An atlas of gene expression variation across the *Caenorhabditis elegans* species

Gaotian Zhang[1], Nicole M. Roberto[1], Daehan Lee[1], Steffen R. Hahnel[1], and Erik C. Andersen[1,*]

1. Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA

ORCID IDs:
0000-0001-6468-1341 (G.Z.)
0000-0002-0546-8484 (D.L.)
0000-0001-8848-0691 (S.R.H.)
0000-0003-0229-9651 (E.C.A.)

[*]Corresponding author:
Erik C. Andersen
Department of Molecular Biosciences
Northwestern University
4619 Silverman Hall
2205 Tech Drive
Evanston, IL 60208
E-mail: erik.andersen@northwestern.edu

Abstract

Phenotypic variation in diverse organism-level traits have been studied in *Caenorhabditis elegans* wild strains, but differences in gene expression and the underlying variation in regulatory mechanisms are largely unknown. Here, we use natural variation in gene expression to connect genetic variants to differences in organismal-level traits, including drug and toxicant responses. We performed transcriptomic analysis on 207 genetically distinct *C. elegans* wild strains to study natural regulatory variation of gene expression. Using this massive dataset, we performed genome-wide association mappings to investigate the genetic basis underlying gene expression variation and revealed complex genetic architectures. We found a large collection of hotspots enriched for expression quantitative trait loci across the genome. We further used mediation analysis to understand how gene expression variation could underlie organism-level phenotypic variation for a variety of complex traits. These results reveal the natural diversity in gene expression and possible regulatory mechanisms in this keystone model organism, highlighting the promise of gene expression variation in shaping phenotypic diversity.

Introduction

Quantitative genetic mapping approaches, such as genome-wide association (GWA) and linkage mapping, have been used in a variety of organisms to disentangle the underlying genetic basis of gene expression variation by considering the expression level of each gene as a quantitative trait[1–9]. Expression quantitative trait loci (eQTL) affecting gene expression are often classified into local eQTL (located close to the genes that they influence) and distant eQTL (located further away from the genes that they influence)[10,11]. Local eQTL are abundant in the genome. For example, over half the genes in yeast and 94.7% of all protein-coding genes in human tissues are hypothesized to have associated local eQTL[7,8]. Genetic variants underlying local eQTL might influence the expression of a specific gene by affecting transcription factor binding sites, chromatin accessibility, other promoter elements, enhancers, or other factors at post-transcriptional levels[12]. Genes encoding diffusible factors, such as transcription factors, chromatin cofactors, and RNAs, are often considered the most likely genes to underlie distant eQTL. Distant eQTL hotspots in several species have been suggested to account for the variation in expression of many genes located throughout the genome[2,3,7,9,13]. Although a substantial amount of eQTL have been identified in different species, it is still largely unknown how gene expression variation relates to organism-level phenotypic differences.

The nematode *Caenorhabditis elegans* is a powerful model to study the genetic basis of natural variation in diverse quantitative traits[14–16]. Genome-wide gene expression variation in different developmental stages and various conditions at the whole-organism or cellular resolution have been discovered and thousands of eQTL have been identified in several studies over the past two decades[3,9,17–23]. However, most of these studies used two-parent recombinant inbred lines derived from crosses of the laboratory-adapted reference strain, N2, and the genetically diverse Hawaiian strain, CB4856. Consequently, the observed variation in gene expression and their identified eQTL were limited to the differences among a small number of *C. elegans* strains and only revealed a tiny fraction of the natural diversity of gene expression and regulatory mechanisms in this species. The *C. elegans* Natural Diversity Resource (CeNDR) has a collection of 540 genetically distinct wild *C. elegans* strains[16,24,25]. Variation in diverse organism-level phenotypes has

been observed among these wild strains, and many underlying QTL, quantitative trait genes (QTGs), and quantitative trait variants (QTVs) have been identified using GWA mappings[15,16]. Therefore, a genome-wide analysis could improve our understanding of the role of gene regulation in shaping organism-level phenotypic diversity, adaptation, and evolution of *C. elegans*.

Here, we investigated the natural variation in gene expression of 207 genetically distinct *C. elegans* wild strains by performing bulk mRNA sequencing on synchronized young adult hermaphrodites. We used GWA mappings to identify 6,545 eQTL associated with variation in expression of 5,291 transcripts of 4,520 genes. We found that local eQTL explained most of the narrow-sense heritability and showed larger effects on expression variation than distant eQTL. We identified 67 hotspots that comprise 1,828 distant eQTL across the *C. elegans* genome. We further found a diverse collection of potential regulatory mechanisms that underlie these distant eQTL hotspots. Additionally, we applied mediation analysis to gene expression and other quantitative trait variation data to elucidate putative mechanisms that can play a role in organism-level trait variation. Our results provide an unprecedented resource of transcriptome profiles and genome-wide regulatory regions that facilitate future studies. Furthermore, we demonstrate efficient methods to locate causal genes that underlie mechanisms of organism-level trait differences across the *C. elegans* species.

Results

Transcriptome profiles of 207 wild *C. elegans* strains

We obtained 207 wild *C. elegans* strains from CeNDR[25] (Fig. 1a). We grew and harvested synchronized populations of each strain at the young adult stage in independently grown and prepared biological replicates (Fig. 1b). We performed bulk RNA sequencing to measure expression levels and aligned reads to strain-specific transcriptomes (Fig. 1b, Supplementary Fig. 1, Supplementary Data 1). We focused on protein-coding genes and pseudogenes and filtered out those genes with low and/or rarely detected expression (See Methods). Because various hyper-divergent regions with

extremely high nucleotide diversity were identified in the genomes of wild *C. elegans* strains[26,27], RNA sequencing reads might be poorly aligned and expression abundances might be underestimated for genes in these regions. For each strain, we filtered out transcripts that fell into the known hyper-divergent regions. We also dropped outlier samples by comparing sample-to-sample expression distances (Supplementary Fig. 1). To further verify the homogeneity of developmental stages of our samples, we evaluated the age of each sample when they were harvested using our expression data and published time-series expression data[28]. We inferred that our animals fit an expected developmental age of 60 to 72 hours post hatching (Fig. 1c), during which time the animal is in the young adult stage. Because we harvested the animals at the first embryo-laying event, the age estimation also reflects natural variation in the duration from hatching to the beginning of embryo-laying of wild *C. elegans*. In summary, we obtained reliable expression abundance measurements for 25,849 transcripts from 16,094 genes (15,364 protein-coding genes and 730 pseudogenes) in 561 samples of 207 *C. elegans* strains (Fig. 1b, Supplementary Fig. 1, Supplementary Data 1), which we used for downstream analyses.

C. elegans geographic population structure has been observed previously[24,27,29]. Wild strains from Hawaii and other regions in the Pacific Rim harbor high genetic diversity and group into distinct clusters using genetic relatedness and principal component analysis[24,27,29]. Other strains that were isolated largely from Europe have relatively low genetic diversity because of the recent selective sweeps[24,27,29]. Similar to the previous results, the 207 strains were classified into three major groups consisting of strains from Hawaii, the Pacific coast of the United States, and Europe, respectively, in the genetic relatedness tree (Fig. 1d). Three Hawaiian strains are extremely divergent from all other strains. However, a tree constructed using transcriptome data only exhibited weak geographic relationships and no highly divergent strains (Fig. 1e), suggesting stabilizing selection has constrained variation in gene expression.

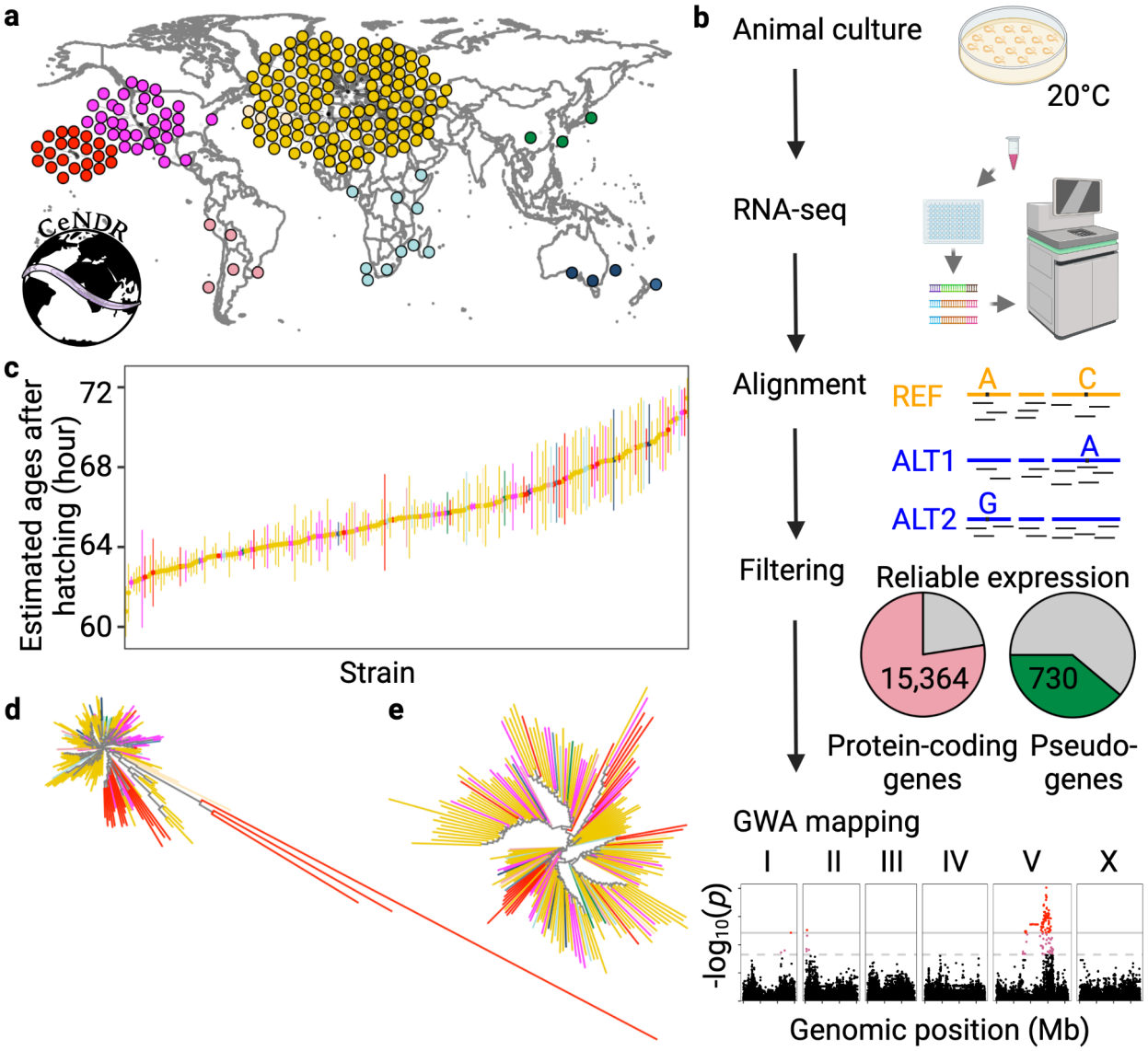


Fig. 1: Overview of species-wide expression analysis in wild *C. elegans*.

a Global distribution of 205 of the 207 wild *C. elegans* strains that were obtained from CeNDR and used in this study. Strains are colored by their sampling location continent, except for Hawaiian strains (in red). The two strains missing on the map are lacking sampling locations. **b** Graphic illustration of the workflow to acquire *C. elegans* transcriptome data. Created using BioRender.com. **c** Estimated developmental age (y-axis) of 561 well clustered samples of the 207 wild *C. elegans* strains (x-axis). Strains on the x-axis are sorted by their mean estimated age from two to three biological replicates. Error bars show standard deviation of estimated age among replicates of each strain. **d, e** Two Neighbor-joining trees of the 207 *C. elegans* strains using 851,105 biallelic segregating sites (**d**) and expression of 22,268 transcripts (**e**) are shown. Strains in **c, d, e** are colored as in **a**.

Complex regulatory genetic architectures in wild *C. elegans* strains

To estimate the association between gene expression differences and genetic variation, we calculated the broad-sense heritability (H^2) and the narrow-sense heritability (h^2) for each of the 25,849 transcript expression traits. We observed a median H^2 of 0.31 and a median h^2 of 0.06 (Fig. 2a, Supplementary Data 1), indicating strong influences from environmental factors, epistasis, or other stochastic factors on transcript expression variation[7,30,31]. Nearly 4,000 traits have a h^2 higher than 0.18, indicating a substantial heritable genetic component of the population-wide expression differences.

We performed marker-based GWA mappings to investigate the genetic basis of expression variation in the 25,849 transcripts (Supplementary Data 1). We determined the 5% false discovery rate (FDR) significance threshold for eQTL detection by mapping 40,000 permuted transcript expression traits using the EMMA algorithm[32] and the eigen-decomposition significance (EIGEN) threshold[33] (See Methods). In total, we detected 6,545 significant eQTL associated with variation in expression of 5,291 transcripts from 4,520 genes (Fig. 2b, Supplementary Data 2). The correlation of h^2 and H^2 among traits with eQTL is much higher than among traits without eQTL (Kendall's τ coefficient, 0.45 and 0.27, respectively) (Fig. 2a), indicating major roles of additive genetic variation on expression variation than other genetic factors. Likely because GWA mappings mainly detect QTL that contribute additively to trait variance, eQTL were detected for 71% of the traits with $h^2 > 0.18$, but only 11% of the remaining traits (Fig. 2a).

In close agreement to previous *C. elegans* eQTL studies using recombinant inbred advanced intercross lines (RIAILs) derived from a cross of the N2 and CB4856 strains[3,9], eQTL in this study were mostly found on chromosome arms (61%) relative to centers (33%), which is likely related to the genomic distribution of genomic variation (Table 1). Of the 4,520 genes with transcript-level eQTL, we found overrepresentation of nonessential genes (Fisher exact test, odds ratio: 1.18, *p*-value: 0.001) and underrepresentation of essential genes (Fisher exact test, odds ratio: 0.75, *p*-value: 0.001), suggesting stronger selection against expression variation in essential genes than nonessential genes[34]. Gene set enrichment analysis (GSEA) on these 4,520 genes showed that proteolysis proteasome-related genes (Fisher Exact Test, Bonferroni FDR

corrected p = 3.76E-20), especially genes encoding E3 ligases containing an F-box domain (Fisher Exact Test, Bonferroni FDR corrected p = 3.73E-15), are the most significantly enriched class (Supplementary Fig. 2, Supplementary Data 3). Other significantly enriched gene classes include metabolism (Fisher Exact Test, Bonferroni FDR corrected p = 2.92E-12), stress response (Fisher Exact Test, Bonferroni FDR corrected p = 7.24E-12), and histones (Fisher Exact Test, Bonferroni FDR corrected p = 3.23E-8). (Supplementary Fig. 2).

