

1 **Supplementary Materials:**

2 Materials and Methods

3 Figures S1-S24

4 Legends for Tables S1-S14 (Tables will be found separated as an excel file)

5 **Supplementary Materials:**

6 **Materials and Methods:**

7 Mammalian cell culture

8 All mammalian cells were cultured at 37°C with 5% CO₂, and were maintained in high
9 glucose DMEM (Gibco cat. no. 11965) supplemented with 10% FBS and 1X Pen/Strep (Gibco
10 cat. no. 15140122; 100U/ml penicillin, 100µg/ml streptomycin). Cells were trypsinized with
11 0.25% trypsin-EDTA (Gibco cat. no. 25200-056) and split 1:10 three times a week.

12

13 Generation of whole *C. elegans* cell suspensions

14 A *C. elegans* strain (RW12139 *stIs11435(unc-120::H1-Wcherry;unc-119(+));unc-*
15 *119(tm4063)*) carrying an integrated Punc-120::mCherry gene in a wild type background was
16 used in all experiments. A synchronized L2 population was obtained by two cycles of bleaching
17 gravid adults to isolate fertilized eggs allowing the eggs to hatch in the absence of food to
18 generate a population of starved L1 animals. Around 150,000 L1 larvae were plated on each 100
19 mm petri plate seeded with NA22 bacteria and incubated at 24°C for 15 hr to produce early L2
20 larvae. Dissociated cells were recovered following a published protocol (62) with modification.

Specifically, L2 stage worms were collected by adding 10 ml sterile ddH₂O to each plate. The collected L2s were pelleted by centrifugation at 1300 g for 1 min. The larval pellet was washed five times with sterile ddH₂O to remove bacteria. The resulting pellet was transferred to a 1.6 ml microcentrifuge tube. Around 40 µl of the final compact pellet was used for each cell dissociation experiment. The worm pellet was treated with 250 µl of SDS-DTT solution (20 mM HEPES pH8, 0.25% SDS, 200 mM DTT, 3% sucrose) for 4 min. Immediately after SDS-DTT treatment, egg buffer (118 mM NaCl, 48 mM KCl, 3 mM CaCl₂, 3 mM MgCl₂, 5 mM HEPES (pH 7.2)) was added to the SDS-DTT treated worms. Worms were pelleted at 500 g for 1 min, then washed 5 times with egg buffer). Pelleted SDS-DTT treated worms were digested with 200 µl of 15 mg/ml pronase (Sigma-Aldrich, St. Louis, MO) for 20 min. The treated worms were broken up to release cells by aspirating up and down through 21G1 ¼ needle. When sufficient single cells were observed the reaction was stopped by adding 900 µl L-15 medium containing 10% fetal bovine serum. Cells were separated from worm debris by centrifuging the pronase-treated worms at 150 g for 5 min at 4°C. The supernatant was transferred to 1.6 ml microcentrifuge tube and centrifuged at 500 g for 5 min at 4°C. The cell pellet was washed twice with egg-buffer containing 1% BSA.

Sample processing

All cell lines were trypsinized, spun down at 300xg for 5 min (4°C) and washed once in 1X PBS. *C. elegans* cells were dissociated as described above.

For sci-RNA-seq on whole cells, 5M cells were fixed in 5 mL ice-cold 100% methanol at -20°C for 10 min, washed twice with 1 ml ice-cold 1X PBS containing 1% diethyl pyrocarbonate (0.1% for *C. elegans* cells) (DEPC; Sigma-Aldrich), washed three times with 1 mL ice-cold PBS

1 containing 1% SUPERase In RNase Inhibitor (20 U/ μ L, Ambion) and 1% BSA (20 mg/ml,
2 NEB). Cells were resuspended in wash buffer at a final concentration of 5000 cells/ μ L. For all
3 washes, cells were pelleted through centrifugation at 300xg for 3 min, at 4°C.

4 For sci-RNA-seq on nuclei, 5M cells were combined and lysed using 1 mL ice-cold lysis
5 buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl₂ and 0.1% IGEPAL CA-630 from
6 (63)), modified to also include 1% SUPERase In and 1% BSA). The isolated nuclei were then
7 pelleted, washed twice with 1 mL ice-cold 1X PBS containing 1% DEPC, twice with 500 μ L
8 cold lysis buffer, once with 500 μ L cold lysis buffer without IGEPAL CA-630, and then
9 resuspended in lysis buffer without IGEPAL CA-630 at a final concentration of 5000 nuclei/ μ L.
10 For all washes, nuclei were pelleted through centrifugation at 300xg for 3 min. at 4°C).

11 For cell-mixing experiments, trypsinized cells were counted and the appropriate number
12 of cells from each cell line were combined prior to fixation or lysis. Fixed cells or nuclei were
13 then distributed into 96- or 384-well plates (Table S1). For each well, 1,000-10,000 cells or
14 nuclei (2 μ L) were mixed with 1 μ L of 25 μ M anchored oligo-dT primer (5'-
15 ACGACGCTCTTCCGATCTNNNNNNNN[10bp
16 index]TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3', where "N" is any base and "V" is
17 either "A", "C" or "G"; IDT) and 0.25 μ L 10 mM dNTP mix (Thermo), denatured at 55°C for 5
18 min and immediately placed on ice. 1.75 μ L of first-strand reaction mix, containing 1 μ L 5X
19 Superscript IV First-Strand Buffer (Invitrogen), 0.25 μ L 100 mM DTT (Invitrogen), 0.25 μ L
20 SuperScript IV reverse transcriptase (200 U/ μ L, Invitrogen), 0.25 μ L RNaseOUT Recombinant
21 Ribonuclease Inhibitor (Invitrogen), was then added to each well. Of note, the RT efficiency was
22 affected by the number of cells (or nuclei) per reaction and too many cells (>4,000) per reaction
23 resulted in lower reaction efficiency and higher impurity. For optimized efficiency, we use 2,000

mammalian cells or 5,000 mammalian nuclei per well for RT reaction. Reverse transcription was carried out by incubating plates at 55°C for 10 min, and was stopped by adding 5 µl 2X stop solution (40 mM EDTA, 1 mM spermidine) to each well. All cells (or nuclei) were then pooled, stained with 4',6-diamidino-2-phenylindole (DAPI, Invitrogen) at a final concentration of 3 µM, and sorted at varying numbers of cells/nuclei per well (depending on experiment; Table S1) into 5 uL buffer EB using a FACS Aria III cell sorter (BD). Cells are gated based on DAPI stain such that singlets are discriminated from doublets and sorted into the each well. 0.5 µl mRNA Second Strand Synthesis buffer (NEB) and 0.25 µl mRNA Second Strand Synthesis enzyme (NEB) were then added to each well, and second strand synthesis was carried out at 16°C for 150 min. The reaction was then terminated by incubation at 75°C for 20 min.

Tagmentation was carried out on double-stranded cDNA using the Nextera DNA Sample Preparation kit (Illumina). Each well was mixed with 5 ng Human Genomic DNA (Promega), as carrier to avoid over-tagmentation and reduce losses during purification, 5 µL Nextera TD buffer (Illumina) and 0.5 µL TDE1 enzyme (Illumina), and then incubated at 55°C for 5 min to carry out tagmentation. Note that because the PCR primers used to amplify libraries are specific to the RT products, tagmented carrier genomic DNA are not appreciably amplified or sequenced. The reaction was then stopped by adding 12 µL DNA binding buffer (Zymo) and incubating at room temperature for 5 min. Each well was then purified using 36 uL AMPure XP beads (Beckman Coulter), eluted in 16 µL of buffer EB (Qiagen), then transferred to a fresh multi-well plate.

For PCR reactions, each well was mixed with 2µL of 10 µM P5 primer (5'-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'; IDT), 2 µL of 10 µM P7 primer (5'-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3'; IDT), and 20 µL

NEBNext High-Fidelity 2X PCR Master Mix (NEB). Amplification was carried out using the following program: 72°C for 5 min, 98°C for 30 sec, 18-22 cycles of (98°C for 10 sec, 66°C for 30 sec, 72°C for 1 min) and a final 72°C for 5 min. After PCR, samples were pooled and purified using 0.8 volumes of AMPure XP beads. Library concentrations were determined by Qubit (Invitrogen) and the libraries were visualized by electrophoresis on a 6% TBE-PAGE gel. Libraries were sequenced on the NextSeq 500 platform (Illumina) using a V2 75 cycle kit (Read 1: 18 cycles, Read 2: 52 cycles, Index 1: 10 cycles, Index 2: 10 cycles).

sci-RNA-seq with three-level indexing

Cells were harvested and processed for reverse transcription following the same procedure as sci-RNA-seq with two-level indexing. After reverse transcription, each well was mixed with 0.66 µL second strand synthesis buffer (NEB), 0.33 µL second strand synthesis enzyme (NEB), and incubated at 16°C for 2 hours. Cells from all wells were pooled and distributed to a new 96 well plate (4.5 µL per well). 5 µL Nextera TD buffer (Illumina) and 0.5 µL indexed TDE1 enzyme (Illumina) were added to each well. Tagmentation was performed at 55°C for 10 min and stopped by adding 5 µL 2X stop solution (40 mM EDTA, 1 mM spermidine) to each well. All cells (or nuclei) were then pooled, stained with 4',6-diamidino-2-phenylindole (DAPI, Invitrogen) at a final concentration of 3 µM, and sorted at varying numbers of cells/nuclei per well (depending on experiment; see Table S1) into 5 µL buffer (4.6 µL EB buffer, 0.2 µL 1% SDS, 0.2 µL BSA (NEB)) using a FACSAria III cell sorter (BD). Cells are gated based on DAPI stain such that singlets are discriminated from doublets and sorted into the each well. After sorting, each well was mixed with 1 µL of 10 µM P7 primer (5'-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3', IDT) and incubated at 55°C for 15 min. Then each well was added with 1 µL 10% Tween-20, 1 µL nuclease-free water,

1 1μL of 10 μM indexed P5 primer (5'-
2 AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTT
3 CCGATCT-3'; IDT), and 10 μL NEBNext High-Fidelity 2X PCR Master Mix (NEB).
4 Amplification program and following steps were the same with sci-RNA-seq with two-level
5 indexing.

6 Read alignments and construction of gene expression matrix

7 Base calls were converted to fastq format and demultiplexed using Illumina's bcl2fastq/
8 2.16.0.10 tolerating one mismatched base in barcodes (edit distance (ED) < 2). Data were
9 processed with GNU Parallel (64). Demultiplexed reads were then adaptor clipped using
10 trim_galore/0.4.1 with default settings. Trimmed reads were mapped to the human reference
11 genome (hg19), mouse reference genome (mm10), *C.elegans* reference genome (PRJNA13758)
12 or a chimeric reference genome of hg19, mm10 and PRJNA13758, using STAR/v 2.5.2b (65)
13 with default settings and gene annotations (GENCODE V19 for human; GENCODE VM11 for
14 mouse, WormBase PRJNA13758.WS253.canonical_gene set for *C.elegans*). Uniquely mapping
15 reads were extracted, and duplicates were removed using the unique molecular identifier (UMI)
16 sequence (ED < 2, including insertions and deletions), reverse transcription (RT) index, and read
17 2 end-coordinate (*i.e.* reads with identical UMI, RT index, and tagmentation site were considered
18 duplicates). Finally, mapped reads were split into constituent cellular indices by further
19 demultiplexing reads using the RT index (ED < 2, including insertions and deletions). For
20 mixed-species experiment, the percentage of uniquely mapping reads for genomes of each
21 species was calculated. Cells with over 85% of UMIs assigned to one species were regarded as
22 species-specific cells, with the remaining cells classified as mixed cells or "collisions". The
23 collision rate was calculated as twice the ratio of mixed cells (as we are blind to collisions

involving cells of the same species). For gene body coverage analysis of exonic reads, the split human and mouse single cell SAM files were concatenated and exonic reads were selected and analyzed using RSEQC/2.6.1, using BED annotation files downloaded from the UCSC Golden Path. For read position analysis for intronic reads, the split human and mouse single cell SAM files were concatenated and intronic reads were selected; the fractional position of each intronic read along the genomic distance between the TSS and transcript terminus was calculated, and these values used to generate a density plot.

To generate digital expression matrices, we calculated the number of strand-specific UMIs for each cell mapping to the exonic and intronic regions of each gene with python HTseq package (66). Generally, fewer than 3% of total UMIs strand-specifically mapped to multiple genes. For multi-mapped reads, reads were assigned to the closest gene, except in cases where another intersected gene fell within 100 bp to the end of the closest gene, in which case the read was discarded. For most analyses we included both expected-strand intronic and exonic UMIs in per-gene single-cell expression matrices.

For sci-RNA-seq with three-level indexing, reads were analyzed with the same procedure, except that RT index was combined with Tn5 index, and thus the mapped reads were split into constituent cellular indices by demultiplexing reads using both the RT index and Tn5 index (ED < 2, including insertions and deletions).

t-SNE visualization of HEK293T cells and HeLa S3 cells

We visualized the clustering of sci-RNA-seq data from populations of pure HEK293T, pure HeLa S3 and mixed HEK293T + HeLa S3 cells using t-Distributed Stochastic Neighbor Embedding (t-SNE). Cells with more than 100,000 UMIs were discarded. The top 3,000 genes

with the highest variance in the digital gene expression matrix for these cells were first given as input to Principal Components Analysis (PCA). The top 10 principal components were then used as the input to t-SNE, resulting in the two-dimensional embedding of the data shown in Fig. 1F. The process was repeated using only intronic reads (fig. S4C). For this analysis, the top 2,000 (instead of 3,000) highly variable genes were used as input to PCA; all other parameters remained unchanged.

Genotyping of single HeLa cells by 3' tag sequences

HeLa S3 cell identity was verified on the basis of homozygous alleles not present in the hg19 assembly, using a callset derived from (67). Single-cell BAM files (with cellular indices encoded in the “read_id” field) were concatenated, and then processed as follows using a python wrapper of the samtools API (*i.e.* pysam). For each homozygous alternate SNV overlapping with a GENCODE V19 defined gene ($n = 865,417$) in the HeLa S3 variant callset, we computed the fraction of matching (*i.e.* HeLa S3 specific) alleles, and computed this value for all cells where at least 1 read containing a polymorphic site. We then re-plotted in R the tSNE visualization shown in fig. S4B, now colored by the relative fraction of homozygous alternate alleles called for each cell.

Comparing sci-RNA-seq and bulk RNA-seq data for HEK293T cells

To compare aggregated sci-RNA-seq single cell transcriptomes with bulk RNA-seq, we performed bulk RNA-seq using a modified protocol (33). In brief, 500 ng total RNA extracted from three biological replicate HEK293T samples (extraction using RNeasy kit (Qiagen)) with the RNeasy kit (Qiagen) were used for reverse transcription following the standard SuperScript II protocol. 500 ng total RNA (in 9 μ L water) was mixed with 2 μ L 25 uM oligo-dT(VN) (5'-

1 ACGACGCTCTTCCGATCTNNNNNNNN[10bp
2 index]TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3', where "N" is any base and "V" is
3 either "A", "C" or "G"; IDT) and 1 µL 10 mM dNTPs, then incubated at 65°C for 5 min
4 Following incubation, 8 µL reaction mix (4 µL 5X Superscript II First-Strand Buffer, 2 µL 100
5 mM DTT, 1 µL SuperScript II reverse transcriptase, 1 µL RnaseOUT) was added. Reactions were
6 incubated at 42°C for 50 min and terminated at 70°C for 15 min. For second strand synthesis, 2
7 µL RT product was mixed with 6.5 µL water, 1 µL mRNA Second Strand Synthesis buffer
8 (NEB) and 0.25 µL mRNA Second Strand Synthesis enzyme (NEB). Second strand synthesis was
9 carried out at 16°C for 150 min, followed by 75°C for 20 min. Tagmentation was carried out by
10 adding 10 µL Nextera TD buffer, 1 µL Nextera Tn5 enzyme and incubating at 55°C for 5 min.
11 Tagmented cDNA was purified using a Clean & ConcentratorTM-100 kit (Zymo) and eluted in 16
12 µL buffer EB. PCR, purification, and quantification were then performed as detailed above.

