Integrating bulk and single cell RNA-seq refines transcriptomic profiles of specific *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], Manasa Basaravaju[1,3], Abigail Poff[4], John A. Tipps[4], Maryam Majeed[5], Berta Vidal[5], Chen Wang[5], Eviatar Yemini[6], Emily A. Bayer[5], HaoSheng Sun[7], Oliver Hobert[5,8]*, David M. Miller III[4,9]*, 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, Columbia University, New York, NY, USA
[6] Neurobiology Department, University of Massachusetts Chan Medical School, Worcester, MA, USA
[7] Department of Cell, Developmental, and Integrative Biology, University of Alabama Birmingham, AL, USA
[8] Howard Hughes Medical Institute, Columbia University, New York, NY, USA
[9] 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 dataset for every anatomical neuron class in the *C. elegans* hermaphrodite. Here we present a complementary set of bulk RNA-seq samples for 41 of the 118 neuron classes in *C. elegans*. We show that the bulk dataset captures both lowly expressed and noncoding RNAs that are missed in the single cell dataset, but also includes false positives due to contamination by other cell types. We present an integrated analytical strategy that effectively resolves both the low sensitivity of single cell RNA-seq data and the reduced specificity of bulk RNA-Seq. We show that this integrated dataset enhances the sensitivity and accuracy of transcript detection and quantification of differentially expressed genes. We propose that our approach provides a new tool for interrogating gene expression, by bridging the gap between old (bulk) and new (single cell) methodologies for transcriptomic studies. We suggest that these datasets will advance the goal of delineating the mechanisms that define neuronal morphology and connectivity in *C. elegans*.

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.

To date, *C. elegans* is the only organism for which goal has been achieved; a gene expression map of the entire nervous system at the resolution of single neuron types. 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 (Brittin et al., 2021; Cook et al., 2019; Moyle et al., 2021; Sulston and Horvitz, 1977; Sulston et al., 1983; White et al., 1986). Recently, the *C. elegans* Neuronal Gene Expression Map & Network project (CeNGEN) (Hammarlund et al., 2018) 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 (Taylor et al., 2021).

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, noncoding RNAs in particular, are poorly represented in the CeNGEN scRNA-seq data. In addition, low abundance transcripts may be under-represented in scRNA-seq data, particularly in clusters with relatively few cells (Taylor *et al.*, 2021). 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 41 individual neuron types from the mature *C. elegans* hermaphrodite nervous system. 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 (Barrett et al., 2021). Importantly, we developed a novel computational approach to integrate the bulk dataset with the existing CeNGEN scRNA-seq dataset. Our new analytical strategy enhanced the accuracy and sensitivity of both data sets for profiles of each neuron type. The resultant integrated data set refines quantitative measures of gene expression and improves accuracy of differential expression calling between neuron types. 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 Supplementary Table S1.

FACS isolation for RNA-seq

Labeled neuron types were isolated for RNA-seq as previously described (Spencer et al., 2014; Taylor et al., 2021). Briefly, synchronized populations of L4 stage larvae were dissociated 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 below).

RNA extraction

RNA extractions were performed as previously described (Taylor *et al.*, 2021). 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.

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* (Barrett *et al.*, 2021). Libraries were sequenced on the Illumina Hiseq 2500 with 150 bp paired end reads. Reads were mapped to the *C. elegans* reference genome from WormBase (version WS281) using STAR (version 2.7.0) with the option `--outFilterMatchNminOverLread` 0.3. Duplicate reads were removed using NuDup (Tecan Genomics, version 2.3.3), and a counts matrix was generated using the featureCounts tool of SubRead (version 1.6.4). FASTQC was used for quality control before alignment, and four samples were removed for failing QC or for a low number of reads.

Pseudobulk aggregation of single-cell data

We downloaded CeNGEN scRNA-seq dataset as a Seurat object from the CeNGEN website (www.cengen.org). Cells from the same cell type and biological replicate (e.g. AFD cluster, replicate eat_4) were aggregated together by summation into a single pseudobulk sample if there were more than 10 cells in the single cell-type-replicate. For this work, single cell clusters of neuron subtypes were collapsed to the resolution of the bulk replicates (ex: VB and VB1 clusters in the single cell data were treated as one VB cluster).

Sample Normalization

Intra-sample normalization (gene length normalization for bulk samples) was performed before integration. 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 3.36.0). TMM Normalizations were performed separately for each

integrated matrix. For differential expression (Figure 3), bulk counts were used as input for integration, as edgeR uses unnormalized counts values as the input. For gene detection (Figure 1), bulk sample counts were normalized to gene length prior to integration, as this intra sample normalization shows improved accuracy for calling gene expression (Supplementary Figure 1C).

Integrating bulk and pseudobulk samples

We integrated bulk and single cell profiles by randomly pairing bulk samples and pseudobulk replicates for the same cell type, and then taking the geometric mean. A value of 0.1 was added to all pseudobulk data sets to obviate zero values (Equation 1). Our analysis was limited to cell types with at least 2 bulk samples and 2 pseudobulk replicates (supplementary table S3).

$$\text{Equation 1: I} = \frac{\log(\text{Bulk} + 0.1) + \log(\text{Pseudobulk} + 0.1)}{2}$$

The random pairing and integration step was performed 50 times. As an example: for AFD, we began with 5 bulk samples, and 3 pseudobulk replicates. For each integration, we randomly selected 3 bulk samples, and paired them with 3 pseudobulk replicates. Each pseudobulk replicate was then scaled to match the total counts in the corresponding bulk sample. Each AFD bulk-pseudobulk pairing was integrated by taking the geometric mean (with an added pseudo-count of 0.1), producing 3 integrated samples. This process was repeated 50 times, across all cell types, producing 50 separate integrated matrices (genes x integrated-replicates), sampling from all possible bulk-pseudobulk pairings across all cell types.

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. These studies used high

174 confidence fosmid fluorescent reporters, CRISPR strains or other methods
175 (Bhattacharya et al., 2019; Harris et al., 2019; Reilly et al., 2020; Stefanakis et
176 al., 2015; Taylor *et al.*, 2021; Yemini et al., 2021).
177 We also curated a list of 445 genes that are exclusively expressed outside
178 the nervous system to assess potential non-neuronal contamination in each
179 sample. This list was curated from published datasets of fluorescent reporters,
180 tissue specific RT-PCR, and transcriptomic studies available on WormBase
181 (Harris *et al.*, 2019). Genes were included if two forms of evidence both
182 suggested expression in the same non-neuronal tissue (non-overlap was allowed
183 so long as at least one tissue was consistent), and there was no evidence
184 available suggesting neuronal expression.
185 Ground truth gene expression is available in supplementary tables S5 &
186 S6.
187
188 **Comparing datasets to ground-truth**
189 When comparing bulk, single cell, and integrated data to “ground truth”
190 gene expression, a static threshold was applied to the average normalized cell
191 profile (arithmetic mean across all cells, or samples). Single cells were
192 normalized to library size prior to averaging to calculate TPM counts (Packer et
193 al., 2019). Bulk samples were normalized using the GeTMM method (Smid et al.,
194 2018), first normalizing to gene length, then to library size using a TMM
195 correction in edgeR (version 3.36.0). Each of the 50 integrated matrices were
196 separately normalized to library size, the average cell profile for each integrant
197 was calculated, then the 50 resultant genes x cell-types matrices were averaged.
198 The area under the curve (AUC) for the Receiver-Operator Characteristic (ROC)
199 and the Precision-Recall (PR) curves were calculated using the auc function with
200 the trapezoid option from the bayestestR package (version 0.11.5).
201
202 **Thresholding lowly expressed genes and noncoding genes**
203 For lowly expressed protein coding genes, and noncoding RNAs, genes
204 were called expressed in a cell type if more than 65% of replicates detect the

gene at or above the threshold. For lowly expressed genes, the threshold (73 normalized counts) was set to match the FDR (14%) for the published single cell analysis (Taylor *et al.*, 2021). For noncoding RNAs, the threshold was set at 5 normalized counts.

Proportion estimates

Contamination estimates were performed for each bulk sample by using non-negative least squares (NNLS) modeling on down-sampled and square root transformed counts, averaging across 100 estimates per sample. Down-sampling was performed to reduce bias against neuron types with small cluster sizes. For each sample (ex: AFD replicate 1), proportions were estimated using only neuronal cells for the corresponding single cell cluster (ex: AFD), and 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 square root transformed before the NNLS calculation. NNLS estimates across all 100 iterations were averaged for the final estimate. NNLS calculations were performed using the nnls package in R (version 1.4).

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, not expression in the target neuron. For genes passing an expression threshold > 2 normalized counts in at least 2 sample, 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 1.2.0), fitting 2 gaussian distributions to the distribution of highest

contaminant correlations. Cutoffs were selected to exclude 98% of the predicted contaminant distribution.

Differential expression and harmonic mean p combination

Differential expression was performed using the quasi-likelihood F-test approach in edgeR (glmQLFit and glmQLFTest functions). Each integrant dataset was fit and tested separately. P-values across integrated tests were treated as dependent, and were combined using the harmonic mean p approach, using the harmonicmeanp package in R (version 3.0) (Wilson, 2019). LogFC values were combined by taking the arithmetic mean across integrated tests. Consensus values were obtained by counting the number of iterations where a gene was called differentially expressed (P-value < 0.05). In the bulk dataset, genes were called differentially expressed if they had a P-value less than 0.05, and an absolute logFC greater than 2. In the integrated dataset, genes were called differentially expressed if they had a consensus value of at least 40 (P-value < 0.05 in 40 out of 50 separate tests), and an absolute average logFC greater than 2.

We used edgeR to perform pairwise differential expression analysis on each of the 50 integrated datasets separately, resulting in 50 edgeR comparisons per neuron pair. As these comparisons are not fully independent, we combined p-values across all 50 tests using the harmonic mean p procedure (Wilson, 2019). We also generated a consensus value based on how often a gene was called differentially expressed in the individual integrated comparisons ($p < 0.05$).

Ground-truth for differential expression

We adapted the binary ground-truth expression matrix to provide a ground truth for continuous differential expression analysis. For all neuron-neuron pairs, we subset the ground-truth genes to genes that are expressed in one of the two cells, and genes expressed in neither cell. We reasoned that genes called expressed in one cell but not the other in the ground truth data should predict differential expression when comparing continuous data from the two neurons.

We also reasoned that genes called unexpressed in both cells in the ground truth data should not be called differentially expressed when comparing continuous data. However, genes called expressed in both cell types in the binary ground-truth data are likely to be a mix of genes that are truly differentially expressed (eg low expression vs high expression), and genes that are not differentially expressed. Therefore, genes expressed in both cell types in the binary ground-truth data were excluded from this analysis. These ground-truth sets for differential expression were designed in a directional manner. For example, when comparing OLQ and PVD, we generated two sets of ground truth genes, and a separate TPR, FPR, and FDR are calculated for OLQ and PVD. For OLQ, the true genes are the genes called expressed in OLQ but not PVD in the ground-truth matrix. The false genes are the genes called unexpressed in both neurons *and* the genes called expressed in PVD alone (we expect those genes to be enriched in PVD, and thus if they are called enriched in OLQ they would be labeled false positives). Thus, we first calculate the genes enriched in OLQ, and compare them to what we expect to see enriched in OLQ, and we separately compare genes enriched in PVD to the genes that we expect to see in PVD.

Accuracy scores were calculated by adding up all true positive (TP) events, and all true negative (TN) events, and dividing by the total number of ground truth genes used (Equation 2).

$$\text{Equation 2: Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Matthew's Correlation Coefficient (MCC) is a metric for evaluating binary true/false classifications that is robust to imbalanced datasets (Chicco and Jurman, 2020; Jurman et al., 2012; Matthews, 1975) (Equation 3). This is useful for evaluating differential expression performance as the ground truth dataset is heavily biased towards actual false values.

$$\text{Equation 3: MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FN) * (TN + FP)}}$$

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 (**Fig. 1**). *C. elegans* is the first metazoan with a completely sequenced genome (Consortium, 1998) and the only animal for which we know every cell division that gives rise to the adult body plan (*i.e.,* "cell lineage") (Sulston and Horvitz, 1977; Sulston *et al.*, 1983), as well as the anatomy of each neuron and all of its connections with other cells (Brittin *et al.*, 2021; Cook *et al.*, 2019; Moyle *et al.*, 2021; Varshney et al., 2011; White *et al.*, 1986). The entire *C. elegans* hermaphrodite nervous system contains 302 neurons with 118 anatomically-defined neuron classes, each comprised of relatively few cells, ranging from 1 to 13 neurons (White *et al.*, 1986). 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. 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. However, the description of gene expression in this atlas is incomplete (Taylor *et al.*, 2021); (1) Lowly-expressed genes, particularly in clusters with few cells, may not be detected; and (2) Non-poly adenylated transcripts are excluded (Taylor *et al.*, 2021).

