scRNA sequencing technology overview

Zinger Yang Loureiro

August 16, 2021

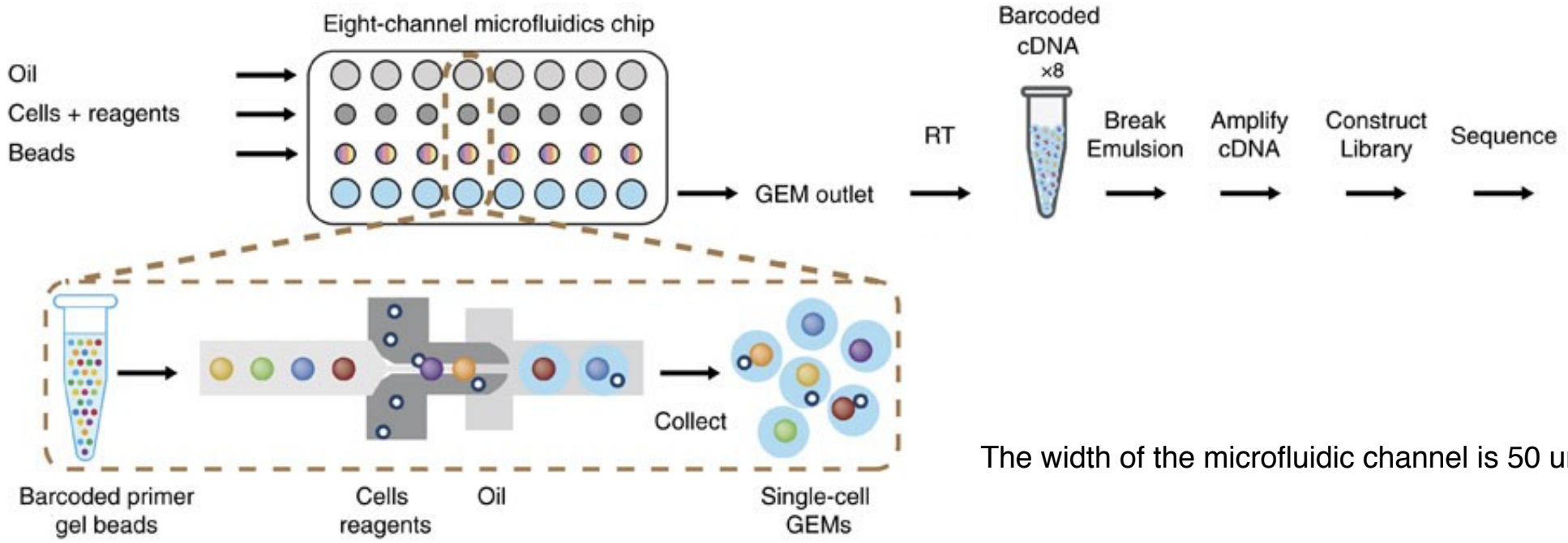
scRNA-seq Library Preparation with 10X Chromium Single Cell **3'** Reagent Kit

Single cells are individually tagged in single vesicles, known as "Gel Beads in Emulsion (GEMs)"