13 For comparing single cell RNA-seq and bulk RNA-seq, single cell gene counts of exonic
14 reads and intronic reads were added for the same gene from sci-RNA-seq of pure HEK293T cells
15 as well as HEK293T cells identified from HEK293T and NIH/3T3 mixed cells. Counts for bulk
16 RNA-seq of HEK293T cells were extracted based on the RT barcode and aggregated separately,
17 again adding exonic and intronic read counts per gene. Transcript counts were converted to
18 transcripts per million (TPM) and then transformed to $\log(\text{TPM} + 1)$. Pearson correlation
19 coefficients were calculated between the aggregated sci-RNA-seq and bulk RNA-seq data using
20 R.

21 Analysis of *C. elegans* whole-organism sci-RNA-seq experiments

22 Both *C. elegans* sci-RNA-seq experiments were processed identically except as noted. A

digital gene expression matrix was constructed from the raw sequencing data as described above. Cells with UMI count for protein-coding genes < 100 (experiment 1) or < 200 (experiment 2; higher threshold to compensate for slightly more leakage between cells) were excluded from the analysis. The dimensionality of this matrix was reduced first with PCA (40 components) and then with t-SNE, giving a two-dimensional representation of the data. This t-SNE was performed using the implementation in Monocle version 2.3.5 (68). Similar to the approach in (69), cells in this two-dimensional representation were clustered using the density peak algorithm (70) as implemented in Monocle 2.3.5. Genes specific to each cluster were identified and compared to microscopy-based expression profiles reported in the literature (Table S2, fig. S15-23), allowing the distinct cell types represented in each cluster to be identified. Based on these results, in experiment 1, we manually merged two clusters that both corresponded to body wall muscle, and manually split two clusters that included hypodermis, somatic gonad cells, and glia. Seven clusters exclusively contained neurons. We identified neuronal subtypes applying PCA, t-SNE, and density peak clustering to this subset of cells using the same approach as for the global cluster analysis.

In addition to neurons, body wall and intestinal/rectal muscle cells, pharyngeal cells, hypodermal cells, glial cells, intestinal cells (from experiment 2), gonad cells, and coelomocytes were each independently sub-clustered. Clusters from these iterative t-SNE analyses that featured expression of marker genes from multiple tissues were identified as likely doublets. These cells, which comprised $\sim 2.5\%$ of the total, were excluded from all downstream analyses.

Consensus expression profiles for each cell type except intestine were constructed by first dividing each column in the gene-by-cell digital gene expression matrix for experiment 1 by the cell's size factor and then for each cell type, taking the mean of the normalized UMI counts for

the subset of cells assigned to that cell type. These mean normalized UMI counts were then re-scaled to transcripts per million. Cells that had a UMI count of less than one quarter of the median for their assigned cell type were excluded from the consensus expression profiles. The intestine consensus expression profile was generated in the same manner, but used cells from experiment 2 instead of experiment 1.

95% confidence intervals for the mean expression of each gene in each cell type were estimated using a normal approximation to the negative binomial distribution. For each cell type, the expression of a given gene was assumed to follow a negative binomial distribution, with a mean μ and dispersion parameter α estimated using Monocle's `estimateDispersions` function (using only cells of that particular cell type). The variance of this random variable is equal to $\mu + \mu^2\alpha$. By the central limit theorem, the values of the estimate for the mean will asymptotically approach a distribution $N(\mu, (\mu + \mu^2\alpha) / n)$, where n is the number of cells of the cell type in question. Confidence intervals for the true value of μ are computed based on this normal approximation.

Genes with expression patterns highly enriched in a single tissue were identified as follows. For each gene (excluding those expressed in fewer than 10 cells), the tissue in which it is expressed highest and the tissue in which it is expressed second-highest (relative to other tissues) are enumerated. The gene is considered enriched in the highest expressing tissue if it is both expressed at a >5-fold greater level than in the second-highest expressing tissue and the differential expression of this gene between the highest and second-highest expressing tissues is non-zero at a false detection rate of < 5%. The differential expression tests are performed with the `differentialGeneTest` function of Monocle 2 (68). The false detection rates are computed based on the tests for all genes, not just the genes with a given highest/second-highest expressing

1 tissue. Genes with expression patterns enriched in a single cell type or a single neuron cluster
2 were identified using the same method (*i.e.* comparing the highest and second-highest expressing
3 cell type instead of tissue).

4 Differential expression tests for analyses presented in Fig. 4F,H and fig. S10B,D,F were
5 also conducted using the differentialGeneTest function of Monocle 2, excluding genes expressed
6 in fewer than 10 cells total among the cell types being compared (*e.g.* when comparing the ASEL
7 vs. ASER neurons, genes are considered if they are expressed in at least 10 ASEL/R cells).

8 Integration of sci-RNA-seq expression profiles and modENCODE (61)/modERN (46) ChIP-seq 9 data

10 Transcription factor (TF) ChIP-seq datasets were downloaded from the ENCODE data
11 portal. The ChIP-seq data included experiments conducted on whole embryos or whole larvae at
12 different developmental stages. ChIP peaks for the same TF were merged if they overlapped and
13 were either both from an embryonic stage experiment or both from a post-embryonic stage
14 experiment. If a TF had both embryonic and post-embryonic data available, only the post-
15 embryonic data was used.

16 A ChIP-seq peak was considered to be associated with a gene if: 1) the peak summit was
17 within 2 kb of the canonical transcription start site (TSS) for the gene, 2) the distance from the
18 peak summit to the second closest TSS (regardless of strand) was at least 50% greater than the
19 distance to the closest TSS, and 3) the peak overlapped peaks for < 20% of assayed TFs from the
20 same broad developmental stage (embryonic or post-embryonic). This excludes so-called “HOT
21 regions” which are likely to reflect either non-sequence-specific TF binding or an artifact of the
22 ChIP-seq assay (71).

Each gene-associated ChIP-seq peak is assigned a score equal to 0.2 minus the proportion of assayed TFs from the same broad developmental stage (embryonic or post-embryonic) that have peaks which overlap the peak in question. This serves to further down-weight peaks in marginally HOT regions. Each gene is assigned a score for each TF that is equal to the maximum peak score of all peaks for the TF that are assigned to the gene (or zero, if no such peaks exist). These scores are referred to as “TF association scores” below.

For each of the 27 cell types with sci-RNA-seq consensus expression profiles, a regression model was constructed to predict the expression levels of genes in the given cell type based on the TF association scores for each individual gene. The response in these models was $\log_2(\text{transcripts per million} + 1)$ for each gene. The features are the TF association scores for each gene; however, only scores for TFs that are expressed with at least 10 transcripts per million in the cell type in question are included as features. The models are fit using elastic net regularization as implemented in the R package *glmnet*. Model coefficients shown in Fig. 5 are from models fit with the largest regularization parameter that gives a mean squared error (MSE) less than 1 standard error from the MSE of a model with the optimal regularization parameter, as inferred by cross validation (“lambda.1se”).

To identify pairs of TFs that have co-localized binding patterns more often than could be expected by chance (fig. S13), peaks were first clustered by recursively merging those with summits within 150 bp of each other. This analysis was limited to TFs with ChIP-seq data from post-embryonic worms, and also included germline-specific ChIP-seq for EFL-1 and DPL-1 produced by (57). Peak clusters that contained peaks for >20% of the TFs (“HOT regions”) were excluded from further analysis. Peak clusters were associated with genes using the same criteria as used for individual peaks (described above, treating the midpoint of the cluster’s genomic

interval as the “summit” of the cluster). Peak clusters that could not be associated with a gene were excluded from further analysis. From the remaining peak clusters, a matrix was constructed where the rows are identifiers for each peak cluster and the columns are binary variables with value 1 if the cluster includes at least one peak for a given TF, 0 otherwise.

This matrix was used as input to the Graphical LASSO (72), an algorithm which provides robust estimates of partial correlations between a set of random variables given a limited number of observations and under the assumption that most variables are conditionally independent from another. In this context, the partial correlation between two columns of the input matrix is equal to the correlation of the events “>0 peaks for TF 1 are present in this peak cluster” and “>0 peaks for TF 2 are present in this peak cluster”, conditioned on the presence or absence of peaks for all other TFs. The Graphical LASSO was applied to either the full matrix (fig. S13D) or the subset of rows in the matrix that corresponded to peak clusters in the promoters of gonad-enriched genes (fig. S13A) or neuron-enriched genes (fig. S13C). From the partial correlations outputted by each Graphical LASSO, we constructed a network where the nodes are TFs (columns in the matrix) and undirected edges connect each pair of TFs for which the partial correlation in either direction ($TF\ 1 \rightarrow TF\ 2$ or $TF\ 2 \rightarrow TF\ 1$) is > 0.01 .

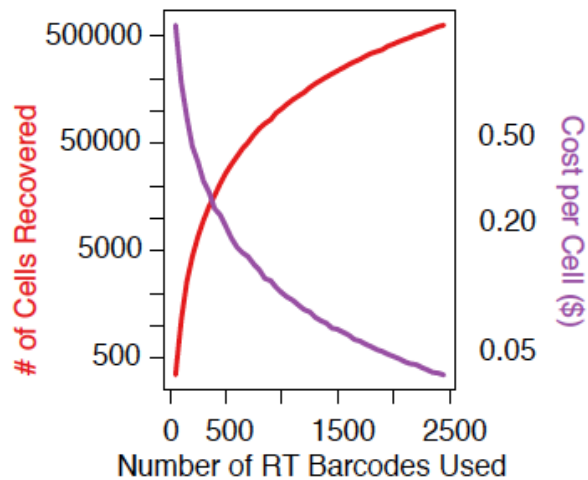
The Graphical LASSO model requires a regularization parameter to be set by the user, with increasing values. We set this parameter to the smallest value that satisfied the requirement that the probability that a non-zero partial correlation in the output is in fact zero in the “true” model—the false detection rate—is less than 5%. To find a mapping between regularization parameter values and the false detection rate, we constructed a null model by shuffling the values of the input matrix in a manner that preserves both row and column sums, using the CurveBall algorithm (73). In a shuffled matrix, all non-zero partial correlations reported by the Graphical

LASSO are false detections. We therefore estimate the false detection rate of a given regularization parameter value to be equal to the mean number of non-zero partial correlations reported by the Graphical LASSO for shuffled matrices (averaging over 50 shuffles) divided by the number of non-zero partial correlations reported by the Graphical LASSO on the unshuffled input data.

Cost estimation

Using the 576 x 960 sci-RNA-seq experiment as an example, reagent costs are largely enzyme-driven and include SuperScript IV reverse transcriptase (\$934), second strand synthesis mix (\$750), Nextera Tn5 enzyme (\$5,000), NEBnext master mix (\$1,150), FACS sorting (\$250) and other reagents and plates (\$250). If we sort 60 cells per well (assuming recovery rate is 100%) for 960 wells (5% collision rate), then the reagent cost of library preparation is around \$0.14 per cell (expected yield of around 55,000 cells). However, it is worth noting that simply increasing the number of cells sorted per well decreases costs (*e.g.* sorting 150 cells to each well would yield around 140,000 cells at a cost of \$0.05 per cell), but also results in an increased collision rate (12%). Alternatively, by increasing to 1,536 barcodes during the first (RT-based) round of indexing, we can sort up to 320 cells per well at a 10% collision rate, thereby reducing the cost per cell to less than \$0.025 per cell. Straightforward reductions in reaction volumes and/or in-house enzyme production at all steps may also lead to further reductions in costs, as would additional rounds of molecular indexing. For example, with 384 x 384 x 384 combinatorial indexing, we can potentially uniquely barcode the transcriptomes of around 12 million cells at a 10% collision rate, corresponding to >200-fold increase in detection capacity relative to the 576 x 960 experiment, without much increase in reagent costs.

1 Supplementary Figures



2

3 Fig. S1

4 **Combinatorial indexing with increasing numbers of reverse transcription (RT) barcodes**

5 **enables sublinear scaling of cost per cell.** Plot assumes two-level indexing and estimates how

6 detection capacity (i.e. the number of cells detected in a sci-RNA-seq experiment, red) and cost

7 per cell (blue) vary as a function of the number of RT barcodes used, assuming a collision rate of

8 5%.

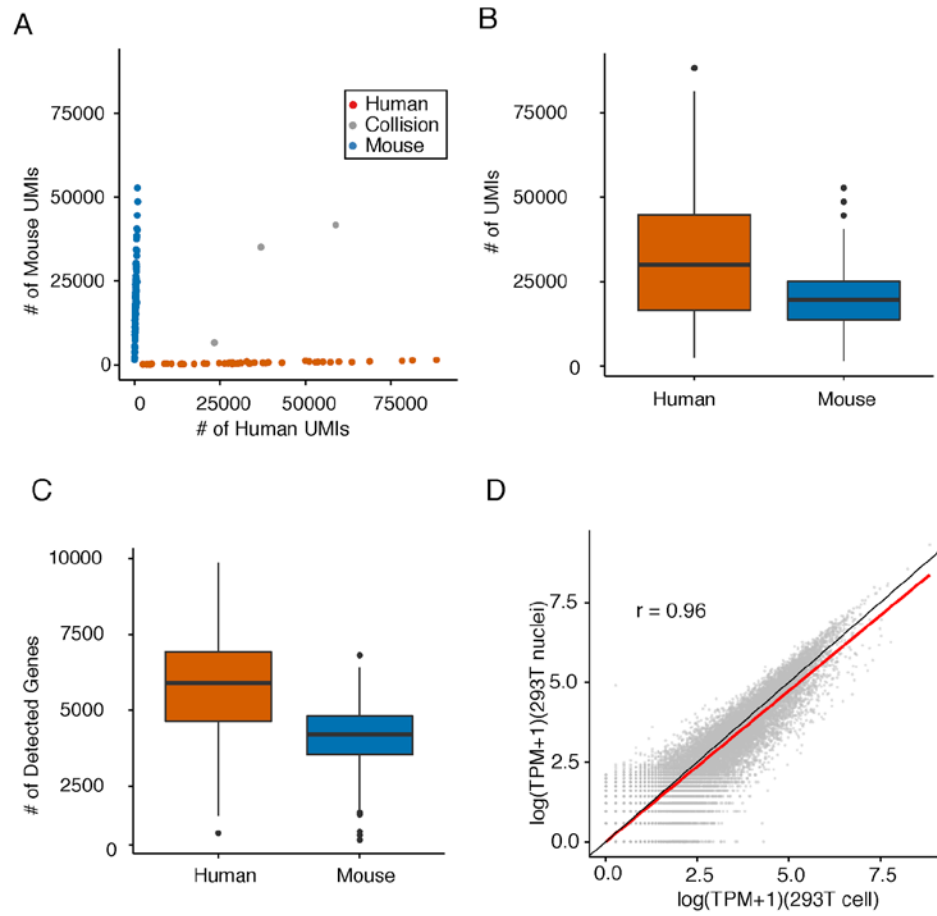
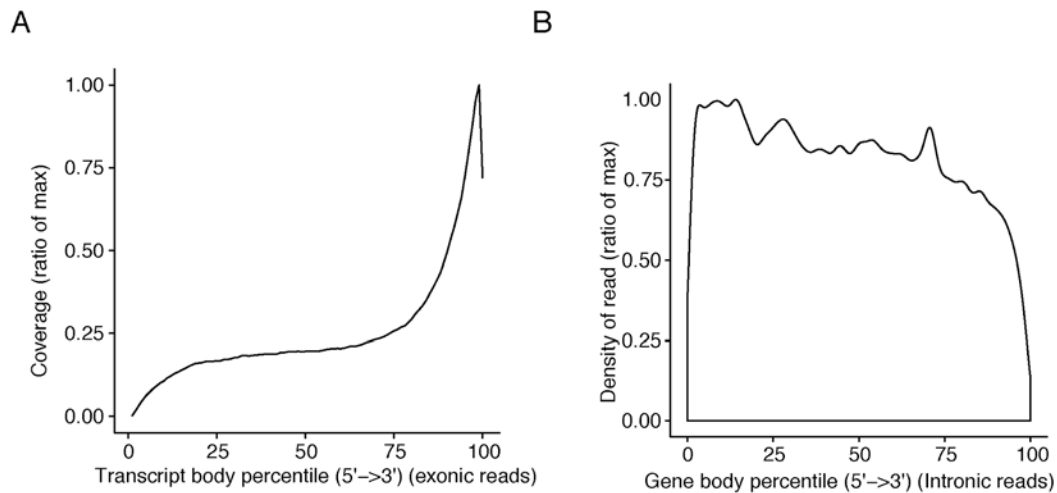


Fig. S2

sci-RNA-seq is compatible with isolated nuclei as starting material. (A) Scatter plot of unique human and mouse nuclei UMI counts from a 96 x 96 sci-RNA-seq experiment. This experiment included different cell populations (Table S1), but only cells originating from a mixture of human (HEK293T) and mouse (NIH/3T3) nuclei are plotted here. Inferred mouse cells ($n = 124$) are colored in blue; inferred human cells ($n = 48$) are colored in red, and “collisions” ($n = 3$) are colored in grey. (B to C) Boxplots showing the number of UMIs (B) and genes (C) detected per cell in nuclear sci-RNA-seq experiments. (D) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HEK293T cells ($n = 328$) vs.

