

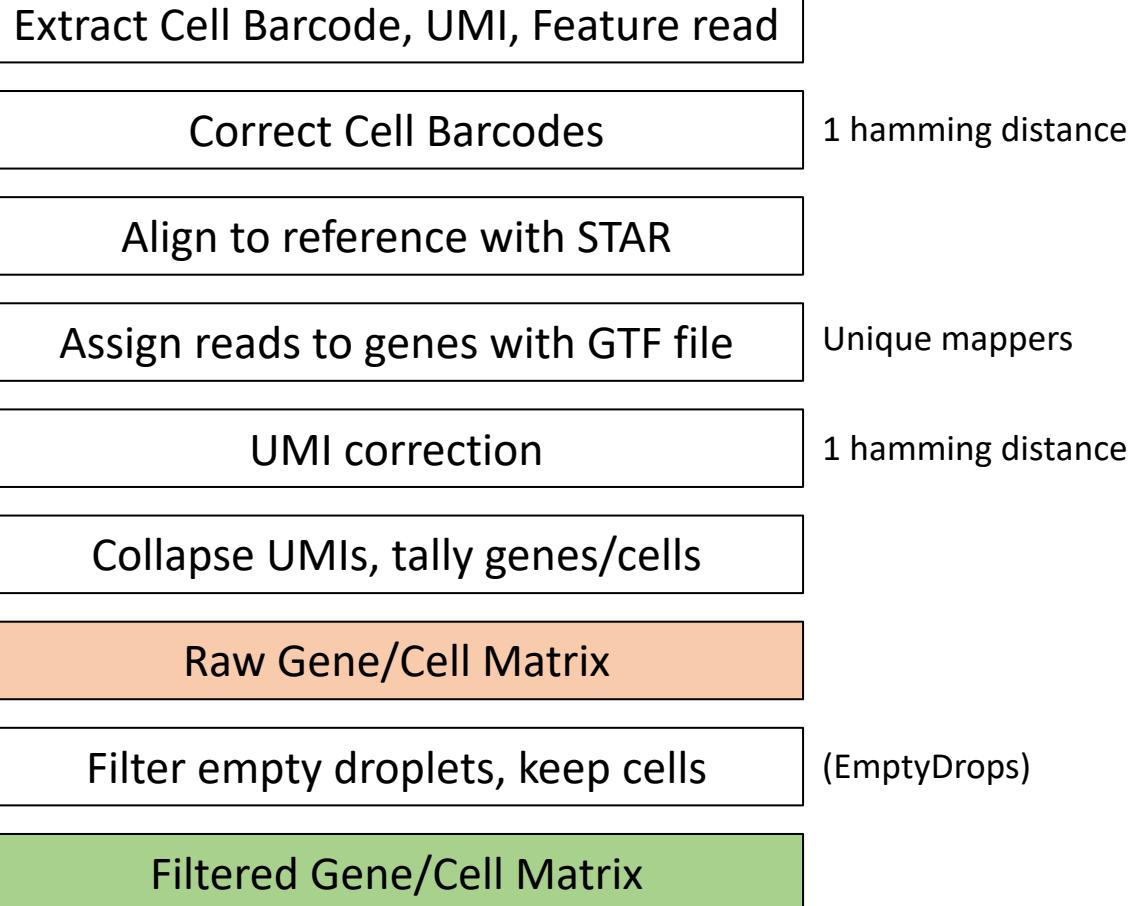
Introduction to scRNAseq Analysis



Jon Preall
Research Associate Professor, CSHL
November 2024

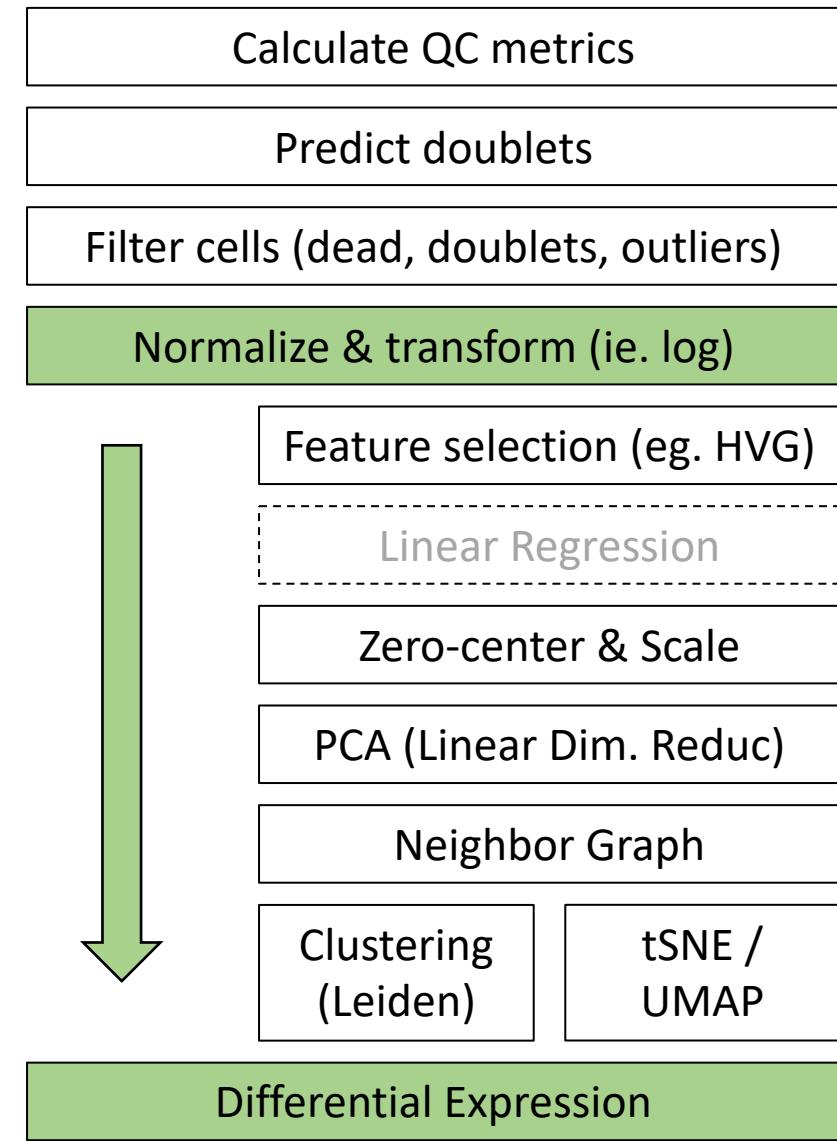
Cellranger / STARSolo / etc

Mapping and Counting



Cell Ranger or Seurat / Scanpy / Liger etc.

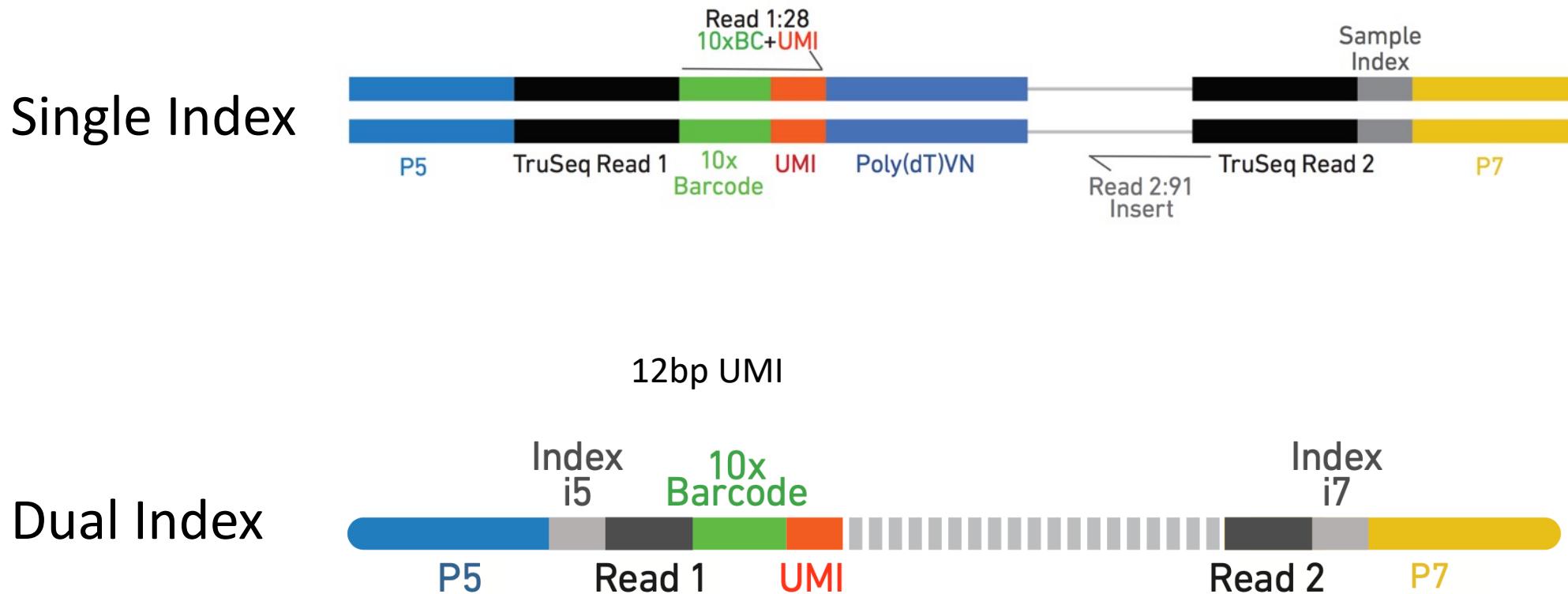
Secondary Analysis



Mapping and Counting

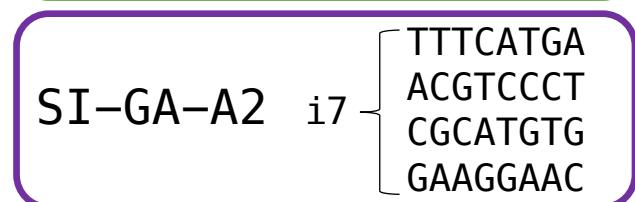
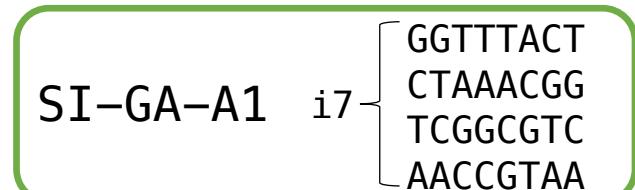
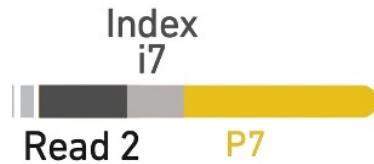
Anatomy of a 10X 3'-Single Cell Amplicon

V3 / NextGem Chemistry



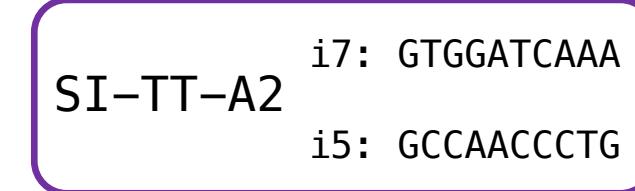
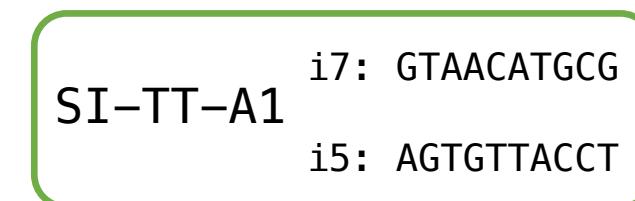
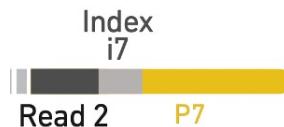
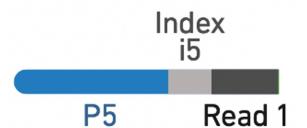
Library Indexing

Single Indexed



- 8bp
- Mix of 4 balanced barcode sequences
- Don't have to worry about how to pool multiple libraries
- Susceptible to INDEX HOPPING

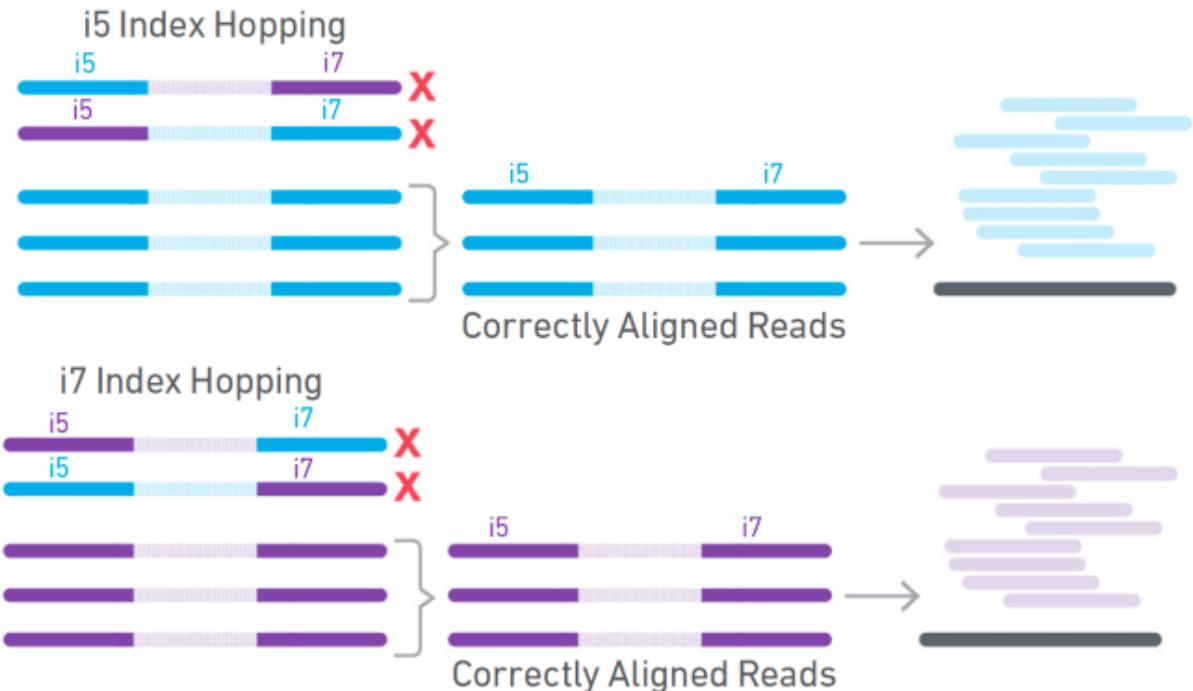
Dual Indexed



- 10bp each (20 cycles total)
- Fixes index hopping
- Pooling with low-plex libraries??

Index Hopping

- Multiple mechanisms can cause chimeric molecules to form during amplification
 - Free adaptors a large culprit
- Mainly a problem on patterned flow-cell instruments using ExAmp technology:
 - NextSeq 2000
 - NovaSeq
 - HiSeq 4000
- < 1% probability of occurring...
- Can have HUGE affect on unique Cell / Barcode / UMI counts!



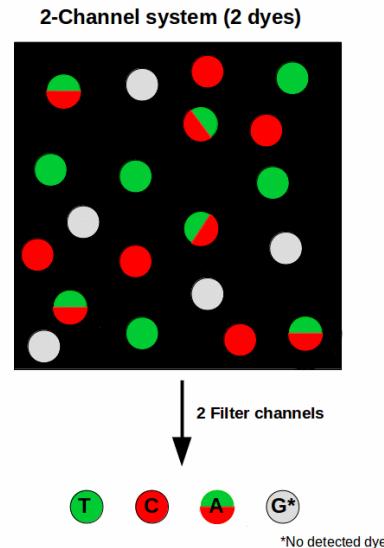
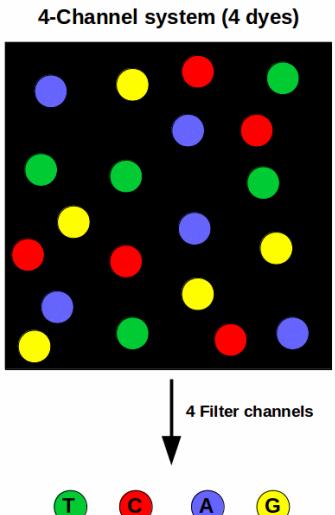
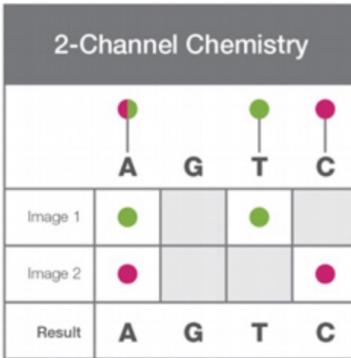
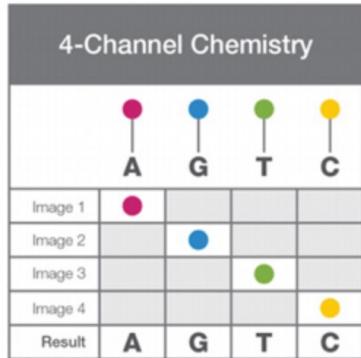
`cellranger mkfastq`



`index-hopping-filter`

Automatically filtered by Cellranger version 4+

Index Diversity



<https://kb.10xgenomics.com/hc/en-us/articles/360045208792-Is-nucleotide-diversity-in-the-index-read-important-for-Illumina-sequencing->

<https://support.illumina.com/bulletins/2016/07/what-is-nucleotide-diversity-and-why-is-it-important.html>

GOOD

GG A C T C C T
T A G G C A T G
T A A G G C G A
C G T A C T A G

✓✓✓✓✓✓✓✓✓✓

BAD

C T C T C T A T
C T C T C T A T
T A T C C T C T
T A T C C T C T

✓✓✓✓XXXX

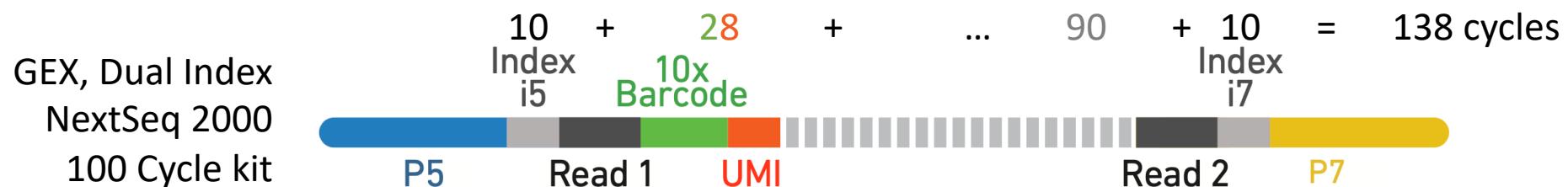
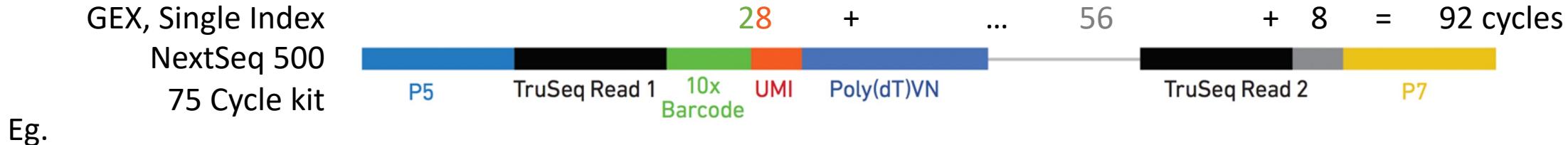
GG
= Blank!
Avoid!

Harder to resolve clusters when ALL spots are the same color

Optimizing your Sequencing run

Instrument	Kit Size	Actual Max. Cycle #	Dark Cycles for Dual Index?	
NextSeq 500/550	High Output or Mid Output	75	92 168	No
		150		
NextSeq 1000/2000	P2 or P3	100	138	No
		200	238	
NovaSeq 6000	v1 (SP-S4)	100	130	Yes – 7 cycles
	v1.5 (SP-S4)	100	138	No

<https://support.illumina.com/bulletins/2016/10/how-many-cycles-of-sbs-chemistry-are-in-my-kit.html>

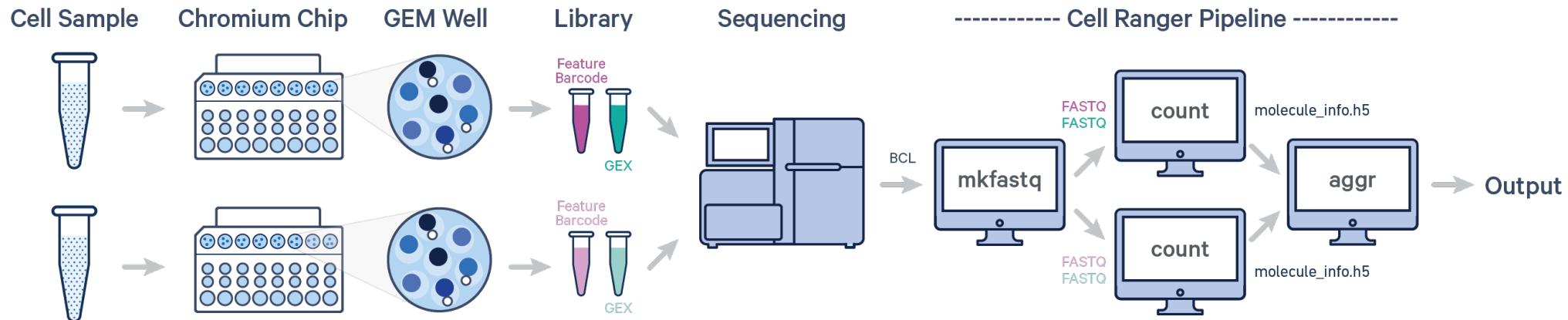


10X Genomics Cell Ranger

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



Sample Indices

Dual_Index_Kit_TT_Set_A.csv

```
# Workflow A = Illumina Forward Strand Sequencing Workflow,,,  
# Workflow B = Illumina Reverse Complement Sequencing Workflow,,,  
# Please contact www.illumina.com if you are unsure which sequencing workflow your  
Illumina instrument & Illumina reagent kit uses,,,
```

```
index_name,index(i7),index2_workflow_a(i5),index2_workflow_b(i5)  
SI-TT-A1,GTAACATGCG,AGTGTACCT,AGGTAACACT  
SI-TT-A2,GTGGATCAAA,GCCAACCCTG,CAGGGTTGGC  
SI-TT-A3,CACTACGAAA,TTAGACTGAT,ATCAGTCTAA  
SI-TT-A4,CTCTAGCGAG,TATCTTCATC,GATGAAGATA
```

A

- NovaSeq 6000 with v1.0 reagent kits
- MiSeq
- MiniSeq with Rapid Reagent kits
- HiSeq 2500, HiSeq 2000

B

- NovaSeq 6000 with v1.5 reagent kits
- iSeq 100
- MiniSeq with Standard reagent kits
- NextSeq Systems
- HiSeq X, HiSeq 4000, HiSeq 3000

... only really necessary if you are not using **cellranger mkfastq** to make FASTQs

Running mkfastq

NOTE: Use 10X's mkfastq. It plays nicely with 10X libraries and downstream pipelines, unlike the default Illumina mkfastq

```
cellranger mkfastq \  
    --localcores=12 \  
    --run=/path/to/basecalls/ \  
    --samplesheet=/path/to/SampleSheet.csv \  
    
```

[Header],,,
IEMFileVersion,4,
Date,5/25/18,
Workflow,GenerateFASTQ,
Application,NextSeqFASTQOnly,
Assay,TruSeqHT,
Description,,
Chemistry,Amplicon,
[Reads],,,
28,,
90,,
[Settings],,,
[Data],,,:
Sample_ID,Sample_Name,Sample_Plate,Sample_Well,I7_Index_ID,index,I5_Index_ID,index2,Sample_Project,Description
PlantCourse2022_10xGEX_control,PlantCourse2022_10xGEX_control,,,SI-TT-A9,SI-TT-A9,SI-TT-A9,SI-TT-A9,PlantCourse2022,,
PlantCourse2022_10xGEX_treatment,PlantCourse2022_10xGEX_treatment,,,SI-TT-A10,SI-TT-A10,SI-TT-A10,SI-TT-A10,PlantCourse2022,

SampleSheet.csv

What do the FASTQ files look like?

i7 index

```
>$ zcat PlantCourse2022_10xGEX_control_S1_L001_I1_001.fastq.gz  
| head -n 4  
@VH00553:6:AAALMHYHV:1:1101:18383:1000  
1:N:0:AAGTGGAGAG+GTAACAGGAA  
AAGTGGAGAG  
+  
CCCCCCCCCC
```

i5 index

```
>zcat PlantCourse2022_10xGEX_control_S1_L001_I2_001.fastq.gz |  
head -n 4  
@VH00553:6:AAALMHYHV:1:1101:18383:1000  
2:N:0:AAGTGGAGAG+GTAACAGGAA  
GTAACAGGAA  
+  
CCCCCCCCCC
```

index_name,	index(i7),	index2_workflow_a(i5),	index2_workflow_b(i5)
SI-TT-A9,	AAGTGGAGAG,	TTCCTGTTAC,	GTAACAGGAA

This was sequenced on a NextSeq, so you can see the i5 read matches workflow b

What do the FASTQ files look like?

Read1

```
>zcat PlantCourse2022_10xGEX_control_S1_L001_R1_001.fastq.gz | head -n 4
@VH00553:6:AAALMHYHV:1:1101:18383:1000 1:N:0:AAGTGGAGAG+GTAACAGGAA
NGCGTATAGGCTGGATGAAGTTAGTCGG
+
#CCC;CCCCCCCCCCCCCCCCCCCC-CC
```

Read2

```
>zcat PlantCourse2022_10xGEX_control_S1_L001_R2_001.fastq.gz | head -n 4
@VH00553:6:AAALMHYHV:1:1101:18383:1000 2:N:0:AAGTGGAGAG+GTAACAGGAA
AAGCAGTGGTATCACGCAGAGTACATGCCAAGTACTACCTGGACGACACGGTGGACGTGGCAAGATGCTGGACGGCCTGGCAGCGC
+
CCCCCCCCCCCC-CCCC;CCCCCCC-C;CCC-CCCCCCCCCCCCCCCC-CCCCCCCCCCCCCCCC-CCCCCCCCCCCC;CCCCCCCC-
```

Note that the i7 and i5 indices of the associated read are written into the @NAME lines here. This means you don't really need to worry about keeping, sharing, or depositing the i5 and i7 FASTQ files, as they are basically just "TMI"

Running `cellranger count`

```
SAMPLE= SeqTech22_RNA_10k
TRANSCRIPTOME=/path/to/CellRanger/references/refdata-gex-GRCh38-2020-A
FASTQPATH=/path/to/folder/containing/your/fastqs/

cellranger count \
--id=$SAMPLE \
--jobmode=local \
--localcores=12 \
--transcriptome=$TRANSCRIPTOME \
--fastqs=$FASTQPATH \
--sample=$SAMPLE \
--include-introns=false
```

Note: If you sequenced your library more than once on different flow cells, Cellranger will accept a comma-separated list of paths to each FASTQ folder

```
FASTQPATH=/path/to/fastq/folder1/,/path/to/fastq/folder2/
```

--include-introns ...?

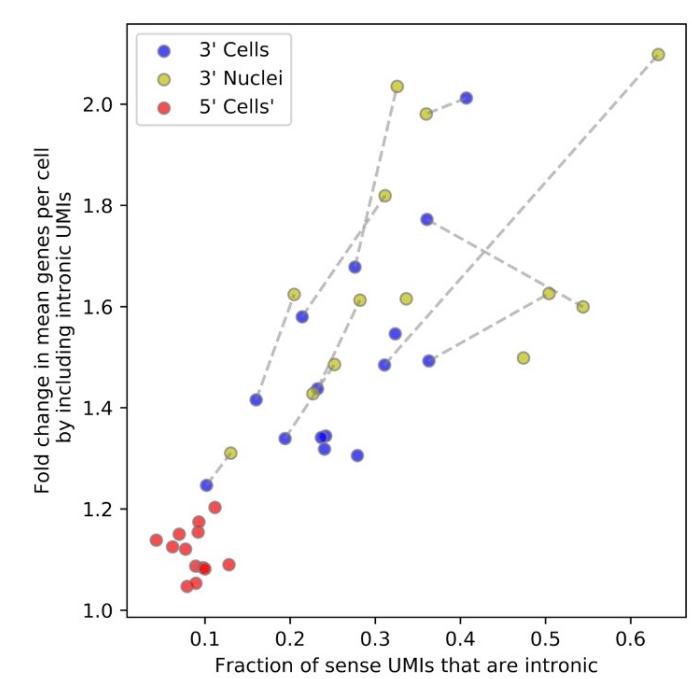
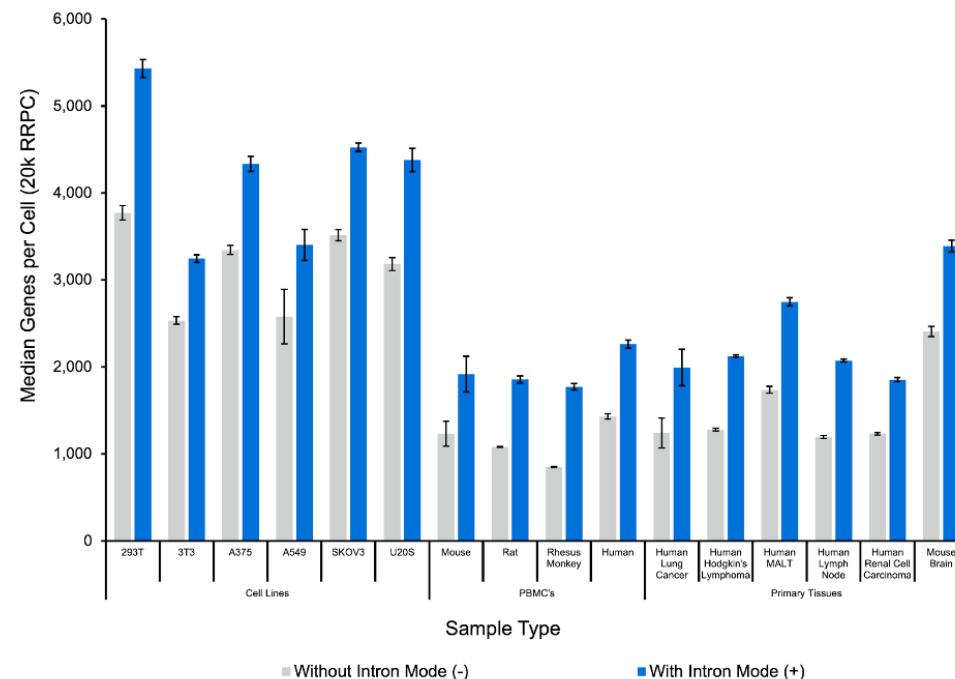
Should intronic reads be counted toward UMI counts?

