Single-Cell RNA Data Analysis Workflow

RNA analysis from single cells using the Illumina Bio-Rad Single-Cell Sequencing Solution with the BaseSpace® SureCell™ RNA Single-Cell App.

Introduction

The behavior of complex biological systems is determined by the coordinated functions of individual cells. A key to understanding these processes is the ability to profile gene expression in a heterogeneous system. To deliver on the promise of single-cell biology, the Illumina Bio-Rad Single-Cell Sequencing Solution (1) combines innovative Droplet Digital™ technology (2) from Bio-Rad with library preparation, next-generation sequencing (NGS), and data analysis technologies from Illumina. This new platform provides a comprehensive, user-friendly workflow for single-cell RNA sequencing (RNA-Seq) that enables controlled experiments with multiple samples, treatment conditions, and time points.

The ddSEQ™ Single-Cell Isolator from Bio-Rad encapsulates and partitions single cells into subnanoliter droplets in a disposable cartridge. Cell lysis and barcoding occur inside individual droplets for tracking of individual cells throughout the workflow on an Illumina sequencing system. This enables transcriptome analysis of hundreds to tens of thousands of single cells in a single experiment.

The simple yet powerful BaseSpace SureCell RNA Single-Cell App provides data analysis options that can resolve heterogeneous cell populations and identify subpopulations of interest using gene expression profiles and data visualization tools. Designed for ease of use, the BaseSpace SureCell RNA Single-Cell App is a push-button solution that only requires the user to select initial inputs and parameters, launch the app, and receive automatically generated results. This tech note describes some of the steps in the analysis workflow, and how the quality of sequencing data is assessed.

Data Analysis Workflow

Upon completion of sequencing, BaseSpace Sequence Hub can be used to demultiplex basecall files (BCLs) to per-sample FASTQ files and perform secondary analysis with the BaseSpace SureCell RNA Single-Cell App. Due to the unique read structure of libraries prepared using the SureCell WTA 3′ Library Prep Kit, bcl2fastq 2.18 (or later) is used with an additional parameter set in the SampleSheet. Customers using the MiniSeq™ or NextSeq® Systems will use the PrepTab to set up their run, which allows the samplesheet to be created automatically by BaseSpace Sequence Hub. After demultiplexing is completed, the FASTQ files are available to analyze.

FASTQ Generation

BCL files are transferred automatically to BaseSpace Sequence Hub during a sequencing run. After a run completes and all BCL files are uploaded, an app called FASTQ Generation, which runs bcl2fastq 2.18, is automatically initiated.

- Bcl2fastq is used for demultiplexing basecall files into FASTQ files that are used in downstream secondary analysis.
- When using BaseSpace Prep Tab to set up a sequencing run, no additional files or parameters are needed to demultiplex.
- If using Illumina Experiment Manager (IEM) version 1.13 or newer, these settings are applied automatically to the SampleSheet when the SureCell WTA 3′ Library Prep Kit is selected.
- When running locally, bcl2fastq requires the following SampleSheet setting: "Read1UMILength,68"

