Introduction to Single-cell RNA-Seq
Make sure your workshop provided computer is connected to the “Broad” or “Broad Internal” wireless network.

**Please do not** connect your personal items (laptop, phone, etc.) to these wireless networks; it will tax the wireless system and make the workshop less effective.

The password for computers is “password”. 
Introduction to single-cell RNA-Seq

Timothy Tickle
Brian Haas
Asma Bankapur
We Know Tissues are Heterogeneous

- **Adipose**
  - Fat
  - Connective tissue

- **Normal Peripheral Blood**
  - Band Neutrophil
  - Eosinophil
  - Segmented Neutrophil
  - Basophil
  - Monocyte
  - Platelet
  - Lymphocyte

- **Small Intestine Mucosa**
  - Epithelial cells
  - Goblet cells
  - Lamina propria
  - Muscularis mucosa (smooth muscle)

Created with figures from library.med.utah/WebPath/HISTHTML/HIST0.html
Cell Identity is More Than Histopathology

A cell participates in multiple cell contexts.

Multiple factors shape a cell’s identity

- Membership in a taxonomy of cell types
- Simultaneous time-dependent processes
- Response to the environment
- Spatial positioning
Before We Get Started

- Single-cell RNA-Seq (scRNA-Seq) analysis methodology is developing.
  - Give you a feel for the data.
  - Perform some analysis together.

- There is a vivid diversity of methodology.
  - These technique will grow as the field does.
  - Why were these specific tools chosen?

- This is a guided conversation through scRNA-Seq analysis.
  - Breadth and targeted depth.
  - There may be other opinions, if you have one, please speak up so we can all learn from it.
Before We Get Started

- Sections will be hands-on.
  - Much can be applied to other analysis.
  - Strengthen those R ninja skills!
  - If you need, cut and pasting is available.
    - cut_and_paste.txt

- There will be many cute corgi pictures.
We Will Attempt to Cover

- Describe scRNA-Seq assays.
- Comparing assays.
- Sequence pipelines.
- How do measured counts behave?
- Concerns over study design.
- Initial data exploration.
- Gene and cell filtering.
- Plotting genes.
- Dimensional Reduction and plotting cells.
- Differential expression.
- Communicating your study.
Section: scRNA-Seq Assays

- There are many scRNA-Seq Assays, each differs:
  - Some commercialized
  - Full transcriptome vs 3’
  - Less or more automated
  - Different levels of throughput
  - Differences in cost
Smart-seq2 for sensitive full-length transcriptome profiling in single cells
Simone Picelli1, Åsa K Björklund1,2, Omid R Faridani2, Sven Sagasser1,2, Gösta Winberg1,3 & Rickard Sandberg1,2

Single-cell gene expression analyses hold promise for characterizing cellular heterogeneity, but current methods compromise on either the coverage, the sensitivity or the throughput. Here, we introduce Smart-seq2 with improved reverse transcription, template switching and preamplification to increase both yield and length of cDNA libraries generated from individual cells. Smart-seq2 transcriptome libraries have improved detection, coverage, bias and accuracy compared to Smart-seq libraries and are generated with off-the-shelf reagents at lower cost.

Several methods exist for constructing full-length cDNAs from template switching, providing more even read coverage across transcripts than poly(A)-tailing methods, consistent with the common use of template switching in applications designed to capture RNA 5’ ends. Despite widespread use of single-cell transcriptome profiling methods, no systematic efforts have been made to improve cDNA library yield and average length from single-cell amounts.

We systematically evaluated a large number of variations in reverse transcription, template-switching oligonucleotides (TSOs) and PCR preamplification (for a total of 457 experiments) and compared the results to those from commercial Smart-Seq (hereafter called SMARTer) in terms of cDNA library yield and length from 1 ng of starting total RNA. (Supplementary Table 1). In particular, exchanging only a single guanilate for a locked nucleic acid (LNA)35 guanilate at the TSO 3’ end (c5G+c6G) led to a two-fold increase in cDNA yield relative to that obtained with the SMARTer II oligo (P = 7.2 x 10^-3, n = 28, Student’s t-test; Fig. 1a, Supplementary Table 2 and Supplementary Fig. 1). This is likely a consequence of the increased thermal stability of LNA-DNA base pairs (1-8 °C per LNA monomer). Additionally, we found that the presence of the methyl group donor betaine32 in combination with higher MgCl2 concentrations significantly increased yield (by two- to fourfold; P ≤ 1.3 x 10^-5, n = 56, Student’s t-test).

