Introduction to Single Cell RNA-Seq Data Analysis

Jenny Wu
Director of Bioinformatics
Genomics Research and Technology Hub
Chao Family Comprehensive Cancer Center
UC Irvine
Outline

• Why single cell
• Single cell RNA-Seq workflow
• Experimental Design
• Data analysis pipeline and applications
• Single cell RNA-Seq data analysis
  ✓ Pre-processing
  ✓ Downstream Analysis
• Pre-processing of 10x Chromium data
• Pre-processing of Parsebio data
Diverse Cell Types, States and Interactions

Skin epithelium

Brain meninges

Blood vessels

Small intestine

Liver cirrhosis

Breast cancer
Advances in Single Cell Technology

Svensson et al., 2018
Single Cell RNA-Seq Applications

• Explore which cell types are present in a tissue
• Identify unknown/rare cell types or states
• Elucidate the changes in gene expression during differentiation processes or across time or states
• Identify genes that are differentially expressed in particular cell types between conditions
• Explore changes in expression among a cell type while incorporating spatial, regulatory, and/or protein information
Experimental Design

• Choose the platforms/protocols based on the biological question you wish to address
• How many samples and how many cells
  • https://satijalab.org/howmanycells/
• Single cell vs Single nucleus
• Isolate RNA and prepare libraries at same time for all samples or alternate sample groups to avoid batch confounding
• Do not confound sample groups by sex, age, or batch
Single Cell RNA-Seq Workflow

1. Pre-processing
2. Downstream Analysis

Pre-Processing of Single Cell Raw Data

Lafzi et al, 2018
Downstream Analysis

Luecken et al, 2019
scRNA-seq Data Analysis Tools

- **Pre-Processing**
  - Quality Control/filtering: FASTQC, RSeQC, QoRTs, Qualimap2, sinQC
  - Read Alignment: STAR, HISAT, RSEM, Kallisto, etc.
  - Normalization: UMI, ERCC, GRM, Census etc.
  - Confounding factor removal: scLVM, OEFinder, cgCorrect

- **Dimension Reduction**
  - PCA, ICA, t-SNE, ZIFA, NMF, MDS, probability PCA

- **Clustering and Visualization**
  - Generic Clustering: Heatmap/Louvian, Leiden etc.

- **Generic Clustering**
  - Heatmap/Louvian, Leiden etc.

- **Differential Expression**
  - scDE, MAST, Monocle3, scDD, MIMOSA

- **Cell-cell communication Network**
  - CellChat, CCinx, CellphoneDB

- **Cell and Gene Network Modeling**
  - WGCNA, SCENIC, IRIS3

- **Lineage inferences/pseudotime ordering**
  - SLICE, Dpath, Monocle3, SCUBA, Oscope Wanderlust

**R packages**:
- Seurat4, Monocle3, Pagoda, scater, scran
- Integrated analysis: Scell, fast Project

**CellRanger**
- bustools, Parsebio pipeline
10X Chromium Data and CellRanger

```
$ cd /home/jdoe/runs
$ cellranger count --id=sample345
    --transcriptome=/opt/refdata-cellranger-GRCh38-1.2.0
    --fastqs=/home/jdoe/runs/HAWT7ADXX/outs/fastq_path
    --sample=mysample
    --expect-cells=1000
```
CellRanger Output: Summary

Estimated Number of Cells
8,627

Mean Reads per Cell
16,889

Median Genes per Cell
2,136

Sequencing
- Number of Reads: 145,709,844
- Valid Barcodes: 98.5%
- Sequencing Saturation: 33.3%
- Q30 Bases in Barcode: 96.8%
- Q30 Bases in RNA Read: 76.4%
- Q30 Bases in UMI: 96.9%

Mapping
- Reads Mapped to Genome: 91.6%
- Reads Mapped Confidently to Genome: 87.1%
- Reads Mapped Confidently to Intergenic Regions: 2.6%
- Reads Mapped Confidently to Intronic Regions: 11.5%
- Reads Mapped Confidently to Exonic Regions: 72.9%
- Reads Mapped Confidently to Transcriptome: 69.0%
- Reads Mapped Antisense to Gene: 1.1%

Sample
- Name: plus_Jan9
- Description: mm10_0.5GF_P_nck
- Chemistry: Single Cell 3' v2
- Cell Ranger Version: 2.1.0
CellRanger Output: Analysis
Parsebio (SPLit-Seq)
Parsebio Raw Reads

**Read 1 - transcript sequence**

@NB551368:195:HCVJVBGXJ:1:11104:5453:1061 1:N:0:ACTTGA
AAGCCNTGTTAATCAACGCAAGTGATNNGGAGATGTCAATNCCTATGACCCTAATGTCATAAAATTGACAGG
+
A/AAA#E#E/EE/EE/EEEEEEEEEEEEA#EE///6/EEE/6A#/<<///E/<<E//E<<E/AA<<A/AA<<E<<E<<<<

**Read 2 - barcodes and degenerate sequence**

@NB551368:195:HCVJVBGXJ:1:11104:13360:1088 2:N:0:ACTTGA
TCAGAGGTATTAGGCTAACGTTTGTGATCTGACGTACGACTGTCTGTCAATCAGTCTGAGACTTGGGGGGCTG
+
AAAAAEEEEEEEEEEEEAAAA/<://EAEAEAEAEAEAEAE/<AE6<<AA/<<AAAE/<AAAE<<<<AAAEAEAEAEAE<EA/<E<<<<

--- Round 3 barcode ---
--- 30bp Linker ---
--- Round 2 barcode ---
--- 22bp Linker ---
--- Round 1 barcode ---

```
singularity exec /difs3a/singularity_containers/rcic/parsebio.simg split-pipe 
--mode all
--kit WT
--genomic_dir ${GPATH}
--fq1 ${IPATH}/Sublibrary1_S1_R1_001.fastq.gz
--fq2 ${IPATH}/Sublibrary1_S1_R2_001.fastq.gz
--output_dir ${OPATH}
--samp_list ${SPATH}
```
Parsebio pre-processing Output: Summary

Sample: S4    Wells: B1-B4

- Estimated Number of Cells: 4,795
- mm10 Number of Cells Detected: 4,795
- mm10 Median Transcripts/Cell: 10,058
- mm10 Median Genes/Cell: 3,377
- Mean Reads/Cell: 41,891
- Number of Reads: 200,668,545
- Sequencing Saturation: 0.504
- BC1 (RT) >Q30: 0.802
- BC2 >Q30: 0.927
- BC3 >Q30: 0.939
- cDNA >Q30: 0.927

Identified Cells

![Graph showing identified cells with logscale y-axis and barcodes on the x-axis.]

- Background
- Cells
Parsebio pre-processing Output: Analysis

Differentially Expressed Genes

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Gene name</th>
<th>Score</th>
<th>Log2 FC</th>
<th>Pval adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A830012C17Rik</td>
<td>33.2</td>
<td>3.4</td>
<td>1.5E-27</td>
</tr>
<tr>
<td>1</td>
<td>Adora2a</td>
<td>26.4</td>
<td>3.1</td>
<td>8.1E-16</td>
</tr>
<tr>
<td>1</td>
<td>Gpr6</td>
<td>23.8</td>
<td>2.6</td>
<td>8.1E-02</td>
</tr>
<tr>
<td>1</td>
<td>Drd2</td>
<td>21.0</td>
<td>3.4</td>
<td>0.0E+00</td>
</tr>
<tr>
<td>1</td>
<td>Penk</td>
<td>19.2</td>
<td>2.9</td>
<td>3.3E-11</td>
</tr>
<tr>
<td>1</td>
<td>Oprd1</td>
<td>18.0</td>
<td>2.9</td>
<td>1.1E-56</td>
</tr>
<tr>
<td>1</td>
<td>Gm30313</td>
<td>14.1</td>
<td>2.6</td>
<td>3.4E-11</td>
</tr>
<tr>
<td>1</td>
<td>Nl5e</td>
<td>13.7</td>
<td>2.5</td>
<td>1.5E-35</td>
</tr>
<tr>
<td>1</td>
<td>Gm39043</td>
<td>12.1</td>
<td>2.4</td>
<td>6.0E-05</td>
</tr>
<tr>
<td>1</td>
<td>Gm37711</td>
<td>10.3</td>
<td>2.1</td>
<td>4.2E-01</td>
</tr>
<tr>
<td>1</td>
<td>Gm19410</td>
<td>10.0</td>
<td>2.3</td>
<td>2.0E-29</td>
</tr>
<tr>
<td>1</td>
<td>Nell1</td>
<td>8.9</td>
<td>2.3</td>
<td>3.4E-238</td>
</tr>
<tr>
<td>1</td>
<td>B230110G15Rik</td>
<td>7.8</td>
<td>2.1</td>
<td>1.8E-15</td>
</tr>
</tbody>
</table>
Alternative Tools for pre-processing

Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data

Dongze He, Mohsen Zakeri, Hiran Sarkar, Charlotte Sone and Rob Patro

The rapid growth of high-throughput single-cell and single-nucleus RNA-sequencing has produced a wealth of data over the past few years. The size, volume and diversity of data make it challenging to quickly and efficiently process and extract meaningful information. Alevin-fry is a new tool that can be used to quantify scRNA-seq data, and also how this data can be efficiently extracted from normal gene expression count matrices.