We classified eQTL located within a two megabase region surrounding each transcript as local eQTL and all other eQTL as distant[3,9] (Fig. 2b, Table 1, Supplementary Data 2). We identified local eQTL for 3,185 transcripts from 2,655 genes (Fig. 2b, Table 1, Supplementary Data 2). The 2,551 local eQTL that passed the Bonferroni 5% FDR threshold explained most of the estimated narrow-sense heritability (Fig. 2c). Additionally, we found 3,360 distant eQTL for 2,553 transcripts from 2,382 genes (Fig. 2b, Table 1, Supplementary Data 2). Compared to local eQTL, distant eQTL generally explained significantly lower variance (Fig. 2c, d). We found that local eQTL and up to six distant eQTL could jointly regulate the expression of transcripts (Fig. 2e). Because substantial linkage disequilibrium (LD) is observed within (r^2 > 0.6) and between (r^2 > 0.2) chromosomes in wild *C. elegans* strains[24,27,35], we calculated LD among eQTL of each of the 861 transcripts with multiple eQTL. We found low LD among most eQTL, with a median LD of r^2 = 0.19 (Supplementary Fig. 3), suggesting complex genetic architectures underlying variation in expression of these transcripts are driven by independent loci.

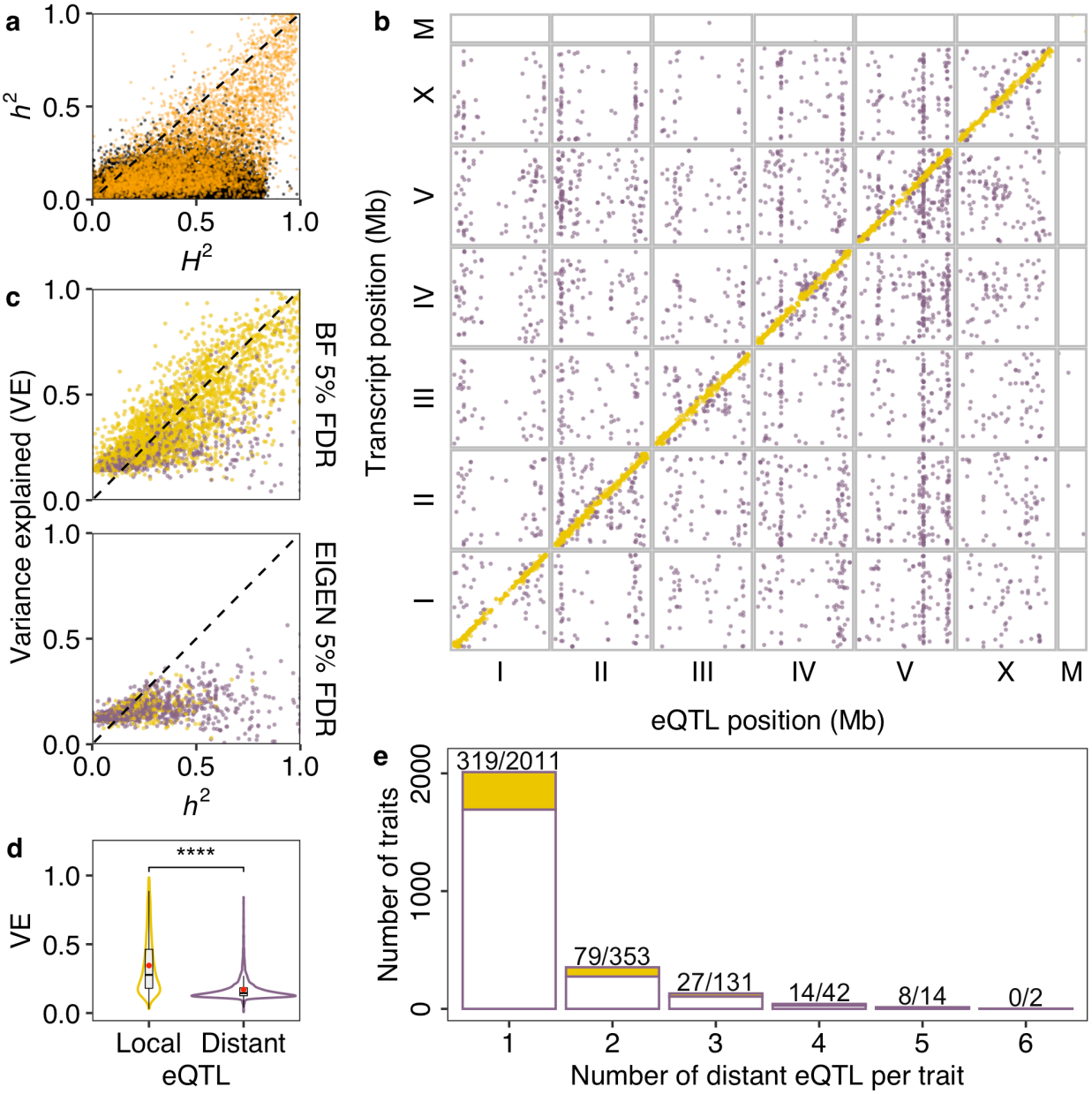


Fig. 2: Expression QTL map of 207 wild *C. elegans* strains.

a Heritability for 25,849 transcript expression traits with (orange) or without (black) detected eQTL. The narrow-sense heritability (h^2, y-axis) for each trait is plotted against the broad-sense heritability (H^2, x-axis). **b** The genomic locations of 6,545 eQTL peaks (x-axis) that pass the genome-wide EIGEN 5% FDR threshold are plotted against the genomic locations of the 5,291 transcripts with expression differences (y-axis). Golden points on the diagonal of the map represent local eQTL that colocalize with the transcripts that they influence. Purple points correspond to distant eQTL that are located further away from the transcripts that they influence. **c** The variance explained (VE) by each detected eQTL (y-axis) that passed Bonferroni (BF) 5% FDR or EIGEN 5% FDR threshold for each trait is plotted against the narrow-sense heritability h^2 (x-axis). The

dashed lines on the diagonal are shown as visual guides to represent $h^2 = H^2$ (**a**) and VE = h^2 (**c**). **d** Comparison of VE between detected local and distant eQTL shown as Violin plots. The mean VE by local or distant eQTL is indicated as red points. Statistical significance was calculated using the Wilcoxon test with *p-value* < 2e-16 indicated as ****. **e** A histogram showing the number of distant eQTL detected per transcript expression trait. One to six distant eQTL were detected for 2,553 transcript expression traits, of which 447 traits also have one local eQTL. Numbers before slashes (indicated as the golden proportion of each bar) represent the number of traits with a local eQTL in addition to their distant eQTL. Numbers after each slash on top of each bar represent the total number of traits in each category.

Table 1: The distribution of eQTL and SNVs.

Genomic domain coordinates were defined previously[36]. Transcript expression traits and SNVs used for eQTL mappings were listed.

Domain	eQTL	Local eQTL	Distant eQTL	Genome	Transcripts	SNVs
Tip	388 (5.93%)	224 (7.03%)	164 (4.88%)	7.37 Mb (7.35%)	1,712 (6.62%)	1,628 (7.76%)
Arm	3,966 (60.60%)	2,027 (63.64%)	1,939 (57.71%)	45.89 Mb (45.76%)	9,503 (36.76%)	12,883 (61.37%)
Center	2,183 (33.35%)	932 (29.26%)	1,251 (37.23%)	47.01 Mb (46.88%)	14,622 (56.57%)	6,429 (30.63%)
MtDNA	8 (0.12%)	2 (0.06%)	6 (0.18%)	0.01 Mb (0.01%)	12 (0.05%)	51 (0.24%)
Total	6,545	3,185	3,360	100.29 Mb	25,849	20,991

Diverse nature of distant eQTL hotspots

Distant eQTL were not uniformly distributed across the genome. Of the 3,360 distant eQTL, 1,828 were clustered into 67 hotspots, each of which affected the expression of 12 to 184 transcripts (Fig. 3). Signatures of selection (Tajima's *D* values) in

hotspots are mostly negative, likely because of the recent selective sweeps (Supplementary Fig. 4)[24].

GSEA on genes with transcript-level distant eQTL in each hotspot revealed potential shared transcriptional regulatory mechanisms across different genes of the same class (Supplementary Fig. 5, Supplementary Data 3). For example, the hotspot at 21.5 cM on chromosome II significantly affected the expression of heat stress related genes (Fisher Exact Test, Bonferroni FDR corrected p = 7.03E-7). Our results also showed that a single hotspot could regulate expression of genes in different classes. The hotspot at 2.5 cM on chromosome II significantly affected the expression of genes in three classes, including metallopeptidases (Fisher Exact Test, Bonferroni FDR corrected p = 1.31E-5), collagen proteins (Fisher Exact Test, Bonferroni FDR corrected p = 3.11E-9), and histones (Fisher Exact Test, Bonferroni FDR corrected p = 1.26E-6) (Supplementary Fig. 5, Supplementary Data 3). Furthermore, different hotspots could affect the expression of the same gene class. For example, the hotspot at 45.5 cM on chromosome III was also enriched with distant eQTL of histones (Fisher Exact Test, Bonferroni FDR corrected p = 8.2E-7) like the hotspot at 2.5 on chromosome II (Supplementary Fig. 5, Supplementary Data 3). Regulatory genes, such as transcription factors and chromatin cofactors, that are located in each hotspot could underlie the regulation of multiple genes. We found previously known or predicted genes encoding chromatin cofactors and transcription factors[37–39] in 24 and 59 of the 67 hotspots, respectively (Supplementary Fig. 6).

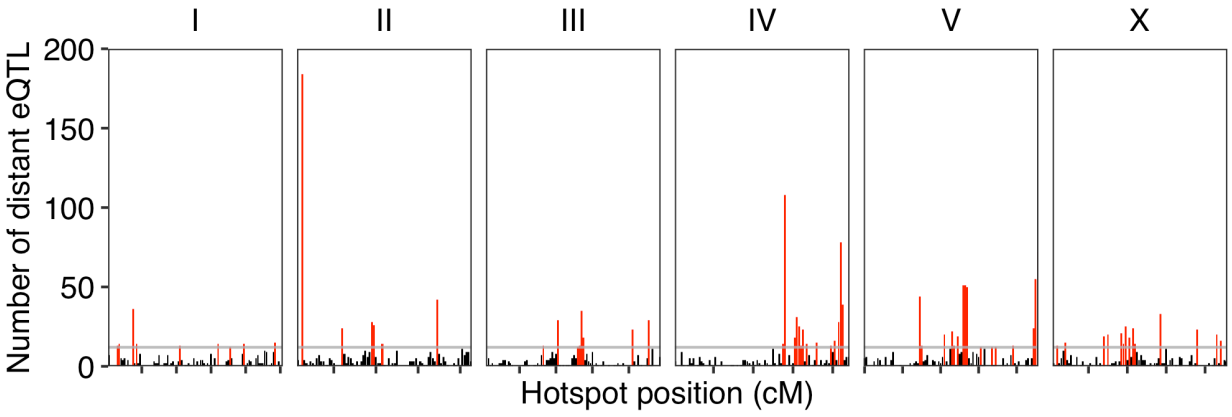


Fig. 3: Distant eQTL hotspots.

The number of distant eQTL (y-axis) in each 0.5 cM bin across the genome (x-axis) is shown. Tick marks on the x-axis denote every 10 cM. The horizontal gray line indicates the threshold of 12 eQTL. Bins with 12 or more eQTL were identified as hotspots and are colored red. Bins with fewer than 12 eQTL are colored black.

To identify causal genes and variants underlying hotspots, we performed fine mapping on distant eQTL in each hotspot and filtered for the most likely candidate variants (see Methods for details) (Supplementary Data 4). Then, we focused on the filtered candidate variants that were mapped for at least four traits in each hotspot and are in genes encoding transcription factors or chromatin cofactors. In total, we identified 36 candidate genes encoding transcription factors or chromatin cofactors for 34 hotspots. For example, the gene *ttx-1*, which encodes a transcription factor necessary for thermosensation in the AFD neurons[40,41], might underlie the expression variation of 97 transcripts with distant eQTL in three hotspots between 44.5 cM and 45.5 cM on chromosome V. TTX-1 regulates expression of *gcy-8* and *gcy-18* in AFD neurons[40,41], but no eQTL were detected for the two genes likely because we measured the expression of whole animals. Additionally, the linker histone gene *hil-2*[39] might underlie the expression variation of 46, 10, 17, and four transcripts with distant eQTL in the hotspots at 28 cM, 30.5 cM, 31 cM and 31.5 cM, respectively, on chromosome IV. We also performed GSEA for groups of transcripts whose expression traits were fine mapped to the 36 candidate genes encoding transcription factors or chromatin cofactors. For instance, the 17 traits that fine mapped to *hil-2* in the hotspot at 31 cM on chromosome IV (Supplementary Fig. 7) were enriched in E3 ligases containing an F-box domain (Fisher Exact Test, Bonferroni FDR corrected p = 0.0003) and transcription factors of the homeodomain class (Fisher Exact Test, Bonferroni FDR corrected p = 0.002). Besides the 36 candidate genes, the hundreds of other fine mapping candidates are not as transcription factors or chromatin cofactors, suggesting other mechanisms underlying distant eQTL. Altogether, as previously implicated in other species[7,11,42], our results indicate that a diverse collection of molecular mechanisms likely cause gene expression variation in *C. elegans*.

Mediation analysis facilitates candidate gene prioritization

Mediation analysis seeks to identify the mechanism that underlies the relationship between an independent variable and a dependent variable via the inclusion of an intermediary mediating variable. Because gene expression has been found to play an intermediate role between genotypes and phenotypes, it could help to identify the causal mediating genes between genotypes and phenotypes in quantitative genetics mapping studies. We have previously identified mediation effects of *scb-1* expression on responses to several chemotherapeutics and *sqst-5* expression on differential responses to exogenous zinc using linkage mapping experiments[9,43]. To validate if our expression and eQTL data can be used to identify candidate genes, we first performed mediation analysis on one published GWA study of variation in responses to the commonly used anthelmintic albendazole (ABZ)[44].

