Title

Whole-organism eQTL mapping at cellular resolution with single-cell sequencing

Authors

Eyal Ben-David^{1,2*}, James Boockock¹, Longhua Guo¹, Stefan Zdravljic¹, Joshua S. Bloom^{1*},
Leonid Kruglyak^{1*}

(*) To whom correspondence should be addressed: eyal.bendavid@mail.huji.ac.il,
jbloom@mednet.ucla.edu, lkruglyak@mednet.ucla.edu

Affiliations

1. Department of Human Genetics, Department of Biological Chemistry, and Howard Hughes
Medical Institute, University of California, Los Angeles, CA 90095, USA.

2. Department of Biochemistry and Molecular Biology, Institute for Medical Research Israel-
Canada, The Hebrew University School of Medicine, Jerusalem, Israel

Abstract

Genetic regulation of gene expression underlies variation in disease risk and other complex traits. The effect of expression quantitative trait loci (eQTLs) varies across cell types; however, the complexity of mammalian tissues makes studying cell-type eQTLs highly challenging. We developed a novel approach in the model nematode *Caenorhabditis elegans* that uses single cell RNA sequencing to map eQTLs at cellular resolution in a single one-pot experiment. We mapped eQTLs across cell types in an extremely large population of genetically distinct *C. elegans* individuals. We found cell-type-specific *trans*-eQTL hotspots that affect the expression of core pathways in the relevant cell types. Finally, we found single-cell-specific eQTL effects in the nervous system, including an eQTL with opposite effects in two individual neurons. Our results show that eQTL effects can be specific down to the level of single cells.

32 Introduction

33

34 Gene expression differences have a strong genetic basis, and genome-wide studies have
35 identified thousands of regions affecting gene expression, termed expression quantitative trait
36 loci (eQTLs) ¹. eQTLs have been proposed to underlie genetic associations with complex traits
37 and diseases ^{2,3}, and genome-wide eQTL mapping holds great promise for uncovering the
38 molecular underpinnings of phenotypic variation.

39 Studies in purified blood cell populations ⁴⁻⁶ and computational analyses in human tissues
40 ^{7,8} indicate that many eQTLs are cell-type specific. Studying eQTLs in relevant tissues and cell
41 types is thus important for understanding their role in trait variation ⁹. However, major challenges
42 remain for cell-type-specific eQTL mapping. First, studying multiple tissues or cell types is
43 laborious and expensive, and has mostly been limited to large consortia ¹⁰. Additionally, diseases
44 can be associated with specific subsets of cells that may not be readily separable from the rest
45 of the tissue, limiting the insight that can be gleaned from studying bulk tissue samples ¹¹.

46 The recently developed technology of single-cell RNA sequencing (scRNA-seq) has the
47 potential to significantly mitigate both of the above issues. In scRNA-seq, gene expression is
48 profiled at the level of individual cells, and different cell populations are identified based on their
49 gene expression profiles. This approach allows simultaneous measurement of expression in
50 multiple tissues and cell types in a “one-pot” experiment, reducing the number of samples that
51 need to be processed and profiled separately. Furthermore, distinct cell types can be
52 interrogated directly, facilitating eQTL mapping in disease-relevant but rare cell types.

53 The size and complexity of mammalian tissues has so far limited the ability to use scRNA-
54 seq for eQTL mapping, and studies have focused on purified cell populations and cell lines ^{12,13}.

To pilot a whole-organism approach, we therefore turned to the model nematode *Caenorhabditis elegans*. One of the mainstays of modern genetic research, *C. elegans* has an invariant cell lineage that leads to each individual having the same number and identity of cells¹⁴. Decades of research have uncovered the functions of many of those cells, as well as expression markers that uniquely identify them^{15,16}. These features and resources make *C. elegans* exceptionally well-suited for using scRNA-seq to map eQTLs across cell types in the natural physiological context of a whole animal, down to cellular resolution.

Results

Cell-type-specific eQTL mapping in a single one-pot experiment

Genome-wide eQTL mapping involves acquiring genotypes and gene expression profiles for a genetically diverse cohort. We recently developed a method, *C. elegans* extreme quantitative trait locus (ceX-QTL) mapping, for genetic analysis of complex traits in extremely large populations of segregants¹⁷. The method takes advantage of a mutation in the gene *fog-2* that forces the normally hermaphroditic *C. elegans* to reproduce via obligate outcrossing, allowing us to propagate a large crossing experiment for multiple generations.

Here we build on ceX-QTL by combining it with single-cell RNA sequencing (scRNA-seq) to carry out eQTL mapping at cellular resolution in a single one-pot experiment (Fig. 1A). In this approach, a large heterogeneous pool of cells from thousands of genetically distinct individuals is profiled using scRNA-seq, cell types are inferred by clustering scRNA-seq profiles and studying known cell-type markers, and genotype information is reconstructed using expressed genetic variants, enabling eQTL mapping in multiple cell types simultaneously.

We propagated a cross between the laboratory strain N2 and a highly divergent isolate from Hawaii, CB4856, for four generations, generating a pool of 200,000 genetically distinct F4 segregants. We dissociated the segregant pool to single cells at the L2 larval stage and profiled the cells with scRNA-seq. We identified clusters in a Uniform Manifold Approximation and Projection (UMAP) of the dataset^{18–20}, and determined their cell-type identities using known markers^{21,22}. Our final dataset comprises 55,508 cells classified into 19 different cell types (Fig. 1B) (Table S1). The observed number of cells of each type was strongly correlated with the known cell-type abundance in L2 larvae (Spearman's $\rho = 0.87$, $p = 2.2 \times 10^{-6}$, Supplementary Fig. 1).

Most of the cells in our sample were expected to carry unique genotypes (Methods). This design is advantageous for eQTL mapping because it maximizes the sample size²³, but requires *de novo* genotype calling, since the genotype of each cell is unknown beforehand. Rather than assign deterministic genotype calls based on sparse scRNA-seq data, we derived genotype probabilities for each cell using a Hidden Markov Model (HMM) (Supplementary Fig. 2). We then performed eQTL mapping with these genotype probabilities in a negative binomial modeling framework (Methods).

eQTL mapping in multiple cell types

We mapped 1,718 *cis* eQTLs in 1,294 genes, and 451 *trans* eQTLs in 390 genes, at a false discovery rate (FDR) of 10% across the different cell types (Fig. 2A-B, Table S2). The number of eQTLs detected in each cell type was strongly correlated with the number of cells of that type (Spearman's $\rho = 0.91$, $p < 2.2 \times 10^{-16}$). In cell types with >1000 cells, we mapped between 52 and 415 eQTLs (Table S1). For 1,071 of the 1,294 genes with a *cis* eQTL (83%), the eQTL was

detected in only one cell type. For 208 of the remaining 223 genes (93%), the direction of the eQTL effect was the same in all cell types in which it was detected.

