A novel, high-throughput full-length scRNA-seq workflow for improved biomarker discovery

Peng Xu*, Joseph Liu[1], Yana Ryan[1], Kazuo Tori[1], Xuan Li[1], Hima Anbunathan[1], Mike Covington[1], Tomoya Uchiyama[1], Mohammad Fallahi[1], Takara Bio USA Engineering, Xuan Qu[2], Xiaoyun Xing[2], Ting Wang[2], Bryan Bell[1], Shuwen Chen[1], Yue Yun[1], Andrew Farmer[1]

1 Takara Bio USA, Inc., 2560 Orchard Pkwy, San Jose, CA 95131 USA 2 Washington University in St. Louis, St. Louis, MO

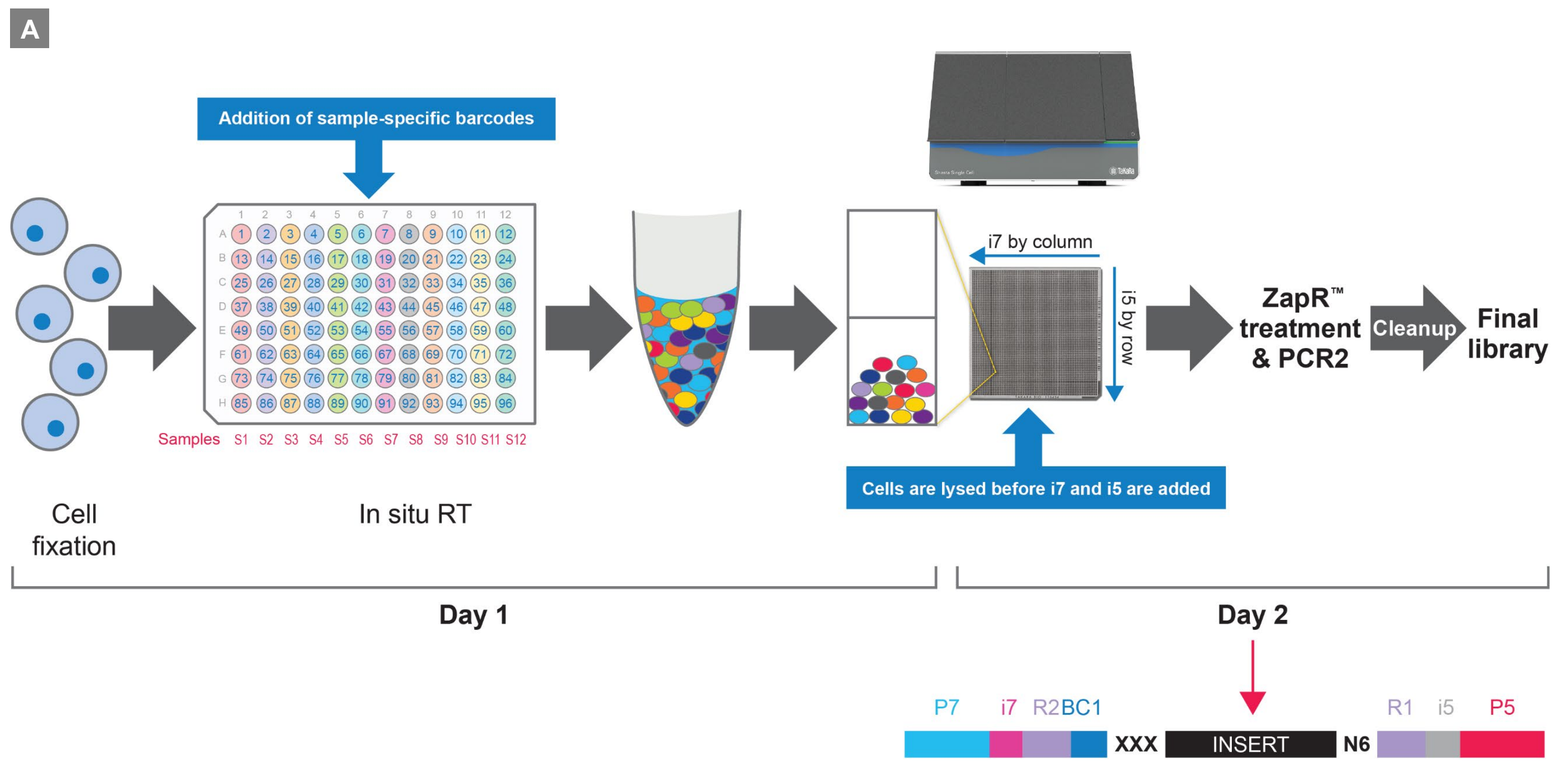
*Corresponding author

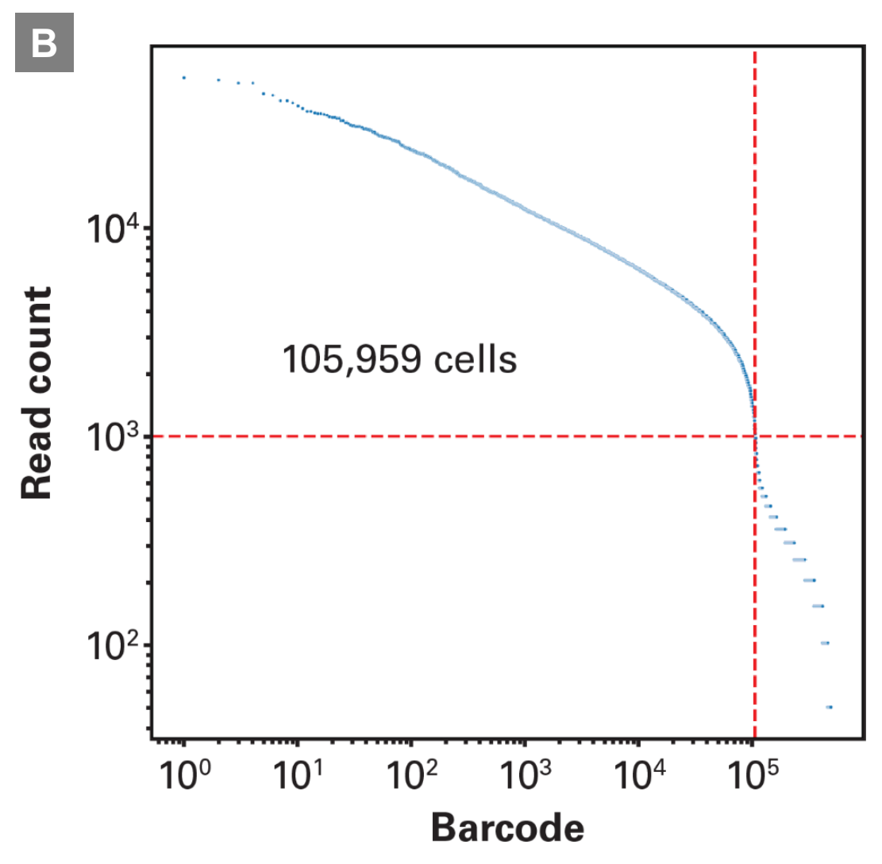
Introduction

Single-cell RNA-seq (scRNA-seq) analysis has been widely applied in oncology research for biomarker discovery. Although droplet-based methods are commonly used for such studies owing to their high throughput, they still miss important insights due to their lack of full-length transcript coverage. While full-length methods are available, to date, they have not been able to meet the throughput demands of many researchers. Moreover, both droplet and full-length scRNA-seq methods do not currently provide adequate readouts for noncoding genes, thereby limiting the investigation of gene regulatory networks to protein-coding genes. To close these gaps, we have developed a new high-throughput full-length scRNA-seq workflow that comprehensively profiles both protein-coding and noncoding genes in up to 100,000 cells within two days.