Full-length RNA-seq from single cells using Smart-seq2
Simone Picelli1, Omid R Faridani1, Åsa K Björklund1,2, Gösta Winberg1,3, Sven Sagasser1,2 & Rickard Sandberg1,2

1Ludwig Institute for Cancer Research, Stockholm, Sweden; 2Department of Cell and Molecular Biology, Karolinska Institute, Stockholm, Sweden. Correspondence should be addressed to R.S. (rickard.sandberg@ki.se).

Published online: 2 January 2014; doi:10.1038/nprot.2014.004

Emerging methods for the accurate quantification of gene expression in individual cells hold promise for revealing the extent, function and origins of cell-to-cell variability. Different high-throughput methods for single-cell RNA-seq have been introduced that vary in coverage, sensitivity and multiplexing ability. We recently introduced Smart-seq for transcriptome analysis from single cells, and we subsequently optimized the method for improved sensitivity, accuracy and full-length coverage across transcripts. Here we present a detailed protocol for Smart-seq2 that allows the generation of full-length cDNA and sequencing libraries by using standard reagents. The entire protocol takes ~2 d from cell picking to having a final library ready for sequencing; sequencing will require an additional 1–3 d depending on the strategy and sequence. The current limitations are the lack of strand specificity and the inability to detect nonpolyadenylated (polyA−) RNA.
Smart-Seq2: Description

Full transcript scRNA-Seq

- Developed for single cell but can performed using total RNA.
- Selects for poly-A tail.
- Full transcript assay.
  - Uses template switching for 5' end capture.
- Standard illumina sequencing.
  - Off-the-shelf products.
- Hundreds of samples.
- Often do not see UMI used.
Smart-Seq2: Assay Overview

- Poly-A capture with 30nt polyT and 25nt 5' anchor sequence.
- RT adding untemplated C
- Template switching
- Locked Nucleic Acid binds to untemplated C
- RT switches template
- Preamplification / cleanup
- DNA fragmentation and adapter ligation together.
- Gap Repair, enrich, purify.
Smart-Seq2: Equipment
Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Graphical Abstract

Authors
Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll

Correspondence
emacosko@genetics.med.harvard.edu (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

In Brief
Capturing single cells along with sets of uniquely barcoded primer beads together in tiny droplets enables large-scale, highly parallel single-cell transcriptomics. Applying this analysis to cells in mouse retinal tissue revealed transcriptionally distinct cell populations along with molecular markers of each type.

1000s of DNA-barcoded single-cell transcriptomes
Drop-seq: Description

- Moved throughput from hundreds to thousands.
- Droplet-based processing using microfluidics.
- Nanoliter scale aqueous drops in oil.
- 3' End.
- Bead based (STAMPs).
- Single-cell transcriptomes attached to microparticles.
- Cell barcodes use split-pool synthesis.
- Uses UMI (Unique Molecular Identifier).
- RMT (Random Molecular Tag).
- Degenerate synthesis.
Drop-seq: Overview

• Click Here for Drop-seq Video Abstract
Drop-seq: Assay Overview

**A** Complex tissue

Use Drop-Seq to analyze the RNA of each individual cell

**B** Barcoded primer bead

- PCR handle
- Cell barcode
- UMI

**C** Synthesis of cell barcode (12 bases)

- Synthesis Round 1
- Synthesis Round 2
- Synthesis Round 12

Number of unique barcodes in pool

**D** Synthesis of UMI (8 bases)

- Millions of the same cell barcode per bead
- $4^8$ different molecular barcodes (UMIs) per bead

- 8 rounds of synthesis
Drop-seq: Assay Overview

1. Cells from suspension
2. Microparticle and lysis buffer
3. Oil
4. Cell lysis (in seconds)
5. RNA hybridization
6. Break droplets
7. Reverse transcription with template switching
8. PCR (STAMPs as template)
9. Sequencing and analysis
   - Each mRNA is mapped to its cell-of-origin and gene-of-origin
   - Each cell’s pool of mRNA can be analyzed