We studied to what degree our *cis* eQTL results were concordant with gene expression differences between the parents. We generated a scRNA-seq dataset from 6,721 N2 and 3,104 CB4856 cells, and used a classifier trained on the segregant dataset to identify cell types in the parental scRNA-seq dataset. We then carried out a differential expression analysis in each cell type. We found 870 differentially expressed genes (at a >2 fold change and FDR of 10%), of which 201 (23%) had a *cis* eQTL in the same tissue (OR = 18.8, $p < 2.2 \times 10^{-16}$, Fisher's Exact Test). 191 of these *cis* eQTL (95%) showed the same direction of effect as the parental difference. Further, the effect sizes of the significant *cis* eQTLs were strongly correlated with the sizes of the parental differences (Spearman's $\rho = 0.66$, $p < 2.2 \times 10^{-16}$) (Supplementary Fig. 3-4). These results provide independent support for our *cis*-eQTL mapping, and show that for a sizable fraction of the genes, those *cis* eQTLs are a major cause of differential gene expression between the strains.

Comparison between bulk and single-cell *cis*-eQTL mapping

To investigate the relationship between single-cell and bulk eQTL mapping, we compared our single-cell eQTLs to those previously identified in a panel of 200 recombinant inbred lines (RILs) generated from crossing N2 and CB4856²⁴. In the bulk study, a large population of whole worms from each RIL was recovered at a late larval stage, L4, and profiled on expression microarrays. We reanalyzed data for 11,535 genes expressed in both datasets and identified 981 *cis* eQTLs in the bulk dataset (at an FDR cutoff of 10%). Despite major differences in experimental design, including the developmental stage of the worms, the overlap with the single-cell *cis* eQTLs was

highly significant, with 335 *cis* eQTLs shared between the studies (Odds Ratio = 7.2, $p < 2.2 \times 10^{-16}$, Fisher's exact test) (Fig. 2C). These shared loci represented 34% of the bulk *cis* eQTLs and 32% of the single-cell *cis* eQTLs. Furthermore, the bulk and single-cell eQTL effect sizes were highly correlated (Spearman's $\rho = 0.64$, $p < 2.2 \times 10^{-16}$) (Fig. 2D). Lastly, single-cell eQTLs detected in multiple cell types were more likely to also be seen in the bulk study: 50% of the genes with *cis* eQTLs detected in multiple cell types were also identified in bulk, compared to 28% of the eQTLs detected in only one cell type (OR = 2.58, $p = 2.1 \times 10^{-8}$) (Fig. 2C). This observation suggests that the single-cell eQTL mapping approach improves the power to detect cell-type specific effects.

Shared and cell-type specific *trans*-eQTL hotspots

We observed that 90 of the 451 *trans* eQTLs clustered at 5 hotspots, each containing 12-31 eQTLs (Supplementary Fig. 5, Table S3). A hotspot on Chr I was identified independently in both neurons and seam cells; the top associated variant (Chr. I:10890182) was the same for both cell types (Supplementary Fig. 5A-B). The other hotspots were identified in the body wall muscle (on Chr. I) (Supplementary Fig. 5C), the intestine (on Chr. V) (Supplementary Fig 5D), and neurons (two distinct hotspots on Chr. III) (Supplementary Fig. 5B).

To test whether the target genes of these 5 hotspots are involved in coherent biological processes, we relaxed the FDR threshold to 20%, which increased the number of genes linked to each hotspot to 21-42, and performed Gene Ontology (GO) enrichment analysis (Table S4). For three of the hotspots, we found significant enrichments that were consistent with the cell-type specificity of the hotspot. The targets of the hotspot detected in intestinal cells were weakly enriched for genes involved in the innate immune response (FDR-corrected $p = 0.042$), a major

role of that tissue²⁵. The targets of the hotspot detected in the body wall muscle were enriched for genes associated with the term *myofilament* (FDR corrected $p = 6.4 \times 10^{-8}$), *actin cytoskeleton* (FDR-corrected $p = 4.2 \times 10^{-6}$), and related terms. The enrichment was driven by the genes *mup-2*, *tnt-1*, *tnt-2*, *mlc-2*, *mlc-3*, *lev-11* and *act-4*. *mlc-2* and *mlc-3* encode a myosin light chain, and *act-4* encodes an actin protein. *lev-11* encodes a tropomyosin, and *mup-2*, *tnt-1*, and *tnt-2* encode 3 of the 4 proteins in *C. elegans* that are expressed in the body-wall muscle and form troponin complexes, highly conserved regulators of muscle contraction²⁶ (Supplementary Fig. 6).

The targets of the neuronal hotspot on the right arm of Chr III were enriched for genes involved in *vesicle localization* (FDR-corrected $p = 7.5 \times 10^{-3}$), as well as for BMP receptor binding genes (FDR-corrected $p = 2.8 \times 10^{-3}$). The latter enrichment was driven by *dbl-1* and *tig-2*, orthologs of human bone morphogenetic protein (BMP) genes BMP5 and BMP8 and ligands of the transforming growth factor beta (TGF- β) pathway²⁷. Notably, *dbl-1* was discovered as a gene that regulates body size in *C. elegans*²⁸, the hotspot peak marker is located <300 kb from the peak of a QTL we previously identified for body size²⁹, and the corresponding confidence intervals overlap (Table S3), suggesting that differential regulation of the TGF- β pathway is involved in variation in body size between N2 and CB4856.

Cell-specific eQTL effects in the *C. elegans* nervous system

C. elegans is a premier model for studying neurobiology at the cellular level, which is aided by its invariant cell lineage and the diverse functions associated with specific individual neurons. Importantly, many of the neurons are highly variable in their gene expression, and express specific gene markers¹⁶. To identify specific subtypes of neuronal cells, we separately clustered

the 12,467 cells identified as neurons and compared the clusters to previous *C. elegans* scRNA-seq datasets, including the recently published *C. elegans* Neuronal Gene Expression Map & Network (CeNGEN)^{21,22,30,31} (Table S5). The neurons fell into 81 distinct clusters, ranging from 17 to 872 cells. We mapped these clusters onto 100 (83%) of the 120 neuronal clusters identified in CeNGEN (Supplementary Fig. 7). We also identified CEM neurons, which are male specific and absent from CeNGEN, based on the expression of the marker *cwp-1*³².

We mapped *cis* eQTLs in each of the single neuronal subtypes (sn-eQTLs) and identified a total of 163 sn-eQTLs in 132 genes at an FDR of 10% (Fig. 3A, Table S6). Of these, 117 (88%) were identified in only a single neuronal subtype. Functional annotation of sn-eQTLs identified 25 genes involved in signaling (FDR-corrected $p = 0.047$), including 12 genes involved in G-protein Coupled Receptor (GPCR) signaling (FDR-corrected $p = 0.047$) and 8 genes involved in neuropeptide signaling (FDR-corrected $p = 9.9 \times 10^{-3}$).

