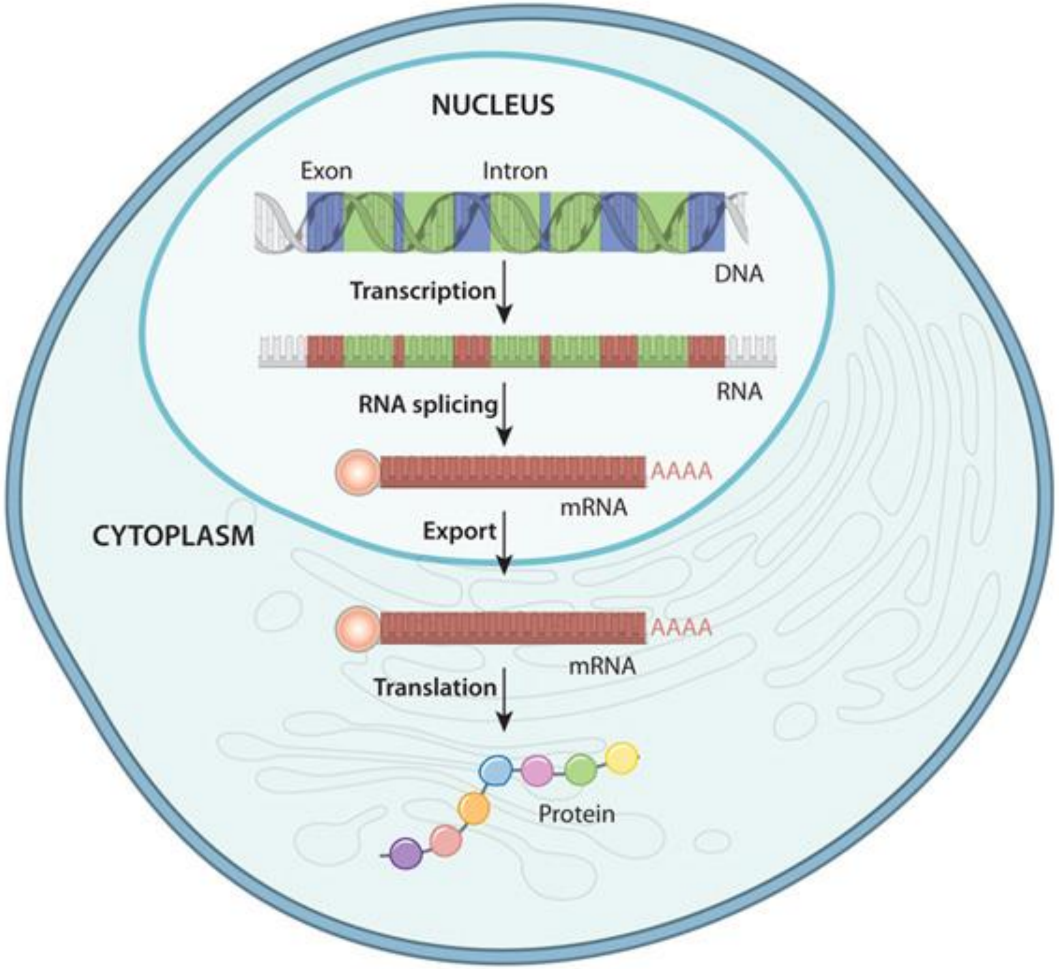
SINGLE CELL TRANSCRIPTOMICS: A CRASH COURSE

UALBERTA HIGH CONTENT ANALYSIS CORE
MIKE WONG

UNIVERSITY OF ALBERTA

Eukaryotic Gene Expression

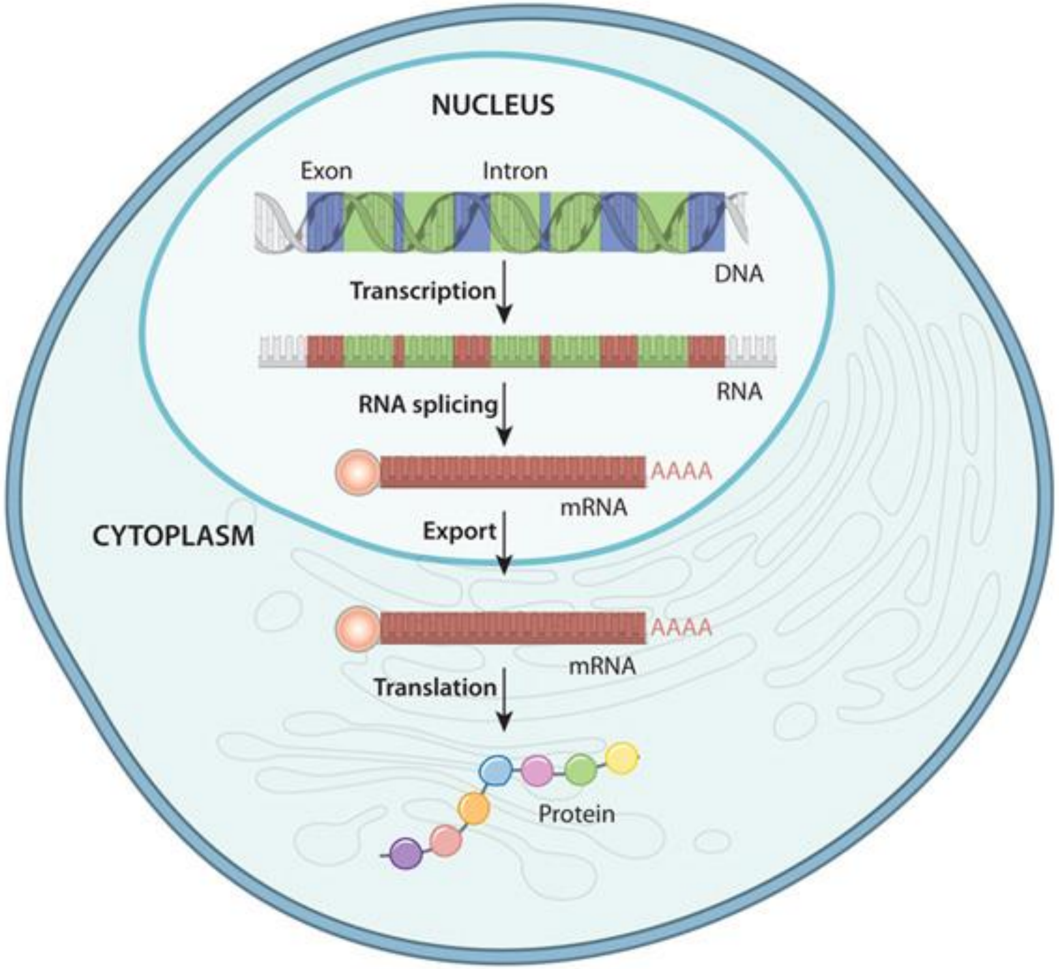


www.nature.com

Eukaryotic Gene Expression

Central Dogma of Biology

- DNA -> mRNA -> Protein
- Gene expression previously measured in ‘bulk’ methods where all RNA from a batch of cells is collected to analyze
  - Quantitative RTPCR
    - Target must be known
  - Bulk RNA Sequencing
    - Can only trace transcripts to whole input tissue or cell suspension
    - All population context lost



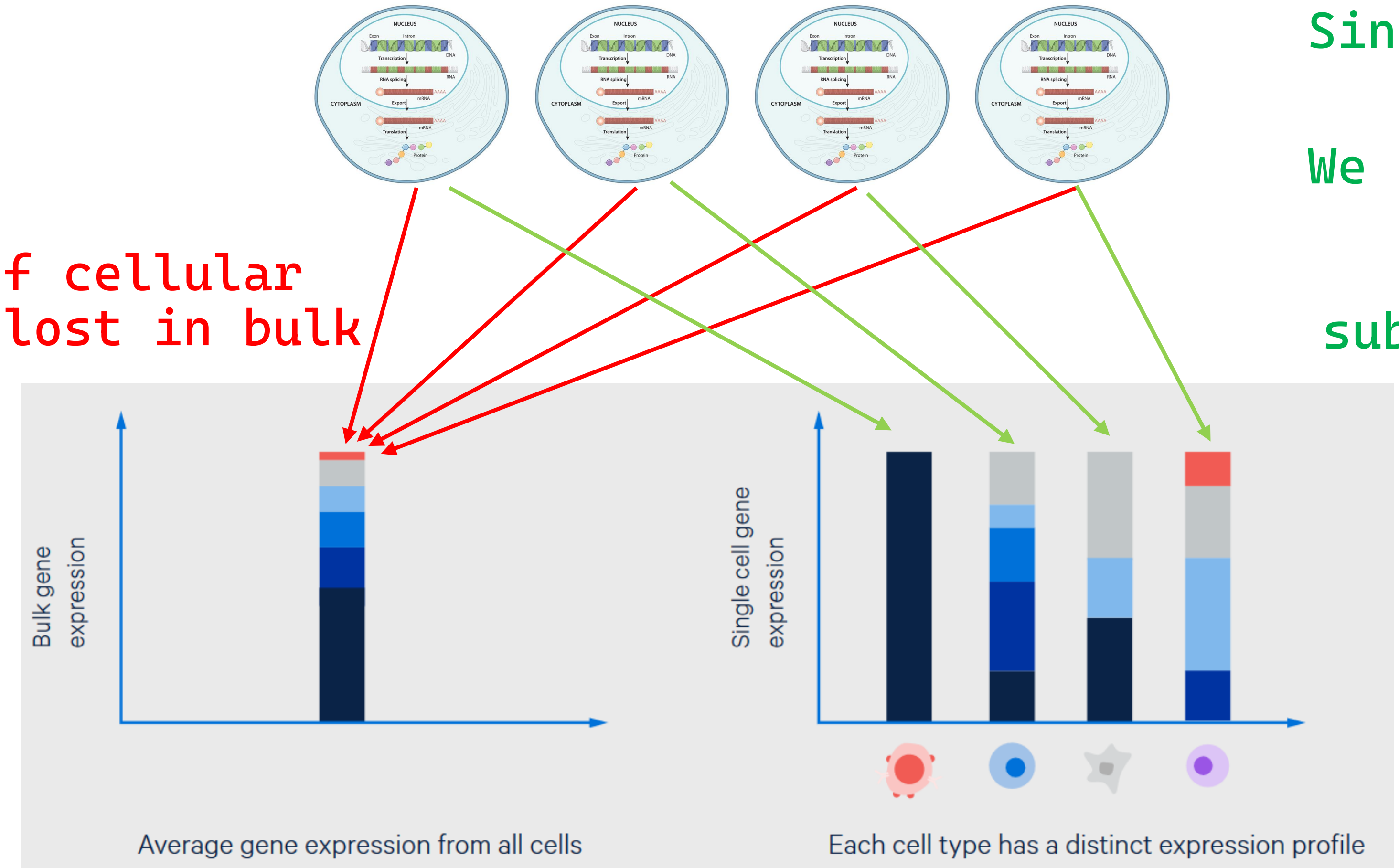
www.nature.com

Eukaryotic Gene Expression

Single Cell Gene Expression

We can now start looking at individual subpopulations!

Complexity of cellular populations lost in bulk analysis



Average gene expression from all cells

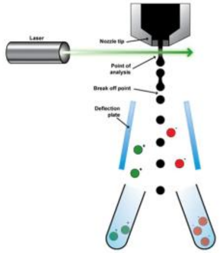
Each cell type has a distinct expression profile

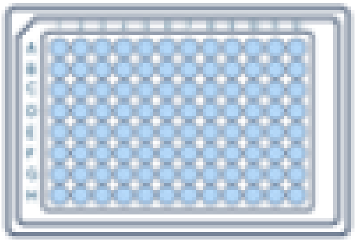
www.nature.com

Single Cell Sequencing Techniques

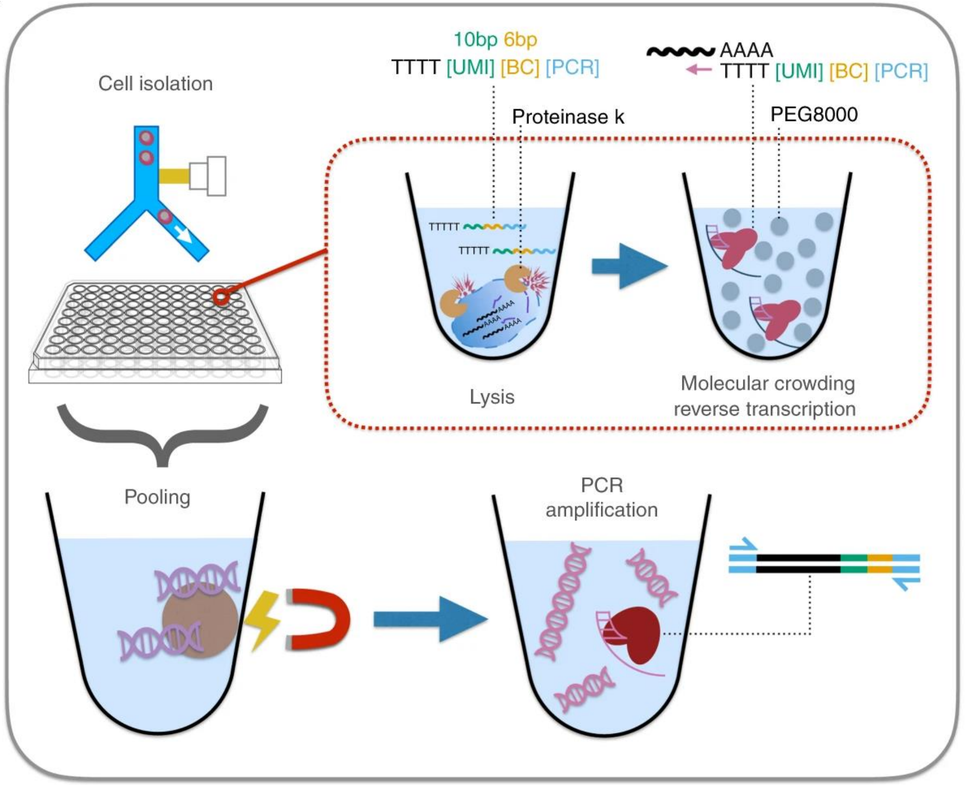
True Single Cell Sequencing

- Glass pipette picked or FACS-Sorted Single Cells
- Similar processing steps to regular bulk RNA seq, but with ultra-low input
- Low throughput, high cost per cell
- Currently used for very specific purposes





