CRISPR/Cas9-based depletion of 16S ribosomal RNA improves library complexity of single-cell RNA-sequencing

Kuang-Tse Wang[1], Carolyn E. Adler[1,2]

1 Department of Molecular Medicine, Cornell University College of Veterinary Medicine, Ithaca, NY, USA

2 Corresponding author: cea88@cornell.edu

Abstract

Single-cell transcriptomics (scRNA-seq) has revolutionized our understanding of cell types and states in various contexts, such as development and disease. To selectively capture protein-coding polyadenylated transcripts, most methodologies rely on poly(A) enrichment to exclude ribosomal transcripts that constitute >80% of the transcriptome. However, it is common for ribosomal transcripts to sneak into the library, which can add significant background by flooding libraries with irrelevant sequences. The challenge of amplifying all RNA transcripts from a single cell has motivated the development of new technologies to optimize retrieval of transcripts of interest. This problem is especially striking in planarians, where a single 16S ribosomal transcript is widely enriched (20-80%) across single-cell methods. Therefore, we adapted the Depletion of Abundant Sequences by Hybridization (DASH) to the standard 10X scRNA-seq protocol. We designed single-guide RNAs tiling the 16S sequence for CRISPR-mediated degradation, and subsequently generated untreated and DASH-treated datasets from the same libraries to enable a side-by-side comparison of the effects of DASH. DASH specifically removes 16S sequences without off-target effects on other genes. By assessing the cell barcodes shared by both libraries, we find that DASH-treated cells have consistently higher complexity given the same amount of reads, which enables the detection of a rare cell cluster and more differentially expressed genes. In conclusion, DASH can be easily integrated into existing sequencing protocols and customized to deplete unwanted transcripts in any organism.

Keywords

scRNA-seq, ribodepletion, CRISPR-Cas9, DASH, planarians

Background

Recent advances in single-cell transcriptomics (scRNA-seq) have greatly facilitated the exploration of cell type complexity. By capturing and barcoding mRNA on a cell-by-cell basis, scRNA-seq enables researchers to identify different cell types in a wide range of species. However, an important aspect of scRNA-seq is quality control in the analysis, which involves removing potential confounding factors such as sequencing depth and the percentage of mitochondrial reads[1,2]. While these factors can be normalized *in silico*, it is optimal to minimize unwanted transcripts during library preparation to improve efficiency and more accurately probe cell complexity.

To maximize retrieval of protein-coding transcripts, most scRNA-seq methods capture mRNA by poly(A) enrichment[3,4]. Ideally, poly(T) primers will selectively bind to polyadenylated protein-coding transcripts to enable reverse transcription. This approach effectively eliminates ribosomal RNA, which comprises >80% of total RNA. Despite its utility, poly(A) enrichment also prevents the capture of some informative RNA species, including histone mRNA, miRNA, enhancer RNA, etc[5]. In addition, unwanted transcripts such as mitochondrial RNA leak out of damaged cells and are carried through into cDNA libraries[2]. These contaminants often appear as ambient RNA, a computational challenge for cell calling algorithms that can introduce bias into normalization[6,7]. Due to these issues, poly(A) enrichment can distort the global transcriptional landscape.

Alternative ribodepletion methods have recently been developed for single-cell transcriptomics. For example, VASA-seq generates cDNA by hexamer primers with a T7 handle to enable

subsequent RNA amplification by in vitro transcription. The ribosomal RNAs will then be removed from amplified RNA by rDNA probes and RNAse-H-mediated digestion[8]. RamDA-seq uses specific hexamers that are not present in rRNA to exclude it from downstream steps[9,10]. However, these methods require new chemistry and reagents, raising the entry barrier for in-house setup.

Depletion of Abundant Sequences by Hybridization (DASH) is a CRISPR/Cas9-based method to remove unwanted DNA sequences from any DNA library[11]. With the customized design of single guide RNA (sgRNA), Cas9 can precisely remove unwanted cDNA sequences, and further enhance the detection of rare transcripts. DASH has not only been shown to efficiently deplete unwanted sequences in 16S sequencing and bulk RNA-seq but has been adapted to single-cell transcriptome methods, including scCLEAN, Smart-seq-total and MATQ-seq[12,13]. However, whether this depletion impacts library complexity or other metrics of sequencing quality has yet to be tested systematically.

Planarian mitochondrial 16S rRNA sequences are known to make up ~30% of bulk RNA-seq libraries generated by poly(A) enrichment[14]. In scRNA-seq experiments, most studies remove 16S from the analysis, but to what extent 16S rRNA contaminates scRNA-seq remains unclear[15,16]. Here, we reanalyzed published datasets and found that the 16S UMIs constitute approximately 20-80% of sequencing reads regardless of the single-cell library generation strategy. We adapted DASH, which leverages CRISPR to remove unwanted sequences, to the 10X Chromium protocol to deplete mitochondrial 16S cDNA from planarian sequencing libraries[14]. We sequenced the same library with and without DASH treatment and carried out a

side-by-side analysis to determine the impact of DASH treatment on overall scRNA-seq performance. We showed that our protocol specifically depletes more than 90% of 16S UMIs from both cells and ambient RNA. This depletion increases the number of genes and non-16S UMIs detected per cell which improves the downstream analysis. We conclude that DASH can enhance library complexity, boosting the information retrieved from scRNA-seq experiments, with significant economic benefits. Importantly, this technique can be easily customized to deplete any abundant transcript after single-cell libraries are generated.

Results

Mitochondrial 16S transcript dominates reads in planarian scRNA-seq datasets

Planarians were one of the first animal models to be profiled at the whole organismal level using scRNA-seq. Because several different sequencing strategies have been applied to planarians, we first asked which method has the highest library complexity and the least 16S rRNA contamination. Thus, we collected 13 datasets from 7 studies using Drop-seq[15,17], Smart-seq2[18], Split-Seq[19], Split-Seq with ACME fixation[20], and 10X Chromium [16,21] (Figure 1). All methods use poly(T) primers to capture transcripts except for Split-Seq, which uses a mixture of random hexamers and poly(T) primers. To standardize these comparisons across libraries, we used the same pipeline to pre-process the barcodes and UMI tagging, align the reads to the genome (including the mitogenome), and generate quality metrics[22,23].

To assess the abundance and quality of these different sequencing strategies, we applied several metrics. First, we analyzed the percentage of reads that mapped to the 16S transcript. Libraries made with 10X Chromium had the highest percentage of UMIs mapping to this transcript (61%

and 74% on average) (Figure 1A). By contrast, SplitSeq libraries were the lowest (ranging from 5% to 8%). DropSeq and ACME-based strategies ranged from 22 to 52%. Another mitochondrial rRNA, 12S UMIs, was consistently low across all datasets (Figure 1B). These findings showed that in various scRNA-seq methods, the single 16S transcript accounted for a significant fraction of UMIs. Although Split-Seq appears to have the lowest percentage of UMIs mapping to the 16S, these libraries also had the highest number of unmappable reads (Table 1). Next, we examined library complexity by assessing the number of genes and UMIs per cell. While 10X Chromium retained the highest levels of 16S UMIs as compared to other tested library methodologies, it also yielded the greatest library complexity (Figure 1C-D). In conclusion, these results show that 16S contamination is a severe problem in scRNA-seq preparation for planarians.

Because all of the scRNAseq library strategies rely on poly(A) enrichment, we hypothesized that 16S may be retained because of either polyadenylation or an internal polyA stretch[24,25]. To distinguish between these two possibilities, we analyzed the coverage of reads across the 16S locus in one of our datasets using 10X Chromium. We found that the reads from a 10X dataset were strongly skewed to the 3' end (Supplementary figure 1A), similar to other protein-coding genes, suggesting that the 16S may be polyadenylated (Supplementary figure 1B). We also identified poly(A) stretches 4 or 5 base pairs long in the middle (Supplementary figure 1B), which could also be responsible for the 16S being pulled down. Still, we sought an alternative method to exclude 16S rRNA for scRNA-seq preparation.

DASH effectively depletes 16S cDNA in scRNA-seq libraries

Due to the widespread use of the 10X Chromium scRNA-seq platform, and because it yields the highest library complexity in planarians, we sought to optimize the 10X protocol by selectively removing the 16S sequence. During single-cell RNAseq library preparation, reverse transcription occurs immediately after cell lysis, concomitant with barcoding, so any depletion must happen during or after this step[4]. Ribodepletion methods that remove unwanted RNA, such as RNAse-H mediated digestion, are difficult to integrate with 10X scRNA-seq because this approach requires RNA-DNA hybrids[26]. Depletion of Abundant Sequences by Hybridization (DASH) is a method that effectively depletes cDNA in a sequence-specific manner, making it a promising method to selectively deplete the 16S transcript after barcoding[11]. Therefore, we designed 30 non-overlapping single guide RNAs (sgRNAs) tiling the entire 16S transcript. We reasoned that degrading 16S as early as possible after cDNA conversion would be optimal to preserve the transcriptional repertoire. We integrated DASH into the 10X Chromium workflow by only performing 10 PCR cycles after cDNA conversion, then incubating the cDNA with pooled sgRNAs complexed with Cas9. After CRISPR/Cas9 degradation, we further amplified cDNA with 10 additional PCR cycles, followed by the standard end repairing and indexing steps (Figure 2A).

To test whether DASH could deplete 16S sequences from cDNA libraries, we first generated three 10X 3' scRNA-seq cDNA libraries from planarian stem cells. Next, we split the cDNA libraries and generated DASH-treated ('DASHed') libraries for each biological replicate (Figure 2B). By sequencing the same biological libraries before and after DASH treatment, we could benchmark the impact of DASH treatment at the single-cell level by analyzing shared cell barcodes (Figure 2C).

To assess the efficiency of 16S depletion, we compared untreated and DASHed libraries on the Bioanalyzer. Fragment analysis showed a strong peak around 1300bp in the untreated cDNA, which was absent in the DASHed cDNA, while the rest of the profile remains comparable. (Figure 2D). After sequencing, differential expression analysis showed that the 16S rRNA was the only downregulated gene in DASHed datasets (Figure 2E). This strong and specific decrease suggests that the highest peak in the untreated library may represent 16S cDNA and that it can be efficiently depleted by DASH.

Overrepresentation of 16S UMIs causes aberrant cell calling

Cell calling algorithms rely on the assumption that cells have significantly more UMIs than empty droplets and ambient RNA, so they can distinguish cells from non-cells by the total UMIs associated with each cell barcode[4]. We hypothesized that the high prevalence of 16S rRNA could potentially interfere with cell calling by inappropriately 'calling' cells that in reality are ambient RNA. To evaluate the effect of DASH treatment on cell calling, we obtained lists of cell barcodes from untreated and DASHed libraries generated by Cell Ranger for each of the 3 biological replicates (Figure 2C). The majority of cell barcodes were shared between untreated and DASHed datasets for each replicate (Supplementary figure 2A). Across all 3 replicates, the untreated datasets consistently had more cell barcodes called by Cell Ranger as cells (ranging from 8% to 20% of the total) (Supplementary Figure 2A). Of the barcodes that were unique to the untreated libraries, the vast majority of UMIs (>90%) mapped to 16S (Supplementary figure 2B). Barcodes shared by both untreated and DASHed libraries had 58-60% of 16S UMIs in untreated datasets, which is comparable to previously published datasets (Supplementary figure

2B; Figure 1A). By contrast, the barcodes that were specific to the DASHed libraries had very low levels of 16S UMIs (<0.2%), indicating that the depletion was thorough (Supplementary figure 3B). Together, these results suggest that 16S comprises a significant proportion of total UMIs in both cell and ambient RNA and misleads the cell calling process.