10X genomics

Chromium Controller

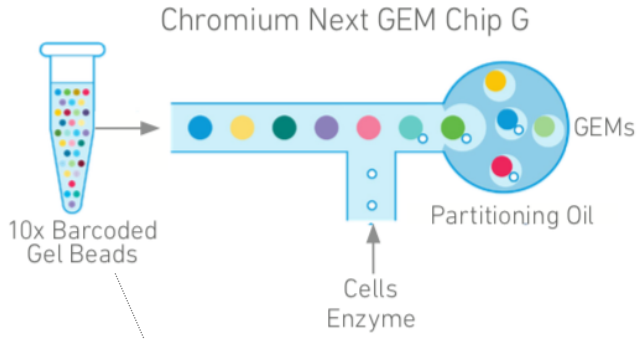
Chip for Chromium Controller

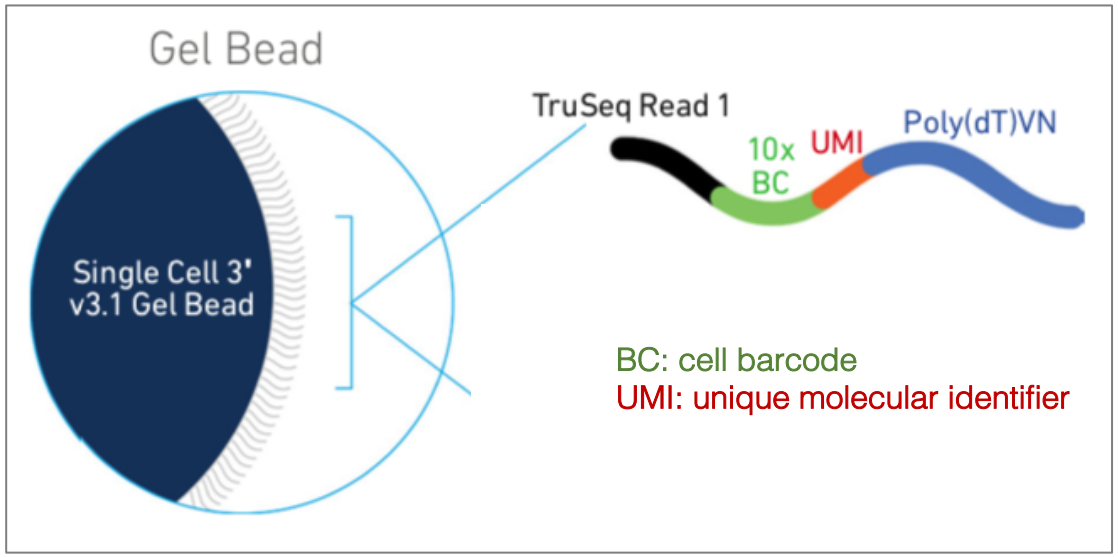


The width of the microfluidic channel is 50 um

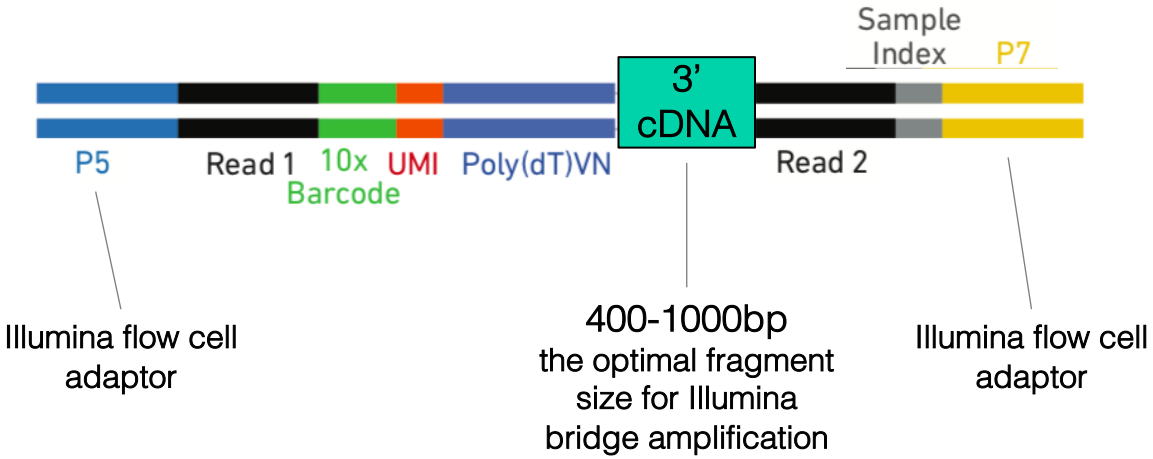
Zheng 2017

scRNA-seq Library Preparation with 10X Chromium Single Cell 3’ Reagent Kit



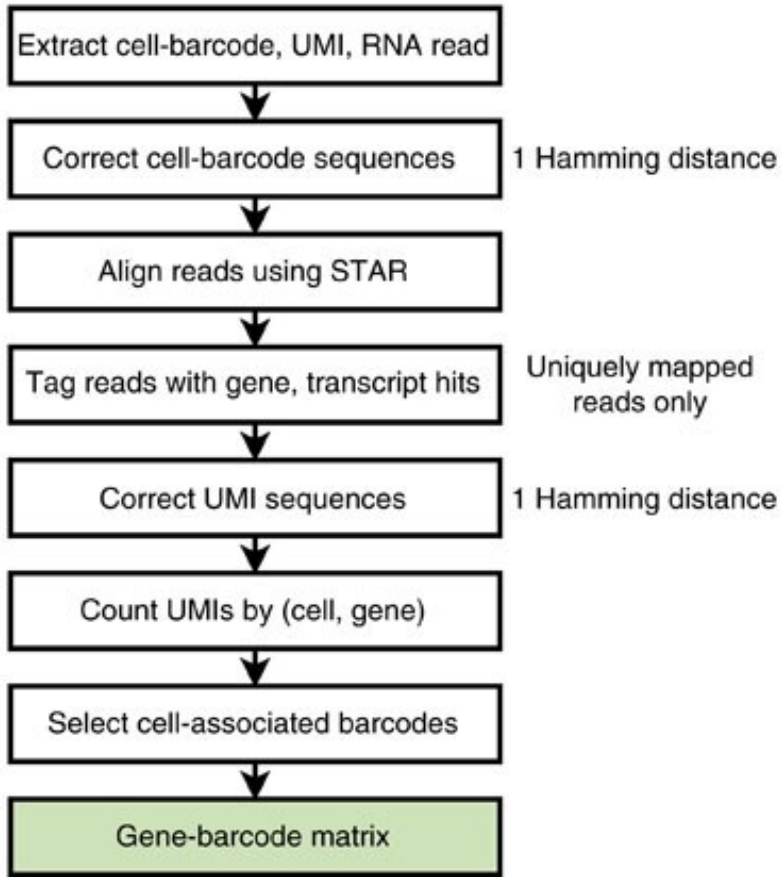


The final library is compatible to Illumina sequencing platform



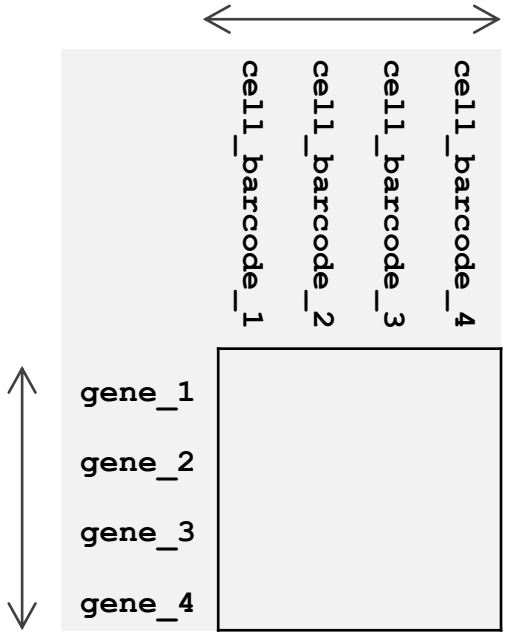
10x Genomics Single Cell Protocols – Cell Preparation Guide Rev C
Chromium Next GEM Single Cell 3′ Reagent Kits v3.1 User Guide Rev C

10X CellRanger processes sequencing results into gene-barcode matrix



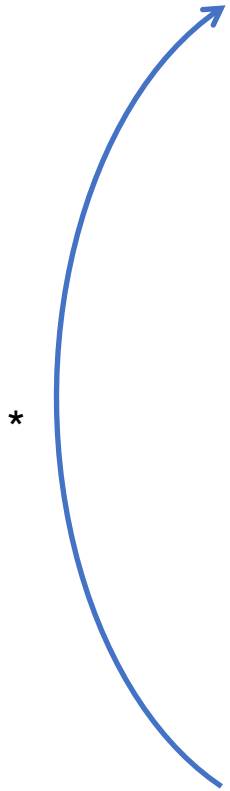
Zheng 2017

- UMI: unique molecular identifier
- STAR: Spliced Transcripts Alignment to a Reference.
  - Further explanation on STAR aligner. https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html



~30000 cells x 20000 genes

Data processing outline *multiple iterations

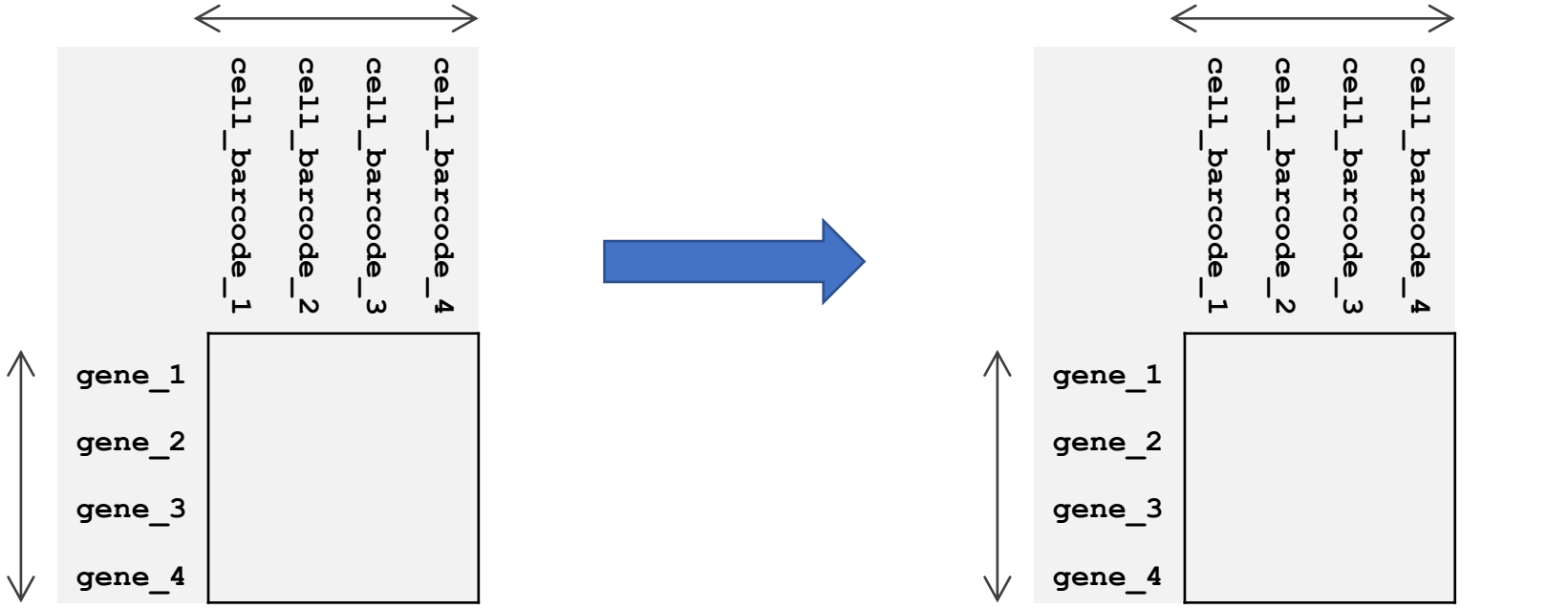


Input: count matrices

1. Quality control
   - Filter out bad quality cells based on number of unique genes detected per cell
   - Filter to genes detected in a minimum number of cells
2. Normalization
   - Normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (default:10,000), and log-transforms the result.
   - Assumption: all cells in the dataset initially contained an equal number of mRNA molecules
3. Feature (gene) selection
   - Principal component analysis (PCA) to highlight greatest sources of variation presented in the dataset
4. Visualization
   - Uniform approximation and projection method (UMAP) projection of cells with the top principal components
5. Cell cluster determination
   - Visual inspection of UMAP local connectivity

Data selection for analysis

Quality control removes cells/cell clusters with technical noise (doublets, incomplete lysis, etc.)
Analysis focuses on data points of top variable genes (adjustable; Seurat default = 2000)
PCA dimension reduction is typically done prior to tSNE/UMAP



~30000 cells x 20000 genes
= 600,000,000 data points

~24000 cells x 2000 genes
= 48,000,000 data points

Seurat provides very good tutorials https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

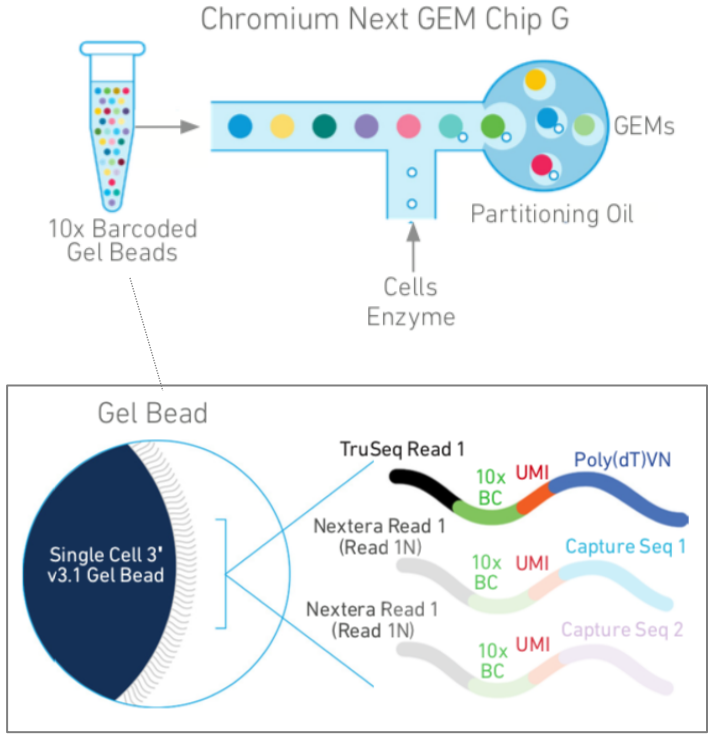
Analysis iterations

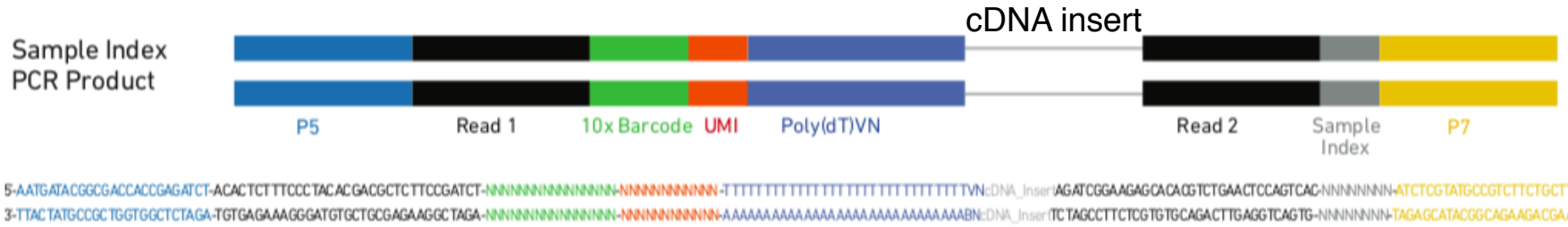
- Iteration 1:
  - Initial data processing with CellRanger
  - Delivery: QC reports, loupe browser file (4 files: BMY/BMR/BMS combined, separated)
- Iteration 2 (whole dataset):
  - Seurat setup
  - Initial data QC with Seurat
  - Delivery: simple R script
- Future iterations:
  - Cell / gene inclusion criteria
  - Subsets:
    - BMY, BMR, BMS separated
    - MSC subtypes
    - Immune cell subtypes
- Loupe browser : https://support.10xgenomics.com/single-cell-gene-expression/software/visualization/latest/tutorial-navigation
- Task and version tracking: https://github.com/zingery/TammyNguyen_SingleCell