Modular, efficient and constant-memory single-cell RNA-seq preprocessing

Páll Melsted, A. Sina Booshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi (Joseph) Min, Eduardo da Veiga Beltrame, Kristján Eldjarn Hjörleifsson, Jase Gehring and Lior Pachter

We describe a workflow for preprocessing of single-cell RNA-sequencing data that balances efficiency and accuracy. Our workflow is based on the kallisto and bustools programs, and is near-optimal in speed with a constant memory requirement providing scalability for arbitrarily large datasets. The workflow is modular, and we demonstrate its flexibility by showing how it can be used for RNA velocity analyses.
Downstream Analysis

Luecken et al, 2019
MatrixMarket Format for Gene Count Matrix

Sparse Gene Count Matrix

<table>
<thead>
<tr>
<th>gene_id, gene_name, genome</th>
<th>gene_id, gene_name, genome</th>
<th>AAACCCAGTTAGCAG-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSFCAG00000000001, INTS6L, felis</td>
<td>ENSFCAG00000000007, HMGR, felis</td>
<td>AAACCCAGTCACCTTCTC-1</td>
</tr>
<tr>
<td>ENSFCAG00000000015, CEP192, felis</td>
<td>ENSFCAG00000000022, RASGRF1, felis</td>
<td>AAACGAAAGAGATGATCTC-1</td>
</tr>
<tr>
<td>ENSFCAG00000000023, GPR39, felis</td>
<td>ENSFCAG00000000024, LYPD1, felis</td>
<td>AAACGAAACCCCTG-1</td>
</tr>
<tr>
<td>ENSFCAG00000000028, RCN3, felis</td>
<td>ENSFCAG00000000029, APOO, felis</td>
<td>AAACGAATCTGAATGC-1</td>
</tr>
<tr>
<td>ENSFCAG00000000030, CXXH2orf58, felis</td>
<td>ENSFCAG00000000031, CB1Eorf19, felis</td>
<td>AAGCCTAGCACCTG-1</td>
</tr>
<tr>
<td>ENSFCAG00000000032, RELL1, felis</td>
<td>ENSFCAG00000000034, DRC1, felis</td>
<td>AAACGCTAGCTTATGC-1</td>
</tr>
<tr>
<td>ENSFCAG00000000035, OTOF, felis</td>
<td>ENSFCAG00000000036, CA3H2orf70, felis</td>
<td>AAACGCTAGCTTATGC-1</td>
</tr>
<tr>
<td>ENSFCAG00000000039, KAT5, felis</td>
<td>ENSFCAG00000000043, PPARGC1A, felis</td>
<td>AAAGAACGAGTGGCC-1</td>
</tr>
<tr>
<td>ENSFCAG00000000044, LRP6, felis</td>
<td>ENSFCAG00000000049, SYNE1, felis</td>
<td>AAAGAACGAGTGGCC-1</td>
</tr>
<tr>
<td>ENSFCAG00000000050, TARDBP, felis</td>
<td>ENSFCAG00000000051, MASP2, felis</td>
<td>AAAGAACGAGTGGCC-1</td>
</tr>
</tbody>
</table>

Cell Info

| AAACCCAGTTAGCAG-1 |
| AAACCCAGTCACCTTCTC-1 |
| AAACGAAAGAGATGATCTC-1 |
| AAACGAAACCCCTG-1 |
| AAACGAATCTGAATGC-1 |
| AAGCCTAGCACCTG-1 |
| AAACGCTAGCTTATGC-1 |
| AAAGAACGAGTGGCC-1 |
| AAAGAACGAGTGGCC-1 |
| AAAGAACGAGTGGCC-1 |
Limitations of Single Cell Data

- Low capture rate, low depth of sequencing per cell, 3’ only
- High biological and technical variability across cells/samples
- Missing spatial information
Useful Links

• CellRanger manual
  https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

• Biojhub3 on HPC3:

• Workshop data:
  /dfs6/pub/ucightf/workshop/

• Seurat manual
  https://satijalab.org/seurat/

• 10x Cloud
  https://www.10xgenomics.com/products/cloud-analysis