Integrating bulk and single cell RNA-seq refines transcriptomic profiles of individual *C. elegans* neurons.

Author list: Alec Barrett[1], Erdem Varol[2], Alexis Weinreb[1,3], Seth R. Taylor[4], Rebecca M. McWhirter[4], Cyril Cros[5], Berta Vidal[5], Manasa Basaravaju[1,3], Abigail Poff[4], John A. Tipps[4], Maryam Majeed[5], Chen Wang[5], Emily A. Bayer[5], Molly Reilly[5], Eviatar Yemini (ORCID: 0000-0003-1977-0761) [5,6], HaoSheng Sun[5,7], Oliver Hobert (ORCID: 0000-0002-7634-2854) [5,8]*, David M. Miller III[4,8]*, Marc Hammarlund[1,3]*

[1] Department of Genetics, Yale University School of Medicine, New Haven, CT, USA

[2] Department of Statistics, Columbia University, New York, NY, USA

[3] Department of Neuroscience, Yale University School of Medicine, New Haven, CT, USA

[4] Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville, TN, USA

[5] Department of Biological Sciences, Howard Hughes Medical Institute Columbia University, New York, NY, USA

[6] Present Address: Department of Neurobiology, University of Massachusetts Chan Medical School, Worcester, MA, USA

[7] Present Address: Department of Cell, Developmental, and Integrative Biology, University of Alabama Birmingham, AL, USA

[8] Program in Neuroscience, Vanderbilt University School of Medicine, Nashville, TN, USA

* Corresponding Authors

Marc Hammarlund: marc.hammarlund@yale.edu

David M. Miller, III: david.miller@vanderbilt.edu

Oliver Hobert: or38@columbia.edu

Abstract

Neuron-specific morphology and function are fundamentally tied to differences in gene expression across the nervous system. We previously generated a single cell RNA-seq (scRNA-Seq) dataset for every anatomical neuron class in the *C. elegans* hermaphrodite. Here we present a complementary set of bulk RNA-seq samples for 52 of the 118 canonical neuron classes in *C. elegans*. We show that the bulk RNA-seq dataset captures both lowly expressed and noncoding RNAs that are not detected in the scRNA-Seq profile, but also includes false positives due to contamination by other cell types. We present an analytical strategy that integrates the two datasets, preserving both the specificity of scRNA-Seq data and the sensitivity of bulk RNA-Seq. We show that this integrated dataset enhances the sensitivity and accuracy of transcript detection and differential gene analysis. In addition, we show that the bulk RNA-Seq data set detects differentially expressed non-coding RNAs across neuron types, including multiple families of non-polyadenylated transcripts. We propose that our approach provides a new strategy for interrogating gene expression by bridging the gap between bulk and single cell methodologies for transcriptomic studies. We suggest that these datasets advance the goal of delineating the mechanisms that define morphology and connectivity in the nervous system.

Introduction

Neurons exhibit an extraordinary range of morphological forms and physiological functions. Because this diversity is largely driven by underlying differences in gene expression, a key goal of neuroscience is to identify the transcripts expressed in each neuron type.

The adult *C. elegans* hermaphrodite contains 302 neurons divided into 118 anatomically distinct neuron types. The structure, connectivity, and lineage are known for each of these neurons(1-6). Recently, the *C. elegans* Neuronal Gene Expression Map & Network project (CeNGEN)(7) used single cell RNA sequencing (scRNA-seq) technology to generate a gene expression atlas that matches the single neuron resolution of the structural map of the mature *C. elegans* nervous system(8).

The CeNGEN scRNA-seq dataset was acquired with 10x Genomics technology and is largely comprised of reads from poly-adenylated transcripts. Thus, major classes of non-poly-adenylated transcripts, especially noncoding RNAs, are under-represented in the CeNGEN scRNA-seq data. In addition, low abundance transcripts may be sparsely detected in scRNA-seq data, particularly in clusters with relatively few cells(8). Both noncoding RNAs and low abundance transcripts are potentially important mediators of neuronal fate. A description of their expression is therefore needed to complement the CeNGEN scRNA-seq map of neuronal poly-adenylated transcripts.

Here, we use FACS to isolate single neuron types for bulk RNA sequencing with the goal of describing neuronal gene expression with high sensitivity and specificity. We generated profiles for 53 individual neuron types from the mature *C. elegans* hermaphrodite nervous system. These 53 types include 52 of the canonical 118 types, with separate data for the subclasses ASEL and ASER. This data set samples a wide range of neuron types including motor neurons, interneurons, and sensory neurons. We built sequencing libraries with random primers for robust detection of both poly-adenylated and non-coding RNAs(9). We developed computational approaches that exploit the bulk dataset to enhance accurate detection of gene expression, including integration with the existing CeNGEN scRNA-seq dataset and methods for measuring non-coding RNAs. The resulting data sets refine quantitative measures of gene expression and improve detection of low-abundance and non-poly-adenylated transcripts. These data provide a unique opportunity for future studies that link gene expression to neuron function, structure, and connectivity.

Methods

Strains

Strains used for FACS isolation of individual neuron classes are listed in Table 1. Strains were either generated by us or are kind gifts of the *C. elegans* community. The lab origin of individual transgenic arrays can be inferred from the allele designation, as detailed at Lab List at the CGC (https://cgc.umn.edu/laboratories). All strains used in this study are available at the CGC.

FACS isolation for RNA-seq

Labeled neuron types were isolated for RNA-seq as previously described(10-12). Populations of synchronized L1s were grown at 23 C until reaching the L4 stage on 150 mM 8P plates inoculated with Na22. The time in culture to reach the L4 stage varied (40.5-49 h) and was determined for each strain. 50-100 animals were inspected with a 40X DIC objective to determine developmental stage as scored by vulval morphology (Mok et al., 2025). Cultures were predominantly composed of L4 larvae but also typically included varying fractions of L3 larvae and adults.Cells were dissociated from a sample of ~300 ul of packed L4 larvae and labeled neuron types isolated by Fluorescence Activated Cell Sorting (FACS) on a BD FACSAria III equipped with a 70-micron diameter nozzle. DAPI was added to the sample (final concentration of 1 mg/mL) to label dead and dying cells. For bulk RNA-sequencing of individual cell types, sorted cells were collected directly into TRIzol LS. At ~15-minute intervals during the sort, the sort was paused, and the collection tube with TRIzol was inverted 3-4 times to ensure mixing. Cells in TRIzol LS were stored at -80C for RNA extractions. See Table S1 for list of samples and associated metadata.

RNA extraction

RNA extractions were performed as previously described(8, 12). Briefly, cell suspensions in TRIzol LS (stored at -80°C) were thawed at room temperature. Chloroform extraction was performed using Phase Lock Gel-Heavy tubes (Quantabio) according to the manufacturer's protocol. The aqueous layer from the chloroform extraction was combined with an equal volume of 100% ethanol and transferred to a Zymo-Spin IC column (Zymo Research). Columns were centrifuged for 30 s at 16,000 RCF, washed with 400 mL of Zymo RNA Prep Buffer, and centrifuged for 16,000 RCF for 30 s. Columns were washed twice with Zymo RNA Wash Buffer (700 mL, centrifuged for 30 s, followed by 400 mL, centrifuged for 2 minutes). RNA was eluted by adding 15 mL of DNase/RNase-Free water to the column filter and centrifuging for 30 s. A 2 µL aliquot was submitted for analysis using the Agilent 2100 Bioanalyzer Picochip to estimate yield and RNA integrity, and the remainder was stored at -80°C. See Table S1 for list of samples and associated metadata.

Bulk sequencing and mapping

Each bulk RNA sample was processed for sequencing using the SoLo Ovation Ultra-Low Input RNaseq kit from Tecan Genomics according to manufacturer instruction, modified to optimize rRNA depletion for *C. elegans*(9). Libraries were sequenced on the Illumina Hiseq 2500 or Novaseq6000 with 150 bp paired end reads. Reads were mapped to the *C. elegans* reference genome from WormBase (version WS289) using STAR (version2.7.7a) with the option --outFilterMatchNminOverLread 0.3. Duplicate reads were removed using UMI-tools (v1.1.4), and a counts matrix was generated using the featureCounts tool of SubRead (v2.0.3). FASTQC was used for quality control before alignment, and four samples were removed for failing QC or for a low number of reads. Samples were excluded after sequencing and mapping if they had: fewer than 1 million read pairs, <1% of uniquely mapping reads to the *C. elegans* genome, > 50% duplicate reads (low UMI diversity), or failed deduplication steps in the nudup package.

Each biological sample contains one or more transgenes, and each transgene generally consists of many copies of a promoter sequence driving expression of a fluorophore (Table 1). To avoid artifacts stemming from spurious transcription of high-copy number transgenes, we masked 5kb of all genes whose promoters were used in transgenes, starting 4kb upstream of the start ATG and extending 1kb past the ATG. This approach removed 231 genes from our analysis (Table S2).

Sample Normalization

Intra-sample normalization (gene length normalization for bulk samples) was performed before integration (see below). Inter-sample normalization (library size normalization) was performed after integration. Library size normalizations were performed using a TMM (trimmed mean of M-values) correction in edgeR (version 4.0.1).

Ground-truth genes

As an independent measure of gene expression, we used a “ground truth” dataset of 160 genes for which expression in individual neuron types is known with high precision across the entire nervous system (Table S3). This expression matrix is based on high confidence fosmid fluorescent reporters, CRISPR strains or other methods(8, 13-17).

We also curated a list of 445 genes that are exclusively expressed outside the nervous system to assess potential non-neuronal contamination in each sample (Table S4). This list was curated from published datasets of fluorescent reporters, tissue specific RT-PCR, and transcriptomic studies available on WormBase(15). Genes were included if two types of evidence documented expression in the same non-neuronal tissue (non-identical expression in other tissues was allowed so long as at least one tissue overlapped), and for which there was no evidence for neuronal expression.

An additional 936 ubiquitously expressed genes were used to evaluate gene expression accuracy in the bulk RNA-Seq samples (Table S5). To compile this list, we used the scRNA-Seq data(11), and defined the expression level of a gene in a given cluster as the proportion of cells in that cluster in which at least one UMI was detected. We then defined a gene as ubiquitous if at least 10 clusters displayed an expression level of $\geq$0.5%, excluding neuron-specific genes.

Comparing datasets to ground-truth

When comparing data to “ground truth” gene expression, a static threshold was applied to the average normalized cell profile (i.e., arithmetic mean across all cells, or samples). Single cells were normalized to library size prior to averaging to calculate TPM counts(18). Bulk samples were normalized with the TMM method using edgeR (version 4.0.1), noncoding RNA analysis was performed with length normalized TMM counts using the GeTMM method(19).

LittleBites Subtraction algorithm

LittleBites is an iterative algorithm for bulk RNA-seq datasets. LittleBites improves the accuracy of cell-type specific bulk RNA-seq sample profiles by removing counts from non-target contaminants (e.g. ambient RNA from dead cells, carry-over non-target cells from FACS enrichment due to imperfect gating). This method leverages single cell reference datasets and ground truth expression information to guide iterative and conservative subtraction to enrich for true target cell-type expression. LittleBites optimizes subtraction by using both a single-cell reference and an orthogonal ground truth reference, moderating potential biases in either reference.

This algorithm first calculates gene level specificity weights in a single cell reference dataset using SPM (Specificity Preservation Method) (22, 23). SPM assigns high weights (approaching 1) to genes expressed in single cell types while applying conservative weights to genes with broader expression patterns, which helps to reduce inappropriate subtraction.

The algorithm proceeds in a loop of three steps.

<u>Step 1: Estimate Contamination.</u> Each bulk sample is modeled as the sum of a linear combination of single-cell profiles (target cell type and likely contaminants) using non-negative least squares (NNLS). The resulting coefficients provide an estimate of how many of the sample’s counts come from the target cell type, and how many come from each contaminant cell type.

<u>Step 2: Weighted Subtraction.</u> Each bulk sample is cleaned by subtracting the weighted sum of contaminant single-cell profiles. This subtraction is attempted multiple times (separately) across a series of learning rate weights (usually ranging from 0-1) which moderate the size of the subtraction step (Equation 1). This produces a range of possible “cleaned” sample options for evaluation.

<u>Step 3: Performance Optimization.</u> For each learning rate, the cleaned result is evaluated against a set of ground truth genes by calculating the area under the receiver operating characteristic curve (AUROC). The learning rate that optimizes the AUROC is then selected. When multiple learning rates yield equivalent AUROC values, the lowest learning rate value is chosen to minimize subtraction.

<u>If the optimal learning rate at Step 3 is 0</u> (no subtraction option beats the baseline) then the loop is halted. Else, the cleaned bulk profile is returned to Step 1, and the loop

continues until the AUROC cannot be improved upon using the single-cell reference modeling.

$$Equation(1)\hat{B}_i = B_i - L * \sum A_j * S_j$$

B = bulk sample, A = estimated composition coefficient, S = Single cell reference, L = learning rate, i = sample, j = cell-type.

Subtraction

bMIND(20) was run by first excluding genes that are invariant in all cell types. Normalized aggregate tissue level scRNA-seq profiles for each cell type were used as the prior reference. Cell-type proportions were estimated using the NNLS option.

ENIGMA(21) was run by first excluding genes that are invariant in all cell types. Normalized aggregate tissue-replicate level scRNA-seq profiles for each cell-type were used as the reference, meaning that each tissue was represented in multiple columns, with different columns for each experimental replicate in the scRNA-seq dataset. Cell-type proportions were estimated using the robust linear regression (RLR) option. ENIGMA was run using the L2 norm, with a log transformation preprocessing step. Trace norm runs did not improve performance on ground truth gene detection (data not shown).

