# A SMARTer solution to stranded single-cell RNA-seq

**Suvarna Gandlur*, Nathalie Bolduc, Simon Lee, Christopher Hardy, Ankita Das, Magnolia Bostick, Andrew Farmer**

Takara Bio USA, Inc., Mountain View, CA 94043, USA *Corresponding Author: suvarna_gandlur@takarabio.com

## Abstract

Our SMARTer® NGS reagent 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 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 both generate transcriptome profiles from mRNA only. Oligo(dT) priming is a very efficient way to capture the transcriptome, with minimal uninformative reads (e.g., 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 nature of the processing (e.g., FFPE samples) or method of isolation. Additionally, these earlier single-cell kits do not preserve strand-of-origin information. All of these factors motivated the development of the SMART-Seq Stranded Kit, which allows for generation of sequencing-ready, stranded Illumina® libraries directly from 1–1,000 sorted cells or an equivalent amount (10 pg–10 ng) of purified total RNA of any quality.

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, as opposed to direct removal of corresponding rRNA molecules prior to reverse transcription. Since cDNA synthesis in the SMART-Seq Stranded Kit relies on random priming, rRNA is also captured, and removal of the resulting cDNA prior to sequencing is essential. The SMART-Seq Stranded Kit protocol can be completed within seven hours, and a convenient pooling option for inputs between one to ten cells facilitates greater ease-of-use by minimizing the number of samples being handled (Figure 1).

## 1 Simple workflow for generation of stranded libraries from cells

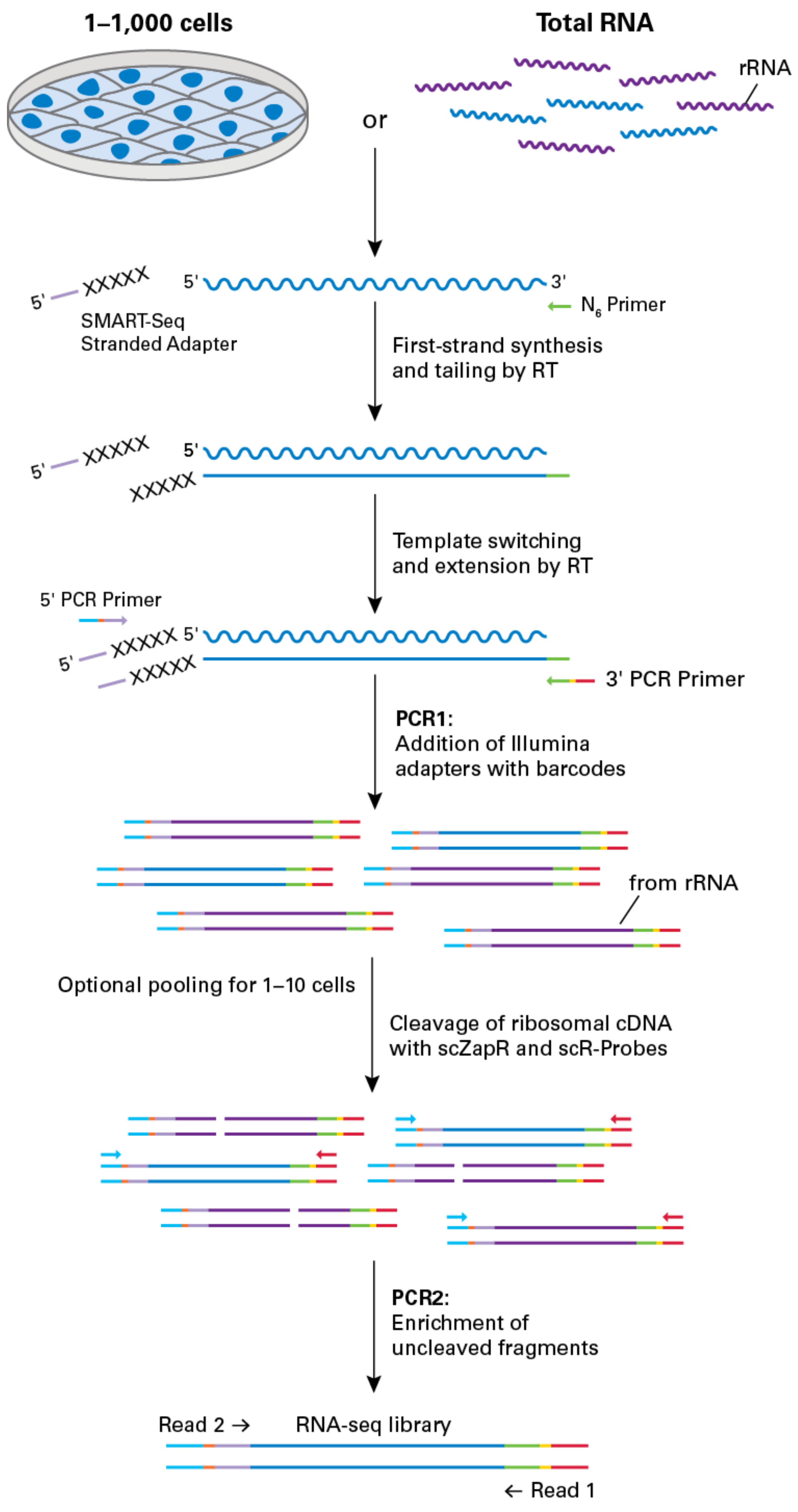
1–1,000 cells
Total RNA
or
rRNA
5'
XXXXX
5'
3'
SMART-Seq Stranded Adapter
N6 Primer
First-strand synthesis and tailing by RT
XXXXX
Template switching and extension by RT
5' PCR Primer
3' PCR Primer
PCR1:
Addition of Illumina adapters with barcodes
from rRNA
Optional pooling for 1–10 cells
Cleavage of ribosomal cDNA with scZapR and scR-Probes
PCR2:
Enrichment of uncleaved fragments
Read 2 →
RNA-seq library
← Read 1