Backup slides

Library Preparation with 10X

1. Preparation of single cell suspension
2. Single cell droplet generation
3. Reverse transcription, cDNA amplification
4. Library construction : Fragmentation, End Repair & A-tailing, Adaptor ligation, Sample Indexing





10x Genomics Single Cell Protocols – Cell Preparation Guide Rev C
Chromium Next GEM Single Cell 3′ Reagent Kits v3.1 User Guide Rev C

Steps

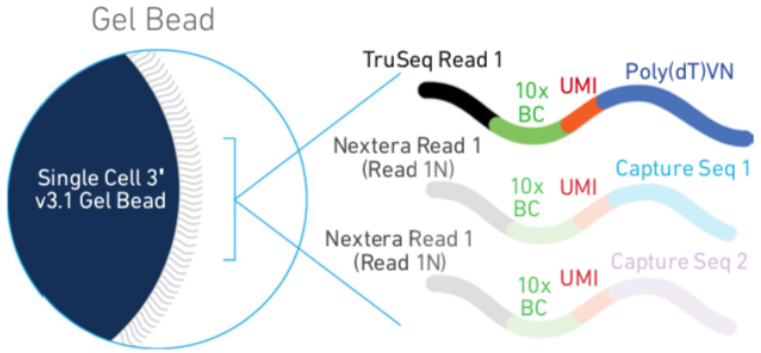
1. Preparation of single cell suspension
2. Single cell droplet generation
3. Reverse transcription
4. Clean up by nucleic acid capture
5. cDNA amplification
6. cDNA cleanup by size selection
7. Library construction : Fragmentation, End Repair & A-tailing
8. Clean up - Double Sided Size Selection
9. Adaptor ligation
10. Post Ligation Cleanup by size selection
11. Sample Index PCR
12. Post Library Construction QC
13. Post Sample Index PCR Double Sided Size Selection

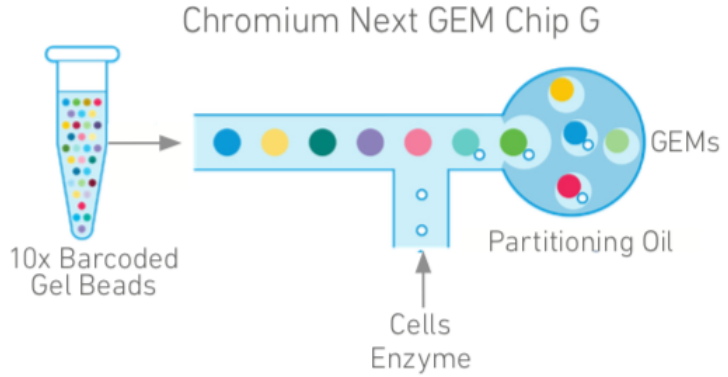
1. Preparation of single cell suspension

- Cell dissociation with Trypsin-EDTA
- Cell resuspension with 1X PBS + 0.04% BSA (400 µg/ml)
- Filter cells with a cell strainer
- Centrifuge at 500g for 5 min. Remove supernatant without disrupting the cell pellet.
- Add the appropriate volume of 1X PBS with 0.04% BSA to achieve the target cell concentration.
- Once the target cell concentration is obtained, place the cells on ice.

2. Single cell droplet generation

- Input:
  - Single Cell 3′ v3.1 Gel Beads
  - Partitioning Oil
  - Cell mix:
    - cells (amount based on targeted cell recovery)
    - reverse transcription (RT) enzyme
    - reagent + template switch oligo (TSO)
- Processing:
  - Chromium Single Cell Controller
- Output:
  - Emulsified droplets ready for RT reaction

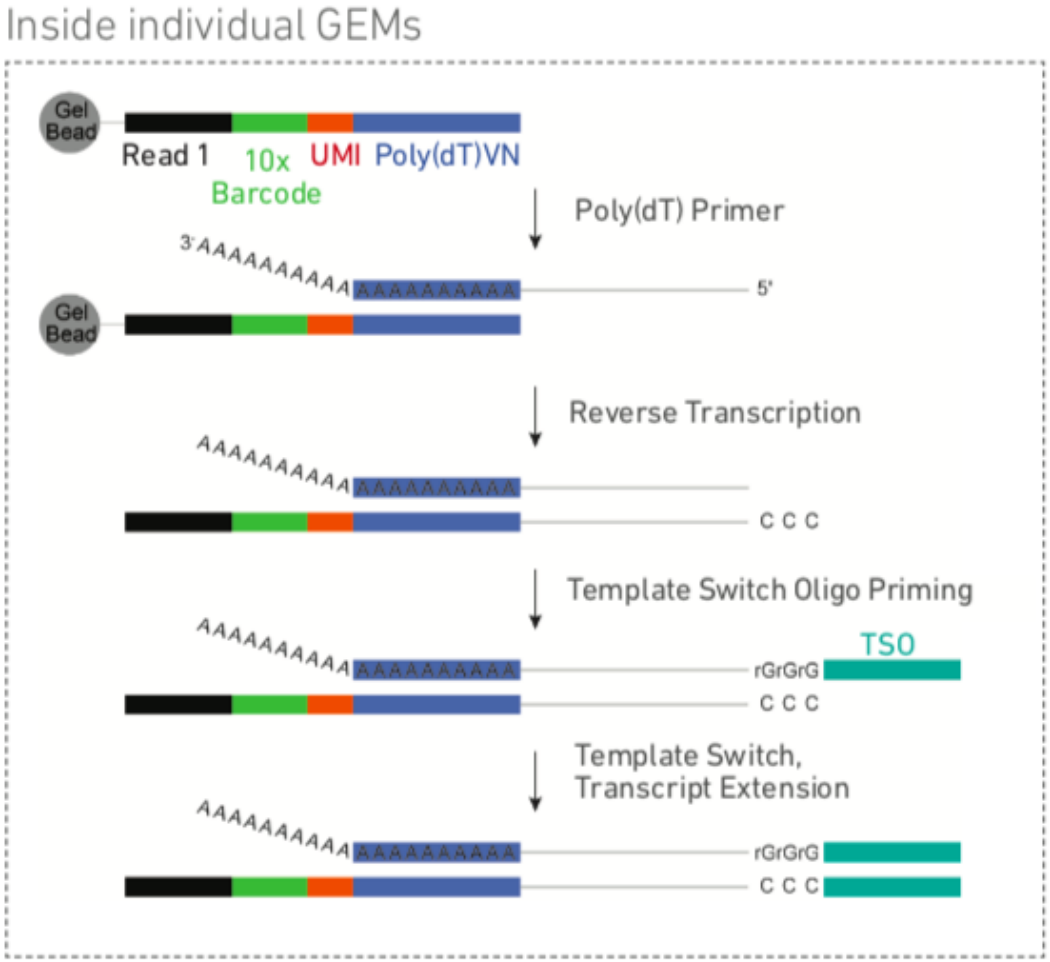




To achieve single cell resolution, cells are delivered at a limiting dilution, such that the majority (~90-99%) of generated GEMs contain no cell, while the remainder largely contain a single cell.
GEM: Gel Beads-in-emulsion

3. Reverse transcription

- Input:
  - Emulsified droplets
- Processing:
  - Thermocycler
    - 53°C for 45min
    - 85°C for 5 min
- Output:
  - Barcoded cDNA from poly-adenylated mRNA in reagent mixture



The TSO (template switch oligo) is an oligo that hybridizes to untemplated C nucleotides added by the reverse transcriptase during reverse transcription. The TSO adds a common 5' sequence to full length cDNA that is used for downstream cDNA amplification.

