TECH NOTE

High-quality stranded RNA-seq libraries from single cells using the SMART-Seq Stranded Kit

Product highlights:

- Simple workflow starts directly from 1–1,000 cells or 10 pg–10 ng total RNA to generate sequencing-ready Illumina libraries in 7 hours
- Unparalleled sensitivity for single-cell, full-length total RNA sequencing with strand-of-origin information
- Reproducible, accurate detection of coding and noncoding transcripts from total RNA and single cells
- Comparable results to our gold-standard single-cell RNA sequencing technology
- Superior performance compared to competitors

Introduction

Our SMARTer NGS portfolio has long included high-performance, cutting-edge solutions for RNA sequencing (RNA-seq). With the growing need for low-input and single-cell NGS library prep solutions, we see that researchers recognize the value in revealing transcriptome profiles from damaged cells as well as noncoding information from single cells and extremely low cell numbers (1–1,000). While we have previously released several industry-leading products that push the limits of sensitivity and reproducibility in RNA-seq from ultra-low inputs as well as single cells (SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing and SMART-Seq HT Kit), they generate transcriptome profiles from mRNA only. Oligo(dT) priming is an efficient way to capture the transcriptome, with minimal uninformative reads (e.g., those from rRNA contamination), but it does not provide a complete view of the transcriptome, as only the polyadenylated fraction can be captured. In addition, for oligo(dT)-primed cDNA synthesis to generate high-quality libraries, one needs to start with high-quality, intact RNA, which excludes the use of this technology with samples damaged or degraded due to the method of isolation or the nature of processing (e.g., FFPE samples). Additionally, these earlier single-cell kits do not preserve stand-of-origin information. All of these factors motivated us to develop the SMART-Seq Stranded Kit, which allows for generation of stranded, sequencing-ready Illumina libraries directly from 1–1,000 cells or an equivalent amount (10 pg–10 ng) of purified total RNA of any quality.

Simple workflow for generation of stranded libraries directly from single cells and total RNA

This kit integrates an innovative technology, already incorporated in our SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian, which enables removal of ribosomal cDNA following cDNA synthesis (Figure 1), as opposed to direct removal of corresponding rRNA molecules prior to reverse transcription. Indeed, since cDNA synthesis in the SMART-Seq Stranded Kit relies on random priming, rRNA is also captured and it is essential to remove the resulting cDNA prior to sequencing. The SMART-Seq Stranded Kit protocol can be completed in just 7 hours, and a convenient pooling option for inputs between one to ten cells facilitates greater ease of use by minimizing the number of samples to be handled.
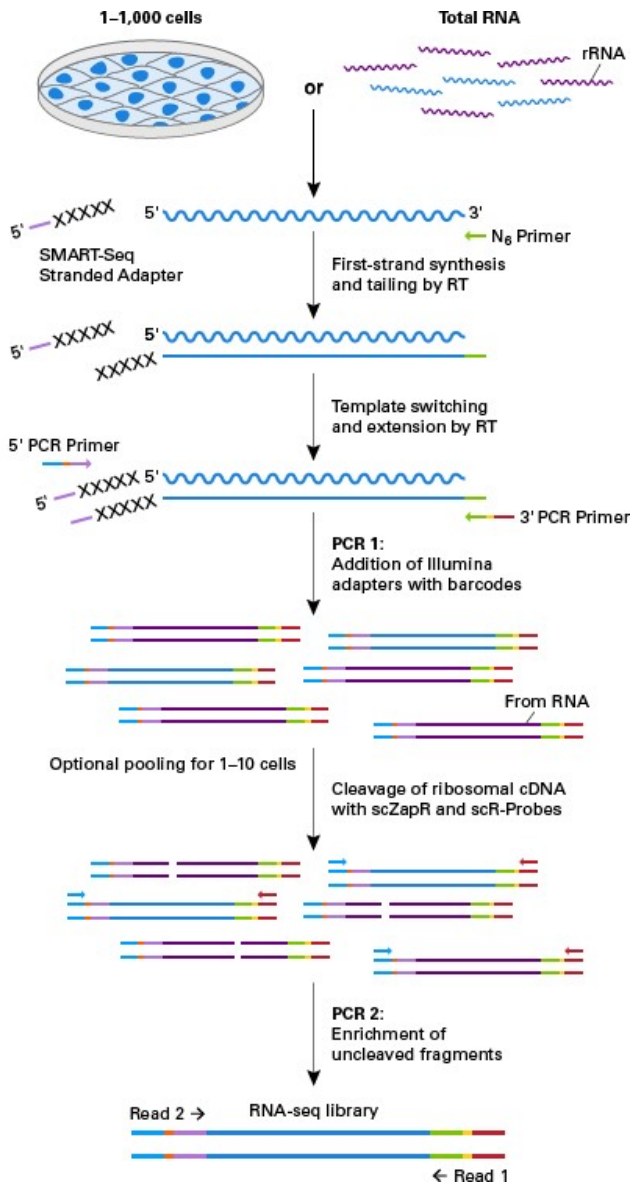
Figure 1. Schematic of technology in the SMART-Seq Stranded Kit. SMART technology is used in a ligation-free protocol to preserve strand-of-origin information. Random priming allows the generation of cDNA from all RNA fragments in the sample, including rRNA. When the SMARTScribe Reverse Transcriptase (RT) reaches the 5' end of the RNA fragment, the enzyme's terminal transferase activity adds a few nontemplated nucleotides to the 3' end of the cDNA (shown as Xs). The carefully designed SMART-Seq Stranded Adapter base-pairs with the nontemplated nucleotide stretch, creating an extended template to enable the RT to continue replicating to the end of the oligonucleotide. The resulting cDNA contains sequences derived from the SMART scN6 Primer and the SMART-Seq Stranded Adapter. In the next step, a first round of PCR amplification (PCR 1) adds full-length Illumina adapters, including barcodes. The 5' PCR Primer binds to the SMART-Seq Stranded Adapter sequence (light purple), while the 3' PCR Primer binds to sequence associated with the SMART scN6 sequence (green). The ribosomal cDNA (originating from rRNA) is then cleaved by scZapR in the presence of the mammalian-specific scR-Probes. This process leaves the library fragments originating from non-rRNA molecules untouched, and available for enrichment in a second round of PCR amplification (PCR2) using primers universal to all libraries. An optional pooling of up to 12 samples after PCR 1 allows for greater ease of use by minimizing the number of samples to be processed downstream.

