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

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

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

*Corresponding Authors

Abstract

Objective: Single-cell RNA-seq (scRNA-seq) has been widely applied in oncology research for biomarker discovery. Although droplet-based methods are commonly used owing to their throughput, they miss insights due to a lack of full gene body coverage. To date, full gene body methods have not met the throughput demands. Moreover, current high-throughput methods do not provide adequate readouts for noncoding genes. We have developed Shasta™ Total RNA-Seq to comprehensively profile both protein-coding and noncoding genes in up to 100,000 cells.

Methods: Shasta Total RNA-Seq employs a unique TSO-based indexing strategy for adding sample-specific barcodes, allowing for 1-12 samples per run. Following sample barcoding, cells were dispensed into a 5,184-well nanochip using the Shasta Single Cell System. As an example, we examined four isogenic A549 samples expressing either WT TP53 or TP53 null, treated and untreated with epigenetic therapy. We also examined human peripheral blood mononuclear cells (PBMCs) to demonstrate the ability of this technology to identify different cell types. Libraries were sequenced on an Illumina® NextSeq® 2000 and analyzed using Cogent NGS Analysis Pipeline (CogentAP) and Discovery Software (CogentDS).

Results: The Shasta Total RNA-Seq method showcased its ability to construct full gene-body libraries from up to 100,000 cells. From A549 samples, we detected ~11,000 genes and 40,000 transcripts per cell at 100,000 reads/cell. UMAP analysis using either protein-coding or noncoding genes revealed four distinct clusters corresponding to four different genotypes and treatment conditions. Differential expression analysis identified transcripts showing significant expression differences. Furthermore, we detected splice junctions and isoforms, highlighting the benefit of full gene-body RNA-seq.

Conclusion: The Shasta Total RNA-Seq enables preparation of high-quality, full gene-body RNA-seq libraries for both cell lines and primary cells through an automated protocol. It demonstrates high sensitivity in gene and transcript detection. This technology enhances the capability to identify new biomarkers by enabling comprehensive profiling of both protein-coding and noncoding genes with full gene-body coverage.