To address these limitations and to provide a broader description of gene expression across the nervous system, we used a bulk RNA sequencing strategy to profile different neuron types. We used a series of *C. elegans* strains, each of which uses one or more fluorescent markers to label an individual neuron type

for isolation by FACS. 21 individual neurons could be uniquely marked with a single neuron-specific promoter. For an additional 19 neuron types, we used an intersectional strategy involving different colored fluorophores to label each target neuron, and for 1 neuron we collected some samples with one fluorophore, and other samples with the intersection of two fluorescent markers (Table S1). For example, we used *flp-22*::GFP and *unc-47*::mCherry to mark the single neuron AVL (Figure 1A).

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 terminally differentiated (Sulston and Horvitz, 1977). Labeled cells were collected in TRIzol LS for RNA extraction (Figure 1B). We isolated a wide range of cells (~700 – 90,000) in each sample across neuron types. Multiple biological replicates (e.g., separately grown cultures) were generated for each neuron class. In total, we sequenced 160 samples across 41 neuron types (Figure 1A; Table S2). The 41 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 (Figure S1A). (A few of these bulk neuron profiles have been previously described, Taylor et al., 2021.)

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) (Barrett *et al.*, 2021). 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) vs motor/interneurons, a result consistent with patterns observed for scRNA-seq data on the same neuron classes (Figure 1B) (Taylor *et al.*, 2021).

A strategy for integrating bulk and single-cell data to improve gene detection accuracy

Bulk RNA-seq and scRNA-seq datasets have complementary strengths and weaknesses. Bulk RNA-seq can enhance sequencing depth and gene detection, capture non-polyadenylated transcripts, and result in uniform coverage of the transcript body (Barrett *et al.*, 2021). Bulk RNA-seq data are typically contaminated, however, 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 (Taylor *et al.*, 2021).

Recent studies have exploited the strengths of these complementary approaches, i.e., the depth of bulk RNA sequencing and the specificity afforded by scRNA-seq, for downstream analysis. These approaches primarily focused on the problem of deconvolution, seeking to infer cell-type expression profiles from tissue level bulk samples, using scRNA-seq references as a guide (Newman et al., 2019; Wang et al., 2021a; Wang et al., 2021b; Zhu et al., 2018). By contrast, our dataset contains bulk RNA-seq reads for individual cell types isolated by FACS, exactly matching cell types identified as scRNA-seq clusters. Thus, our data present an opportunity to directly integrate bulk and scRNA-seq profiles for individual cell types, with the goal of combining both datasets to increase depth and accuracy.

We constructed pseudobulk samples from the scRNA-seq data for the subset of overall neuron types represented in our bulk RNA-seq data set. Each pseudobulk sample was generated by aggregating scRNA-seq data from individual biological replicates for each annotated cell type. For example, for the AFD cluster, we generated 3 pseudobulk samples, each containing cells from a different single cell experiment, with cell numbers ranging from 27 to 141, and total read counts across all genes ranging from 28,781 to 126,778 (Table S3). We adopted the approach of generating separate pseudobulk data sets for scRNA-seq data from independent single cell experiments because biological

replicates have been shown to improve the accuracy of differential expression analysis of scRNA-seq datasets (Crowell et al., 2020; Squair et al., 2021; Thurman et al., 2021).

For integrating bulk and scRNA-seq data sets, we adopted the straightforward approach of calculating the geometric mean for each transcript of randomly paired bulk and pseudobulk replicates. This pairing was performed across 50 iterations to sample all possible bulk-pseudobulk arrangements and averaged for comparison to the ground truth genes (see Methods for details, Figure 2A).

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

Accurately detecting gene expression (distinguishing between true signal vs noise) is a central goal for RNA-seq experiments. We first set out to assess our bulk datasets by comparison to ground truth genes (see Methods, Supplementary Table S5) (Taylor *et al.*, 2021). We also used published expression data to curate a list of 445 ground truth genes in non-neuronal cells that are likely not expressed in neurons (Supplementary Table S6).

For the bulk, scRNA-seq, and integrated datasets, expression calling was performed by setting a single threshold at the average normalized counts values for each cell type. Thus, all genes in all cell types that meet or exceed the threshold are called “expressed”, and all genes in all cells that fall below the threshold are called “unexpressed”. These binary expression values were then compared to the ground-truth datasets for neuronal and non-neuronal cells. This treatment determined that the bulk samples show a high (FPR) (False Positive Rate) versus combined ground truth genes for neuron and non-neuronal cells across all thresholds (Figure 2B-D). These results suggest that bulk data set contains non-neuronal transcripts from a low level of contaminating cells in the FACS preparation. By contrast, the clustering algorithms used to generate the scRNA-seq data (before pseudobulk aggregation) effectively exclude unwanted cell types and thus result in fewer false positives in the scRNA-seq data.

Interestingly, at relatively low precision (or high FPR), the bulk data approached a TPR (True Positive Rate) of 100% (Figure 2B-D). By contrast, the scRNA-seq pseudobulk data peak at a 91.9% TPR, suggesting that the single cell dataset fails to detect some genes. Together, this analysis indicates that bulk and single-cell approaches both afford robust approximations of gene expression, but that they have different disadvantages: the bulk approach is prone to contaminating data from other cell types, whereas the single-cell approach is limited in detection.

Measured against the neuronal ground truth genes, the integrated dataset shows a similar sensitivity to the bulk data at low thresholds, while matching the scRNA-seq ratio of specificity and sensitivity across most thresholds and improving on the scRNA-seq performance for some thresholds (Figure 2B-C). The scRNA-seq data still outperforms the integrated dataset for non-neuronal ground truth genes, but the integrated dataset performs nearly as well at thresholds above 10 normalized counts (Figure 2D). 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 and high specificity across a wide range of thresholds.

Integration of bulk and single-cell data enhances the accuracy of differential expression analysis.

To determine the effect of integration on the accuracy of differential expression analysis, we compared differential expression (DE) analysis of our bulk vs integrated data sets. For both cases, we performed DE analysis for all possible pairwise combinations of different neuron types (595 in total). Genes were called differentially expressed in bulk data for p-values < 0.05 and an absolute value log2 fold-change (logFC) > 2, (i.e. 4-fold enrichment) in either cell type. Genes were scored as differentially expressed in the integrated data if they were called significant in at least 40/50 iterations (consensus >= 40) and had an average absolute value logFC > 2 (see Methods).

We scored the accuracy of differential expression of bulk and integrated data by comparison to neuronal ground truth data. For each pair of neuron types A and B, the ground truth data give rise to one of four possible outcomes for each gene: (1) expressed in both neurons A and B; (2) not expressed in either neuron; (3) expressed in A only; (4) expressed in B only. We assessed accuracy in a directional fashion, such that we examined separately genes called expressed only in A and genes called expressed only in B. For example, for genes called expressed only in A, true positives are ground truth genes with expression only in A, whereas false positives include ground truth genes with expression only in B, as well as ground truth genes that are not expressed in either cell. (Ground truth genes expressed in both neurons A and B were excluded as they could correspond to genes that are not truly differentially expressed between the two cell types). Non-neuronal ground truth genes were used to calculate a separate FPR.

We calculated TPR, FPR, and FDR (False Discovery Rate) values for every pair of neurons in both the bulk and integrated datasets. In addition, we calculated Accuracy scores (total true calls / all possible calls, see Methods), and the Matthew's Correlation Coefficient (MCC) (the Pearson product-moment correlation coefficient of the observed and expected results, see Methods) (Chicco and Jurman, 2020) (Figure S3C-F). These results indicate that the integrated dataset is more accurate overall than the bulk dataset (Figure 3A). In addition, on for each neuron-neuron pair, integration results in more improvement than degradation in differential expression accuracy (mean = 0.026, 95.conf.int ± 0.003) (Figure 3B). A similar relationship was observed for MCC scores. Specifically, the number of comparisons with MCC scores near 0 was lower in the integrated data set (figure 3C), which represents the expected performance of a coin toss (Chicco and Jurman, 2020). The difference in MCC scores for each neuron-neuron pair also showed higher scores in the Integrated dataset (mean = 0.089, 95.conf.int ± 0.010). Together, these analyses indicate that integration improves the accuracy of differential gene expression.

Next, we examined whether integration could improve differential expression analysis even when scRNA-seq data are limited. The lowest abundance single cell clusters show reduced gene detection (Taylor *et al.*, 2021), suggesting that they might not perform as well for integration. Of the 41 neuron types for which we performed bulk sequencing, PVD and OLQ were the neuron types with the fewest cells per cluster in the single cell dataset (62 cells and 85 cells, respectively). In the bulk data, a majority of the genes expected to be enriched in PVD from the neuronal ground truth dataset are correctly called, but none of the expected OLQ genes are called enriched. For example, the gene *gar-1* is expected to be enriched in OLQ but is instead enriched in PVD in the bulk RNA-seq data. After integration, *gar-2* is called enriched in OLQ, and all but one gene that was enriched in PVD or showed mild enrichment in PVD now show mild enrichment towards OLQ, though only *gar-2* passes both the logFC and significance cutoffs (Figure 3F). Considering all true positive genes for both PVD and OLQ, we see a modest increase in the TPR for this comparison (Figure 3G), along with a sharp drop in the FPR for neuronal ground-truth genes (Figure 3H), and non-neuronal ground-truth genes (Figure 3I). Similar results were observed for other comparisons (Figure S3H-K, although there are also rare instances in which integration decreased the TPR (Figure S3J). Thus, integration with scRNA-seq data improves the accuracy of differential gene expression in bulk RNA samples, even when scRNA-seq data are limited.

Non-neuronal contamination in FACS-isolated neuronal Bulk RNA-seq samples varies between samples, and between cell types (Figure S4A-B). This variance could lead to non-neuronal genes being erroneously called significantly enriched in some neuron-neuron comparisons. Most neuron-neuron comparisons in both the bulk and integrated datasets show low but detectable false positive rates for non-neuronal ground truth genes (Figure S3Gi-ii). In addition, some neuron-neuron comparisons in the bulk dataset show low specificity scores for non-neuronal ground truth genes, suggesting that differences in non-neuronal contamination are influencing differential expression calling (Figure S3Giii). The integrated dataset shows much higher specificity scores for the same neuron-

neuron pairs, and modest specificity improvements overall (Figure S3Giv). When comparing I5 and BAG neurons in the bulk analysis, 26.3% of non-neuronal ground truth genes are called enriched in either I5 or BAG. In the integrated analysis, only 4.5% of non-neuronal genes are called enriched in either cell type. We conclude that our analysis of the systematic pairwise differential expression among all cell types shows that integration improves differential expression by reducing false positives, both for genes expressed in the nervous system and non-neuronal genes, while maintaining the overall true positive rate.

Bulk sequencing powers detection of low-abundance transcripts

scRNA-seq analysis of the *C. elegans* neuronal transcriptome generated a map of protein coding gene expression for a total of 128 transcriptionally distinct neuron types. However, this map contains some false negatives—ground truth genes that are known to be expressed in the neuron type but are not detected in the scRNA-seq data. Two factors that contribute to these dropouts are low gene expression and small cluster size (clusters with few neurons tend to detect fewer genes) (Mereu et al., 2020; Taylor *et al.*, 2021).

We tested whether bulk RNA-seq data might provide this missing information. We collected a minimum of 701 cells per bulk sample (Table S1), and sequenced each sample to high depth, suggesting that even low-expressed genes might be represented in bulk data. A comparison of protein coding genes between bulk and single-cell data showed a mean Spearman coefficient of 0.612 (95.conf.int ± 0.027), with a sharp drop off in the Spearman coefficient for the smallest single cell clusters (Figure 4A). (This analysis used all protein-coding genes detected in a minimum of 3 cells in the single cell dataset.) This result matches previous analysis of the scRNA-seq data, which showed that gene detection is reduced for clusters with < 500 cells (Taylor *et al.*, 2021). Together these results indicate that bulk data contain gene expression information that is missing from scRNA-seq clusters that contain few cells.

Although bulk sequencing typically includes lowly-expressed genes, at least some of them may represent false positives derived from non-neuronal tissue contamination. Since these genes are typically not included in the scRNA-seq data, the integration strategy described above does not ameliorate this problem. Previous studies have shown that correlations between gene expression and tissue level proportion estimates can be used to deconvolve the profiles of multiple tissues from one mixed bulk profile (Wang *et al.*, 2021a). We utilized a similar approach to enrich for genes that are truly expressed in our cell types of interest. First, we estimated contamination in each bulk sample using a non-negative least squares regression (NNLS). We used 100 bootstraps to reduce bias against lowly abundant single cell clusters (see Methods, Supplementary Figure S4A-B). We then calculated per-gene Spearman correlations to each contaminant type (e.g., the correlation of *pgl-1* to reproductive cell contamination across all samples). We validate this approach by observing that contaminant correlations for non-neuronal ground-truth genes are higher than the contaminant correlations for all other protein coding genes (Figure S4C). Using the highest correlation per gene, we modeled this data as a mixture of two Gaussian distributions, one distribution of low contamination correlation scores representing truly expressed neuronal genes, and a second distribution of higher contamination correlation scores representing genes likely present due to contamination from non-neuronal tissues. (Figure S4D). Setting a threshold which removes all genes with a contaminant correlation higher than 0.3 excludes 98% of the predicted contaminant distribution profile.

