Single Cell RNASeq

Theory and Practice

Overview

Bulk RNASeq

- Library preparation
- Analysis methods
  - Normalization strategies
  - Differential gene expression
  - Linear models

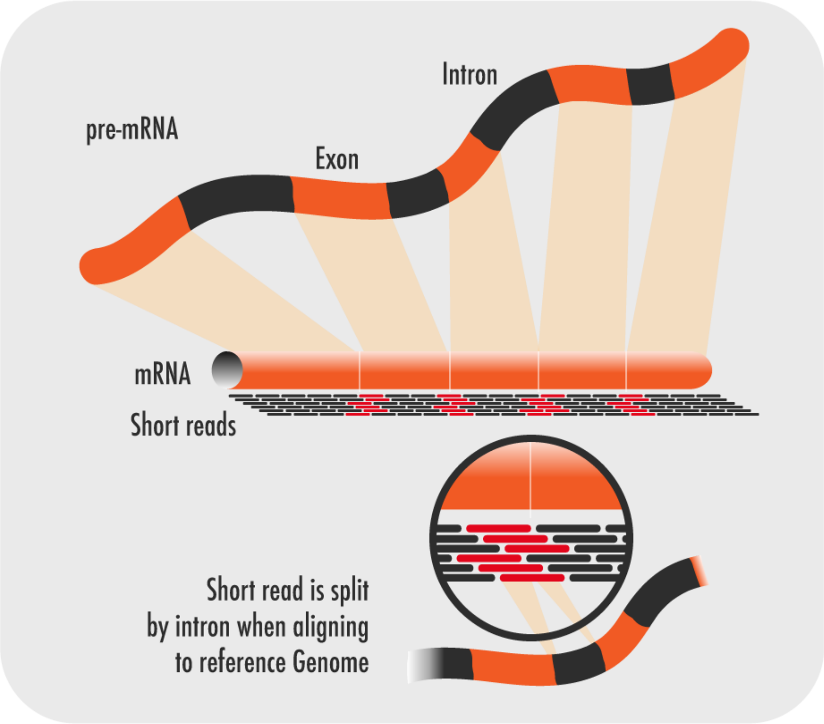
Single cell RNASeq

- Library preparation
  - Demux and barcoding
- Analysis methods
  - Clustering
  - Cluster marker identification
  - Differential gene expression

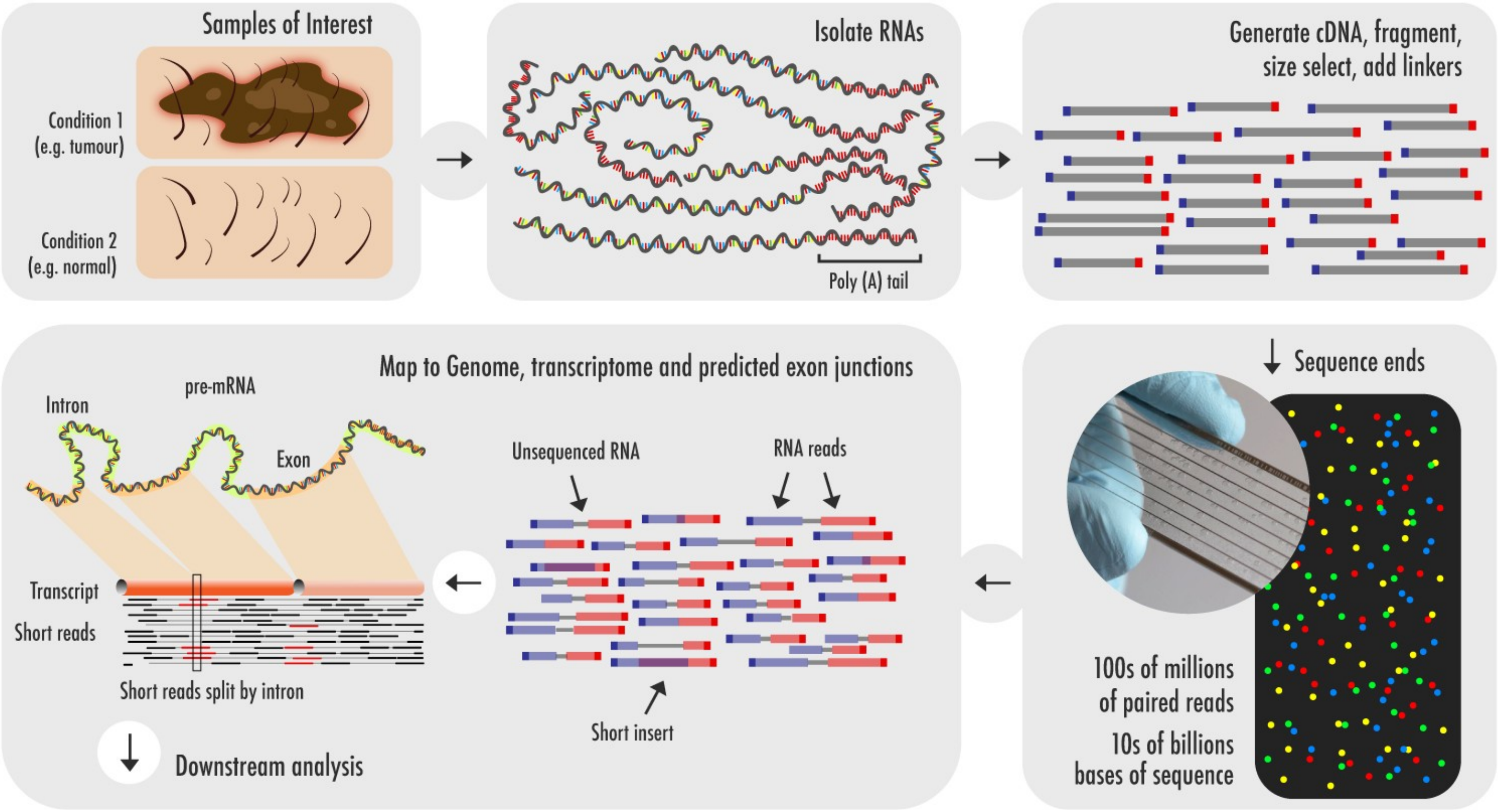
To understand single cell RNASeq,
you need to first understand bulk RNASeq.

RNASeq

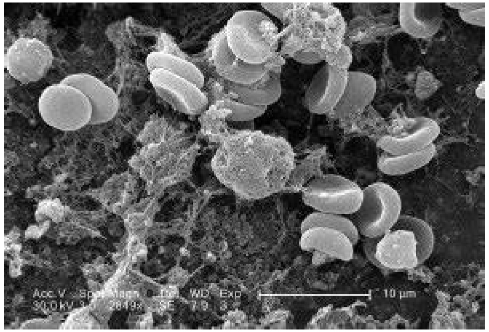
- Using NGS technology to sequence RNA transcripts
- Typically refers to the sequencing of mRNA
- Different RNA species (i.e. miRNA, snoRNA, tRNA) require different preparation protocol
- Any type of RNA from any sample sources, such as cell, body fluid, stool, water, etc. can be the sequenced
- Sample from different sample sources, such as cell, body fluid, stool, water, etc, require different extraction method



