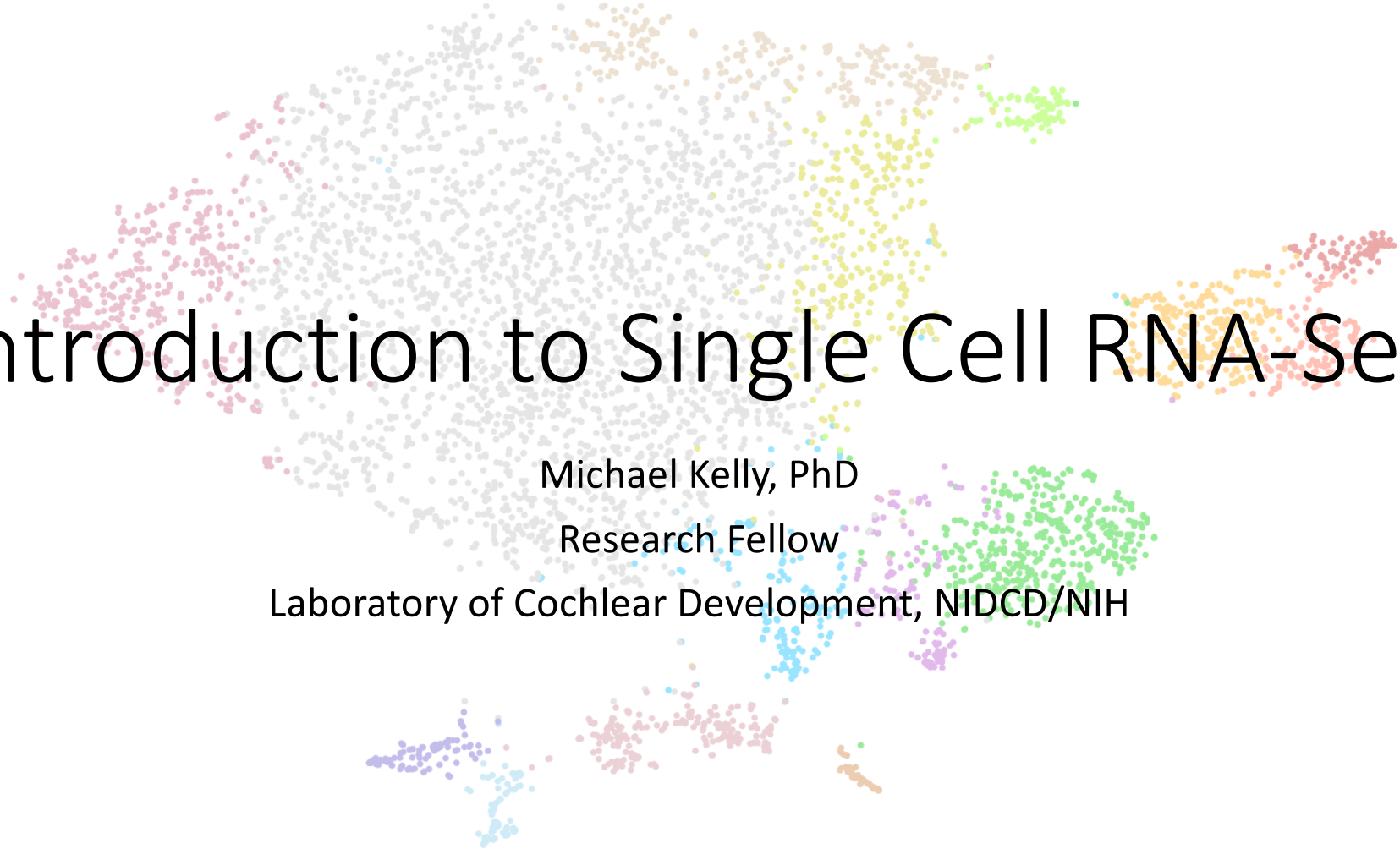


Introduction to Single Cell RNA-Seq

Michael Kelly, PhD

Research Fellow

Laboratory of Cochlear Development, NIDCD/NIH

Outline

Part I

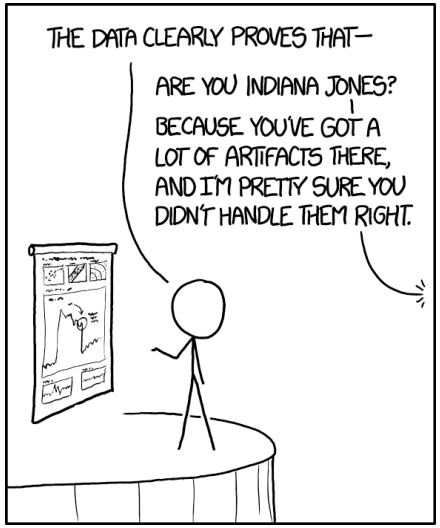
- How is single cell RNA-Seq (scRNA-Seq) data different than bulk RNA-Seq data?
- Different scRNA-Seq methods and the data type they generate
- Unique molecular identifiers (UMI's) and how they help
- Greater number of samples vs. greater depth?

Part II

- First steps in handling scRNA-Seq data
- What can you do with single cell data?
- Ongoing (open-source) development of analysis tools
- Web-based queries and analysis of published scRNA-Seq data

Part III

- What's next for scRNA-Seq?



Is single cell RNA-seq just RNA-seq with more samples? Not really.