1 HEK293T nuclei (n = 48), together with a linear regression line (red) and y=x line (black).

2



3

4

5 **Fig. S3**

6 **Positional bias of exonic and intronic sci-RNA-seq reads.** (A) Density plot showing that as

7 expected, sci-RNA-seq reads mapping to exons are strongly biased to originate near the 3' ends

8 of transcripts (intronic regions excluded from percentile scaling). (B) Density plot showing that

9 in contrast, sci-RNA-seq reads mapping to introns do not exhibit 3' bias (intronic regions

10 included in percentile scaling). Y-axis is scaled to the ratio of max.

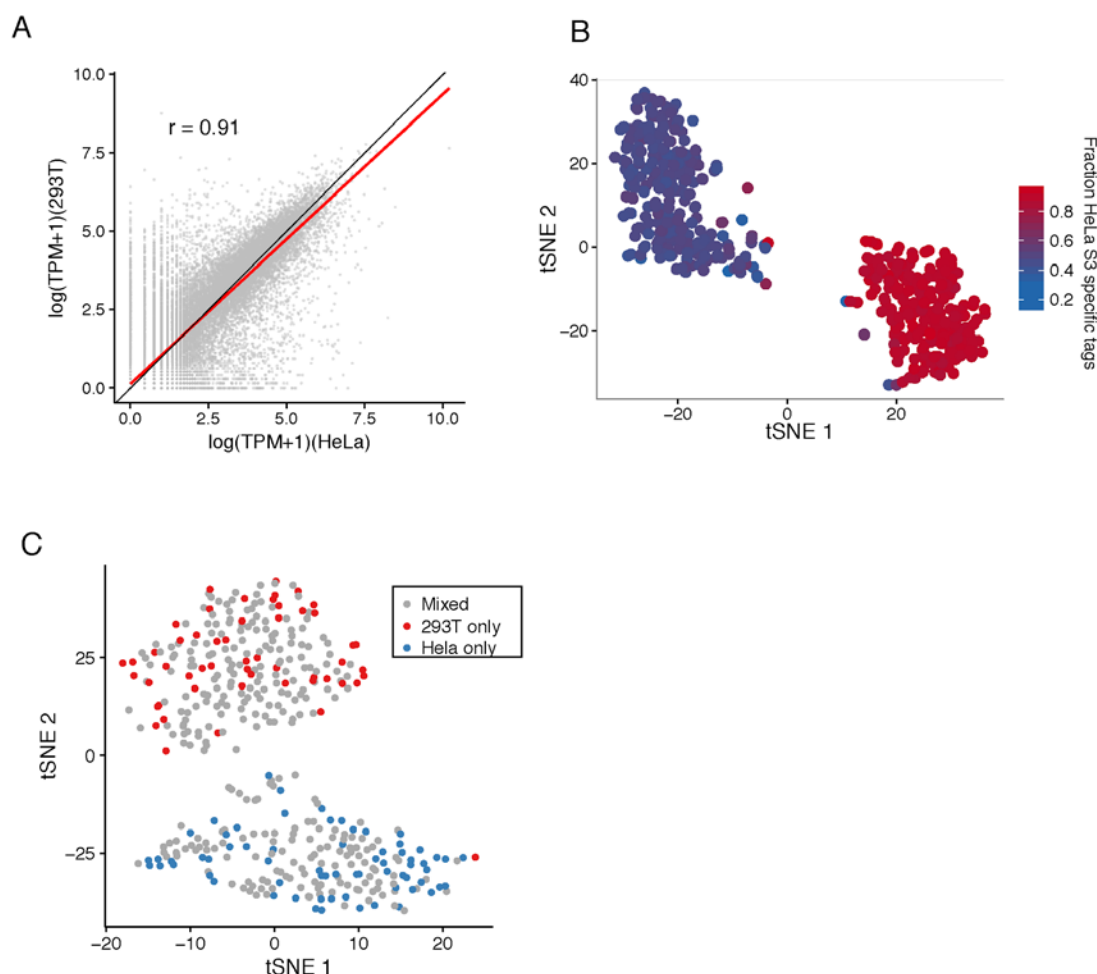
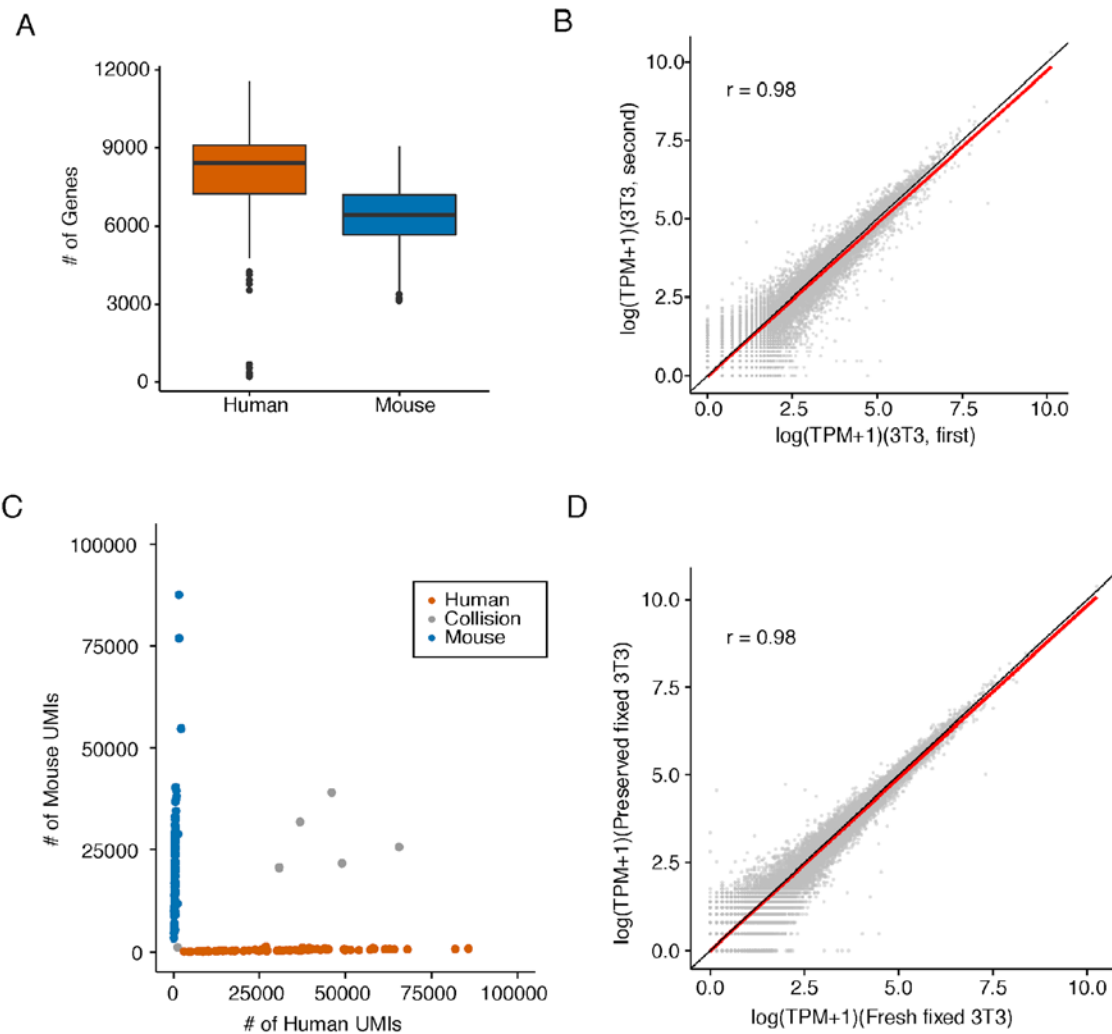


Fig. S4

Quality control for sci-RNA-seq on mixed populations of HeLa S3 and HEK293T cells. (A)

Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HeLa S3 vs. HEK293T cells, together with a linear regression line (red) and y=x line (black). **(B)** tSNE plot (as in Fig. 1F), with cells colored by fraction of reads harboring HeLa S3 specific SNVs (single nucleotide variants) relative to hg19 assembly. **(C)** tSNE using digital gene expression matrices constructed from only intronic reads. Cells are colored by the population from which they derived, with pure HEK293T in red, pure HeLa S3 in blue, and mixed cells in grey.

1



2

3 Fig. S5

4 **sci-RNA-seq shows robust gene expression measurements.** (A) Boxplots showing the number
 5 of genes detected per cell in a 16 x 84 well sci-RNA-seq experiment. (B) Correlation between
 6 gene expression measurements in aggregated sci-RNA-seq profiles of NIH/3T3 cells from two
 7 sci-RNA-seq experiments, performed two months apart and on independently grown and fixed
 8 cells, together with a linear regression line (red) and $y=x$ line (black). (C) Scatter plot of unique

1 human and mouse UMI counts from a 16 x 84 sci-RNA-seq experiment on mixed HEK293T and
2 NIH/3T3 cells after methanol fixation and freezing at -80°C for 4 days. Inferred mouse cells (n =
3 90) are colored in blue; inferred human cells (n = 89) are colored in red, and “collisions” (n = 6)
4 are colored in grey. **(D)** Correlation between gene expression measurements in aggregated sci-
5 RNA-seq profiles of fixed-fresh vs. fixed-frozen NIH/3T3 cells, together with a linear regression
6 line (red) and y=x line (black).

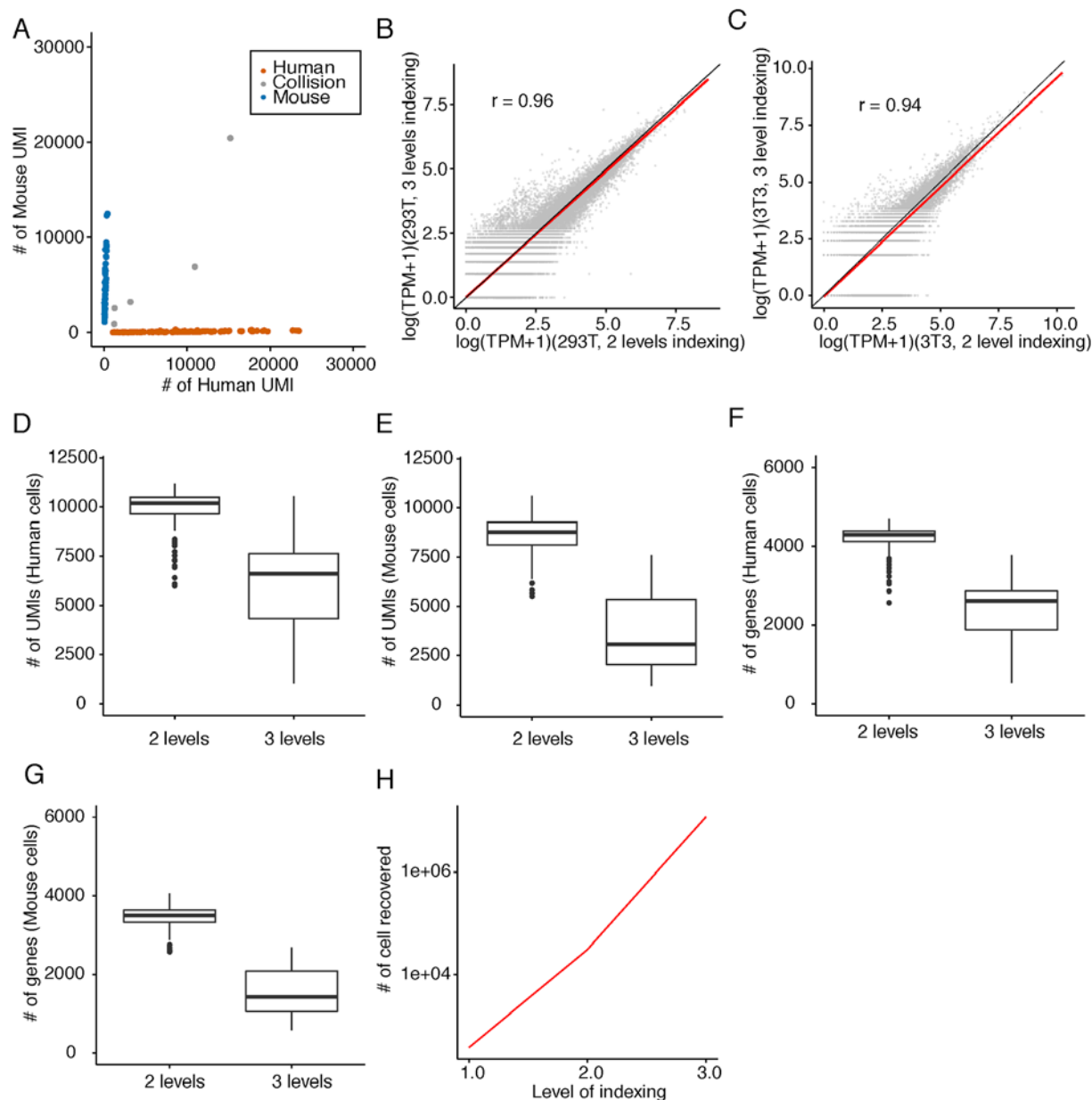


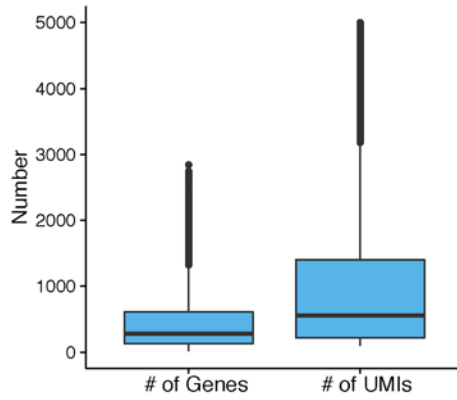
Fig. S6

Representative result from sci-RNA-seq with 3-level indexing. (A) Scatter plot of unique human and mouse UMI counts from a 16 x 6 x 16 sci-RNA-seq experiment on mixed HEK293T and NIH/3T3 cells. Inferred mouse cells ($n = 62$) are colored in blue; inferred human cells ($n = 119$) are colored in red, and “collisions” ($n = 5$) are colored in grey. (B) Correlation between

1 gene expression measurements in aggregated sci-RNA-seq profiles of HEK293T cells with 2-
2 level vs. 3-level indexing, together with a linear regression line (red) and $y=x$ line (black). (C)
3 Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of
4 NIH/3T3 cells in sci-RNA-seq with 2-level vs. 3-level indexing, together with a linear regression
5 line (red) and $y=x$ line (black). (D to E) Boxplots showing the number of UMIs detected per
6 HEK293T cell (D) and NIH/3T3 cell (E) in sci-RNA-seq with 2-level or 3-level indexing,
7 sampling 15,000 total reads per cell. (F to G) Boxplots showing the number of genes detected
8 per HEK293T cell (F) and NIH/3T3 cell (G) in sci-RNA-seq with 2-level or 3-level indexing,
9 sampling 15,000 total reads per cell. (H) Plot illustrating how estimated detection capacity (*i.e.*
10 the number of cells detected in a sci-RNA-seq experiment, red) varies as a function of number of
11 rounds of indexing used, assuming a collision rate of 10% and 384 indexes at each level.