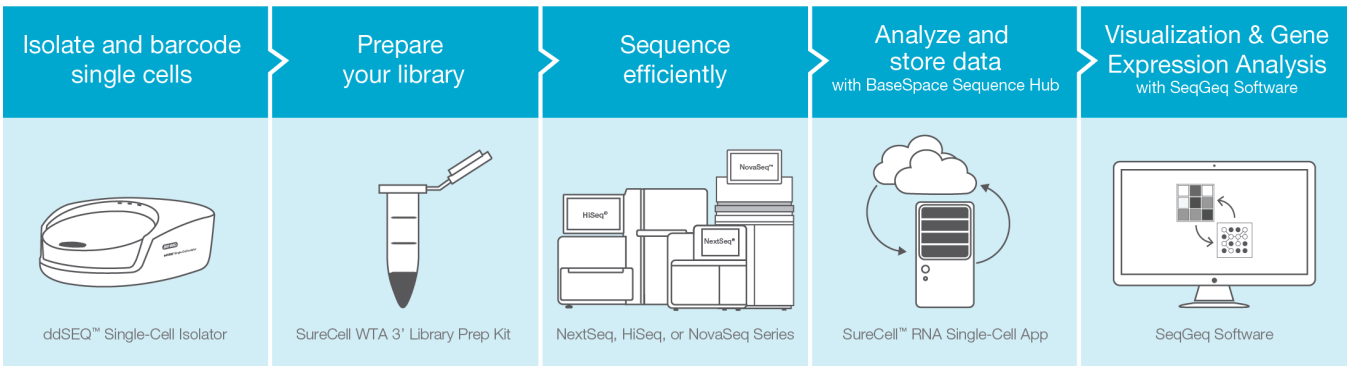

Fig. 1. Single-Cell RNA-Seq Workflow—The workflow integrates proven cell isolation using the Bio-Rad ddSEQ Single-Cell Isolator, followed by library preparations using the SureCell WTA 3′ Library Prep Kit with Nextera® technology, Illumina sequencing, and data analysis.

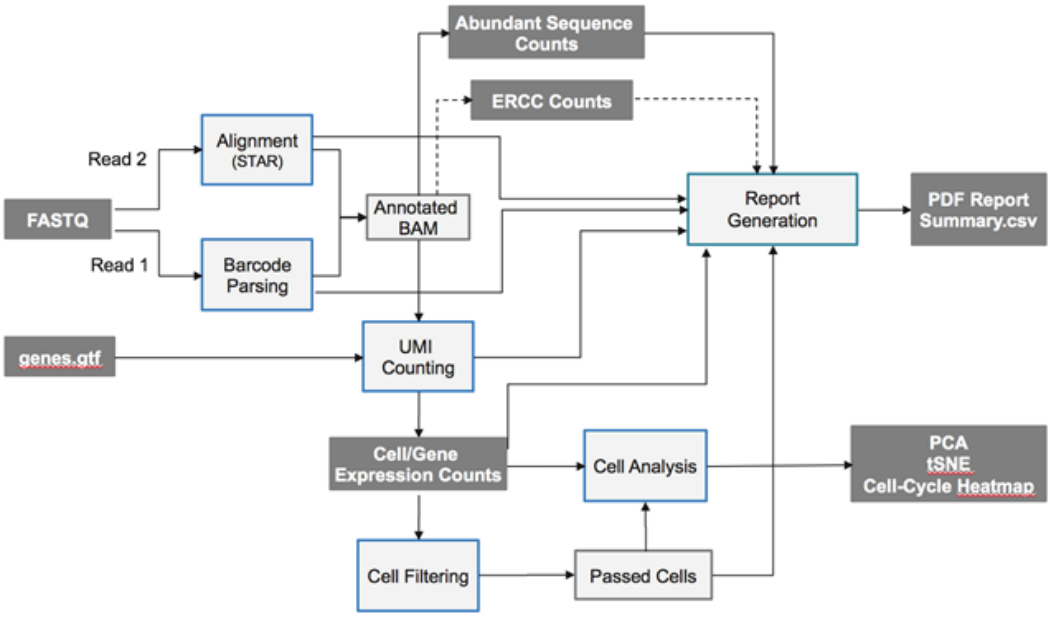

Fig. 2. Workflow Diagram—The major outputs from the workflow include PDF reports for each sample, global PCA and t-SNE plotes, and heatmaps.

Single-Cell RNA Analysis Workflow

The BaseSpace SureCell RNA Single-Cell App is designed to analyze samples prepared using the SureCell WTA 3′ Library Prep Kit. This app performs cell counting, gene counting, filtering, metric calculations, and report generation (Figure 2). The Single-Cell RNA workflow consists of the following major steps:

1. Align read 2 against reference genome using the STAR aligner.
   - Only read 2 is aligned, treating it as a single-end read
   - Alignment is done against a pregenerated STAR indexed reference genome
   - Index includes known splice-junctions and extra contigs
   - Output is one BAM file per sample
2. Identify the cell barcode and UMI for each read, based on the corresponding read 1, and add them to the BAM file as custom tags (XB and XU).
   - BAMs are indexed with samtools to generate .bam.bai files
3. Count UMIs for each gene in each cell files and associate statistics using Gene UMI counter.
   - Produce UMI, gene count files, and associated statistics
   - Read count for UMI and cell, read count for gene and cell
   - Duplicate UMIs are removed
   - Read counts are not duplicated
4. Identify good barcodes that likely correspond to single cells and which barcodes are empty beads or noise (cell filtering, or knee calling).
   - The purpose of knee calling is the find the transition from “good” cells (barcodes) to empty beads in the distribution of number of UMI counts per cell barcode (Figure 3)
   - The purpose of the UMI per cell plot is to indicate the total number of cells passing filter (Figure 4)
5. Calculate alignment cell & gene metrics.
   - Generate coverage statistics
   - Compute crosstalk metrics:
     - Doublets are a function of cell type and cell isolation methods can affect doublet formation
     - A mixed species experiment can be used to assess doublets
6. Report, with four major outputs from the analysis pipeline
   - Per sample pdf reports
   - Principle component analysis (PCA) plots (Figure 5)
   - T-Distributed Stochastic Neighbor Embedding (t-SNE) analysis (Figure 6)
   - A gene table listing total counts for each gene in each cell (Figure 7)