Using this decontaminated data, we tested our detection of poorly represented genes. We first interrogated the expression of all genes that are detectable in scRNA-seq experiments, by virtue of being called expressed in at least one cell type (by thresholding on the proportion of cells detecting the gene, see Methods). We tested whether our decontaminated bulk data might provide evidence for expression in additional neuron types. Using a minimum normalized count threshold in the bulk data to match the FDR of “threshold 2” from the published single cell analysis (Taylor *et al.*, 2021), we detected 5 to 169 genes

per cell type that were missed in that single cell cluster (mean = 36.9, 95.conf.int ± 9.4) (Figure 4B). Plotting the number of newly detected genes against the single cell cluster size reveals that bulk sequencing detects more protein coding genes for cell types with low coverage in the single cell dataset vs cell types with larger numbers of cells in each cluster (Figure 4C). We used GO term enrichment to evaluate genes called expressed in bulk that were missing in the scRNA-seq data. Most cell types show enrichment for neuron-associated terms, chiefly neuropeptide signaling (Figure 4D, S3). Several cell types also show enrichment for synaptic signaling, dendritic morphology, and receptor regulator activity. Thus, we detect genes in the bulk dataset that are missing from some single cell clusters with the greatest improvement biased towards clusters with low coverage in the scRNA-seq dataset.

Next, we tested whether decontaminated bulk data might yield expression information about genes that were undetected in the scRNA-seq dataset. Thresholding the scRNA-seq data results in 3,567 protein coding genes that are identified as not expressed in all cell types, including non-neuronal tissues (see Methods). Additionally, 873 protein coding genes were excluded from analysis in the scRNA-seq dataset because they were detected in fewer than 3 of the 100,955 cells sequenced. We combined these gene sets to generate a list of 4,440 ‘unexpressed’ genes that were not detected in the single cell analysis (Supplementary table S7).

To examine expression of these unexpressed genes in the bulk data, we first ‘decontaminated’ the data by removing genes with strong correlations to any contaminants as described above. We used the non-neuronal ground-truth genes to set a minimum normalized counts threshold for calling expression, which was set to a non-neuronal FPR of 0%. Using this threshold on the remaining decontaminated unexpressed genes, we detected between 9 and 150 protein coding genes per cell type (mean = 25.9, 95.conf.int = ± 7.8) (Figure 4E). Using ADL as an example, we performed Tissue Enrichment Analysis on the 150 new genes (Angeles-Albores et al., 2016). The most enriched term is “ADL genes”, as expected, followed by the “amphid sensillum” and “lateral ganglion”, structures

603 that include the ADL neuron (Figure 4F) (Inglis et al., 2007). Thus, these results
604 suggest that our analysis of bulk data reveals truly expressed genes that were
605 not detected by scRNA-seq.
606
607 **Bulk RNA-seq reveals both broadly expressed and neuron-specific**
608 **noncoding RNAs**
609 A significant benefit of our bulk RNA-seq approach is its sensitivity to non-
610 poly-adenylated transcripts, which include many species of non-coding RNA
611 (Barrett *et al.*, 2021). However, we do not have a ground-truth data set of non-
612 coding genes to evaluate accuracy. In addition, most non-coding RNAs are
613 expressed at lower levels than protein coding genes, making it unreasonable to
614 apply a static threshold using the protein coding FDR (Figure S5A). Thus, we
615 opted to apply a uniform threshold for “expressed” genes and selected the
616 criteria of > 5 normalized counts in at least 65% of samples within a cell type. We
617 again used gene level correlation to contamination estimates as a procedure to
618 eliminate genes that were likely detected due to contamination from other tissues
619 in the bulk samples. First, we estimated contamination for each sample using a
620 bootstrapped NNLS regression (see Methods, Supplementary figure S4A-B), and
621 then calculated per-gene Spearman correlations to each contaminant type. We
622 applied a threshold on the gene level correlation to contamination estimates for
623 each sample by fitting a Gaussian mixture model to the maximum correlation
624 score for each gene. We selected a cutoff of 0.23, which excludes 98% of the
625 estimated contamination distribution (Figure 5A). With these thresholds, an
626 average of 603 noncoding RNAs were identified as “expressed” per cell type (95
627 CI ± 54.5). By RNA type, we detected 23.0 ± 1.7 lincRNAs, 55.6 ± 7.1
628 pseudogenes, 62.6 ± 12.5 tRNAs, 49.3 ± 2.1 snRNAs, 148.9 ± 2.4 snoRNAs, and
629 266.6 ± 39.1 uncategorized ncRNAs per cell type (Figure 5B).
630 Next, we sought to identify noncoding RNAs with broad expression across
631 multiple neuron types. This approach detected 266 non-coding genes that are
632 called expressed in > 90% of neuron classes defined by bulk RNA-seq (Figure
633 5C, D). These broadly expressed noncoding RNAs, include 128 (48%) snoRNAs

and 37 (13.9%) snRNAs, both tenfold greater than the expected proportion assuming a random distribution (Fisher's exact test, P-value < 0.01) (Figure S5B). In contrast, pseudogenes and otherwise uncategorized ncRNAs were significantly depleted (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 (Fafard-Couture et al., 2021; Isakova et al., 2020), and is consistent with their key roles in rRNA processing and splicing (Bratkovič et al., 2019; Valadkhan, 2013; Wassarman and Steitz, 1992).

We also sought to identify cell-type-specific noncoding RNAs. We calculated tissue specificity scores for each noncoding RNA called expressed in at least one cell type using the Preferential Expression Measure (PEM) score (Huminiecki et al., 2003; Kryuchkova-Mostacci and Robinson-Rechavi, 2016). We called these genes cell-type specific according to three criteria: (1) Called expressed in $\geq$ one cell type (see above); (2) PEM score > 0.65; (3) > 2 normalized counts in a maximum of 10/41 cell types. Using these thresholds, we identified 561 cell-type-specific noncoding RNAs (Figure 5E). By RNA type, 347 (61.8%) of cell type-specific noncoding RNA genes are uncategorized ncRNAs, 186 (33.2%) are pseudogenes, 15 (2.6%) are tRNAs, 8 (1.4%) are lincRNAs, 3 (0.5%) are snoRNAs, and 2 (0.3%) are snRNAs (Figure S5C). We observed significant enrichment of pseudogenes, and a subtle but significant depletion of ncRNAs, snoRNAs, and tRNAs (P-value < 0.01). Clustering by genes and cell type modalities revealed clear enrichment for noncoding RNAs in individual neuron types (Figure 5F). The number of specific noncoding RNAs per cell type ranged from 0 (PVC) to 120 (ADL), with a mean of 14 (± 8.5) (Supplementary Table S8). These data 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 41 neuron classes or about 1/3 of all known neuron types in the *C. elegans* nervous system (Figure 1A-B). We describe a new method of integrating these bulk RNA-seq data with previously obtained single-cell RNA-seq data (Taylor et al., 2021) that improves gene detection accuracy for both data sets (Figure 1D-F). Integrated data sets also outperform the original bulk samples in accurately calling differential gene expression across all pairwise comparisons (Figure 3), with a clear reduction in false positives (Figure S3D, G). With the rapid growth of scRNA-seq atlases that complement bulk RNA-seq datasets for individual tissues, our results offer a timely and useful opportunity to improve the accuracy of cell and tissue-specific transcriptional profiles. Furthermore, our computational integration approach is general and can be applied to combine additional sequencing modalities to further incorporate complementary gene expression signals to amplify the depth of sequencing.

In addition to enhancing the accuracy of differential gene expression, the integrated bulk RNA-seq dataset detects lowly expressed protein coding genes that were not detected by scRNA-seq (Figure 4B-C,E) and thus could reveal new drivers of neuron-specific traits. 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 (Barrett *et al.*, 2021; Taylor *et al.*, 2021) (Figure 5B). Some of these noncoding RNAs are broadly expressed in the nervous system (Figure 5C-D) which is suggestive of shared functions across different types of neurons. Interestingly, a subset of non-coding RNAs are expressed in a limited number of neuron types (Figure 5E-F) pointing to potentially important roles in determining key neuron-specific functions. In addition, the bulk RNA-seq dataset contains transcript information across the gene body, which might yield information about mRNA splicing that is not found in the scRNA-seq dataset.

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 also 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.

Supplementary Tables

Supplementary tables S1-8 are available on figshare (https://doi.org/10.6084/m9.figshare.19522096.v1).

Data Availability

Bulk raw data are in the process of being posted at GEO, and the linking information will be posted to the CeNGEN website when available. Single cell raw data are available at Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo, GEO: GSE136049). Counts data and additional supporting files can be downloaded from the CeNGEN website (https://www.cengen.org) and code is available at GitHub (https://www.github.com/cengenproject).

DECLARATION OF INTERESTS

The authors declare no competing interests.

Figure legends

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 replicates 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: Bulk RNA sequencing encompasses a broad range of neuron types and correlates with scRNA-seq results. 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) (Taylor *et al.*, 2021).

Figure 2: Integrating bulk RNA-seq and scRNA-seq data sets improves gene detection accuracy. A) Individual pseudobulk scRNA-seq replicates and bulk RNA-seq samples from the same neuron type (NSM neuron samples illustrated) are randomly paired and integrated (50X for each neuron type) using the geometric mean (see Methods) to generate 50 integrated matrices (genes x integrated-replicate). The average integrated profile was used to call gene expression. Pairwise neuron-neuron differential expression (edgeR) was performed for each of the 50 integrated matrices which were then combined to generate consensus sets of differentially expressed genes. Bulk RNA-seq datasets are used to identify genes that are not detected in scRNA-seq data,

including noncoding RNAs and lowly expressed mRNAs. B) 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). 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) The non-neuronal FPR across a range of thresholds for bulk, single-cell, and integrated datasets compared to non-neuronal ground-truth genes. The x-axis shows the $\log_{10}$-transformed threshold used for each point; the y-axis shows the FPR. A pseudocount of 1 was added for the $\log_{10}$-transformation. Each point represents a static threshold applied to all genes in all samples (e.g., expressed >= 10 normalized counts); Bulk RNA-seq data (green), scRNA-seq (blue), average integrated data (red).

Supplementary Figure 2: Intra-sample normalization improves the FPR for non-neuronal genes in bulk RNA-seq samples: The non-neuronal FPR across a range of thresholds for bulk RNA-seq datasets with different normalizations compared to non-neuronal ground-truth genes. The x-axis shows the log10 transformed threshold for each point, the y-axis shows the FPR. Each point represents a static threshold applied to all genes in all samples (e.g., expressed >= 10 normalized counts). Bulk data with only inter-sample normalization using TMM factors (trimmed mean of M-values, used by edgeR) (green) vs bulk data with both intra-sample and inter-sample normalization (GeTMM) (red). AUC = Area Under Curve. A pseudocount of 1 was added for the $\log_{10}$-transformation.

Figure 3: Integrated samples show improved accuracy in detecting differentially expressed genes. A) Density histograms of the accuracy score for all pairwise differential expression comparisons in bulk RNA-seq (blue) vs integrated (orange) datasets. B) Density histogram of the difference (integrated

minus bulk) in the accuracy score for each pairwise differential expression comparison, vertical dashed line at 0 represents no difference between the datasets. C) Density histograms of the Matthew's Correlation Coefficient (MCC) score for all pairwise differential expression comparisons in the bulk RNA-seq (blue) vs integrated (orange) datasets. D) Density histogram of the difference (integrated minus bulk) in the MCC score for each pairwise differential expression comparison, vertical dashed line at 0 represents no difference between the datasets. E) Volcano plot for the differential expression profile of the bulk RNA-seq PVD samples vs OLQ samples. Dots represent individual genes. X-axis is log2 fold change (logFC), and the Y-axis is -log10(P-value). Grey dots are genes that are not called significant, and black dots are genes that pass significance thresholds (P-value < 0.05, and |logFC| > 2, red lines). F) Volcano plot for the differential expression profile of the Integrated PVD samples vs OLQ samples. X-axis is the log2 fold change (logFC), and the Y-axis is the -log10(harmonic mean P value) (p.hmp). Grey dots are genes that are not called significant, and black dots are genes that pass significance thresholds (P-value < 0.05 in >=80% of edgeR runs across all 50 integrations, and |logFC| > 2). Magenta squares mark genes expected to be enriched in PVD from the neuronal ground-truth dataset, and orange triangles denote genes expected to be enriched in OLQ. *gar-1* and *gar-2* are expected to be enriched in OLQ. G) Bar plot showing the differential expression True Positive Rate (TPR) for genes expected to be expressed in OLQ or PVD but not both. H) Bar plot showing the differential expression false positive rate (FPR) for genes expected to be expressed in neither OLQ nor PVD, and genes that were called enriched in the wrong neuron type. I) Bar plot showing the differential expression FPR for genes expected to be expressed only in non-neuronal tissues.