We compared the sn-eQTLs to those identified when all neurons were analyzed jointly (“pan-neuronal mapping”), and found that a sizable fraction of the sn-eQTLs did not have evidence for a pan-neuronal signal: 92 were not identified pan-neuronally at an FDR of 10%, and 69 were not identified even at a highly permissive FDR of 50%, suggesting that they exert their effects only in specific neuronal subtypes (Fig. 3A). Regardless of statistical significance, pan-neuronal eQTLs should have consistent effect directions across neuronal subtypes, while subtype-specific eQTLs should not. We therefore compared the direction of effect of each sn-eQTL in the subtype in which it was detected with its direction of effect in the set of all neurons excluding that subtype. Among the 69 sn-eQTLs with no signal in the pan-neuronal mapping even at the permissive FDR, the direction of the effect was concordant for 33 and discordant for 36, not significantly different from chance ($p > 0.5$; binomial test), as would be expected if these

effects are truly subtype-specific (Fig. 3B). In contrast, among the 94 that had a pan-neuronal signal at an FDR of 50%, the direction of the effect was concordant for 88 and discordant for only 6 ($p < 0.000001$; binomial test), consistent with differences in detection arising from limited statistical power.

In a striking case, we observed an sn-eQTL in the neuropeptide gene *nlp-21* that showed significant and opposing effects in two neurons (Fig. 3C-D). In the RIC neuron, higher *nlp-21* expression was associated with the CB4856 allele ($\beta = 4.4$, FDR-corrected $p = 0.03$), while in the RIM neuron, higher *nlp-21* expression was associated with the N2 allele ($\beta = -5.4$, FDR-corrected $p = 9.8 \times 10^{-7}$). In the pan-neuronal mapping, no significant effect is observed for this gene. We identified the RIC and RIM neurons in the parental dataset, and although the small number of cells in each group (35 and 27, respectively, with only 9 and 5 of them from CB4856) was insufficient for statistical testing, the directions of the differences agreed with the eQTL effects (Fig. 3E). These results provide direct evidence that eQTLs can be specific down to the cellular level.

Discussion

We used scRNA-seq to map eQTLs in *C. elegans* across cell types in a single one-pot experiment. Earlier scRNA-seq eQTL mapping studies were limited in sample size to at most ~100 individuals, but nevertheless highlighted the potential of this approach to identify cell-type¹³ and developmental¹² eQTLs, as well as loci affecting expression variance³³. Our novel approach allowed us to map eQTLs in tens of thousands of genotypically distinct individuals and enabled detection of both *cis* and *trans* eQTLs, as well as resolution of their effects down to the level of specific cells.

One of the major factors affecting gene expression studies is variation resulting from uncontrolled environmental differences between individuals that are grown or processed separately. By using scRNA-seq, we were able to process all individuals jointly. After the initial parental cross, all subsequent steps carried out over the course of five *C. elegans* life-cycles (three weeks) were performed in bulk, limiting any confounding environmental factors. To minimize the influence of genotype on development, we synchronized the worms at the first larval stage, L1, and collected samples at the L2 stage, limiting the time for differences to accumulate post-synchronization. Even careful synchronization is not expected to completely remove the effects of genetic variation on developmental timing, and such variation can be combined with gene expression time course data collected during development to increase the power of eQTL mapping and to study the developmental dynamics of eQTLs³⁴. This raises the possibility that future scRNA-seq studies of *C. elegans* across developmental stages would open the door to a similar analysis of our single-cell eQTL dataset.

Previous work suggested the existence of cell-type specific eQTL hotspots in *C. elegans* based on the expression patterns of hotspot targets³⁴. We discovered three hotspots that are cell-type specific, with targets that are involved in core functions performed by these cell types. Recently, eQTL hotspots have been identified in human blood cells^{35,36}, as well as in cell lines³⁷. These results suggest hotspot and *trans* eQTL discovery is facilitated by expression studies that can distinguish cell types and point to a larger role of hotspots in the genetics of gene expression in animals.

A comparison of the single-cell *cis* eQTLs to those mapped in a previous whole-worm eQTL study from our laboratory showed a highly significant overlap despite major differences in experimental design. These results join accumulating evidence that a sizable fraction of *cis*

eQTLs have robust, consistent effects^{10,38}, and show that many of the effects are conserved across worm development. The strong overlap of *cis* eQTLs mapped by scRNA-seq and by whole-worm analysis also suggests that the effect of many *cis* eQTLs is conserved across cell types. By generating an extremely large number of unique segregants, our method enables scaling up the number of studied individuals simply by sequencing a larger number of cells. Thus, the increasing throughput of single-cell technologies and sequencing platforms will enable future work to study cell-type specificity of *cis* and *trans* eQTLs in greater detail.

Lastly, we discovered *cis* eQTLs that act in single subtypes of *C. elegans* neurons, including many that were not found when all neurons were analyzed jointly. Importantly, we also discovered an eQTL that influences expression of the gene *nlp-21* in opposing directions in two different neurons. Such antagonism could result from context-dependent effects of the same variant, or from different variants acting in each cell. Our study joins a growing body of work that identified eQTLs that are specific to cell types^{5–8,13,39,40}, as well as developmental^{12,34} and environmental^{41–43} contexts. Distinct genetic effects are found across cell types and conditions, and large-scale study of such effects is therefore crucial for gaining a comprehensive understanding of regulatory variation.

256 Figures

257

258 **Figure 1. Whole-organism eQTL mapping with single-cell RNA-sequencing (scRNA-seq).** (A) A

259 large population of segregants is dissociated to single cells. Each cell in the suspension has an

260 unknown genotype and cell-type identity. The suspension is profiled using scRNA-seq. Cell-type

261 identity is inferred by clustering cells and comparing the expression of known marker genes.

262 Genotypes are reconstructed from expressed single-nucleotide variants (SNVs). (B) The UMAP

263 projection of 55,508 scRNA-seq expression profiles from approximately 200,000 *C. elegans* F4

264 segregants collected at the L2 larval stage is shown. Cells are colored based on the inferred cell type.

265

266

267

268

269 **Figure 2. eQTL mapping in cell types.** (A) A genome-wide map of eQTLs across all cell types is
 270 shown. The position of the eQTLs is shown on the x-axis, while the y-axis shows the position of the
 271 associated transcripts. Points along the diagonal are *cis* eQTLs (those mapping to nearby genes). (B)
 272 The number of *cis* and *trans* eQTLs mapped in each cell type. (C) The overlap between a previous
 273 study that mapped eQTLs in whole worms in a panel of recombinant inbred lines (RIL) and our dataset.

