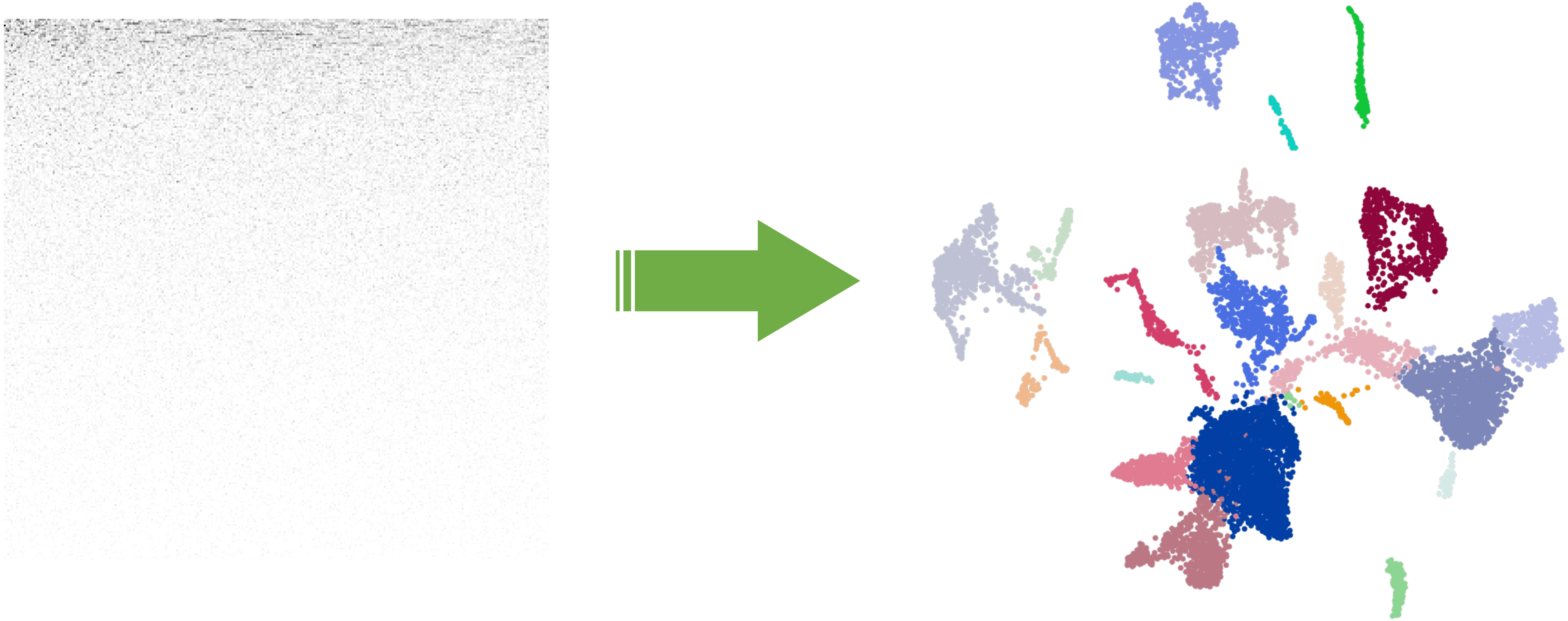
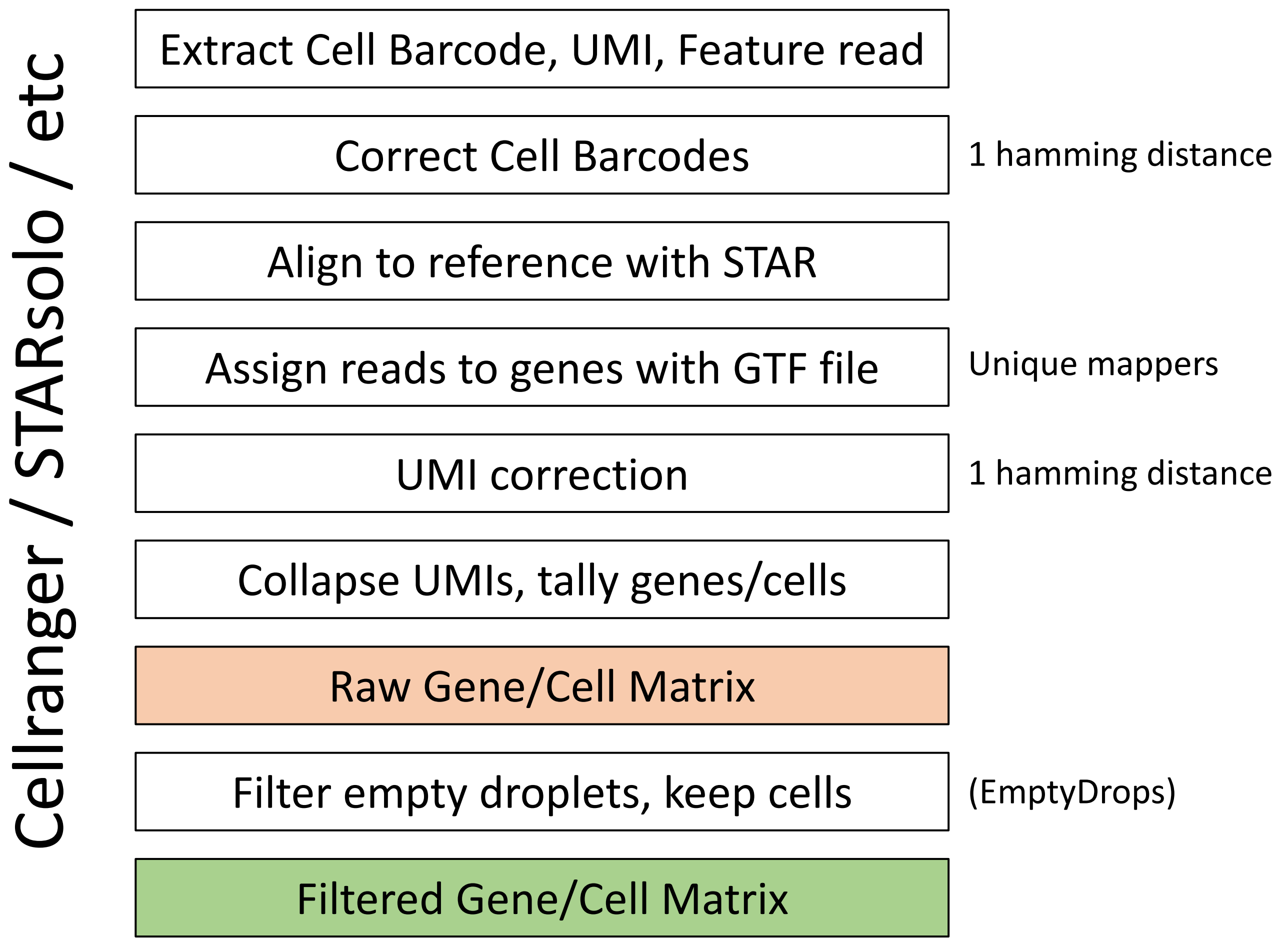
Introduction to scRNAseq Analysis