12

A



B

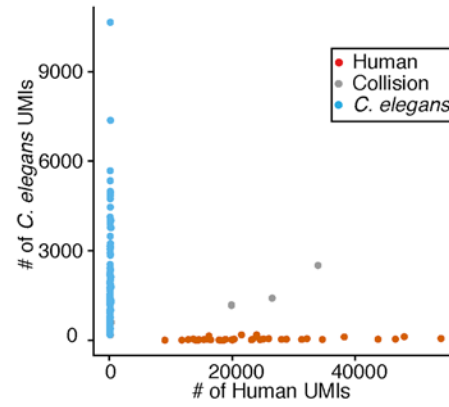


Fig. S7

Quality control metrics for *C. elegans* sci-RNA-seq experiments. (A) Distribution of number of protein-coding genes and UMI counts (mapping to protein-coding genes) detected per *C. elegans* cell. (B) Scatter plot of unique UMI counts per cell from a sci-RNA-seq experiment performed on mixture of HEK293T (human) and *C. elegans* cells.

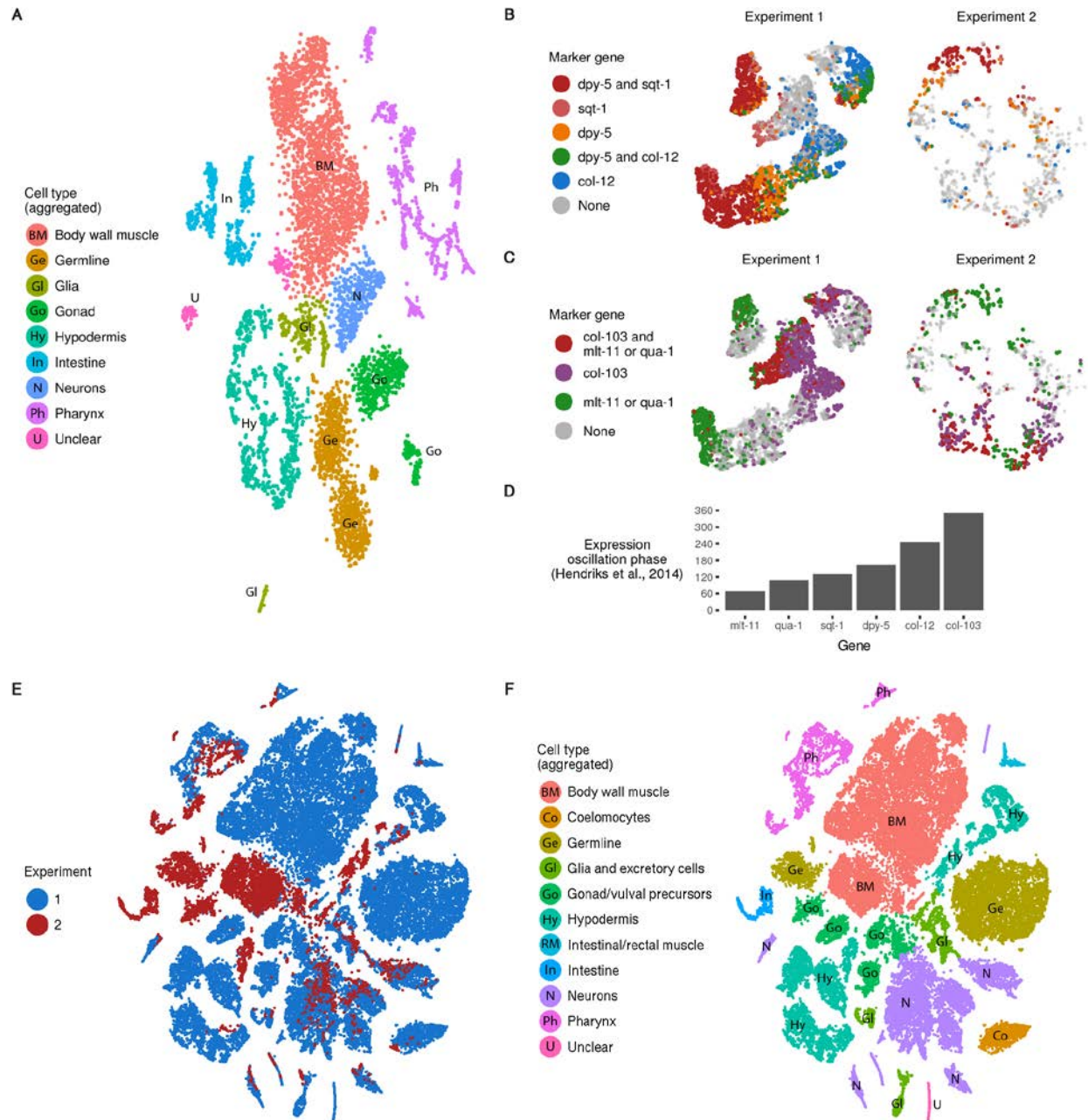


Fig. S8

A second *C. elegans* sci-RNA-seq experiment recovers intestine cells. (A) t-SNE visualization of cells from the second *C. elegans* experiment, which included all cells (96 wells) or only cells with high DAPI stain (48 wells). 511 intestine cells were successfully recovered. (B) Expression of the cuticle collagens *dpy-5*, *sqt-1*, and *col-12* in cells from experiments 1 and 2. t-SNE

coordinates for cells are the same as in Fig. 3A (for experiment 1) and (A) (for experiment 2), but only hypodermal cells are shown. *dpy-5* and *sqt-1* are expressed during the synthesis of new cuticle preceding each larval molt, while *col-12* is expressed during molting and ecdysis (74). (C) Expression of the signaling gene *qua-1*, the protease inhibitor *mlt-11*, and the collagen *col-103*, in experiments 1 and 2. *qua-1* and *mlt-11* are expressed at the initiation of new cuticle synthesis (75). *col-103* is expressed in the intermolt, after ecdysis but before new cuticle synthesis begins (36). Taken together with (B), the expression patterns suggest that the worms in experiment 1 spanned a range of developmental sub-stages from late L2 to around the L3 molt, while worms from experiment 2 had greater synchrony and were mostly from the early L2 stage. (D) Phase of the molting-cycle associated gene expression oscillations of selected genes, as reported by (36). The values are modulo 360, *i.e.* 360 is the same as 0 and equidistant from 90 and 270. (E to F) t-SNE visualizations of cells from both *C. elegans* experiments processed together.

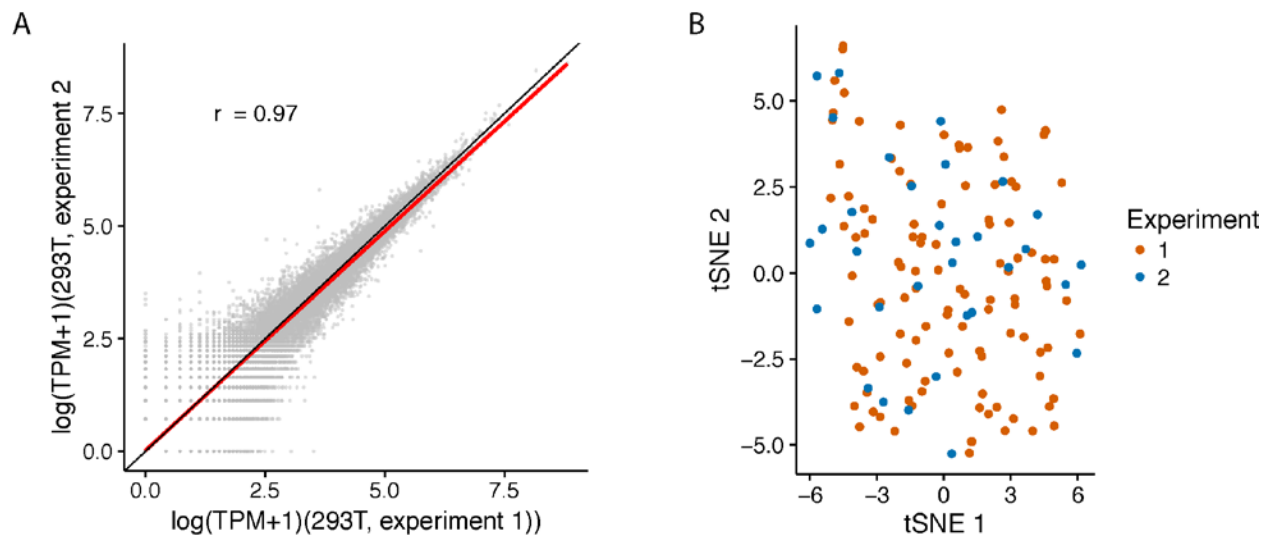


Fig. S9

1 **Evaluation of technical variance between the two *C. elegans* experiments.** (A) Correlation
2 between gene expression measurements in aggregated sci-RNA-seq profiles of HEK293T cells
3 spiked in with in the first *C. elegans* experiment (n = 32) vs. the second experiment (n = 111),
4 together with a linear regression line (red) and y=x line (black). (B) t-SNE clustering of
5 HEK293T cells recovered from the two experiments. Cells are colored by the experiment from
6 which they derived.

7

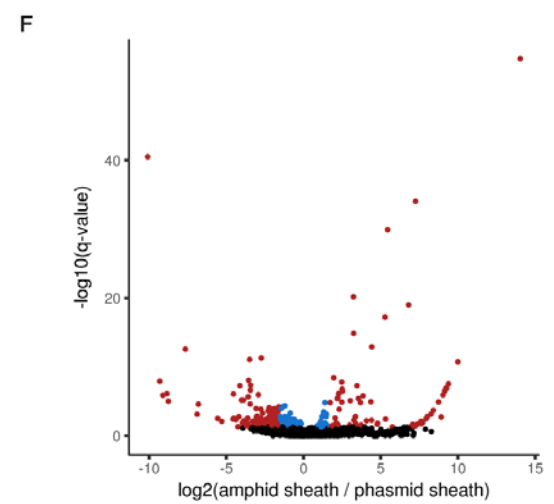
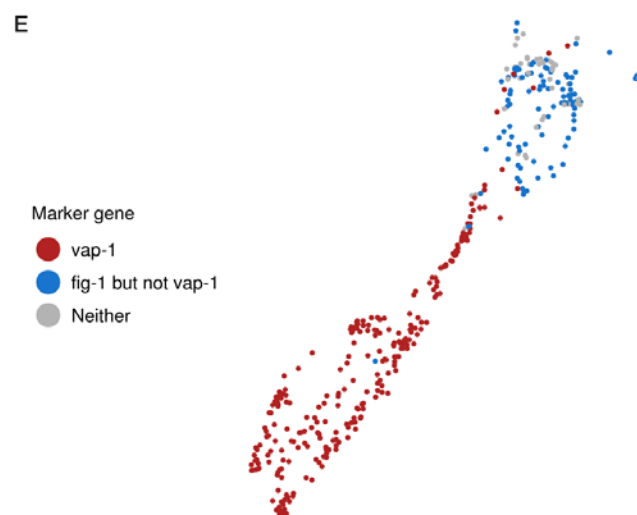
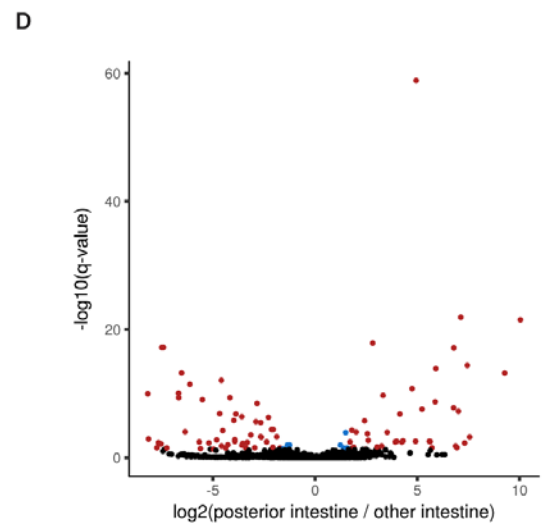
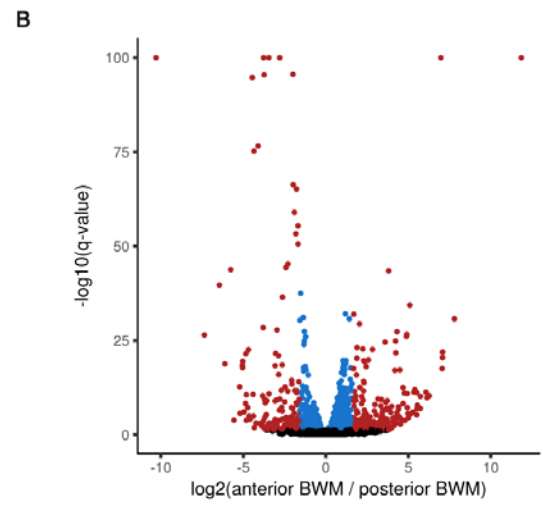
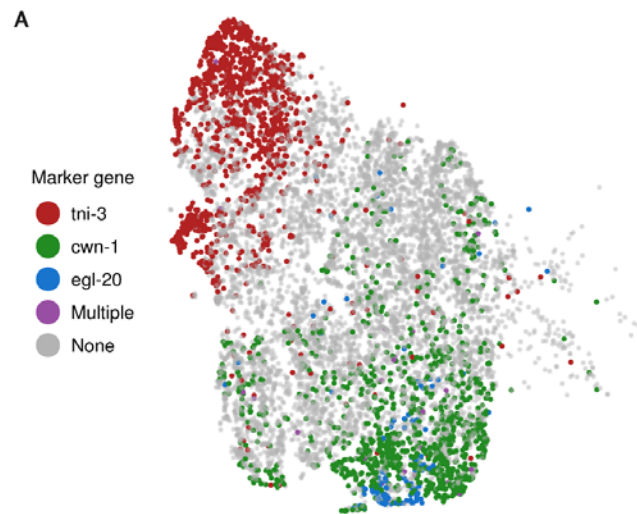


Fig. S10

sci-RNA-seq reveals genes differentially expressed between anterior and posterior cells for

three cell types. (A) Expression of anterior/posterior marker genes in body wall muscle cells.

