Introduction to single-cell RNA-seq analysis

Inma Barrasa
BaRC Hot Topics
Bioinformatics and Research Computing
Whitehead Institute
Feb 27th 2020
http://barc.wi.mit.edu/hot_topics/

Whitehead Institute

BaRC Hot Topics

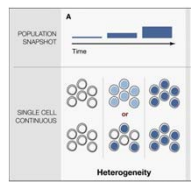
Outline

- Introduction to single-cell RNA-seq data analysis
  - Overview of scRNA-seq technology, cell barcoding, UMIs
  - Experimental design
  - Analysis pipeline
    - Preprocessing and quality control
    - Normalization
    - Dimensionality reduction
    - Clustering of cells
    - Trajectory inference
    - Differential expression and functional annotation
- Hands-on analysis using the R package Seurat

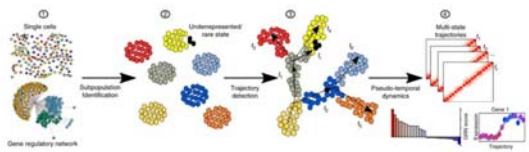
Why do single cell RNA-seq?

Access to expression profiles of individual cells allows us to:

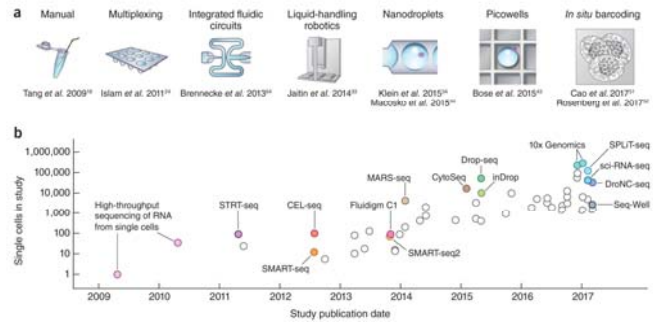
- Learn about cellular heterogeneity
- Discover of new cell populations
- Order cells within a developmental trajectory



Etzrodt, *Cell Stem Cell* 2014

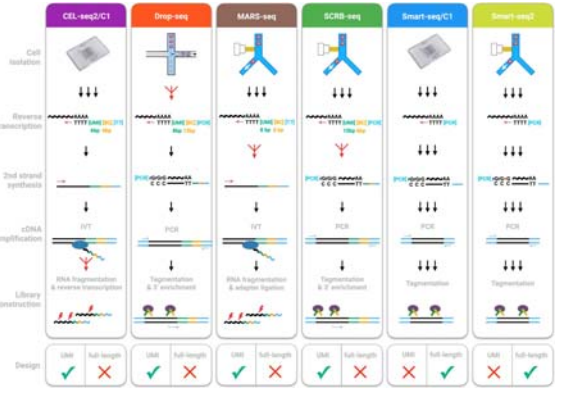

Lummertz da Rocha, *Nature Communications* 2018

Exponential scaling of single-cell RNA-seq in the past decade



Svensson, V., Vento-Tormo, R. & Teichmann, S.
Nat Protoc **13,** 599–604 (2018). https://doi.org/10.1038/nprot.2017.149

Library preparation steps



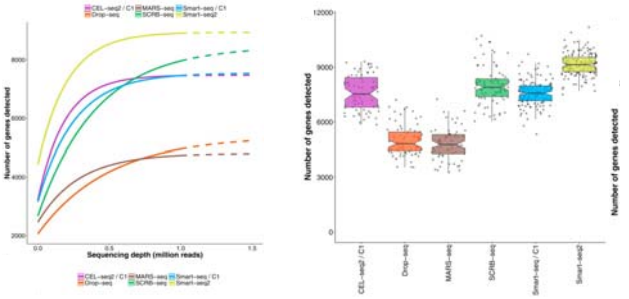
Comparative Analysis of Single-Cell RNA Sequencing Methods
Ziegenhain et. al, Molecular Cell
Volume 65, Issue 4, 16 February 2017,

Features of scRNA-seq methods

Name	Transcript coverage	Strand specificity	Positional bias	UMI possible?
Tang method	Nearly full-length	No	Strongly 3´	No
Smart-seq	Full-length	No	Medium 3´	No
Smart-seq2	Full-length	No	Weakly 3´	No
STRT-seq & STRT/C1	5´-only	Yes	5´-only	Yes
CEL-seq	3´-only	Yes	3´-only	No
CEL-seq2	3´-only	Yes	3´-only	Yes
MARS-seq	3´-only	Yes	3´-only	Yes
CytoSeq	Pre-defined genes only	Yes	3´-only	Yes
Drop-seq/InDrop	3´-only	Yes	3´-only	Yes