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->

Alternative Aligners

- STARsolo: <https://github.com/alexdobin/STAR/>
 - scumi: <https://bitbucket.org/jerry00/scumi-dev/src/master/>
- kallisto bustools: <https://www.kallistobus.tools/>
 - Pseudoalignment
- Alevin (salmon) <https://salmon.readthedocs.io/en/latest/alevin.html>
- Alevin-fry <https://github.com/COMBINE-lab/alevin-fry>

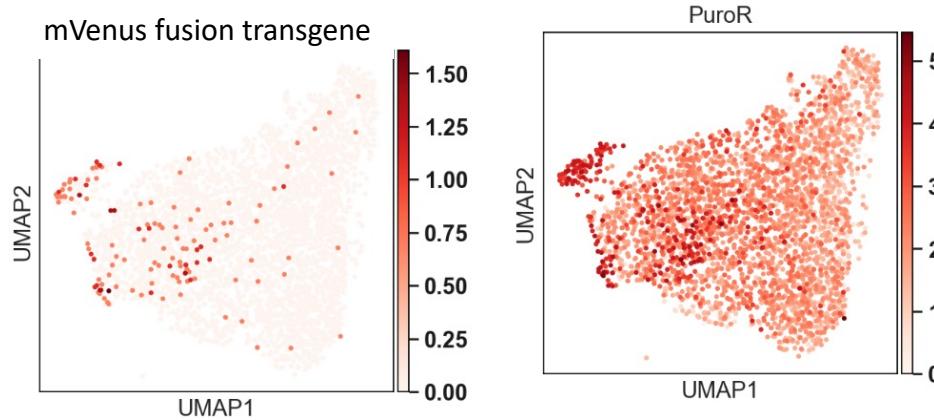
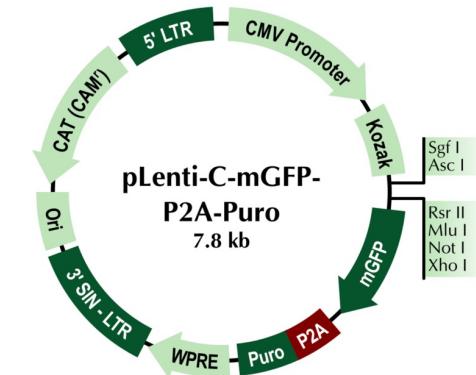
Summary

	Cell Ranger	STARsolo	Alevin	Alevin-fry	Kallisto
Mapping performance	Longest runtime	- Short runtime - Comparable results with Cell Ranger	- Whitelisting causes loss or gain of barcodes	- Faster mapping in comparison with Alevin. - Pseudoalignment (sketch mode) further decreases runtime	- Shortest runtime - highest mapping rate
Barcode correction and filtering			- Detected barcodes that are not in the whitelist	- More barcodes are retained than in Alevin	- Reports more cells
Gene discovery				- Lower detection of Vmn and Olfr gene family than in Alevin	- Highest detection rate of genes - Highest UMI count for genes not expressed in studied tissue
Differences between filtered and unfiltered annotation	- Multi-mapped reads are discarded	- Multi-mapped reads are discarded - EM-algorithm can be used (optional)	- Counts of multi-mapped reads split with EM-algorithm	- Multi-mapped reads are discarded - EM-algorithm can be used (optional)	- Multi-mapped reads are discarded - EM-algorithm can be used (optional)
Clustering	- Highest Overlap with SCINA classification	- Very similar to Cell Ranger with minor differences	- Cell types contain lower amount of cells with SCINA classification		- High amount of barcodes not detected
DEG	- No difference detected	- No difference detected	- Lower detection rate than STARsolo and Alevin-fry	- Improved concordance (than Alevin) with Cell Ranger	- Lowest concordance with Cell Ranger
Practical Recommendation	- Replacement with STARsolo is recommended	- Recommended as a general purpose mapper		- Pseudoalignment is especially suitable for huge datasets	- Fast mapper - qualitative issues with gene detection

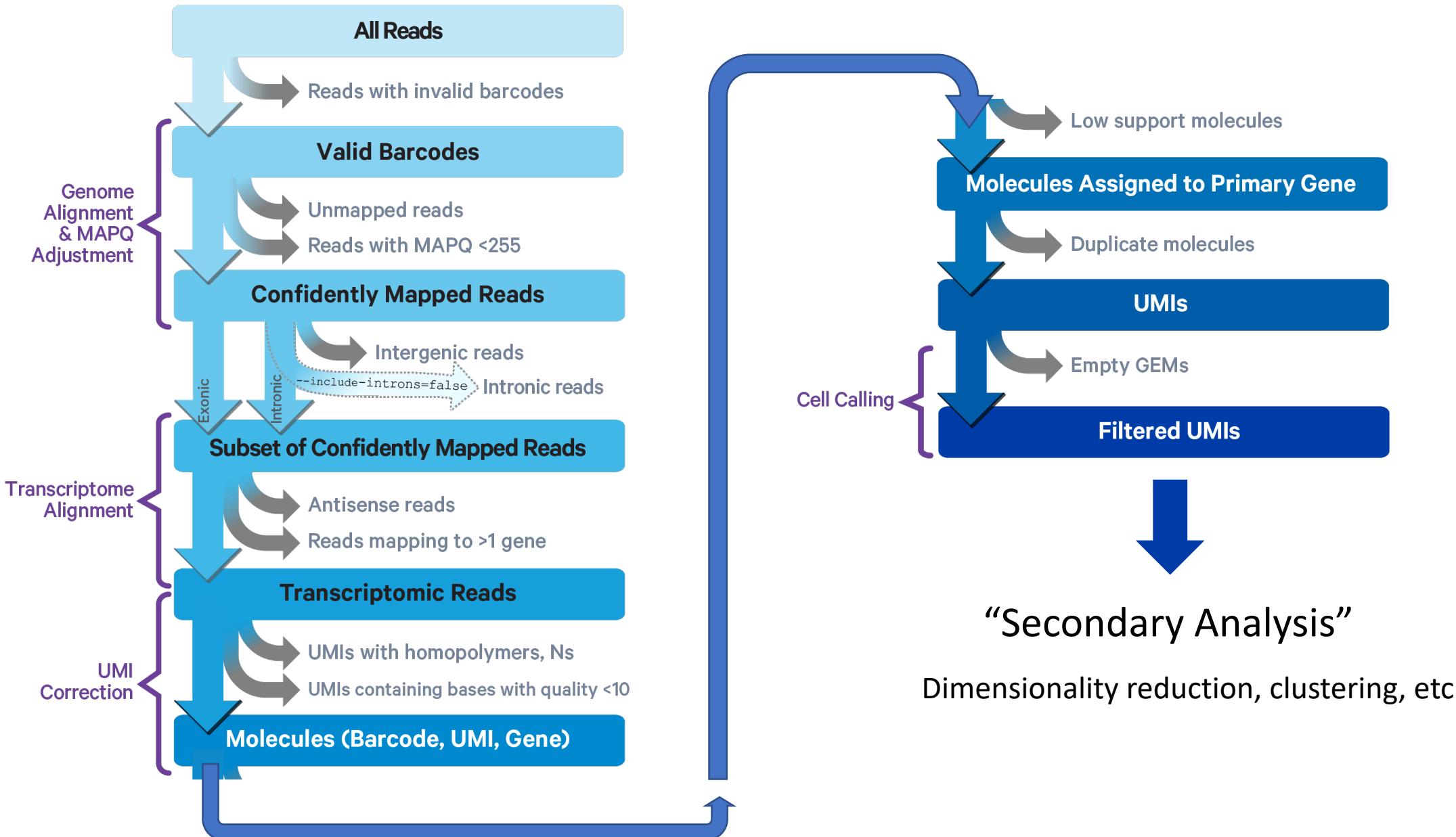
Custom Genomes w/ cellranger mkref

<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



Overview of cellranger count



Cell Ranger Pipeline Output



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

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

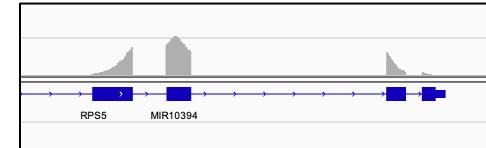
BAM file

You should Archive this.

```
samtools view possorted_genome_bam.bam | head -n 1

NB501555:883:HKHCNBGXH:1:13203:17140:14586 16 1 3000079 255 56M * 0 0
AAACCATTGGTCCCTTCTTTTTTTTTTTTTGGGTGGGAGAC
EE//A//////////6EEEEEEEEEEEEEEEEEEEEAAAAA 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:AGCTTCCTCTTCCGA CY:Z:AAAAAEEEEEEEEE CB:Z:AGCTTCCTCTTCCGA-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 used to extract per-cell genotype / allelic expression using [Vartrix](#)

Vartrix

```
##fileformat=VCFv4.0

##contig=<ID=chr1,length=248956422>
##contig=<ID=chr10,length=133797422>
##contig=<ID=chr11,length=135086622>
... (more header)

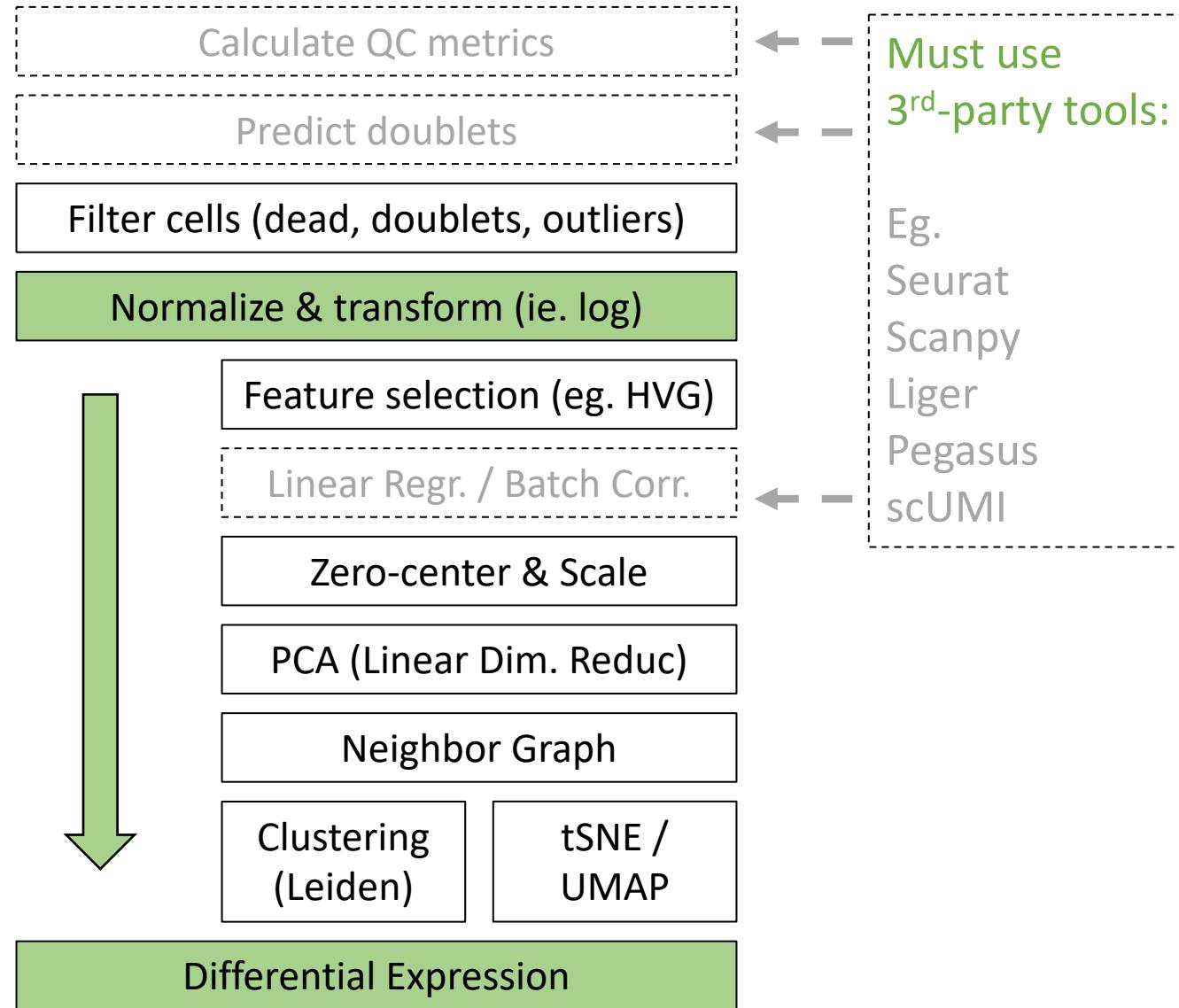
1 226064434 H3.3-K27M A T . PASS DP=100
```

Which Cell Ranger pipeline to use?

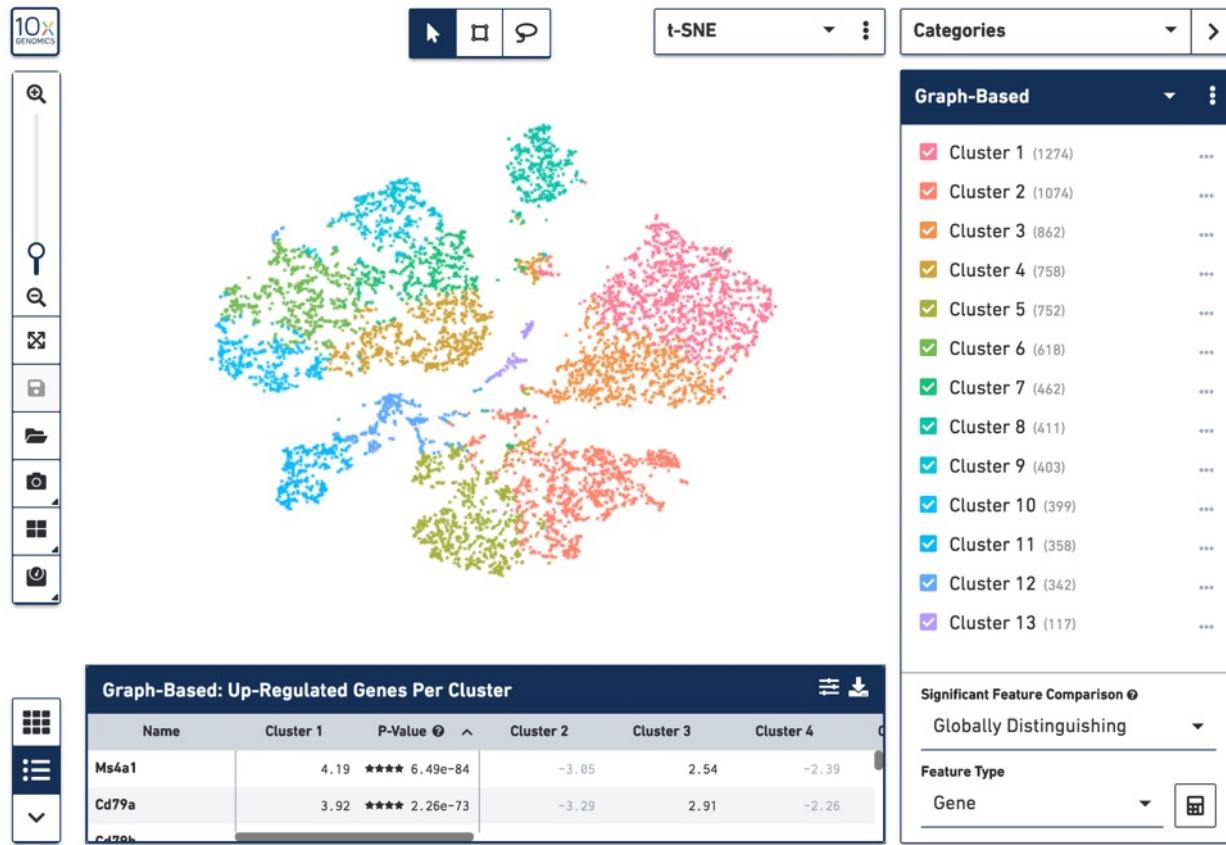
Library type	Cell Ranger pipeline
<u>3/5' Gene Expression</u>	count
<u>3' Targeted Gene Expression</u>	count
<u>3/5' Gene Expression + Antibody/CRISPR Guide Capture</u>	count
<u>Antibody Capture only</u>	count
<u>3/5' Gene Expression + Cell Multiplexing (+ Antibody/CRISPR Guide Capture)</u>	multi
<u>Fixed RNA Profiling Gene Expression (+ Antibody Capture)</u>	multi
<u>ATAC</u>	atac
<u>RNA + ATAC Multiome</u>	arc

Cell Ranger

Secondary Analysis



Loupe Browser



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.html

FTPS22_Ctrl

Summary Analysis

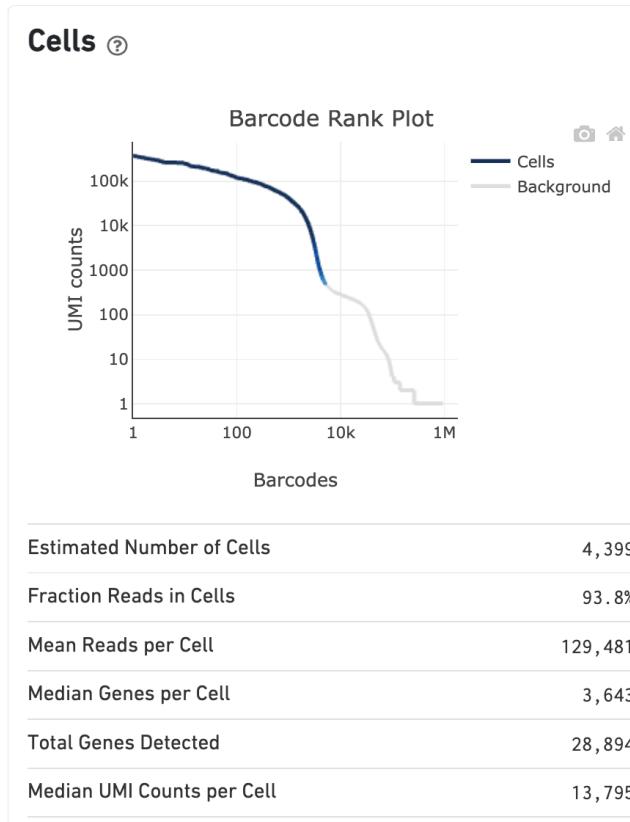
4,399
Estimated Number of Cells

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%



Mapping ?

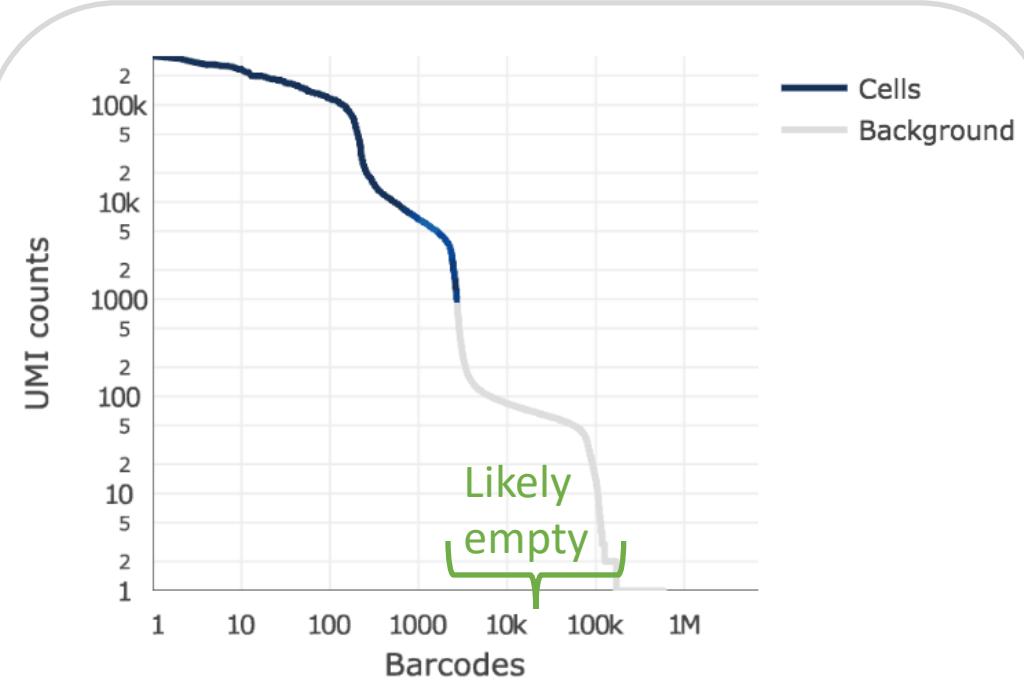
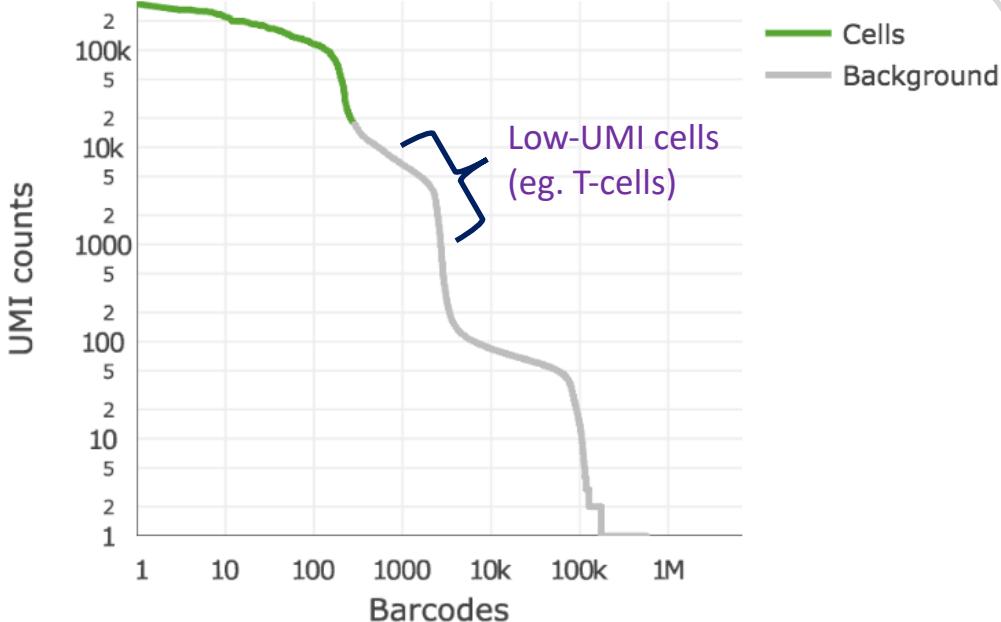
Reads Mapped to Genome	64.5%
Reads Mapped Confidently to Genome	58.7%
Reads Mapped Confidently to Intergenic Regions	4.4%
Reads Mapped Confidently to Intronic Regions	0.6%
Reads Mapped Confidently to Exonic Regions	53.7%
Reads Mapped Confidently to Transcriptome	51.2%
Reads Mapped Antisense to Gene	0.5%

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

Cell Calling Algorithm

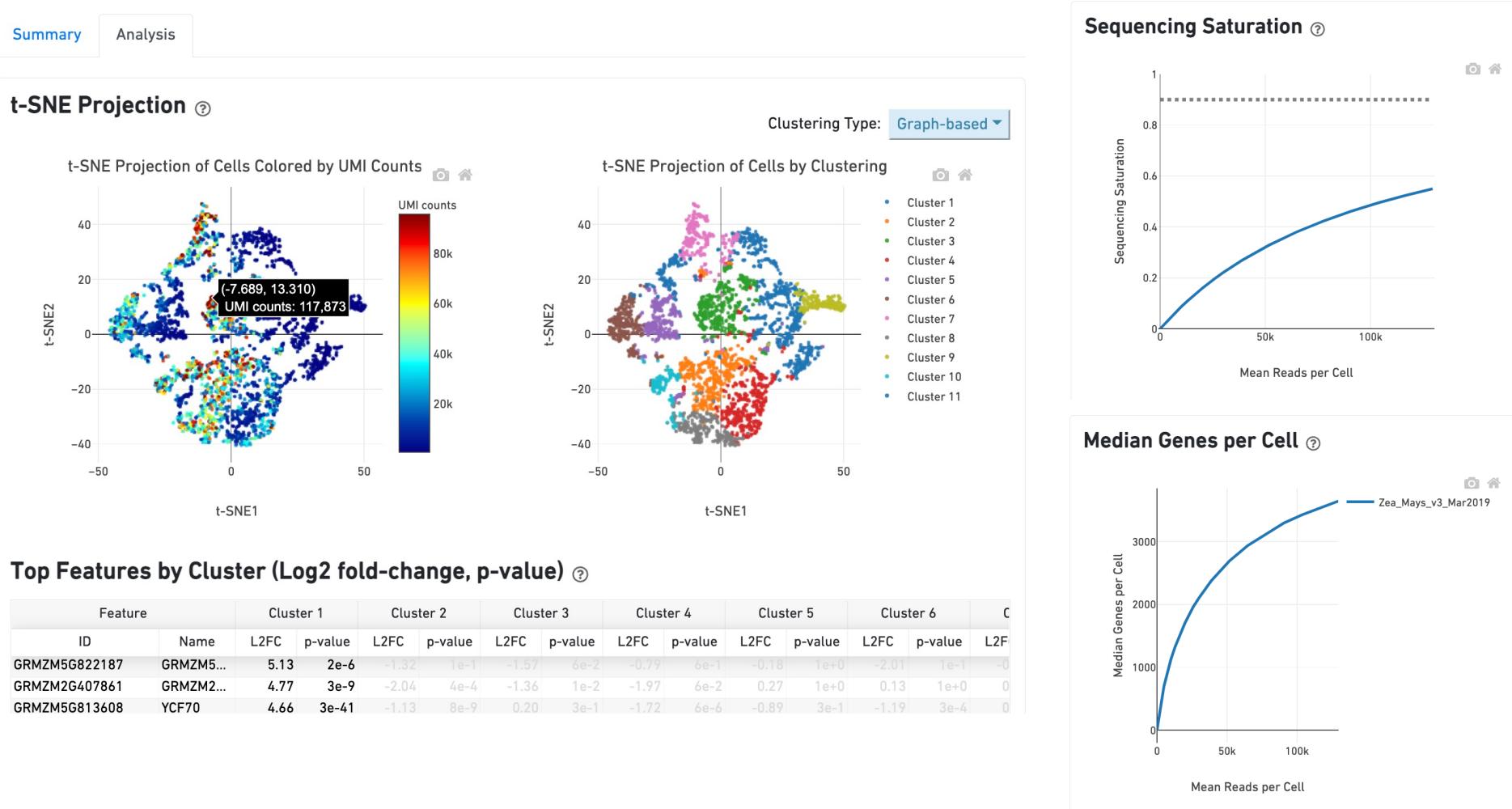
Based on **EmptyDrops** <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1662-y>



- 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”

Sequencing Saturation

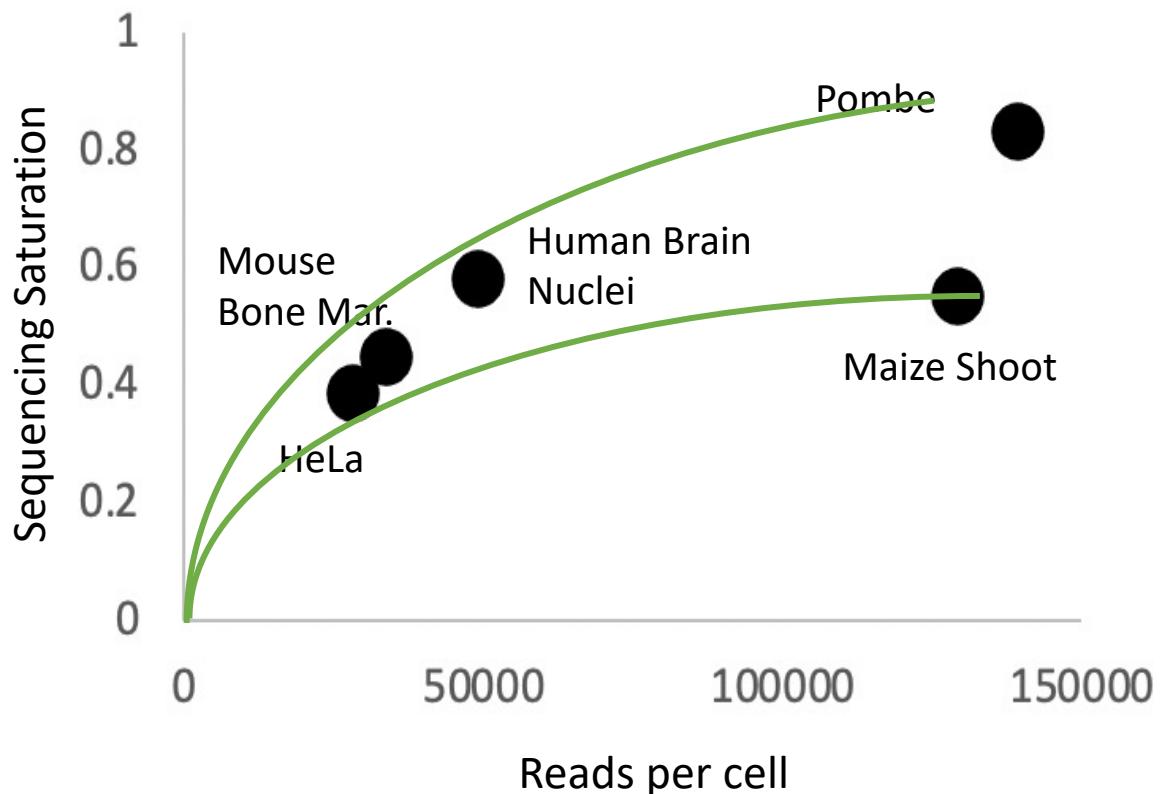


Putting up the numbers

	Maize Shoot Meristem	Mouse Bone Marrow	Human Brain Nuclei	HeLa cells	S. Pombe
Genome Mapping %	58.70%	87.10%	90.40%	94.90%	52.10%
Transcript Mapping %	51.20%	67.10%	77.40%	72.20%	50.50%
UMIs/cell	13,795	4,278	4,042	10,596	3,266
Genes/cell	3,643	1,770	2,028	2,857	982
Intronic %	0.60%	14.10%	66.80%	15.20%	0.10%
Antisense	0.50%	1.50%	4.30%	1.20%	0.40%
Reads per cell	129,481	34,018	49,303	28,421	139,430
Saturation	55%	44.90%	58.20%	38.50%	83.30%



Samples saturate at different rates



Function of:

Sample Composition:

of cells

of UMIs/cell

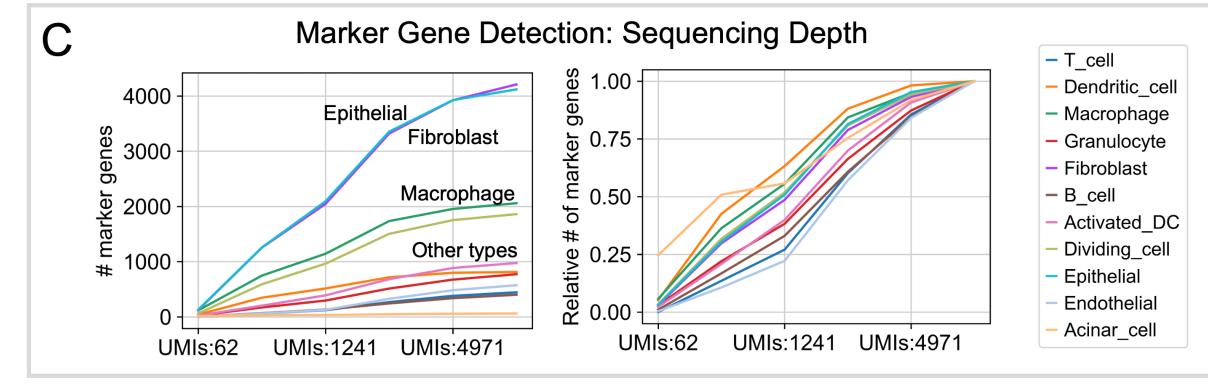
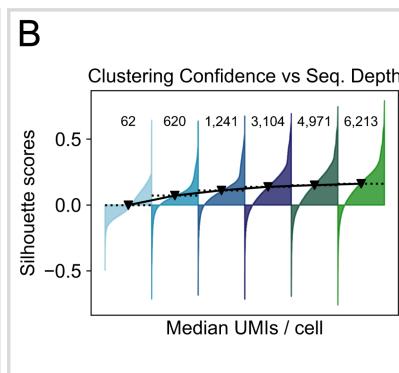
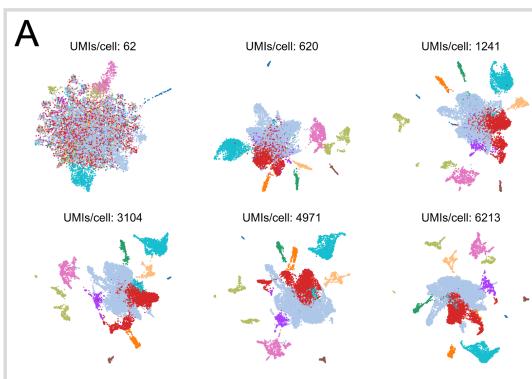
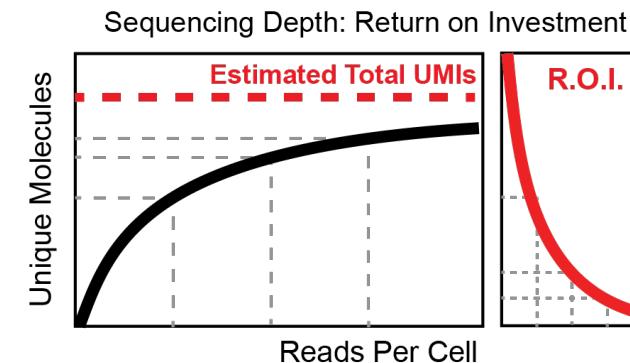
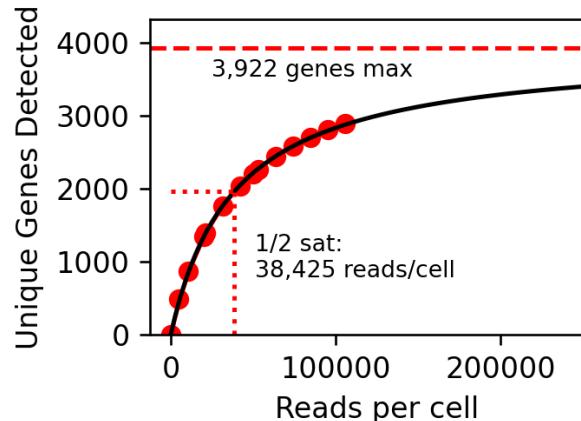
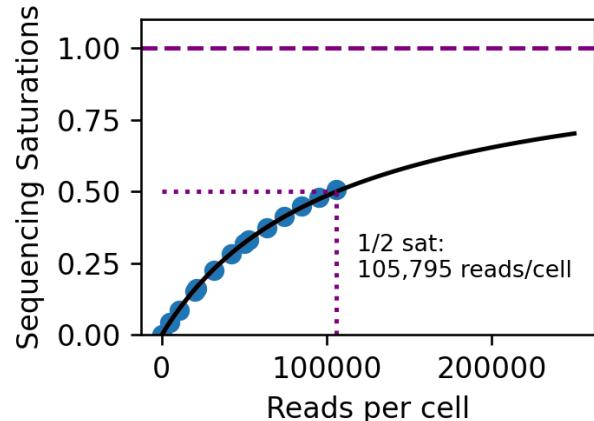
Library Quality:

PCR duplication

Adapter dimers

Contaminating DNA (mapping rate)

Sequencing Saturation – How deep?



Combine samples with `cellranger aggr`

First, make an `aggr.csv` file

```
sample_id,molecule_h5
SeqTech22_RNA_10k,/fake/path/SqTech22/count/SqTech22_RNA_10k/outs/molecule_info.h5
SeqTech22_RNA_12k,/fake/path/SqTech22/count/SqTech22_RNA_12k/outs/molecule_info.h5
```

```
PROJECTDIR=/fake/path/SqTech22/
cellranger aggr --id=SeqTech \
    --jobmode=local \
    --csv=$PROJECTDIR/aggr.csv \
    --normalize=none \ # or --normalize=mapped
    --localcores=16 \
    --localmem=64
```

`--normalize=mapped` will subsample reads from higher-depth libraries until all have roughly similar numbers of reads per cell. It isn't frequently used in publications. Rather, try your best to just sequence all samples to the same depth.

`--normalize=None` simply concatenates the two matrices together, and gives each samples a unique barcode suffix, in the order they appear in the `aggr.csv` file (-1, -2, -3, etc)

Matrix format: HDF5 vs MTX (MEX)

Market Exchange (MEX) Format

- Simple, deprecation-resistant
- Slow, lazy about metadata

barcodes.csv

```
AAACCCAAGAATCCCT-1  
AAACCCAAGCAACCAG-1  
AAACCCAAGGTCTGC-1  
AAACCCACAATAGTCC-1
```

features.csv

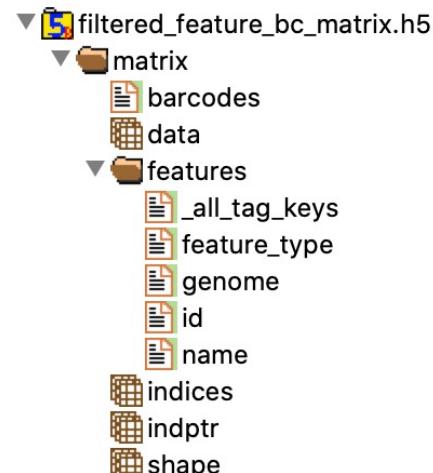
ENSG00000243485	MIR1302-2HG	Gene Expression	
ENSG00000237613	FAM138A	Gene Expression	
ENSG00000186092	OR4F5	Gene Expression	
ENSG00000238009	AL627309.1	Gene Expression	

matrix.mtx