**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 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.
- In the next step, a first round of PCR amplification (PCR1) adds full-length Illumina adapters, including barcodes.
- 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, with priming sites available on both 5' and 3' ends for further PCR amplification.
- These fragments are enriched via a second round of PCR amplification (PCR2) using primers universal to all libraries. The final library contains sequences allowing clustering on any Illumina flow cell.
- For inputs below 100 ng or 10 cells, an optional pooling of up to 12 samples after PCR1 allows for greater ease-of-use by minimizing the number of samples to be processed downstream.

## 2 Consistent sequencing metrics across RNA input amounts

**Sequencing alignment metrics for 10 pg–10 ng total RNA**

| RNA source | **Human brain total RNA** | | | | |
|---|---|---|---|---|---|
| Input amount (ng) | **10** | **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 correlations | 0.99 | 0.99 | 0.99 | 0.97 | 0.92 |
| 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 |

**Figure 2. 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. Reproducibility between replicates was high at every input level, including the single-cell equivalent of 10 pg of total RNA, as demonstrated by the high Pearson correlations between technical replicates. The data show that even within this single-cell input rang, over 97% of the reads match the correct strand, as determined per biological annotation.

## 3 The SMART-Seq Stranded Kit outperforms the Pico v2 kit for ultra-low inputs

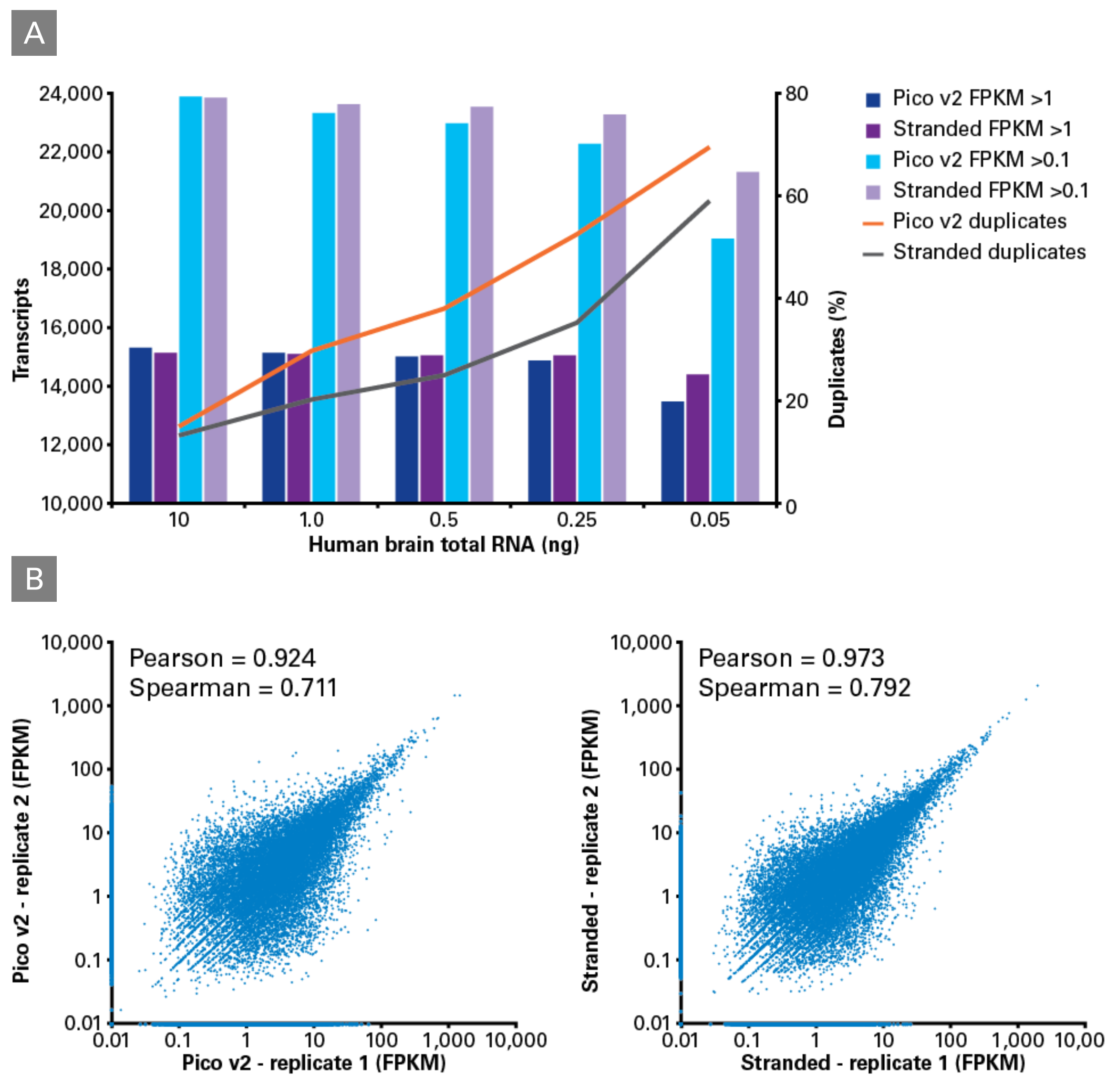

**Figure 3. The SMART-Seq Stranded Kit outperforms the Pico v2 kit for ultra-low inputs. Panel A.** Human brain total RNA (50 pg–10 ng) was used to generate RNA-seq libraries in triplicate with the SMART-Seq Stranded Kit (Stranded) and the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Pico v2). For both kits, sequencing data were down-sampled to 2.5 million paired-end reads prior to analysis. The SMART-Seq Stranded Kit shows a higher sensitivity (i.e., identifies more transcripts) for inputs <500 pg. **Panel B.** Comparison of transcript expression level from libraries generated with 50 pg of total RNA (Panel A) shows tighter Pearson and Spearman correlations with the SMART-Seq Stranded Kit. FPKM values are shown on a log10 scale. Transcripts represented in only one sample (dropouts) can be seen along the X- and Y-axes of the scatter plots. For the SMART-Seq Stranded Kit, the dropout transcripts are restricted to expression levels close to or below 10 FPKM, while with the Pico v2 kit, a higher proportion of the dropouts are >10 FPKM.

