G3, 2021, 11(7), jkab121
DOI: 10.1093/g3journal/jkab121
Advance Access Publication Date: 15 April 2021
Investigation

A head-to-head comparison of ribodepletion and polyA selection approaches for *Caenorhabditis elegans* low input RNA-sequencing libraries

Alec Barrett,[1] Rebecca McWhirter,[2] Seth R. Taylor,[2] Alexis Weinreb,[1,3] David M. Miller III ,[2,4] and Marc Hammarlund [1,3,*]

[1]Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA
[2]Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[3]Department of Neuroscience, Yale University School of Medicine, New Haven, CT 06510, USA
[4]Program in Neuroscience, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

*Corresponding author: Yale University School of Medicine, BCMM 436E, 295 Congress Avenue, New Haven, CT 06510, USA. Email: marc.hammarlund.@yale.edu
Sequence files and counts tables can be accessed under GEO accession GSE165793.

Abstract

A recent and powerful technique is to obtain transcriptomes from rare cell populations, such as single neurons in *Caenorhabditis elegans*, by enriching dissociated cells using fluorescent sorting. However, these cell samples often have low yields of RNA that present challenges in library preparation. This can lead to PCR duplicates, noisy gene expression for lowly expressed genes, and other issues that limit end-point analysis. Furthermore, some common resources, such as sequence-specific kits for removing ribosomal RNA, are not optimized for nonmammalian samples. To advance library construction for such challenging samples, we compared two approaches for building RNAseq libraries from less than 10 nanograms of *C. elegans* RNA: SMARTSeq V4 (Takara), a widely used kit for selecting poly-adenylated transcripts; and SoLo Ovation (Tecan Genomics), a newly developed ribodepletion-based approach. For ribodepletion, we used a custom kit of 200 probes designed to match *C. elegans* rRNA gene sequences. We found that SoLo Ovation, in combination with our custom *C. elegans* probe set for rRNA depletion, detects an expanded set of noncoding RNAs, shows reduced noise in lowly expressed genes, and more accurately counts expression of long genes. The approach described here should be broadly useful for similar efforts to analyze transcriptomics when RNA is limiting.

Keywords: *C. elegans*; transcriptomics; RNAseq; low input RNAseq; Ribodepletion; Poly-adenylated; library preparation

Introduction

RNA sequencing (RNAseq) is a well-established method for assessing gene expression. In *Caenorhabditis elegans*, RNAseq can be combined with cell enrichment by fluorescence activated cell sorting (FACS) for studies of gene expression in individual cell types, enabling a high-resolution view of cell type-specific transcription (Spencer *et al.* 2014; Kaletsky *et al.* 2016, 2018; Ahn *et al.* 2017; Taylor *et al.* 2019; Warner *et al.* 2019). Sequencing library construction is a significant variable in RNAseq experiments from FACS-enriched *C. elegans* samples, as well as from other samples with low amounts of input RNA. Comparing library construction techniques to maximize data recovery is therefore an important goal.

At approximately 90% of total RNA in most cells, the prevalence of ribosomal RNA (rRNA) constitutes a major challenge for RNAseq profiling of other RNA species (O'Neil *et al.* 2013). To efficiently sequence the remaining 10% of cellular RNA, rRNA is typically excluded during library construction. One common approach to meet this goal is the use of poly-d(T) primers to favor cDNA synthesis from poly-adenylated (polyA) RNA *vs* rRNA, which is typically not poly-adenylated. In an alternative strategy, known as ribodepletion, oligonucleotides complementary to specific rRNA sequences are used to deplete rRNA transcripts from the library by either bead affinity extraction (Petrova *et al.* 2017; Culviner *et al.* 2020) or directed enzymatic cleavage (Herbert *et al.* 2018). Overall, polyA approaches can be more efficient at excluding rRNA compared to ribodepletion strategies (Zhao *et al.* 2014). However, polyA methods require the RNA input to be largely free from degradation, tend to bias coverage toward the 3' end of transcripts, and exclude ncRNA (noncoding RNA) species that lack polyA tails. In contrast, ribodepletion is better suited for low-quality samples, as random primer amplification is more likely to capture fragmented RNAs. Importantly, ribodepletion preserves ncRNAs and allows library construction methods that favor more uniform gene body coverage (Zhao *et al.* 2014; Ching *et al.* 2015; Cao *et al.* 2019).

Although recent advancements in library preparation methods have demonstrated that both polyA and ribodepletion methods are feasible for ultra-low input samples from *C. elegans* (Spencer *et al.* 2014; Tintori *et al.* 2016), most published *C. elegans* studies have used polyA approaches (Lim and Brunet 2013; Camacho *et al.* 2018; Serra *et al.* 2018; Posner *et al.* 2019; Zullo *et al.*

Received: January 30, 2021. **Accepted:** March 25, 2021

2019), and a ribodepletion approach was found to retain high levels of rRNA in the final library (Spencer *et al.* 2014), potentially because these rRNA ribodepletion oligonucleotides were optimized for mammalian rRNA.

Here, we introduce a ribodepletion approach that uses oligonucleotides specifically designed to match *C. elegans* rRNA sequences and test the idea that this ribodepletion strategy can produce favorable results on low input RNA samples. We compared polyA (SMARTseq V4) and ribodepletion (SoLo Ovation) approaches, using <10 ng of total RNA input prepared from FACS-enriched *C. elegans* neurons. We constructed multiple libraries with each strategy and sequenced the resulting libraries at high depth. Detailed comparisons of the results indicated that although high-quality libraries can be obtained with either method, SoLo ribodepletion with *C. elegans* specific probes has significant advantages over the commonly used SMARTseq polyA approach, including increased detection of noncoding RNAs, reduced noise for lowly expressed genes, and more accurate counts for long genes.

Materials and methods

Sample collection

The wild-type N2 strain and OH10689 *otIs355 [rab-3(prom1)::2xNLS-TagRFP]* IV were used for this study. Worms were grown on 8 P nutrient agar 150 mm plates seeded with *E. coli* strain NA22. To obtain synchronized cultures of L4 worms, embryos obtained by hypochlorite treatment of adult hermaphrodites were allowed to hatch in M9 buffer overnight (16–23 h) and then grown on NA22-seeded plates for 45–48 h. The developmental age of each culture was determined by scoring vulval morphology (>75 worms) (Mok *et al.* 2015). Single-cell suspensions were obtained as described (Zhang *et al.* 2011; Spencer *et al.* 2014; Kaletsky *et al.* 2018) with some modifications. Worms were collected and separated from bacteria by washing twice with ice-cold M9 and centrifuging at 150 rcf for 2.5 min. Worms were transferred to a 1.6 mL centrifuge tube and pelleted at 16,000 rcf for 1 min. 250 µl pellets of packed worms were treated with 500 µl of SDS-DTT solution (20 mM HEPES, 0.25% SDS, 200 mM DTT, 3% sucrose, pH 8.0) for 2 min.

Following SDS-DTT treatment, worms were washed five times by diluting with 1 mL egg buffer and pelleting at 16,000 rcf for 30 s. Worms were then incubated in pronase (15 mg/mL, Sigma-Aldrich P8811, diluted in egg buffer) for 23 min. During the pronase incubation, the solution was triturated by pipetting through a P1000 pipette tip for four sets of 80 repetitions. The status of dissociation was monitored under a fluorescence dissecting microscope at 5-min intervals. The pronase digestion was stopped by adding 750 µl L-15 media supplemented with 10% fetal bovine serum (L-15-10), and cells were pelleted by centrifuging at 530 rcf for 5 min at 4°C. The pellet was resuspended in L-15-10, and single-cells were separated from whole worms and debris by centrifuging at 100 rcf for 2 min at 4°C. The supernatant was then passed through a 35-micron filter into the collection tube. The pellet was resuspended a second time in L-15-10, spun at 100 rcf for 2 min at 4°C, and the resulting supernatant was added to the collection tube.

FACS was performed on a BD FACSAria™ III equipped with a 70-micron diameter nozzle. DAPI was added to the sample (final concentration of 1 µg/mL) to label dead and dying cells. Cells were sorted under the "4-way Purity" mask. Sorted cells were collected directly into TRIzol LS. At ~15-min 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 −80°C for RNA extractions (see below).