Results

High-quality, stranded NGS libraries from 10 pg–10 ng total RNA

In order to test overall kit performance, we started with inputs ranging between 10 pg–10 ng of human brain total RNA. Sequencing alignment metrics for the resulting libraries were consistent across inputs, including exonic, intronic, and intergenic reads (Table I). Reproducibility between replicates was high at every input level, including the single-cell equivalent of 10 pg total RNA, as demonstrated by the high Pearson correlations between technical replicates. The data show that even within this single-cell input range, over 97% of the reads match the correct strand, as determined per biological annotation.

Sequencing alignment metrics for 10 pg–10 ng total RNA					
RNA source	Human brain total RNA				
Input amount (ng)	1 0	1	0.25	0.05	0.01
Number of reads (paired-end)	2,500,000	2,500,000	2,500,000	2,500,000	1,000,000
Number of transcripts >1 FPKM	15,128	15,097	15,066	14,394	13,151
Number of transcripts >0.1 FPKM	23,864	23,631	23,274	21,335	16,700
Pearson/Spearman correlations	0.99/0.87	0.99/0.85	0.99/0.82	0.97/0.68	0.92/0.46
Correct strand per biological annotation (%)	97.7	97.8	97.6	97.5	97.1
Proportion of reads (%):					
Exonic	37.1	36.5	41.5	39.7	34.1
Intronic	36.1	35.6	36.7	35.4	30.6
Intergenic	8.6	8.5	8.8	8.7	7.4
rRNA	9.7	9.6	3.6	4.1	6.7
Mitochondrial	5.2	6.3	6.4	6.4	5.9
Overall mapping (%)	96.8	96.4	97.0	94.4	84.7
Duplicate rate (%)	13.3	20.2	35.2	59.0	62.4

Table I. Consistent sequencing metrics across RNA input amounts. Human brain total RNA (10 pg–10 ng) was used to generate RNA-seq libraries with the SMART-Seq Stranded Kit. Data shown are the average of three technical replicates and exhibit exceptionally high Pearson and Spearman correlations between replicates, even with as little as 10 pg of input material. Sequences were analyzed as described in the Methods.

Superior performance for ultra-low-input amounts of total RNA

When we compared the SMART-Seq Stranded Kit (Stranded) against the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Pico v2), we found similar performance across most of the Pico v2 input range (250 pg–10 ng total RNA; Figure 2A). However, below 1 ng, the SMART-Seq Stranded kit identifies more unique reads (fewer duplicates) than the Pico v2 kit, which correlates to a higher number of transcripts identified. This is most obvious for inputs of 250 pg and 50 pg, for which up to 920 additional transcripts with an FPKM >1 were detected using the Stranded kit. In addition, a considerably higher number of low-abundance transcripts were identified with the SMART-Seq Stranded kit. For those with an FPKM >0.1, 572 more transcripts were found with the 500-pg input, while as many as 2,289 more transcripts were found with the 50-pg input.

Reproducibility with the SMART-Seq Stranded Kit was also shown to be higher for inputs below 500 pg, as evidenced by the tighter correlations for 50-pg inputs (Figure 2B; Pearson of 0.97 for the SMART-Seq Stranded kit, compared to 0.92 for the Pico v2 kit). Additionally, we can see in Figure 2B that for the SMART-Seq Stranded kit, the transcripts identified in only one of the replicates (dropouts) are restricted to expression levels close to or