- scRNA-Seq is zero-heavy data
  - Depending on method, you could have 500 genes of 40,000 have non-zero values
  - Analysis is a combination of discrete and continuous math (10 vs 0, and 1000 vs 1)

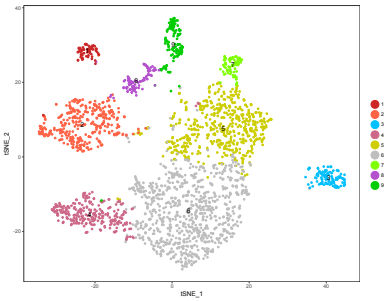
- Differential expression usually starts with defining which samples to compare
  - May require identification of outlier samples, normalization, and clustering of data
  - Ability to select samples in each comparison groups makes data very flexible

- Don't trust any one gene. Dimensionality reduction provide more reliable "meta-genes"
  - Both "drop-out events and noise/over-amplification can give the wrong impression
  - Biologically relevant principle components can represent "meta-genes" that can help sort out cell types

- Protocols are limited by the low-input amount of RNA
  - scRNA-Seq relies on quite a bit of PCR
  - Stranded protocols or total RNA methods generally not supported
  - Reverse transcription usually happens in the presence of the lysate (not ideal conditions)

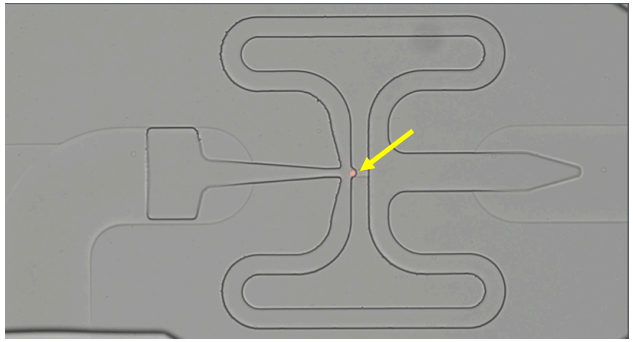
Cell # 1 ... 20

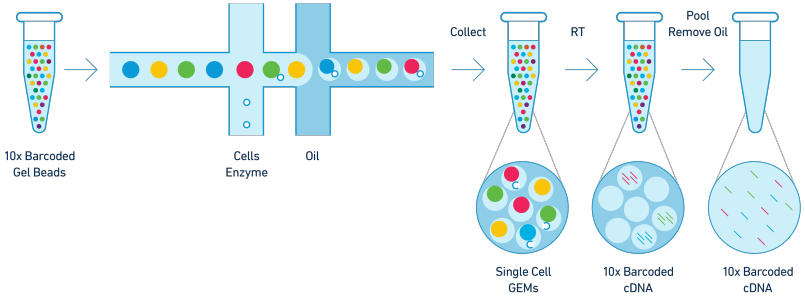
```
Xkr4      . . . . . .
Gm1992    . . . . . .
Gm37381   . . . . . .
Rp1       . . . . . .
Rp1.1     . . . . . .
Sox17     . . . . . 1 . . . . . .
Gm37323   . . . . . .
Mrpl15    . . . . . 1 . . . . 2 . . . . . .
Lypla1    . 1 2 . . . . . 2 . . . . 1 .
Gm37988   . . . . . .
```



- Manage expectations
- Don't assume bulk RNA-Seq analysis tools are appropriate for scRNA-Seq data

Overview of common single cell RNA-seq methods





Single Cell-Per-Well Protocols

Individual samples remain partitioned until library prep indexing

-Fluidigm C1
-FACs-based protocols
-Hand-picking
-Wafergen iCell8

Droplet-Based Massively Parallel Protocols

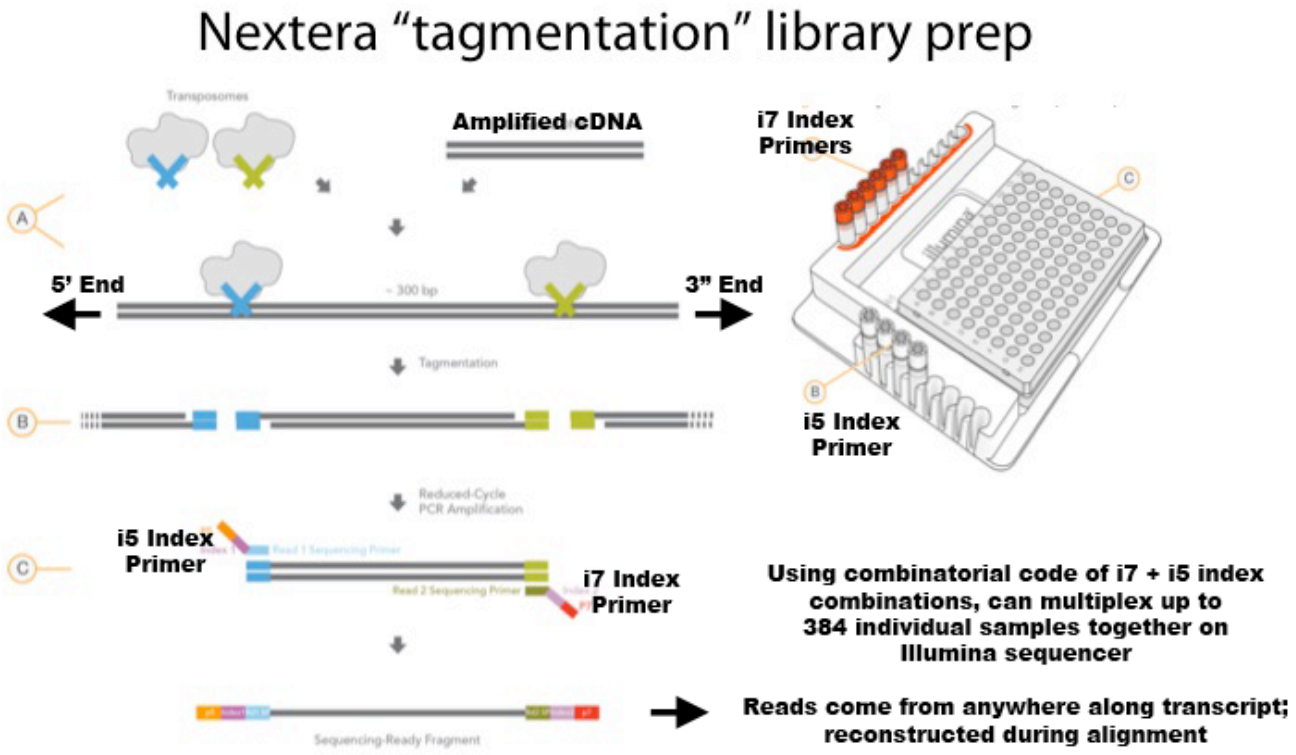
Cells and barcodes partitioned in droplets; indexing occurs at RT