```
%%MatrixMarket matrix coordinate integer general  
%metadata_json: {"software_version": "cellranger-  
6.0.0", "format_version": 2}  
36601 4500 13943571  
24 1 5  
49 1 1  
54 1 1  
58 1 1  
60 1 4  
63 1 7  
72 1 2
```

Hierarchical Data Format 5 (HDF5)

- Fast, stashes arbitrary data/metadata
- Require software to open – vulnerable to bit rot



Both store data
in SPARSE format

Name	Value[50]...
chemistry_description	Single Cell 3' v2
filetype	matrix
library_ids	ST22_int
original_gem_groups	1
software_version	cellranger-7.0.0
version	2

Metadata!

Sparse vs Dense Matrices

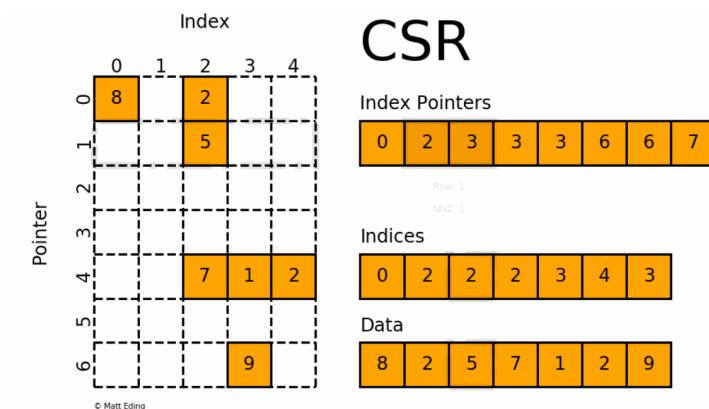
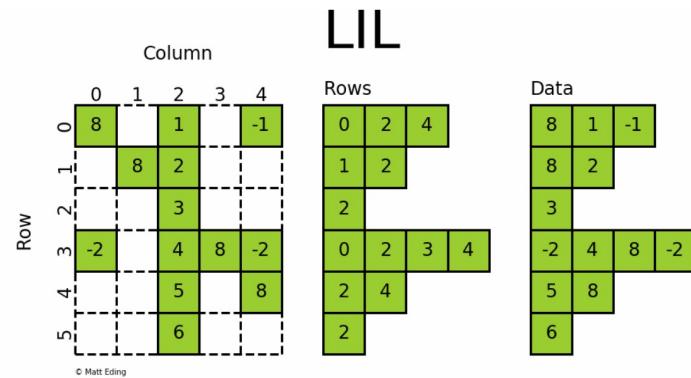
DENSE

0	7	0	0	0	0	6
0	7	6	3	0	4	0
0	4	3	0	0	0	0
4	2	0	0	0	0	0
0	0	0	0	3	2	4

s p a r s e

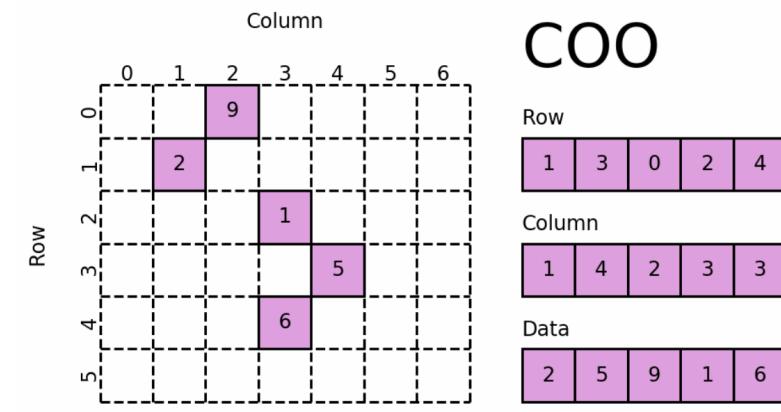
7	6	3	4	2	4
7	6	3	4		
4	3				
4	2				

© Matt Eding



Different ways to encode sparse matrices
Use packages like `scipy.sparse`

Each encoding is optimized for different types
of computations, memory usage, I/O, etc



Calculate QC metrics	A-	
Predict doublets	F	
Filter cells (dead, doublets, outliers)	F	Cell Ranger's Report Card
Normalize & transform (ie. log)	B	
Feature selection (eg. HVG)	D	Batch Correction D
Regress unwanted variation	F	Cell Type Identification F
PCA (Linear Dim. Reduc)	A	Sharing Data with Biologist Colleagues A++
Neighbor Graph	A	
Clustering (Leiden)	B	Quickly visualizing markers A++
tSNE / UMAP	A	
Differential Expression	B+	Quickly finding diffex. genes A-

Demo Datasets

<https://www.10xgenomics.com/resources/datasets>

Products

- Single Cell Gene Expression**
- Single Cell Immune Profiling
- Single Cell ATAC
- Spatial Gene Expression
- De Novo Assembly
- Single Cell CNV
- Targeted Gene Expression
- Genome & Exome
- Single Cell Multiome ATAC + Gene Expression
- Fixed RNA Profiling

Single Cell Gene Expression

Chemistry Version

All

Pipeline Version

All

[Reset Filters](#)

Application Note - Alternative Transcript Isoform Detection With Single Cell and Spatial Resolution (v3.1 Chemistry) +

Nuclei Isolation for Single Cell Gene Expression (v3.1 Chemistry) +

Comparing 3' Datasets With and Without Intronic Reads (v3.1 Chemistry) +

Chromium X Series 3' Demonstration (v3.1 Chemistry) +

EXERCISE 1

- Interactively explore a PBMC dataset from SeqTech 2023
- https://www.dropbox.com/scl/fi/riq835zvulmhnm1fk6mcb/iMac_package.zip?rlkey=1yljwzq5elratwn8pq7sq4bgz&dl=0

EXERCISE 2

Secondary Analysis

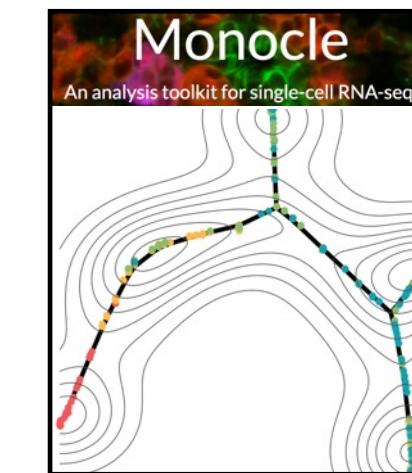


Fabian Theis - München



<https://scanpy.readthedocs.io/en/latest/>

Python



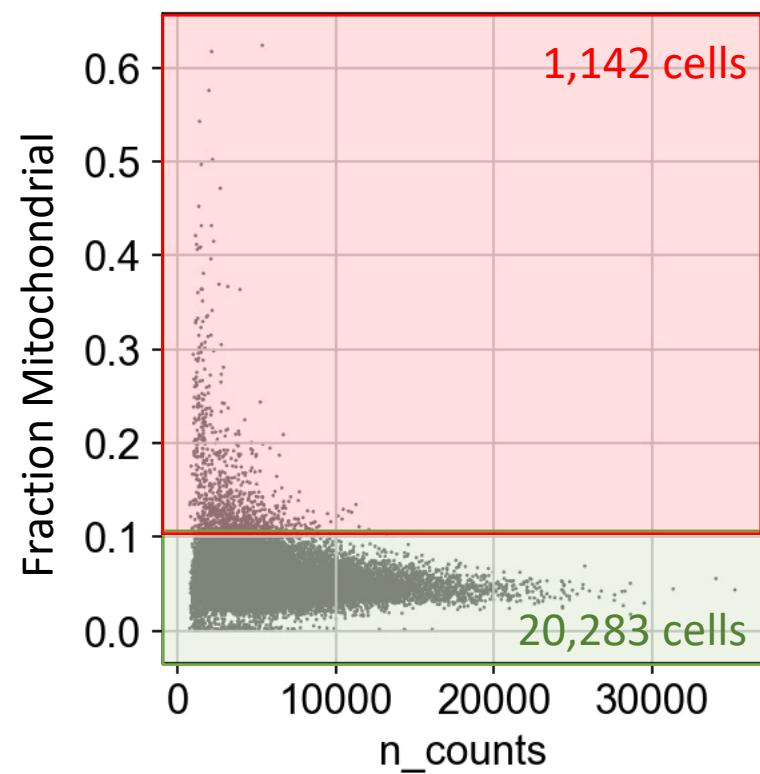
Cole Trapnell –WashU



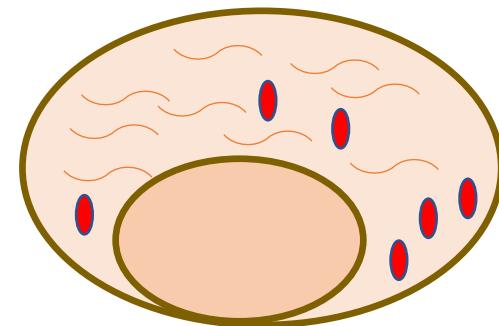
Macosko lab

QC metrics – Cell Death

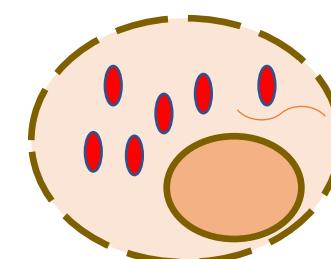
Filter cells by Fraction Mitochondrial Reads



Healthy Cell



mRNA
Mitochondrion

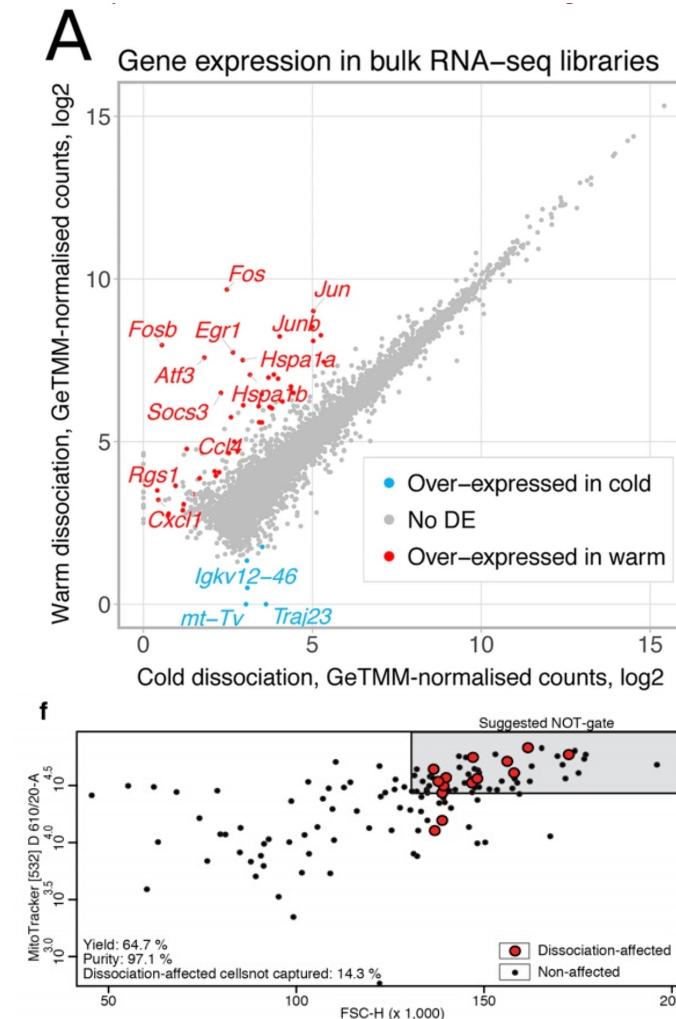
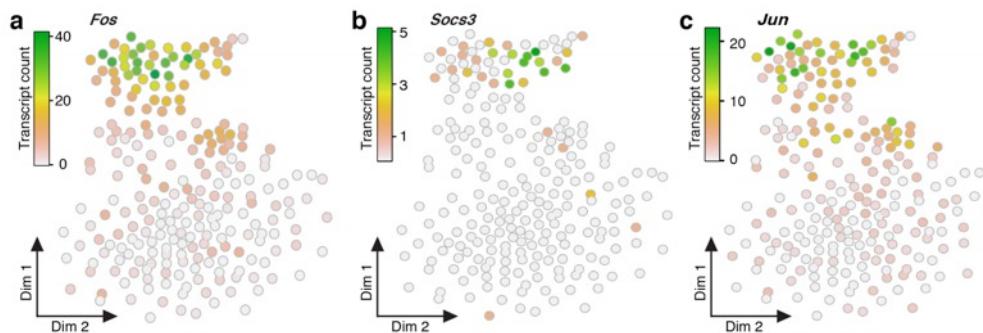


Dying / Ruptured Cell

QC metrics – Dissociation Stress

Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations

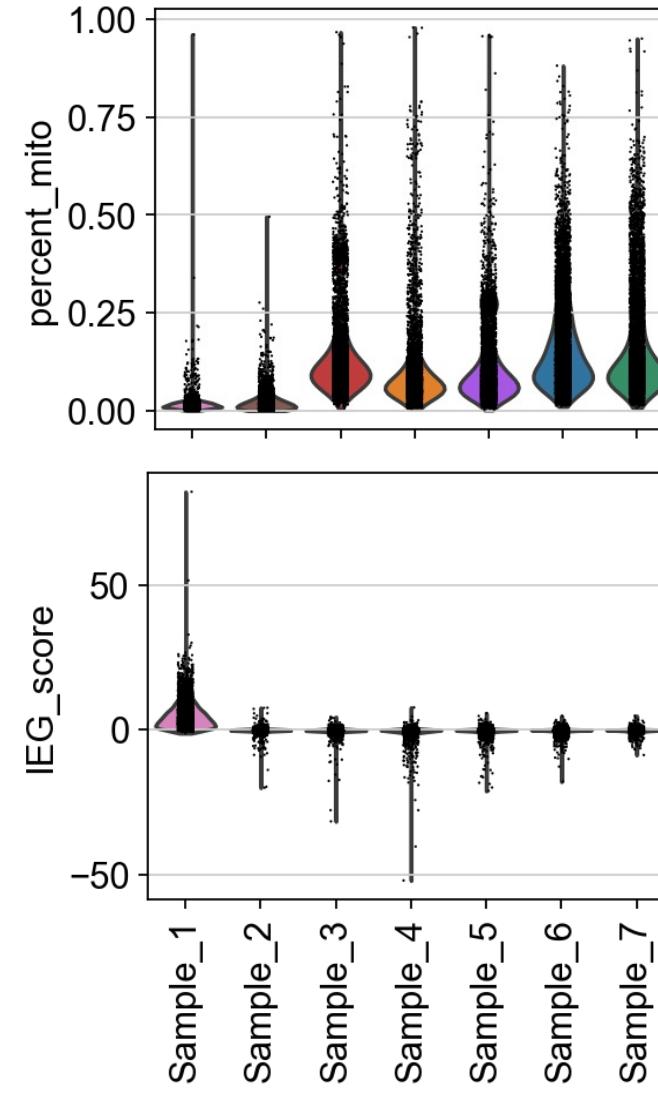
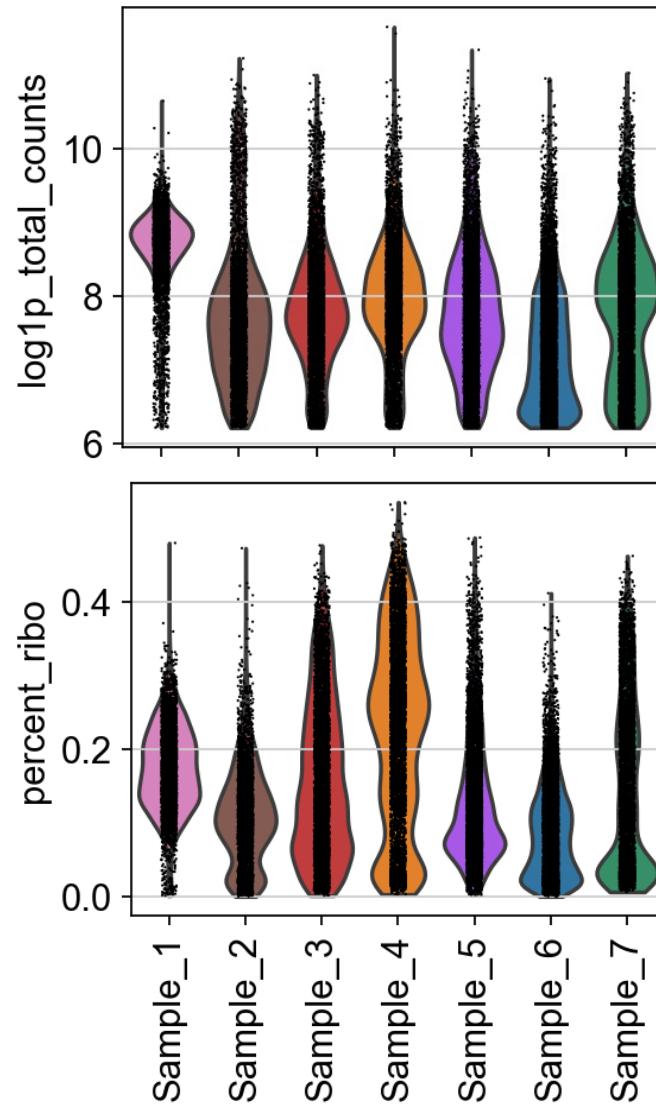
- Enzymatic treatment and suspension at 37C lead to induction of immediate-early genes (IEG)
- Artifactual, post-dissociation transcription (and degradation) can obscure underlying transcriptome



QC metrics don't always correspond to artifacts

UMI counts:
Cell types
Viability
Sequencing depth
Library handling

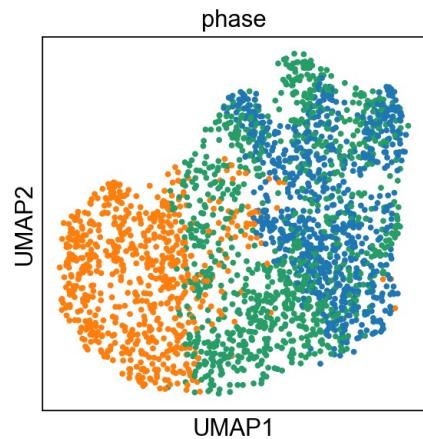
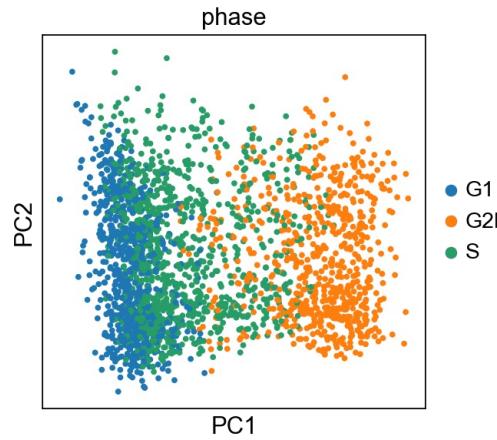
Ribosomal %
Cell types
Activation / Metabolism
Sequencing depth
(normalization artifact)



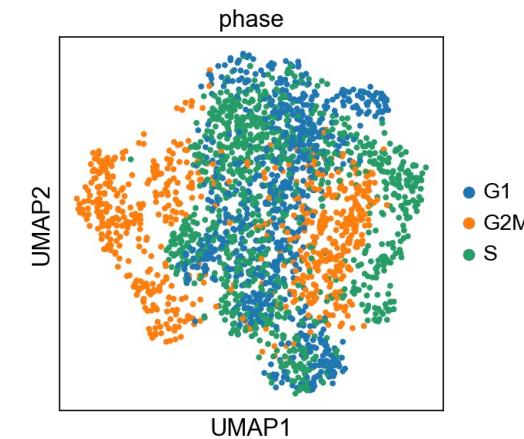
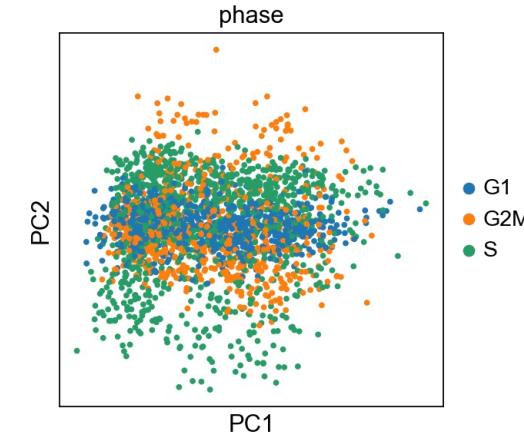
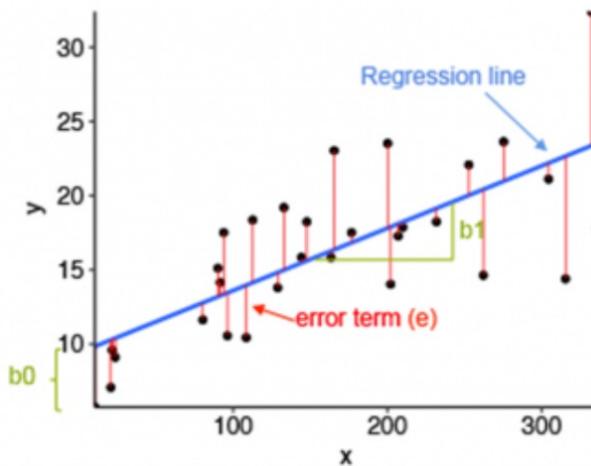
Mitochondrial %
Dying cells
Respiration activity!
(some cells have > 50%
but are alive!)