Supplementary Figure 3 A) Table showing an example ground-truth matrix for OLQ and PVD neurons. Here we expect Gene a to be differentially enriched in OLQ over PVD, so it would be considered a positive ground-truth for OLQ and would be used to calculate the TPR for OLQ vs PVD. All other genes

807 shown are expected not to be enriched in OLQ and would thus be used as
808 negative ground-truth, to calculate the FPR and FDR for OLQ vs PVD. When
809 calculating the ground-truth for PVD vs OLQ, we expect Gene b to be enriched in
810 PVD, and so it is treated as a positive ground-truth gene, and all other genes
811 shown are treated as negative ground-truth. B) Example heatmap showing the
812 MCC score for directional OLQ and PVD differential expression. In the OLQ row,
813 we use edgeR to compare genes enriched in OLQ vs expected enrichment using
814 the ground truth data. In the PVD row, we perform the same function, looking
815 instead for enrichment in PVD. Thus, we have 595 neuron-neuron comparisons,
816 with two entries for each pair. For OLQ vs PVD, we have an OLQ entry showing
817 the scores for genes enriched in OLQ, and a PVD entry showing the scores for
818 genes enriched in PVD. C-G) Heatmaps and density plots, showing scores for
819 differential expression compared to neuronal ground-truth genes (C-F) and non-
820 neuronal genes (G) across all neuron types. C) Recall, D) Specificity (1-FPR), E)
821 Accuracy, F) MCC score, and G) non-neuronal specificity. i) Heatmap of the
822 score for the Bulk samples. ii) Heatmap of the score for the Integrated samples.
823 iii) Heatmap of the difference in the scores (Integrated minus Bulk). iv) Density
824 plot for the difference in the scores, black line at 0 indicates no difference
825 between integrated and bulk comparisons. H-J) Bar plots showing neuronal
826 ground-truth TPR and FPR, and the non-neuronal FPR, for four pairs of neurons.
827 All bar graphs of TPR and FPR are shown for both directions of the comparison.
828
829 **Figure 4: Bulk RNA-seq samples detect protein coding genes that are**
830 **not detected in scRNA-seq clusters:** A) Scatter plot showing the relationship
831 between the size of a scRNA-seq cluster (i.e., the number of cells in the cluster)
832 and the Spearman correlation between the average bulk RNA-seq profile and the
833 average scRNA-seq for all protein coding genes. Each dot represents one cell
834 type. Red dashed line shows a Michaelis-Menten fit (see Methods), gmax =
835 0.675, beta = 29.507. Blue dashed lines show the 97.5% confidence interval of
836 the fit. B) Bar plot showing the number of protein coding genes detected per cell
837 type in the bulk dataset. Genes plotted are: 1) called unexpressed in the

corresponding single cell cluster; 2) have a maximum correlation to any contaminant tissue less than 0.3; and 3) are expressed above 73 normalized counts in the average bulk profile for that cell type. C) Scatter plot showing the relationship between the size of a scRNA-seq cluster and the number of additional protein coding genes detected per cell type (as defined in panel B). Each dot represents one cell type. Red dashed line shows an exponential decay fit (see Methods), M = 140.2, m = 26.5, alpha = 89.1. Blue dashed lines show the 97.5% confidence interval of the fit. D) GO enrichment analysis for protein coding genes detected in bulk IL1 samples that were not detected in the IL1 scRNA-seq cluster. GO enrichment performed using WormBase. E) Bar plot showing the number of protein coding genes detected per cell type in the bulk dataset. Restricted to genes that are never called expressed in any scRNA-seq cluster, have a contaminant correlation less than 0.3, and are expressed above 16 normalized counts (determined by setting the non-neuronal FPR threshold to 0). F) Tissue enrichment analysis for protein coding genes detected in the ADL bulk samples but never called expressed in any scRNA-seq cluster (Angeles-Albores *et al.*, 2016).

Supplementary Figure 4 A-B) Scatter plots with a linear fit showing the relationship between the $\log_{10}$ transformed single cell cluster size and the estimated neuronal proportion of each bulk sample. Estimates were made using an NNLS regression (non-negative least squares, see Methods). A) Estimates with all single cells in each cluster. Neuronal proportion = 0.081 * log10(sc_size) + 0.149. R^2 = 0.05489, p = 0.001666 B). Estimates taken from the average Neuronal proportion estimate across 100 bootstraps, down-sampled to 30 cells for all clusters before each bootstrap. Neuronal proportion = 0.029 * log10(sc_size) + 0.268. R^2 = 0.003752, p = 0.2079. C) Density plot of the gene level correlation to contaminant estimates. Only the highest correlation per gene is used. Distribution for all protein coding genes (red) vs distribution for non-neuronal ground-truth protein coding genes (blue). D) Density plot of the gene level correlation to contaminant estimates for all genes that are detected in single

cell but called unexpressed in one of the 41 cell types covered by bulk sequencing. Only the highest correlation per gene is used. Blue and black dashed lines represent a Gaussian mixture model, used to threshold against contaminant genes. Red line at 0.3 indicates the cutoff, all protein coding genes with a maximum correlation above 0.3 were removed from analysis. E-H) GO term enrichment plots for genes called expressed in the bulk dataset which were called unexpressed in the corresponding scRNA-seq cluster for neurons OLL (E), RIS (F), PVD (G) and PVM (H).

Figure 5: Bulk analysis reveals noncoding RNA expression pattern: A) 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. B) Stacked bar graph showing the number of noncoding RNAs called expressed in each neuron type. Colors represent RNA classes. Genes were called expressed in a cell if they were detected above 5 normalized counts in greater than 65% of samples for that cell. C) Bar plot showing number of cell types in which each noncoding RNA is detected. The x axis shows the number of cells, and the y axis shows the number of genes detected in that many cells. Genes to the right of the red line are called expressed in more than 90% of the sequenced cell types. D) Heatmap of log transformed GeTMM values of the pan-neuronal genes identified in panel C, columns are annotated by neuron modality. E) Histogram showing the distribution of Preferential Enrichment Measure (PEM) scores per gene, a metric for cell type specificity. Genes are considered cell type specific if they have a PEM greater than 0.65 (red line) and are expressed above 2 normalized counts in fewer than 10 cell types. F) Heatmap of average normalized counts per cell type, for genes considered cell type specific, columns are annotated by neuron modality, and rows are grouped by RNA class.

Supplementary Figure 5 A) Density plot showing the relative expression of noncoding RNAs (purple) and protein coding RNAs (orange), x-axis is maximum normalized counts per gene. B) Pie chart showing proportions of classes of pan-neuronal noncoding RNAs. C) Pie chart showing proportions of cell type specific noncoding RNAs. D) Box plot showing the number of cell type specific noncoding RNAs per cell type, grouped by neuron modality.

Supplementary Table S1: All cell types sorted for bulk RNA-seq experiments, with the strain names and allele information.

Supplementary Table S2: Replicate metadata for bulk RNA-seq experiments, with replicate names, strain names, and the number of cells collected.

Supplementary Table S3: All cell types used for integrating bulk and single cell RNA-seq data, with the number of replicates in the bulk and single cell datasets for each cell type.

Supplementary Table S4: Metadata for each single cell replicate, including the replicate name, the total UMI counts in the replicate, the number of individual cells included in the replicate, the cell type, and the experimental replicate name.

Supplementary Table S5: Ground Truth expression for 160 genes in the *C. elegans* nervous system using fosmid and CRISPR/Cas reporter lines (see methods).

Supplementary Table S6: Ground Truth expression for 445 genes that are expressed exclusively outside the *C. elegans* nervous system, curated from published data (see methods).

Supplementary Table S7: Genes called unexpressed in all single cell clusters.

Supplementary Table S8: An annotated heatmap of highly specific noncoding RNA genes and their log10 transformed expression values in each cell type.

References

Angeles-Albores, D., N. Lee, R.Y., Chan, J., and Sternberg, P.W. (2016). Tissue enrichment analysis for C. elegans genomics. BMC Bioinformatics *17*, 366. 10.1186/s12859-016-1229-9.

Barrett, A., McWhirter, R., Taylor, S.R., Weinreb, A., Miller, D.M., III, and Hammarlund, M. (2021). A head-to-head comparison of ribodepletion and polyA selection approaches for Caenorhabditis elegans low input RNA-sequencing libraries. G3 Genes|Genomes|Genetics *11*. 10.1093/g3journal/jkab121.

Bhattacharya, A., Aghayeva, U., Berghoff, E.G., and Hobert, O. (2019). Plasticity of the Electrical Connectome of C. elegans. Cell *176*, 1174-1189.e1116. 10.1016/j.cell.2018.12.024.

Bratkovič, T., Božič, J., and Rogelj, B. (2019). Functional diversity of small nucleolar RNAs. Nucleic Acids Research *48*, 1627-1651. 10.1093/nar/gkz1140.

Brittin, C.A., Cook, S.J., Hall, D.H., Emmons, S.W., and Cohen, N. (2021). A multi-scale brain map derived from whole-brain volumetric reconstructions. Nature *591*, 105-110. 10.1038/s41586-021-03284-x.

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics *21*, 6. 10.1186/s12864-019-6413-7.

Consortium, T.C.e.S. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012-2018. 10.1126/science.282.5396.2012.

Cook, S.J., Jarrell, T.A., Brittin, C.A., Wang, Y., Bloniarz, A.E., Yakovlev, M.A., Nguyen, K.C.Q., Tang, L.T.H., Bayer, E.A., Duerr, J.S., et al. (2019). Whole-animal connectomes of both Caenorhabditis elegans sexes. Nature *571*, 63-71. 10.1038/s41586-019-1352-7.

Crowell, H.L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., and Robinson, M.D. (2020). muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nature Communications *11*, 6077. 10.1038/s41467-020-19894-4.

Fafard-Couture, É., Bergeron, D., Couture, S., Abou-Elela, S., and Scott, M.S. (2021). Annotation of snoRNA abundance across human tissues reveals complex snoRNA-host gene relationships. Genome Biology *22*, 172. 10.1186/s13059-021-02391-2.

Hammarlund, M., Hobert, O., Miller, D.M., 3rd, and Sestan, N. (2018). The CeNGEN Project: The Complete Gene Expression Map of an Entire Nervous System. Neuron *99*, 430-433. 10.1016/j.neuron.2018.07.042.

Harris, T.W., Arnaboldi, V., Cain, S., Chan, J., Chen, W.J., Cho, J., Davis, P., Gao, S., Grove, C.A., Kishore, R., et al. (2019). WormBase: a modern Model Organism Information Resource. Nucleic Acids Research *48*, D762-D767. 10.1093/nar/gkz920.

Huminiecki, L., Lloyd, A.T., and Wolfe, K.H. (2003). Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. BMC Genomics *4*, 31. 10.1186/1471-2164-4-31.

Inglis, P.N., Ou, G., Leroux, M.R., and Scholey, J.M. (2007). The sensory cilia of Caenorhabditis elegans. WormBook, 1-22. 10.1895/wormbook.1.126.2.

Isakova, A., Fehlmann, T., Keller, A., and Quake, S.R. (2020). A mouse tissue atlas of small noncoding RNA. Proceedings of the National Academy of Sciences *117*, 25634-25645. 10.1073/pnas.2002277117.

Jurman, G., Riccadonna, S., and Furlanello, C. (2012). A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. PLOS ONE *7*, e41882. 10.1371/journal.pone.0041882.

Kryuchkova-Mostacci, N., and Robinson-Rechavi, M. (2016). A benchmark of gene expression tissue-specificity metrics. Briefings in Bioinformatics *18*, 205-214. 10.1093/bib/bbw008.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure *405*, 442-451. https://doi.org/10.1016/0005-2795(75)90109-9.

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Álvarez-Varela, A., Batlle, E., Sagar, Grün, D., Lau, J.K., et al. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nature Biotechnology *38*, 747-755. 10.1038/s41587-020-0469-4.

Moyle, M.W., Barnes, K.M., Kuchroo, M., Gonopolskiy, A., Duncan, L.H., Sengupta, T., Shao, L., Guo, M., Santella, A., Christensen, R., et al. (2021). Structural and developmental principles of neuropil assembly in C. elegans. Nature *591*, 99-104. 10.1038/s41586-020-03169-5.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology *37*, 773-782. 10.1038/s41587-019-0114-2.

Packer, J.S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., et al. (2019). A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. Science *365*. 10.1126/science.aax1971.

Reilly, M.B., Cros, C., Varol, E., Yemini, E., and Hobert, O. (2020). Unique homeobox codes delineate all the neuron classes of C. elegans. Nature *584*, 595-601. 10.1038/s41586-020-2618-9.

Smid, M., Coebergh van den Braak, R.R.J., van de Werken, H.J.G., van Riet, J., van Galen, A., de Weerd, V., van der Vlugt-Daane, M., Bril, S.I., Lalmahomed, Z.S., Kloosterman, W.P., et al. (2018). 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. 10.1186/s12859-018-2246-7.

Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nature Communications *12*, 5692. 10.1038/s41467-021-25960-2.

Stefanakis, N., Carrera, I., and Hobert, O. (2015). Regulatory Logic of Pan-Neuronal Gene Expression in C. elegans. Neuron *87*, 733-750. 10.1016/j.neuron.2015.07.031.

Sulston, J.E., and Horvitz, H.R. (1977). Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev Biol *56*, 110-156. 10.1016/0012-1606(77)90158-0.

Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol *100*, 64-119. 10.1016/0012-1606(83)90201-4.

Taylor, S.R., Santpere, G., Weinreb, A., Barrett, A., Reilly, M.B., Xu, C., Varol, E., Oikonomou, P., Glenwinkel, L., McWhirter, R., et al. (2021). Molecular topography of an entire nervous system. Cell *184*, 4329-4347.e4323. https://doi.org/10.1016/j.cell.2021.06.023.

Thurman, A.L., Ratcliff, J.A., Chimenti, M.S., and Pezzulo, A.A. (2021). Differential gene expression analysis for multi-subject single-cell RNA-sequencing studies with aggregateBioVar. Bioinformatics *37*, 3243-3251. 10.1093/bioinformatics/btab337.

Valadkhan, S. (2013). Chapter Six - The Role of snRNAs in Spliceosomal Catalysis. In Progress in Molecular Biology and Translational Science, G.A. Soukup, ed. (Academic Press), pp. 195-228. https://doi.org/10.1016/B978-0-12-381286-5.00006-8.

Varshney, L.R., Chen, B.L., Paniagua, E., Hall, D.H., and Chklovskii, D.B. (2011). Structural properties of the Caenorhabditis elegans neuronal network. PLoS Comput Biol *7*, e1001066. 10.1371/journal.pcbi.1001066.

Wang, J., Roeder, K., and Devlin, B. (2021a). Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data. Genome Research *31*, 1807-1818. 10.1101/gr.268722.120.

Wang, W., Yao, J., Wang, Y., Zhang, C., Tao, W., Zou, J., and Ni, T. (2021b). Improved estimation of cell type-specific gene expression through deconvolution of bulk tissues with matrix completion. bioRxiv, 2021.2006.2030.450493. 10.1101/2021.06.30.450493.

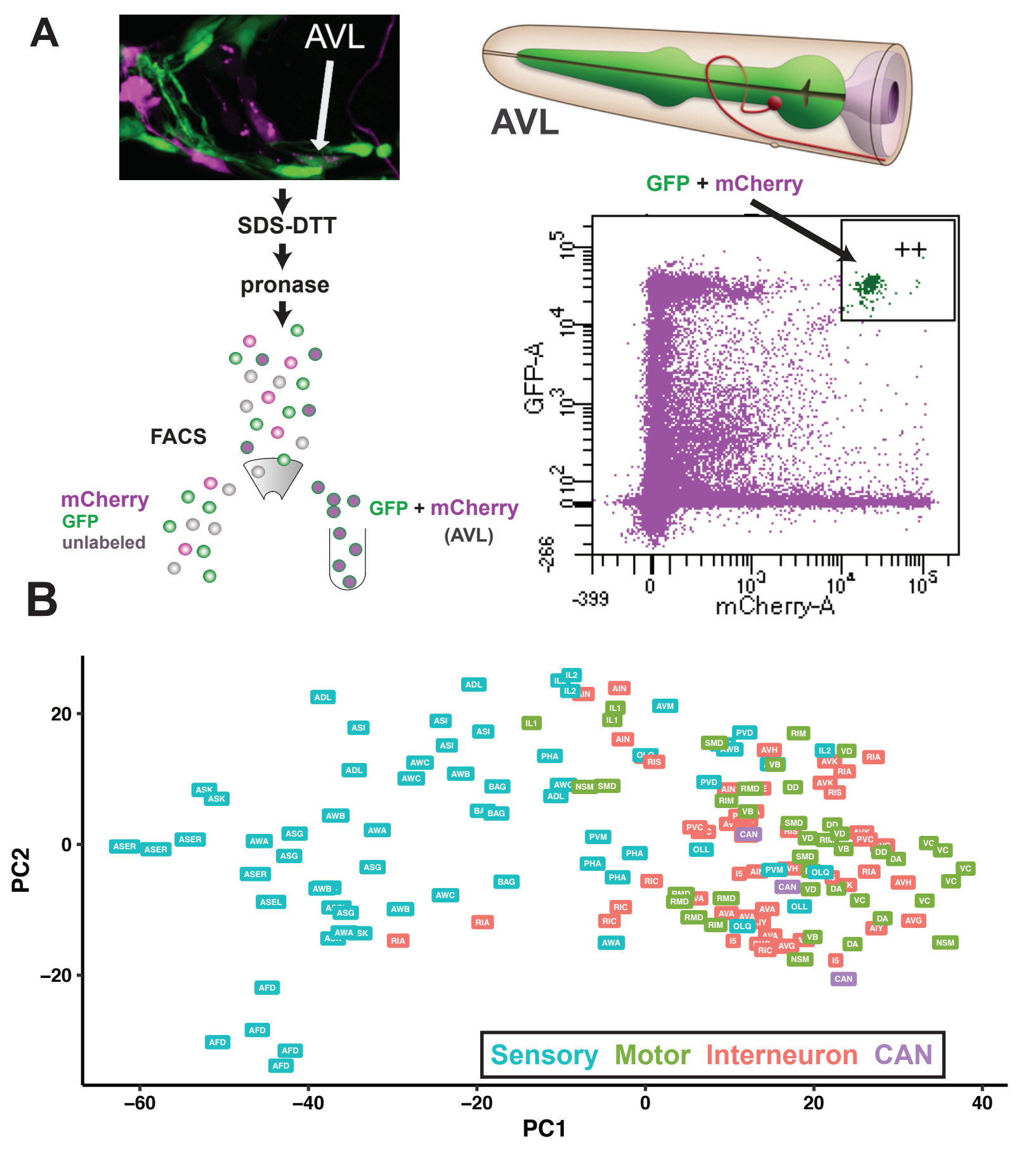
Wassarman, D.A., and Steitz, J.A. (1992). INTERACTIONS OF SMALL NUCLEAR RNAS WITH PRECURSOR MESSENGER-RNA DURING INVITRO SPLICING. Science *257*, 1918-1925. 10.1126/science.1411506.

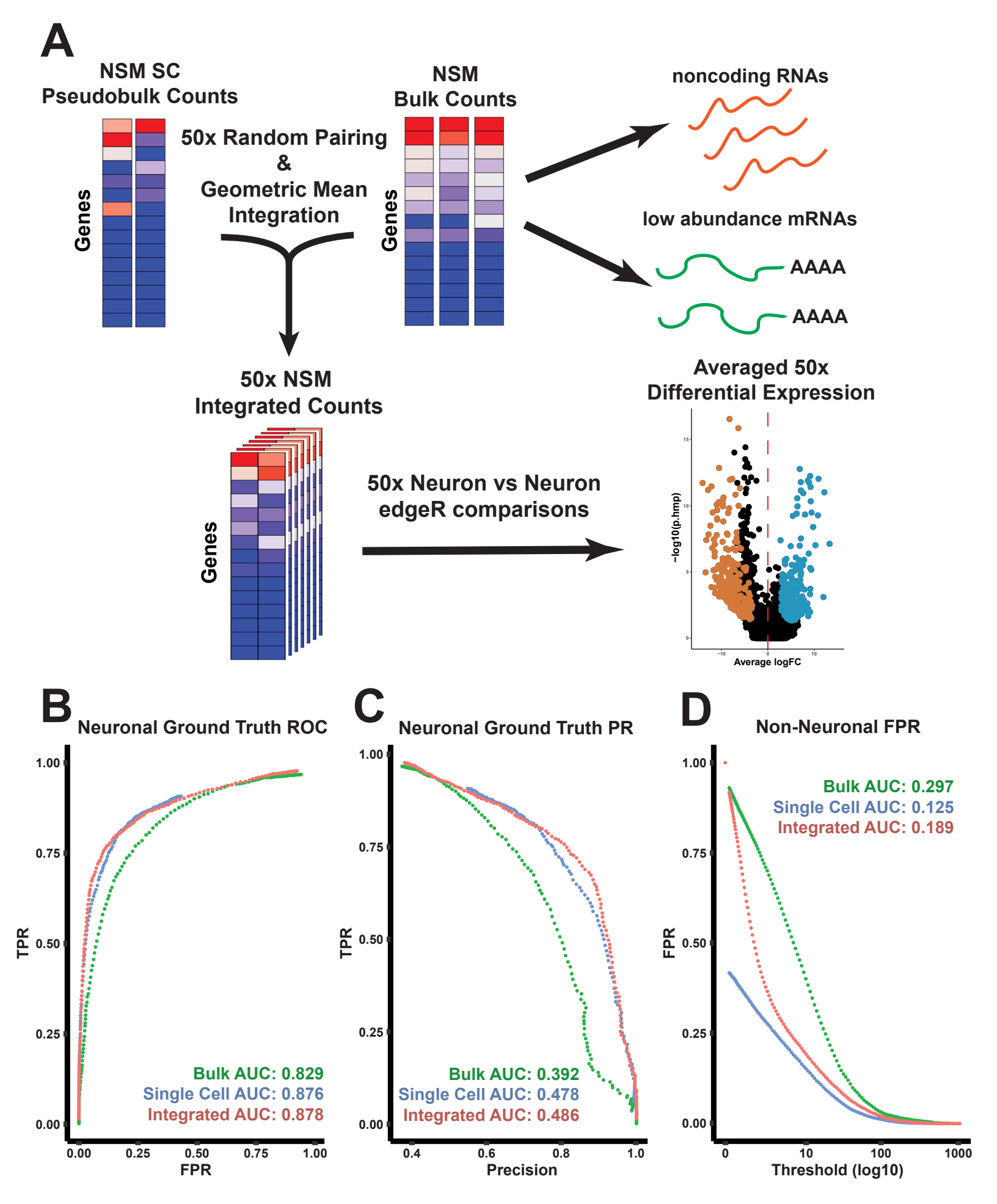
White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. Philos Trans R Soc Lond B Biol Sci *314*, 1-340. 10.1098/rstb.1986.0056.

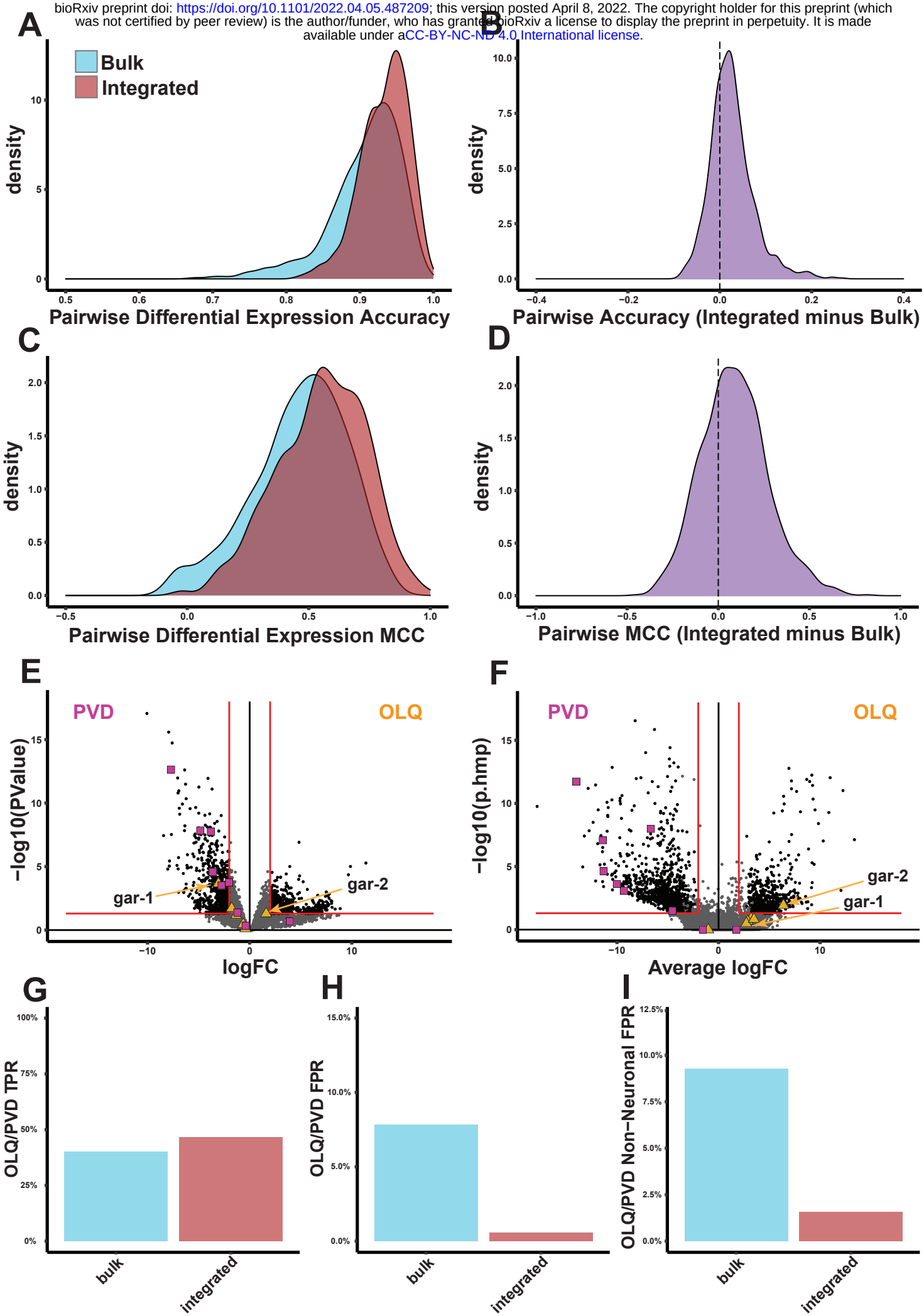
Wilson, D.J. (2019). The harmonic mean *p*-value for combining dependent tests. Proceedings of the National Academy of Sciences *116*, 1195-1200. 10.1073/pnas.1814092116.