-DropSeq
-InDrop (1CellBio)
-10X Genomics Chromium
-BioRad/Illumina SureCell on ddSEQ

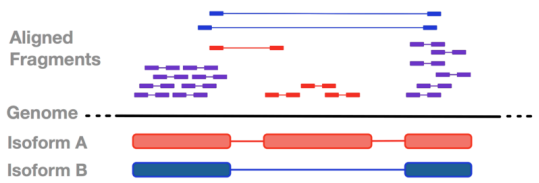
Method	$ system	$ per cells	No. cells	Doublets	Transcript type	UMIs	Capture Efficiency
DROP-seq	$50000	$0.65	up to 50000	0.36-11.3%	3' mRNA	Yes	~2%
Fluidigm C1	$150,000	$1.5-10	96, 800 (10k?)	10-23%	mRNA	No	~10%
10X Genomics	$125,000	$0.20-1.00	1000-6000	1-5%	3' mRNA	Yes	65%
Wafergen	$200,000	$1.5-2.5	~1800	1-5%?	3' mRNA	Yes	?

Modified from core-genomics.blogspot.com

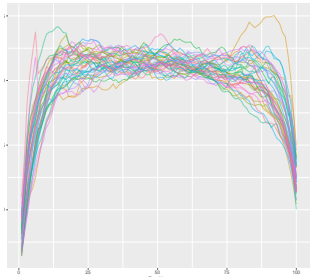
Single cell-per-well methods allow full-length scRNA-seq on Illumina NGS sequencing platforms



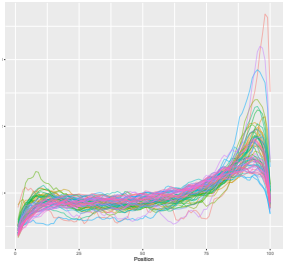
Sample 1: i7=N708; i5=S510
Sample 2: i7=N712; i5=S511
Sample 3: i7=N708; i5=S512