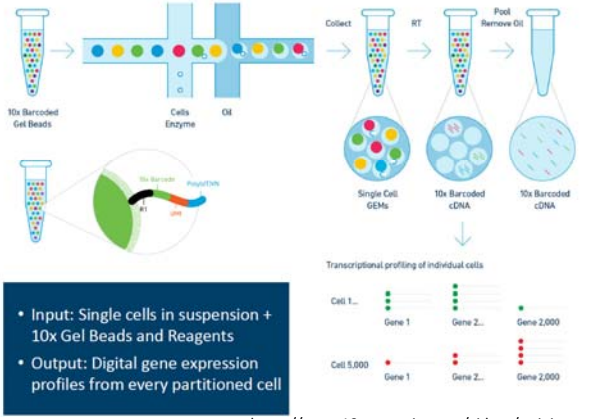
Single-cell RNA-sequencing: The future of genome biology is now
Simone Picelli, RNA Biology, Volume 14, 2017 - Issue 5

Sensitivity of scRNA-seq methods



Comparative Analysis of Single-Cell RNA Sequencing Methods Ziegenhain et. al, Molecular Cell
Volume 65, Issue 4, 16 Feb 2017

Single Cell Digital Gene Expression



- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell

https://www.10xgenomics.com/videos/training-modules/

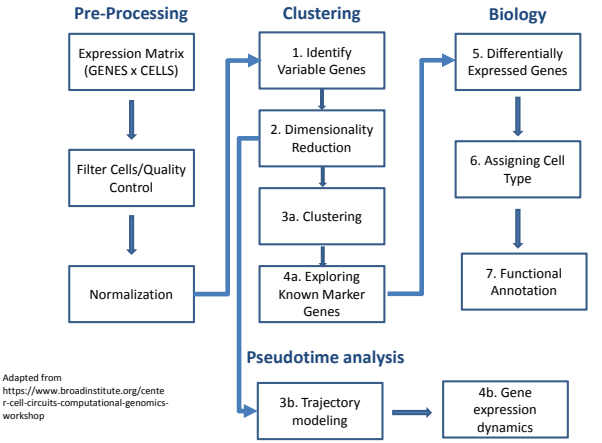
Libraries Compatible with Illumina® Sequencers 10x GENOMICS

Sequencing Read	Description	Length	Notes
Read 1	Cell Barcode+UMI read	26 nt	Extra cycles are unused
i7 Index	Sample Index read	8 nt	Extra cycles are unused
Read 2	RNA read	98nt (recommended)	Length may vary

Validated on: HiSeq 2500 (RR and HO), HiSeq 3000/4000, NextSeq 500/550, MiSeq

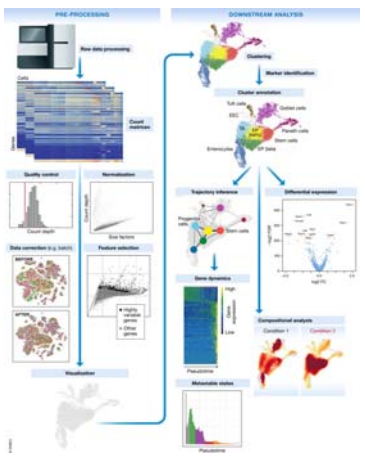
https://www.10xgenomics.com/videos/training-modules/

Analysis pipeline



Adapted from https://www.broadinstitute.org/center-cell-circuits-computational-genomics-workshop

Analysis pipeline



Current best practices in single-cell RNA-seq analysis: a tutorial

Mol Syst Biol, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

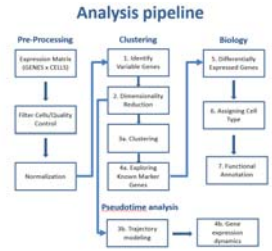
Technical challenges

- Data is noisy due to
  - cDNA amplification bias
  - mRNA capture efficiency
  - drop outs: large number of genes with 0 counts due to limiting mRNA. Zero expression doesn't mean the gene isn't on.
- Cells can change or die during isolation.