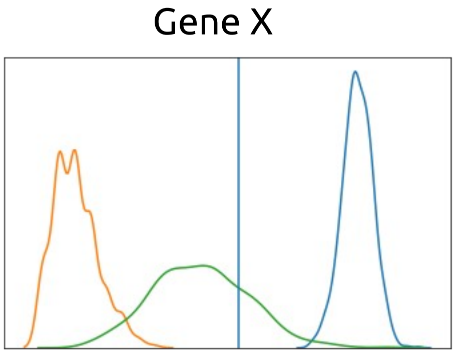
RNASeq Experiment Workflow

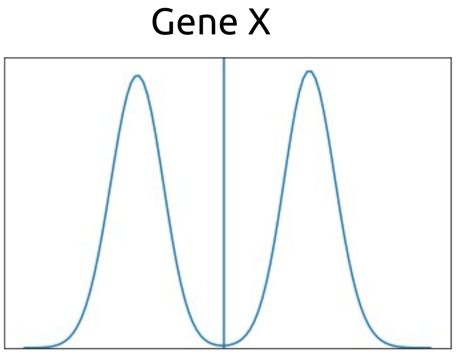


RNASeq: mean expression



1 single
expression value



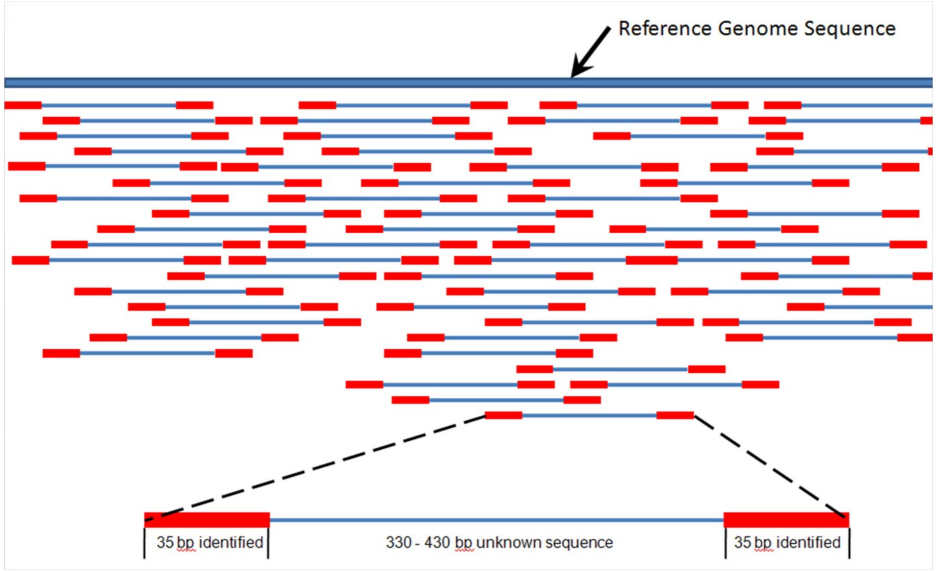


Mean expression

Methods to calculate counts

Alignment based

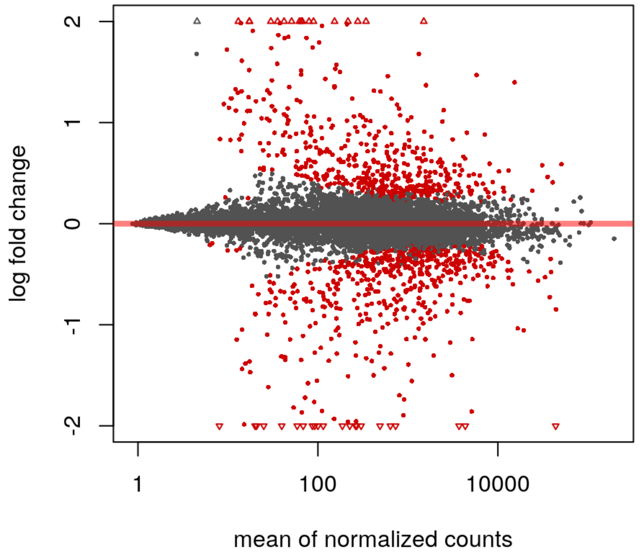
- Align reads
  - STAR, GSNAP, HISAT2
- Count reads aligned to genes
  - RSEM, featureCounts



DGE

Differential Gene Expression (DGE)

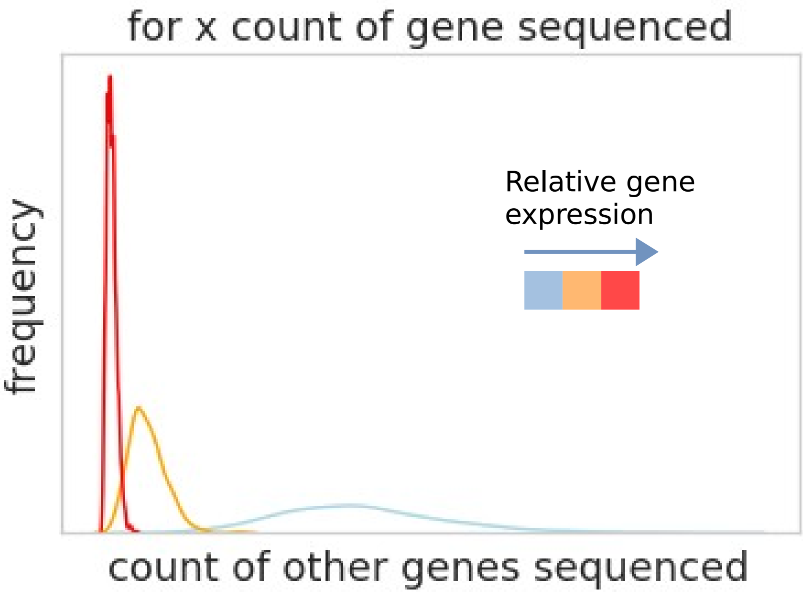
- DeSEQ2, VOOM, EdgeR, etc.
- For inter-sample comparison
  - Relative expression between groups
  - Library size normalization
    - So all libraries are on the same scale



Model count distribution as a negative binomial

Distribution of expression as discrete events

To sequence a read, you did not sequence another read.



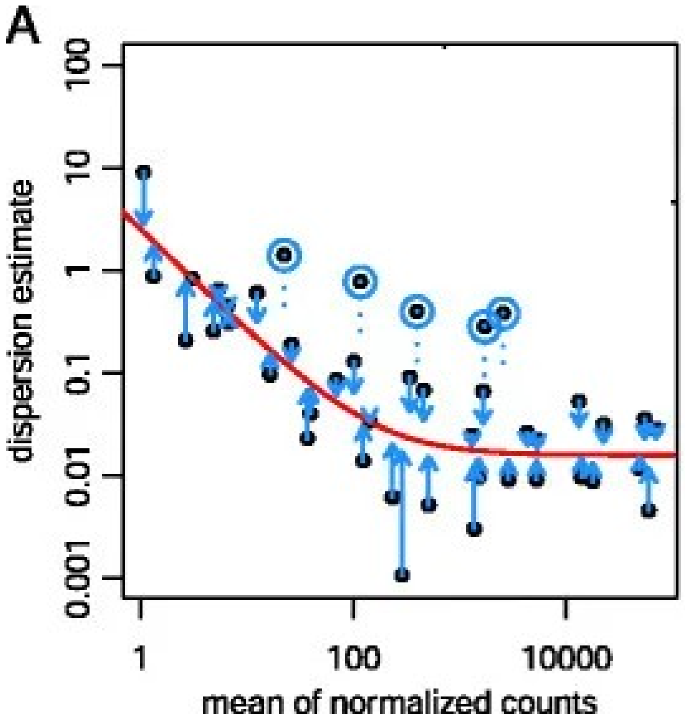
In use, not much different

$$T(\underset{\text{mean}}{\mu}, \underset{\text{standard deviation}}{\sigma^2})$$

$$NB(\underset{\text{mean (counts)}}{\mu}, \underset{\text{dispersion (i.e. variance)}}{\alpha})$$

Shrinkage (of variance)

- Individual genes have high variance
  - *n* is small
  - High variance = poor statistical power
- Reduce the calculated variance (black dots)
  - Use information from other genes
    - Fit a mean dispersion curve (red)
  - Adjust (shrink) variance with this new piece of information (blue arrows)
- How shrinkage is done is major differentiator between DGE algorithms

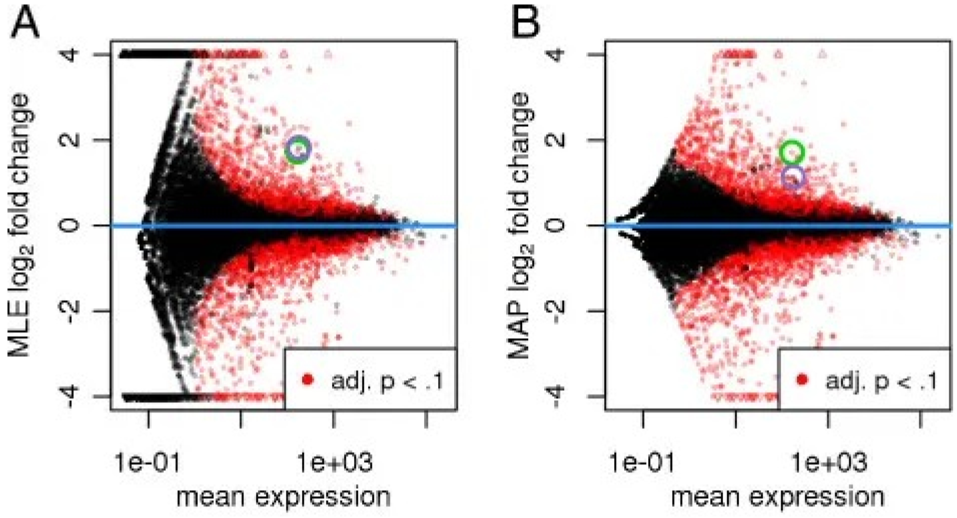


Love MI, Genome Biology, 2014

Weighted shrinkage for low counts

- Lower counts have intrinsically higher variance

- Weight shrinkage **more** for low count genes



no shrinkage

shrinkage

Love MI, Genome Biology, 2014

Linear modeling of expression

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Simple model

KO vs WT: 2 samples, 1 each condition, 1 gene

$$\beta_0 + \beta_1 x_1$$

Gene expression at baseline (i.e. WT)

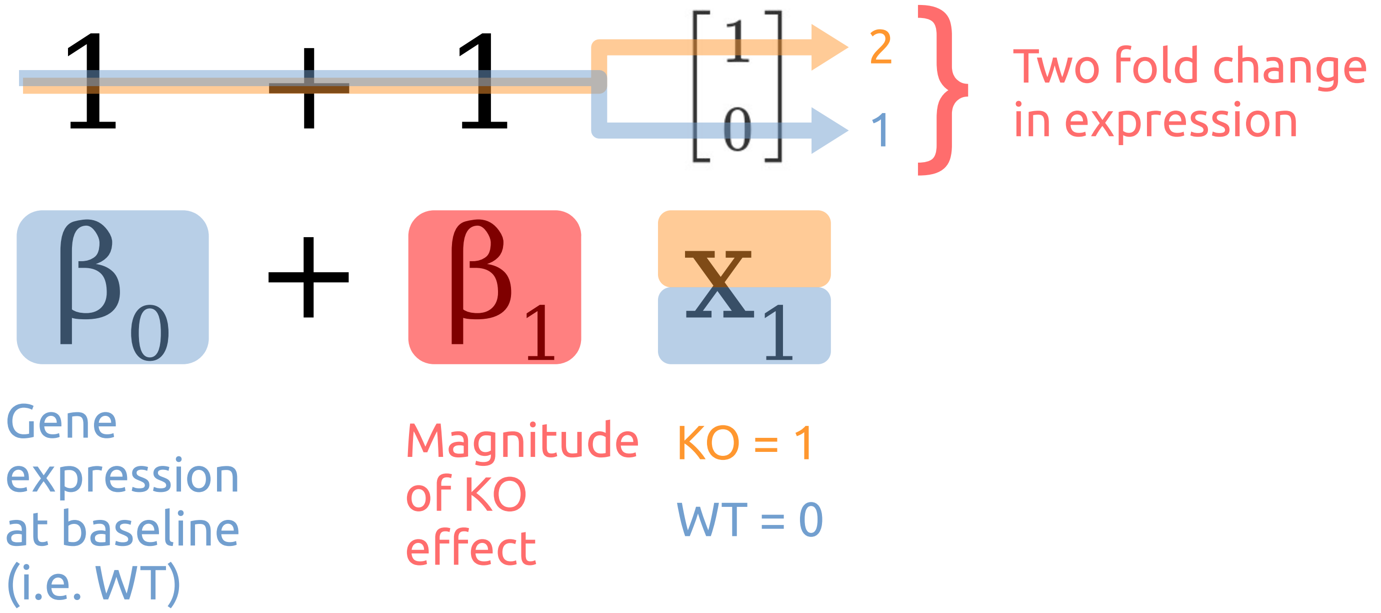
Magnitude of KO effect

KO = 1

WT = 0

Simple model

KO vs WT: 2 samples, 1 each condition, 1 gene



$$1 + 1 \begin{bmatrix}1\\0\end{bmatrix} \rightarrow \begin{matrix}2\\1\end{matrix}$$