1 Shasta Total RNA-Seq overview

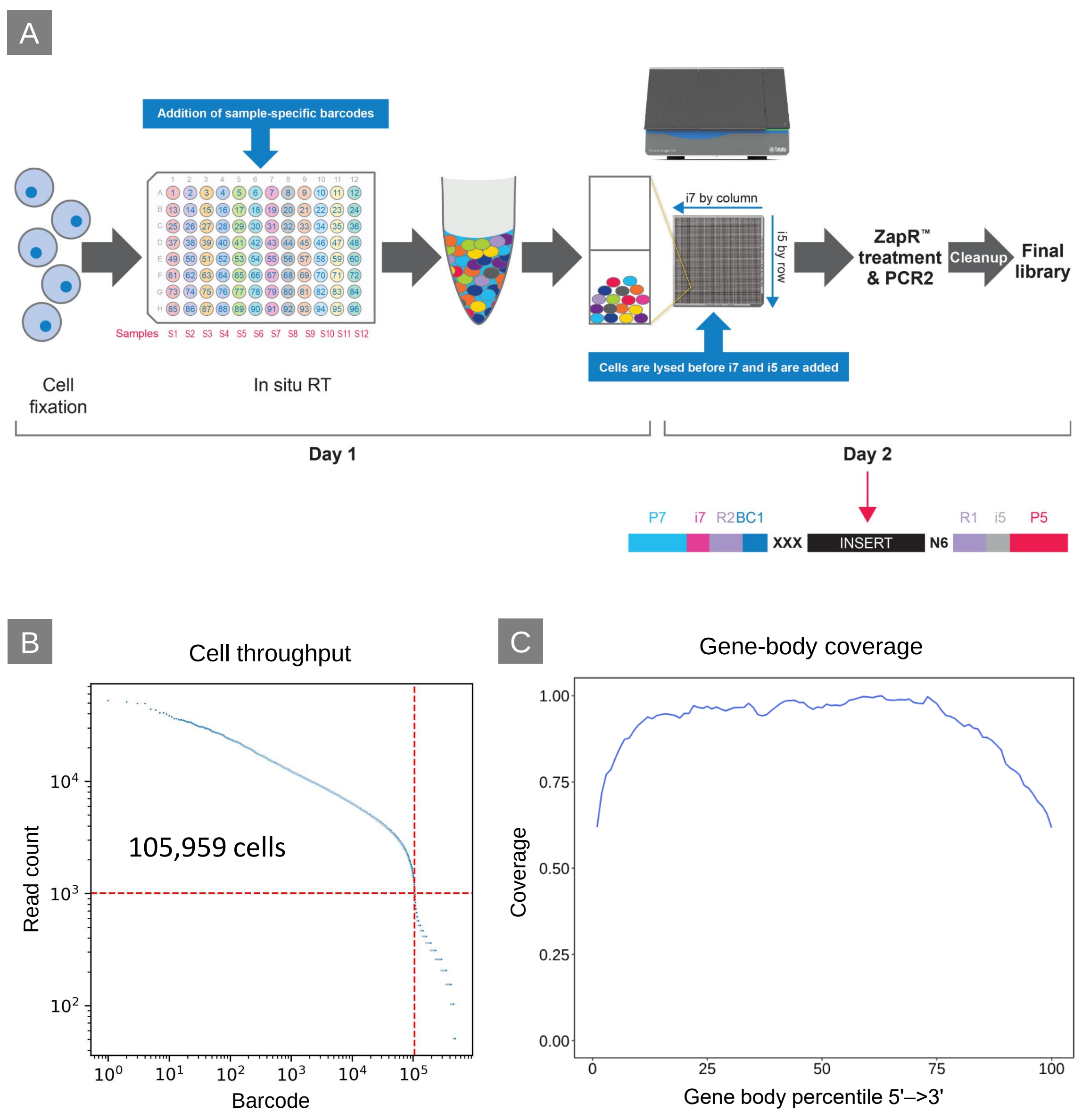
A
Addition of sample-specific barcodes
Samples S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 S11 S12
i7 by column
i5 by row
Cells are lysed before i7 and i5 are added
ZapR™ treatment & PCR2
Cleanup
Final library
Cell fixation
In situ RT
Day 1
Day 2
P7 i7 R2BC1 XXX INSERT N6 R1 i5 P5
B
Cell throughput
105,959 cells
Read count
Barcode
C
Gene-body coverage
Coverage
Gene body percentile 5'–>3'

Figure 1. Shasta Total RNA-Seq overview. Panel A. The 2-day workflow from cell fixation to final library. The ribosomal cDNA (originating from rRNA) is removed using ZapR technology (Takara Bio). **Panel B.** The knee plot shows the high-throughput ability of Shasta Total RNA-Seq with over 100,000 K562 cells analyzed from one experiment. **Panel C.** The gene-body coverage plot shows the uniformity of full gene-body coverage by Shasta Total RNA-Seq.