Experimental design

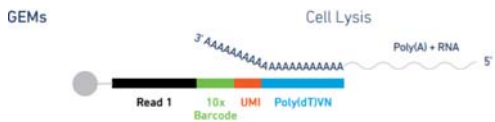
- Process your samples in a way that the condition can not be confounded with a batch effect, like processing date, facility, or reagents used.
  - *i.e.* If you have to process your cells in several batches, each batch should contain an equal number of cells from each condition.
- If you are comparing your data to published data you may have to remove batch effects.
  - R packages like Combat can be used for this (https://www.rdocumentation.org/packages/sva/versions/3.20.0/topics/ComBat)
  - See "Dealing with confounders" section of the "Analysis of single cell RNA-seq data" course (Hemberg Group).
  - Tran, H.T.N., Ang, K.S., Chevrier, M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21,** 12 (2020).

Preprocessing for Smart-seq2



- Demultiplexing: assign all the reads with the same cell barcode to the same cell. Done at the sequencing facility.
- We can check the quality of the reads with *FastQC* and the library composition with *FastQ Screen* as we would do with bulk RNA-seq.

Preprocessing for technologies using Unique Molecular Identifiers (UMIs)



- Demultiplexing: assign all the reads with the same cell barcode to the same cell.
- Remove PCR duplicates: if several reads have the same UMI and map to the same location in the genome, keep only one.
  - Cell ranger software for 10x data (run by the genome technology core)
  - Drop-seq tools for drop-seq and seq-well data

https://www.10xgenomics.com/videos/training-modules/

Demultiplexing and counting 10x data

Cell Ranger™ Pipelines

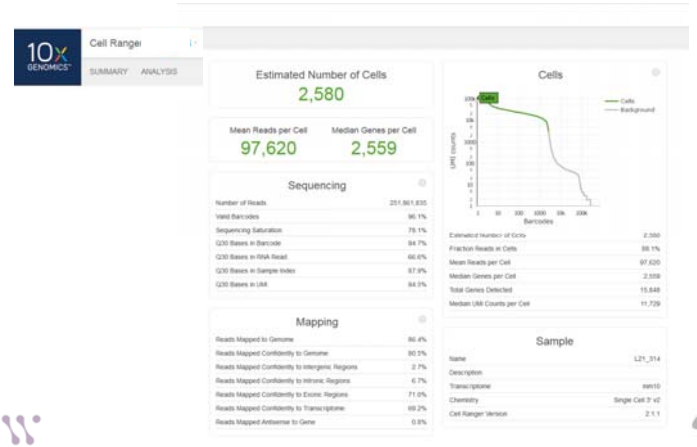
10x GENOMICS

Pipeline	Functionality
cellranger mkfastq	Barcode-aware demultiplexing from BCL to FASTQ

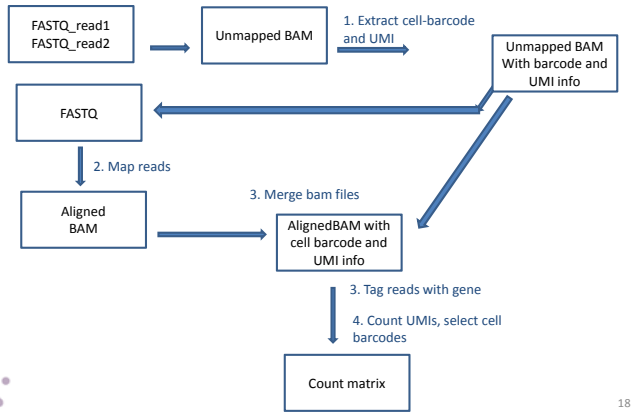
cellranger count	• Read-level analysis of a single library
	  • Transcriptome alignment with STAR
	  • Barcode processing
	  • Gene counting
	• Produces gene/cell matrix
	• Produces expression analysis and static visualizations
	• Produces .cloupe file for Loupe™ Cell Browser

https://www.10xgenomics.com/videos/training-modules/

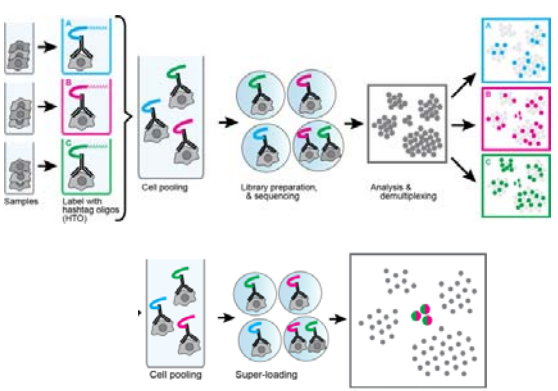
CellRanger web summary



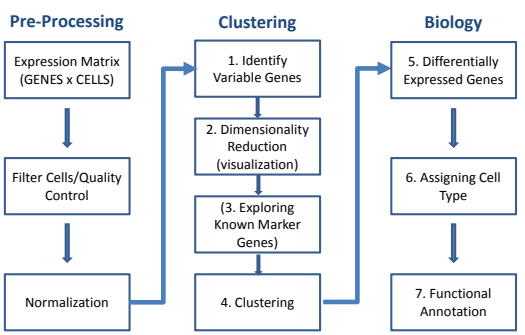
Demultiplexing and counting Drop-seq or Seq-well data



