Resolving tumor heterogeneity by uncovering novel genomic and transcriptomic events with a new scaled and automated workflow

Shuwen Chen[1], Peng Xu[1], Xuan Li[1], Joseph Liu[1], Yana Ryan[1], Kazuo Tori[1], Hima Anbunathan[1], Alan Du[1], Mike Covington[1], Raymond Mendoza[1], Samantha Leong[1], Tomoya Uchiyama[1], Mohammad Fallahi[1], Xuan Qu[2], Xiaoyun Xing[2], Bryan Bell[1], Patricio Espinoza[1], Ting Wang[2], Yue Yun[1], Andrew Farmer[1]

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

Poster LB047

Abstract

Single-cell omics has been widely applied in oncology research for biomarker discovery, providing an in-depth understanding of cancer heterogeneity. While bulk sequencing methods lack the specificity afforded by single-cell studies, single-cell applications miss insights due to trade offs for sensitivity at scale. Current high-throughput single-cell DNA-seq applications are limited to targeted-sequencing approaches, while scaled single-cell RNA-seq applications are limited to 3' or 5' end-counting methods or only capture polyadenylated RNA transcripts. To address these challenges, we have developed a nanoliter dispensing instrument, library prep chemistries, and a bioinformatics analyses suite that scales both single-cell genomics and transcriptomics assays while maintaining whole genome and whole transcriptome coverage, respectively. To demonstrate the ability to scale a non-targeted, single-cell whole genome amplification (WGA) application, we applied our new WGA workflow to cancer cell lines and primary Clear Cell Renal Cell Carcinoma (ccRCC) samples. The data revealed segmental aneuploidies and both germline and putative somatic variants in thousands of single cancer cells in a single day, at shallow sequencing depths of approximately 300,000 paired-end reads per single cell. Addressing the limitation of scaled single-cell transcriptomic solutions, our new total RNA-seq workflow is capable of generating data on up to 100,000 single cells at a time in two days. We applied this high-throughput workflow on cancer cells treated and untreated with epigenetic therapy and selected 11,000 cells to reach a deeper sequencing depth. The results demonstrate the ability of the new WGA workflow to identify new biomarkers through comprehensive profiling of both protein-coding and noncoding genes with full gene-body coverage, revealing significant expression differences across multiple RNA biotypes as well as identifying splice junction isoforms. Overall, our data highlight the advantages of complex and rich datasets generated from single-cell workflows, which, when paired with an unbiased, non-targeted approach, enable the discovery of novel genomic and transcriptomic events in oncological samples.

1 Shasta™ WGA overview

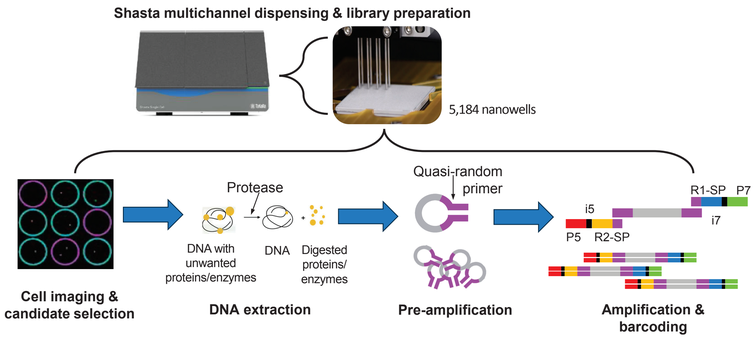

Figure 1. Shasta WGA workflow. Hoechst-stained single cells are dispensed into a 5,184-nanowell chip and screened by imaging. Reagents are deposited into nanowells in equal-volume dispenses: a DNA extraction mix, a pre-amplification mix of quasi-random primers, a PCR mix, and two indexing primer dispenses. The pooled barcoded libraries are ready for Illumina® sequencing after off-chip purification.