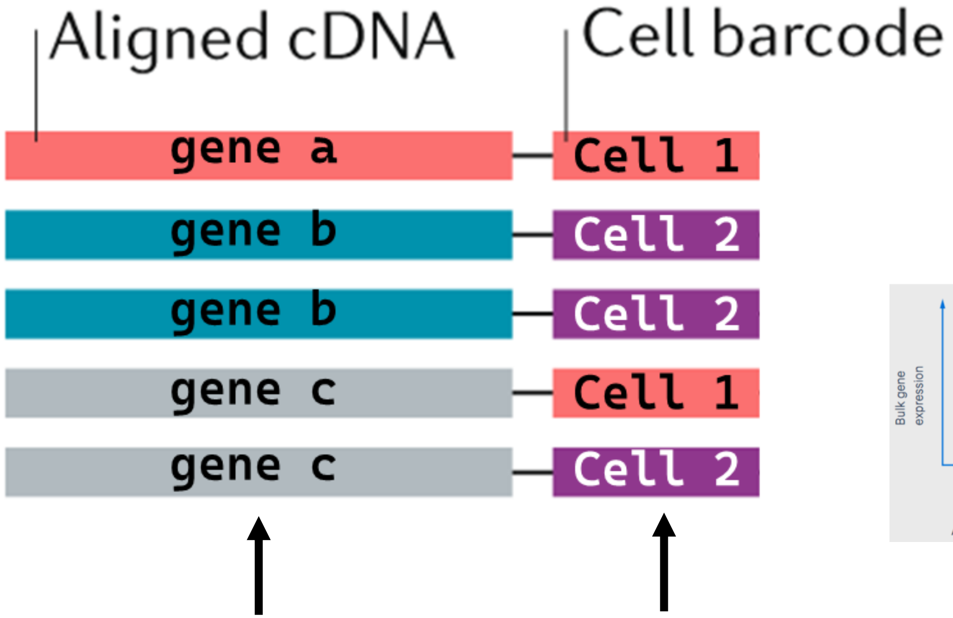
1-384 Cells



Single Cell Sequencing Techniques

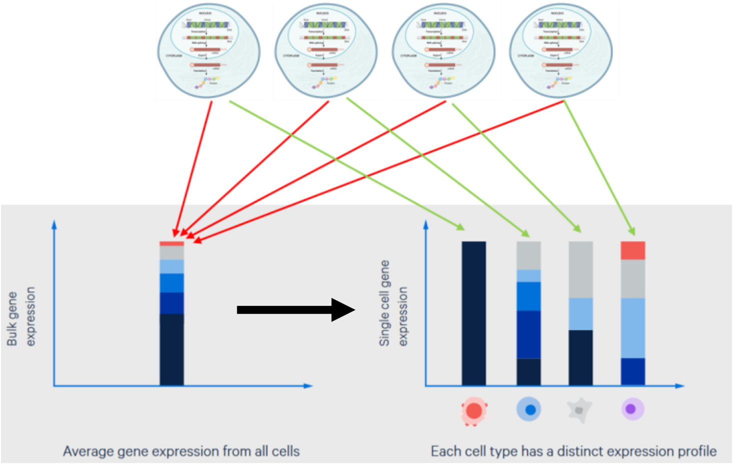
Single Cell Barcoding techniques

- <u>Barcode the cDNAs from each cell with a cell identifier</u>
- Process as bulk RNA
- <u>Use the identifier to assign transcript counts to individual cells</u>
- Moderate to high throughput
- Expensive, but low cost/cell



cDNA created from mRNA transcripts

Cell identifier barcode


Bulk gene expression
Single cell gene expression
Average gene expression from all cells
Each cell type has a distinct expression profile

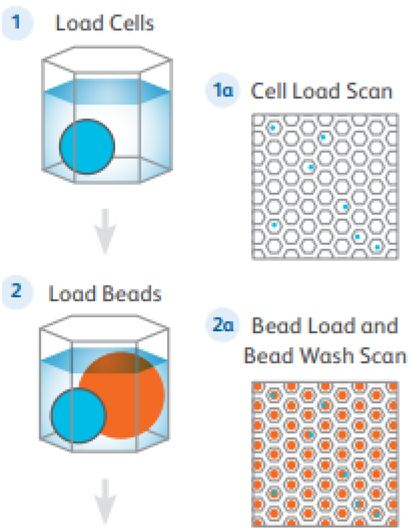
10X Genomics

Single Cell Sequencing Techniques

Single Cell Barcoding techniques

- Use various methods to give cells unique barcodes when mRNA is captured
- Moderate to high throughput
- Expensive, but low cost/cell

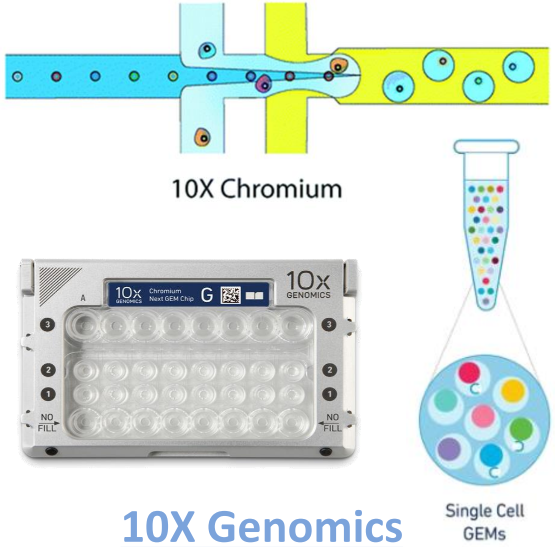
Gravity Loaded Nano-well Arrays



BD Rhapsody

Up to 100 000 cells/sample

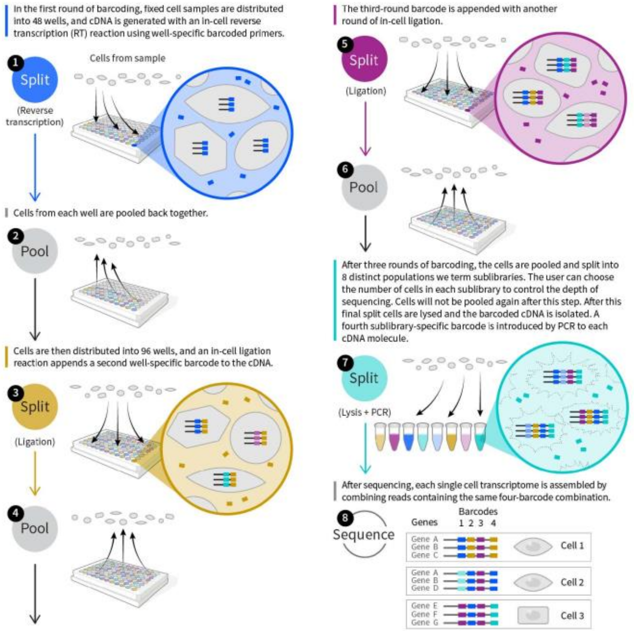
Microfluidic Droplet Based Partitioning



10X Genomics Chromium

Up to 10 000 cells/ sample

Combinatorial Barcoding


In the first round of barcoding, fixed cell samples are distributed into 48 wells, and cDNA is generated with an in-cell reverse transcription (RT) reaction using well-specific barcoded primers.

Cells from each well are pooled back together.

Cells are then distributed into 96 wells, and an in-cell ligation reaction appends a second well-specific barcode to the cDNA.

The third-round barcode is appended with another round of in-cell ligation.

After three rounds of barcoding, the cells are pooled and split into 8 distinct populations we term sublibraries. The user can choose the number of cells in each sublibrary to control the depth of sequencing. Cells will not be pooled again after this step. After this final split cells are lysed and the barcoded cDNA is isolated. A fourth sublibrary-specific barcode is introduced by PCR to each cDNA molecule.

After sequencing, each single cell transcriptome is assembled by combining reads containing the same four-barcode combination.

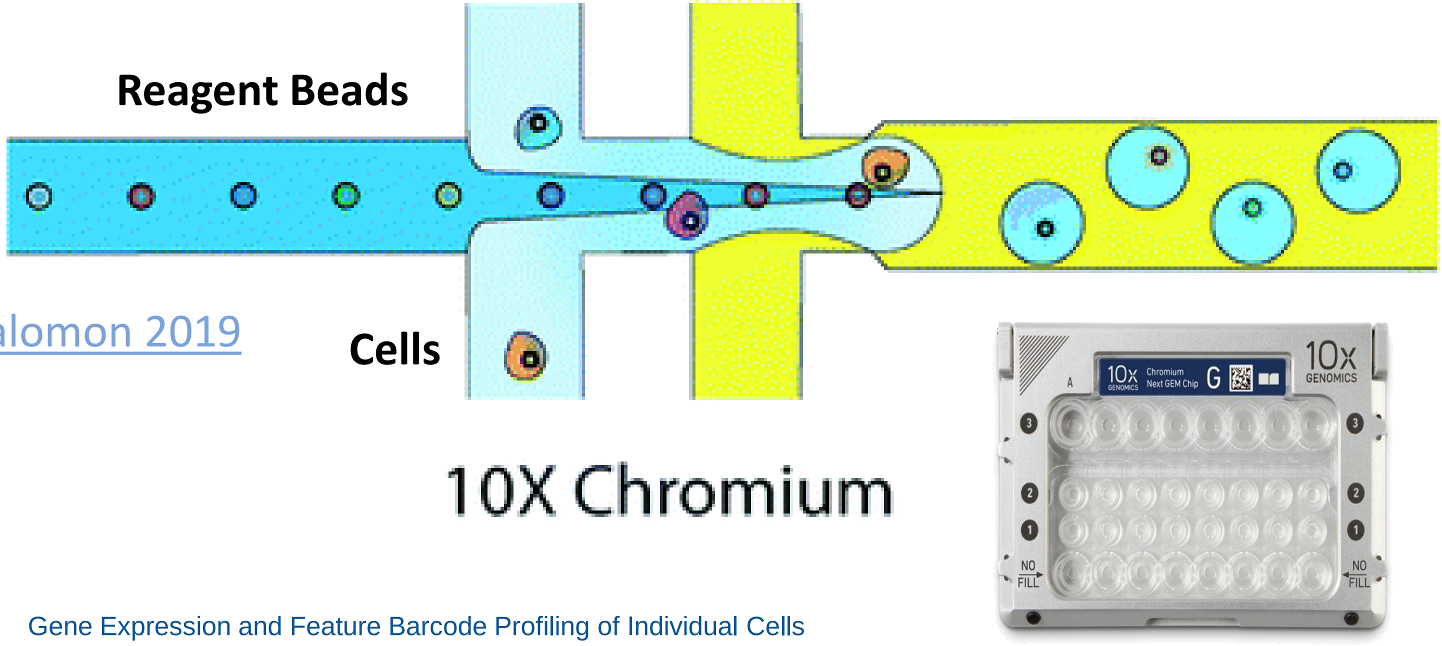
Parse Biosciences

Up to 1 000 000 cells/sample

Droplet Microfluidic Partitioning scRNASeq

Fluidic Partitioning

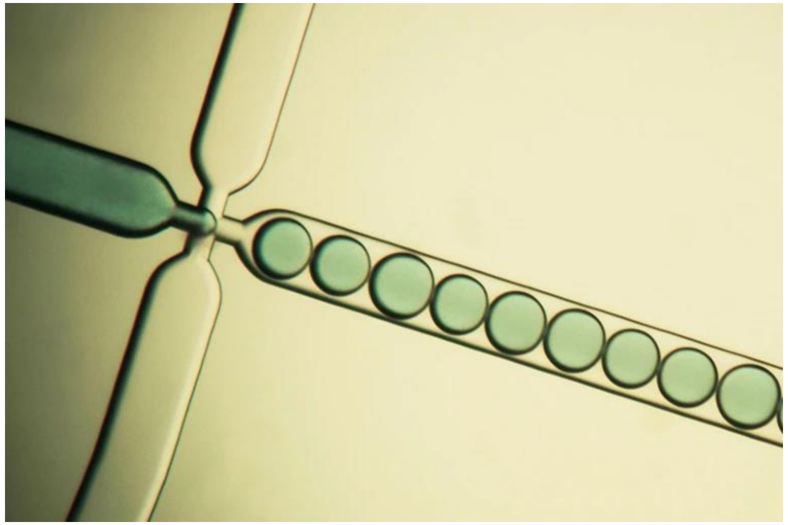
Configuration of the 10X Genomics Chromium fluidic System



Salomon 2019

Gene Expression and Feature Barcode Profiling of Individual Cells

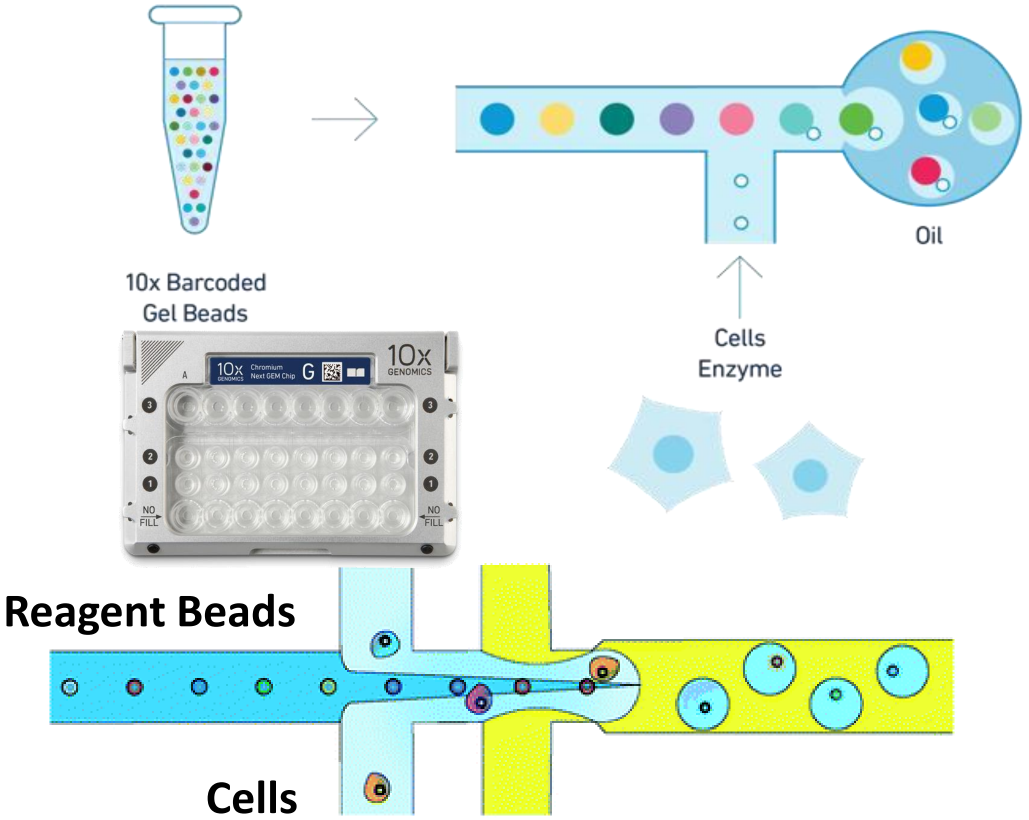
Microscopic photo of a similar microfluidic device