274 (Left) The proportion of genes with a *cis* eQTL in at least one dataset, out of all genes tested. (Middle)
 275 Of the 1,688 significant *cis*-eQTL genes, 355 had a *cis* eQTL in both datasets, representing a highly
 276 significant enrichment. (Right) Hits mapped in more than one cell type were more likely to also be
 277 found in the whole-worm (“bulk”) dataset. (D) Quantitative comparison between normalized effect
 278 sizes in our dataset and in the whole-worm dataset.
 279

Figure 3. Neuron-specific eQTL mapping. (A) *cis*-eQTLs mapped in single neuronal subtypes (sn-eQTLs) are shown. The top three rows indicate whether the eQTL was mapped pan-neuronally at a 10% FDR threshold (row 1), at a 50% FDR threshold (row 2), and whether the sign of the effect estimate (“effect direction”) was the same in the pan-neuronal and single cell mapping (row 3). (B) Comparing the effect direction between the sn-eQTL mapping and mapping in a set of neurons excluding the sn-eQTL neuron shows evidence for subtype-specific effects. The number of genes showing the same (“purple”) or opposite (“turquoise”) effect directions is shown for genes with pan-neuronal FDR > 50% (top) and < 50% (bottom). (C-D) An eQTL with antagonistic effects in two

neurons. Higher expression of the gene *nlp-21* in the RIM neuron is associated with the N2 allele (C), while higher expression in the RIC neuron is associated with the CB4856 allele (D). In C and D, a linear fit is shown for illustration. All p values are FDR-corrected. Read counts were normalized to the number of UMIs in each cell and log-transformed. (E) Expression of *nlp-21* in the parental dataset. The direction of effect is concordant between the left panel and (C) (RIM neuron) and between the middle panel and (D) (RIC neuron). Horizontal lines are averages.

297 **Methods**

298 ***C. elegans* culturing**

299 *C. elegans* strains were cultured at 20°C using standard conditions with the exception that the
 300 agar in the nematode growth media (NGM) was replaced with a 4:6 mixture of agarose and agar
 301 (NGM+agarose), to prevent burrowing of the CB4856 strain. Parental strains used were QX2314
 302 (N2 *fog-2*(q71) V; *hsp-90p*::GFP II) and PTM299 (CB4856 *fog-2*(kah89)). Large segregant
 303 panels were generated as before ¹⁷. Briefly, 500 L4 males from PTM299 and 500 L4
 304 hermaphrodites from QX2314 were seeded on a plate for 30 hours, and gravid worms and eggs
 305 were collected and bleached. Eggs were synchronized to L1 larvae for 24h, and seeded on 10cm
 306 NGM+agarose plates. In each generation, gravid worms were bleached, their progeny
 307 synchronized for 24h and seeded. The entire process was repeated up to F4, with 3-4 days per
 308 generation.

310 **Cell extraction and sequencing**

311 192,000 F4 were seeded on four 10cm NGM+agarose plates seeded with OP50. L2 were
 312 recovered after 24 hours, and staging was validated under a stereomicroscope. L2 cell
 313 dissociation was carried out as previously described ⁴⁴, implementing modifications from a later
 314 study ⁴⁵, as well as our own. Worms were recovered off the plates and washed three times in
 315 M9. Lysis was then done with an SDS-DTT solution (200 mM DTT, 0.25% SDS, 20 mM HEPES,
 316 pH 8.0, 3% sucrose) in a hula mixer set on low speed to prevent worms from settling. The lysate
 317 was observed under the stereoscope every two minutes, and lysis was stopped when a blunted
 318 head shape appeared in the majority of worms ⁴⁵, after ~4 minutes. Worms were then washed
 319 quickly three times in 1ml of M9, and two additional times in 1ml of egg buffer (118 mM NaCl,

48 mM KCl, 2 mM CaCl₂, 2 mM MgCl₂, 25 mM HEPES, pH 7.3, osmolarity adjusted to 340 mOsm with sucrose). Worms were then resuspended in 0.5ml of 20mg/ml Pronase E that was freshly prepared in L15 media supplanted with 2% fetal bovine serum (L15-FBS) and adjusted to 340 mOsm with sucrose. Worm dissociation was done by continuous pipetting on the side of the tube, and monitored every 2-3 minutes on a microscope equipped with a 40x phase contrast objective lens. Dissociation was stopped when few intact worms remained and a high density of cells was visible. 0.5ml of L15-FBS was added to stop the reaction, and the lysate was spun for 6 minutes at 500g at 4°C. The cell pellet was resuspended in PBS (PBS was adjusted to 340 mOsm with sucrose). Cell suspension was spun for 1 minute in 100g at 4°C to remove remaining undigested worms, counted and diluted to 1M cells/ml in osmolarity adjusted PBS, and loaded directly onto 5 lanes of 3' Chromium single-cell RNA-sequencing flow cells (10x Genomics), targeting 10,000 cells on each lane. Library prep was carried out according to manufacturer's protocol. Prepared libraries were sequenced together on an S4 lane of Novaseq 6000. A paired-end 2x150 run was done, to maximize the recovery of single-nucleotide variants. In all downstream processing, each of the five 3' Chromium lanes processed concurrently was treated as a separate "batch", and lane identity corresponds to the "batch" identity for the rest of the methods.

Single-cell RNA-sequencing data processing

Raw sequencing reads were analyzed using *Cell Ranger* (Ver 3.0.2). We used a gene transfer format (GTF) file that was corrected for misannotation of 3' untranslated regions (3'UTR) that was generated in a previous study²². *C. elegans* cell-types differ widely in the number of UMIs that are recovered using scRNA-seq. Therefore, a simple UMI cutoff, as is commonly used, may

be biased for cell-types with more UMIs. We therefore implemented an iterative pipeline to recover clusters of *bona fide* cells and remove cell doublets as well as degraded cells. We took 20,000 cells with the most UMIs in each cluster (twice the targeted number of cells, 100,000 overall), and processed them in *Monocle* (Ver 3) ⁴⁶. Default parameters were used, with the exception that 100 dimensions were used for reduction, and batch was added as a covariate. Leiden clustering identified a total of 154 clusters, and we used the *top_markers* function in *Monocle* to identify the genes upregulated in each. We then removed clusters whose top genes included any ribosomal genes or the mitochondrial genes *ndfl-4*, *nduo-6*, *atp-6*, *ctc-2*, *ctc-3*, *ctc-1*, which we noticed were usually found together as the most upregulated genes in clusters that did not specifically express any known markers for *C. elegans* cell-types. This removed a total of 30,980 cells (31%). For the remaining 69,020 cells, raw counts were processed using the R package *SoupX* to reduce ambient RNA contamination ⁴⁷. We then normalized, reduced dimensions and clustered the background corrected cell profiles in *Monocle* using the same parameters as above.