Jon Preall
Research Associate Professor, CSHL
November 2022

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

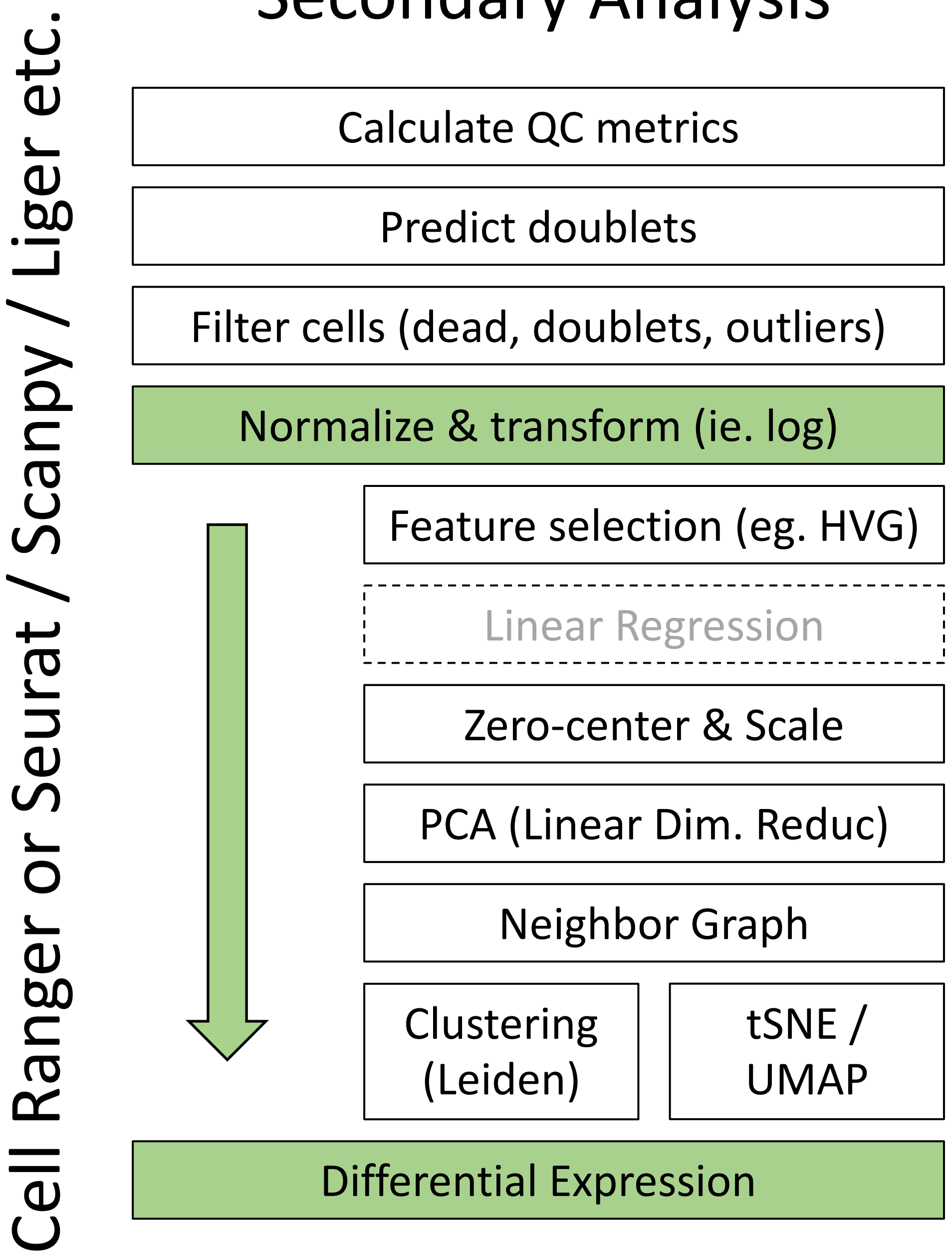
Collapse UMIs, tally genes/cells

Raw Gene/Cell Matrix

Filter empty droplets, keep cells — (EmptyDrops)

Filtered Gene/Cell Matrix

Secondary Analysis



Cell Ranger or Seurat / Scanpy / Liger etc.

Calculate QC metrics

Predict doublets

Filter cells (dead, doublets, outliers)

Normalize & transform (ie. log)

Feature selection (eg. HVG)

Linear Regression

Zero-center & Scale

PCA (Linear Dim. Reduc)

Neighbor Graph

Clustering (Leiden)

tSNE / UMAP

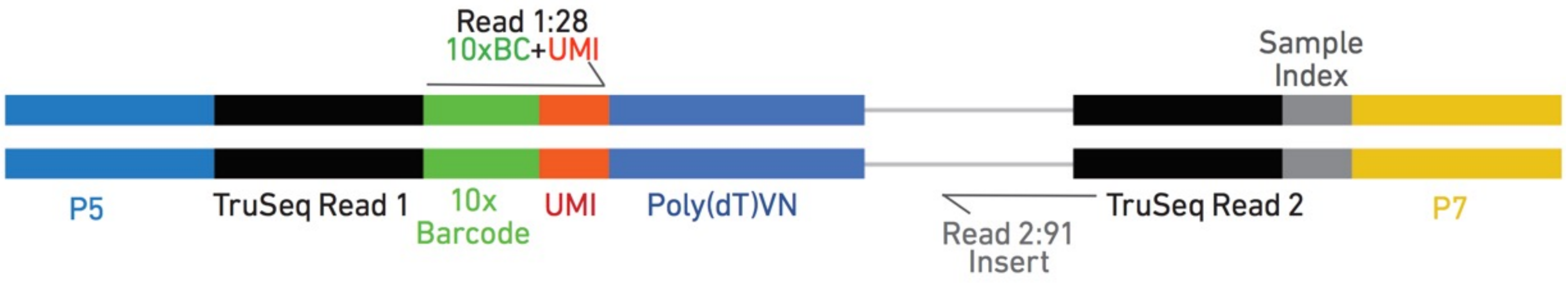
Differential Expression

Mapping and Counting

Anatomy of a 10X 3'-Single Cell Amplicon

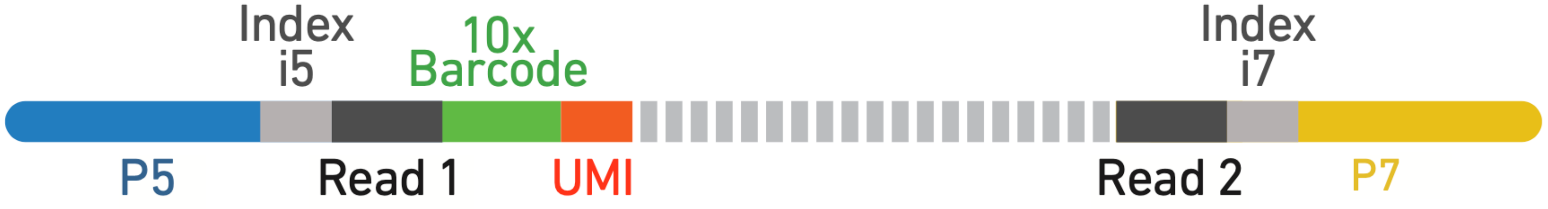
V3 / NextGem Chemistry

Single Index



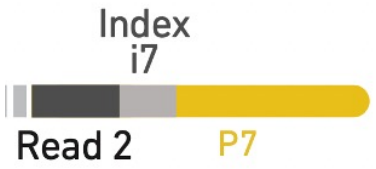
12bp UMI

Dual Index



Library Indexing

Single Indexed



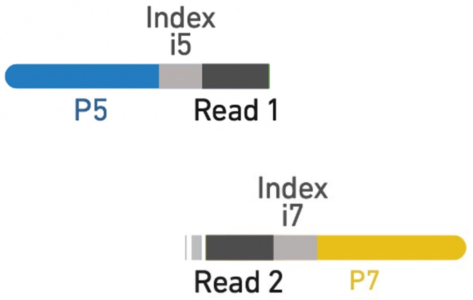
SI-GA-A1 i7
GGTTTACT
CTAAACGG
TCGGCGTC
AACCGTAA

SI-GA-A2 i7
TTTCATGA
ACGTCCCT
CGCATGTG
GAAGGAAC

- 8bp
- Mix of 4 balanced barcode sequences
- Don’t have to worry about how to pool multiple libraries