RNA extraction

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. RNA in the aqueous layer was cleaned and concentrated using the RNA Clean and Concentrator Kit (Zymo Research, R1013). The aqueous layer from the chloroform extraction was combined with an equal volume of 100% ethanol and transferred to a Zymo-Spin IC column. Columns were centrifuged for 30 s at 16,000 rcf. Samples 2 and 4 were then treated in-column with DNase I for 15 min (Supplementary Table S1). Samples 1 and 3 were not treated with DNase I. All samples were then washed with 400 µl of Zymo RNA Prep Buffer and centrifuged for 16,000 rcf for 30 s. Columns were washed twice with Zymo RNA Wash Buffer (700 µl, centrifuged for 30 s, followed by 400 µl, centrifuged for 2 min). RNA was eluted by adding 15 µl 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 Pico chip to estimate yield and RNA integrity and the remainder was stored at −80°C.

rRNA probe optimization

To generate a probe set that targets *C. elegans* rRNA sequences, fasta sequences of all *C. elegans* rRNA genes were downloaded from wormmine (version WS235). Any exact duplicate sequences longer than 60 bases were reduced to a single copy. Tecan Genomics used these sequences to generate a set of 200 probes, proprietary to and available from, Tecan Genomics. Most rRNA genes were well covered, with the exception of a 150 bp A/T rich region of *MTCE.33* and a 400 BP A/T region at the 3' end of *MTCE.7*.

The probe set is available as a custom order under the name Ovation® SoLo® RNA-Seq System with Custom AnyDeplete® for the depletion of *C. elegans* rRNA (Tecan Genomics, Inc., Redwood City, CA, USA), part number: PN 30185717.

Library preparation and sequencing

SoLo and SMARTseq libraries were constructed for all samples according to manufacturer's instructions. RNA concentrations are provided in Supplementary Table S1.

SoLo samples are treated with DNase I prior to first-strand cDNA synthesis with a mix of oligo d(T) and random primers. First-strand cDNA then undergoes hydrolysis, fragmentation, and strand selection before second strand synthesis, adapter ligation, Ampure bead purification, and amplification of the cDNA library with another round of Ampure bead purification. The amplified library is then incubated with the rRNA probe set, and rRNA fragments are selected against by nuclease-mediated cutting of the adapter sequence, followed by a final round of amplification and a final Ampure bead cleanup. Adapters in the Tecan SoLo Ovation kit use a single 8 base index, followed by an 8 base unique molecular identifier sequence (UMI).

SMARTseq samples are amplified using oligo d(T) primers to create the first strand of cDNA. A primer binding sequence is then appended to the first strand to allow for template switching and second strand synthesis. Long Distance PCR (LD-PCR) is then used to amplify the full-length cDNA into the final library before Ampure bead cleanup. cDNA libraries were then processed using the Illumina Nextera XT kit to create the sequencing library. Samples first underwent fragmentation and adapter ligation by

Tn5 enzymatic tagmentation before final amplification and Ampure bead cleanup. Nextera uses dual-barcode adapters, but lack a UMI sequence.

All samples were checked for adapters dimers on an Agilent Bioanalyzer using the High Sensitivity DNA chip. Libraries were sequenced to a depth of 15–32 million read pairs on an Illumina HiSeq2500 machine, with paired end 75 bp reads. SoLo libraries sequences on HiSeq4000 machines experienced run failures when not mixed with ∼40% non-SoLo cDNA (data not shown). HiSeq2500 runs worked optimally when multiplexed samples were run at 5% higher concentration than standard.

Read mapping, deduplication, sub feature base counting, and gene body coverage

Reads were mapped to the WBcel235 reference genome assembly using STAR (version 2.7.0). Duplicate reads were marked and removed using SAMtools (version 1.9.0). Counts files were generated using the featureCounts program from the SubRead package (version 1.6.4). Genes encoding rRNAs were removed from counts files prior to downstream normalization and gene detection steps.

Bases that map to exons, UTRs, introns, and intergenic regions were calculated using the CollectRNAseqMetrics program in PICARD (version 2.23.8).

Gene body coverage curves were obtained using the geneBody_coverage tool in RSeQC (version 2.6.4). To compare the coverage of the 5' end between techniques, only protein-coding genes called expressed in both techniques were considered. Gene models were split and each gene's coverage was run separately. Within each gene, coverage was normalized to the highest coverage value. Average coverage across the first 20% of each gene was calculated for all replicates, and then averaged within each library building technique. Significance in 5' coverage was calculated using a paired *t*-test in Scipy (version 1.5.0), *P*-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg correction.

Gene expression normalization, thresholding, dispersion, and transcript length analysis

Count matrices were normalized first to transcript length, and then to library size adjusted to the Trimmed-Mean of M (TMM) using edgeR (version 3.28.1) to obtain GeTMM values. 95% Confidence Intervals (CI) were calculated within each technique using the CI function in gmodels package in R (version 2.18.2). Normalized CI size was calculated for each gene by dividing the size of the interval by the average GeTMM value (separately within each technique) to allow for easier comparison of CI sizes across expression levels. Genes were called "not expressed" if the CI upper bound was <5 GeTMM, genes were called "ambiguous" if the CI overlapped 5 GeTMM, and genes were called expressed if the CI lower bound was >5 GeTMM. Dispersion estimates were generated for each library approach in edgeR using only samples from that library. Genes <0.5 average GeTMM showed identical gene-level dispersion and were excluded. Differential expression analysis was calculated using a generalized linear model in edgeR. Analyses that focused on protein-coding gene expression used only protein-coding genes to calculate TMM values and library size. Transcript lengths were used from the output of featureCounts. Statistical tests comparing quintiles by GeTMM value were performed by first filtering to only include genes expressed >0.5 average GeTMM in both SMARTseq and SoLo samples. Genes were then ranked within each sample set, and split into five groups of equal size. When plotting log10 scale GeTMM counts, pseudocounts of 1 were added. When plotting log2 ratios of 5' coverage, gene dispersion, or normalized CI size between SoLo and SMARTseq samples, pseudocounts of 1×10^{-5} were added to the numerator and denominator.

For comparison to previously published data, tables of expressed genes were obtained from supplemental materials of that publication (Kaletsky *et al.* 2018).

Data availability

Raw fastq files and a processed counts table for SoLo and SMARTseq samples are available at GEO under accession number GSE165793. Strains are available upon request. Supplemental Material available at figshare: https://doi.org/10.25387/g3.14394353.

Results

Designing a custom set of *C. elegans* rRNA depletion probes

Ribodepletion strategies are sequence dependent, and mismatches or gaps between the oligonucleotide probe set and rRNA sequences may allow substantial rRNA sequences to remain in the finished sequencing library. rRNA depletion probe sets designed for mammalian samples perform poorly in *C. elegans* (Spencer *et al.* 2014). We, therefore, designed a probe set to match *C. elegans* rRNA. We collected fasta files from all *C. elegans* rRNA genes in the WS235 genome assembly and removed duplicate sequences to identify 200 unique rRNA probes for use in our ribodepletion experiments (Methods).

Collecting pan-neuronal samples via FACS

We tested different methods for library construction on low abundance RNA inputs typically provided by FACS-sorted samples from *C. elegans*. We dissociated L4-stage larvae with SDS and protease treatments (Spencer *et al.* 2014; Kaletsky *et al.* 2018; Taylor *et al.* 2019), and then used FACS to enrich for cells expressing the neuron-specific reporter *Prab-3*::RFP (Nonet *et al.* 1997). For each RNA sample, ∼25,000 cells were collected directly into TRIzol, and RNA was extracted by adding chloroform, mixing, and collecting the aqueous layer. RNA was concentrated prior to library preparation using a Zymo-Spin IC column. For this study, we used 4 independently generated RNA samples, with one of those samples split into two technical replicates during purification, one DNase pre-treated, and one not (see Materials and Methods, Supplementary Table S1). Yields of total RNA ranged from 2.2 to 7.2 ng for each sample (Supplementary Table S1).