LittleBites (this work) was run by first calculating gene level specificity weights in a single cell reference dataset using the SPM method(22, 23). The SPM method weights genes highly (values close to 1) only if they are present exclusively in one cell type, and the method is robust to different numbers of cell types. SPM specificity scores are conservative, and result in less subtraction for any gene that is expressed even in just a few cell types. The LittleBites algorithm runs in 4 steps: 1) Model each bulk sample as a linear combination of single cell reference profiles of target and putative contaminant tissues, using a non-negative least squares approach. The coefficients are used as estimates of the bulk sample composition. 2) Subtract the non-target single cell reference profiles from the bulk sample, using the modeled coefficients as weights for each corresponding profile (Equation 1). For each subtraction, set a range of learning rates from 0 to 1, and perform the subtraction for each learning rate supplied. 3) For each learning rate, compare the resulting subtracted values against known ground truth genes, and calculate the AUROC, and return the subtracted sample that maximizes the AUROC value. If more than one learning rate returns the same AUROC value, return the subtracted dataset with the lowest learning rate. 4) If there is no learning rate that produces a higher AUROC than a learning rate of 0, halt subtraction. Else, return to step 1 using the subtracted bulk sample as the input.

LittleBites (this work) was run using cell-type level aggregate data from the scRNA-seq dataset were as a reference, with just the neuron of interest (e.g. ADL) and putative non-neuronal comtaminants (e.g. muscle, intestine, germline) used to model each sample. Ground truth genes were taken from a list of ubiquitously expressed genes, and

non-neuronal genes (see “Ground Truth Genes” methods second above). LittleBites was run with a log transformation preprocessing step for bulk and single cell references. LittleBites code is available at https://github.com/alecbarrett/LittleBites.

Bulk sample composition estimates

Contamination estimates were performed for each bulk sample by using non-negative least squares (NNLS) modeling on down-sampled and log transformed counts, averaging across 100 estimates per sample. Down-sampling was performed to reduce bias against neuron types with small cluster sizes. For each bulk sample (ex: AFD replicate 1), proportions were estimated using only neuronal cells for the corresponding single cell cluster (ex: AFD), and 7 identified non-neuronal clusters (Glia, Excretory, Hypodermis, Intestine, Muscle-mesoderm, Pharynx, and Reproductive). For each iteration, all 8 single cell clusters were down-sampled to 30 cells each, and average TPM counts were calculated using the arithmetic mean for each gene in the 30 cells. Gene level variance was calculated using the averaged TPM values, and low variance genes were removed. Bulk sample counts and single cell TPMs were log transformed before the NNLS calculation with a pseudocount of one. NNLS estimates across all 100 iterations were averaged for the final estimate. NNLS calculations were performed using the nnls package in R (version 1.5).

Correlating gene expression to non-neuronal contaminants

Each gene was correlated to non-neuronal contamination across all samples using Spearman’s correlation test. High correlation to any contaminant was used to indicate that the gene is likely detected because of contamination, and not due to expression in the target neuron. For genes passing an expression threshold > 2 normalized counts in at least 2 samples, their highest correlation value to any contaminant tissue was collected, and cutoffs were determined by fitting a gaussian mixture model using the normalmixEM2comp function in mixtools (version 2.0.0), fitting 2 gaussian distributions to the distribution of highest contaminant correlations. Cutoffs were selected to exclude 95% of the predicted contaminant distribution.

Defining new gene expression for lowly expressed genes and for noncoding genes

For protein coding genes that are under the detection threshold for specific single cell clusters, but are detectable elsewhere in the single cell dataset, we identified expanded gene expression profiles using the Integrated dataset. These genes were thresholded to match the FDR (14%) for the published single cell analysis(8). For protein coding genes that are never detected in the single cell dataset and for noncoding RNAs, genes with high correlation to non-neuronal contaminant estimates were removed, and for the remaining genes the threshold was set at 5 normalized counts, and genes were called expressed if they passed that threshold in >65% of samples.

Pseudobulk aggregation of single-cell data

We downloaded the CeNGEN scRNA-seq dataset as a Seurat object from the CeNGEN website (www.cengen.org). Cells from the same cell type and experiment (e.g. AFD cluster) were aggregated together by summation into a single pseudobulk sample. For this work, single cell clusters of neuron subtypes were collapsed to the resolution of the bulk replicates (example: VB and VB1 clusters in the single cell data were treated as one VB cluster). Pseudobulk samples from cluster-replicates with fewer than 10 cells were excluded.

Modeling Single Cell cluster proportions and counts

To simulate a single cell pseudobulk-cluster, we sampled from a negative binomial distribution for each gene using the rnbinom function from the stats package in R (Version 4.3.2). We then modeled the number of counts per gene as the sum of the log odds-ratio and the log of the total cells in the simulated cluster (equation 2). For a given cluster k and gene g, $C_{k,g}$ = total counts for the gene in the cluster, $P_{k,g}$ = proportion of cells with at least 1 count for a given gene, and n_k = the total cells in the cluster

$$Equation(2): log\left(C_{k,g}\right) = log\left(\frac{P_{k,g}}{1 - P_{k,g}}\right) + log\ (n_k)$$

We applied this formula to our real single cell dataset and used this equation to transform proportion measures of gene expression into a count space to generate the Prop2Count dataset for downstream analysis and integration with bulk datasets. This procedure allows for proportions data to be used in downstream analyses that work with counts datasets. However, it does limit the range of potential values that each gene can have, with the potential values set as:

$$p \in \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\right\}$$

As n approaches 0, the number of potential values decreases, which can be incompatible with some downstream models. Thus, caution should be used when applying this transformation to datasets with few cells.

Prop2Count code is available at https://github.com/alecbarrett/Prop2Count.

Results

Bulk sequencing of individual neuron types

The model organism *C. elegans* is uniquely suitable for the task of defining gene expression in the nervous system at high resolution and genome scale. *C. elegans* is the first metazoan with a completely sequenced genome(24) and the only animal for which we know every cell division that gives rise to the adult body plan (*i.e.,* "cell lineage")(1, 2), as well as the anatomy of each neuron and all of its connections with other cells(3-6, 25). The entire *C. elegans* hermaphrodite nervous system contains 302 neurons with 118 canonical anatomically-defined neuron classes, each comprised of relatively few cells, ranging from 1 to 13 neurons(3). Most of these neuron classes are either a bilateral pair of anatomically similar cells (70 classes) or single neurons (26 classes) with unique morphological and functional characteristics. The rich array of distinct neuron classes in *C. elegans,* combined with the fact that these types are invariant among individuals, means that each neuron class can be analyzed in depth to reveal the genetic programs that define neuronal diversity.

We previously generated a gene expression atlas for the entire *C. elegans* nervous system at the resolution of single neuron types(8). We completed this atlas with single-cell techniques by adopting the strategy of using FACS to enrich for specific groups of neurons for a series of scRNA-seq experiments. The single-cell atlas provides a detailed description of gene expression across the *C. elegans* nervous system. However, the single-cell atlas may fail to detect lowly-expressed genes, particularly in clusters with few cells. Also, due to its reliance on poly-dT priming for reverse transcription, the single-cell atlas excludes non-coding transcripts that are not poly adenylated(11). Finally, because the scRNA-seq data is biased toward the 3' ends of transcripts, it does not contain information about most splicing events(26, 27).

To address these limitations and to provide a broader description of gene expression across the nervous system, we developed a bulk RNA sequencing strategy to profile individual neuron types(10). We hypothesized that bulk RNA-seq and scRNA-seq datasets might have complementary strengths and weaknesses. Bulk RNA-seq can enhance sequencing depth and gene detection, capture non-polyadenylated transcripts, and provide uniform coverage of the transcript body(9). However, bulk RNA-seq data are typically contaminated with transcripts from non-target cell types which can limit specificity for some genes. By contrast, scRNA-seq datasets allows for high specificity in gene detection, as contaminating cells can be identified post-hoc, but can show reduced transcript sensitivity, especially for low abundance cell types(8).

Our bulk sequencing strategy uses a series of *C. elegans* strains, each of which contains one or more fluorescent markers that label an individual neuron type for isolation by FACS. For example, we used *flp-22*::GFP and *unc-47*::mCherry, which together uniquely mark mark the single neuron AVL, for FACS (Fig. 1A). We selected strains based on the following criteria: 1) In the desired neuron type, the fluorophore(s) is expressed brightly enough to be clearly visible under a stereo scope. 2) The fluorophore, or combination of fluorophores, is uniquely expressed in the desired neuron

type. These strains thus represent a rich resource for studying individual neuron types in *C. elegans*; all are deposited at the CGC (Table 1).

For each strain, we used FACS to isolate neurons from synchronized populations of hermaphrodites at the L4 stage, by which time all neurons have been born and are differentiated. Labeled cells were collected in TRIzol LS for RNA extraction. We isolated a wide range of cells (~700 – 90,000) in each sample across neuron types. Multiple biological replicates (e.g., separately grown cultures that were then dissociated and sorted separately) were generated for each neuron class. In total, we sequenced 206 samples across 53 neuron types (Table S1). The 53 neurons that we profiled sample a wide range of anatomical locations (head ganglia, ventral cord, mid-body and tail neurons, pharyngeal neurons) functional modalities (sensory, inter- and motor neurons), neurotransmitter usage (glutamatergic, GABAergic, cholinergic, aminergic) and lineage history (Fig. S1A). A few of these bulk neuron profiles have been previously described(11).

We used a ribodepletion strategy combined with random priming for cDNA synthesis. This approach optimized whole transcript coverage for each gene and also captured non-polyadenylated RNAs (see Methods)(9). The resultant datasets comprise a high-resolution view of RNA expression across the *C. elegans* nervous system. A distribution of neuron-specific data sets for the first two principal components shows separation between sensory neurons (especially ciliated sensory neurons, e.g. ASK, ADL, AFD, AWA) vs motor and interneurons, a result consistent with patterns observed for scRNA-seq data on the same neuron classes (Fig. 1B)(8). The average bulk data for each neuron type also correlates with the single-cell data for that neuron type, with the exception of the OLL neuron, which is an outlier in the single cell dataset with few sampled cells, and may therefore be less accurate (Fig. S1B). Overall, these new bulk datasets enable new studies of neuronal gene expression in individual neuron types in *C. elegans.*

A novel subtraction approach to clean bulk data using information from single-cell data

Our bulk datasets rely on FACS for enrichment of each specific neuron type. Although FACS-isolated cell populations are highly enriched for the target cell type, these samples are also typically contaminated with additional RNA from other cell types—either from errant cells or from ambient RNA in the cell suspension(28, 29). To address this problem, we developed an approach, ‘LittleBites’, that uses pre-existing and reliable information on gene expression to remove contamination from bulk sample data.

LittleBites performs iterative linear subtraction on bulk RNA-Seq data using an scRNA-seq reference and a set of ground truth genes. In brief, LittleBites models each bulk sample as a mix of single cell expression profiles, estimates the cell-type proportions, and then subtracts the non-target profiles. The subtraction is compared to a set of ground truth genes, and repeated iteratively until no further improvements can be made (Fig. 2A; see Methods). To evaluate the post-subtraction datasets we used the area under the Receiver-Operator Characteristic (AUROC) score. In brief, we set a wide range of thresholds to call genes expressed or unexpressed. For each threshold within

that range, we compared the resulting gene expression calls to expected expression from a set of ground truth genes. This comparison produces a true positive rate (TPR, the percentage of truly expressed genes that are called expressed), a false positive rate (FPR, the percentage of truly not expressed genes that are called expressed), and a false discovery rate (FDR, the percentage of genes called expressed that are truly not expressed) for each threshold. The Receiver-Operator Characteristic (ROC) is the graph of the line produced by the TPR and FPR values across the range of thresholds tested, and the AUROC is calculated as the sum of the area under that line. A “random” model of gene expression is expected to have an AUROC value of 0.5, and a “perfect” model is expected to have an AUROC value of 1. Thus, AUROCs below 0.5 are worse than a random guess, and values closer to 1 indicate higher accuracy.

For example, the bulk VB sample VBr31 had an initial composition estimate of: 21% neuronal, 0% excretory, 7% glia, 11% hypodermis, 17% intestine, 5% pharynx, 5% muscle, and 33% reproductive, and its AUROC for ubiquitous and non-neuronal testing genes was 0.87. After LittleBites subtraction, the composition estimates were: 97% neuronal, 0% excretory, 0% glia, 0% hypodermis, 0% intestine, <1% pharynx, 2% muscle, and <1% reproductive, and the AUROC score for the testing gene set increased to 0.92. On average, the neuronal estimate for each sample increased 2.8-fold, and the testing gene AUROC increased by 0.064 (7.7% increase in mean AUROC).

To assess the performance of LittleBites on our bulk data, we compared it to two recently described tools designed to extract cell-type specific expression from bulk RNA-Seq data sets. bMIND uses a Bayesian approach to model cell type-specific expression profiles with a pre-calculated proportion matrix, leveraging averaged single cell profiles for each cell type as an informative prior for a gaussian distribution, estimating parameters using a Markov Chain Monte Carlo (MCMC)(30). ENIGMA uses a regularized matrix completion approach (using either an L2 regularization or a Trace Norm regularization to estimate a low matrix rank constraint) to estimate expression profiles for each sample(21).

Two types of ground truth genes were used for assessment. 1) We used a set of 160 ground truth genes with individual cell-type resolution in the nervous system so that we could also assess how well each technique can distinguish signals expressed in some, but not all, neurons (see Methods; Table S3). 2) We used negative ground truth genes that are expressed exclusively outside of the nervous system (‘non-neuronal’), paired with positive ground truth genes that are expressed in all neuron types (‘ubiquitous’), to assess each method’s ability to accurately remove non-neuronal counts without removing genes expressed in all neuron types (see Methods; Table S4 and S5). LittleBites was run with a testing subset of ubiquitous and non-neuronal genes as a ground truth comparison in the algorithm, and all assessments between methods were performed using a reserved set of testing genes.