Two fold change in expression

$$\beta_0 + \beta_1 x_1$$

Gene expression at baseline (i.e. WT)

Magnitude of KO effect

KO = 1

WT = 0

Model more effects in experiment

$$\beta_0 + \underbrace{\beta_1 x_1}_{\text{Experimental condition}} + \underbrace{\beta_2 x_2}_{\text{Batch effect}} + \ \ldots \ + \underbrace{\beta_n x_n}_{\text{Time series}}$$

Model more effects in experiment

$$\beta_0 + \underset{\text{Experimental condition}}{\boxed{\beta_1 x_1}} + \underset{\text{Batch effect}}{\boxed{\beta_2 x_2}} + \dots + \underset{\text{Time series}}{\boxed{\beta_n x_n}}$$

In R: ~ condition + batch + … + time

Hypothesis testing

Consider a drug response experiment with:
DMSO
0.5 mg/kg
2 mg/kg

Wald

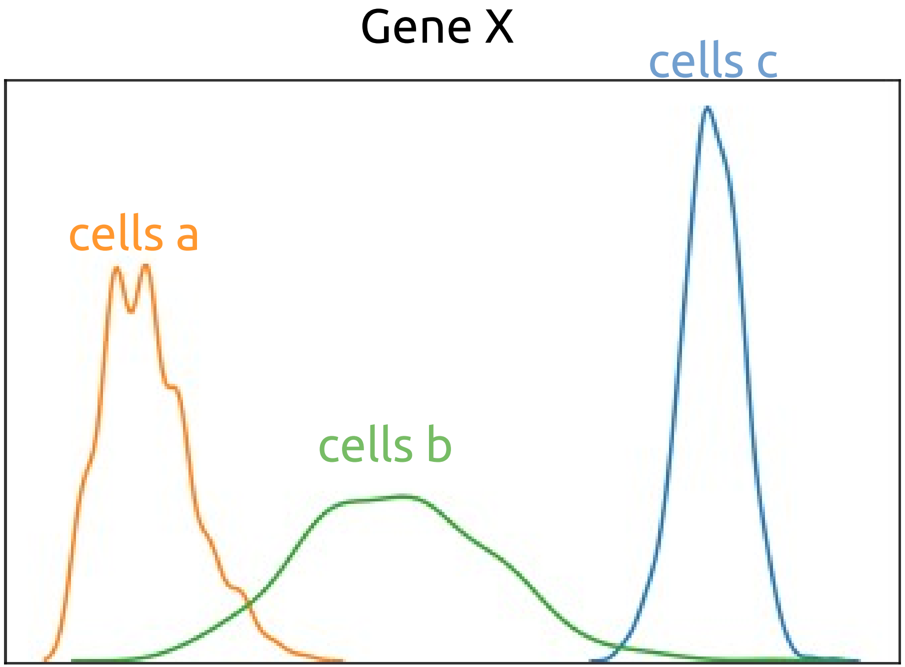
- Is there a statistically significant effect of 2mg/kg on some genes
  - Compare to DMSO
  - Log fold change
  - p-values

Likelihood ratio test

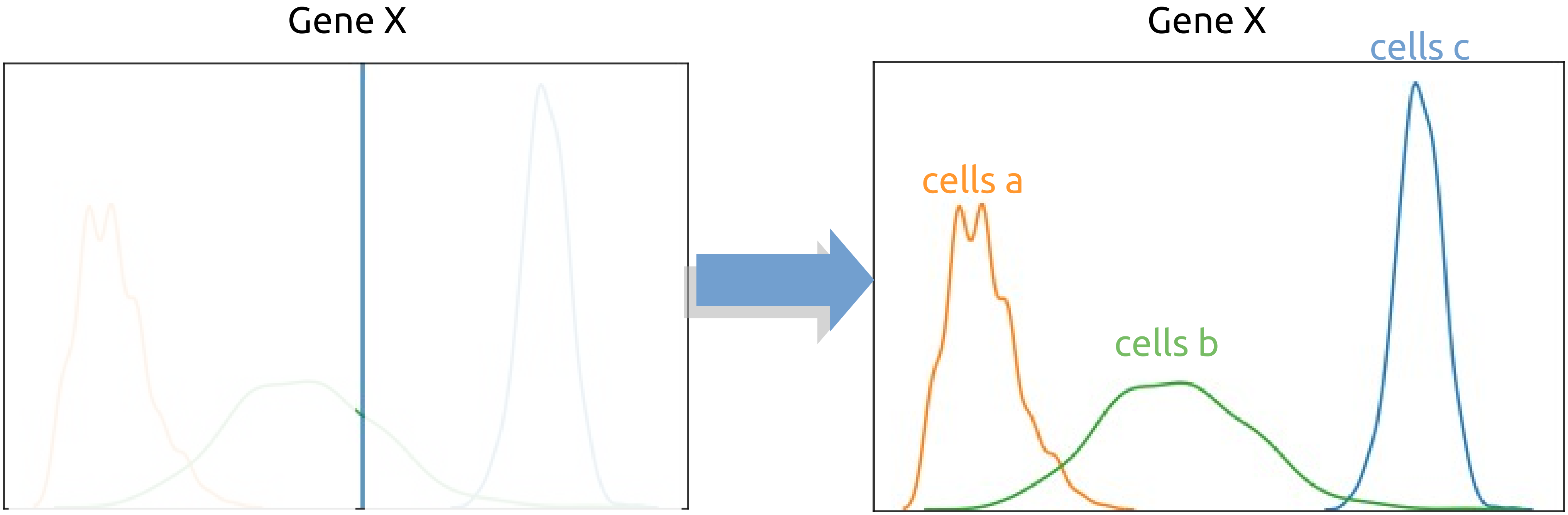
- Is there some statistically effect from the drug on gene expression on some genes
  - DMSO is baseline
  - p-values

In R
Full model: ~ batch + drug amount
Reduced model: ~ batch

Cells have individual expression profiles



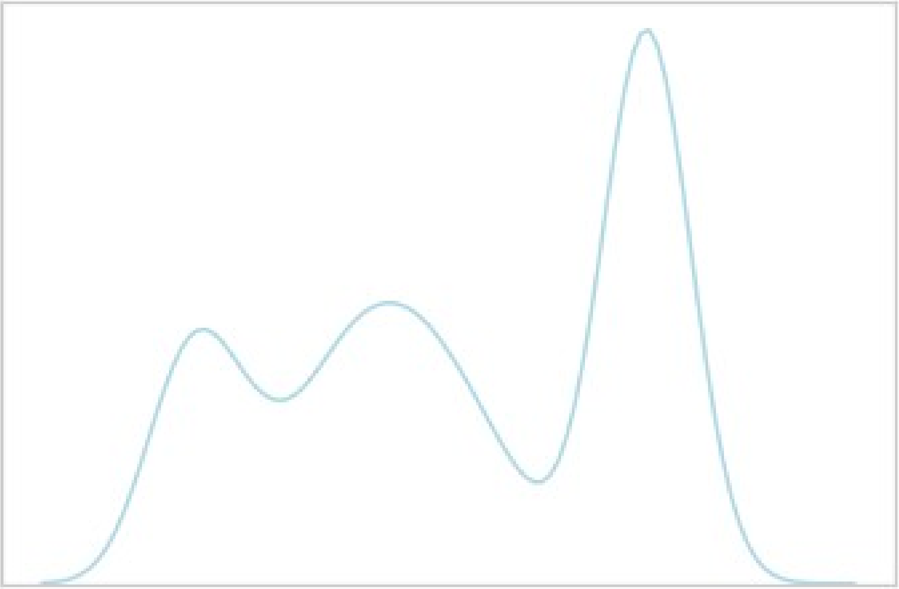
How do we go from mean to individual expression?



Single Cell RNASeq

The reality of scRNASeq

Gene X



All the different expression distributions blend together.

What is scRNASeq good for?