SMARTseq excludes more rRNA whereas SoLo shows fewer duplicate reads

Next, for direct comparisons of polyA and ribodepletion approaches, we split each RNA sample in half, resulting in two sets of matched samples, with five total replicates in each. We built sequencing libraries for all samples, using either a polyA approach (SMARTseq V4) or a ribodepletion strategy (SoLo Ovation) for each pair of matched samples. Briefly, for SMARTseq libraries, cDNA was prepared with poly-d(T) primers and then amplified. Fragmentation, adapter ligation, and final library preparation were performed using the Illumina Nextera XT kit. For the SoLo approach, cDNA libraries were prepared using random primers prior to fragmentation, adapter ligation, and amplification. After isothermal amplification, the *C. elegans* custom probe set was used to direct cleavage of rRNA fragment adapters prior to a final round of amplification. We sequenced all libraries on an Illumina

Hiseq 2500 machine with paired end 75 bp reads to a depth of 15 to 37 million read pairs per library. After sequencing, fastq files were checked for read quality using fastQC. One SMARTseq sample failed quality control (low per base quality, and highly 3' biased gene coverage). This SMARTseq library and the corresponding SoLo library were removed from further analysis. Of the four remaining paired sets of samples, SoLo sequenced libraries had an average of 17.96 million read pairs, and SMARTseq libraries had an average of 31.77 million read pairs (Supplementary Table S1).

We assessed the basic properties of the ribodepletion and polyA libraries using metrics that reflect the relative number of useful reads. We defined useful reads as those that are not PCR duplicates and that do not map to rRNA genes. We used STAR to map all reads to the WS235 genome with default parameters (Dobin *et al.* 2013), and observed that SMARTseq samples had slightly higher unique mapping rates than SoLo samples, averaging 67% and 59%, respectively (Figure 1A). After mapping we marked and removed duplicate reads based on position using SAMtools (Li *et al.* 2009). The percentage of reads remaining after deduplication provides a measure of duplicate reads. Using this metric, SoLo libraries had consistently lower rates of PCR duplicates (mean 33.7% reads remaining after deduplication) than SMARTseq libraries (mean 26.5% reads remaining after deduplication) (Figure 1B).

To assess where reads mapped, we compared the percentage of bases that mapped to exons, untranslated regions (UTRs), introns, and intergenic regions. SMARTseq libraries mapped an average of 73.9% of bases to exons, 19.0% to UTRs, 3.0% to introns, and 4.1% to intergenic regions. SoLo libraries mapped an average of 57.4% of bases to exons, 29.3% to UTRs, 6.8% to introns, and 6.5% to intergenic regions (Figure 1C). As UTRs are included in the gene models used for assigning counts, the average total percent of gene-feature mapped reads is 92.9% for SMARTseq and 86.7% for SoLo. Next, we assessed the fraction of rRNA reads in both original and deduplicated libraries. Prior to deduplication, SMARTseq samples had an average of 3.0% (range = 2.5–3.6%), while SoLo samples had an average of 22.7% rRNA reads (range = 17.9–25.7%). In deduplicated libraries, SMARTseq samples had an average of 2.1% (range = 1.7–2.3%), while SoLo samples had an average of 13.3% rRNA reads (range = 11.7–15.6%) (Figure 1D). Assuming rRNA is 90% of the cellular RNA, these data indicate that SMARTseq removed an average of 99.8% of the rRNA, whereas SoLo removed an average of 98.3% of the rRNA. Overall, these data indicate that both techniques result in efficient selection against rRNA but that the SMARTseq polyA-based approach performed better than SoLo in rRNA removal (Figure 1D).

SoLo and SMARTseq detect largely overlapping gene sets

Next, we compared the overall number of expressed genes between ribodepletion and polyA libraries. Each sample was normalized using the GeTMM method, first to gene length and then to the Trimmed Mean of M-values (TMM) corrected library size to account for intra- and inter-sample variation (Smid *et al.* 2018). Average GeTMM values for all genes were generally correlated between SMARTseq and SoLo samples, with a Spearman correlation coefficient of 0.79. Within each technique, replicates were also highly correlated (Supplementary Figure S1, A–C). We then calculated 95% CI for all genes within SMARTseq and SoLo samples. We defined "expressed" genes as those genes where the lower bound of the 95% CI is >5 GeTMM. Using this definition, we called 6146 genes "expressed" in SMARTseq, and 7108 genes expressed in SoLo. The majority of expressed genes (5104) were called "expressed" in both approaches (Figure 2A). Similarly, we defined "not expressed" genes as those genes where the upper bound of the CI is <5 GeTMM. The remaining genes, with CI that overlap 5 GeTMM, we consider to be genes for which expression is "ambiguous." Interestingly, "ambiguous" genes were more common in SMARTseq samples (Figure 2B). For the 6146 genes called "expressed" in SMARTseq, 788 were called "ambiguous" in SoLo (12.9%), and 254 were called "not expressed" in SoLo (4.1%). Of the 7108 genes called expressed in SoLo, 1618 were called "ambiguous" in SMARTseq (22.8%), and 386 were called "not expressed" (5.4%). Of the 7827 genes called "ambiguous" in SMARTseq 3229 were called "not expressed" in SoLo (41.3%). Similarly, of the 5824 genes called "ambiguous" in SoLo 2056 were called "not expressed" in SMARTseq (35.3%). Ribodepletion and polyA priming approaches resulted in broadly similar results for gene expression. However, substantial differences exist between the two approaches with respect to confidently calling genes "expressed" or "not expressed." The existence of these differences raises the question of how differences in RNA capture or amplification affect specific RNA types.

SoLo detects an expanded set of noncoding RNA species

A potential source of differences between the datasets is the role of polyA tails in cDNA synthesis. Many noncoding RNAs lack 3' polyA tails and are thus unlikely to be efficiently captured by SMARTseq cDNA synthesis, which depends on poly-d(T) priming. To test for this possibility, we compared the detection rates of six classes of noncoding RNAs between SoLo and SMARTseq, using the 5 GeTMM threshold for calling genes expressed. Of these classes, pseudogenes and long intergenic noncoding RNAs (lincRNAs) often have polyA tails (Pink *et al.* 2011; Ransohoff *et al.* 2018), and we found that these classes are detected at similar frequencies between SoLo and SMARTseq (Figure 2C). By contrast, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), and transfer RNAs (tRNAs) usually lack polyA tails (Cech and Steitz 2014), and we found that these classes are detected at much higher frequencies in SoLo than SMARTseq libraries (Figure 2C). In comparison to SMARTseq, SoLo calls 6.9 times as many snRNAs "expressed," 24 times as many snoRNAs, and 13 times as many tRNAs (Figure 2C). In addition, of the genes called "expressed" by SoLo, 57.6% of snRNAs, 33.3% of snoRNAs, and 82.2% of tRNAs have zero counts in any SMARTseq replicate. A final class of RNAs, uncategorized noncoding RNAs (ncRNAs), was detected at high levels in both approaches, although SoLo detects ~50% more ncRNA transcripts, suggesting that this category contains a mix of poly-adenylated and nonpoly-adenylated transcripts.

To determine the contribution of noncoding RNAs to discrepancies in gene detection between techniques, we focused on the genes that are confidently called "expressed" in one technique, but "not expressed" in the other, not considering "ambiguous" genes. Breaking down those sets by biotype, we see that 67 (31.5%) of the 254 SMARTseq "expressed" exclusive genes are noncoding RNAs, whereas 213 (55.1%) of the 386 SoLo "expressed" exclusive genes are noncoding RNAs (Figure 2D). This analysis indicates that a majority of the genes confidently detected by the SoLo method but not by SMARTseq are noncoding RNAs.