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

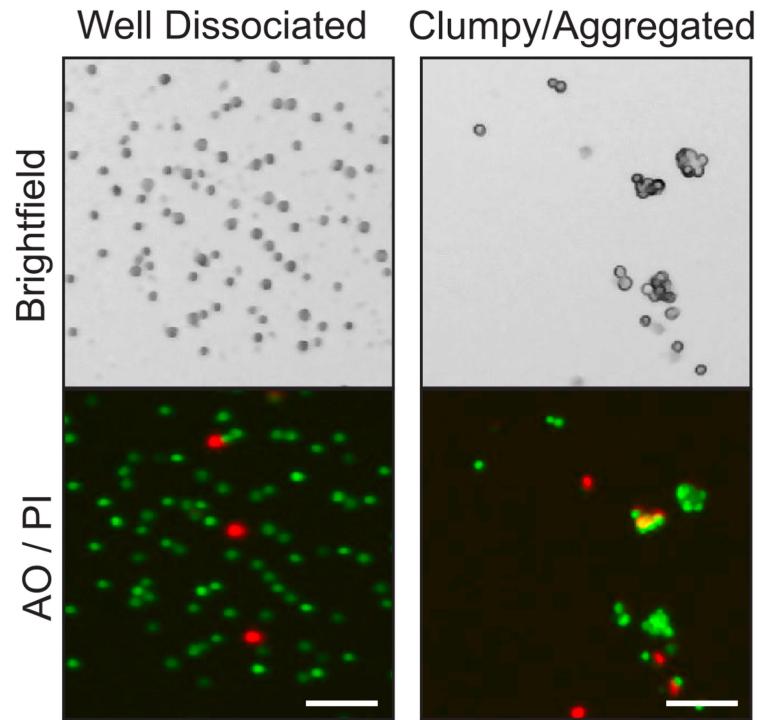
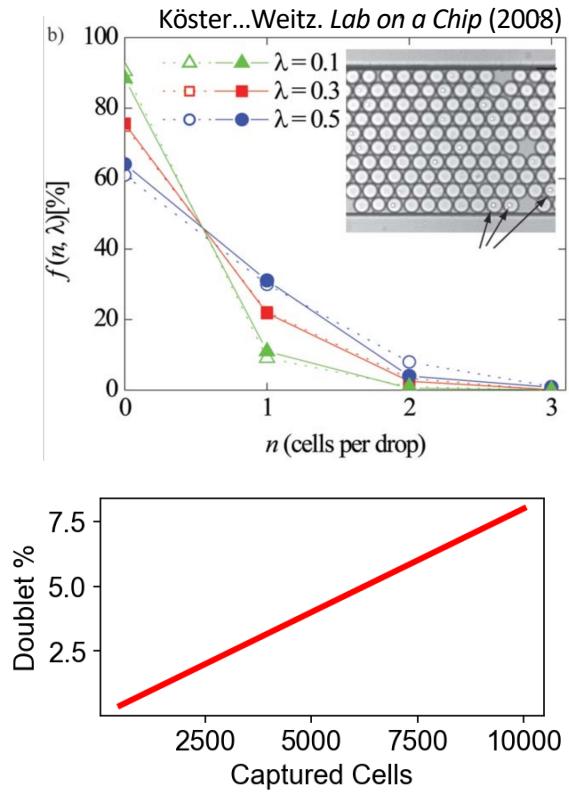
Linear Regression of Unwanted Variation



- Fit a linear model of each gene's expression vs confounder variable
 - Eg. cell cycle gene score, % mito, dissociation-stress response, library size
- Take model residuals as new scaled values
- **This often doesn't work well.**



Doublets / Multiplets



Doublet Filtering

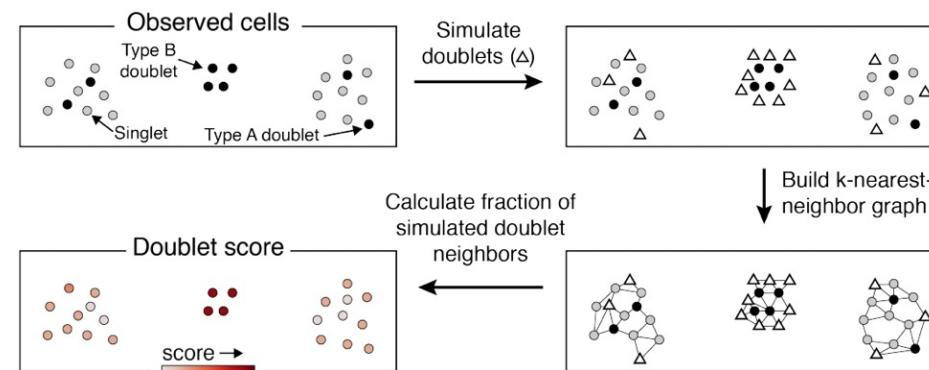
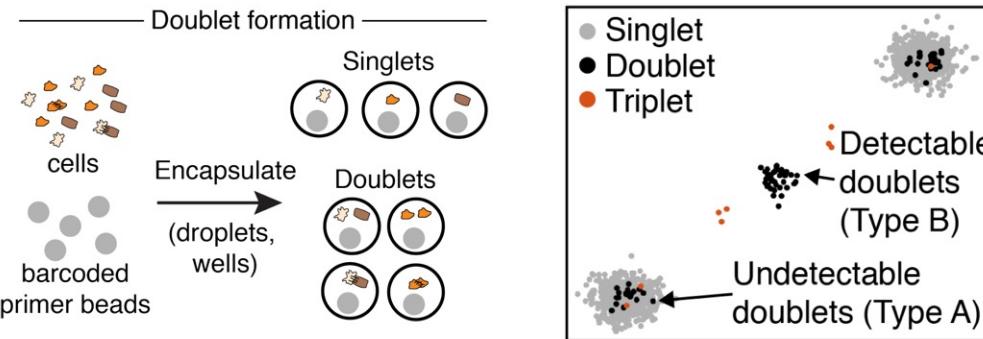
Scrublet

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

• [DoubletDecon](#) - [R] - Cell-State Aware Removal of Single-Cell RNA-Seq Doublets. [\[BioRxiv\]](#)(DoubletDecon: Cell-State Aware Removal of Single-Cell RNA-Seq Doublets)

• [DoubletDetection](#) - [R, Python] - A Python3 package to detect doublets (technical errors) in single-cell RNA-seq count matrices. An [R implementation](#) is in development.

• [Scrublet](#) - [Python] - Computational identification of cell doublets in single-cell transcriptomic data. [BioRxiv](#)



Wolock et al. (2018) bioRxiv

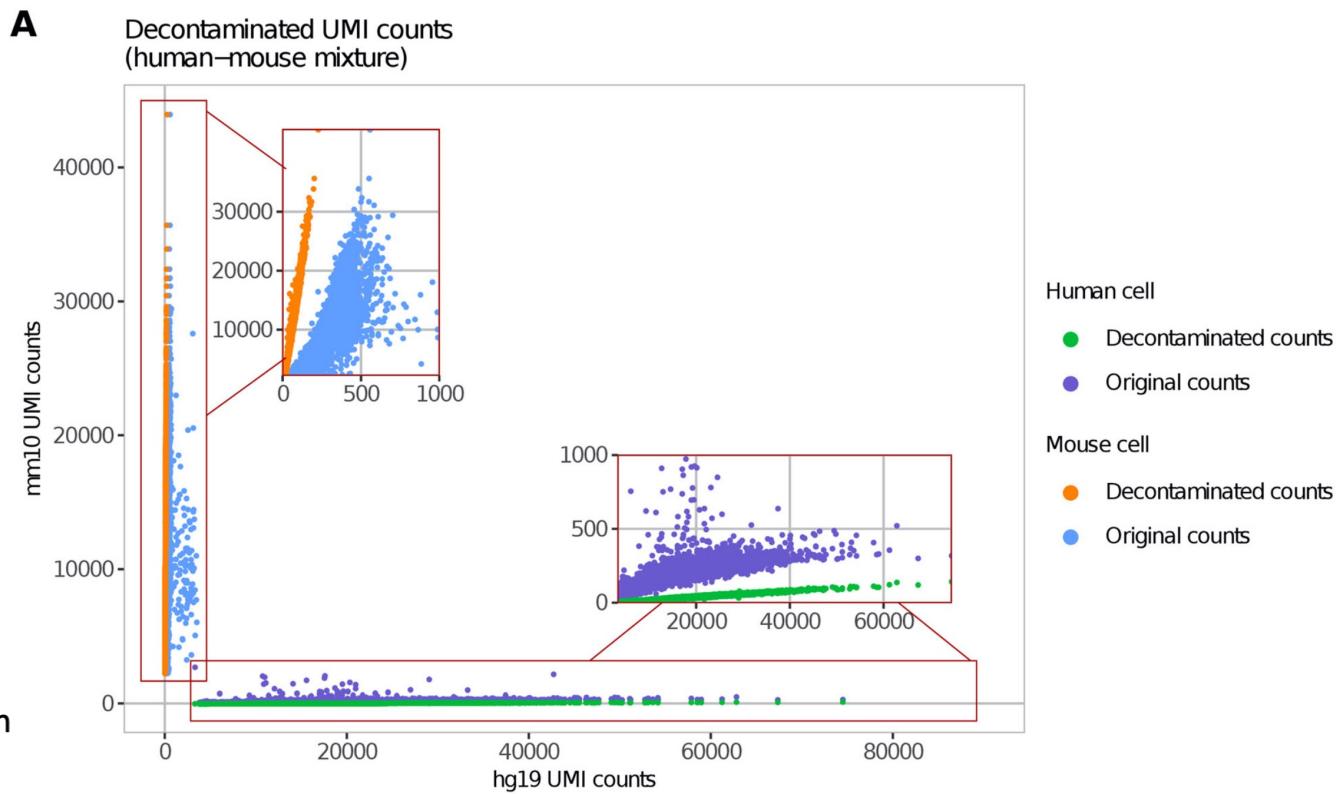
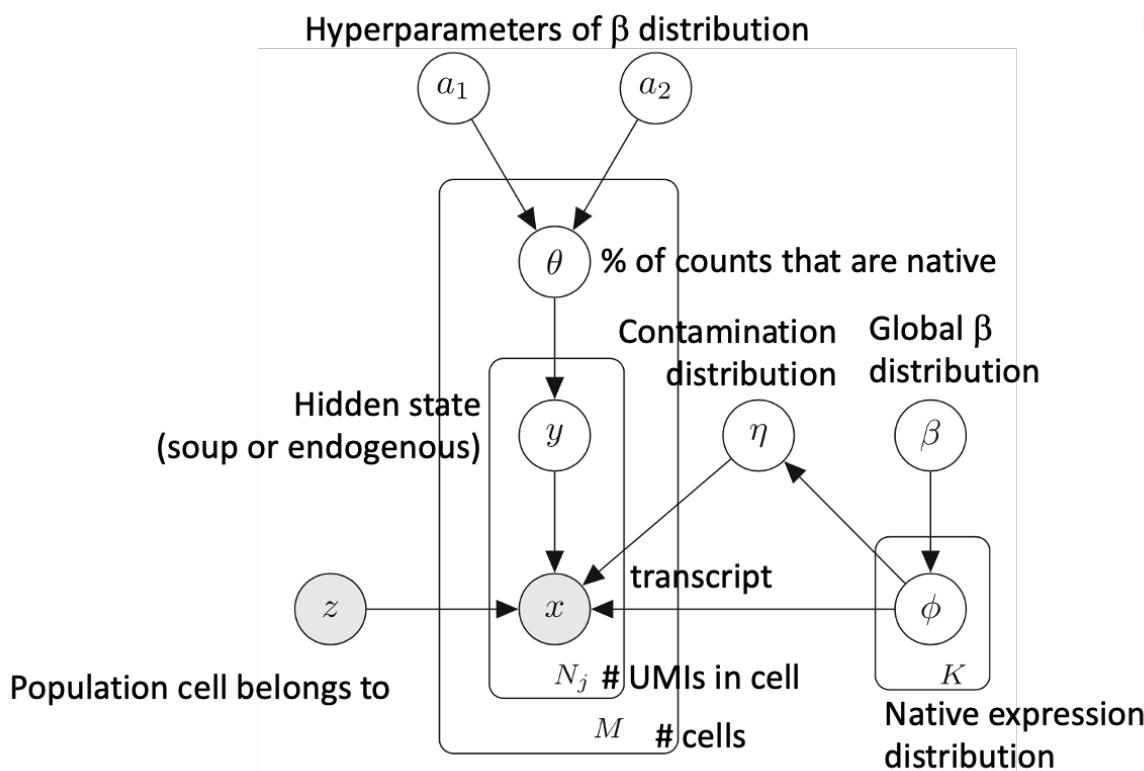


Decontamination of ambient RNA in single-cell RNA-seq with DecontX

Shiyi Yang¹ , Sean E. Corbett¹, Yusuke Koga¹, Zhe Wang¹ , W Evan Johnson¹, Masanao Yajima² and Joshua D. Campbell^{1*}

Uses Variational Bayes Inference
Similar to Latent Dirichlet Allocation (LDA)

Models soup as a weighted combination of other cell types in the population



Normalization & log transformation

Basic Method:

$$\frac{\text{Gene(i)Counts}}{\text{All Gene Counts}}$$

* 10,000 (ie. transcripts per 10k, “tp10k”)

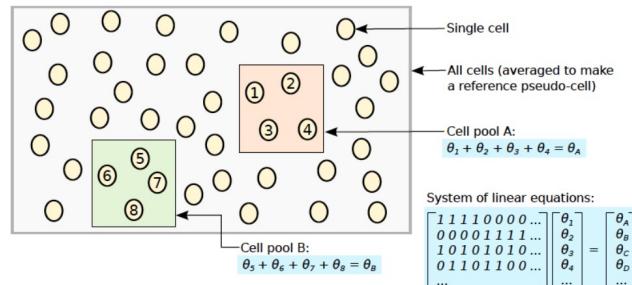
Alternatively, though less commonly:

* 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!

Other options, eg scran normalization

- Estimate cell-specific size factors.
- Handles sparsity and is robust to DE.

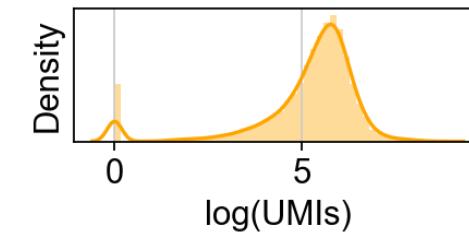
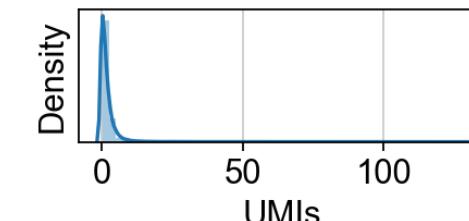


Lun et al., Genome Biology (2016)

<http://bioconductor.org/packages/release/bioc/html/scran.html>



Add 1



Logarithmize

Basic variance stabilizing transformation
Other options:

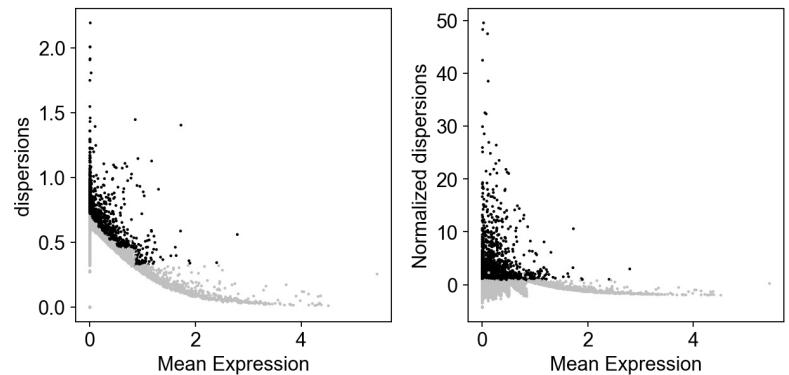
Arcsin(h)

SCTransform (Seurat)

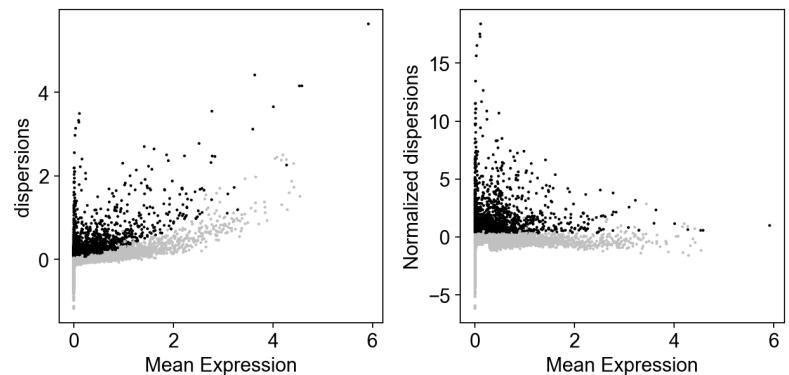
Pearson residuals

Feature Selection: Highly Variable Genes (HVGs)

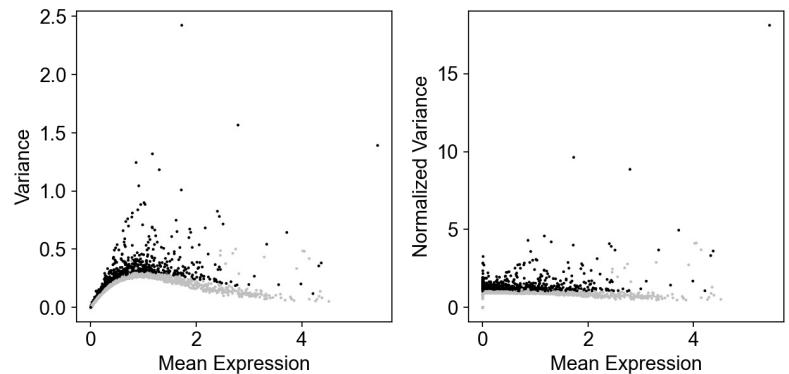
Cell
Ranger



Seurat



Seurat
v3

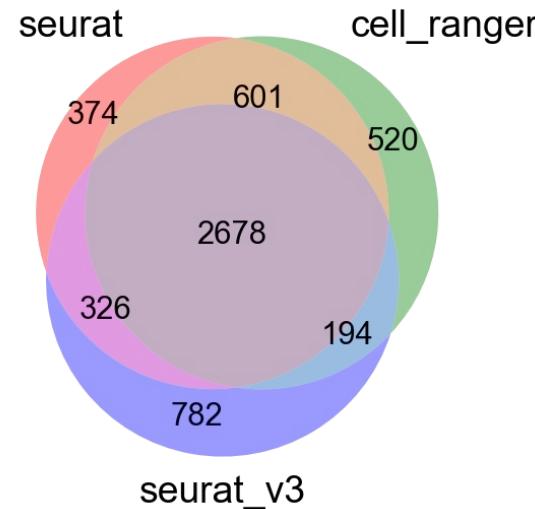


Assumption: genes that are interesting in the data will have higher variance

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

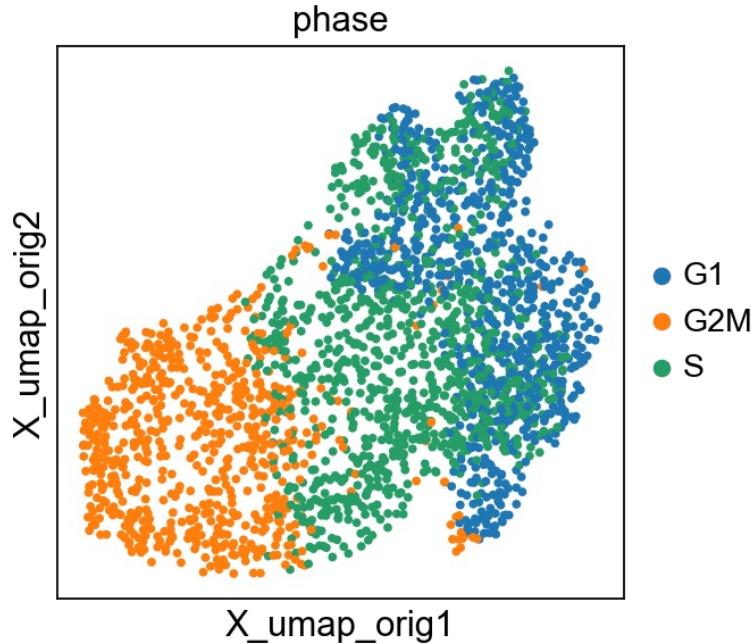
Thus, different approaches to select HVGs exist to select relevant HVGs to account for their expression ‘bin’

Alternatively: do PCA on ALL genes (computationally intensive)



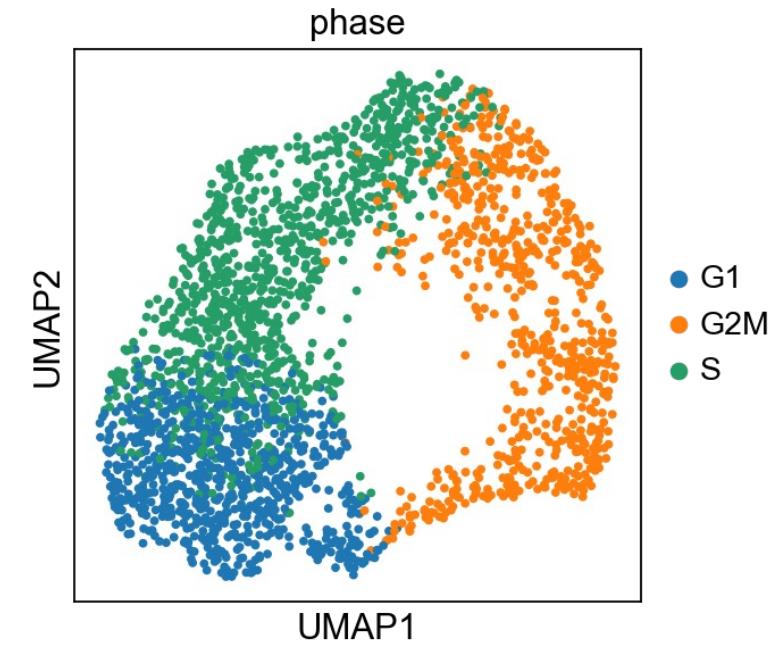
Feature Selection: Supervised gene sets

Siha cervical cancer cell line



529 curated cell-cycle genes
Macosko et al. (Cell) 2015 May 21;
161(5): 1202–1214.

PCA, UMAP



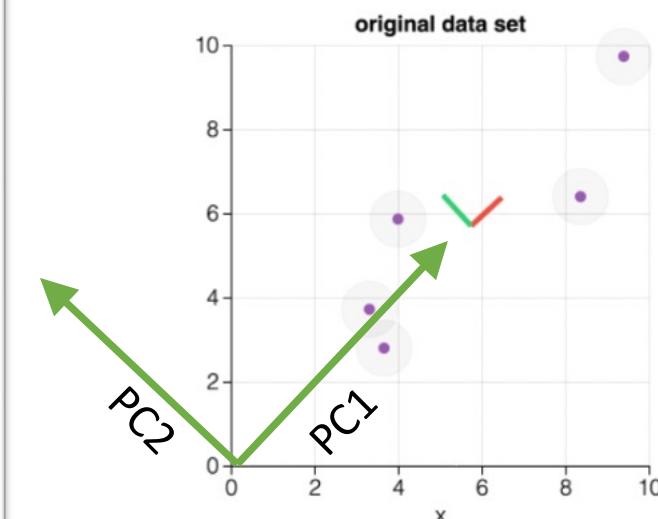
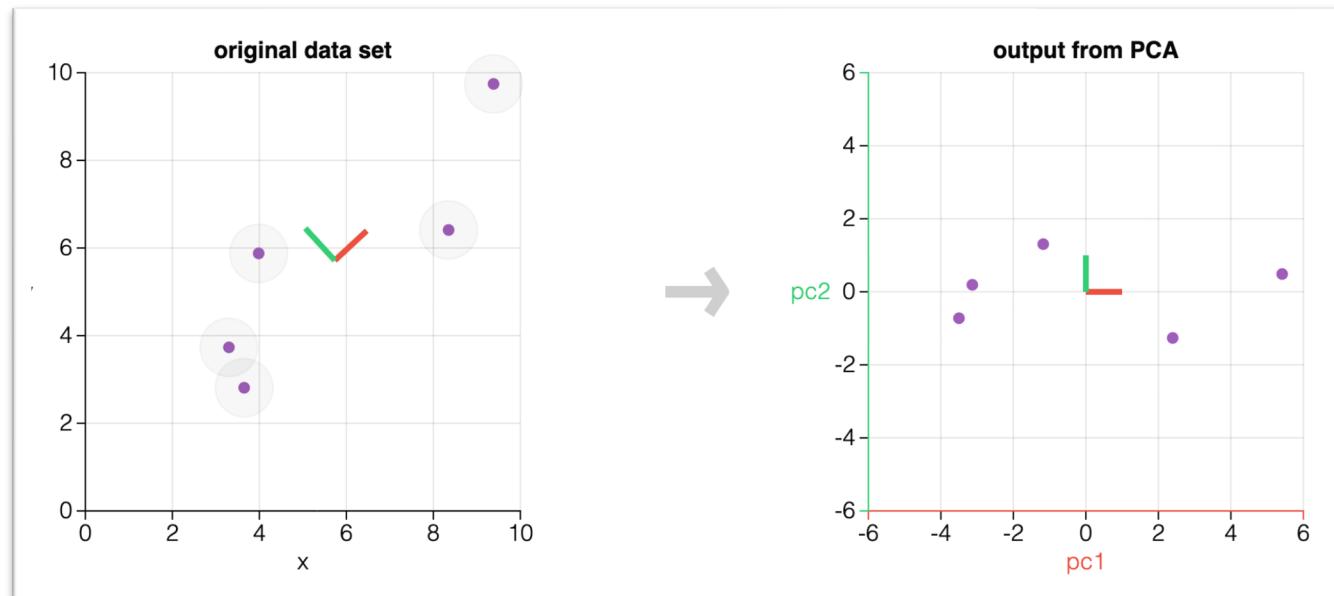
~1D circular manifold!

Principal Component Analysis

NOT a dimensionality reduction method, *per se*. It is a ROTATION.

1. Zero-center data (subtract mean)
 2. Scale Uniformly (ie. Z-score by gene)
 3. Produce covariance matrix
 4. Eigen decomposition
- } Singular value decomposition

- Identifies linear combinations of genes that explain the most variation in the data
- PCs ~ analogous to **gene expression programs**
- PCs are *orthogonal* (**gene expression programs** might not be)



Cool interactive visualization

<https://setosa.io/ev/principal-component-analysis/>

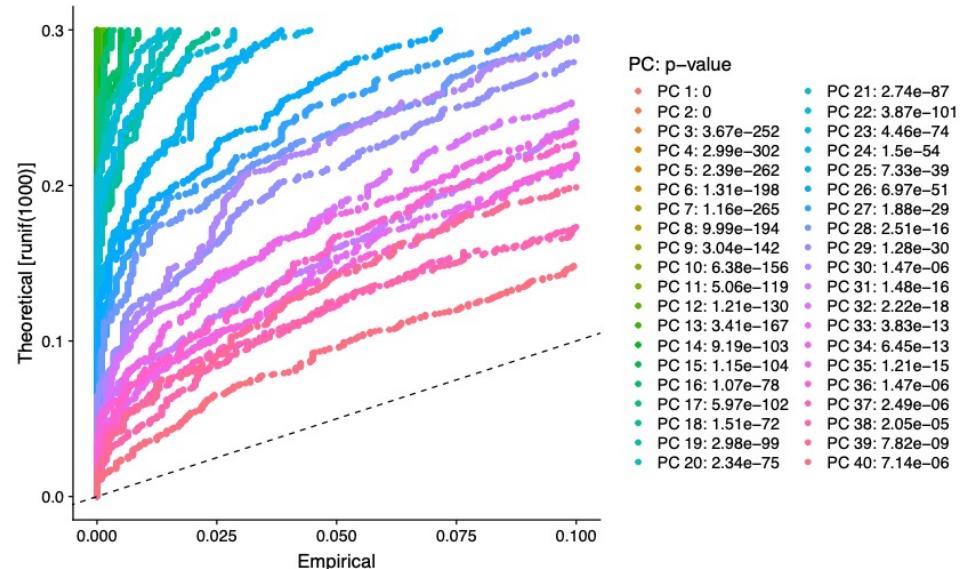
How Many PCs?

Option 1:

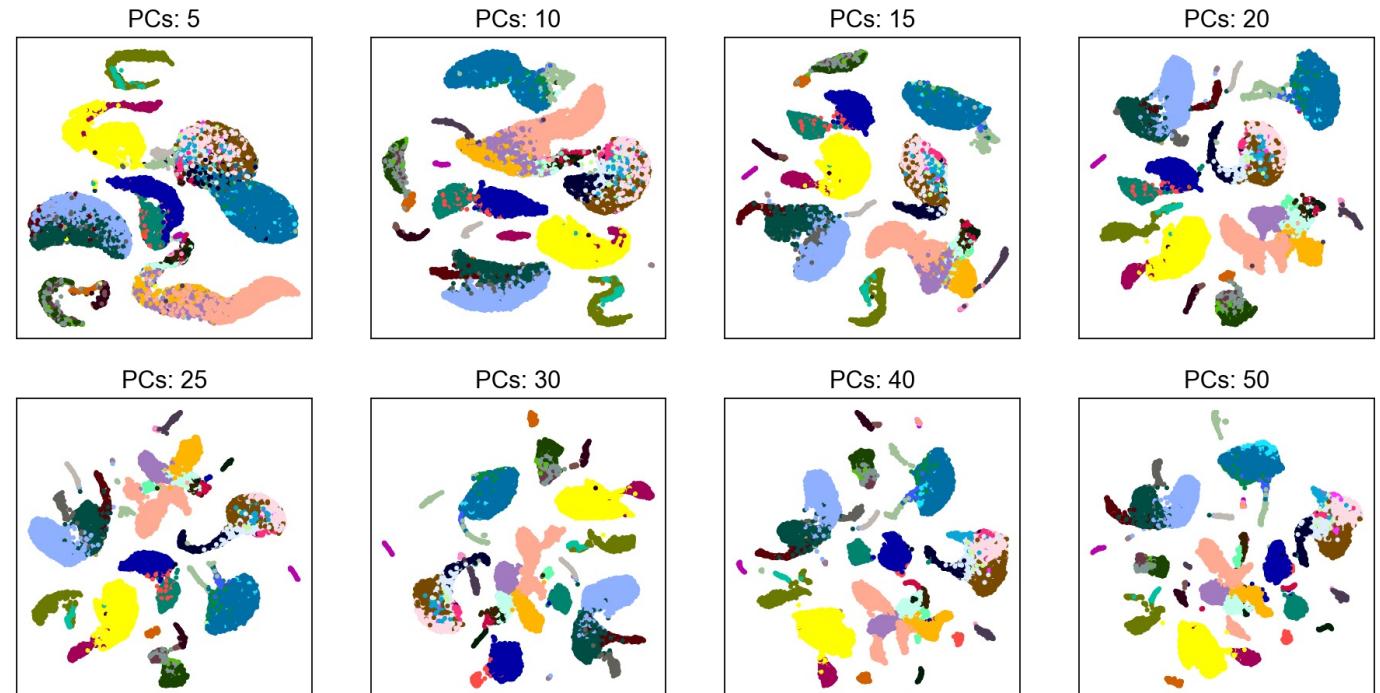