Read 1 is TruSeq Read 1 primer

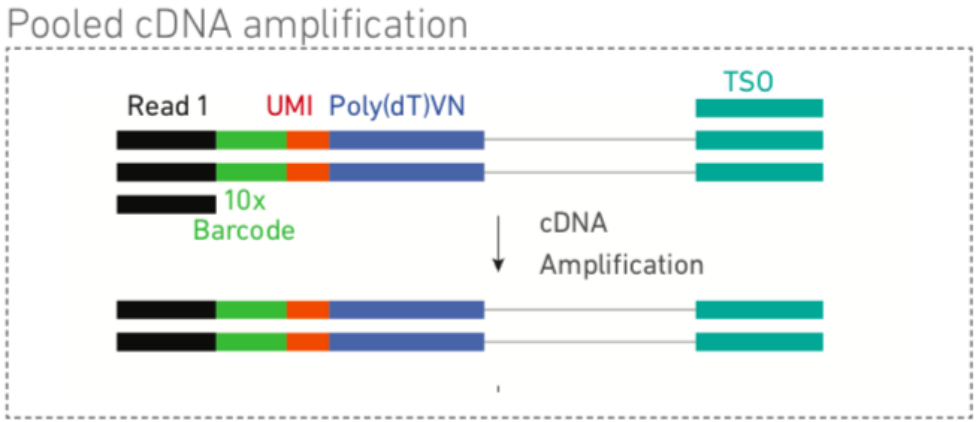
4. Clean up by nucleic acid capture

- Input:
  - cDNA in reagent mixture containing partition oil, recovery reagent, primers

- Processing:
  - Magnetic nucleic acid capture.
    - Dynabeads® MyOne™ SILANE are uniform, monosized magnetic beads, 1 µm in diameter. They are composed of highly crosslinked polystyrene with evenly distributed magnetic material.
- Output:
  - Barcoded cDNA

5. cDNA amplification

- Input:
  - Barcoded cDNA
  - Primers
- Processing:
  - Thermocycler
    - Initialization, denaturation 98°C
    - Annealing 63°C
    - Elongation 72°C
- Output:
  - Sufficient barcoded cDNA for library construction

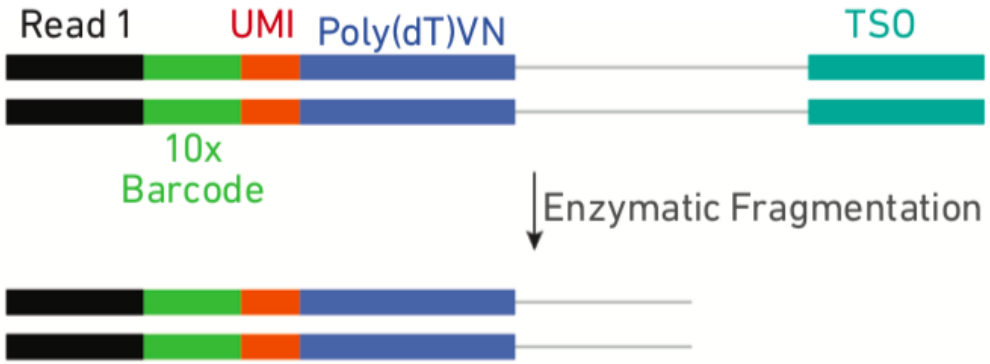


6. cDNA cleanup

- Input:
  - Sufficient barcoded cDNA for library construction
- Processing:
  - 0.6X SPRI bead selection
- Output:
  - Cleaned up barcoded cDNA for library construction

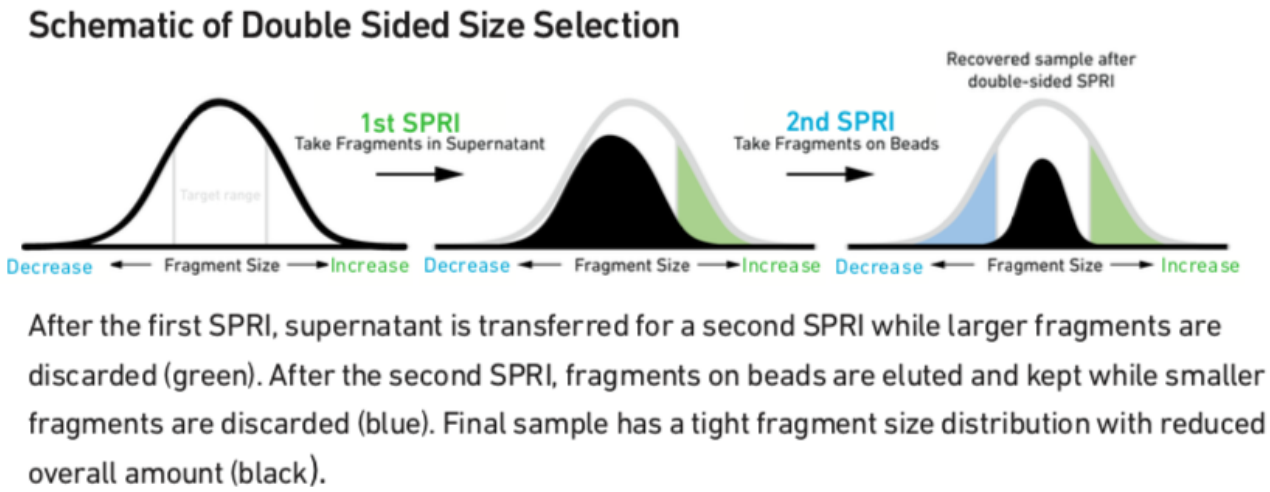
7. Library construction : Fragmentation, End Repair & A-tailing

- Input:
  - 10 µl purified cDNA sample
- Processing:
  - Thermocycler
    - Fragmentation at 32°C for 5 min
    - End Repair & A-tailing at 65°C for 30 min
- Output:
  - 3’ cDNA fragments



8. Clean up - Double Sided Size Selection

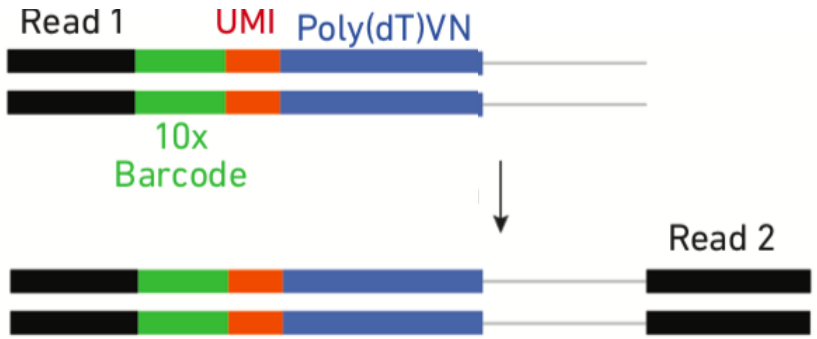
- Input:
  - 3’ cDNA fragments
- Processing:
  - 0.6X SPRI supernatant and 0.8X SPRI select
- Output:
  - Cleaned up 3’ end cDNA fragment



We look for traces that range between 400 – 1000bp with a significant amount of inserts that are **400 – 600bp** in length. Inserts in this size range are optimal for cluster formation in Illumina® flowcells.
(10x Genomics. CG000061 Rev A Technical Note – SPRIselect:DNA Ratios Affect the Size Range of Library Fragments (2016))

9. Adaptor ligation

- Input:
  - 3’ cDNA fragments
  - Adaptor oligos
- Processing:
  - Thermocycler
    - Ligation at 20°C for 15 min
- Output:
  - cDNA fragments with adaptor

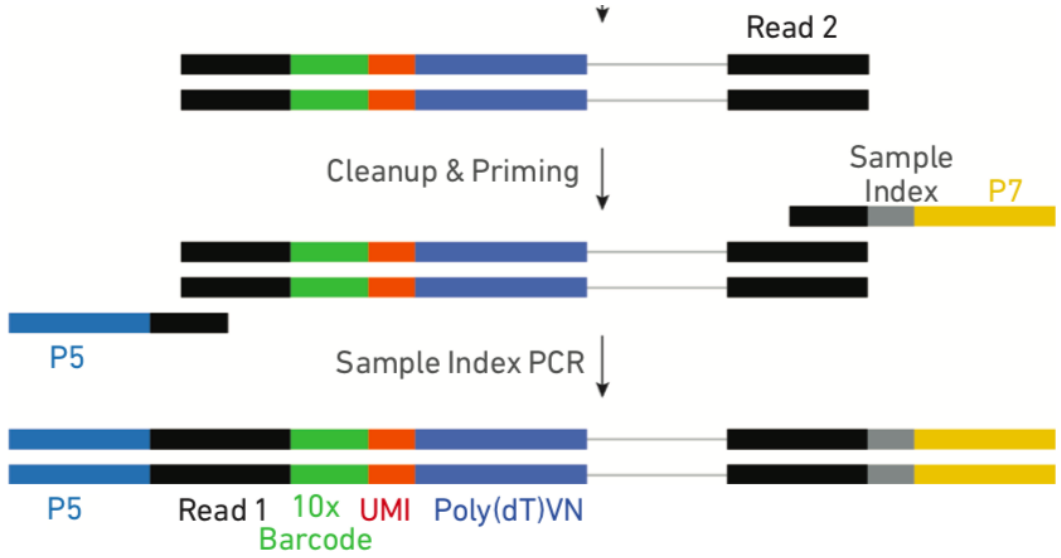


10. Post Ligation Cleanup

- Input:
  - cDNA fragments
- Processing:
  - 0.8X SPRI bead selected
- Output:
  - Cleaned up cDNA fragment

11. Sample Index PCR

- Input:
  - cDNA fragment
  - Sample Index
- Processing:
  - Thermocycler
- Output:
  - cDNA with sample index and Illumina P5/P7 primers



12. Post Library Construction QC

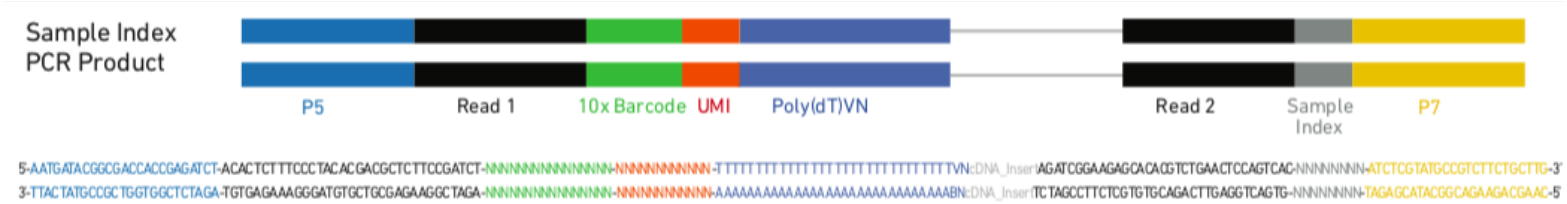
- Processing:
  - Agilent Bioanalyzer : obtain measurement of the cDNA concentration, average fragment size, as the insert size for library quantification.