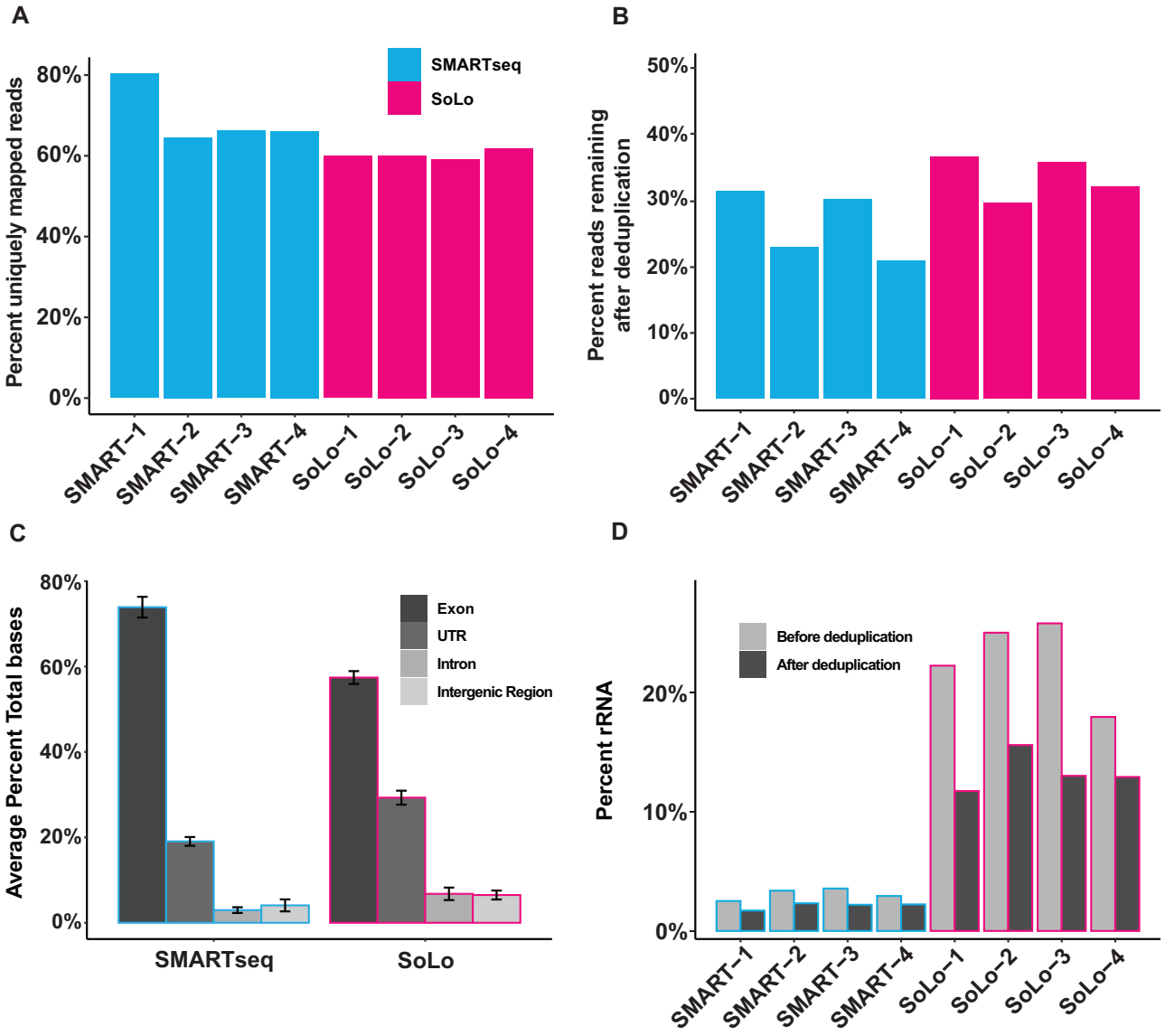

Figure 1 Mapping summary (A) Bar graph showing the percent of all reads mapped from raw fastq files for SMARTseq replicates (blue) and SoLo replicates (magenta). (B) Bar graph showing the percent of mapped reads that were not marked as likely PCR duplicates. (C) Bar graph showing the percent of bases that mapped to exons, UTRs, introns, and intergenic regions. Error bars show 95% CI. (D) Bar graph showing the percent of mapped reads counted in rRNA genes before and after deduplication.

SoLo samples show reduced variance among lowly expressed protein-coding genes

SMARTseq has many more genes for which expression was "ambiguous" than SoLo (Figure 2B). These data could indicate a difference in noise between the techniques, which might account for some of the remaining differences in apparent gene expression that are not explained by differences in noncoding RNA detection. To explore this possibility, we generated GeTMM values using only counts that map to protein-coding genes. Using the thresholding method described above, SMARTseq libraries yielded 5899 "expressed" protein-coding genes, 8320 "not expressed" protein-coding genes, and 6288 "ambiguous" protein-coding genes. SoLo libraries yielded 6625 "expressed" protein-coding genes, 10,091 "not expressed" protein-coding genes, and 3731 "ambiguous" protein-coding genes (Figure 3A). Thus, SoLo libraries yielded more protein-coding genes that are confidently called either "expressed" or "not expressed," whereas SMARTseq showed almost twice as many "ambiguous" protein-coding genes.

Other than this important difference, the results from the two techniques for protein-coding genes were similar. Similar to results analyzing expression data for all genes (Figure 2B), the majority of protein-coding genes that were confidently called "expressed" in either technique were called "expressed" in both techniques (Figure 3B). In addition, mRNAs called "expressed" in one technique were very rarely called "not expressed" in the other (SMARTseq: 187, SoLo: 173) (Figure 3C). GeTMM values for protein-coding genes were highly correlated between the two techniques, with a Spearman correlation coefficient of 0.88 (Supplementary Figure S1D). Among the 6288 "ambiguous" SMARTseq protein-coding genes, 1453 (23.1%) were called "expressed" in SoLo, and 2339 (37.2%) were called "not expressed" in SoLo. For SoLo samples, of the 3731 "ambiguous" protein-coding genes, 713 (19.1%) were called "expressed" in SMARTseq and 582 (15.6%) were called "not expressed" in SMARTseq. Overall numbers of "ambiguous" protein-coding genes were higher in SMARTseq. (Figure 3, D and E).