2 CNP heatmaps of single cells disassociated from tumor tissue and adjacent normal tissue

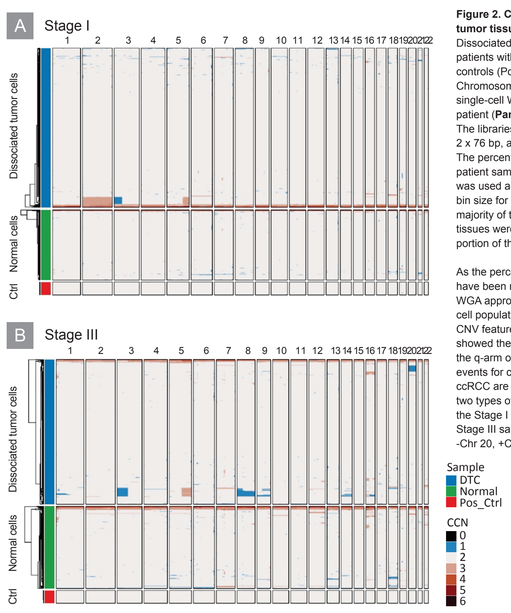

Figure 2. CNP heatmaps of single cells disassociated from the tumor tissues and the normal adjacent tissues of ccRCC. Dissociated cells from the tumor and normal adjacent tissues of two patients with ccRCC were processed via Shasta WGA. Positive controls (Pos_Ctrl) are in the bottom track of both panels. Chromosome copy number (CCN) is color coded in the legend. The single-cell WGA library was generated for 815 cells from the Stage I patient (**Panel A**), and 858 cells from the Stage III patient (**Panel B**). The libraries were sequenced on NextSeq® 500, at a read length of 2 x 76 bp, and an average read depth of 370,000 PE reads per cell. The percent of mapped and unique reads is 91.07% for the Stage I patient sample and 92.57% for the Stage III patient sample. Ginkgo was used as the pipeline for CNV variation calling, and the average bin size for segmentation was 1 Mb. As shown in the heatmap, the majority of the single cells from either tumors or the normal adjacent tissues were diploid cells with no extreme CNVs. Intriguingly, a portion of the tumor cells showed unique CNV features.

As the percentage of cells with CNVs was low, these features could have been masked in bulk DNA-seq data. However, the single-cell WGA approach allowed for discovery of heterogeneity in the tumor cell population and for the identification of rare cells with abnormal CNV features. The tumor cells in the Stage I and Stage III samples showed the loss of the p-arm on chromosome 3 (-3p) and gain of the q-arm on chromosome 5 (+5q), which are characteristic CNV events for ccRCC. The four most commonly mutated genes in ccRCC are included in -3p: *VHL*, *PBRM1*, *BAP1*, and *SETD2*. The two types of tumor cells also had unique CNV events. Tumor cells in the Stage I sample showed +Chr 2. Whereas the tumor cells in the Stage III sample showed other CNV events such as -8p, -Chr 8, -Chr 20, +Chr 16, -9q, -Chr9, etc.

3 Example copy number profiles for single tumor cells

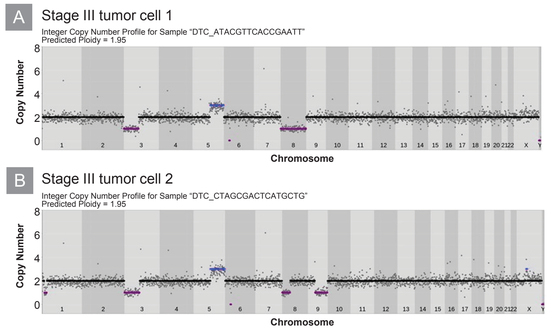

Figure 3. Example CNPs generated from individual ccRCC tumor cells with the Shasta WGA kit identify CNVs commonly associated with ccRCC. Tumor Cell 1 (**Panel A**) and Tumor Cell 2 (**Panel B**) both had loss of the p-arm on chromosome 3 (-3p), and gain of the q-arm on chromosome 5 (+5q). A comprehensive study from The Cancer Genome Atlas (TCGA) found that -3p is a cytogenic hallmark of ccRCC that was present in 91% of 417 patients (Cancer Genome Atlas Research Network 2013). This region encompasses the four most commonly mutated genes in ccRCC. In the same study, 67% of the 417 patients showed +5q, which was associated with better patient survival. Tumor cell 1 had loss of an entire chromosome 8, and tumor cell 2 had loss of a p-arm on chromosome 8 (-8p). The partial or complete loss of chromosome 8 has been reported in previous studies of ccRCC patients and is typically associated with *TCEB1* mutations (Sato et al. 2013).