5-3' Transcript Coverage

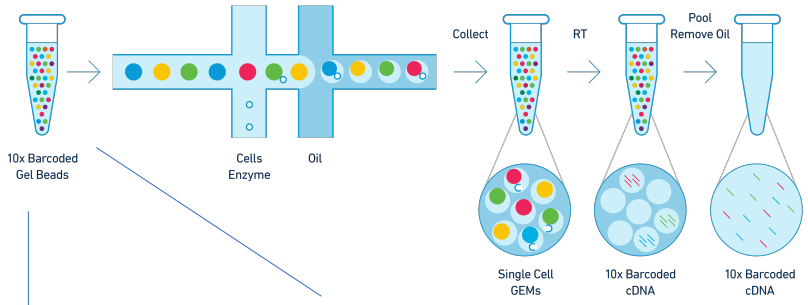


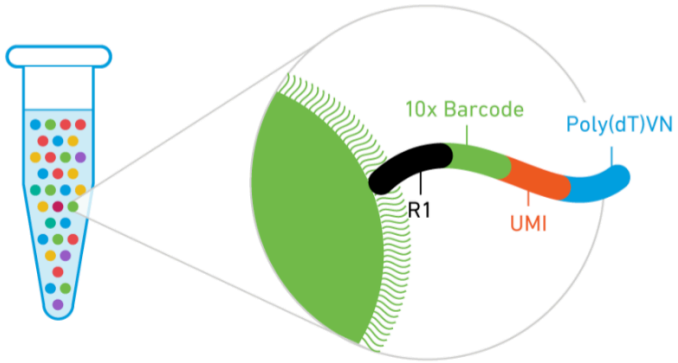
Good



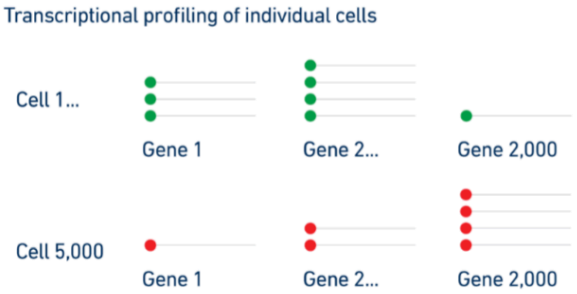
Bad

Droplet-based barcoding allows high-throughput scRNA-Seq gene counting



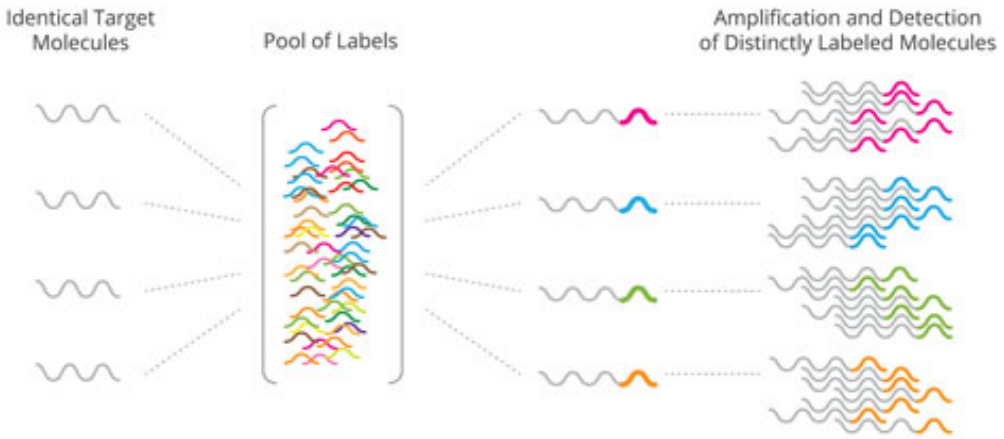


- Barcode added during reverse transcription
- 3’ (or 5’ end) of transcript is selectively enriched by PCR
- Interestingly, these methods give inherent strand information
- Originally lower sensitivity than single cell-per-well protocols, but now approaching similar levels



From 10X Genomics Promotional Material

Unique Molecular Indices Help Reduce Undesirable Biases From PCR Amplification

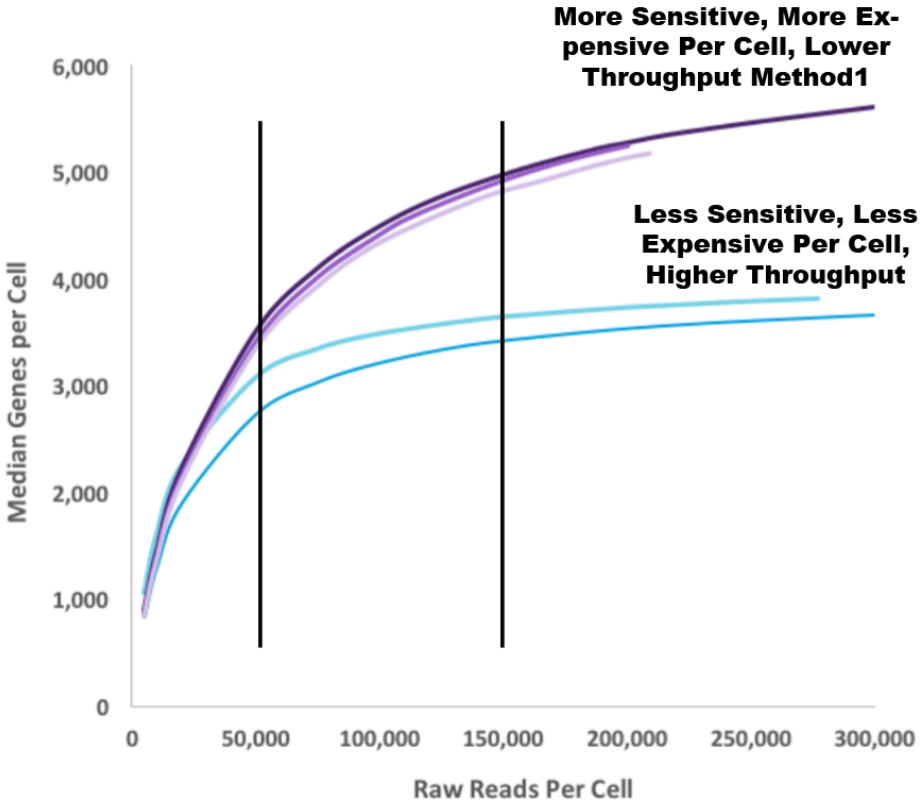


From: http://www.genengnews.com/gen-articles/molecular-indexing-with-precise-assays/5607

Note: Unique molecular identifiers are currently only possible with 5' or 3' end only methods on Illumina sequencers

Islam et al 2014 Nature Methods PMID: 24363023

Which is better – more cells or greater depth?



Modified from 10X Genomics material

- More information can be gained by sequencing to greater depth – especially using sensitive methods
  - More genes detected; fewer "drop-outs"
  - Better isoform discrimination (when full-length libraries sequenced)
- More independent observation (more cells) is better for cell identity classification – averages out noise
- Classic scientific non-answer: it depends on what you are looking for
  - Broad survey of cell types or dynamics processes best modeled by higher-throughput data
  - Investigation of presumably low-expressed (or specific isoforms) requires greater depth

First steps in handling scRNA-Seq data: Primary Analysis (Processing & Alignment)

- Demultiplexing of individual samples based on barcodes
  - Single cell-per-well protocols generally use Illumina indices
  - Droplet-based systems use custom scripts to extract and demultiplex
- Trimming and alignment
  - Removal of adapter sequence
  - Alignment of reads to transcriptome or genome with transcript coordinates
  - Full-length libraries can handle some multi-mapping; 5 or 3’ end libraries usually on utilize non-ambiguously mapped reads
- Assessment of alignment metrics
  - Percentage of reads mapped
  - Percentage exon vs intron vs intergenic
  - For full-length: gene body coverage and detection of splice sites