B. Microfluidic device
   - Barcoded bead primers
   - Droplet formation

C. Sample sequencing read-pair
   - Read 1
   - Read 2
   - Cell barcode
   - UMI
   - Gene identity
Drop-seq: Assay Overview
Drop-seq: Equipment
Drop-seq: Pointers

- Droplet-based assays can have leaky RNA.
- Before library generation wash off any medium (inhibits library generation).
- Adding PBS and BSA (0.05-0.01%) can protect the cell.
  - Too much produces a residue making harvesting the beads difficult.
- Filter all reagent with a 80 micron strainer before microfluidics.
- Some purchased devices add a hydrophobic coating.
  - Can deteriorate (2 months at best).
  - Recoating does work (in-house).
New Results

Massively parallel digital transcriptional profiling of single cells


doi: http://dx.doi.org/10.1101/065912

This article is a preprint and has not been peer-reviewed [what does this mean?].
10X: Description

- Droplet-based, 3' mRNA.
  - GEM (Gel Bead in Emulsion)
- Standardized instrumentation and reagents.
- More high-throughput scaling to tens of thousands.
- Less processing time.
- Cell Ranger software is available for install.
10X: Assay Overview

Diagram showing the steps involved in the assay:

- 8-Channel Microfluidics Chip
  - Oil
  - Cells + Reagents
  - Beads

- Barcoded cDNA
  - Break Emulsion
  - Amplify cDNA
  - Construct Library
  - Sequence

- Barcoded Primer Gel Beads
  - Cells
  - Oil
  - Single Cell GEMs

Flowchart illustrating the process from input to output.
10X: Assay Overview
10X: Equipment

- Outlet wells
- Gel bead wells
- Master Mix wells
- Partitioning Oil wells
A Word on Sorting

- After disassociating cells cells can be performed.
- Know your cells, are they sticky, are they big?
  - Select an appropriate sized nozzle.
- Don't sort too quickly (1-2k cells per second or lower)
  - The slower the more time cells sit in lysis after sorting
  - 10 minutes max in lysis (some say 30 minutes)
- Calibrate speed of instrument with beads
  - Check alignment every 5-6 plates
- Afterwards spin down to make sure cells are in lysis buffer
  - Flash freeze
- Chloe Villani on sorting [click here]
Section: Comparing scRNA-Seq Assays
Power Analysis of Single Cell RNA-sequencing Experiments

Authors
Valentine Svensson*1,2, Kedar Nath Natarajan*1,2, Lam-Ha Ly2, Ricardo J Miragaia3,5, Charlotte Labalette2,3,4, Iain C Macaulay2, Ana Cvejic2,3,4 & Sarah A Teichmann1,2

Affiliations
1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
2 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
3 Wellcome Trust – Medical Research Council Cambridge Stem Cell Institute, Cambridge, CB2 1QR, UK
4 Department of Haematology, University of Cambridge, CB2 0PT, UK
5 Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
ERCC-based Benchmarking

- Based on ERCC spike-ins.
  - Exogenous RNA-Spikins
  - No secondary structure
  - 25b polyA Tail
    - May be a conservative measurement given endogenous mRNA will have ~250b polyA.
- Accuracy
  - How well the abundance levels correlated with known spiked-in amounts.
- Sensitivity
  - Minimum number of input RNA molecules required to detect a spike-in.
Sensitivity and Specificity

Accuracy

Great!

Poor

Sensitivity

Great!

CEL-Seq2

Drop-Seq

10X

10 molecules

1 molecule

Bulk

Smart-Seq2

protocols

Sensitivity

Molecule detection limit

Full length

Tag counting

Followup experiments
Final Thoughts

- Different assays have different throughput.
  - SmartSeq2 < Drop-seq < 10X
- SmartSeq2 is full transcript.
- Plate-based methods get lysed in wells and so do not leak.
  - Droplet-based can have leaky RNA.
- In Drop-seq assays RT happens outside the droplets
  - Can use harsher lysis buffers.
  - 10X needs lysis buffers compatible with the RT enzyme.
- 10X is more standardized and comes with a pipeline.
  - Drop-seq is more customizable but more hands-on.
- Cost per library varies greatly.
Section: scRNA-Seq Pipelines
Sequences Differ So Pipelines Differ

- scRNA-Seq assays are different and produce different sequences
  - The sequence pipelines must be tailored to the sequence of interest.
  - Many pipelines are NOT compatible but many show similarities.
Start with FASTQ Sequences