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

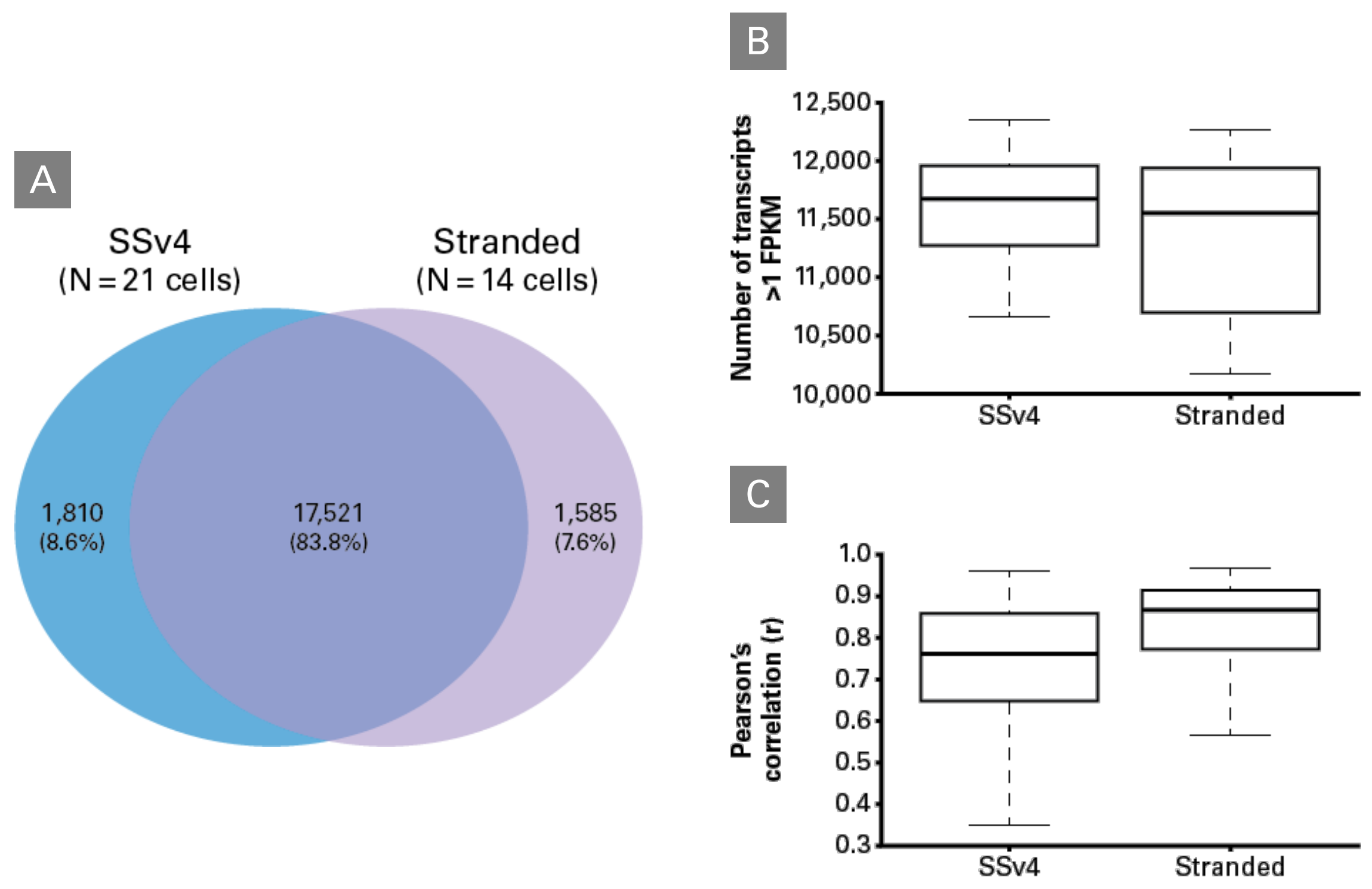

**Figure 4. Comparison between SMART-Seq v4 and SMART-Seq Stranded kits using cells isolated by FACS.** Single cells (K562) isolated by FACS were used to generate RNA-seq libraries with the SMART-Seq Stranded Kit (Stranded) and a SMART-Seq v4 kit (SSv4; cDNA from this kit was further processed with Nextera® XT). **Panel A.** The overlap in the total number of transcripts identified (FPKM >1) by each kit was analyzed and shown to be 84%. **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.