- Gene expression profile heterogeneity
  - Heterogeneity of expression
  - Demographic shifts in cell population

- Conditional testing
  - Differential expression in select subpopulations
  - Cell type transition / differentiation

scRNASeq: process overview

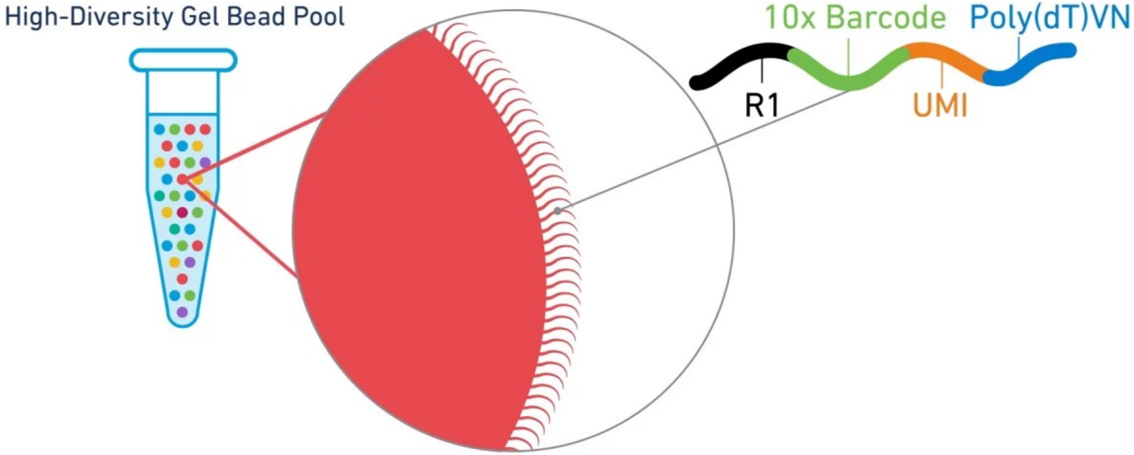
- Beads
  - Barcoded primers
  - Material construction of beads is the major differentiator between technologies
- Cell sorting
  - Attaches to bead
- Encapsulate cell and bead in oil droplet
- Library construction in isolated bead
- Remove oil
- Sequence

scRNASeq major technologies

- 10x (Chromium)
  - Performance in both capture and sequencing
  - Relative to the below platforms, more expensive
- InDrop
  - Completely open-source
  - Amenable to modification
- Drop-seq
  - Performance in sequencing
  - Poor capture

10x Chromium chemistry

10x GemCode™ Technology samples a pool of ~750,000 10x Barcodes
to separately index each cell's transcriptome



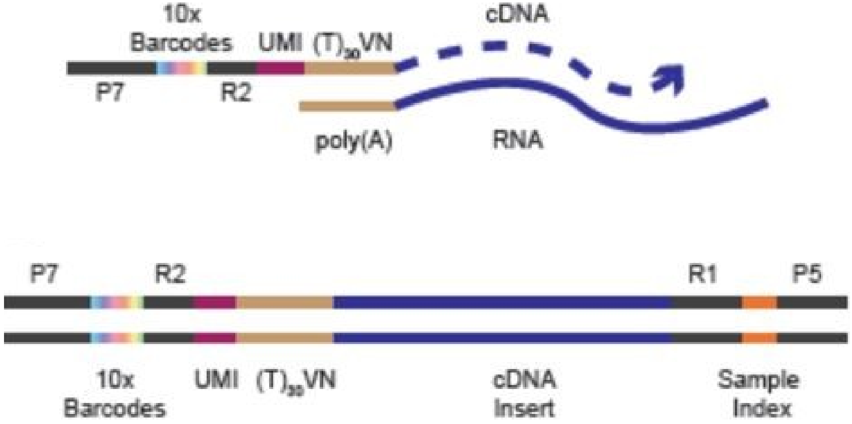
https://www.10xgenomics.com/, 3/2020

10x Chromium chemistry

- 3’ poly(T) tagging

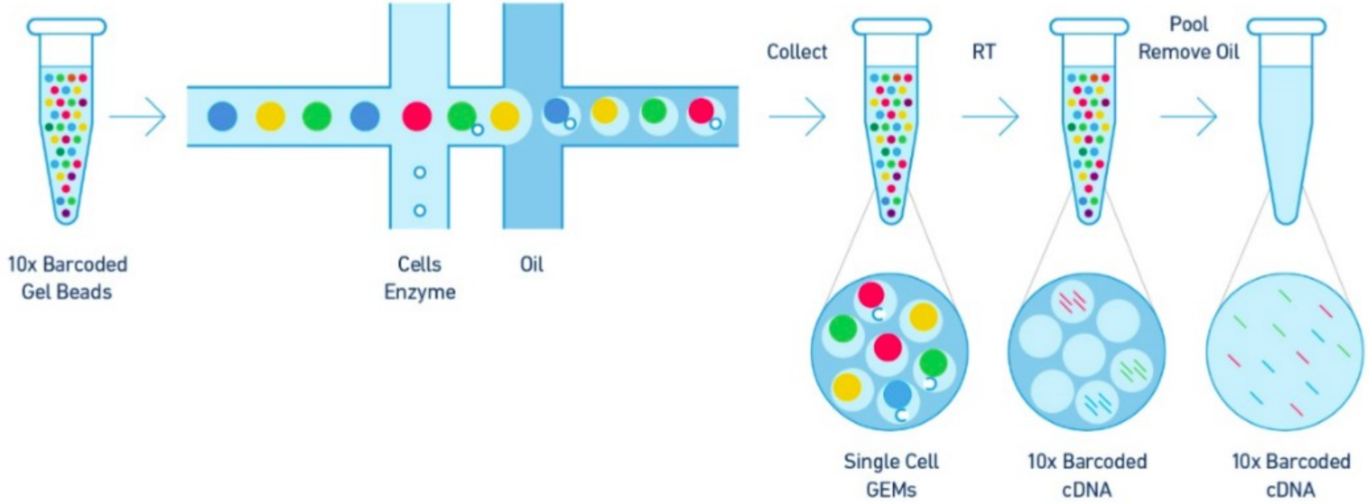
- Only sequences the 3’ end

- Mature mRNA only



Zheng GXY, Nature Communications, 2017

10x Chromium chemistry

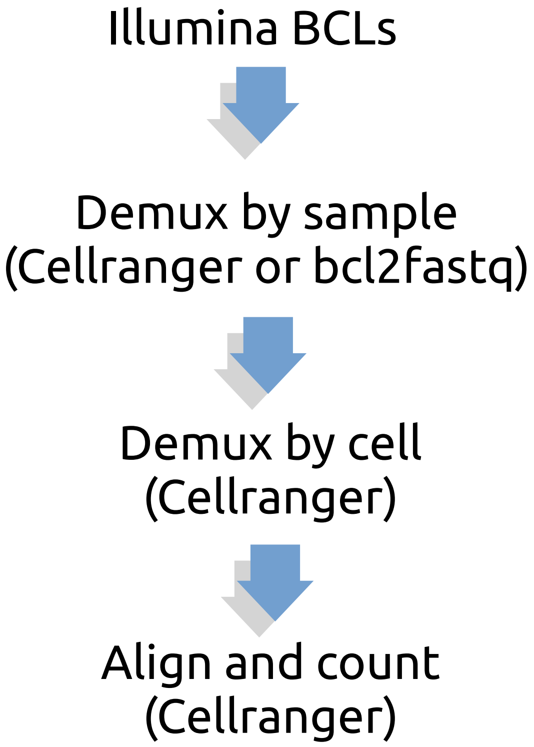


https://www.10xgenomics.com/, 3/2020

scRNASeq sequencing

- Same as with bulk RNASeq
  - Illumina sequencing
  - BCLs → FASTQ

- Counts
  - Unique Molecular Index (UMI)
    - Identifies reads from a unique RNA molecule
    - Commonly used and highly suggested
  - Not a typical alignment result
    - 3’ end of mRNA only
    - Still aligns; ex: STAR for 10x and InDrop



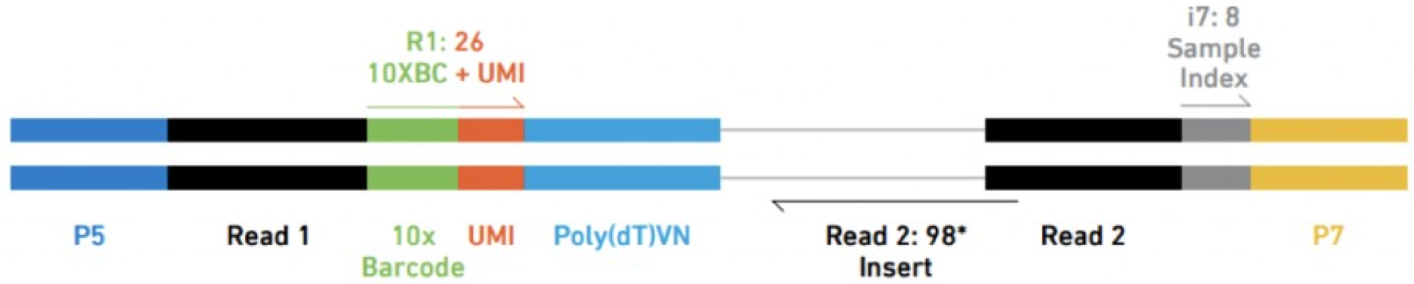
Illumina BCLs

Demux by sample
(Cellranger or bcl2fastq)

Demux by cell
(Cellranger)

Align and count
(Cellranger)

Demultiplexing (demux) 10x



1. Split the sequences by associated sample-level bar codes

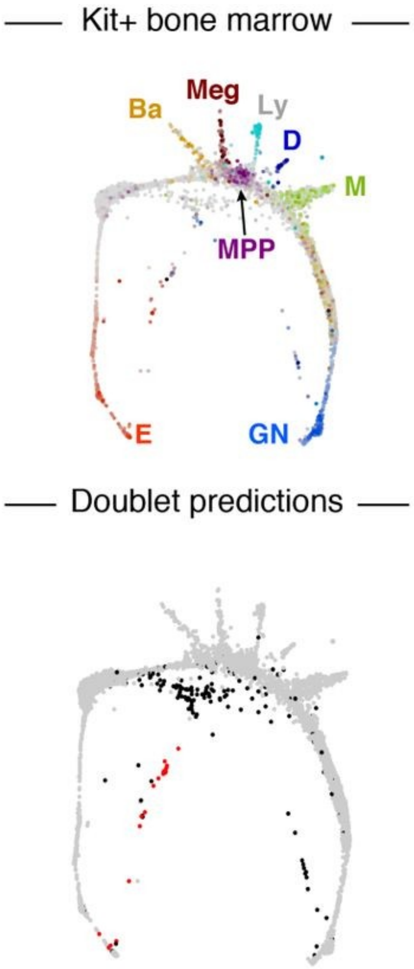
3. Remove duplicate molecules

2. Split the sequences by cell

Multiplets

- Multiple cells trapped in a droplet
  - Can segregate as separate cluster
    - False positive cluster
  - Multiple cells show mixed expression profile of the cells