2 High-throughput and sensitivity for biomarker discovery

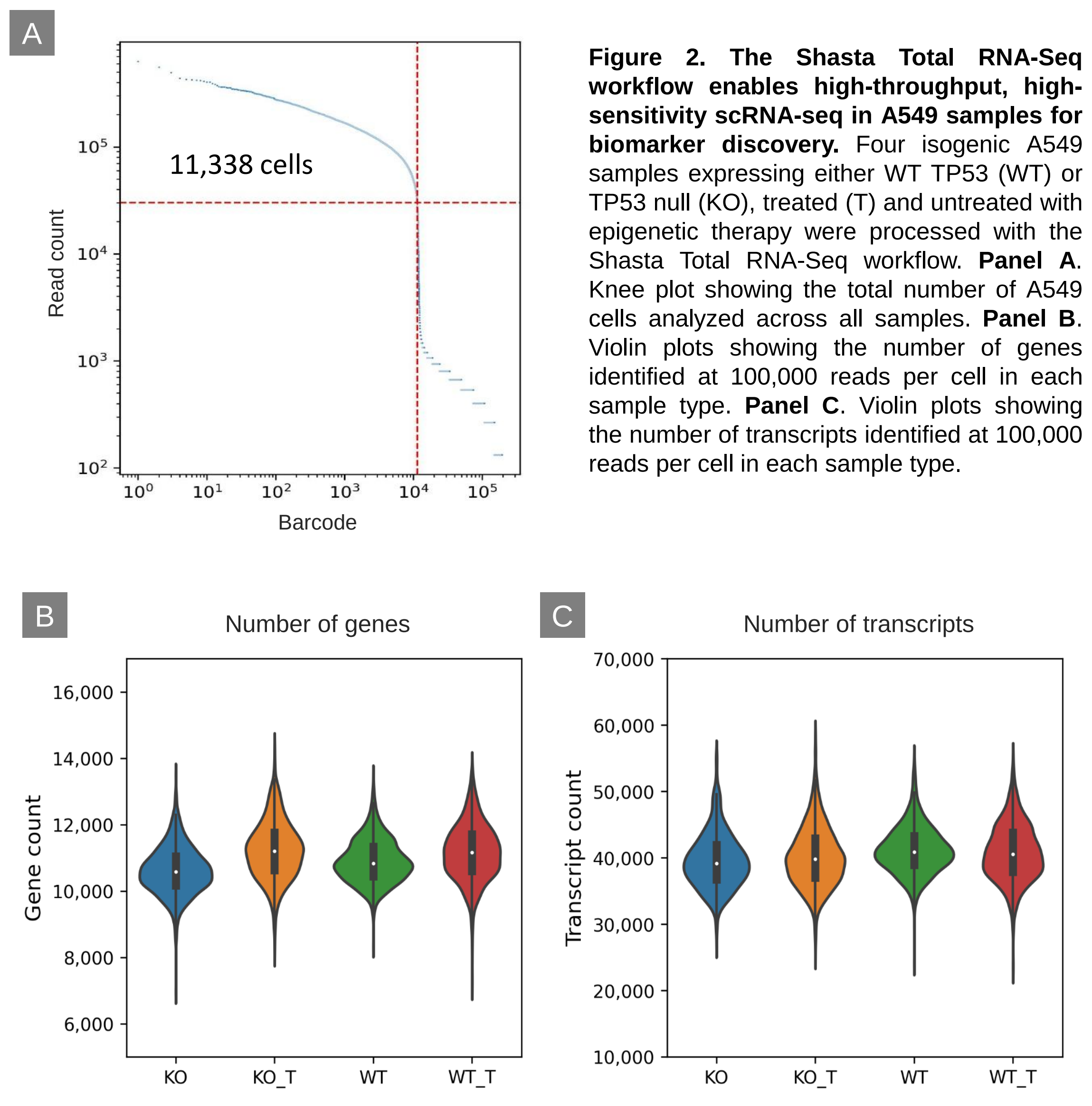

Figure 2. The Shasta Total RNA-Seq workflow enables high-throughput, high-sensitivity scRNA-seq in A549 samples for biomarker discovery. Four isogenic A549 samples expressing either WT TP53 (WT) or TP53 null (KO), treated (T) and untreated with epigenetic therapy were processed with the Shasta Total RNA-Seq workflow. **Panel A.** Knee plot showing the total number of A549 cells analyzed across all samples. **Panel B.** Violin plots showing the number of genes identified at 100,000 reads per cell in each sample type. **Panel C.** Violin plots showing the number of transcripts identified at 100,000 reads per cell in each sample type.

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
FOR RESEARCH USE ONLY. NOT FOR USE IN DIAGNOSTIC PROCEDURES.