Previously, wild *C. elegans* strains were exposed to ABZ and measured for effects on development to identify genomic regions that contribute to variation in ABZ resistance. A single-marker GWA mapping was performed first to detect two QTL on chromosomes II and V, but no putative candidate gene was identified. Using a burden mapping approach, prior knowledge of ABZ resistance in parasitic nematodes, and manually curation of raw sequence read alignment files, the gene *ben-1* was found to underlie natural variation in ABZ resistance variation[44]. The single-marker GWA mapping was not able to detect an association between ABZ resistance and *ben-1* variation because of high allelic heterogeneity caused by rare SNVs and structural variants (Supplementary Fig. 8). However, rare SNVs or structural variants might lead to changes in *ben-1* expression and ABZ resistance. We found two distant eQTL, in regions overlapping with the two organism-level ABZ QTL, for *ben-1* expression variation. Therefore, these results provided an excellent opportunity to test the effectiveness of mediation analysis among organism-level phenotypes, genotype, and gene expression. The mediation estimate for *ben-1* expression was the second strongest hit in the analysis on the phenotype (animal length in response to ABZ), the genotype (GWA QTL of the phenotype), and the expression of 1,157 transcripts (Fig. 4a). We found a moderate negative correlation between the expression of *ben-1* and animal length and almost no

correlation after we regressed animal length by the expression of *ben-1* (Fig. 4b), suggesting that expression variation impacts differences in ABZ responses. We further examined genetic variants across strains and found that those strains with relatively low *ben-1* expression and high ABZ resistance all harbor SNVs or structural variants with different predicted effects (Fig. 4b), suggesting that the extreme allelic heterogeneity at the *ben-1* locus might affect ABZ response variation by reducing the abundance of this beta-tubulin. To test the impact of expression variation on phenotypic variation, we regressed animal length by expression of every transcript in our data and performed GWA mappings. Then, we compared the GWA mapping significance value after regression to the original GWA mapping significance value at a pseudo variant marker that represents all the variants in *ben-1* (Fig. 4c, Supplementary Fig. 8)[45]. We found animal length regressed by the expression of *ben-1* showed one of the largest drops in significance, and significance in most of the other mappings was approximately equal to the original significance value (Fig. 4c, Supplementary Fig. 8). These results indicated that increasing *ben-1* expression decreases resistance to ABZ and suggested the applicability of mediation analysis using the expression and eQTL data for other *C. elegans* quantitative traits.

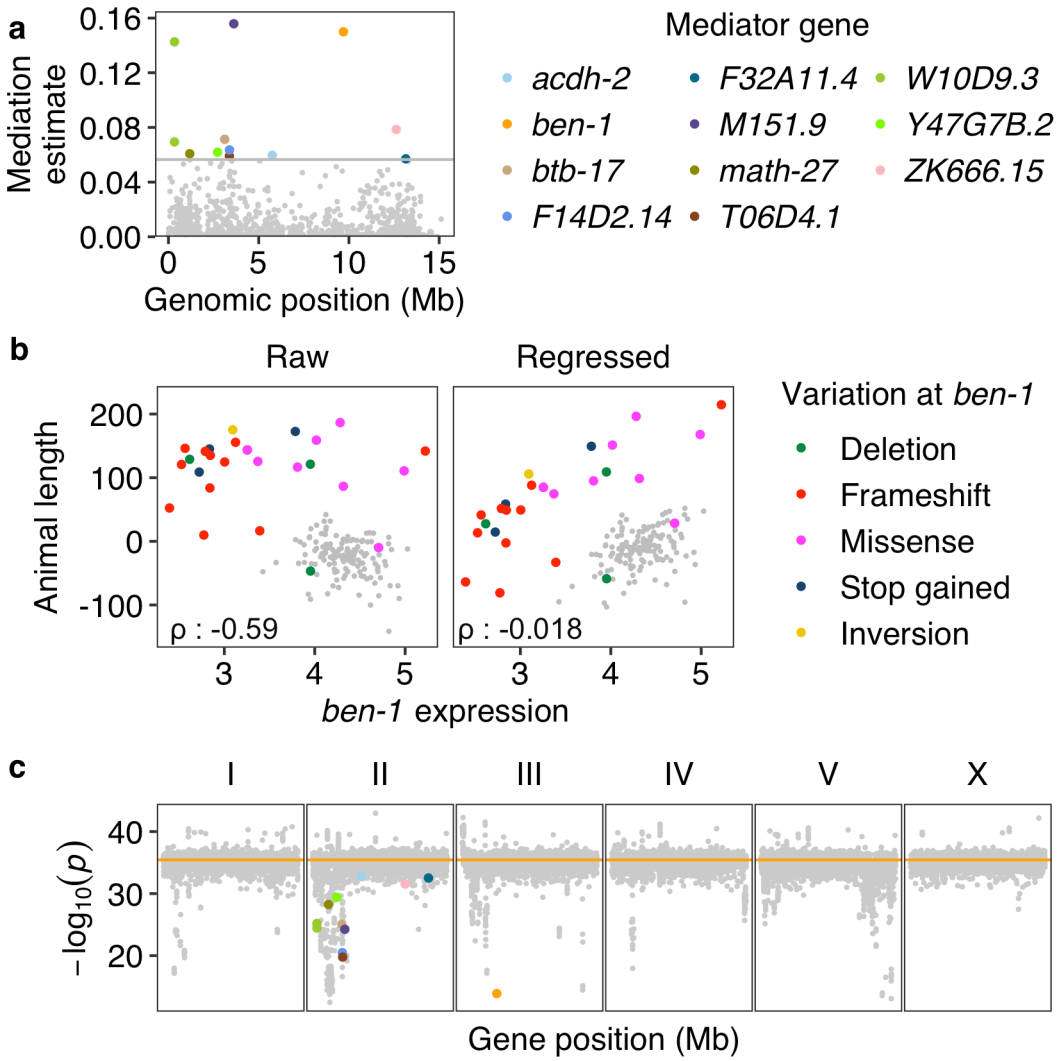


Fig. 4: Mediation effects of *ben-1* expression on *C. elegans* resistance to ABZ.

a Mediation estimates (y-axis) calculated as the indirect effect that differences in expression of each gene play in the overall phenotype are plotted against the genomic position of the eQTL (x-axis) on chromosome II. The horizontal gray line represents the 99th percentile of the distribution of mediation estimates. Significant mediators are colored other than gray by their genes as shown in the legend. **b** The correlation of *ben-1* expression (x-axis) to raw animal length and to animal length regressed by *ben-1* expression on y-axis. The Pearson's coefficient ρ for each correlation was indicated at bottom left. Strains are colored by the type of their genetic variants in *ben-1*. Strains without identified variants are colored gray. **c** Significance at the pseudo variant marker of 25,837 GWA mappings. Each point represents a GWA mapping that is plotted with its -log$_{10}$(*p*) value (y-axis) at the pseudo variant marker (III: 3,539,640) against the genomic locations (x-axis) of the transcript of which the expression was used in regression for animal length. Points for traits regressed by expression of transcripts identified as significant mediators are colored as in (**a**). The orange horizontal line represents the significance at the pseudo variant marker using the raw animal length of 167 strains

(Supplementary Fig. 8). GWA mapping results of 12 traits regressed by expression of mitochondrial genes were excluded but all with significance close to the horizontal line.

We further applied mediation analysis to another eight previously published studies of *C. elegans* natural variation and GWA mappings in diverse traits, including telomere length[46] (Fig. 5a), responses to arsenic[47] (Fig. 5b), zinc[43] (Fig. 5c), etoposide[48] (Fig. 5d), propionate[49] (Fig. 5e), abamectin[50] (Fig. 5f), dauer formation in response to pheromone[51], and lifetime fecundity[52] (Fig. 5g). Causal variants and genes that partially explained the phenotypic variation in all the eight traits, except for lifetime fecundity, have been identified using fine mappings and genome-editing experiments[43,46–52]. Only one causal gene, *dbt-1* (for arsenic response variation[47]), has eQTL detected and its expression was tested in mediation analysis for arsenic response variation[47] (Fig. 5b). No significant mediation effects were found on arsenic response variation by the expression of *dbt-1*. We also did not observe significant differential expression between strains with different alleles at the previously validated causal *dbt-1* QTV (II:7944817)[47]. Therefore, this causal variant possibly causes arsenic response variation only by affecting enzymatic activity[47] and not the abundance of the *dbt-1* transcript. Instead, we identified *bath-15* as a significant mediator gene for arsenic response variation (Fig. 5b). For the other seven organism-level traits, putative genes whose expression likely mediated the phenotypic variation were detected for six of the traits (Fig. 5). For example, the top mediator gene for the variation in responses to abamectin was *cyn-7*, which is predicted to have peptidyl-prolyl cis-trans isomerase acitivity (Fig. 5f)[53]. For the variation in lifetime fecundity (Fig. 5g), one of the putative mediator genes was *ets-4*, which is known to affect larval developmental rate, egg-laying rate, and lifespan[54]. Mediator genes suggest candidate genes in addition to those genes identified in fine mappings or linkage mappings. Taken together, we concluded that mediation analysis using the newly generated expression and eQTL data facilitates candidate gene prioritization in GWA studies.

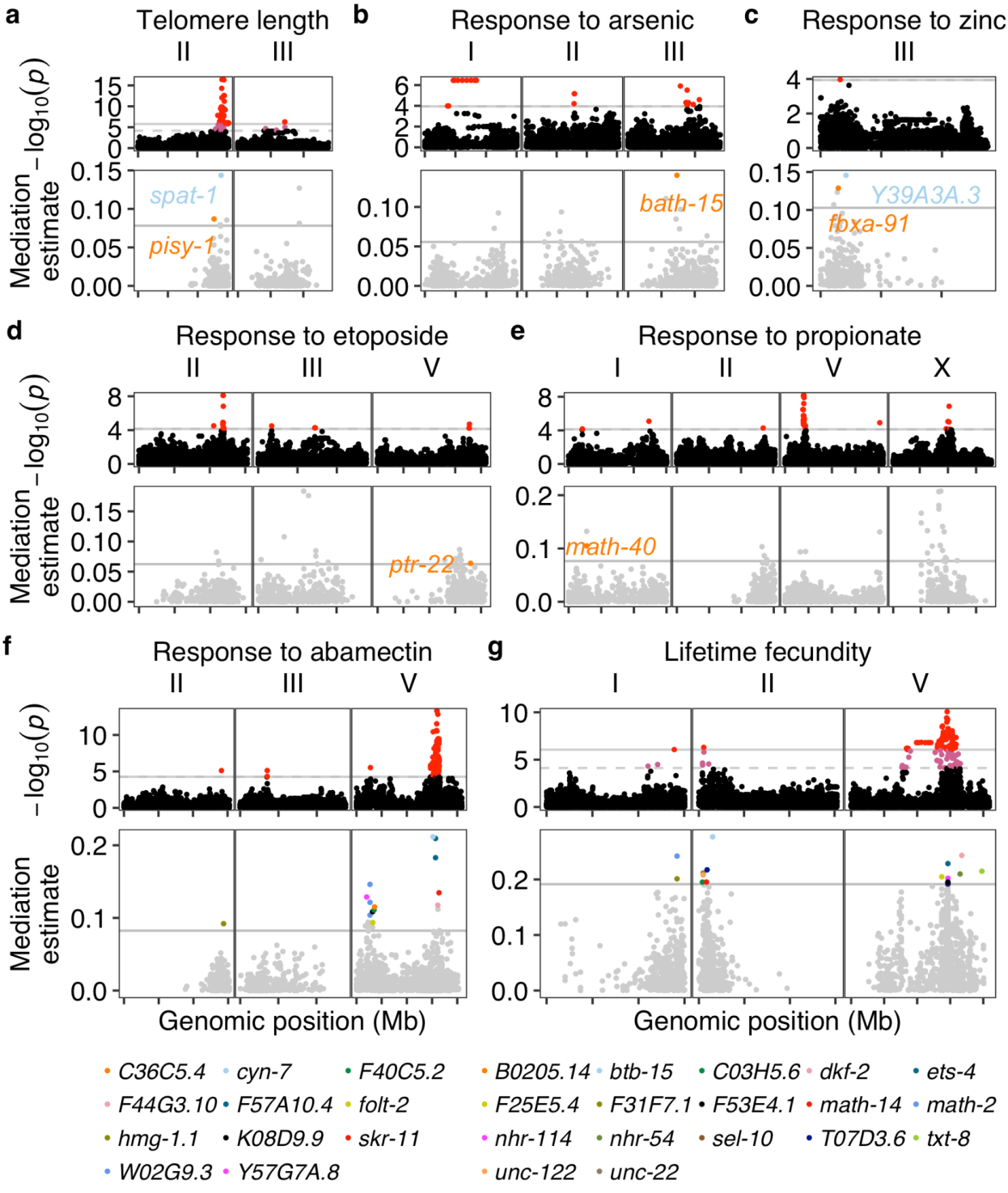


Fig. 5: Mediation effects of gene expression on variation in seven organism-level phenotypes of *C. elegans*.

GWA mapping and mediation analysis results of natural variation in *C. elegans* telomere length (**a**), responses to arsenic (**b**), zinc (**c**), etoposide (**d**), propionate (**e**), abamectin (**f**),

and lifetime fecundity (**g**). Top panel: A Manhattan plot indicating the GWA mapping result for each phenotype is shown. Each point represents an SNV that is plotted with its genomic position (x-axis) against its -$\log_{10}(p)$ value (y-axis) in mapping. SNVs that pass the genome-wide EIGEN threshold (the dotted gray horizontal line) and the genome-wide Bonferroni threshold (the solid gray horizontal line) are colored pink and red, respectively. QTL were identified by the EIGEN (**c,d,e,f**) or Bonferroni (**a,b,g**) threshold. Only chromosomes with identified QTL were shown. Bottom panel: Mediation estimates (y-axis) calculated as the indirect effect that differences in expression of each gene plays in the overall phenotype are plotted against the genomic position (x-axis) of the eQTL. The horizontal gray line represents the 99th percentile of the distribution of mediation estimates. The mediator genes with adjusted $p < 0.05$ and interpretable mediation estimate > the 99th percentile estimates threshold are colored other than gray and labeled in the panel (**a-e**) or below the panel (**f, g**). Tick marks on x-axes denote every 5 Mb.