- Susceptible to INDEX HOPPING

Dual Indexed



SI-TT-A1
i7: GTAACATGCG
i5: AGTGTTACCT

SI-TT-A2
i7: GTGGATCAAA
i5: GCCAACCCTG

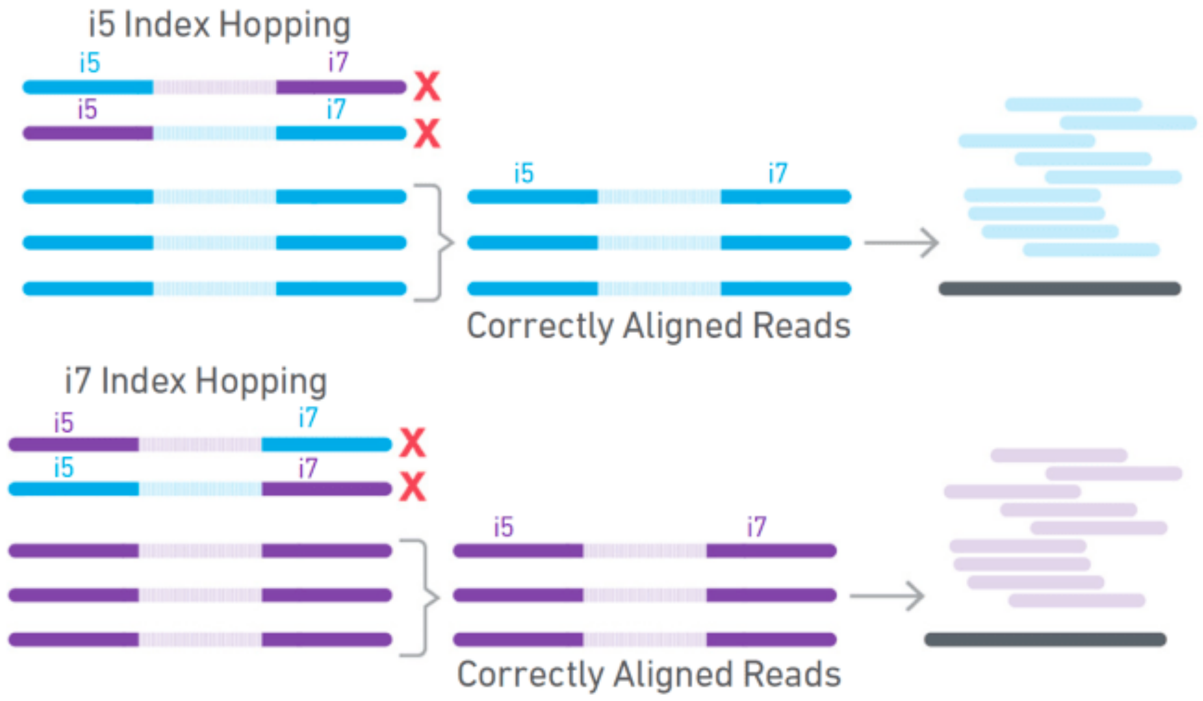
- 10bp each (20 cycles total)
- Fixes index hopping
- Pooling with low-plex libraries??

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

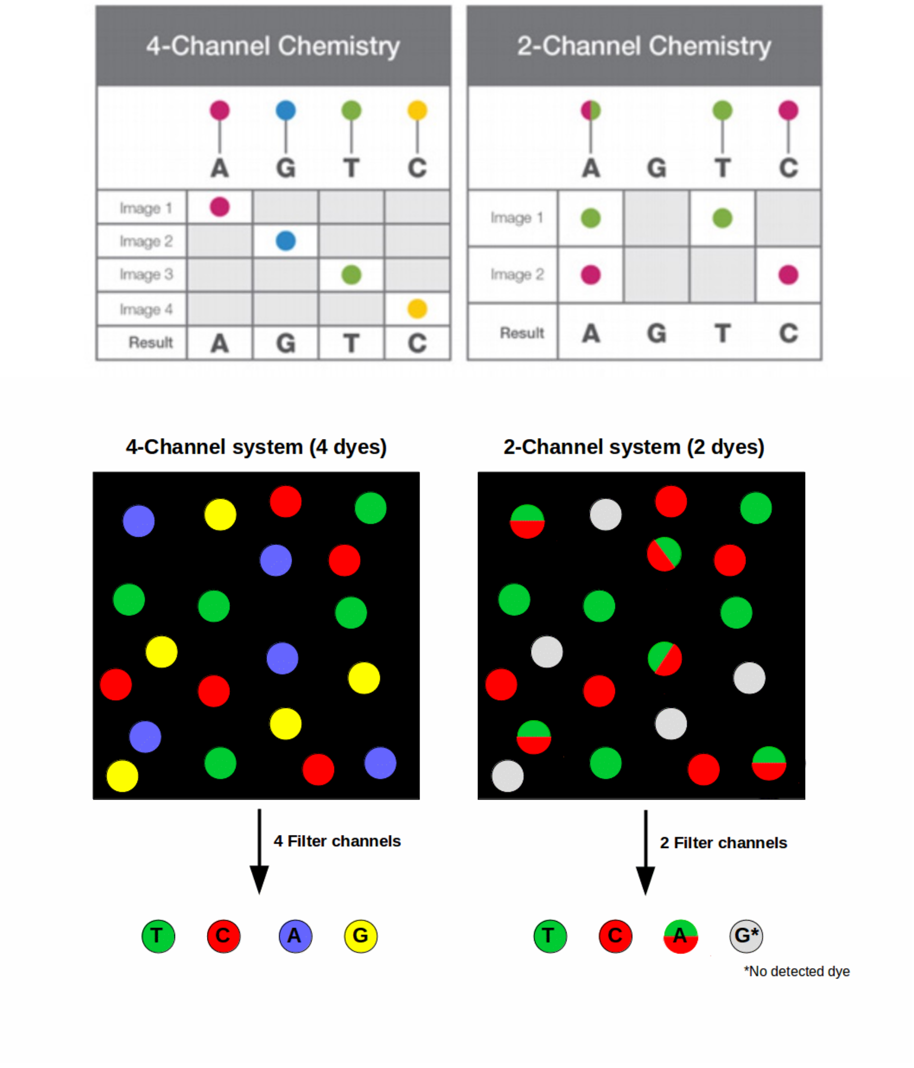
- < 1% probability of occurring…
- Can have HUGE affect on unique Cell / Barcode / UMI counts!



`cellranger mkfastq` ➡ `index-hopping-filter`

Automatically filtered by Cellranger version 4+

Index Diversity

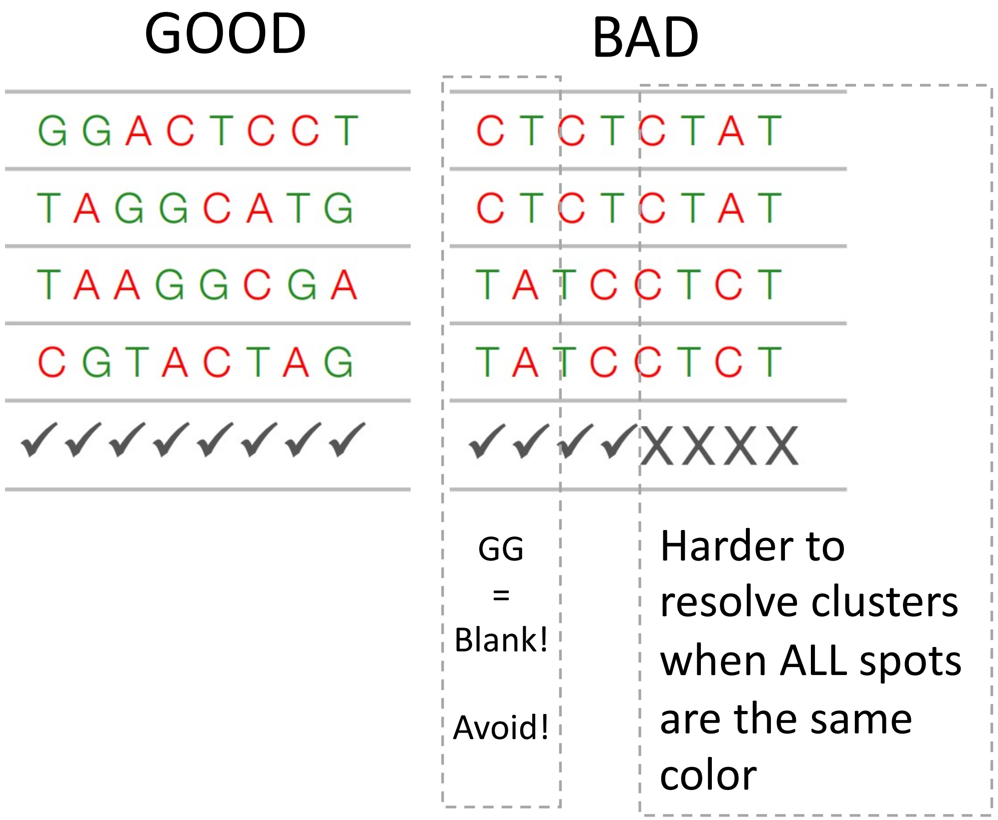


4-Channel Chemistry	A	G	T	C
Image 1	●			
Image 2		●		
Image 3			●	
Image 4				●
Result	A	G	T	C

2-Channel Chemistry	A	G	T	C
Image 1	●		●	
Image 2	●			●
Result	A	G	T	C

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	BAD
GGACTCCT	CTCTCTAT
TAGGCATG	CTCTCTAT
TAAGGCGA	TATCCTCT
CGTACTAG	TATCCTCT
✓✓✓✓✓✓✓✓	✓✓✓✓XXXX

GG = Blank!

Avoid!