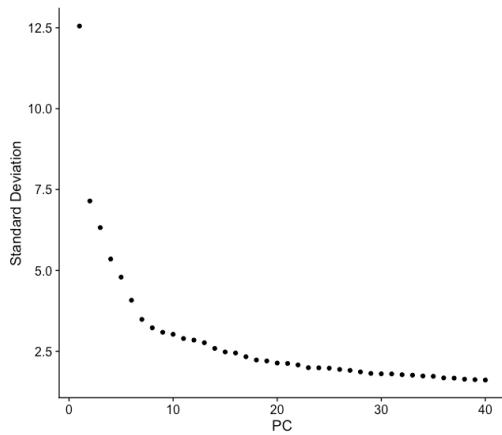
Jackstraw analysis: Determine “significant” PCs

Randomly permute data & recalculate PCs
Compare “Real” PCs with random noise.

- slow
- still subject to weakness of PCA



Option 2: Elbow Plot

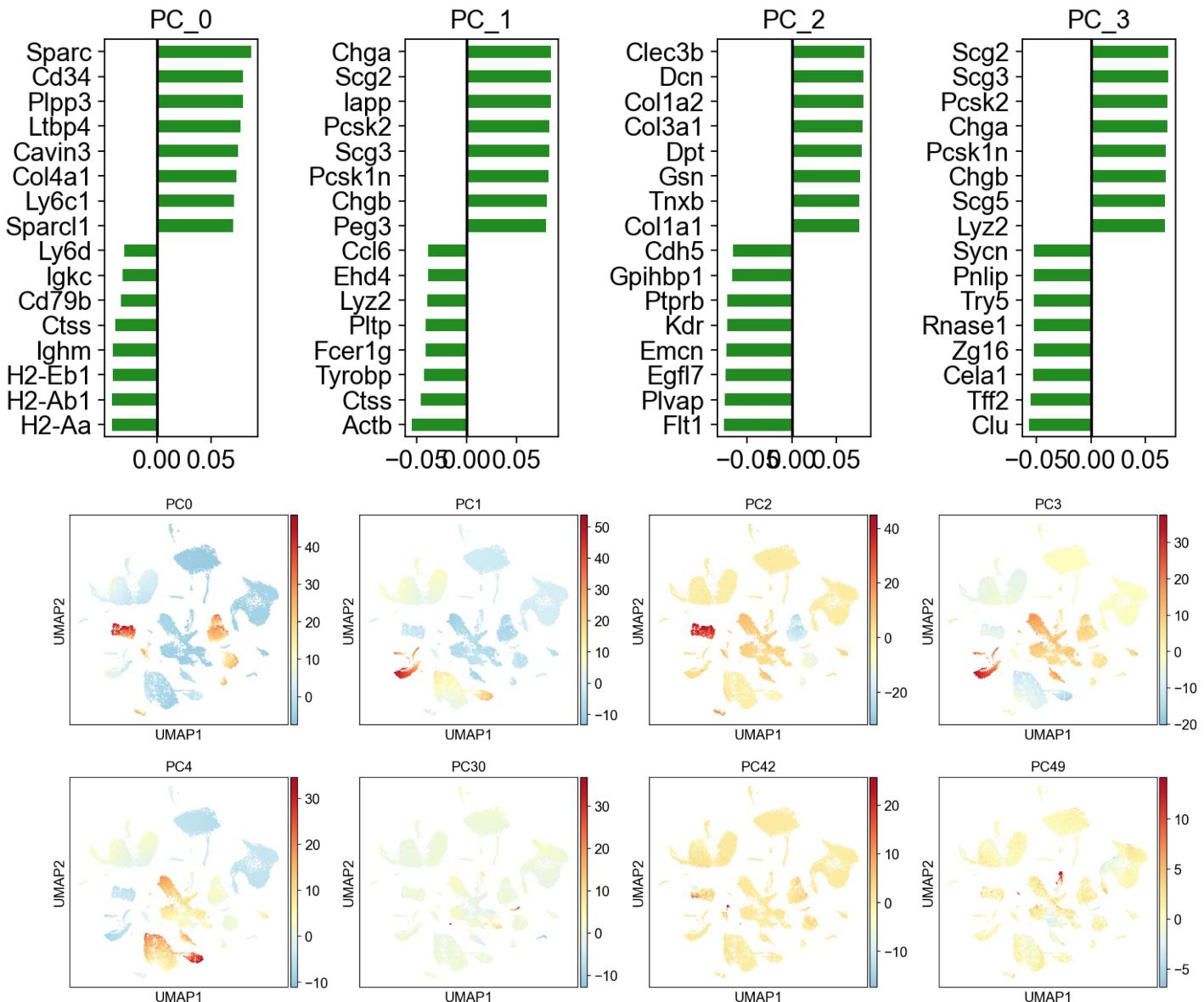
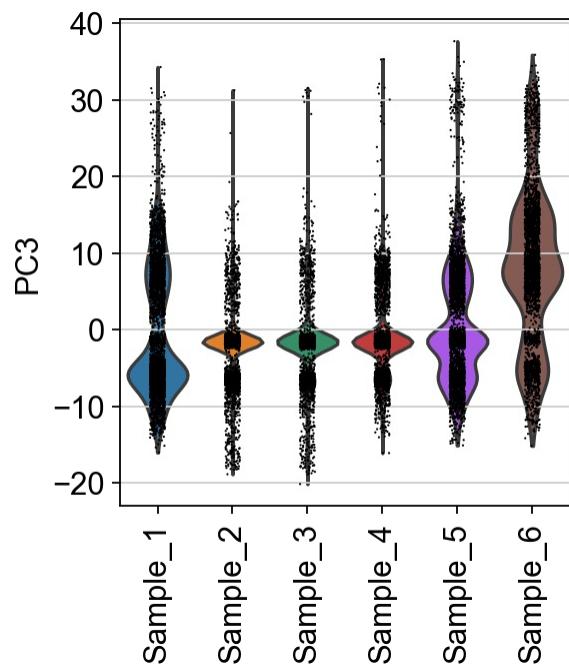


Option 3:

Pick a number that gives you results that you can interpret and defend.

PC “Loadings”

- Semi-interpretable as gene expression signatures
- Check what they contain
- Eg. PC3 here is correlated with Sample → Batch effect?



Other means of Dimensionality Reduction

Linear Transformation

PCA

ICA

Matrix Factorization

NMF

CoGAPS

HPF

Model-based

ZIFA

Ensemble

SIMLR

... any many, many more

Deep Learning

scVI

scVAE

SAUCIE

VASC

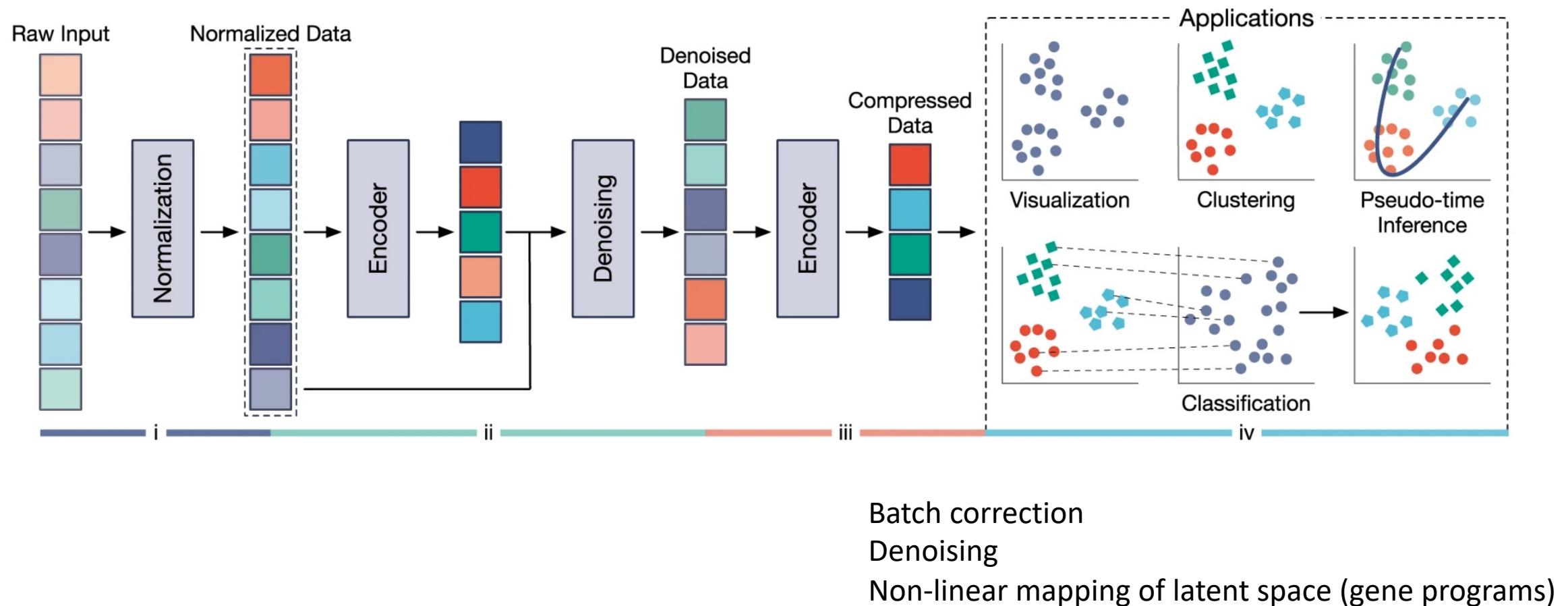
DCA

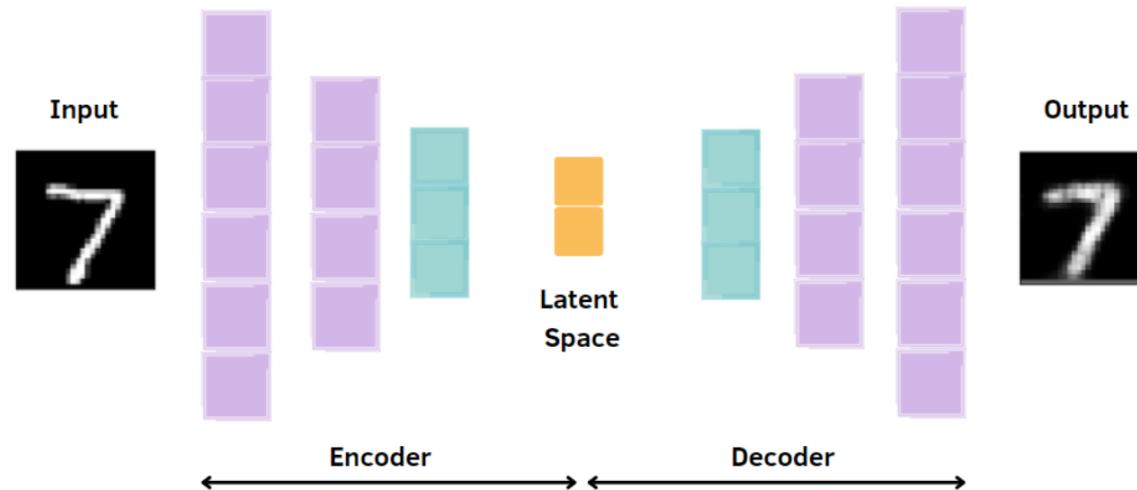
scvis

scSemiGAN

Most methods seek to find a latent / embedding space that is biologically interpretable

Deep Learning methods: Autoencoders

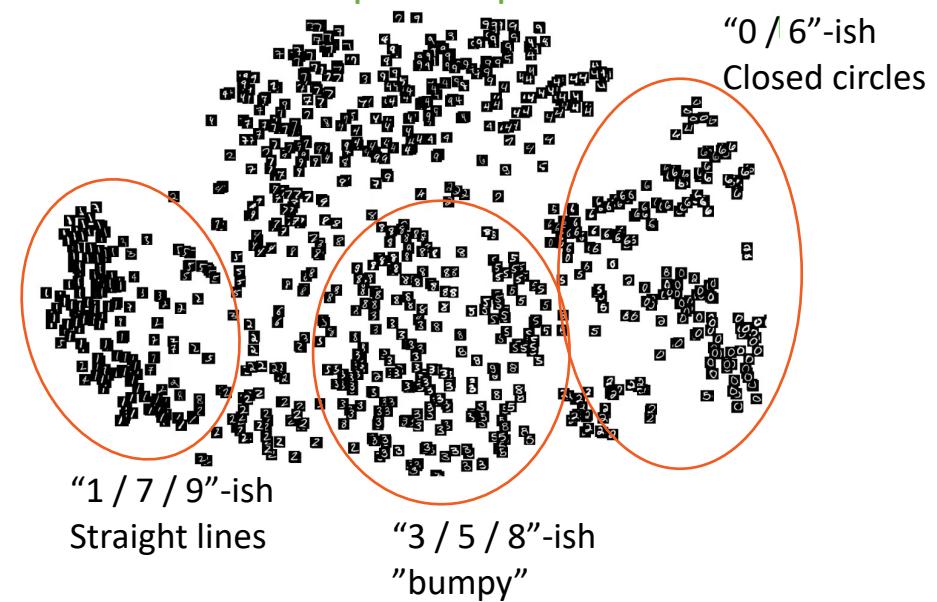




Training set of handwritten numbers

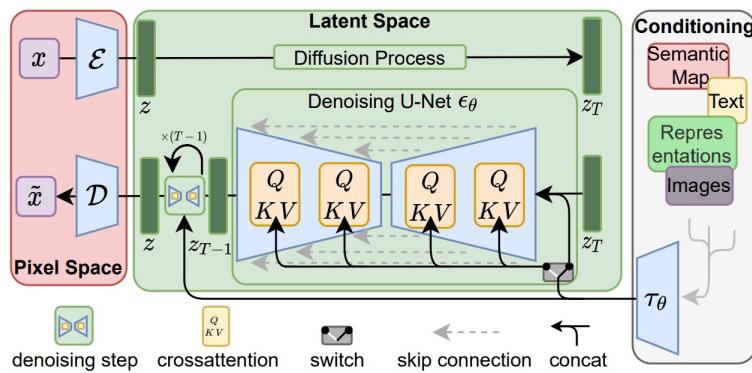
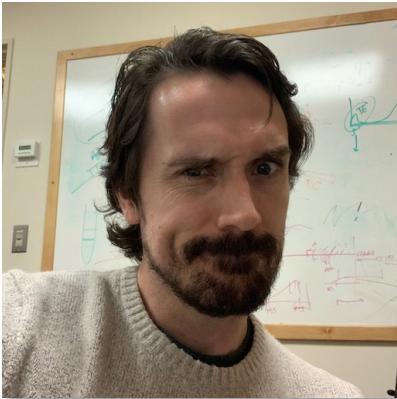


2D Latent-space representation

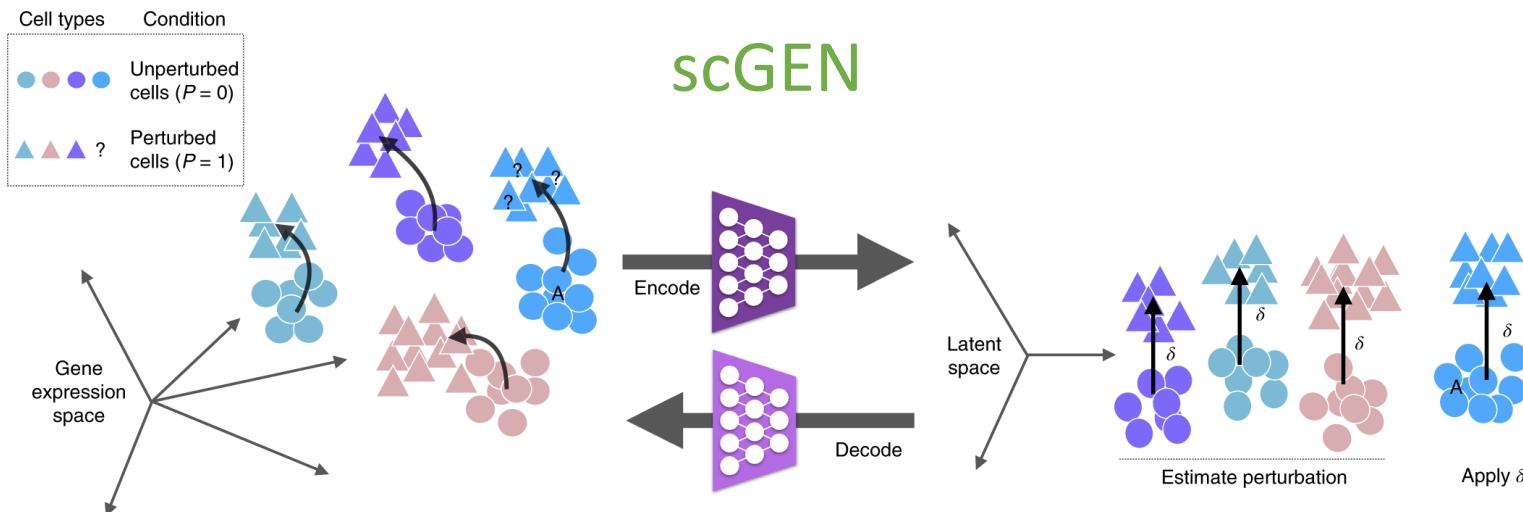


Variational Autoencoders

Stable Diffusion



Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752

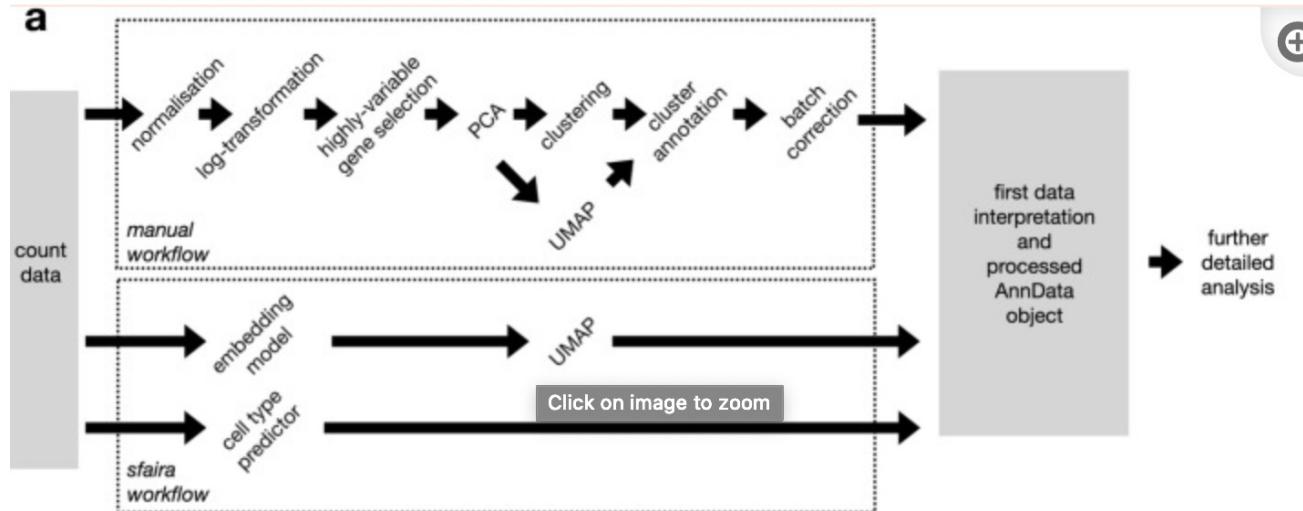


Towards sharable and reproducible embeddings

Using generative, deep-learning based models

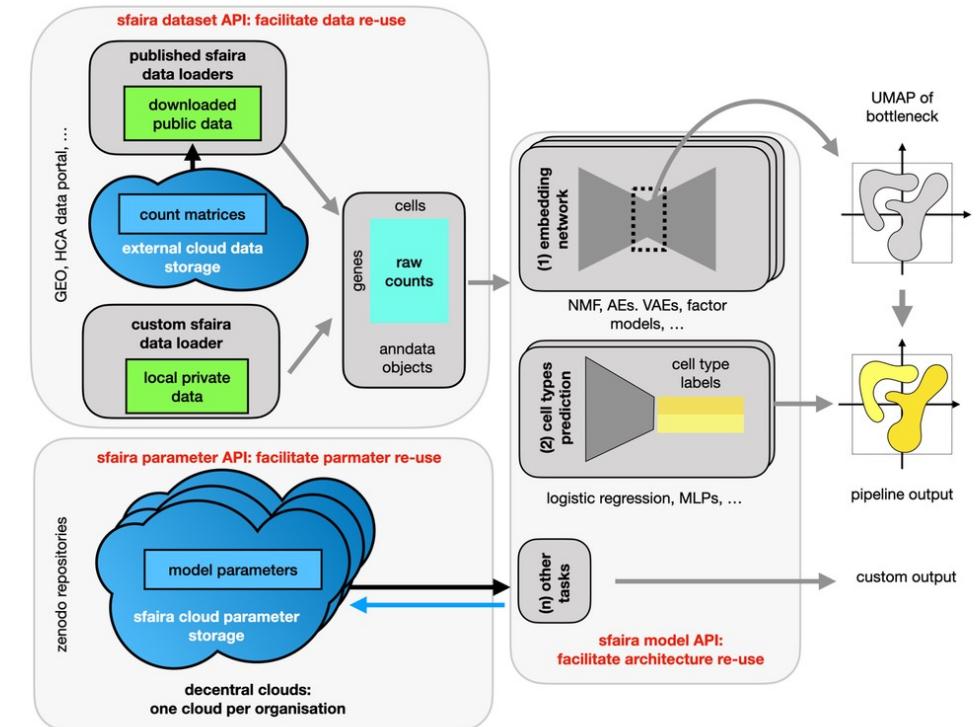
Eg. Sfaira (Theis lab)

Online ‘zoo’ of hundreds of datasets and models



Project your own data into shared embedding
transfer cell type labels

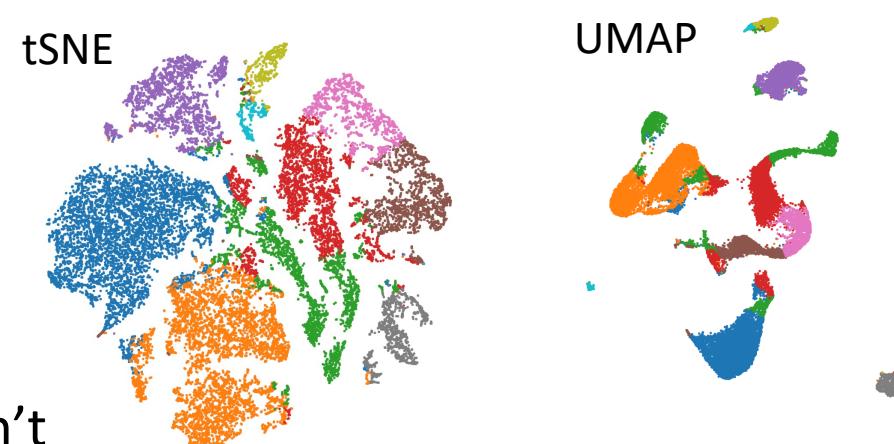
a



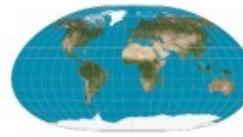
Visualization with tSNE/UMAP

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



No one projection is “correct”

	Lawful	Neutral	Chaotic
Good			
Neutral			
Evil			

Clustering

Generally run on latent space (eg. PC space)

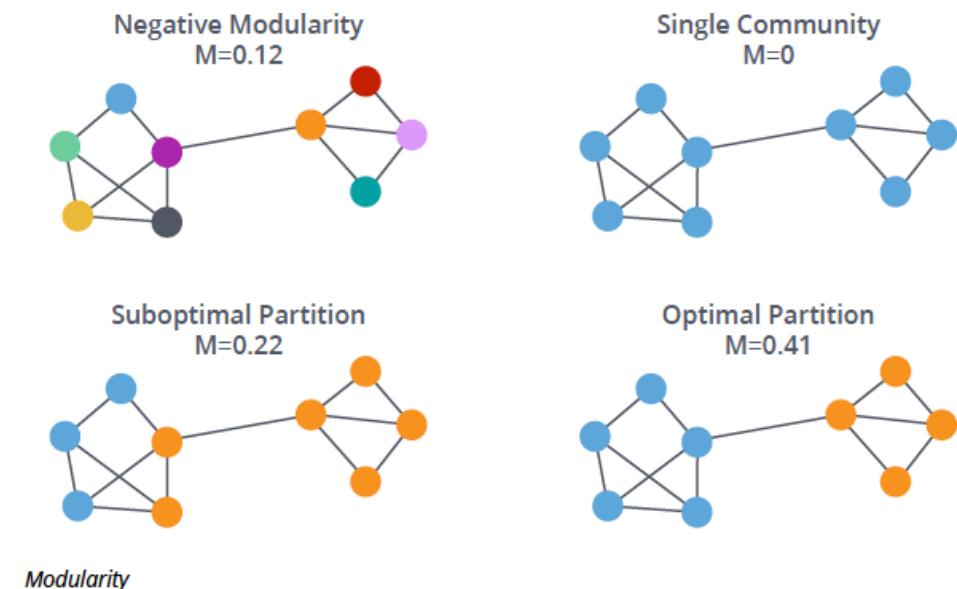
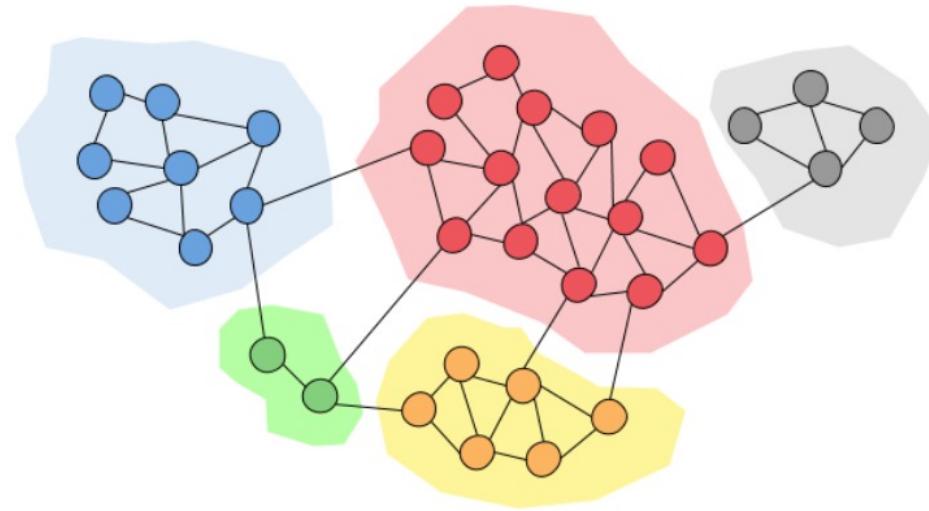
Built into Cell Ranger, most 3rd party tools:

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 (eg. Louvain or Leiden Modularity Optimization)

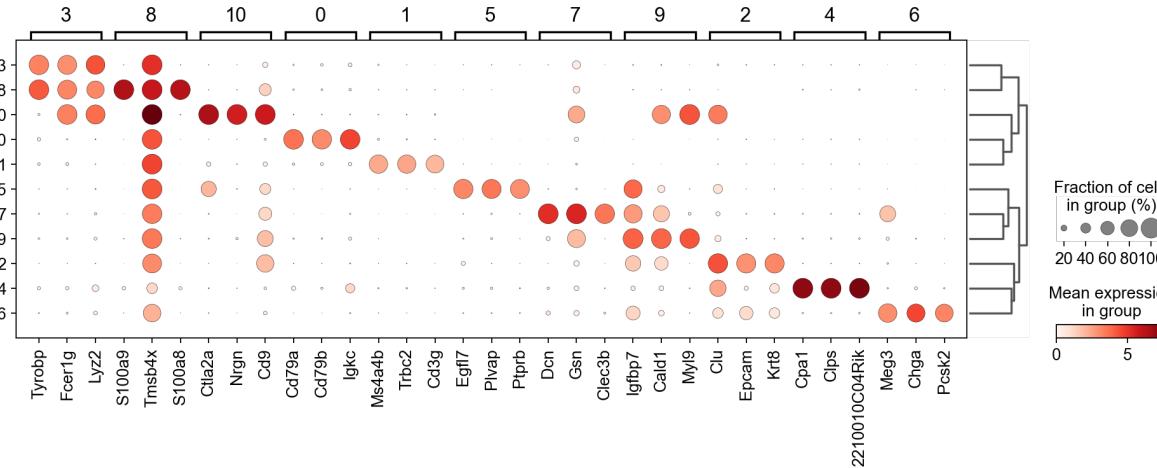
- Compute all pairwise Euclidean distances in latent space
- Trim graph only keep each cells 'K' nearest neighbors* (k
 - "kNN Graph"
- Draw boundaries around "communities" of connected cells to optimize modularity (in-group connections vs out-group connections)
- Variants: Shared Nearest Neighbors, Weighted Nearest Neighbors



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

Clustering Resolution

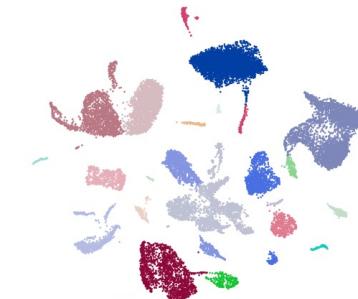
- **Resolution** parameter tunes clustering sensitivity
- Optimal clustering is **subjective**
- Guided sub-clustering may provide best results
 - Ie. **globally** changing resolution parameter may nicely identify meaningful subtypes of one kind of cell, while breaking others into meaningless arbitrary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



Cluster_res_0.02



Cluster_res_0.1



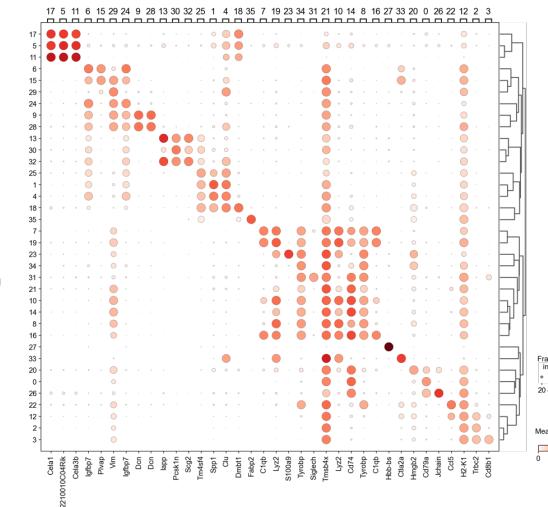
Cluster_res_0.5



Cluster_res_1.0



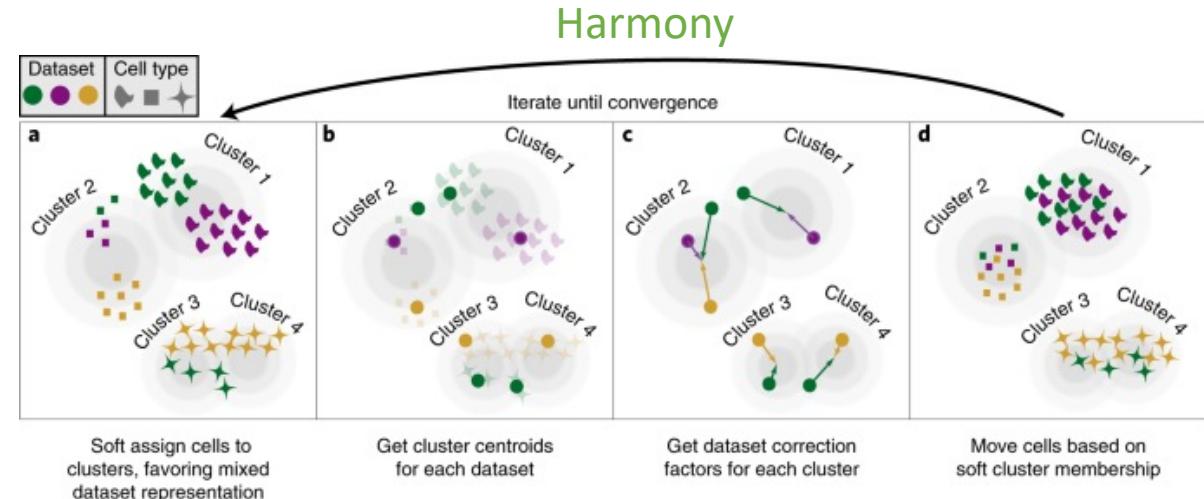
Too Low: miss subtypes, phenotypic states



Batch Correction:

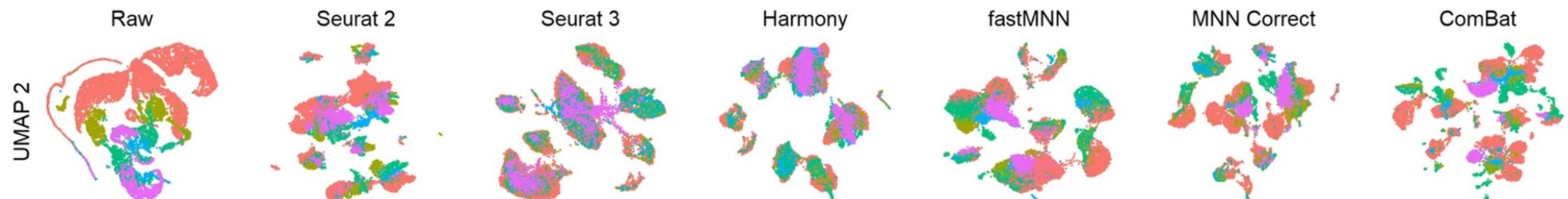
“Batch” corresponds to global differences resulting from:

- Samples run on different days
- Samples run in by different people / labs
- Samples handled slightly differently
- Order of sample processing
- Chemistry differences (kit version, 3' vs 5'), etc.



Good batch correction will find a joint latent space that prioritizes and aligns intra-sample distances

Often only operates on latent dimensions / clustering. Most approaches not directly modify data matrix