13. Post Sample Index PCR Double Sided Size Selection

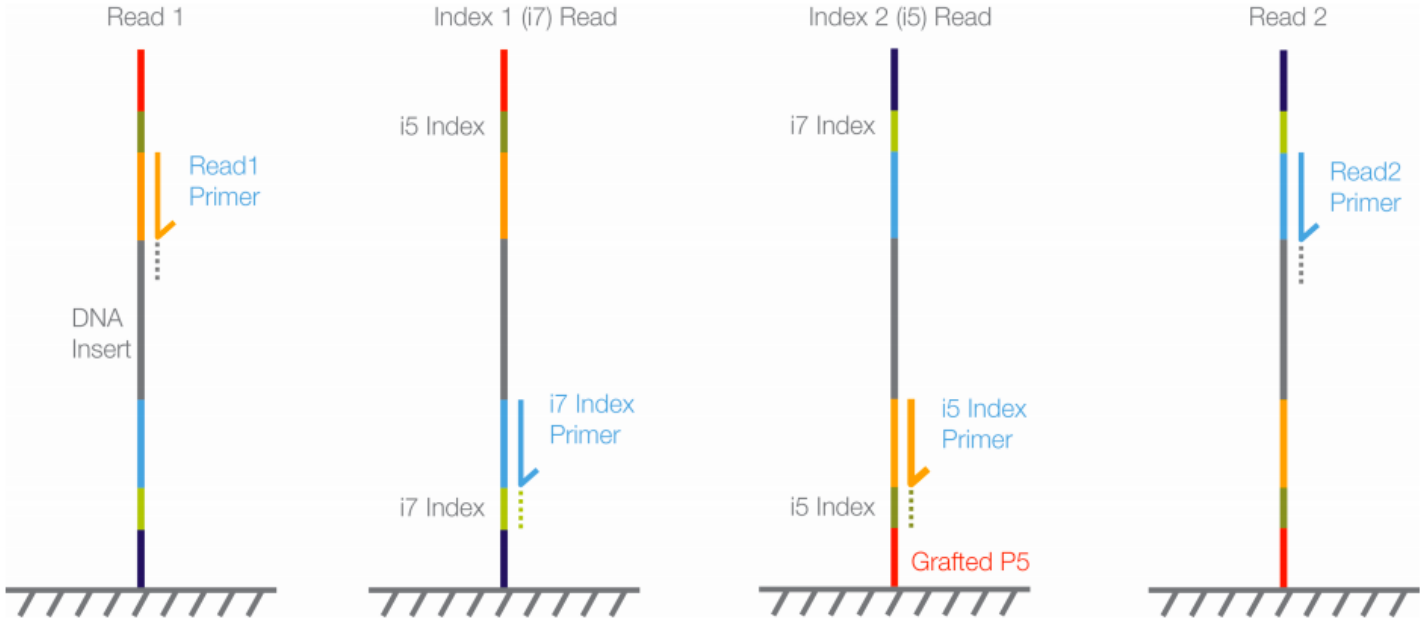
- Input:
  - Post sample index cDNA
- Processing:
  - 0.6X SPRI supernatant and 0.8X SPRI select
- Output:
  - Cleaned up cDNA

14. Sequencing

10X output is compatible with specified Illumina sequencing platforms



Dual-Indexed Workflow on an Illumina NextSeq or HiSeq 3000/4000 Paired-End Flow Cell



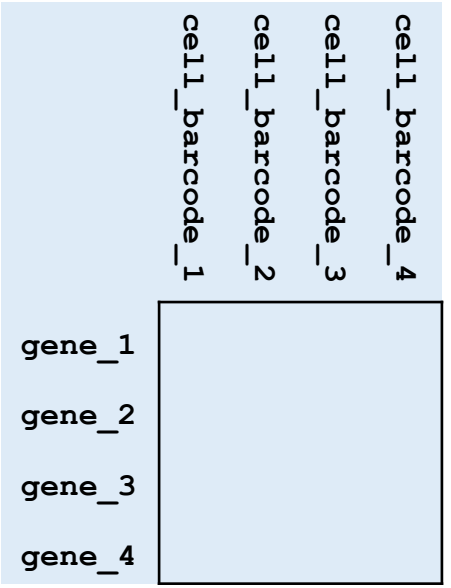
- Cluster Generation
  - P5 or P7 hybridization
  - Template copying
  - Bridge amplification
  - Keeping forward strand
- Sequencing
  - Sequencing by synthesis (# cycles = # reads)
    - Read1 with primer, read index1 with primer
    - Folds over, add reverse strand
    - Keeping reverse strand
    - Read 2 with primer
- Base calling to fastq (bcl2fastq)

Overview of Indexed Sequencing on the NextSeq, MiSeq, and HiSeq Platforms (Part # 15057455 Rev. B)

Data processing outline *multiple iterations

1. FASTQ processing to count matrices (10X genomics CellRanger pipeline)
2. Quality control
   - Filter out bad quality cells based on number of unique genes detected per cell
   - Filter to genes detected in a minimum number of cells
3. Normalization
   - Normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (default:10,000), and log-transforms the result.
   - Assumption: all cells in the dataset initially contained an equal number of mRNA molecules
4. Feature (gene) selection
   - Principal component analysis (PCA) to highlight greatest sources of variation presented in the dataset
5. Visualization
   - Uniform approximation and projection method (UMAP) projection of cells with the top principal components
6. Cell cluster determination
   - Visual inspection of UMAP local connectivity

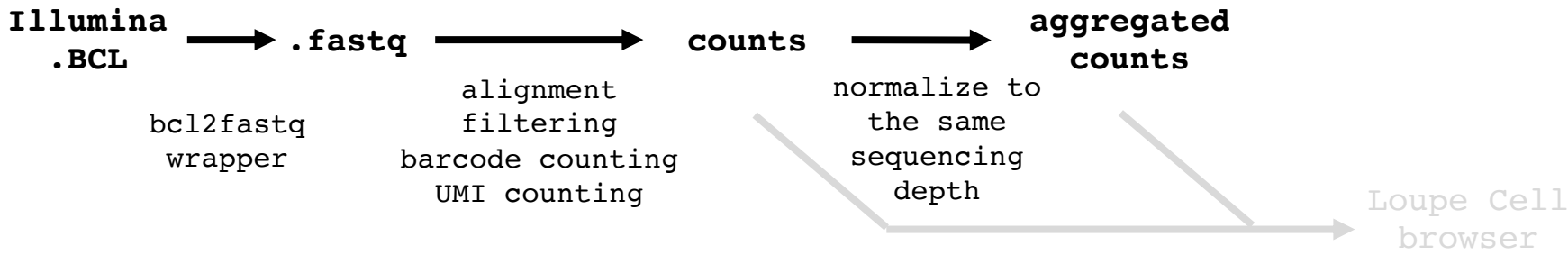
*



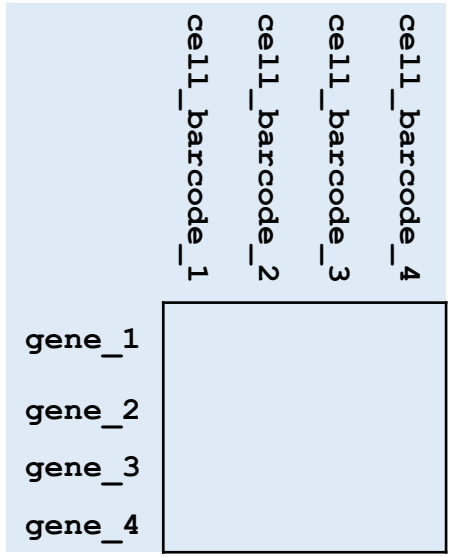
Raw data processing

(CellRanger data process (cellranger/3.1.0))

Cell Ranger is a set of analysis pipelines that process Chromium single-cell RNA-seq output to align reads, generate feature-barcode matrices and perform clustering and gene expression analysis.