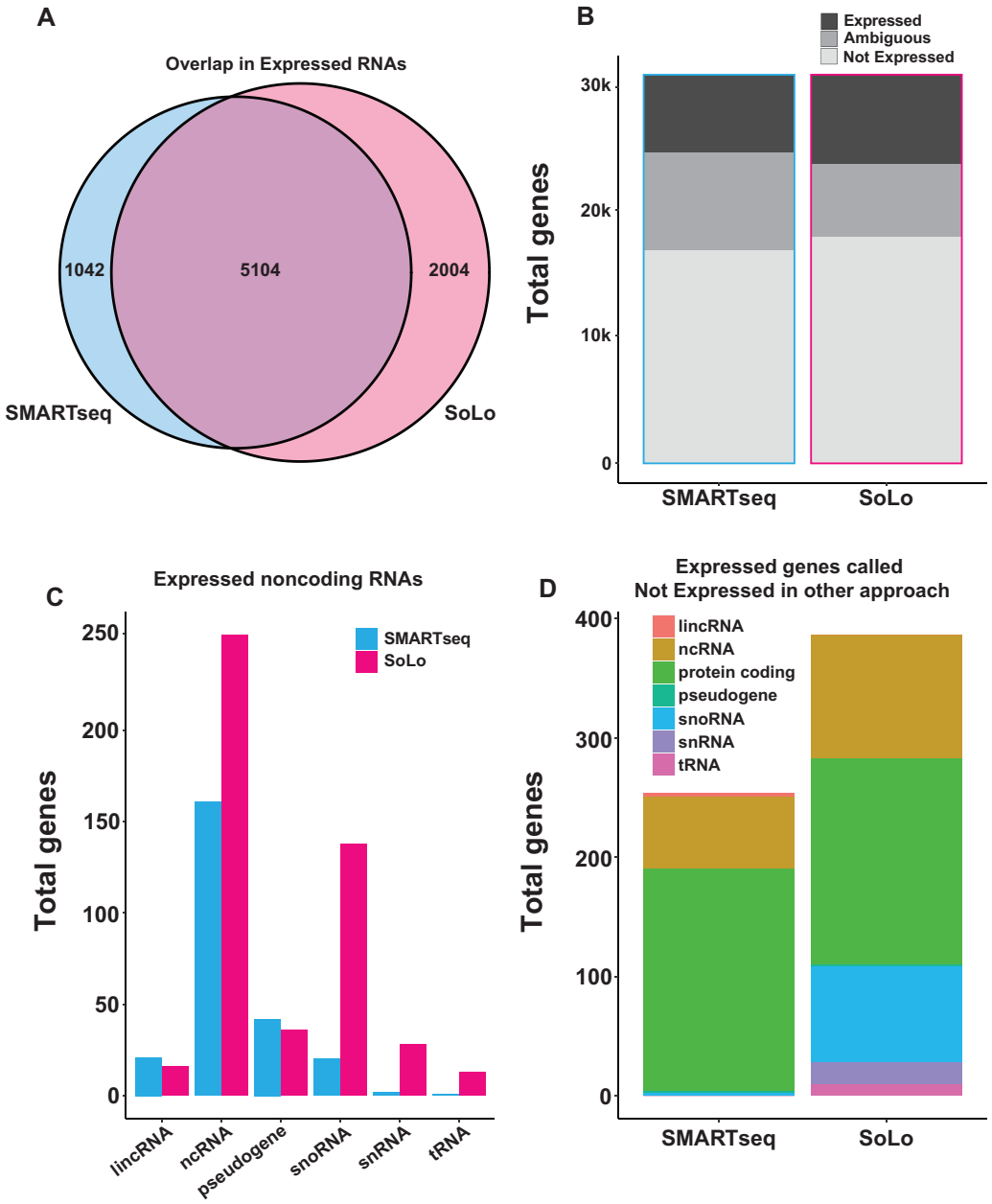

Figure 2 SoLo detects classes of noncoding RNAs missed in SMARTseq (A) Venn diagram showing the overlap between "expressed" genes (lower bound of CI >5 GeTMM) in SMARTseq and SoLo. (B) Bar graph showing gene detection for SMARTseq and SoLo using all genes, with three levels based on CI (CI): "expressed" (lower bound of CI >5 GeTMM), "ambiguous" (CI overlaps 5 GeTMM), and "not expressed" (upper bound of CI <5 GeTMM). (C) Bar graph showing the number of noncoding RNAs called "expressed" (lower bound of CI >5 GeTMM) in SMARTSeq (blue) and SoLo (magenta), in separate categories for each RNA type. (D) Bar graph showing genes called "expressed" (lower bound of CI >5 GeTMM) in one technique, and "not expressed" (upper bound of CI <5 GeTMM) in the other, broken down by gene type.

We considered the possibility that ambiguity in gene expression might correlate with lower gene expression, since lowly expressed genes might be more prone to noise. Dispersion is a measure of variance calculated when fitting expression data to a negative binomial model. We estimated the library-to-library dispersion of each protein-coding gene within each technique using edgeR, and plotted these values against average gene expression. We observed that SMARTseq had more dispersion than SoLo across all GeTMM values. However, this difference is strongest among lowly expressed genes (Figure 3, F and G, Supplementary Figure S2A). This difference was tested by first grouping genes into quintiles of expression within the SMARTseq and SoLo samples (lowest 20 to highest 20%) and performing Wilcoxon rank sum tests for each quintile. Tests for all five quintiles showed significant differences in dispersion scores between SMARTseq and SoLo ($P < 0.0001$). The difference in the average dispersion per quintile was highest for the lowest 20% of expressed genes (Supplementary Table S2). Comparing CI size (another indicator of variance), on a gene-by-gene level, reveals a similar trend (Supplementary Figure S2B). Wilcoxon tests comparing CI interval size by quintile of expression as above showed significant differences between SMARTseq and SoLo in all quintiles ($P < 0.0001$). These data suggest that SoLo produces consistent values for protein-coding genes than SMARTseq across a wider range of expression levels, and that at least some of the difference in genes confidently called "expressed" or "not expressed" is explained by intra-technique variance among low expressed genes.

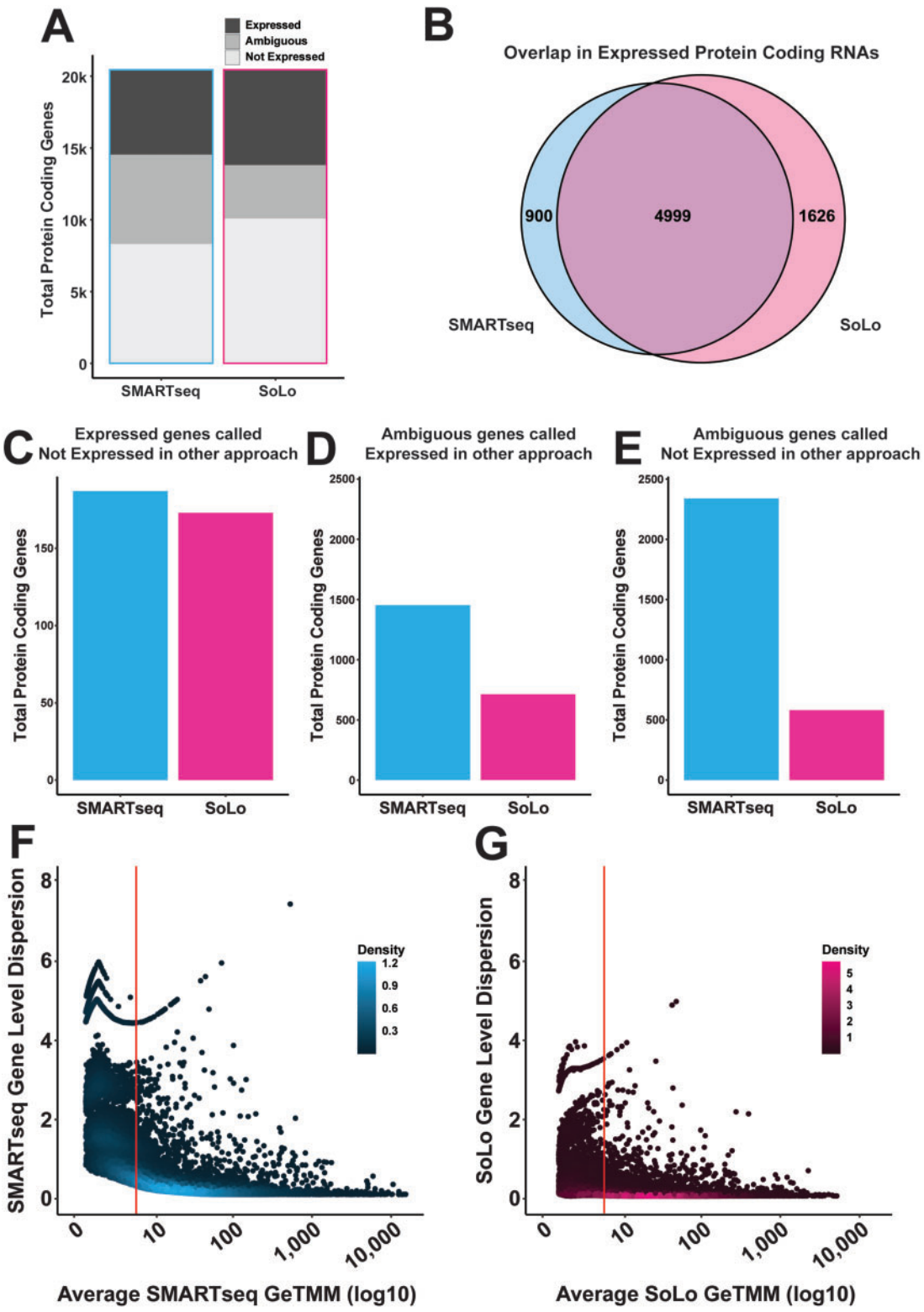

Figure 3 SMARTseq protein-coding genes show higher dispersion (A) Venn diagram showing the overlap between "expressed" protein-coding genes in SMARTseq and SoLo. (B) Bar graph showing gene detection for SMARTseq and SoLo using protein-coding genes, with three levels based on CI: "expressed" (lower bound of CI >5 GeTMM), "ambiguous" (CI overlaps 5 GeTMM), and "not expressed" (upper bound of CI <5 GeTMM). (C) Bar graph showing the number of protein-coding genes called "expressed" (lower bound of CI >5 GeTMM) in one technique and "not expressed" (upper bound of CI <5 GeTMM) in the other. (D) Bar graph showing the number of protein-coding genes called "ambiguous" (CI overlaps 5 GeTMM) in one technique and "expressed" (lower bound of CI >5 GeTMM) in the other. (E) Bar graph showing the number of protein-coding genes called "ambiguous" (CI overlaps 5 GeTMM) in one technique and "not expressed" (upper bound of CI <5 GeTMM) in the other. (F) Scatter plot showing the edgeR gene level dispersion estimate against average log10 GeTMM levels for SMARTseq protein-coding genes (average SMARTseq GeTMM >0.5). Red line shows 5 GeTMM. (G) Scatter plot showing the edgeR gene level dispersion estimate against average log10 GeTMM levels for SMARTseq protein-coding genes (average SoLo GeTMM >0.5). Red line shows 5 GeTMM. Wilcoxon tests comparing dispersion estimates for quintiles of expressed protein-coding genes were significant ($P < 0.0001$) for all comparisons.