Harder to resolve clusters when ALL spots are the same color

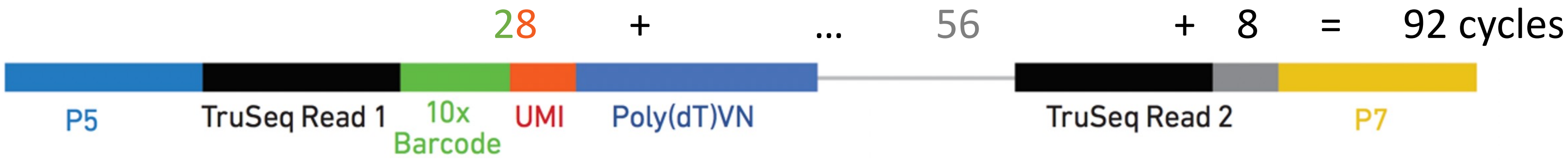
Optimizing your Sequencing run

Instrument		Kit Size	Actual Max. Cycle #	Dark Cycles for Dual Index?
NextSeq 500/550	High Output or Mid Output	75 150	92 168	No
NextSeq 1000/2000	P2 or P3	100 200	138 238	No
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

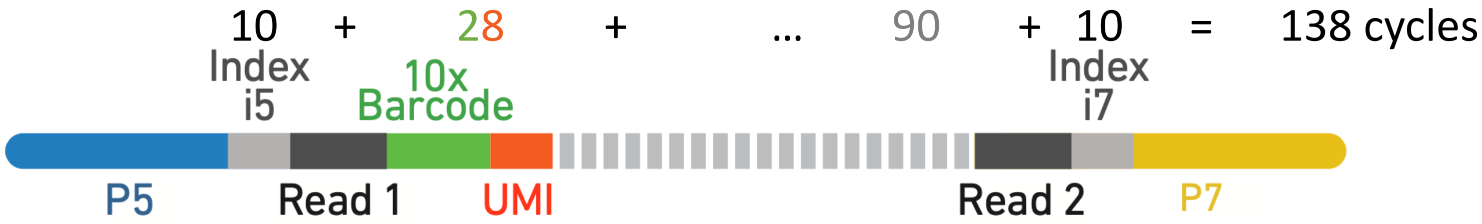
Eg.

GEX, Single Index
NextSeq 500
75 Cycle kit



28 + … 56 + 8 = 92 cycles

GEX, Dual Index
NextSeq 2000
100 Cycle kit



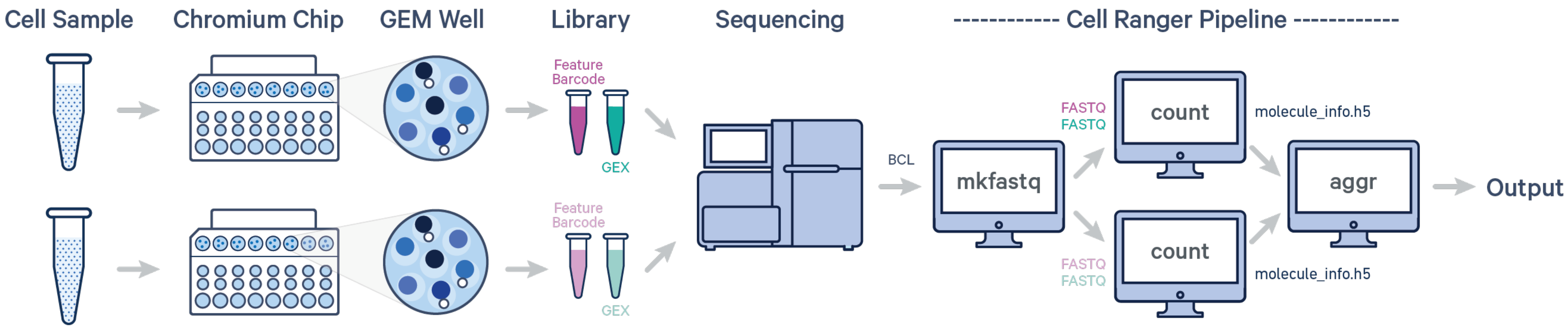
10 Index i5 + 28 + … 90 + 10 Index i7 = 138 cycles

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

https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/sequencing/sample-index-sets-for-single-cell-3

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,AGTGTTACCT,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 \
```

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,Descriptio
n
PlantCourse2022_10xGEX_control,PlantCourse2022_10xGEX_control,,,SI-TT-A9,SI-TT-A9,SI-TT-A9,SI-TT-
A9PlantCourse2022,,
PlantCourse2022_10xGEX_treatment,PlantCourse2022_10xGEX_treatment,,,SI-TT-A10,SI-TT-A10,SI-TT-A10,SI-TT-
A10,PlantCourse2022,
```

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
AAGCAGTGGTATCAACGCAGAGTACATGGCCAAGTACTACCTGGACGACACGGTGGACGTGGTCAAGATGCTGGACGGCCTGGCCAGCGC
+
CCCCCCCCCCCC-CCCCC;CCCCCCC-C;CCC-CCCCCCCCCCCCCCCC-CCCCCCCCCCCCCCCCC-CCCCCCCCCCC;CCCCCCCCC-
```

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).

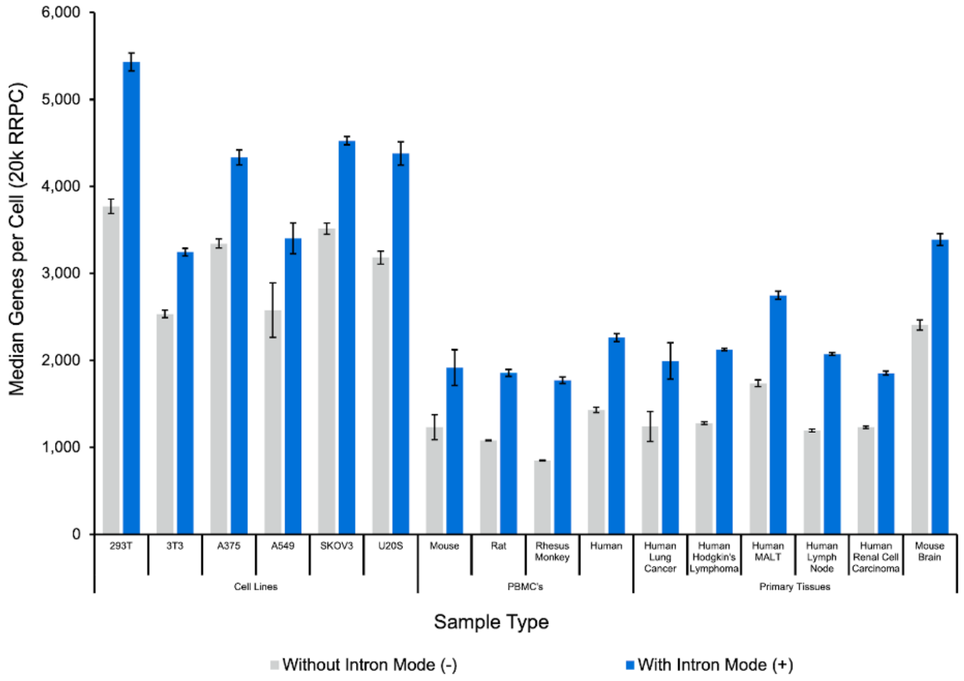
Single cell RNAseq:
Debatable.
More UMIs, but less comparable to legacy datasets

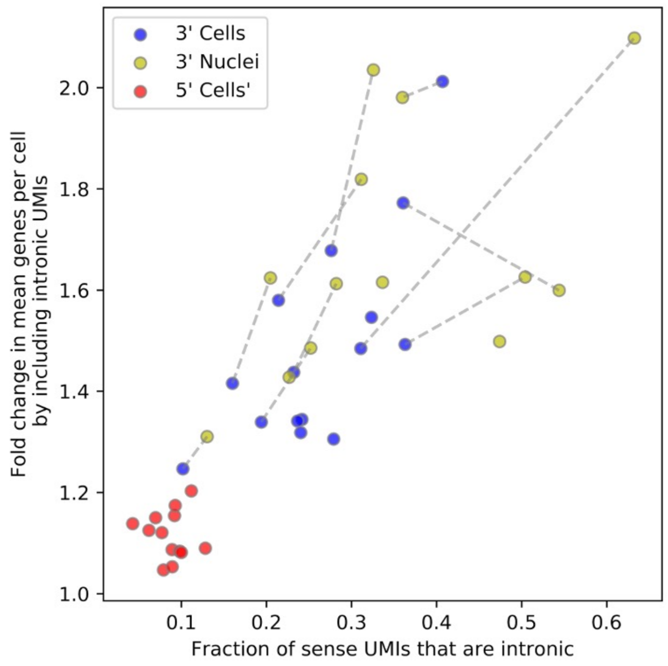
Default Behavior:
Pre-Cell Ranger 7:

```
--include-introns=False
```

Cell Ranger 7+

```
--include-introns=True
```

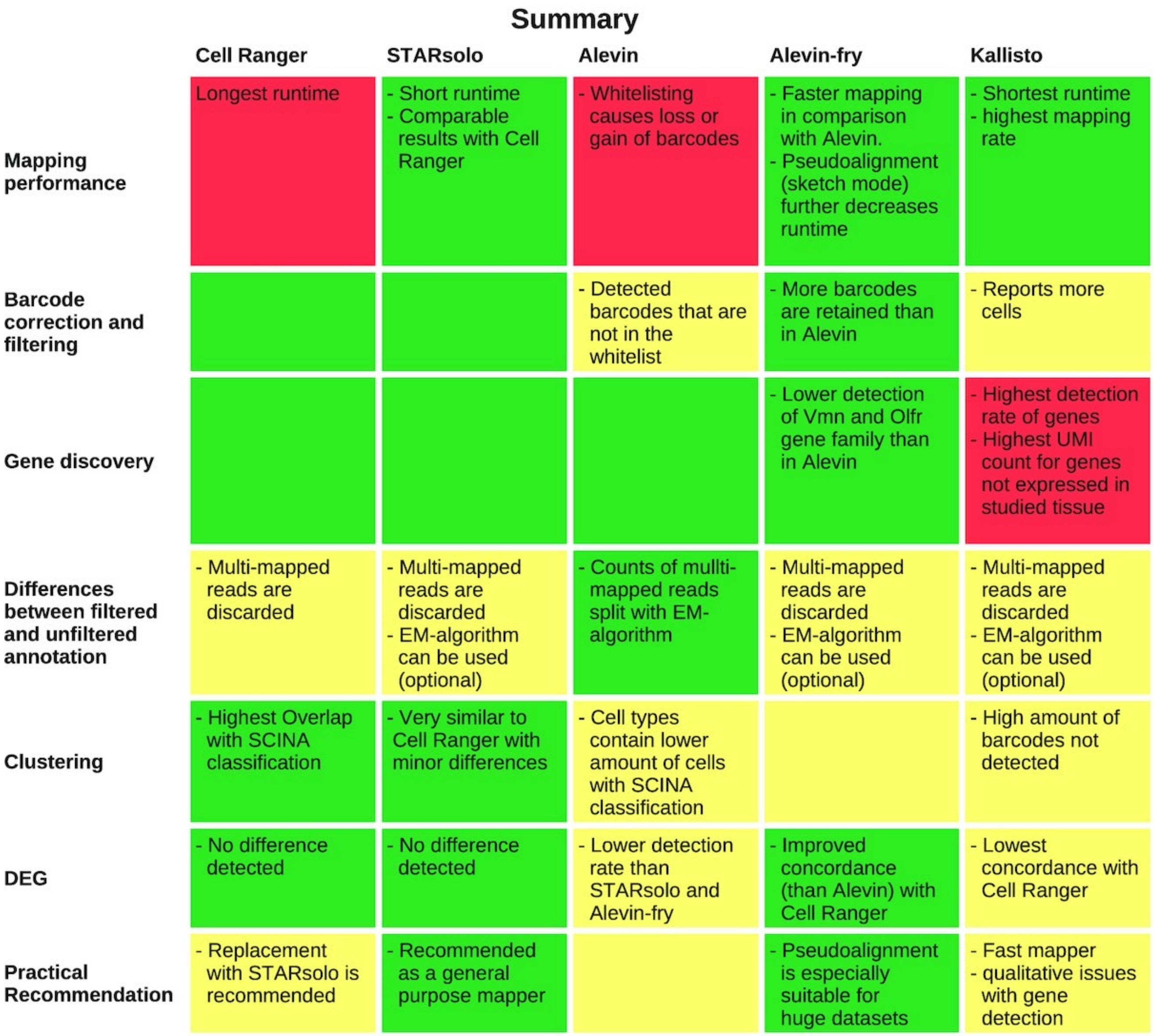




https://kb.10xgenomics.com/hc/en-us/articles/4998628924429-Why-should-I-include-introns-for-my-single-cell-whole-transcriptome-Gene-Expression-data-analysis-

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

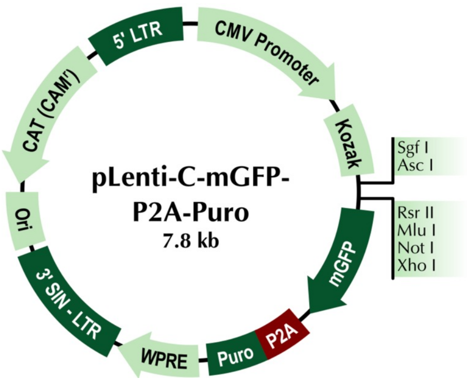
	Cell Ranger	STARsolo	Alevin	Alevin-fry	Kallisto
Mapping performance	🟥 Longest runtime	🟩 - Short runtime - Comparable results with Cell Ranger	🟥 - Whitelisting causes loss or gain of barcodes	🟩 - Faster mapping in comparison with Alevin. - Pseudoalignment (sketch mode) further decreases runtime	🟩 - Shortest 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 mullti-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

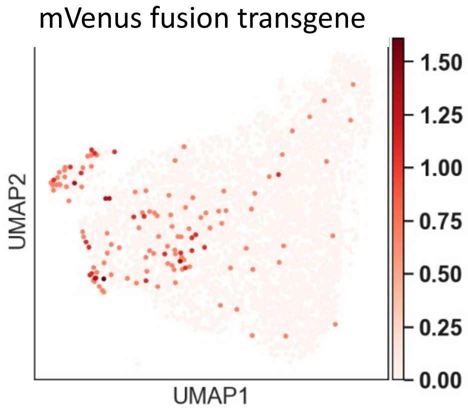
Gigascience, Volume 11, , 2022, giac001, https://doi.org/10.1093/gigascience/giac001

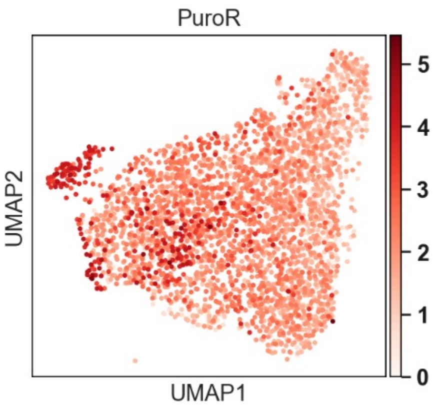
Custom Genomes w/ `cellranger mkref`

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

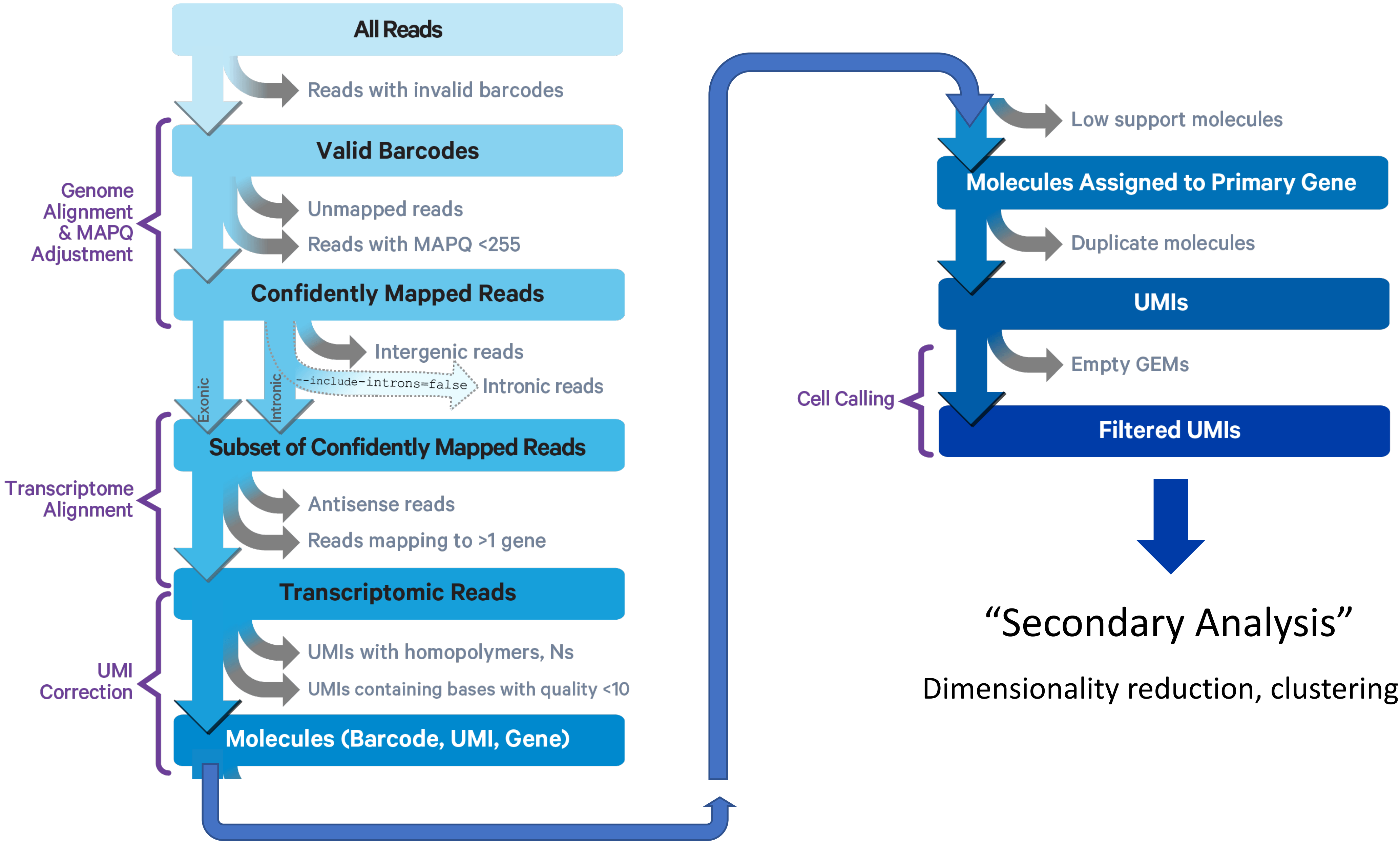
- 10X Provides Prebuilt references for:
  - Human (hg19 and GRCh38)
  - Mouse (mm9 and mm10)
- Why do I need a new reference genome?
  - Expanded annotations (eg. GENCODE, ncRNAs, etc.)
  - Additional species (eg. Maize)
  - Monitoring custom transgenes