below 10 FPKM, while in the Pico v2 kit, a higher proportion of the dropouts are >10 FPKM. Taken together, these data indicate that the SMART-Seq Stranded Kit outperforms the Pico v2 kit when starting with less than 500 pg of total RNA.
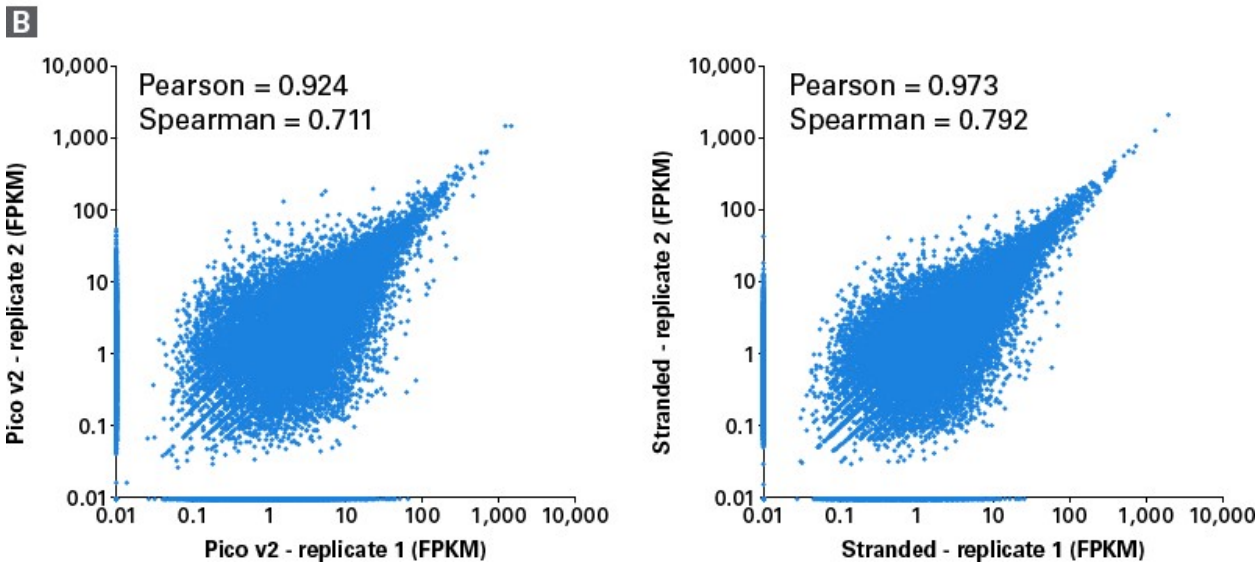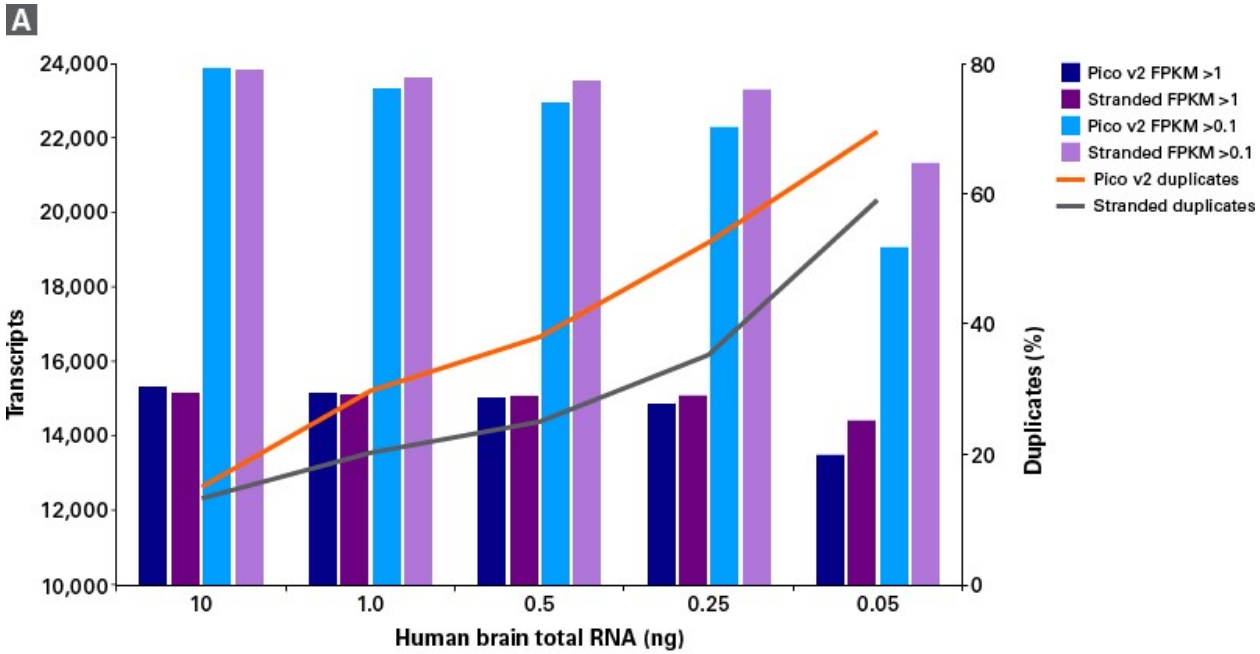
Figure 2. The SMART-Seq Stranded Kit outperforms the Pico v2 kit for ultra-low inputs. Panel A. Human brain total RNA (50 pg–10 ng) was used to generate RNA-seq libraries in triplicate with the SMART-Seq Stranded Kit (Stranded) and the Pico v2 kit. For both kits, sequencing data were down-sampled to 2.5 million paired-end reads prior to analysis. Panel B. Comparison of transcript expression level from libraries generated with 50 pg of total RNA (as shown in Panel A). FPKM values are shown on a $\log_{10}$ scale. Transcripts represented in only one sample (dropouts) can be seen along the X- and Y-axes of the scatter plots.