3 Separation of distinct cell populations based on gene expression

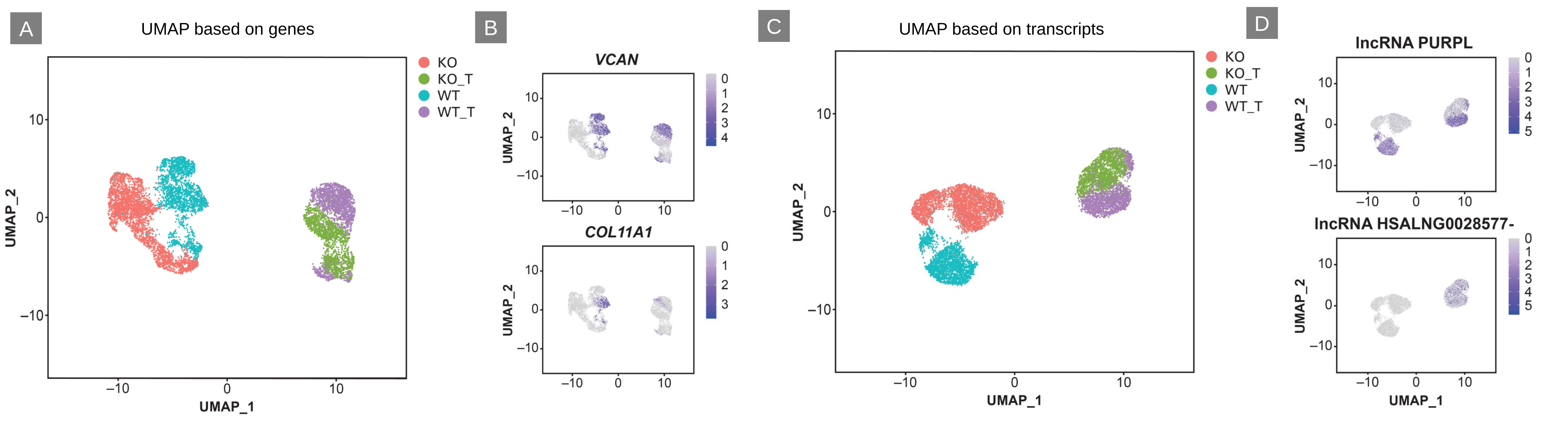

Figure 3. Shasta Total RNA-Seq separates different cell populations by profiling gene expression changes. Panel A. UMAP plot showing four major distinct clusters correlating with the four A549 samples based on gene expression profile using the CogentAP/DS standard pipeline. Untreated wild-type A549 cells (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 epitherapy treatment (KO_T) – green. **Panel B.** Feature plots showing the expression of two representative genes, *VCAN* (upper) and *COL11A1* (lower). Both genes were enriched in WT cells but not in the KO population. **Panel C.** UMAP plot showing the same four samples forming four major distinct clusters using the Cogent AP lncRNA pipeline. **Panel D.** Feature plots showing the expression of two representative lncRNA genes, PURPL and HSALNG0028577-. PURPL is highly expressed in WT A549 cells (treated and untreated) (upper), and HSALNG0028577- is highly expressed in treated WT and p53 KO A549 cells (lower).