Discussion

C. elegans was the first metazoan to have its genome sequenced and has been subjected to numerous genetic screens to identify the genes that underlie diverse traits, including programmed cell death, drug responses, development, and behaviors. Despite huge efforts by a large research community, over 60% of its genes have not been curated with functional annotations or associated with defined mutant phenotypes[55]. A likely reason is that most *C. elegans* research uses the reference strain N2 under laboratory conditions, and the functions of many genes might only be revealed in natural environments or in different genetic backgrounds[56]. In the last decade, wild *C. elegans* strains have exhibited diverse phenotypic variation in natural ecology studies[16,25,29,57–59]. Here, we provide an unprecedentedly large resource of transcriptome profiles from wild *C. elegans* strains. We used these data and GWA mappings to study gene regulation variation and detected 6,545 eQTL associated with variation in expression of 5,291 transcripts of 4,520 genes. These genes are enriched in processes, including the proteasome, metabolism, stress response, etc., suggesting gene expression regulation plays an important role in adaptation of natural *C. elegans* strains to various environments[60,61]. We identified local eQTL that explained most of the narrow-sense heritability (h^2) and significantly larger variance than distant eQTL, likely because of

higher possibilities of pleiotropy and thus stronger selection pressures. We also observed lower variation in gene expression than in genome sequence and underrepresentation of essential genes among all of the genes identified with eQTL, suggesting stabilizing selection against gene expression as previously observed in *C. elegans* and other species[5,12,62,63].

Although previous *C. elegans* eQTL studies using recombinant inbred lines have revealed rich information on the genetic basis of gene expression variation, mapping using 207 genetically distinct wild strains has the advantage of much greater genetic diversity. We reanalyzed results of one previous study that used linkage mapping to identify eQTL from the young adult stage of N2xCB4856 recombinant inbred lines[3,9]. We reclassified 1,208 local eQTL and 1,179 distant eQTL for 2,054 microarray probes of 2,003 genes (Supplementary Fig. 9a). Both the eQTL GWA and linkage mappings detected overlapping local eQTL for 454 genes and distant eQTL for 19 genes, indicating that the CB4856 strain carries the common alternative alleles among wild *C. elegans* strains for these 473 loci. However, among the 6,545 eQTL that we detected, the strains N2 and CB4856 shared the same genotypes in 4,476 eQTL, which could not be discovered using N2xCB4856 recombinant inbred lines. Alternatively, RIAILs might have less linkage disequilibrium between nearby variants and thus smaller eQTL regions of interest than eQTL in wild *C. elegans* strains. The GWA eQTL in this study have a median region of interest of 2.1 Mb (ranged from 12 kb to 18 Mb), whereas the N2xCB4856 RIAILs eQTL showed a median size of 0.55 Mb (ranged from 149 bp to 6.8 Mb), which might make the identification of underlying causal variants easier. We further found 17 distant eQTL hotspots overlapped between the two studies (Supplementary Fig. 9b). However, these shared hotspots comprise different genes between the two studies, indicating that variation in regulatory factors is not common between the linkage and association mapping studies. Future research should leverage both types of mapping studies to identify common regulatory mechanisms, focusing on local eQTL.

In addition to the high linkage disequilibrium across the *C. elegans* genome, the recently discovered hyper-divergent genomic regions made this eQTL study challenging. Approximately 20% of the genomes in some wild *C. elegans* strains were found to have

444 extremely high diversity compared to the N2 reference genome[27]. Short-sequence reads
445 of wild *C. elegans* strains often fail to align to the N2 reference genome in these regions
446 and showed lower coverage than in other regions[27]. Similarly, expression levels of genes
447 in hyper-divergent regions could be underestimated because of the poor alignment of
448 RNA-seq reads. Therefore, we only used expression of transcripts in non-divergent
449 regions to map eQTL and flagged the loci that are in common hyper-divergent regions,
450 where we are less confident in the genotypes of wild strains (Supplementary Data 2).
451 Furthermore, we only used distant eQTL that are not in common hyper-divergent regions
452 to identify hotspots. Because hyper-divergent regions were suggested to be under long-
453 term balancing selection, our estimates of Tajima's *D* in hotspots are probably biased
454 towards lower values. Future efforts using long-read sequencing are necessary to study
455 the sequence, expression, natural selection, and evolution of genes in hyper-divergent
456 regions.

457 Variation in gene expression was suggested to impact organism-level phenotypic
458 variation[7,64–66]. Combining previous GWA studies in *C. elegans* with expression of genes
459 with eQTL, we used mediation analysis to search for organism-level phenotypic variation
460 that can be explained by variation in gene expression. Compared to previous studies
461 using mediation analysis on gene expression and eQTL data from the N2xCB4856
462 recombinant inbred lines[9,43], we added a multiple testing correction procedure to our
463 mediation analysis. We performed mediation analysis on ABZ response variation[44]. The
464 causal gene *ben-1* underlying the trait was identified using a burden mapping approach[44]
465 along with prior knowledge[67,68] about the role of beta-tubulin in this drug response.
466 Although two GWA QTL on chromosomes II and V were found, they were identified likely
467 because of their interchromosomal linkage disequilibrium to variants in the *ben-1* locus[44]
468 (Supplementary Fig. 8). The single-marker GWA mapping could not associate ABZ
469 response variation because of the extreme allelic heterogeneity at the *ben-1* locus.
470 However, we used mediation analysis to identify *ben-1* without consideration of prior
471 knowledge or burden mapping results, demonstrating the power of the approach (Fig.
472 4a). We further identified significant mediators for seven other organism-level traits (Fig.

5). The expression of these mediator genes could affect the corresponding phenotypic variation, which should be validated in the future.

Mediation analysis provides an efficient hypothesis-generating approach to be performed in parallel to fine mappings. Additionally, mediator genes could contribute to organism-level phenotypic variation in addition to causal genes identified using fine mappings. One limitation of fine mappings is that searching for causal genes and variants is restricted to the QTL region of interest. Mediation analysis can make statistical connections between the organism-level phenotypes and expression of genes far away from the QTL. As mentioned above, large GWA QTL regions of interest make it difficult to identify causal genes, which require validation using genome editing. Future *C. elegans* GWA studies should use both fine mappings and mediation analysis to prioritize candidate genes. If the candidate genes overlap between the two approaches, then validation approaches can be initiated using genome editing. In cases where the two approaches identify different candidate genes, prioritization using prior knowledge across all genes identified by both approaches can inform which genes should be tested for validation using genome editing. Previous studies using fine mappings prioritized candidate genes harboring coding variants predicted to have strong functional impacts. In mediation analysis, noncoding variants that likely affect expression of mediator genes could also be nominated as candidates. For example, upstream variants were suggested to underlie expression variation of the gene *scb-1*, which mediated differences in responses to bleomycin and three other chemotherapeutics[9,69]. To summarize, we recommend using both fine mappings and mediation analyses to nominate candidate genes and variants.

The goal of quantitative genetics is to understand the genetic basis and mechanisms underlying phenotypic variation. Here, we showed that mediation analysis, which uses expression and eQTL data to search connections between genetic variants and complex traits, provides additional loci that might further explain phenotypic variation. The framework we developed for mediation analysis complements marker-based GWA mappings and is also applicable using various other intermediate traits, such as small RNAs, proteins, and metabolites. Any genes and variants underlying

503 variation in these factors can be nominated as candidates for phenotypic validation.
504 Furthermore, we could measure all of these data and complex traits from the exact same
505 samples using *C. elegans*, which can be easily grown at large scale to have synchronized
506 isogenic populations. Analyses using measurements of mRNAs, small RNAs, proteins,
507 and metabolites could strengthen conclusions about causal genes and mechanisms
508 underlying complex traits using a more holistic perspective of organismal phenotypic
509 variation. We foresee this strategy will greatly improve the powers of quantitative genetic
510 mappings in the future.

Methods

C. elegans strains

We obtained 207 wild *C. elegans* strains from *C. elegans* Natural Diversity Resource (CeNDR)[25]. Animals were cultured at 20°C on modified nematode growth medium (NGMA) containing 1% agar and 0.7% agarose to prevent burrowing and fed *Escherichia coli* (*E. coli*) strain OP50[70]. Prior to each assay, strains were grown for three generations without starvation or encountering dauer-inducing conditions[70].

Animal growth and harvest

We grew and harvested synchronized populations of each strain at the young adult stage with independently grown and prepared biological replicates. Specifically, L4 larval stage hermaphrodites were grown to the gravid adult stage on 6 cm plates and were bleached to obtain synchronized embryos. Approximately 1,000 embryos were grown on each 10 cm plate to the young adult stage and were harvested after the first embryo was observed. M9 solution was used to wash harvested animals twice to remove *E. coli*. Animals were then pelleted by centrifugation (2000 rpm for one minute) and Trizol reagent (Ambion) was added to maintain RNA integrity before storage at -80°C.

RNA extraction

Frozen samples in Trizol were thawed at room temperature and 100 µL acid-washed sand (Sigma, catalog no. 274739) was added to help to disrupt animal tissues. Then chloroform, isopropanol, and ethanol were used for phase separation, precipitation, and washing steps, respectively. Total RNA pellets were resuspended in nuclease-free water. The concentration of total RNA was determined using the Qubit RNA XR Assay Kit (Invitrogen, catalog no. Q33224). RNA quality was measured using the 2100 Bioanalyzer (Agilent). RNA samples with a minimum RNA integrity number (RIN) of 7 were used to construct Illumina sequencing libraries.

RNA library construction and sequencing

Illumina RNA-seq libraries were prepared in 96-well plates. Replicates of the same strain were prepared in different 96-well plates. For each sample, mRNA was purified and enriched from 1 µg of total RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, catalog no. E7490L). RNA fragmentation, first and second strand cDNA synthesis, and end-repair processing were performed with the NEBNext Ultra II RNA Library Prep with Sample Purification Beads (New England Biolabs, catalog no. E7775L). The cDNA libraries were adapter-ligated using adapters and unique dual indexes in the NEBNext Multiplex Oligos for Illumina (New England Biolabs, catalog no. E6440, E6442) and amplified using 12 PCR cycles. All procedures were performed according to the manufacturer's protocols. The concentration of each RNA-seq library was determined using Qubit dsDNA BR Assay Kit (Invitrogen, catalog no. Q32853). Approximately 96 RNA-seq libraries were pooled and quantified with the 2100 Bioanalyzer (Agilent) at Novogene, CA, USA. Each of the pools of libraries were sequenced on a single lane of an Illumina NovaSeq 6000 platform, yielding 150-bp paired-end (PE150) reads.

In total, RNA-seq data of 608 samples from 207 wild *C. elegans* strains in seven pooled libraries were obtained with an average of 32.6 million reads per sample and a minimum of 16.6 million reads. Of the 207 strains, 194 strains with three replicates and 13 strains with two replicates.

Sequence processing and expression abundance quantification

Adapter sequences and low-quality reads in raw sequencing data were removed using *fastp* (v0.20.0)[71]. *FastQC* (v0.11.8) analysis (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) was performed on trimmed FASTQ files to assess read quality (adapter content, read-length distribution, per read GC content, etc.). For RNA-seq mapping, SNV-substituted reference transcriptomes for each of the wild *C. elegans* strains were generated using *BCFtools* (v.1.9)[72], *gffread* (v0.11.6)[73], the N2 reference genome (WS276), a GTF file (WS276)[53], and the hard-filtered isotype variant call format (VCF) 20200815 CeNDR release (Supplementary Fig. 1).

Transposable element (TE) consensus sequences of *C. elegans* were also extracted from Dfam (release 3.3)[74] using scripts (https://github.com/fansalon/TEconsensus). We used *Kallisto* (v0.44.0) to (1) pseudoalign trimmed RNA-seq reads from each sample to the transcriptome index built from the strain-specific SNV-substituted reference transcriptome (65,173 transcripts) and TE consensus sequences (157 TEs) and (2) quantify expression abundance at the transcript level[75]. On average, 31.3 million reads pseudoaligned to the transcriptome index per sample with a minimum of 15.5 million reads, which were sufficient to capture the expression of more than 70% of the *C. elegans* reference genome genes. We used the 608 samples of 207 strains and 39,008 transcripts of protein-coding genes and pseudogenes in our analysis.

Selection of reliably expressed transcripts

We first normalized the raw counts of transcript expression abundances without the default filtering of low abundance transcripts using the R package *sleuth* (v0.30.0)[76]. Then, we filtered reliably expressed transcripts (26,043) by requiring at least five normalized counts in all the replicates of at least ten strains (Supplementary Fig. 1). We also filtered out 3,775 transcripts of 2,597 genes that are in hyper-divergent genomic regions of at least one strain. We further excluded 194 transcripts in hyper-divergent regions of more than 107 of the 207 strains. In summary, we collected reliable expression abundance for 25,849 transcripts of 16,094 genes (15,364 protein-coding genes and 730 pseudogenes).

Selection of well clustered samples

We used sample-to-sample distance to select well clustered samples (Supplementary Fig. 1). We first summarized raw counts of reliably expressed transcripts into gene-level abundances using the R package *tximport* (v1.10.1)[77]. Then, we performed variance stabilizing transformations on the gene expression profile using the *vst()* function in the R package *DESeq2* (v1.26.0), which generated $\log_2$ scale normalized expression data[78]. Sample-to-sample pairwise Euclidean distances among the 608 samples were calculated using the generic function *dist()* in R[79]. Our basic assumption is that intra-

strain distances among replicates should be smaller than inter-strain distances. Because the majority of the 207 wild *C. elegans* strains exhibit low overall genetic diversity (Fig. 1d)[24,35,80], we required that the intra-strain distances of replicates be smaller than the median of inter-strain distances of the strain to other strains. Specifically, for each strain, if all of its intra-strain distances were smaller than the median of its inter-strain distances, then all of its replicates were kept. If none of its intra-strain distances were smaller than the median of its inter-strain distances, then all samples of the strain were removed. For strains with three replicates, if one or two of its three intra-strain distances were smaller than the median of its inter-strain distances, then the two replicates with the minimum distances were kept. After removal of some outlier samples, the median of inter-strain distances would change. Therefore, we repeatedly performed the procedures of data transformation, sample-to-sample distance calculation, and filtering by comparing inter- and intra-strain distances until no more samples were removed. Eventually, 561 samples of 207 strains were selected as well clustered samples, which comprised 147 strains with three replicates and 60 strains with two replicates.