We assessed the performance of LittleBites, ENIGMA, and bMIND against these ground truth genes using the AUROC for the average resulting bulk profiles as the primary metric (Fig. 2B,C; S2A,B). We found that LittleBites improved the ubiquitous and non-neuronal AUROC by 7.0%, ENIGMA improved the neuronal AUROC by 6.89%,

and bMIND improved the AUROC by 3.2% ($p < 0.0001$). Further, LittleBites improved the neuronal AUROC by 7.5%, while ENIGMA and bMIND had a lesser effect. LittleBites is the only method that uses ground truth genes to guide cleanup, and we observed that training and testing gene splits have similar AUROC improvement, suggesting that LittleBites is not overfitting to the training genes (Fig. S2C). Comparing results on all individual RNA samples showed consistent improvements across samples from LittleBites (Fig. S2D,E). After LittleBites cleanup, all samples have higher AUROC scores for ubiquitous and non-neuronal testing genes, and all but one sample have higher AUROC scores for the neuronal testing genes compared to unaltered bulk samples. ENIGMA cleanup improves ubiquitous and non-neuronal AUROC for all samples but has little effect on neuronal AUROC for each sample. bMIND improves ubiquitous and non-neuronal AUROC for most samples, but some samples show little effect, and bMIND has a mixed effect on neuronal AUROC, improving some samples while reducing the score for others. We conclude that, for cases where robust ground truth data are available, LittleBites improves the accuracy of cell type expression data in bulk RNA samples.

To test the stability of LittleBites cleanup across different single-cell reference dataset qualities, we ran the algorithm on a set of bulk samples by first subsetting the corresponding single-cell cluster's population to 10, 50, 100, or 500 cells. We performed this process 500 times for each subsampling rate for each sample (2000 total runs per sample). We found that testing gene AUROC values are stable across reference cluster sizes (Fig. 2D), suggesting that even if the target cell type is rarely represented in a single cell reference, accurate cleaning is still possible. However, comparing gene level stability across target cluster population levels reveals that low abundance references have higher gene level variance (Fig. 2E), lower purity estimates (Fig. S2F), higher variance in the mean expression across genes (Fig. S2G), and they tend to have lower overall expression (suggesting more aggressive subtraction) (Fig. S2H). This indicates that while binary gene calling is improved even if the reference cluster is small, users should be cautious when using fewer than 100 cells in their single cell reference cluster as the resulting cleanup is less stable.

A model for the relationship between gene counts and proportions in pseudobulk single-cell RNAseq data

To integrate our previous scRNAseq data(11) with our bulk data, we first considered how to analyze the single cell data in each cluster. Gene expression in single cell data can be measured in two ways: as *counts,* averaging or aggregating across all cells in each cluster, or as *proportions,* assessing the proportion of cells in the cluster that detect the gene (Fig. 3A)(31-33). Comparing these two measurement systems against a set of ground truth genes in the *C. elegans* nervous system, we found that using the proportion yields a more accurate measurement of gene expression than the alternative of relying on counts (Fig. 3B)(11).

To exploit the improved accuracy that proportion measures provide, we aimed to transform these proportion values into units of counts for comparison to our bulk data. Building on previous work(32), we simulated a single cell dataset, using a negative binomial distribution to model counts for individual cells in a cluster. We found that this simulated dataset could be well-modeled using the log-likelihood ratio (logit) of the non-zero proportion of cells, scaled by the total number of cells in the dataset (Fig. 3C). Using our real dataset, we observed a similar fit, with the exception that some genes with low non-zero proportions values showed very high counts (Fig. 3D). These exceptions may be caused by high ambient RNA background in some droplets, misassigned cells in the clustering step, doublets that escaped early filtering steps, or rare bursts of expression for lowly expressed genes. The presence of these outliers suggests why the non-zero proportion value may be a more robust measure for gene expression in single-cell clusters than counts: it may be more resistant to rare occurrences of contaminating information.

We used the logit model to generate a denoised counts estimate for each single cell cluster, which we call the scProp2Count dataset. As a final test, we compared the accuracy of the scProp2Count dataset to a standard approach using counts on pseudobulk data, performing pairwise differential expression analysis using edgeR, and then comparing the differentially expressed genes to ground truth expression patterns. We found that the scProp2Count dataset slightly outperforms standard single-cell counts in detection sensitivity and correlation to ground truth, while matching the single-cell counts in precision (Fig. S3). However, caution should be taken when using this approach in scRNAseq cases where all replicates of a cell type contain few cells. scProp2Count values are limited to the space of possible proportion values, and so replicates with low numbers of cells will have fewer potential expression “levels” which may break some model assumptions in downstream applications (see Methods).

Integrating bulk and single-cell data improves gene detection accuracy.

To integrate the bulk and single-cell data sets, for each neuron type in the bulk data set (Table 1) we computed the arithmetic mean of normalized gene counts across all biological replicates (‘Bulk’ dataset). Next, we integrated the Bulk data set with the scProps2Counts dataset by taking the geometric mean for each gene. The result of this analysis was three data sets, each containing normalized gene count data for *C. elegans* neuron classes: Bulk, scProps2Counts, and Integrated. Finally, for each of these three data sets we thresholded the gene counts to call every gene as either expressed or not expressed in each cell type. For this purpose, a single threshold was used for each data set, consisting of the average gene count value for all cells and all genes in each data set. Genes with count values above this threshold were defined as expressed.

To assess whether integration improved the accuracy of expression, we compared expression calls to ground-truth data for both individual neuron types and non-neuronal cell types (see Methods; Table S3,4,5)(8). We found that the Integrated dataset generally outperforms the LittleBites-processed Bulk and scProps2Counts scRNA

datasets (Fig. 4). Like the LittleBites-processed Bulk dataset, the Integrated dataset approaches a True Positive Rate (TPR) of 100% at relatively low precision (or high FPR), whereas the scProp2Count scRNA dataset has a maximum TPR of 91.9%. The Integrated dataset improves the overall AUROC value by 3.8% compared to the scProp2Count scRNA data, and 2.4% compared to the LittleBites-processed Bulk data ($P < 0.0001$) (Fig. 4A,B). Like the scProp2Count scRNA dataset, the Integrated dataset has a relatively low False Positive Rate (FPR) for neuronal ground truth genes (Fig. 4C). At a false discovery rate (FDR) of 5%, the Integrated dataset shows substantially better sensitivity than the LittleBites-cleaned Bulk and scProps2Counts scRNA datasets (Fig. 4D). At the level of individual ground truth homeobox and neurotransmitter genes expressed in neurons(34, 35), specific and broad improvements can be seen across the Integrated dataset, compared to all other dataset versions (Fig. S4A,B).

In addition to using ground truth genes to assess the accuracy of the Integrated dataset, we also evaluated gene expression differences with the original sc-RNAseq dataset(11). We thresholded (see Methods) both datasets and determined that the integration strategy results in fewer expressed genes—for some neuron types, over 1,000 fewer genes (Fig. 4E). To further investigate the difference between the Integrated dataset and the original sc-RNAseq dataset, we examined the 'unique' genes for each neuron type: i.e., genes called expressed in that neuron type in one dataset but not the other. We analyzed these 'unique' genes using Tissue Enrichment Analysis and Gene Ontology (GO) for two representative neuron types, ADL and VB, both of which are well-represented in the sc-RNAseq dataset(11). Importantly, Tissue Enrichment Analysis maps genes to specific *C. elegans* cell types. We found that overall, contamination by other cell types was much lower in the Integrated dataset. Further, the 'unique' genes in the sc-RNAseq dataset were more enriched in non-neuronal cell types, such as germline, and for non-neuronal GO terms. By contrast, the 'unique' genes in the Integrated dataset were more enriched for other neuronal cell types and neuronal functions (Fig. 4F,G, S4C). Testing against a curated list of non-neuronal genes from fluorescent reporters and genomic enrichment studies, we found that of 445 non-neuronal markers, each gene was detected in an average of 12.5 cells or a median of 3 cells in the single-cell dataset, and an average of 8.7 cells or a median of 1 cell in the integrated dataset, at a 14% FDR threshold. Thus, integration appears to successfully remove most contaminating transcripts.

Together, these results show that geometric mean integration of bulk RNA-seq and scRNA-seq datasets combines the strengths of both approaches, providing high sensitivity (bulk RNA-Seq) and high specificity (scRNA-Seq) across a wide range of thresholds. For most purposes, we propose that the Integrated dataset will give the best results. However, for certain applications (such as genes not detected in scRNA-seq data, or non-polyadenylated transcripts, see below) other approaches are needed. To make the Integrated data easily accessible to the community, we provide tables of gene expression for each neuron type based on the Integrated data set, both the values for all genes (Table S6) and values binarized at four thresholds (Tables S7-10).

Bulk RNA-seq reveals both broadly expressed and neuron-specific noncoding RNAs

A significant benefit of our bulk RNA-seq approach is its sensitivity to non-poly-adenylated transcripts, which include many species of non-coding RNA(9). However, similar to coding genes not found in the scRNA-seq dataset (see above), the scRNA-seq dataset is largely uninformative about these transcripts. Further, we do not have a ground-truth data set of non-coding genes to evaluate accuracy of expression calling for these transcripts. Finally, most non-coding RNAs are expressed at lower levels than protein coding genes (Fig. S5A). We therefore applied a generous, uniform threshold for calling gene expression, consisting of > 5 normalized counts in 2/3 or more of the samples within a cell type. As we did for our analysis of coding genes not found in scRNA-seq, we used gene level correlation to contamination estimates as a procedure to eliminate non-coding genes that were likely detected due to contamination from other tissues in the bulk samples. First, we estimated contamination for each sample using a bootstrapped NNLS regression (Fig. S5B; see Methods), and then calculated per-gene Spearman correlations to each contaminant type. We applied a threshold on the gene level correlation to contamination estimates for each sample by fitting a Gaussian mixture model to the maximum correlation score for each gene. We selected a cutoff of 0.23, which excludes 98% of the estimated contamination distribution (Fig. S5B). With these thresholds, an average of 512 noncoding RNAs were identified as “expressed” per cell type (95 CI ± 40.6) (Table S11). By RNA type, we detected 16.2 ± 1.2 lincRNAs, 39.0 ± 5.3 pseudogenes, 54.8 ± 9.7 tRNAs, 41.8 ± 2.0 snRNAs, 131.5 ± 1.8 snoRNAs, and 228.9 ± 29.0 uncategorized ncRNAs per cell type (Table S12).

Next, we sought to identify noncoding RNAs with broad expression across multiple neuron types. We defined broad expression as genes expressed in > 80% of neuron classes in our dataset (Fig. S5C). This approach defined 247 non-coding genes as broadly expressed, or “pan-neuronal”. These pan-neuronal noncoding RNAs include 108 (43.8%) snoRNAs and 33 (13.4%) snRNAs, both tenfold greater than the expected proportion assuming a random distribution (Fisher’s exact test, P-value < 0.01) (Fig. 5A). By contrast, pseudogenes and otherwise uncategorized ncRNAs were significantly depleted from the list of pan-neuronal noncoding RNAs (P-value < 0.001). These results indicate that snoRNAs and snRNAs are widely expressed, which matches studies showing broad expression of many snoRNAs and snRNAs in other systems(39, 40), and is consistent with their key roles in rRNA processing and splicing(41-43).

We also sought to identify neuron-type-specific noncoding RNAs. We calculated tissue specificity scores for each noncoding RNA called expressed in at least one neuron type using the Preferential Expression Measure (PEM) score(23, 44). We called these genes neuron-type specific according to three criteria: (1) Called expressed in $\geq$ one neuron type (see above); (2) PEM score > 0.65; (3) > 2 normalized counts in a maximum of 10/53 neuron types. Using these thresholds, we identified 523 cell-type-specific noncoding RNAs (Fig. 5B). By RNA type, 293 (56.0%) of cell type-specific noncoding RNA genes are uncategorized ncRNAs, 93 (17.8%) are general pseudogenes 112 (21.4%) are GPCR pseudogenes, 14 (2.7%) are tRNAs, 10 (1.9%) are lincRNAs, 1 (0.2%) is a snoRNA, and none are snRNAs. We observed significant enrichment of pseudogenes, and a small but significant depletion of ncRNAs, snoRNAs, and tRNAs (P-value < 0.01). Clustering by genes and cell type modalities revealed clear enrichment for specific noncoding RNAs in individual neuron types (Fig. 5C), although the overall proportions of noncoding RNA types were similar between individual neuron classes

(Fig. S5D). The number of specific noncoding RNAs per cell type ranged from 0 (PVC) to 139 (ADL), with a mean of 10.2 (± 5.8).

Interestingly, we observed that in ADL—the neuron that expresses the highest number of specific ncRNAs—many of these specific ncRNAs are GPCR pseudogenes (Fig. 5C,D, Table S13). ADL also expresses the highest number of bona fide GPCRs among neurons in our dataset (Fig. 5E, Table S14). Looking across neuron types, we observe a rough correspondence between expression of specific GPCR pseudogenes and expression of bona fide GPCRs (Fig. 5D,E). This analysis suggests that pseudogenes retain their regulatory information, despite losing their coding potential. Overall, our data on expression of non-coding genes reveal a wide diversity of noncoding RNA expression across the nervous system and open the door to in depth studies of noncoding RNA contributions to individual neuron function.

Discussion

In this work, we present bulk RNA-seq data for 53 neuron classes in the *C. elegans* nervous system. We describe new methods for integrating these bulk RNA-seq data with previously obtained single-cell RNA-seq data(11). We find that this approach improves gene detection accuracy in comparison to each individual data set. We use the bulk data sets to detect two important classes of genes that were underrepresented in the previous single-cell atlas: (1) lowly expressed genes, and (2) non-poly-adenylated transcripts, chiefly non-coding RNAs. Overall, our results provide a rich resource for studies of gene expression in the *C. elegans* nervous system.