4 Identification of differentially expressed protein-coding and noncoding transcripts

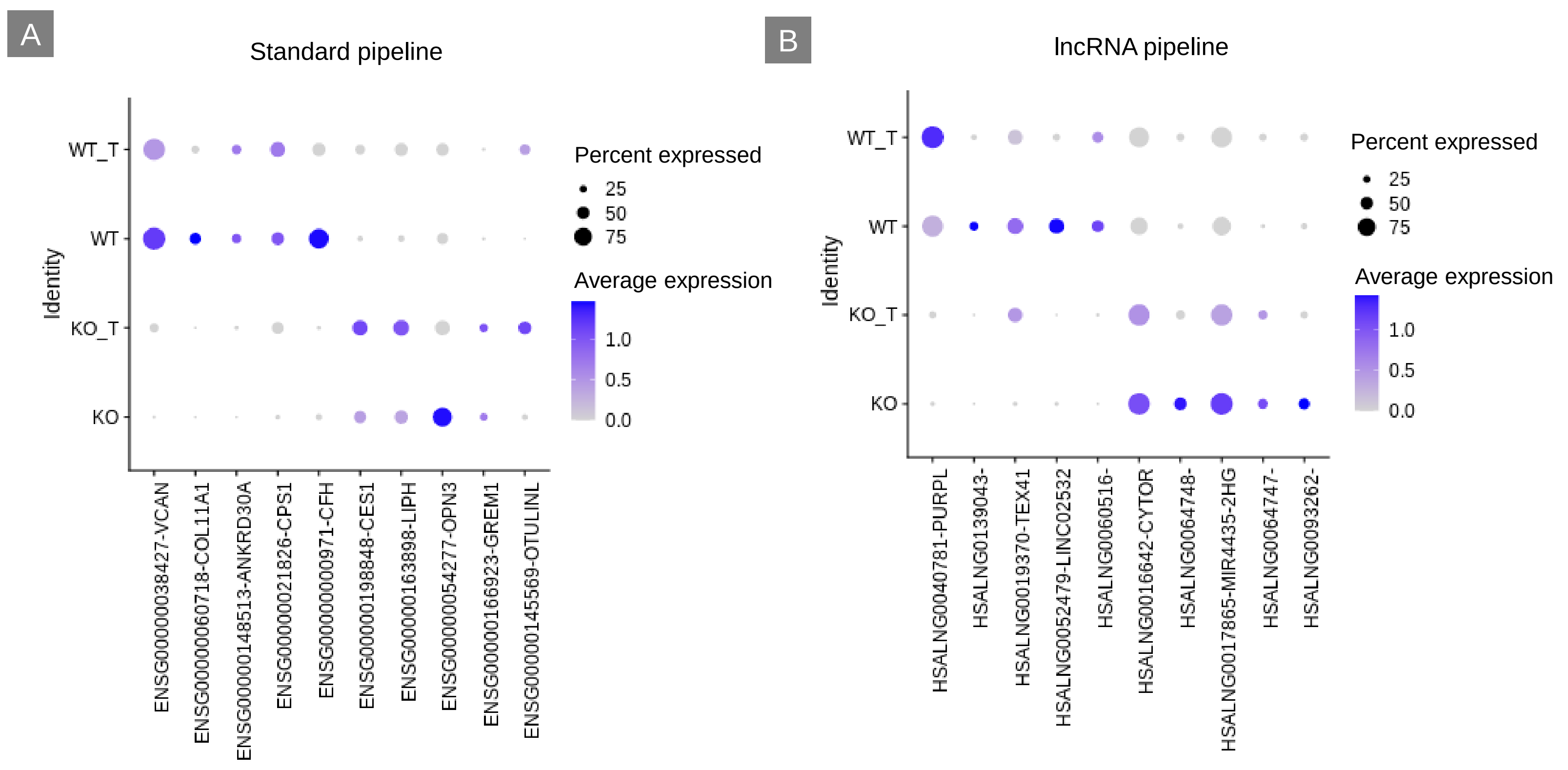

Figure 4. Shasta Total RNA-Seq identified differentially expressed genes using the Cogent AP standard pipeline and lncRNA pipeline. The top 10 differentially expressed transcripts were identified between TP53 null (KO) and WT TP53 (WT) A549 samples. Expression levels across all four A549 sample types (WT and KO, treated (T), and untreated with epigenetic therapy) are displayed in the dot plots. The dot size represents the percentage of the cell population expressing the genes. The color intensity of the dot indicates gene expression level. **Panel A**. Analysis using the CogentAP/DS standard pipeline. **Panel B**. Analysis using the CogentAP/DS lncRNA pipeline. The CogentAP/DS standard pipeline focuses on protein-coding genes but also identifies a subset of lncRNAs. The lncRNA pipeline focuses specifically on lncRNAs.