To annotate cell-types, we used the markers described in a previous study (Table S12 in Ref ²²) that reanalyzed a previous L2 single-cell dataset ²¹. Our cell-type annotation corresponds to the “UMAP” column in that table, with the following exceptions: (a) we separated hypodermis from seam cells, somatic gonad from sex muscle cells, and glia from excretory cells, since those groups were not clustering together in our data. (b) cells identified as “Miscellaneous” in that table were annotated as individual cell-type identifications in our data, with the exception of the sphincter and anal muscles which were not differentiated from each other in our data. Finally, we re-evaluated our cell type identifications, and filtered cell doublets as well as dead cell or debris that may still contaminate *bona fide* cell-type clusters. We trained a classifier using our

manually curated cell-type classifications with a L2-penalized multinomial logistic regression framework, as implemented in the *Scikit-learn* Python package (v0.22)⁴⁸. We read the raw gene expression matrices into python using scanpy (v1.4.2). We removed 2,582 genes that were expressed in less than 10 cells. The gene expression levels of each cell were corrected so that the total gene expression counts added up to 10,000. Per gene, these corrected counts were normalized using a $\log(1+x)$ transformation. To speed up the computation of the multinomial logistic regression, we only used the 2,037 genes with a mean expression between 0.0125 and 3, and a minimum dispersion of 0.5. We scaled the gene expression matrix so that the expression level of each gene across cells had a mean of 0 and a variance of 1, after scaling expression values over 20 were set to 20. We fit a multinomial logistic regression model using the scaled gene expression values for the 2,037 highly variable genes from the complete set of 69,020 cells to obtain an estimate for the inverse regularization strength (C). Using the estimated C of 7.74×10^{-04} , we performed 5 fold cross validation to estimate the probability that each cell belongs to one of the manually curated cell-type classifications. Any cell with a probability higher than 0.2 of belonging to 2 or more cell types (9,198) was classified as a doublet. Any cell which did not belong to a cell-type with probability ≥ 0.4 (5,547) was classified as low quality. In total, we removed 11,398 cells that were classified as a doublet or low quality. We removed an additional 2,114 cells classified as Neurons as described in the section “Neuronal cell-type classification”. For the remaining final list of 55,508 cells, we used the output of the classifier as the final cell-type classification. The final classification is shown in Figure 1. For display purposes the plot in Figure 1 was generated by rerunning umap on the finalized dataset with *euclidean* distance metric and *umap.min_dist*=0.5, resulting in a more compressed visualization of the dataset.

Estimating the number of unique genotypes

We note that calculating the number of expected unique genotypes is akin to the well-known “Birthday problem” in statistics. Given C cells sampled from I individuals, the expected number of cells with a unique genotype is $C(1-1/I)^{C-1}$. Assuming 50%-90% of worms were successfully dissociated (a conservative range), we expect 31,134-40,257 unique genotypes.

Single-nucleotide variant counting

We used a list of single-nucleotide variants (SNVs) we previously curated for CB4856 compared to the N2 reference¹⁷. We derived genotype informative UMI counts for N2 and CB4856 variants using *Vartrix version 1.0* (<https://github.com/10XGenomics/vartrix>) directly on the output of *Cell Ranger*. To reduce SNV counts that result from SNVs in the ambient RNA background, we only kept SNVs that resided in genes with positive counts in the background corrected matrix.

Genotype inference using a hidden Markov model

We set up a hidden Markov model (HMM) to infer the genotypes of the recombinant progeny^{49,50}. The HMM is used to calculate the probability of underlying genotypes for each individual and requires three components: (1) prior probabilities for each of the possible genotypes, (2) emission probabilities for observing variant informative reads given each of the possible genotypes, (3) and transition probabilities - the probabilities of recombination occurring between adjacent genotype informative sites.

For the autosomal chromosomes we defined prior genotype probabilities as 0.25 for homozygote N2, 0.5 for heterozygote CB4856/N2, and 0.25 for homozygote CB4856. For the

sex chromosome we defined prior genotype probabilities as 0.44 for homozygotes N2, 0.44 for the heterozygote CB4856/N2, and 0.11 for homozygote CB4856. These values were chosen because to generate the segregant population, N2 hermaphrodites were crossed to CB4856 males, and thus contributed twice as many X chromosomes to the progeny as CB4856.

Emission probabilities were calculated as previously described for low coverage sequencing data^{51,52} under the assumption that the observed counts of reads for both possible variants (Y) at a genotype informative site (g) arise from a random binomial sampling of the alleles present at that site and that sequencing errors (e) occur independently between reads at a rate of 0.002:

$$p(Y|g = NN) = \binom{D}{r}(1-e)^r(1-(1-e))^{D-r}$$

$$p(Y|g = NC) = \binom{D}{r}\frac{1}{2}^D$$

$$p(Y|g = CC) = \binom{D}{r}(e)^r(1-e)^{D-r}$$

Where (D) is the total read depth at a genotype informative site for a given individual, (r) is the total read depth for the N2 variant at that site, and N represents the N2 variant and C represents the CB4856 variant.

Transition probabilities were derived from an existing N2 x CB4856 genetic map⁵³. We linearly interpolated genetic map distances from the existing map to all genotype informative sites in our cross progeny. We scaled these genetic map distances, multiplying them by a factor of 0.4, to account for the fact that the previous genetic map was built using ten generations of intercrossing whereas progeny from our cross are derived from four generations of intercrossing⁵⁴.

For QTL mapping we used an additive coding, summing the probability that the genotype was homozygote CB4856 with one half the probability that the genotype was heterozygote N2/CB4856.

eQTL mapping

Genotype probabilities were standardized, and markers in very high LD ($r > .9999$) were pruned. This LD pruning is approximately equivalent to using markers spaced 5 centimorgans (cm) apart. For each transcript, we counted the number of cells for which at least one UMI count was detected in each cell type. Transcripts with non-zero counts in at least 20 cells in a cell type were considered expressed in that cell type and used for downstream analyses.

As has been previously described for droplet scRNA-seq, counts of UMIs can be adequately parameterized by a gamma-poisson distribution, which is also known as the negative binomial distribution⁵⁵. Thus we used a negative binomial regression framework for eQTL mapping here. We also note that simpler approaches using $\log(\text{counts}+1)$ with ordinary least squares behave pathologically, especially in regard to behavior with multiple partially correlated covariates, and simulations (not shown) showed such models lead to inflated false positive rates.