- Scrublet
  - Program for identification of multiplets

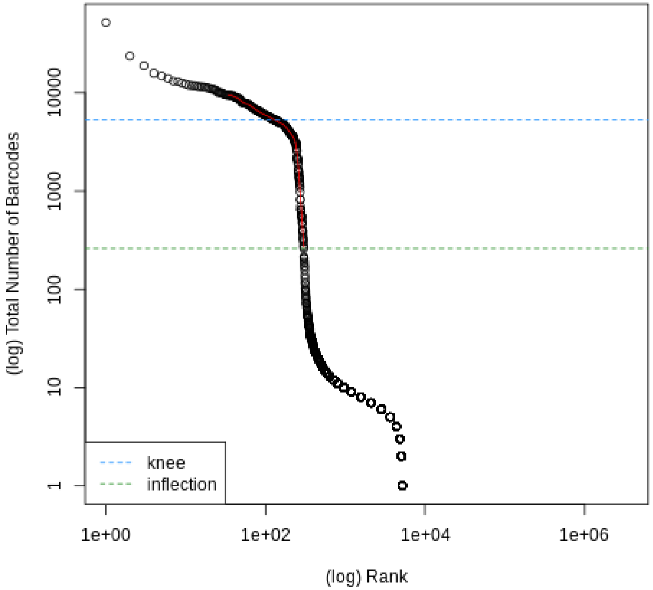


Wolock SL, Cell Systems, 2018

Quality of cell sequencing capture

- Cumulative plot
  - *i.e.* y-axis is interpreted as % of reads at *x* value
  - Used to estimate number of cells sequenced
- Know what your expectation for number of cells sequenced prior
- A sharp, long declination preferred
  - Indicates the dropoff from sequencing cell to background

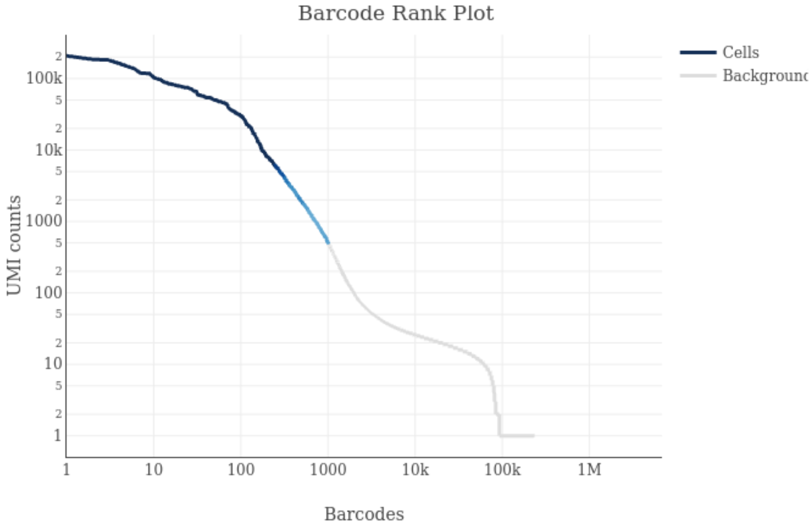
Idealized outcome



Quality of cell sequencing capture

- Cumulative plot
  - *i.e.* y-axis is interpreted as % of reads at *x* value
  - Used to estimate number of cells sequenced
- A sharp, long declination preferred
  - Indicates the dropoff from sequencing cell to background

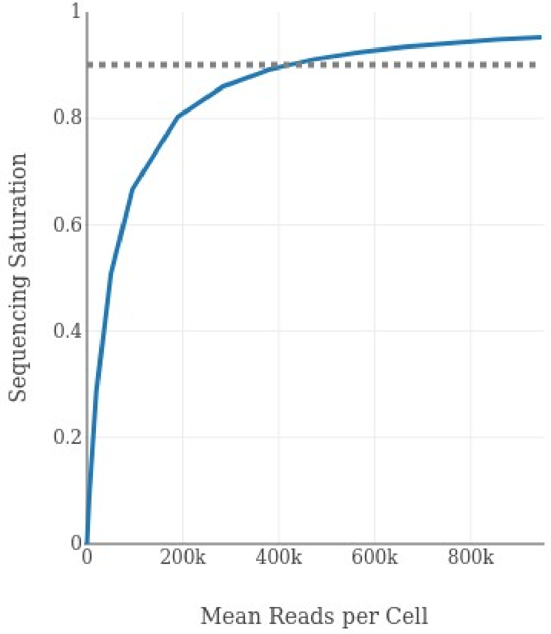
Realistic outcome



Sequencing depth and breadth

- Depth
  - Sequencing saturation
  - Ideally, you would not waste too many reads on diminishing returns
- Breadth
  - Sequencing coverage of genes
    - *i.e.* how many genes are sequenced
  - How many genes are being sequenced
  - Does gene coverage track well with sequencing coverage
    - Does coverage rise quickly with depth

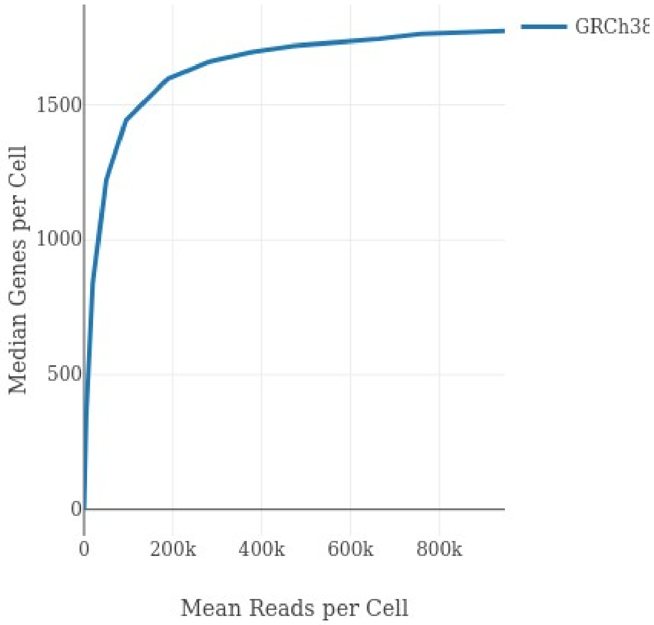
Depth of sequencing



Sequencing depth and breadth

- Depth
  - Sequencing saturation
  - Ideally, you would not waste too many reads on diminishing returns
- Breadth
  - Sequencing coverage of genes
    - *i.e.* how many genes are sequenced
  - How many genes are being sequenced
  - Does gene coverage track well with sequencing coverage
    - Does coverage rise quickly with depth

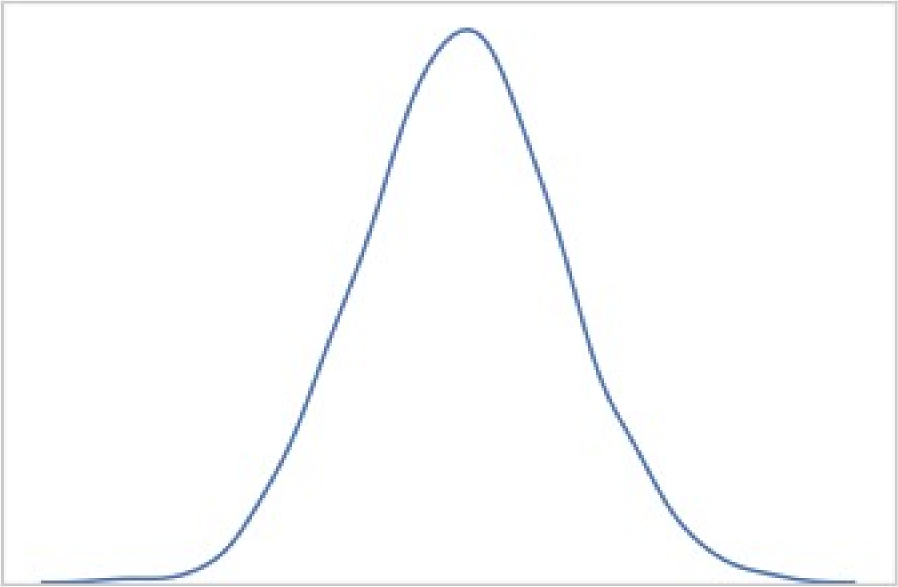
Breadth of gene coverage



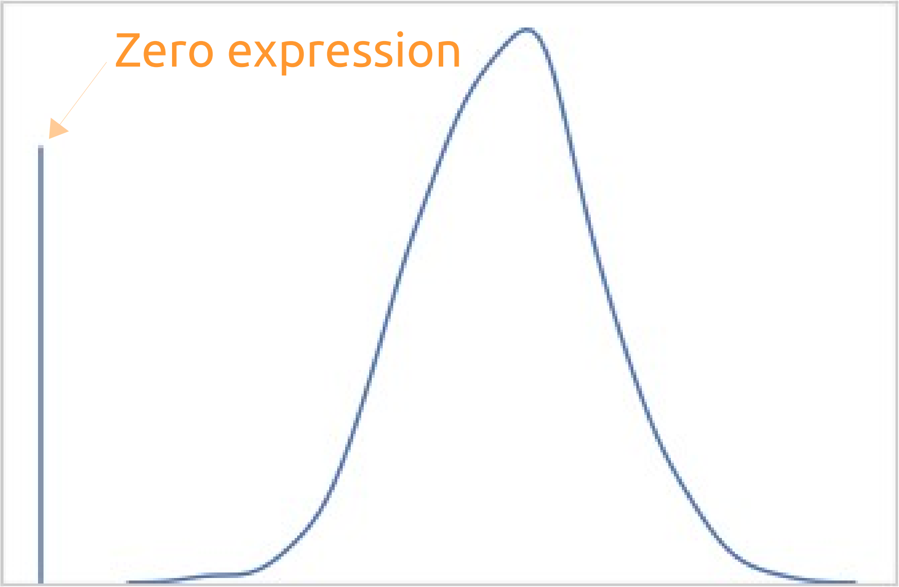
Why can we not use bulk RNASeq techniques for scRNASeq?