4 Pseudo-bulk SNV analysis for cell clusters

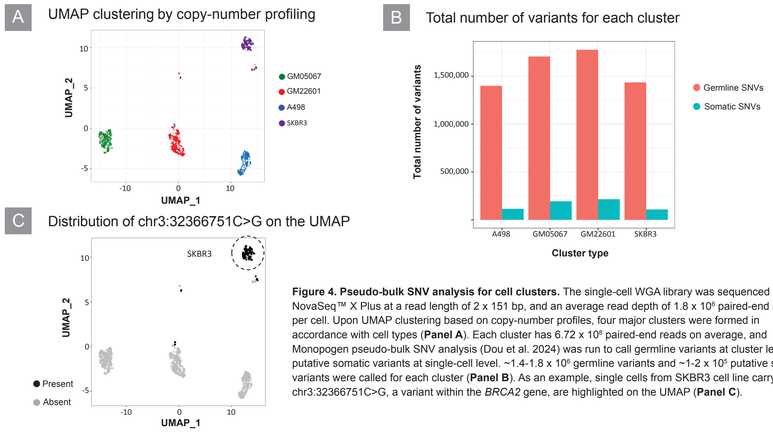

Figure 4. Pseudo-bulk SNV analysis for cell clusters. The single-cell WGA library was sequenced on NovaSeq™ X Plus at a read length of 2 x 151 bp, and an average read depth of 1.8 x 10^6 paired-end reads per cell. Upon UMAP clustering based on copy-number profiles, four major clusters were formed in accordance with cell types (**Panel A**). Each cluster has 6.72 x 10^5 paired-end reads on average, and Monopogen pseudo-bulk SNV analysis (Dou et al. 2024) was run to call germline variants at cluster level and putative somatic variants at single-cell level. ~1.4-1.8 x 10^6 germline variants and ~1-2 x 10^5 putative somatic variants were called for each cluster (**Panel B**). As an example, single cells from SKBR3 cell line carrying chr3:32366751C>G, a variant within the *BRCA2* gene, are highlighted on the UMAP (**Panel C**).

5 Shasta Total RNA-Seq overview

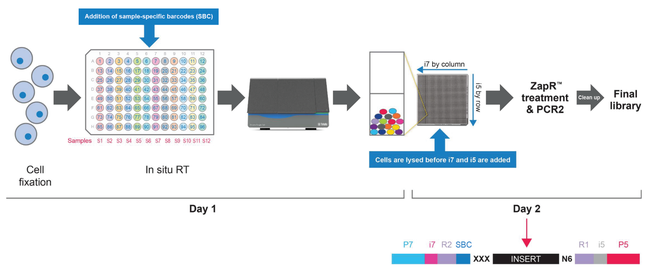

Figure 5. Shasta Total RNA-Seq overview. The two-day workflow from cell fixation to final library. The ribosomal cDNA (originating from rRNA) is removed using ZapR technology.

6 High sensitivity with high-throughput single-cell total RNA profiling

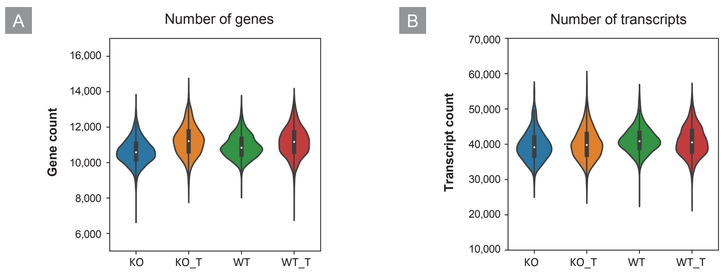

Figure 6. The Shasta Total RNA-Seq workflow enables high-throughput, high-sensitivity scRNA-seq in A549 samples for biomarker discovery. Four isogenic A549 samples expressing either wildtype TP53 (WT) or TP53 null (knockout [KO]), treated (T), and untreated with epigenetic therapy were processed with the Shasta Total RNA Seq workflow. Violin plots show the number of genes (**Panel A**) and transcripts (**Panel B**) identified at 100,000 reads per cell.

