P., Geppert, B. M. & Liu, Y. Using single-worm RNA sequencing to study C. elegans responses to pathogen infection. *BMC Genomics* **23**, 653, doi:10.1186/s12864-022-08878-x (2022).
26 Perez, M. F., Francesconi, M., Hidalgo-Carcedo, C. & Lehner, B. Maternal age generates phenotypic variation in Caenorhabditis elegans. *Nature* **552**, 106-109, doi:10.1038/nature25012 (2017).
27 Gouvea, D. Y., Aprison, E. Z. & Ruvinsky, I. Experience Modulates the Reproductive Response to Heat Stress in C. elegans via Multiple Physiological Processes. *PLoS One* **10**, e0145925, doi:10.1371/journal.pone.0145925 (2015).
28 Hirsh, D., Oppenheim, D. & Klass, M. Development of the reproductive system of Caenorhabditis elegans. *Dev Biol* **49**, 200-219, doi:10.1016/0012-1606(76)90267-0 (1976).
29 Stumpf, C. R., Kimble, J. & Wickens, M. A Caenorhabditis elegans PUF protein family with distinct RNA binding specificity. *RNA* **14**, 1550-1557, doi:10.1261/rna.1095908 (2008).
30 Lublin, A. L. & Evans, T. C. 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, doi:10.1016/j.ydbio.2006.12.004 (2007).
31 Janes, J. *et al.* Chromatin accessibility dynamics across C. elegans development and ageing. *Elife* **7**, doi:10.7554/eLife.37344 (2018).
32 Ahringer, J. & Gasser, S. M. Repressive Chromatin in Caenorhabditis elegans: Establishment, Composition, and Function. *Genetics* **208**, 491-511, doi:10.1534/genetics.117.300386 (2018).
33 Evans, K. J. *et al.* Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. *Proc Natl Acad Sci U S A* **113**, E7020-E7029, doi:10.1073/pnas.1608162113 (2016).
34 Serizay, J. *et al.* Distinctive regulatory architectures of germline-active and somatic genes in C. elegans. *Genome Res* **30**, 1752-1765, doi:10.1101/gr.265934.120 (2020).
35 Rea, S. L., Wu, D., Cypser, J. R., Vaupel, J. W. & Johnson, T. E. A stress-sensitive reporter predicts longevity in isogenic populations of Caenorhabditis elegans. *Nat Genet* **37**, 894-898, doi:10.1038/ng1608 (2005).
36 Kinser, H. E., Mosley, M. C., Plutzer, I. B. & Pincus, Z. Global, cell non-autonomous gene regulation drives individual lifespan among isogenic C. elegans. *Elife* **10**, doi:10.7554/eLife.65026 (2021).
37 Pincus, Z., Smith-Vikos, T. & Slack, F. J. MicroRNA predictors of longevity in Caenorhabditis elegans. *PLoS Genet* **7**, e1002306, doi:10.1371/journal.pgen.1002306 (2011).
38 Perez-Lluch, S. *et al.* Absence of canonical marks of active chromatin in developmentally regulated genes. *Nat Genet* **47**, 1158-1167, doi:10.1038/ng.3381 (2015).
39 Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**, e108, doi:10.1093/nar/gkt214 (2013).
40 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).
41 Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & Team, R. C. Linear and nonlinear mixed effects models. *R package version* **3**, 1-89 (2007).
42 Revelle, W. & Revelle, M. W. Package 'psych'. *The comprehensive R archive network* **337** (2015).
43 Friedman, J. *et al.* Package 'glmnet'. *CRAN R Repositary* (2021).
44 Wilson, R., Le Bourgeois, M., Perez, M. & Sarkies, P. 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, doi:10.1371/journal.pgen.1010647 (2023).
45 Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-34, doi:10.1002/0471250953.bi1112s47 (2014).
46 Kamath, R. S. & Ahringer, J. Genome-wide RNAi screening in Caenorhabditis elegans. *Methods* **30**, 313-321, doi:10.1016/s1046-2023(03)00050-1 (2003).
47 Ahringer, J. in *WormBook: The Online Review of C. elegans Biology [Internet]* (WormBook, 2006).

606 48 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079,
607 doi:10.1093/bioinformatics/btp352 (2009).
608