FASTQ File Format

- Sequence Header
- cDNA Sequence
- Base Quality

4 Lines are 1 sequence
Assays Differ in FASTQ Contents

<table>
<thead>
<tr>
<th>Assay</th>
<th>Left (R1)</th>
<th>Right (R2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SmartSeq2</strong></td>
<td>Paired Transcript Sequence</td>
<td>Paired Transcript Sequence</td>
</tr>
<tr>
<td><strong>Dropseq</strong></td>
<td>Cell Barcode UMI PolyT</td>
<td>Transcript Sequence</td>
</tr>
<tr>
<td><strong>10X</strong></td>
<td>Cell Barcode</td>
<td>Transcript Sequence UMI</td>
</tr>
</tbody>
</table>
SmartSeq2: Pipeline Overview
Drop-seq: Pipeline Overview

- Common functionality: trimming, alignment, generating count matrix.
- Adds book keeping for cell barcodes and UMIs, bead error detection, cell barcode collapsing, UMI collapsing.
Welcome to Drop-seq!

Drop-seq is a technology that allows biologists to analyze genome-wide gene expression in thousands of individual cells in a single experiment. This work is described in Macosko et al., Cell, 2015. This site provides interested users with resources to implement Drop-seq in their own labs. We hope you do amazing things with Drop-seq. Tell us about it!

If you would like to be informed of new protocol optimizations, discussion forums, etc., please send an email to dropseq@gmail.com and we'll put you on the list.
• Steps conceptually similar to the Drop-seq pipeline.
Overview of Single Cell Software

The Chromium Single Cell Software Suite is a complete package for analyzing and visualizing single cell 3’ RNA-seq data produced by the 10x Chromium Platform. The package includes:

- **Cell Ranger: Pipelines**
  A set of analysis pipelines that perform sample demultiplexing, barcode processing, and single cell 3’ gene counting.

- **Cell Ranger: R Kit**
  R package for secondary analysis of Cell Ranger matrix data, including PCA and t-SNE projection, and k-means clustering. A user-friendly guided tutorial is also available.
Much of the QC that is performed is using traditional tools.
Pipeline Section Summary

- Single-cell RNA-Seq is a diverse ecosystem of assays.
  - Each assay has pros and cons.
- Sequences derived from these assays are complex and vary.
- Different pipelines are needed to address different sequence formats.
  - Common steps include:
    - Aligning
    - QC
    - Read counting.
Section: scRNA-Seq Count Data
This is an Expression Matrix

<table>
<thead>
<tr>
<th>Gene</th>
<th>Cell 1</th>
<th>Cell 2</th>
<th>Cell 3</th>
<th>Cell 4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Gene 2</td>
<td>24</td>
<td>0</td>
<td>41</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Gene 3</td>
<td>175</td>
<td>284</td>
<td>93</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>Gene 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Gene 5</td>
<td>36</td>
<td>0</td>
<td>32</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Genes Have Different Distributions

Distribution of Expression of a Gene throughout a Study

Housekeeping Genes

Gene Expression (Per Cell)

Density

Log Normal

Log Expression

0
Genes Have Different Distributions

Distribution of Expression of a Gene throughout a Study

Different Populations

Density

Bimodal

0

Log Expression

Gene Expression (Per Cell)
Genes Have Different Distributions

Distribution of Expression of a Gene throughout a Study

Rare Populations

Density

Zero Inflated

0 Log Expression

Gene Expression (Per Cell)
Genes Have Different Distributions

Distribution of Expression of a Gene throughout a Study

- Noise

Technical (Should filter)

Density

0 Log Expression

Gene Expression (Per Cell)
Genes Have Different Distributions

Distribution of Expression of a Gene throughout a Study

- Noise
- Zero Inflated
- Bimodal
- Log Normal

Log Expression

Gene Expression (Per Cell)
• Zero inflation.
  – Drop-out event during reverse-transcription.
  – Genes with more expression have less zeros.
  – Complexity varies.
• Transcription stochasticity.
  – Transcription bursting.
  – Coordinated transcription of multigene networks.
  – Over-dispersed counts.
• Higher Resolution.
  – More sources of signal
Expression has Many Sources per Cell

- **Technical variation**
  - Batch effect
  - Library quality
  - Cell-specific capture efficiency
  - Amplification bias

