Open Workflows to Enable and Understand Single-Cell RNA-Seq Data

Nathan Salomonis
Department of Biomedical Informatics, CCHMC
Goals for this session

1. Discuss what is needed to enable scientists to analyze and interpret single-cell genomics data.

2. Overview the technologies, best practices and infrastructure needed to accomplish the emerging questions afforded by these new methods.

3. Provide workflows which you and your users can apply on their own or in collaboration.
About Me

• Computational genomicist – our lab focused on software development, outreach and analysis.

• We automate initial single-cell RNA-Seq analyses for folks at our institution at no cost, but encourage labs to perform downstream analyses on their own (AltAnalyze, ICGS, cellHarmony, DoubletDecon, Schrödinger).
Evolution of Single-Cell Genomics

- **2010**: 16 cells within 6 days
- **2012**: 96 cells within 1 day
- **2014**: Hundreds of cells
- **2015**: 10k cells within 1 day
- **2016**: CRISPR Screens in 200k cells
- **2017**: >1 million neurons

Publications per year:
- **2016**: 270+
Applications

- 100-15,000 cells measured per experiment.
- Unique molecular indexes or UMI, 3’ gene biased.
- Typically low depth-seq (<200k reads) and lower confidence gene exp.
- 1-10% doublets (QC challenging to detect).
- No splicing, polyA containing genes only.

http://core-genomics.blogspot.com/2016/07/10x-genomics-single-cell-3mrna-seq.html
Emerging Technologies for Single-Cell Genomics

- **Single-Cell RNA-Seq**
  - Smarter low throughput (Fluidigm)
  - High-throughput 3’ or 5’ biased (10x, Drop-Seq, Microwell-Seq)

- **Single-Cell ATAC-Seq**
  - Low resolution, TF activities inferred

- **CITE-Seq, BCR/TCR Single-Cell Analysis**
  - Couple scRNA-Seq with the detection of specific RNAs or labelled proteins
Which Platforms to Choose and Which Give Comparable Results

E14 Kidney Dissociation

10X Genomics

DropSeq

Fluidigm

Supervised Classification

Unsupervised Analysis

Interactive Technology Comparison

Cellular Interactions

Population Description

Magela et al. 2017 Developmental Biology
Different Technologies Provide Quantitative and Qualitatively Different Results

Magela et al. 2017 Developmental Biology
Not all scRNA-Seq Analyses Are the Same

Each experiment has its own technical challenges and needed bioinformatics expertise.

- **Defining notably different cell states.**
  - Numerous well-established automated workflows exist.

- **Defining subtle heterogeneity or transitional cell states.**
  - Testing of many tools likely required (no perfect methods).

- **Comparing different scRNA-Seq Datasets**
  - Requires prior bioinformatics experience and novel tools.

- **Evaluating BCR/TCR subsets**
  - Requires immunology and informatics expertise.

- **CRISPR Screens**
  - Significant bioinformatics and investment expertise required.
Developing Reliable Methods to Interface Multi-Omics Single-Cell Data

Megakaryocyte

HSCP-1

Granulocyte

Mpl

Map4k4
How Can Core Facilities Aid in the Processing and Interpretation of Massive Datasets?

• Communicating Quality Control

• Establishing Workflows that Automate Primary Analyses

• Partnering with Computational Genomics Collaborators/Vendors

• Providing Documentation on User-Friendly Workflows Users Can Run (e.g., AltAnalyze)
Analysis of 10x Genomics Datasets

• Ideally requires a cluster computing environment.

• Software from 10x Genomics is open source and is run on the command-line (Cell Ranger).

• Works with input bcl2 sequence files directly from the sequencing core.

• Runs in 1-2 days, requiring 200GB.

• Provides an interactive viewer to explore their results (Loupe Browser).

• Starting point for deeper analysis with various tool kits.
Analysis of 10x Genomics Datasets

Estimated Number of Cells
2,295

Mean Reads per Cell
41,916
Median Genes per Cell
2,556

Sequencing
Number of Reads
96,199,414
Valid Barcodes
97.7%
Reads Mapped Confidently to Transcriptome
67.0%
Reads Mapped Confidently to Exonic Regions
71.1%
Reads Mapped Confidently to Intrinsic Regions
16.4%
Reads Mapped Confidently to Intergenic Regions
3.3%
Sequencing Saturation
48.5%
Q30 Bases in Barcode
99.2%
Q30 Bases in RNA Read
95.8%
Q30 Bases in Sample Index
98.9%
Q30 Bases in UMI
99.2%

Sample
Name
10x_WT
Description

Transcriptome
mm10
Chemistry
Single Cell 3' v2
Cell Ranger Version
1.3.0
Analysis of 10x Genomics Datasets

Cell Ranger - 10x WT

SUMMARY ANALYSIS

Clustering Type: Graph-based

Top Genes By Cluster (Log2 fold-change, p-value)