SoLo shows enhanced detection of long genes

Besides noise, a potential source of the differences in protein-coding gene expression between the techniques is bias that depends on gene length. In general, RNAseq expression analysis counts the number of reads per gene, which is dependent on the number of cDNA fragments from that gene in the sequencing library. Longer genes have the potential to be represented in the library by more fragments, and thus accumulate more reads than short genes with the same number of RNA molecules in the sample. Thus, read counts must be normalized to the known length of the transcript in the genome assembly—in essence, reads per gene are divided by transcript length. However, this normalization approach assumes that read abundance increases linearly with read length, across all lengths. Length-dependent bias can occur if, for example, reads are depleted from the 5' end of long genes, but not short genes. In polyA-primed approaches such as SMARTseq, this depletion can occur due to RNA degradation, particularly of longer transcripts, or due to incomplete processivity during reverse transcription. To test the idea that gene length correlates with differences between the two approaches, we examined the gene length distribution for protein-coding genes called "expressed" only in SMARTseq or SoLo and compared these unique genes to the distribution of gene lengths for all protein-coding genes (Figure 4A). We found that SMARTseq exclusive genes (median 1.07 kb) are generally very close to the distribution for all protein-coding genes (median 1.09 kb), whereas SoLo exclusive genes are enriched for longer genes (median 1.74 kb) (SMARTseq: $P = 0.022$; SoLo: $P = 1.07 \times 10^{-15}$). This analysis suggests that differences in library construction may result in length-dependent biases in the data.

As an additional test of this result, we examined read coverage across the length of each gene in all samples. Using the RSeQC suite to measure average coverage (Wang *et al.* 2012), we found that SMARTseq coverage drops off near the 5' end. SoLo samples show no such drop-off and generally cover the entire length of the gene (Figure 4B). Given this difference, we hypothesized that longer genes may be especially prone to low 5' end coverage in SMARTseq libraries. We defined the 5' end as the first 20% of each gene and found that 5' end coverage tends to decrease for both SMARTseq and SoLo as gene length increases, but that the drop-off is much steeper for SMARTseq (Figure 4, C and D). Comparing the 5' coverage at the gene level shows that 178 genes have higher coverage in SMARTseq, whereas 1641 genes have significantly higher coverage in SoLo (paired *T*-test, *P*-adjusted < 0.05). Plotting the ratio of 5' coverage between SoLo and SMARTseq similarly shows that as gene length increases, SoLo tends to have better coverage of the 5' end than SMARTseq [Linear model: 0.872 × log10(Length)—2.084, $R^2 = 0.073$, *P*-value < 0.001] (Figure 4E).

To determine whether these differences in length and coverage translate to differences in GeTMM levels, we compared log fold change values between SoLo and SMARTseq of the shortest 2000 mRNAs to the longest 2000 mRNAs, and found that the shortest genes had a wide distribution centered close to zero (median log2 fold change = −0.118), and the longest genes had a narrower distribution with most genes showing higher values in SoLo (median log2 fold change = 0.68) ($P < 0.001$) (Figure 4F). The relationship between dispersion and gene length is also different in SoLo and SMARTseq. Plotting gene level dispersion against average intra-technique GeTMM values reveals that gene dispersion is much lower in SoLo samples in the longest mRNAs (Supplementary Figure S3). The longest 2000 genes are overrepresented in genes called "expressed" in SoLo but "not expressed" in SMARTseq (2.5x expected), and underrepresented in the reverse comparison (0.27x expected). These results suggest that the longest genes make up ∼20% of the SoLo exclusive genes. Together these data demonstrate that SoLo library preparation method results in both higher expression and better detection for longer genes.

SoLo and SMARTseq show strong overlap with previous pan-neuronal gene sets

We benchmarked the two techniques against a published *C. elegans* pan-neuronal dataset to assess how well they replicate previous results. The Kaletsky dataset includes 8437 protein-coding genes called expressed in *C. elegans* neurons (Kaletsky *et al.* 2018). Although the Kaletsky dataset was also derived from FACS-enriched neurons, the starting strain and the library construction methods differ (see Supplementary Table S3). Of the 8437 Kaletsky expressed genes, 5215 (61.8%) were called "expressed" in SMARTseq (Figure 5A). Of the remaining 3222 genes called expressed in Kaletsky, 617 were called "not expressed" in SMARTseq, while another 2605 (30.9%) were ambiguous. SoLo called 6231 (73.9%) of the Kaletsky expressed protein-coding genes "expressed" (Figure 5B). Of the remaining 2206 Kaletsky expressed protein-coding genes, 456 were called "not expressed" in SoLo, while 1740 (20.6%) were "ambiguous." Thus, both techniques have a broad agreement with previous data, with some minor differences.

The Kaletsky dataset also defines 867 neuronal enriched protein-coding genes when compared to muscle, hypodermis, and intestinal cells. Of these, 792 (91.1%) were called "expressed" in SMARTseq (Figure 5C), and 808 (92.9%) were called "expressed" in SoLo (Figure 5D). Of the remaining genes found to be neuronal enriched in the Kaletsky gene set, 70 genes were called "ambiguouos in SMARTseq (8.1%), and 56 were called "ambiguous" in SoLo (6.5%). These data show that for neuronal protein-coding genes in the Kaletsky gene set, >90% of expressed genes and >99% of enriched genes are called either "expressed" or "ambiguous" in both SoLo and SMARTseq, with minor differences explained by a mix of slight differences in contamination from other cell types during FACS, and different experimental parameters (Supplementary Table S3). Thus, our data from SoLo and SMARTseq approaches appear to strongly replicate previous findings for neuronal gene enrichment.

Discussion

Here, we performed a head-to-head comparison of ribodepletion and polyA selection approaches for RNAseq library preparation using low input samples from *C. elegans*. Using RNA from FACS-isolated neurons, we evaluated the performance of SoLo Ovation (Tecan Genomics) and SMARTseq V4 (Takara) library preparation methods for rRNA depletion efficiency, overall library complexity and gene detection. Our results indicate that both techniques efficiently removed rRNA from the final libraries, although SMARTseq performed better than SoLo. SoLo libraries had fewer PCR duplicates than SMARTseq and detected more reads in UTRs. It is somewhat surprising that SoLo libraries contained fewer putative PCR duplicates than SMARTseq libraries from the same RNA inputs, considering that SoLo preparations also produced an order of magnitude more cDNA (Supplementary Table S1). This effect may be due to differences in where rRNA depletion occurs in the protocol. In the SMARTseq protocol, rRNA is selected against in the first step through preferential cDNA