- Filtered feature-barcode matrices containing only cellular barcodes



Summarized information from https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

Analysis iterations

- Iteration 1:
  - Initial data processing with CellRanger
  - Export loupe browser file
- Iteration 2:
  - Seurat setup
  - Initial data QC with Seurat

- Task tracking:
  - https://github.com/zingery/TammyNguyen_SingleCell

Quality Control

1. Assess number of reads, number of unique genes detected per cell
2. Assess the percentage of reads mapping to mitochondrial genes
3. Filter out bad quality cells based on (1) an (2)
- Additional: Filter to genes detected in a minimum number of cells

- Likely reasons for poor quality:
  - Dying cells
  - Other lysed cells whose mRNA contaminated the cell suspension prior to library construction
  - Doublets (unexpectedly high counts and a large number of detected genes)

- Note:
  - Cells with a comparatively high fraction of mitochondrial counts may be involved in respiratory processes.
  - Cells with low counts and/or genes may correspond to quiescent cell populations
  - Cells with high counts may be larger in size
  - It may be necessary to revisit quality control decisions multiple times during analysis

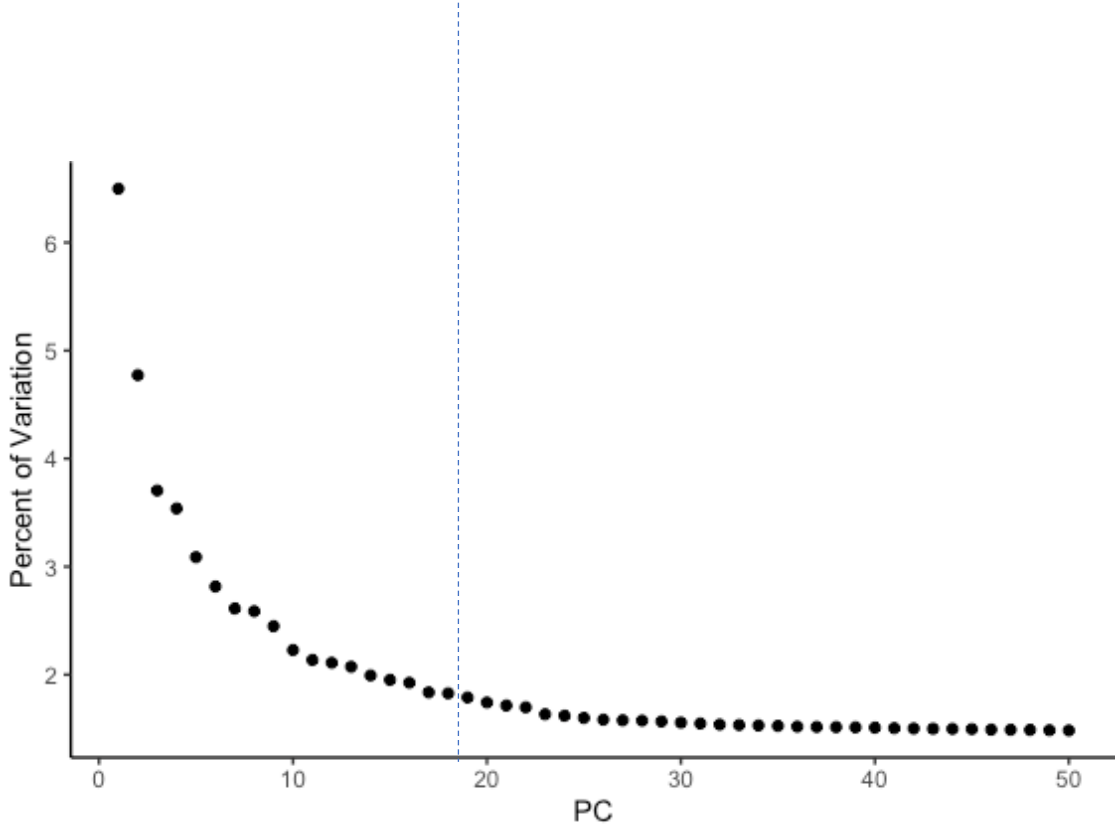
Normalization

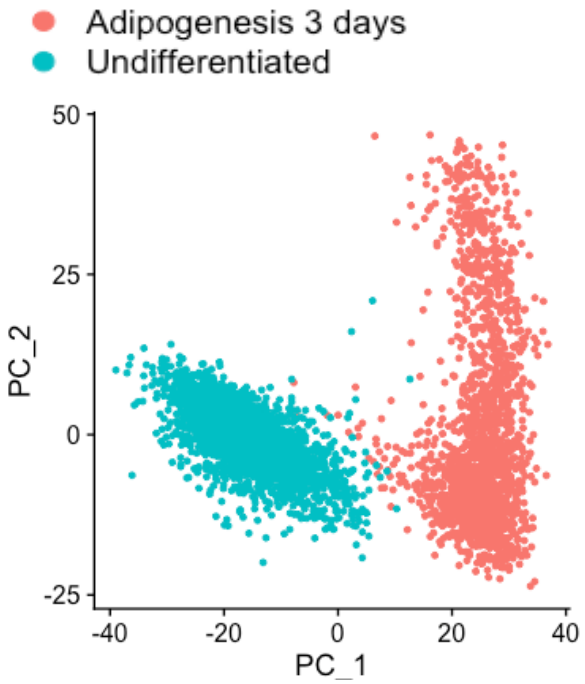
- Seurat: (Stuart… Satija, 2019)
  - Normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms the result.
  - Assumption: all cells in the dataset initially contained an equal number of mRNA molecules

Gene scaling

- gene counts can be scaled to improve comparisons between genes. Gene normalization constitutes scaling gene counts to have zero mean and unit variance (z scores)

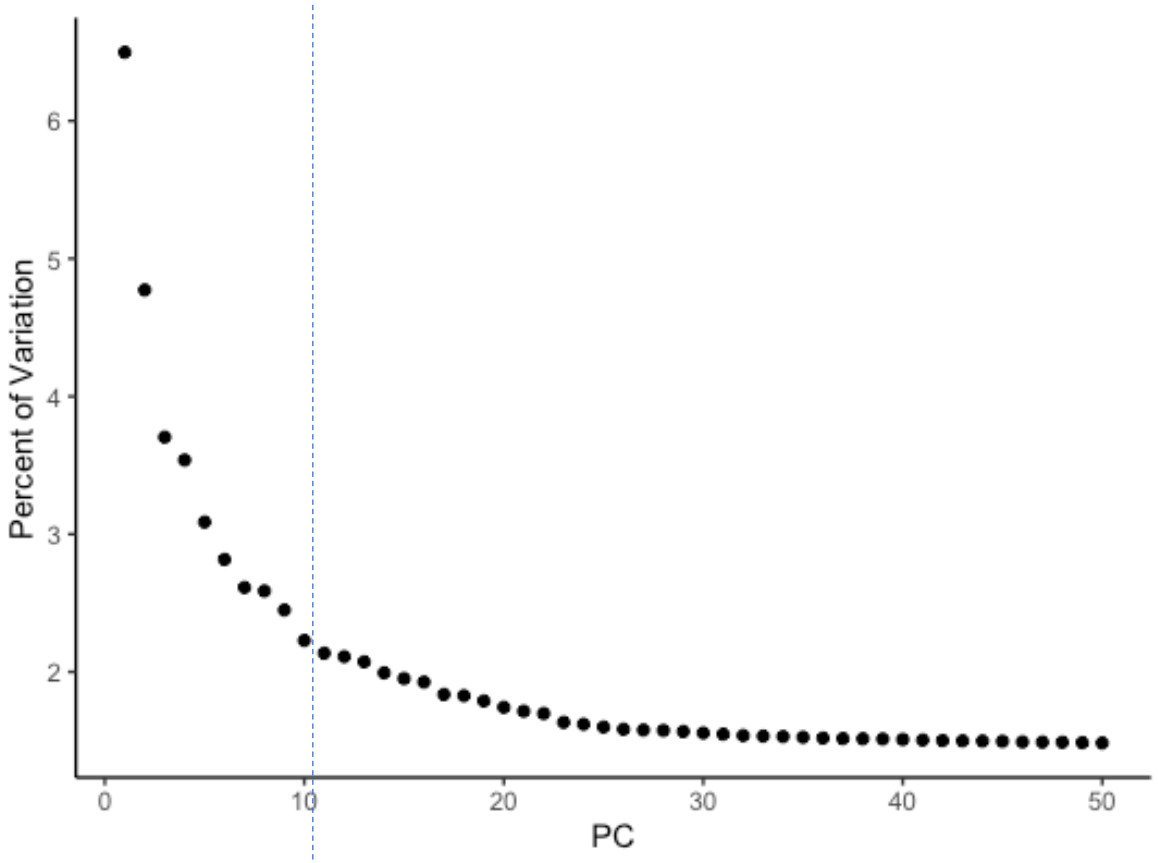
Choosing number of PC's
(Undifferentiated + Adipogenesis 3day)

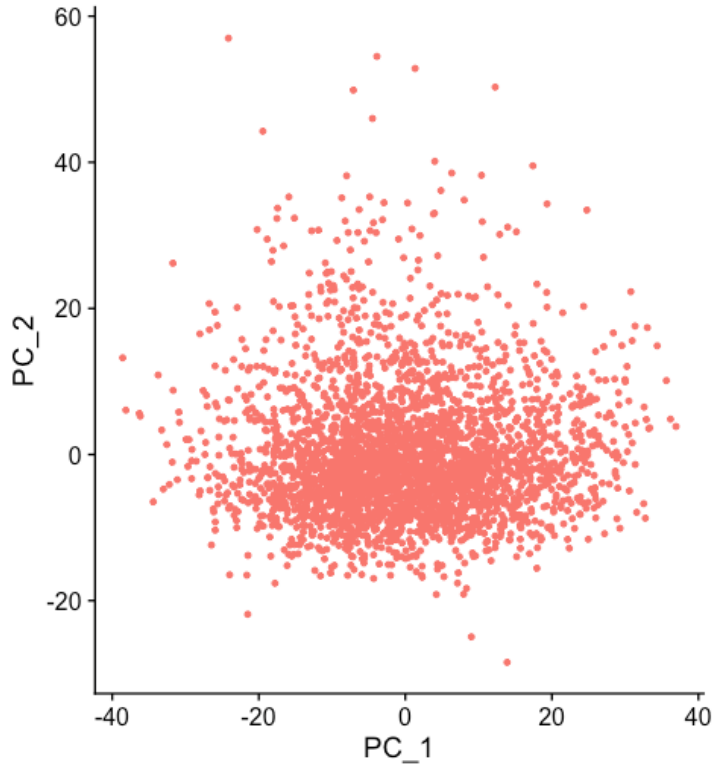




PCA was performed using all features

Choosing number of PC's
(Undifferentiated)