Transcript expression abundance normalization

We used the function *norm_factors()* in the R package *sleuth* (v0.30.0)[76] to compute the normalization factors for each sample using the raw transcripts per million reads (TPM) of 22,268 reliably expressed transcripts in non-divergent regions of the 207 strains and their well clustered samples. Then, we normalized the raw TPM of all the 25,849 reliably expressed transcripts of each sample with the normalization factors and used $\log_2$(normalized TPM + 0.5) for downstream analysis unless indicated otherwise.

Sample age estimation

To further verify the homogeneous developmental stage of our samples, we evaluated the age of each sample when they were harvested using the R package *RAPToR* (v1.1.3)[28] (Supplementary Fig. 1). As the requirement of the package, we first generated gene-level expression abundances. Raw TPM of 22,268 reliably expressed transcripts in non-divergent regions were summarized into abundances of 13,637 genes using the R

package *tximport* (v1.10.1)[77]. Normalization factors for each sample using gene-level abundances were calculated as described for transcript level and were used to normalize gene level TPM. Correlation of $\log_2$(normalized TPM + 0.5) of our data against the reference gene expression time series (Cel_YA_2) in *RAPToR* was computed using the function *ae()* in *RAPToR* with 10,489 intersected genes and default parameters.

Genetic and expression relatedness

Genetic variation data for 207 *C. elegans* isotypes were acquired from the hard-filtered isotype variant call format (VCF) 20200815 CeNDR release. These variants were pruned to the 851,105 biallelic single nucleotide variants (SNVs) without missing genotypes. We converted this pruned VCF file to a PHYLIP file using the *vcf2phylip.py* script[81]. Expression distance among the 207 wild strains was calculated based on the mean expression of 22,268 transcripts without missing data using the function *dist()* in R. The unrooted neighbor-joining trees for genetic and expression relatedness were made using the R packages *phangorn* (v2.5.5)[82] , *ape* (v5.6)[83] and *ggtree* (v1.14.6)[84].

eQTL mapping

Input phenotype and genotype data

For the 25,849 transcripts, we summarized the expression abundance of replicates to have the mean expression for each transcript of each strain as phenotypes used in GWA mapping (Supplementary Data 1). Genotype data for each of the 207 strains were acquired from the hard-filtered isotype VCF (20200815 CeNDR release).

Permutation-based FDR threshold

We performed GWA mapping using the pipeline *cegwas2-nf* (https://github.com/AndersenLab/cegwas2-nf). The pipeline uses the eigen-decomposition significance (EIGEN) threshold or the more stringent Bonferroni-corrected significance (BF) threshold to correct for multiple testing because of the large number of genetic markers (SNVs). To further correct for false positive QTL because of the large number of transcript expression traits, we computed a permutation-based

False Discovery Rate (FDR) at 5%. We randomly selected 200 traits from our input phenotype file and permuted each of them 200 times. These 40,000 permuted phenotypes were used as input to call QTL using *cegwas2-nf* with EIGEN and BF threshold, respectively, as previously described[47,49,52]. Briefly, we used *BCFtools*[72] to filter variants that had any missing genotype calls and variants that were below the 5% minor allele frequency. Then, we used *-indep-pairwise 50 10 0.8* in *PLINK* v1.9[85,86] to prune the genotypes to 20,991 markers with a linkage disequilibrium (LD) threshold of $r^2 < 0.8$ and then generated the kinship matrix using the *A.mat()* function in the R package *rrBLUP* (v4.6.1)[87]. The number of independent tests (N_{test}) within the genotype matrix was estimated using the R package *RSpectra* (v0.16.0) (https://github.com/yixuan/RSpectra) and *correlateR* (0.1) (https://github.com/AEBilgrau/correlateR). The eigen-decomposition significance (EIGEN) threshold was calculated as $-\log_{10}(0.05/N_{test})$. We used the *GWAS()* function in the *rrBLUP* package to perform the genome-wide mapping with the EMMA algorithm[32]. QTL were defined by at least one marker that was above the EIGEN or BF threshold. The EIGEN and BF %5 FDR was calculated as the 95 percentile of the significance of all the detected QTL under each threshold. The EIGEN and BF 5% FDR thresholds were 6.11 and 7.76, respectively.

eQTL mapping

We performed GWA mapping on the expression traits of the 25,849 transcripts as for permuted expression traits but using the EIGEN 5% FDR (6.11) as the threshold. We identified QTL with significance that also passed the Bonferroni 5% FDR threshold to locate the best estimate of QTL positions with the highest significance. We used the generic function *cor()* in R and Pearson correlation coefficient to calculate the phenotypic variance explained by each QTL. We used the *LD()* function from the R package *genetics* (v1.3.8.1.2) (https://cran.r-project.org/package=genetics) to calculate the LD correlation coefficient r^2 among QTL for traits with multiple eQTL.

eQTL classification

Local eQTL were classified if the QTL was within a 2 Mb region surrounding the transcript. All other QTL were classified as distant.

Heritability calculation

Heritability estimates were calculated for each of the 25,849 traits used for eQTL mapping as previously described[88]. Narrow-sense heritability (h^2) was calculated with the phenotype file and pruned genotypes in eQTL mapping using the functions *mmer()* and *pin()* in the R package *sommer* (v4.1.2)[89]. Broad-sense heritability (H^2) was calculated using expression of replicates of each strain and the *lmer* function in the R package *lme4* (v1.1.21) with the model phenotype ~ 1 + (1|strain)[90].

Hotspot identification

We first filtered out distant eQTL in common hyper-divergent genomic regions of wild C. elegans strains. Common hyper-divergent regions were defined among our 206 strains (reference N2 excluded) as described previously[27]. Briefly, we divided the genome into 1 kb bins and calculated the percentage of 206 strains that are classified as hyper-divergent in each bin. Common hyper-divergent regions were defined as bins ≥ 5%[27].

Distant eQTL hotspots were identified by dividing the genome into 0.5 cM bins and counting the number of non-divergent distant eQTL that mapped to each bin. Significance was determined as bins with more eQTL than the 99th percentile of a Poisson distribution using the maximum likelihood method and the function eqpois() in the R package EnvStats[1,3,9,91].

Reanalysis of RIAILs eQTL

We reclassified eQTL detected in a previous study using microarray expression data from synchronized young adult populations of 208 recombinant inbred advanced intercross lines (RIAILs) derived from N2 and CB4856[9,36]. A total of 2,540 eQTL from 2,196 probes were identified using linkage mappings[9]. We selected 2,387 eQTL of 2,054 probes that are in 2,003 live genes based on the probe-gene list in the R package

719 *linkagemapping* (https://github.com/AndersenLab/linkagemapping) and the GTF file
720 (WS276)[53]. We classified 1,208 local eQTL and 1,179 distant eQTL as described above.
721 We further identified hotspots as above for 1,124 distant eQTL that are not in the hyper-
722 divergent regions of CB4856.
723
724 **Population genetics**
725 We use 851,105 biallelic SNVs with no missing calls among the 207 strains from the
726 hard-filtered VCF 20200815 CeNDR release to calculate population genomic statistics.
727 Tajima's D, Watterson's θ, and pi were all calculated using *scikit-allel*[92]. Each of these
728 statistics was calculated for non-overlapping 1,000-bp windows across the genome.
729
730 **Fine mapping for causal genes underlying hotspots**
731 For transcript expression traits with distant eQTL in hotspots, we performed fine
732 mapping using the pipeline *cegwas2-nf* as previously described[47]. Briefly, we defined
733 QTL regions of interest from the GWA mapping as +/- 100 SNVs from the rightmost and
734 leftmost markers above the EIGEN 5% FDR significance threshold. Then, using
735 genotype data from the imputed hard-filtered isotype VCF (20200815 CeNDR release),
736 we generated a QTL region of interest genotype matrix that was filtered as described
737 above, with the one exception that we did not perform LD pruning. We used *PLINK*
738 v1.9[85,86] to extract the LD between the markers used for fine mapping and the QTL peak
739 marker identified from GWA mappings. We used the same command as above to
740 perform fine mappings. To identify causal genes and variants that affect expression of
741 several transcripts underlying hotspots, we retained the fine-mapped candidate variants
742 that passed the following per trait filters: top 5% most significant markers; out of
743 common hyper-divergent genomic regions; with negative BLOSUM[93] scores as
744 characterized and annotated in CeNDR[25].
745
746 **Enrichment analysis**

Gene set enrichment analyses were carried out for all genes found with transcript-level eQTL and for genes with distant eQTL in each hotspot using the web-based tool *WormCat*[39].

Mediation analysis

GWA mapping of diverse C. elegans *phenotypes*

We obtained nine different phenotype data used in previous *C.elegans* natural variation and GWA studies[43,44,46–52]. We filtered genetically distinct isotype strains for each trait based on CeNDR (20200815 release) and performed GWA mapping as for permuted expression traits but mostly using EIGEN or BF as the threshold according to the original studies. GWA was performed under EIGEN for two studies originally using BF as the threshold[48,49].

Mediation analysis

For each QTL of the above phenotypes, we used the genotype (*Exposure*) at the phenotype QTL peak, transcript expression traits (*Mediator*) that have eQTL overlapped with the phenotype QTL, and the phenotype (*Outcome*) as input to perform mediation analysis using the *medTest()* function and 1,000 permutations for *p*-value correction in the R package *MultiMed* (v2.6.0) (https://bioconductor.org/packages/release/bioc/html/MultiMed.html). For mediation, we used only strains with all of the three input data types available and where variation was found. For instance, between the 202 strains used in the study of ABZ resistance[44] and the 207 strains used in this study, 167 strains overlapped. Although we searched overlapped eQTL against QTL in the GWA mapping for ABZ resistance using 202 strains (Supplementary Fig. 8), 167 strains at maximum were used in mediation analysis. Furthermore, because some transcripts were found in hyper-divergent regions in certain strains and their expression data were filtered out, the rest of the strains with all of the data types available might contain no variation in one or all of the three data types and were not used in mediation analysis. For example, we found 1,193 eQTL overlapped with

776 the QTL on chromosome II in the GWA mapping for ABZ resistance, but only 1,157
777 mediation analyses were performed.
778
779 For mediators with adjusted $p < 0.05$ or mediation estimate greater than the 99th
780 percentile of the distribution of mediation estimates, we performed a second mediation
781 analysis as described previously[9] using the *mediate()* function from the R package
782 *mediation* (version 4.5.0)[94] to filter out the uninterpretable results where the proportion of
783 the total effect (the estimated effect of genotype on phenotype, ignoring expression) that
784 can be explained by the mediation effect (the estimated effect of expression on
785 phenotype) is negative or larger than 100%.
786
787 *GWA of traits regressed by transcript expression*
788 We regressed the trait animal length (q90.TOF)[44] by expression of every transcript using
789 the generic function *residuals()* in R, which fits a linear model with the formula (*phenotype*
790 ~ *expression*) to account for any differences in phenotype parameters present in
791 transcript expression. Then GWA was performed for each regressed trait as for permuted
792 expression traits using BF as the threshold.
793
794
795

Acknowledgements

We would like to thank Stefan Zdraljevic and Samuel J. Widmayer for helpful suggestions. G.Z. is supported by the NSF-Simons Center for Quantitative Biology at Northwestern University (awards Simons Foundation/SFARI 597491-RWC and the National Science Foundation 1764421). S.R.H. was funded by a DFG fellowship (HA 8449/1-1) from the Deutsche Forschungsgemeinschaft (www.dfg.de). E.C.A. is supported by a National Science Foundation CAREER Award (IOS-1751035) and a grant from the National Institutes of Health R01 DK115690. The *C. elegans* Natural Diversity Resource is supported by a National Science Foundation Living Collections Award to R.E.T. and E.C.A. (1930382). We also like to thank WormBase for which these analyses would not have been possible.

Author contributions

E.C.A. conceived of and designed the study. D.L., S.R.H., and G.Z. prepared *C. elegans* cultures. G.Z. and N.M.R. performed RNA-seq experiments. G.Z. analyzed the data. G.Z. and E.C.A. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Data and code availability

The raw RNA-seq data generated in this study are available at the NCBI Sequence Read Archive (Project PRJNA669810). The raw expression counts and TPM quantified in this study are available at NCBI's Gene Expression Omnibus (Series GSE186719). The RNA-seq mapping pipeline can be found at https://github.com/AndersenLab/PEmRNA-seq-nf. The mediation analysis pipeline can be found at https://github.com/AndersenLab/mediation_GWAeQTL. The datasets and code for generating all figures can be found at https://github.com/AndersenLab/WI-Ce-eQTL.

References

1. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
2. West, M. A. L. *et al.* Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* **175**, 1441–1450 (2007).
3. Rockman, M. V., Skrovanek, S. S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in C. elegans. *Science* **330**, 372–376 (2010).
4. Zan, Y., Shen, X., Forsberg, S. K. G. & Carlborg, Ö. Genetic Regulation of Transcriptional Variation in Natural Arabidopsis thaliana Accessions. *G3* **6**, 2319–2328 (2016).
5. Kita, R., Venkataram, S., Zhou, Y. & Fraser, H. B. High-resolution mapping of cis-regulatory variation in budding yeast. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E10736–E10744 (2017).
6. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
7. Albert, F. W., Bloom, J. S., Siegel, J., Day, L. & Kruglyak, L. Genetics of trans-regulatory variation in gene expression. *Elife* **7**, 1–39 (2018).
8. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
9. Evans, K. S. & Andersen, E. C. The Gene scb-1 Underlies Variation in Caenorhabditis elegans Chemotherapeutic Responses. *G3* **10**, 2353–2364 (2020).
10. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).
11. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
12. Hill, M. S., Vande Zande, P. & Wittkopp, P. J. Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* **22**, 203–215 (2021).

13. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
14. Gaertner, B. E. & Phillips, P. C. Caenorhabditis elegans as a platform for molecular quantitative genetics and the systems biology of natural variation. *Genet. Res.* **92**, 331–348 (2010).
15. Snoek, B. L. *et al.* WormQTL2: an interactive platform for systems genetics in Caenorhabditis elegans. *Database* **2020**, (2020).
16. Evans, K. S., van Wijk, M. H., McGrath, P. T., Andersen, E. C. & Sterken, M. G. From QTL to gene: C. elegans facilitates discoveries of the genetic mechanisms underlying natural variation. *Trends Genet.* **0**, (2021).
17. Li, Y. *et al.* Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. *PLoS Genet.* **2**, e222 (2006).
18. Viñuela, A., Snoek, L. B., Riksen, J. A. G. & Kammenga, J. E. Genome-wide gene expression regulation as a function of genotype and age in C. elegans. *Genome Res.* **20**, 929–937 (2010).
19. Li, Y. *et al.* Global genetic robustness of the alternative splicing machinery in Caenorhabditis elegans. *Genetics* **186**, 405–410 (2010).
20. Sterken, M. G. *et al.* Ras/MAPK Modifier Loci Revealed by eQTL in Caenorhabditis elegans. *G3* **7**, 3185–3193 (2017).
21. Snoek, B. L. *et al.* Contribution of trans regulatory eQTL to cryptic genetic variation in C. elegans. *BMC Genomics* **18**, 500 (2017).
22. Ben-David, E. *et al.* Whole-organism eQTL mapping at cellular resolution with single-cell sequencing. *Elife* **10**, (2021).
23. Snoek, B. L. *et al.* The genetics of gene expression in a Caenorhabditis elegans multiparental recombinant inbred line population. *G3* **11**, (2021).
24. Andersen, E. C. *et al.* Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat. Genet.* **44**, 285–290 (2012).
25. Cook, D. E., Zdraljevic, S., Roberts, J. P. & Andersen, E. C. CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic Acids Res.* **45**, D650–D657 (2017).

881 26. Thompson, O. A. *et al.* Remarkably Divergent Regions Punctuate the Genome
882 Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856. *Genetics* **200**,
883 975–989 (2015).
884 27. Lee, D. *et al.* Balancing selection maintains hyper-divergent haplotypes in
885 Caenorhabditis elegans. *Nat Ecol Evol* 1–14 (2021).
886 28. Bulteau, R. & Francesconi, M. Real age prediction from the transcriptome with
887 RAPToR. *bioRxiv* 2021.09.07.459270 (2021) doi:10.1101/2021.09.07.459270.
888 29. Crombie, T. A. *et al.* Deep sampling of Hawaiian Caenorhabditis elegans reveals
889 high genetic diversity and admixture with global populations. *Elife* **8**, e50465
890 (2019).
891 30. Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V. & Kruglyak, L. Finding the
892 sources of missing heritability in a yeast cross. *Nature* **494**, 234–237 (2013).
893 31. Chen, A., Liu, Y., Williams, S. M., Morris, N. & Buchner, D. A. Widespread epistasis
894 regulates glucose homeostasis and gene expression. *PLoS Genet.* **13**, e1007025
895 (2017).
896 32. Kang, H. M. *et al.* Efficient control of population structure in model organism
897 association mapping. *Genetics* **178**, 1709–1723 (2008).
898 33. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues
899 of a correlation matrix. *Heredity* **95**, 221–227 (2005).
900 34. Campos, T. L., Korhonen, P. K., Sternberg, P. W., Gasser, R. B. & Young, N. D.
901 Predicting gene essentiality in Caenorhabditis elegans by feature engineering and
902 machine-learning. *Comput. Struct. Biotechnol. J.* **18**, 1093–1102 (2020).
903 35. Cutter, A. D. Nucleotide polymorphism and linkage disequilibrium in wild
904 populations of the partial selfer Caenorhabditis elegans. *Genetics* **172**, 171–184
905 (2006).
906 36. Rockman, M. V. & Kruglyak, L. Recombinational landscape and population
907 genomics of Caenorhabditis elegans. *PLoS Genet.* **5**, e1000419 (2009).
908 37. Araya, C. L. *et al.* Regulatory analysis of the C. elegans genome with
909 spatiotemporal resolution. *Nature* **512**, 400–405 (2014).
910 38. Kudron, M. M. *et al.* The ModERN Resource: Genome-Wide Binding Profiles for

Hundreds of Drosophila and Caenorhabditis elegans Transcription Factors. *Genetics* **208**, 937–949 (2018).

39. Holdorf, A. D. *et al.* WormCat: An Online Tool for Annotation and Visualization of Caenorhabditis elegans Genome-Scale Data. *Genetics* **214**, 279–294 (2020).
40. Satterlee, J. S. *et al.* Specification of thermosensory neuron fate in C. elegans requires ttx-1, a homolog of otd/Otx. *Neuron* **31**, 943–956 (2001).
41. Kagoshima, H. & Kohara, Y. Co-expression of the transcription factors CEH-14 and TTX-1 regulates AFD neuron-specific genes gcy-8 and gcy-18 in C. elegans. *Dev. Biol.* **399**, 325–336 (2015).
42. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
43. Evans, K. S. *et al.* Natural variation in the sequestosome-related gene, sqst-5, underlies zinc homeostasis in Caenorhabditis elegans. *PLoS Genet.* **16**, e1008986 (2020).
44. Hahnel, S. R. *et al.* Extreme allelic heterogeneity at a Caenorhabditis elegans beta-tubulin locus explains natural resistance to benzimidazoles. *PLoS Pathog.* **14**, e1007226 (2018).
45. Churchill, G. A., Gatti, D. M., Munger, S. C. & Svenson, K. L. The Diversity Outbred mouse population. *Mamm. Genome* **23**, 713–718 (2012).
46. Cook, D. E. *et al.* The Genetic Basis of Natural Variation in Caenorhabditis elegans Telomere Length. *Genetics* **204**, 371–383 (2016).
47. Zdraljevic, S. *et al.* Natural variation in C. elegans arsenic toxicity is explained by differences in branched chain amino acid metabolism. *Elife* **8**, e40260 (2019).
48. Zdraljevic, S. *et al.* Natural variation in a single amino acid substitution underlies physiological responses to topoisomerase II poisons. *PLoS Genet.* **13**, e1006891 (2017).
49. Na, H., Zdraljevic, S., Tanny, R. E., Walhout, A. J. M. & Andersen, E. C. Natural variation in a glucuronosyltransferase modulates propionate sensitivity in a C. elegans propionic acidemia model. *PLoS Genet.* **16**, e1008984 (2020).
50. Evans, K. S. *et al.* Two novel loci underlie natural differences in Caenorhabditis

elegans abamectin responses. *PLoS Pathog.* **17**, e1009297 (2021).

51. Lee, D. *et al.* Selection and gene flow shape niche-associated variation in pheromone response. *Nat Ecol Evol* **3**, 1455–1463 (2019).
52. Zhang, G., Mostad, J. D. & Andersen, E. C. Natural variation in fecundity is correlated with species-wide levels of divergence in Caenorhabditis elegans. *G3* (2021) doi:10.1093/g3journal/jkab168.
53. Harris, T. W. *et al.* WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res.* **48**, D762–D767 (2020).
54. Thyagarajan, B. *et al.* ETS-4 is a transcriptional regulator of life span in Caenorhabditis elegans. *PLoS Genet.* **6**, e1001125 (2010).
55. Petersen, C., Dirksen, P. & Schulenburg, H. Why we need more ecology for genetic models such as C. elegans. *Trends Genet.* **31**, 120–127 (2015).
56. Sterken, M. G., Snoek, L. B., Kammenga, J. E. & Andersen, E. C. The laboratory domestication of Caenorhabditis elegans. *Trends Genet.* **31**, 224–231 (2015).
57. Frézal, L. & Félix, M.-A. The natural history of model organisms: C. elegans outside the Petri dish. *Elife* **4**, e05849 (2015).
58. Schulenburg, H. & Félix, M.-A. The Natural Biotic Environment of Caenorhabditis elegans. *Genetics* **206**, 55–86 (2017).
59. Crombie, T. A. *et al.* Local adaptation and spatiotemporal patterns of genetic diversity revealed by repeated sampling of Caenorhabditis elegans across the Hawaiian Islands. *bioRxiv* 2021.10.11.463952 (2021) doi:10.1101/2021.10.11.463952.
60. Thomas, J. H. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. *Genome Res.* **16**, 1017–1030 (2006).
61. de Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nat. Rev. Genet.* **12**, 833–845 (2011).
62. Denver, D. R. *et al.* The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. *Nat. Genet.* **37**, 544–548 (2005).
63. Rifkin, S. A., Houle, D., Kim, J. & White, K. P. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* **438**, 220–

223 (2005).

64. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
65. Oliver, F. *et al.* Regulatory variation at glypican-3 underlies a major growth QTL in mice. *PLoS Biol.* **3**, e135 (2005).
66. Courtier-Orgogozo, V., Arnoult, L., Prigent, S. R., Wiltgen, S. & Martin, A. Gephebase, a database of genotype-phenotype relationships for natural and domesticated variation in Eukaryotes. *Nucleic Acids Res.* **48**, D696–D703 (2020).
67. Driscoll, M., Dean, E., Reilly, E., Bergholz, E. & Chalfie, M. Genetic and molecular analysis of a Caenorhabditis elegans beta-tubulin that conveys benzimidazole sensitivity. *J. Cell Biol.* **109**, 2993–3003 (1989).
68. Kwa, M. S., Veenstra, J. G. & Roos, M. H. Benzimidazole resistance in Haemonchus contortus is correlated with a conserved mutation at amino acid 200 in beta-tubulin isotype 1. *Mol. Biochem. Parasitol.* **63**, 299–303 (1994).
69. Brady, S. C. *et al.* A Novel Gene Underlies Bleomycin-Response Variation in Caenorhabditis elegans. *Genetics* **212**, 1453–1468 (2019).
70. Andersen, E. C., Bloom, J. S., Gerke, J. P. & Kruglyak, L. A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology. *PLoS Genet.* **10**, e1004156 (2014).
71. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
72. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
73. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res.* **9**, 304 (2020).
74. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
75. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-

seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

76. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687–690 (2017).
77. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).
78. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
79. Core Team, R. & Others. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. *Available* (2013).
80. Barrière, A. & Félix, M.-A. Natural variation and population genetics of Caenorhabditis elegans. *WormBook* 1–19 (2005).
81. Ortiz, E. M. *vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis*. (2019). doi:10.5281/zenodo.2540861.
82. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
83. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
84. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
85. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
86. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
87. Endelman, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **4**, 250–255 (2011).
88. Wit, J., Rodriguez, B. C. & Andersen, E. C. Natural variation in Caenorhabditis elegans responses to the anthelmintic emodepside. *Int. J. Parasitol. Drugs Drug*

Resist. **16**, 1–8 (2021).

89. Covarrubias-Pazaran, G. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLoS One* **11**, e0156744 (2016).
90. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles* **67**, 1–48 (2015).
91. Millard, S. P. *EnvStats: An R Package for Environmental Statistics*. (Springer-Verlag New York, 2013).
92. Miles, A., Ralph, P., Rae, S. & Pisupati, R. *cggh/scikit-allel: v1.2.1*. (2019). doi:10.5281/zenodo.3238280.
93. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).
94. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software, Articles* **59**, 1–38 (2014).

Supplementary Information

An atlas of gene expression variation across the *Caenorhabditis elegans* species

Gaotian Zhang[1], Nicole M. Roberto[1], Daehan Lee[1], Steffen R. Hahnel[1], and Erik C. Andersen[1,*]

1. Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA

Description of Additional Supplementary Files

File Name: Supplementary Data 1

Description: Expression abundances and heritabilities of 25,849 transcripts in 207 wild *C. elegans* strains

File Name: Supplementary Data 2

Description: GWA mapping results of 5,291 transcript expression traits, eQTL classification, and distant eQTL hotspots