High reproducibility with inputs ranging from 1 to 1,000 cells

The SMART-Seq Stranded Kit was developed specifically to directly accommodate cells as input, as opposed to only purified total RNA. Kit performance with cells was verified by generating libraries from 1–1,000 cells (Figure 3). For comparison, total RNA samples were purified from aliquots of 1,000 cells and processed in parallel with the cell inputs. Sequencing alignment metrics for the resulting libraries were consistent across all inputs, including reads mapping to exons, introns, intergenic regions, mitochondrial sequences, and rRNA (Figure 3A). Importantly, proportions of reads mapping to introns and intergenic regions were similar for cells and purified RNA, indicating that gDNA contamination is not a concern for library preparation directly from intact cells. In contrast, we observed that compromised (dead) cells exhibit very low exonic mapping and very high intergenic mapping (data not shown). 7–10% of reads mapped to lncRNA, regardless of the number of cells used, and a consistent number of lncRNA transcripts were detected with inputs ranging from 5–1,000 cells.

Further analysis of the sequencing data indicated very high reproducibility across all inputs. The hierarchical clustering heat map in Figure 3B shows that most single cells tend to cluster together, yet display very high correlation with higher inputs. Indeed, the Pearson correlations between any given sample range from 0.85 to 0.99, regardless of input amount. The libraries generated from purified total RNA display extremely high correlations with libraries generated directly from various cell inputs, particularly those generated from 5–100 cells. The library yields from 500 and 1,000 cells were slightly lower than anticipated (data not shown) and, as such, the correlations to the libraries made from purified total RNA or 5–100 cells are not as high. This suggests some inefficiency in the reverse transcription step for the higher inputs, possibly due to contaminants associated with cells and culture media. The cells for this experiment were isolated by FACS, which is not ideal for inputs of 500–1,000 cells, considering the small volume of sorting solution. Independent experiments for libraries generated with manually counted and aliquoted cells led to higher reproducibility relative to purified total RNA (data not shown). Taken together, these data show that the SMART-Seq Stranded Kit exhibits consistent performance from 1–1,000 cells.

A

| Sequencing alignment metrics for A375 total RNA and cells							
Input	Total RNA	1,000 cells	500 cells	100 cells	10 cells	5 cells	1 cell
Number of reads (pairs)	6,000,000	6,000,000	6,000,000	6,000,000	6,000,000	6,000,000	5,873,974
Number of transcripts >1 FPKM	13,260	13,294	13,583	13,520	12,726	12,602	11,540
Number of transcripts >0.1 FPKM	21,334	21,113	21,365	21,145	20,550	18,888	15,815
Proportion of reads (%):							
Exonic	34.7	36.4	39.2	42.7	36.7	36.2	37.3
Intronic	29.6	29.3	27.7	28.3	34.0	30.4	21.1
Intergenic	14.2	13.4	12.2	12.9	16.7	16.8	10.1
rRNA	7.0	11.4	11.5	6.3	3.6	4.9	7.1
Mitochondrial	4.1	3.5	3.7	4.9	3.8	4.4	4.6
Overall mapping (%)	89.6	93.9	94.3	95.1	94.9	92.7	80.2
Duplicate rate (%)	37.3	45.2	40.3	46.1	52.5	72.2	78.5
lncRNA mapping:							
Number of mapped reads (%)	7.2	10.4	10.8	9.4	8.7	8.6	7.3
lncRNA transcripts detected	5,395	4,687	4,565	5,439	5,440	4,983	2,802
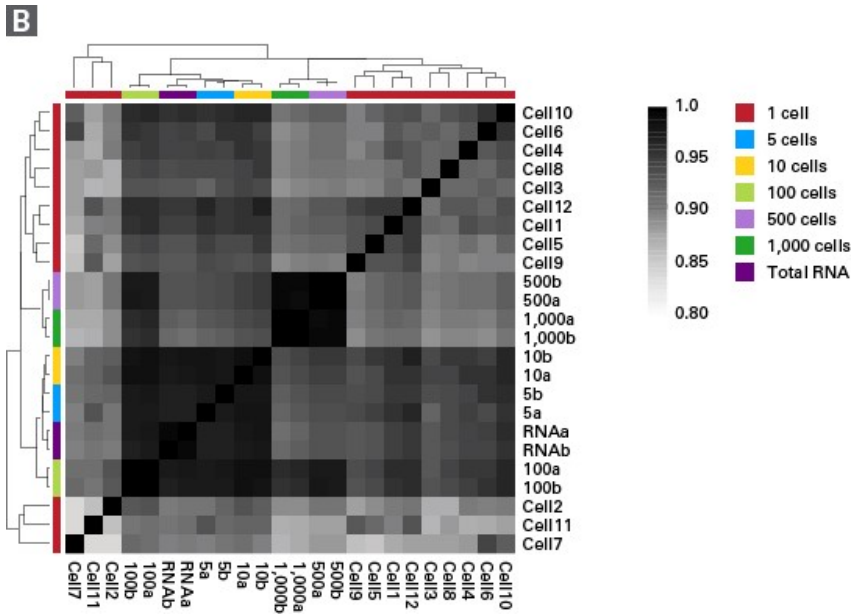
Figure 3: High reproducibility across cell input amounts. A375 cells isolated by FACS were used to generate RNA-seq libraries with the SMART-Seq Stranded Kit. Input varied from 1 cell to 1,000 cells, with two replicates per input of 5–1,000 cells and 12 replicates for the single cells. For comparison, two aliquots of 1,000 cells were used for total RNA purification and then used for library preparation. Panel A. Consistent sequencing metrics across 1–1,000 cells. Panel B. Hierarchical clustering heat map displaying Euclidean distance between all the samples shown in Panel A, and reporting Pearson correlations ranging from 0.85 to 0.99. Single cells are labelled Cell1–Cell12; replicates for other inputs are labelled a–b.