Bimodality proves problematic for current bulk RNASeq algorithms

Gene$_i$ expression among samples in bulk RNASeq

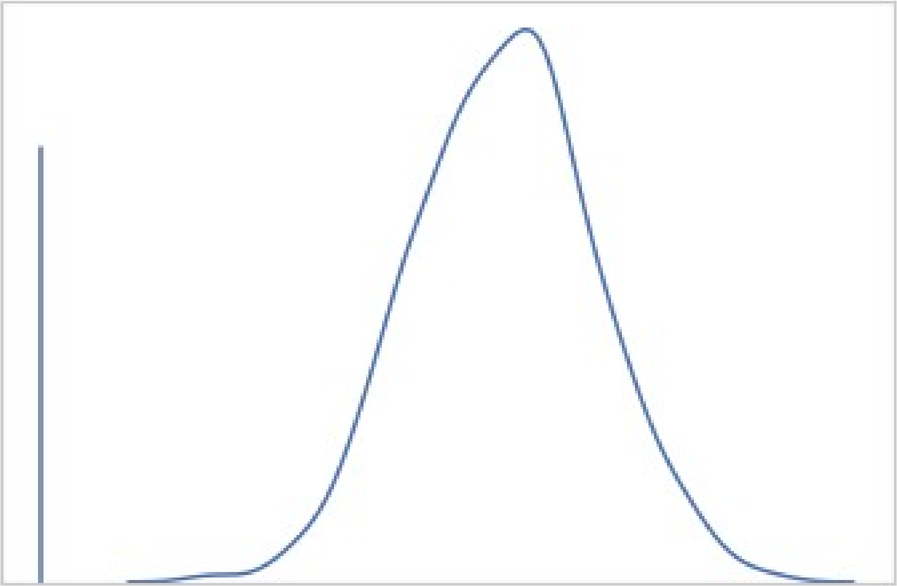


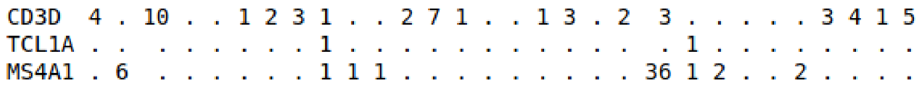
Gene$_i$ expression among cell in scRNASeq



Bimodal expression distribution

Sparsity is a problem



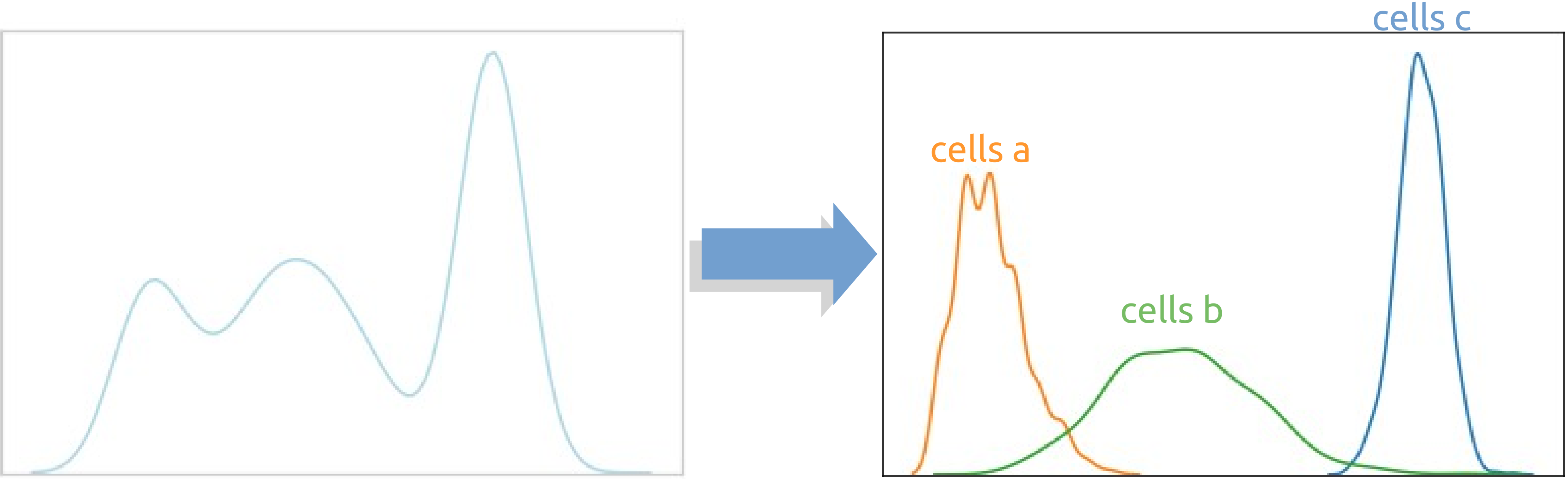


```
CD3D  4 . 10 . . 1 2 3 1 . . 2 7 1 . . 1 3 . 2  3 . . . . . 3 4 1 5
TCL1A . .  . . . . . . 1 . . . . . . . . . . .  . 1 . . . . . . . .
MS4A1 . 6  . . . . . . 1 1 1 . . . . . . . . . 36 1 2 . . 2 . . . .
```

How to get the constitutive parts?

Gene X

Gene X



All the different expression distributions blend together.

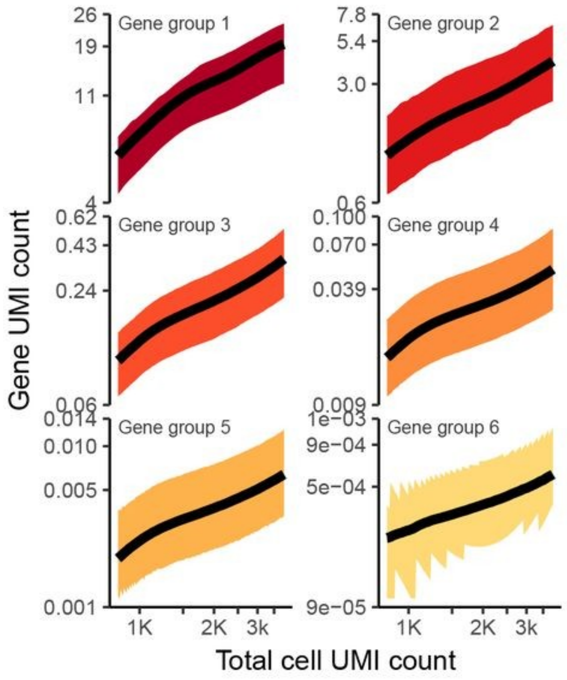
Normalization

- Attempts to scale data all onto the same scale
  - To adjust for:
    - Technical bias / variation in capture and sequencing
    - Library preparation variance
- Adjusted scale allows comparison of cells and samples to other cells and samples

Normalize the data

Total RNA per cell

- Cells do not intrinsically have identical RNA quantities
- Total RNA per cell can be a feature unto itself
  - New factor to consider from bulk RNASeq
- How to compare when gene expression and total sequencing depth is confounding?



Hafemeister C, Genome Biology, 2019

Normalization methods

Fixed scale

- Fix all the cells to a fixed range
  - e.g. 10,000 UMI counts max
  - Adjust the UMI counts to be relative to this range
- Log transform the data
- Does not preserve the total sequencing depth as a feature

Alternative example: scTransform

- Preserves the information of total sequencing depth
  - While still mitigating it's effect on PCA
- Given a total sequencing depth, models the expected gene expression
  - Reports the normalized result as a value relative to its expectation

Determining cell types

- Label with known biomarkers

- Determine the biomarkers from the data

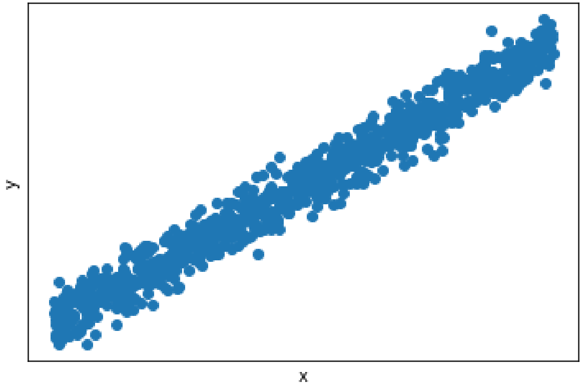
- Both are aided with:
  - Visualization
  - Clustering

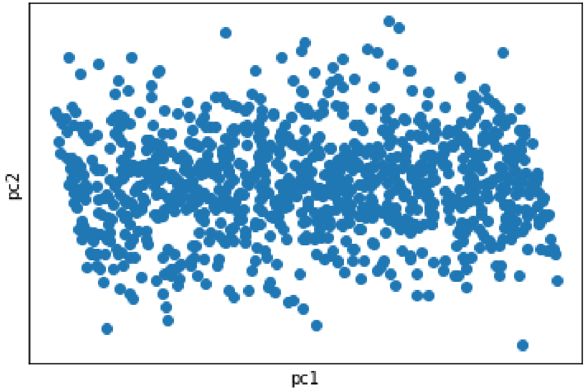
Dealing with high dimensionality

- High dimensionality to the data
  - Lots of cells
  - Lots of genes
- Difficult to visualize
- Reduce the dimensions
  - PCA
  - t-SNE / UMAP