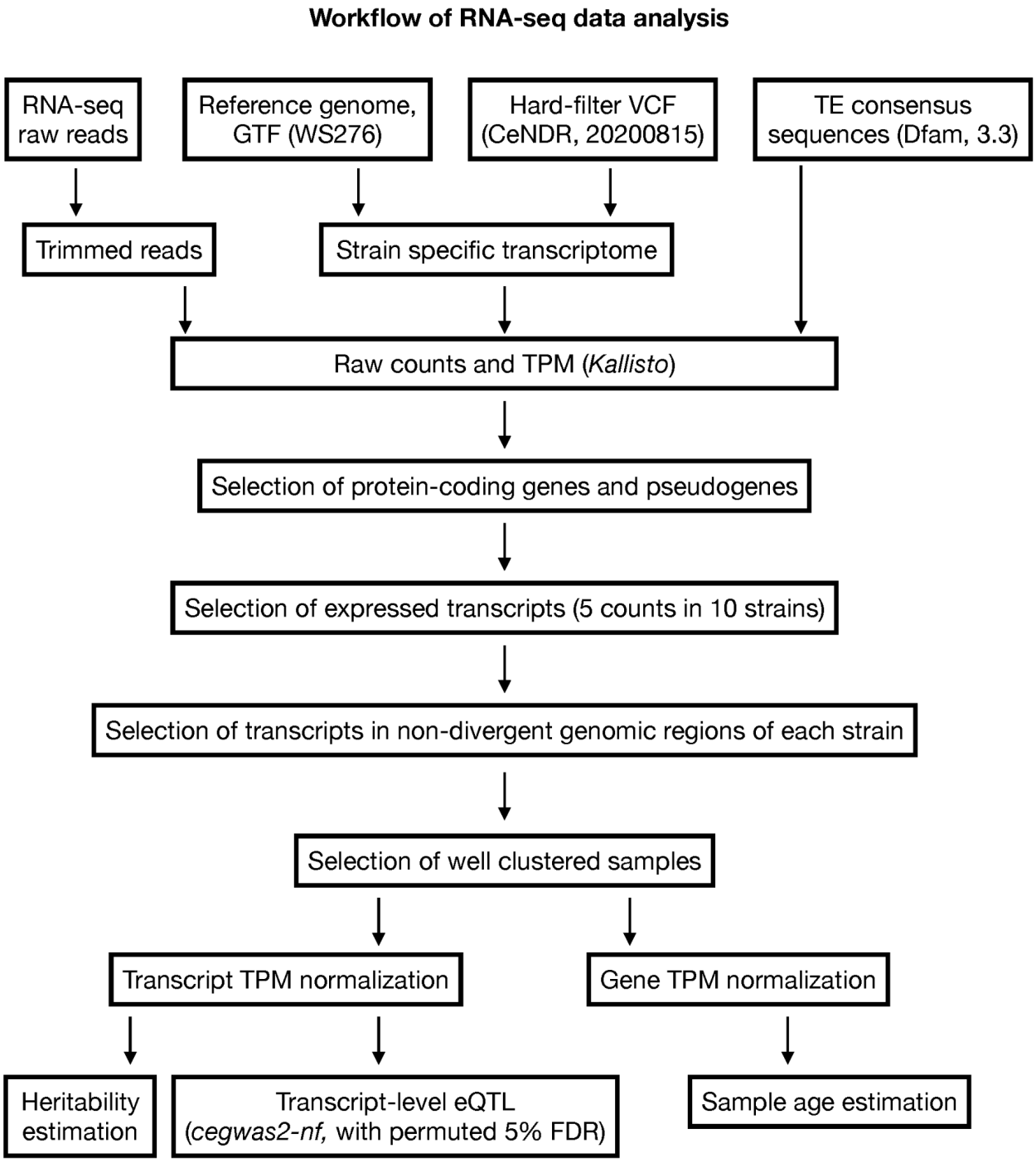
File Name: Supplementary Data 3

Description: Enrichment of genes with eQTL, distant eQTL in hotspots

File Name: Supplementary Data 4

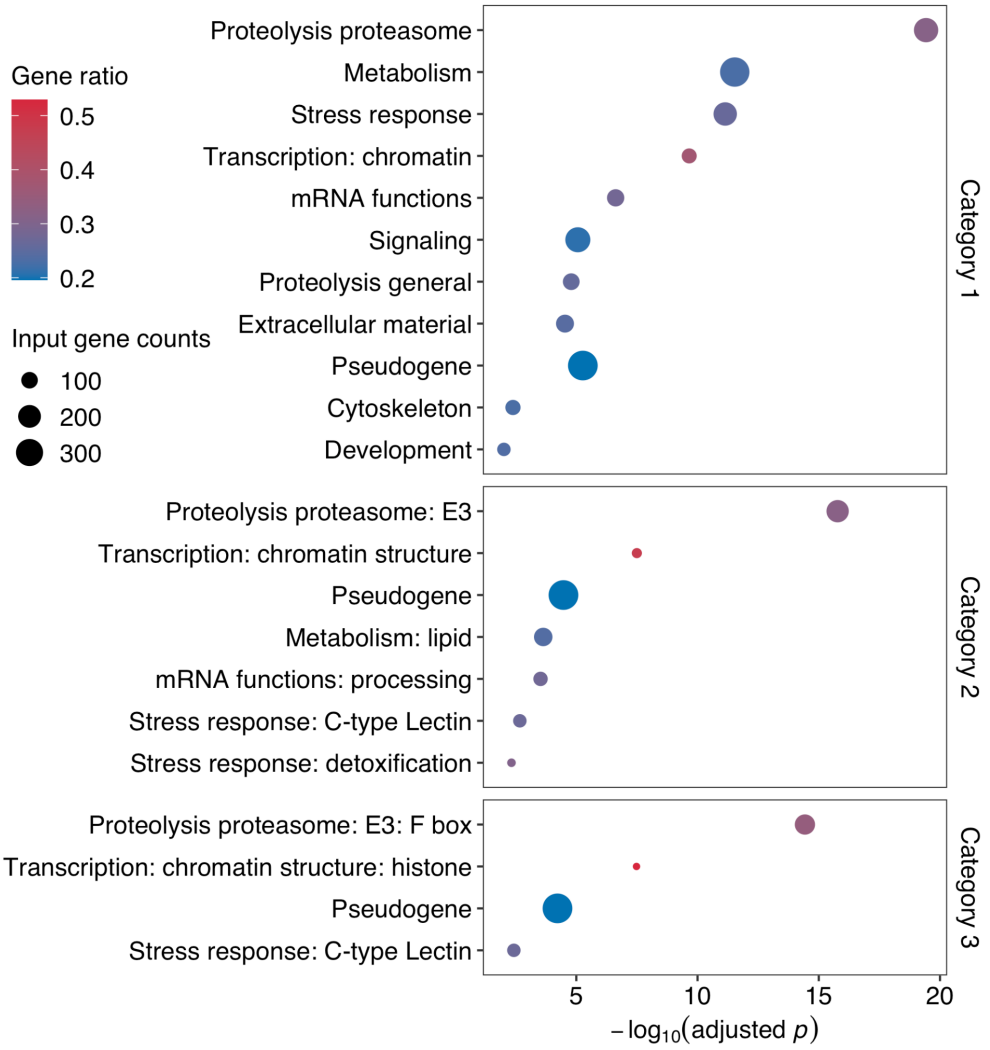
Description: Fine mappings for distant eQTL in hotspots

Supplementary Figures



Supplementary Fig. 1

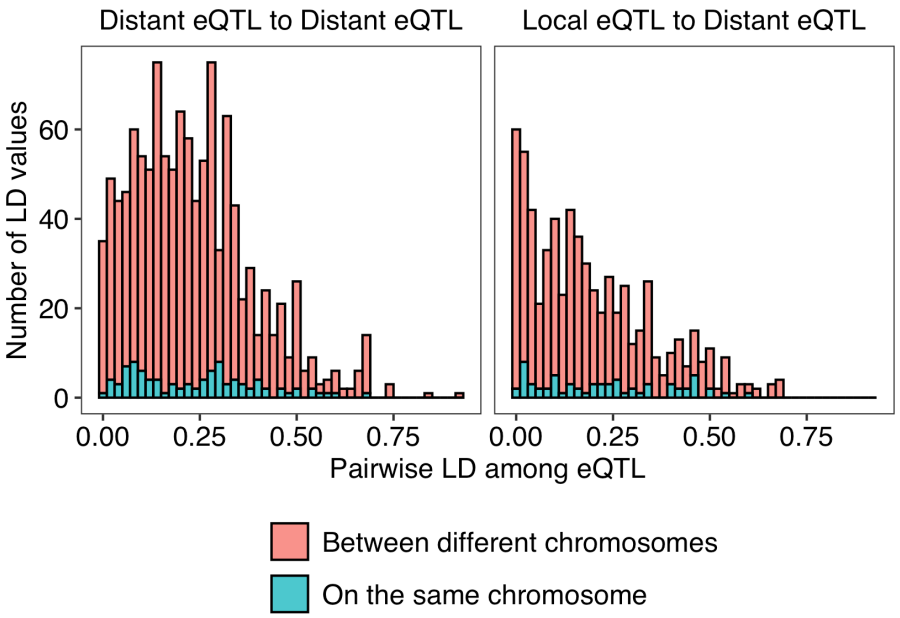
Workflow of RNA-seq data processing.



37

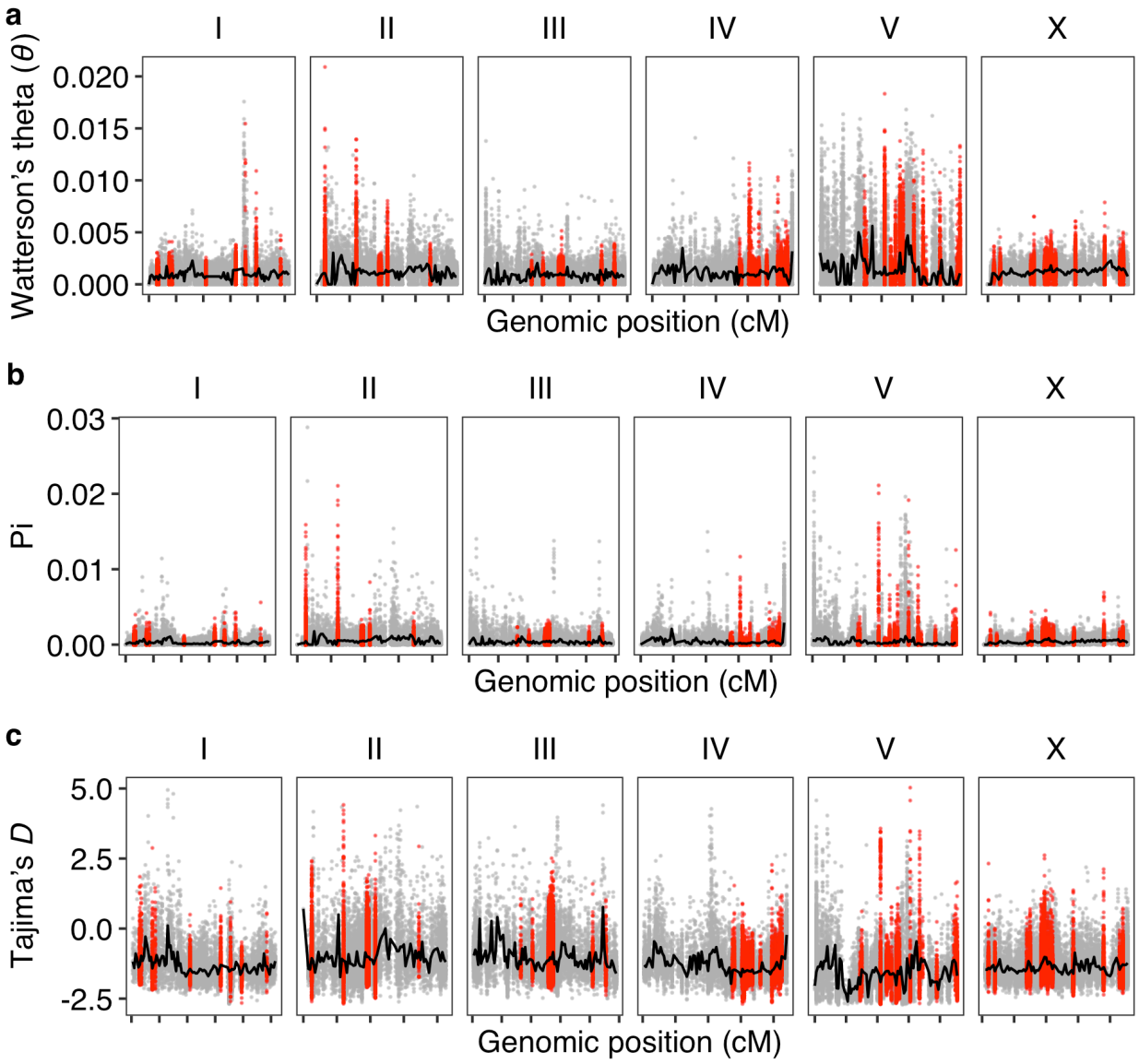
38 **Supplementary Fig. 2**

39 Gene set enrichment analysis for genes with transcript level eQTL. Enriched gene
40 classes of broad and specific categories (Category 1 to 3)[39] are shown on the y axis.
41 Bonferroni FDR corrected significance values using Fisher Exact Test for gene set
42 enrichment analysis are shown on the x axis. The sizes of the circles correspond to the
43 input gene counts of the annotation and the colors of the circles correspond to the gene
44 ratio of input gene counts to total gene counts of the annotation.



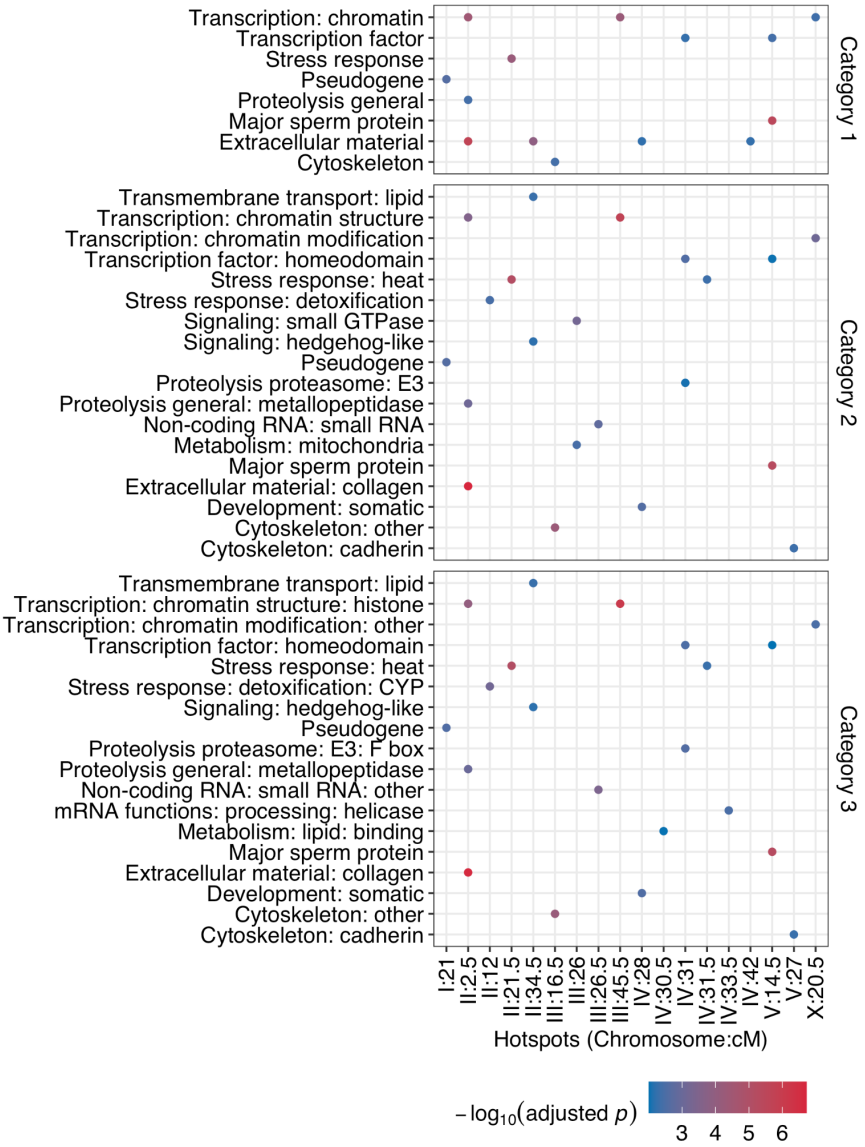
Supplementary Fig. 3

A histogram showing the distribution of linkage disequilibrium (LD) values (x-axis) among QTL of multiple eQTL of transcript expression traits. A total of 861 traits were found with multiple eQTL. LD of eQTL from the same chromosome and different chromosomes are colored salmon and blue, respectively.