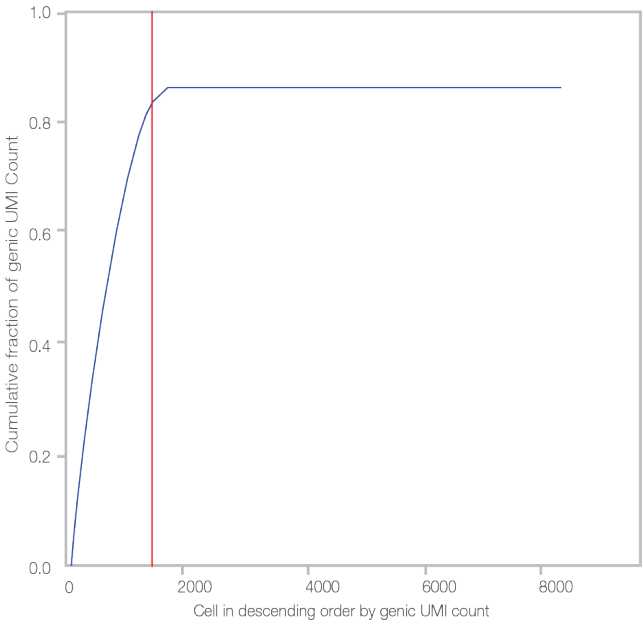

Fig. 3. Knee Calling—The x-axis shows number of cells in descending order by genic UMI count. The y-axis shows the cumulative fraction of genic UMI counts. Cells located to the left of the red line in the center of the bend in the plot (the knee) are marked as valid cells and pass filter. To the right of the red line are total observable barcodes.

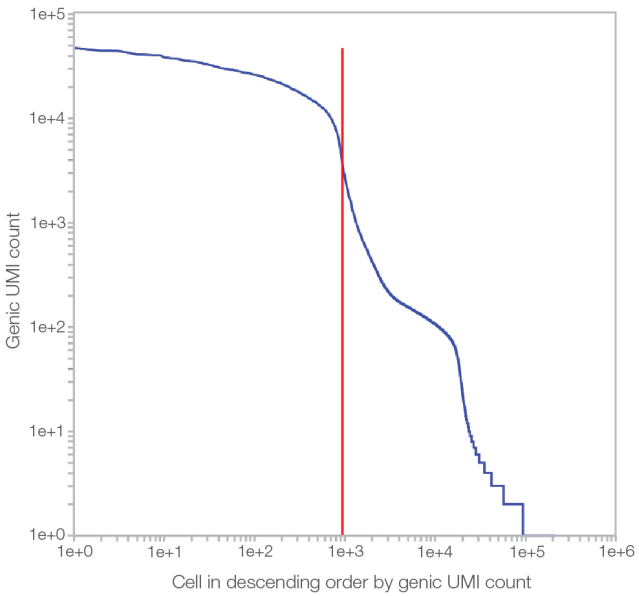

Fig. 4. UMI per Cell Plot—The UMI per cell plot indicates the total number of cells passing filter. The red line must pass through the first curve (knee). The defining features are the two distinct curves (knees) and the threshold (red line), which indicate the amount of valid cells detected in the sample. If a double knee plot contains more than two knees, the knee-calling algorithm will fail. In these cases, it is important to confirm the level of crosstalk in the sample by checking the number of doublet cells, and cells detected per species.