Similar sensitivity and reproducibility between the SMART-Seq v4 and SMART-Seq Stranded kits

The new SMART-Seq Stranded Kit was compared side by side with our industry-standard single-cell RNA-seq kit, the SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (SSv4). Libraries were generated with both kits from K562 single cells isolated by FACS and then sequenced. The total number of transcripts identified by each kit (FPKM >1) was very similar, with 19,331 transcripts identified with the SSv4 kit and 19,106 transcripts with the SMART-Seq Stranded Kit (Figure 4A). Importantly, most transcripts were identified by both kits, with an impressive overlap of 84%. The number of transcripts identified in individual cells was also similar between the two kits, although these numbers were more consistent across cells processed with the SSv4 kit (Figure 4B). Reproducibility across all cells from each kit is similar, although slightly higher and more consistent for the SMART-Seq Stranded Kit, as demonstrated by the Pearson correlations in Figure 4C. The wider range observed for the SSv4 kit could be due to the higher number of cells analyzed. Overall, these data show that the SMART-Seq Stranded Kit can achieve comparable sensitivity and reproducibility to the SSv4 kit.
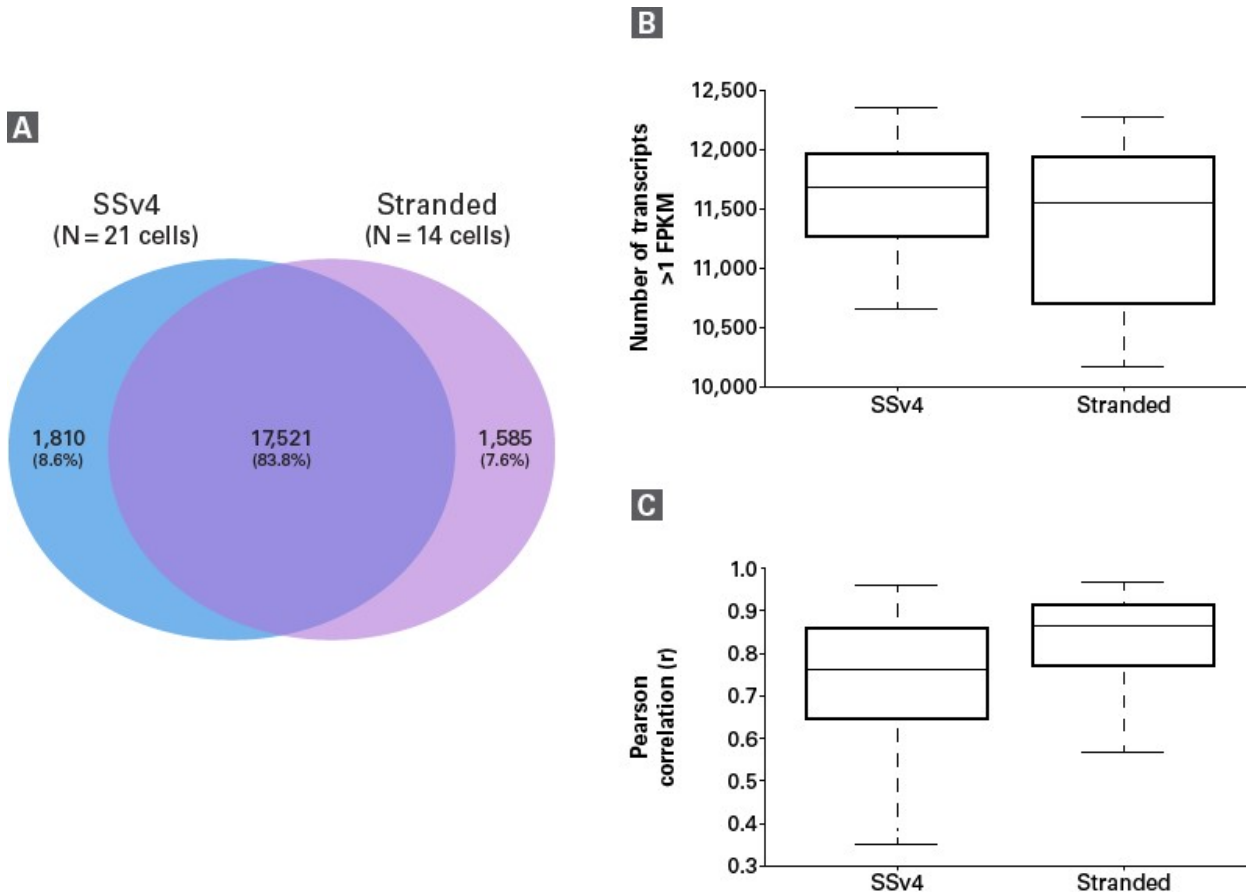
Figure 4. Comparison between SMART-Seq v4 and SMART-Seq Stranded kits. Single cells (K562) isolated by FACS were used to generate RNA-seq libraries with the SMART-Seq Stranded Kit (Stranded) and a SMART-Seq v4 Kit (SSv4; cDNA from this kit was further processed with a Nextera® XT DNA Library Preparation Kit). Panel A. The overlap in the total number of transcripts identified (FPKM >1) by each kit was analyzed and shown to be 83.8%. Panel B. The number of transcripts identified (FPKM >1) in individual cells was similar between the two kits, with a tighter range across cells processed with the SSv4 kit. Panel C. The reproducibility (Pearson correlation) of transcript expression levels across all cells from each kit was similar, although slightly higher and more consistent across cells processed with the Stranded kit.