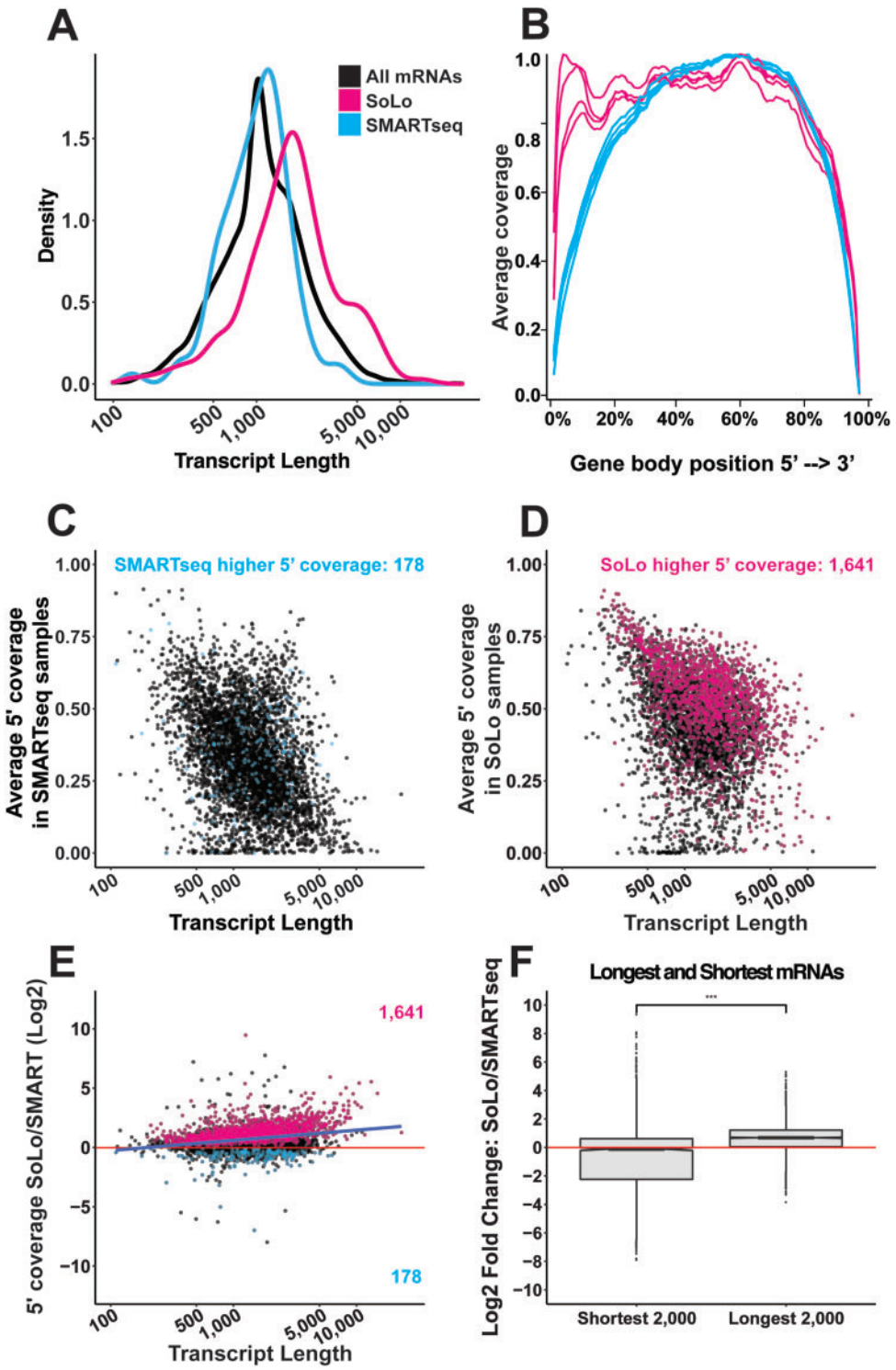

Figure 4 SoLo shows higher expression for long genes (A) Density graph showing the transcript length distribution for all protein-coding mRNAs (black), mRNAs "expressed" (lower bound of CI >5 GeTMM) in SMARTseq but "not expressed" (upper bound of CI <5 GeTMM) in SoLo (blue), and mRNAs "expressed" in SoLo but "not expressed" in SMARTseq (magenta). (B) Line plot showing the average normalized gene body coverage for all protein-coding genes >100 bp. Left to right, 5' to 3'. SMARTseq replicates shown in blue, Solo replicates shown in magenta. (C) Scatterplot showing the average SMARTseq coverage of the 5' end of all protein-coding genes called "expressed" in both SMARTseq and SoLo. One hundred and seventy-eight genes found with significantly higher 5' coverage in SMARTseq colored blue, paired *t*-test, BH adjusted *P*-value < 0.05. (D) Scatterplot showing the average SoLo coverage of the 5' end of all protein-coding genes called expressed in both SMARTseq and SoLo. One thousand six hundred and forty-one genes found with significantly higher 5' coverage in SoLo colored magenta, paired *t*-test, BH adjusted *P*-value < 0.05. (E) Scatter plot showing the log2 fold ratio of SoLo and SMARTseq 5' gene coverage for protein-coding genes called "expressed" in both SMARTseq and SoLo. Significant genes called higher in SoLo colored magenta, genes called higher in SMARTseq colored blue. Paired *t*-test, BH adjusted *P*-value < 0.05. Linear model: 0.872* log10(Length)—2.084, $R^2 = 0.073$. (F) Box plot showing the edgeR log fold change for gene expression among the 2000 shortest protein-coding genes (length >100 bp) and 2000 longest protein-coding genes, SoLo/SMARTseq. Wilcoxon test, ***: *P*-value < 0.001.

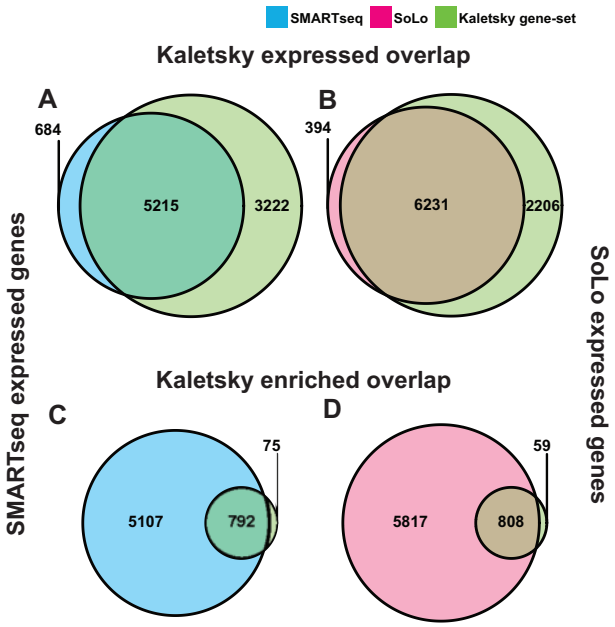

Figure 5 SoLo and SMARTseq detect neuronal genes (A, B) Overlap of "expressed" protein-coding genes (lower bound of CI >5 GeTMM) between SMARTseq (blue) or SoLo (magenta) and Kaletsky gene-set neuronal expressed genes (green). (C, D) Overlap of "expressed" protein-coding genes (lower bound of CI >5 GeTMM) between SMARTseq (blue) or SoLo (magenta) and Kaletsky gene-set neuronal enriched genes (green).

synthesis from polyA RNA. Downstream steps amplify only the targeted molecules across 16–20 rounds of PCR depending on the sample input, and any stochastic over-amplification will directly affect the targeted RNA species. For SoLo samples, all RNA is reverse transcribed to cDNA prior to initial amplification. In these initial 11–17 rounds of PCR amplification, as rRNA sequences comprise the bulk of the cDNA, they are more likely to be represented in over-amplified products. Selected cleavage of rRNA adapters targets sequences that are already prone to stochastic over-amplification by virtue of their abundance. Thus, rRNA depletion after the first round of cDNA amplification in SoLo may partially protect against duplicate reads dominating the library during amplification of low input samples.

Noncoding RNAs play critical roles in gene regulation and cell function, and detecting ncRNAs is key to fully understanding the transcriptome of any cell type. As many noncoding RNAs lack polyA tails, they may differ substantially in detection between SMARTseq and SoLo techniques. As expected, poly-adenylated noncoding RNAs, lincRNAs and pseudogenes, are detected at least as well by SMARTseq as they are by SoLo. Four classes of noncoding RNAs lacking polyA tails are detected at much higher rates in SoLo than SMARTseq: tRNAs, snRNAs, snoRNAs, and uncategorized ncRNAs. Of these, tRNAs present their own challenges in sequencing, given their highly modified structure that often impedes RT-PCR. The robust detection is seen here and the stark difference shown between approaches suggests that SoLo may be better suited to tRNA detection than SMARTseq, although further experiments are needed to confirm this finding. Overall, our results show that SoLo outperforms SMARTseq at detecting noncoding RNA species.

Given that each approach should theoretically treat polyadenylated protein-coding transcripts roughly the same, we set out to investigate whether these RNAs were detected at the same rate. The majority of protein-coding genes called "expressed" in SoLo samples were also called "expressed" in SMARTseq, however, SMARTseq was more prone to calling genes "ambiguous" than SoLo. We explain this difference by observing that estimated gene-level dispersion is markedly higher among lowly expressed genes in SMARTseq compared to SoLo. This result suggests that SoLo may provide more confidence in calling genes expressed or not expressed, especially for genes that are expressed at low values.