Input:
Raw sequencing files
Output:
Gene expression matrix

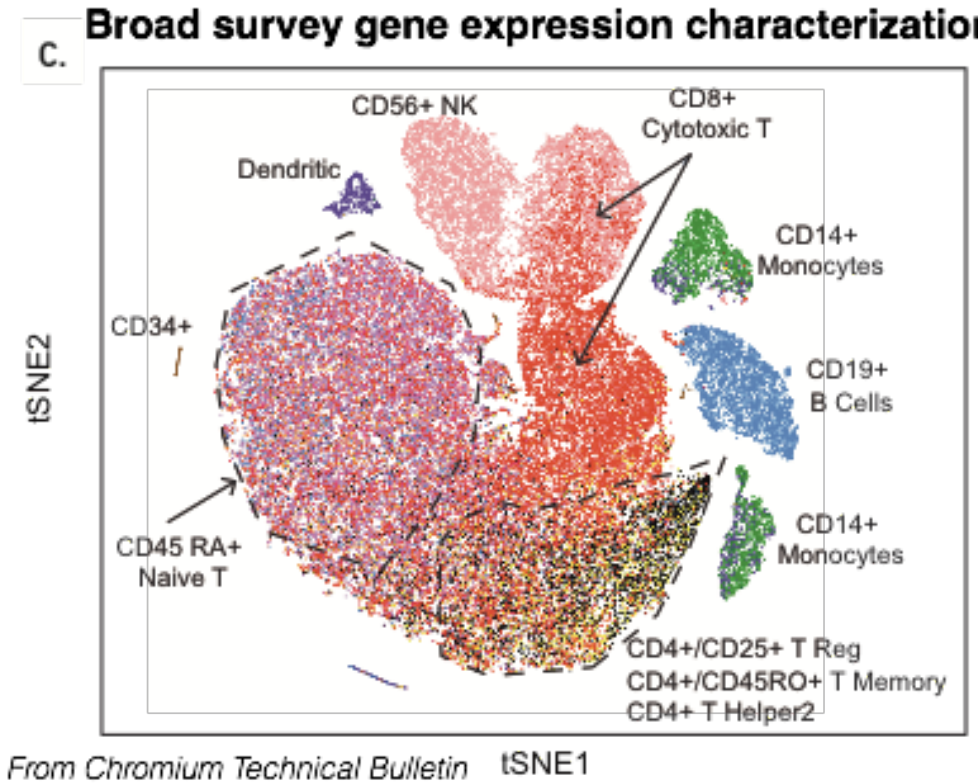
First steps in handling scRNA-Seq data: Secondary Analysis

- Initial QC and filtering
  - Outlier identification
  - Thresholding based on read depth, UMI counts, and/or genes detected
- Cross-sample normalization
  - Adjustment for library size, etc.
- Variance thresholding and stabilization
  - Selection of variable genes (non-"housekeepers")
  - Dispersion (variance over mean) threshold often used
  - Data transformed to reduce statistical weight of huge expression values (e.g. log-transformation)
- Dimensionality reduction
  - Principle component analysis (or similar) to look for structure in data
- Define relationships between individual samples
  - Clustering (hierarchical, k-means, graph-based)
  - Trajectory modeling
- Differential expression testing

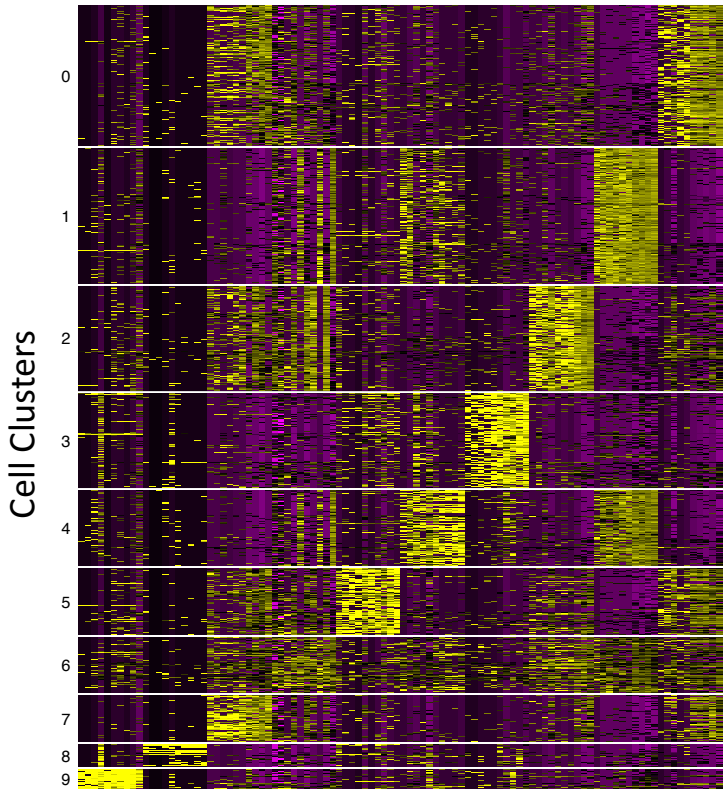
Input:
Gene expression matrix
Output:
Analyzed data

What can you do with single cell data?