Cell t-SNE coordinates are the same as in Fig. 3A, except only BWM cells are shown. *tni-3* (red)

is specific to the head (39), while *cwn-1* (green) and *egl-20* (blue) are specific to the posterior

and tail respectively (76). **(B)** Volcano plot showing genes differentially expressed between

anterior [*tni-3*(+)] and posterior [*cwn-1*(+) or *egl-20*(+)] body wall muscle. -log₁₀ q-values (y-

axis) are capped at 100. Genes with differential expression q-value < 0.05 are colored red if the

fold difference in expression is >3, blue otherwise. **(C)** Expression of posterior marker genes in

intestine cells. Cell t-SNE coordinates are the same as in fig. S10A, except only intestine cells

are shown. *pbo-4* and *nob-1* are specific to the posterior (77, 78). **(D)** Volcano plot showing

genes differentially expressed between posterior [*pbo-4*(+) or *nob-1*(+)] intestine and other

intestine. Colors are the same as in (B). **(E)** Expression of amphid/phasmid (anterior/posterior)

marker genes in amphid/phasmid sheath cells. Cell t-SNE coordinates are the same as in Fig. 3A,

except only amphid/phasmid sheath cells are shown. *fig-1* is expressed in both amphid and

phasmid sheath cells, while *vap-1* is specific to the amphid sheath cells. (79, 80). **(F)** Volcano

plot showing genes differentially expressed between amphid [*vap-1*(+)] and phasmid [*fig-1*(+)

vap-1(-)] sheath cells. Colors are the same as in **(B)**.

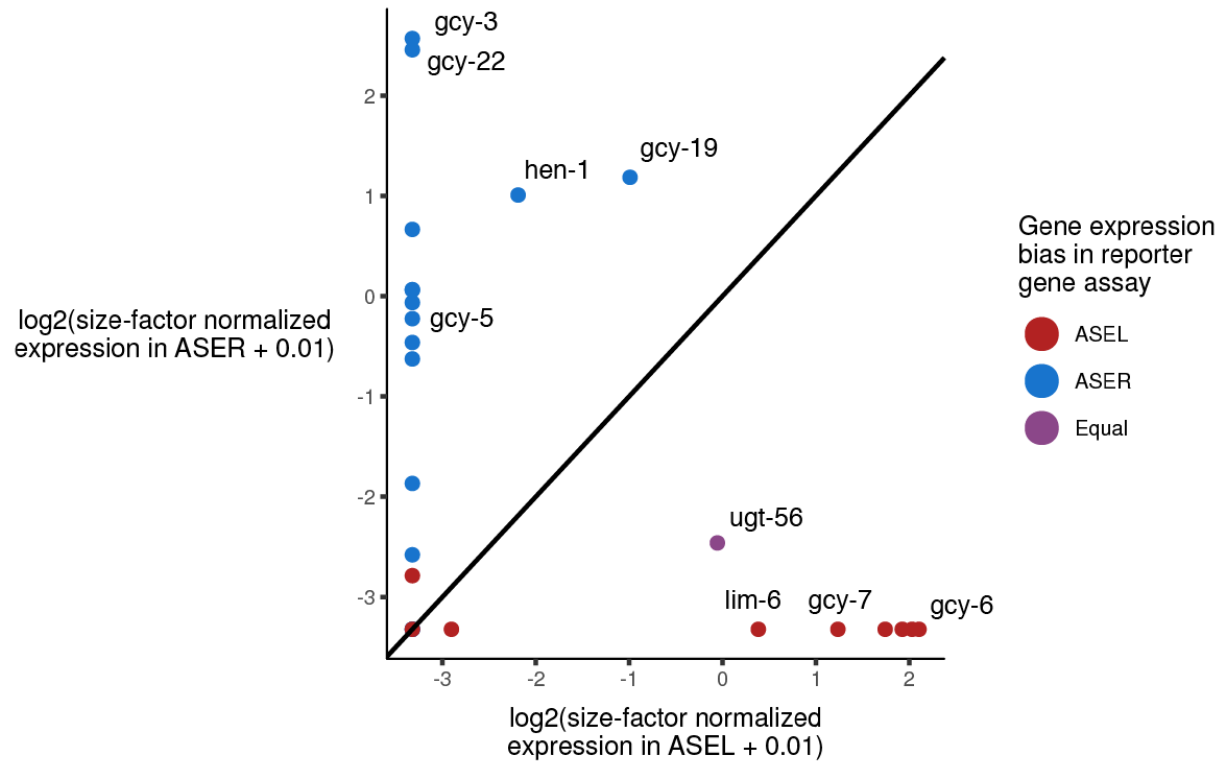


Fig. S11

sci-RNA-seq expression profiles for the ASEL and ASER neurons are consistent with reporter gene assays for asymmetric gene expression. Points represent genes which were tested for asymmetric expression between the ASEL and ASER neurons in promoter-fusion reporter gene assays, as reported by (43). Point colors show the expression bias observed in the reporter gene assay for a given gene. The x-axis and y-axis show the log-transformed, size-factor normalized mean number of unique molecular identifiers observed for a given gene per ASEL and ASER cell respectively in the sci-RNA-seq data.

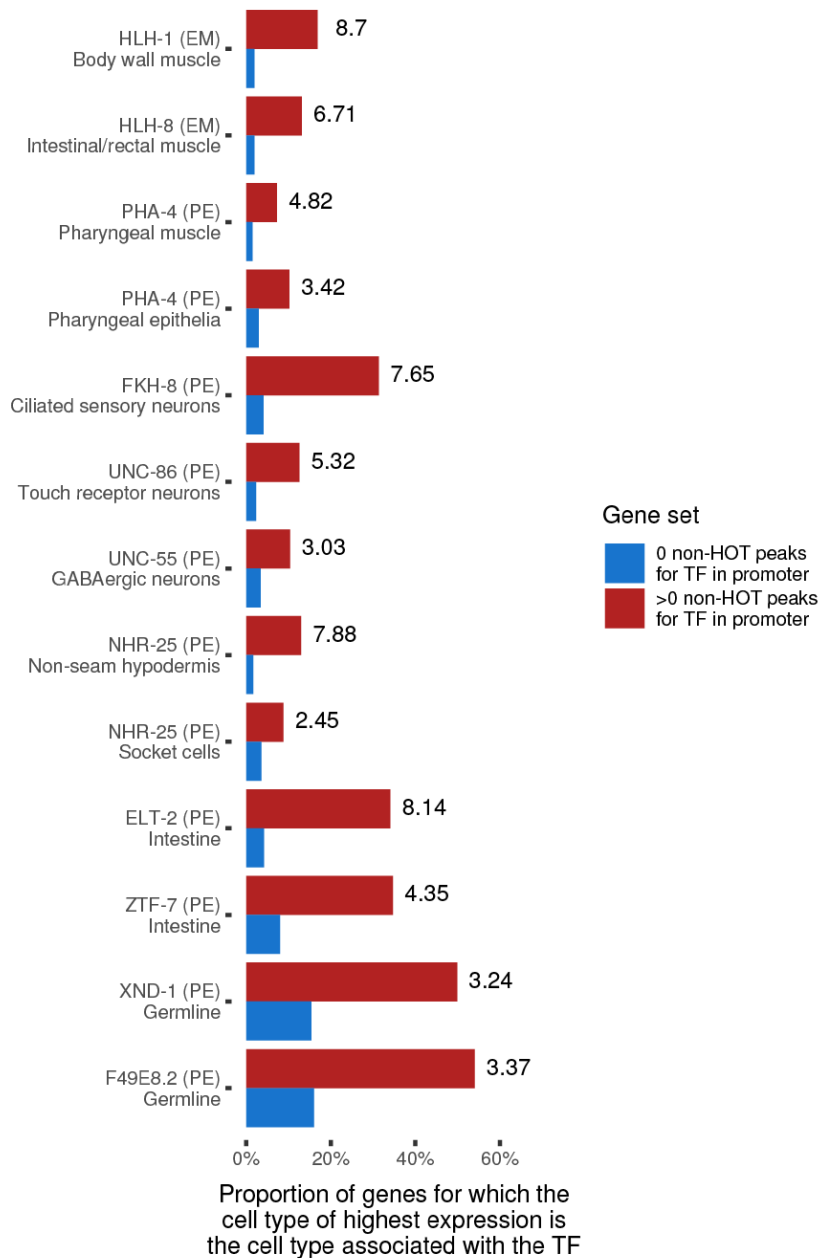


Fig. S12

Transcription factor ChIP-seq peaks predict cell type enriched gene expression. For many TF-to-cell-type associations, the presence of a ChIP-seq peak for the TF in the promoter of a given gene substantially increases the likelihood of the associated cell type being the cell type in which the gene is most highly expressed. Red bars show this probability for genes with at least

1 one peak for the listed TF in their promoter; blue bars show the probability for genes with no
2 peak for the TF in their promoter. Numbers next to the red bars show the ratio of the
3 probabilities for genes with >0 vs. 0 peaks for the TF in their promoter. The associations here are
4 selected examples, each having a positive coefficient in Fig 5. A “PE” following a TF name
5 indicates that the ChIP-seq dataset(s) for that TF are from post-embryonic worms; “EM”
6 indicates that they are from embryos. “HOT region” peaks, defined as those which overlap
7 peaks >20% of all TFs assayed in the same broad developmental stage (embryonic or post-
8 embryonic), are excluded from the analysis.

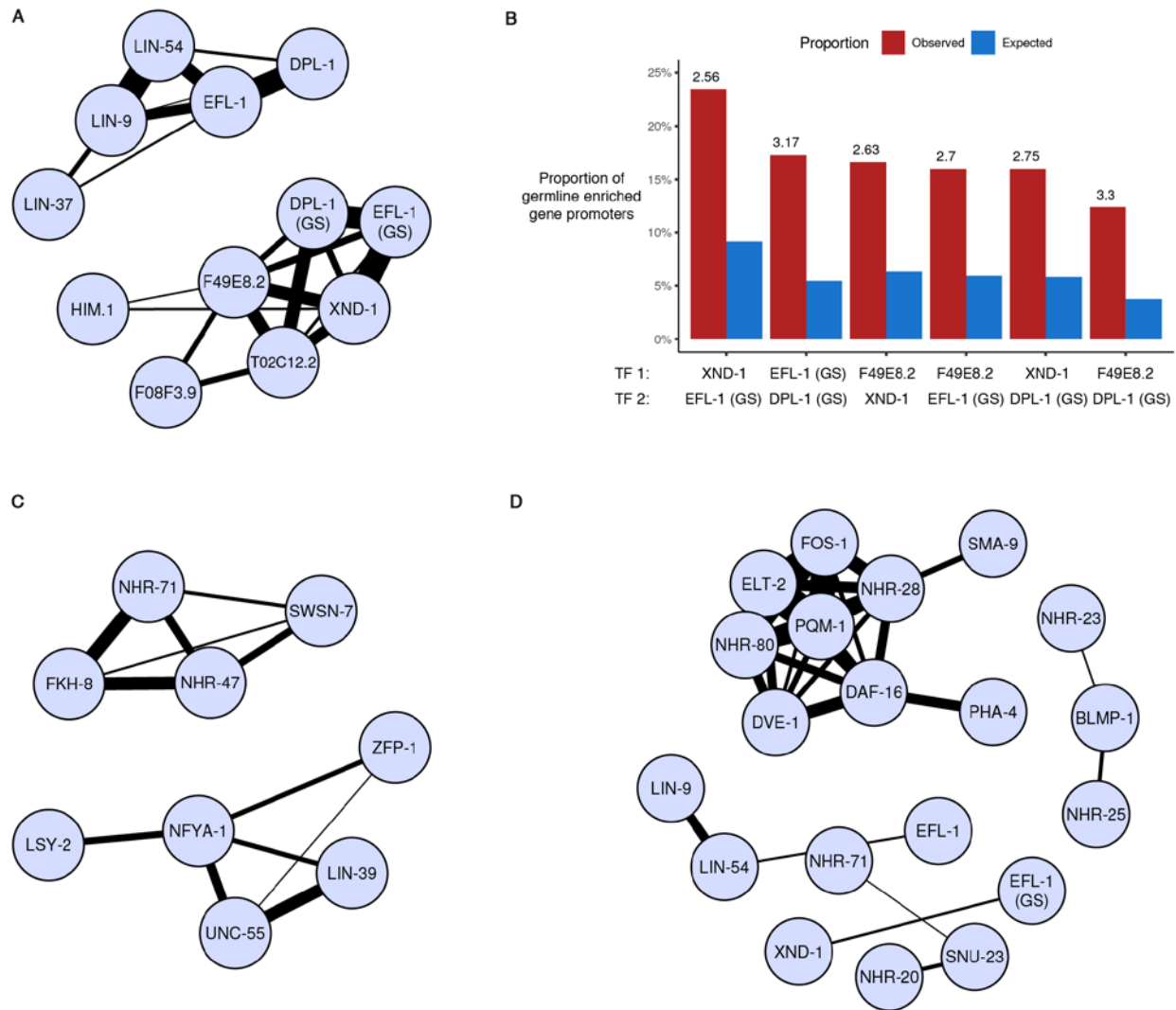


Fig. S13

Transcription factor ChIP-seq peaks have distinct co-localization patterns in the promoters of genes with tissue-enriched expression patterns. (A, C and D) A Graphical LASSO model (Methods) is used to find pairs of transcription factors which have overlapping ChIP-seq peaks more often than could be expected by chance, in the context of (A) the promoters of genes with gonad-enriched expression (>5-fold greater in gonad than in any other tissue), (C) the promoters of genes with neuron-enriched expression, or (D) the promoters of all genes. All TF ChIP-seq in this analysis is from post-embryonic stages. EFL-1 (GS) and DPL-1 (GS) refer to peaks from

1 germline-specific ChIP-seq datasets from (57). EFL-1, DPL-1, LIN-9, LIN-37, and LIN-54 are
2 members of the DRM complex (*C. elegans* ortholog of the mammalian DREAM complex),
3 which activates a subset of genes in the germline while repressing them in soma (57, 81–83). **(B)**
4 The observed proportion of germline-enriched genes (those with germline expression >5-fold
5 higher than in any other cell type) that have peaks for both listed TFs in their promoter (in red),
6 compared to the proportion that would be expected if the TF binding patterns were independent
7 conditional on being in a germline-enriched gene promoter (in blue). The numbers above each
8 red bar is the ratio of observed / expected. The conditioning of these statistics on the context of
9 being in a germline-enriched gene promoter rules out the possibility that the co-localizations
10 observed in (A) are simply due to each TF independently being associated with germline-specific
11 genes.

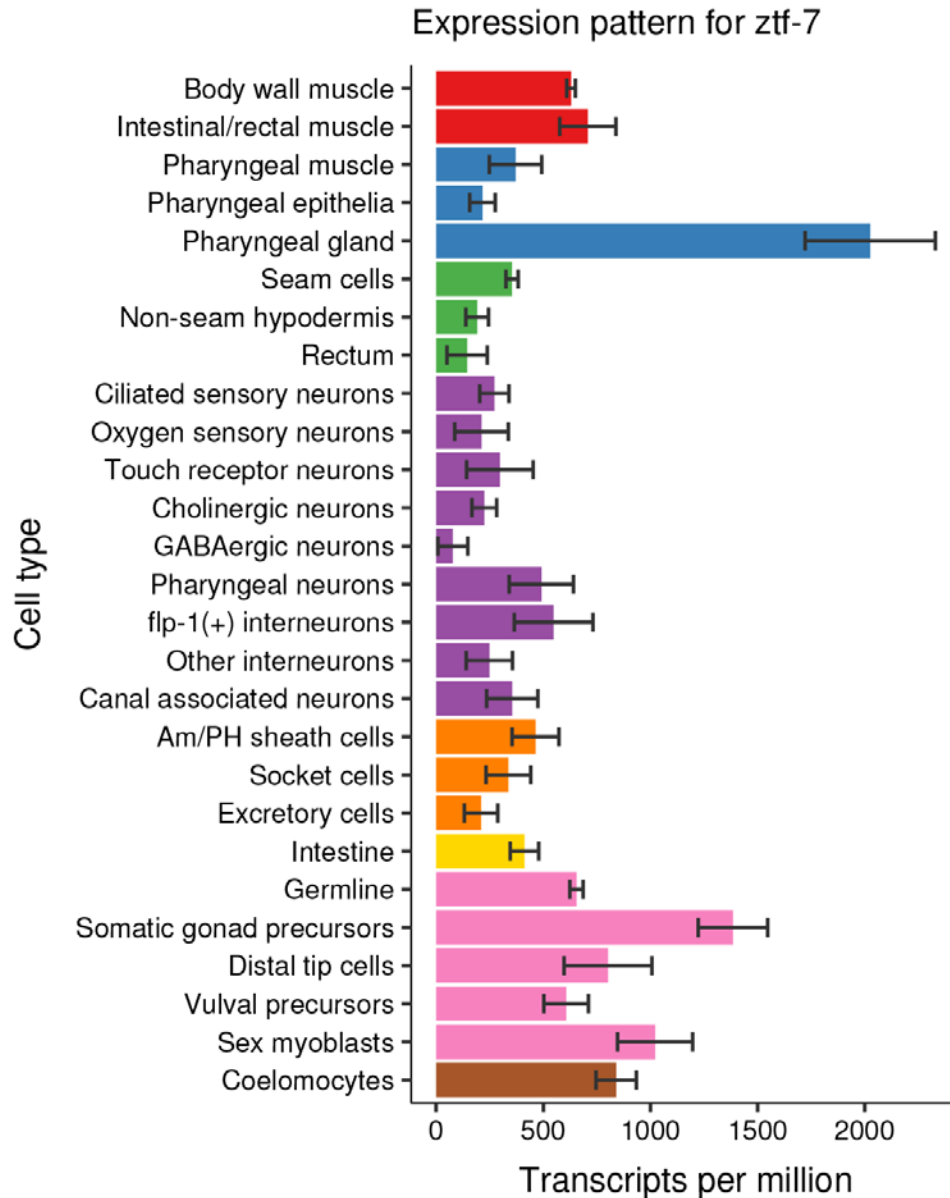


Fig. S14

Example of “gene expression report” image, with full set hosted on GExplore. For a given gene, mean expression values are shown for each of 27 cell types. Black bars indicate the 95% confidence interval. All gene profiles are viewable at:
http://genome.sfu.ca/gexplore/gexplore_search_tissues.html.