Cell Hashing



Analysis pipeline



Adapted from
https://www.broadinstitute.org/center-cell-circuits-computational-genomics-workshop

Quality control and filtering

- Quality control
  - Number of reads per cell
  - Number of genes detected per cell
  - Proportion of reads mapping to mitochondrial reads
- Remove cells with poor quality
  - Filter out cells with percentage of mitochondrial reads higher than a cut off
  - Filter out cells with less than a lower threshold on the number of genes or counts per cell
- Remove doublets (two cells captured with one bead in the droplet)
  - Filter out cells with more than an upper threshold on the number of genes or counts per cell in your data
  - More sophisticated way of removing doublets
    - https://github.com/JonathanShor/DoubletDetection
    - https://github.com/AllonKleinLab/scrublet
    - DoubletFinder
      https://www.sciencedirect.com/science/article/pii/S2405471219300730?via%3Dihub

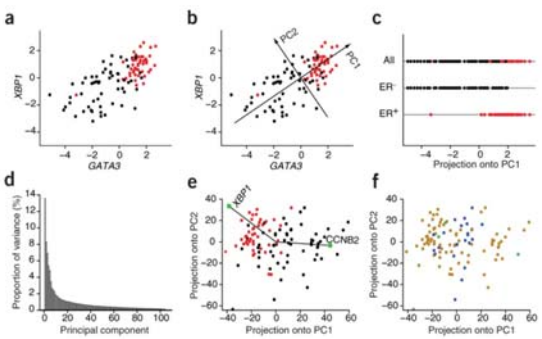
Normalization

Correct for sequencing depth (i.e. library size) of each cell so we can compare across cells

1. Normalize gene expression for each cell by total expression
2. Multiply by a scale factor (*i.e.* 10,000).
3. Log transform the scaled counts

This is the log normalization implemented in Seurat

Visualization: Principal Component Analysis



Ringnér, M. What is principal component analysis?. *Nat Biotechnol 26, 303–304 (2008).* https://doi.org/10.1038/nbt0308-303

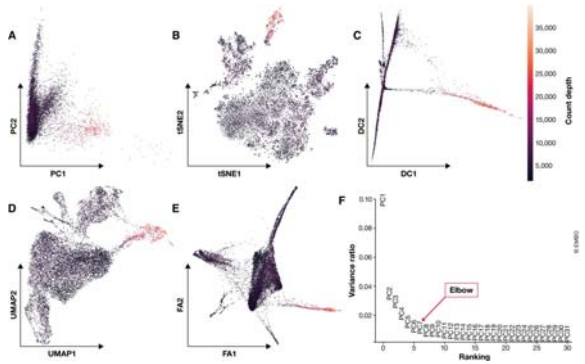
Other dimensionality reduction methods

Cells in 20000 (genes) dimensional space → PCA → Cells in 10-50 principal components space

How can we further summarize these multiple PCAs into just 2 dimensions?

Cells in 10-50 principal components space → tSNE, UMAP, other → Cells in 2D space

Visualization: dimensionality reduction



Current best practices in single-cell RNA-seq analysis: a tutorial

Mol Syst Biol, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

t-Distributed Stochastic Neighbor Embedding (tSNE)

- Takes a set of points in a high-dimensional space and finds a faithful representation of those points in a lower-dimensional space, typically the 2D plane.
- The algorithm is **non-linear** and adapts to the underlying data, performing different transformations on different regions.
- The t-SNE algorithm adapts its notion of "distance" to regional density variations in the data set. As a result, it naturally expands dense clusters, and contracts sparse ones, evening out cluster sizes.
- Distances between clusters might not mean anything.

https://distill.pub/2016/misread-tsne/

UMAP

Uniform manifold approximation and projection

- It is a non linear dimensionality reduction algorithm.
- Preserves the local structure but also the global structure and the continuity of the cell subsets better.
- See PMID: 30531897 for comparison of Seurat and UMAP.