5 Identification of splicing isoforms from full gene-body coverage data

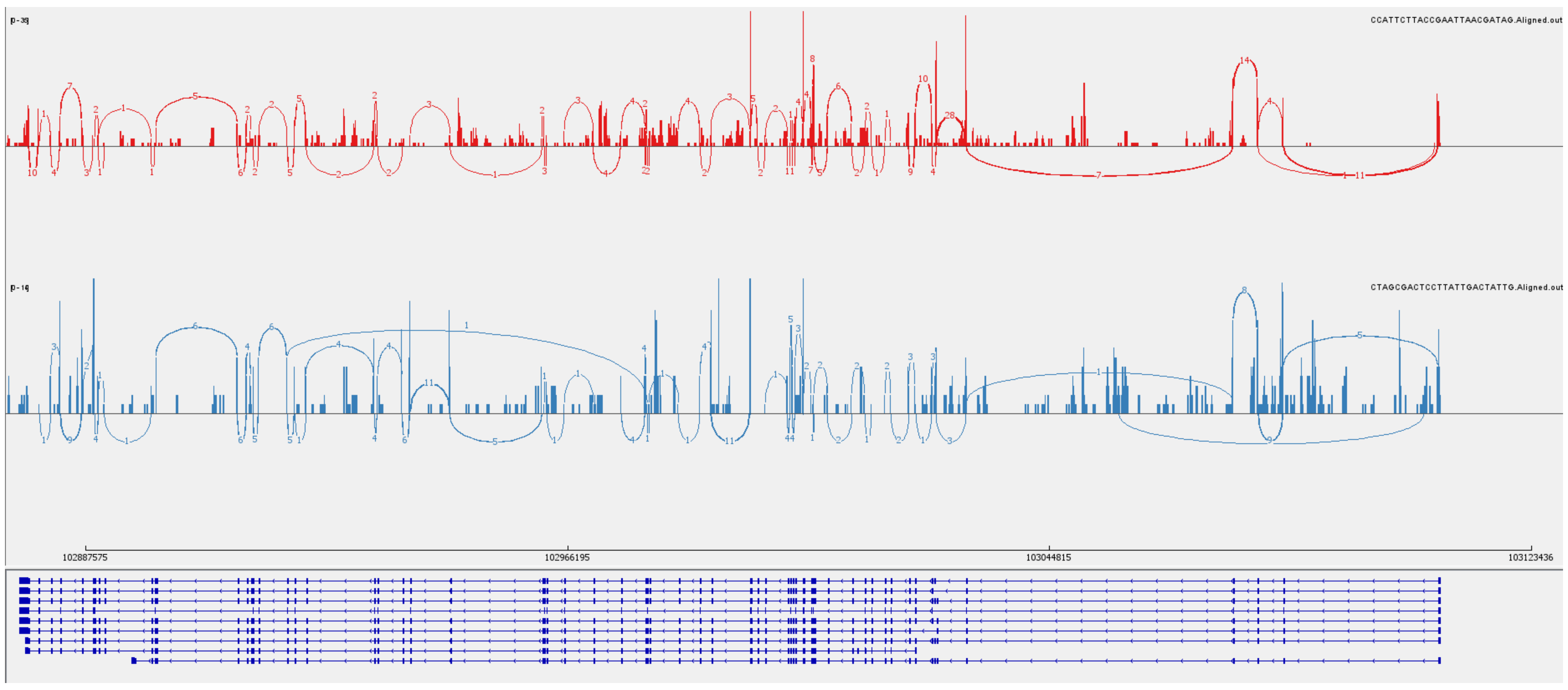

Figure 5. Shasta Total RNA-Seq identified splicing isoforms of *COL11A1*. Libraries created from two representative A549 cells were analyzed for splicing isoforms in the gene *COL11A1*. The sashimi plot was generated using Integrative Genomics Viewer. The full gene-body coverage data produced using the Shasta Total RNA-Seq kit allowed for the identification of splicing isoforms at the single-cell level.