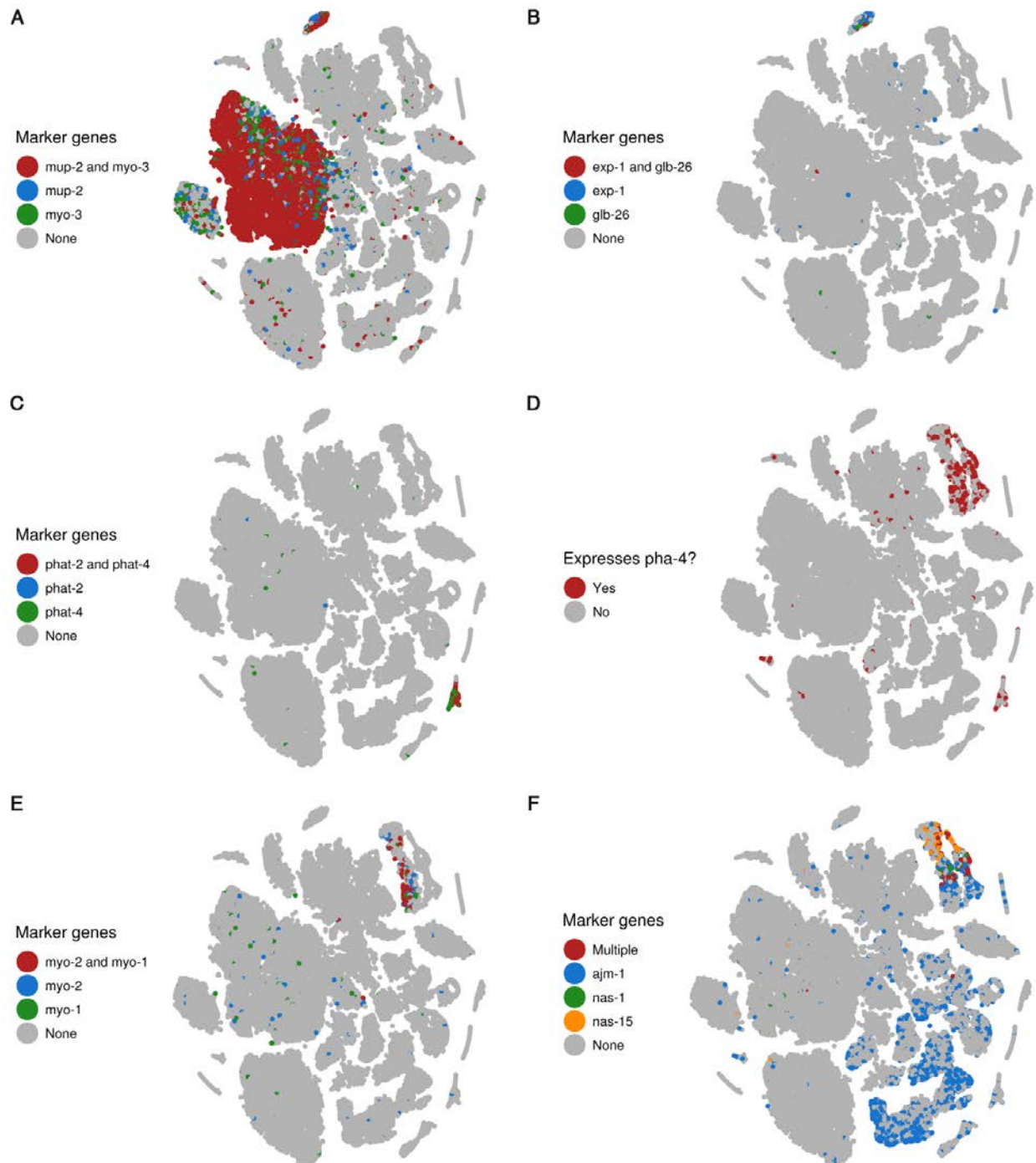


Fig. S15

Expression patterns of marker genes for body wall muscle, intestinal/rectal muscle, and

pharynx. (A) *mup-2* (troponin T) and *myo-3* (myosin heavy chain A) expression identifies body

1 wall muscle and intestinal/rectal muscle cells (84). The cluster to the left of the large muscle
2 cluster are low UMI-count cells that we believe to be damaged body wall muscle cells. They
3 were excluded from downstream analysis. **(B)** *exp-1* and *glb-26* expression distinguishes
4 intestinal/rectal muscle cells from body wall muscle (85, 86). **(C)** *phat-2* and *phat-4* expression
5 identifies pharyngeal gland cells (87). **(D)** *pha-4* expression identifies a cluster (top right) of non-
6 gland pharyngeal cells (48). The small *pha-4*(+) cluster on the left are distal tip cells (see fig.
7 S18B). **(E)** *myo-1* and *myo-2* expression identifies pharyngeal muscle cells (88). For the purpose
8 of constructing consensus expression profiles, cells in this t-SNE cluster were considered
9 pharyngeal muscle if they expressed at least two of *myo-1*, *myo-2*, *myo-5*, *tnt-4*, *mlc-1* or *mlc-2*.
10 **(F)** *ajm-1*, *nas-1*, and *nas-15* expression identifies non-muscle epithelial cells in the pharyngeal
11 t-SNE cluster. *ajm-1* is expressed in all epithelial cells, while *nas-1* and *nas-15* are specific to the
12 pharynx (89, 90). For the purpose of constructing consensus expression profiles, cells in the
13 pharyngeal muscle/epithelial t-SNE cluster were considered to be epithelial if they do not
14 express any of the markers listed in (E) and expressed at least one of *ajm-1*, *sma-1*, *nas-1*, *nas-*
15 *15*, or *ifa-1*.

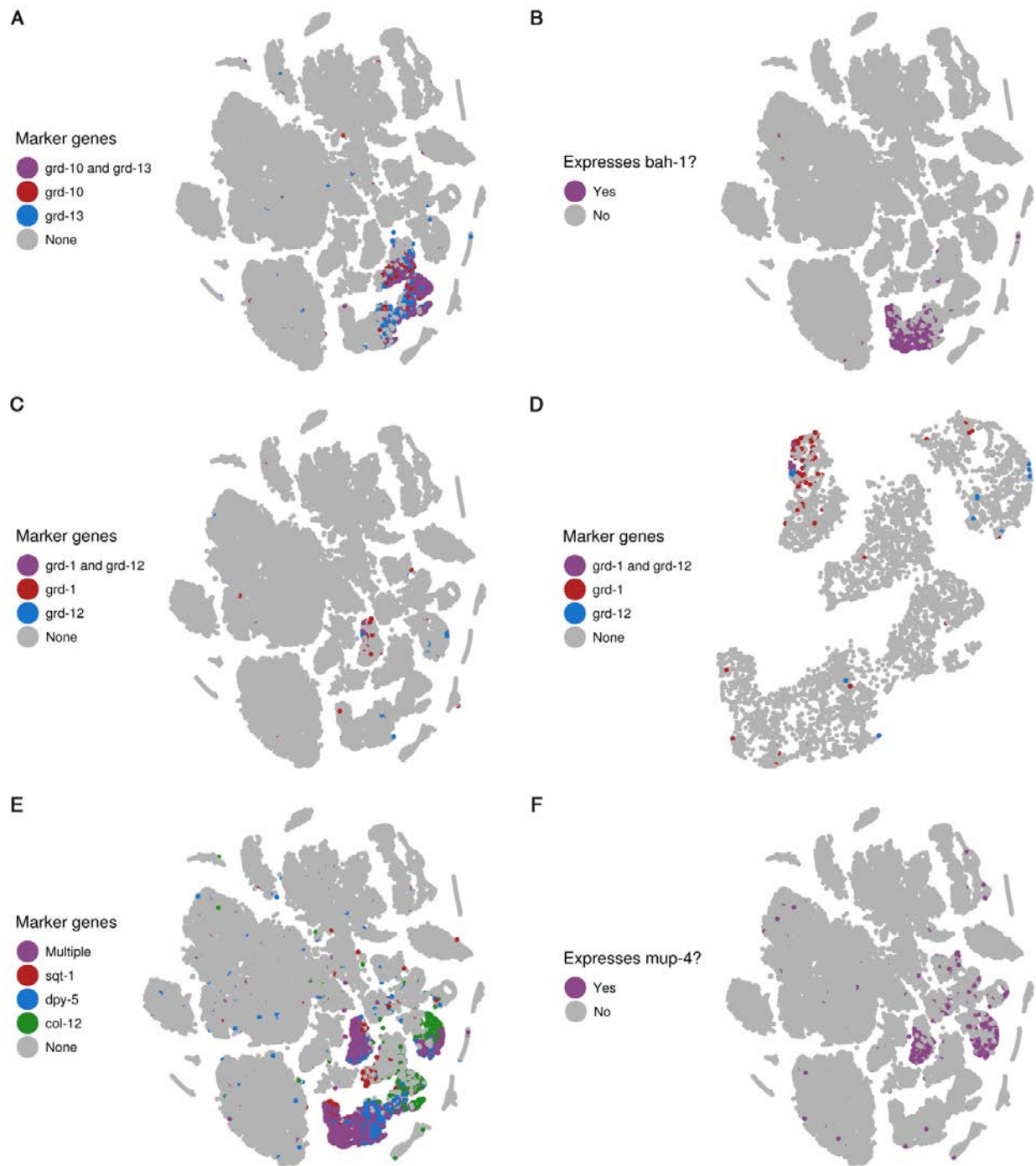


Fig. S16

Expression patterns for marker genes for hypodermis and the rectum. (A) *grd-10* and *grd-13* expression identifies seam cells (91). (B) *bah-1* expression identifies additional seam cells

1 (92) and shows that the t-SNE cluster with *grd-10/13* expression is likely to be entirely seam
2 cells. This cluster also expresses seam cell specific transcription factors including *ceh-18* and
3 *nhr-73*. (C to D) *grd-1* and *grd-12* expression identifies rectal cells. *grd-1* is expressed in the
4 rectal gland cells (93), while *grd-12* is expressed in the B and Y rectal epithelial cells (91) (D) is
5 a zoomed-in view of the hypodermal cell clusters in (C). E) Expression of the cuticle collagen
6 genes *sqt-1*, *dpy-5*, *col-12* identify hypodermal cells (94), including two clusters of non-seam
7 hypodermal cells. We were unable to clearly identify the anatomical differences between the
8 cells in the two non-seam hypodermal clusters. F) Expression of *mup-4* is exclusive to non-seam
9 hypodermis and glia, consistent with previous reports of its expression in the circumferential
10 rings of the cuticle (95).

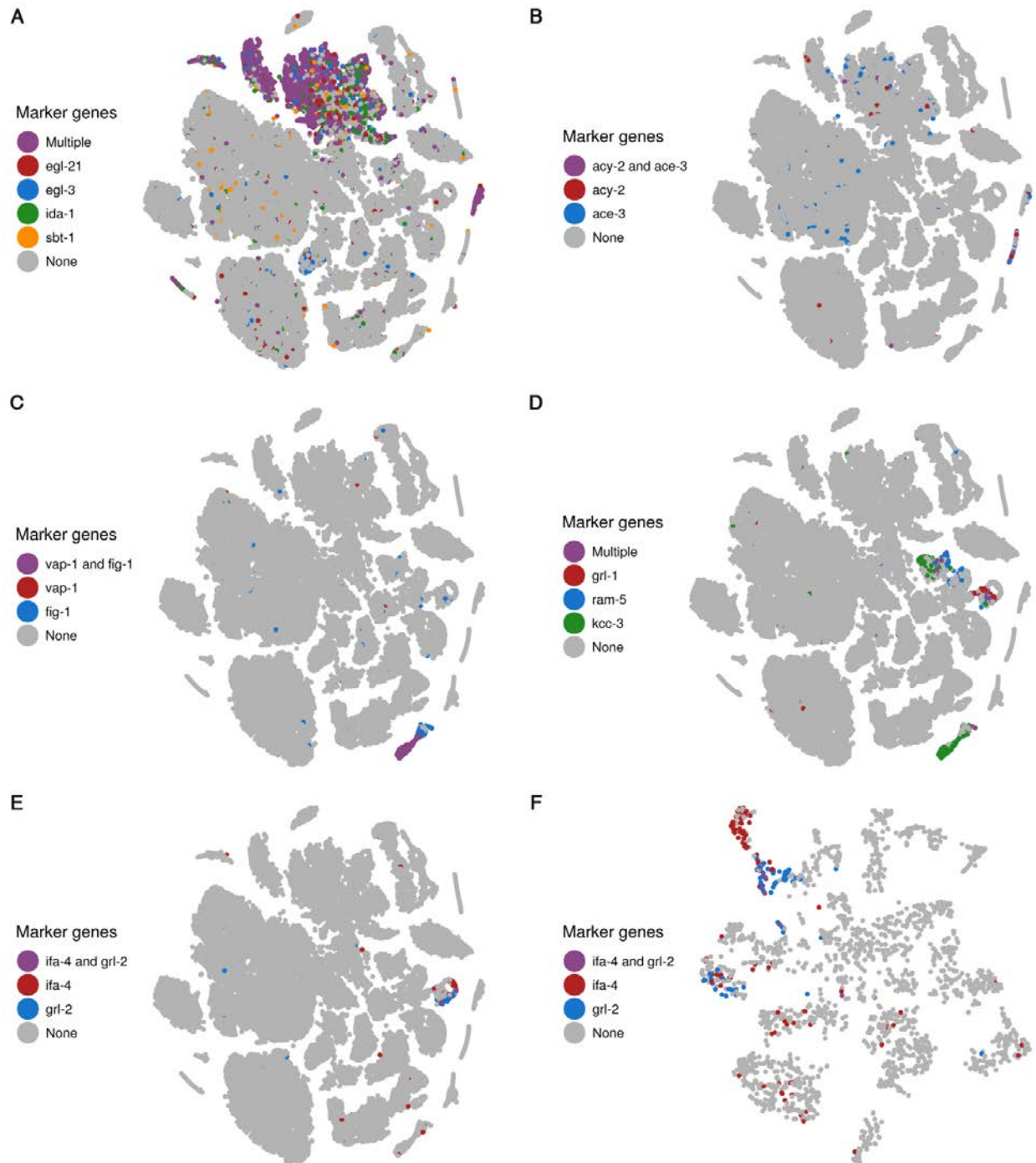


Fig. S17

Expression patterns of marker genes for neurons, glia, and excretory cells. (A) Expression of *egl-21*, *egl-3*, *ida-1*, and *sbt-1* identifies neuronal cells (96–99). (B) The canal associated

1 neurons do not express the marker genes listed in (A), but are identified by their expression of
2 *acy-2* and *ace-3* (100, 101). (C) Expression of *vap-1* and *fig-1* identifies the amphid and phasmid
3 sheath cells (79). (D) Expression of *grl-1* and *ram-5* identifies socket cells (91, 102). Expression
4 of *kcc-3* outside the amphid/phasmid sheath cell cluster identifies additional sheath cells (103).
5 For the purpose of constructing consensus expression profiles, cells in the non-amphid/phasmid-
6 sheath glial t-SNE clusters were considered to be socket cells if they were not identified to be
7 excretory cells, expressed at least one of *grl-1*, *grd-15*, *daf-6*, or *ram-5*, and did not express *kcc-*
8 3. (E) Expression of *ifa-4* and *grl-2* identifies excretory cells (91, 104). (F) *ifa-4(+)* and *grl-2(+)*
9 cells cluster together in a t-SNE of only cells from the glial/excretory cell clusters. We suspect
10 that the *ifa-4(+)* cluster at the top corresponds to the excretory canal cell, while the *grl-2(+)*
11 cluster corresponds to the excretory duct, pore, and/or gland cells.