Methods

Our new high-throughput workflow uses a unique indexing strategy, starting with a 96-well-plate format for the addition of sample-specific barcodes, followed by automated addition of nanowell-specific barcodes after cells are lysed in a 5,184-well nanochip using our Shasta™ Single-Cell System. Initial testing demonstrated that our method could handle up to 100,000 cells without generating significant levels of doublets due to barcode collisions. To further illustrate the capacity of the new scRNA-seq approach, we profiled a total of approximately 11,000 isogenic A549 cells that either express WT TP53 or are TP53 null. In addition, both isogenic cell lines were treated with epigenetic therapy or mock treatment. Libraries were generated and sequenced using an Illumina® NextSeq® 2000 sequencer. The sequencing data was then analyzed to define differential gene expression for both protein-coding and noncoding transcripts as a function of TP53 genotype and treatment condition, using Cogent™ NGS software.





Shasta Total RNA-Seq workflow. Panel A shows an easy workflow featuring a two-day protocol from cell fixation to a sequencing-ready library. With in situ RT to add sample-specific barcodes in a 96-well-plate format, users can load up to 96 different samples. **Panel B.** The knee plot shows the high-throughput feature of the Shasta Total RNA-Seq workflow with >100,000 cells analyzed from one experiment.

Results

Preliminary analysis showed that, on average, approximately 11,000 genes and 40,000 transcripts were detected per single cell at a read depth of 100,000 reads per cell. UMAP-based clustering confidently separated the cells according to their genotypes and treatment conditions using either protein-coding genes or noncoding genes. Furthermore, differential expression analysis identified both protein-coding and noncoding transcripts with significant expression differences, underscoring biological significance.