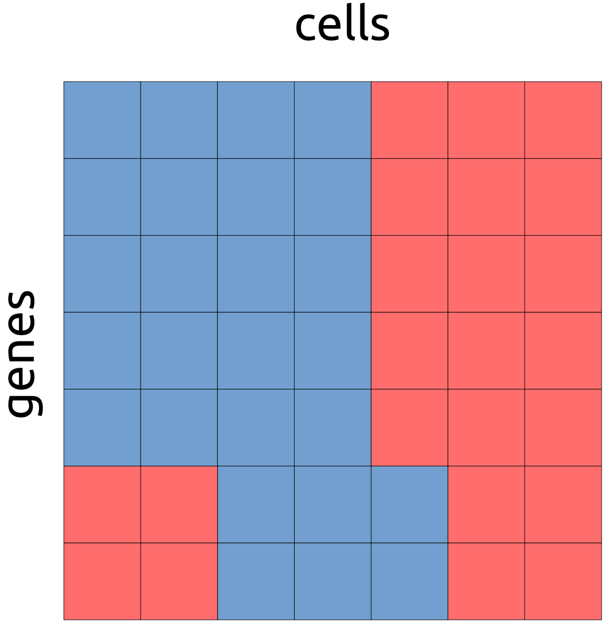
Principal component analysis

- Reduces dimensionality
  - Compresses information
- Removes correlations from its dimensions
  - Useful mathematically
- So we can reduce 20,000 genes, to 50 components
  - However, input to PCA is often a limited subset of the top ~2000 most variable genes
  - 50 components is sufficient to capture the majority of the variance





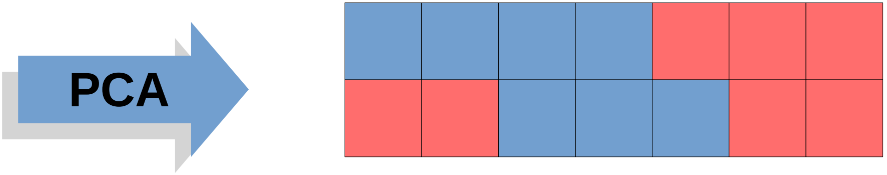
Advantages of compressing correlating features with PCA



🟦	🟦	🟦	🟦	🟥	🟥	🟥
🟦	🟦	🟦	🟦	🟥	🟥	🟥
🟦	🟦	🟦	🟦	🟥	🟥	🟥
🟦	🟦	🟦	🟦	🟥	🟥	🟥
🟦	🟦	🟦	🟦	🟥	🟥	🟥
🟥	🟥	🟦	🟦	🟦	🟥	🟥
🟥	🟥	🟦	🟦	🟦	🟥	🟥

The sheer number of correlated genes skews the clustering toward their representation.

Now clustering methods may find 3-4 clusters instead of just 2.



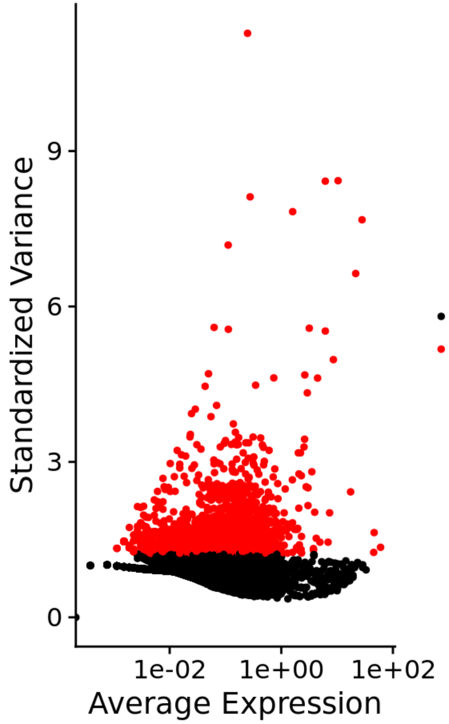
🟦	🟦	🟦	🟦	🟥	🟥	🟥
🟥	🟥	🟦	🟦	🟦	🟥	🟥

PCA attempts to compress the correlated genes together.

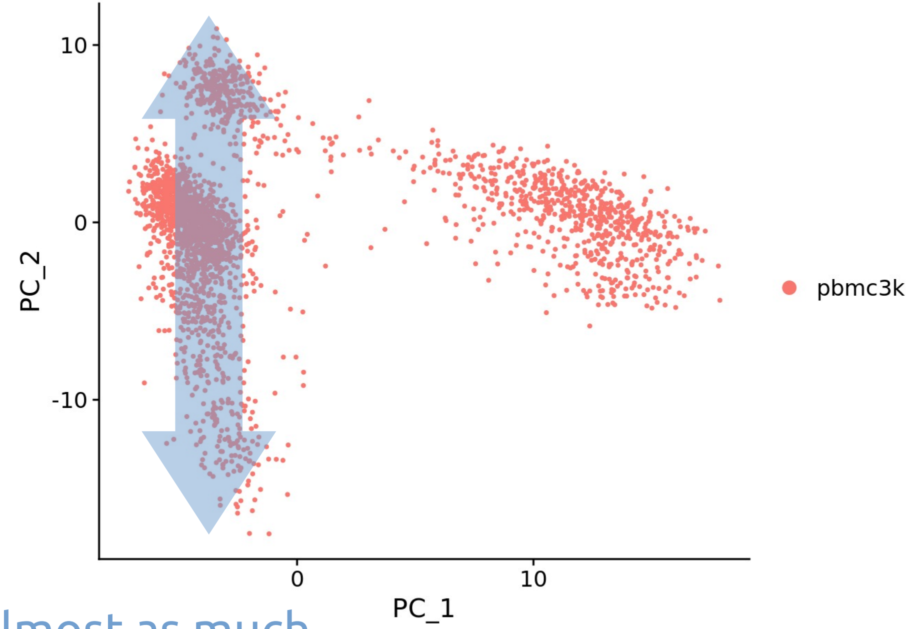
Now the unique variations are more evenly weighted.

Principal component analysis

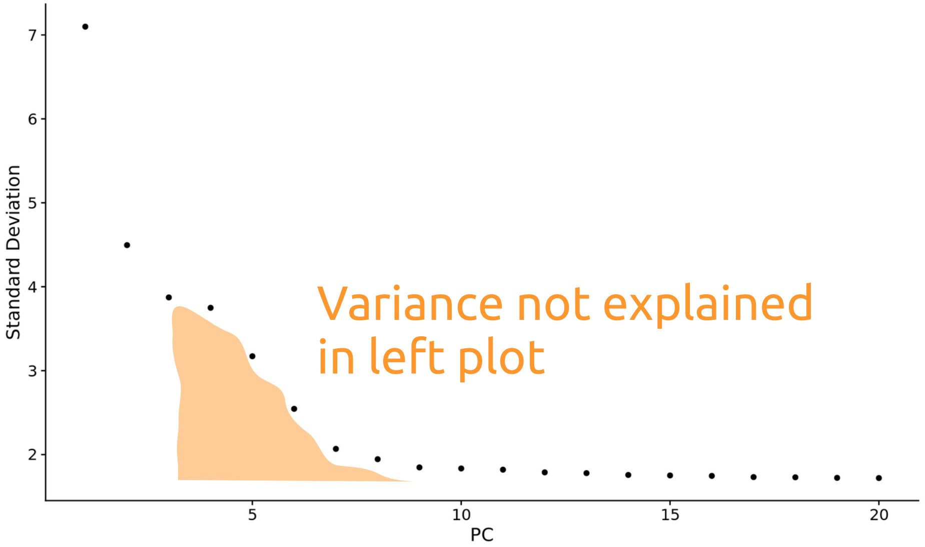
- Reduces dimensionality
  - Compresses information
- Removes correlations from its dimensions
  - Useful mathematically
- So we can reduce 20,000 genes, to 50 components
  - However, input to PCA is often a limited subset of the top ~2000 most variable genes
  - 50 components is sufficient to capture the majority of the variance



PCA is not sufficient dimensionality reduction for visualization

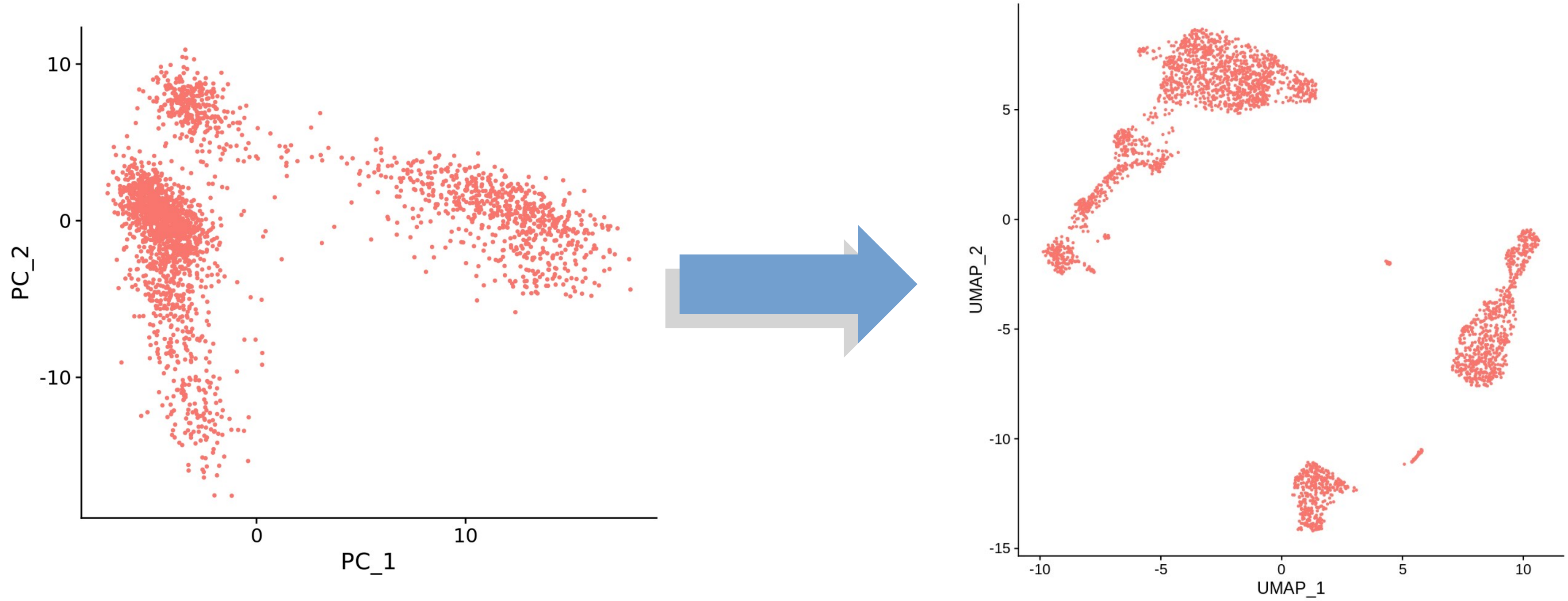


Almost as much spread in PC3 and PC4 as seen in PC2.