Superior performance of the SMART-Seq Stranded Kit over other single-cell RNA-seq kits

The SMART-Seq Stranded Kit is designed to capture all RNAs through random priming, while our gold-standard SSv4 kit only captures the polyadenylated fraction through oligo(dT) priming. Among other commercially available kits for single-cell RNA-Seq (Table II), NuGEN's Ovation SoLo RNA-Seq System is promoted for the capture of total RNA transcripts, but features over 30 different components and a strenuous 15-hour protocol (more than double that of the SMART-Seq Stranded Kit). In addition, the requirement for a custom read1 primer makes the sequencing logistics complicated. QIAGEN's QIAseq FX Single Cell RNA Library Kit (QIAseq FX) is also promoted for the capture of total RNA transcripts (if using the optional combination of random and oligo(dT) priming), and features a short and simple workflow similar to that of the SMART-Seq Stranded Kit. However, libraries generated with the QIAseq FX kit do not retain strand-of-origin information, which is a major drawback for the study of lncRNA or more precise mapping in general.

Capabilities of commercially available single-cell RNA-seq kits						
Kit	Strand-specificity	Generates sequencing-ready libraries	Captures polyA or total transcripts	Total time	Accommodates degraded samples	Pearson correlation at 10 pg, 1 ng
SMART-Seq Stranded Kit	Yes	Yes	Both	7 hr	Yes	0.90, 0.99
QIAseq FX Single Cell RNA Library Kit	No	Yes	Both	5.5 hr	No	0.57, NA
Ovation SoLo RNA-Seq System	Yes	Yes	Both	15 hr	Yes	0.8, 0.9*
SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing	No	No	PolyA	4 hr + Nextera processing	No	0.97, 0.99

*Data obtained from NuGEN's website (unknown RNA origin).

Table II. Feature comparison of commercially available single-cell RNA-seq kits. All Pearson correlation data were generated by TBUSA using human brain total RNA, unless otherwise noted.

Even though the QIAseq FX kit does not retain strand-of-origin information, the vendor claims to generate libraries from single cells in a relatively short amount of time, so we compared performance for the SMART-Seq Stranded Kit and QIAseq FX kit in a side-by-side experiment. Libraries were generated using human brain total RNA and Jurkat cells. For total RNA samples, 50 pg was chosen in order to match the manufacturer's lowest recommended input amount for the QIAseq FX kit. Jurkat cells were isolated by FACS and dispensed as 50 cells per tube. To minimize variability, the same batch of sorted cells were used for testing each kit. In all cases, the sequencing metrics were fairly consistent between technical replicates (Figure 5A). However, 40–47% of the reads generated from the QIAseq FX kit were essentially wasted rRNA reads, while the SMART-Seq Stranded Kit generated only 3–4% rRNA reads—one tenth that of the QIAseq FX kit. Overall mapping was 94–97% for the SMART-Seq Stranded Kit, while only 82–87% of the reads generated with the QIAseq FX kit could be mapped. This may be because the QIAseq FX kit workflow involves a ligation of all cDNA followed by multiple displacement amplification, which generates hybrid cDNA junctions that do not exist in nature and cannot be mapped properly. Nevertheless, the high number of uninformative reads can in theory be compensated by higher sequencing depth. This does not appear to be the case with the QIAseq FX kit, as even with 8M reads used for analyzing the libraries from Jurkat cells, the number of transcripts detected was considerably below the number of transcripts detected in the SMART-Seq Stranded Kit. The substantially higher sensitivity of the SMART-Seq Stranded Kit was observed regardless of the sequencing depth (Figure 5B).