- **Allele-intrinsic variation**
  - Bursts of transcription
    - Stochastic initiation
    - Stochastic duration
  - Varying rates of RNA processing

- **Allele-extrinsic variation** (cell types and states)
  - Fixed cell identity
    - Discrete
    - Continuous
  - Temporal progression/oscillation
  - Spatial location
    - Niche environments

---

Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner¹, Aviv Regev²,³,⁵ & Nir Yosef¹,⁴,⁵
Data Analysis with UMIs

<table>
<thead>
<tr>
<th>Read Counts</th>
<th>Counts by UMI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A4GALT</strong></td>
<td>0</td>
</tr>
<tr>
<td>AAAS</td>
<td>20</td>
</tr>
<tr>
<td>AACS</td>
<td>15</td>
</tr>
<tr>
<td>AADAT</td>
<td>14</td>
</tr>
<tr>
<td>AAED1</td>
<td>33</td>
</tr>
<tr>
<td>AAGAB</td>
<td>19</td>
</tr>
<tr>
<td>AAK1</td>
<td>5</td>
</tr>
<tr>
<td>AAMDC</td>
<td>90</td>
</tr>
<tr>
<td>AAMP</td>
<td>56</td>
</tr>
<tr>
<td>AANAT</td>
<td>0</td>
</tr>
</tbody>
</table>

Collapsed but Not Linear
Summary of the Data

• We are still understanding scData and how to apply it.
  – Data can be NOT normal.
  – Data can be Zero-inflated.
  – Data can be very noisy.
  – Cells vary in library complexity.
  – Can represent many “basis vectors” or sources of expression simultaneously.

• Keeping these characteristics in analysis assumptions.
• Tend to filter more conservatively with UMIs.
Section: Study Design and scRNA-Seq
scRNA-Seq Study Design

- How many cells?
  - Can change depending on the variability of the biology and the expectation of finding rare populations.
- How to design cell capture?
  - Single cell RNA-Seq is especially prone to technical batch affects (due to processing).
- Use of UMIs
- Use of ERCC spike-ins
How Many Cells?

- Satija lab online tool
  - satijalab.org/howmanycells

![Probability of seeing at least 5 cells from each cluster](image-url)
New Results

On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data

Stephanie C Hicks, Mingxiang Teng, Rafael A Irizarry

doi: http://dx.doi.org/10.1101/025528

This article is a preprint and has not been peer-reviewed [what does this mean?].
What is Study Confounding?

<table>
<thead>
<tr>
<th>Cell</th>
<th>Site</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Main</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>Main</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>Main</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>Main</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>Remote</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>Remote</td>
<td>B</td>
</tr>
<tr>
<td>7</td>
<td>Remote</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>Remote</td>
<td>B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cell</th>
<th>Site</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Main</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>Main</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>Main</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>Main</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>Remote</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>Remote</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>Remote</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>Remote</td>
<td>B</td>
</tr>
</tbody>
</table>
Confounding by Design
Section: Initial Data Analysis
Motivation: Why Am I Using R?

- A lot of method development is happening in R.
- Free / open source / open science.
- Many supplemental computational biology packages.
- Data science is an art.
  - Data often requires one to create and manipulate analysis.
- This will allow you to experience key concepts in analysis.
RStudio (IDE)
Initial Data Exploration
Today’s Data

To generate a comprehensive, validated classification scheme for the bipolar cells of the mouse retina.

- Cone or rod type, ON or OFF, 9-12 subtypes (morphological)

- ~44k cells from a transgenic mouse line marking BCs
  - After filtering 27k (we use 5k)
Logistics: Importing Code Libraries

- R Exercise
Representing Sparse Matrices

- R Exercise
What is a Sparse Matrix?

- Sparse Matrix
  - A matrix where most of the elements are 0.
- Dense Matrix
  - A matrix where most elements are not 0.
- Many ways to efficiently represent a sparse matrix in memory.
  - Here, the underlying data structure is a coordinate list.
2D Arrays vs Coordinate Lists

Can be optimal for dense matrices

More optimal for sparse matrices

2D Arrays vs Coordinate List

2D Arrays

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Coordinate List

<table>
<thead>
<tr>
<th>2</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>
Spatial reconstruction of single-cell gene expression data

Rahul Satija¹,⁷,⁸, Jeffrey A Farrell²,⁸, David Gennert¹, Alexander F Schier¹,⁵,⁹ & Aviv Regev¹,⁶,⁹

Cell

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

https://github.com/satijalab/seurat
Create a Seurat Object

- R exercise
In bulk RNA-Seq we learned counts are not expression.