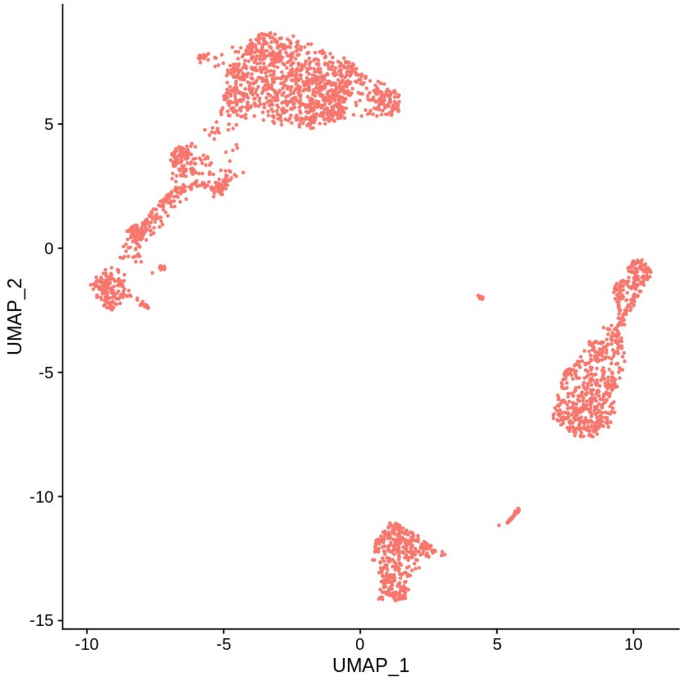
Variance not explained in left plot

Transforming further into low dimensional space



t-SNE and UMAP

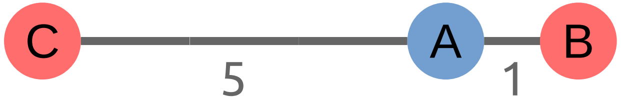
- Further compresses the data points
  - Takes as input PCA
  - Usually 2 dimensions (in t-SNE / UMAP) is enough
- Non-deterministic
  - No two runs (with different seedings) are the same
- For data visualization
  - Not typically for clustering
  - Can help with clustering



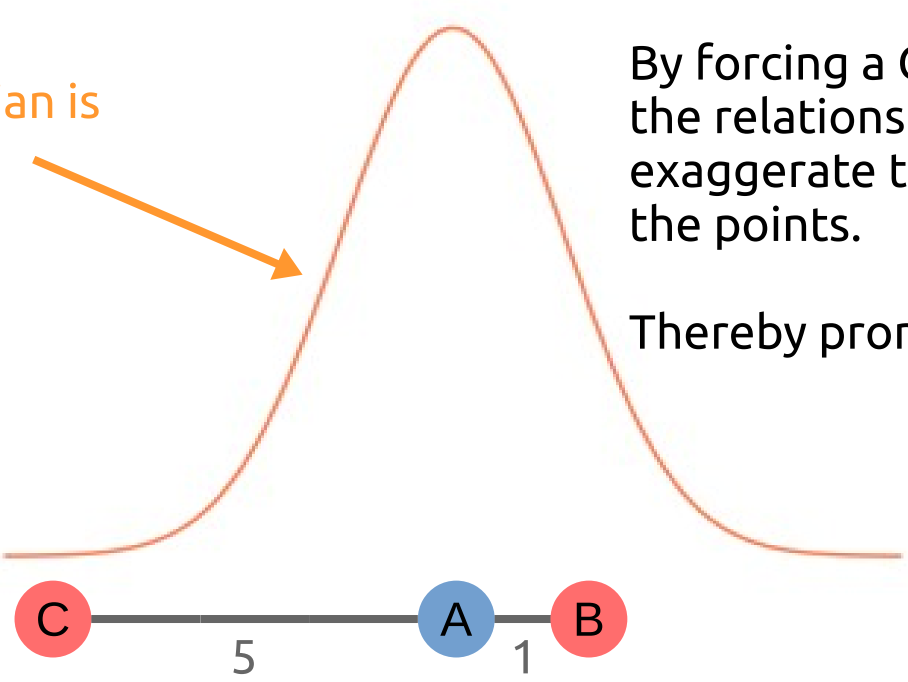
t-SNE

We have three points.

What's the relationship between them?
They're related by distance linearly.



t-SNE



“Distance” in Gaussian is now exponential.

By forcing a Gaussian probability over the relationship between points, we exaggerate the relationships between the points.

Thereby promoting local relationships.

t-SNE

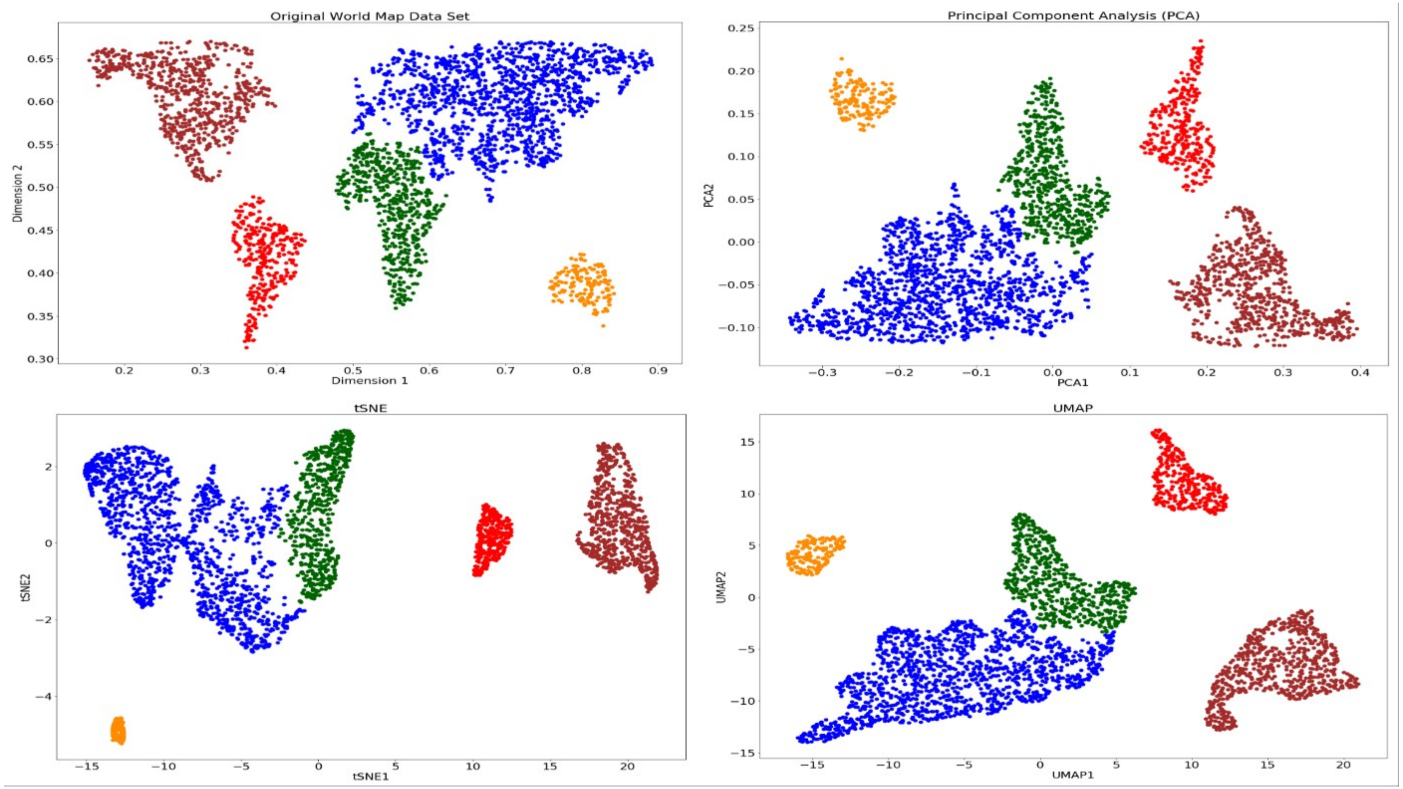
- Exaggerates local relationships
  - Expands dense clusters
  - Contracts sparse clusters

- Requires a lot of parameter tuning
  - Testing lots of different parameters to find the “best” options

UMAP

- Improvement over t-SNE in calculating the probabilities
- Determining the “true” distribution of probabilities for both t-SNE and UMAP
  - Machine learning to identify the distribution
- Faster computation speed
- More components
- Greater preservation of global relationships
  - Inter-cluster distances have more meaning
  - Meaningful organization between clusters
- More suitable for clustering if PCA variance dimensionality is too high

“What You're Seeing…
Is Not What's Happening.”



https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17

Bulk vs. single cell RNASeq

Bulk RNASeq

- Measures an average snapshot of the population of cells
- Well established methodology
  - Technology
  - Algorithms
- Requires extra work in cell sorting for cell type specific expression
  - Still does not have enough resolution

scRNASeq

- Addresses the inadequacies of bulk RNASeq as regards cell specific expression
- Shares much of the same tooling and methods as bulk RNASeq
  - Library preparation and sequencing
  - Alignment methods
  - Counting
- Introduces its own new problems
  - From its own chemistry
  - From the basic premise of what is asked