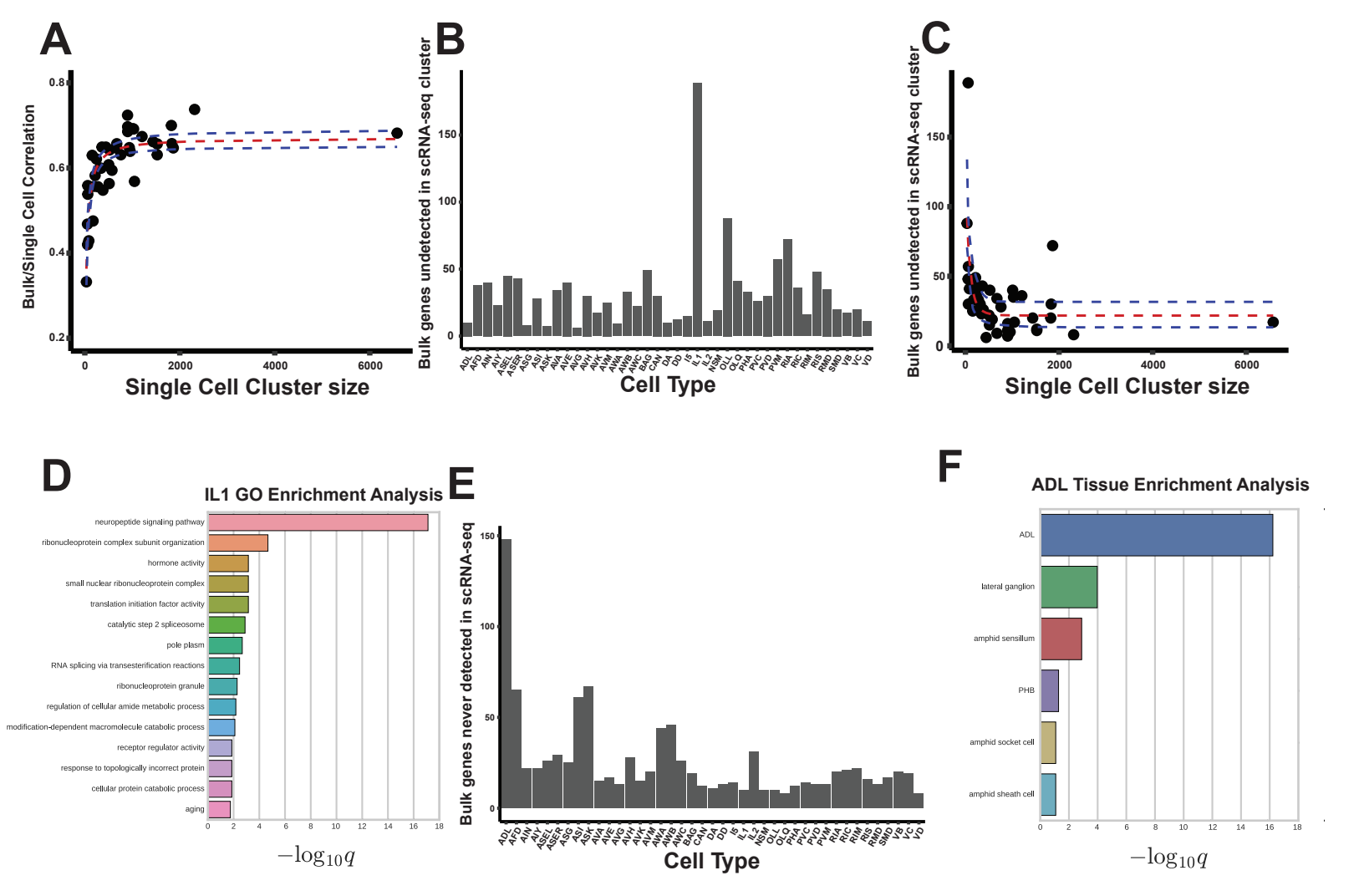
Yemini, E., Lin, A., Nejatbakhsh, A., Varol, E., Sun, R., Mena, G.E., Samuel, A.D.T., Paninski, L., Venkatachalam, V., and Hobert, O. (2021). NeuroPAL: A Multicolor Atlas for Whole-Brain Neuronal Identification in C. elegans. Cell *184*, 272-288 e211. 10.1016/j.cell.2020.12.012.

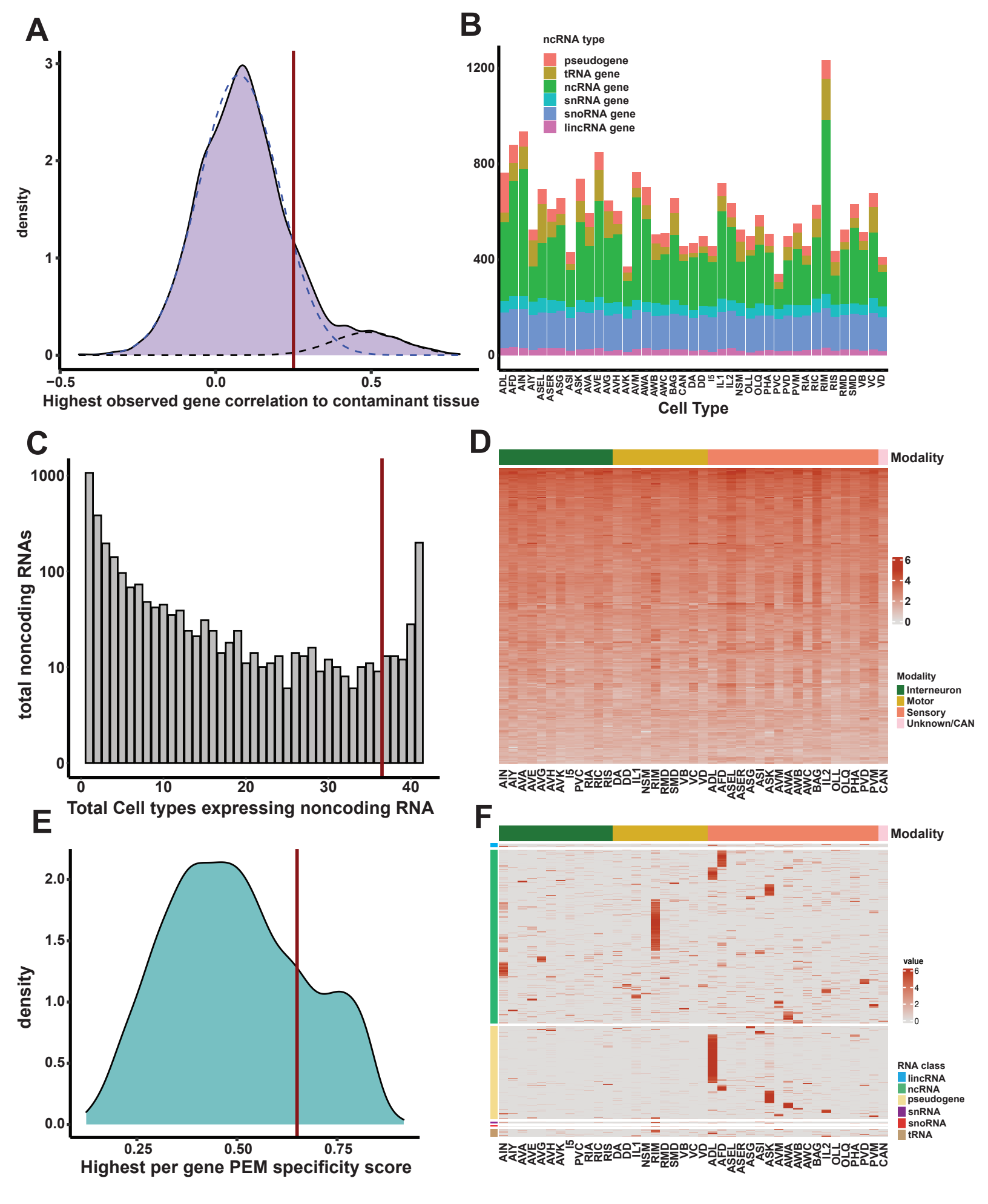
Zhu, L., Lei, J., Devlin, B., and Roeder, K. (2018). A UNIFIED STATISTICAL FRAMEWORK FOR SINGLE CELL AND BULK RNA SEQUENCING DATA. Ann Appl Stat *12*, 609-632. 10.1214/17-AOAS1110.

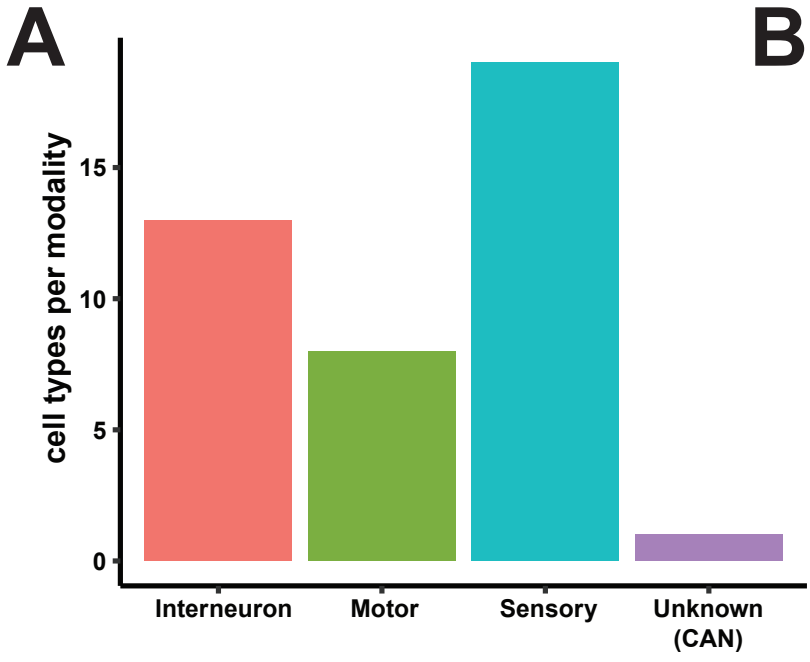

A
AVL
AVL
SDS-DTT
pronase
FACS
mCherry
GFP
unlabeled
GFP + mCherry
(AVL)
GFP + mCherry
++
GFP-A
mCherry-A
-266
-399
0
B
PC2
PC1
20
0
−20
−60
−40
−20
0
20
40
Sensory
Motor
Interneuron
CAN

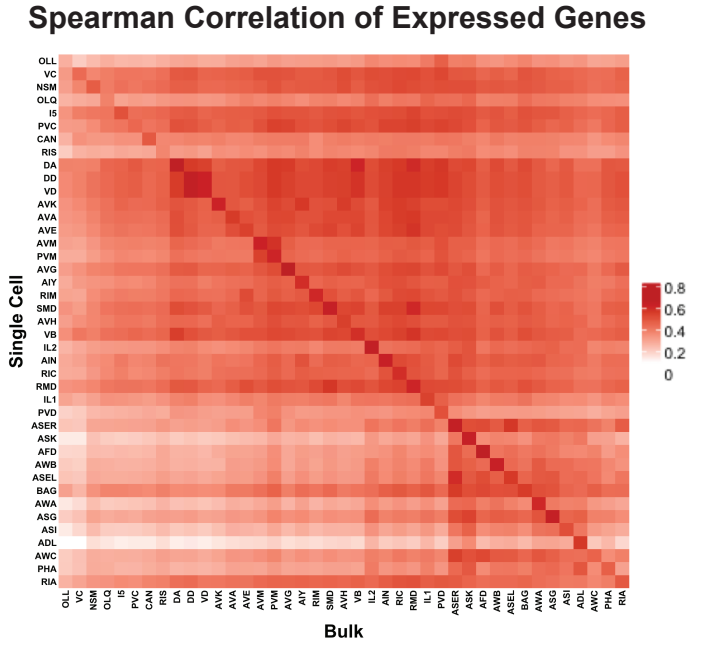

A
NSM SC
Pseudobulk Counts
Genes
50x Random Pairing
&
Geometric Mean
Integration
NSM
Bulk Counts
Genes
noncoding RNAs
low abundance mRNAs
AAAA
AAAA
50x NSM
Integrated Counts
Genes
50x Neuron vs Neuron
edgeR comparisons
Averaged 50x
Differential Expression
−log10(p.hmp)
Average logFC
B
Neuronal Ground Truth ROC
TPR
FPR
Bulk AUC: 0.829
Single Cell AUC: 0.876
Integrated AUC: 0.878
C
Neuronal Ground Truth PR
TPR
Precision
Bulk AUC: 0.392
Single Cell AUC: 0.478
Integrated AUC: 0.486
D
Non-Neuronal FPR
FPR
Threshold (log10)
Bulk AUC: 0.297
Single Cell AUC: 0.125
Integrated AUC: 0.189