While noise appears to drive much of the difference in the confidence of gene expression calling, we also investigated whether gene length bias drove differences in protein-coding gene detection between the techniques. In studying the length distribution for high confidence and exclusive genes for each technique, we found that SMARTseq shows no clear deviation from the distribution of all protein-coding transcripts, but SoLo shows a ~700 bp increase in median transcript length. This finding corresponds with data on average gene body coverage which shows SoLo having much more uniform coverage, especially at the 5' end of the gene. On the basis of these findings, we hypothesized that long transcripts may be especially prone to reduced 5' coverage in the polyA SMARTseq approach. The longer the gene, the more opportunity for degradation that severs the 5' end from the 3' polyA priming site. In addition, longer genes may be more vulnerable to losing coverage of the 5' end if the reverse transcriptase enzyme falls off of the RNA molecule prior to reaching the end. To test this idea, we measured coverage of the 5' section of each gene and found that, for genes detected in both techniques, the ratio of SoLo coverage to SMARTseq coverage tends to increase with transcript length. By comparing the edgeR log fold changes for the shortest and the longest protein-coding transcripts we also found that while the shortest genes showed a wide range of fold changes centered close to zero, the longest genes were primarily enriched in SoLo, and were overrepresented in genes detected in SoLo and not detected in SMARTseq. Taken together these results suggest that expression of longer genes is generally prone to being underestimated, and that ribodepletion based techniques like SoLo are less vulnerable to this deficit. This disparity is unlikely to affect comparisons that focus solely on relative expression of a given transcript between conditions. However, the relative abundance of longer *vs* shorter genes within each condition could be underestimated due to this bias.

Kaletsky *et al.* (2018) published protein-coding gene expression and enrichment lists for several *C. elegans* tissues, including a pan-neuronal dataset using an *unc-119* fluorescent reporter. This data set provided an opportunity to assess how well our SoLo and SMARTseq *rab-3* pan-neuronal libraries reproduce previous results. The comparison showed substantial overlap of SoLo/SMARTseq "expressed" protein-coding genes with the Kaletsky dataset (Figure 5, A and B), similar to our comparisons between the two library preparation techniques which use identical RNA samples (Figure 3A). Other differences between the Kaletsky data set and our results could be due to different fluorescent markers; the use of animals at different developmental stages; and differences in library preparation and gene thresholding procedures.

Overall, our findings suggest that SoLo Ovation, using a custom probe set to deplete *C. elegans* rRNA, outperforms SMARTseq with ultra-low input RNAseq samples by detecting an expanded set of noncoding RNAs, providing reduced noise for lowly expressed genes, and more accurate counts for long genes. Application of this technique, for example in efforts to profile all *C. elegans* neurons (Hammarlund *et al.* 2018), should result in increased knowledge of cellular expression of diverse RNA molecules.

Acknowledgments

Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440).

Funding

This work was funded by the National Institutes of Health grant R01NS100547 to MH, OH, DMM, and NS and by Vanderbilt Trans-Institutional Program funds to DMM. Flow Cytometry experiments were performed in the Vanderbilt Flow Cytometry Shared Resource which is supported by the Vanderbilt Ingram Cancer Center (P30 CA68485) and the Vanderbilt Digestive Disease Research Center (DK058404). The Vanderbilt VANTAGE Core provided technical assistance for this work and is supported by CTSA Grant (5UL1 RR024975-03), the Vanderbilt Ingram Cancer Center (P30 CA68485), the Vanderbilt Vision Center (P30 EY08126), and NIH/NCRR (G20 RR030956).

Conflicts of interest

None declared.

Literature cited

Ahn RS, Taravati K, Lai K, Lee KM, Nititham J, *et al.* 2017. Transcriptional landscape of epithelial and immune cell populations revealed through FACS-seq of healthy human skin. Sci Rep. 7:1343.

Camacho J, Truong L, Kurt Z, Chen Y-W, Morselli M, *et al.* 2018. The memory of environmental chemical exposure in *C. elegans* is dependent on the jumonji demethylases jmjd-2 and jmjd-3/utx-1. Cell Reports. 23:2392–2404.

Cao S, Ma T, Ungerleider N, Roberts C, Kobelski M, *et al.* 2019. Circular RNAs add diversity to androgen receptor isoform repertoire in castration-resistant prostate cancer. Oncogene. 38: 7060–7072.

Cech TR, Steitz JA. 2014. The noncoding RNA revolution—trashing old rules to forge new ones. Cell. 157:77–94.

Ching T, Masaki J, Weirather J, Garmire LX. 2015. Non-coding yet non-trivial: a review on the computational genomics of lincRNAs. BioData Mining. 8:44–44.

Culviner PH, Guegler CK, Laub MT. 2020. A simple, cost-effective, and robust method for rRNA depletion in RNA-sequencing studies. mBio. 11:e00010-20.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, *et al.* 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England). 29:15–21.

Hammarlund M, Hobert O, Miller III DM, Sestan N. 2018. The CeNGEN project: the complete gene expression map of an entire nervous system. Neuron. 99:430–433.

Herbert ZT, Kershner JP, Butty VL, Thimmapuram J, Choudhari S, *et al.* 2018. Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction. BMC Genomics. 19:199.

Kaletsky R, Lakhina V, Arey R, Williams A, Landis J, *et al.* 2016. The *C. elegans* adult neuronal IIS/FOXO transcriptome reveals adult phenotype regulators. Nature. 529:92–96.

Kaletsky R, Yao V, Williams A, Runnels AM, Tadych A, *et al.* 2018. Transcriptome analysis of adult *Caenorhabditis elegans* cells reveals tissue-specific gene and isoform expression. PLoS Genet. 14:e1007559.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.* 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25: 2078–2079.

Lim JP, Brunet A. 2013. Bridging the transgenerational gap with epigenetic memory. Trends Genet. 29:176–186.

Mok DZL, Sternberg PW, Inoue T. 2015. Morphologically defined sub-stages of *C. elegans* vulval development in the fourth larval stage. BMC Dev Biol. 15:26.

Nonet ML, Staunton JE, Kilgard MP, Fergestad T, Hartwieg E, *et al.* 1997. *Caenorhabditis elegans* rab-3 mutant synapses exhibit impaired function and are partially depleted of vesicles. J Neurosci. 17:8061–8073.

O'Neil D, Glowatz H, Schlumpberger M. 2013. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Prot Mol Biol. 103: 4–19.

Petrova OE, Garcia-Alcalde F, Zampaloni C, Sauer K. 2017. Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. Sci Rep. 7:41114.

Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, *et al.* 2011. Pseudogenes: pseudo-functional or key regulators in health and disease? RNA (New York, NY). 17:792–798.

Posner R, Toker IA, Antonova O, Star E, Anava S, *et al.* 2019. Neuronal Small RNAs Control Behavior Transgenerationally. Cell. 177: 1814–1826.

Ransohoff JD, Wei Y, Khavari PA. 2018. The functions and unique features of long intergenic non-coding RNA. Nat Rev Mol Cell Biol. 19:143–157.

Serra L, Chang DZ, Macchietto M, Williams K, Murad R, *et al.* 2018. Adapting the smart-seq2 protocol for Robust single Worm RNA-seq. Bio-protocol. 8:e2729.

Smid M, Coebergh van den Braak RRJ, van de Werken HJG, van Riet J, van Galen A, *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.

Spencer WC, McWhirter R, Miller T, Strasbourger P, Thompson O, *et al.* 2014. Isolation of specific Neurons from *C. elegans* larvae for gene expression profiling. PLoS One. 9:e112102.

Taylor SR, Santpere G, Reilly M, Glenwinkel L, Poff A, *et al.* 2019. Expression profiling of the mature C. elegans Nervous system by single-Cell RNA-sequencing. bioRxiv: 737577.

Tintori SC, Nishimura EO, Golden P, Lieb JD, Goldstein B. 2016. A Transcriptional lineage of the early *C. elegans* embryo. Dev Cell. 38:430–444.

Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 28:2184–2185.

Warner AD, Gevirtzman L, Hillier LW, Ewing B, Waterston RH. 2019. The *C. elegans* embryonic transcriptome with tissue, time, and alternative splicing resolution. Genome Res. 29:1036–1045.

Zhang S, Banerjee D, Kuhn JR. 2011. Isolation and culture of larval cells from *C. elegans*. PLoS One. 6:e19505.

Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, *et al.* 2014. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics. 15:419–419.

Zullo JM, Drake D, Aron L, O'Hern P, Dhamne SC, *et al.* 2019. Regulation of lifespan by neural excitation and REST. Nature. 574:359–364.

Communicating editor: J. Ward