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene name</th>
<th>Cluster 1 L2FC</th>
<th>p-value</th>
<th>Cluster 2 L2FC</th>
<th>p-value</th>
<th>Cluster 3 L2FC</th>
<th>p-value</th>
<th>Cluster 4 L2FC</th>
<th>p-value</th>
<th>Cluster 5 L2FC</th>
<th>p-value</th>
<th>Cluster 6 L2FC</th>
<th>p-value</th>
<th>Cluster 7 L2FC</th>
<th>p-value</th>
<th>Cluster L2FC</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSMUSG00000025902</td>
<td>Sox17</td>
<td>8.08</td>
<td>5e-67</td>
<td>-5.91</td>
<td>2e-10</td>
<td>-5.10</td>
<td>1e-05</td>
<td>-5.08</td>
<td>2e-09</td>
<td>-6.00</td>
<td>4e-08</td>
<td>-5.74</td>
<td>1e-06</td>
<td>-5.19</td>
<td>8e-01</td>
<td>-5.60</td>
<td>1e-02</td>
</tr>
<tr>
<td>ENSMUSG00000094686</td>
<td>Cdc21a</td>
<td>7.74</td>
<td>1e-06</td>
<td>-5.50</td>
<td>1e+00</td>
<td>-5.11</td>
<td>1e+00</td>
<td>-6.42</td>
<td>1e+00</td>
<td>-5.08</td>
<td>1e+00</td>
<td>-5.08</td>
<td>1e+00</td>
<td>-5.75</td>
<td>1e+00</td>
<td>-5.19</td>
<td>1e+00</td>
</tr>
<tr>
<td>ENSMUSG00000050493</td>
<td>Fam167b</td>
<td>7.38</td>
<td>3e-76</td>
<td>-5.10</td>
<td>3e-11</td>
<td>-4.03</td>
<td>2e-05</td>
<td>-5.76</td>
<td>1e-12</td>
<td>-5.99</td>
<td>5e-10</td>
<td>-5.87</td>
<td>3e-08</td>
<td>-5.13</td>
<td>3e-01</td>
<td>-3.33</td>
<td>3e-01</td>
</tr>
<tr>
<td>ENSMUSG00000054335</td>
<td>Gimap4</td>
<td>7.07</td>
<td>2e-92</td>
<td>-4.71</td>
<td>8e-13</td>
<td>-4.73</td>
<td>3e-08</td>
<td>-6.02</td>
<td>1e-16</td>
<td>-6.69</td>
<td>3e-14</td>
<td>-4.95</td>
<td>3e-09</td>
<td>-3.60</td>
<td>7e-01</td>
<td>-4.01</td>
<td>7e-01</td>
</tr>
<tr>
<td>ENSMUSG00000062515</td>
<td>Fabp4</td>
<td>7.04</td>
<td>1e-96</td>
<td>-4.99</td>
<td>6e-14</td>
<td>-4.02</td>
<td>2e-09</td>
<td>-5.38</td>
<td>1e-16</td>
<td>-6.47</td>
<td>9e-16</td>
<td>-6.03</td>
<td>3e-12</td>
<td>-2.97</td>
<td>1e+00</td>
<td>-4.39</td>
<td>1e+00</td>
</tr>
<tr>
<td>ENSMUSG000000623960</td>
<td>Kdr</td>
<td>6.87</td>
<td>2e-80</td>
<td>-4.42</td>
<td>3e-11</td>
<td>-4.03</td>
<td>4e-06</td>
<td>-6.38</td>
<td>1e-15</td>
<td>-6.58</td>
<td>3e-12</td>
<td>-6.87</td>
<td>2e-09</td>
<td>-2.12</td>
<td>1e+00</td>
<td>-6.13</td>
<td>1e+00</td>
</tr>
</tbody>
</table>
Analysis of 10x Genomics Datasets
Seurat Analysis Workflow

- Fast
- Handles >100k cells
- Command-Line
- Well-supported
- Many options
- Can visualize the expression of genes on the t-SNE plot.

Using Single-Cell Genomics to Understand the Cellular Origins of Disease

Healthy

Neutropenia

Bone Marrow
Single-Cell Progenitors
Seurat Canonical Correlation Analysis

- Can align cells independent of batch
- Can align cells between species
- Requires similar cell types to align.

Cell Classification and Molecular Impact Analysis with cellHarmony

Find Your Best Match
Reference ICGS States

Find Out How You’re Different
ICGS States Assigned States

Differential Population % Differential Expression

DePasquale, Salomonis in preparation
Defining Mis-Expressed Genes with CellHarmony

AltAnalyze
Cell States
(wild-type)

CellHarmony
Mapped
(mutant)

Row = gene
Column = single cell
Height = expression amplitude

Muench, Salomonis, Grimes in preparation
AltAnalyze: Comprehensive and Intuitive Application for Genome Research

- Among the most easy-to-use freely available bioinformatics tools.
- Extensive documentation, blogs, tutorials, online help.

http://www.altanalyze.org

- Over 18,000 downloads
- >200 peer-reviewed citations
- Available for Mac, Windows, Linux and source-code.
Nathan Salomonis
Cincinnati Children’s Hospital
email: Nathan.Salomonis@cchmc.org
email: altanalyze@gmail.com

website: http://altanalyze.org

tutorials: YouTube AltAnalyze