Another remarkable difference between the two kits is the reproducibility between replicates. Using 50 pg of total RNA, the Pearson correlations between replicates prepared with the QIAseq FX kit was only 0.66 compared to 0.97 for the SMART-Seq Stranded Kit (Figure 5A). When using a true single-cell equivalent of 10 pg, the Pearson correlation was only 0.57 for the QIAseq FX kit (Table II and data not shown), while the SMART-Seq Stranded Kit shows a correlation of 0.92 (Table I). When testing higher inputs (50 Jurkat cells), both kits presented a high Pearson correlation of 0.99 between biological replicates, but the Spearman correlation was considerably lower for the QIAseq FX kit (Figure 5A). Since correlations are only calculated for transcripts identified in the two replicates, we asked how many transcripts this represented. The SMART-Seq Stranded Kit identified a total of 13,354 transcripts with FPKM >1, of which 80% were common to the two replicates (Figure 5C). The QIAseq FX kit identified almost 1,000 fewer transcripts, and only 66% of overlap between the two replicates. Taken together, these data show that the SMART-Seq Stranded Kit exhibits higher sensitivity and more consistent performance than the QIAseq FX kit.

A

| Sequencing alignment metrics for SMART-Seq Stranded Kit compared to QIAseq FX kit | | | | | | | | |
|---|---|---|---|---|---|---|---|
| RNA source | Human brain total RNA - 50 pg | | | | Jurkat cells - 50 cells | | | |
| Protocol | SMART-Seq Stranded | | QIAseq FX | | SMART-Seq Stranded | | QIAseq FX | |
| Number of transcripts >1 FPKM | 14,398 | 14,364 | 13,300 | 12,594 | 11,973 | 12,067 | 10,949 | 10,232 |
| Number of transcripts >0.1 FPKM | 21,314 | 21,346 | 20,610 | 19,744 | 19,053 | 19,242 | 16,198 | 15,373 |
| Pearson/Spearman correlations | 0.97/0.67 | | 0.66/0.63 | | 0.99/0.90 | | 0.99/0.74 | |
| Proportion of reads (%) | | | | | | | | |
| *Exonic* | 39.9 | 39.5 | 22.9 | 20.7 | 36.8 | 36.2 | 17.8 | 18.2 |
| *Intronic* | 35.8 | 35.4 | 9.8 | 10.5 | 45.3 | 45.0 | 8.4 | 8.2 |
| *Intergenic* | 8.8 | 8.7 | 4.5 | 7.1 | 9.9 | 10.6 | 9.8 | 10.1 |
| *rRNA* | 3.8 | 4.4 | 40.8 | 42.6 | 2.5 | 2.9 | 47.8 | 47.2 |
| *Mitochondrial* | 6.3 | 6.4 | 4.4 | 5.5 | 2.8 | 2.4 | 3.2 | 3.2 |
| Overall mapping (%) | 94.5 | 94.4 | 82.4 | 86.4 | 97.3 | 97.0 | 87.0 | 86.9 |

B
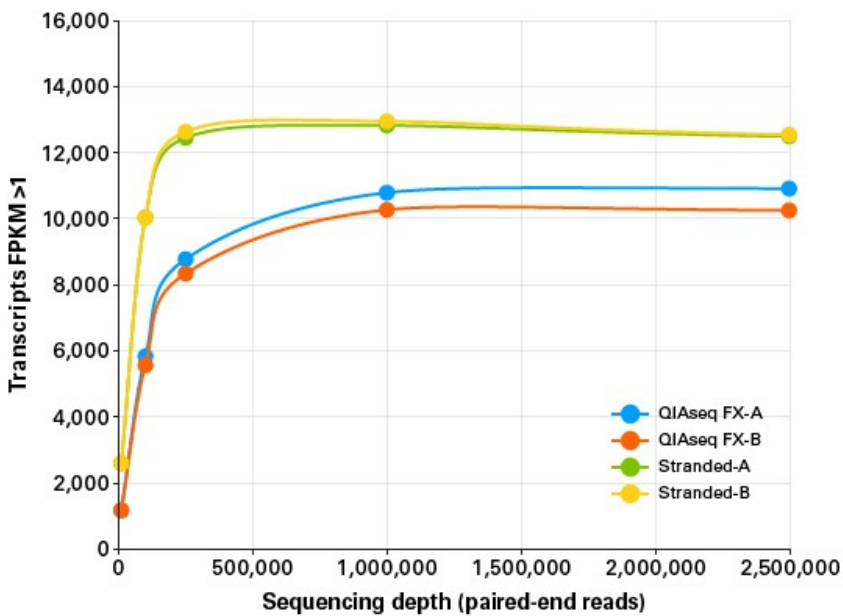
C
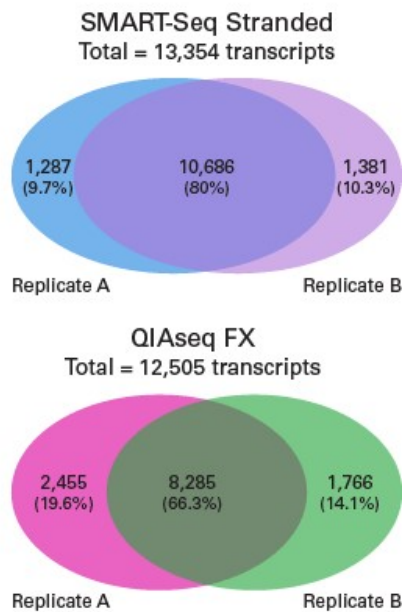
Figure 5. Comparison between the SMART-Seq Stranded Kit and the QIAseq FX Single Cell RNA Library Kit. Panel A. Human brain total RNA (50 pg) and cells isolated by FACS (Jurkat cell line, 50 cells) were used to generate RNA-seq libraries in duplicate with the SMART-Seq Stranded Kit (Stranded) and the QIAseq FX Single Cell RNA Library Kit (QIAseq FX). Following sequencing, human brain total RNA data were analyzed using 2.5M paired-end reads for the Stranded kit, and 5M paired-end reads for the QIAseq FX kit. All data generated from Jurkat cells were normalized to 8M paired-end reads. The Pearson and Spearman correlations were determined between the replicates shown. Panel B. A downsampling experiment with the data generated from Jurkat cells clearly shows that the higher sensitivity observed for the Stranded kit is maintained with lower sequencing depth. Panel C. Assessment of reproducibility between the two kits for the libraries generated from Jurkat cells. Using an expression level cutoff of FPKM >1, the total number of transcripts identified and the overlap between the replicates is

