Interactive scRNA-Seq analysis with the Single Cell Toolkit (SCTK)

Evan Johnson and Josh Campbell
Division of Computational Biomedicine
Boston University School of Medicine

May 1, 2020
Single Cell Toolkit (SCTK)

- Standard R package w/Shiny toolkit
- Interacts and operates on an SCE (SingleCellExperiment) object
- SCE object in/out of Shiny any stage
- Great for common tasks:
  - Interactive clustering/visualization
  - Differential gene/pathway analysis
- (Also works for bulk RNA-seq analysis)
“inDrops” is custom microfluidics device that can process large numbers of cells with high capture rates.

Goals of parent R33:

- **Aim 1.** Optimize inDrops for microscopic samples and fixed cells.
- **Aim 2.** Optimize inDrops for low-cost, high-throughput, high sensitivity targeted transcriptomics.
- **Aim 3.** Integrate single cell genomics with histopathology.
Aims of ITCR-IMAT collaboration

**IMAT (HMS - Klein)**
Aim 1: Develop a Total-seq protocol for the inDrop system.

Aim 2: Develop single cell ATAC-seq for inDrop.

**ITCR (BU – Johnson/Campbell)**
Aim 1: Develop computational modules for analysis and display of single-cell Total/CITE-seq and ATAC-seq data generated from tumor specimens.

Aim 2: Develop computational modules for comprehensive assessment and correction of batch effects or sample-specific effects across tumor specimens.
Aim 1: Develop computational modules for analysis and display of single-cell CITE-seq and ATAC-seq data generated from tumors.

Total-Seq/CITE-seq antibody derived tags (ADTs) can be used to measure protein levels at single cell resolution.

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a technique used in molecular biology to assess genome-wide chromatin accessibility.
Comprehensive scRNA-seq QC pipeline for the Human Tumor Atlas Network (HTAN)

1. **Raw Data**
   - Single cell/nuc RNA-seq
     - 10X
     - inDrops
     - CEL-seq2
     - Drop-seq
     - SMART-seq2

2. **Preprocessing**
   - CellRanger
   - STARsolo
   - BUStools
   - dropEST
   - HCA Optimus
   - SEQC
   - **SCTK**
     - Import into R
     - SingleCellExperiment
     - MultiAssayExperiment

3. **Quality control**
   - Standard metrics
   - Doublets
   - Ambient RNA
   - Empty Drop Detection
   - Interactive visualization and analysis

**SEQC**
Comprehensive scRNA-seq QC pipeline for the Human Tumor Atlas Network (HTAN)
Expansion to include import functions for Total-seq/CITE-seq or scATAC-seq data in R data containers.

1. Raw Data
   - Single cell/nuc RNA-seq
     - 10X
     - inDrops
     - CEL-seq2
     - Drop-seq
     - SMART-seq2
   - CITE-seq/Total-seq
     - 10X
     - inDrops
   - scATAC-Seq
     - inDrops
     - 10X

2. Preprocessing
   - CellRanger
   - STARsolo
   - BUSTools
   - dropEST
   - HCA Optimus
   - SEQC

3. Quality control
   - Standard metrics
   - Doublets
   - Ambient RNA
   - Empty Drop Detection

Interactive visualization and analysis

Import into R

SCTK

SingleCellExperiment
MultiAssayExperiment

SCTK
Developing novel statistical approaches for joint clustering of ADT (protein) and scRNA-seq data.

https://github.com/campbio/celda/
Develop computational modules for analysis and display of single-cell ATAC-seq data generated from tumors.

**Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types**

Rongxin Fang, Sebastian Preissl, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, Eran A. Mukamel, Yanxiao Zhang, M. Margarita Behrens, Joseph Ecker, Bing Ren

![Barcode Selection](image1)

**A** Barcode Selection

![Selection of Principal Components](image2)

**B** Selection of Principal Components

![Visualization of clusters](image3)

**C** Visualization of clusters

![Peak visualization](image4)

**D** Peak visualization
A benchmark of batch-effect correction methods for single-cell RNA sequencing data

Aim 2: Develop computational modules for comprehensive assessment and correction of batch effects or sample-specific effects across tumors.