Supplementary Fig. 4

Genome-wide pattern of (a) Watterson's theta (θ), (b) nucleotide diversity (pi), and (c) Tajima's *D*. Each point represents the value (y-axis) for a non-overlapping 1 kb genomic window and is plotted against the genome position (x-axis) with tick marks denoting every 10 cM. Points for genomic windows in distant eQTL hotspots are colored red. Other points are colored gray. Median values of each statistic in each 0.5 cM bin were colored black. Tajima's D values that suggest purifying selection are outliers for most values within a hotspot.

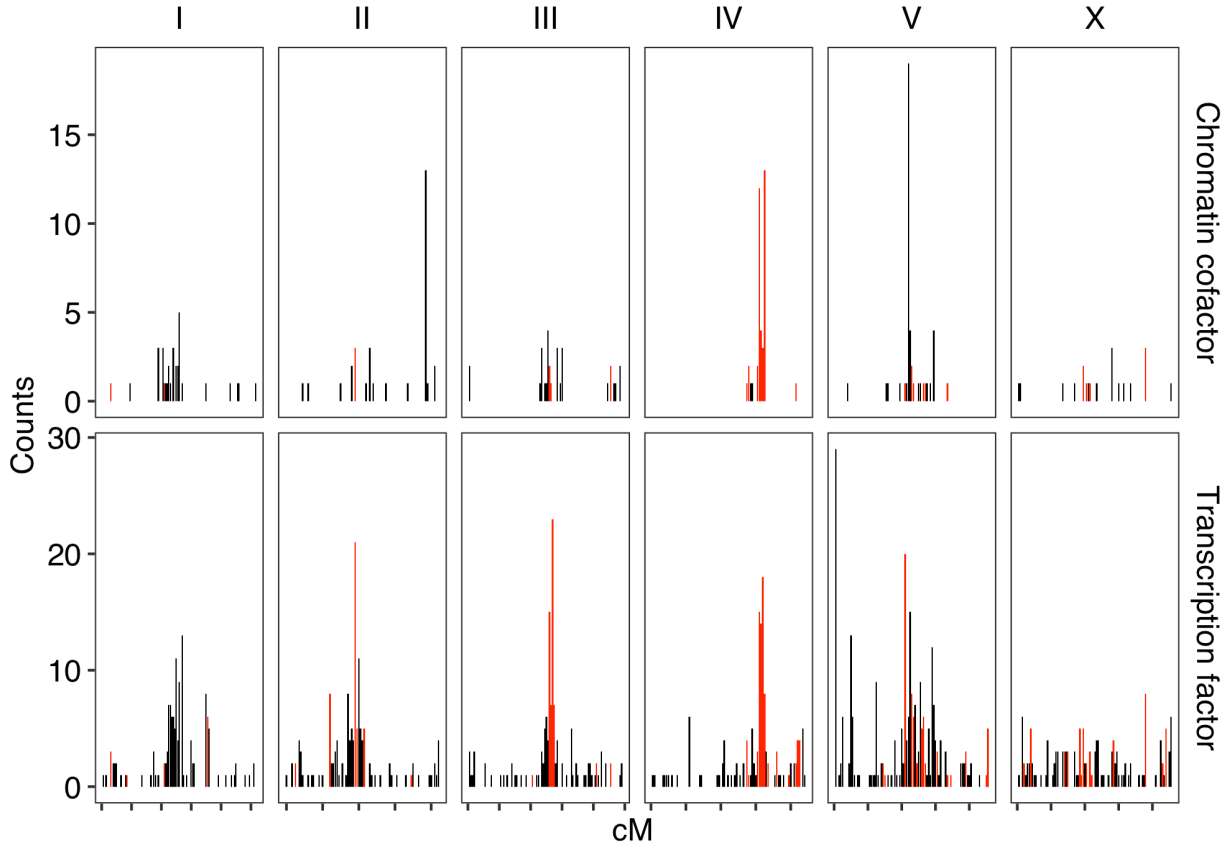


62

63

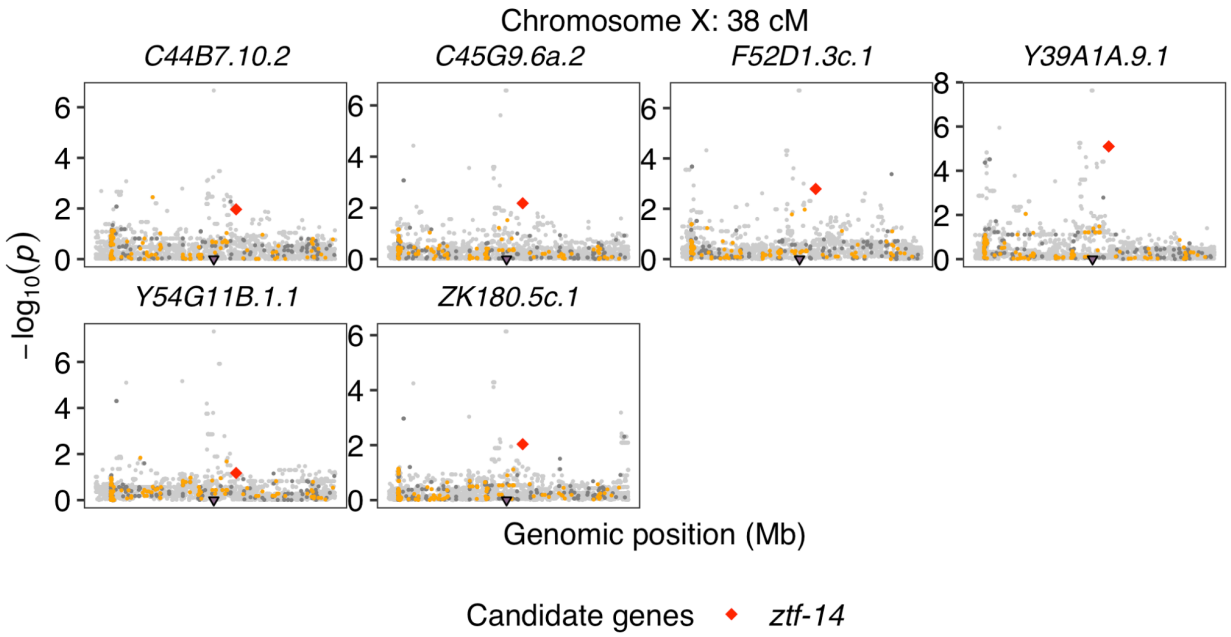
64 **Supplementary Fig. 5**

65 Gene set enrichment analysis for genes with transcript level distant eQTL in each
66 hotspot. Broad and specific categories of enriched gene (Category 1 to 3)[39] are shown
67 on the y axis. Distant eQTL hotspots with significant gene set enrichment are shown on
68 the x axis. The colors of the circles correspond to Bonferroni FDR corrected significance
69 values using Fisher Exact Test.



Supplementary Fig. 6

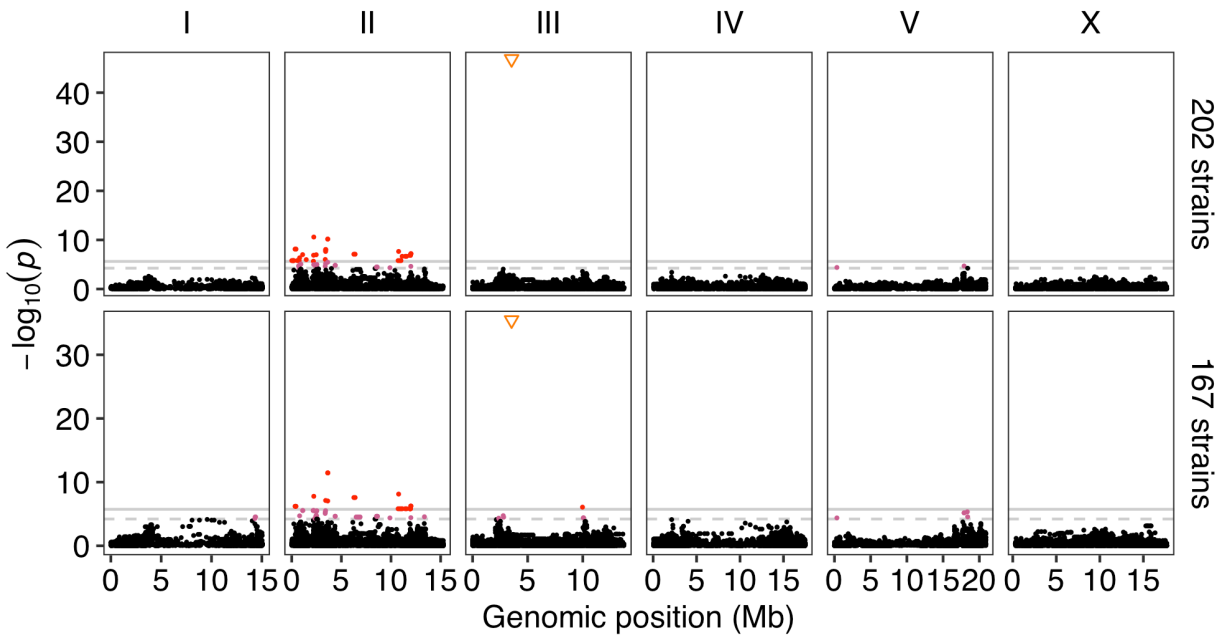
Number of genes encoding chromatin cofactors and transcription factors in each 0.5 cM bin of the *C. elegans* genome. Bins that were identified as distant eQTL hotspots are colored red. Other bins are colored black.



132

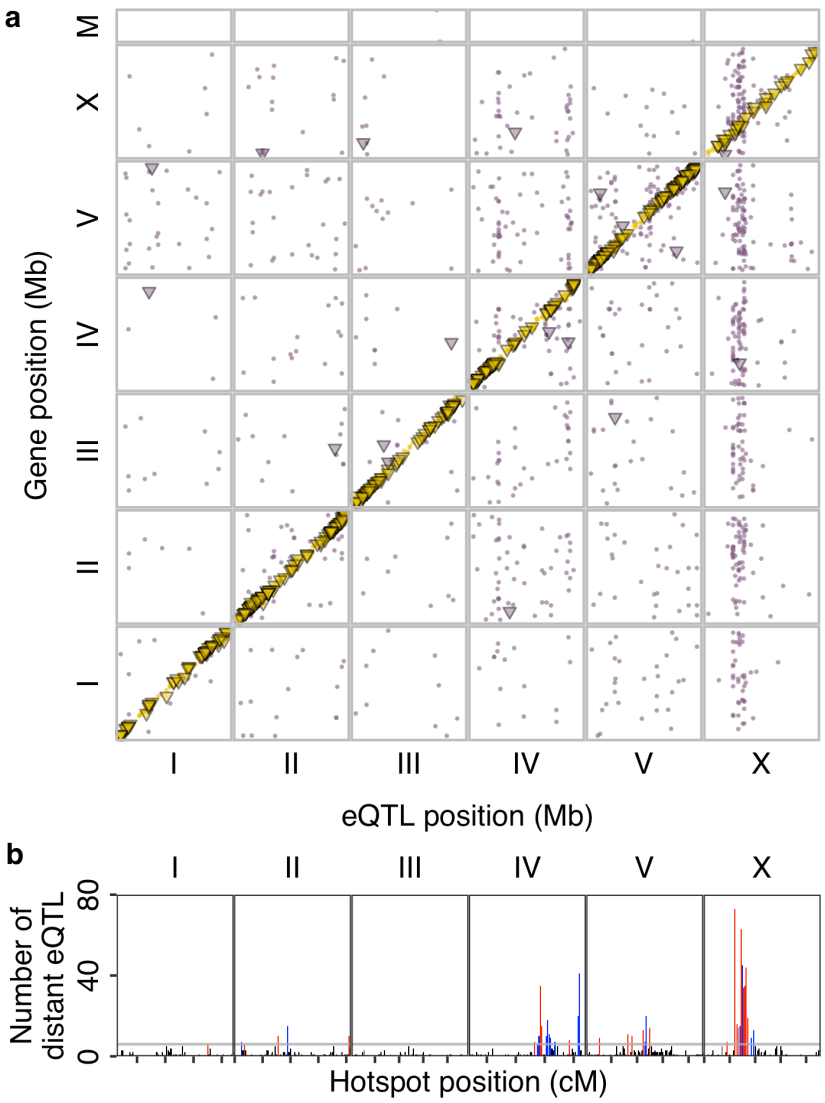
133 **Supplementary Fig. 7**

134 Fine mapping of transcript expression traits with distant eQTL in different hotspots is
135 shown. Genomic position (x-axis) is plotted against the -log10(*p*) values (y-axis) for each
136 variant. Purple triangles on the x-axis represent eQTL positions. Candidate variants with
137 negative BLOSUM scores in genes encoding transcription factors or chromatin cofactors
138 are indicated as red diamonds. Other variants that are with negative BLOSUM scores,
139 with non-negative BLOSUM scores or intergenic are colored orange, dark gray, and light
140 gray, respectively. Transcript names of each trait are indicated above each panel.



Supplementary Fig. 8

Manhattan plots indicating the GWA mapping result for animal length (q90.TOF) of 202 (top panel) and 167 (bottom panel) *C. elegans* wild strains in response to ABZ[44] are shown. Each point represents an SNV that is plotted with its genomic position (x-axis) against its -$\log_{10}(p)$ value (y-axis) from the GWA mapping. Real SNVs that pass the genome-wide EIGEN threshold (the dotted gray horizontal line) and the genome-wide Bonferroni threshold (the solid gray horizontal line) are colored pink and red, respectively. The pseudo SNV marker representing high allelic heterogeneity in the gene *ben-1* at position 3,539,640 on chromosome III is indicated as an orange inverted triangle.



Supplementary Fig. 9

RIAILs eQTL.

a, The genomic locations of 2,387 eQTL peaks (x-axis) in the RIAILs eQTL studies[3,9] are plotted against the genomic locations of the 2,003 genes with expression differences (y-axis). Golden points or triangles on the diagonal of the map represent local eQTL. Purple points or triangles correspond to distant eQTL. Triangles represent eQTL that were also found in our study. **b**, The number of distant eQTL (y-axis) in each 0.5 cM bin across the genome (x-axis) is shown. Tick marks on the x-axis denote every 10 cM. The horizontal gray line indicates the threshold of 6 eQTL. Bins with 6 or more eQTL were identified as hotspots and are colored red or blue. Bins with fewer than 6 eQTL are colored black. Blue bins represent hotspots that were also found in our study.