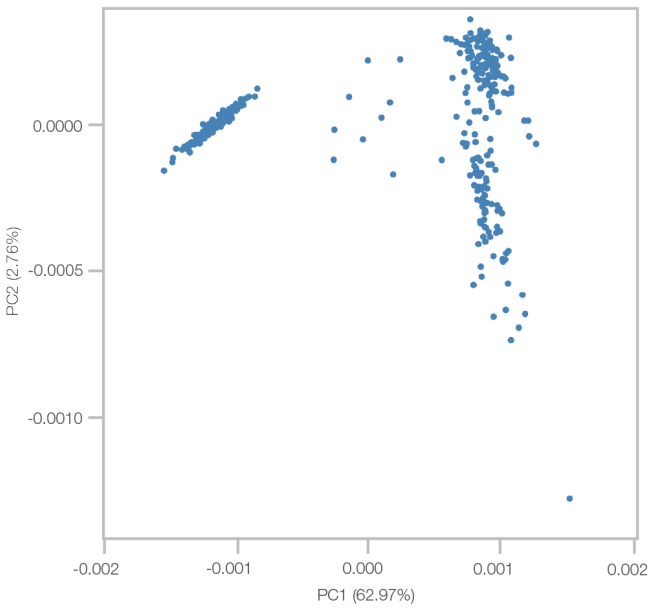

Fig. 5. Global PCA Plot—The principal components analysis (PCA) implementation consumes the gene expression (UMI count) per cell matrix normalized to UMI counts per cell, and excludes cells passing filter. PCA is performed using linear algebra calls and returns a projection of cells onto the top two principal components and the percentage of variance explained by each principal component.

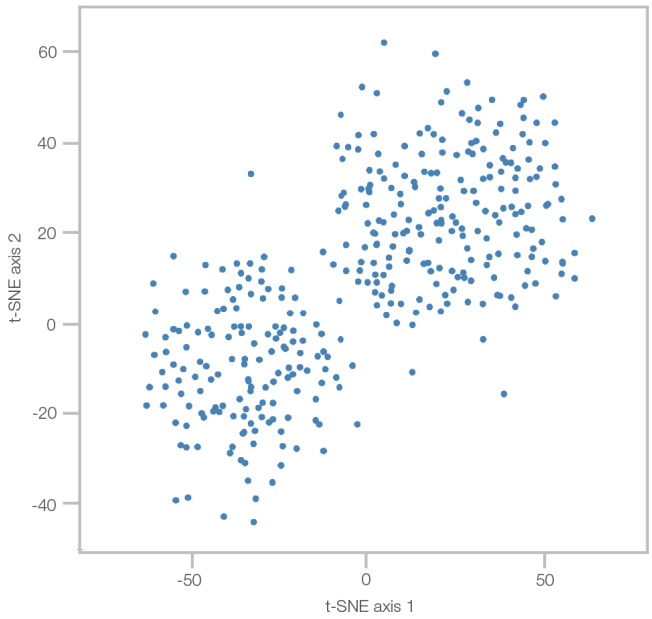

Fig. 6. The t-Distributed Stochastic Neighbor Embedding (t-SNE) Analysis—The t-SNE plot is a two-dimensional projection of cells illustrating potential clusters (populations) of neighboring cells with similar expression profiles. The t-SNE algorithm consumes the UMI counts per cell matrix normalized to total UMI counts per cell, excludes cells not passing filter, and computes all pairwise Euclidean distances between cells.

CellId	DF4	B3GALT6	FAM132A	UBE2J2	SCNN1D	ACAP3	MIR6726	PUSL1
atacttggattgaaagaa	8	1	0	4	0	0	0	6
caccacatacttttaaga	5	1	0	4	0	0	0	4
cttgaagtttcaacccaa	10	1	0	7	0	2	0	7
gagtgacggtcccaggag	3	2	0	1	0	1	0	2
tcgggaagtctgcaaccg	6	0	0	6	0	0	0	3
tcgcctaacgtgctgtgg	11	1	0	13	0	0	0	2
gcttgtatgaagggtgct	3	0	0	6	0	1	0	6
tcgcctgaaatacaagtc	1	1	0	3	0	0	0	1
gaagggtggcaggaaata	1	3	0	7	0	2	0	2
cgcatagcttgtcgaaag	4	0	0	6	0	0	0	8
cctctatccaaggccgtt	4	2	0	8	0	1	0	4
tcagtgacccaaggaggcc	10	1	0	4	0	1	0	9
ggccataattgggccgtt	2	0	0	6	0	0	0	2
attagtatccggagatgt	0	0	0	7	0	0	0	0
taccgacgaaaggtggtg	5	1	0	4	0	0	0	3
attagttaatagtgctaa	4	2	0	2	0	1	0	2
gtggtgcgcatatatttc	1	1	0	6	0	0	0	4
gctcccatacttgaggcc	2	1	0	8	0	0	0	2
tatttcgttaacggtgct	2	0	0	4	0	3	0	2
actgcaacccaatttggg	3	1	0	9	0	2	0	1

Fig. 7. Cell-Gene Table—An example table is shown that is provided in the automatically generated *counts.umiCounts.zip file. The table contains cell barcodes and observed UMI counts for each transcript for cells passing the knee threshold. The same information is provided for all cells, including those not passing the knee threshold. Rows include the cell barcodes and the columns include gene names. The counts associated with each gene for each cell are provided.