Batch correction tools implemented in SCTK:
1) ComBat/ComBat-Seq
2) Seurat3 Integration
3) Harmony
4) scMerge
5) FastMNN
6) MNncorrect
7) BBKNN
8) LIGER
9) scGEN
10) Scanorama
11) ZinB-wave

Tran et al, Genome Biology, 2020
Mean-variance dependence in RNA-seq counts:
- Over-dispersion (variance > mean)
- Genes with smaller counts tend to have larger variance

Negative Binomial (NB):

\[ y \sim NB(\mu, \phi) \]

- Negative Binomial regression used in edgeR & DESeq2
- Variance is a function of mean
  \[ \text{var}(y) = \mu + \phi \mu^2 \]

Law, Charity W., et al. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." Genome biology 2014
ComBat-Seq algorithm

1. Fit Negative Binomial regression model
2. Obtain batch effect estimates

1. Compute "batch-free" distribution
2. Quantile adjustment
ComBat-Seq algorithm: Model

*Negative Binomial regression*

Gene-wise model: for a certain gene \( g \), count in sample \( j \) from batch \( i \)

\[ y_{gij} \sim NB(\mu_{gij}, \phi_{gi}) \]

\[
\log \mu_{gij} = \alpha_g + X_j\beta_g + \gamma_{gi} + \log N_j \\
Var(y_{gij}) = \mu_{gij} + \phi_{gi}\mu_{gij}^2
\]

Decompose scaled counts into 3 components

\[
\begin{align*}
\alpha_g & \quad \text{Average level for gene } g \text{ (in “negative” samples)} \\
X_j\beta_g & \quad \text{Biological condition of sample } j \\
\gamma_{gi} & \quad \text{Mean batch effect} \\
\phi_{gi} & \quad \text{Variance batch effect}
\end{align*}
\]
ComBat-Seq algorithm: Adjust

Adjust the data

- Calculate parameters for “batch-free” distribution: $y_{gj}^* \sim NB(\mu_{gj}^*, \phi_g^*)$

\[
\log \mu_{gj}^* = \log \hat{\mu}_{gij} - \hat{\gamma}_{gi}
\]

\[
\phi_g^* = \frac{1}{N_{batch}} \sum_i \hat{\phi}_{gi}
\]

- Map quantiles from empirical distribution to the batch-free distribution
ComBat-Seq algorithm: Adjust

Adjust the data

Original count matrix

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>G2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G3</td>
<td>112</td>
<td>11</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted count matrix

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>G2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G3</td>
<td>60</td>
<td>47</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Empirical distribution of original counts:

\[ y_{gij} \sim NB(\hat{\mu}_{gij}, \hat{\phi}_{gi}) \]

Batch-free distribution for adjusted counts:

\[ y_{gj}^* \sim NB(\mu_{gj}^*, \phi_g^*) \]
ComBat Methods for scRNA-seq Analysis

**ComBat-Cell-Seq**
- Include cell types in ComBat design model

**ComBat-Seq**
- Okay if combining batches with ‘balanced’ cell types

**ComBat-SVA-Seq**
- Use SVA to identify surrogate cell-type variability

**Flowchart**
1. Are cell types known? (or discoverable)
   - Yes: ComBat-Cell-Seq
   - No: ComBat-Seq
2. Are cell types expected to be balanced?
   - Yes: ComBat-Seq
   - No: ComBat-SVA-Seq
ComBat-Seq for Balanced Designs (ComBat-Cell-Seq, ComBat-Seq)
ComBat-Seq for Balanced Designs (ComBat-Cell-Seq, ComBat-Seq)
ComBat-Seq for Unbalanced Designs (ComBat-Cell-Seq, ComBat-SVA-Seq)
ComBat-Seq for Unbalanced Designs (ComBat-Cell-Seq, ComBat-SVA-Seq)
ComBat-seq Summary

For balanced designs:
- **ComBat-Seq, ComBat-Cell-Seq, ComBat-SVA-Seq** all work well!

For unbalanced designs:
- **ComBat-Seq**: May remove cell-type specific variation
- **Combat-Cell-Seq**: Performs extremely well
- **Combat-SVA-Seq**: Not quite as good as Combat-Cell-Seq, but performs well in both simulated and real-data examples