1 Shasta Total RNA-Seq enables high-throughput, full-length single-cell RNA-seq in oncology samples

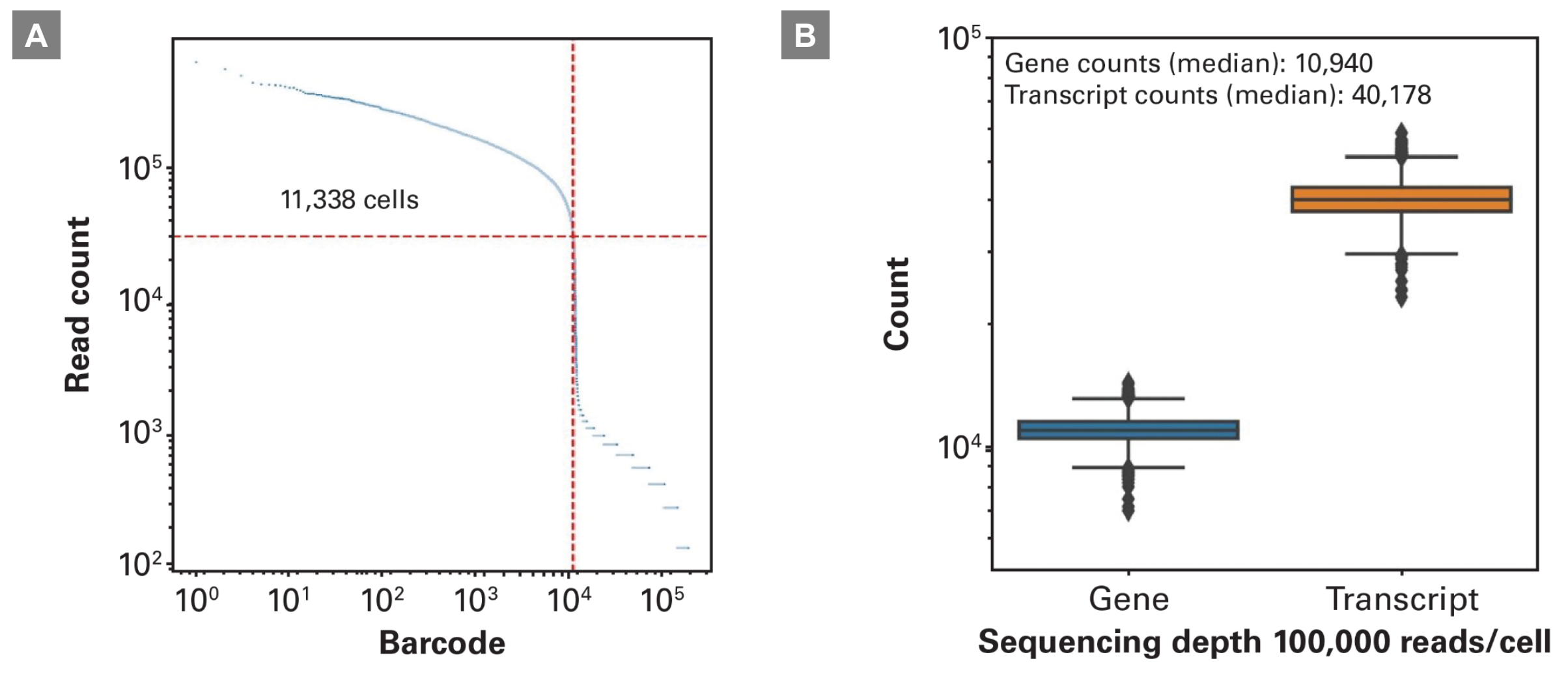


Figure 1. Panel A. Knee plot showing number of A549 cells passing QC thresholds in the Cogent NGS Analysis Pipeline (CogentAP). **Panel B.** Number of genes (blue, median 10,940 genes/cell at 100,000 reads/cell) and transcripts (orange, median 40,178 transcripts/cell at 100,000 reads/cell) identified are shown in box plots.