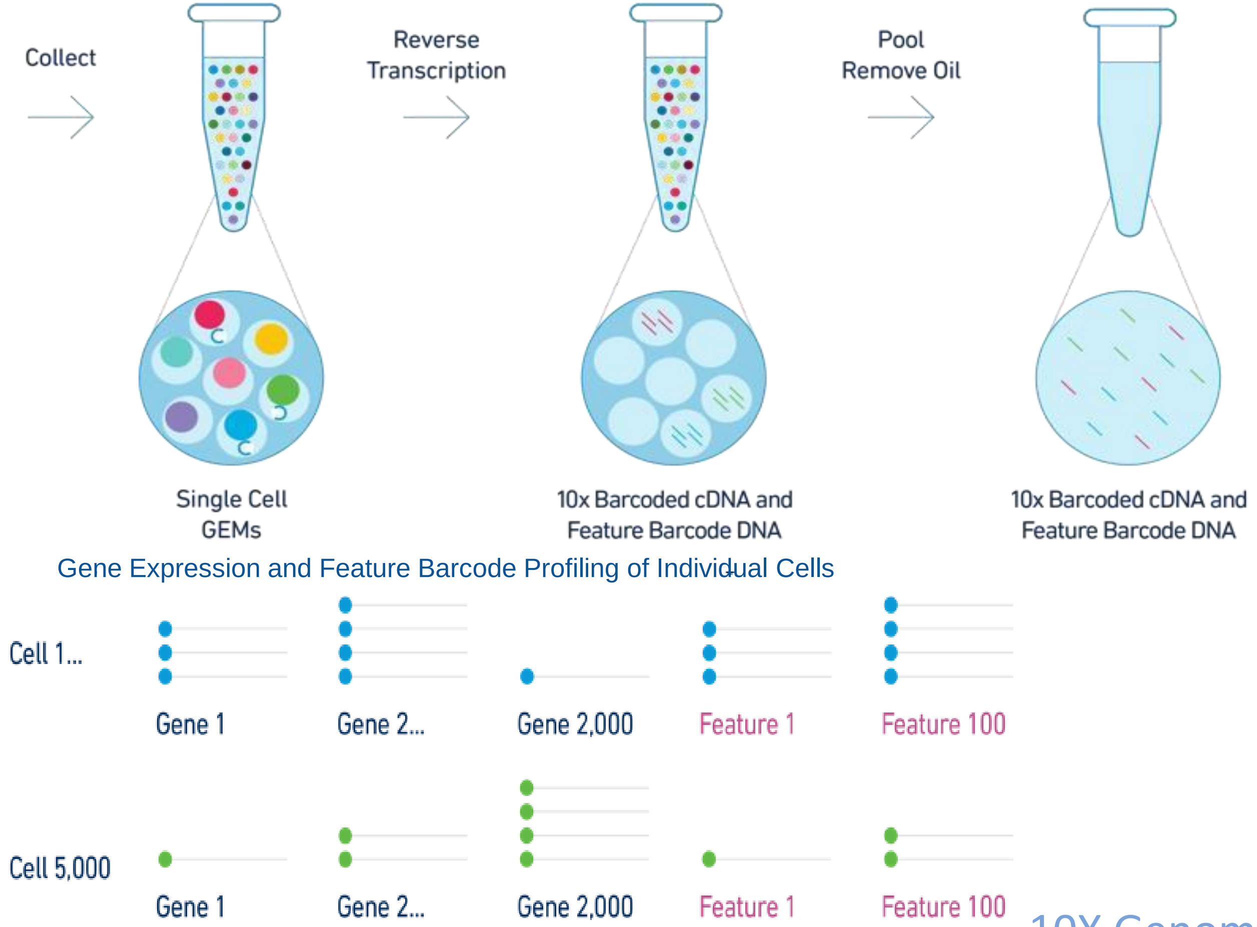


Nature

Droplet Microfluidic Partitioning scRNASeq

Fluidic Partitioning





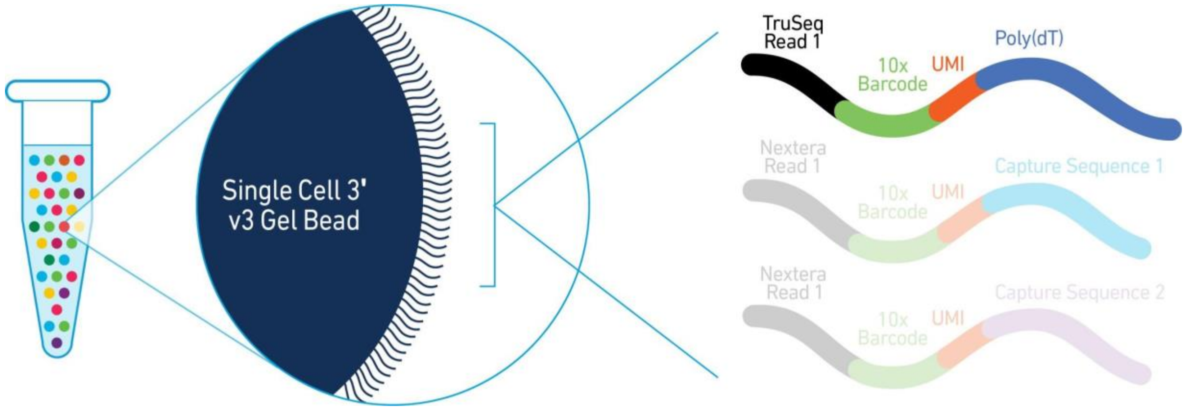
Salomon 2019

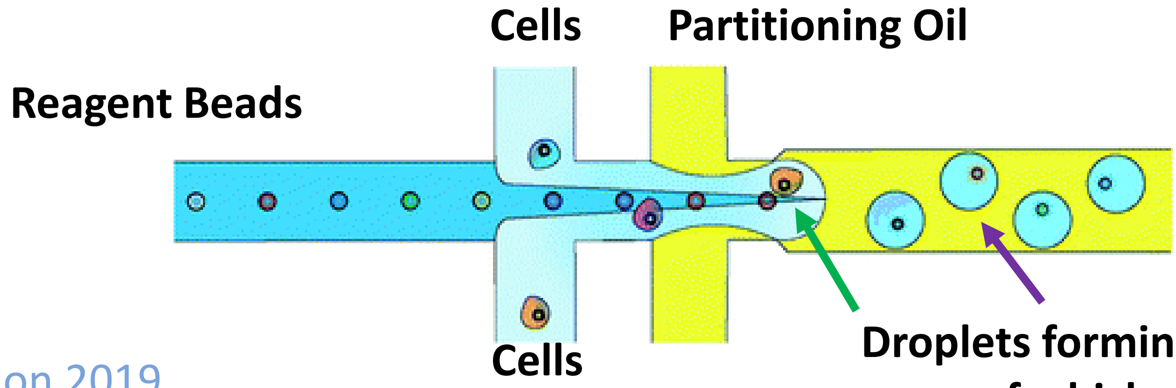
10X Chromium

10X Genomics

Droplet Microfluidic Partitioning scRNASeq

Fluidic Partitioning Reagent Beads with Individual Cells





Droplets forming with Reagent Beads, some of which will also capture a single cell

Create oil droplet ‘chambers’ that will contain single cell reactions

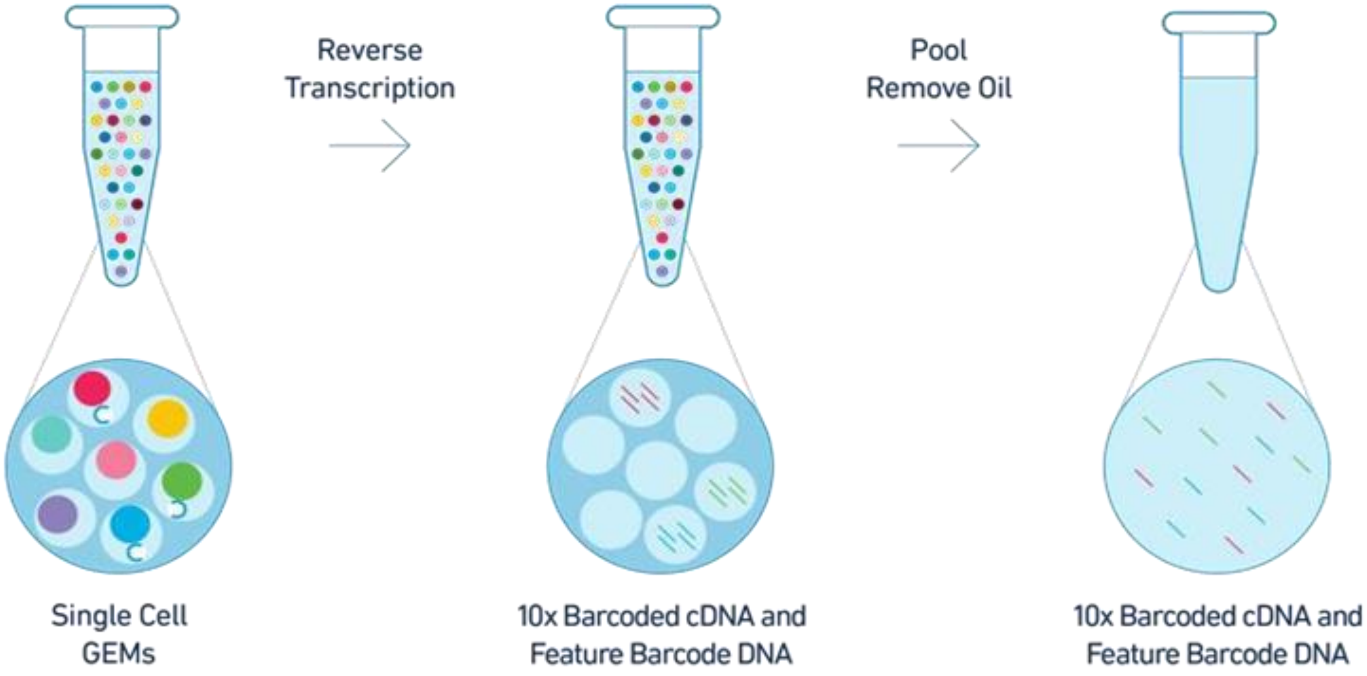
- Beads contain Poly-dT primers with a barcode unique to each
  - TruSeq for Illumina Read Sequencing
  - 10x Single Cell Barcode
  - Unique Molecule Identifier (UMI)
- Polymerases and buffer components

Salomon 2019

10X Genomics

Droplet Microfluidic Partitioning scRNASeq

Fluidic Partitioning Reagent Beads with Individual Cells


Reverse Transcription
Pool Remove Oil
Single Cell GEMs
10x Barcoded cDNA and Feature Barcode DNA
10x Barcoded cDNA and Feature Barcode DNA

Emulsion Partitioning

Melting of beads And cell lysis **<u>Single cell barcoding</u>**

Remove partition process the sample the same as bulk RNA

Create oil droplet ‘chambers’ that will contain single cell reactions

- Beads contain Poly-dT primers with a barcode unique to each to create identifiable cDNA copies of each transcript
- Polymerases and buffer components
- Once droplets are formed, samples are heated which melts the beads and releases the components, allowing partitioned barcoding reactions
- Partitions are destroyed and oil removed to yield a bulk pool of barcoded cDNA transcript copies.

10X Genomics

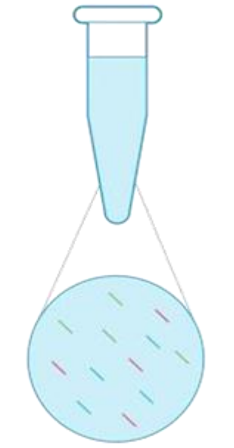
Library Preparation and Sequencing

Library prep:

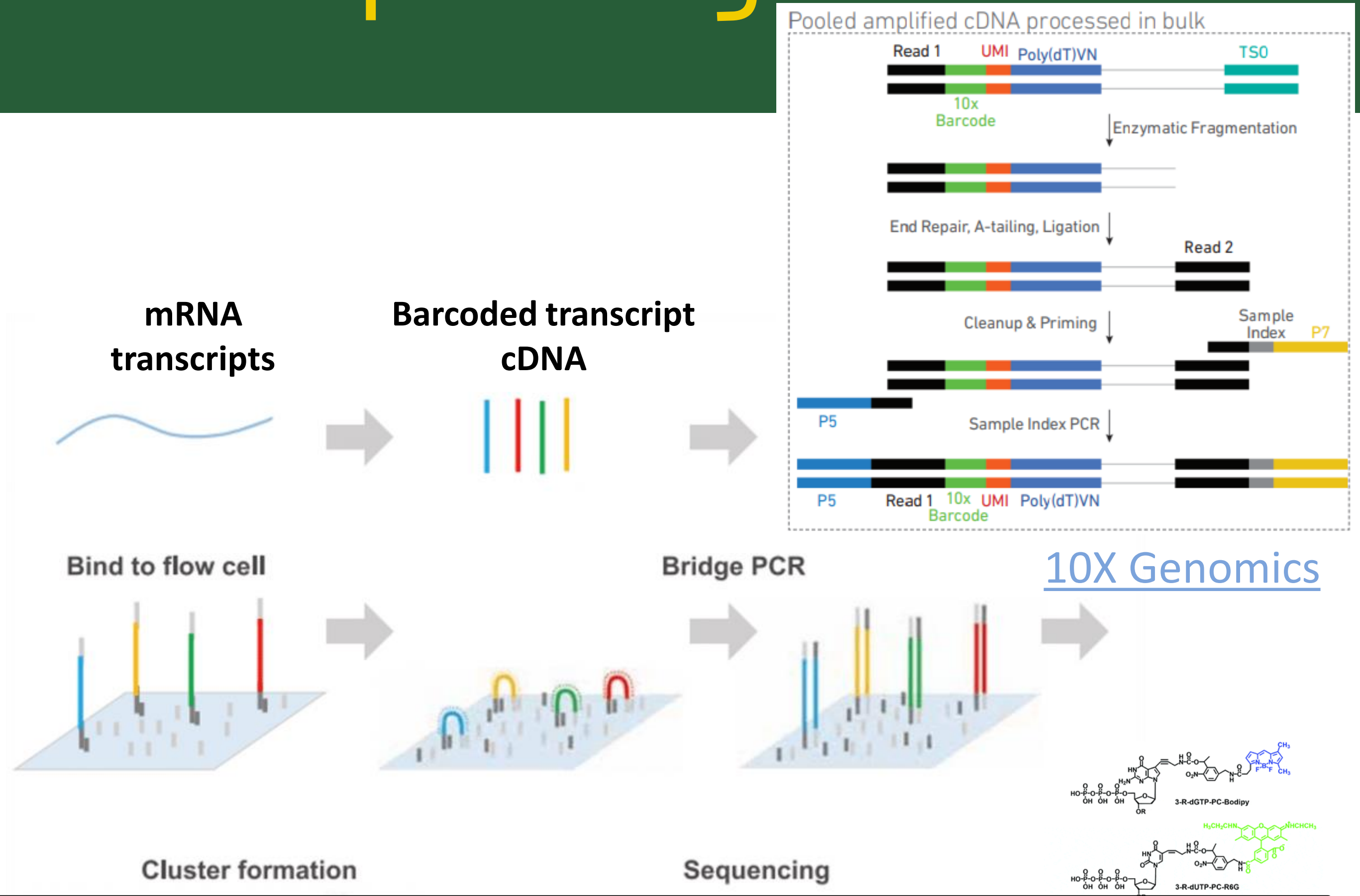
- Fragment and prepare the cDNA for sequencing.