DASH depletes 16S rRNA and increases mRNA and gene recovery

In general, mitochondrial and ribosomal sequences are computationally removed from downstream analysis [15,16]. Because depletion from DASH is so specific, it may achieve the same effect as this computational processing step. Alternatively, the depletion of 16S at an early step in library generation could recover more sequencing information than computational removal after sequencing is complete. To assess the efficacy of 16S depletion on library complexity, we normalized untreated and DASHed scRNA-seq datasets to 100 million reads by downsampling and examined metrics of library quality. The 16S UMIs dropped from 60% to less than 0.2% in DASHed datasets, suggesting that the 16S cDNA was efficiently depleted (Figure 3A). Moreover, the total number of genes detected per cell increased (ranging from 27% to 40%) (Figure 3B), and the non-16S UMIs increased from 46% to 70% in DASHed datasets as compared to untreated libraries (Figure 3C). These findings suggest that DASH treatment can enrich library complexity further than computational depletion alone (Figure 3C-D).

To investigate whether the loss of complexity in the untreated library could be overcome by increasing sequencing reads, we performed a rarefaction analysis. This analysis assesses library complexity by measuring how diversity increases with sequencing reads until it reaches saturation. We downsampled the datasets to 10, 50, and 100 million reads for replicates 2 and 3,

and 150 and 200 million reads for replicate 1, and examined the library quality metrics in shared cell barcodes. We found that the number of genes detected per cell and non-16S UMIs were consistently higher in DASHed libraries, even at the lowest read depth (Figure 3D-E). These results indicate that the depletion of 16S rRNA consistently enhances library complexity given the same amount of reads, beyond what can be achieved by computational ribodepletion.

Post-CRISPR amplification is necessary and improves library complexity

In scRNA-seq methods, the number of PCR cycles to amplify cDNA is adjusted based on estimated cell numbers, tested empirically by the 10X manufacturer[4]. Over- or under-amplification leads to a decrease in library complexity. We asked if the post-CRISPR amplification is necessary and if so, what are the optimal PCR cycles that maximize library diversity without changing the overall fragment distribution. We sequenced replicate 1 without any re-amplification (0 cycles), or with 5, 10, and 15 PCR cycles, and assessed library complexity by downsampling into the same total reads. We excluded the cDNA amplified with 15 cycles from further sequencing because the cDNA trace changed drastically as compared to 5 and 10 cycles (Supplementary figure 3). After sequencing, the number of UMIs and genes per cell increased with more PCR cycles (Figure 4A-B). Similarly, the rarefaction analysis also showed that library complexity was consistently higher in the 10-cycle condition (Figure 4C-D). This finding suggests that post-CRISPR amplification is necessary for improving library complexity, with an optimal cycle number of 10, at least for these samples.

Since the extra PCR steps were necessary for improving library complexity, we asked whether this might introduce bias that enriches certain transcripts or cells that are abundant in the library.

Therefore, we compared the numbers of genes and UMIs in shared cells in untreated and DASHed libraries. We found that the number of genes and non-16S UMIs in each cell with and without DASH treatment showed strong linear correlations, suggesting the relative UMI abundance across cells maintained after DASH treatment (R^2>0.97) (Figure 5A-B). Slopes greater than 1 in this analysis are indicative of a global increase in library complexity across cells. Moreover, we asked if these increases exist locally across different read depths. When we binned the cells into different groups based on the number of genes or UMIs, the levels of increase were consistent across different groups (Figure 5C-D). Together, these findings show that DASH increases cell complexity in an unbiased manner.

DASH treatment improves downstream single-cell analysis

By reducing contamination and ambient transcripts, the ultimate goal of DASH depletion is to reduce background and improve detection of rare cell types or more potential marker genes. To assess the benefit of DASH treatment as compared to the computational removal of 16S, we removed 16S UMIs computationally from all libraries and then pooled the replicates from either untreated or DASHed libraries. We used the shared nearest neighbor (SNN) method in Seurat to cluster the cells in both datasets separately (Supplemental Figure 4A-B). Clusters were annotated and aligned between two datasets based on previously described marker genes [15,16](Supplementary figure 4C), except that cluster 13 could not be annotated due to the high percentage of 16S sequence (Supplementary figure 4D). Other clusters, with one exception, were present in both groups (Figure 6A-B). The unique cluster (C15) that appeared in the DASHed dataset represented *cathepsin*$^+$ cells that were dispersed in the untreated dataset (Figure 6A),

suggesting that the clustering is more sensitive due to the increased library complexity after DASH treatment.

Differential expression analysis in scRNA-seq often selects genes expressed in at least 10-25% of the cells within a cluster to obtain robust results [27]. On average, across the 17 DASHed clusters, 3090 genes were expressed in at least 25% of the cells, while in the 16 untreated clusters, 2101 genes were expressed in at least 25% of the cells, suggesting that the DASH treatment likely reduced dropout rates (Figure 6C). To identify differentially expressed (DE) genes for each cluster, we used the Wilcoxon Rank Sum test. Of the 17 clusters, 14 had more DE genes in DASHed than untreated datasets (Figure 6D). We conclude that DASH treatment reduces the dropout rate of genes, enabling the clustering algorithm and differential expression analysis to perform better under the same parameters.

Discussion

Sparsity of reads in single-cell sequencing is a technical challenge that arises from flooding of transcripts of housekeeping genes and ambient RNA that may swamp out detection of biologically informative genes[7,28]. In this study, we analyzed several different scRNA-seq methodologies used in planarians and show that the single transcript for the planarian 16S rRNA accounts for 20-80% of all reads. To eliminate this contaminant, we integrated a DASH depletion step into the protocol for scRNA-seq library generation, and demonstrate that depletion of this sequence benefits overall library complexity. Removing the 16S transcript early during library preparation enhances the discovery of mRNAs that differentiate potential cell types, which we showed by performing a parallel analysis of untreated and depleted libraries. In

conclusion, by eliminating unwanted reads, our approach improves detection of gene expression and economizes sequencing yield. Our approach is also customizable, and can be adapted to any system where similar contamination is evident.

An unusually high fraction of 16S ribosomal RNA in planarian datasets

The prevalence of mitochondrial UMIs has been widely used as quality control for identifying healthy cells in human and mouse datasets, where acceptable maximum levels range from 5-10% [29,30]. In the planarian *S. mediterranea*, mitochondrial 16S rRNA overflows RNA-seq experiments even after poly(A) enrichment[14]. Following poly(A) enrichment, 16S still makes up 11-32% of total reads in bulk RNA-seq, and worsens in various single-cell methods, comprising 20-80% of total UMIs. While in most studies, the abundance of mitochondrial RNA is thought to arise from damaged cells, this is an unlikely source of 16S in planarians because most studies used either fresh or FACS-purified live cells[14,15,17]. Ribosomal RNA transcribed in the nucleus is typically not polyadenylated, but reports have shown that both human and *Drosophila* mitochondrial 16S and 12S rRNA do get polyadenylated [25,31]. Our coverage plots show that sequencing reads skew toward the 3' end of 16S transcript, a trend resembling that of protein-coding genes. Alternatively, 3 poly(A) stretches in the middle of the sequences might also contribute to the abundance, but it is unlikely because if the reverse transcription starts in the middle, the 5' end would have higher coverage than the 3' end. Thus, we speculate that the planarian 16S might have an exceptionally long poly(A) tail, so it is captured significantly in scRNA-seq, which remains to be tested experimentally. Overall, we capitalize on the overrepresentation of 16S to investigate the impact of ribodepletion on scRNA-seq.

Alternative strategies for depleting abundant transcripts in single-cell RNA-seq

Recent advances in the ribodepletion of scRNA-seq have been shown to improve library complexity, but most of them require using new set-ups or novel reagents[8,10]. Our DASH-mediated approach is straightforward to implement and highly customizable to any system[11]. DASH has been shown to enhance library complexity in general, but has been incorporated into single-cell methods in different ways[11,12,32,33]. A common workflow of single-cell transcriptomes includes these steps in order: concomitant barcoding and reverse transcription, fragmentation of cDNA, and library indexing[3,4]. Here, we digest the 16S rRNA before the fragmentation and indexing steps of the 10X Chromium protocol. To minimize PCR bias and loss of rare transcripts, we used 10 PCR cycles, 2 cycles fewer than recommended by 10X Genomics, to amplify cDNA before CRISPR/Cas9 digestion. After CRISPR digestion, we conducted a post-CRISPR amplification to enrich cDNA diversity, which is necessary, resulting in less than 0.5% of total UMIs belonging to 16S. Although other single-cell total transcriptome methods include CRISPR/Cas9-mediated depletion strategies to remove ribosomal sequences, the depletion is not as complete as what we observe here (10% of reads for Smart-seq-total[12] and 34% for scDASH[32] mapped to the target genes). More recently, a technique called scCLEAN targets 255 housekeeping and ribosomal genes for removal during library preparation. Even after depletion, these reads still make up 8% of the total reads[33]. The strong depletion that we observe here may result from several factors (1) that the contaminant in planarians is just a single transcript, and therefore easier to remove, (2) that we used 30 sgRNAs to target one gene, compared to others that use fewer sgRNAs per gene, and (3) that eliminating target genes as early as possible in the library preparation may be beneficial.

Impact of ribodepletion by DASH

While other studies have shown the overall benefit of DASH on library quality, we have performed a parallel analysis of the same library before and after depletion to show the impact of DASH at single-cell resolution. This analysis reveals two key benefits of depletion. First, our data demonstrate that not depleting the 16S transcript leads to aberrant cell calling. The extra cells in untreated libraries are highly enriched in 16S UMIs, indicating that ambient RNA can falsify the cell calling process[34]. Moreover, the distribution of transcripts during initial PCR steps required for library generation is distorted by their presence. Second, depletion of 16S does not appear to introduce bias in the analysis. We conclude this because the increase in library complexity and fold change of genes increased uniformly across cells with different UMIs. Furthermore, clustering analysis showed overall agreement between untreated and DASH-treated datasets and improvement in detecting differentially expressed genes by reducing the dropout rates. These findings are important because most scRNA-seq normalization uses a “size factor”, which equalizes the cell read depth[35]. If the depletion happens disproportionately to the read depth across cells, normalization outcomes would be significantly altered by DASH. Future work using any depletion of a new panel of genes should look for any biased or off-target depletion.