The rapid growth of scRNA-seq and bulk RNA-seq datasets create opportunities to combine these two data types to address diverse goals. For example, recent studies have combined data from bulk RNA sequencing and scRNA-seq to address the problem of deconvolution. The goal of deconvolution is to infer cell-type expression profiles from tissue level bulk samples, using scRNA-seq references as a guide(30, 45-47). In this study we combine these two types of data, both collected from the *C. elegans* nervous system, for a different purpose: to improve the accuracy of cell and tissue-specific transcriptional profiles. We developed three approaches: a subtraction approach, LittleBites, that uses information from scRNA-seq data, combined with ground truth gene expression data, to identify and remove contaminating counts from bulk RNAseq data; Prop2Count, that expands on the theoretical link between gene expression and single-cell dropout rates to provide a denoising approach for pseudobulk scRNA-seq; and an integration approach that directly combines information from bulk and single-cell approaches to improve sensitivity and specificity. Our computational approaches may be useful in other cases where scRNA-seq and bulk RNA-seq datasets are available to improve information on gene expression.

In addition to enhancing the accuracy of gene expression, the integrated bulk RNA-seq dataset detects lowly expressed protein coding genes that were not detected by scRNA-seq. Further, because our library construction methods were designed to capture non-polyadenylated transcripts, our bulk RNA-seq data set detects noncoding RNAs that were not revealed by previous scRNA-seq results(8, 9). Some of these noncoding RNAs are broadly expressed in the nervous system, suggestive of shared functions across different types of neurons. Other non-coding RNAs are expressed in a limited number of neuron types, suggesting neuron type-specific functions. In addition, the bulk RNA-seq dataset contains transcript information across the gene body, which parallel efforts have used to identify mRNA splicing patterns that are not found in the scRNA-seq dataset(26, 27).

Overall, our approach achieves a comprehensive representation of all classes of transcripts expressed in individual neuron types. These data can now drive analysis of mechanisms that control gene expression across the genome in individual neuron types, and support identification of differentially expressed genes that define neuron-type specific differences in morphology and function. Public access to these data (described below) will enable further analysis into the regulation and function of differential gene expression in *C. elegans* neurons.

Data Availability

Bulk raw data and single cell raw data are available at Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo). Bulk data identifier: GSE229078. Single cell data identifier: GSE136049. Counts data and additional supporting files can be downloaded from the CeNGEN website (https://www.cengen.org). Code is available at GitHub (https://www.github.com/cengenproject).

Declaration of Interests

The authors declare no competing interests.

Figures and Tables

Table 1:

All strains used for bulk RNA-seq experiments, including strain names, allele information, and transgenic fluorescent labels for sorting individual neuron types.

Supplementary Table 1:

Replicate metadata for bulk RNA-seq experiments, with replicate names, strain names, and the number of cells collected.

Supplementary Table 2:

List of genes that were removed from analysis because they overlap with promoter regions used in the construction of transgenic markers for FACS sorting.

Supplementary Table 3:

Ground Truth expression matrix for 160 genes in the *C. elegans* nervous system using fosmid and CRISPR/Cas reporter lines (see Methods).

Supplementary Table 4:

List of 445 genes that are expressed exclusively outside the *C. elegans* nervous system (see Methods).

Supplementary Table 5:

List of 936 ubiquitous genes (see Methods).

Supplementary Table 6:

Integrated gene expression matrix.

Supplementary Tables 7-10:

Thresholded gene-by-cell expression matrices using the Integrated data set, computed at four thresholds, values below the threshold are set to 0.

Supplementary Table 11:

List of ncRNA genes expressed in individual neuron types.

Supplementary Table 12:

List of ncRNA gene types expressed in individual neuron types.

Supplementary Table 13:

List of GPCR pseudogenes included in this analysis.

Supplementary Table 14:

List of protein coding GPCR genes included in this analysis.

Figure 1: Single neuron bulk RNA-seq via targeted marker expression and FACS isolation.

A) Labeling, tissue dissociation, and FACS-enrichment schemes for capturing individual neuron types. Intersecting *flp-22*::GFP and *unc-47*::mCherry markers uniquely label AVL for isolation by FACS from dissociated L4 stage larval cells. RNA from this pool of AVL-enriched cells was used for bulk RNA sequencing (see Methods). B) PCA plot showing all bulk RNA-seq data labeled by cell type and colored according to functional modality; Sensory neurons (blue), motor neurons (green), interneurons (red), and CAN neurons (purple).

Supplementary Figure 1.

A) Number of cell types sequenced per functional modality. B) Heatmap of Spearman Correlations between average single cell RNA-seq (row) and Bulk RNA-seq (column) profiles for each neuron type. For each row, correlations were calculated for genes called expressed in that single cell cluster (from single cell thresholding)(*11*). The single cell data did not separate the DD and VD GABA motor neuron types(*11*); these rows are identical.

Figure 2: Identifying and removing contamination from bulk data with LittleBites.

A) Schematic representation of the LittleBites algorithm. Each iteration of the LittleBites algorithm contains 3 steps: 1) Use an NNLS regression to estimate the cell-type proportions of each bulk sample with aggregated single cell data as a reference. 2) Subtract the estimated contamination profiles, constrained by the proportion estimates, a gene level specificity weight, and a variable learning rate. 3) Compare the result of every tested learning rate against ground truth, and select the learning rate with the highest area under the receiver operator characteristic (AUROC). This loop repeats until further subtraction cannot improve the AUROC. B) Line plots of the receiver operator characteristic (ROC) curve for the unaltered bulk data and corrected datasets, as well as bMIND and ENIGMA analyses, assessed against ubiquitous genes and genes exclusively expressed in non-neuronal cells. C) Line plots of the ROC for the unaltered bulk data and corrected datasets, as well as bMIND and ENIGMA analyses, assessed against 160 genes expressed within the nervous system. D) Violin plot of AUROC values for bulk samples after LittleBites cleaning using a random subset of target cells in the single cell reference. X-axis indicates the number of target cells used in the reference. Each subsampling rate was run 500 times. e.g. for a VB sample, it was modeled using 10, 50, 100, or 500 VB neurons in the reference. E) Violin plot of the average coefficient of variation (CV) per gene after LittleBites cleanup. The Y-axis indicates the average CV across all genes in a single sample, and the the X-axis indicates the number of target cells used in the single-cell reference.

Supplementary Figure 2.

A) Bar graph showing the AUROC values for each dataset assessed against ubiquitous genes and genes exclusively expressed in non-neuronal cells. Error bars represent bootstrap estimated confidence intervals. B) Bar graph showing the AUROC values for each dataset, assessed against 160 genes expressed within the nervous system. Error bars represent bootstrap estimated confidence intervals. C) Scatterplot of the training and testing AUROC improvement after LittleBites cleanup. Each dot represents one sample. Red line indicates x=y. D, E) Scatter plots of pairwise comparisons of sample-level AUROC values for unaltered and corrected bulk datasets. Ubiquitous and non-neuronal AUROCs are represented in panel C, and neuronal AUROCs are represented in panel D. F) Violin plot of unaltered bulk sample estimates for the proportion of the counts that come from neurons, vs the number of target cells used in the single cell reference. G) Violin plot in the coefficient of variation (CV) in the mean value of LittleBites cleaned samples, vs the number of target cells used in the single cell reference. H) Violin plot of the mean gene expression value after LittleBites cleanup, vs the number of target cells used in the single cell reference. Model was fit with slopes calculated for each neuron type separately.

Figure 3: A model for the relationship between cell proportions and gene counts in sc-RNAseq data.

A) Aggregate count and proportion measures of scRNA-seq data. UMAP representations of AFD neurons captured in the scRNA-seq dataset(*11*). A) Left, log scale representation of *ttx-1* counts in all AFD neurons. Right, binarized representation of *ttx-1* expression in all AFD neurons. B) ROC for *C. elegans* neuronal ground-truth genes in scRNA-seq counts and proportions. The teal line indicates the proportion of non-zero cells, and the red line indicates the normalized counts. C) Scatter plot of fitted log-counts for simulated data, with log-counts on the Y-axis, and the sum of the log likelihood ratio of the proportion of non-zero cells and the log of the number of cells in the cluster on the X-axis. Red line indicates Y = X. D) Scatter plot of fitted log-counts for one SIA neuron replicate from the *C. elegans* neuronal scRNA-seq dataset(*11*), with log-counts on the Y-axis, and the sum of the log likelihood ratio of the proportion of non-zero cells and the log of the number of cells in the cluster on the X-axis. Red line indicates Y = X.

Supplementary Figure 3.

Box plots of edgeR differential expression metrics for single-cell counts and Prop2Count inputs. 1176 neuron-neuron pairs were compared in edgeR, and differentially expressed genes were called as any

gene with a p.adj value < 0.05, and a log2 effect size > 2. Metrics were generated by comparing differentially expressed genes with 160 differentially expressed ground truth genes. A) Box plots of TPR values for neuron-neuron edgeR comparisons. B) Box plots of FPR values for neuron-neuron edgeR comparisons. C) Box plots of FDR values for neuron-neuron edgeR comparisons. D) Box plots of MCC values for neuron-neuron edgeR comparisons. Red box plots represent single-cell counts edgeR comparisons, and teal box plots represent Prop2Count edgeR comparisons. All comparisons were made with a two-way permutation test (* = p < 0.05, ** = p < 0.01).

Figure 4: Integrating bulk RNA-seq and scRNA-seq data sets improves gene detection accuracy. A) Receiver Operator Characteristic (ROC) curve for bulk, single-cell, and integrated datasets compared to neuronal ground-truth genes. The x-axis shows the False Positive Rate (FPR), and the y-axis shows the true positive rate (TPR). B) Area under the receiver operator characteristic (AUROC). Error bars represent bootstrap 95% confidence intervals of the AUROC value. Pairwise statistical testing was performed using the deLong method for AUROC scores. *** = p < 0.0001. C) Precision-Recall (PR) curve for bulk, single-cell, and integrated datasets compared to neuronal ground-truth genes. The x-axis shows the Precision (1 – False Discovery Rate/FDR), and the y-axis shows the TPR (Recall). D) Boxplot of bootstrapped sensitivity scores for each dataset at a 5% FDR cutoff. E) Bar chart of the number of genes detected at the medium threshold (FDR = 0.14) for the integrated dataset (red) and the dynamic proportions single cell dataset (blue). Bars are ordered by the number of cells in the corresponding single cell cluster from lowest to highest. F,G) Bar plots of significant terms in Tissue enrichment analyses for genes detected only in the single-cell and integrated datasets in ADL neurons. Dynamic thresholding was used for single cell data, and medium thresholds were used to binarize both datasets.

Supplementary Figure 4.

A) Comparison between 'ground truth' fluorescent reporters and RNAseq gene expression for all datasets, shown for a set of homeobox genes(*34*). Fluorescent reporter expression is binarized and is shown as small central squares, with black = expressed and white = not expressed. RNA transcripts shown as natural log transformed TPM counts, color-mapped according to the scale. B) Comparison between 'ground truth' fluorescent reporters and RNAseq gene expression for the Integrated dataset, shown for a set of genes encoding neurotransmitter machinery(*35*). Fluorescent reporter expression is binarized and is shown as small central squares, with black = expressed and white = not expressed. RNA transcripts shown as natural log transformed TPM counts, color-mapped according to the scale. C) Bar plots of significant terms in GO and Tissue enrichment analyses for genes detected only in the single-cell and integrated datasets in ADL and VB neuron types. D) scatterplot of newly detected genes in the Integrated dataset that were previously undetected in all tissues using the single cell reference. Y-axis is the number of genes detected, X-axis is the number of cells in the single cell reference. Red line indicates the line of best fit (log model), and blue lines indicate 95% confidence intervals of the model. E) Boxplot of the number neurons that each ground truth non-neuronal gene was called expressed in at the 14% or 8.4% FDR threshold. Integrated data is filled red, single cell data is filled teal. Notches indicate 95% confidence intervals. Dots represent outliers.

Figure 5: Bulk analysis reveals noncoding RNA expression patterns.

A) Proportions of classes of pan-neuronal noncoding RNAs. B) Proportions of classes of cell type specific noncoding RNAs C) Heatmap of average normalized counts per cell type for ncRNA genes considered cell type specific, rows are grouped by RNA class. Note that genes within individual neuron types are clustered by expression patterns, not similar gene function. D) Number of GPCR pseudogenes with specific expression per neuron type, grouped by neuron function. E) Number of GPCR genes per neuron type, grouped by neuron function.

Supplementary Figure 5.

A) Log-scaled expression distribution of protein coding genes (blue) and non-coding genes (pink). B) Density plot showing the distribution of gene level correlation to contaminant estimates (purple); values plotted are the highest correlation per gene. Genes plotted were called expressed in at least one cell type. Blue and black dashed lines represent a gaussian mixture model, used to threshold against contaminant genes. All noncoding genes with a maximum correlation above 0.22 (vertical red line) were removed from analysis. C) Histogram of ncRNA genes binned by number of neuron types in which they

are expressed; genes to the right of the red line are expressed in >90% of neuron types in this study. D) Barplot of the number of noncoding RNAs detected in each neuron type, colored by noncoding RNA class.

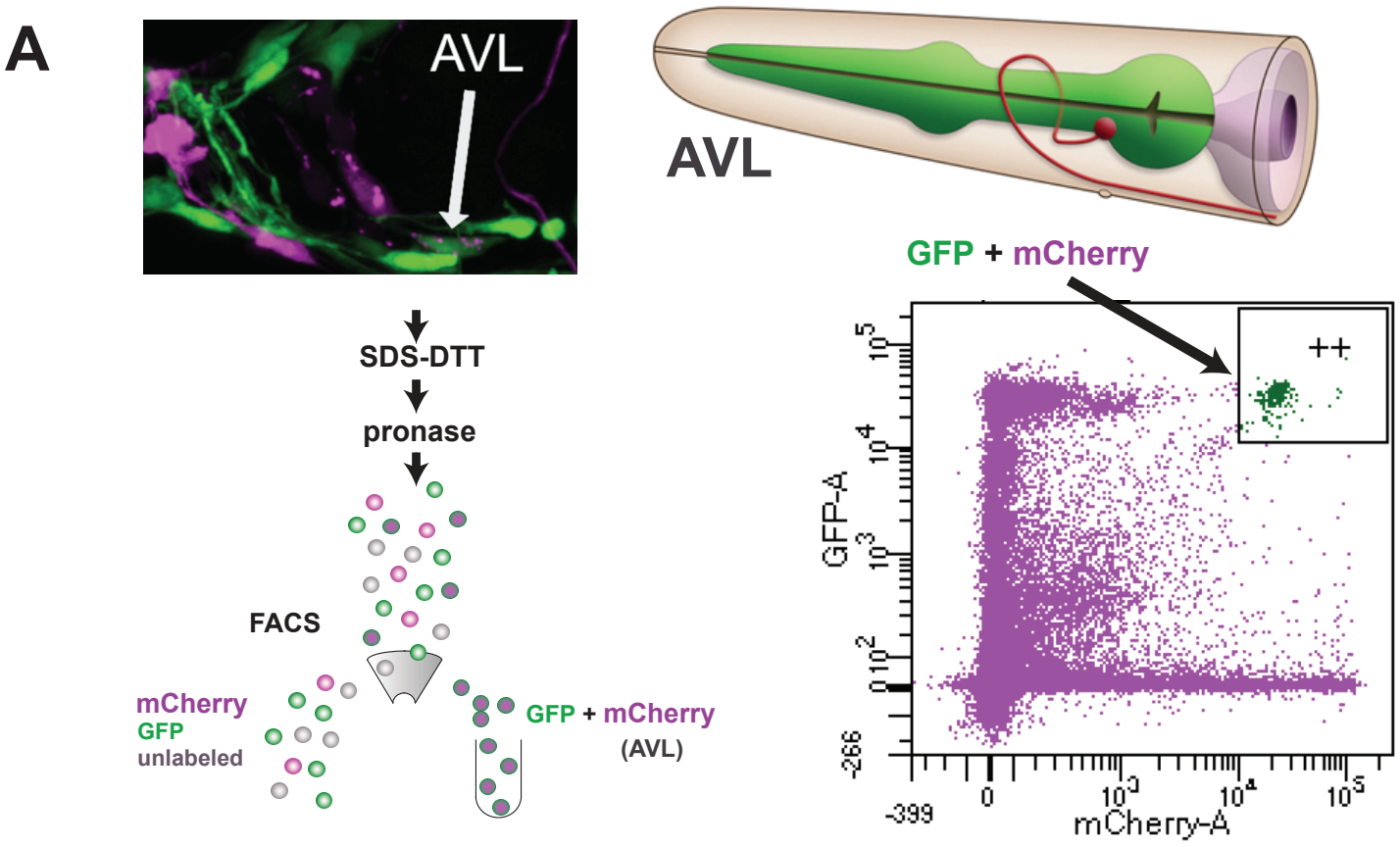
References

1. J. E. Sulston, H. R. Horvitz, Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev Biol* **56**, 110-156 (1977).
2. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol* **100**, 64-119 (1983).
3. J. G. White, E. Southgate, J. N. Thomson, S. Brenner, The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci* **314**, 1-340 (1986).
4. S. J. Cook *et al.*, Whole-animal connectomes of both Caenorhabditis elegans sexes. *Nature* **571**, 63-71 (2019).
5. C. A. Brittin, S. J. Cook, D. H. Hall, S. W. Emmons, N. Cohen, A multi-scale brain map derived from whole-brain volumetric reconstructions. *Nature* **591**, 105-110 (2021).
6. M. W. Moyle *et al.*, Structural and developmental principles of neuropil assembly in C. elegans. *Nature* **591**, 99-104 (2021).
7. M. Hammarlund, O. Hobert, D. M. Miller, 3rd, N. Sestan, The CeNGEN Project: The Complete Gene Expression Map of an Entire Nervous System. *Neuron* **99**, 430-433 (2018).
8. S. R. Taylor *et al.*, Molecular topography of an entire nervous system. *Cell* **184**, 4329-4347.e4323 (2021).
9. A. Barrett *et al.*, A head-to-head comparison of ribodepletion and polyA selection approaches for Caenorhabditis elegans low input RNA-sequencing libraries. *G3 Genes|Genomes|Genetics* **11**, (2021).
10. W. C. Spencer *et al.*, Isolation of specific neurons from C. elegans larvae for gene expression profiling. *PLoS One* **9**, e112102 (2014).
11. S. R. Taylor *et al.*, Molecular topography of an entire nervous system. *Cell* **184**, 4329-4347 e4323 (2021).
12. S. R. Taylor, R. D. McWhirter, B. K. Matlock, D. K. Flaherty, D. M. Miller, 3rd, Protocol for isolating specific C. elegans neuron types for bulk and single-cell RNA sequencing. *STAR Protoc* **5**, 103439 (2024).
13. N. Stefanakis, I. Carrera, O. Hobert, Regulatory Logic of Pan-Neuronal Gene Expression in C. elegans. *Neuron* **87**, 733-750 (2015).
14. A. Bhattacharya, U. Aghayeva, E. G. Berghoff, O. Hobert, Plasticity of the Electrical Connectome of C. elegans. *Cell* **176**, 1174-1189.e1116 (2019).
15. T. W. Harris *et al.*, WormBase: a modern Model Organism Information Resource. *Nucleic Acids Research* **48**, D762-D767 (2019).
16. M. B. Reilly, C. Cros, E. Varol, E. Yemini, O. Hobert, Unique homeobox codes delineate all the neuron classes of C. elegans. *Nature* **584**, 595-601 (2020).

17. E. Yemini *et al.*, NeuroPAL: A Multicolor Atlas for Whole-Brain Neuronal Identification in C. elegans. *Cell* **184**, 272-288 e211 (2021).
18. J. S. Packer *et al.*, A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science* **365**, (2019).
19. M. Smid *et al.*, Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons. *BMC Bioinformatics* **19**, 236 (2018).
20. J. Wang, K. Roeder, B. Devlin, Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res* **31**, 1807-1818 (2021).
21. W. Wang *et al.*, Approximate estimation of cell-type resolution transcriptome in bulk tissue through matrix completion. *Brief Bioinform* **24**, (2023).
22. S. J. Xiao, C. Zhang, Q. Zou, Z. L. Ji, TiSGeD: a database for tissue-specific genes. *Bioinformatics* **26**, 1273-1275 (2010).
23. N. Kryuchkova-Mostacci, M. Robinson-Rechavi, A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics* **18**, 205-214 (2016).
24. T. C. e. S. Consortium, Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
25. L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, D. B. Chklovskii, Structural properties of the Caenorhabditis elegans neuronal network. *PLoS Comput Biol* **7**, e1001066 (2011).
26. A. Weinreb *et al.*, Alternative splicing across the C. elegans nervous system. *bioRxiv*, 2024.2005.2016.594567 (2024).
27. Z. Wolfe, D. Liska, A. Norris, Deep Transcriptomics Reveals Cell-Specific Isoforms of Pan-Neuronal Genes. *bioRxiv*, 2024.2005.2016.594572 (2024).
28. H. Tripathi *et al.*, Isolation Methods for Human CD34 Subsets Using Fluorescent and Magnetic Activated Cell Sorting: an In Vivo Comparative Study. *Stem Cell Rev Rep* **16**, 413-423 (2020).
29. S. R. Ocanas *et al.*, Minimizing the Ex Vivo Confounds of Cell-Isolation Techniques on Transcriptomic and Translatomic Profiles of Purified Microglia. *eNeuro* **9**, (2022).
30. J. Wang, K. Roeder, B. Devlin, Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data. *Genome Research* **31**, 1807-1818 (2021).
31. T. H. Kim, X. Zhou, M. Chen, Demystifying "drop-outs" in single-cell UMI data. *Genome Biol* **21**, 196 (2020).
32. A. S. Booeshaghi, L. Pachter, Normalization of single-cell RNA-seq counts by log(x + 1)dagger or log(1 + x)dagger. *Bioinformatics* **37**, 2223-2224 (2021).
33. G. A. Bouland, A. Mahfouz, M. J. T. Reinders, Differential analysis of binarized single-cell RNA sequencing data captures biological variation. *NAR Genom Bioinform* **3**, lqab118 (2021).
34. M. B. Reilly *et al.*, Widespread employment of conserved C. elegans homeobox genes in neuronal identity specification. *PLoS Genet* **18**, e1010372 (2022).
35. C. Wang *et al.*, A neurotransmitter atlas of C. elegans males and hermaphrodites. *bioRxiv*, 2023.2012.2024.573258 (2024).

36. E. Mereu *et al.*, Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology* **38**, 747-755 (2020).
37. D. Angeles-Albores, R. Y. N. Lee, J. Chan, P. W. Sternberg, Tissue enrichment analysis for C. elegans genomics. *BMC Bioinformatics* **17**, 366 (2016).
38. P. N. Inglis, G. Ou, M. R. Leroux, J. M. Scholey, The sensory cilia of Caenorhabditis elegans. *WormBook*, 1-22 (2007).
39. A. Isakova, T. Fehlmann, A. Keller, S. R. Quake, A mouse tissue atlas of small noncoding RNA. *Proceedings of the National Academy of Sciences* **117**, 25634-25645 (2020).
40. É. Fafard-Couture, D. Bergeron, S. Couture, S. Abou-Elela, M. S. Scott, Annotation of snoRNA abundance across human tissues reveals complex snoRNA-host gene relationships. *Genome Biology* **22**, 172 (2021).
41. D. A. Wassarman, J. A. Steitz, INTERACTIONS OF SMALL NUCLEAR RNAS WITH PRECURSOR MESSENGER-RNA DURING INVITRO SPLICING. *Science* **257**, 1918-1925 (1992).
42. S. Valadkhan, in *Progress in Molecular Biology and Translational Science,* G. A. Soukup, Ed. (Academic Press, 2013), vol. 120, pp. 195-228.
43. T. Bratkovič, J. Božič, B. Rogelj, Functional diversity of small nucleolar RNAs. *Nucleic Acids Research* **48**, 1627-1651 (2019).
44. L. Huminiecki, A. T. Lloyd, K. H. Wolfe, Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics* **4**, 31 (2003).
45. L. Zhu, J. Lei, B. Devlin, K. Roeder, A UNIFIED STATISTICAL FRAMEWORK FOR SINGLE CELL AND BULK RNA SEQUENCING DATA. *Ann Appl Stat* **12**, 609-632 (2018).
46. A. M. Newman *et al.*, Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37**, 773-782 (2019).
47. W. Wang *et al.*, Improved estimation of cell type-specific gene expression through deconvolution of bulk tissues with matrix completion. *bioRxiv*, 2021.2006.2030.450493 (2021).

Fig. 1


A
AVL
AVL
SDS-DTT
pronase
FACS
mCherry
GFP
unlabeled
GFP + mCherry
(AVL)
GFP + mCherry
++
GFP-A
mCherry-A
-266
-399
0
10^2
10^3
10^4
10^5

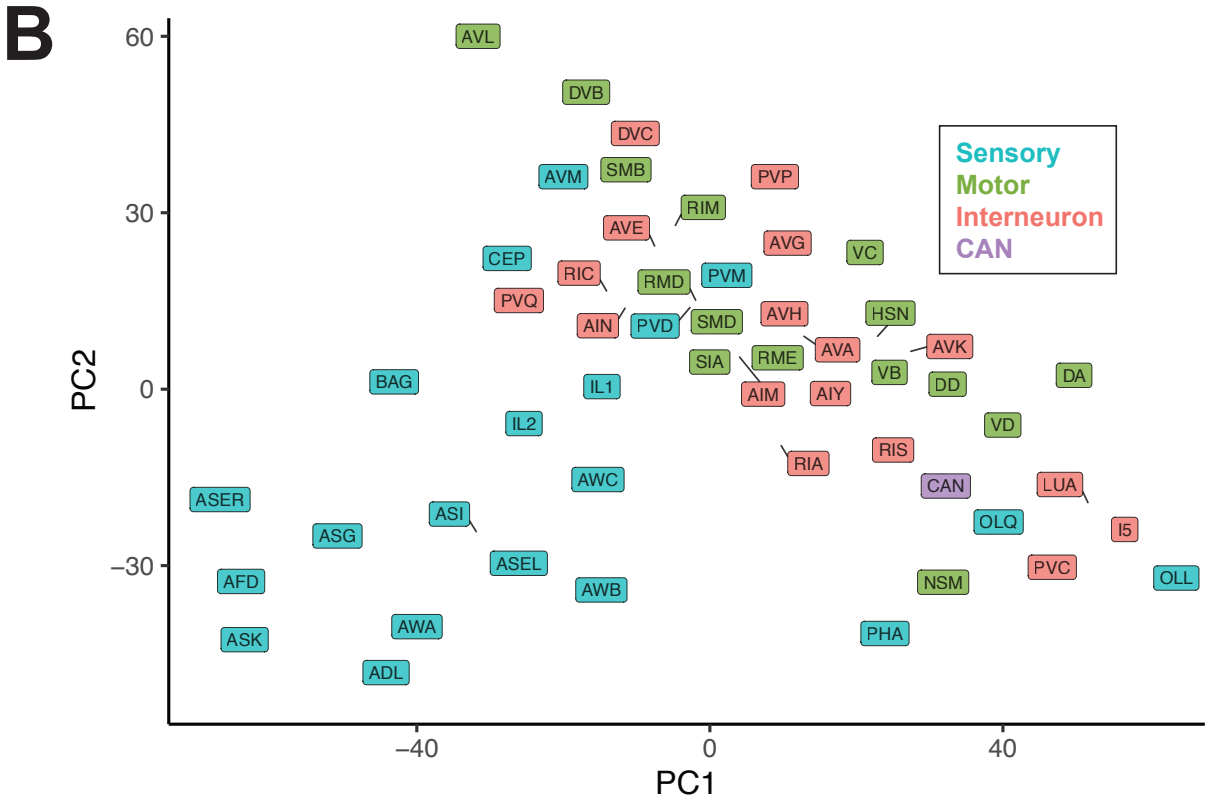

B
Sensory
Motor
Interneuron
CAN
PC1
PC2
60
30
0
−30
−40
0
40
AVL
DVB
DVC
AVM
SMB
PVP
RIM
AVE
AVG
VC
CEP
RIC
RMD
PVM
PVQ
AIN
PVD
SMD
AVH
HSN
AVA
AVK
SIA
RME
VB
DD
DA
BAG
IL1
AIM
AIY
IL2
VD
RIS
RIA
AWC
CAN
LUA
ASER
ASI
OLQ
I5
ASG
ASEL
PVC
AFD
AWB
NSM
OLL
AWA
PHA
ASK
ADL

Fig. S1

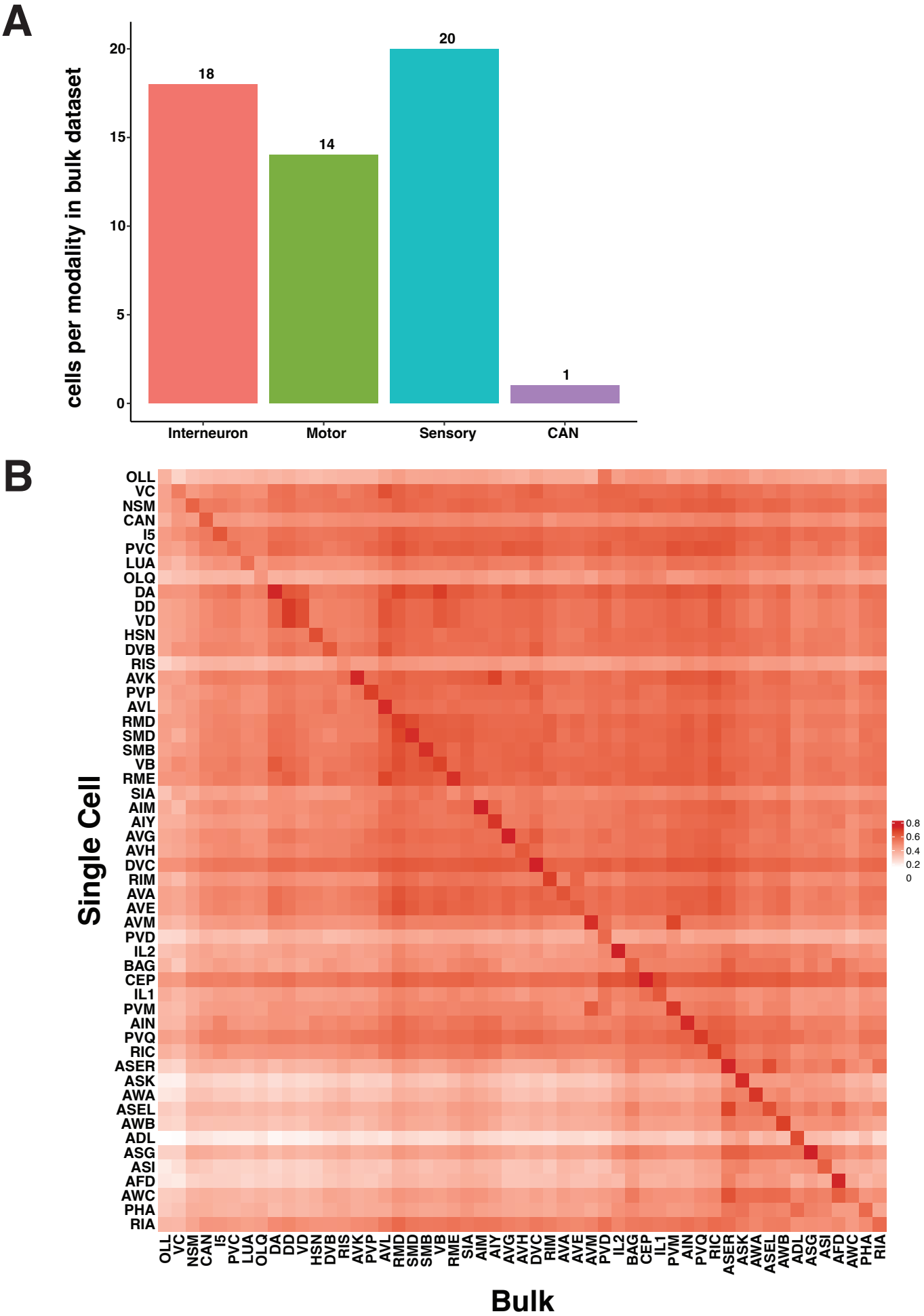

A
cells per modality in bulk dataset
0
5
10
15
20
18
14
20
1
Interneuron
Motor
Sensory
CAN
B
Single Cell
Bulk
OLL
VC
NSM
CAN
I5
PVC
LUA
OLQ
DA
DD
VD
HSN
DVB
RIS
AVK
PVP
AVL
RMD
SMD
SMB
VB
RME
SIA
AIM
AIY
AVG
AVH
DVC
RIM
AVA
AVE
AVM
PVD
IL2
BAG
CEP
IL1
PVM
AIN
PVQ
RIC
ASER
ASK
AWA
ASEL
AWB
ADL
ASG
ASI
AFD
AWC
PHA
RIA
0.8
0.6
0.4
0.2
0

Fig. 2

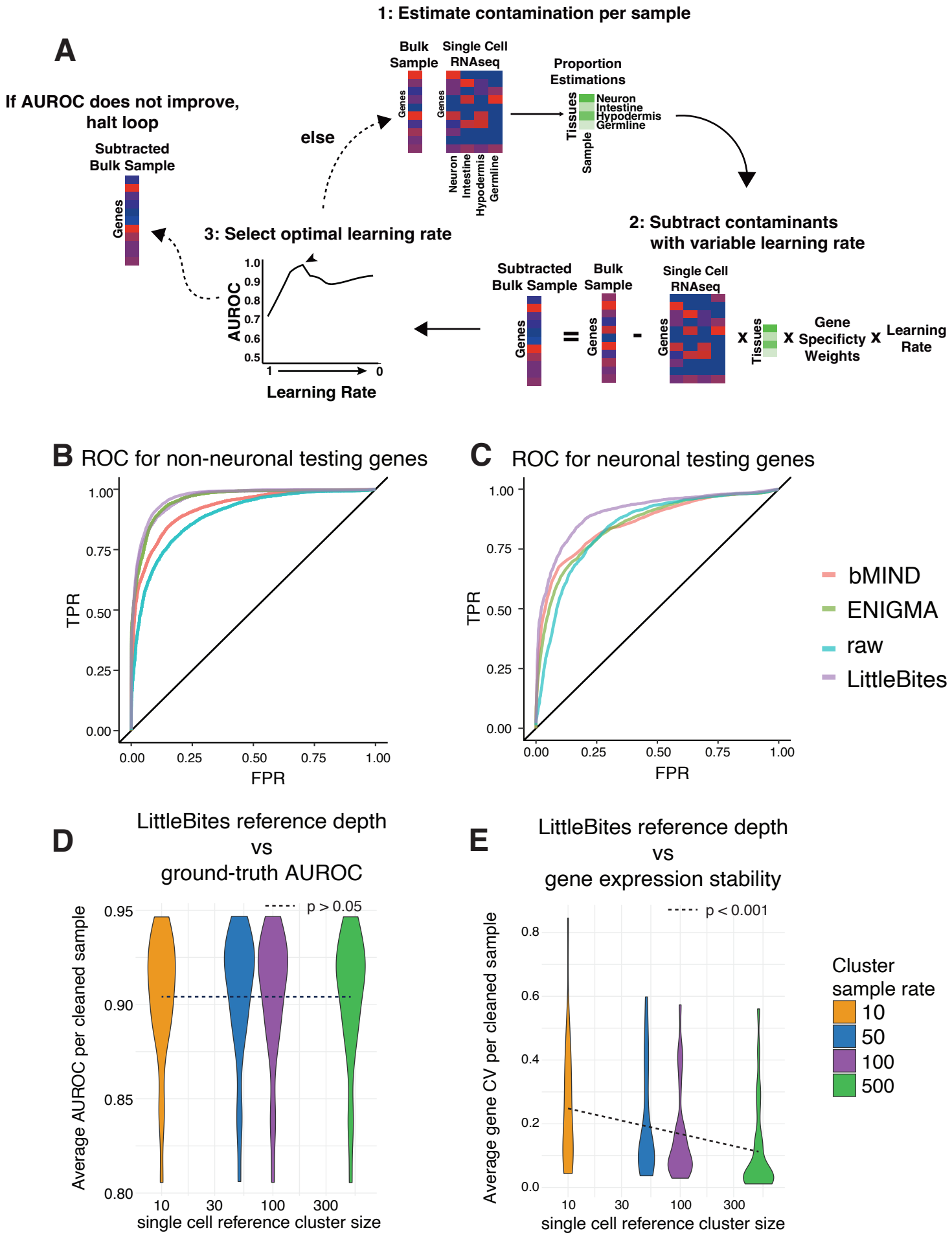
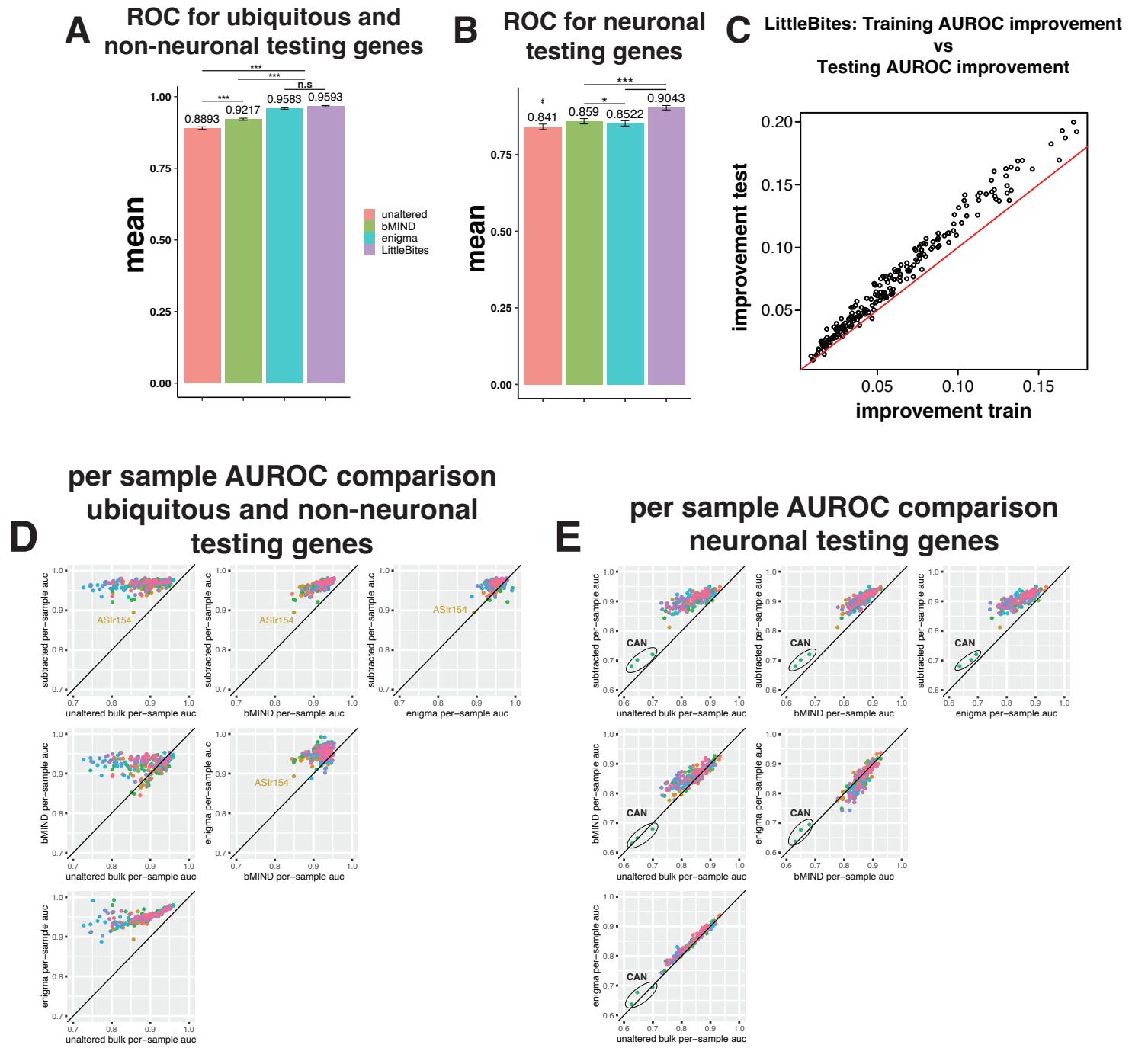

A
1: Estimate contamination per sample
Bulk Sample
Single Cell RNAseq
Genes
Proportion Estimations
Tissues
Neuron
Intestine
Hypodermis
Germline
Sample
If AUROC does not improve, halt loop
else
Subtracted Bulk Sample
3: Select optimal learning rate
AUROC
Learning Rate
2: Subtract contaminants with variable learning rate
Subtracted Bulk Sample
Bulk Sample
Single Cell RNAseq
Gene Specificty Weights
Learning Rate
B ROC for non-neuronal testing genes
C ROC for neuronal testing genes
TPR
FPR
bMIND
ENIGMA
raw
LittleBites
D
LittleBites reference depth vs ground-truth AUROC
Average AUROC per cleaned sample
p > 0.05
single cell reference cluster size
E
LittleBites reference depth vs gene expression stability
Average gene CV per cleaned sample
p < 0.001
single cell reference cluster size
Cluster sample rate
10
50
100
500

Fig. S2


A
ROC for ubiquitous and non-neuronal testing genes
mean
0.8893
0.9217
0.9583
0.9593
n.s
unaltered
bMIND
enigma
LittleBites
B
ROC for neuronal testing genes
0.841
0.859
0.8522
0.9043
C
LittleBites: Training AUROC improvement vs Testing AUROC improvement
improvement test
improvement train
D
per sample AUROC comparison ubiquitous and non-neuronal testing genes
ASlr154
subtracted per-sample auc
unaltered bulk per-sample auc
bMIND per-sample auc
enigma per-sample auc
E
per sample AUROC comparison neuronal testing genes
CAN