  - Viruses, pathogens
  - known rearrangements,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







Overview of `cellranger count`



All Reads

Reads with invalid barcodes

Valid Barcodes

Genome Alignment & MAPQ Adjustment

Unmapped reads

Reads with MAPQ <255

Confidently Mapped Reads

Intergenic reads

Exonic

Intronic

--include-introns=false

Intronic reads

Subset of Confidently Mapped Reads

Transcriptome Alignment

Antisense reads

Reads mapping to >1 gene

Transcriptomic Reads

UMI Correction

UMIs with homopolymers, Ns

UMIs containing bases with quality <10

Molecules (Barcode, UMI, Gene)

Low support molecules

Molecules Assigned to Primary Gene

Duplicate molecules

UMIs

Cell Calling

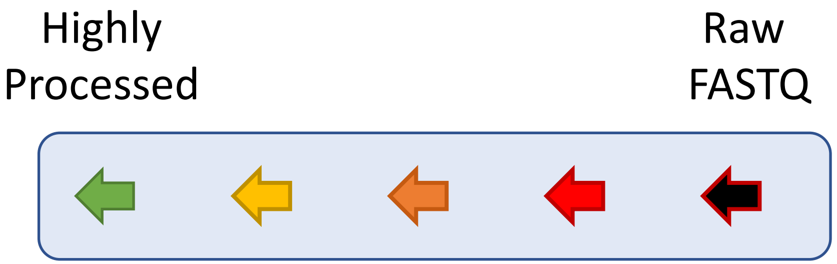
Empty GEMs

Filtered UMIs

“Secondary Analysis”

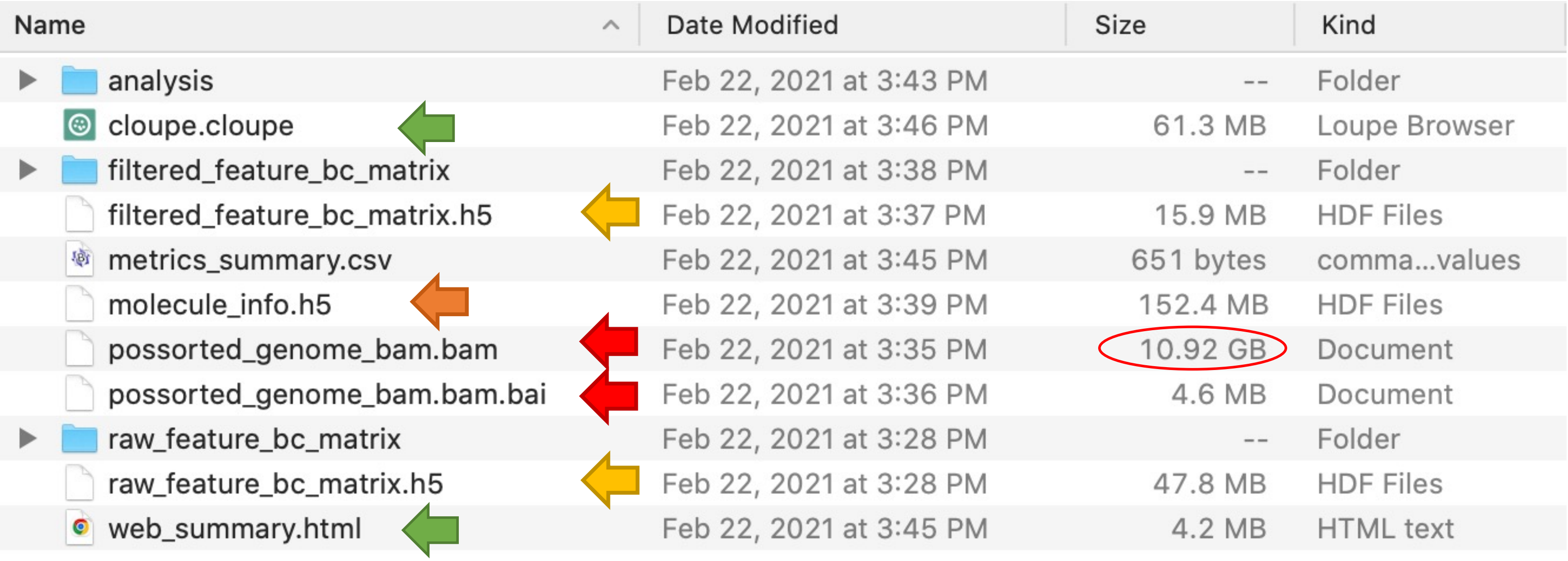
Dimensionality reduction, clustering, etc

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

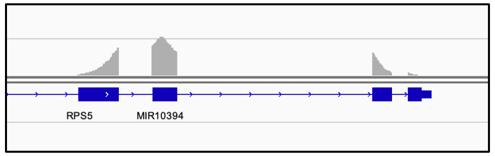
BAM file

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
AAACCATTTGGTCCCTTTCTTTTTTTTTTTTTTTTTTTTTTTTTTGGGTGGGAGAC
EE//A//////////////6EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAAAAA 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:AAAAAEEEEEEEEEEE CB:Z:AGCTTCCTCTTCCCGA-1 UR:Z:AGTTATTCCCAA UY:Z:EEEEEEEEEEEE
UB:Z:AGTTATTCCCAA
```