In summary, we showcased the efficiency and robustness of DASH in scRNA-seq by demonstrating ribodepletion of 16S in planarians. The customizability of DASH will benefit any model organisms that may suffer from contamination of ambient RNA, or overabundance of irrelevant transcripts in important scRNA-seq experiments. The integration of DASH is a simple add-on to the current 10X protocol and therefore requires little expertise in developing new

single-cell protocols. In addition, because the depletion can be done at any time, our protocol offers significant flexibility for sequencing after library generation.

Conclusions

This study describes and benchmarks the ribodepletion of 16S rRNA by CRISPR-based treatment to improve scRNA-seq of planarians. Ribodepletion enhances the library complexity and performance of single-cell analysis. This demonstrates the benefit of ribodepletion in the cDNA library over *in silico* removal of ribosomal RNA.

METHODS

Worm care

Schmidtea mediterranea asexual clonal line CIW4 was raised in a recirculating water system supplemented with water containing Montjuïc salts (planaria water)[36,37]. Animals were fed with beef liver and cleaned once a week. Animals were transferred to static culture containing 50 µg/mL gentamicin for at least a week prior to use.

Cell sorting

For each biological replicate, 10 animals were dissociated into single-cell suspensions by dicing in CMFB buffer [calcium-magnesium-free solution with 1% BSA (400mg/L NaH_2PO_4, 800 mg/L NaCl, 1200 mg/L KCl, 800 mg/L $NaHCO_3$, 240 mg/L glucose, 1% BSA, 15 mM HEPES, pH7.3)] and nutating for 2 hours at room temperature. Cells were centrifuged at 500g for 5 min, resuspended, and filtered through a 30µm cell strainer (BD Biosciences, cat.no 340628) to remove debris. The concentration of filtered cells was calculated using a TC20 automated cell counter (Bio-Rad). After centrifugation, cell concentration was adjusted to

100,000 cells/mL with staining buffer [CMFB containing DRAQ5 (5 µM) and Calcein-AM (0.4 µM)] and nutated at room temperature for 5 min. X1 cells were gated for vital 4N cells (DRAQ5$^+$ Calcein-AM$^+$) on a Sony MA900 Cell Sorter. 100,000 cells were sorted and diluted to a concentration of 1000 cells/µL for subsequent library preparation.

sgRNA design and *in vitro* transcription

Detailed protocol for DASH is in Supplementary file 1. The 16S ribosomal RNA sequence was retrieved from the mitochondrial genome of *Schmidtea mediterranea* (NCBI accession number: NC_022448.1) and uploaded to IDT's Alt-R Custom Cas9 crRNA Design Tool. 30 non-overlapping seed regions were selected from the output of the Design Tool to ensure complete digestion of the entire 905 bp transcript. The primer sequences are listed in Supplementary file 1. To synthesize T7-flanking templates for sgRNA, PCR reactions were assembled following the Phusion High-Fidelity (NEB) protocol with final concentrations of primers: one of the sgRNA primers (0.2µM), T7RevLong (0.2µM), T7FwdAmp (1µM), T7RevAmp (1µM)[38,39]. PCR reactions were carried out as follows: 98°C 30 sec, repeating the steps of 98°C for 10 sec, 51°C for 10 sec, 72°C for 10 sec 30 times, and then 72°C for 2min. PCR products were run on agarose gels to determine whether primers remained. If they did, the PCR products were gel purified. To synthesize sgRNA, the concentration of templates of each sgRNA was measured by nanodrop and pooled equivalently. In vitro transcription reactions were assembled as follows: sgRNA templates (4µg), 10X transcription buffer [0.1 M $MgCl_2$, 0.4 M Tris (pH 8.0), 0.1 M DTT, 20 mM spermidine] (10µL), 25mM rNTPs(Promega) (8µL), T7 polymerase (in-house) (2µL), TIPP(NEB) (2µL), rRNAsin(Promega) (1µL) and nuclease-free water (adjusting the total volume to 100 µL)[40]. In vitro transcription reactions were incubated at 37°C overnight. The next

day, 2 µL RQ1 DNase was added to remove templates and incubated at 37°C for 20 min. To precipitate the sgRNAs, 250 µL ice-cold 100% ethanol was added to each reaction and incubated at -20°C for 1 hour. sgRNAs were pelleted by centrifugation at 4□ for 2 minutes at 17,000g, and the supernatant was removed. To wash, 250 µL 70% ice-cold ethanol was added, followed by centrifugation for 2 minutes at 17,000g twice. The sgRNAs were resuspended in 10 µL nuclease-free water.

10X single-cell library preparation and DASH

Sorted cells were counted and checked for viability on a Countess 3 Automated Cell Counter (ThermoFisher) with Trypan blue (0.4%) staining. Sorted cells showing >85% viability were used for 10X single-cell library preparation. To aim for recovery of 5000 cells after sequencing, 8250 cells were loaded onto the 10X Genomics Chromium Controller for subsequent library preparation using Chromium Next GEM Single Cell 3□ Reagent Kits v3.1. Samples were amplified with 10 PCR cycles with cDNA primers (R1+TSO) after clean-up. For DASH, 29 µL CRISPR master mix [NEBuffer 3.1(3µL), 300nM sgRNAs (3µL), 1 µM Cas9 Nuclease (NEB, M0386S)(1µL), nuclease-free water (22µL)] was mixed and pre-incubated at 37□ for 10 min. Then, 1ng of cDNA was added to the master mix and incubated at 37□ overnight. After CRISPR treatment, the cDNA was cleaned up and eluted into 15µL using AMPure beads(Beckman) following the manufacturer's protocol. Then, the cDNA was diluted to 30 µL and amplified with cDNA primers(R1+TSO) again with 10 cycles unless otherwise specified in this study. After PCR amplification, the cDNA was processed as specified in the 10X Genomics protocol for enzymatic fragmentation and indexing. Three biological replicates were used in this study. Libraries were pooled and sequenced using the NextSeq 2000 platform (Illumina). To assess

whether 16S cDNA was removed, we ran individual samples on the Fragment Analyzer (Agilent).

Library quality metrics for published datasets

Parallel-fastq-dump(0.6.7) was performed to retrieve fastq files from NCBI, and the SRA accession numbers were in Table 1. All the preprocessing and alignment in Figure 1 use Drop-seq tools(2.3.0) except Smart-seq2[3]. Smart-seq2 method doesn't have unique molecular modifiers (UMIs), so Smart-seq2 dataset was processed differently [41]. Reads of Smart-seq2 were aligned to a customized genome file containing both the chromosomal-level genome (Smed_chr_ref_v1)[23] and mitochondrial genome of *Schmidtea mediterranea* (NCBI accession number: NC_022448.1) using STAR (2.7.10)[42], and reads mapped to exon regions annotated by SMESG gene model were extracted and segregated into a gene expression matrix. In other single-cell RNA-seq libraries, reads were aligned to the same genome file described above using Drop-seq tools. Cell barcodes and unique molecular identifiers (UMIs) were extracted and tagged to the reads in BAM format and further segregated into a gene expression matrix. To calculate library quality metrics, 'CreateSeuratObject' function of Seurat(4.3.0) in R(4.2.0) was used to import the count matrix from Drop-seq tools or Smart-seq2 pipeline, which calculates total UMI counts and the number of genes expressed associated with each cell barcode[27]. Further, custom R codes were used to extract 16S UMIs and non-16S UMIs (total UMIs - 16S UMIs). The workflow was adapted from TAR-scRNA-seq and compiled in snakemake (7.18.2)[43].

Single-cell analysis for untreated and DASHed libraries

Figure 2B demonstrates the analysis workflow. The untreated and DASHed libraries from three biological replicates were processed by Cell Ranger (6.1.2)[4]. The lists of cell barcodes were retrieved from both libraries and split into "untreated-specific", "shared" and "DASHed-specific" categories (Supplementary figure 3). The percentage of 16S UMIs was calculated for each category. For the rest of the analysis, only "shared" cells were used.

For rarefaction analysis, we downsampled the datasets into 10, 50, and 100 million reads for replicate pairs 2 and 3, and 150 and 200 million reads for replicate pair 1. Downsampled datasets were processed independently in CellRanger.

For clustering analysis, the shared cells were further selected if the cells had (1) more than 200 genes and (2) *piwi-1* expression ≥ 2.5 [ln(UMI-per-10,000 + 1)] to remove cells with low complexity and non-stem cells. If a cell didn't meet the criteria in the untreated dataset, the same cells would be also eliminated in the DASHed dataset. Subsequently, all three replicates of either untreated or DASHed libraries were pooled separately. The libraries were normalized and scaled in Seurat and then integrated by Harmony(0.1.1)[44]. The first fifty Harmony coordinates were used to calculate UMAP embedding. The clustering used the first twenty Harmony coordinates with a resolution of 0.5 for the FindClusters function, resulting in 16 clusters in untreated and 17 clusters in DASHed datasets [27]. The list of markers from Zeng et al. (2018) was used to manually annotate and align the clusters[16]. To detect differentially expressed genes, Wilcoxon Rank Sum adjusted by Bonferroni correction was used for each cluster. Differentially

expressed genes were genes that (1) have an adjusted p-value <0.05, (2) were expressed by at least 25% of cells within the clusters and (3) log2 fold change >0.25 compared to other clusters.

Figure Legends

Figure 1. Planarian 16S rRNA is highly enriched in various scRNA-seq platforms.

A-B, Boxplots show the percentage of 16S UMIs (**A**) and 12S UMIs (**B**) across different published scRNA-seq datasets.

C-D, Boxplots show the number of genes (**C**) and UMI (**D**) per cell.** Note, SmartSeq2 does not implement UMI. The values graphed indicate total read counts for SmartSeq2.The boxes show the 25th, 50th, and 75 quantiles and the whiskers show 1.5 times interquartile range. DropSeq.1-5: "head" datasets from ref.15; DropSeq.6:gfp(RNAi) from ref.17; SmartSeq2 from ref.18; SplitSeq.1-3: 0 days post-amputation (dpa), 1dpa and 2dpa from ref.19; SplitSeq.ACME from ref.20; 10X.1: post-pharyngeal wound fragments, 0h from ref.21; 10X.2 from ref.16.

Figure 2. Workflow and depletion of 16S sequence with DASH.

A, Schematic of DASH protocol. Probes containing poly(T)VN, where V is any base except T, and N is any base, are used for poly(A) enrichment. cDNA is reverse-transcribed in the 10X Chromium Controller and amplified by cDNA primers, Read1 and template switching oligo (TSO). After generation of cDNA, 16S cDNA is depleted by incubation with Cas9 and sgRNAs targeting 16S sequence, followed by post-CRISPR amplification with the same cDNA primers. The re-amplified cDNA is then fragmented and indexed for subsequent sequencing.

B, Schematic of experimental design to benchmark the performance of DASH. Three biological replicates of stem cells (X1 cells) are sorted and processed for cDNA preparation. cDNA is then split into "untreated" and "DASHed" libraries. Sequencing reads are processed by CellRanger and then the cell barcodes recovered from both groups are used for downstream analysis.

C, Venn diagram of cell barcodes from untreated and DASHed libraries. The shared cell barcodes are used to assess library quality.