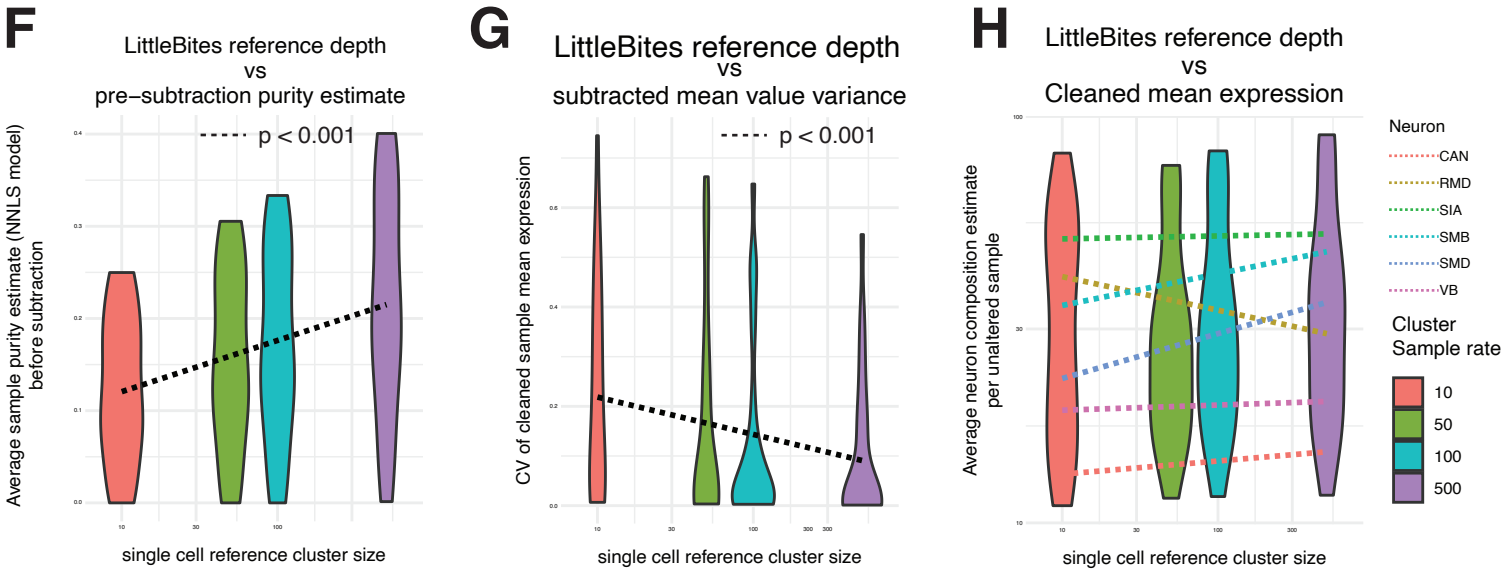

F
LittleBites reference depth vs pre-subtraction purity estimate
p < 0.001
Average sample purity estimate (NNLS model) before subtraction
single cell reference cluster size
G
LittleBites reference depth vs subtracted mean value variance
p < 0.001
CV of cleaned sample mean expression
single cell reference cluster size
H
LittleBites reference depth vs Cleaned mean expression
Average neuron composition estimate per unaltered sample
single cell reference cluster size
Neuron
CAN
RMD
SIA
SMB
SMD
VB
Cluster Sample rate
10
50
100
500

Fig. 3

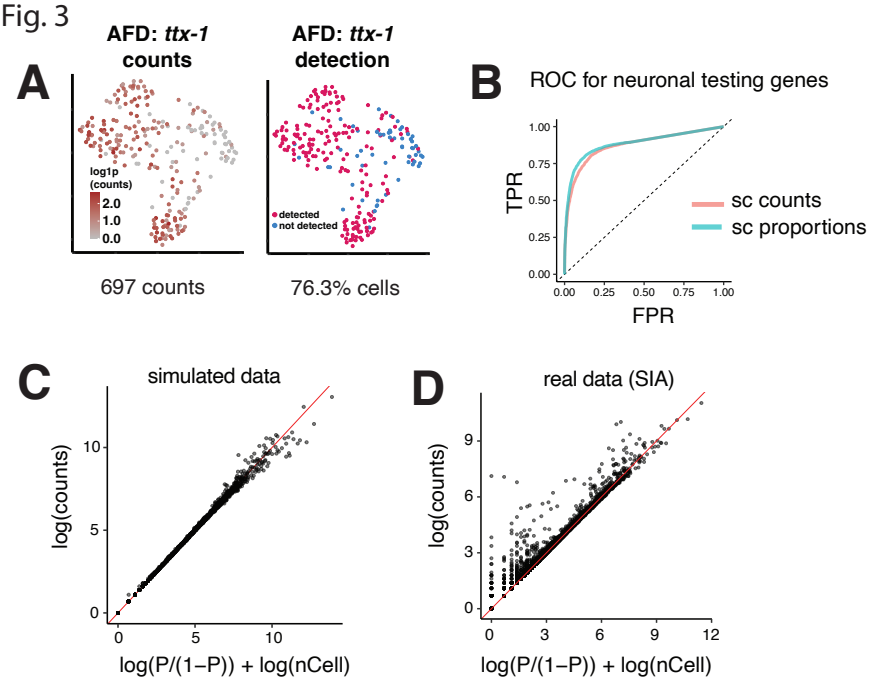

A
AFD: ttx-1 counts
AFD: ttx-1 detection
log1p (counts)
2.0
1.0
0.0
detected
not detected
697 counts
76.3% cells
B
ROC for neuronal testing genes
TPR
FPR
sc counts
sc proportions
C
simulated data
log(counts)
log(P/(1−P)) + log(nCell)
D
real data (SIA)
log(counts)
log(P/(1−P)) + log(nCell)

Fig. S3

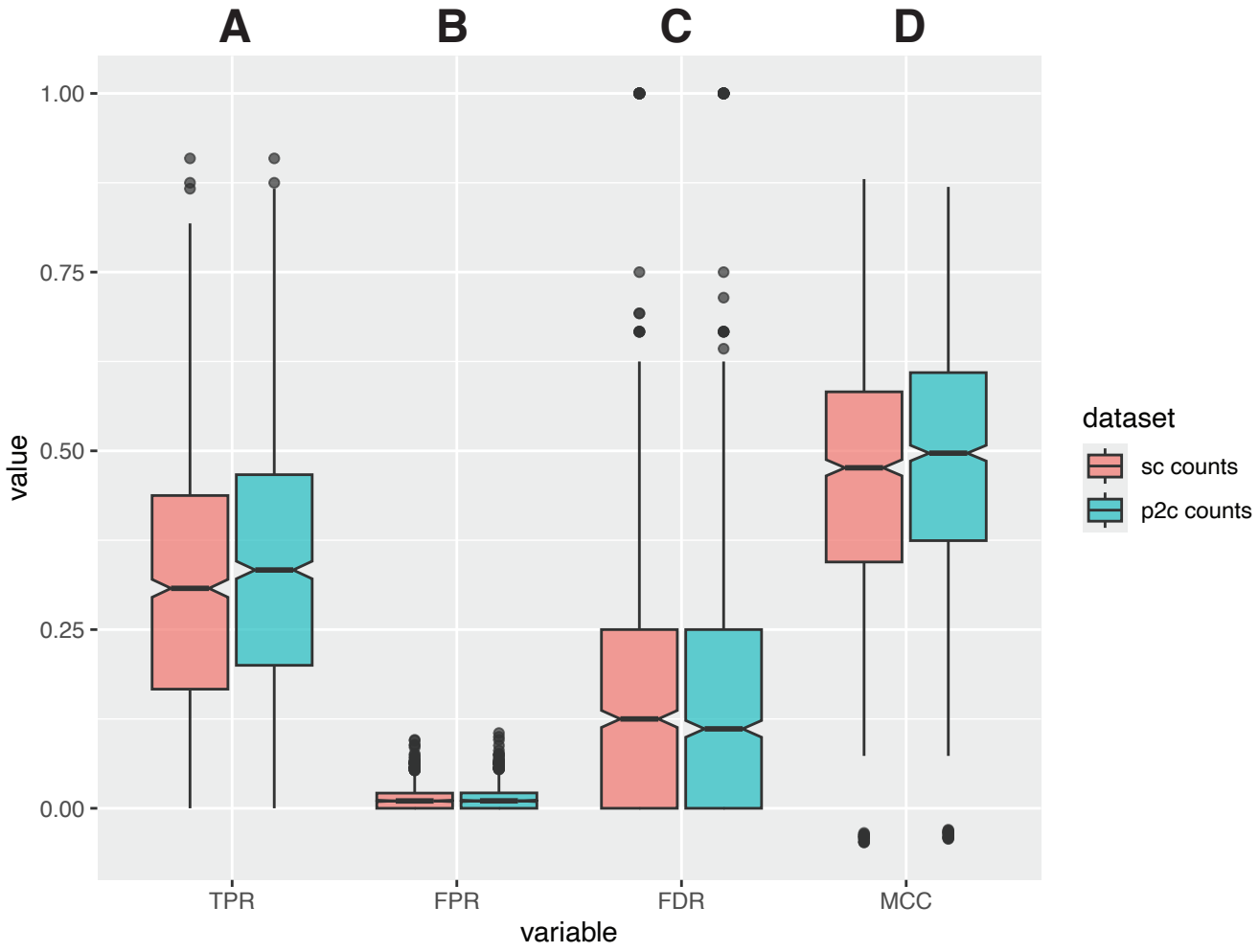

A
B
C
D
1.00
0.75
0.50
0.25
0.00
value
TPR
FPR
FDR
MCC
variable
dataset
sc counts
p2c counts

Fig. 4

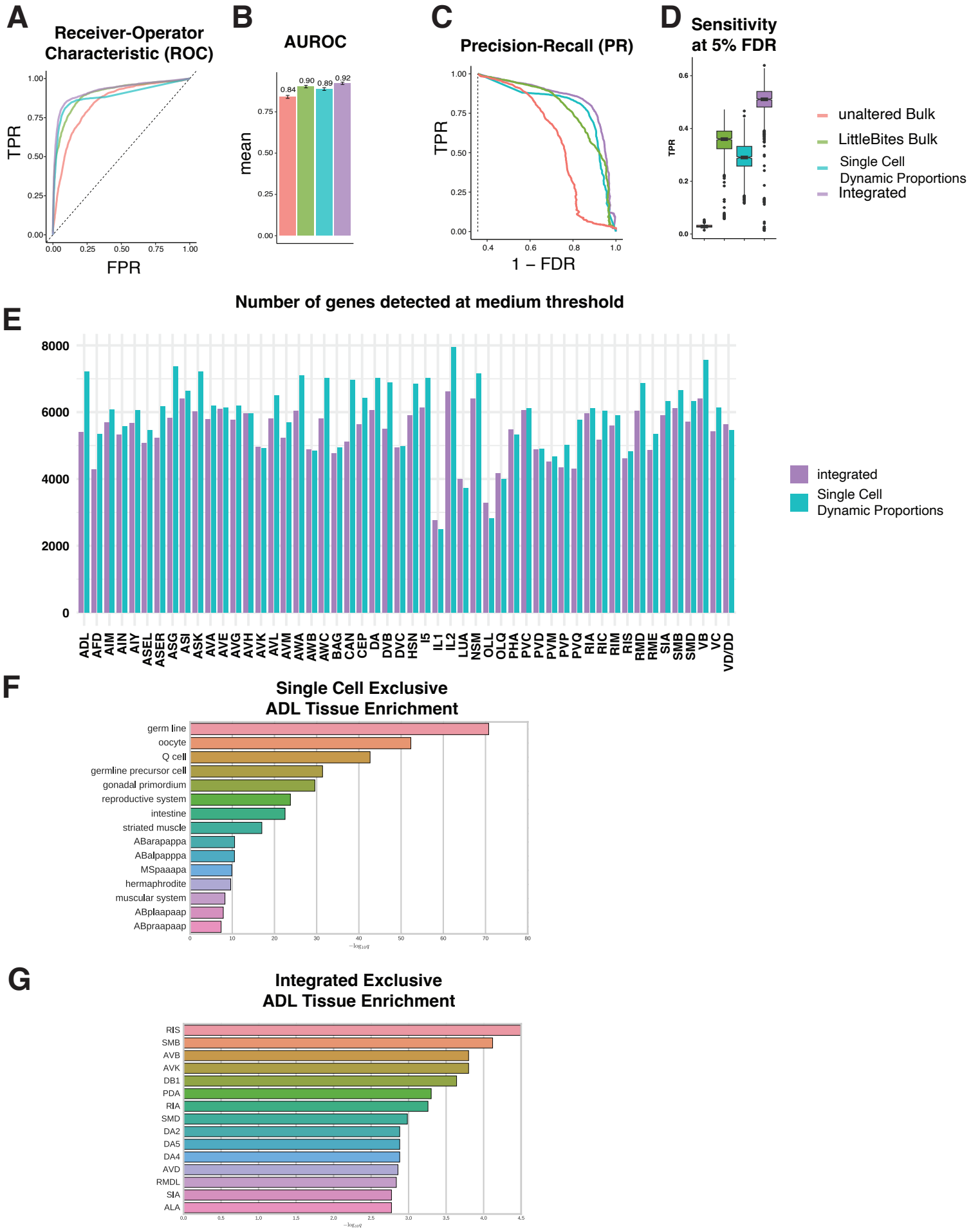

A
Receiver-Operator Characteristic (ROC)
TPR
FPR
B
AUROC
mean
0.84
0.90
0.89
0.92
C
Precision-Recall (PR)
TPR
1 – FDR
D
Sensitivity at 5% FDR
TPR
unaltered Bulk
LittleBites Bulk
Single Cell Dynamic Proportions
Integrated
E
Number of genes detected at medium threshold
integrated
Single Cell Dynamic Proportions
F
Single Cell Exclusive ADL Tissue Enrichment
germ line
oocyte
Q cell
germline precursor cell
gonadal primordium
reproductive system
intestine
striated muscle
ABarapappa
ABalpapppa
MSpaaapa
hermaphrodite
muscular system
ABplaapaap
ABpraapaap
G
Integrated Exclusive ADL Tissue Enrichment
RIS
SMB
AVB
AVK
DB1
PDA
RIA
SMD
DA2
DA5
DA4
AVD
RMDL
SIA
ALA