ie. often can't be used for differential expression

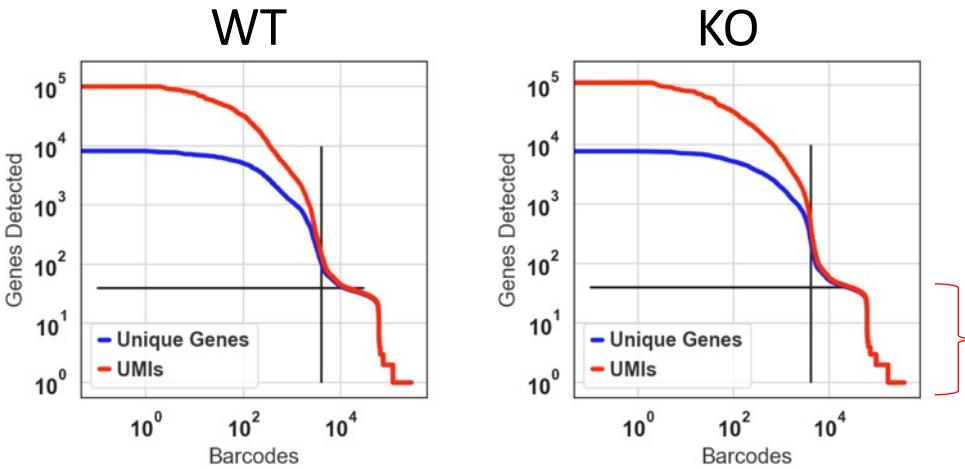


“The best batch correction happens at the bench” – J Preall

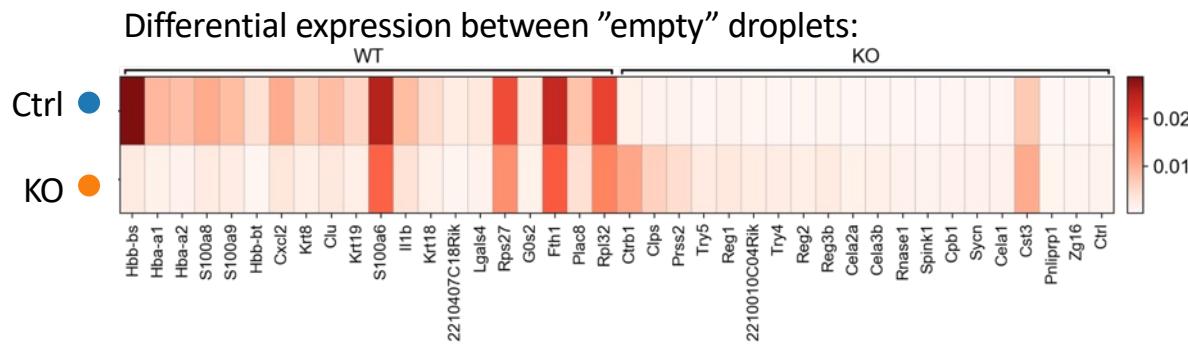
Harmony: Korsunsky, I., Millard, N., Fan, J. et al. Nat Methods 16, 1289–1296 (2019).

Benchmarking: Tran, H.T.N., Ang, K.S., Chevrier, M. et al. Genome Biol 21, 12 (2020). <https://doi.org/10.1186/s13059-019-1850-9>

Controlling for batch effects



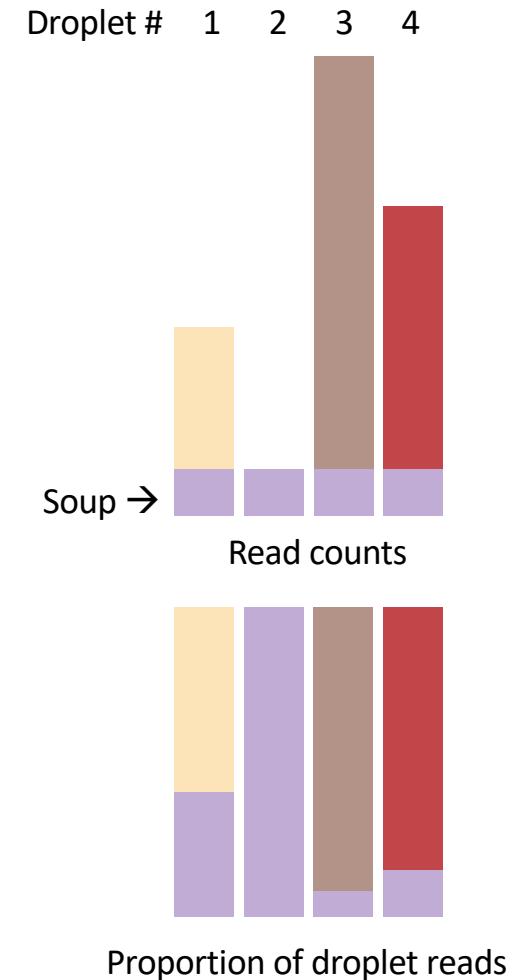
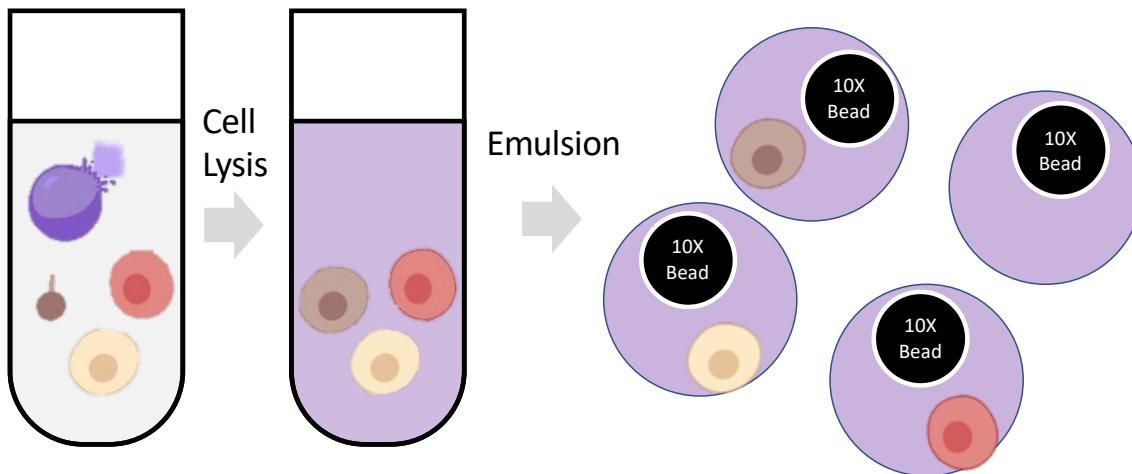
Ambient RNA
in droplets



Significant sources of
contaminating mRNA:

WT:	Erythrocytes
WT:	Epithelial
WT:	Granulocytes
KO:	Acinar cells

Ambient RNA: "SOUP"

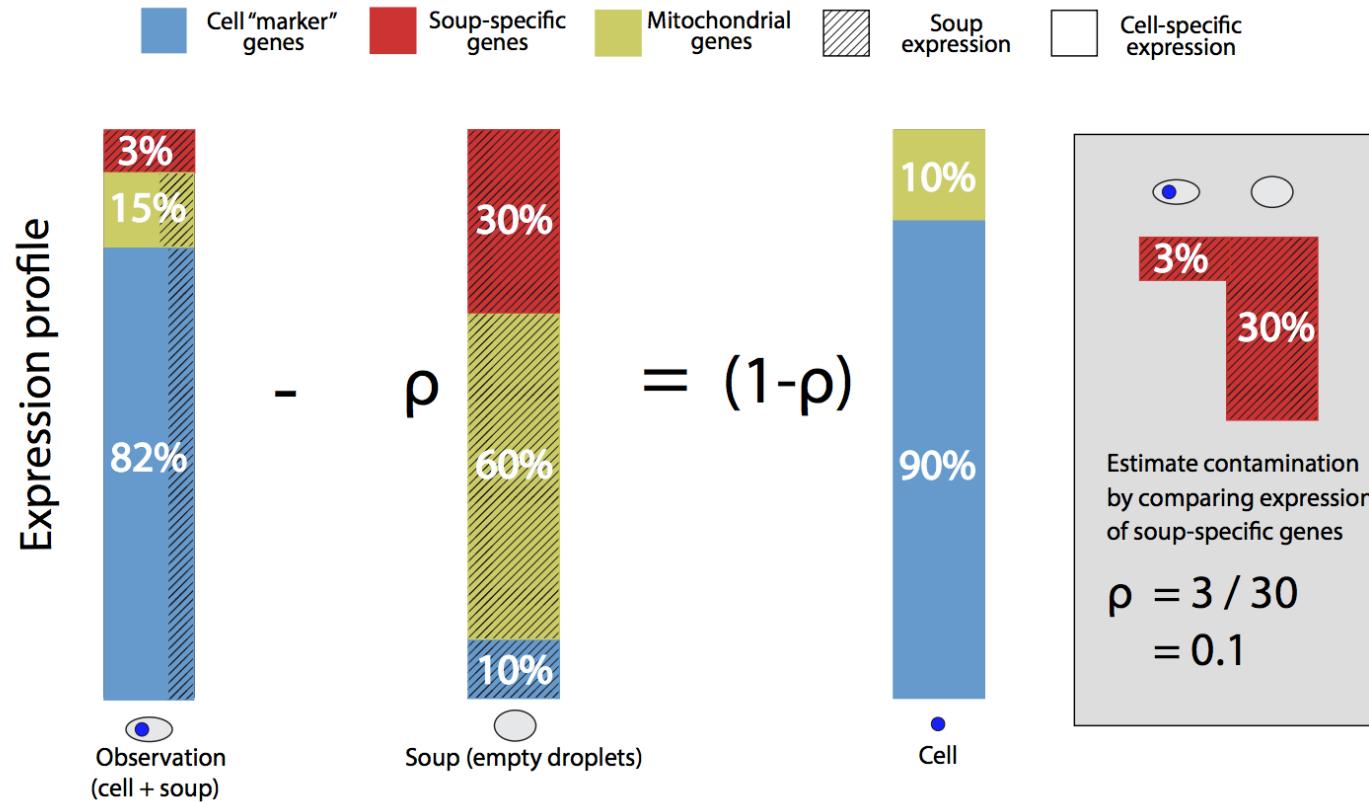


Computational methods to mitigate (but not totally fix) Soup:

- *SoupX*
- *DecontX*
- *Cellbender*

The best time to fix Ambient RNA problems is at the bench!

SoupX



Young and Behati (bioRxiv) 2018.

<https://www.biorxiv.org/content/10.1101/303727v1>

METHOD

Open Access

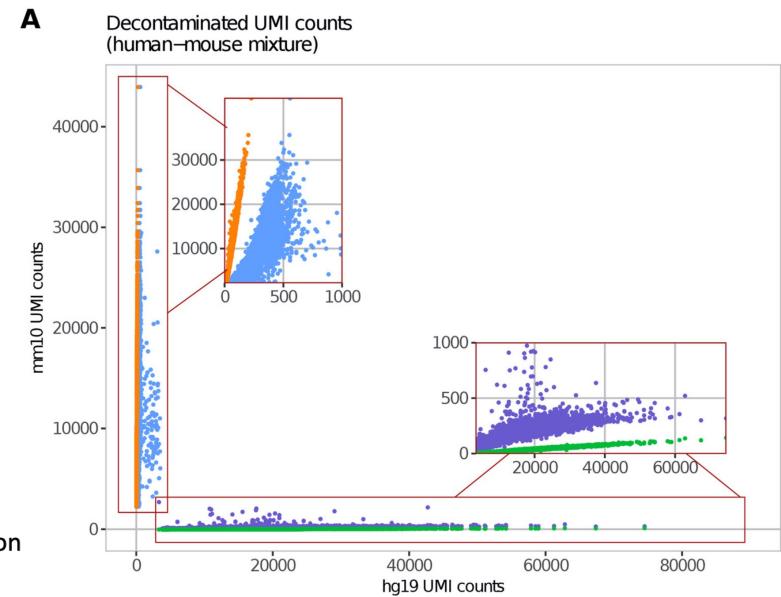
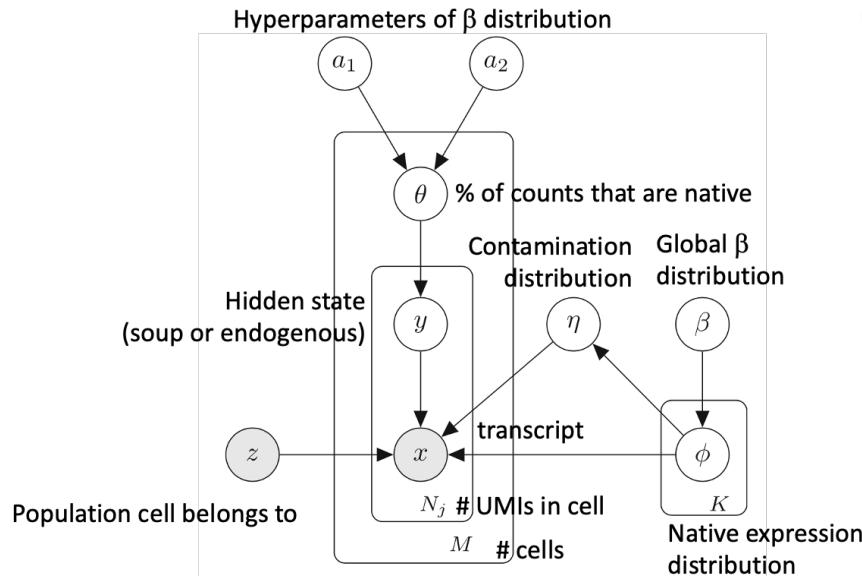


Decontamination of ambient RNA in single-cell RNA-seq with DecontX

Shiyi Yang¹ , Sean E. Corbett¹, Yusuke Koga¹, Zhe Wang¹ , W Evan Johnson¹, Masanao Yajima² and Joshua D. Campbell^{1*}

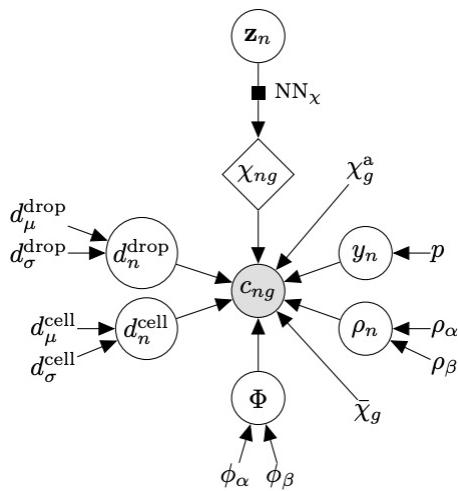
Uses Variational Bayes Inference
Similar to Latent Dirichlet Allocation (LDA)

Models soup as a weighted combination of other cell types in the population



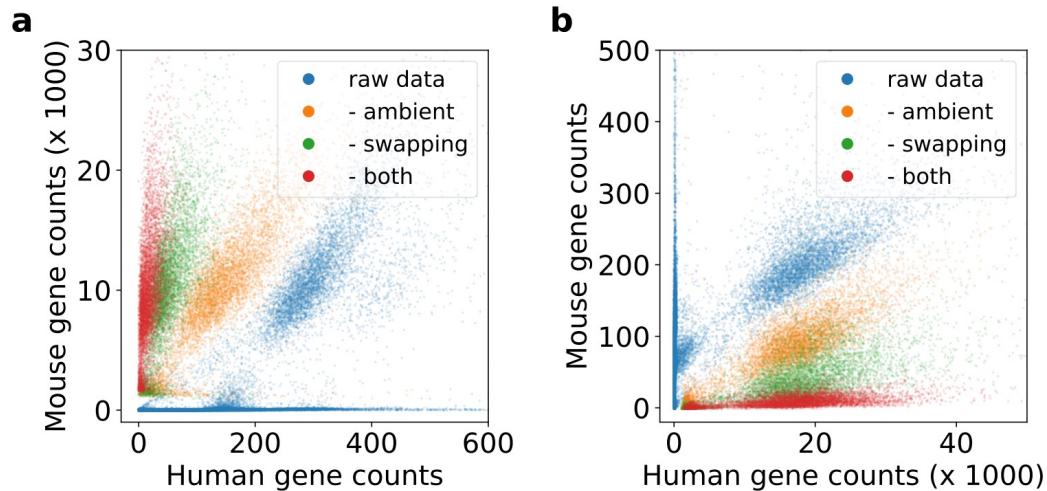
CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets

Stephen J. Fleming^{1,2}, John C. Marioni^{3,4}, and Mehrtash Babadi^{1,2}



Variational Inference method
GPU-optimized

Corrects 2 problems:
1. Soup / Ambient RNA
2. Barcode swapping during PCR



Getting started with your own analyses

Rahul Satija -

R

<https://satijalab.org/seurat>

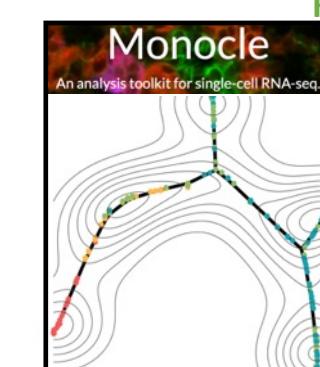


Fabian Theis - München



<https://scanpy.readthedocs.io/en/latest/>

Python



Cole Trapnell – WashU

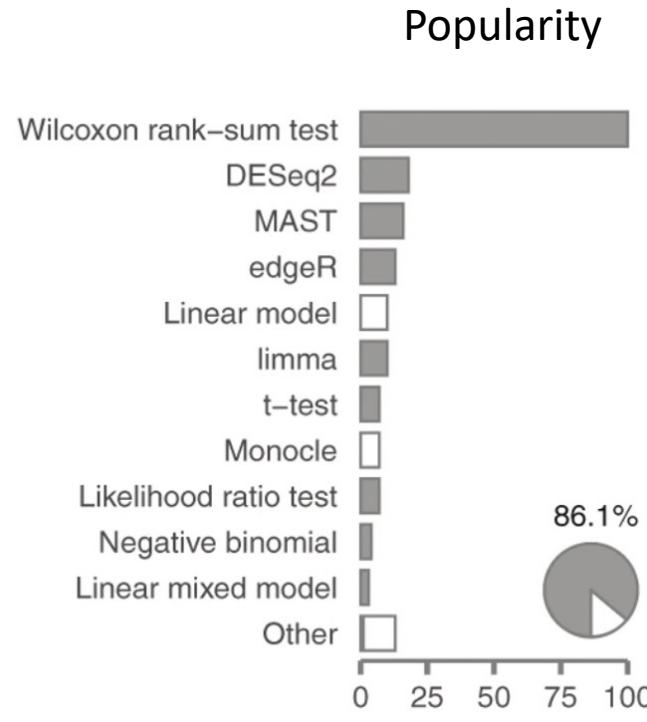


Macosko lab

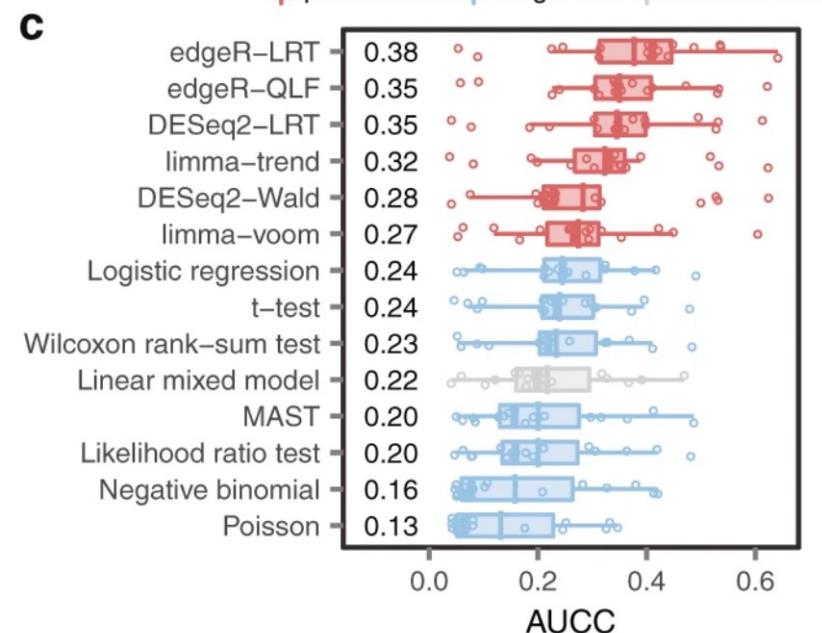
AWESOME SINGLE CELL RESOURCE

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

Differential Gene Expression



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

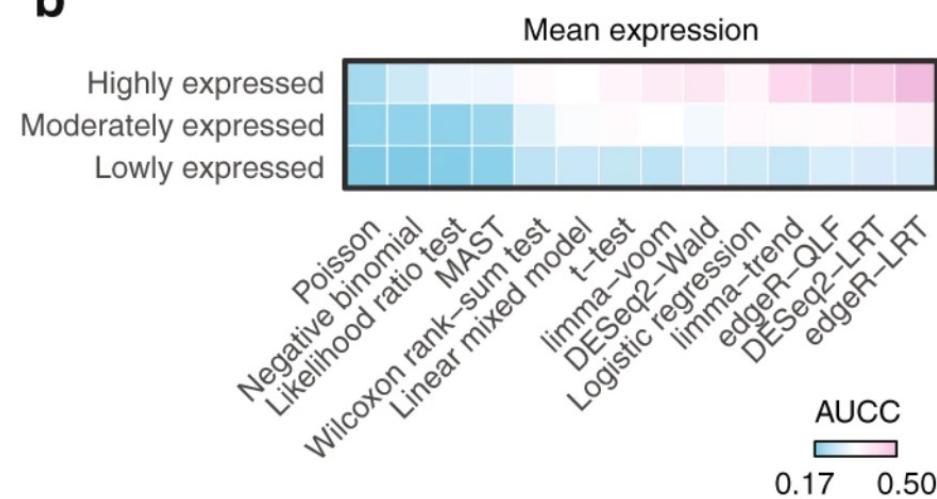
Binomial (low cell counts)

EdgeR NB (high cell counts)

Differential Gene Expression

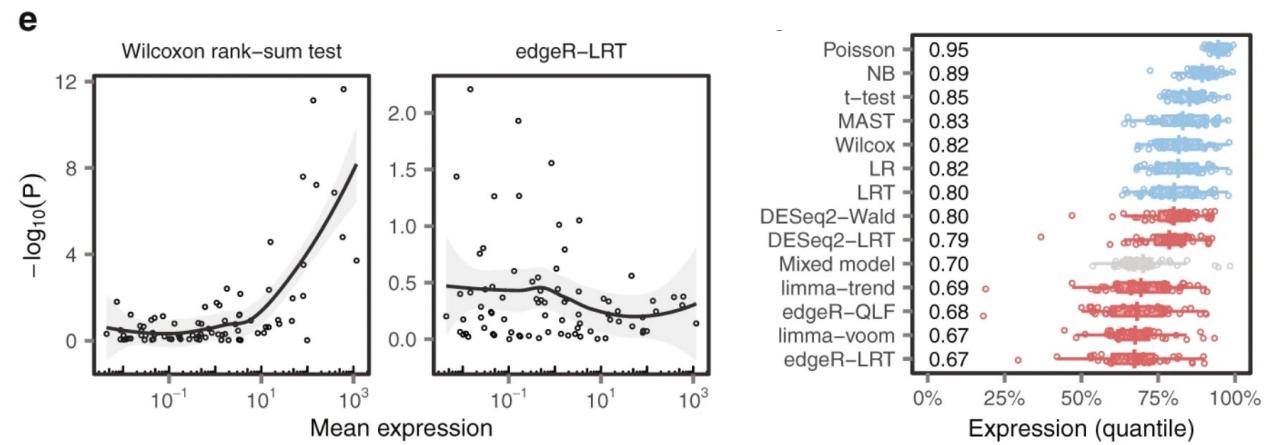
b

Single-cell methods biased towards highly expressed genes



Pseudobulk methods make fewer false discoveries of highly expressed genes

(ie. 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	[22]
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]

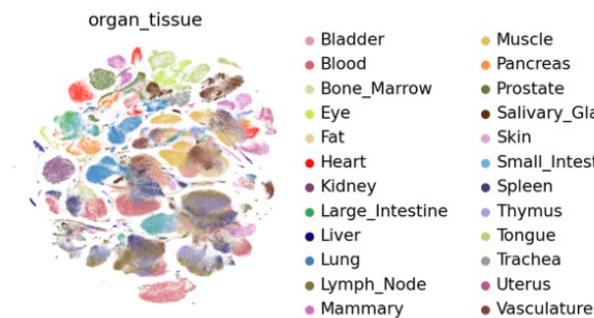
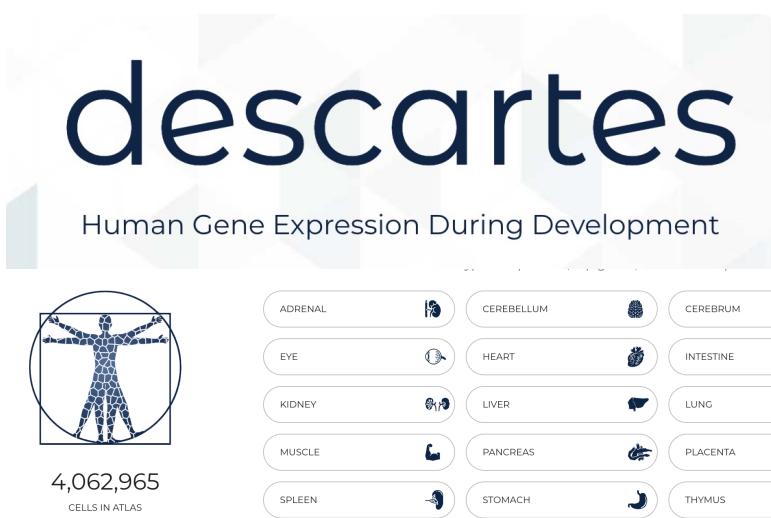
- A **hard** problem
- What even is a cell type?
- Who is curating these?
- Cell types versus states
 - normal vs disease / perturbed
- What type of model is being used?
 - Marker-based
 - Reference dataset label transfer



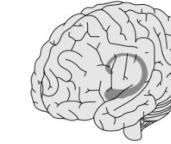
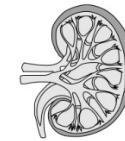
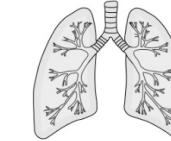
<https://www.scrna-tools.org/tools?sort=name&cats=Classification>

MANY more tools available

Atlas Efforts building Cell Type Classifiers



OnClass
Single cell typing based on cell ontology.
[Paper](#) [Read the Docs](#) [GitHub](#)

**Blood** [UMAP \(CELLxGENE\)](#) [Tree plot](#) [Download data](#) [Download models](#)**Bone marrow** [UMAP \(CELLxGENE\)](#) [Tree plot](#) [Download data](#) [Download models](#)**Heart** [UMAP \(CELLxGENE\)](#) [Tree plot](#) [Download data](#) [Download models](#)**Hippocampus** [UMAP \(CELLxGENE\)](#) [Tree plot](#) [Download data](#) [Download models](#)**Intestine****Kidney****Liver****Lung**

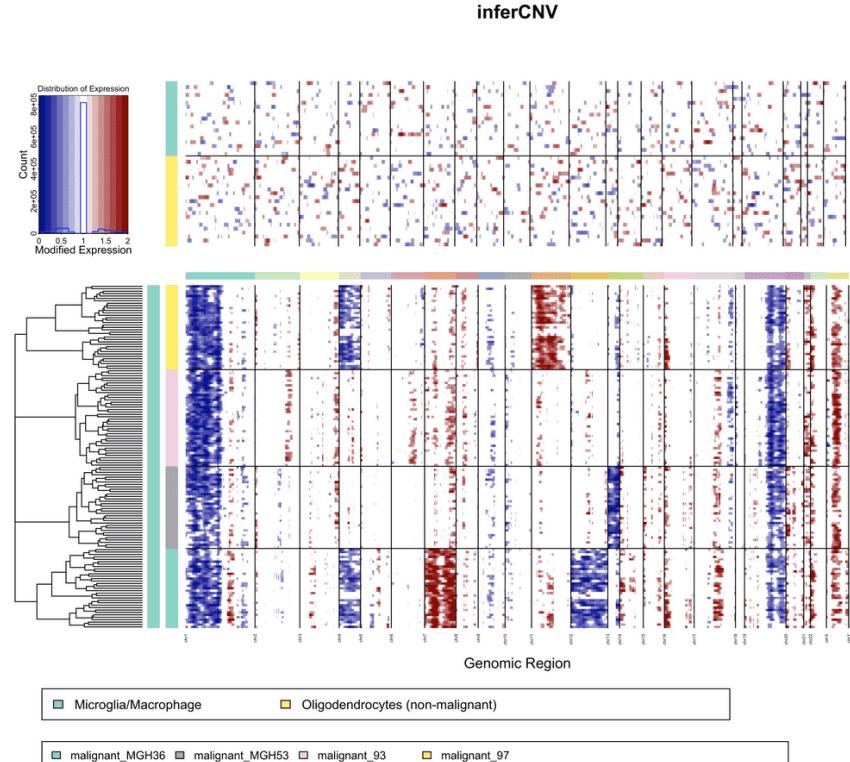
Going Beyond Gene Expression