Example of scRNA-Seq Analysis: Unbiased Identification of New Cell Types and Markers

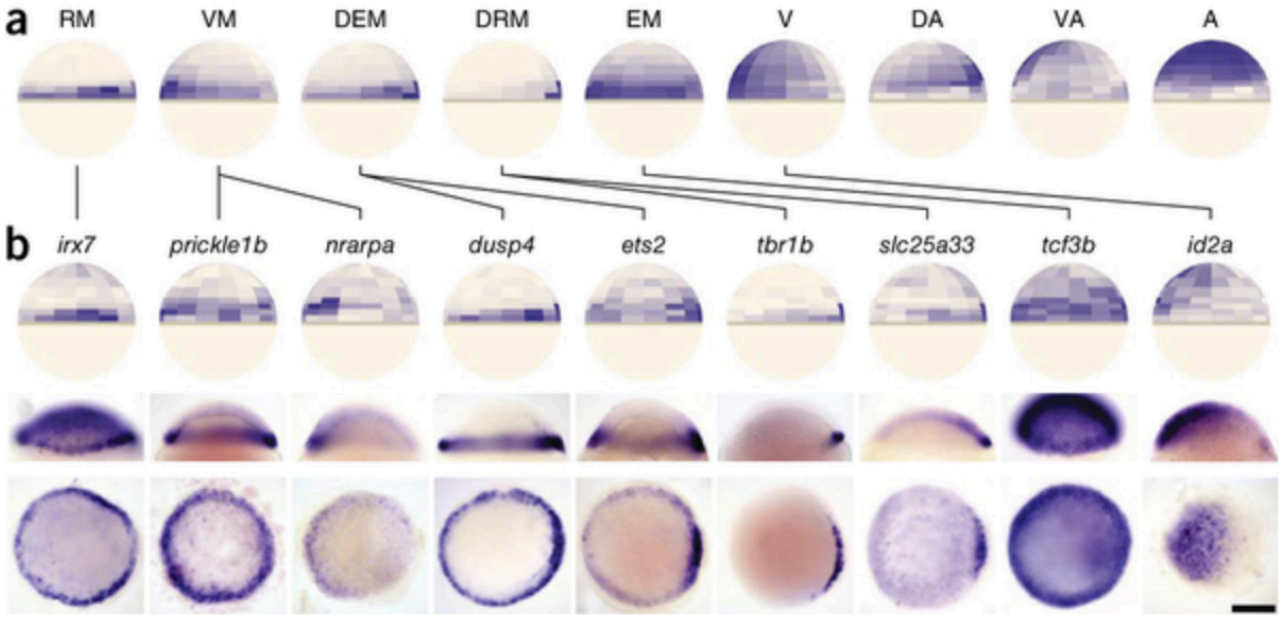


From Chromium Technical Bulletin



Marker Genes for Each Cluster

Example of scRNA-Seq Analysis: Spatial Inference & In Silico In Situ



Spatial reconstruction of scRNA-Seq data provides comprehensive expression location in developing xenopus

Satija et al 2015

Example of scRNA-Seq Analysis: Unbiased survey of cell ratio and transcriptional phenotypes changes

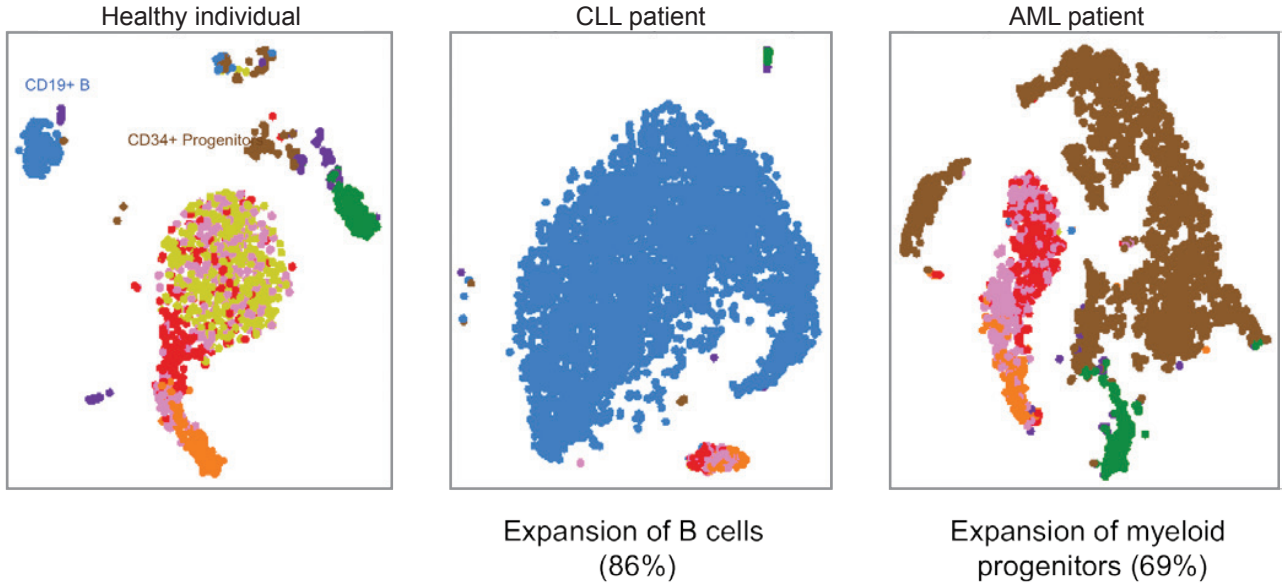
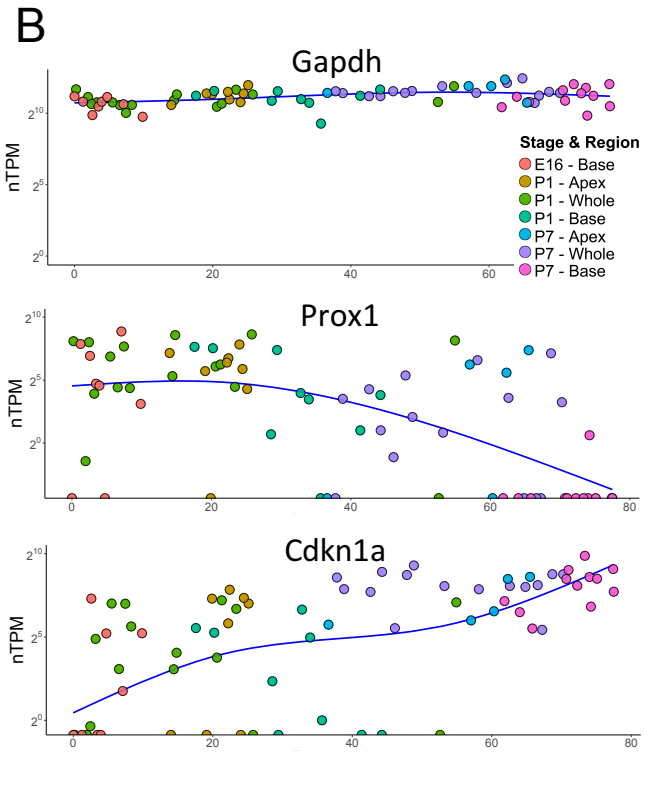
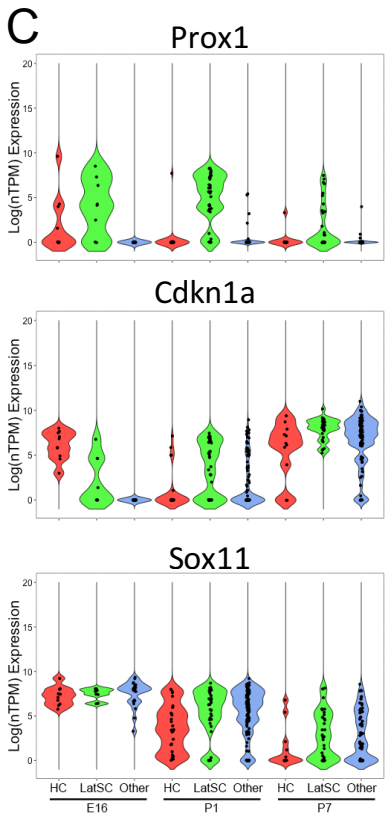