D, Fragment analysis of untreated cDNA (**A**) and DASHed cDNA (**B**). x-axis shows fragment size in base pairs (bp), and y-axis is relative fluorescence units (RFU). Red peak marks the lower marker (1bp) and blue peak marks the upper marker (6000bp).

E, Volcano plot of differential expression analysis of DASHed versus untreated datasets. x-axis is the average log2 fold change across all cells. y-axis is -log10 of adjusted p-value of Wilcoxon Rank Sum test based on Bonferroni correction. The cutoff for significant genes is adjusted p-value < 0.05 and absolute average log2 fold change > 1.

Figure 3. DASH depletes 16S UMIs and enhances library complexity.

A-C, Boxplots show the percentage of 16S UMI (**A**), numbers of genes (**B**) and non-16S UMIs (**C**) per cell in three replicates before and after DASH treatment. The medians of ratio in DASHed versus untreated are shown on the top. All samples from 3 biological replicates (rep) contain 100 million reads.

D-E, Rarefaction analysis of library complexity comparing the libraries between untreated and DASHed, shown are the medians of genes (**D**) and non-16S UMI per cell (**E**). Each dot represents a downsampled replicate with indicated total reads.

Figure 4. Post-CRISPR amplification increases library complexity.

A-B, Boxplots showing the UMI (**A**) and genes per cell (**B**) among the libraries with different numbers of PCR cycles post-CRISPR. The medians are labeled in the 50th percentile in the box.

C-D, Rarefaction analysis of library complexity comparing the libraries with different numbers of PCR cycles post-CRISPR, shown are the medians of genes (**C**) and non-16S UMIs per cell (**D**). Each dot represents a downsampled replicate with indicated total reads.

Figure 5. DASH treatment boosts library complexity independent to read depth.

A-B, Comparison of genes per cell (**A**) and mRNA per cell (**B**) between untreated and DASHed libraries. Each dot represents a cell shared by untreated and DASHed. Orange lines are the linear regression models, and formula of linear regression and R-squared (R^2) of the models are indicated in top left.

C-D, Boxplots show fold changes of DASHed v.s. untreated in numbers of genes (**C**) and non-16S UMIs (**D**) per cell across ranges.

Figure 6. DASH treatment benefits clustering and differential expression analysis.

A-B, UMAP plots of untreated (**A**) and DASHed (**B**) samples. Each dot represents a single cell. Cells belonging to cluster 15 (C15) in DASHed sample are labeled in red.

C, Bar plot shows numbers of genes that are expressed in at least 25% of cells in the same cluster. Black box indicates the cluster that is unique to DASHed dataset.

D, Bar plot shows numbers of differentially expressed (DE) genes across clusters, tested by Wilcoxon Rank Sum test and adjusted by Bonferroni correction. Black box indicates the cluster that is unique to DASHed dataset.

Supplementary figure 1. 16S reads are enriched at the 3' end of the locus.

A, Genomic tracks showing the raw read depth surrounding 12S and 16S rRNA loci, The bottom track shows 0-10000 to highlight the 12S locus. Arrows show the orientations of gene loci.

B, Gene body coverage plot showing the coverage of 16S rRNA locus (red) and average coverage of all non-16S gene loci (purple) in planarians. The coverage is retrieved from the untreated library, replicate 2. x-axis represents the percentage of the gene body, from 0% (5' end) to 100% (3' end). y-axis represents the relative coverage of reads that map to each part of the gene body. Dashed lines indicate the positions of A stretches in 16S locus.

Supplementary figure 2. Overrepresentation of 16S UMI causes aberrant cell calling.

A, Venn diagrams show the cell barcodes shared or specific to "untreated" or "DASHed" across biological replicates. Table shows the number of cell barcodes in each area and replicate. The percentages show the proportion of the numbers over the total cell numbers within the replicate.

B, Box plots depict the percentage of 16S UMIs from the cells of indicated area. Enlarged box plots from B show cells from "DASHed" libraries.

Supplementary figure 3. 10 PCR cycles post-CRISPR is optimal.

Fragment analysis of cDNA with 0 (**A**), 5 (**B**), 10 (**C**) and 15 (**D**) cycles post-CRISPR.

x-axis shows fragment size in base pairs (bp), and y-axis is relative fluorescence units (RFU). Red peak marks the lower marker (1bp) and blue peak marks the upper marker (6000bp).

Supplementary figure 4. DASH treatment benefits clustering and differential expression analysis.

A-B, UMAP plots of untreated (**A**) and DASHed (**B**) samples. Each dots represent a single cell. Dots are color-coded by clusters.

C, Dot plot of the marker gene expression across annotated clusters. The size of the dots represents the percentages of cells within each cluster that express the marker gene. The color gradient represents the expression level. Red box shows the cluster unique to DASHed dataset.

D, Violin plot shows the percentage of 16S UMIs in each cluster, where C13 has the highest (red box). Each dot represents a single cell.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data is available upon request. scRNA-seq data have been deposited into NCBI Accession number GSE231548. Codes for generating the figures have been deposited on Github (https://github.com/kw572/DASH_figures).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by a National Institutes of Health grant R01GM139933 (to CEA), and Cornell University Mong Fellowship (to K-TW).

Authors' contributions

K-TW and CEA conceived and designed the study and drafted the manuscript; K-TW acquired data and performed computation analyses; CEA supervised the study. All authors read and approved the final manuscript.

Acknowledgments

We thank the Cornell University Biotechnology Resource Center's Flow Cytometry (RRID:SCR_021740), Imaging (RRID:SCR_021741), and Genomics (RRID:SCR_021727) cores, including Peter Schweitzer who generated libraries. We would also like to thank members of the Adler laboratory for input on this project, and David McKellar, Charles Danko, Leslie Babonis, and Bhargav Sanketi for comments on the manuscript.

References

1. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).
2. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
3. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
4. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
5. McKellar, D. W. *et al.* Spatial mapping of the total transcriptome by in situ polyadenylation. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01517-6.
6. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).
7. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, (2020).
8. Salmen, F. *et al.* High-throughput total RNA sequencing in single cells using VASA-seq. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01361-8.
9. Hayashi, T. *et al.* Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.* **9**, 619 (2018).
10. Armour, C. D. *et al.* Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat. Methods* **6**, 647–649 (2009).
11. Gu, W. *et al.* Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to

remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* **17**, 41 (2016).

12. Isakova, A., Neff, N. & Quake, S. R. Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
13. Sheng, K., Cao, W., Niu, Y., Deng, Q. & Zong, C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods* **14**, 267–270 (2017).
14. Kim, I. V. *et al.* Efficient depletion of ribosomal RNA for RNA sequencing in planarians. *BMC Genomics* **20**, 909 (2019).
15. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* **360**, (2018).
16. Zeng, A. *et al.* Prospectively Isolated Tetraspanin+ Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. *Cell* **173**, 1593–1608.e20 (2018).
17. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, (2018).
18. Molinaro, A. M. & Pearson, B. J. In silico lineage tracing through single cell transcriptomics identifies a neural stem cell population in planarians. *Genome Biol.* **17**, 87 (2016).
19. Benham-Pyle, B. W. *et al.* Identification of rare, transient post-mitotic cell states that are induced by injury and required for whole-body regeneration in *Schmidtea mediterranea*. *Nat. Cell Biol.* **23**, 939–952 (2021).
20. García-Castro, H. *et al.* ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics. *Genome Biol.* **22**, 89 (2021).

21. Scimone, M. L., Cloutier, J. K., Maybrun, C. L. & Reddien, P. W. The planarian wound epidermis gene equinox is required for blastema formation in regeneration. *Nat. Commun.* **13**, 2726 (2022).

22. Grohme, M. A. *et al.* The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms. *Nature* **554**, 56–61 (2018).

23. Guo, L. *et al.* Island-specific evolution of a sex-primed autosome in a sexual planarian. *Nature* **606**, 329–334 (2022).

24. Chang, J. H. & Tong, L. Mitochondrial poly(A) polymerase and polyadenylation. *Biochim. Biophys. Acta* **1819**, 992–997 (2012).

25. Slomovic, S., Laufer, D., Geiger, D. & Schuster, G. Polyadenylation and degradation of human mitochondrial RNA: the prokaryotic past leaves its mark. *Mol. Cell. Biol.* **25**, 6427–6435 (2005).

26. Phelps, W. A., Carlson, A. E. & Lee, M. T. Optimized design of antisense oligomers for targeted rRNA depletion. *Nucleic Acids Res.* **49**, e5 (2021).

27. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

28. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).

29. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).

30. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**, 963–967 (2021).

31. Bratic, A. *et al*. Mitochondrial Polyadenylation Is a One-Step Process Required for mRNA Integrity and tRNA Maturation. *PLoS Genet.* **12**, e1006028 (2016).

32. Loi, D. S. C., Yu, L. & Wu, A. R. Effective ribosomal RNA depletion for single-cell total RNA-seq by scDASH. *PeerJ* **9**, e10717 (2021).

33. Pandey, A. C. *et al.* A CRISPR/Cas9-based enhancement of high-throughput single-cell transcriptomics. *bioRxiv* 2022.09.06.506867 (2022) doi:10.1101/2022.09.06.506867.

34. Lun, A. T. L. *et al*. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol*. **20**, 63 (2019).

35. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14**, 565–571 (2017).

36. Arnold, C. P. *et al*. Pathogenic shifts in endogenous microbiota impede tissue regeneration via distinct activation of TAK1/MKK/p38. *Elife* **5**, (2016).

37. Merryman, M. S., Alvarado, A. S. & Jenkin, J. C. Culturing Planarians in the Laboratory. *Methods Mol. Biol.* **1774**, 241–258 (2018).

38. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).

39. Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **3**, e04766 (2014).

40. Rouhana, L. *et al.* RNA interference by feeding in vitro-synthesized double-stranded RNA to planarians: methodology and dynamics. *Dev. Dyn.* **242**, 718–730 (2013).

41. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

42. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21

(2013).

43. Wang, M. F. Z. *et al*. Uncovering transcriptional dark matter via gene annotation independent single-cell RNA sequencing analysis. *Nat. Commun.* **12**, 2158 (2021).