- Some counts belong to sequences which could go to many genes.
- Some transcripts are longer than other so they get sequenced more.
- Some samples are more deeply sequenced.
- The data is not normally distributed.

Depending on the scRNA-Seq assay these may be important.

Seurat has assumptions it makes with it’s defaults
  - More appropriate for 3 prime assays.
Count Preparation is Different Depending on the Source

RSEM
KALLISTO
TPM

Correct for Sequencing Depth

Seurat
Log2() + 1
Seurat
Log2() + 1
Prepping Counts For Seurat

3 prime-
- Expected by Seurat.
- Counts collapsed with UMIs.
- Log2 transform (in Seurat).
- Account for sequencing depth (in Seurat).

Full Transcript Sequencing-
- Can be used in Seurat.
- TPM +1 transformed counts.
- Log2 transform (in Seurat).
- Sequencing depth is already accounted.
Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be perfectly comfortable. –Bobby Bragan
Filtering Genes: Averages are Less Useful

0s pull down average

5 Cells at 75
Average $\sim 2.7$
95 Cells at 0

Amount of 0s is arbitrary (study size, diversity)

5 Cells at 75
Average $\sim 0.42$
1000 Samples
Filtering Genes: Using Prevalence
Filtering Genes: Using Prevalence

Filter: 5 cells must have 10 expression

5 Cells at 75

95 Cells at 0

1000 Cells at 0
Filtering Using Metadata
What is Metadata?

Other information that describes your measurements.

- Patient information.
  - Life style (smoking), Patient Biology (age), Comorbidity
- Study information.
  - Treatment, Cage, Sequencing Site, Sequencing Date
- Sequence QC on cells.
  - Useful in filtering.
Filtering Cells: Removing Outlier Cells

• Bulk RNA-Seq studies often do not remove outliers cells
  – scRNA-Seq often removes “failed libraries”.
• Outlier cells are not just measured by complexity
  • Percent Reads Mapping
  • Percent Mitochondrial Reads
  • Presence of marker genes
  • Intergenic/ exonic rate
  • 5' or 3' bias
  • other metadata …
• Useful Tools
  – Picard Tools and RNASEQC
Seurat: Filtering on Metadata

- R Exercise
Seurat: Viewing Specific Genes

- R Exercise
Normalization and Batch Affect Correction

- The nature of scRNA-Seq assays can make them prone to confounding with batch affects.
  - Normalization and batch affect correction can help.
- Some are moving away from relying on a specific method.
  - Exploring the idea of combining or selecting from a collection of normalization or correction methods best for a specific study.
- Some believe UMI based analysis need not be normalized between samples given the absolute count of the molecules are being reported.
  - Be careful not to remove biological signal with good experimental design (avoiding confounding by design).
Using linear models one can regress covariates.
  - `scale.data` hold the residuals after regressing (z-scored)
Dimensionality reduction and clustering.
We use metadata we have.
  - One could imagine creating a metadata for cell cycle.
Seurat and Batch Affect Correction

- R exercise
Section: Dimensionality Reduction and Plotting Samples
Dimensionality Reduction

• Start with many measurements (high dimensional).
  – Want to reduce to few features (lower-dimensional space).
• One way is to extract features based on capturing groups of variance.
• Another could be to preferentially select some of the current features.
  – We have already done this.
• We need this to plot the cells in 2D (or ordinate them)
• In scRNA-Seq PC1 may be complexity.
PCA: in Quick Theory

- Eigenvectors of covariance matrix.
- Find orthogonal groups of variance.
- Given from most to least variance.
  - Components of variation.
  - Linear combinations explaining the variance.
PCA: an Interactive Example

- PCA Explained Visually
PCA: in Practice

Things to be aware of-

• Data with different magnitudes will dominate.
  – Zero center and divided by SD.
    • (Standardized).
• Can be affected by outliers.
• Data is often first filtered to remove noise.
t-SNE: Nonlinear Dimensional Reduction
t-SNE: Collapsing the Visualization to 2D
t-SNE: How it works.
PCA and t-SNE Together