much greater for the Stranded kit.

Conclusions

The new SMART-Seq Stranded Kit generates sequencing-ready, stranded RNA-Seq libraries from 1–1,000 cells or 10 pg–10 ng of total RNA. It provides an excellent alternative to SMART-Seq v4 kits for researchers interested in acquiring single-cell, whole-transcriptome data with strand-of-origin information. It can also be a valuable tool for analyzing cell input samples with a partially degraded transcriptome due to the harsh conditions required for isolation of single cells (e.g., from tumor tissues). In addition, the SMART-Seq Stranded Kit is a better option than the Pico v2 kit for purified total RNA inputs below 500 pg.

Methods

Cell sorting and library preparation

Sequencing libraries were generated using the SMART-Seq Stranded Kit (Cat. # 634442, 634443, 634444), the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Cat. # 634411, 634412, 634413, 634414), or the QIAseq FX Single Cell RNA Library Kit (Qiagen, Cat. # 180733) as specified in the respective user manuals. For the comparison against the SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing, the cDNA synthesis and amplification from the SMART-Seq v4 cDNA protocol was performed with the assistance of a SMARTer Apollo instrument, as described in the SMART-Seq v4 Reagent Kit for the SMARTer Apollo System user manual. The Apollo System was also used to prepare the Nextera libraries (with the Nextera XT DNA Library Preparation Kit; Illumina, Cat. # FC-131-1024) from the SMART-Seq v4 cDNA. PCR cycling parameters varied for the respective input types and amounts as specified in the user manuals. For the QIAseq FX kit, the options for "Amplification of Total RNA from Single Cells", involving the use of a mixture of Oligo dT Primer and Random Primer, were followed.

For preparation of libraries directly from cells, aliquots of 1–1,000 cells were obtained using FACS. Sorting was done using a BD FACSJazz Cell Sorter. Cells were labeled with an anti-CD81 (K562, Jurkat) or CD45 (A375) antibody and sorted in 7 µl of 1X PBS buffer (DPBS without calcium chloride and magnesium chloride; Sigma Aldrich, Cat. # D8537) in 8-tube PCR strips. Following sorting, cells were subjected to a quick spin and immediately flash frozen on dry ice, then stored at –80°C until use (up to three months after sorting). For the comparison against the SMART-Seq v4 Reagent Kit for the SMARTer Apollo System, additional 8-tube PCR strips were prepared containing 12 µl of FACS Dispensing Solution (prepared by mixing 0.95 µl of 10X lysis buffer, 0.05 µl of RNase Inhibitor, 1 µl of 3' SMART-Seq CDS Primer II A) and 10.5 µl of water. A single batch of cells was sorted either in PBS or FACS Dispensing Solution; this same batch of cells was used for testing both kits. For the comparison between the Stranded kit and QIAseq FX kit, cells from the Jurkat cell line were sorted in 7 µl of PBS as described above (50 cells per tube), and two aliquots from this batch were used for each kit.

For comparison to purified total RNA, aliquots of 1,000 A375 cells were used to extract RNA using NucleoSpin RNA XS (Cat. # 740902) in accordance with the manufacturer's instructions (including the optional on-column DNase step). Purified total RNA was eluted in a volume of 12 µl, from which 7 µl was used as input for library construction.

Sequencing and analysis

Libraries were sequenced on a NextSeq® 500 instrument using 2 x 75 bp paired-end reads. Reads from all libraries were trimmed to remove Illumina adapters and polyA sequences, and mapped to human rRNA and mitochondrial genomes using CLC Genomics Workbench. The remaining reads were subsequently mapped using CLC to the human genome with RefSeq annotation. All percentages shown, including the number of reads that map to introns, exons, or intergenic regions, are percentages of total reads in each library. The number of transcripts identified for each library was determined based on the number of transcripts with an FPKM ≥1 or 0.1, as specified. The number of reads mapping to the correct strand (as defined in the current genome annotation) was determined using Picard analysis. For analysis of lncRNA, reads were mapped against the "Long non-coding RNA transcript sequences" fasta file included in the GENCODE GRCh38-release 26 (containing 27,720 loci transcripts). The number of lncRNAs detected is based on a cutoff of 10 unique counts or more. Scatter plots in Figure 5 were generated using FPKM values from CLC mapping to the transcriptome. To highlight transcripts found in only one replicate (dropouts), 0.01 was added to each value prior to graphing (Figure 2B).