2 Shasta Total RNA-Seq separates different cell populations by profiling gene expression changes

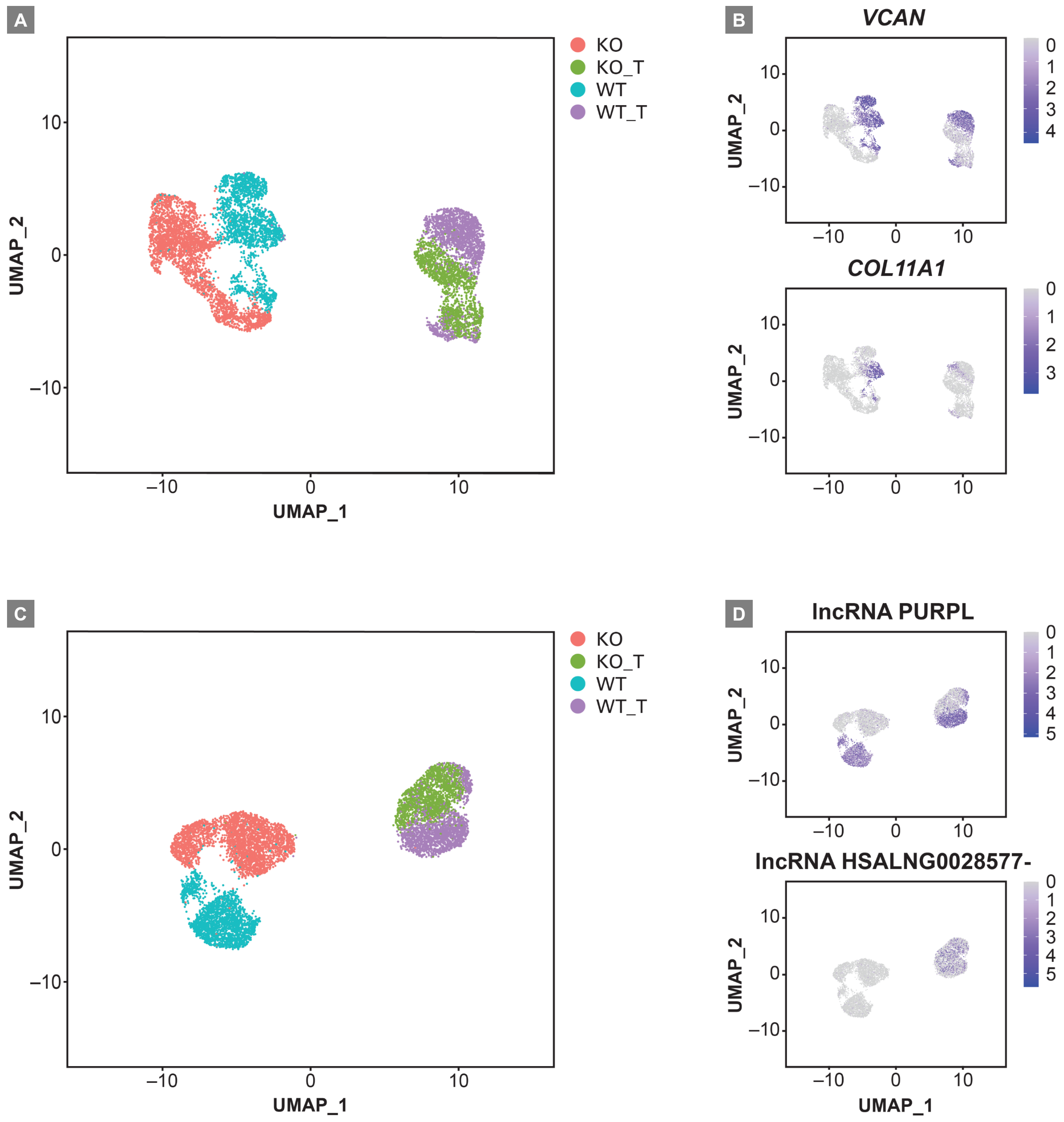


Figure 2. Shasta Total RNA-Seq separates different cell populations by profiling the expression of protein-coding genes and noncoding genes such as lncRNAs. Panel A. UMAP plot showing four distinct clusters for four A549 samples based on gene expression profile using the CogentAP protein-coding pipeline: untreated wildtype A549 cells (WT, cyan); untreated A549 cells with p53 knockout (KO, orange); WT A549 cells with epitherapy treatment (WT_T, purple); A549 cells with p53 KO with epithapy treatment (KO_T, green). **Panel B.** Expression of identified signature genes in different cell populations using the CogentAP protein-coding pipeline. UMAP plots show the expression of *VCAN* (upper plot) and *COL11A1* (lower plot) enriched in WT cells but not in the KO population. **Panel C.** UMAP plot showing the above different A549 cell populations form distinct clusters with different lncRNA expression profiles. **Panel D.** Full-length coverage of our new Shasta Total RNA-Seq workflow unraveled enriched expression of lncRNA PURPL in WT A549 cells (treated and untreated) (upper plot) and lncRNA HSALNG0028577- in treated WT and p53 KO A549 cells (lower plot).