PCA was performed using all features

Seurat identification of highly variable features

- Choosing genes solely based on their log-normalized single-cell variance fails to account for the mean-variance relationship that is inherent to single-cell RNA-seq. Therefore, (Seurat) applied a **variance-stabilizing transformation (vst)** to correct for this
  - computed the mean and variance of each gene using the unnormalized data
  - applied log10-transformation to both
  - fit a curve to get an estimator of variance given the mean of a feature
  - perform the variance stabilizing transformation using the global fit to obtain a standardized feature count

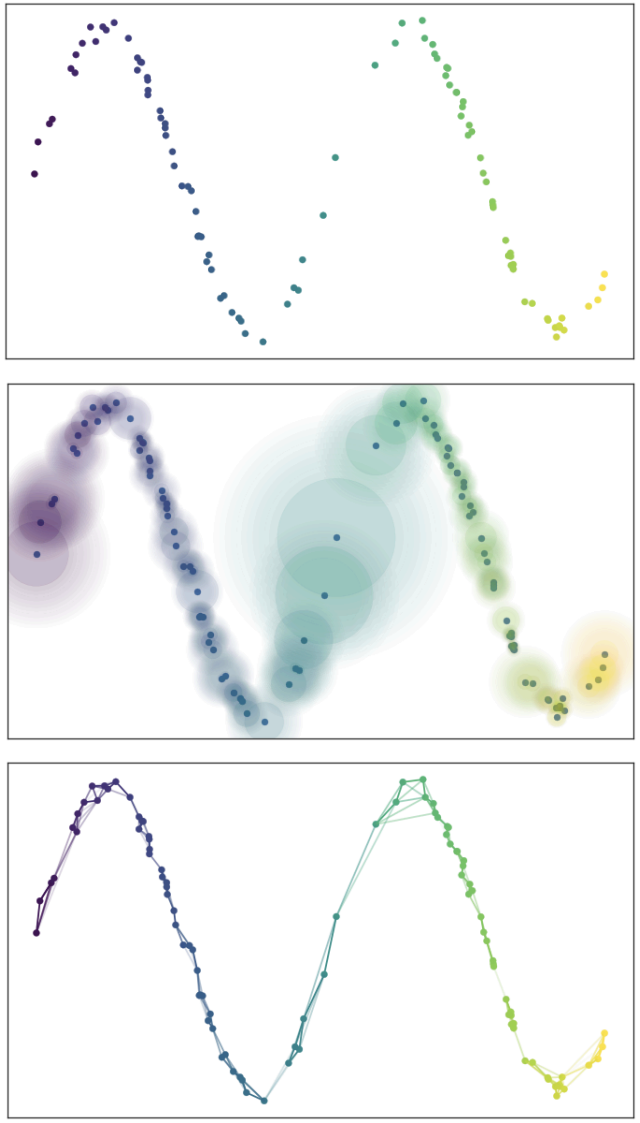
Given the expected variances, we performed the transformation

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i},$$

where z_{ij} is the standardized value of feature i in cell j, x_{ij} is the raw value of feature i in cell j, $\bar{x}_i$ is the mean raw value for feature i, and σ_i is the expected standard deviation of feature i derived from the global mean-variance fit.

  - computed the variance of standardized values across all cells
  - rank the features to identify the highest standardized variance

UMAP is a topological data analysis technique for dimension reduction



Starting dataset

Find a unit ball stretched to the k-th nearest neighbors of the point, with the fuzzy confidence decay in terms of distance beyond the first nearest neighbor.

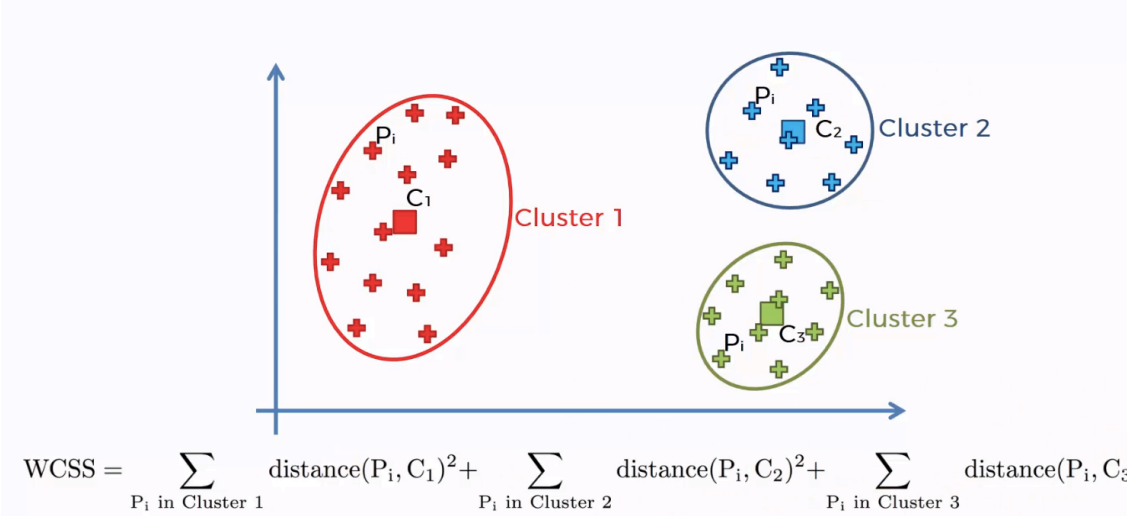
(UMAP prioritizes local distance over long range distances)

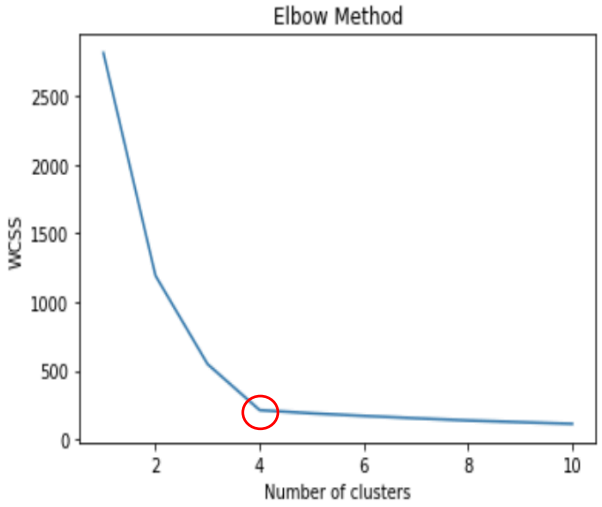
Build a weighted graph with the local metric. Low dimensional representation via a force directed layout algorithm

McInnes, L., Healy, J., and Melville, J. arXiv, (2018)
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

Within-Cluster-Sum of Squared Errors

- WCSS is the sum of squares of the distances of each data point in all clusters to their respective centroids (the arithmetic mean position of all the points in the cluster).

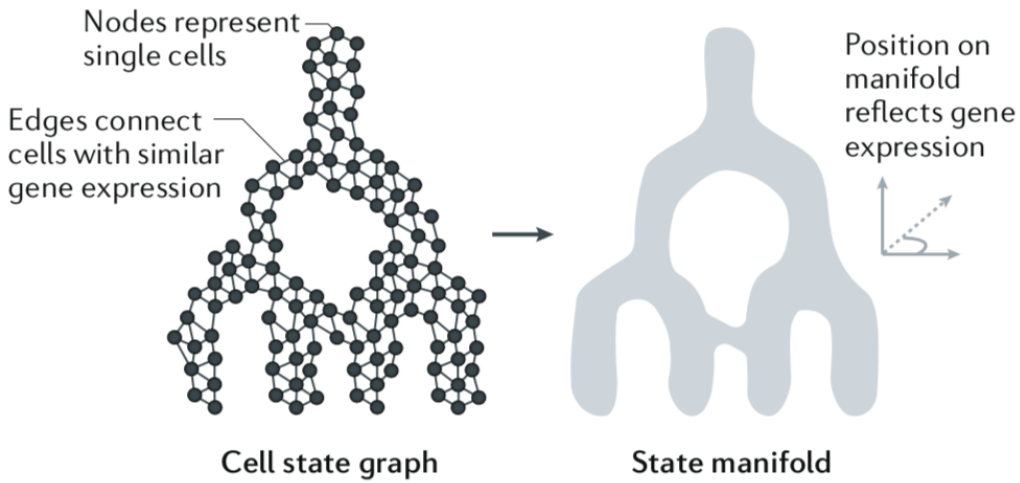




- Using the "elbow" as a cutoff point is a common heuristic in mathematical optimization to choose a point where diminishing returns are no longer worth the additional cost. In clustering, this means one should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. (https://en.wikipedia.org/wiki/Elbow_method_(clustering))

Trajectory inferences

- Trajectory inference methods interpret single-cell data as a snapshot of a continuous process. This process is reconstructed by finding paths through cellular space that minimize transcriptional changes between neighbouring cells