- Copy Number Inference:
 - InferCNV: Tirosh et al (2016) *Science* doi: [10.1126/science.aad0501](https://doi.org/10.1126/science.aad0501)
 - KopyKAT: Gao R et al. (2021) *Nat Biotechnol.* doi: [10.1038/s41587-020-00795-2](https://doi.org/10.1038/s41587-020-00795-2)
- Alternative TSS Utilization:
 - scRCAT-seq: Hu Y et al (2020) *Nat Commun.* [10.1038/s41467-020-18976-7](https://doi.org/10.1038/s41467-020-18976-7)
- Alternative Polyadenylation:
 - scAPA: Shulman et al (2019) *Nucleic Acids Res.* doi: [10.1093/nar/gkz781](https://doi.org/10.1093/nar/gkz781)
 - Sierra: Patrick et al (2021) *Genome Biology*. doi: [10.1186/s13059-020-02071-7](https://doi.org/10.1186/s13059-020-02071-7)
- eQTL Analysis
 - Kang H et al. (2018) *Nat Biotechnol.* doi: [10.1038/nbt.4042](https://doi.org/10.1038/nbt.4042)
 - van der Wijst et al (2018) *Nat Genet.* doi: [10.1038/s41588-018-0089-9](https://doi.org/10.1038/s41588-018-0089-9)
 - Neavin et al (2021) *Genome Biology*. doi: [10.1186/s13059-021-02293-3](https://doi.org/10.1186/s13059-021-02293-3)
- Alternative Splicing
 - Various methods involving long-read sequencing

CNV inference

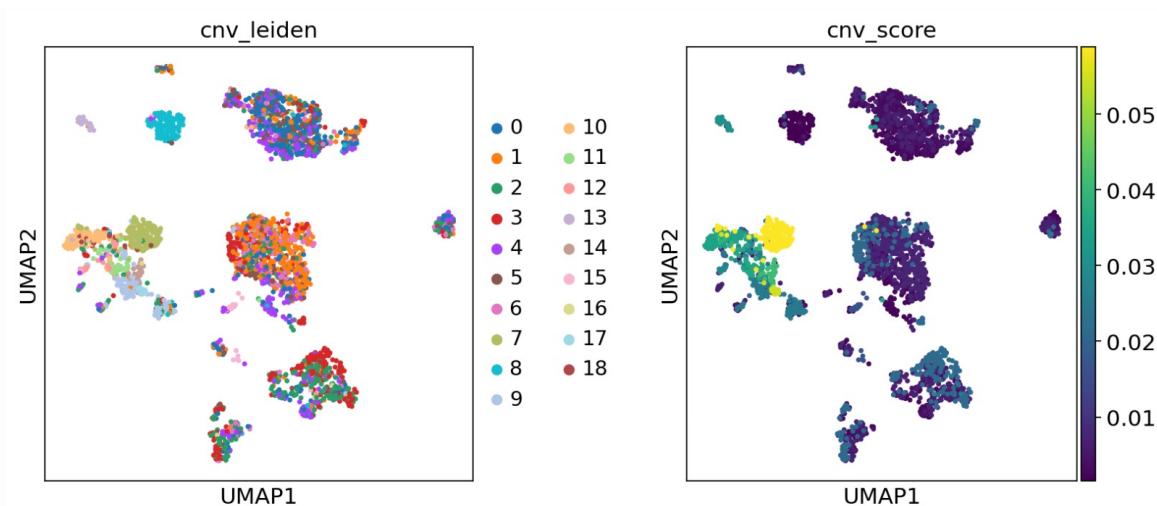
Eg. inferCNV, CaSpER, CopyKAT, SCYN

- Counts binned expression data across chromosomes
- Builds 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



Transcriptional Dynamics: RNA Velocity

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

Quantifies spliced / unspliced

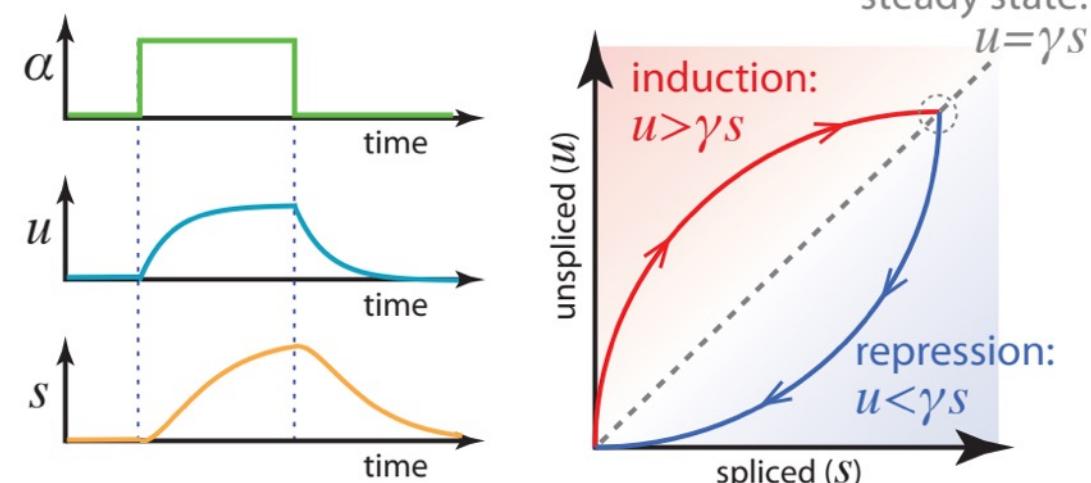
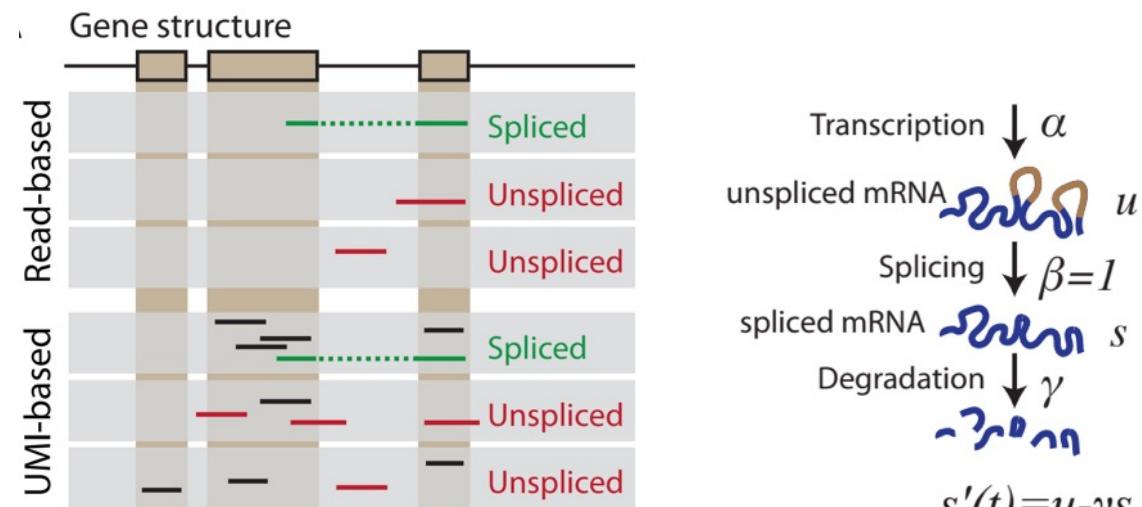
Models dynamics

CAVEATS: Gene annotations

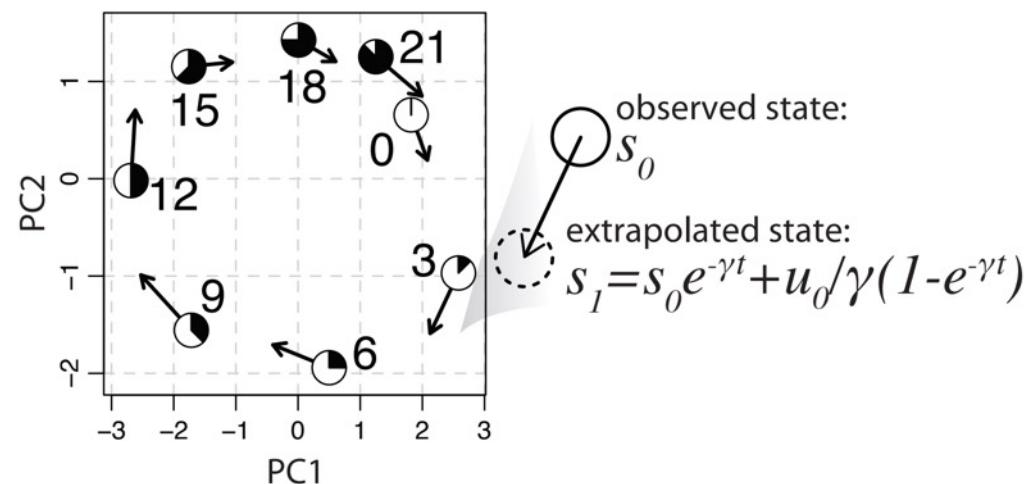
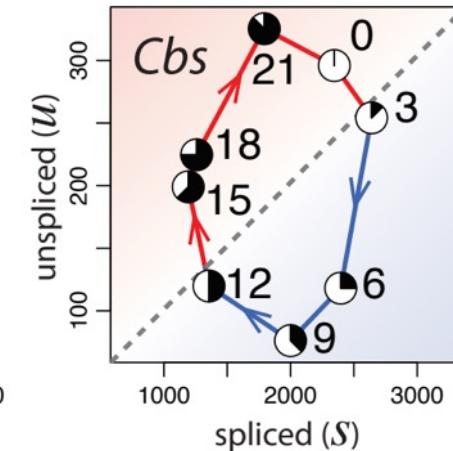
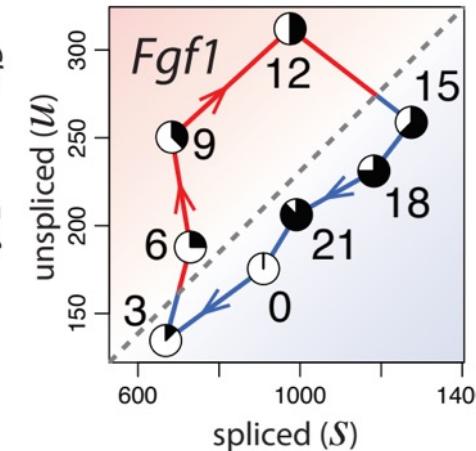
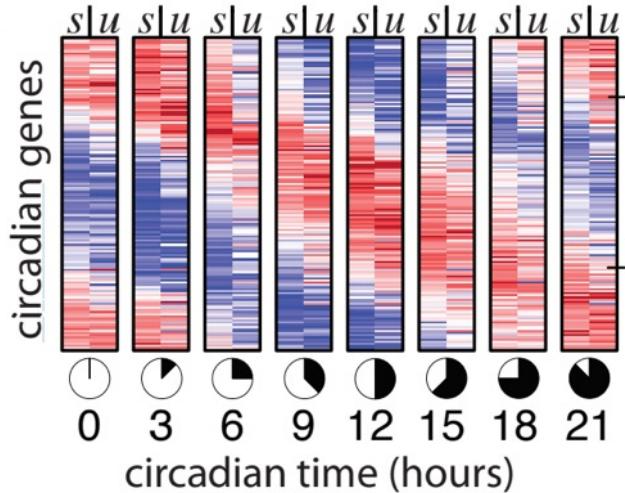
Cryptic exons

unannotated intronic genes

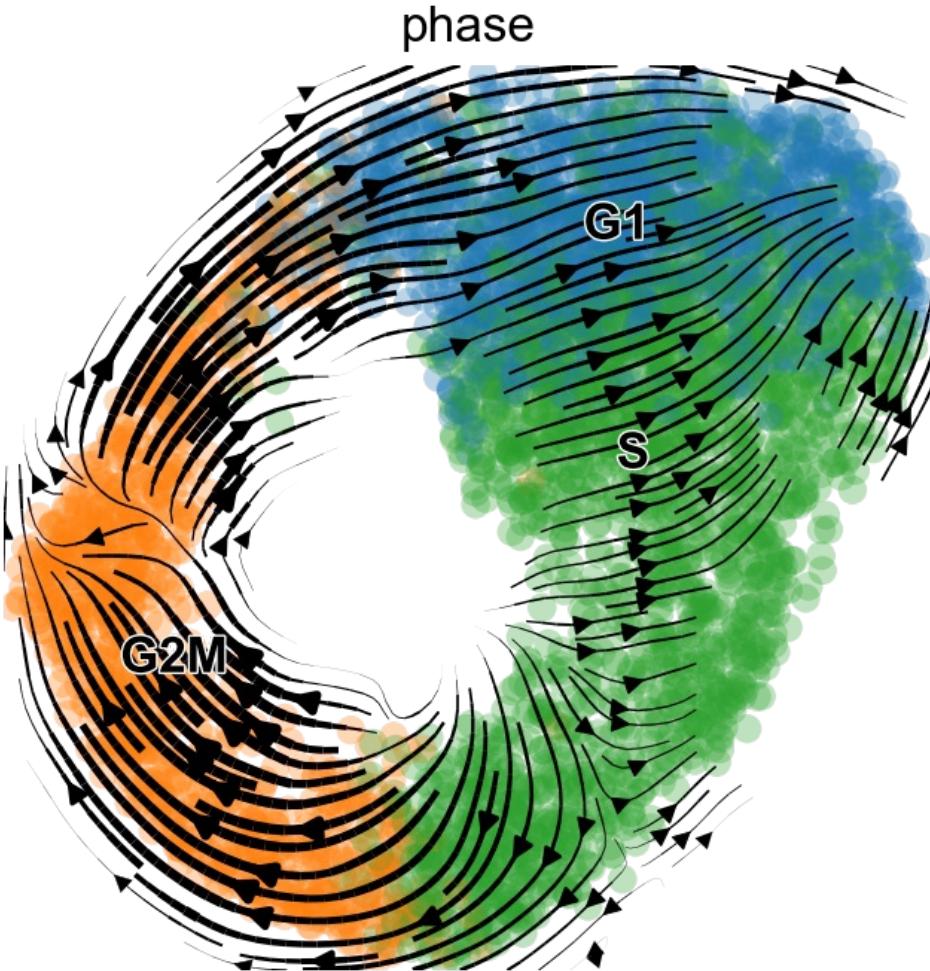
repetitive elements



Bulk RNAseq from mouse circadian rhythm data



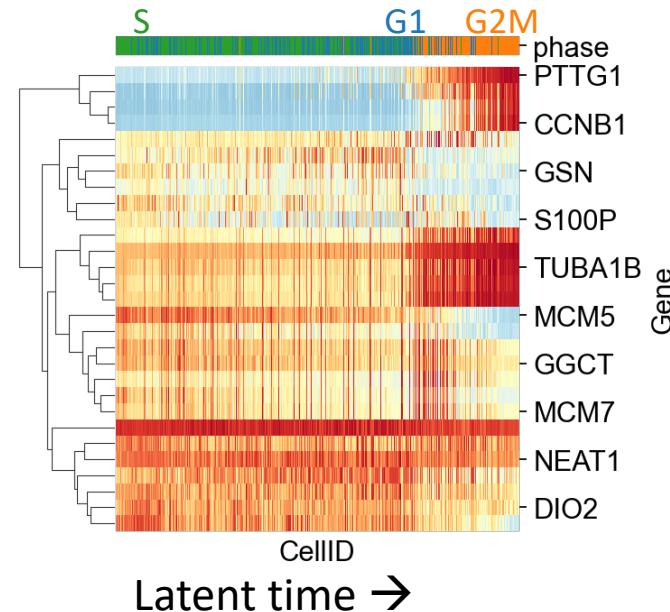
Cultured Cells – RNA Velocity



Siha (cervical epithelium cells)

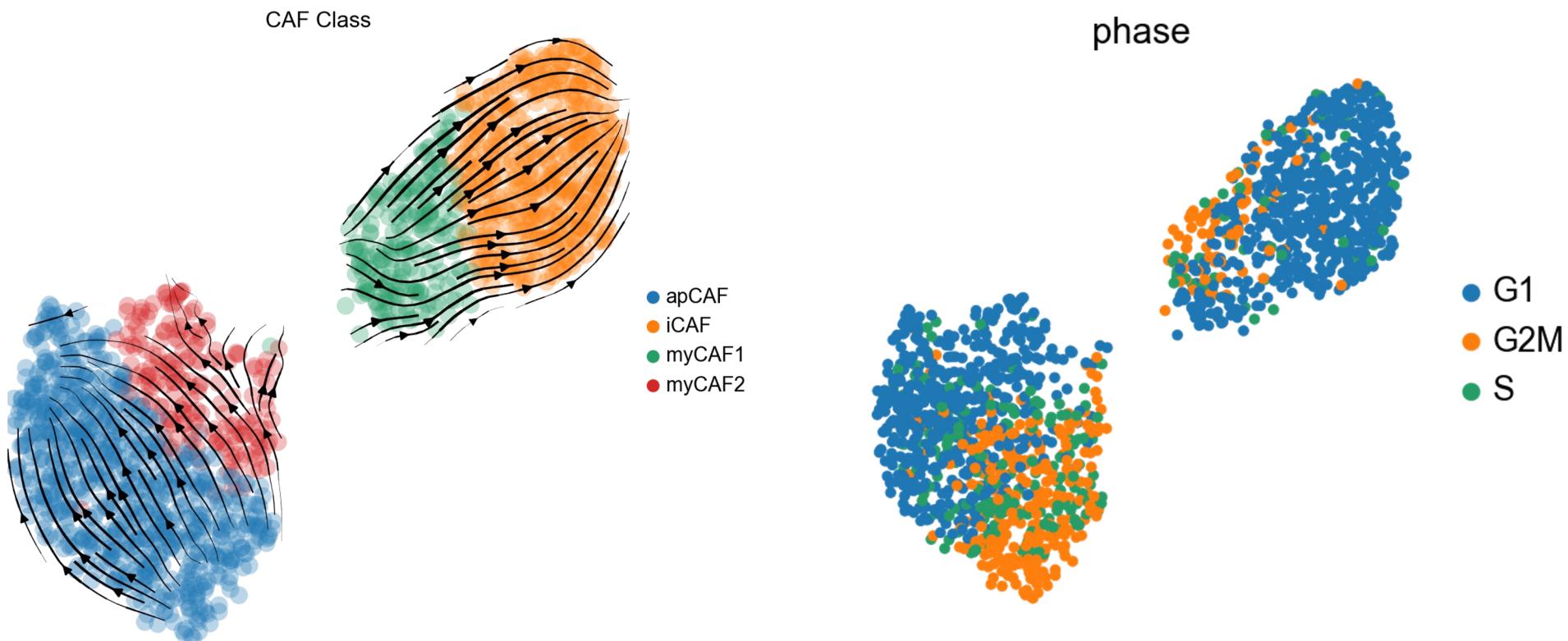
Projected on to subspace using only cell-cycle relevant genes

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



Caveat:

Don't Confuse Developmental Trajectories with Cell Cycle Kinetics



SCENIC single-cell regulatory network inference and clustering

Infers Transcription Factor Activity

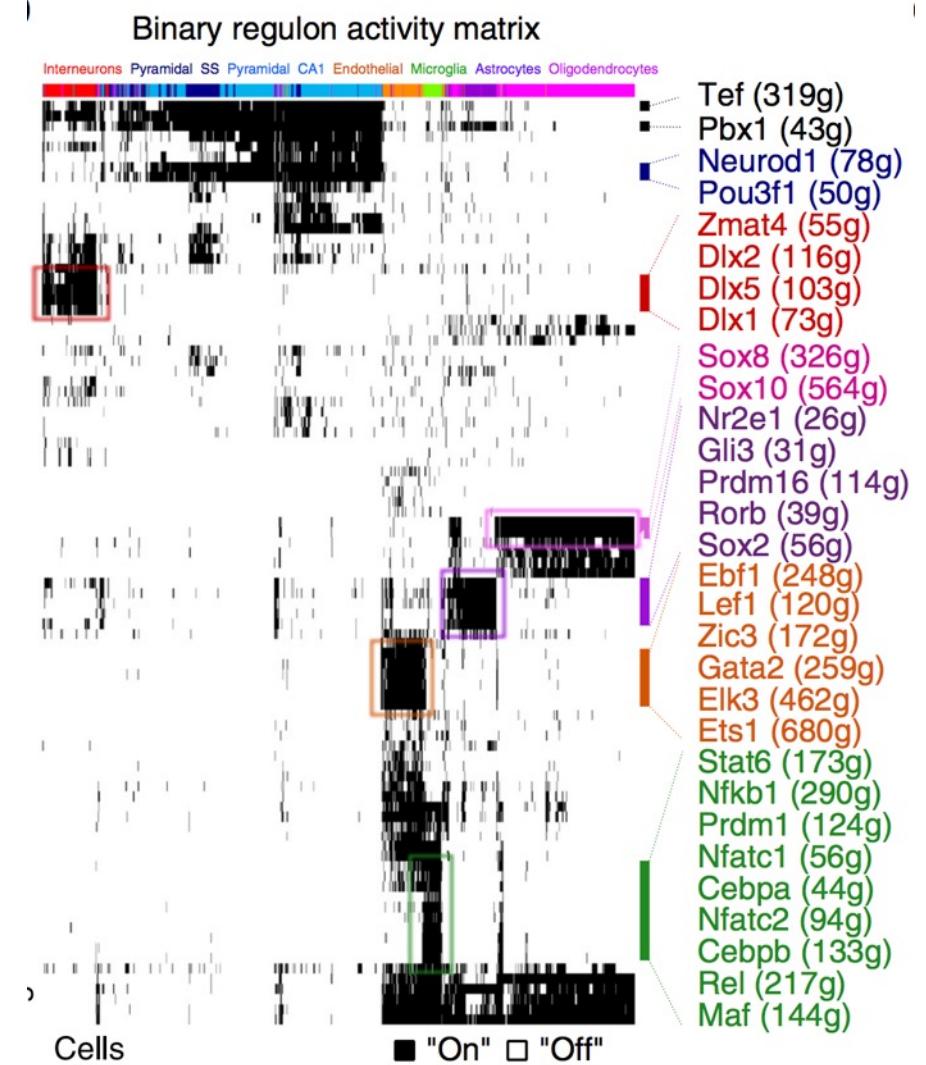
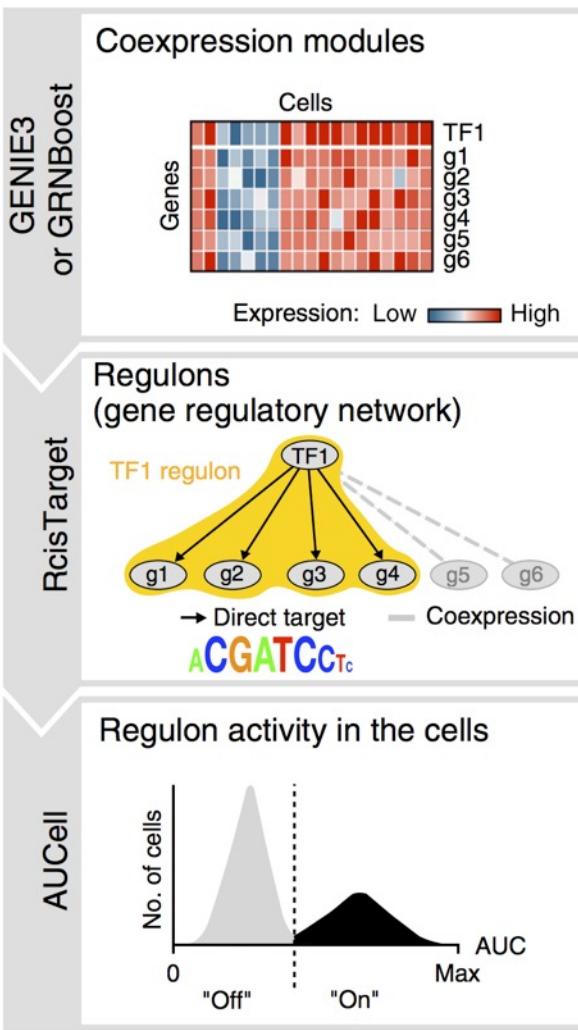
Gene Co-expression network



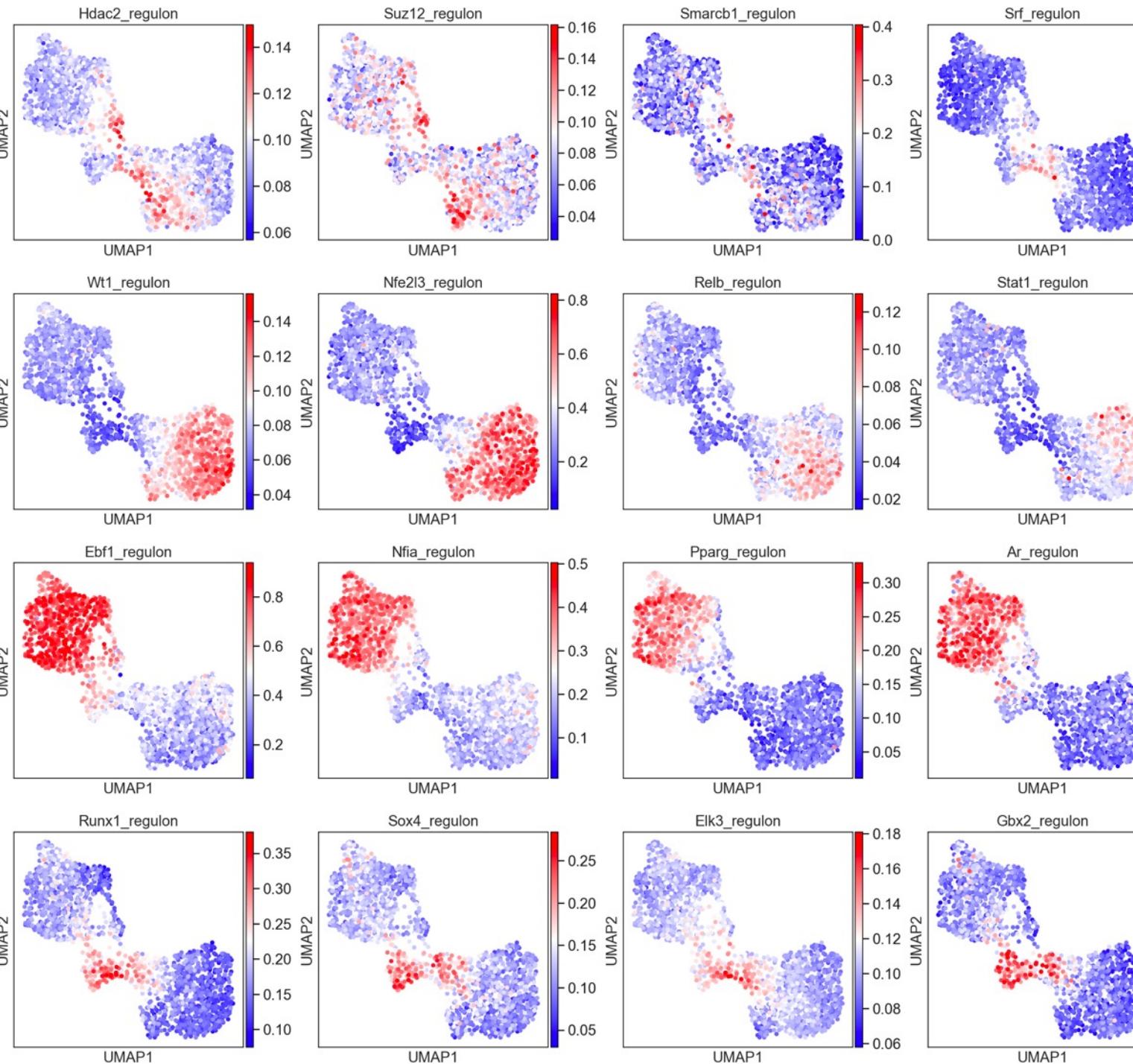
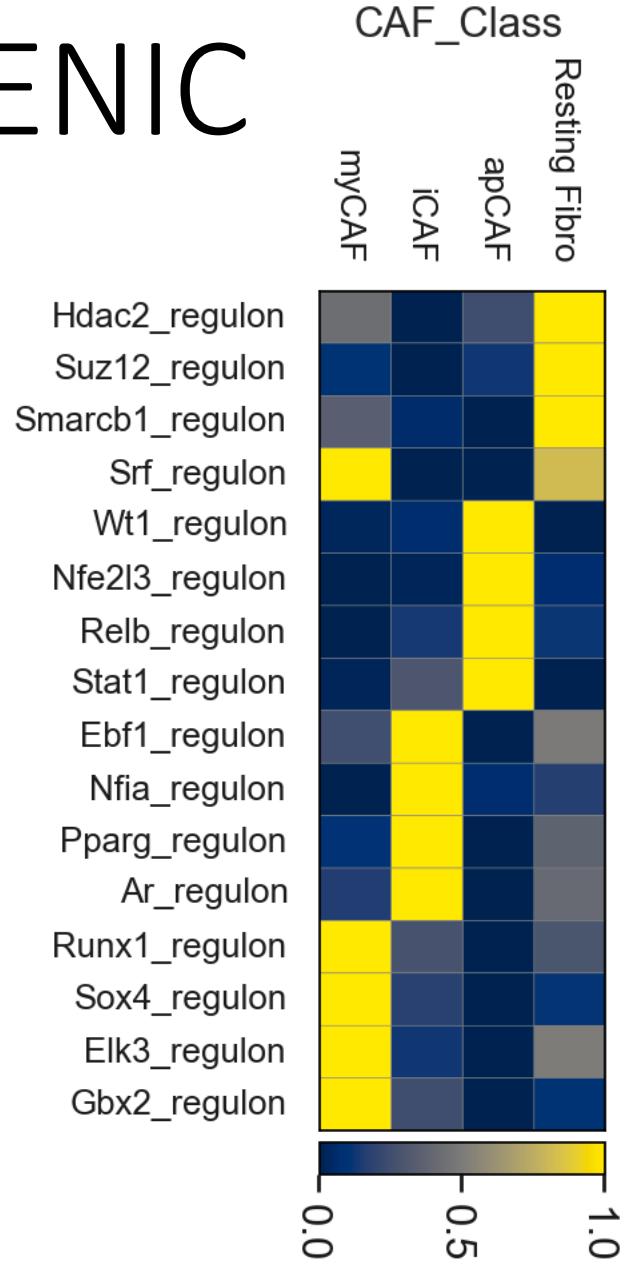
Motif search



Regulon activity



SCENIC





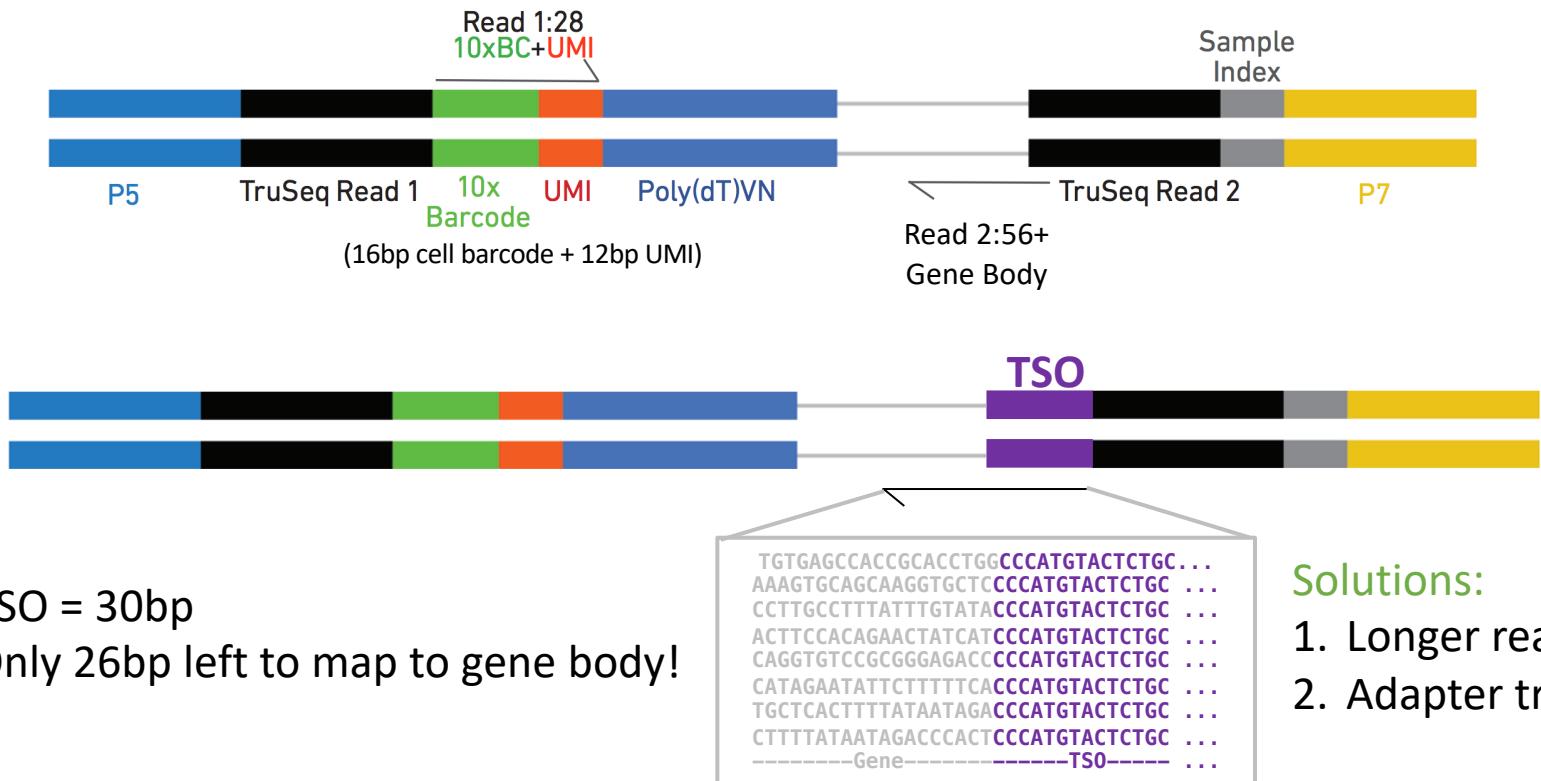
Problem with mapping rates:

Mapping	?
Reads Mapped to Genome	62.4%
Reads Mapped Confidently to Genome	56.0%

Unmapped reads

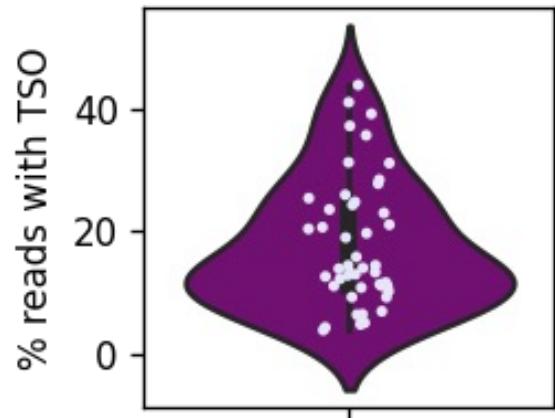
NB501555:820:HV3HNBGXG:3:23410:18728:10576 AAGCAGTGGTATCAACGCAGAGTACATGGGCCAGTTACTAAGATATCATTCCAGA
NB501555:820:HV3HNBGXG:2:11108:26684:1988 AAGCAGTGGTATCAACGCAGAGTACATGGGTGTTAGGATAGATCTAACGAGTAATC
NB501555:820:HV3HNBGXG:1:21106:4541:17390 AAGCAGTGGTATCAACGCAGAGTACATGGGTGCAAACAAAGATGTTAAGGTATGCT
NB501555:820:HV3HNBGXG:2:12211:18559:6225 AAGCAGTGGTATCAACGCAGAGTACATGGGTGGATCTCCTGTATTCCATGCATACT
NB501555:820:HV3HNBGXG:2:13201:21902:7842 AAGCAGTGGTATCAACGCAGAGTACATGGGCAATTAAATCAGAATCTCTCAGAGTA
NB501555:820:HV3HNBGXG:1:13210:15292:11964 AAGCAGTGGTATCAACGCAGAGTACATGGGTGTGCTCTCACCGACTACGGCCTGG
NB501555:820:HV3HNBGXG:1:23206:15499:9560 AAGCAGTGGTATCAACGCAGAGTACATGGGTAGTGACAGTTGAACACAGGTGACAA
NB501555:820:HV3HNBGXG:2:22204:9702:2224 AAGCAGTGGTATCAACGCAGAGTACATGGGTGGAATAGAAAACCTCCTCCAAAAC
NB501555:820:HV3HNBGXG:2:12111:18322:2738 AAGCAGTGGTATCAACGCAGAGTACATGGGGAATATCACATGACCAAATATAGAAA
NB501555:820:HV3HNBGXG:4:21502:4318:13345 AAGCAGTGGTATCAACGCAGAGTACATGGTTGGGTGTATGTACATATA

Anatomy of a 10X 3'-Single Cell Amplicon