3 Shasta Total RNA-Seq identifies both protein-coding and noncoding transcripts with significant expression differences

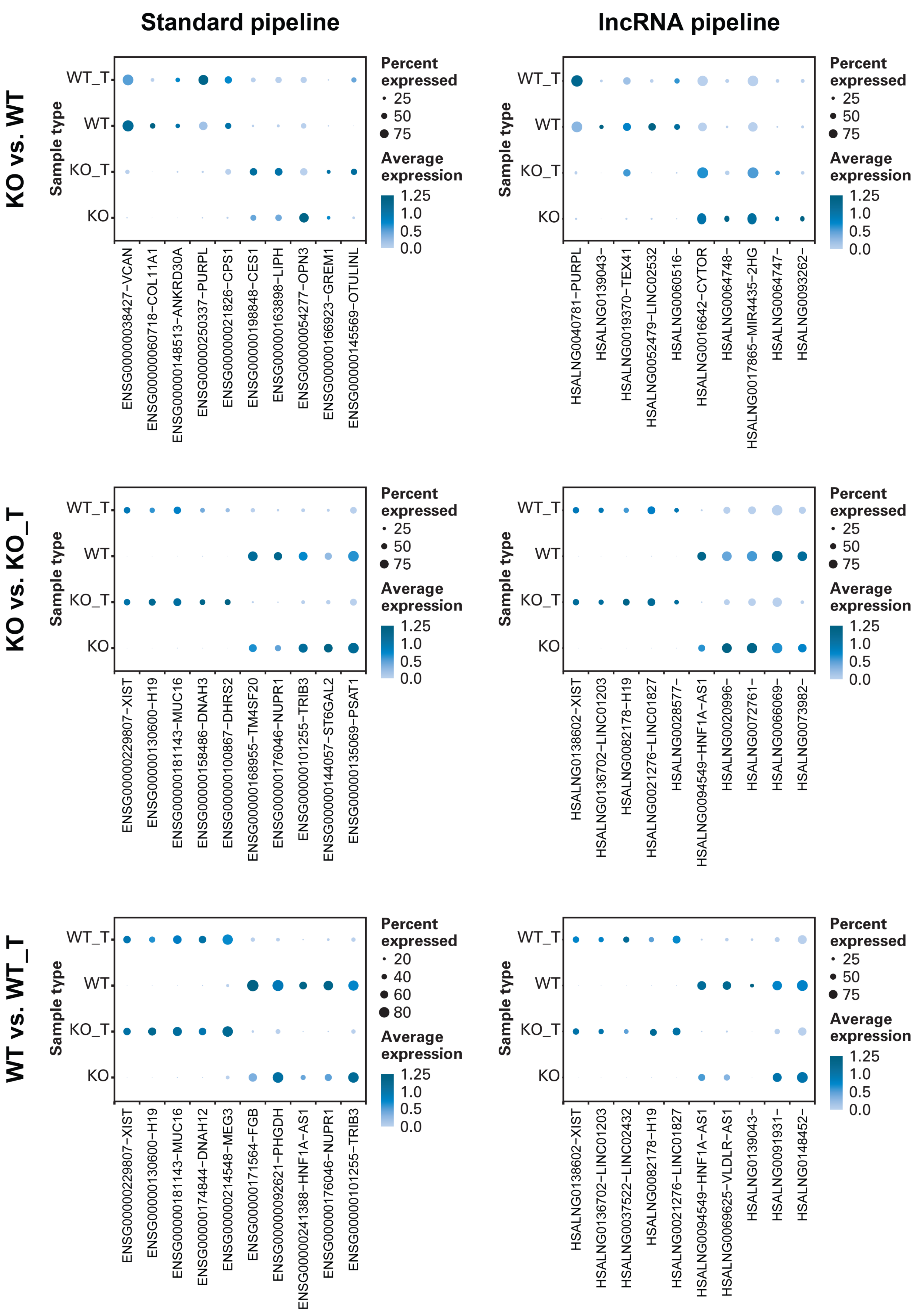


Figure 3. Dot plots show the top 10 differentially expressed genes between different samples identified by the CogentAP standard pipeline (left) and the lncRNA pipeline (right). The CogentAP standard pipeline focuses on protein-coding genes but also identifies some ncRNAs. The lncRNA pipeline focuses specifically on the analysis of lncRNAs. The dot size represents the percentage of the cell population expressing the genes. The color intensity of the dot indicates the gene expression level.

Conclusions

- Our new high-throughput full-length scRNA-seq workflow enables the preparation of high-quality full-length RNA-seq libraries for up to 100,000 cells with a unique indexing strategy and shows high sensitivity and specificity in gene/transcript detection and quantification.
- The technology significantly improves the ability to identify new biomarkers by enabling comprehensive profiling of both protein-coding and noncoding full-length transcripts.

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 get priority access to product and technical information, or visit takarabio.com/shasta2024

800.662.2566
takarabio.com