Clustering and Biology: What do you want to learn from the experiment?

- Classify cells and discover new cell populations
- Compare gene expression between different cell populations
- Reconstruct developmental 'trajectories' to reveal cell fate decisions of distinct cell subpopulations

Lots of software available to analyze single-cell RNA-seq data

- **Seurat**
- Monocle
- Scanpy
- Destiny
- See

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

Seurat

https://satijalab.org/seurat/

- Seurat is an R package designed for QC, analysis, and exploration of single cell RNA-seq data.
- Developed and by the Satija Lab at the New York Genome Center.
- It is well maintained and well documented.
- It has a built in function to read 10x Genomics data. It can de-multiplex hash tag data.
- It has implemented most of the steps needed in common analyses.

Analysis Demo

- Goal:
  - To walk you through an example analysis of scRNA-seq data.
    - Exploring the data
    - Performing quality control
    - Identifying cell type subsets.
  - To introduce you to scRNA-seq analysis using the Seurat package.
- We will be analyzing the a dataset of Non-Small Cell Lung Cancer Cells (NSCLC) freely available from 10X Genomics (https://support.10xgenomics.com/single-cell-vdj/datasets/2.2.0/vdj_v1_hs_nsclc_5gex)

Links to Seurat tutorials
Seurat demo

- https://satijalab.org/seurat/vignettes.html
- https://scrnaseq-course.cog.sanger.ac.uk/website/seurat-chapter.html

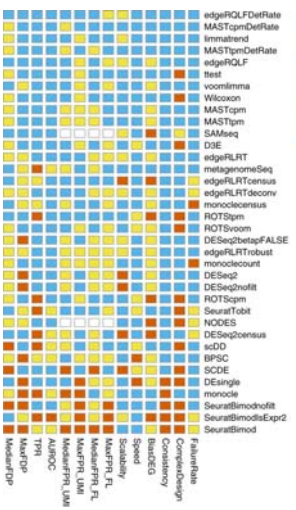
Links to Scanpy tutorials
Scanpy demo

- https://icb-scanpy.readthedocs-hosted.com/en/stable/tutorials.html
- https://github.com/theislab/single-cell-tutorial/blob/master/supplementary_scripts/Splatter-marker-genes-random-data.ipynb
- https://github.com/theislab/single-cell-tutorial/blob/master/latest_notebook/Case-study_Mouse-intestinal-epithelium_1906.ipynb

Differential expression analysis between clusters

- Finds marker genes that will help determine the identity of the clusters
- Since the expression data used to find the clusters and the markers is the same, the P-values are inflated and can lead to an overestimation of marker genes.
- The ranking of genes based on P-values is unaffected and it is a better way of selecting marker genes

Differential expression analysis between conditions



Soneson, C., Robinson, M. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods 15, 255–261 (2018). https://doi.org/10.1038/nmeth.4612

Clustering and Biology:

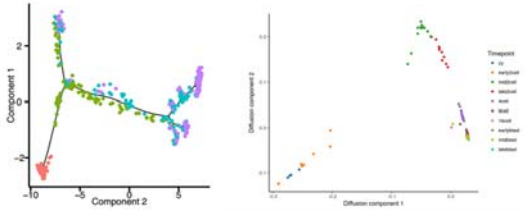
What do you want to learn from the experiment?

- Classify cells and discover new cell populations
- Compare gene expression between different cell populations
- Reconstruct developmental 'trajectories' to reveal cell fate decisions of distinct cell subpopulations

Reconstructing 'trajectories' Pseudotime analysis

Applicable when studying a process where cells change continuously. For example cell differentiation during development, or cell response to a stimulus.

- Monocle
- TSCAN
- Slicer
- Slingshot
- Diffusion map
  - ✓ Scanpy (pytho
  - ✓ destiny (R)



References and resources

- A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. PMID: 28821273
- Current best practices in single-cell RNA-seq analysis: a tutorial. PMID: 31217225
- "Analysis of single cell RNA-seq data" course (Hemberg Group).
- Single cell RNA sequencing - NGS Analysis - NYU
- 2017/2018 Single Cell RNA Sequencing Analysis Workshop (UCD,UCB,UCSF)
- seandavi/awesome-single-cell
- Broad Institute single cell portal
- Tabula Muris (https://tabula-muris.ds.czbiohub.org/)

Upcoming Hot Topics

- Excel tips and tricks: March 12th
- Dimensionality reduction: March