• Often t-SNE is performed on PCA components
  – Liberal number of components.
  – Removes mild signal (assumption of noise).
  – Faster, on less data but, hopefully the same signal.
Learn More About t-SNE

- Awesome Blog on t-SNE parameterization
  - [http://distill.pub/2016/misread-tsne](http://distill.pub/2016/misread-tsne)

- Publication

- Nice YouTube Video
  - [https://www.youtube.com/watch?v=RJVL80Gg3lA](https://www.youtube.com/watch?v=RJVL80Gg3lA)

- Code
  - [https://lvdmaaten.github.io/tsne/](https://lvdmaaten.github.io/tsne/)

- Interactive Tensor flow
  - [http://projector.tensorflow.org/](http://projector.tensorflow.org/)
Plotting Cells
Plotting Cells and Gene Expression

- R exercise.
Defining Clusters through Graphs

- Smart Local Moving (SLM) algorithm for community (cluster) detection in large networks.
  - Can be applied to 10s of millions cells, 100s of millions of relationships.
  - Evolved from the Louvain algorithm

http://www.ludowaltman.nl/slm/
Local Moving Heuristic
Section Summary

- Dimensionality reduction help reduce data while hopefully keeping important signal.
  - t-SNE on PCA is often used in analysis

- Created several types of plot often seen in publications.
  - Plotting genes (through subgroups).
  - Ordinating cells in t-SNE space.
  - Heat maps of genes associated with PC components.
  - Plotting metadata on projects of data is an important QC tool.

- Cluster of cells are currently defined through graph, separate from the ordination (t-SNE / PCA).
Seurat: Differential Expression

- Default if one cluster again many tests.
  - Can specify an ident.2 test between clusters.
- Adding speed by excluding tests.
  - Min.pct - controls for sparsity
  - Min percentage in a group
  - Thresh.test - must have this difference in averages.
Seurat: Many Choices for DE

- **bimod**
  - Tests differences in mean and proportions.
- **roc**
  - Uses AUC like definition of separation.
- **t**
  - Student's T-test.
- **tobit**
  - Tobit regression on a smoothed data.
Seurat: DE and Plotting DE Genes

• R Exercise.
Dot plots

Size of circle
• Gene prevalence in cluster.

Color of circle
• More red, more expressed in cluster.

Scales well with many cells.
Mast

• Uses hurdle model
  – Two part generalized linear model to address both rate of expression (prevalence) and expression.
  – GLM means covariates can be used to control for unwanted signal.

• CDR: Cellular detection rate
  – Cellular complexity
  – Values below a threshold are 0

• Additionally introduces a GSEA method.

https://github.com/RGLab/MAST
Mast: Hurdle Models

Distribution of Expression of a Gene throughout a Study

Logistic Regression

Gaussian Linear Model

Difference in number of zeroes

Difference in distributions
Mast: DE and Plotting DE Genes

- R Exercise.
Section: Communicating Results to Collaborators

- Designing a study.
- Writing a grant.
- Performing experiments.
- Collecting data.
- Running sequencing pipelines.
- Performing some preliminary analysis.
- **Sharing ideas with private collaborators.**
- Refining analysis.
- Completing a paper.
- **Sharing analysis publicly.**
The Single Cell Portal

https://portals.broadinstitute.org/single_cell
The Single Cell Portal

Study Descriptions Can Be Created

Study: Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus (sNuc-Seq)

Overview:

Single nucleus RNA-seq provides rich information about cell types and states. However, it is difficult to capture rare dynamic processes, such as adult neurogenesis, because isolation of rare neurons from adult tissue is challenging and markers for each phase are limited. Here, we develop Div-Seq, which combines scalable single-nucleus RNA-Seq (sNuc-Seq) with pulse labeling of proliferating cells by EdU to profile individual dividing cells. sNuc-Seq and Div-Seq can sensitively identify closely related hippocampal cell types and track transcriptional dynamics of newborn neurons within the adult hippocampal neurogenic niche, respectively. This study contains the sNuc-Seq analysis performed as a part of the Div-Seq method development.