Sequencing by synthesis

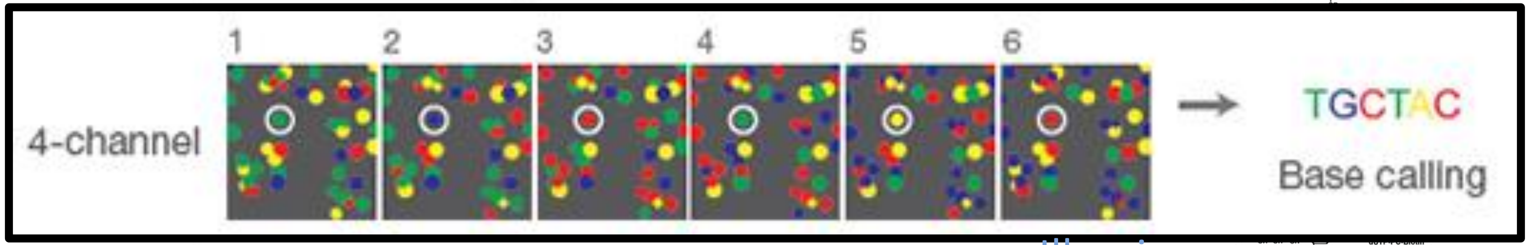
- Adapters are added to barcoded cDNA to support bridge amplification
- Build clusters of identical fragments
- Change the nucleotides to fluorescent tagged versions with chemical stops
  - Clusters emit the fluorescence of the current nucleotide and calls the appropriate base
  - Chemically cleave the fluorescent and repeat



10x Barcoded cDNA and Feature Barcode DNA



10X Genomics



Primers on Next Gen Sequencing (illumina)
https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina

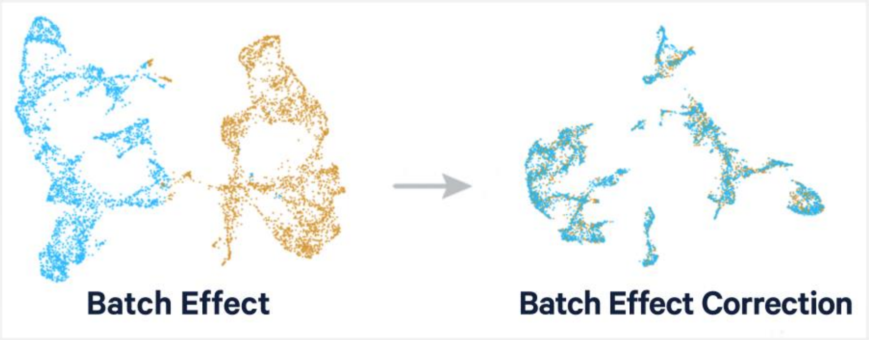
National Center for Genome Analysis Support Blog

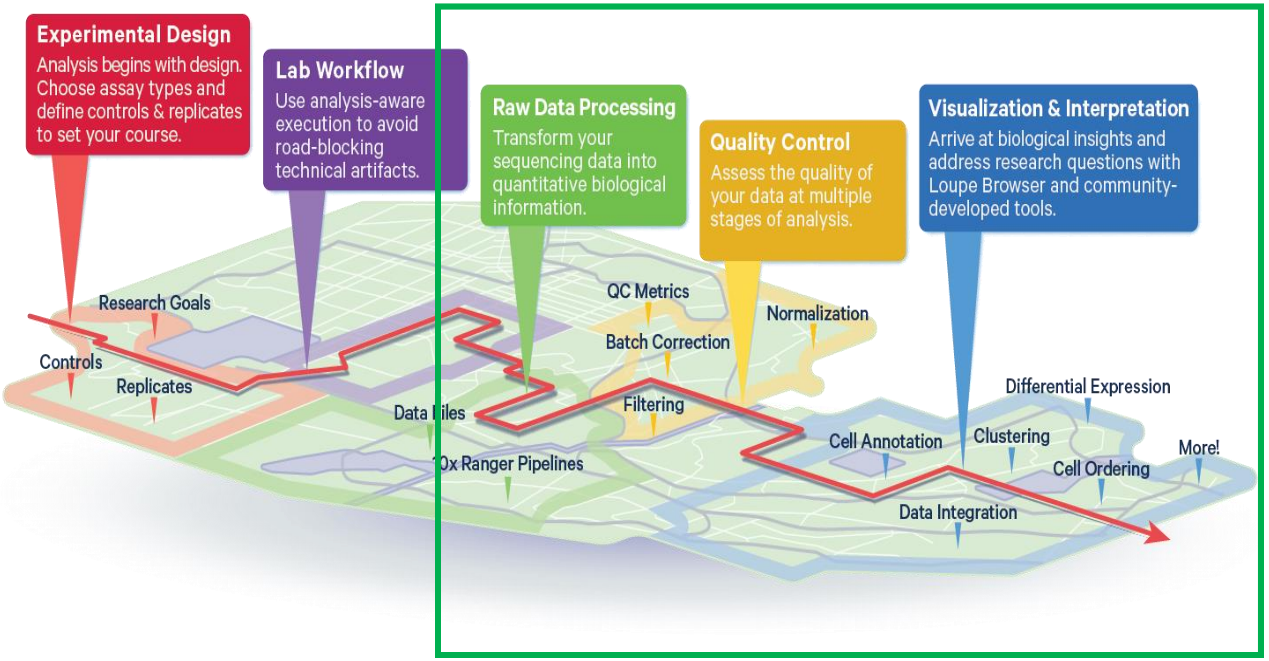
Data Processing and Quality Control

Once Sequencing is complete, still a long way to go!

Quality control and downstream processing is a huge part of scRNAseq

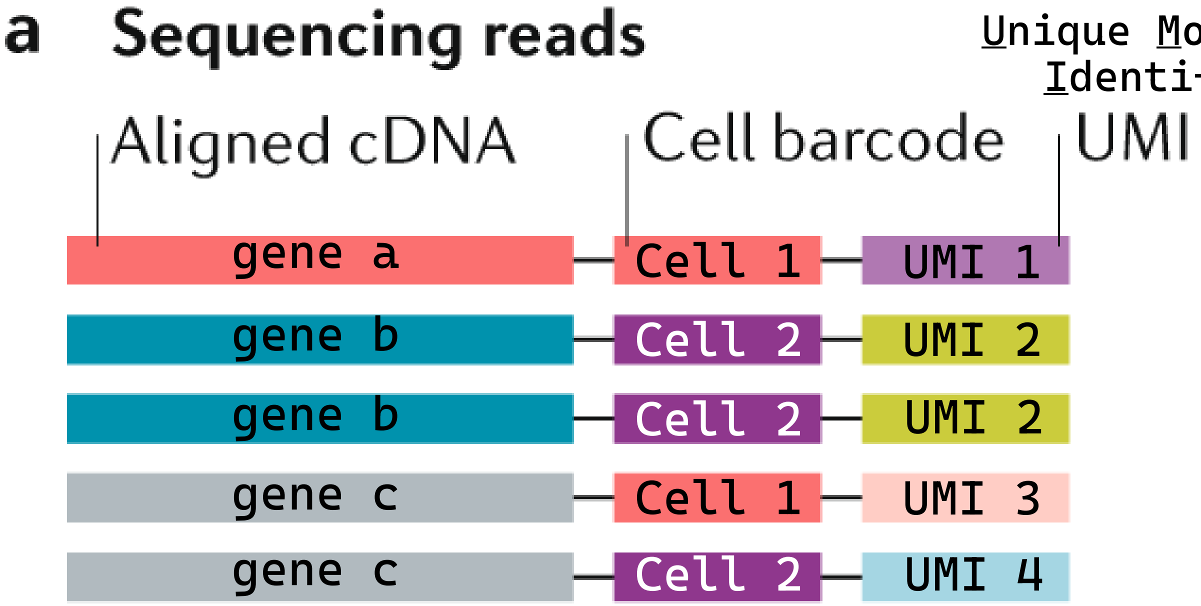
- Align sequences to identify genes
- De-multiplex all barcodes (samples, cells, UMI) and to create gene expression matrices
- Remove dead cells
- Empty Droplet Detection
- Adjust for ambient RNA
- Correct batch effects
    - Unsupervised clustering can be heavily impacted by batch effects before correction and aggregation




Experimental Design
Analysis begins with design. Choose assay types and define controls & replicates to set your course.
Lab Workflow
Use analysis-aware execution to avoid road-blocking technical artifacts.
Raw Data Processing
Transform your sequencing data into quantitative biological information.
Quality Control
Assess the quality of your data at multiple stages of analysis.
Visualization & Interpretation
Arrive at biological insights and address research questions with Loupe Browser and community-developed tools.
Research Goals
Controls
Replicates
Data Files
10x Ranger Pipelines
QC Metrics
Batch Correction
Filtering
Normalization
Cell Annotation
Clustering
Differential Expression
Cell Ordering
Data Integration
More!

10X Genomics

Sequence Alignment and Binning

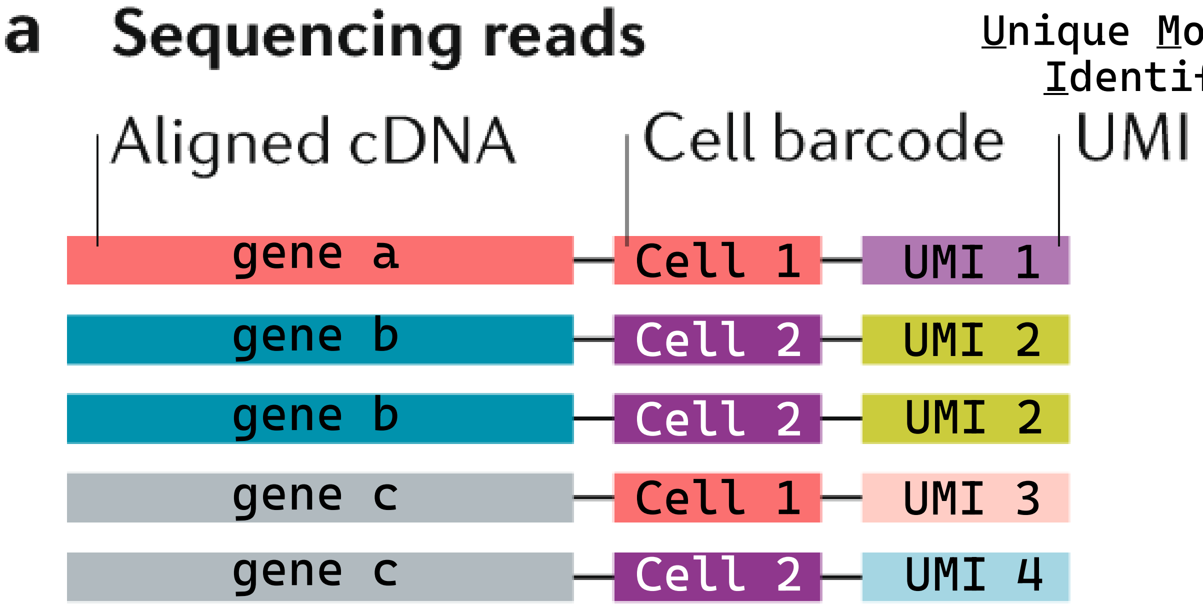


Unique Molecule Identifier

Aligned cDNA	Cell barcode	UMI
gene a	Cell 1	UMI 1
gene b	Cell 2	UMI 2
gene b	Cell 2	UMI 2
gene c	Cell 1	UMI 3
gene c	Cell 2	UMI 4