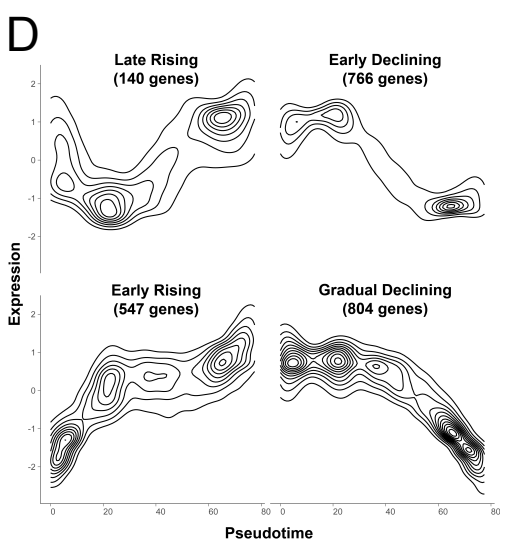
Figure 5. Single cell profiling from healthy and malignant tumor cell samples. Single cell profiling of BMMCs from healthy, CLL and AML patients. ~30,000 reads/cell in this experiment.

From 10X Genomics Chromium Technical Bulletin

Example of scRNA-Seq Analysis: Developmental Trajectory Analysis







E

Rising	
Gene	***ANOVA Rank (p-value)***
Plekhb1	1st (~0)
Enho	2nd (5.22×10^{-13})
Sdc4	3rd (2.10×10^{-12})
Cdkn1a	7th (9.71×10^{-10})
Car14	13th (2.00×10^{-8})

Declining	
Gene	***ANOVA Rank (p-value)***
Fn1	1st (4.41×10^{-11})
Chst15	2nd (6.91×10^{-10})
Epha7	3rd (1.53×10^{-8})
Sox11	17th (4.23×10^{-06})
Prox1	759th (0.01)

Pseudotime ordering of single cell samples according to differentiation stage (using Monocle package)

What analysis tools are available?

Note: Single cell RNA-Seq data, in particular, is best handled with active communication between the subject matter expert and computational biologist. For efficiency and effective, it helps to have subject matter expert get their hands dirty in the bioinformatics

A rich set of analysis tools for single cell RNA-seq datasets continues to expand (and they're open-source)

SATIJA LAB

HOME NEWS PEOPLE RESEARCH PUBLICATIONS

SEURAT R toolkit for single cell genomics

Monocle Trapnell Lab

Single-cell differential expression and trajectory analysis

Single Cell Differential Expression

SCDE pacakge provides routines for analysis of single-cell RNA-seq data. It is based on the probabilistic mixture error model, which is used to implement differential expression, subpopulation analysis and other tasks on the data.