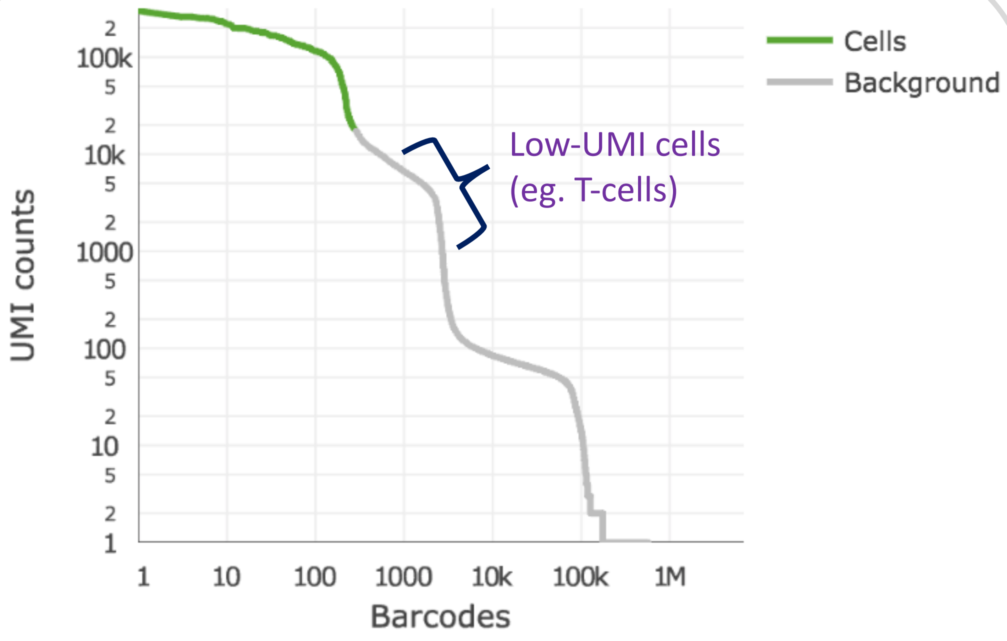
- Stores alignment features, cell barcode, overlapping genes, UMI
- Contains complete record of sequencing data
- FASTQs can be faithfully recreated (eg. for publication) using `bam2fastq`
- Can be viewed as a browser track



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

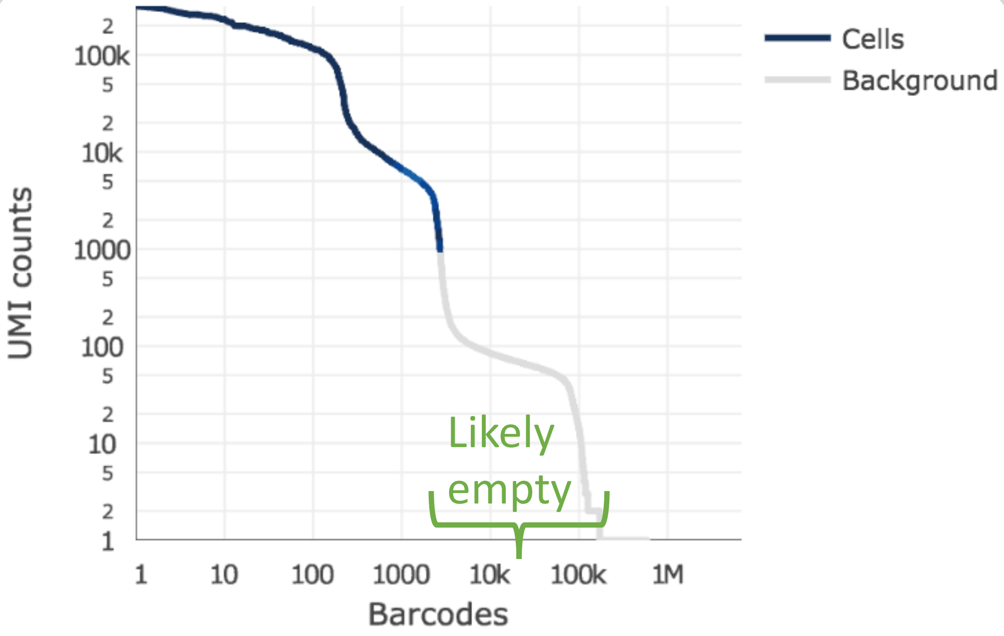
Cell Calling Algorithm

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



Low-UMI cells (eg. T-cells)

Cellranger 2



Likely empty

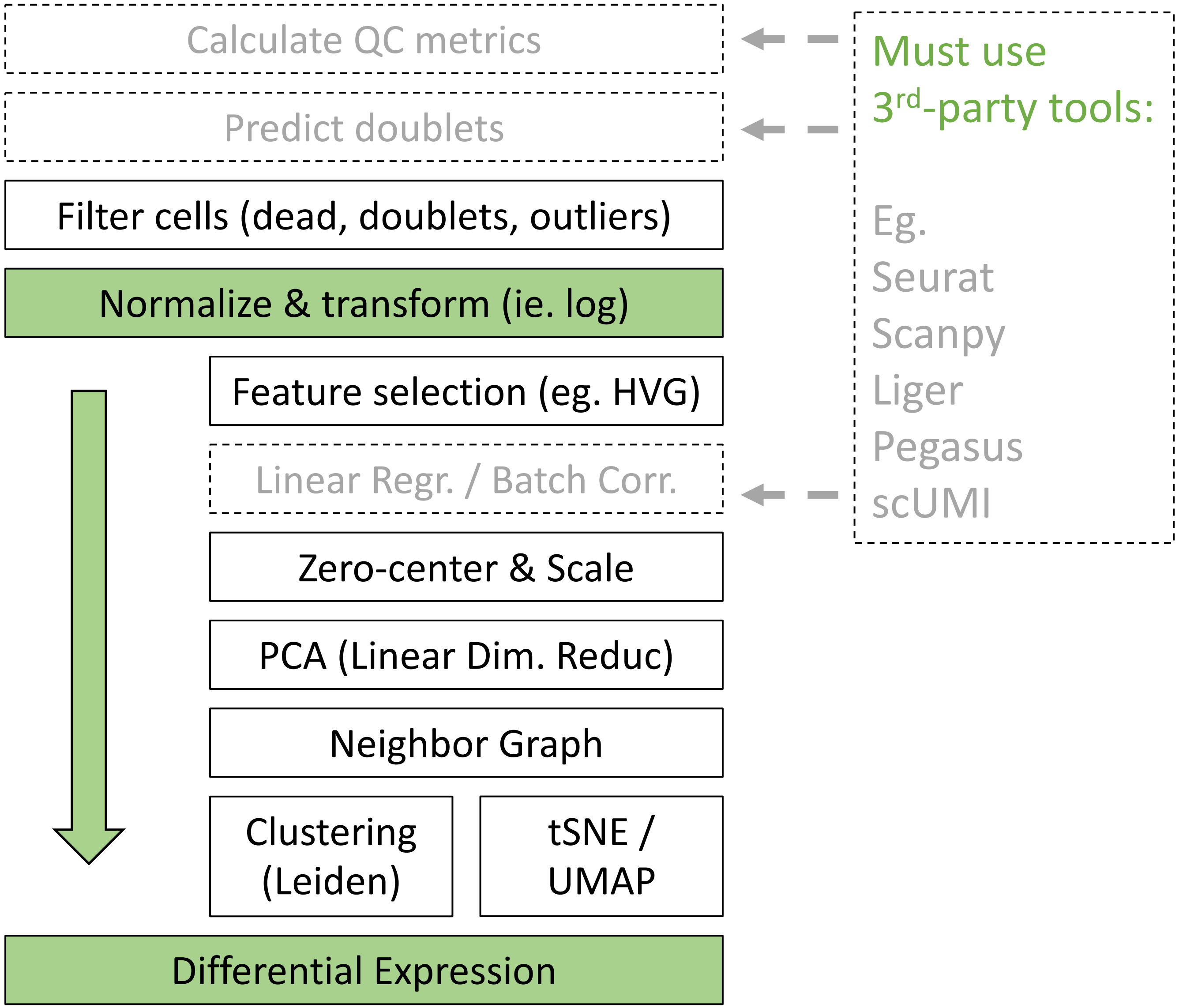
Cellranger 3+

- 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”

Secondary Analysis

Cell Ranger



Calculate QC metrics

Predict doublets

Filter cells (dead, doublets, outliers)

Normalize & transform (ie. log)

Feature selection (eg. HVG)

Linear Regr. / Batch Corr.