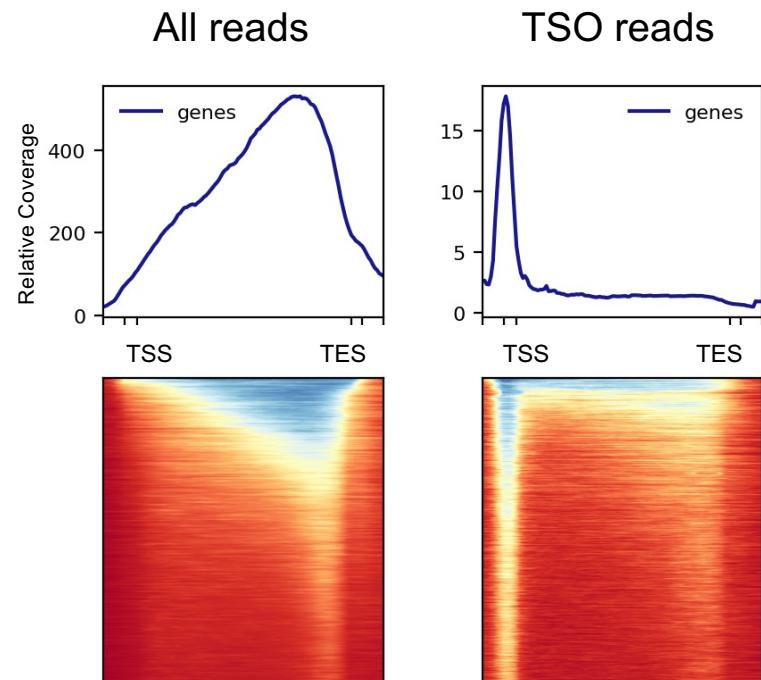
Libraries vary in abundance of TSO reads

- Optimize Fragmentation Step?
- Improve size selection



Where do these
TSO-containing
reads come
from?

TSO reads are largely 5'-ends of short genes

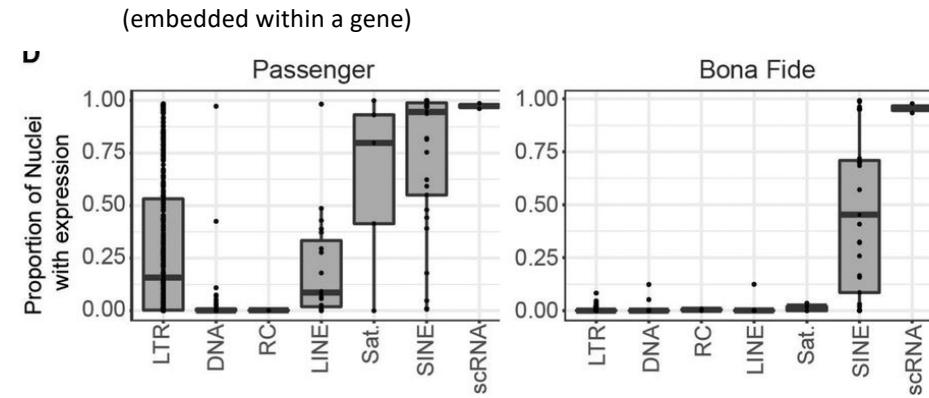
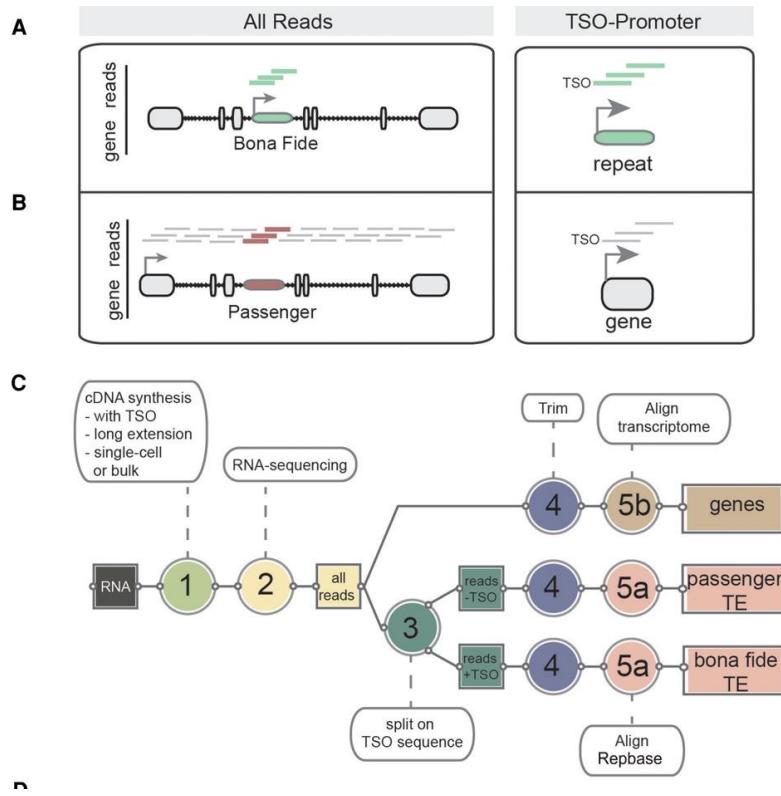


Sparse, non-reliable measure of promoter / TSS utilization

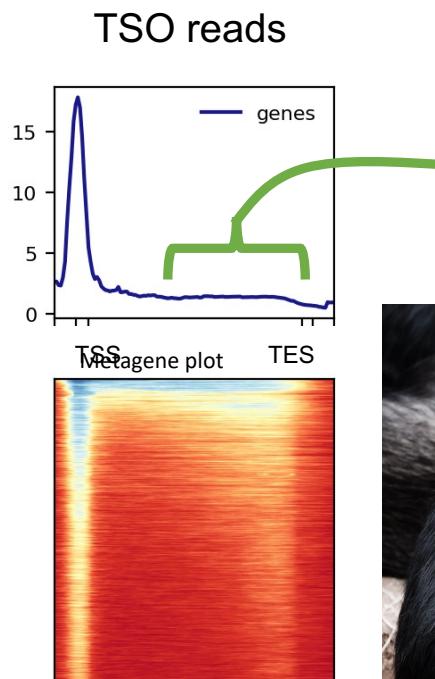
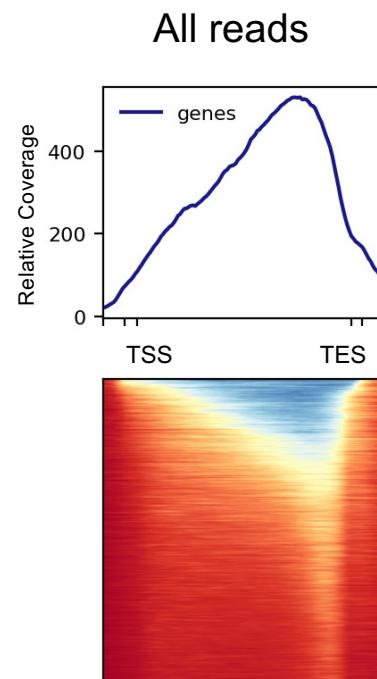
Making use of TSO-containing reads

Identification of bona fide B2 SINE retrotransposon transcription through single-nucleus RNA-seq of the mouse hippocampus

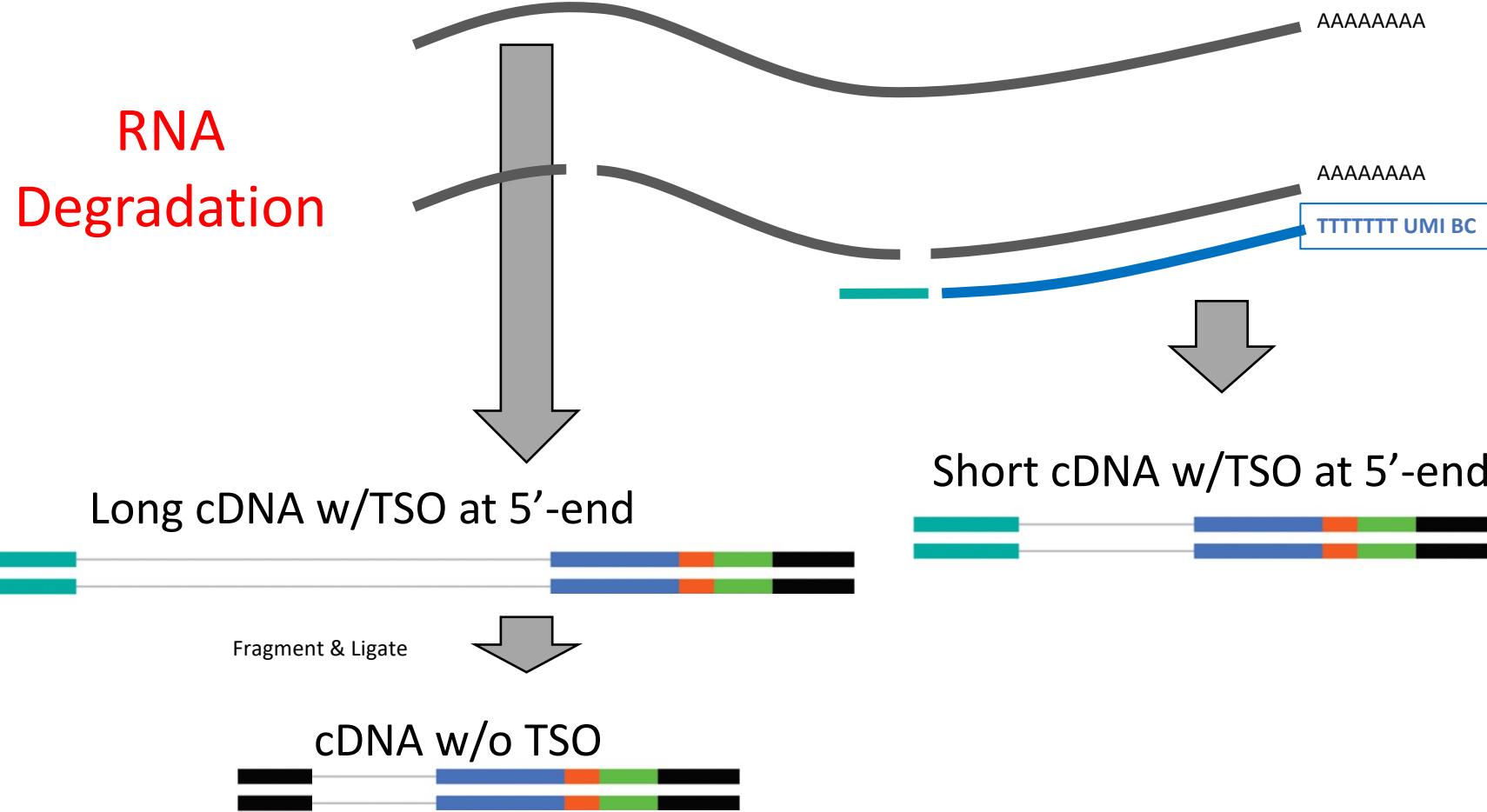
Sara B. Linker,¹ Lynne Randolph-Moore,¹ Kalyani Kottilil,¹ Fan Qiu,¹ Baptiste N. Jaeger,² Jerika Barron,³ and Fred H. Gage¹



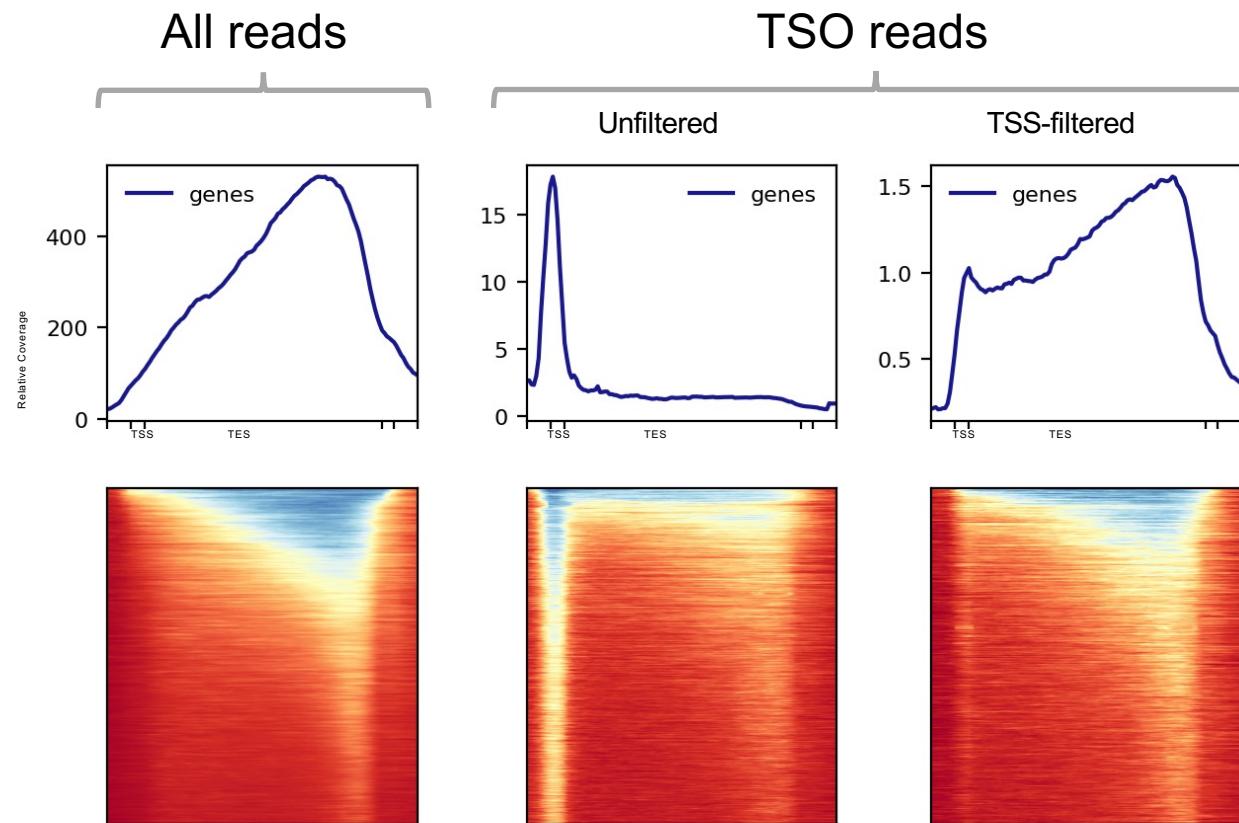
TSO reads are largely 5'-ends of short genes



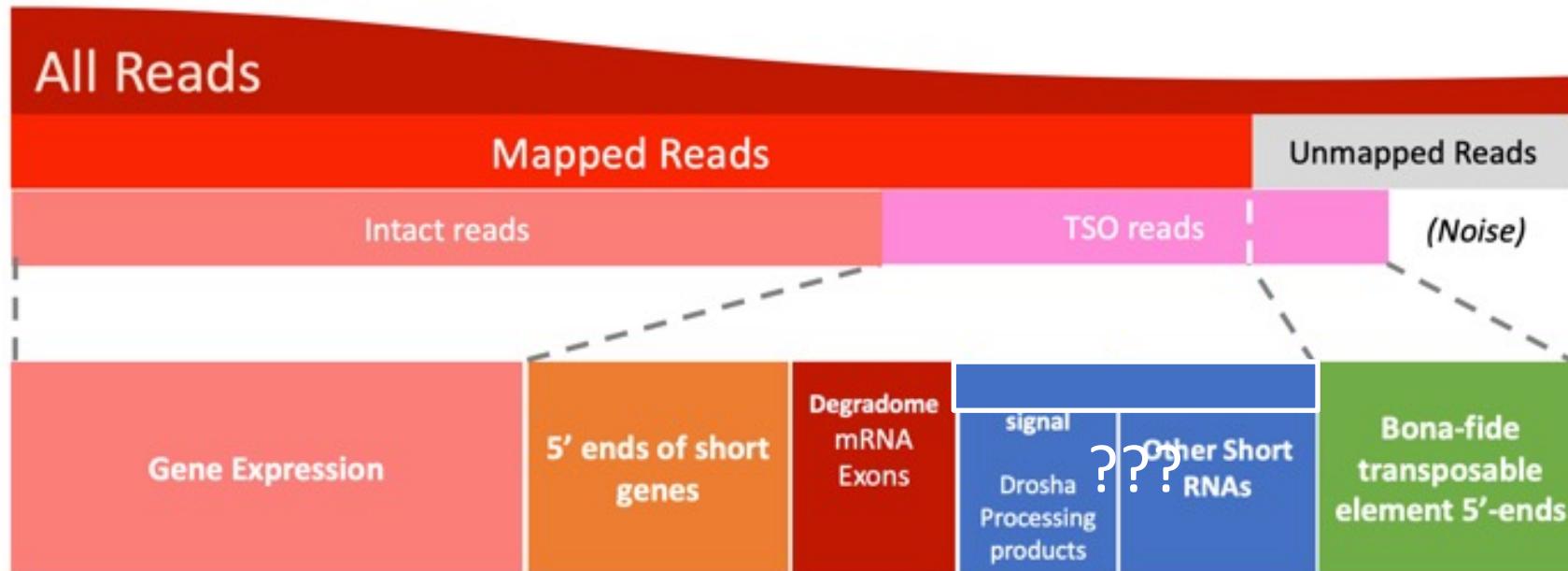
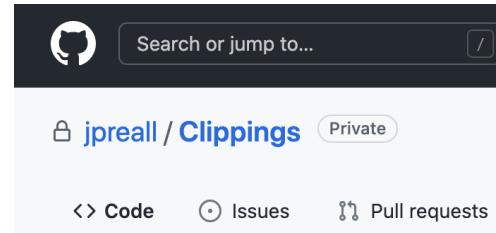
What about reads
from nearer to the
3' end?



Estimate RNA degradation by filtering TSS reads

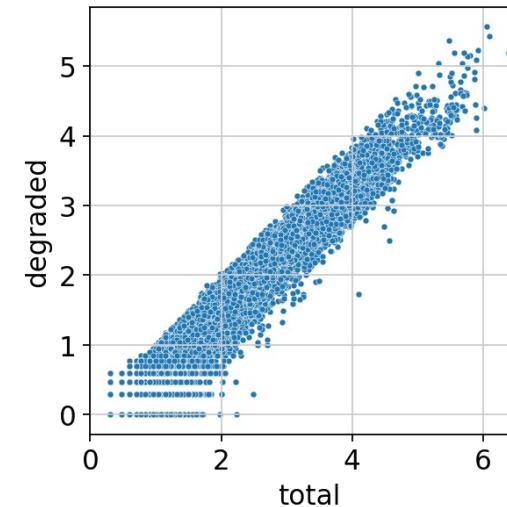
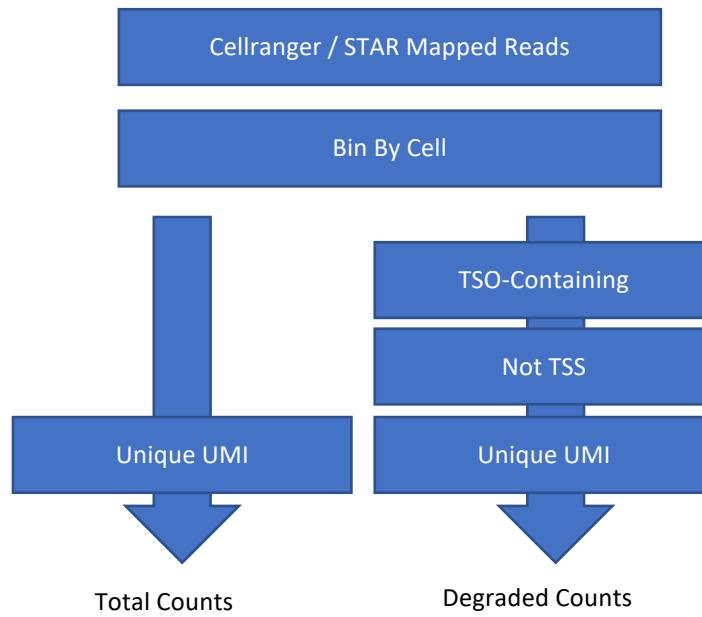


“Clippings” Software in development



Clippings Function #1

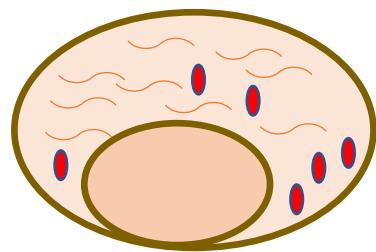
Degraded Gene Expression Matrix Strongly Resembles
Total Gene Expression Matrix



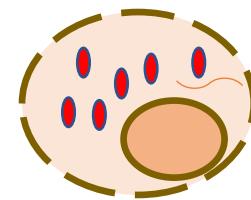
No evidence (yet) of
gene-set-specific
degradation
programs

Cell Quality Control: Mitochondrial Reads

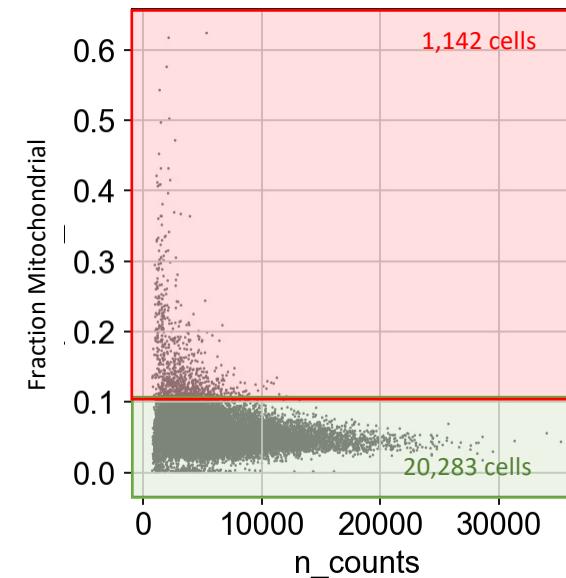
Healthy Cell



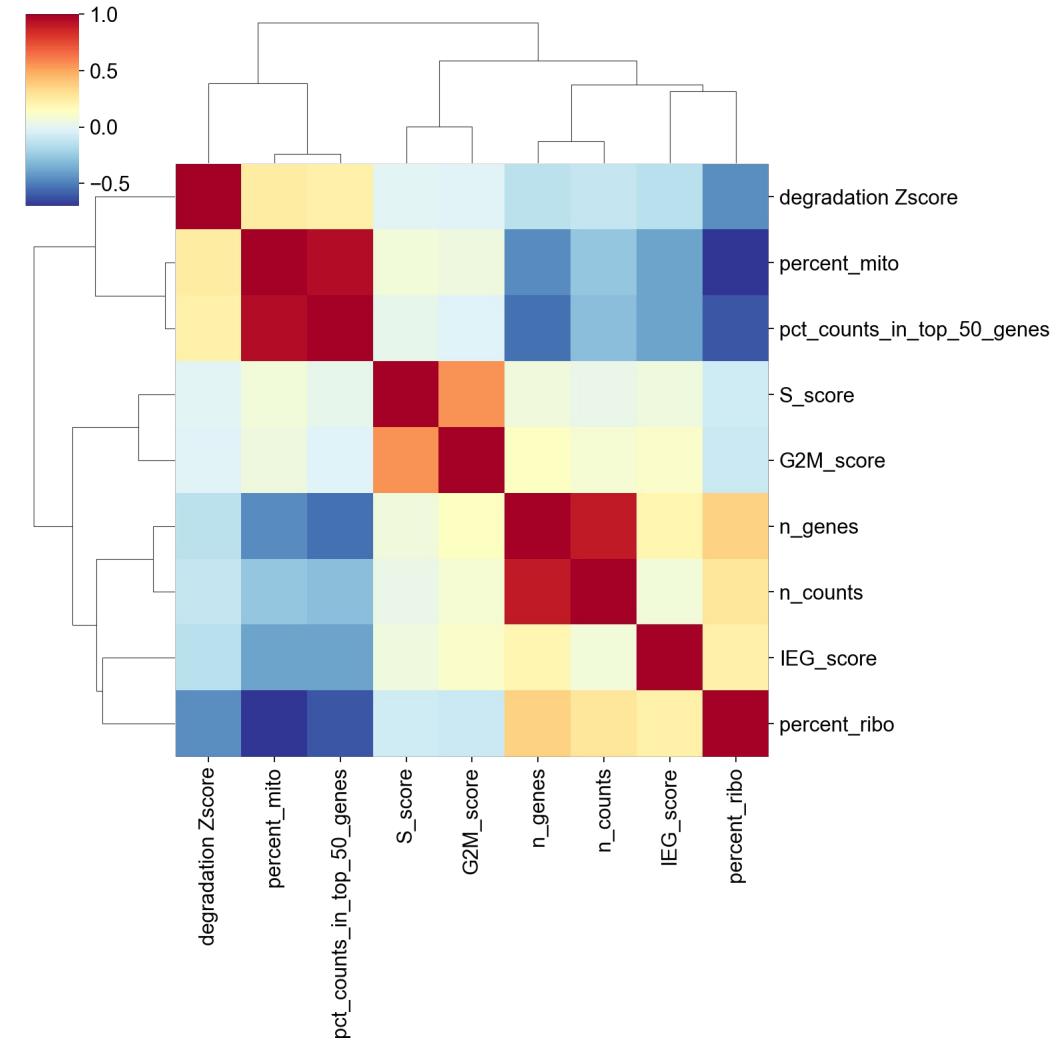
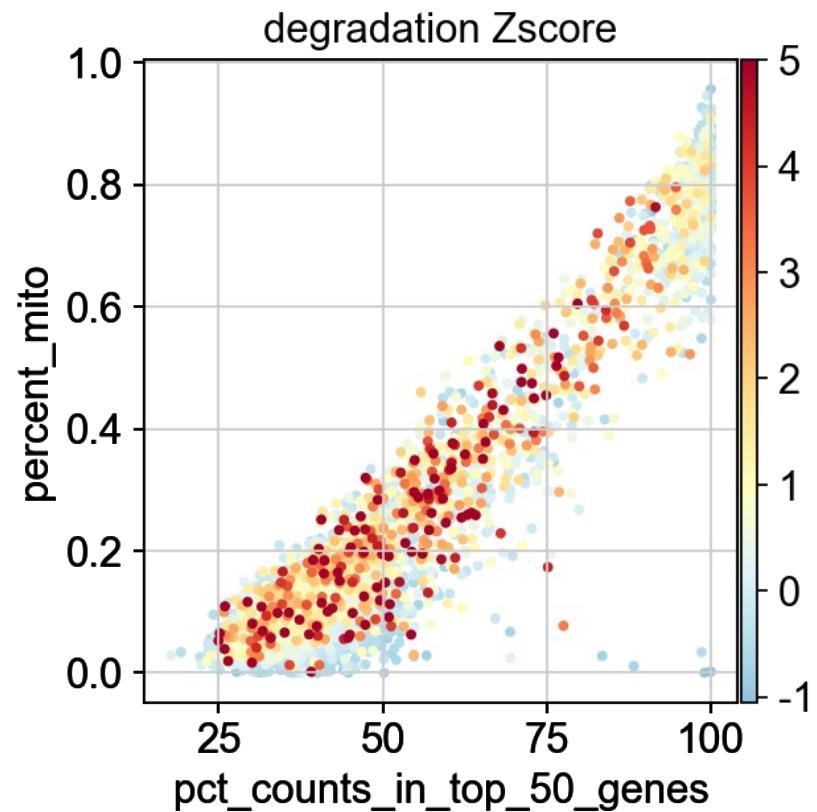
mRNA
Mitochondrion



Dying / Ruptured Cell



RNA degradation as an additional QC metric

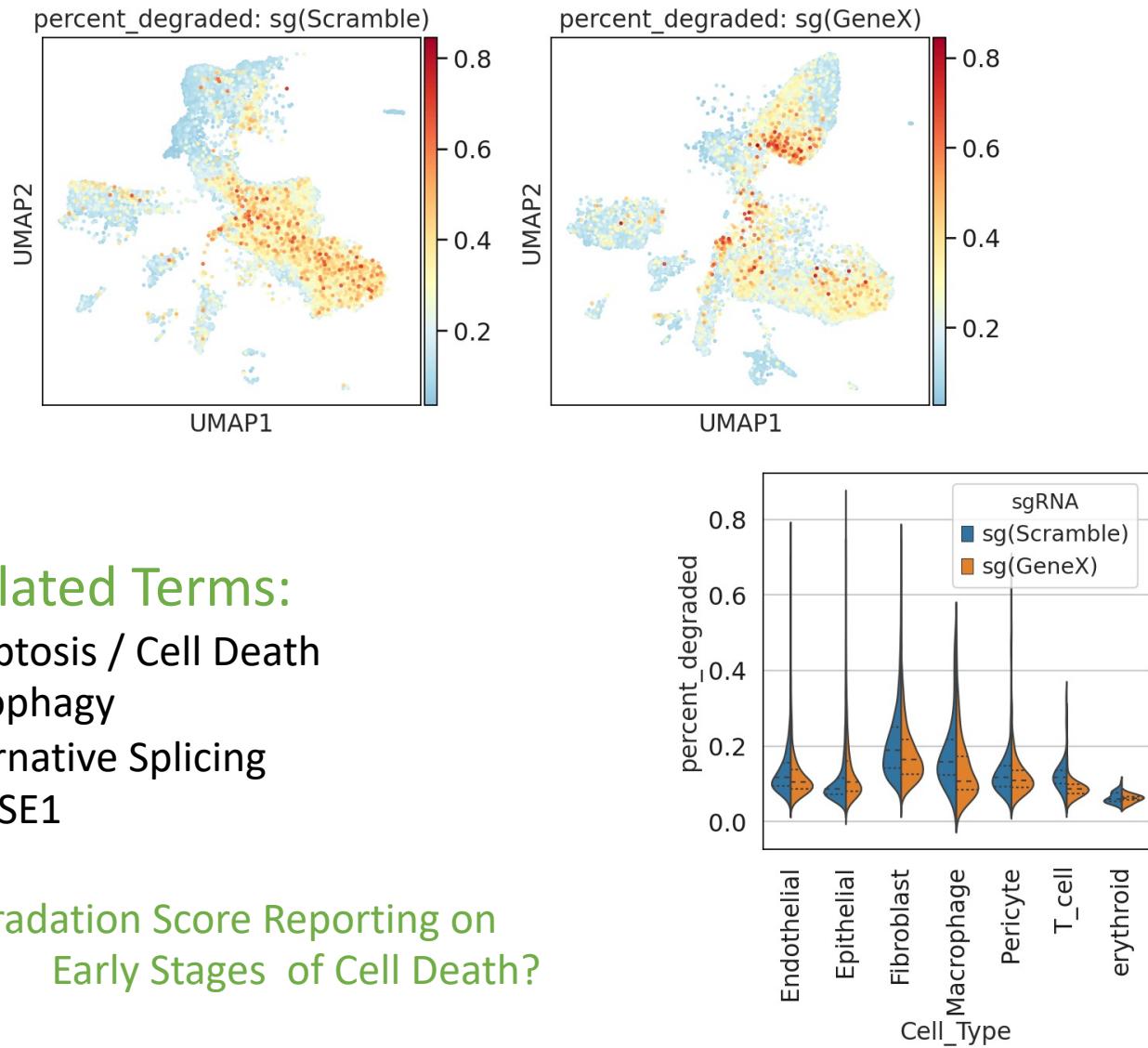
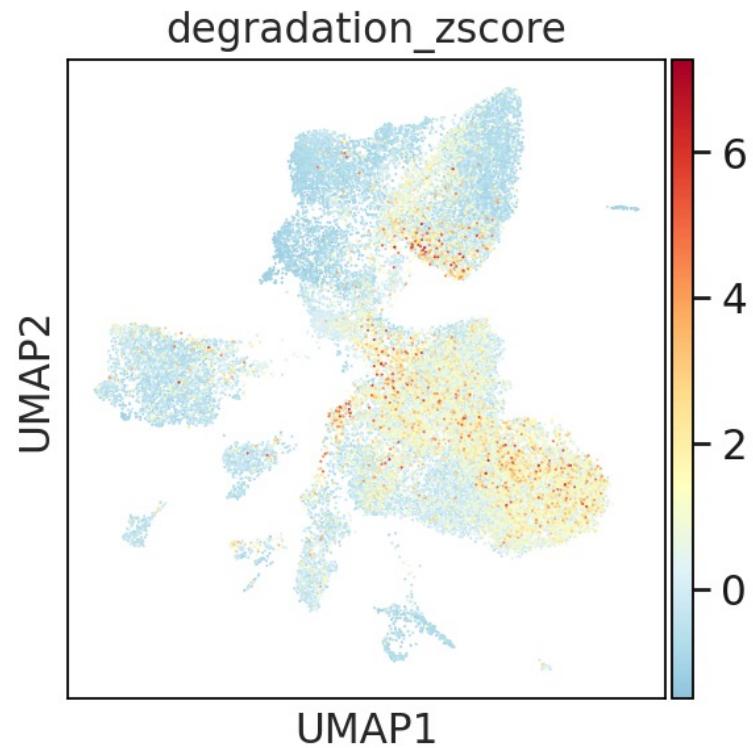


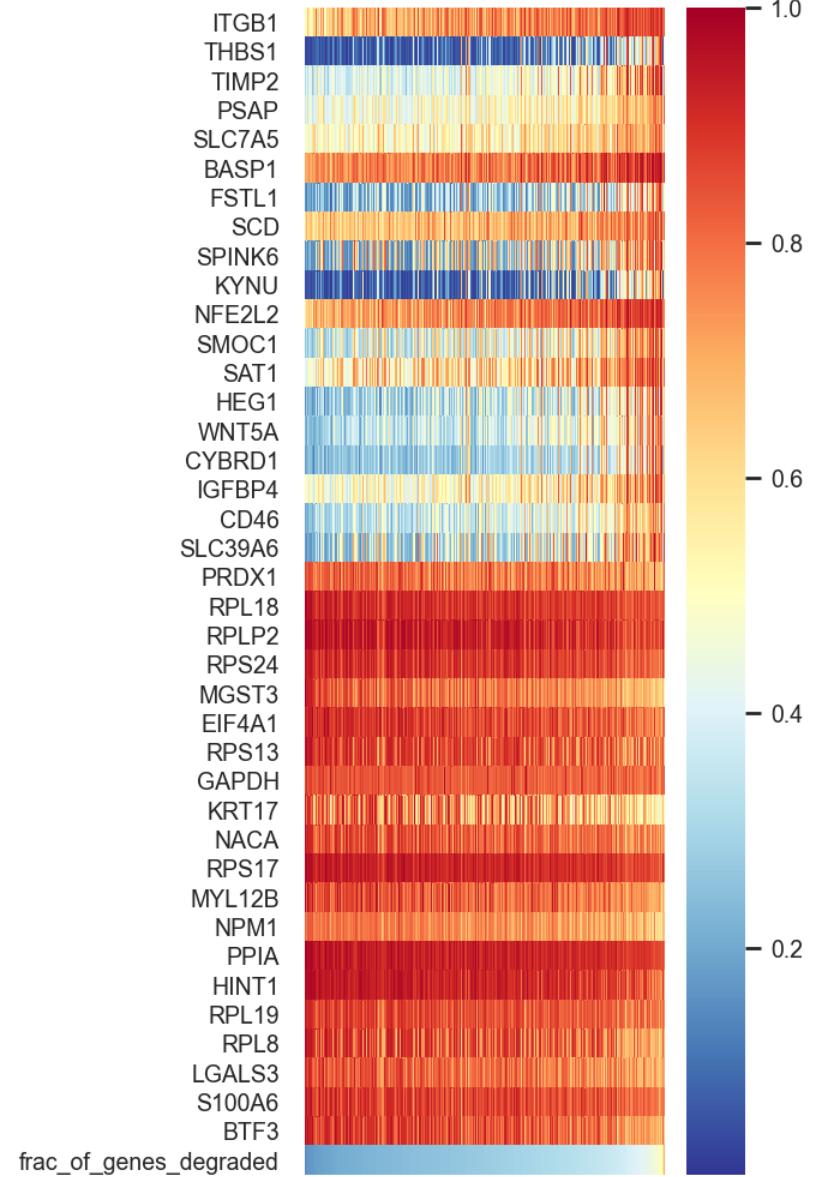
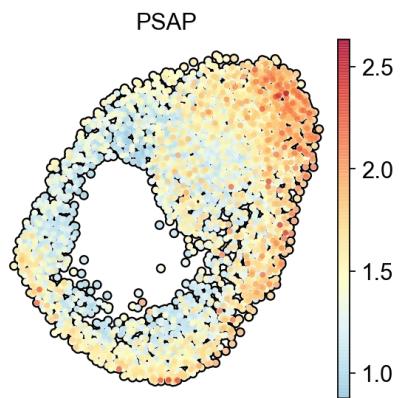
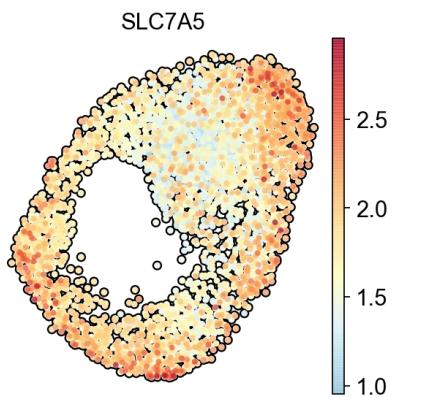
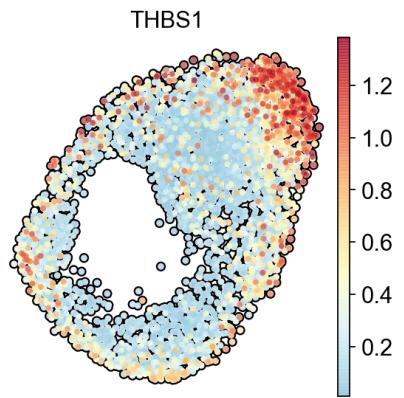
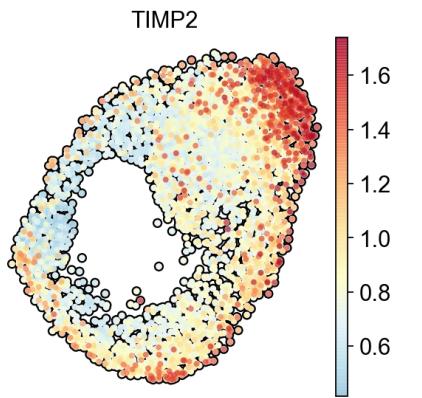
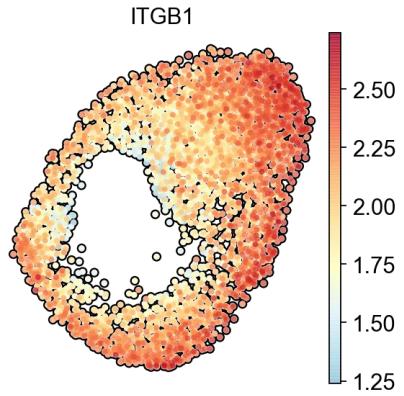
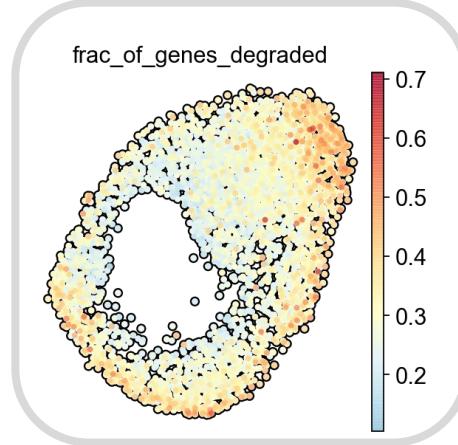
Transplanted Tumor Model: Genetic Perturbation

Organoid Orthotopic Transplant:

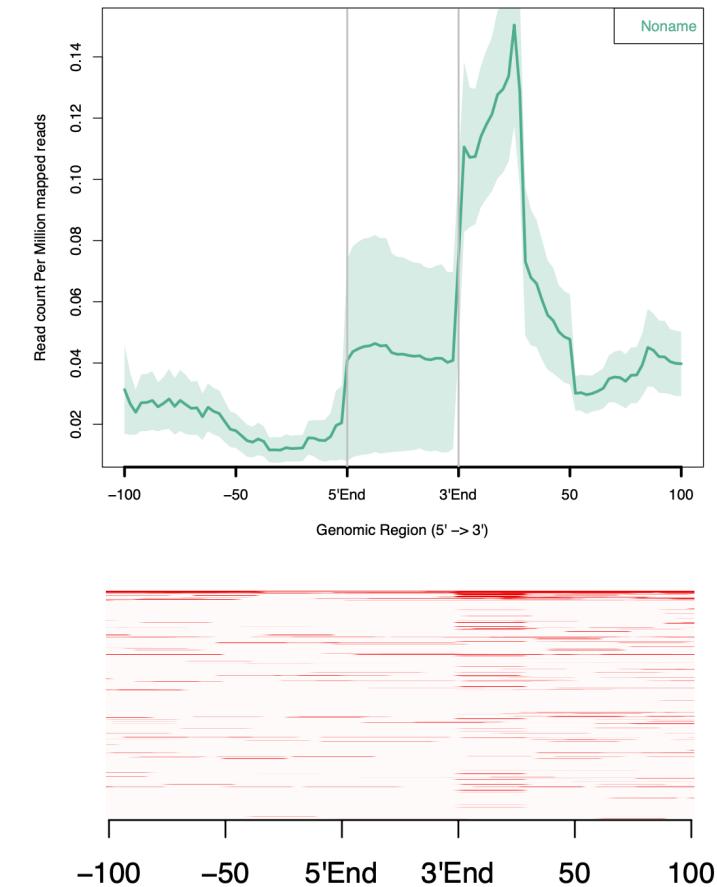
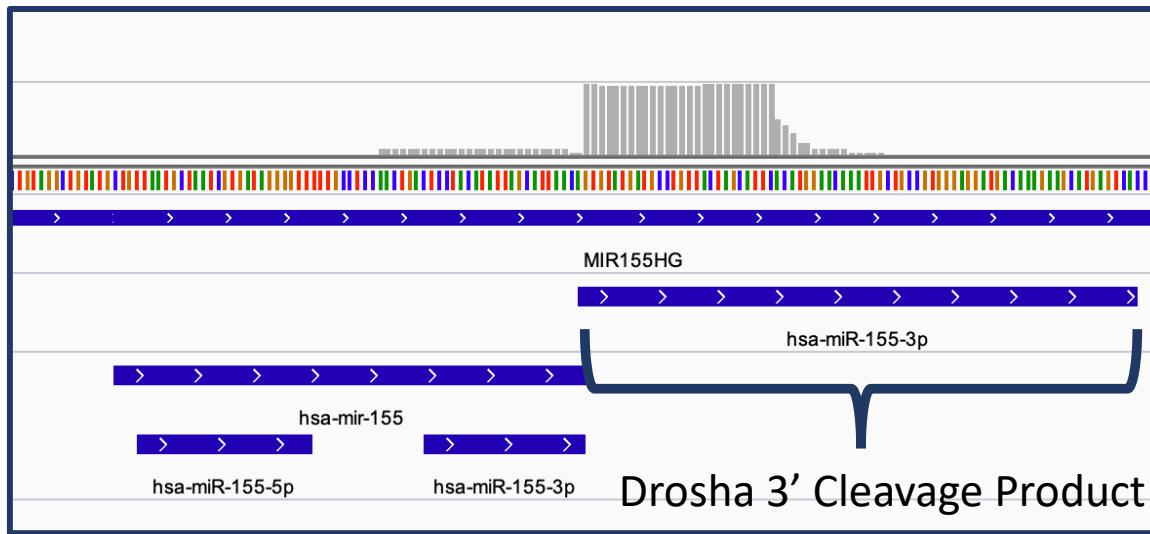
CRISPR manipulated ex vivo

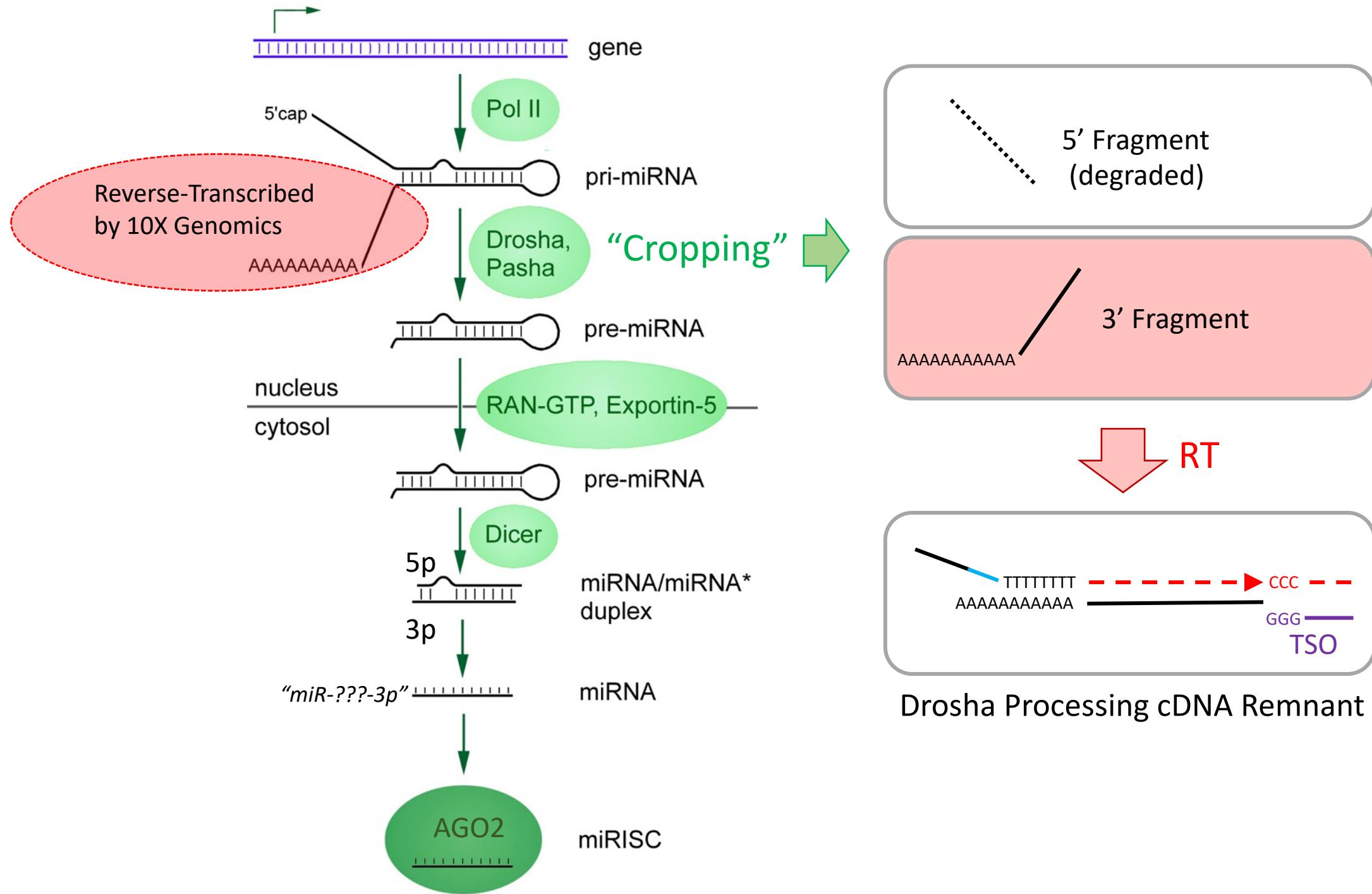
- sgRNA (Scramble)
- sgRNA (GeneX)





Another Signal in TSO Reads: microRNAs





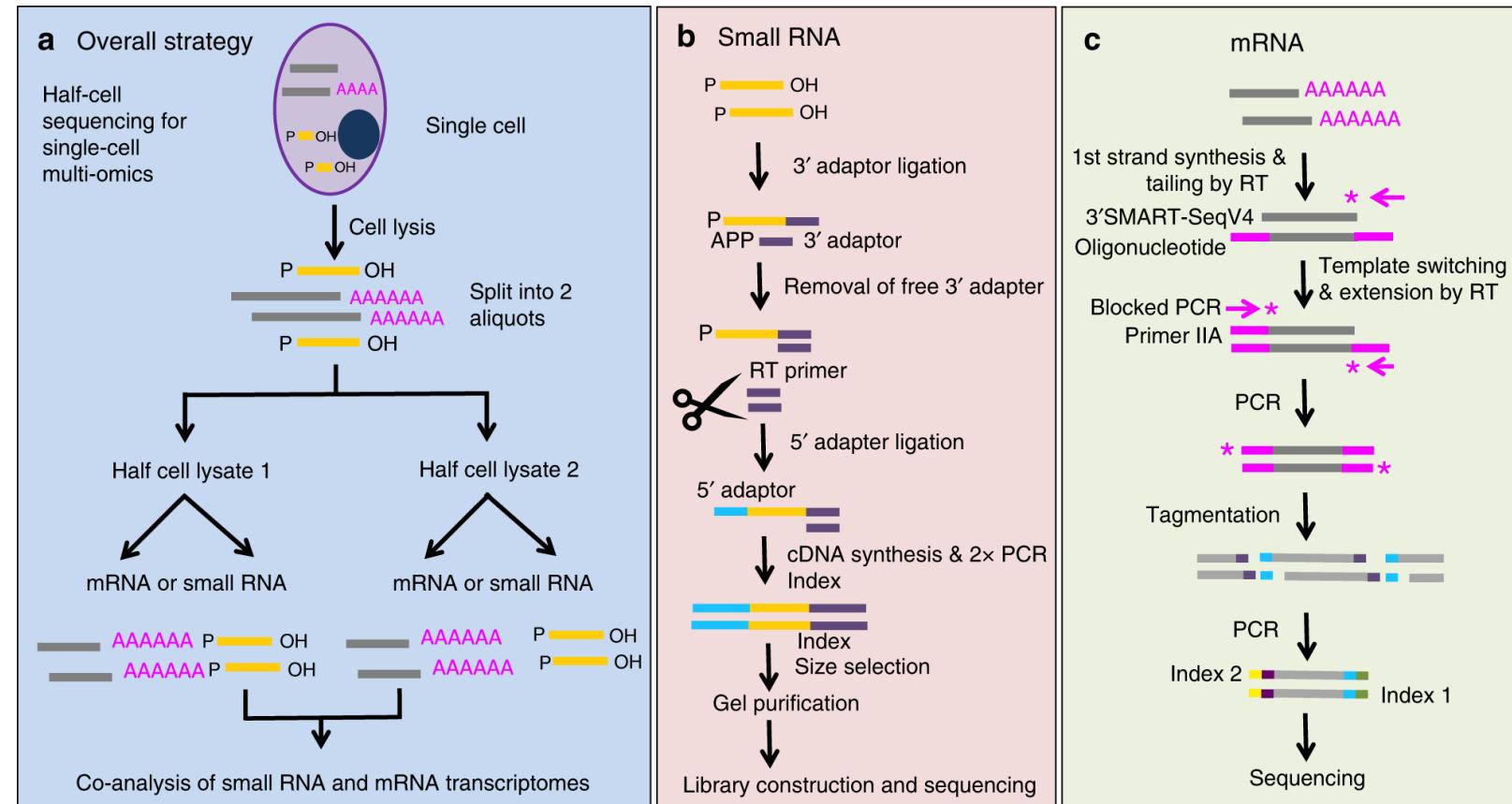
Existing sc-miRNA Methods Are Low Throughput

Advantages:

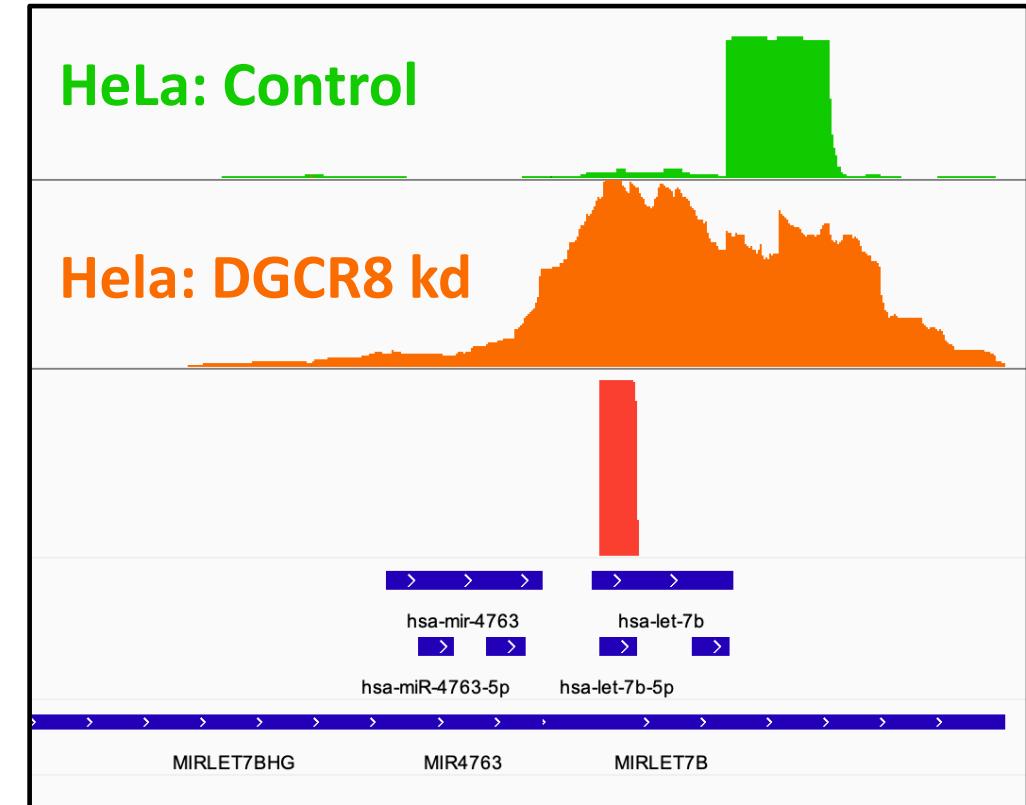
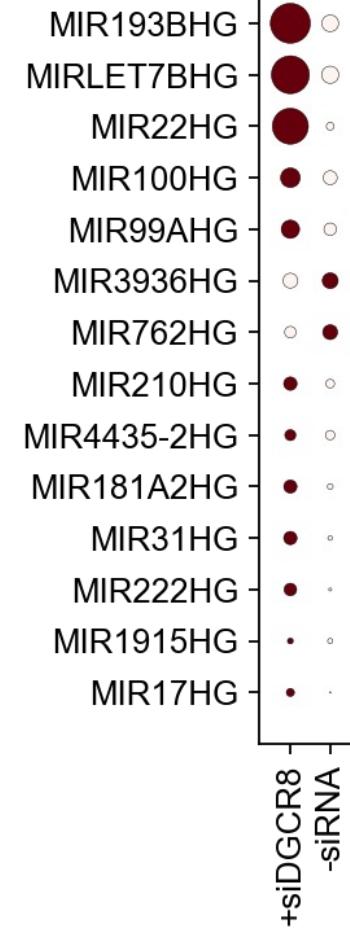
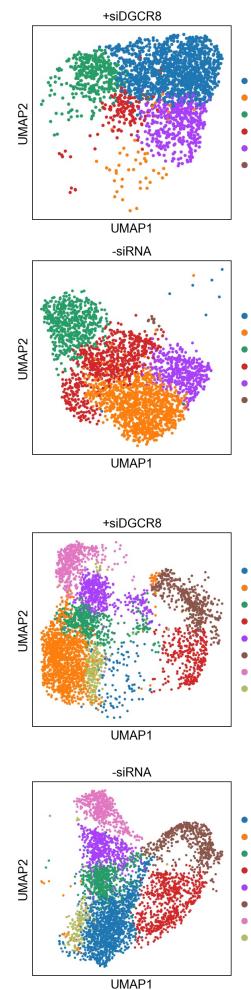
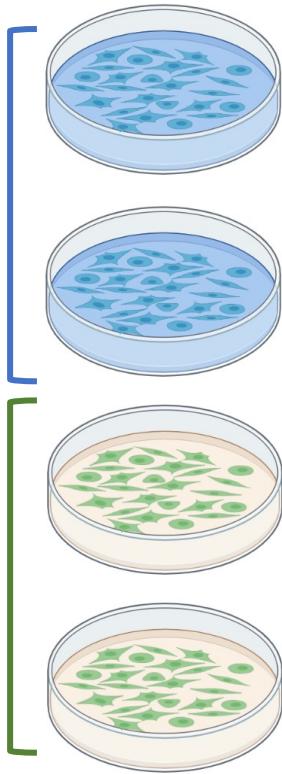
- Direct miRNA measurement
- Pan-small RNA cloning protocol

Disadvantages:

- Plate-based, manual handling
- Low throughput
 - 19 cells in P.O.C. study
- Split cDNA: lose $\frac{1}{2}$ signal each
- High reagent cost
 - RNA ligase
 - APP-adaptor
- Not adaptable to microfluidics

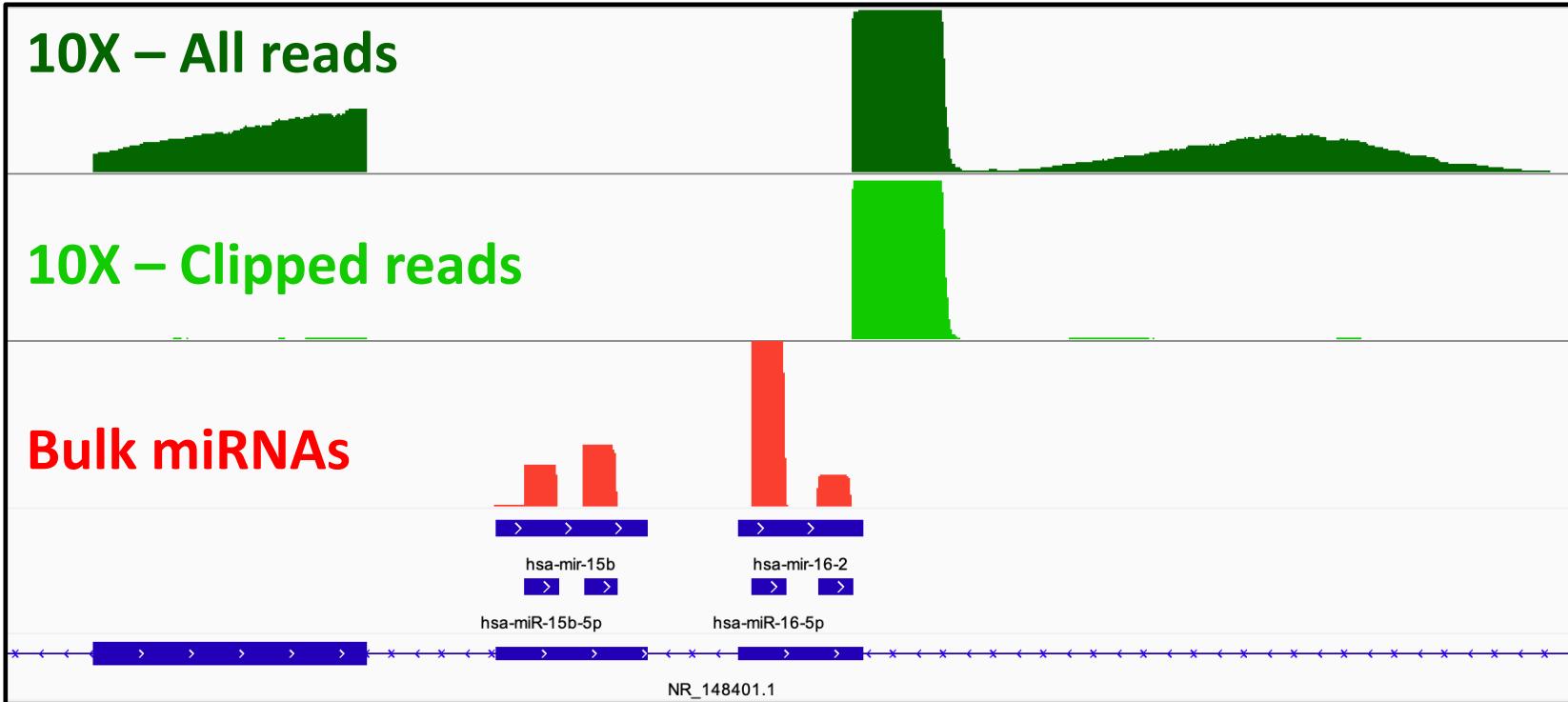


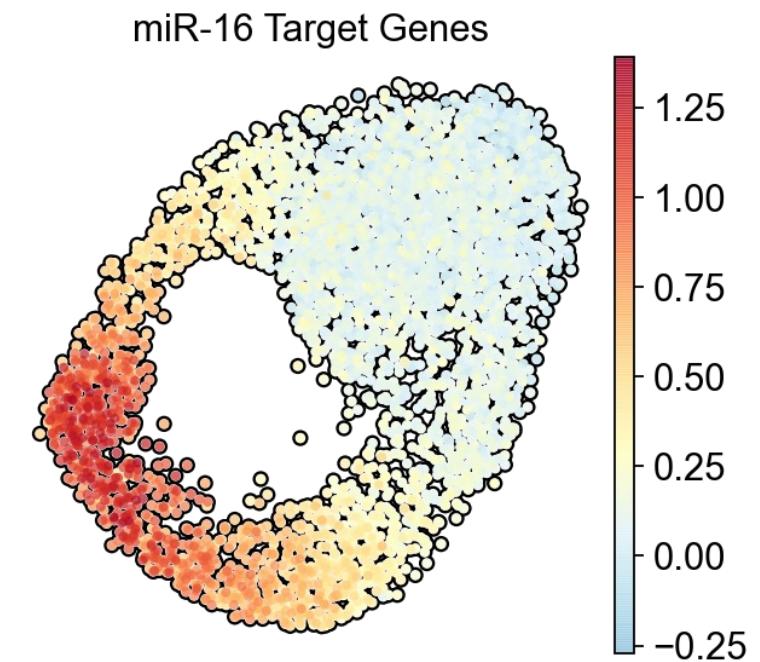
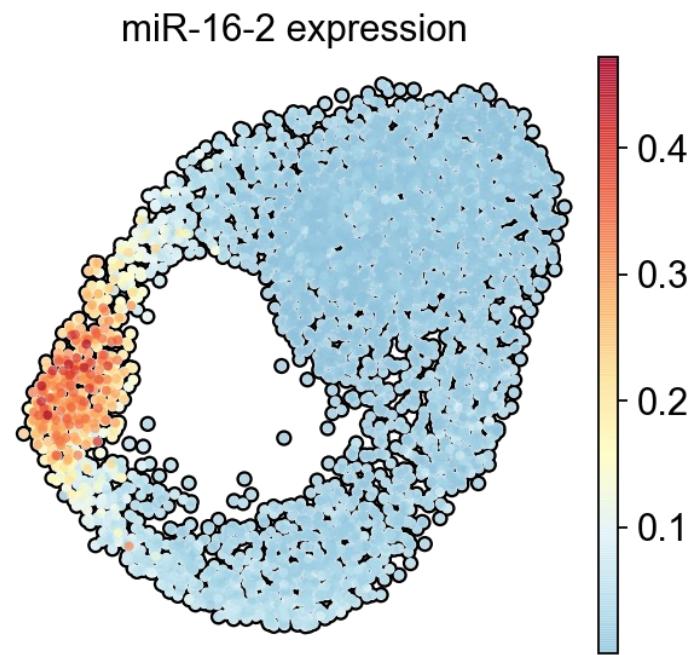
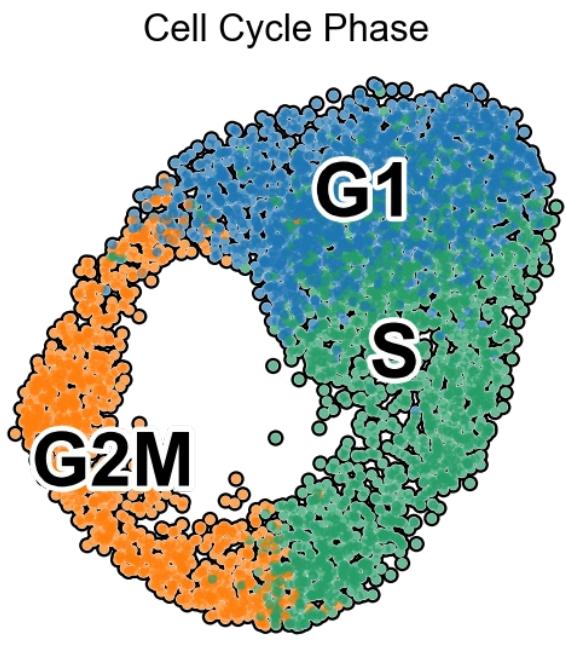
HeLa



10X

Bulk miRNA cloning





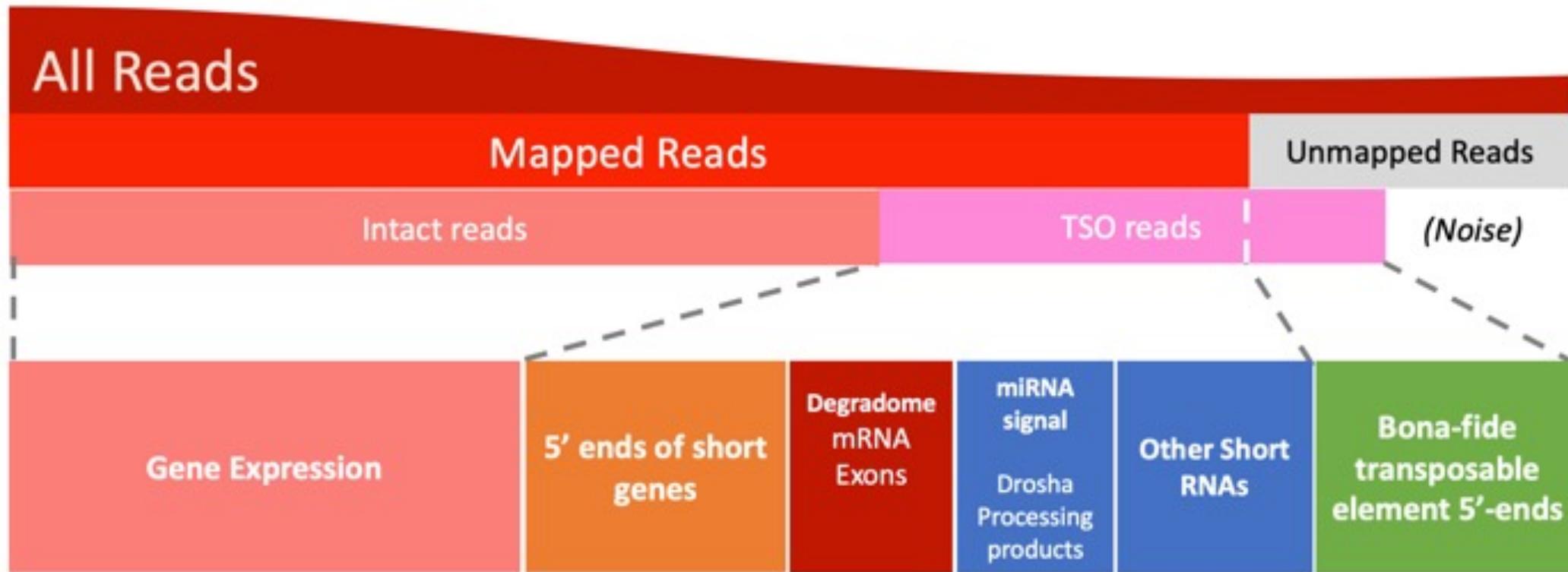
Software in development

Mining existing 10X mapping data for miRNA,
degradome, and TE expression

Clippings



A screenshot of a GitHub repository page. The repository name is 'jpreall / Clippings' and it is marked as 'Private'. Below the repository name are three navigation links: '<> Code', 'Issues', and 'Pull requests'. The background of the page features a faint watermark of the cartoon illustration from above.



Acknowledgements

Preall Lab:

Claire Regan
Brian He
Caleb Mallery
Rad Utama

Tuveson Lab

Ela Elyada
Giulia Biffi

Dick McCombie / NGS

Sara Goodwin
Elena Ghiban

Beyaz Lab

Fearon Lab

Rena Yan



Funding:

Northwell Health
CSHL Cancer Center Support Grant