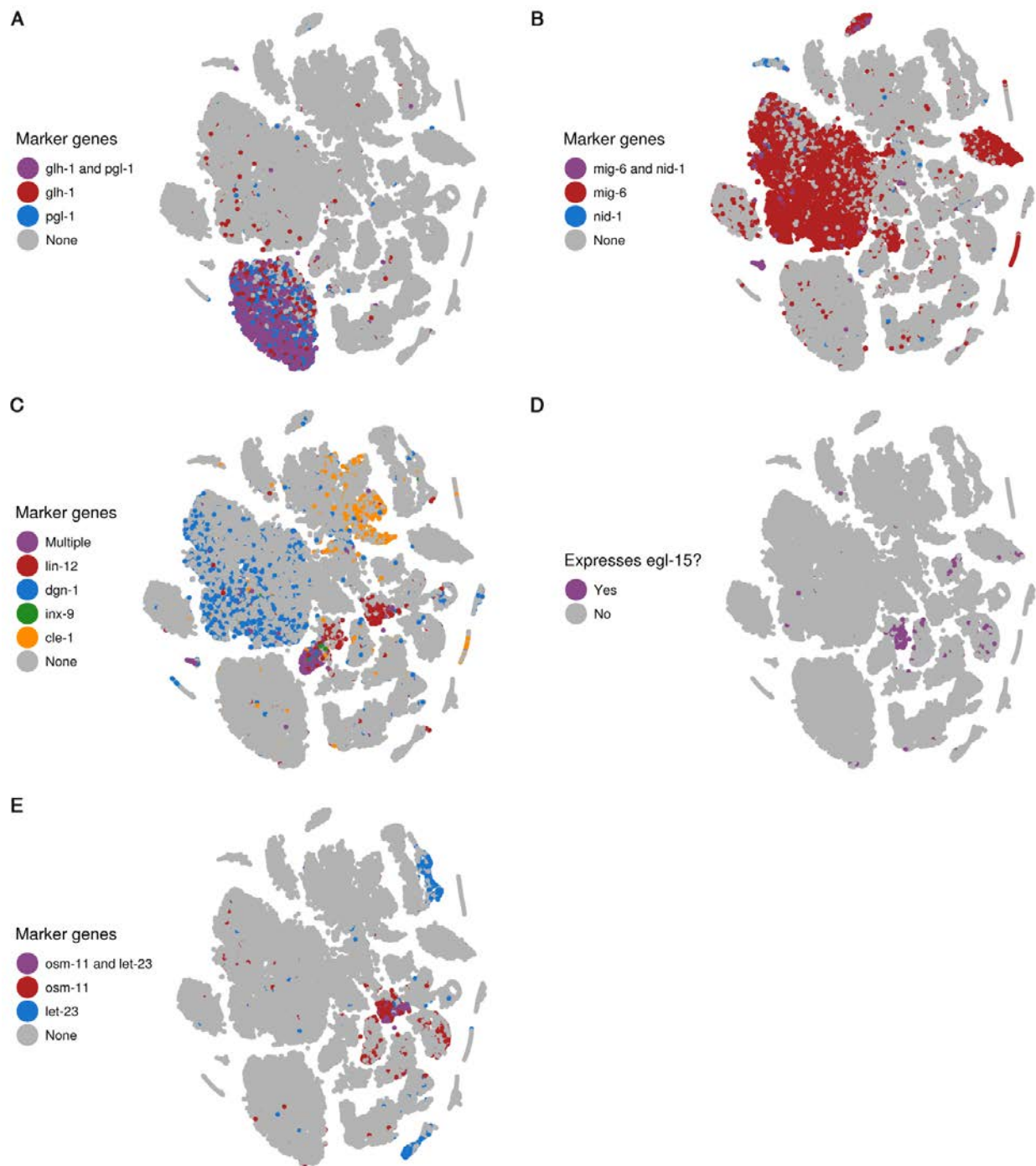
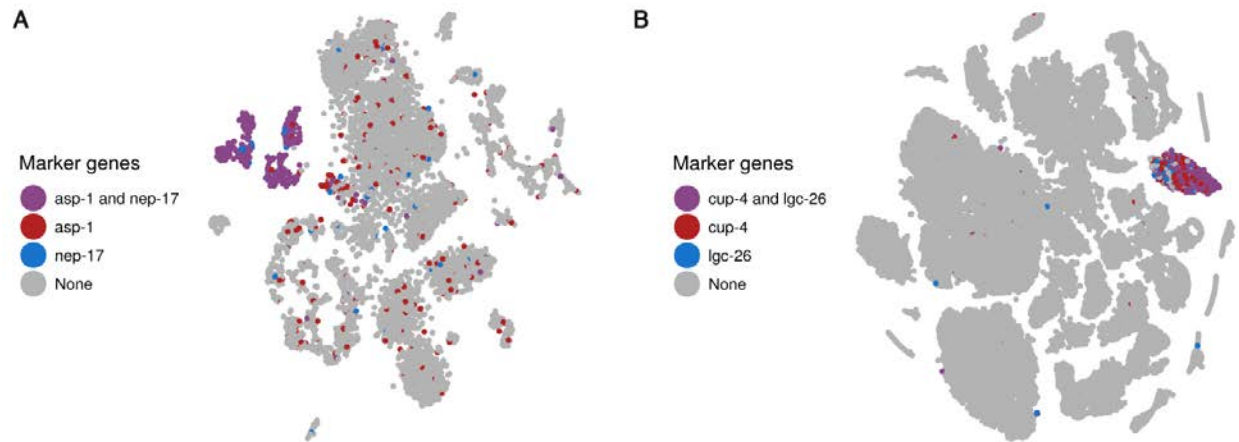


Fig. S18

Expression of marker genes for the germline, somatic gonad, and other sex-related tissues.

(A) Expression of *glh-1* and *pgl-1* identifies germline cells (105, 106). **(B)** Co-expression of *mig-*

1 6 and *nid-1* identifies the distal tip cells of the somatic gonad (small purple cluster on the lower
2 left; (107, 108)). (C) Co-expression of at least two of *lin-12*, *dgn-1*, *inx-9*, and *cle-1* identifies the
3 somatic gonad precursor cells (109–112). (D) Expression of *egl-15* identifies sex myoblasts
4 (113). (E) Expression of *osm-11* and *let-23* identifies vulval precursor cells (114, 115).



1

2 **Fig. S19**

3 **Expression of marker genes for the intestine and coelomocytes.** (A) Expression of *asp-1* and
 4 *nep-17* identifies intestine cells from the second *C. elegans* experiment (experiment 7 in Table
 5 S1). (116, 117). (B) Expression of *cup-4* and *lgc-26* identifies coelomocytes (118).

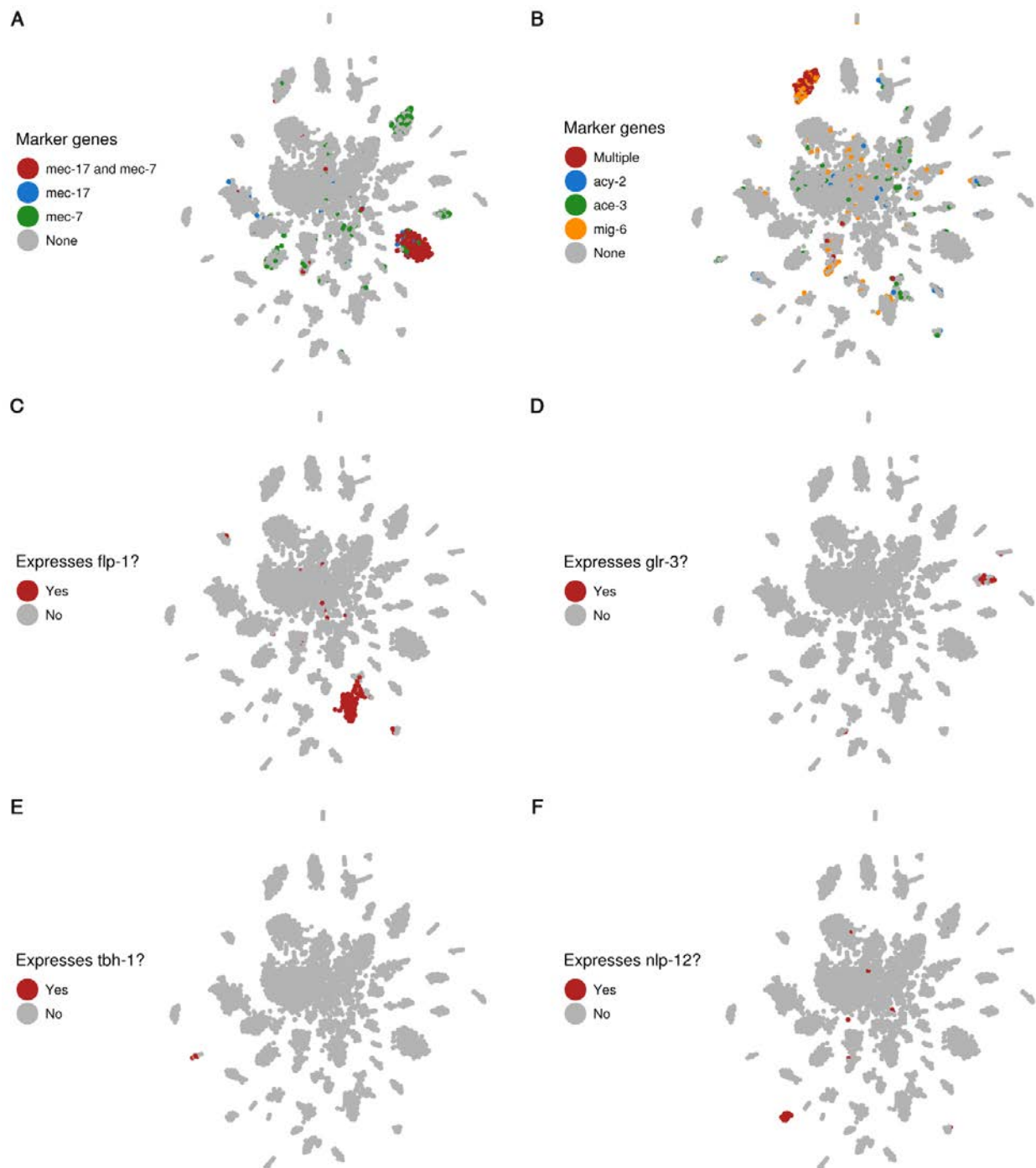


Fig. S20

Expression patterns of marker genes for touch receptor neurons and interneuron subtypes.

t-SNE plots shown are from a clustering of just neuronal cells (identified in fig. S17A,B). (A)

1 Expression of *mec-17* and *mec-7* identifies touch receptor neurons (119). **(B)** Expression of *acy-2*
2 and *ace-3* identifies canal associated neurons (100, 101). The canal associated neurons are also
3 the only neuron class that expresses *mig-6* (120). **(C)** *flp-1* expression identifies interneurons of
4 the anatomical classes AVK, AVA, AVE, RIG, RMG, AIY, AIA (121). *flp-1* has also been
5 reported to be expressed in the M5 pharyngeal motor neuron. **(D)** *glr-3* is expressed exclusively
6 in the RIA interneurons (122). **(E)** Among neurons, *tbh-1* is expressed exclusively in the RIC
7 interneurons (123). **(F)** *nlp-12* expression identifies the DVA tail interneuron (124).

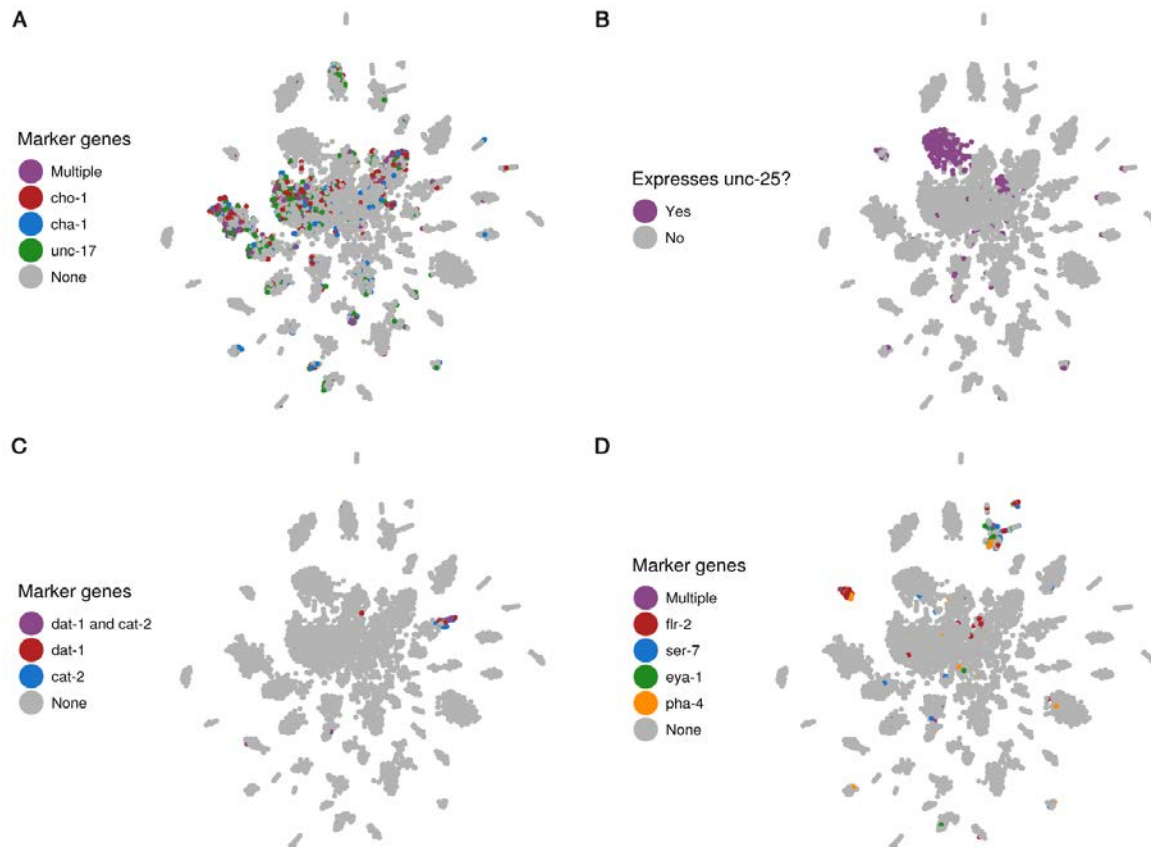


Fig. S21

Expression of marker genes for cholinergic, GABAergic, dopaminergic, and pharyngeal

neurons. t-SNE plots shown are from a clustering of just neuronal cells (identified in fig.