Wu & Zhang Nature 2020

Sequence Alignment and Binning



a Sequencing reads

Unique Molecule Identifier

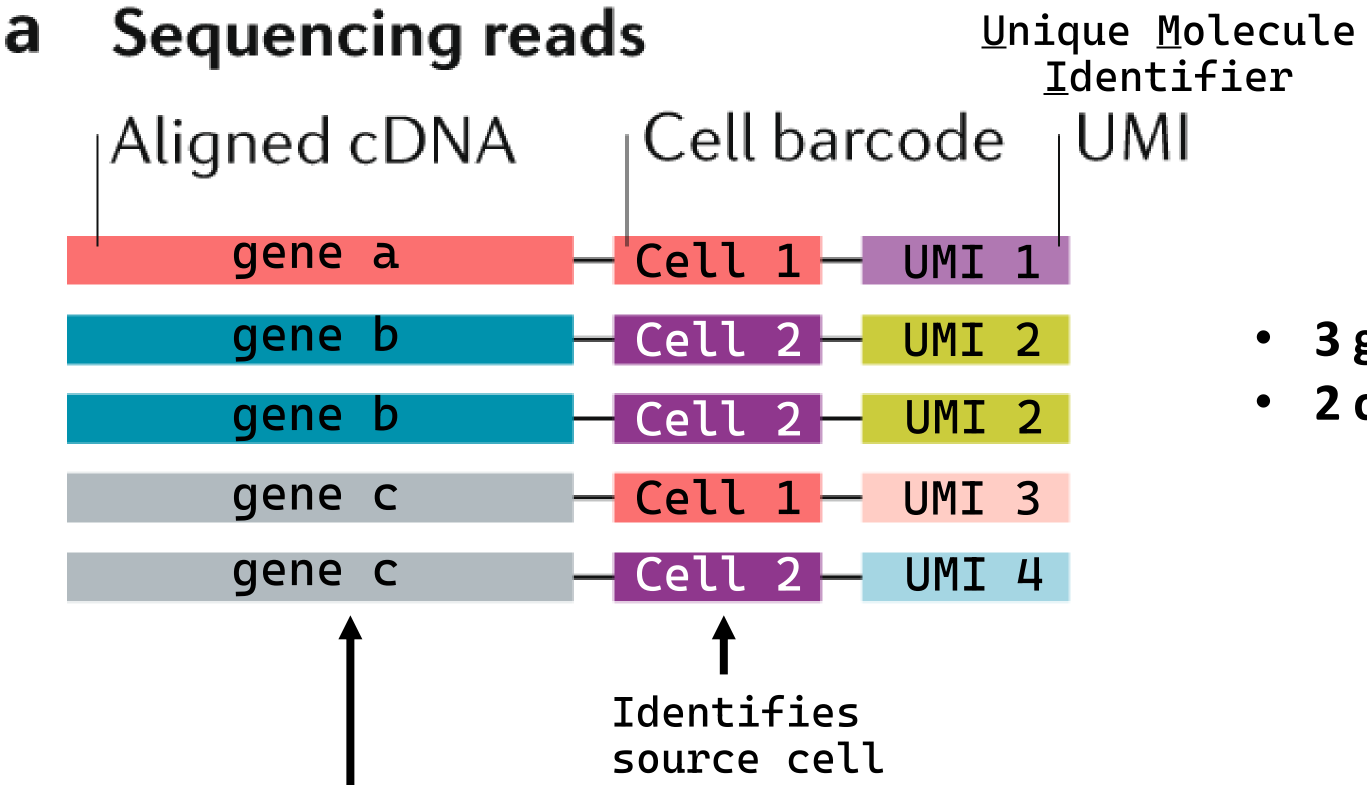
Aligned cDNA	Cell barcode	UMI
gene a	Cell 1	UMI 1
gene b	Cell 2	UMI 2
gene b	Cell 2	UMI 2
gene c	Cell 1	UMI 3
gene c	Cell 2	UMI 4

cDNA fragment is aligned
against a reference genome

- **3 genes detected**

Wu & Zhang Nature 2020

Sequence Alignment and Binning



Unique Molecule Identifier

Aligned cDNA	Cell barcode	UMI
gene a	Cell 1	UMI 1
gene b	Cell 2	UMI 2
gene b	Cell 2	UMI 2
gene c	Cell 1	UMI 3
gene c	Cell 2	UMI 4

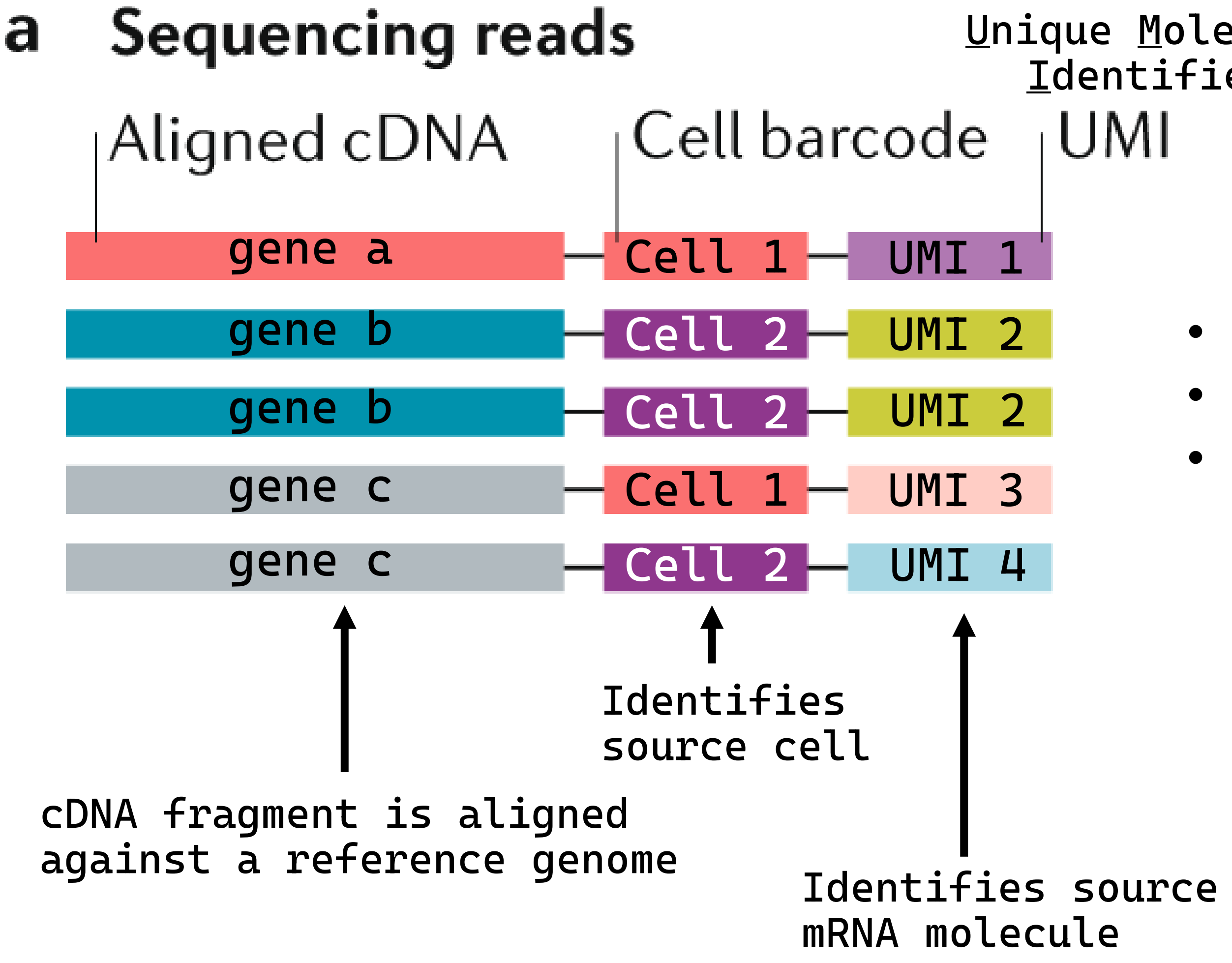
Identifies source cell

cDNA fragment is aligned against a reference genome

- **3 genes detected**
- **2 cells detected**

Wu & Zhang Nature 2020

Sequence Alignment and Binning



Unique Molecule Identifier

Aligned cDNA	Cell barcode	UMI
gene a	Cell 1	UMI 1
gene b	Cell 2	UMI 2
gene b	Cell 2	UMI 2
gene c	Cell 1	UMI 3
gene c	Cell 2	UMI 4

cDNA fragment is aligned against a reference genome

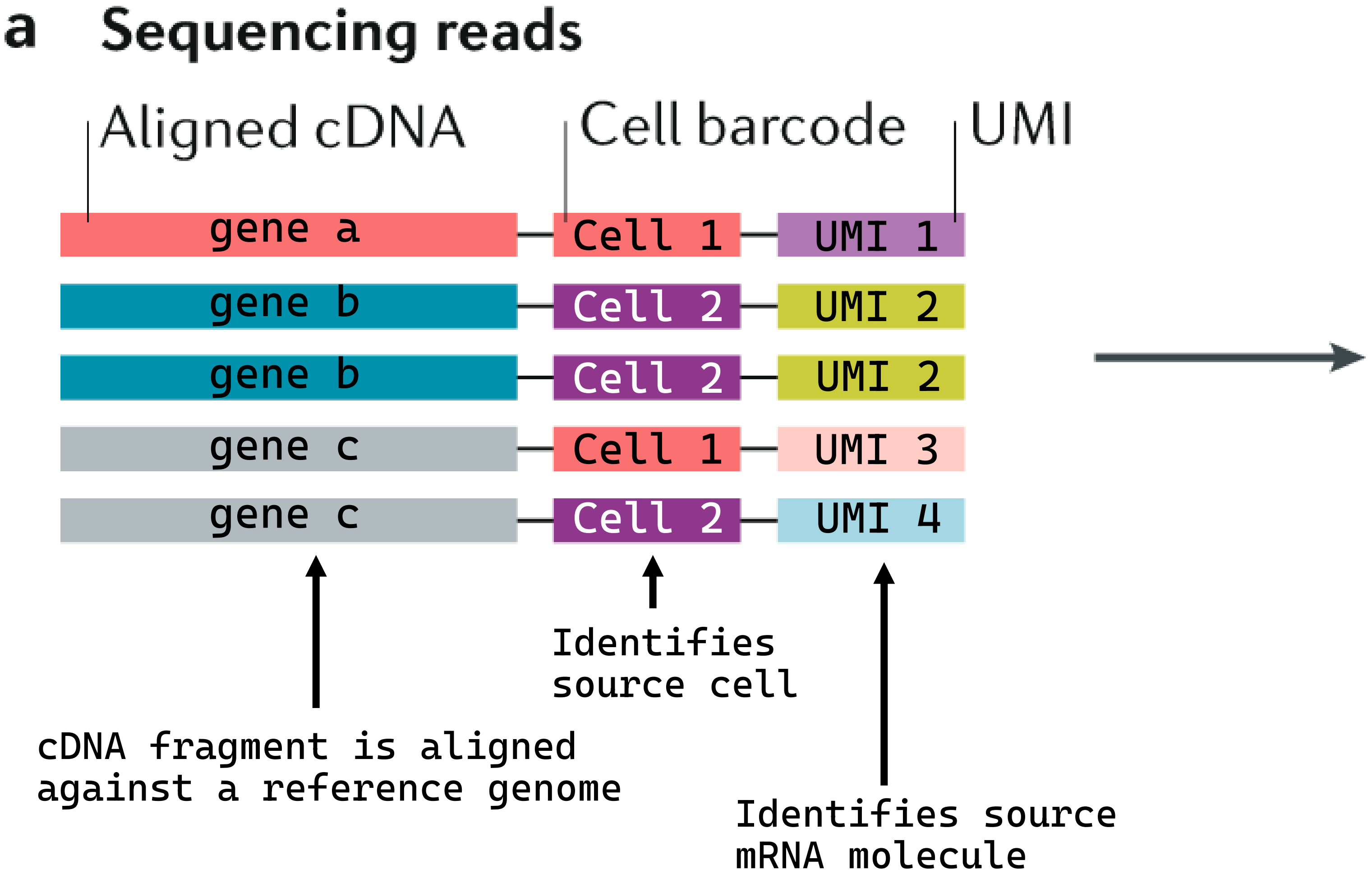
Identifies source cell

Identifies source mRNA molecule

- **3 genes detected**
- **2 cells detected**
- **4 unique original transcript molecules**

Wu & Zhang Nature 2020

Sequence Alignment and Binning



a Sequencing reads

Aligned cDNA | Cell barcode | UMI

gene a — Cell 1 — UMI 1

gene b — Cell 2 — UMI 2

gene b — Cell 2 — UMI 2

gene c — Cell 1 — UMI 3

gene c — Cell 2 — UMI 4

cDNA fragment is aligned against a reference genome

Identifies source cell

Identifies source mRNA molecule

Gene expression matrix

Cells

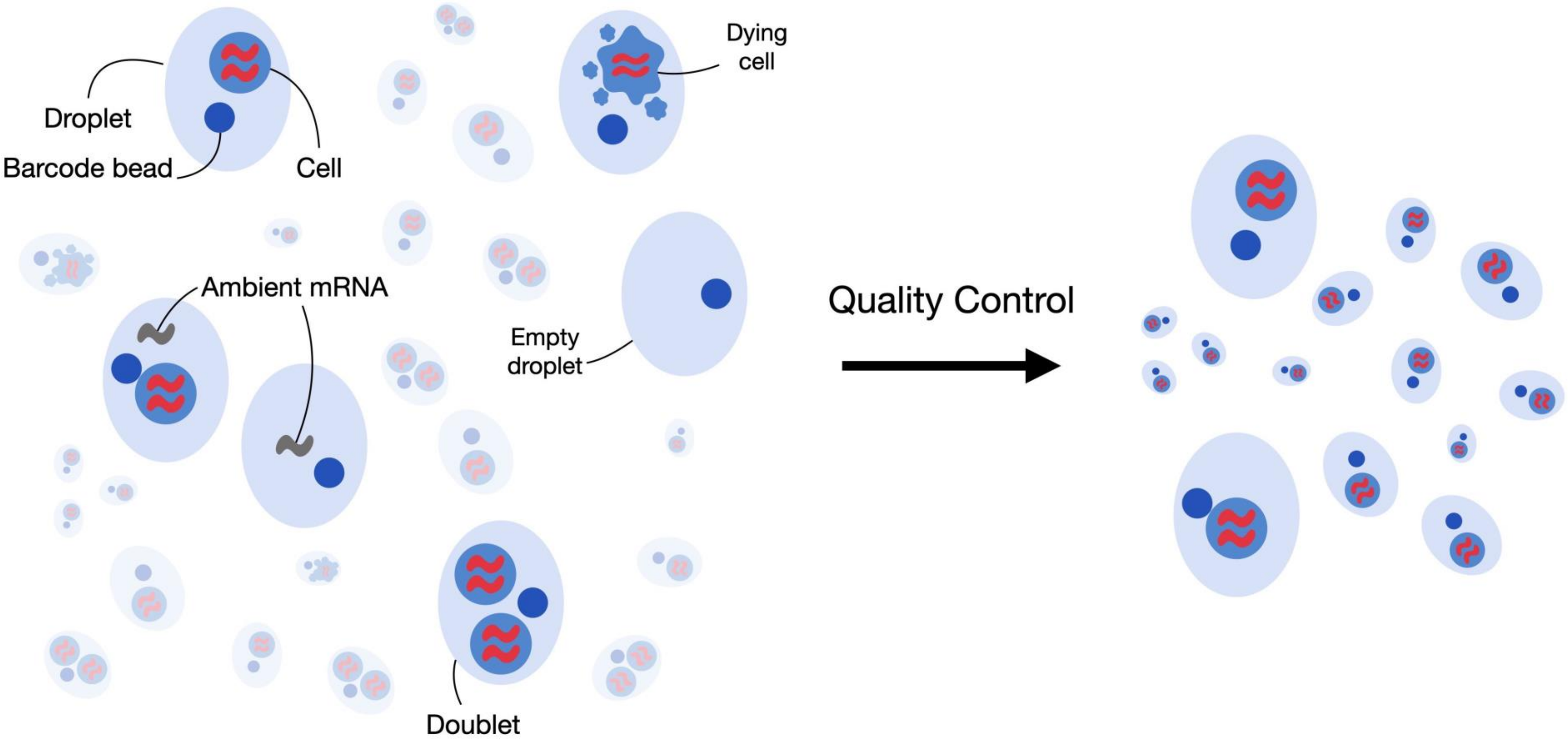
Genes	1	2	3	4 → n Cells	
a	1.2	0.3	2.1	3.6	.....
b	3.2	1.9	5.2	1.1	.....
c	2.6	4.6	0.8	2.2	.....
d	0.6	3.3	0.9	4.4	.....
↓ n genes	⋮	⋮	⋮	⋮	

20 000 gene dimensions!!!
At 10 000 cells/ sample:
500 million data points per sample!!!

Wu & Zhang Nature 2020

Quality Control

Many scenarios of imperfect droplets that can affect your data!

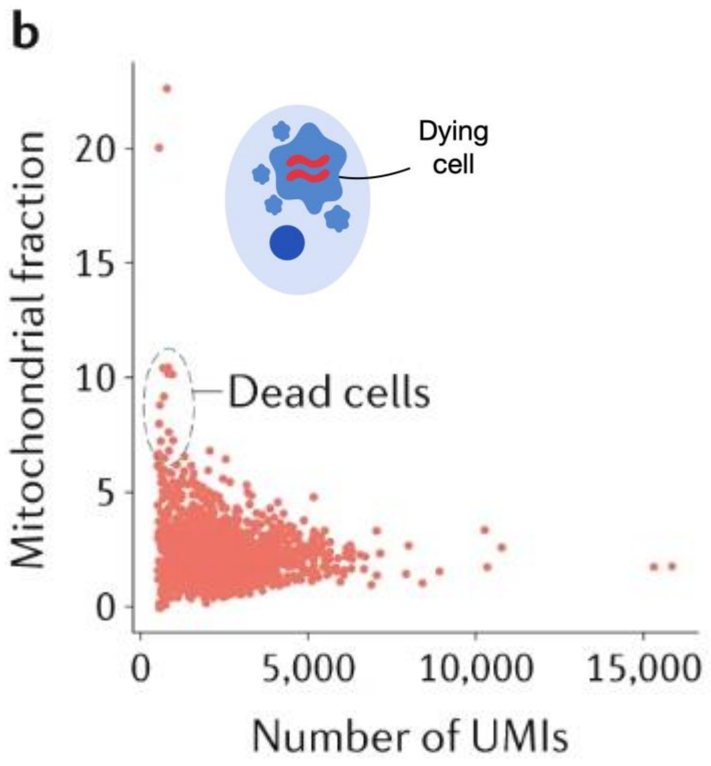


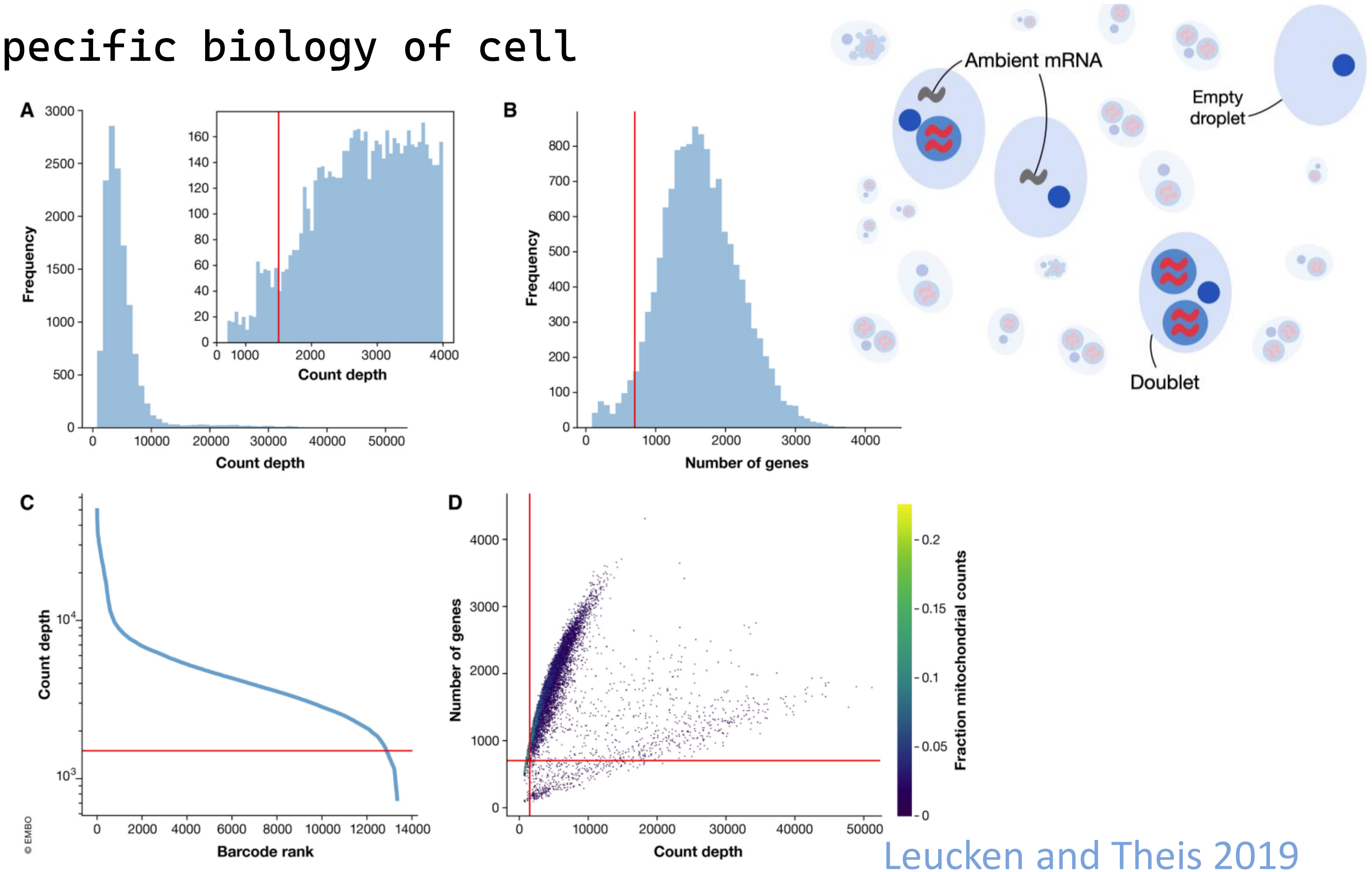
sc-best-practices.org

Doublet and Empty Droplet Removal

Gene counts, identified genes and mitochondrial gene expression all important

*Can be heavily impacted by specific biology of cell populations





Leucken and Theis 2019

Single Cell Gene Expression

Dimension Reduction

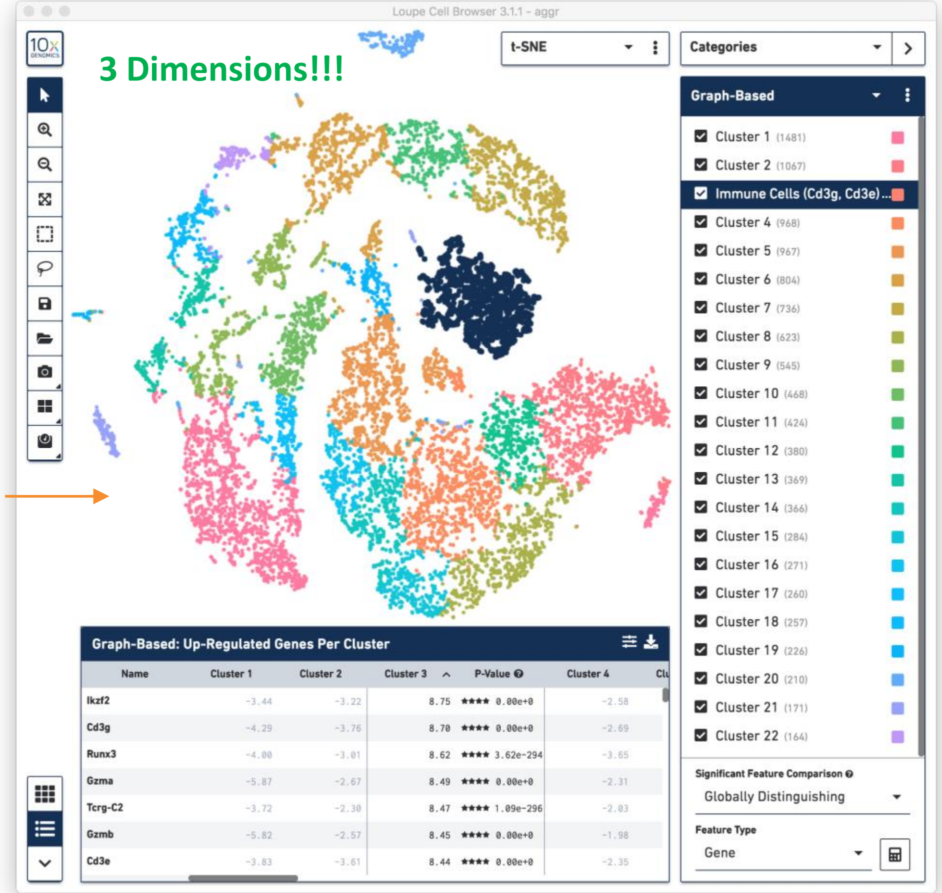
- Very difficult to visualize thousands of dimensions of data at once
- Use fancy stats and data science techniques (clustering) to find patterns and associations within the data to group
- Supervised Clustering
    - Takes user input on cluster number
- Unsupervised
    - Will make as many clusters as it thinks exists, depending on variance limits for specific algorithms

Gene Count Matrices

	Cell 1	Cell 2	Cell *x* *-> 10000*
Gene 1	3	170	…
Gene 2	500	30	…
Gene *x* *-> 20000*	…	…	…

20 000 genes *10 000 cells/ sample:
500 million data points per sample!!!

Clustering type: k-mer clustering (unsupervised)



3 Dimensions!!!

Graph-Based

- ☑ Cluster 1 (1481)
- ☑ Cluster 2 (1067)
- ☑ Immune Cells (Cd3g, Cd3e)…
- ☑ Cluster 4 (968)
- ☑ Cluster 5 (967)
- ☑ Cluster 6 (804)
- ☑ Cluster 7 (736)
- ☑ Cluster 8 (623)
- ☑ Cluster 9 (545)
- ☑ Cluster 10 (468)
- ☑ Cluster 11 (424)
- ☑ Cluster 12 (380)
- ☑ Cluster 13 (369)
- ☑ Cluster 14 (366)
- ☑ Cluster 15 (284)
- ☑ Cluster 16 (271)
- ☑ Cluster 17 (260)
- ☑ Cluster 18 (257)
- ☑ Cluster 19 (226)
- ☑ Cluster 20 (210)
- ☑ Cluster 21 (171)
- ☑ Cluster 22 (164)

Significant Feature Comparison: Globally Distinguishing

Feature Type: Gene

Graph-Based: Up-Regulated Genes Per Cluster

Name	Cluster 1	Cluster 2	Cluster 3	P-Value	Cluster 4	Clu
Ikzf2	-3.44	-3.22	8.75	**** 0.00e+0	-2.58	
Cd3g	-4.29	-3.76	8.70	**** 0.00e+0	-2.69	
Runx3	-4.00	-3.01	8.62	**** 3.62e-294	-3.65	
Gzma	-5.87	-2.67	8.49	**** 0.00e+0	-2.31	
Tcrg-C2	-3.72	-2.30	8.47	**** 1.09e-296	-2.03	
Gzmb	-5.82	-2.57	8.45	**** 0.00e+0	-1.98	
Cd3e	-3.83	-3.61	8.44	**** 0.00e+0	-2.35	

Not necessary to understand the underlying data science mathematics in order to understand what the algorithm functions. This is an excellent resource for learning about data science and machine learning techniques without needing any coding or advanced mathematic knowledge:

https://machinelearningmastery.com/start-here/#algorithms

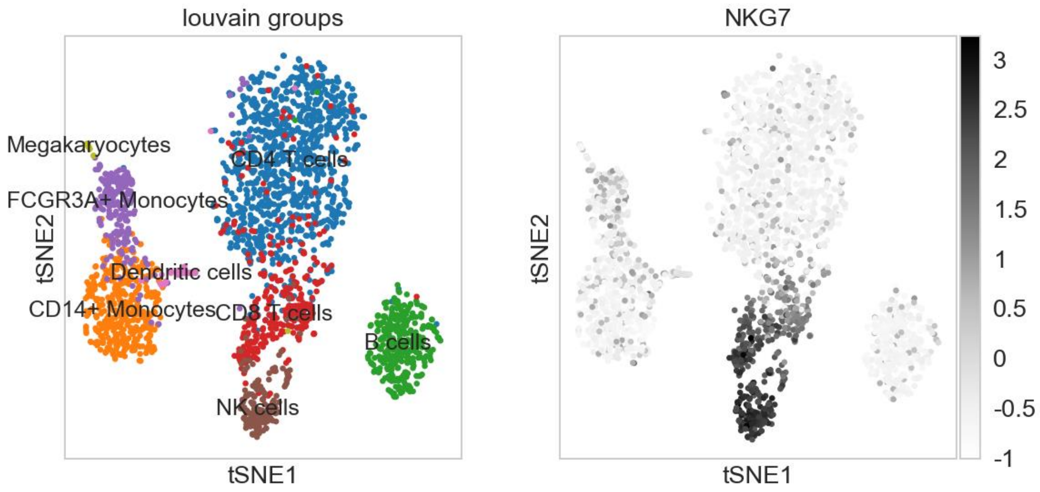
10X Genomics

Single Cell Gene Expression

Now can do some really interesting analyses!