## 5 High reproducibility across cell input amounts

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 |

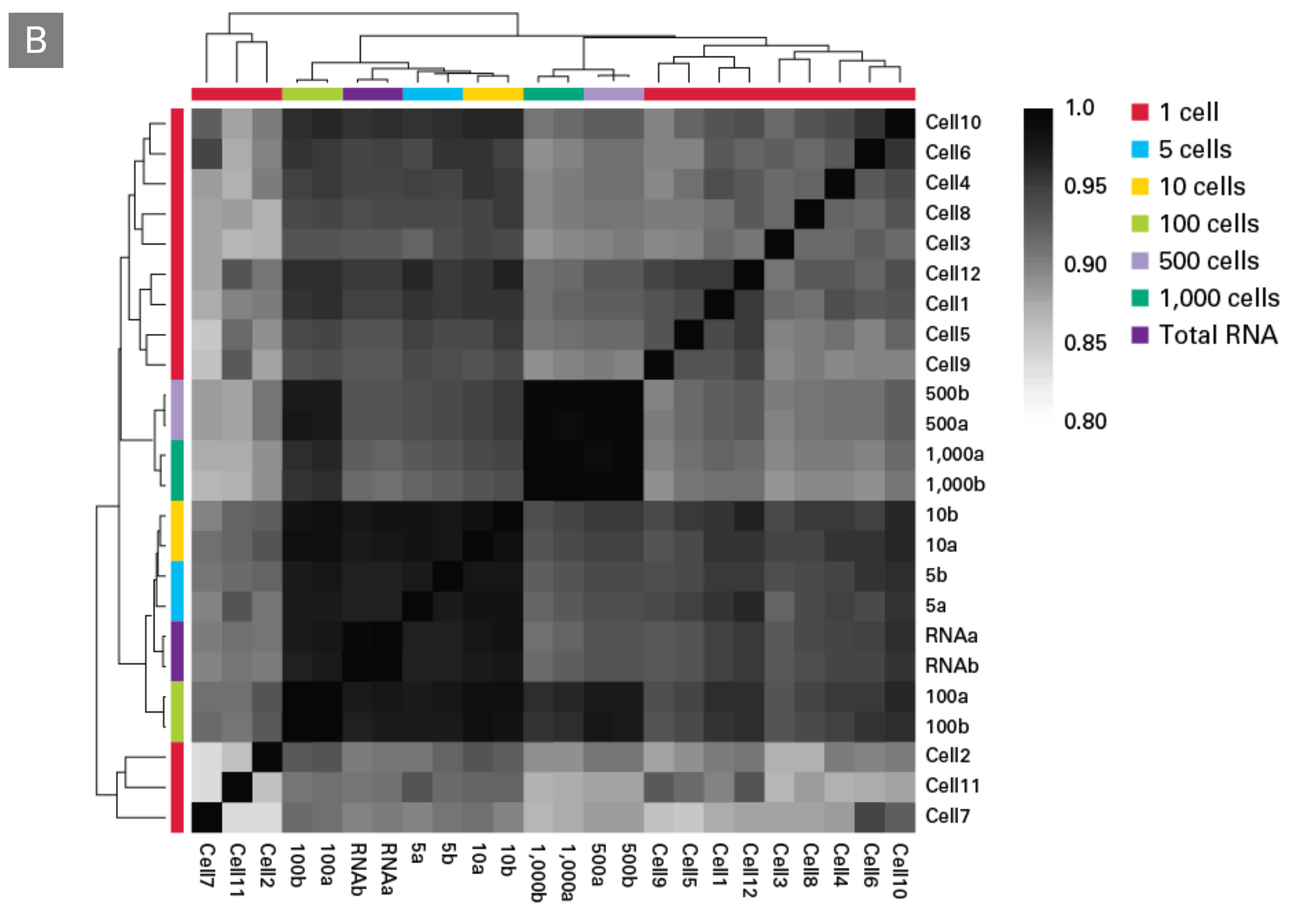

**Figure 5. High reproducibility across cell input amounts.** A375 cells isolated by FACS were used to generate RNA-seq libraries with the SMART-Seq Stranded Kit. Input varied from 1 cell to 1,000 cells, with two replicates per input of 5–1,000 cells and 12 replicates for the single cells. For comparison, 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, 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 cells. 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. **Panel B.** Hierarchical clustering heatmap displaying the Euclidean distance between all the samples shown in Panel A, and reporting Pearson correlations ranging from 0.85 to 0.99. Single cells are labeled Cell1–Cell12; replicates for other inputs are labeled a–b.

## Methods

Sequencing libraries were generated using the SMART-Seq Stranded Kit, the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian, and the SMART-Seq v4 Reagent Kit for the SMARTer™ Apollo™ System, as specified in the respective user manual.

For the 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.

Libraries were sequenced on a NextSeq® 500 instrument using 2 x 75 bp paired-end reads, and analysis was performed using CLC Genomics Workbench (mapping to the to the human (hg19) 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. Scatter plots in Figure 3 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 3B).

For analysis of lncRNA, reads were mapped against the lncRNA data set from GENCODE GRCh38-v26. The number of lncRNAs detected is based on a cutoff of 10 unique counts or more.

## Conclusions

- Simple workflow starts directly from 1–1,000 cells or 10 pg–10 ng total RNA to generate sequencing-ready Illumia libraries in 7 hours
- Unparalleled sensitivity for single-cell, full-length total RNA sequencing with strand-of-origin information
- Reproducible, sensitive detection of coding and noncoding transcripts from total RNA and single cells
- Comparable sensitivity to our gold-standard single-cell RNA sequencing NGS technology used in the SMART-Seq v4 kit

**Takara Bio USA, Inc.**
United States/Canada: +1.800.662.2566 • Asia Pacific: +1.650.919.7300 • Europe: +33.(0)1.3904.6880 • Japan: +81.(0)77.565.6999

**Scan to download your copy of this poster visit:**
**https://go.takarabio.com/ABRF-2018.html**

800.662.2566
Visit us at www.takarabio.com