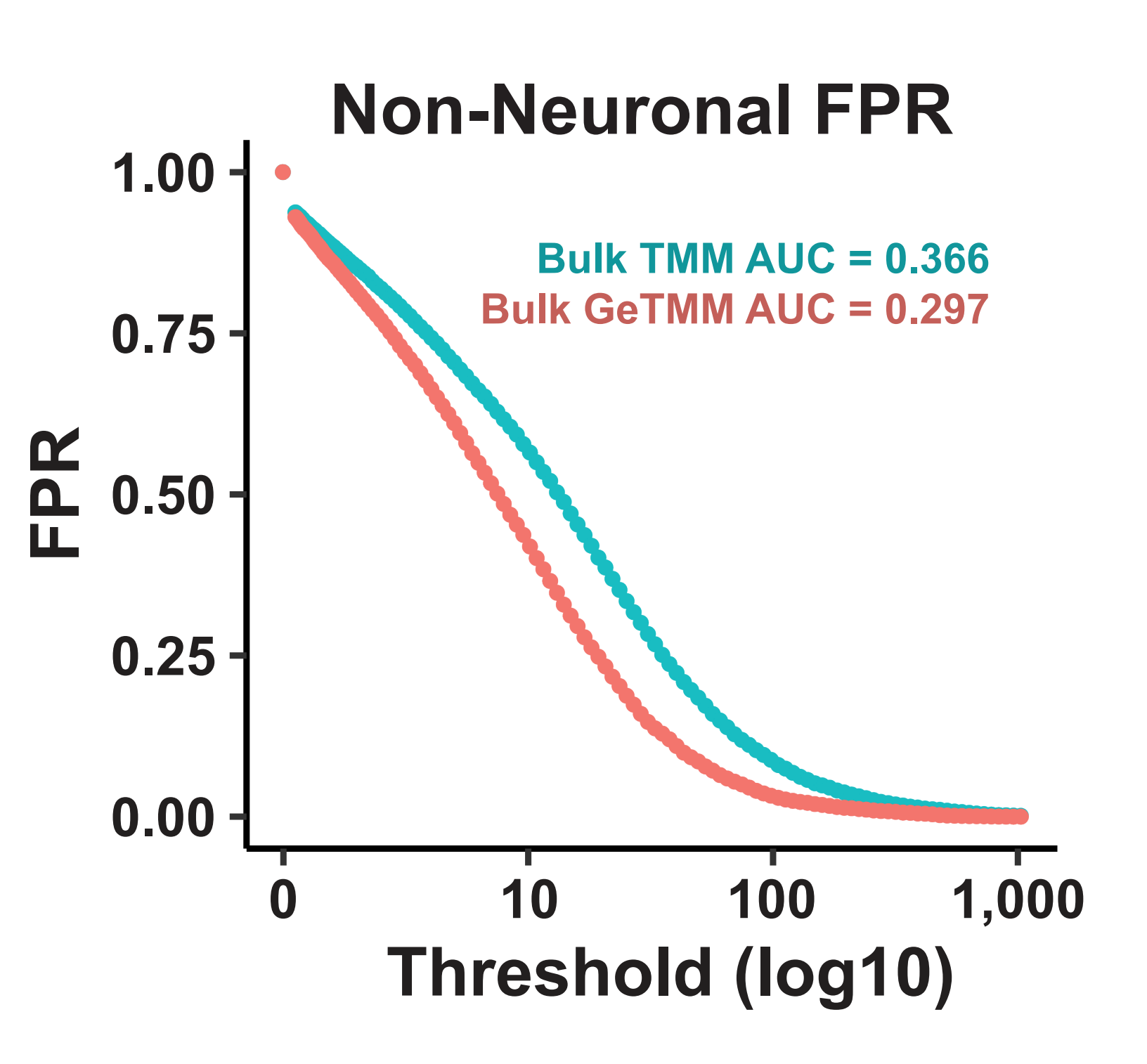



A
Bulk/Single Cell Correlation
Single Cell Cluster size
B
Bulk genes undetected in scRNA-seq cluster
Cell Type
C
Bulk genes undetected in scRNA-seq cluster
Single Cell Cluster size
D
IL1 GO Enrichment Analysis
neuropeptide signaling pathway
ribonucleoprotein complex subunit organization
hormone activity
small nuclear ribonucleoprotein complex
translation initiation factor activity
catalytic step 2 spliceosome
pole plasm
RNA splicing via transesterification reactions
ribonucleoprotein granule
regulation of cellular amide metabolic process
modification-dependent macromolecule catabolic process
receptor regulator activity
response to topologically incorrect protein
cellular protein catabolic process
aging
$-\log_{10} q$
E
Bulk genes never detected in scRNA-seq
Cell Type
F
ADL Tissue Enrichment Analysis
ADL
lateral ganglion
amphid sensillum
PHB
amphid socket cell
amphid sheath cell
$-\log_{10} q$


A
density
Highest observed gene correlation to contaminant tissue
B
ncRNA type
pseudogene
tRNA gene
ncRNA gene
snRNA gene
snoRNA gene
lincRNA gene
Cell Type
C
total noncoding RNAs
Total Cell types expressing noncoding RNA
D
Modality
Interneuron
Motor
Sensory
Unknown/CAN
E
density
Highest per gene PEM specificity score
F
Modality
value
RNA class
lincRNA
ncRNA
pseudogene
snRNA
snoRNA
tRNA

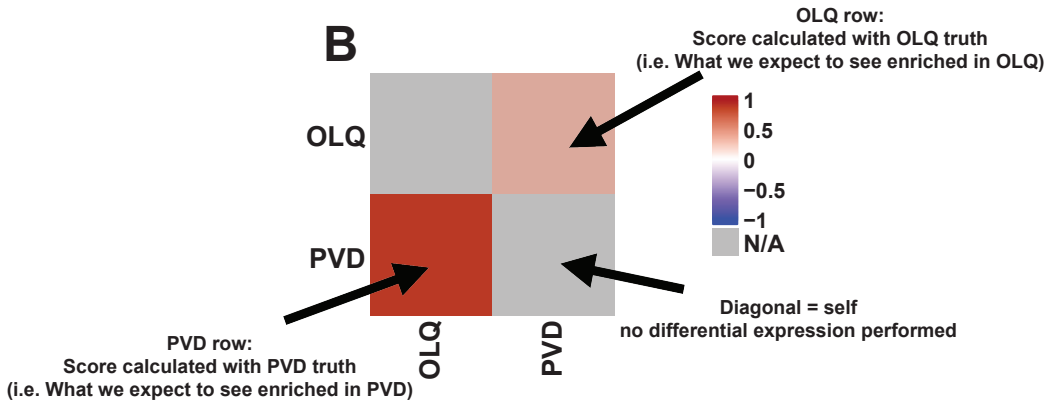

A
cell types per modality
15
10
5
0
Interneuron
Motor
Sensory
Unknown
(CAN)

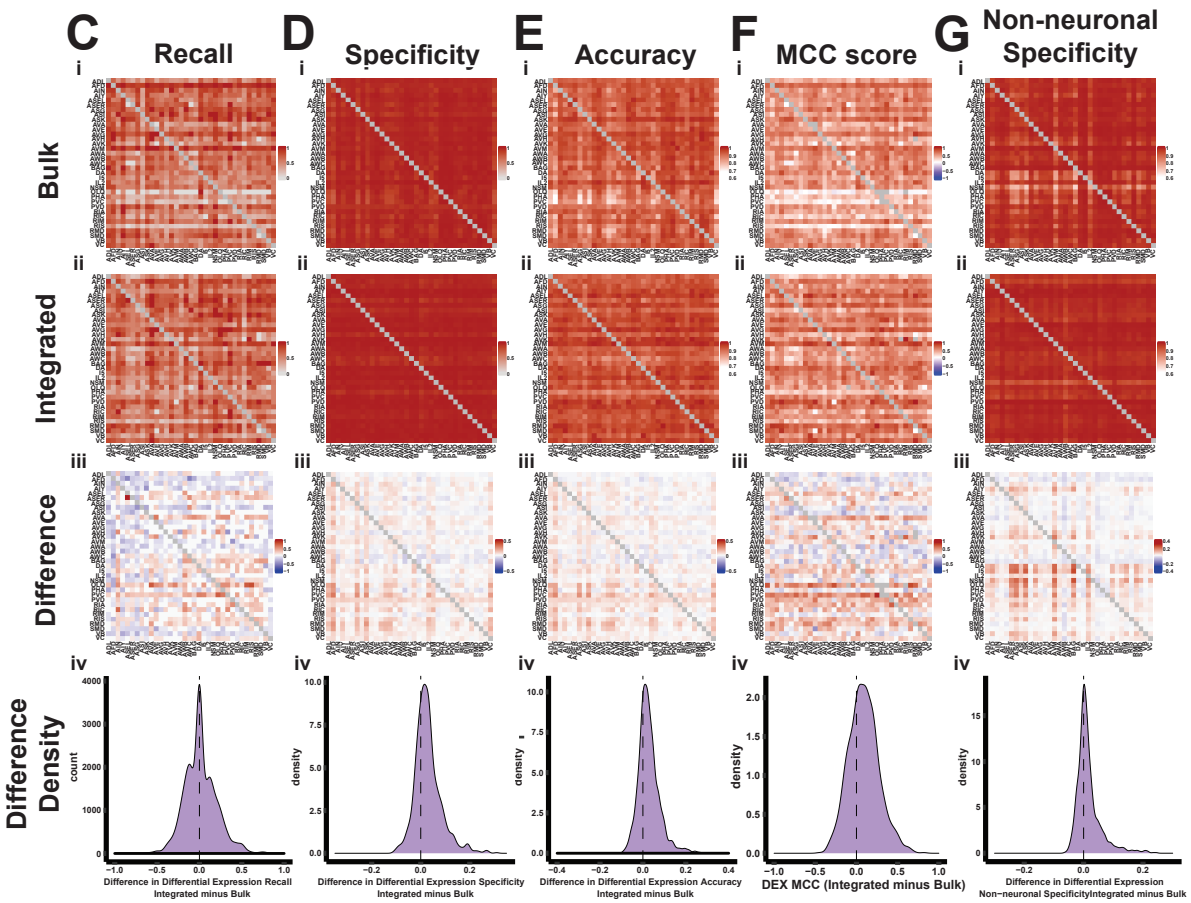

B
Spearman Correlation of Expressed Genes
Single Cell
Bulk
OLL
VC
NSM
OLQ
IS
PVC
CAN
RIS
DA
DD
VD
AVK
AVA
AVE
AVM
PVM
AVG
AIY
RIM
SMD
AVH
VB
IL2
AIN
RIC
RMD
IL1
PVD
ASER
ASK
AFD
AWB
ASEL
BAG
AWA
ASG
ASI
ADL
AWC
PHA
RIA
0.8
0.6
0.4
0.2
0

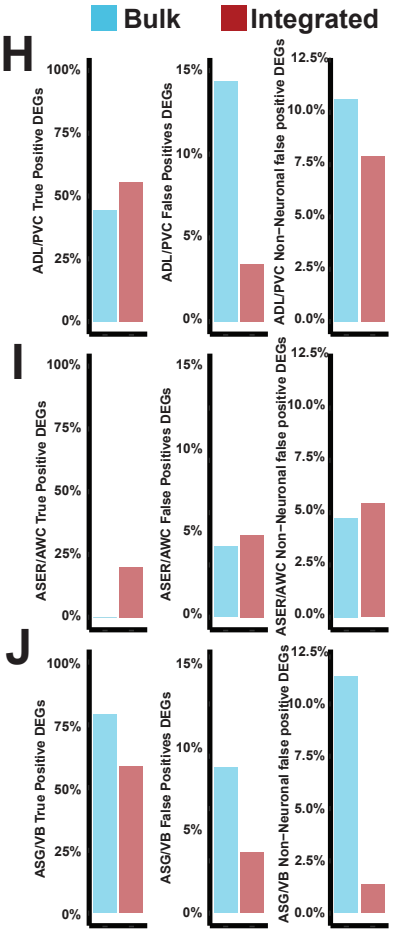

Non-Neuronal FPR
Bulk TMM AUC = 0.366
Bulk GeTMM AUC = 0.297
FPR
1.00
0.75
0.50
0.25
0.00
0
10
100
1,000
Threshold (log10)