Using sNuc-Seq, we analyzed 1,367 single nuclei from hippocampal anatomical sub-regions (DG, CA1, CA2, and CA3) from adult mice, including enrichment of genetically-tagged low abundance GABAergic neurons (9). sNuc-Seq robustly generated high quality data across animal age groups (including 2 years old mice), detecting 5,100 expressed genes per nucleus on average, with comparable complexity to single neuron RNA-Seq from young mice (1, 2, 3). Analysis of sNuc-Seq data revealed distinct nuclei clusters (Fig. 1B-D shown below) corresponding to known cell types and anatomical distinctions in the hippocampus.
# The Single Cell Portal

## Data Can Be Shared

<table>
<thead>
<tr>
<th>Filename</th>
<th>Description</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipolar1_R1.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 1, left fastq file</td>
<td>6.54 GB</td>
</tr>
<tr>
<td>Bipolar1_R2.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 1, right fastq file</td>
<td>19.5 GB</td>
</tr>
<tr>
<td>Bipolar2_R1.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 2, left fastq file</td>
<td>8.75 GB</td>
</tr>
<tr>
<td>Bipolar2_R2.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 2, right fastq file</td>
<td>16.6 GB</td>
</tr>
<tr>
<td>Bipolar3_R1.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 3, left fastq file</td>
<td>5.07 GB</td>
</tr>
<tr>
<td>Bipolar3_R2.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 3, right fastq file</td>
<td>11.8 GB</td>
</tr>
<tr>
<td>Bipolar4_R1.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 4, left fastq file</td>
<td>6.96 GB</td>
</tr>
<tr>
<td>Bipolar4_R2.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 4, right fastq file</td>
<td>16 GB</td>
</tr>
<tr>
<td>Bipolar5_R1.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 5, left fastq file</td>
<td>6.9 GB</td>
</tr>
<tr>
<td>Bipolar5_R2.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 5, right fastq file</td>
<td>17.3 GB</td>
</tr>
<tr>
<td>Bipolar6_R1.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 6, left fastq file</td>
<td>6.54 GB</td>
</tr>
<tr>
<td>Bipolar6_R2.fastq.gz</td>
<td>Bipolar cell Drop-seq experiment 6, right fastq file</td>
<td>16.8 GB</td>
</tr>
<tr>
<td>clust_retinal_bipolar.txt</td>
<td>Louvain-Jaccard cluster assignments (CLUSTER) and Infomap assignments (SUB-CLUSTER)</td>
<td>1.57 MB</td>
</tr>
<tr>
<td>coordinates_retinal_bipolar.txt</td>
<td>Primary coordinates</td>
<td>1.56 MB</td>
</tr>
<tr>
<td>exp_matrix.txt</td>
<td>median normalized, log transformed values</td>
<td>1.1 GB</td>
</tr>
</tbody>
</table>
The Single Cell Portal

One Can Interact with Cell Clusters
The Single Cell Portal

Gene Expression Can be Viewed Across Clusters
The Single Cell Portal

Gene Expression Can be Viewed Across Clusters
The Single Cell Portal

Multiple Clustering Can be Used
The Single Cell Portal

Genes Can Be Viewed in Many Clusters
The Single Cell Portal

Expression Can Be Shown in Many Clusterings
The Single Cell Portal

Expression in Clusters Can Also Be Shown as Heatmaps
The Single Cell Portal

• Studies can be ...
  – Private
  – Private but shared privately
  – Public but with data inaccessible
  – Public
Section: Wrapping Up
What Did We Miss (So Much)?

So much more to learn!

We covered this
Awesome List

https://github.com/seandavi/awesome-single-cell

List of software packages for single-cell data analysis, including RNA-seq, ATAC-seq, etc.

awesome-single-cell

List of software packages (and the people developing these methods) for single-cell data analysis, including RNA-seq, ATAC-seq, etc. Contributions welcome...

Software packages

RNA-seq

- anchor - [Python] - Find bimodal, unimodal, and multimodal features in your data
- BackSPIN - [Python] - Biclustering algorithm developed taking into account intrinsic features of single-cell RNA-seq experiments.
Thank You

Aviv Regev
Brian Haas
Adam Haber
Anindita Basu
Asma Bankapur
Chloe Villani
Karthik Shekhar
Kristine Schwenck

Matan Hofree
Michel Cole
Monika Kowalczyk
Nir Yosef
Sean Simmons
Regev Single Cell Working Group
Today's Attendees