For each expressed transcript in each cell type we first fit the negative binomial generalized linear model:

$$E[Y] = \mu(1)$$

$$Var(Y) = \mu + \frac{1}{\theta} \mu^2 \quad (2)$$

$$\mu = \exp(\beta_i + X_t \beta_t + \mathbf{X}_b \beta_b + X_c \beta_c) \quad (3)$$

458 Which has the following log-likelihood:

459

$$\ell(\beta, \theta) = - \sum_{n=1}^N [(y_n + \theta) \log(\mu_n + \theta) - y_n \log(\mu_n) + \log(\Gamma(y_n + 1)) - \theta \log(\theta) + \log(\Gamma(\theta)) - \log(\Gamma(\theta + y_n))] \quad (4)$$

461

462 And where Y is a vector of UMI counts per cell, X_t is a vector of the log(total UMIs per cell) and
463 controls for compositional effects, X_b is an indicator matrix assigning cells to batches, and X_c is
464 the vector of standardized genotype probabilities across cells for the closest genotypic marker
465 to each transcript from the pruned marker set. In addition, β is a vector of estimated coefficients
466 from the model, μ_n is the expected value of Y for a given cell n , N is the total number of cells in
467 the given cell type, and θ is a negative binomial overdispersion parameter. Model parameters
468 were estimated using iteratively-reweighted least squares as implemented in the *negbin.reg*
469 function in the *Rfast2* R package. If the model did not converge, model parameters were
470 estimated with the *gam* function in the *mgcv* package⁵⁶, which opts for certainty of convergence
471 over speed. We note that due to the computational burden of fitting so many GLMs in the context
472 of sc-eQTL mapping, we chose to estimate θ once for each transcript in each cell type for this
473 model, and use that estimate of θ in the additional models for that transcript within the cell type,
474 as described below. This approach is conservative, as the effects of unmodeled factors (for
475 example *trans* eQTLs) will be absorbed into the estimate of overdispersion, resulting in larger
476 estimated overdispersion $\left(\frac{1}{\theta}\right)$ and lower model likelihoods. Computational approaches that re-
477 estimate θ for each model, that jointly model all additive genetic effects, or that regularize θ
478 across models and transcripts⁵⁷, may further increase statistical power to identify linkages.

To evaluate the statistical significance of *cis* eQTLs, a likelihood ratio statistic $-2(\ell_{nc} - \ell_{fc})$, was calculated comparing the log-likelihood of this model described above (ℓ_{fc}) to the log-likelihood of the model where β is re-estimated while leaving out the covariate X_c for the *cis* eQTL marker (ℓ_{nc}). A p-value was derived under the assumption that this statistic is χ^2 distributed with one degree of freedom. This p-value was used for the evaluation of significance of *cis* eQTLs for the neuronal subtypes. Within each neuronal subtype, FDR adjusted p-values were calculated using the method of Benjamini and Hochberg⁵⁸. For the other cell types (with typically much larger cell numbers) and for the genome-wide scans for eQTLs, a permutation procedure was used to calculate FDR adjusted p-values, and is described further below.

For each expressed transcript in each cell type we also scanned the entire genome for eQTLs, enabling detection of *trans* eQTLs. A similar procedure was used as for *cis* eQTL except that equation (3) was replaced with:

$$\mu = \exp(\beta_i + X_t\beta_t + \mathbf{X}_b\beta_b + X_g\beta_g) \quad (5)$$

where X_g is a vector of the scaled genotype probabilities at the g th genotypic marker, and the model is fit separately, one at a time, for each marker across the genome for each transcript. A likelihood ratio statistic for each transcript, within each cell type, for each genotypic marker is calculated by comparing this model to the model where β is re-estimated while leaving out the covariate X_g . The likelihood ratio statistic was transformed into a LOD score, by dividing it by $2\log_e(10)$. We also used functions in the *fastglm* R package for this scan, again re-using estimates of θ obtained as described above for each transcript for each cell type. For each transcript and each chromosome, QTL peak markers were identified as the marker with the

highest LOD score. The 1.5 LOD-drop procedure was used to define approximate 95% confidence intervals for QTL peaks⁵⁹.

FDR-adjusted p-values were calculated for QTL peaks. They were calculated as the ratio of the number of transcripts expected by chance to show a maximum LOD score greater than a particular LOD threshold vs. the number of transcripts observed in the real data with a maximum LOD score greater than that threshold, for a series of LOD thresholds ranging from 0.1 to 0.1+the maximum observed LOD for all transcripts within a cell type, with equal-sized steps of 0.01. Per chromosome, the number of transcripts expected by chance at a given threshold was calculated by permuting the assignments of segregant identity within each batch relative to segregant genotypes, calculating LOD scores for all transcripts across the chromosome as described above, and recording the maximum LOD score for each transcript. In each permutation instance, the permutation ordering was the same across all transcripts. We repeated this permutation procedure 10 times. Then, for each of the LOD thresholds, we calculated the average number of transcripts with maximum LOD greater than the given threshold across the 10 permutations. We used the *approxfun* function in R to interpolate the mapping between LOD thresholds and FDR and estimate an FDR-adjusted p-value for each QTL peak⁶⁰.

The same procedure was performed for *cis* eQTL analysis, with the difference being that the expected and observed number of transcripts at a given LOD threshold were calculated only at the marker closest to the transcript. We note that, as expected, Benjamini and Hochberg adjusted p-values, and FDR adjusted p-values from this permutation procedure for *cis* eQTLs were nearly identical.

scRNA-seq in the parental strains

The parental QX2314 and PTM299 strains were grown separately for 4 generations on 10cm plates, with recurrent cycles of bleaching and synchronization as was done for the segregant population. For single cell preparation, synchronized L1 from both strains were seeded together in equal numbers on 10cm plates, and they were processed together from that point onwards, to limit any environmental effects. We believe that differences in efficiency of the cell preparation procedure between N2 and CB4856 could explain the imbalanced representation in the final dataset (6,721 N2 and 3,104 CB4856 cells). We took advantage of the different parental genotypes when processing the cells, and called cells as those with at least 50 SNV counts supporting one genotype, and less than 50 supporting the other. Cell-type identification was automated by using the logistic regression model trained on the segregants which is discussed above. Differential expression analysis was carried out using the *DEsingle* R package in each cell type, as well as globally in all cells combined ⁶¹.

To compare differential expression results with our *cis* eQTL results, we first normalized the effect size of each *cis* eQTL by its standard error. Those were used directly in comparisons done within each cell type. To compare with global differential expression, those standardized effects were combined across all cell types in which an eQTL was identified using Stouffer's weighted-Z method ⁶².

Processing whole-worm eQTL data

Microarray genotype and gene expression data for our published expression QTL data were acquired from the gene expression omnibus (GEO) ²⁴. Probe sequences were realigned to the WBcel235 transcriptome using BWA, and uniquely mapping probes were used. Expression probes that were present in less than 2/3 of the sample were removed. The genotype and

expression matrices were standardized. To map eQTLs, we calculated the Pearson correlation between each probe and every genotype. Correlation coefficients were transformed to LOD scores using $-n \cdot \frac{\ln(1 - R^2)}{2\ln(10)}$. To assess significance and account for multiple testing, we permuted the sample identities 100 times and calculated the average number of transcripts with an identified eQTL at different LOD scores. We compared these results to the unpermuted LOD scores to estimate the false-discovery rate (FDR), and selected a cutoff corresponding to a rate of 10% (LOD = 4.2), equivalent to the single-cell mapping. *cis* eQTLs were derived by calculating the Pearson correlation between transcript expression and the normalized genotypes in the variant nearest to a given transcript, transforming to LOD score and comparing against the global threshold.