Fig. S4

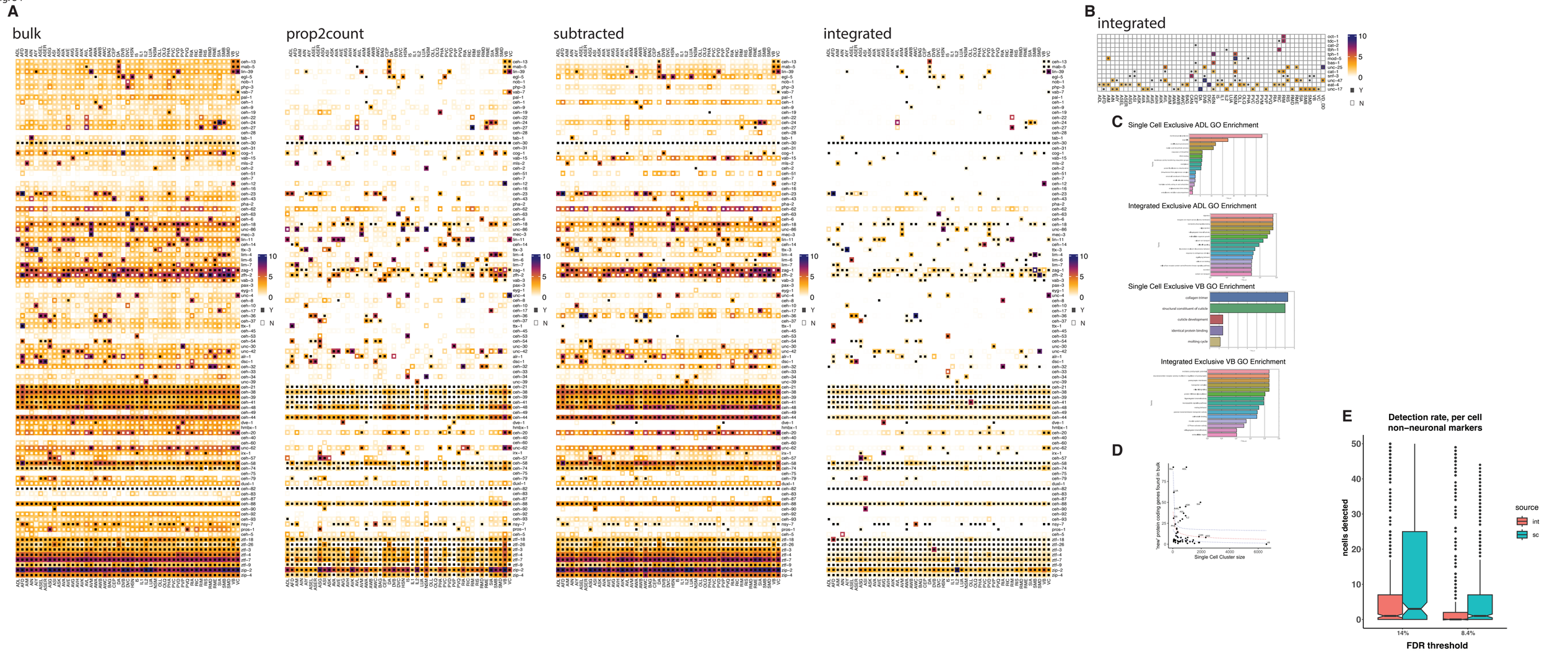

A
bulk
prop2count
subtracted
integrated
B
integrated
C
Single Cell Exclusive ADL GO Enrichment
Integrated Exclusive ADL GO Enrichment
Single Cell Exclusive VB GO Enrichment
collagen trimer
structural constituent of cuticle
cuticle development
identical protein binding
molting cycle
Integrated Exclusive VB GO Enrichment
D
'new' protein coding genes found in bulk
Single Cell Cluster size
E
Detection rate, per cell
non−neuronal markers
ncells detected
source
int
sc
14%
8.4%
FDR threshold

Fig. 6

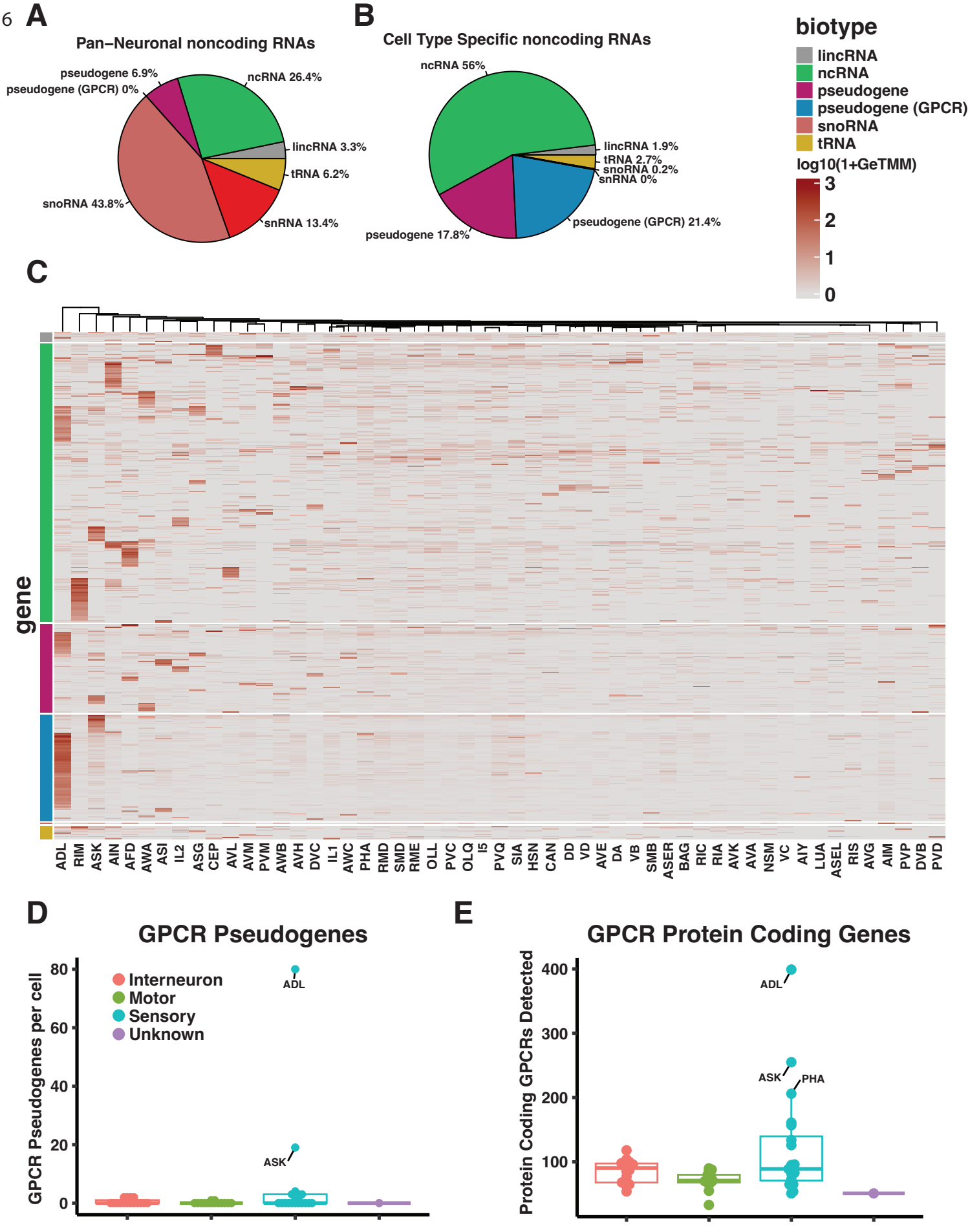
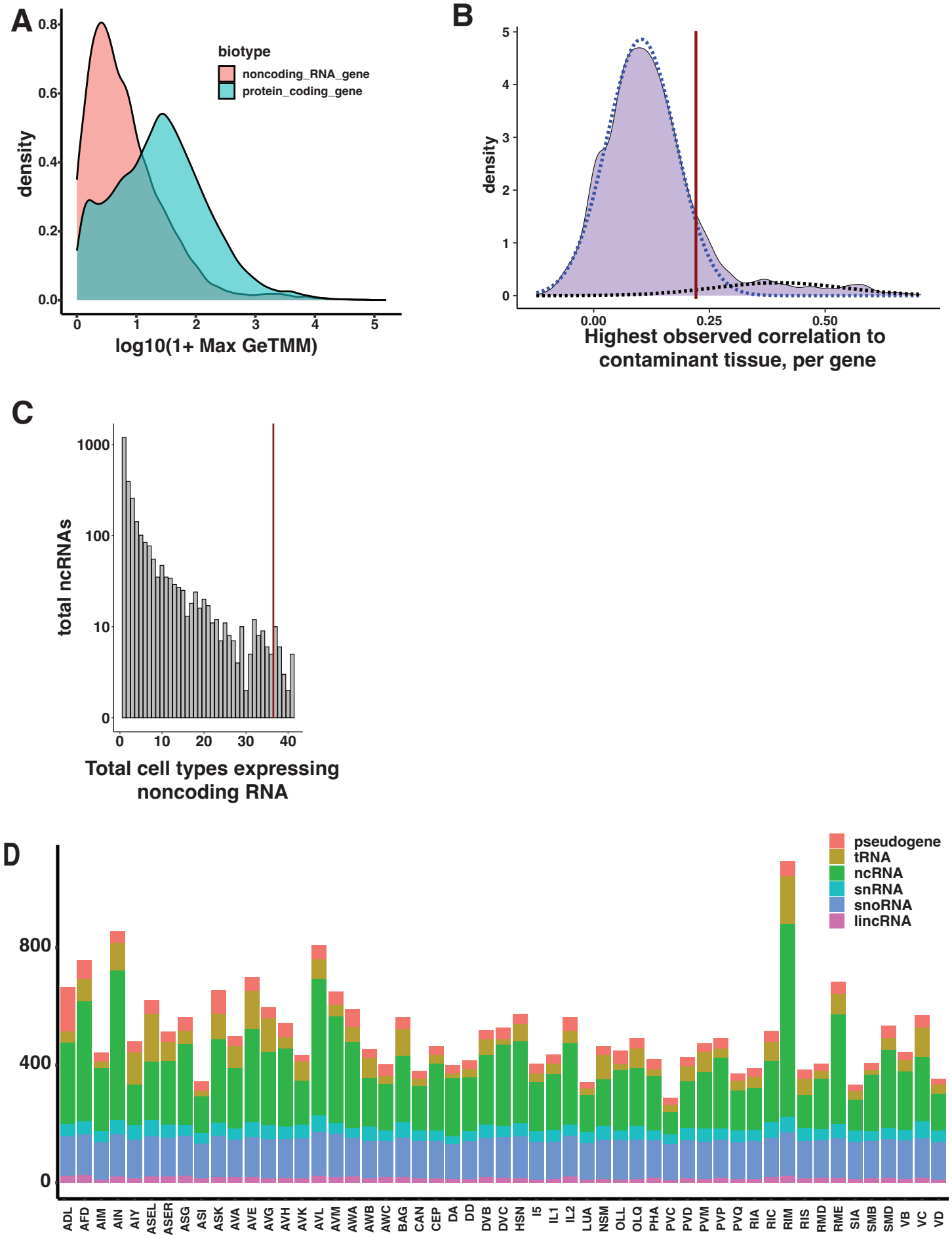

A
Pan–Neuronal noncoding RNAs
pseudogene 6.9%
pseudogene (GPCR) 0%
ncRNA 26.4%
lincRNA 3.3%
tRNA 6.2%
snRNA 13.4%
snoRNA 43.8%
B
Cell Type Specific noncoding RNAs
ncRNA 56%
lincRNA 1.9%
tRNA 2.7%
snoRNA 0.2%
snRNA 0%
pseudogene (GPCR) 21.4%
pseudogene 17.8%
biotype
lincRNA
ncRNA
pseudogene
pseudogene (GPCR)
snoRNA
tRNA
log10(1+GeTMM)
3
2
1
0
C
gene
ADL RIM ASK AIN AFD AWA ASI IL2 ASG CEP AVL AVM PVM AWB AVH DVC IL1 AWC PHA RMD SMD RME OLL PVC OLQ I5 PVQ SIA HSN CAN DD VD AVE DA VB SMB ASER BAG RIC RIA AVK AVA NSM VC AIY LUA ASEL RIS AVG AIM PVP DVB PVD
D
GPCR Pseudogenes
GPCR Pseudogenes per cell
80
60
40
20
0
Interneuron
Motor
Sensory
Unknown
ADL
ASK
E
GPCR Protein Coding Genes
Protein Coding GPCRs Detected
400
300
200
100
ADL
ASK
PHA

Fig. S6


A
biotype
noncoding_RNA_gene
protein_coding_gene
density
log10(1+ Max GeTMM)
B
density
Highest observed correlation to contaminant tissue, per gene
C
total ncRNAs
Total cell types expressing noncoding RNA
D
pseudogene
tRNA
ncRNA
snRNA
snoRNA
lincRNA
ADL AFD AIM AIN AIY ASEL ASER ASG ASI ASK AVA AVE AVG AVH AVK AVL AVM AWA AWB AWC BAG CAN CEP DA DD DVB DVC HSN I5 IL1 IL2 LUA NSM OLL OLQ PHA PVC PVD PVM PVP PVQ RIA RIC RIM RIS RMD RME SIA SMB SMD VB VC VD