44. Korsunsky, I. *et al*. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

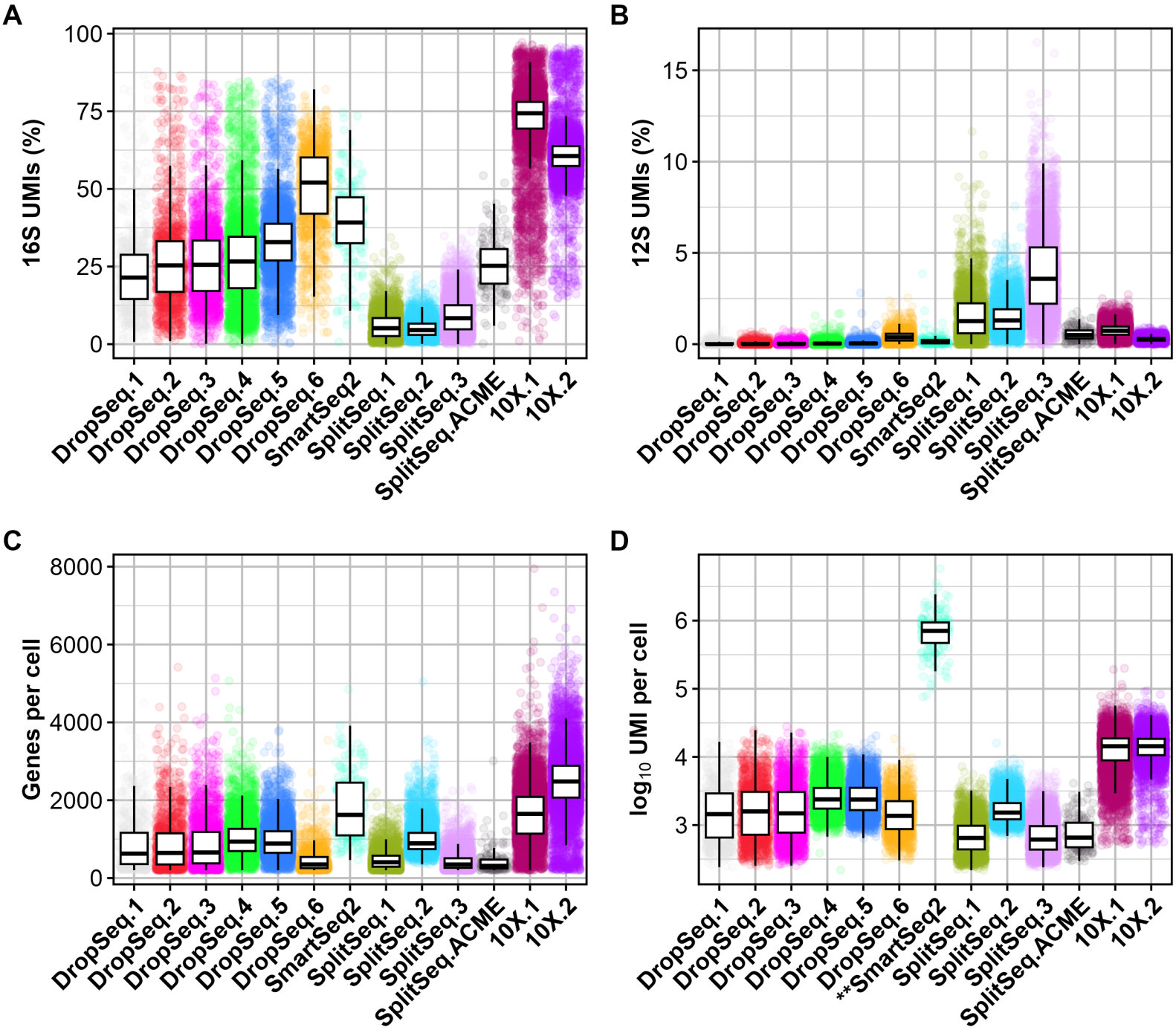


Figure 1. Planarian 16S rRNA is highly enriched in various scRNA-seq platforms.
A-B, Boxplots show the percentage of 16S UMIs (A) and 12S UMIs (B) across different published scRNA-seq datasets.
C-D, Boxplots show the number of genes (C) and UMI (D) per cell.** Note, SmartSeq2 does not implement UMI. The values graphed indicate total read counts for SmartSeq2.
The boxes show the 25th, 50th and 75 quantiles and the whiskers show 1.5 times interquartile range. DropSeq.1-5: “head” datasets from ref.15; DropSeq.6:*gfp*(RNAi) from ref.17; SmartSeq2 from ref.18; SplitSeq.1-3: 0 days post-amputation (dpa), 1dpa and 2dpa from ref.19 ; SplitSeq.ACME from ref.20; 10X.1: post-pharyngeal wound fragments, 0h from ref.21; 10X.2 from ref.16.

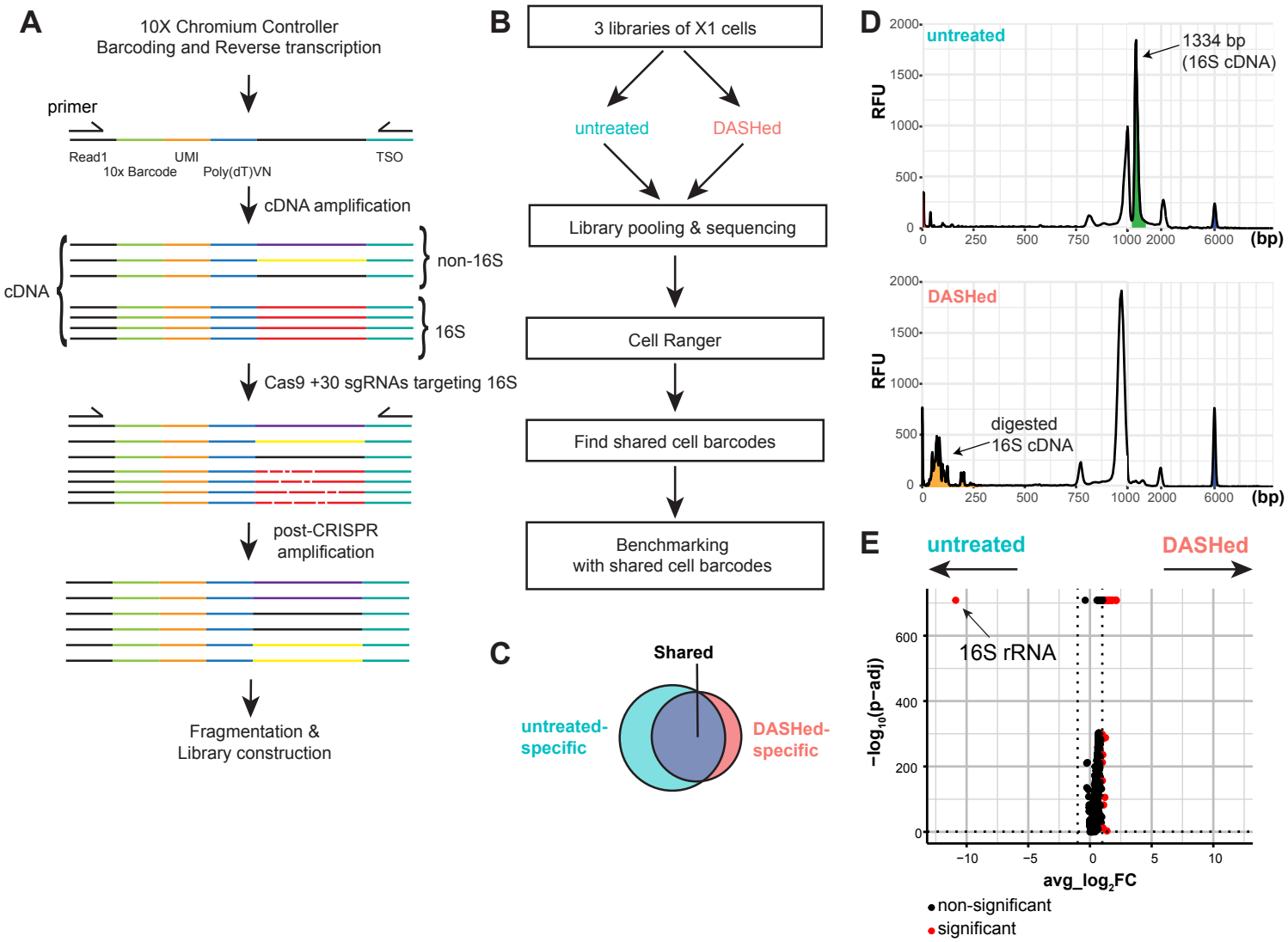


Figure 2. Workflow and depletion of 16S sequence with DASH.
A, Schematic of DASH protocol. Probes containing poly(T)VN, where V is any base except T, and N is any base, are used for poly(A) enrichment. cDNA is reverse-transcribed in the 10X Chromium Controller and amplified by cDNA primers, Read1 and template switching oligo (TSO). After generation of cDNA, 16S cDNA is depleted by incubation with Cas9 and sgRNAs targeting 16S sequence, followed by post-CRISPR amplification with the same cDNA primers. The re-amplified cDNA is then fragmented and indexed for subsequent sequencing.
B, Schematic of experimental design to benchmark the performance of DASH. Three biological replicates of stem cells (X1 cells) are sorted and processed for cDNA preparation. cDNA is then split into "untreated" and "DASHed" libraries. Sequencing reads are processed by CellRanger and then the cell barcodes recovered from both groups are used for downstream analysis.
C, Venn diagram of cell barcodes from untreated and DASHed libraries. The shared cell barcodes are used to assess library quality.
D, Fragment analysis of untreated cDNA (A) and DASHed cDNA (B). x-axis shows fragment size in base pairs (bp), and y-axis is relative fluorescence units (RFU). Red peak marks the lower marker (1bp) and blue peak marks the upper marker (6000bp).
E, Volcano plot of differential expression analysis of DASHed versus untreated datasets. *x*-axis is the average log2 fold change across all cells. *y*-axis is -log10 of adjusted p-value of Wilcoxon Rank Sum test based on Bonferroni correction. The cutoff for significant genes is adjusted p-value < 0.05 and absolute average log2 fold change > 1.

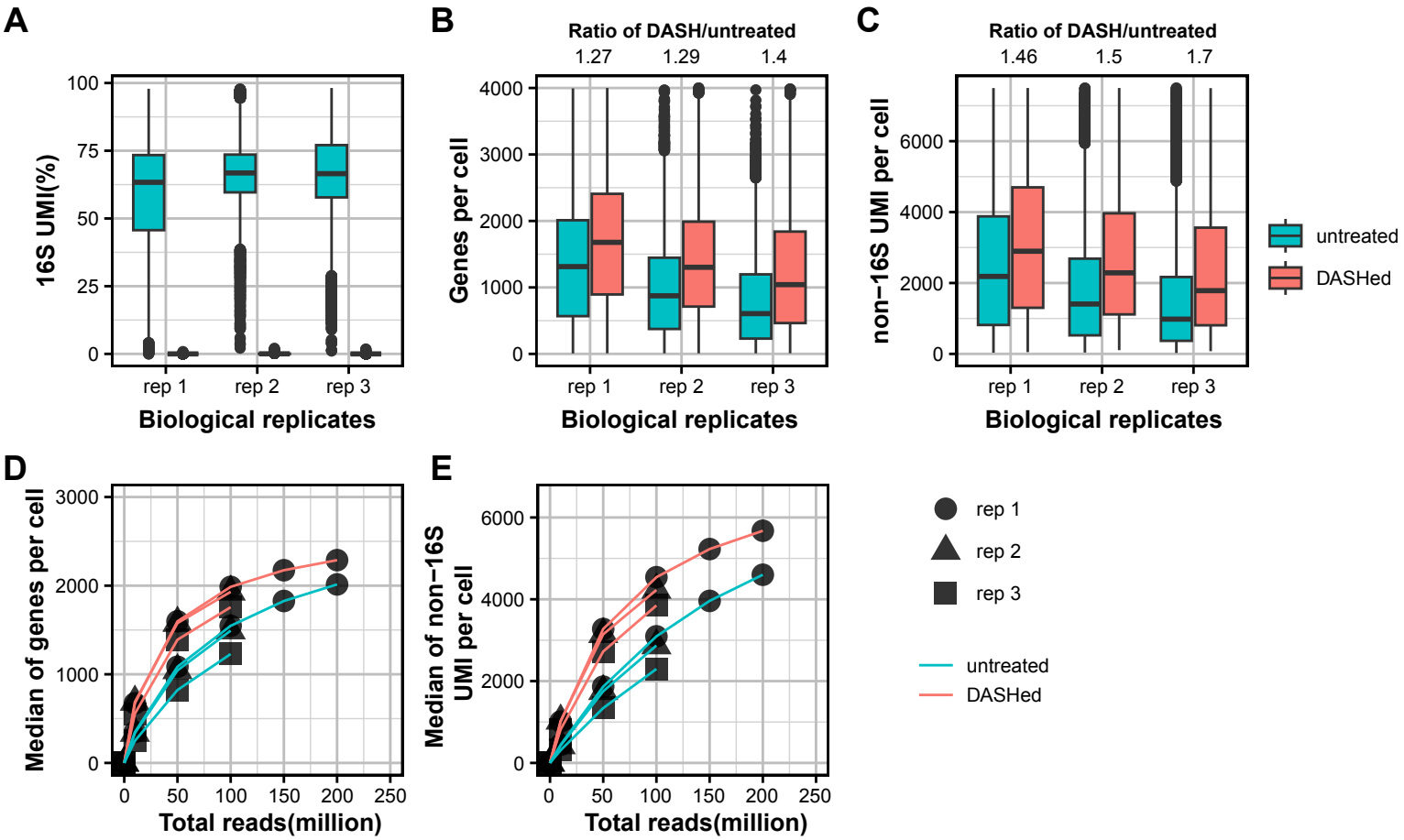


Figure 3. DASH depletes 16S UMIs and enhances library complexity.
A-C, Boxplots show the percentage of 16S UMI (A), numbers of genes (B) and non-16S UMI (C) per cell in three replicates before and after DASH treatment. The medians of ratio in DASHed versus untreated are shown on the top. All samples from 3 biological replicates (rep) contain 100 million reads.
D-E, Rarefaction analysis of library complexity comparing the libaries between untreated and DASHed, shown are the medians of genes (D) and non-16S UMI per cell (E). Each dot represents a downsampled replicate with indicated total reads.

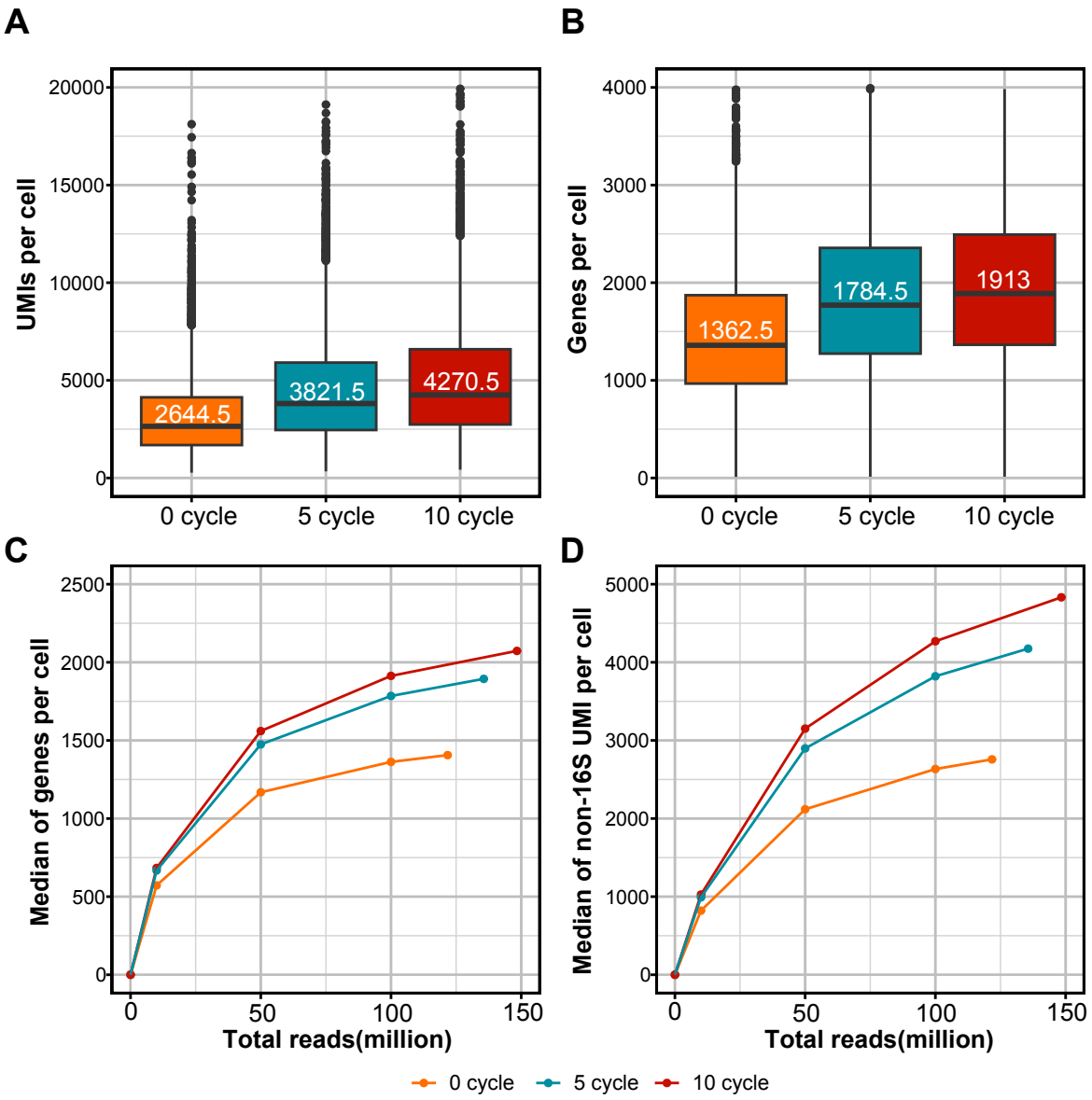


Figure 4. Post-CRISPR amplification increases library complexity.
A-B, Boxplots showing the UMI (A) and genes per cell (B) among the libraries with different numbers of PCR cycles post-CRISPR. The medians are labeled in the 50th percentile in the box.
C-D, Rarefaction analysis of library complexity comparing the libraries with different numbers of PCR cycles post-CRISPR, shown are the medians of genes (C) and non-16S UMI per cell (D). Each dot represents a downsampled replicate with indicated total reads.

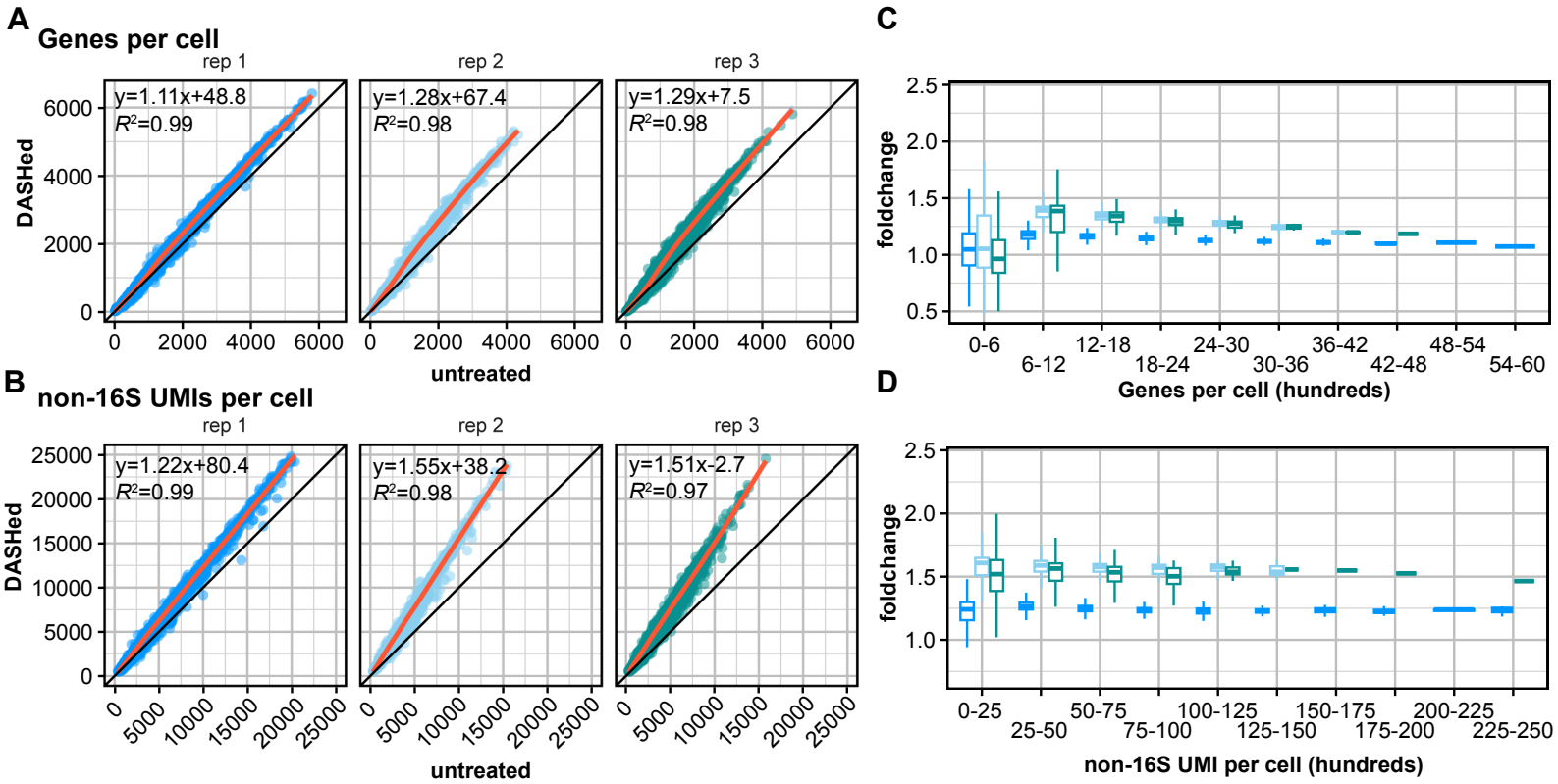


Figure 5. DASH treatment boosts library complexity independent to read depth.
A-B, Comparison of genes per cell (A) and mRNA per cell (B) between untreated and DASHed libraries. Each dot represents a cell shared by untreated and DASHed. Orange lines are the linear regression models, and formula of linear regression and R-squared (R^2) of the models are indicated in top left.
C-D, Boxplots show fold changes of DASHed v.s. untreated in numbers of genes (C) and non-16S (D) per cell across ranges.