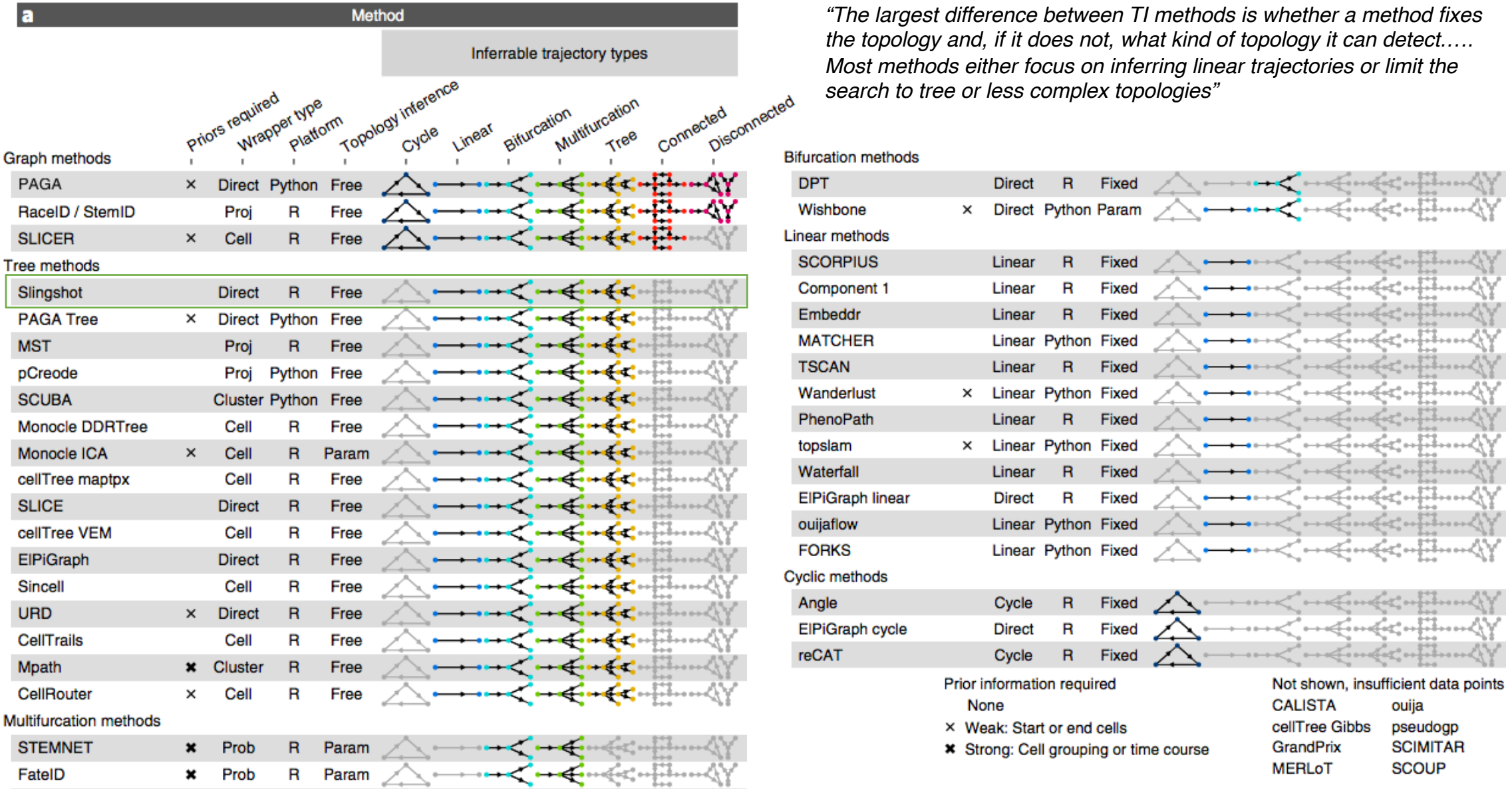


"Two-dimensional representations… can be misleading, as they distort high-dimensional structures upon 'flattening' them, and in some cases algorithms force tree-like visual layouts that may further distort the original structure.… ***The dynamics predicted from cell state snapshots should thus be considered hypotheses***"

Wagner, D.E. & Klein,A.M., *Nat Rev Genetics*, (2020)

There are 70+ trajectory inference software supports various trajectory types

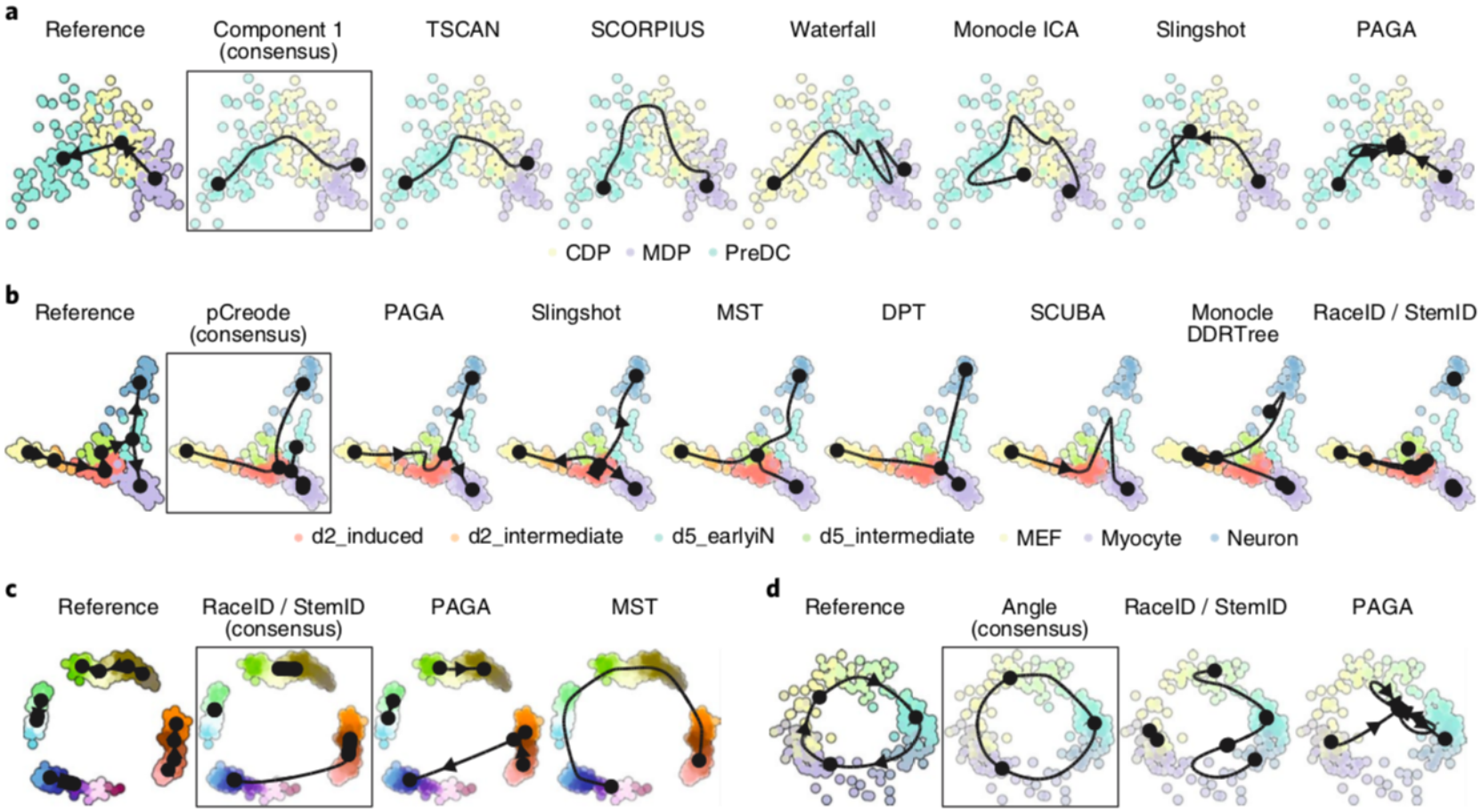
"The largest difference between TI methods is whether a method fixes the topology and, if it does not, what kind of topology it can detect….. Most methods either focus on inferring linear trajectories or limit the search to tree or less complex topologies"



Saelens, W…. Saeys, Y. *Nat Biotechnology,* (2019)

A large scale evaluation of 45 trajectory inference methods revealed disagreeing predictions

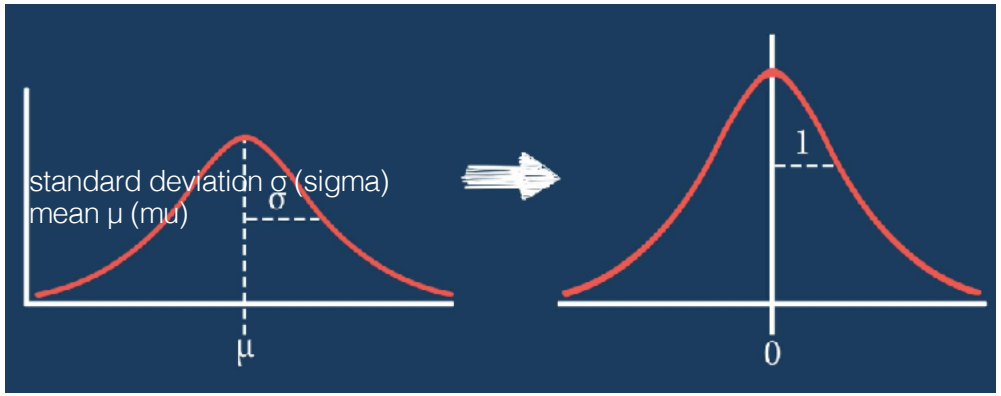
“method performance was very variable across datasets, indicating that there is no ‘one-size-fits-all’ method that works well on every dataset “
“almost every method returned a unique set of outputs”



Saelens, W…. Saeys, Y. *Nat Biotechnology,* (2019)

Unit variance scaling

- “Rows are centred; unit variance scaling is applied to rows.“
  - Row centering: subtracting mean from each row
  - Unit variance scaling: a difference of 1 means that the values are one standard deviation away from each other



Picture credit: https://medium.com/@swethalakshmanan14/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff