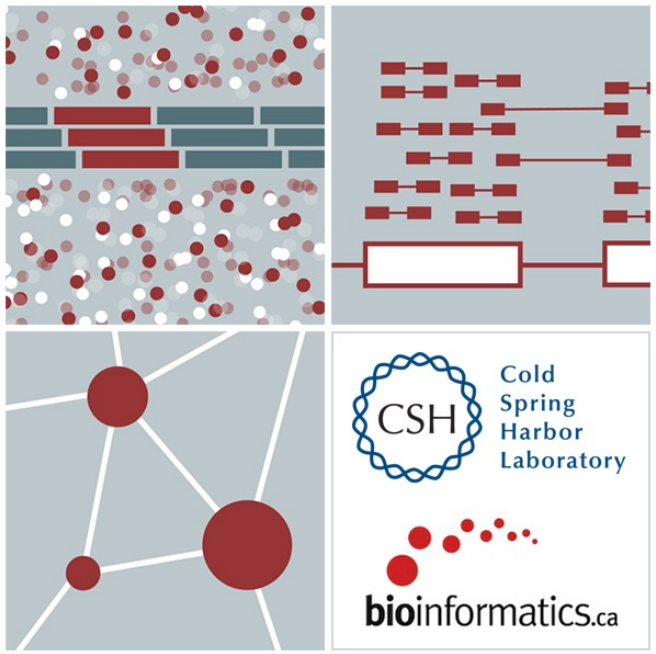
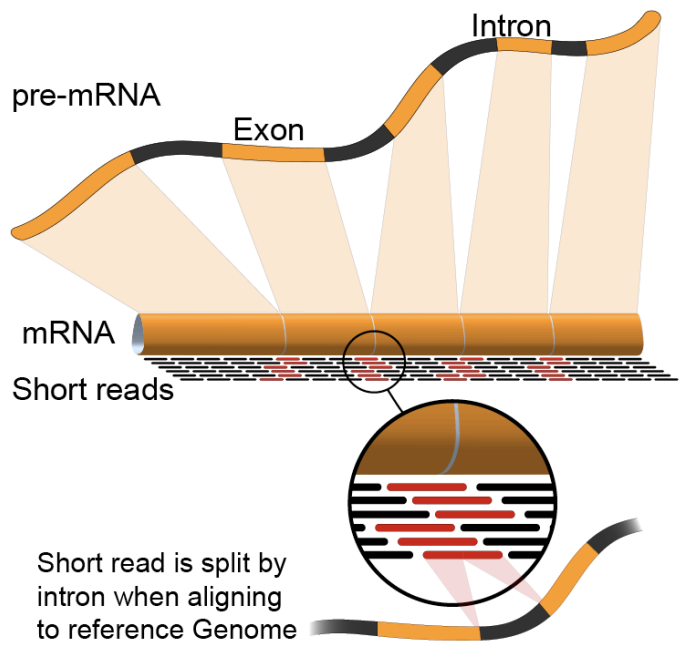
Introduction to single cell RNA sequencing and analysis

Kelsy Cotto, Malachi Griffith, Obi Griffith, Charles Hayfron-Benjamin, Evelyn Schmidt, Kartik Singhal, Zach Skidmore

CRI Bioinformatics Workshop. Apr 27-May 2, 2024





Washington University in St. Louis
School of Medicine

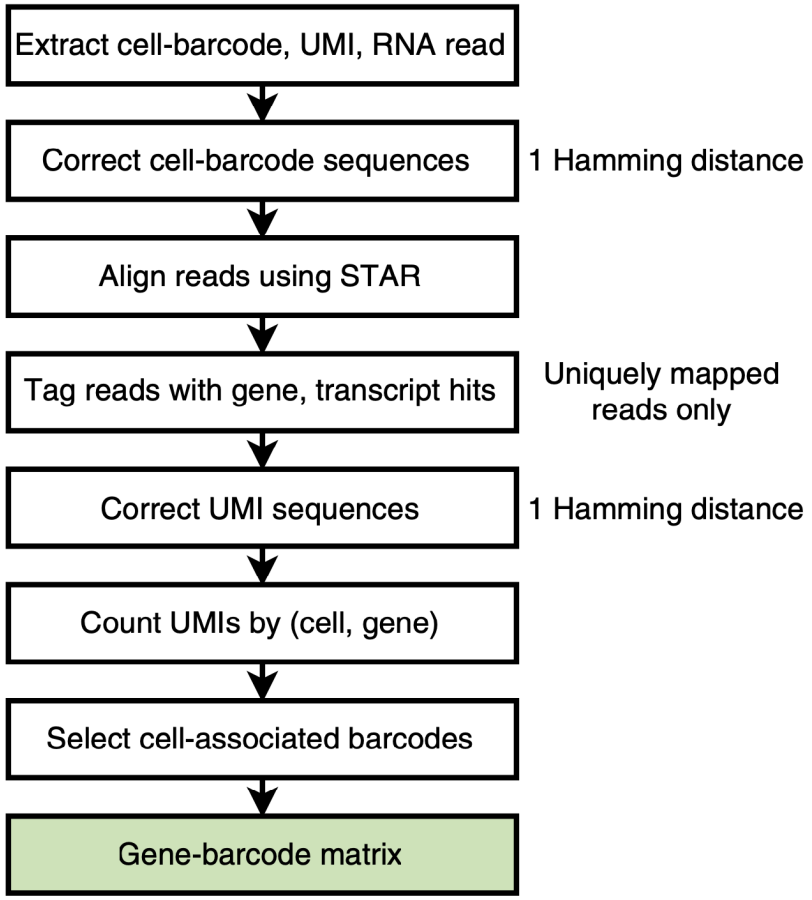
Acknowledgements

- Jon Preall, CSHL, SEQTEC
- Jennifer Foltz, WashU, Genomics in Medicine
- Trevor Pugh, PMH, CBW
- Allegra Petti, MGH/Harvard
- Brian Haas, Broad Institute

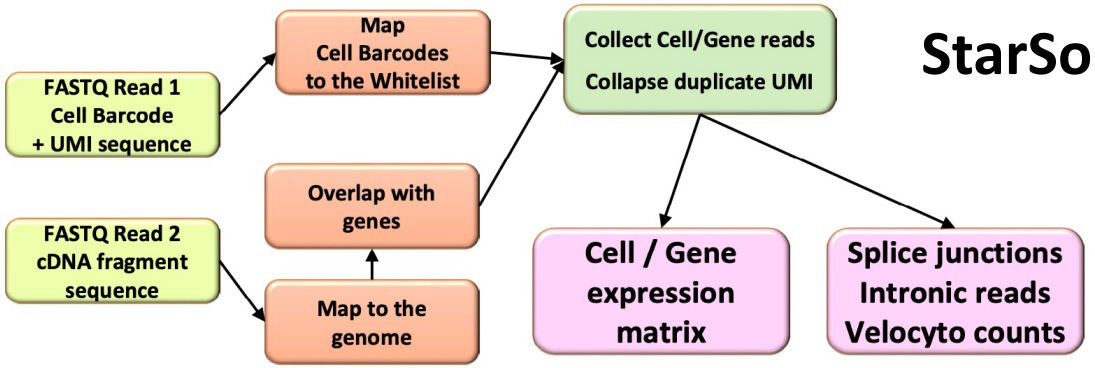




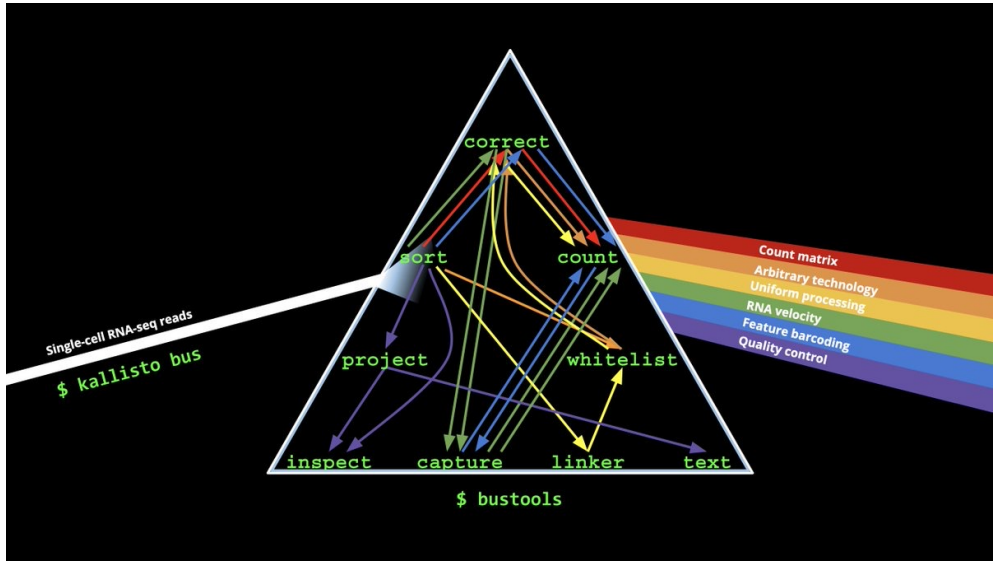
scRNA tools (primary analysis)



Cell Ranger



StarSolo



kallisto/ bustools

Mapping and Counting

Cellranger / STARsolo / etc

- Extract Cell Barcode, UMI, Feature read
- Correct Cell Barcodes — 1 hamming distance
- Align to reference with STAR
- Assign reads to genes with GTF file — Unique mappers
- UMI correction — 1 hamming distance
- Collapse UMIs, tally genes/cells
- Raw Gene/Cell Matrix
- Filter empty droplets, keep cells — (EmptyDrops)
- Filtered Gene/Cell Matrix

Secondary Analysis

Seurat / Scanpy / Liger etc.

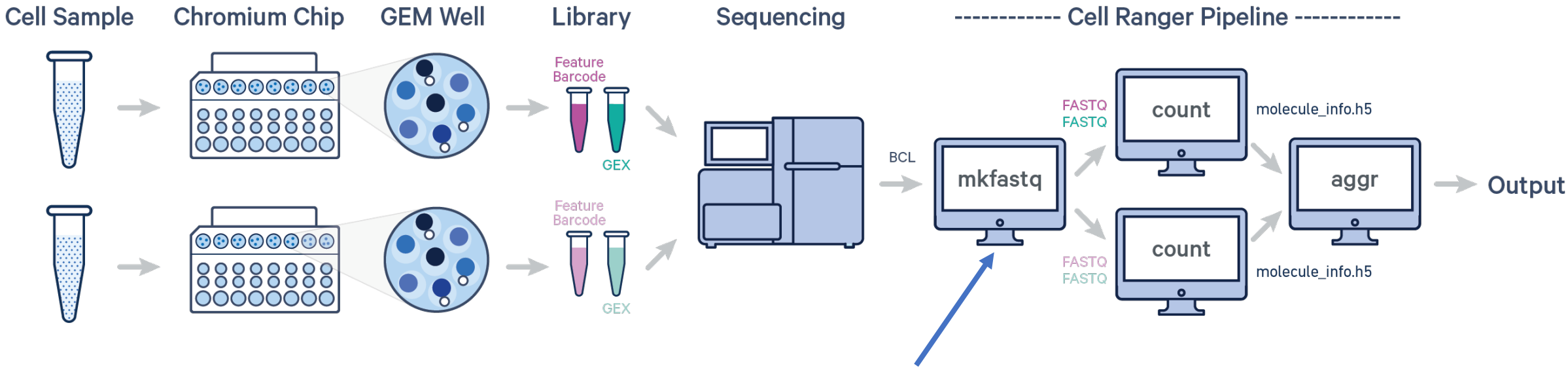
- Calculate QC metrics
- Predict doublets
- Filter cells (dead, doublets, outliers)
- Normalize & transform (i.e. log)
- Feature selection (eg. HVG)
- Linear Regression
- Zero-center & Scale
- PCA (Linear Dim. Reduc)
- Neighbor Graph
- Clustering (Leiden)
- tSNE / UMAP
- Differential Expression

10X Genomics Cell Ranger

Support page: https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

Option 1: Use 10X Genomics Cloud
credits included with reagent purchase

Option 2: Install and run on Linux system:
a. local mode (single computer)
b. cluster mode

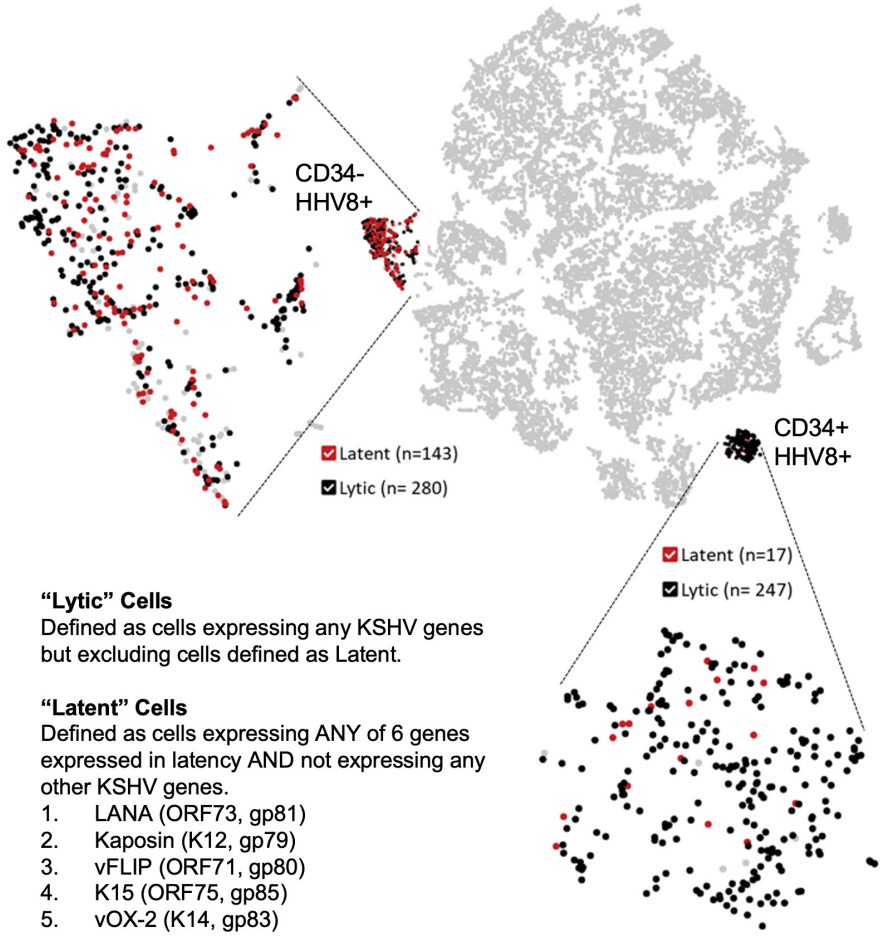


Probably run by sequence provider

Custom Genomes w/ `cellranger mkref`

https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/advanced/references

- 10X Provides Prebuilt references for:
  - Human (hg19 and GRCh38)
  - Mouse (mm9 and mm10)
- Why do I need a new reference genome?
  - Expanded annotations (eg. GENCODE, ncRNAs, etc.)
  - Additional species (eg. Maize)
  - Monitoring custom transgenes
  - Viruses, pathogens
  - Known rearrangements, model specific genomes
  - Unconventional gene annotations
- What do I need?:
  - Genome FASTA file
  - GTF file containing feature coordinates



Lee Ratner & Daniel Rauch

Running `cellranger multi (v7.0.0)`

```
SAMPLE=Rep1_ICB

cellranger multi \
   --localmem 64 \
   --localcores 4 \
   --id Rep1_ICB \
   --csv ICB_run.csv
```

icb_run

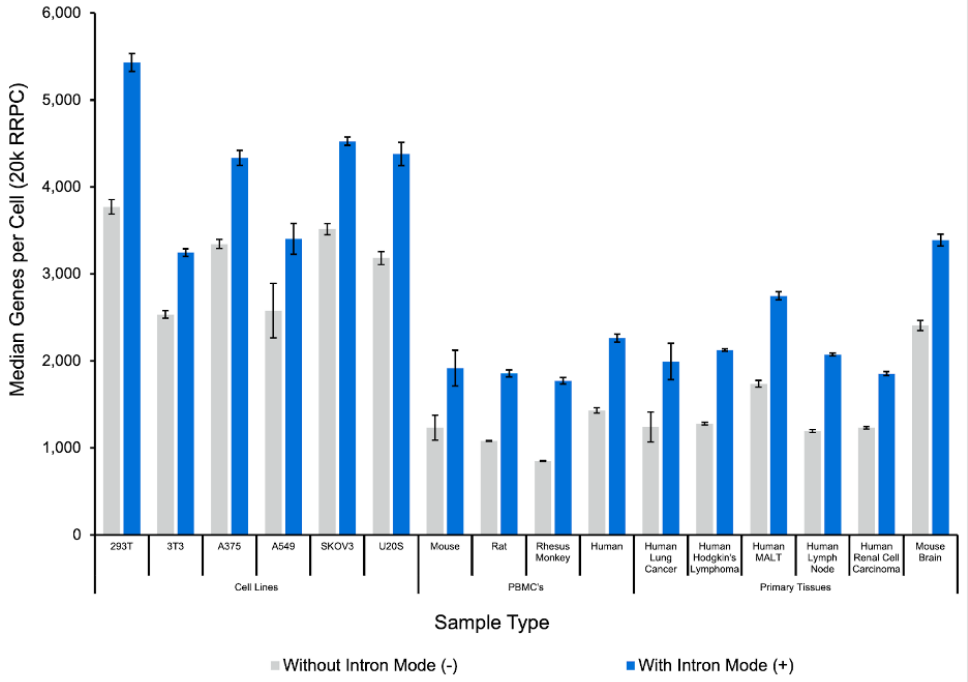
[gene-expression]		
reference	~/refdata-gex-mm10-2020-A	
[vdj]		
reference	~/refdata-cellranger-vdj-GRCm38-alts-ensembl-7.0.0	
[libraries]		
fastq_id	fastqs	feature_types
ICB_MCB6C_Rep_1-lib1	~/fastqs/Rep1/ICB/GEX	Gene Expression
ICB_MCB6C_Rep_1-lib4	~/fastqs/Rep1/ICB/BCR	VDJ-B
ICB_MCB6C_Rep_1-lib3	~/fastqs/Rep1/ICB/TCR	VDJ-T

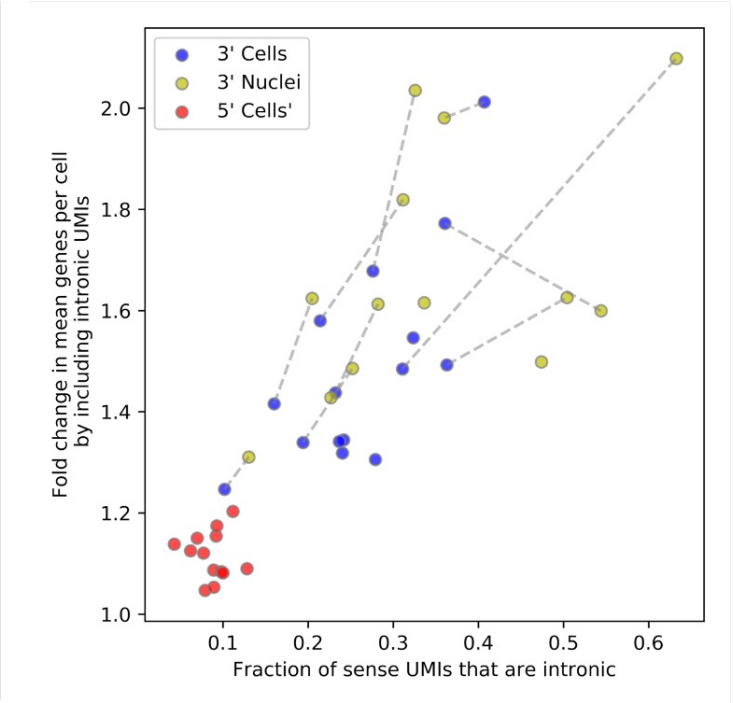
--include-introns …?

Should intronic reads be counted toward UMI counts?

Single nucleus RNAseq:
YES (duh).

Single cell RNAseq:
Debatable.
More UMIs, but less comparable to legacy datasets



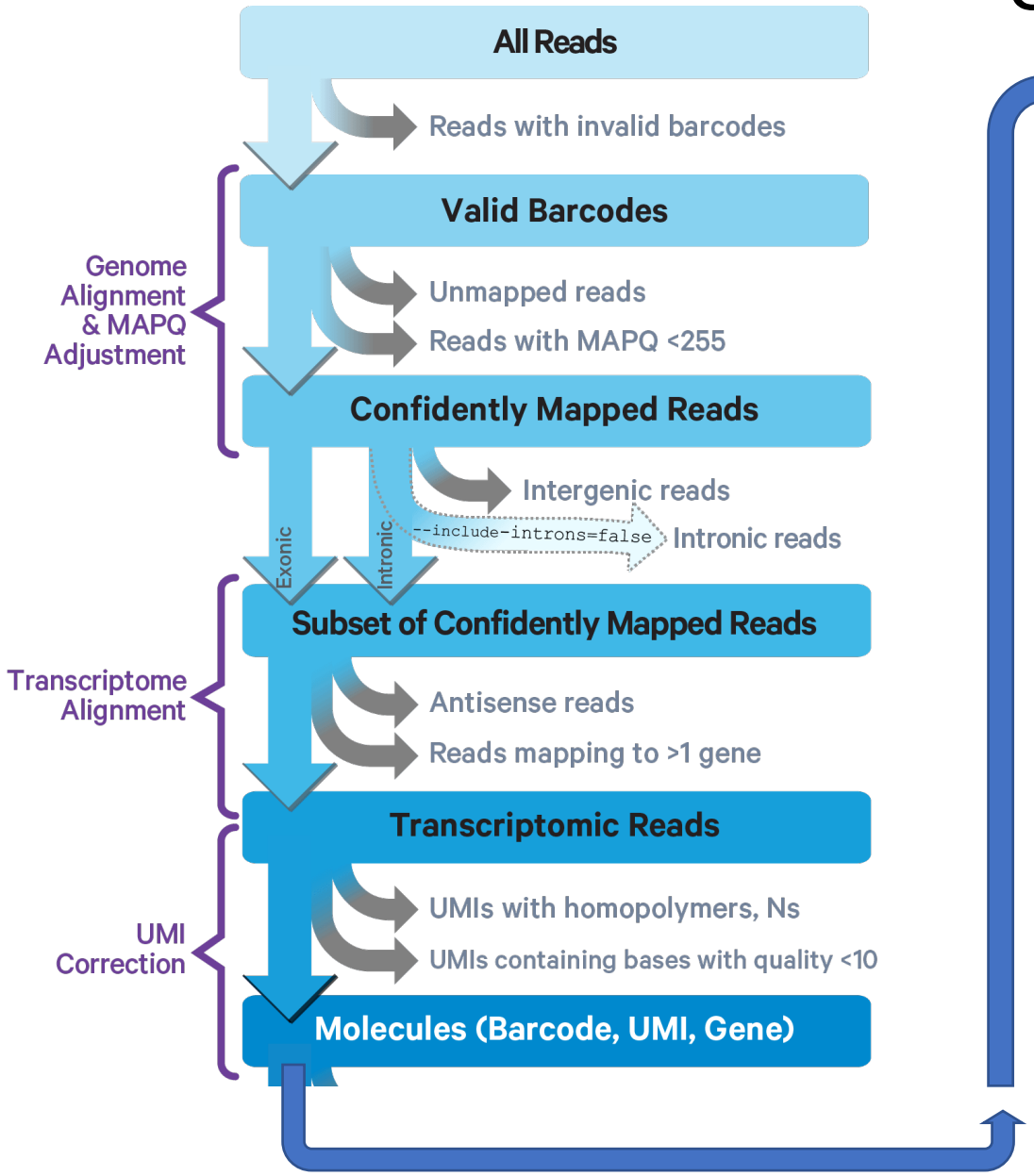


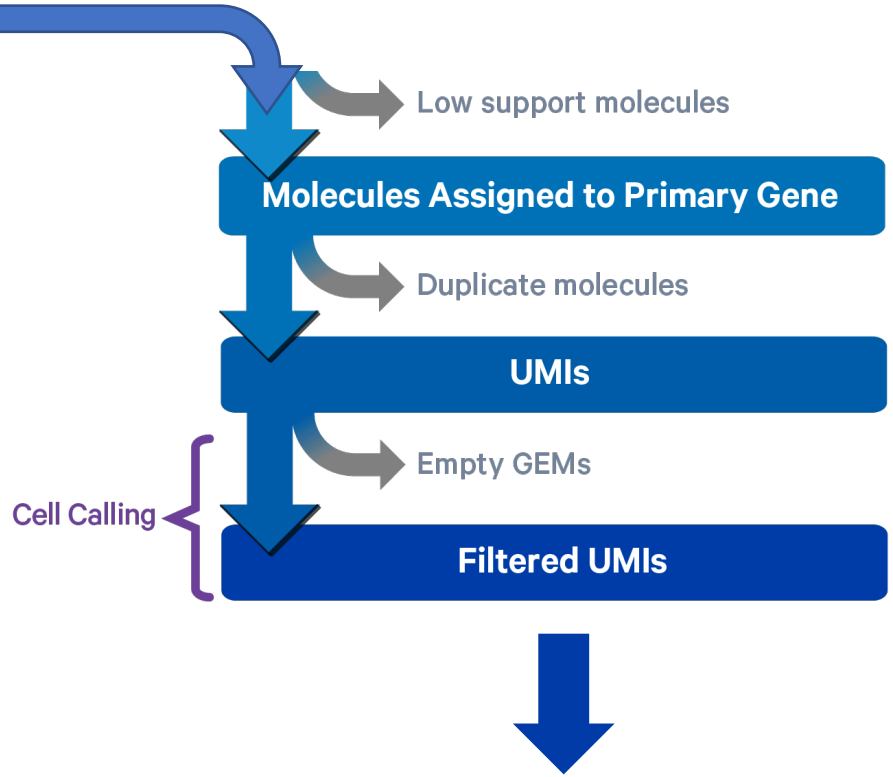
Default Behavior:
Pre-Cell Ranger 7:
`--include-introns=False`

Cell Ranger 7+
`--include-introns=True`

https://kb.10xgenomics.com/hc/en-us/articles/4998628924429-Why-should-I-include-introns-for-my-single-cell-whole-transcriptome-Gene-Expression-data-analysis-

Overview of `cellranger count`





"Secondary Analysis"

Dimensionality reduction, clustering, etc

Cell Ranger Pipeline Output

Highly Processed

Raw FASTQ

Name	Date Modified	Size	Kind
analysis	Feb 22, 2021 at 3:43 PM	--	Folder
cloupe.cloupe	Feb 22, 2021 at 3:46 PM	61.3 MB	Loupe Browser
filtered_feature_bc_matrix	Feb 22, 2021 at 3:38 PM	--	Folder
filtered_feature_bc_matrix.h5	Feb 22, 2021 at 3:37 PM	15.9 MB	HDF Files
metrics_summary.csv	Feb 22, 2021 at 3:45 PM	651 bytes	comma...values
molecule_info.h5	Feb 22, 2021 at 3:39 PM	152.4 MB	HDF Files
possorted_genome_bam.bam	Feb 22, 2021 at 3:35 PM	10.92 GB	Document
possorted_genome_bam.bam.bai	Feb 22, 2021 at 3:36 PM	4.6 MB	Document
raw_feature_bc_matrix	Feb 22, 2021 at 3:28 PM	--	Folder
raw_feature_bc_matrix.h5	Feb 22, 2021 at 3:28 PM	47.8 MB	HDF Files
web_summary.html	Feb 22, 2021 at 3:45 PM	4.2 MB	HTML text

/seq/Illumina_runs/NextSeqData/NextSeqOutput/181221_NB551387_0127_AHHL52BGX9/HHL52BGX9/outs/fastq_path

Note: 10X has a bamtofastq tool that can reconstruct a publishable, lossless FASTQ directly from the mapping output bam file

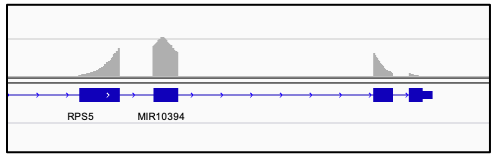
BAM file

You should Archive this!

```
(base) [jpreall@bamdev2 outs]$ samtools view possorted_genome_bam.bam | head -n 1

NB501555:883:HKHCNBGXH:1:13203:17140:14586 16 1 3000079 255 56M * 0 0
AAACCATTTGGTCCCTTTCTTTTTTTTTTTTTTTTTTTTTTTTTGGGTGGGAGAC
EE//A/////////////6EEEEEEEEEEEEEEEEEEEEEEEEEEEEAAAAA NH:i:1 HI:i:1 AS:i:43 nM:i:6
RG:Z:Vakoc_YH02_trachea:0:1:HKHCNBGXH:1 RE:A:I xf:i:0 CR:Z:AGCTTCCTCTTCCCGA
CY:Z:AAAAAEEEEEEEEEE CB:Z:AGCTTCCTCTTCCCGA-1 UR:Z:AGTTATTCCCAA UY:Z:EEEEEEEEEEE
UB:Z:AGTTATTCCCAA
```

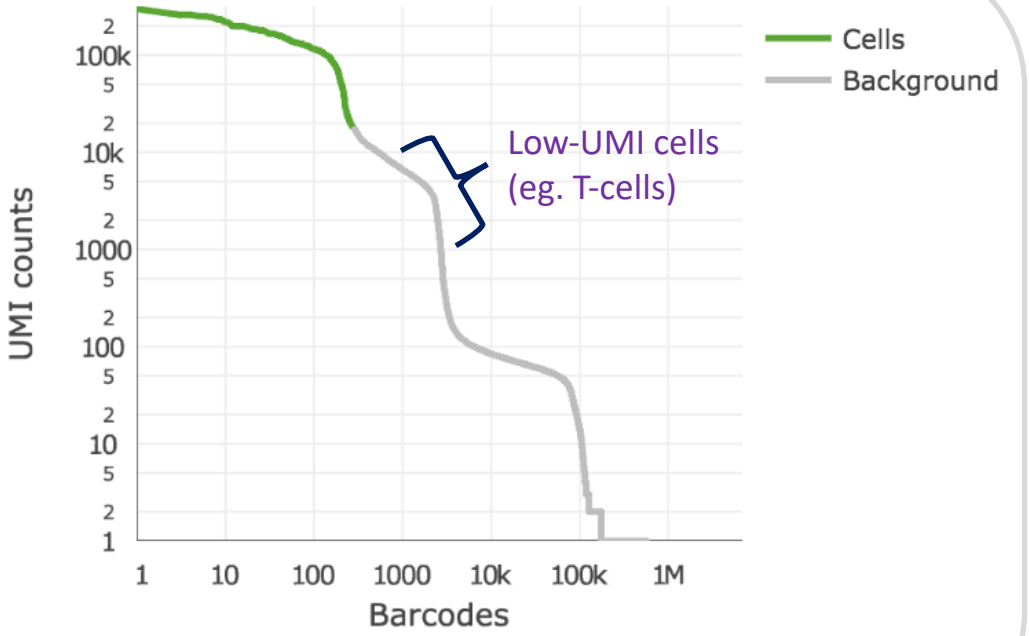
- Stores alignment features, cell barcode, overlapping genes, UMI
- Contains complete record of sequencing data
- FASTQs can be faithfully recreated (eg. for publication) using `bam2fastq`
- Can be viewed as a browser track



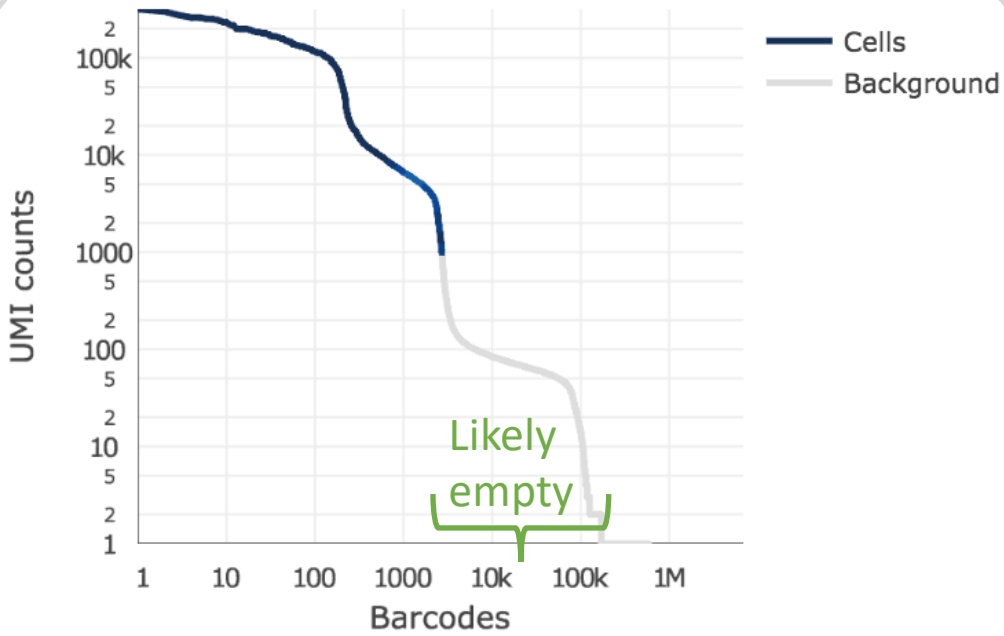
- Can be be used to extract per-cell genotype / allelic expression using `Vartrix`

Cell Ranger: Cell Calling Algorithm

Based on EmptyDrops https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1662-y



Cellranger 2

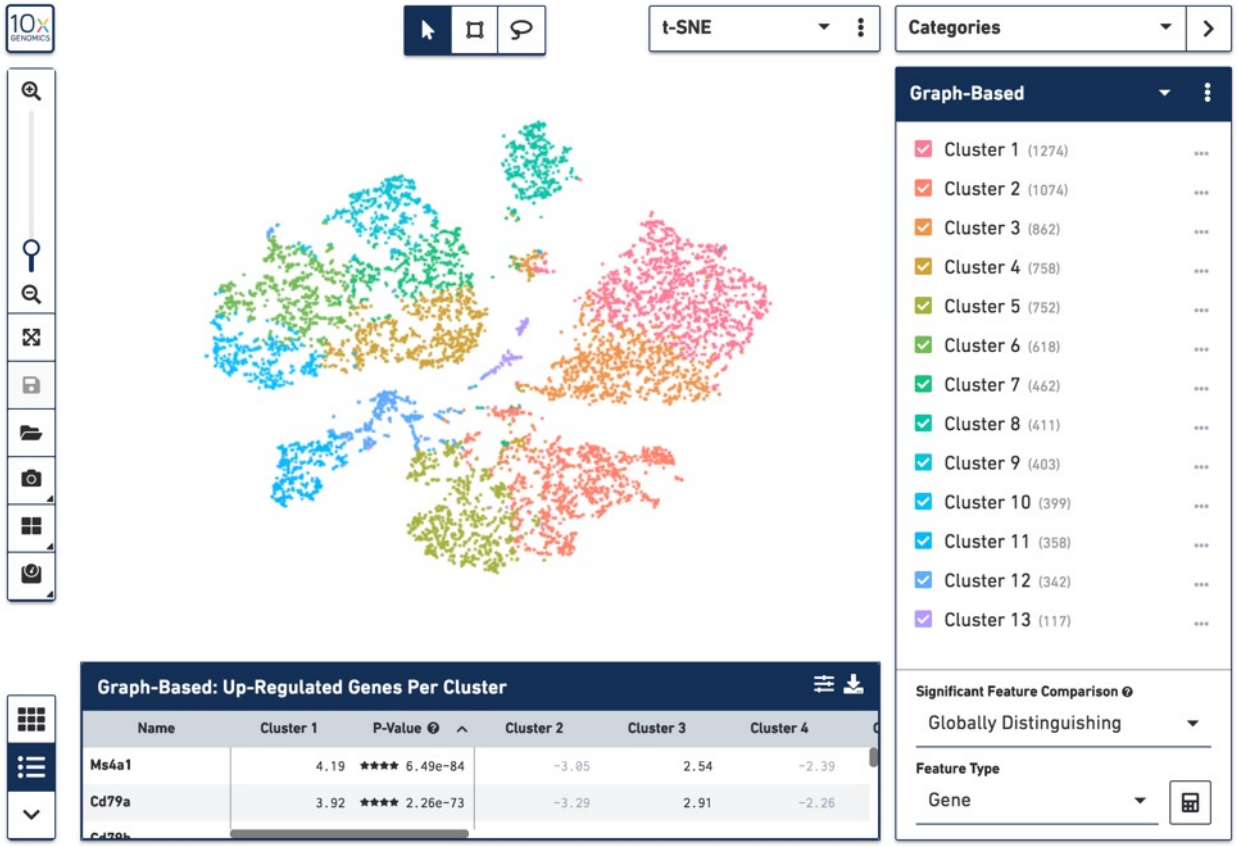


Cellranger 3+

- Inflection point of knee plot
- Missed Low-UMI cell types

- Compares each barcode with likely empty droplets
- Much more permissive – keeps dead cells any anything remotely “cellish”

Loupe Browser (brief live demo)



Tutorial: https://support.10xgenomics.com/single-cell-gene-expression/software/visualization/latest/tutorial

Can:

- Quickly visualize genes
- Do guided clustering via marker genes / tSNE selections
- Calculate Differential Expression
- Export cells and gene sets for reanalysis on Cellranger (cluster)

Can't

- Compute & track composite features (eg. %mito, cell cycle)
- Perform linear regression
- Filter Doublets
- Properly batch correct
- Pseudotime, other fancy things

Quality Control: web_summary **(brief live demo)**

FTPS22_Ctrl

Summary | Analysis

4,399
Estimated Number of Cells

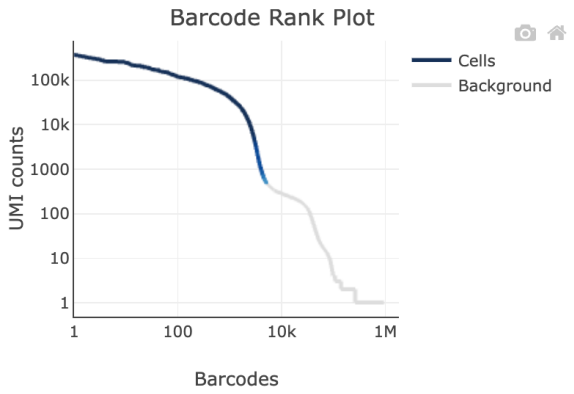
129,481
Mean Reads per Cell

3,643
Median Genes per Cell

Sequencing

Number of Reads	569,587,142
Number of Short Reads Skipped	0
Valid Barcodes	93.9%
Valid UMIs	99.9%
Sequencing Saturation	55.0%
Q30 Bases in Barcode	96.6%
Q30 Bases in RNA Read	96.3%
Q30 Bases in UMI	96.6%

Cells



Estimated Number of Cells	4,399
Fraction Reads in Cells	93.8%
Mean Reads per Cell	129,481
Median Genes per Cell	3,643
Total Genes Detected	28,894
Median UMI Counts per Cell	13,795

Mapping

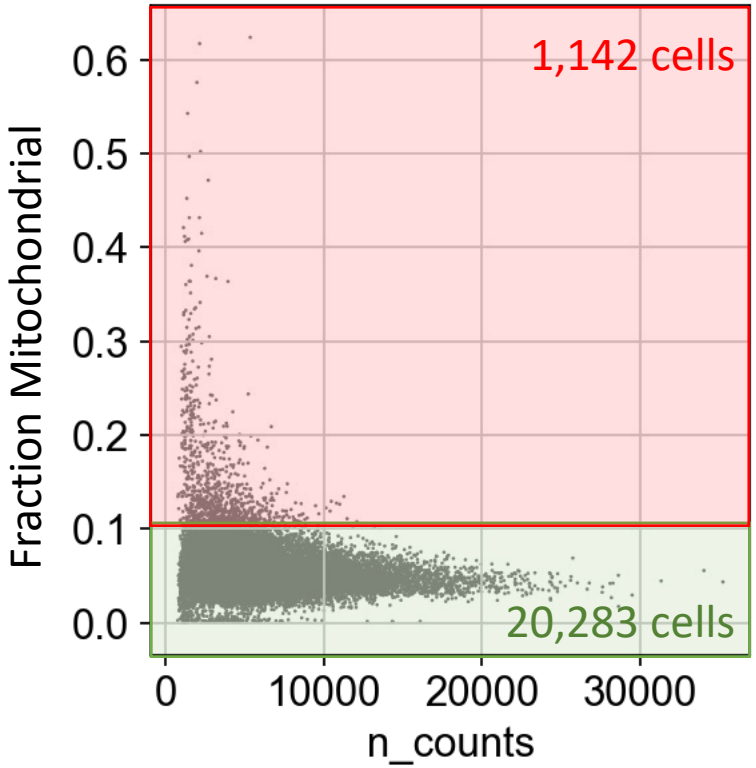
Reads Mapped to Genome	64.5%
Reads Mapped Confidently to Genome	58.7%
Reads Mapped Confidently to Intergenic Regions	4.4%
Reads Mapped Confidently to Intronic Regions	0.6%
Reads Mapped Confidently to Exonic Regions	53.7%
Reads Mapped Confidently to Transcriptome	51.2%
Reads Mapped Antisense to Gene	0.5%

Sample

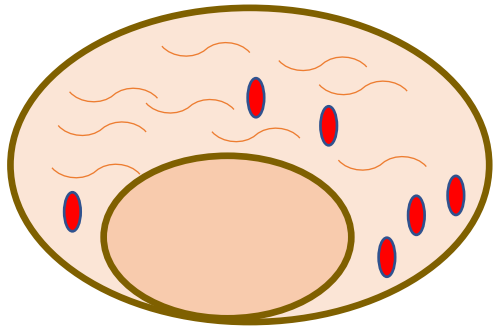
Sample ID	FTPS22_Ctrl
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	False
Reference Path	…anger/references/Zea_Mays_v3_Mar2019
Transcriptome	Zea_Mays_v3_Mar2019-
Pipeline Version	cellranger-6.0.0

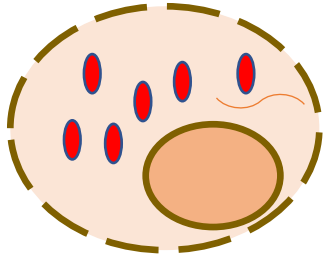
Quality Control: QC metrics – Cell Death

Filter cells by Fraction Mitochondrial Reads



Healthy Cell

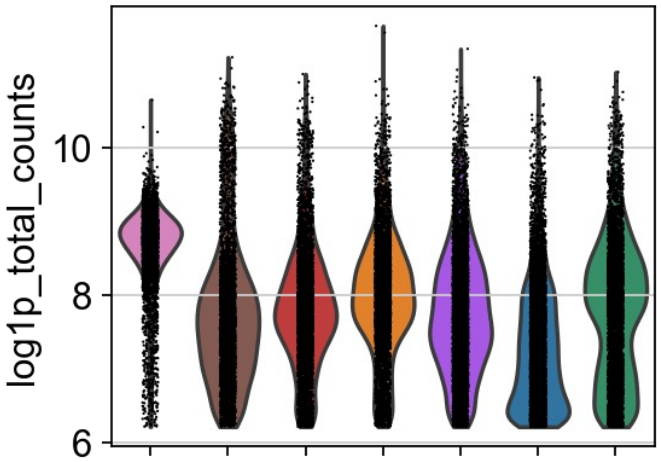




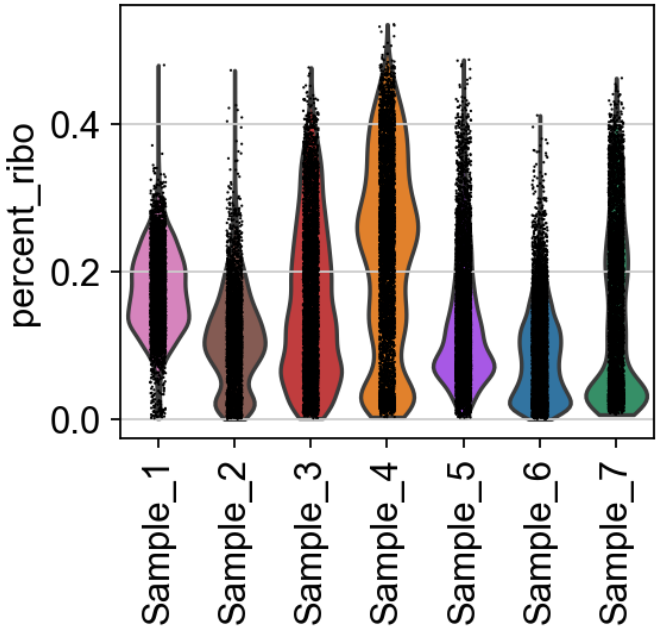
Dying / Ruptured Cell

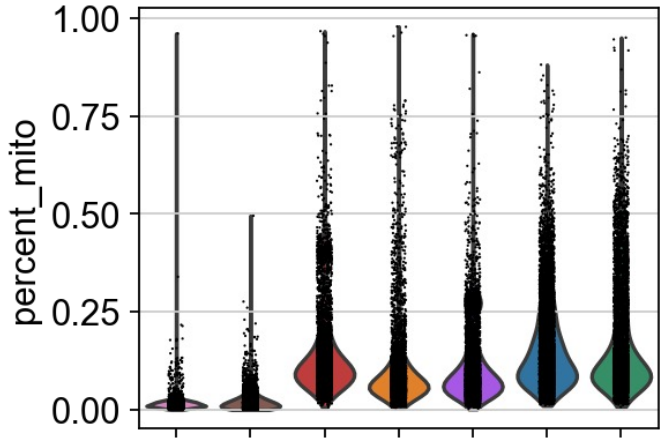
Quality Control: QC metrics -> interpret cautiously

UMI counts:
Cell types
Viability
Sequencing depth
Library handling

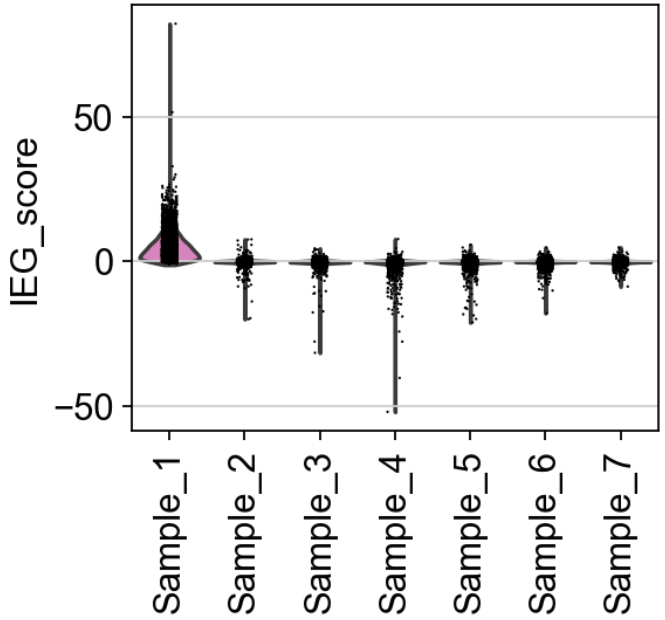


Ribosomal %
Cell types
Activation / Metabolism
Sequencing depth
(normalization artifact)





Mitochondrial %
Dying cells
Respiration activity!
(some cells have > 50% but are alive!)



Immediate-Early Genes
Post-dissoc. stress
In vivo acute activation

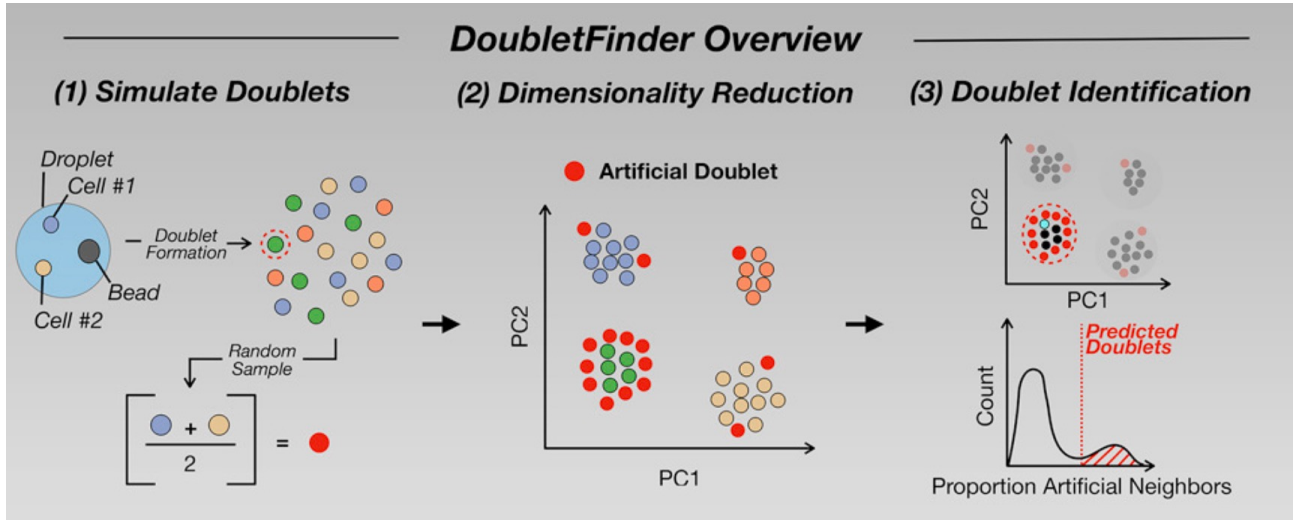
Quality Control: Doublet Filtering

DoubletFinder

•DoubletFinder - [R] - Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. BioRxiv

Doublet Finder – simplified explanation:

- Relies solely on gene expression data
- Begins by simulating artificial doublets and incorporating these "cells" into existing scRNA-seq data processed with "Seurat"
- Then distinguishes real doublets from singlets by identifying real cells with high proportions of artificial neighbors in gene expression space.
- In other words, real cells that have a similar expression patterns to the simulated doublets and cluster together with them



•Other methods:

•DoubletDecon , DoubletDetection , Scrublet

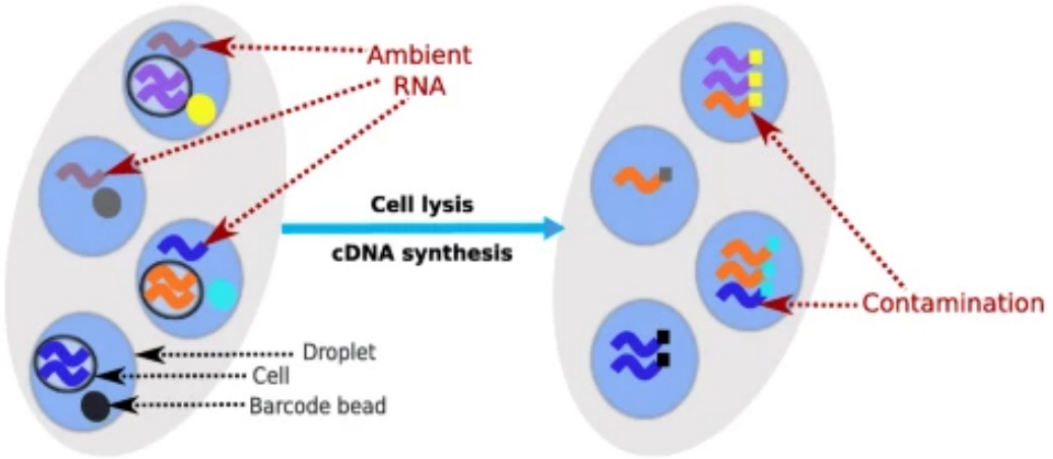
Quality Control: Ambient RNA (soup)

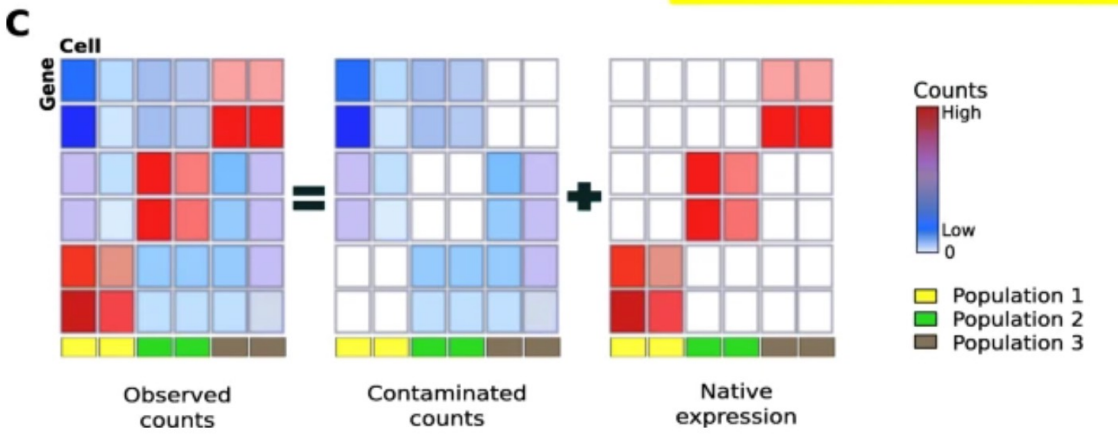
METHOD Open Access

Decontamination of ambient RNA in single-cell RNA-seq with DecontX

Shiyi Yang[1], Sean E. Corbett[1], Yusuke Koga[1], Zhe Wang[1], W Evan Johnson[1], Masanao Yajima[2] and Joshua D. Campbell[1*]

- Models “soup” as a weighted combination of other cell types in the population
- DecontX assumes that each cell is a mixture of two distributions: (1) a distribution of native transcripts from the cell’s true population and (2) a distribution of contaminating transcripts from all other cell populations captured in the assay.
- Produces a matrix of contamination counts and a matrix of “native” counts which can be used in downstream analyses



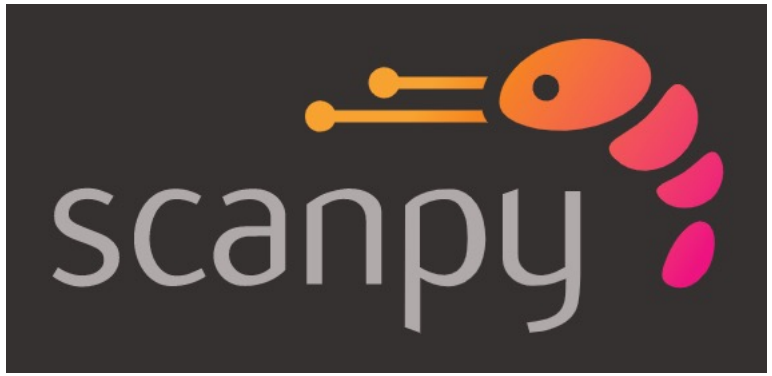


Secondary Analysis

R

Rahul Satija – NYGC / NYU

Python



Fabian Theis - München

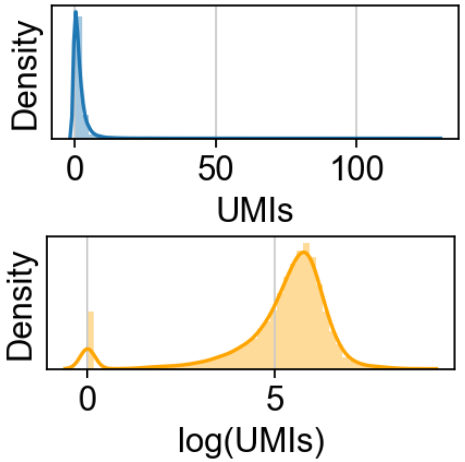
Normalization & log transformation

Basic Method:

$$\frac{\text{Gene(i)Counts}}{\text{All Gene Counts}} * 10{,}000$$ (ie. transcripts per 10k, "tp10k")

Alternatively, though less commonly:
* 1e6 (ie. transcripts per million, "tpm")

Does not account for stochastic variation in droplet performance, batch, other noise OR innate difference in UMI counts between cell types!



Add 1

Logarithmize

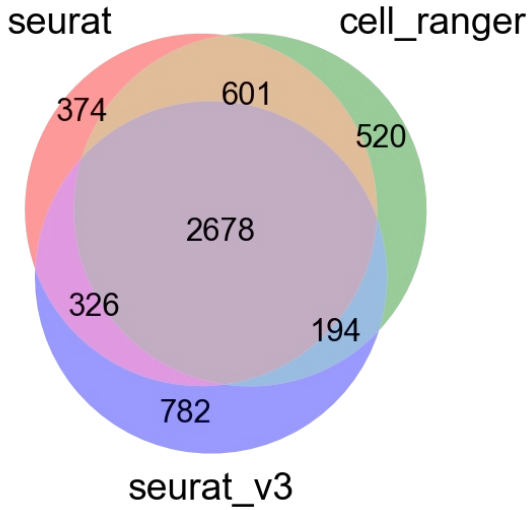
Basic variance stabilizing transformation

Feature Selection: Highly Variable Genes (HVGs)

Assumption: genes that are interesting in the data will have higher variance

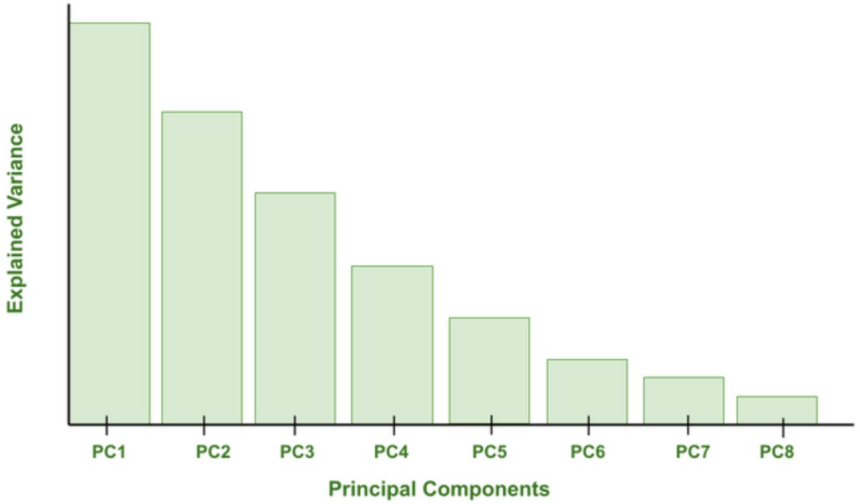
Count-based RNA-seq data has higher variance with higher expression (mean-variance relationship)

Thus, different approaches to select HVGs exist to select relevant HVGs to account for their expression ‘bin’



Principal Component Analysis

- “Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.”
  - https://builtin.com/data-science/step-step-explanation-principal-component-analysis

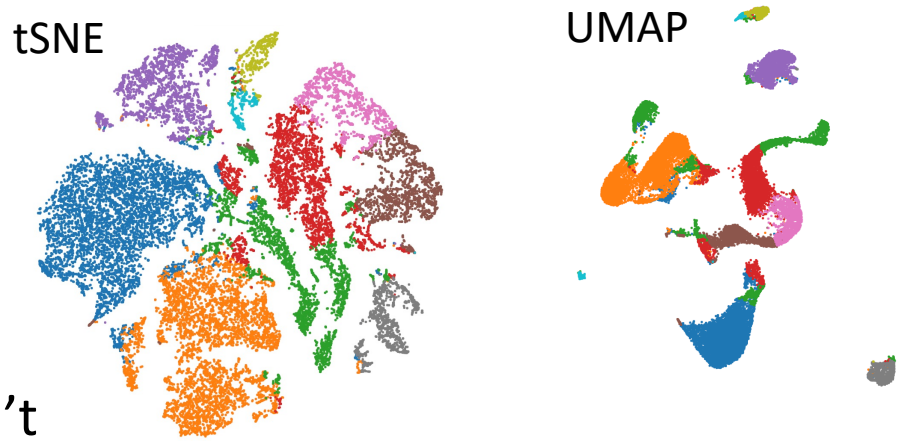


Explanation of theory of PCA:
https://hbctraining.github.io/scRNA-seq_online/lessons/05_theory_of_PCA.html

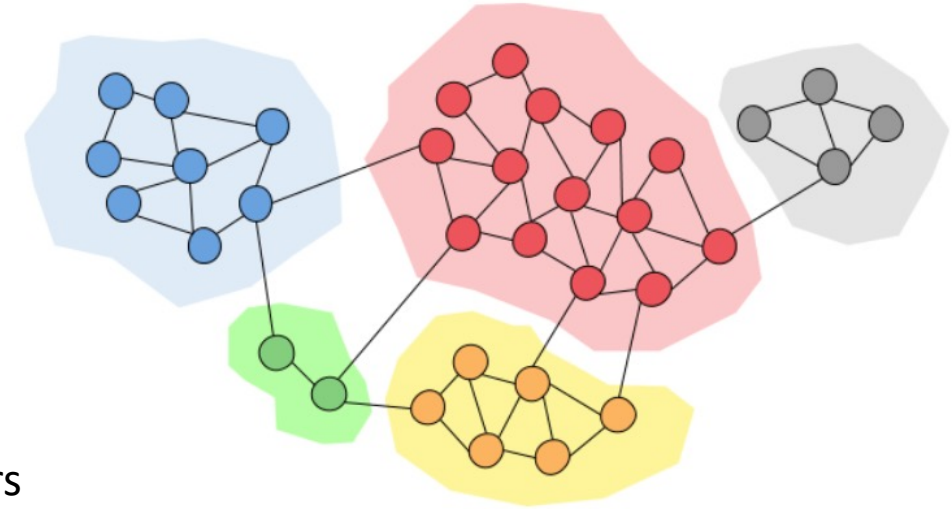
Visualization with tSNE/UMAP

T-distributed stochastic neighbor embedding
Uniform Manifold Approximation

- Non-linear dimensionality reduction best suited to visualization
- PCA space → Neighbor Graph →
- Not a good way to cluster cells
  - But clusters *should* correlate visually
- Very Similar Algorithms
  - tSNE runs a normalization on distance graph in PCA space, UMAP doesn't
  - tSNE favors fine / local structure
  - UMAP "smooths" clusters, favors global structure
  - Speed: UMAP >> tSNE
- Proximity roughly corresponds to similarity



Clustering



Generally, run on latent space (e.g., PC space)

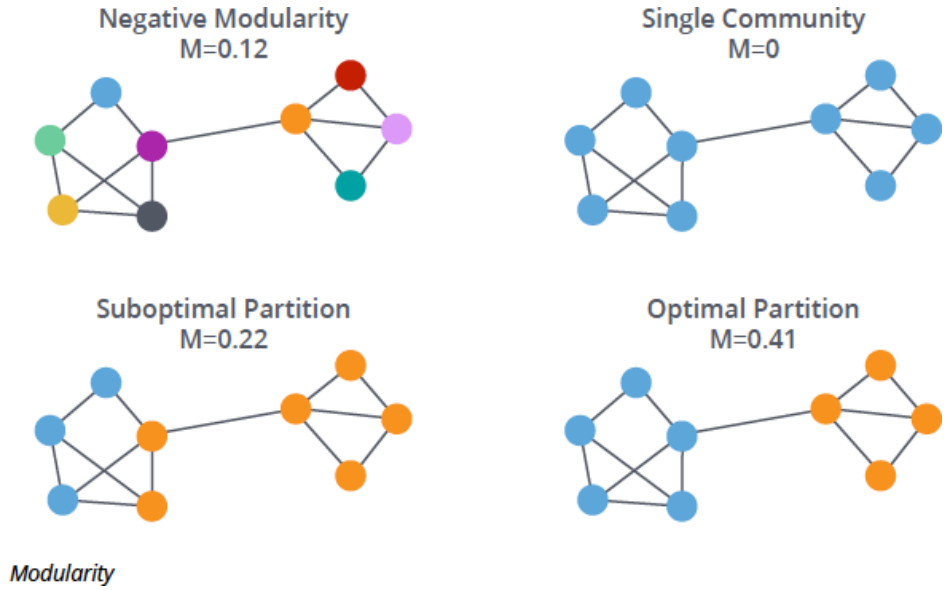
Built into Cell Ranger, most 3rd party tools:

K-means:

- Ask algorithm to break data into ‘K’ optimally distinct clusters
- Usually run iteratively over a range of ‘K’ based on expectations of cell types, state, etc. in data

Graph-Based (e.g., Louvain or Leiden Modularity Optimization)

- Compute all pairwise Euclidean distances in latent space
- Trim graph only keep each cells ‘K’ nearest neighbors* (k
  - “kNN Graph”
- Draw boundaries around “communities” of connected cells to optimize modularity (in-group connections vs out-group connections)



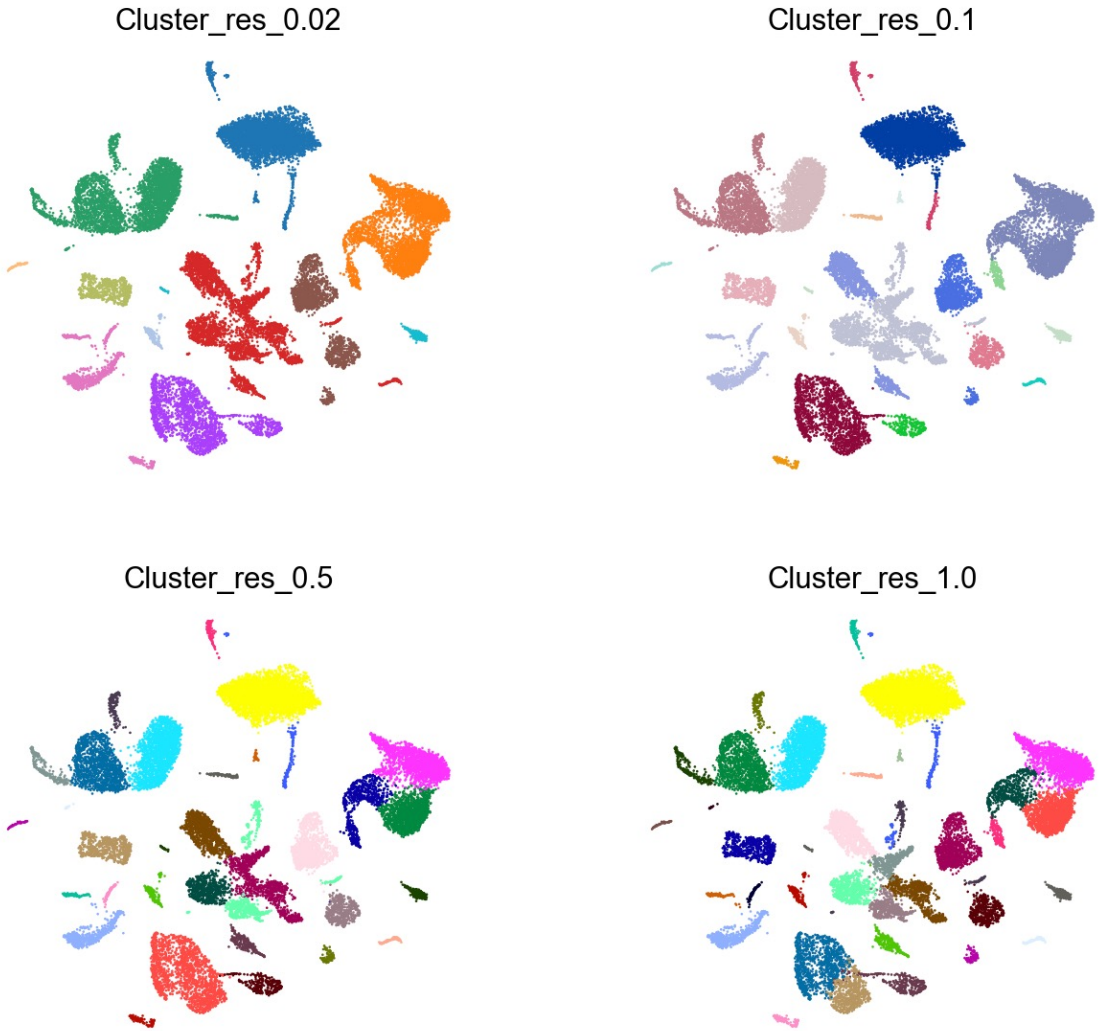
*(Completely unrelated to K in K-means!)

Clustering Resolution

- Resolution parameter tunes clustering sensitivity
- Optimal clustering is subjective
- Guided sub-clustering may provide best results
  - i.e., globally changing resolution parameter may nicely identify meaningful subtypes of one kind of cell, while breaking others into meaningless arbitrary blobs.
- Note: Clustering is probabilistic: reproducibility requires manually setting a random seed!

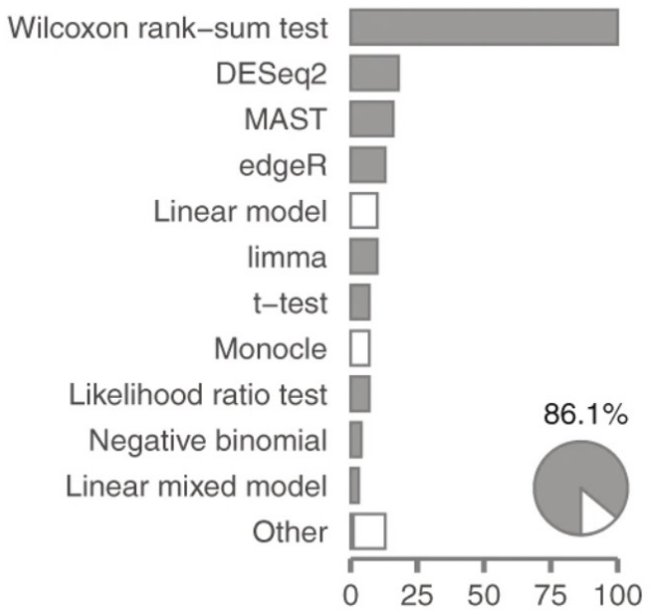
Too Low: miss subtypes, phenotypic states

Too High: over-clustered, no meaningful distinguishing marker genes

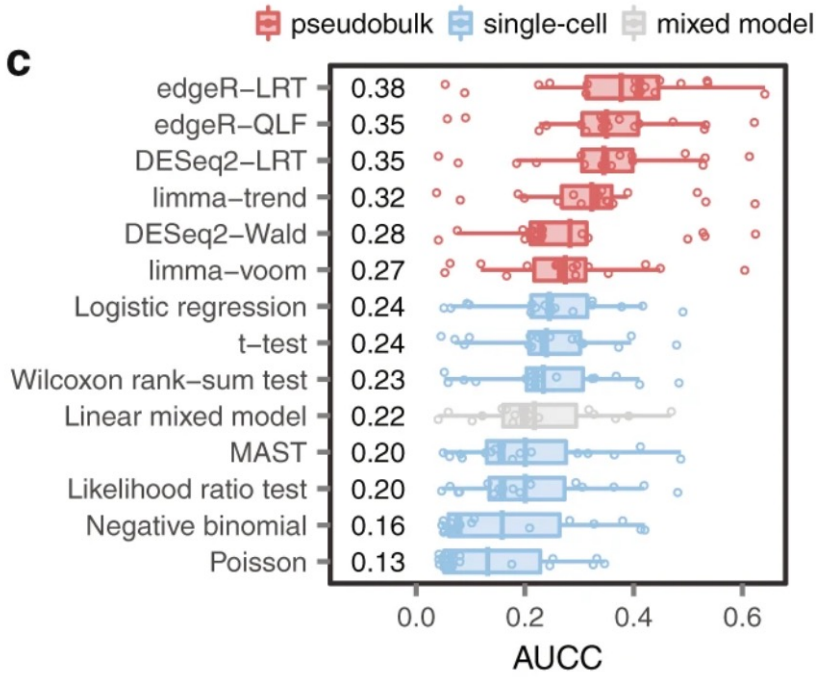


Differential Gene Expression

Popularity



Performance vs 'Ground Truth' datasets



Seurat:

Default: Wilcoxon
Many other tests built in

Scanpy:

Default: t-test
use scDE or other plugins for additional tests

Cell Ranger:

Defaults: Exact Negative Binomial (low cell counts)
EdgeR NB (high cell counts)

Squair, J.W., Gautier, M., Kathe, C. et al. Confronting false discoveries in single-cell differential expression. Nat Commun 12, 5692 (2021). https://doi.org/10.1038/s41467-021-25960-2

- Single-cell methods biased towards highly expressed genes
- Pseudobulk methods make fewer false discoveries of highly expressed genes
  - (i.e. bin clusters or similar cells to reduce zeros and control for high variance in highly expressed genes)

Cell Type Identification

2019

Name	Version	Language	Underlying classifier	Prior knowledge	Rejection option	Reference
Garnett	0.1.4	R	Generalized linear model	Yes	Yes	[14]
Moana	0.1.1	Python	SVM with linear kernel	Yes	No	[15]
DigitalCellSorter	GitHub version: e369a34	Python	Voting based on cell type markers	Yes	No	[16]
SCINA	1.1.0	R	Bimodal distribution fitting for marker genes	Yes	No	[17]
scVI	0.3.0	Python	Neural network	No	No	[18]
Cell-BLAST	0.1.2	Python	Cell-to-cell similarity	No	Yes	[19]
ACTINN	GitHub version: 563bcc1	Python	Neural network	No	No	[20]
LAmbDA	GitHub version: 3891d72	Python	Random forest	No	No	[21]
scmapcluster	1.5.1	R	Nearest median classifier	No	Yes	[22]
scmapcell	1.5.1	R	kNN	No	Yes	[22]
scPred	0.0.0.9000	R	SVM with radial kernel	No	Yes	[23]
CHETAH	0.99.5	R	Correlation to training set	No	Yes	[24]
CaSTLe	GitHub version: 258b278	R	Random forest	No	No	[25]
SingleR	0.2.2	R	Correlation to training set	No	No	[26]
scID	0.0.0.9000	R	LDA	No	Yes	[27]
singleCellNet	0.1.0	R	Random forest	No	No	[28]
LDA	0.19.2	Python	LDA	No	No	[29]
NMC	0.19.2	Python	NMC	No	No	[29]
RF	0.19.2	Python	RF (50 trees)	No	No	[29]
SVM	0.19.2	Python	SVM (linear kernel)	No	No	[29]
$SVM^{rejection}$	0.19.2	Python	SVM (linear kernel)	No	Yes	[29]
kNN	0.19.2	Python	kNN (k = 9)	No	No	[29]

- A hard problem
- What even is a cell type?
- Who is curating these?
- Cell types versus states
  - normal vs disease / perturbed
- What type of model is being used?
  - Marker-based
  - Reference dataset label transfer

scRNA-tools

https://www.scrna-tools.org/tools?sort=name&cats=Classification

MANY more tools available

Abdelaal, T., Michielsen, L., Cats, D. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 20, 194 (2019). https://doi.org/10.1186/s13059-019-1795-z

Celldex - reference index for cell types

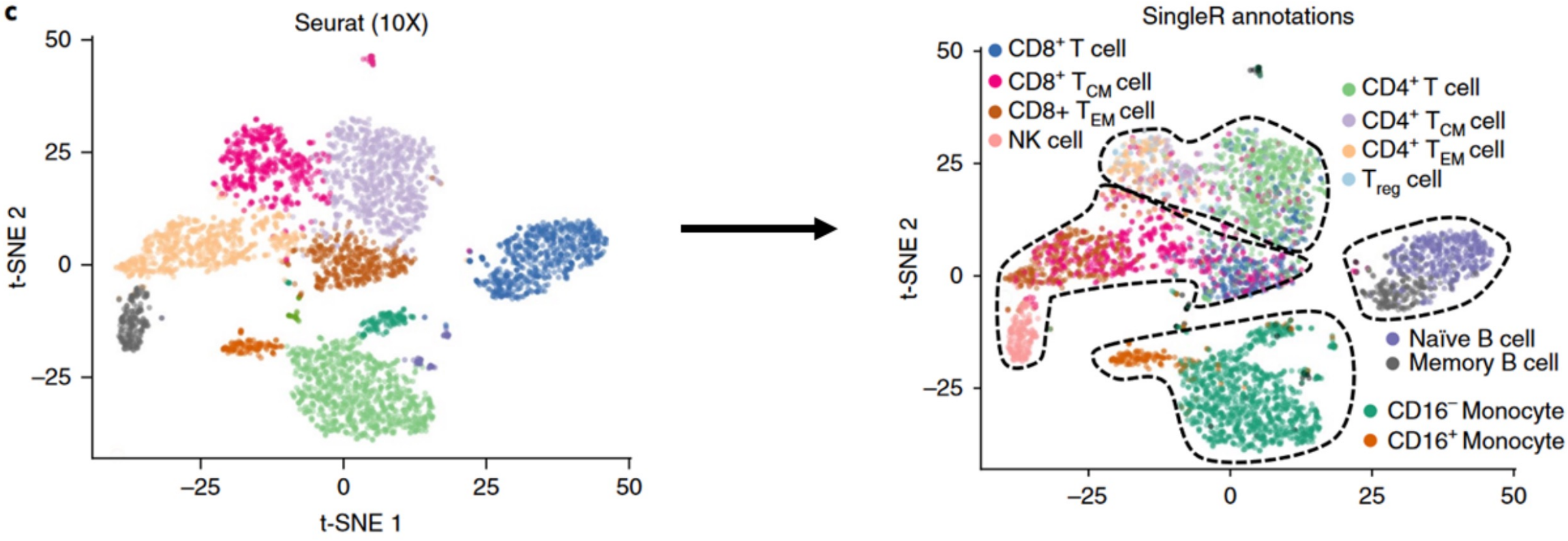
	label.main	label.fine
CD38.negative.naive.B.cell	B-cells	naive B-cells
memory.B.cell	B-cells	Memory B-cells
class.switched.memory.B.cell	B-cells	Class-switched memory B-cells
hematopoietic.stem.cell	HSC	HSC
hematopoietic.multipotent.progenitor.cell	HSC	MPP
common.lymphoid.progenitor	HSC	CLP
granulocyte.monocyte.progenitor.cell	HSC	GMP
macrophage	Macrophages	Macrophages
CD8.positive..alpha.beta.T.cell	CD8+ T-cells	CD8+ T-cells
erythroblast	Erythrocytes	Erythrocytes

```
          B cell1  B cell 2  B cell 3 T cell 1  T cell 2 T cell 3
TSPAN6   0.000000 0.0000000 0.0000000 2.121015 0.0000000 0.000000
TNMD     0.000000 0.0000000 0.0000000 0.000000 0.0000000 0.000000
DPM1     4.242603 3.8718436 4.0188122 5.930974 3.9250500 4.036503
SCYL3    2.594549 1.1176950 2.8599695 0.000000 1.1826923 1.163499
C1orf112 0.000000 0.9107327 0.4436067 1.700440 0.3785116 1.144046
```

https://github.com/LTLA/celldex

- BlueprintEncodeData.R
- DatabaseImmuneExpressionData.R
- HumanPrimaryCellAtlasData.R
- ImmGenData.R
- MonacoImmuneData.R
- MouseRNAseqData.R
- NovershternHematopoieticData.R

singleR - reference-based single-cell RNA-seq annotation



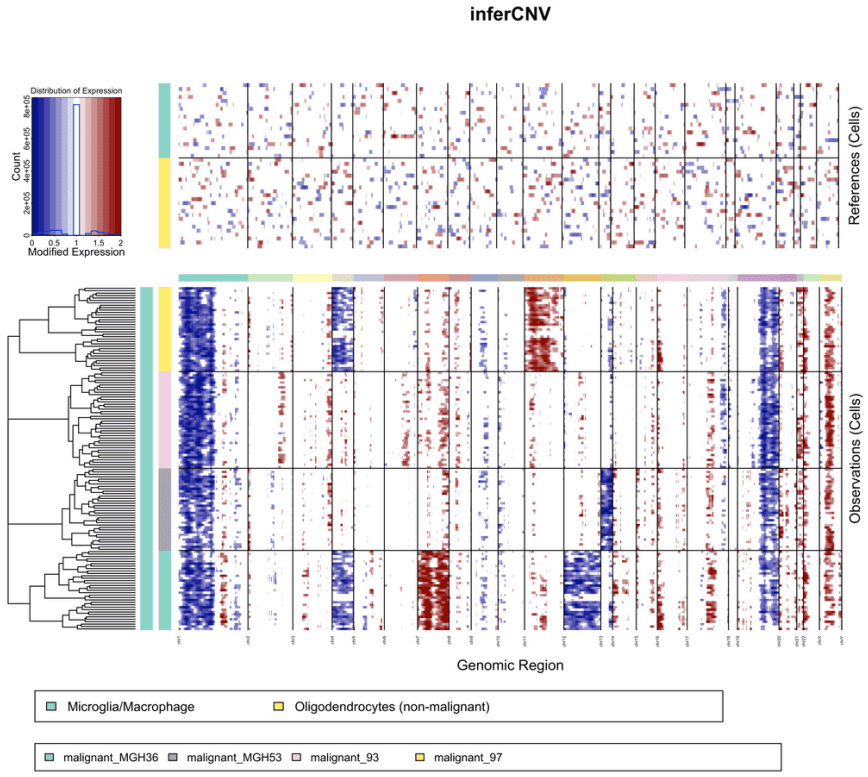
https://biostatsquid.com/singler-tutorial/

CNV inference

Eg. inferCNV, CaSpER, CopyKAT, SCYN, CONICS

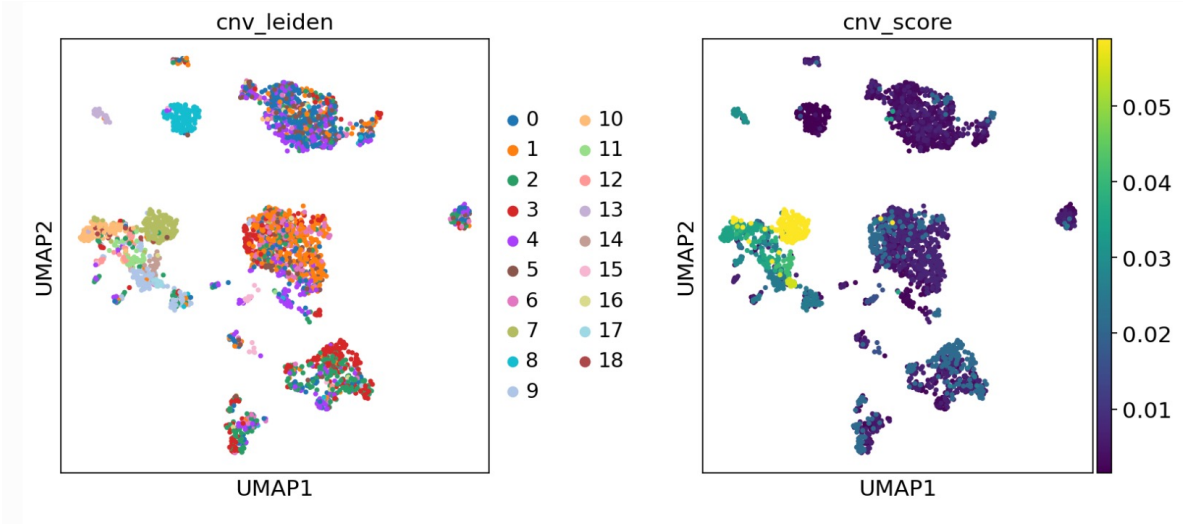
- Counts binned expression data across chromosomes
- Build a background model based on provided “Normal” reference
- Identifies regions with higher-than-expected expression across entire window

- Low resolution (multi-megabase) for scRNAseq
- Can still resolve large-scale clonal copy number loss / gain in chromosome arms, etc.

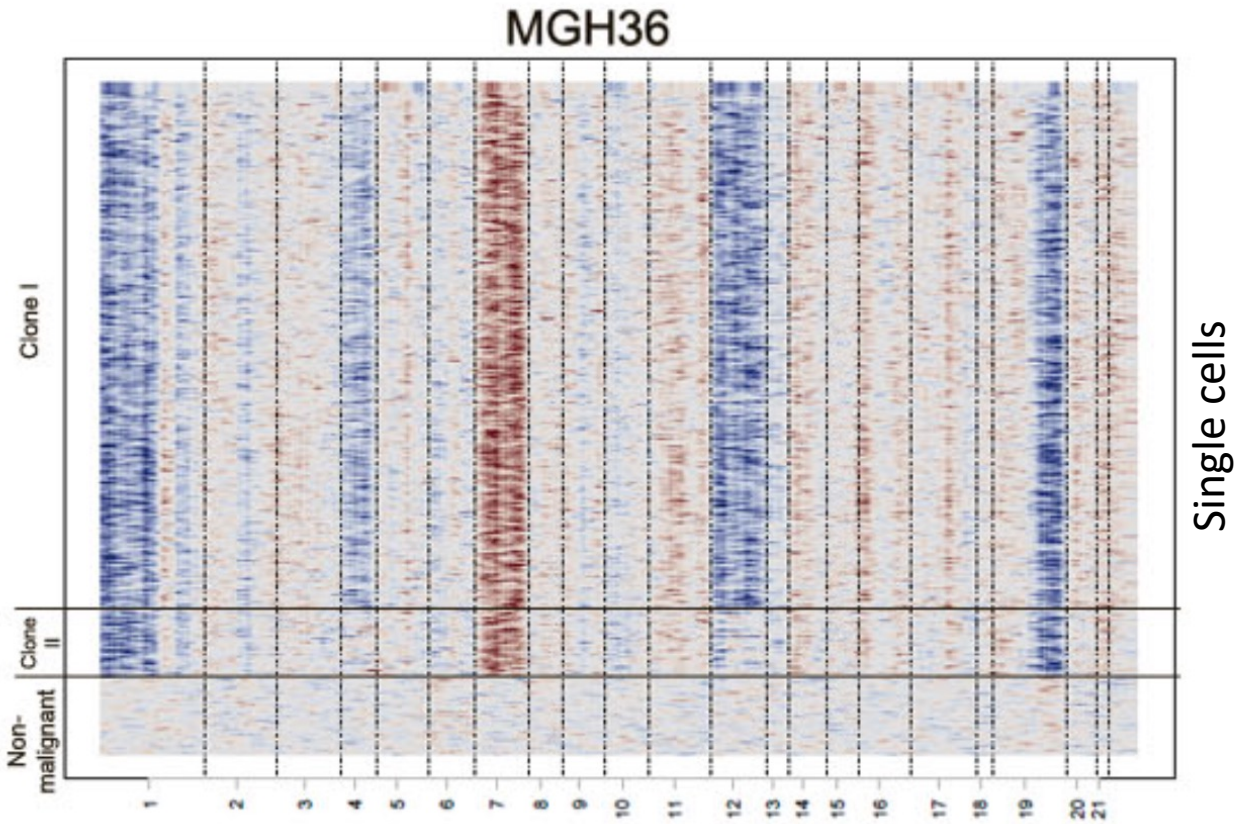


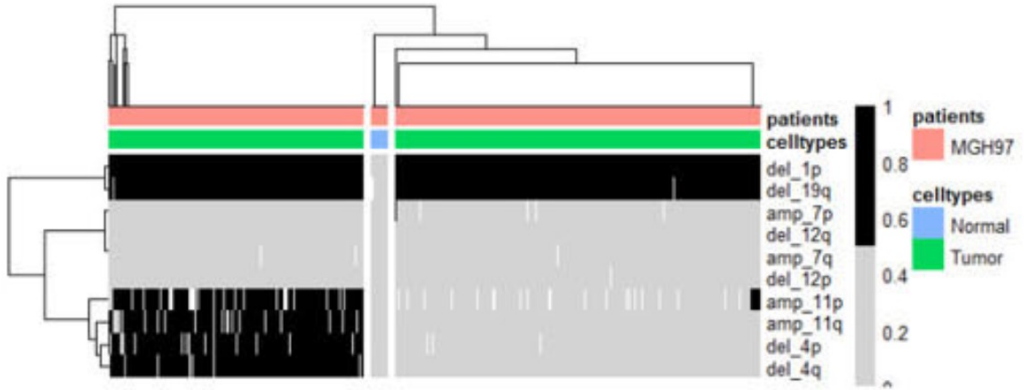
”Normal” Reference cells

Query Cells Eg. Polyclonal Cancer



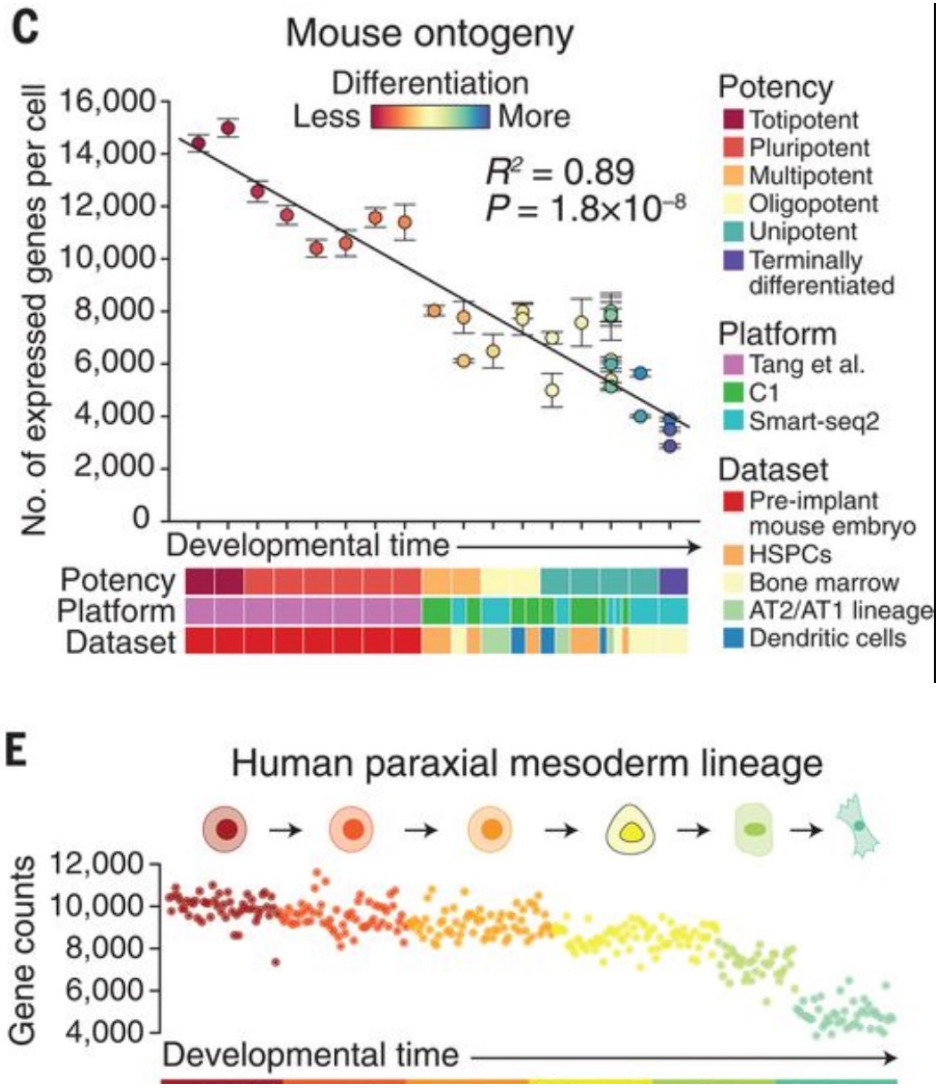
CONICS: COpy-Number analysis In single-Cell RNA-Sequencing



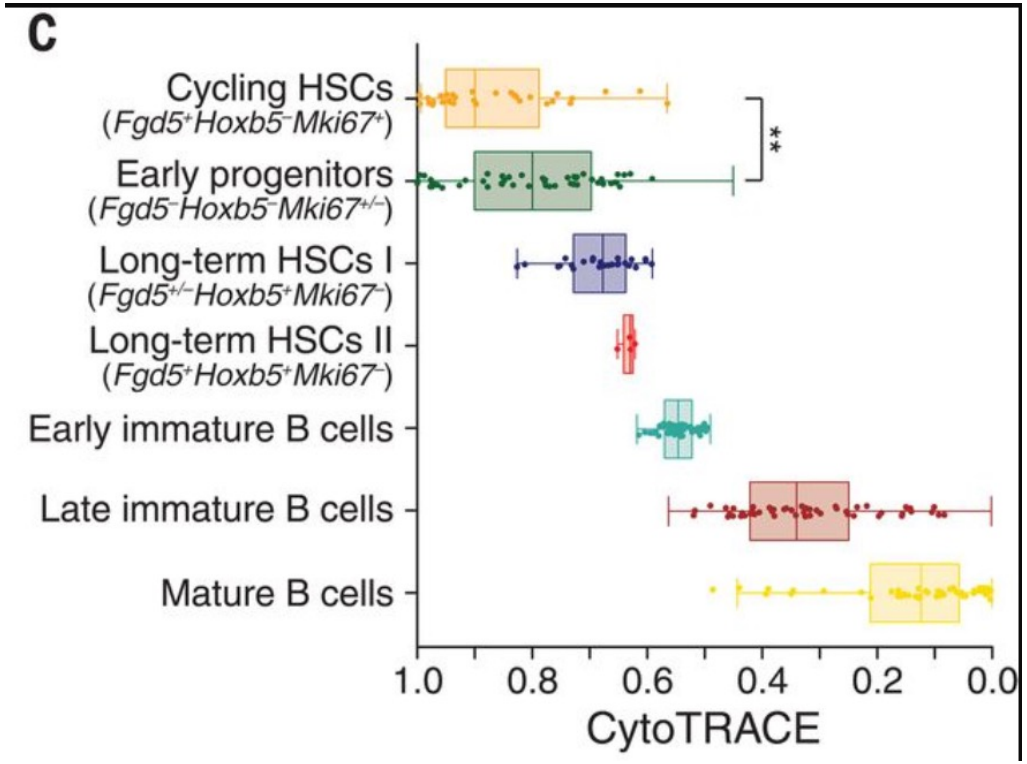


Single cells

Cytotrace - Trajectory Reconstruction Analysis using gene Counts and Expression

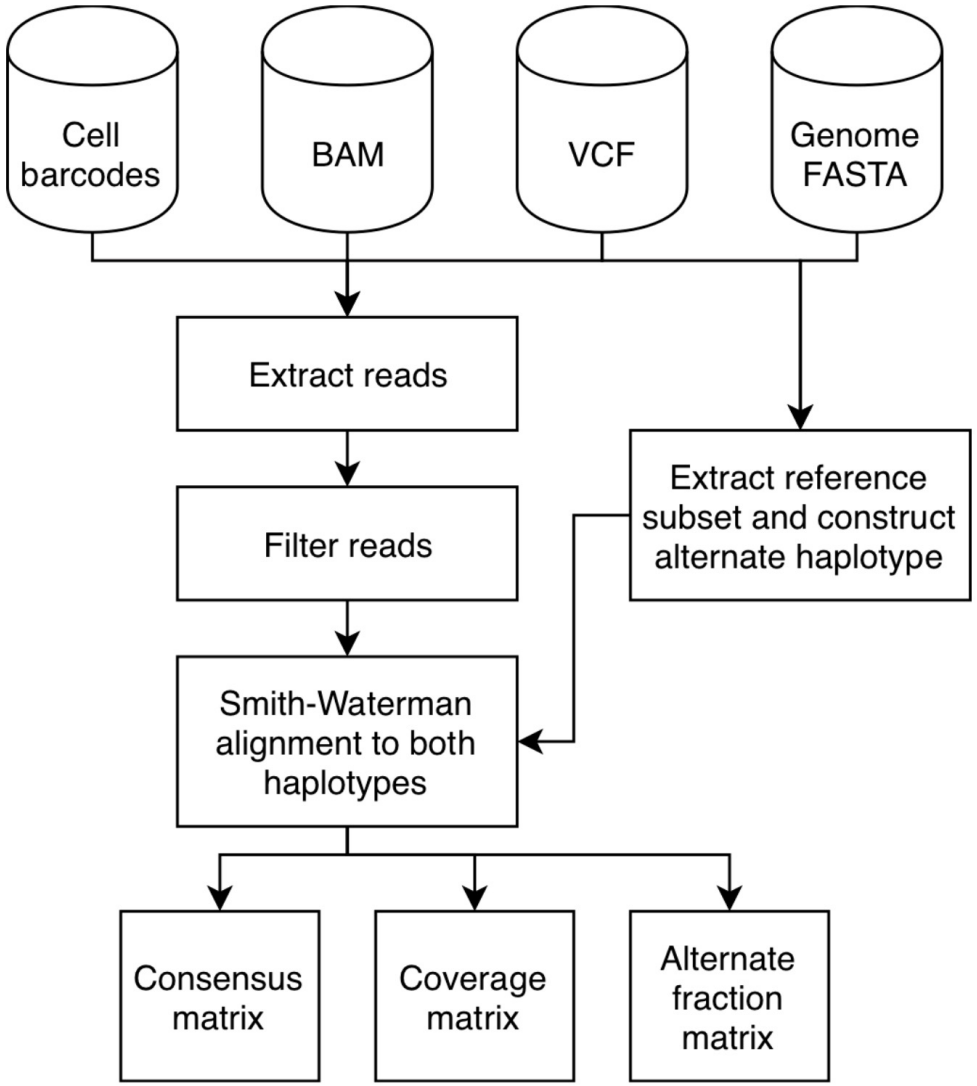


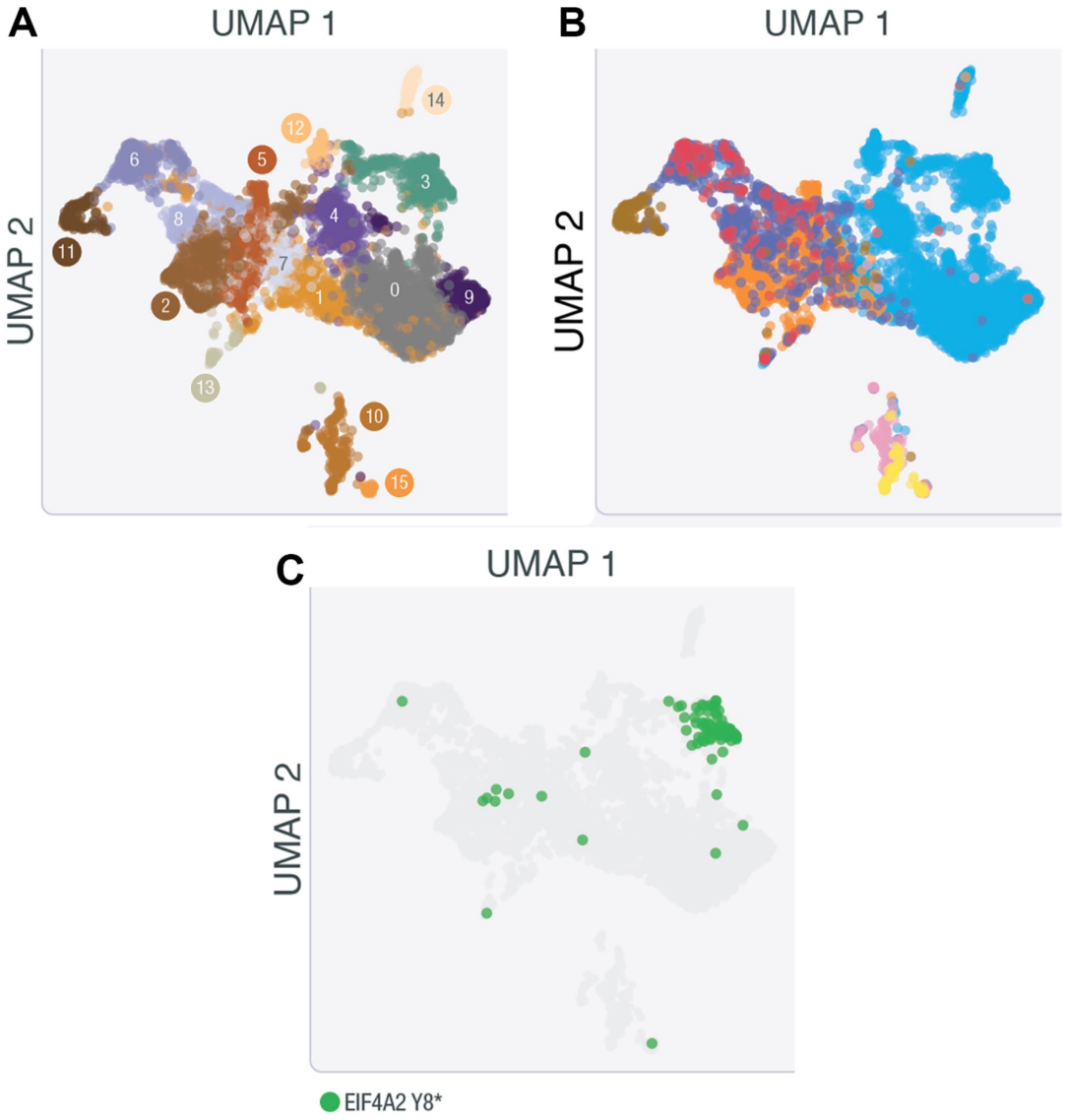
Based on a simple idea that actually seems to work.
As cells become more differentiated, they express fewer genes



https://doi.org/10.1126/science.aax0249

VarTrix – single cell genotyping tool





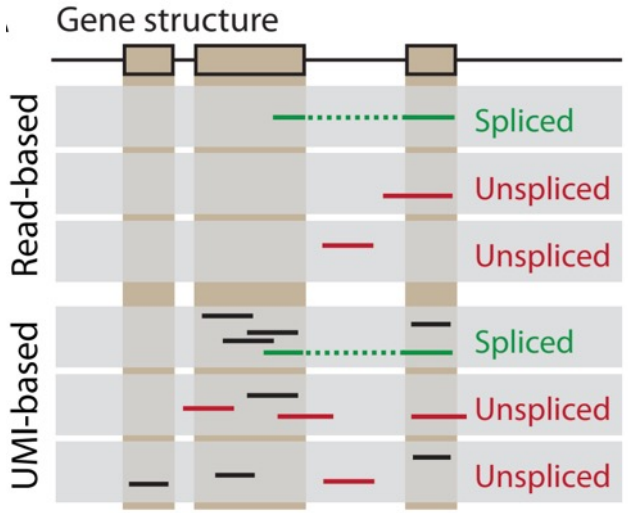
Transcriptional Dynamics: RNA Velocity

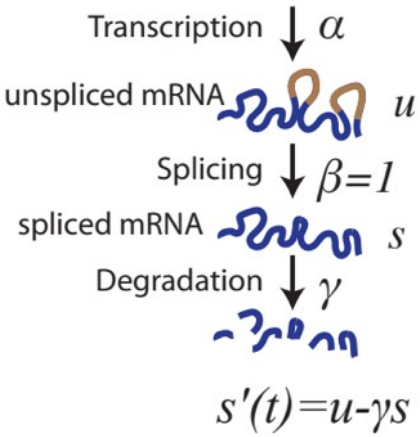
Estimates rates of change in mRNA levels by modeling nascent RNA synthesis

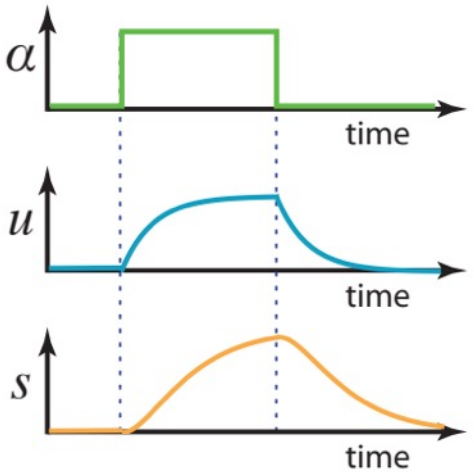
Quantifies spliced / unspliced

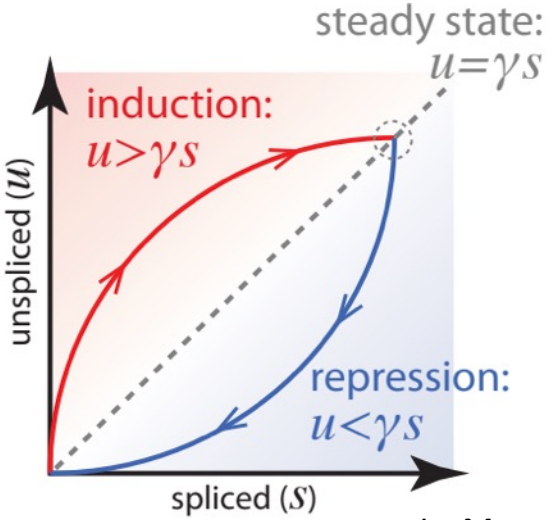
Models dynamics

CAVEATS: Gene annotations
- Cryptic exons
- unannotated intronic genes
- repetitive elements



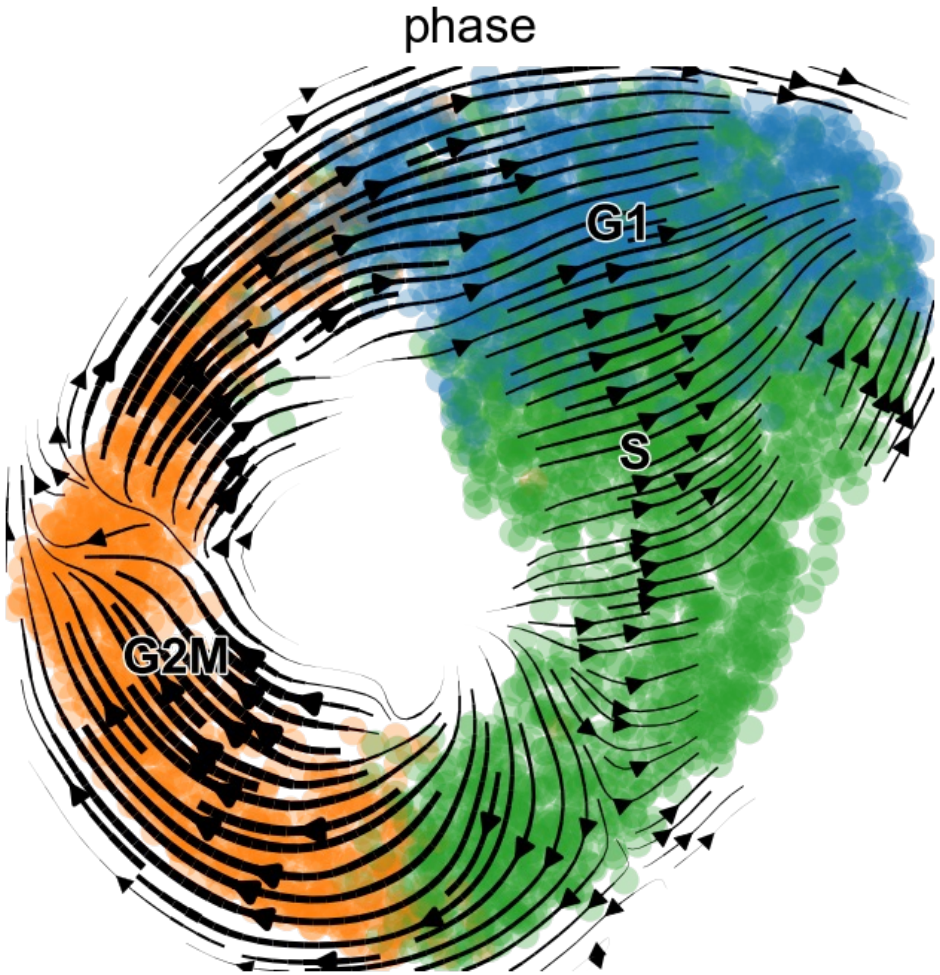






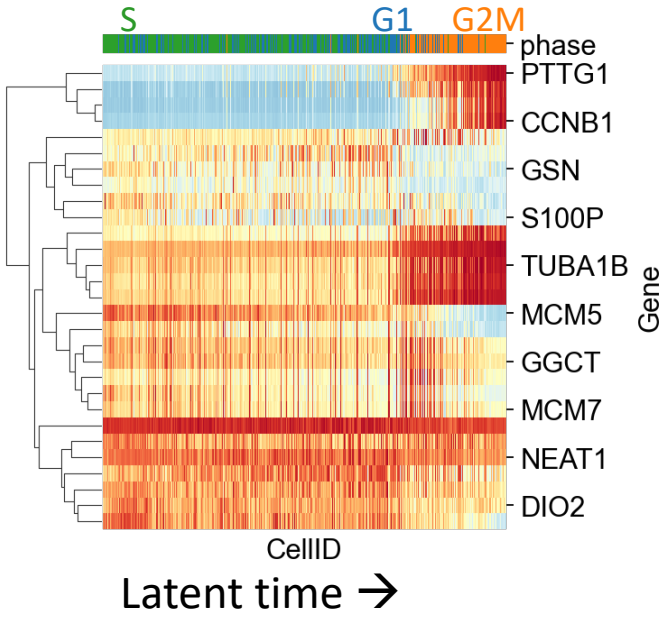
Le Manno et al. (2018) *Nature*

Cultured Cells – RNA Velocity



Siha (cervical epithelium cells)
Projected on to subspace using only cell-cycle relevant genes

RNA Velocity is relevant over very short time scales, when transition states are abundant in the population



So many tools … others not covered yet …

- Monocle
- Enrichr
- scRepertoire
- Cell – cell communication
- Alternative splicing
- Variant discovery
- eQTL analysis

We are on a Coffee Break & Networking Session