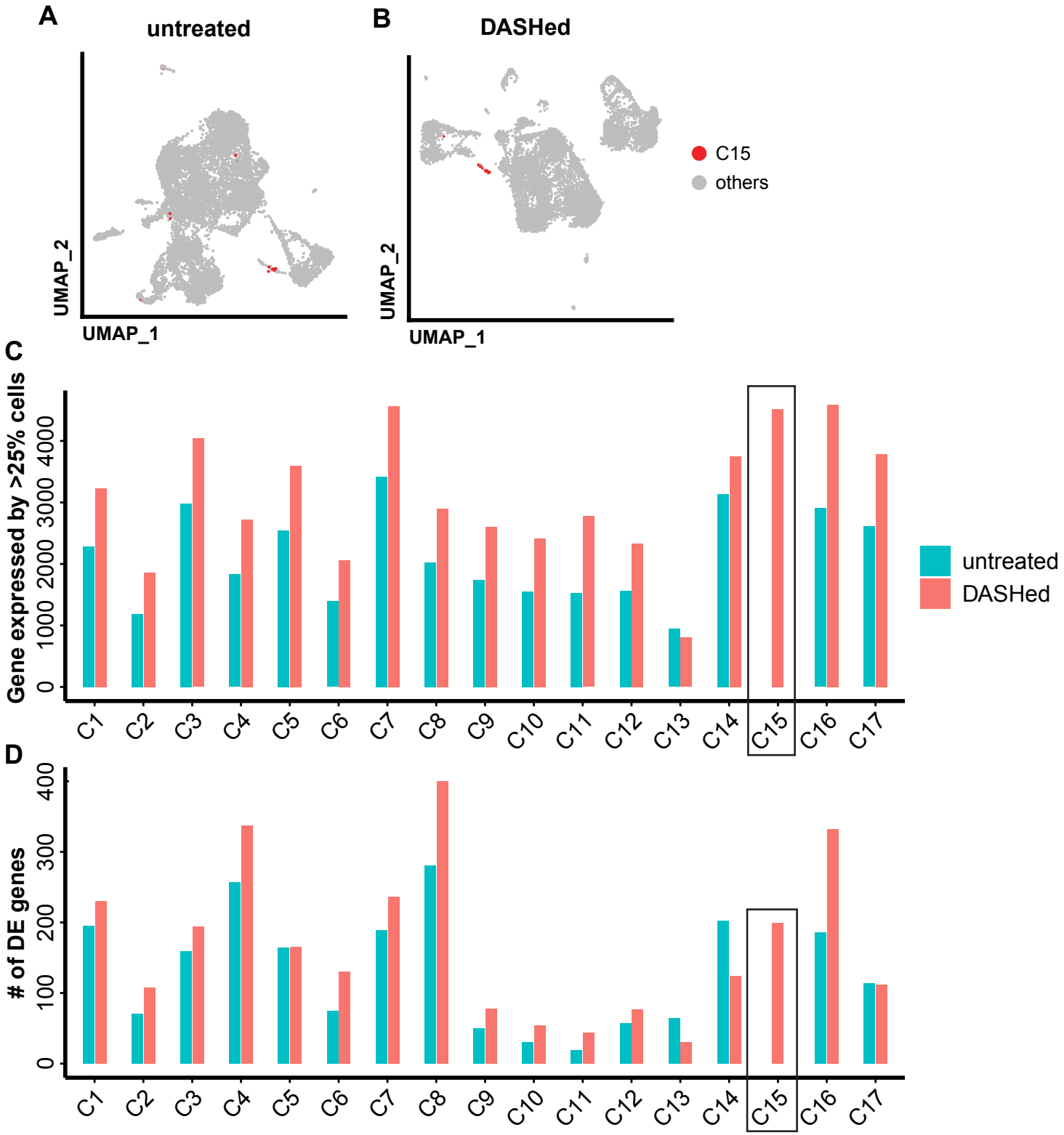
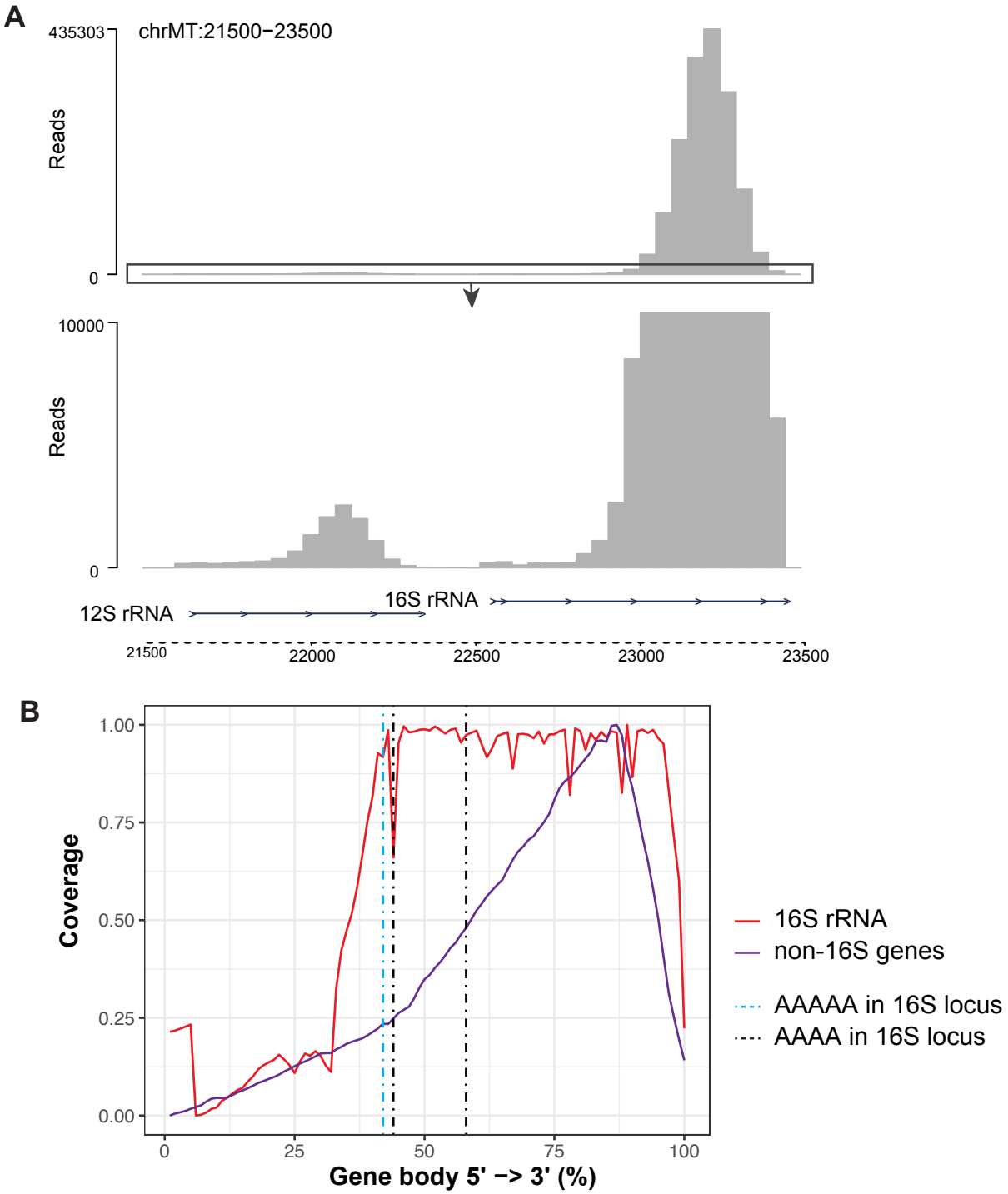
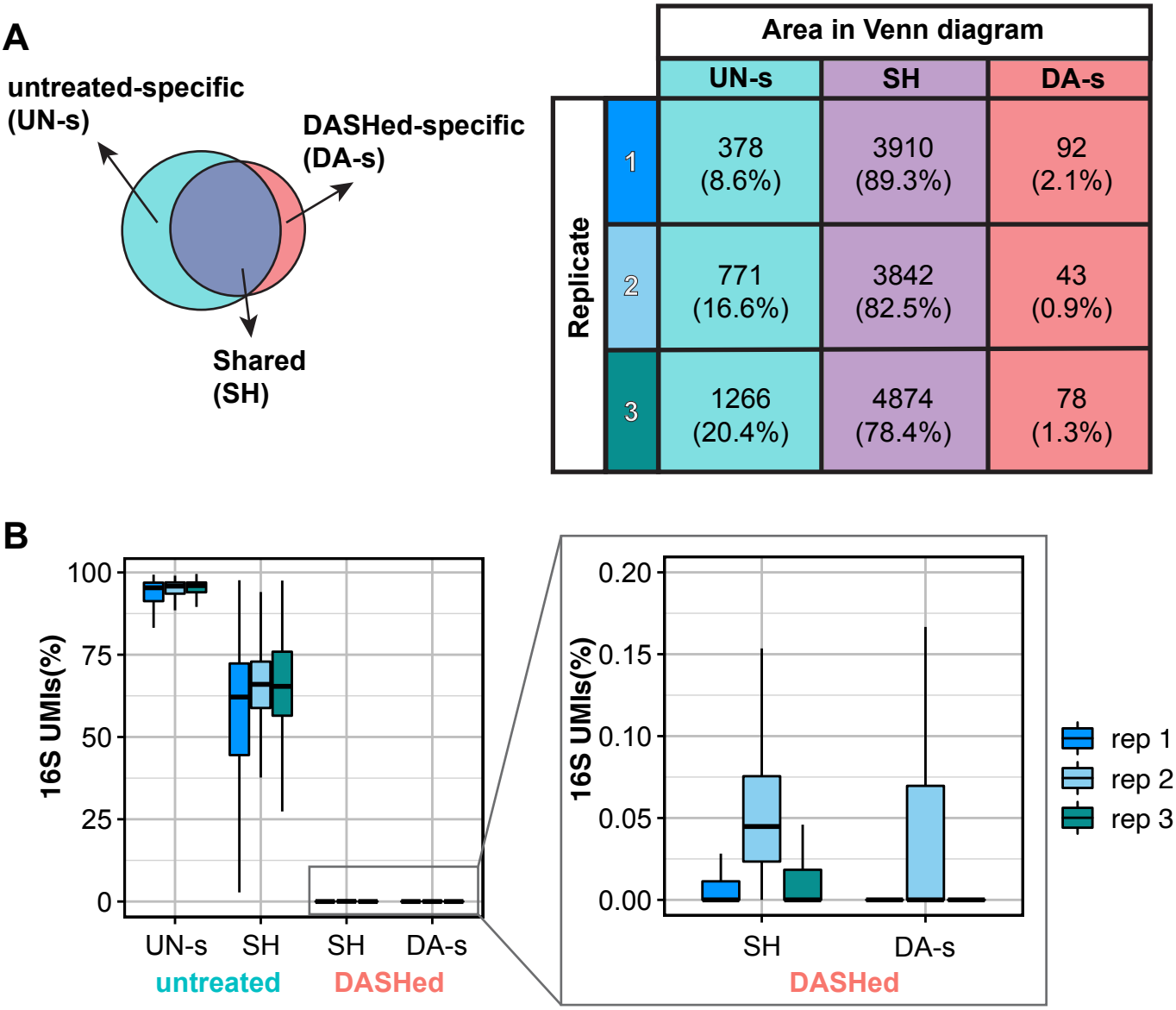