7 Identification of differentially expressed protein-coding and noncoding transcripts

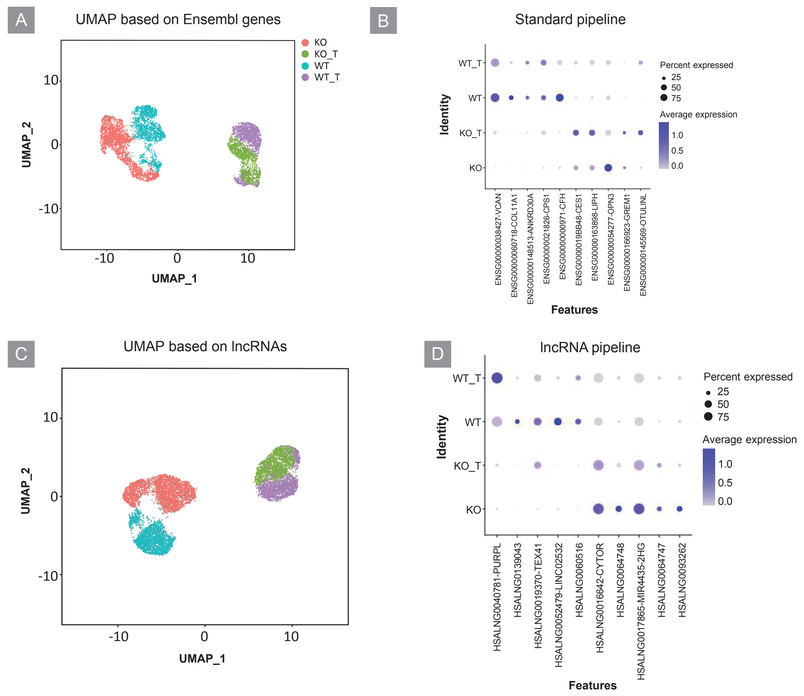

Figure 7. Shasta Total RNA Seq technology identified both protein-coding and lncRNA biomarkers. Panel A. UMAP plot showing four distinct clusters correlating with the four A549 samples based on gene expression profile using the Cogent™ Discovery Software and Analysis Pipeline standard pipeline (KO, KO_T, WT, WT_T). **Panel B.** Dot plot showing expression levels of the top 10 differentially expressed protein-coding genes that were identified between KO and WT samples. **Panel C.** UMAP plot showing the same four samples forming four major distinct clusters using the Cogent AP lncRNA pipeline. **Panel D.** Dot plot showing expression levels of the top 10 differentially expressed lncRNA genes that were identified between KO and WT samples.

8 Detect more splice junctions with full gene-body coverage

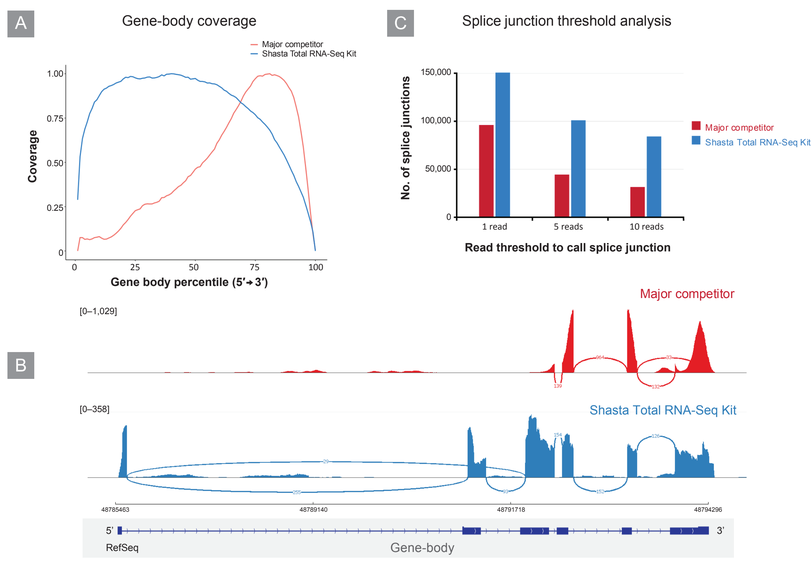

Figure 8. Detect more splice junctions with full gene body coverage. Panel A. Gene-body coverage of GENCODE protein-coding genes for pseudo-bulked K562 data from Shasta Total RNA-Seq Kit (1,300 cells) and a major end-counting-based, single-cell competitor (1,500 cells). **Panel B.** IGV genome browser view of pseudo-bulk K562 data for Shasta Total RNA-Seq Kit and the competitor at the *GATA1* locus. Higher signals corresponds to more aligned reads. Arcs represent detected splice junctions with the number indicating the number of reads supporting the splice junction. Read signal scale is located in the upper left corner of each track. **Panel C.** Number of annotated GENCODE splice junctions found with one, five, or 10 reads of support for pseudo-bulk data from Shasta Total RNA-Seq Kit and the competitor. ~20 x 10^6 read pairs were used for the comparison.

References

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-9 (2013).
Dou, J. et al. Single-nucleotide variant calling in single-cell sequencing data with Monopogen. *Nature Biotechnology* **42**, 803-12 (2024).
Sato, Y. et al. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nature Genetics* **45**, 860-7 (2013).

See more of what the Shasta technology can do:
takarabio.com/learnmore

800.662.2566
www.takarabio.com