Hotspot analysis

To discover hotspots, we split the genome into 130 bins of 5 centimorgans each. We then counted the number of eQTLs in each bin identified in each cell type (applying a 10% FDR significance threshold), after removing all *cis* linkages. *Cis* linkages were defined here as those where the transcribed gene falls within the 95% eQTL confidence-interval range extended by 1MB on both sides. A bin was considered to have an excess of linkages if the number of linkages exceeded the number expected by chance from a Poisson distribution, given the average number of linkages per bin for that cell type and a Bonferroni correction for the total number of bins ($p < 3.8e-4$)⁶³. The *findpeaks* function in the *pracma* R package was used to identify peak hotspot bins and prevent identifying sets of adjacent bins as hotspots.

For gene ontology (GO) analysis, we identified hotspot targets using the same procedure above, but relaxed the significance threshold to 20% FDR. We then used the R package *topGO* to identify enriched terms, with the genes expressed in the cell type used as background.

Neuronal cell-type classification

Neuronal classification was carried out using a combination of available *C. elegans* scRNA-seq datasets, including a published L2 dataset^{21,22}, and the *C. elegans* Neuronal Gene Expression Map & Network (CeNGEN) project³¹. Neuronal cells were processed separately using *monocle3* with default parameters, with the exception that 100 dimensions were specified for the *preprocess_cds* step. The analysis was carried out in two passes. In the first pass, we processed all cells identified by our classifier as neurons. Following Leiden clustering, we removed 2,114 cells that were in clusters whose top genes were mostly mitochondrial and ribosomal genes, similar to the analysis described above for the global dataset. We then processed the pruned dataset in *monocle3* as described above. To annotate the final neuronal clusters, we first used the list of marker genes from two previous publications^{22,30}, to derive candidate clusters that uniquely express marker genes. We next used the *top_markers* function in *monocle3* to identify upregulated genes in each cluster compared to the rest. These were compared with the data available in the online SCeNGEN *Shiny* application (<https://cengen.shinyapps.io/SCeNGEA>) for the candidate cluster. The full list of genes used for classification is found in Table S5. In the final dataset, clusters Unknown_1 - Unknown_4 are of unknown identity and do not correspond to the clusters of the same name in CeNGEN, while the clusters Unknown_touch and Unknown_glut_2 do correspond to cell clusters of the same names in CeNGEN.

Single neuronal subtype eQTL (sn-eQTL) analysis

sn-eQTL mapping is described above (section “eQTL mapping”). GO annotation of genes with sn-eQTL was done in *topGO*, with the genes expressed in neurons (determined using the criteria for inclusion in eQTL mapping) used as background. A heatmap was plotted using the *ComplexHeatmap* package⁶⁴. To determine the consistency in effect direction between the sn-eQTL neuron and the rest of the neurons, we repeated the eQTL mapping, aggregating cells from all neuronal cell-types but omitting the neuron with the sn-eQTL. The RIC and RIM neurons in the parental datasets were identified using the same gene markers used in the segregant eQTL dataset, as described in Table S5.

Acknowledgments

The work was supported by the Howard Hughes Medical Institute and NIH grants R01 HG004321 (L.K.) and K99 HG010369 (E.B-D). Figure 1A was created with Biorender.com. We thank Michael Mashock, Marco De Simone, and other members of the Technology Center for Genomics & Bioinformatics (TCGB) at UCLA for preparing and sequencing 10x 3’ chromium single-cell libraries.

Author contributions

Conceptualization: E.B-D, J.B., J.S.B, and L.K.; Methodology: E.B-D, J.B. and J.S.B.; Investigation: E.B-D, J.B., L.G., S.Z. and J.S.B.; Formal analysis: E.B-D, J.B., and J.S.B; Writing – original draft: E.B-D and L.K.; Supervision: J.S.B and L.K; Funding acquisition: E.B-D. and L.K. All authors reviewed the manuscript.

Data availability

Raw sequencing data are available under NCBI bioproject PRJNA658829.

Code availability

Code for the HMM and eQTL mapping can be found at https://github.com/joshsbloom/single_cell_eQTL. Data and code to recreate the figures in the manuscript can be found at https://github.com/eyalbenda/worm_sceQTL.

Competing interests

The authors declare no competing interests

Materials & Correspondence

Should be addressed to E.B-D, JS.B or L.K

Supplementary Figures

Supplementary Figure 1. Observed representation of cell-types in our dataset compared to expected. The expected number of cells was calculated by manually curating the cellular lineage information available at <https://www.wormatlas.org/> for the L2 stage.

Supplementary Figure 2. Probabilistic genotyping using a hidden markov model (HMM).

A cell with the median (69) number of unique genotype-informative SNV UMI counts is shown for illustration. The trace is a summation of the probability of a CB4856 homozygous genotype and half the probability of CB4856 heterozygous genotype at each position. Each vertical line is a count for an SNV, and colors correspond to the count depth. Vertical lines pointing upwards denote counts supporting the CB4856 variant, while lines pointing downwards are counts supporting the N2 variant.

Supplementary Figure 3. *cis* eQTLs reflect gene expression differences in the parent strains. Comparison between *cis* eQTLs mapped across all cell types and gene expression differences in a dataset of 6,721 N2 and 3,104 CB4856 cells. For *cis* eQTLs mapped in multiple cell-types, effect sizes were combined using Stouffer's weighted-Z method⁶². Differential gene expression between N2 and CB4856 was calculated using R package DEsingle.

Supplementary Figure 4. Correlation between cell-type *cis*-eQTL signal and differential gene expression between the parental strains. The data corresponds to Figure S3, split out by each cell type separately.

672

673

674 **Supplementary Figure 6. Cell-type expression of genes that form troponin complexes.**

675 Size of circles corresponds to the percentage of cells expressing each gene in each cell-type,
676 and the color corresponds to average expression. Of the four troponin genes strongly expressed
677 in the body wall muscle, three are affected by a *trans*-eQTL hotspot on Chr. I.

678

bioRxiv preprint doi: <https://doi.org/10.1101/2020.08.23.263798>; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
2. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
3. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
4. Raj, T. *et al.* Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. *Science* **344**, 519–523 (2014).
5. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
6. Ishigaki, K. *et al.* Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. *Nat. Genet.* **49**, 1120–1125 (2017).
7. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* **11**, 955 (2020).
8. Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, (2020).
9. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
10. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
11. van der Wijst, M. *et al.* The single-cell eQTLGen consortium. *eLife* **9**, e52155 (2020).
12. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 1–14 (2020).

13. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
14. Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
15. Hall, D. H. & Altun, Z. F. C. *elegans atlas*. (Cold Spring Harbor Laboratory Press, 2007).
16. Hobert, O., Glenwinkel, L. & White, J. Revisiting Neuronal Cell Type Classification in *Caenorhabditis elegans*. *Curr. Biol. CB* **26**, R1197–R1203 (2016).
17. Burga, A., Ben-David, E., Vergara, T. L., Boocock, J. & Kruglyak, L. Fast genetic mapping of complex traits in *C. elegans* using millions of individuals in bulk. *Nat. Commun.* **10**, 2680 (2019).
18. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2018).
19. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 1–12 (2019).
20. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
21. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
22. Packer, J. S. *et al.* A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**, (2019).
23. Mandric, I. *et al.* Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nat. Commun.* **11**, 5504 (2020).
24. Rockman, M. V., Skrovanek, S. S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* **330**, 372–376 (2010).
25. Pukkila-Worley, R. & Ausubel, F. M. Immune defense mechanisms in the *Caenorhabditis elegans* intestinal epithelium. *Curr. Opin. Immunol.* **24**, 3–9 (2012).

26. Ono, K. & Ono, S. Tropomyosin and Troponin Are Required for Ovarian Contraction in the
Caenorhabditis elegans Reproductive System. *Mol. Biol. Cell* **15**, 2782–2793 (2004).
27. Gumienny, T. L. TGF- β signaling in C. elegans. *WormBook* 1–34 (2013)
doi:10.1895/wormbook.1.22.2.
28. Suzuki, Y. *et al.* A BMP homolog acts as a dose-dependent regulator of body size and male tail
patterning in Caenorhabditis elegans. *Dev. Camb. Engl.* **126**, 241–250 (1999).
29. Andersen, E. C. *et al.* A Powerful New Quantitative Genetics Platform, Combining Caenorhabditis
elegans High-Throughput Fitness Assays with a Large Collection of Recombinant Strains. *G3*
Genes Genomes Genet. **5**, 911–920 (2015).
30. Hammarlund, M., Hobert, O., Miller, D. M. & Sestan, N. The CeNGEN Project: The Complete Gene
Expression Map of an Entire Nervous System. *Neuron* **99**, 430–433 (2018).
31. Taylor, S. R. *et al.* Expression profiling of the mature C. elegans nervous system by single-cell
RNA-Sequencing. *bioRxiv* 737577 (2019) doi:10.1101/737577.
32. Portman, D. S. & Emmons, S. W. Identification of C. elegans sensory ray genes using whole-
genome expression profiling. *Dev. Biol.* **270**, 499–512 (2004).
33. Sarkar, A. K. *et al.* Discovery and characterization of variance QTLs in human induced pluripotent
stem cells. *PLOS Genet.* **15**, e1008045 (2019).
34. Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during
development. *Nature* **505**, 208–211 (2014).
35. Yao, C. *et al.* Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits.
Am. J. Hum. Genet. **100**, 571–580 (2017).
36. Kolberg, L., Kerimov, N., Peterson, H. & Alasoo, K. Co-expression analysis reveals interpretable
gene modules controlled by trans-acting genetic variants. *eLife* **9**, e58705 (2020).
37. Brynedal, B. *et al.* Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of
Transcriptional Co-regulation. *Am. J. Hum. Genet.* **100**, 581–591 (2017).

- 759 38. Smith, E. N. & Kruglyak, L. Gene–Environment Interaction in Yeast Gene Expression. *PLOS Biol.* **6**,
760 e83 (2008).
- 761 39. Zhang, T. *et al.* Cell-type–specific eQTL of primary melanocytes facilitates identification of
762 melanoma susceptibility genes. *Genome Res.* **28**, 1621–1635 (2018).
- 763 40. Westra, H.-J. *et al.* Cell Specific eQTL Analysis without Sorting Cells. *PLOS Genet.* **11**, e1005223
764 (2015).
- 765 41. Orozco, L. D. *et al.* Unraveling Inflammatory Responses using Systems Genetics and Gene-
766 Environment Interactions in Macrophages. *Cell* **151**, 658–670 (2012).
- 767 42. Fairfax, B. P. & Knight, J. C. Genetics of gene expression in immunity to infection. *Curr. Opin.*
768 *Immunol.* **30**, 63–71 (2014).
- 769 43. Kim-Hellmuth, S. *et al.* Genetic regulatory effects modified by immune activation contribute to
770 autoimmune disease associations. *Nat. Commun.* **8**, 266 (2017).
- 771 44. Zhang, S., Banerjee, D. & Kuhn, J. R. Isolation and Culture of Larval Cells from *C. elegans*. *PLOS*
772 *ONE* **6**, e19505 (2011).
- 773 45. Kaletsky, R. *et al.* The *C. elegans* adult neuronal IIS/FOXO transcriptome reveals adult phenotype
774 regulators. *Nature* **529**, 92–96 (2016).
- 775 46. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods*
776 **14**, 309–315 (2017).
- 777 47. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet based single-
778 cell RNA sequencing data. *bioRxiv* 303727 (2020) doi:10.1101/303727.
- 779 48. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830
780 (2011).
- 781 49. Broman, K. W. The Genomes of Recombinant Inbred Lines. *Genetics* **169**, 1133–1146 (2005).
- 782 50. Arends, D., Prins, P., Jansen, R. C. & Broman, K. W. R/qtl: high-throughput multiple QTL mapping.
783 *Bioinforma. Oxf. Engl.* **26**, 2990–2992 (2010).

51. Dodds, K. G. *et al.* Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics* **16**, 1047 (2015).
52. Bilton, T. P. *et al.* Accounting for Errors in Low Coverage High-Throughput Sequencing Data When Constructing Genetic Maps Using Biparental Outcrossed Populations. *Genetics* **209**, 65–76 (2018).
53. Rockman, M. V. & Kruglyak, L. Recombinational Landscape and Population Genomics of *Caenorhabditis elegans*. *PLOS Genet.* **5**, e1000419 (2009).
54. Rockman, M. V. & Kruglyak, L. Breeding Designs for Recombinant Inbred Advanced Intercross Lines. *Genetics* **179**, 1069–1078 (2008).
55. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol* **38**, 147–150 (2020).
56. Wood, S. N. *Generalized Additive Models: An Introduction with R*. (Chapman and Hall/CRC, 2017).
57. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**, 4288–4297 (2012).
58. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* (1995).
59. Dupuis, J. & Siegmund, D. Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**, 373–386 (1999).
60. Albert, F. W., Bloom, J. S., Siegel, J., Day, L. & Kruglyak, L. Genetics of trans-regulatory variation in gene expression. *eLife* **7**, e35471 (2018).
61. Miao, Z., Deng, K., Wang, X. & Zhang, X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* **34**, 3223–3224 (2018).
62. Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18**, 1368–1373 (2005).
63. Smith, E. N. & Kruglyak, L. Gene–Environment Interaction in Yeast Gene Expression. *PLOS Biol.* **6**, e83 (2008).
64. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinforma. Oxf. Engl.* **32**, 2847–2849 (2016).