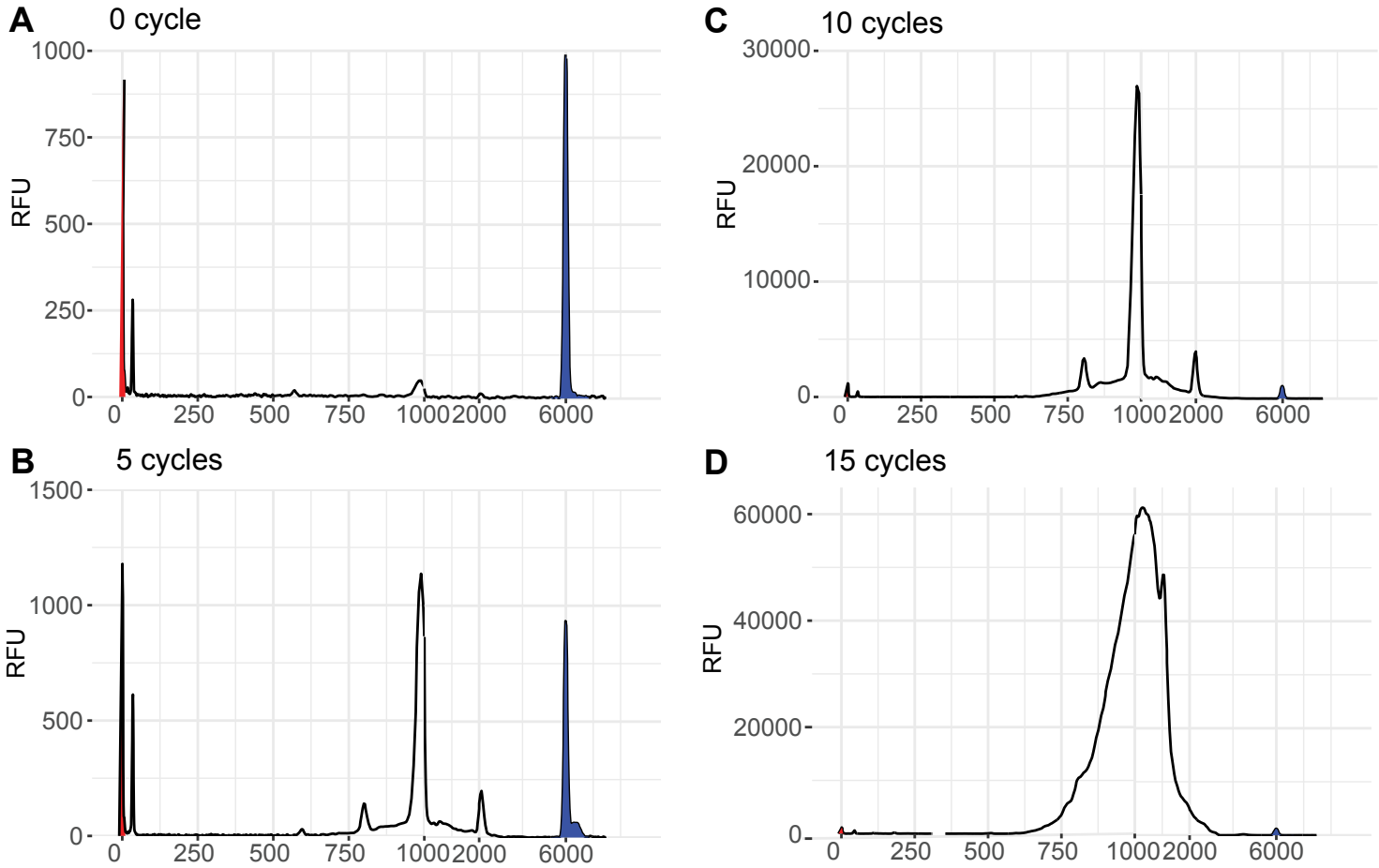
Figure 6. DASH treatment benefits clustering and differential expression analysis.
A-B, UMAP plots of untreated (A) and DASHed (B) samples. Each dot represents a single cell. Cells belonging to cluster 15 (C15) in DASHed sample are labeled in red.
C, Bar plot shows numbers of genes that are expressed at least 25% of cells in the same cluster. Black box indicates the cluster that is unique to DASHed dataset.
D, Bar plot shows numbers of differentially expressed (DE) genes across clusters, tested by Wilcoxon Rank Sum test and adjusted by Bonferroni correction. Black box indicates the cluster that is unique to DASHed dataset.



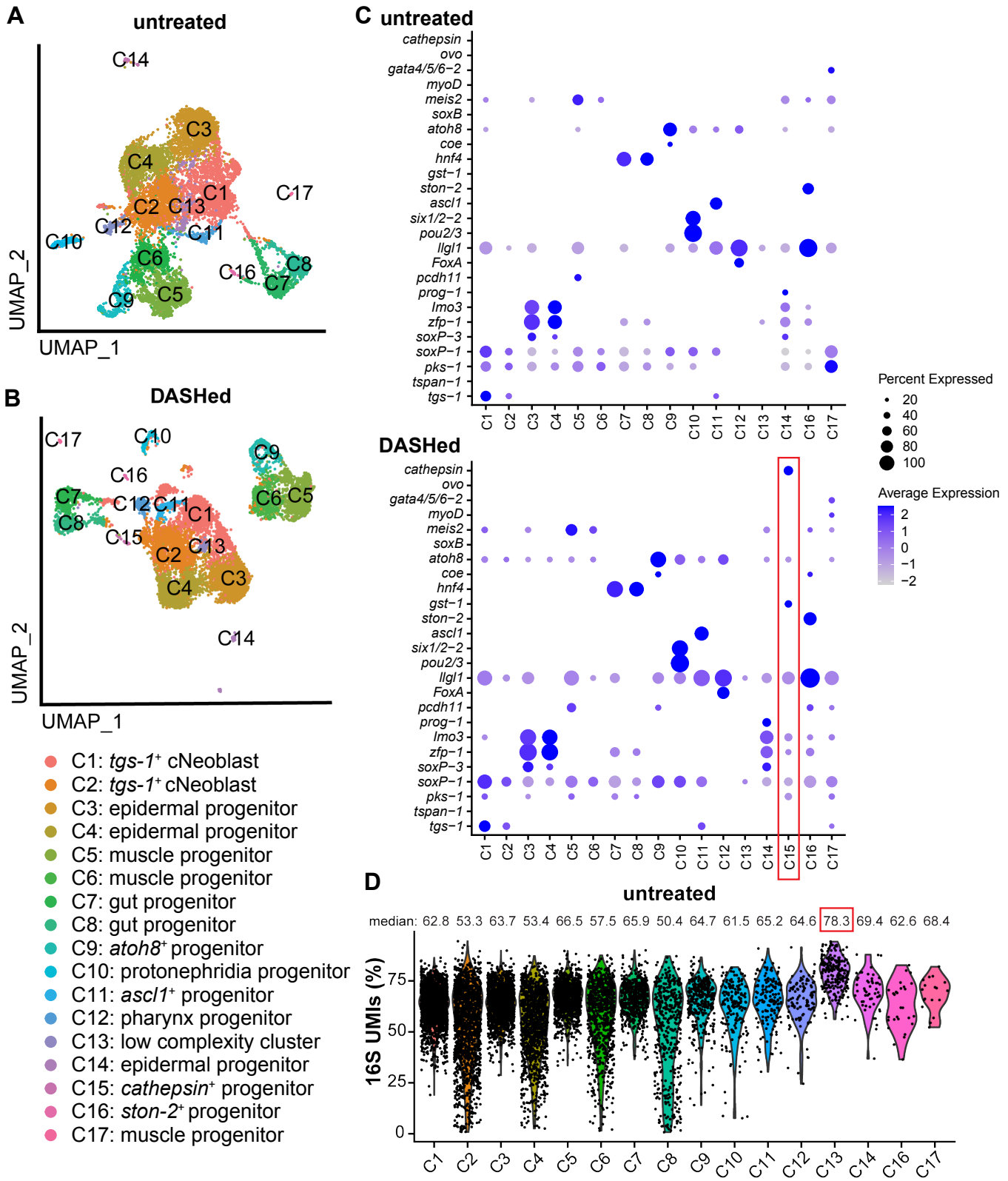
Supplementary figure 1. 16S reads are enriched at the 3' end of the locus.
A, Genomic tracks showing the raw read depth surrounding 12S and 16S rRNA loci, The bottom track shows 0-10000 to highlight the 12S locus. Arrows show the orientations of gene loci.
B, Gene body coverage plot showing the coverage of 16S rRNA locus (red) and average coverage of all non-16S gene loci (purple) in planarians. The coverage is retrieved from the untreated library, replicate 2. x-axis represents the percentage of the gene body, from 0% (5' end) to 100% (3' end). y-axis represents the relative coverage of reads that map to each part of the gene body. Dashed lines indicate the positions of A stretches in 16S locus.



Supplementary figure 2. Overrepresentation of 16S UMI causes aberrant cell calling.
A, Venn diagrams show the cell barcodes shared or specific to “untreated” or “DASHed” across biological replicates. Table shows the number of cell barcodes in each area and replicate. The percentages show the proportion of the numbers over the total cell numbers within the replicate.
B, Box plots depict the percentage of 16S UMI from the cells of indicated area. Enlarged box plots from B show cells from “DASHed” libraries.



Supplementary figure 3. 10 PCR cycles post-CRISPR is optimal.
Fragment analysis of cDNA with 0 (A), 5 (B), 10 (C) and 15 (D) cycles post-CRISPR.
x-axis shows fragment size in base pairs (bp), and y-axis is relative fluorescence units (RFU). Red peak marks the lower marker (1bp) and blue peak marks the upper marker (6000bp).



Supplementary figure 4. DASH treatment benefits clustering and differential expression analysis.

A-B, UMAP plots of untreated (A) and DASHed (B) samples. Each dots represent a single cell. Dots are color-coded by clusters.

C, Dot plot of the marker gene expression across annotated clusters. The size of the dots represents the percentages of cells within each cluster that express the marker gene. The color gradient represents the expression level. Red box shows the cluster unique to DASHed dataset.

D, Violin plot shows the percentage of 16S UMIs in each cluster, where C13 has the highest (red box). Each dot represents a single cell.