Merging Samples

Every port on the chip with a different sample index during library prep will become a different sample in BaseSpace Sequence Hub. For small datasets (< 2000 cells), combine the samples (FASTQ files) prior to analysis with the SureCell RNA Single-Cell App. For larger experiments, it is recommended that each sample be input into the BaseSpace App individually, and that the UMI output files be combined for downstream analysis (eg, in SeqGeq Software from FlowJo, LLC) using the app output for each sample.

Identifying a Good-Quality Dataset

It is important to understand and identify, based on the output of the app, whether or not the sequencing experiment is successful. Several key metrics can be used to asses the experimental outcome. All of these metrics should be judged relative to other datasets with similar cell-type and similar number of reads per cell.

- The # valid barcodes is provided in the first table of the analysis results. This metric shows how well read 1 performed, and is essential for identifying cells.
- The # aligned reads shows how well read 2 aligned to genes in the selected reference genome.
- The percent of reads aligned to unique genes shown in the report provides insight into the read utilization for the sample. This metric represents the number of reads with valid barcode that passed QC and aligned to unique genes with a mapping quality of > 11.
- Cells passing filter represents the number of cells with a UMI count above the knee threshold. This will exclude cells with few UMIs (low RNA content or low library efficiency).
- Median genes detected in cells passing filter + Median UMIs per cell passing filter.

Evaluating Crosstalk

Crosstalk represents the percentage of doublet cells, or the cells identified to contain UMIs assigned to >1 species, in a sample (Figure 8). In all cases where mixed samples are sequenced together, it is necessary to determine the level of crosstalk present in the sequenced sample and identify whether the amount of crosstalk is high or low.

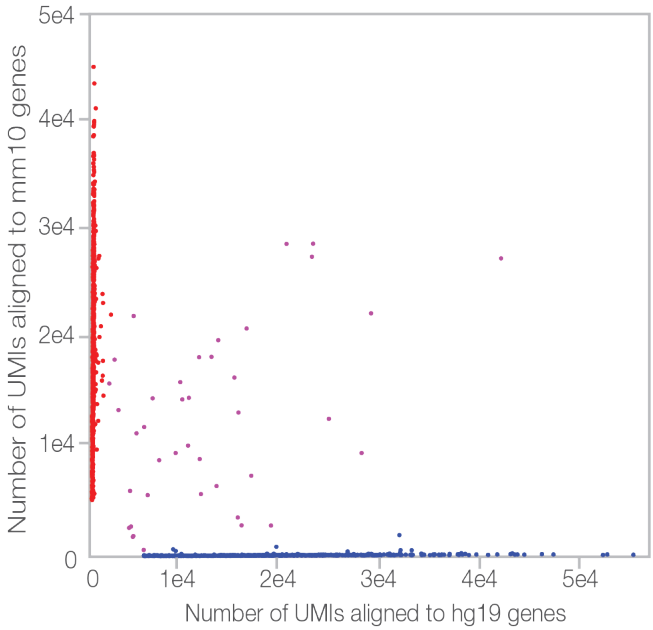

Fig. 8. Evaluating Crosstalk—Crosstalk represents the percentage of doublet cells, or the cells identified to contain UMIs assigned to >1 species, in a sample.

Learn More

Visit illumina.com/surecell to learn more about the Illumina Bio-Rad Single-Cell Sequencing Solution.

Visit illumina.com/basespace to learn more about BaseSpace Sequence Hub.

References

1. Illumina (2016). Bio-Rad SureCell WTA 3′ Prep Kit for the ddSEQ System. www.illumina.com/surecell, accessed January 12, 2007.
2. Hindson BJ et al. (2011). High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem 83, 8,604–8,610.

Pub. No. 1070-2017-001-B
Bio-Rad Bulletin 6939 Ver A 17-0609 04.17