Kharchenko Lab

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

★ 142 ⑂ 68

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.
- BASiCS - [R] - Bayesian Analysis of single-cell RNA-seq data. Estimates cell-specific normalization constants. Technical variability is quantified based on spike-in genes. The total variability of the expression counts is decomposed into technical and biological components. BASiCS can also identify genes with differential expression/over-dispersion between two or more groups of cells.
- bonvoyage - [Python] - 📐 Transform percentage-based units into a 2d space to evaluate changes in distribution with both magnitude and direction.
- BPSC - [R] - Beta-Poisson model for single-cell RNA-seq data analyses
- Cellity - [R] - Classification of low quality cells in scRNA-seq data using R
- cellTree - [R] - Cell population analysis and visualization from single cell RNA-seq data using a Latent Dirichlet Allocation model.
- clusterExperiment - [R] - Functions for running and comparing many different clusterings of single-cell sequencing data. Meant to work with SCONE and slingshot.

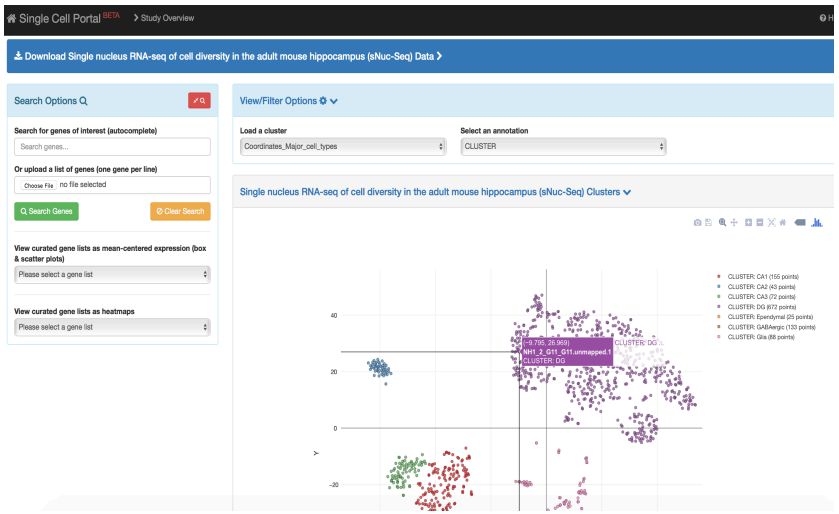
Continued...

Are there single cell RNA-Seq datasets to play with before collecting my own?

- Most single cell publications have data deposited in GEO
  - Can download raw data and usually processed expression matrices
- Some commercial platforms provide example datasets to view and analysis
  - 10X Genomics (https://support.10xgenomics.com/single-cell/datasets)
- Some analysis package developers provide example datasets
  - Seurat (http://satijalab.org/seurat/get_started.html)
- Some data can also be viewed in web-based portals

Single Cell Dataset Portals

Broad Single Cell Portal



Linnarsson Lab Data Viewers

Molecular Diversity of Midbrain Development in Mouse, Human and Stem Cells. Gioele La Manno, Daniel Gyllborg, Simone Codeluppi, Kaneyasu Nishimura, Carmen Salto, Amit Zeisel, Lars E. Borm, Simon R.W. Stott, Enrique M. Toledo, J. Carlos Villaescusa, Peter Lönnerberg, Jesper Ryge, Roger A. Barker, Ernest Arenas and Sten Linnarsson

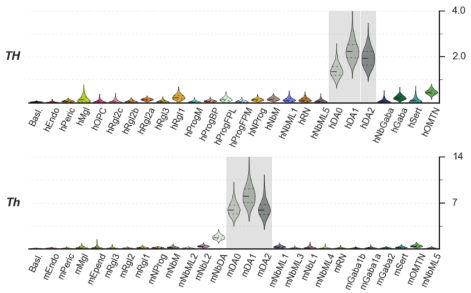
Raw data available from GEO.

Download all the data and Python Notebooks from GitHub to reproduce the main figures.

Gene browser

Th | Show gene

Bayesian posterior probability of expression for *Th* (molecules/cell)



What's next for single cell RNA-Seq?

- Better sensitivity and less technical noise
- Improved methods for fixed or frozen tissue
- Isoform information from high-throughput methods
- Preservation or encoding of spatial information
- Parallel detection of other –omic modalities along with transcription (e.g. epigenetics)
- Encoding of lineage relationships along with transcription
- Handling data from multiple methods and sources to build a model of all possible cell states (complete cell atlas)
- Many cell from one genome – reconstructing gene regulatory networks

Improvements in methods and throughput will have to be done hand-in-hand with computational methods; computational tools are also improving independently

Wagner et al Nature Biotech PMID: 27824854; Tanay & Regev Nature PMID: 28102262

Who to watch:
Regev, Linnarsson, Satija, Trapnell, and Shendure Labs
Twitter: @scell_papers

Summary

- Single Cell RNA-Seq has some differences from bulk RNA-Seq
  - Nuances to methods and the data make it important to make sure you're handling it correctly (consult with a bioinformatician that is familiar with scRNA-Seq data)
  - Variety of analysis tools allow you to harness very flexible scRNA-Seq datasets

- The methods of data collection vary in sensitivity and throughput
  - Choose a method that is best suited to primary experimental design/question (and what you can afford…)
  - UMI's help and the more cells the better

- It helps to know the biology when doing scRNA-Seq data analysis
  - Initial alignment and gene / isoform counting is similar to bulk RNA-Seq
  - Secondary analysis is a combination of determining what is statistically significant and what the biological question is (helps to know cell types and what genes are expressed)

- Single Cell RNA-Seq analysis is a fast-growing field that holds great promise
  - Many open-source analysis tools exists and are being developed
  - Computational tools are utilizing large sample size and flexibility – only getting better