A Neuronal Gound-Truth example

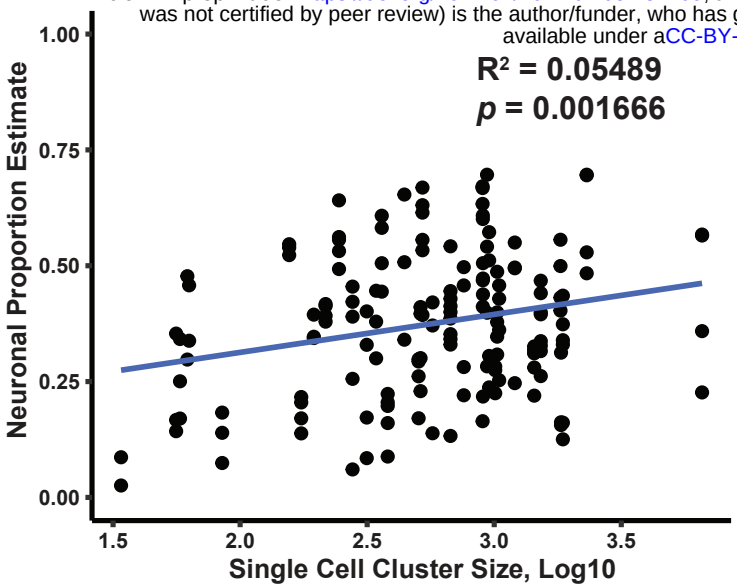
Gene	OLQ Truth	PVD Truth
a	1	0
b	0	1
c	0	0
d	0	0
e	0	0
f	0	0



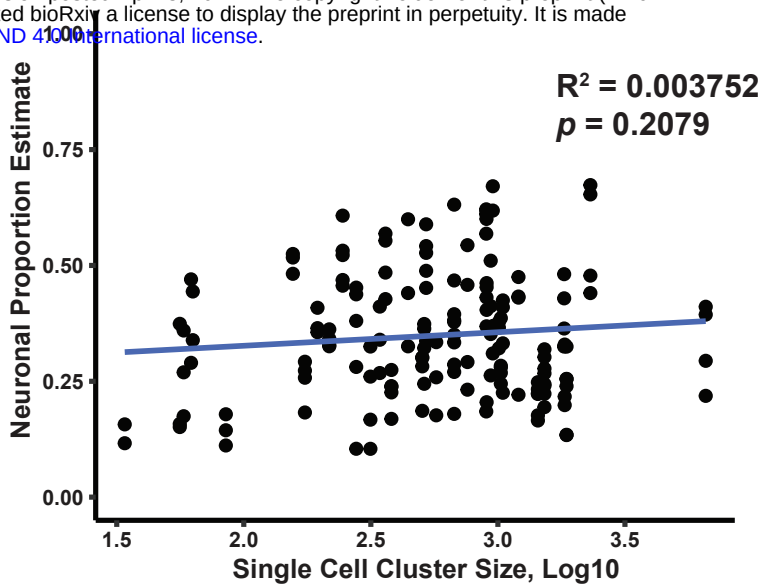




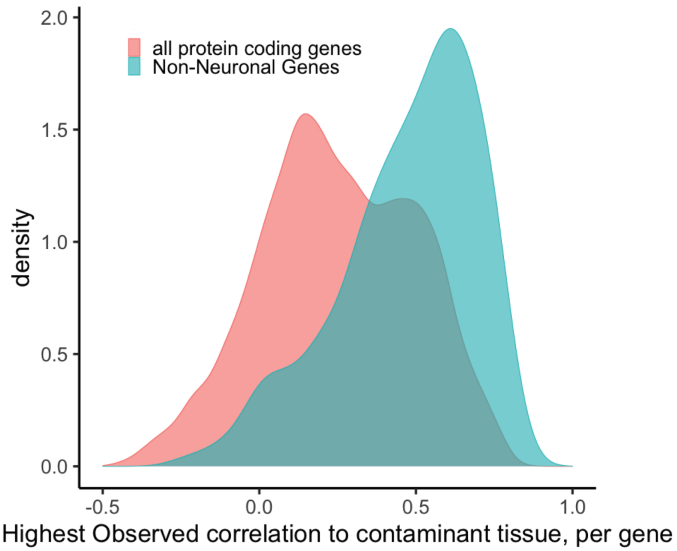
A



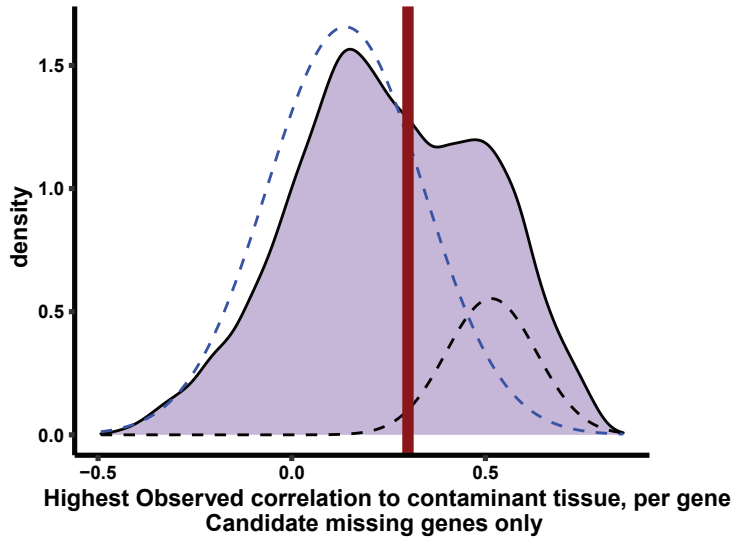
B



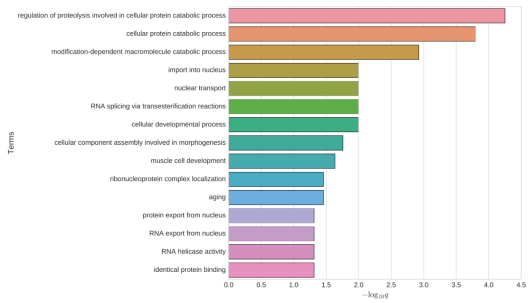
C



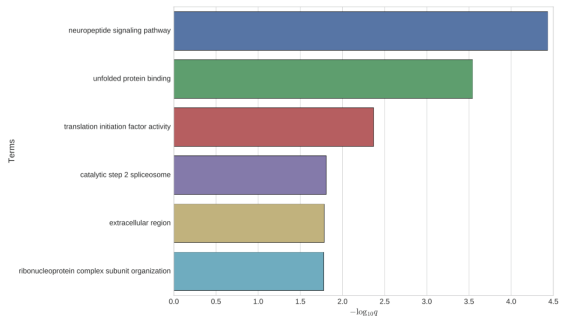
D



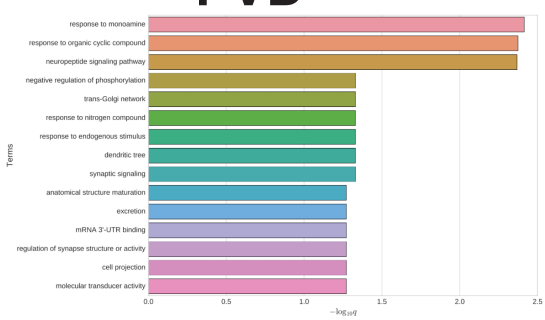
E



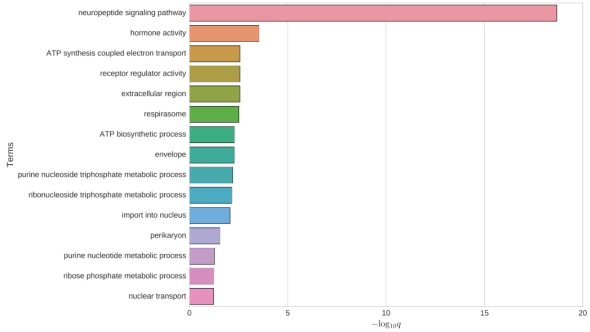
F

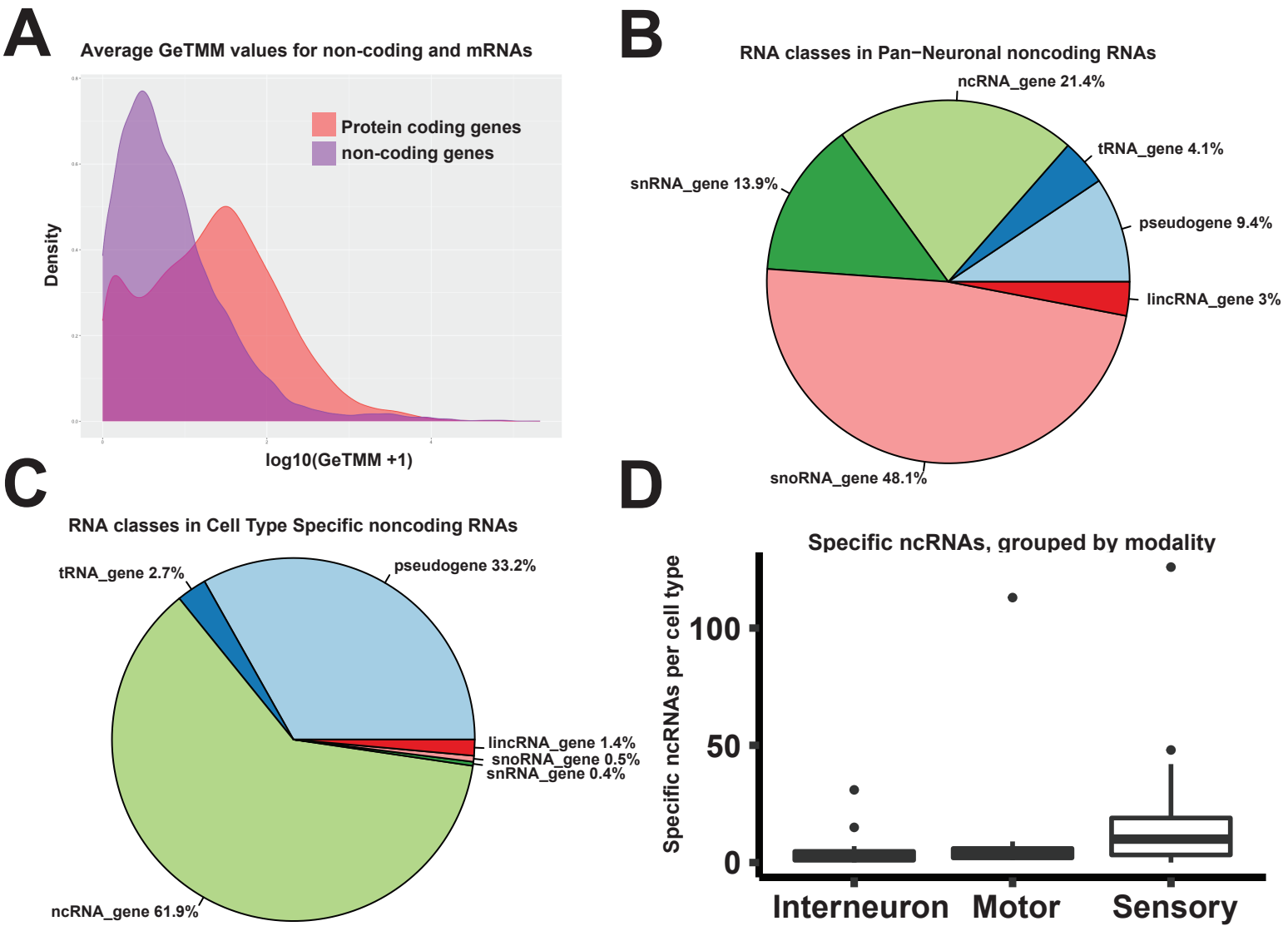


G



H




A
Average GeTMM values for non-coding and mRNAs
Protein coding genes
non-coding genes
Density
log10(GeTMM +1)
B
RNA classes in Pan−Neuronal noncoding RNAs
ncRNA_gene 21.4%
tRNA_gene 4.1%
pseudogene 9.4%
lincRNA_gene 3%
snoRNA_gene 48.1%
snRNA_gene 13.9%
C
RNA classes in Cell Type Specific noncoding RNAs
tRNA_gene 2.7%
pseudogene 33.2%
lincRNA_gene 1.4%
snoRNA_gene 0.5%
snRNA_gene 0.4%
ncRNA_gene 61.9%
D
Specific ncRNAs, grouped by modality
Specific ncRNAs per cell type
100
50
0
Interneuron
Motor
Sensory