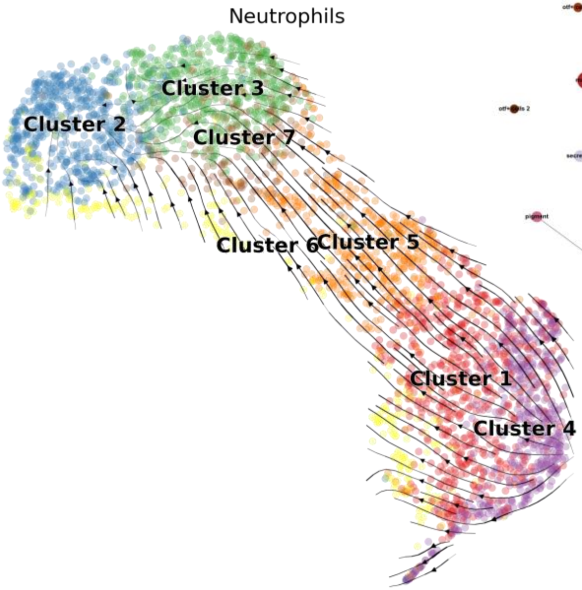
Analysis

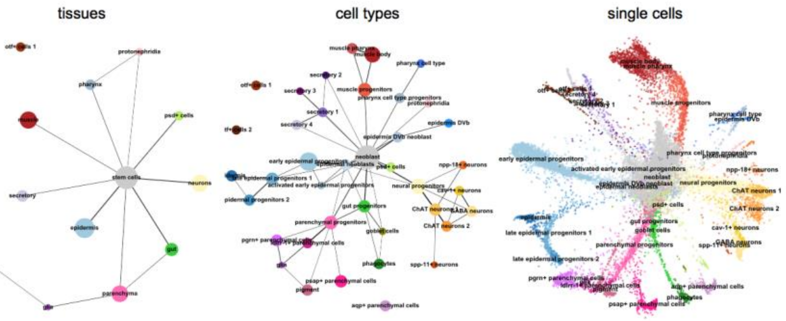
- Clustering
- Differential expression
- Associated genes
- Leads on pathways and mechanisms for novel markers
- Endless data mining
  - Revisit previous experiments with new genes of interest



Trajectory Analysis

- What direction are progenitor cells differentiating to, in what proportions? How is this affected by experimental conditions?



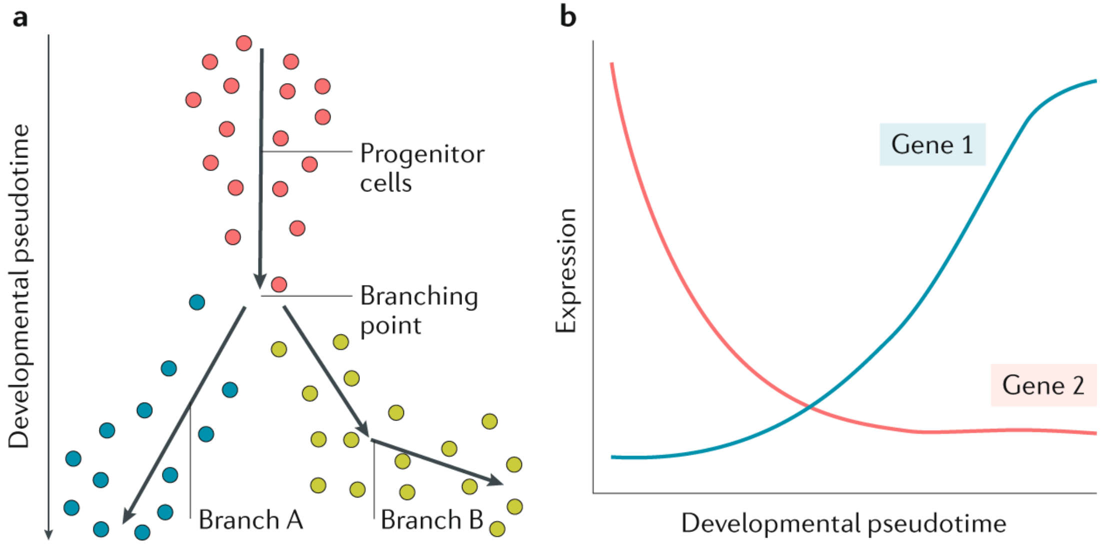


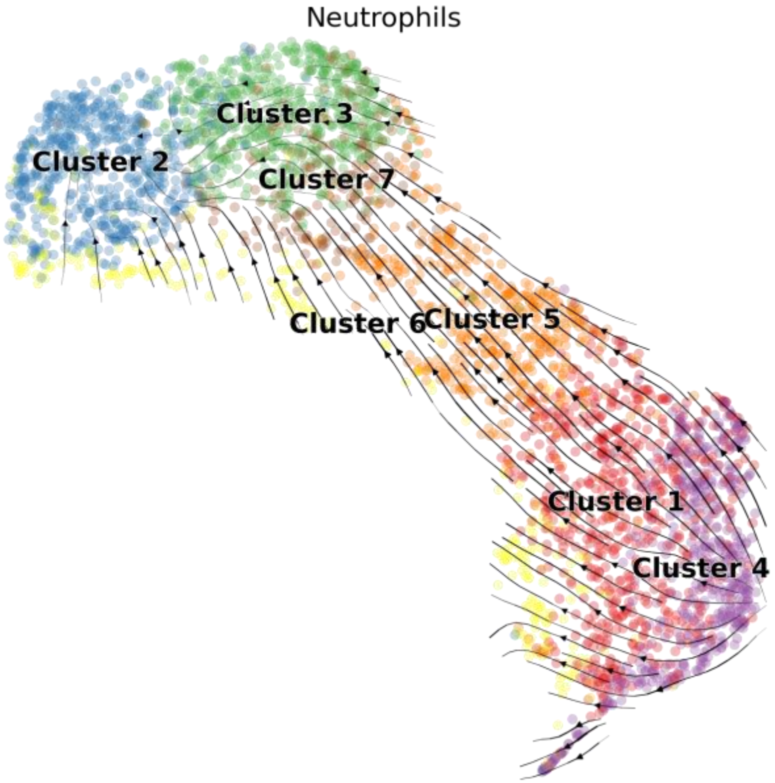
[10X Genomics](#)

Single Cell Gene Expression

Trajectory Inference Analysis

- What direction are progenitor cells differentiating to, in what proportions? How is this affected by experimental conditions?





- Vaccine immunity development
- In vitro stem cell differentiation

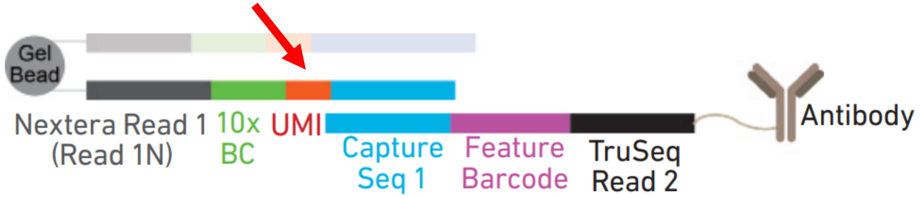
Nature

Additional Single Cell Assays

Protein Expression

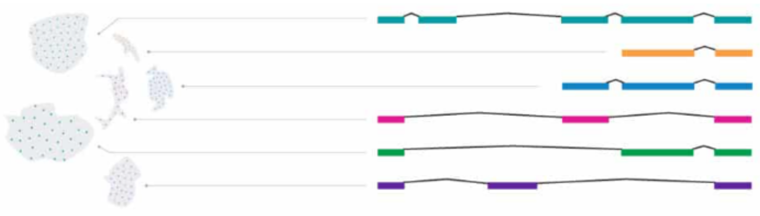
- Tag cell surface moieties with antibodies that are bound to oligos carrying a Capture Sequence and Feature Barcode
- Acts in place of a transcript once in the partitioned barcoding reaction

UMI – Allows Quantification



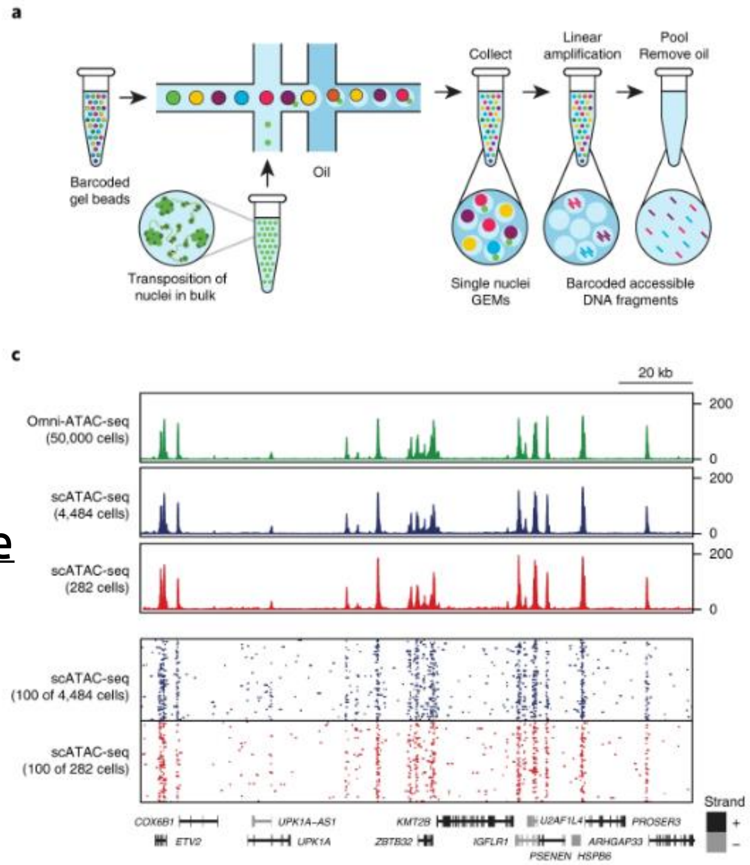
B Cell and T Cell receptor sequencing

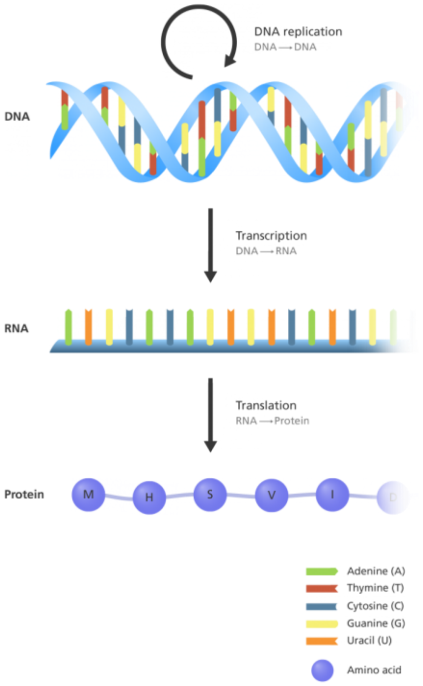
Full length transcript sequencing for splice variants (PacBio long read sequencing)



DNA availability

Transposase-accessible chromatin with sequencing (ATAC-Seq)





10X Genomics

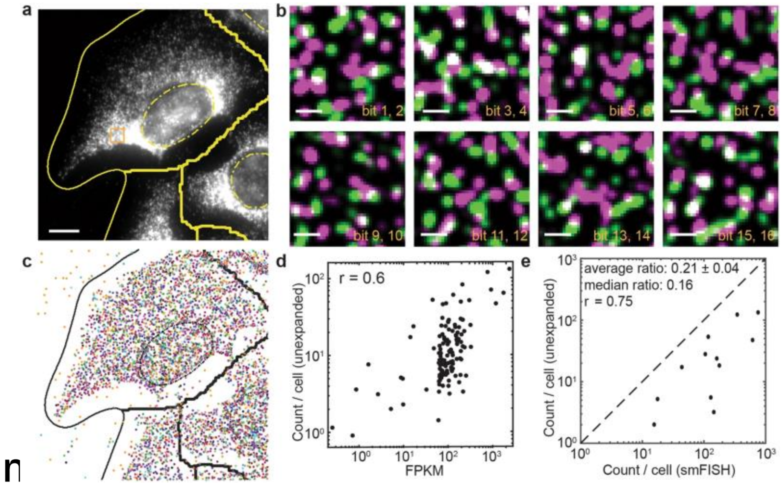
Spatial Transcriptomics

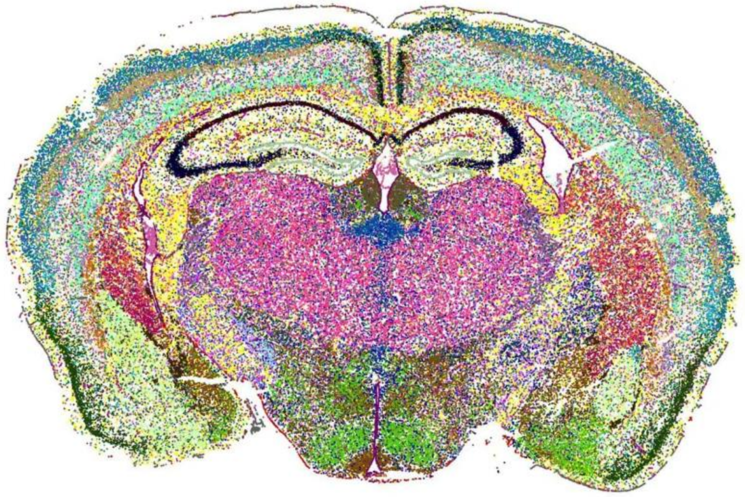
- The next step in single cell analysis – context within a tissue
- Analysis on whole tissue sections of over 1cm^2, down to subcellular resolution

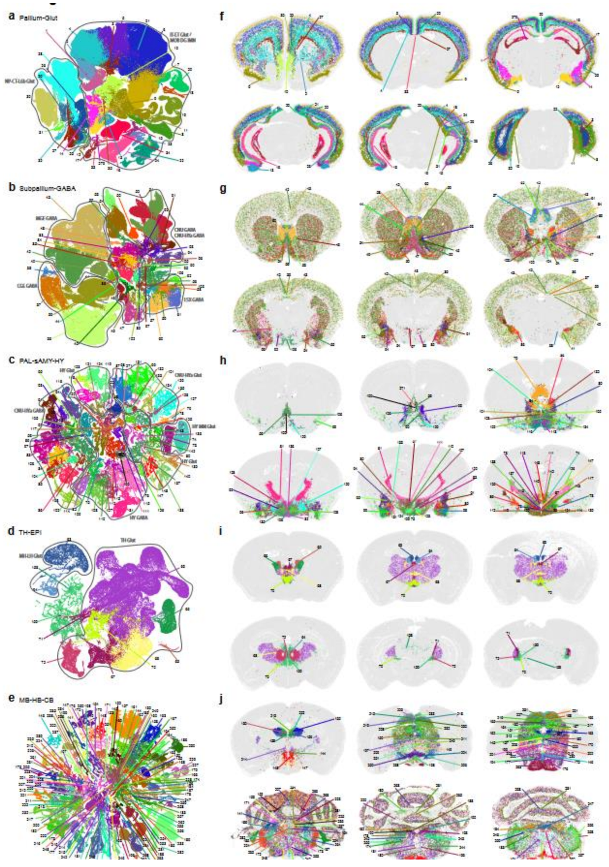
<u>Techniques</u>

- Full transcriptomic sequencing (array)
- Cyclic in-situ hybridization
- Multiplexed error-robust fluorescence in situ hybridization (MERFISH) optical barcoding
  - MERSCOPE technology of 500 gene in-situ hybridization of ~5 million cells
  - Combined with 10 million cell neuron scRNA seq datasets to create the first mouse brain atlas