S17A,B). (A) Expression of *cho-1*, *cha-1*, and *unc-17* identifies cholinergic neurons (125). For

the purpose of constructing consensus expression profiles, neuronal cells were identified as

cholinergic if they were not part of a t-SNE cluster identified as any other neuronal subtype and

they expressed at least one of *cho-1*, *cha-1*, *unc-17*, *acr-15*, or *acr-18*. (B) *unc-25* expression

identifies GABAergic neurons (126). (C) Expression of *dat-1* and *cat-2* identifies dopaminergic

neurons (127, 128). (D) While no single marker is both highly expressed and specific to

pharyngeal neurons, the expression patterns of *flr-2*, *ser-7*, *eya-1*, and *pha-4* together identify

two clusters as highly likely to correspond to pharyngeal neurons (48, 129–131).

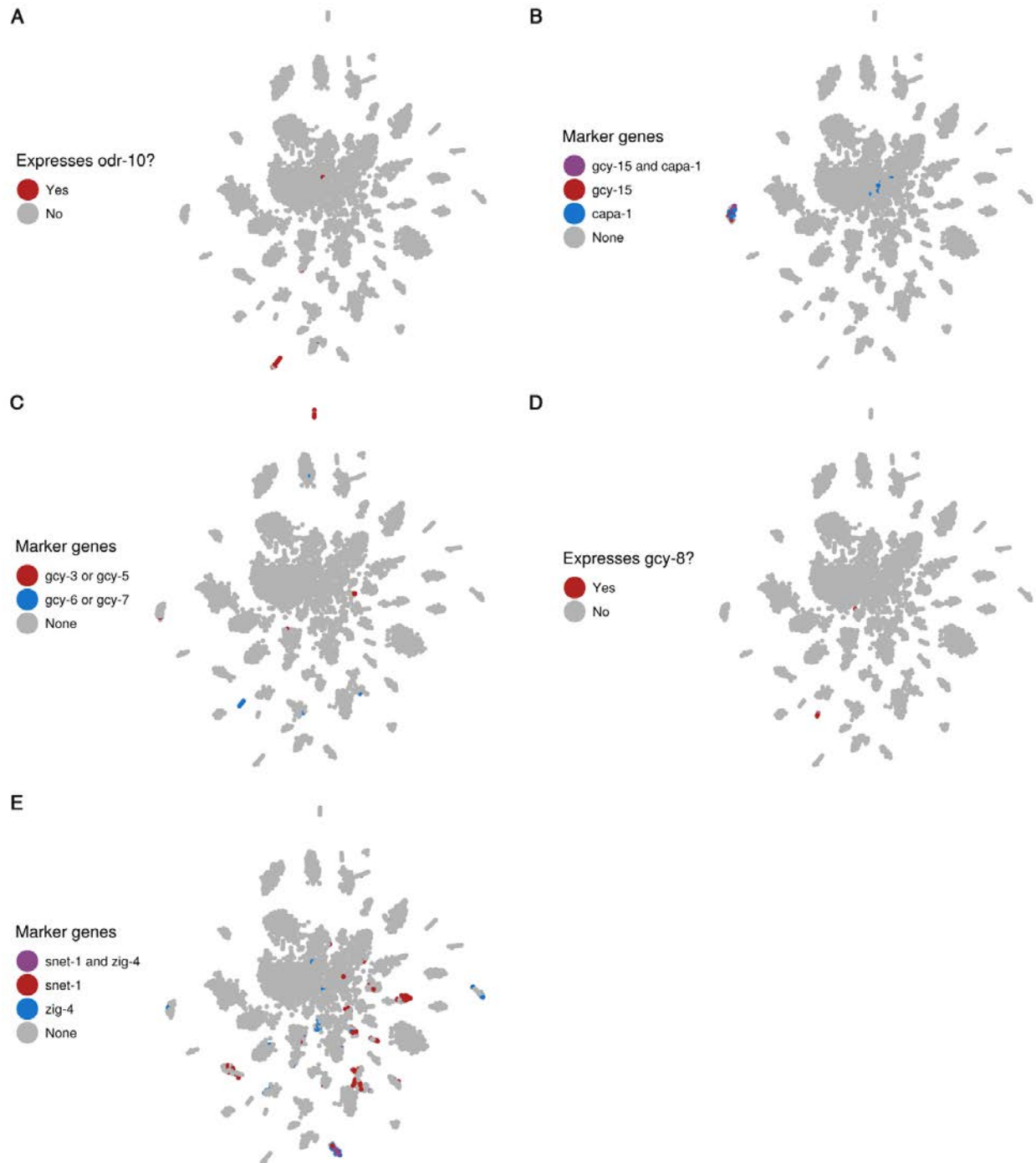


Fig. S22

Expression patterns of marker genes for the AWA, ASG, ASE, AFD, and ASK neurons. t-

SNE plots shown are from a clustering of just neuronal cells (identified in fig. S17A,B). (A) *odr-*

1 *10* expression identifies the AWA neurons (132). **(B)** *gcy-15* expression identifies the ASG
2 neurons (133). *capa-1* has also been reported to be expressed in two specific but unidentified
3 pairs of neurons in the head (134); in our data it is expressed predominantly in the same cluster
4 as *gcy-15*. **(C)** Expression of *gcy-3* and *gcy-5* identifies the ASER neuron, while expression of
5 *gcy-6* and *gcy-7* identifies the ASEL neuron (42, 43). **(D)** *gcy-8* expression identifies the AFD
6 neurons (135). **(E)** Co-expression of *snet-1* and *zig-4* identifies the ASK neurons (136, 137).

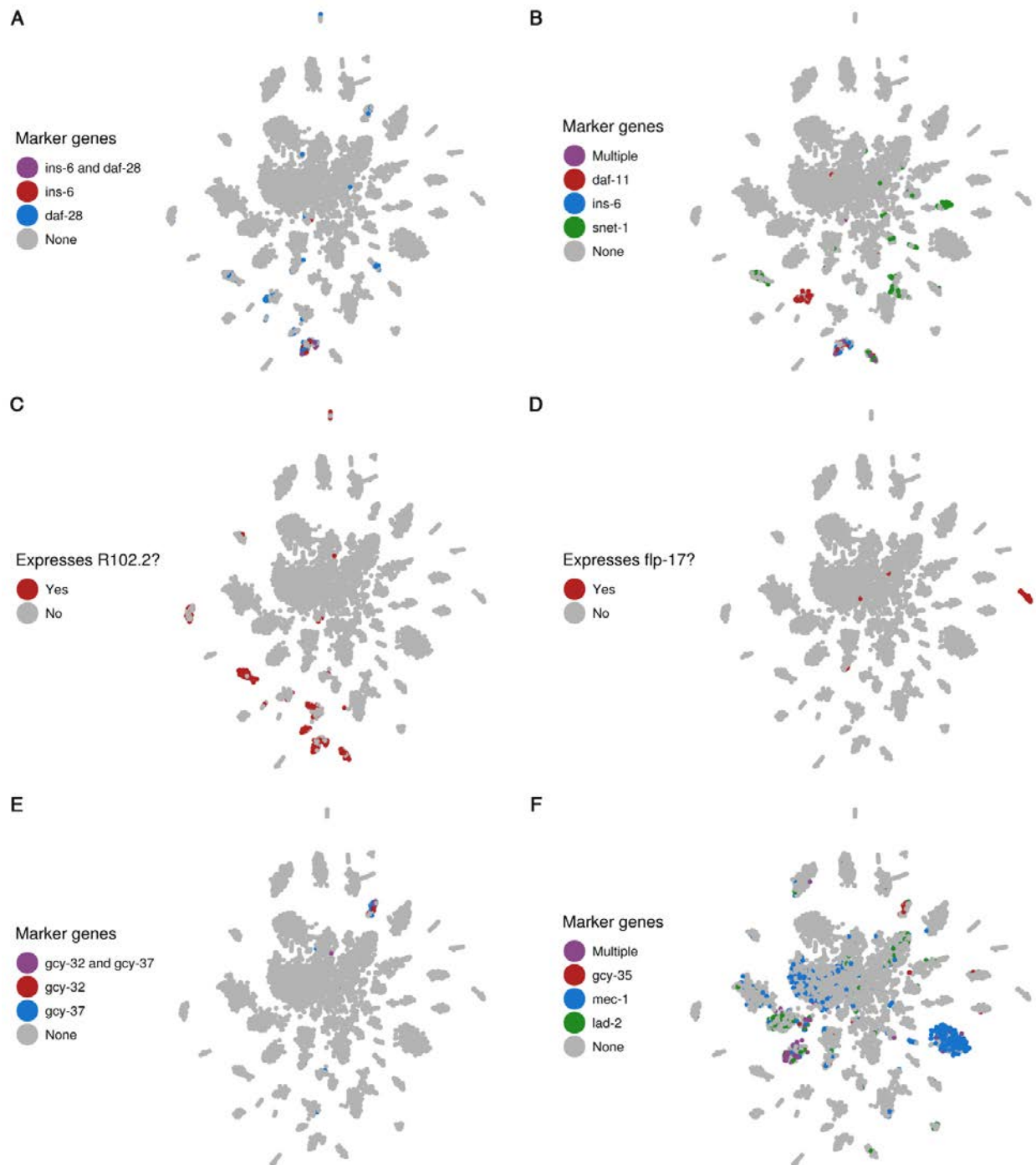


Fig. S23

Expression patterns of marker genes for ASI/ASJ, AWB/AWC, BAG, URX, SDQ, and other ciliated sensory neurons. t-SNE plots shown are from a clustering of just neuronal cells

(identified in fig. S17A,B). (A) Expression of *ins-6* and *daf-28* identifies a neuron cluster that consists of the ASI and ASJ neurons (138, 139). (B) Based on reported expression patterns, a neuron cluster that expresses *daf-11* but not *ins-6* or *snet-1* can only correspond to the AWB and/or AWC neurons (136, 138, 140). (C) Beyond those identified in fig. S22, and (A) of this figure, three additional neuron clusters express *R102.2*. Based on the expression patterns reported by (141), these clusters correspond to the ciliated sensory neurons classes ADF, ASH, PHA, and/or PHB. We could not precisely identify them however. For the purpose of constructing consensus expression profiles, neuronal cells were considered ciliated sensory neurons if they either were part of a cluster that was identified as a ciliated sensory neuron class or were part of a cluster that could not be conclusively identified but expressed high levels of *R102.2*, *dyf-2*, *che-3*, or *nphp-4*. (D) *flp-17* expression identifies the BAG neurons (121). (E) Expression of *gcy-32* and *gcy-37* identifies a neuron cluster that consists of the URX, AQR, and PQR neurons (142, 143). (F) Among neurons, *gcy-35* is expressed in the URX, AQR, PQR, SDQ, ALN, PLN O₂-sensory neurons, as well as the AVM and BDU neurons (143). *mec-1* was reported to be expressed in the touch receptor neurons, SDQ/ALN/PLN O₂-sensory neurons, and PVT neurons (144). *lad-2* was reported to be expressed in the SDQ/ALN/PLN O₂-sensory neurons and some sublateral motor neurons (145). Based on these expression patterns, a neuron cluster enriched for expression of all three of these genes is likely to correspond to the SDQ/ALN/PLN O₂-sensory neurons.

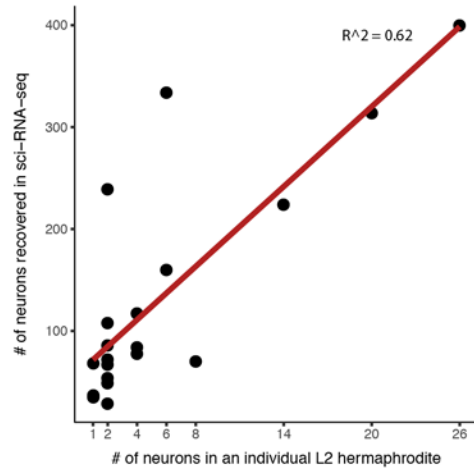


Fig. S24

Recovery rates of neuron types in sci-RNA-seq. The observed number of cells identified in sci-RNA-seq for a given neuron type (y axis) is compared to the number of neurons of that type in an individual L2 hermaphrodite *C. elegans* (x-axis). The plot includes all specific neuron types that we were able to identify, excluding cholinergic neurons, which were not limited to distinct t-SNE clusters and therefore may be under-counted as we only considered a cell cholinergic if we observed expression of at least one cholinergic marker gene (see **Fig. S21**). The neuron types included in the plot are: ASEL, ASER, DVA, AFD, ASG, ASK, AWA, BAG, CAN, RIA, RIC, ASI/ASJ, AWB/AWC, URX/AQR/PQR, SDQ/ALN/PLN, touch receptor neurons (ALM/PLM/AVM/PVM), dopaminergic neurons (CEP/ADE/PDE), flp-1(+) neurons (excluding the pharyngeal neuron M5), pharyngeal neurons, and GABAergic neurons.

Legends for Tables S1-13 (Tables will be found separated as an excel file)

Table S1. Summary of experiments

Table S2: Summary statistics for cell type consensus expression profiles constructed in this study.

Table S3: Tissue-level consensus expression profiles. Values listed are transcripts per million.

Table S4: Cell type consensus expression profiles. Values listed are transcripts per million.

Table S5: Neuron cluster consensus expression profiles. Values listed are transcripts per million.

Table S6: Differential expression test results for the identification of tissue-enriched genes.

See Methods. There is a row for each gene that is expressed in at least 10 cells in the analysis dataset. “Max tissue” is the tissue that the gene is expressed highest in. “Tissue 2” is the tissue the gene is expressed second highest in. “q-val” is the false detection rate at which the differential expression between the highest and second-highest expressing tissues can be called as non-zero.

Table S7: Differential expression test results for the identification of cell type enriched genes. See Methods. There is a row for each gene that is expressed in at least 10 cells in the analysis dataset. “Max cell type” is the cell type that the gene is expressed highest in. “Cell type 2” is the cell type the gene is expressed second highest in. “q-val” is the false detection rate at which the differential expression between the highest and second-highest expressing cell types can be called as non-zero.

Table S8: Differential expression test results for the identification of neuron cluster enriched genes. See Methods and Fig. 4. There is a row for each gene that is highly enriched (>5-fold) in neurons relative to other tissues (as reported in **Table S6**). “Max cluster” is the neuron cluster that the gene is expressed highest in. “Cluster 2” is the neuron cluster the gene is expressed second highest in. “q-val” is the false detection rate at which the differential expression between the highest and second-highest expressing neuron clusters can be called as non-zero.

Table S9: Differential expression test results for anterior vs. posterior body wall muscle. “moderated log₂(anterior / posterior)” is equal to log₂(anterior TPM+1) - log₂(posterior TPM+1).

Table S10: Differential expression test results for posterior vs. other intestine. “moderated log₂” is defined as in **Table S9**.

Table S11: Differential expression test results for amphid vs. phasmid sheath cells. “moderated log₂” is defined as in **Table S9**.

- 1 **Table S12: Differential expression test results for the ASEL vs. ASER neuron.** “moderated
- 2 log2” is defined as in **Table S9.**
- 3 **Table S13: Differential expression test results for AWA vs. ASG neurons.** “moderated log2”
- 4 is defined as in **Table S9.**
- 5 **Table S14: List of genes used in heatmaps in Fig. 3F and Fig. 4C.**
- 6