6 Identification of distinct cell populations from a PBMC sample

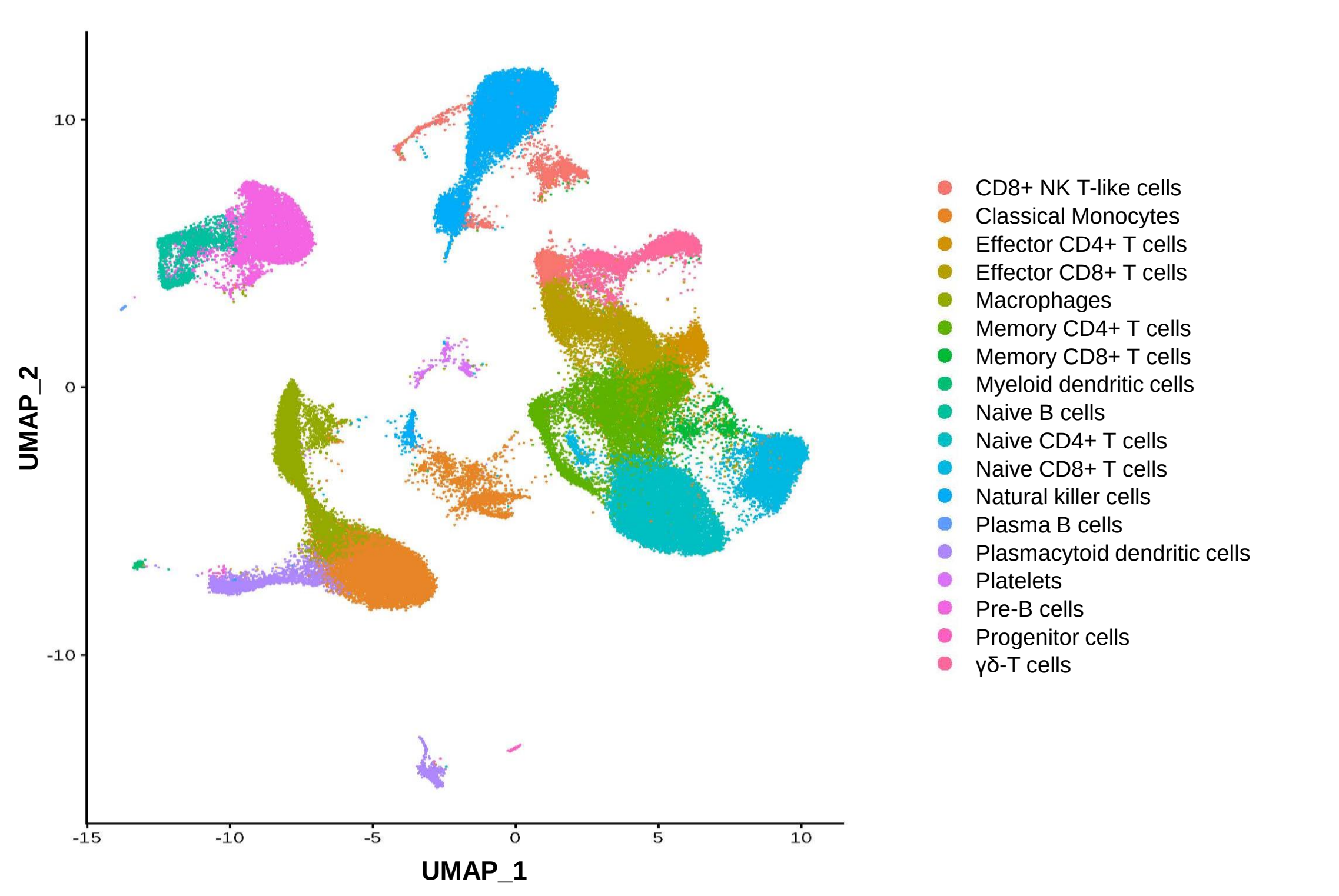

Figure 6. Shasta Total RNA-seq enables identification of distinct cell populations in a heterogeneous PBMC sample. 80,000 human PBMCs from a single sample were processed using the Shasta Total RNA-Seq workflow and sequencing data was analyzed using CogentAP/DS. The UMAP plot shows cell clusters that correlate to distinct cell populations within the PBMC sample.

Conclusions

- The Shasta Total RNA-Seq workflow enables high-quality, full gene-body coverage RNA-seq libraries for up to 100,000 cells with an easy, automated protocol.
- Automation with nanoliter reactions enabled by the Shasta Single Cell System reduces human labor and reagent costs.
- Data generated using Shasta Total RNA-Seq allows for identification of distinct cell populations within a highly complex PBMC sample.
- When paired with free-to-use Cogent NGS analysis tools, Shasta Total RNA-Seq easily identifies gene isoforms, fusions, and differentially expressed genes, transcripts, and lncRNAs.
- The full gene-body coverage technology significantly improves the ability to identify isoform information that end-counting technologies miss.

Learn more about Shasta Total RNA-Seq:
takarabio.com/shasta-total-rna-seq

800.662.2566
takarabio.com