Wang *et al* 2018







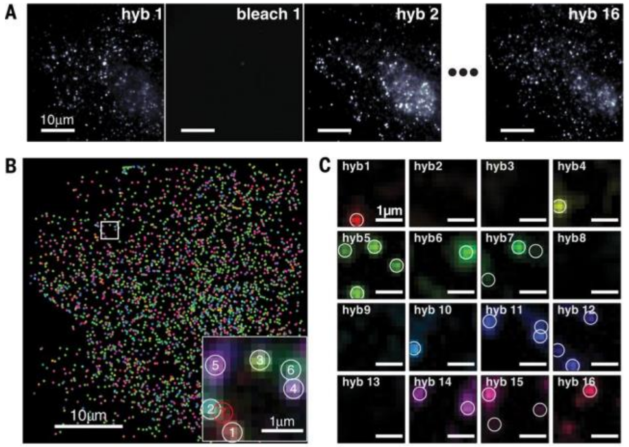
Yao *et al* 2023 Preprint

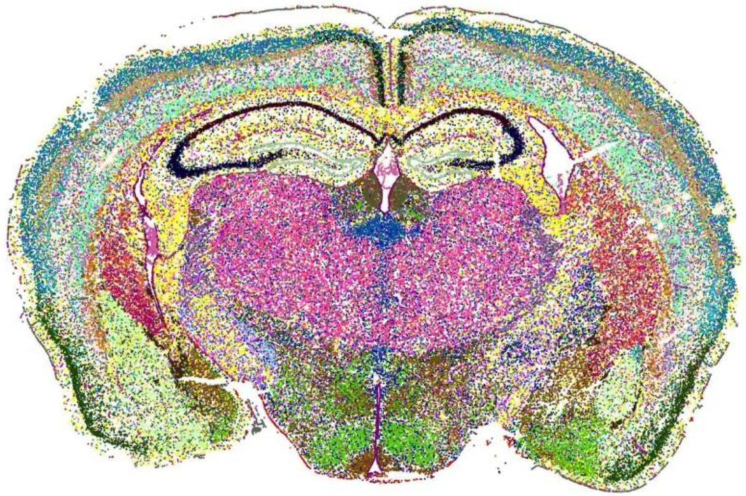
Spatial Transcriptomics

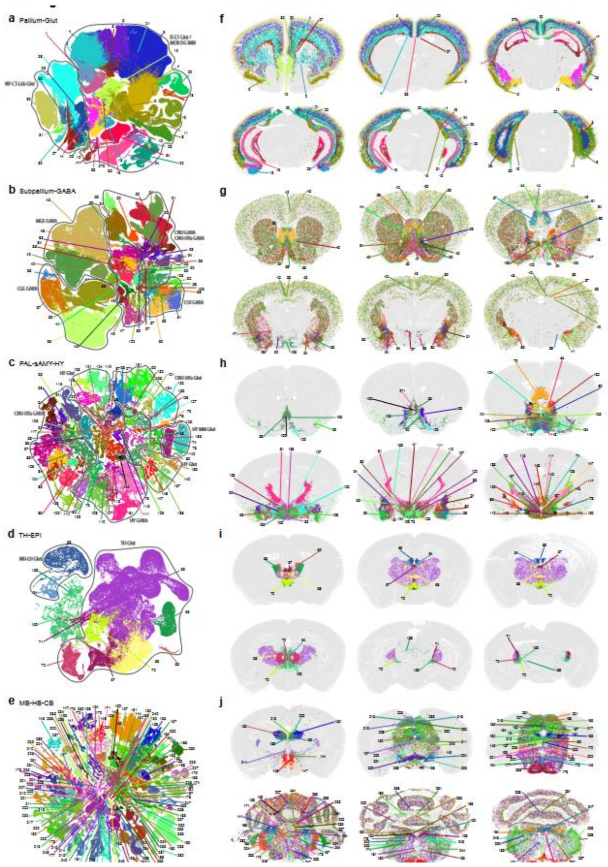
- The next step in single cell analysis – context within a tissue
- Analysis on whole tissue sections of over $1cm^2$, down to subcellular resolution

<u>Techniques</u>

- Full transcriptomic sequencing (array)
- Cyclic in-situ hybridization
- Multiplexed error-robust fluorescence in situ hybridization (MERFISH)
  - MERSCOPE technology of 500 gene in-situ hybridization of ~5 million cells
  - Combined with 10 million cell neuron scRNA seq datasets to create the first mouse brain atlas

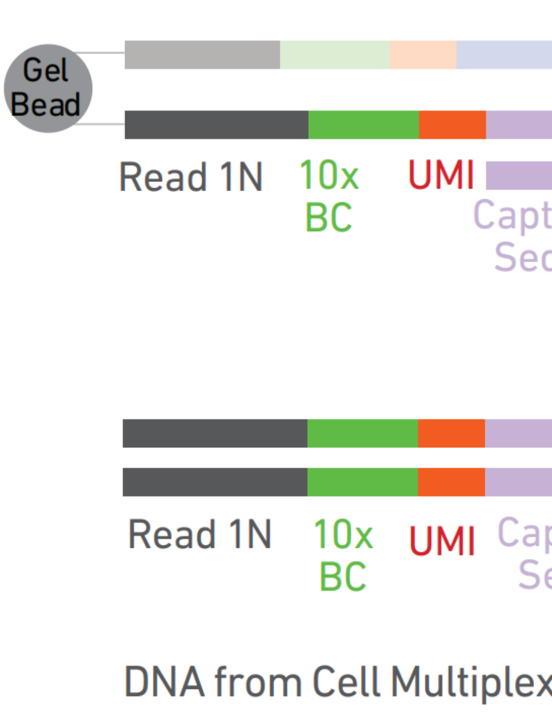






Yao *et al* 2023 Preprint

Single Cell ATAC + Gene Expression



Layers of single cell Interrogation

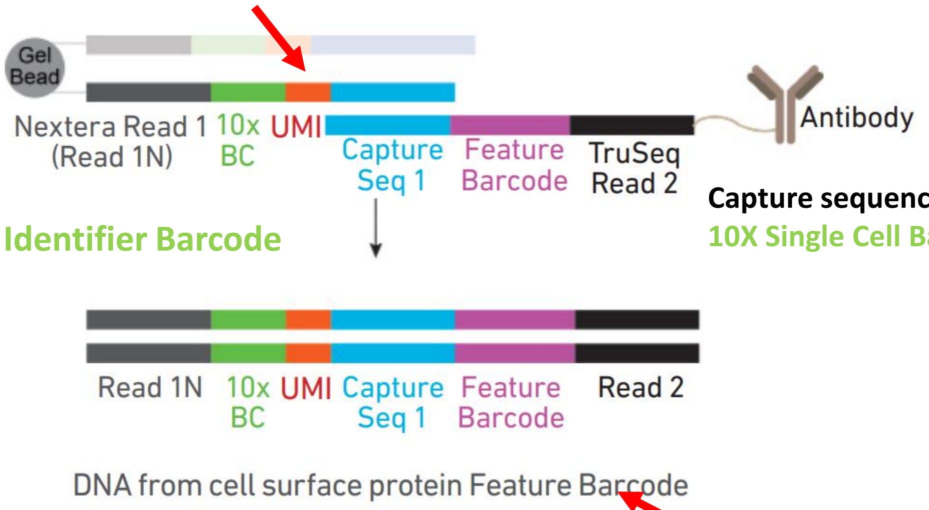
Cell Surface Feature Barcoding

- Tag cell surface moieties with antibodies that are bound to Oligos carrying a Capture Sequence and Feature Barcode
- Acts in place of a transcript once in the partitioned barcoding reaction

Trajectory Analysis

- What direction are proge differentiating to, in w proportions? How is this experimental conditions?

UMI – Allows Quantification



Single Cell Identifier Barcode

Capture sequence allows barcoding with 10X Single Cell Barcode and UMI

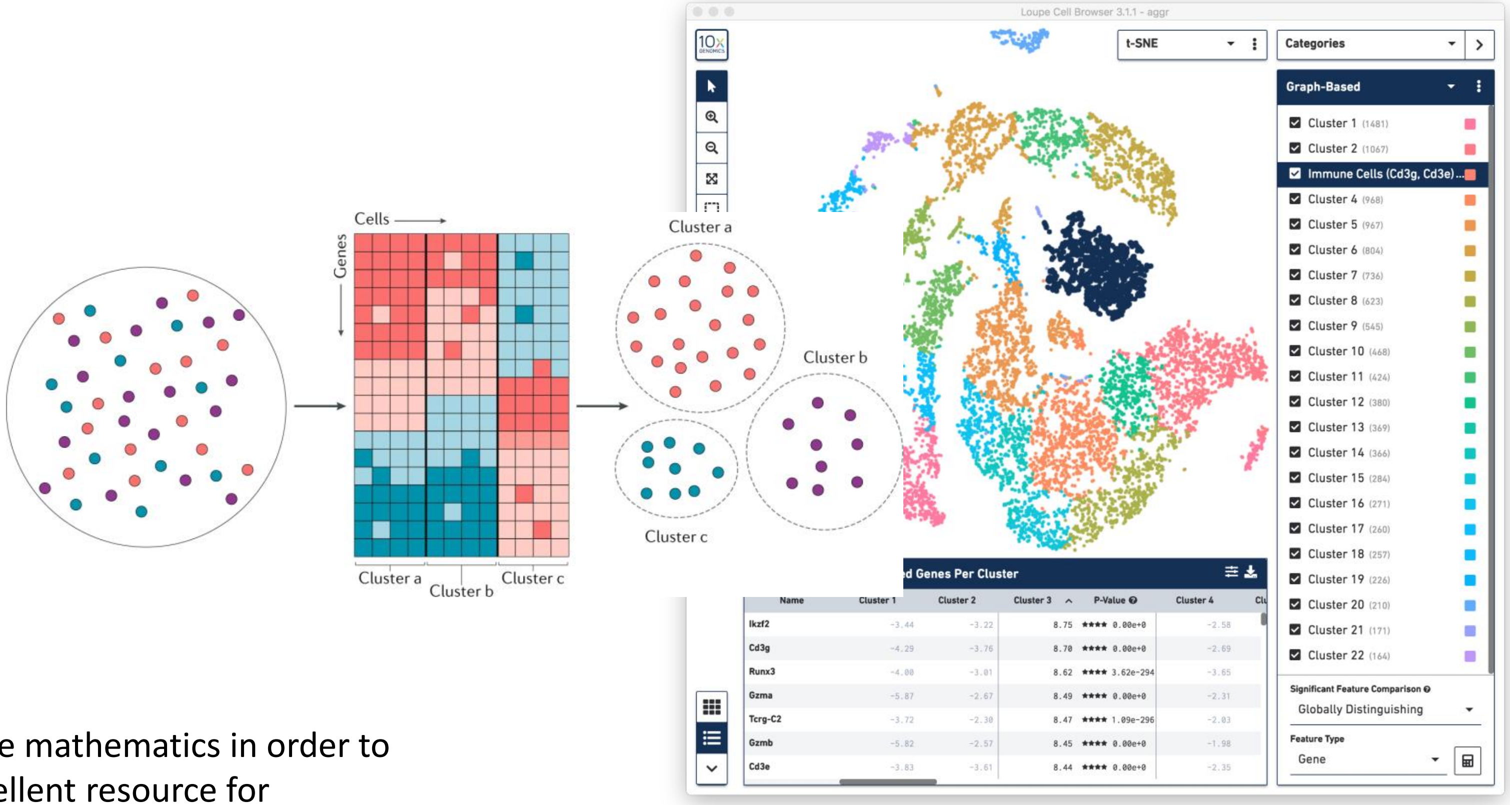
Feature Barcode – Moiety Identification

10X Genomics

Single Cell Gene Expression

Dimension Reduction

- Very difficult to visualize thousands of dimensions of data at once
- Use fancy stats and data science techniques (clustering) to find patterns and associations within the data to group
- Supervised Clustering
  - Takes user input on cluster number
- Unsupervised
  - Will make as many clusters as it thinks exists, depending on variance limits for specific algorithms



Categories

Graph-Based

- ☑ Cluster 1 (1481) 🟥
- ☑ Cluster 2 (1067) 🟥
- ☑ Immune Cells (Cd3g, Cd3e)... 🟥
- ☑ Cluster 4 (968) 🟧
- ☑ Cluster 5 (967) 🟧
- ☑ Cluster 6 (804) 🟧
- ☑ Cluster 7 (736) 🟨
- ☑ Cluster 8 (623) 🟨
- ☑ Cluster 9 (545) 🟩
- ☑ Cluster 10 (468) 🟩
- ☑ Cluster 11 (424) 🟩
- ☑ Cluster 12 (380) 🟩
- ☑ Cluster 13 (369) 🟩
- ☑ Cluster 14 (366) 🟦
- ☑ Cluster 15 (284) 🟦
- ☑ Cluster 16 (271) 🟦
- ☑ Cluster 17 (260) 🟦
- ☑ Cluster 18 (257) 🟦
- ☑ Cluster 19 (226) 🟦
- ☑ Cluster 20 (210) 🟦
- ☑ Cluster 21 (171) 🟪
- ☑ Cluster 22 (164) 🟪

Significant Feature Comparison: Globally Distinguishing

Feature Type: Gene

d Genes Per Cluster

Name	Cluster 1	Cluster 2	Cluster 3	P-Value	Cluster 4	Clu
Ikzf2	-3.44	-3.22	8.75	★★★★ 0.00e+0	-2.58	
Cd3g	-4.29	-3.76	8.70	★★★★ 0.00e+0	-2.69	
Runx3	-4.00	-3.01	8.62	★★★★ 3.62e-294	-3.65	
Gzma	-5.87	-2.67	8.49	★★★★ 0.00e+0	-2.31	
Tcrg-C2	-3.72	-2.30	8.47	★★★★ 1.09e-296	-2.03	
Gzmb	-5.82	-2.57	8.45	★★★★ 0.00e+0	-1.98	
Cd3e	-3.83	-3.61	8.44	★★★★ 0.00e+0	-2.35	

Not necessary to understand the underlying data science mathematics in order to understand what the algorithm functions. This is an excellent resource for learning about data science and machine learning techniques without needing any coding or advanced mathematic knowledge

https://machinelearningmastery.com/start-here/#algorithms

10X Genomics