Zero-center & Scale

PCA (Linear Dim. Reduc)

Neighbor Graph

Clustering (Leiden)

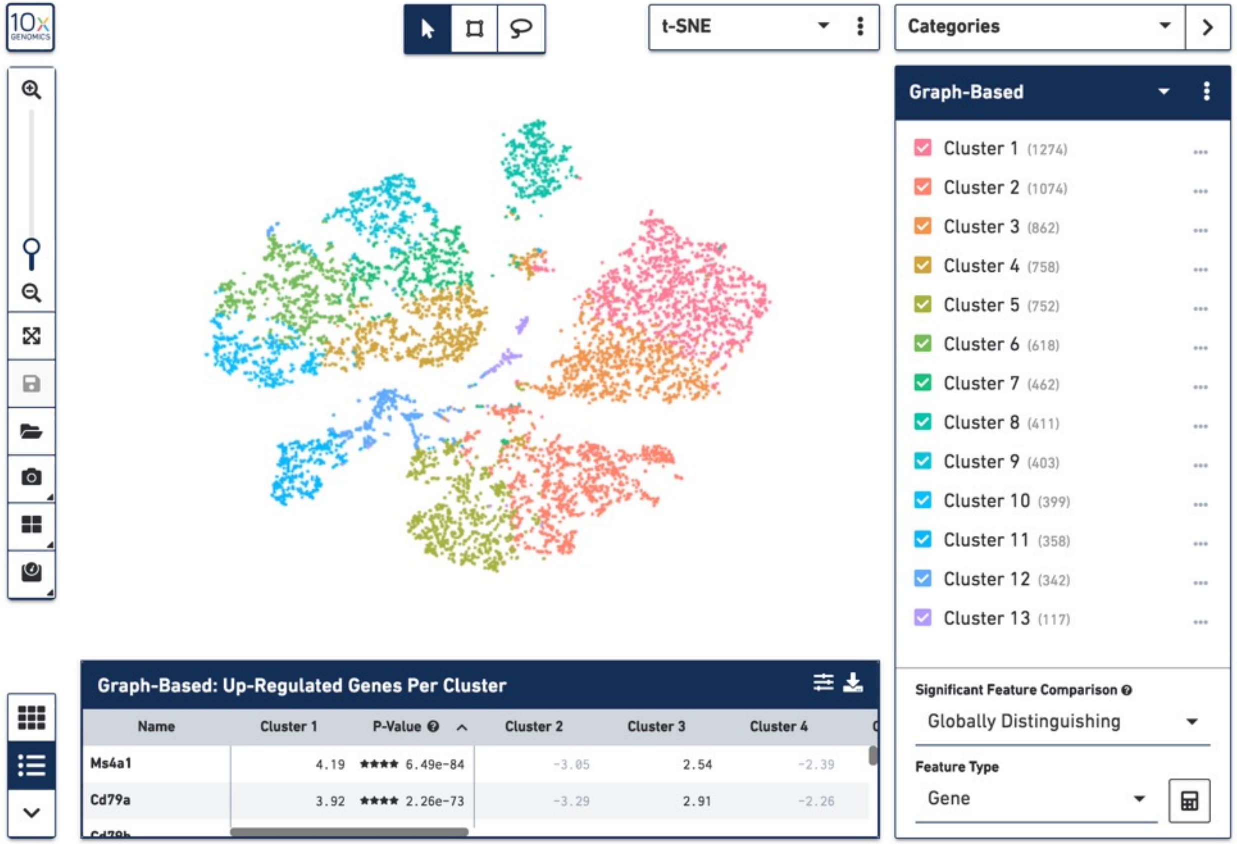
tSNE / UMAP

Differential Expression

Must use 3rd-party tools:

Eg.
Seurat
Scanpy
Liger
Pegasus
scUMI

Loupe Browser



t-SNE

Categories

Graph-Based

- ☑ Cluster 1 (1274)
- ☑ Cluster 2 (1074)
- ☑ Cluster 3 (862)
- ☑ Cluster 4 (758)
- ☑ Cluster 5 (752)
- ☑ Cluster 6 (618)
- ☑ Cluster 7 (462)
- ☑ Cluster 8 (411)
- ☑ Cluster 9 (403)
- ☑ Cluster 10 (399)
- ☑ Cluster 11 (358)
- ☑ Cluster 12 (342)
- ☑ Cluster 13 (117)

Significant Feature Comparison: Globally Distinguishing

Feature Type: Gene

Graph-Based: Up-Regulated Genes Per Cluster

Name	Cluster 1	P-Value	Cluster 2	Cluster 3	Cluster 4
Ms4a1	4.19	**** 6.49e-84	-3.05	2.54	-2.39
Cd79a	3.92	**** 2.26e-73	-3.29	2.91	-2.26
Cd79b					

Can:

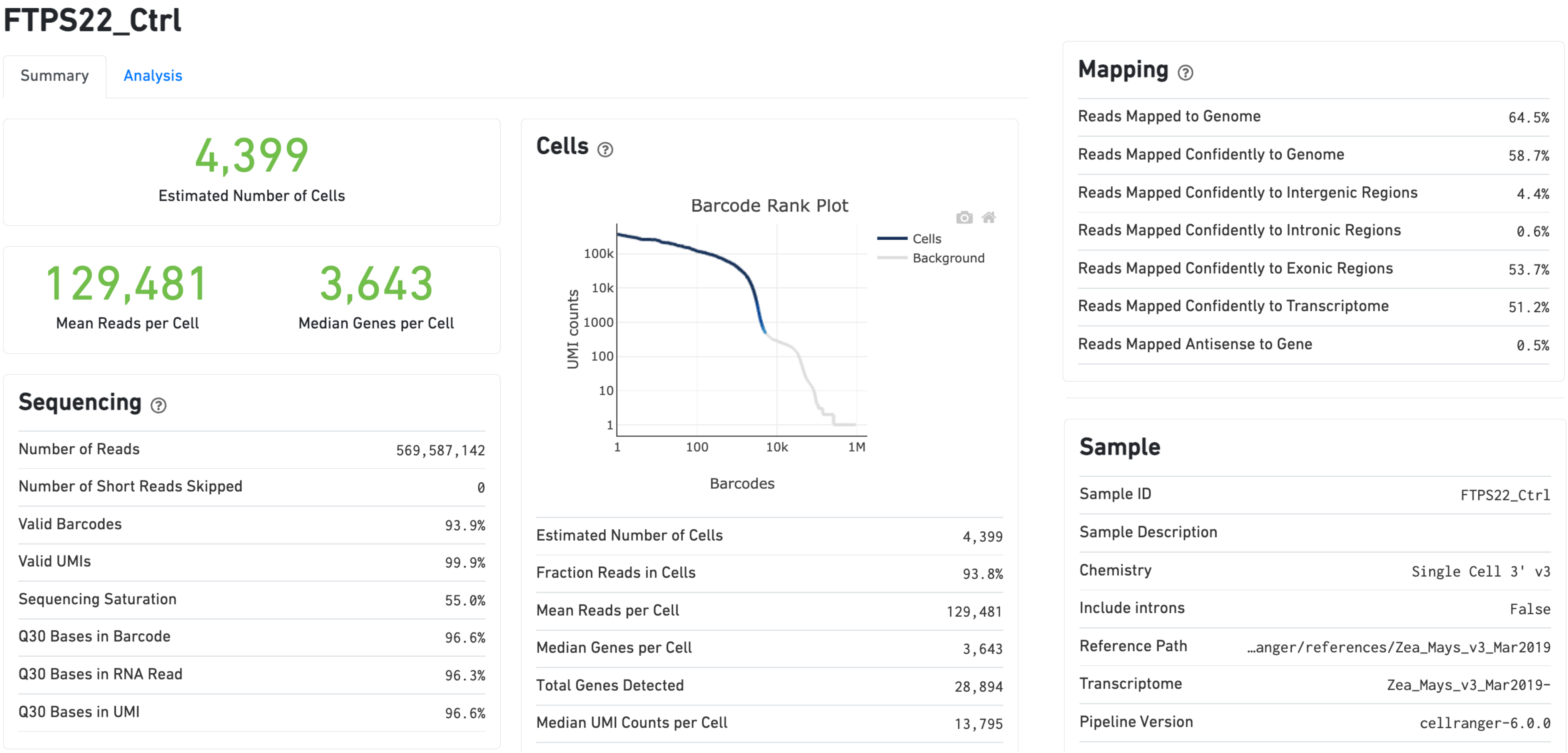
- Quickly visualize genes
- Do guided clustering via marker genes / tSNE selections
- Calculate Differential Expression
- Export cells and gene sets for reanalysis on Cellranger (cluster)

Can't

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

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

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%

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

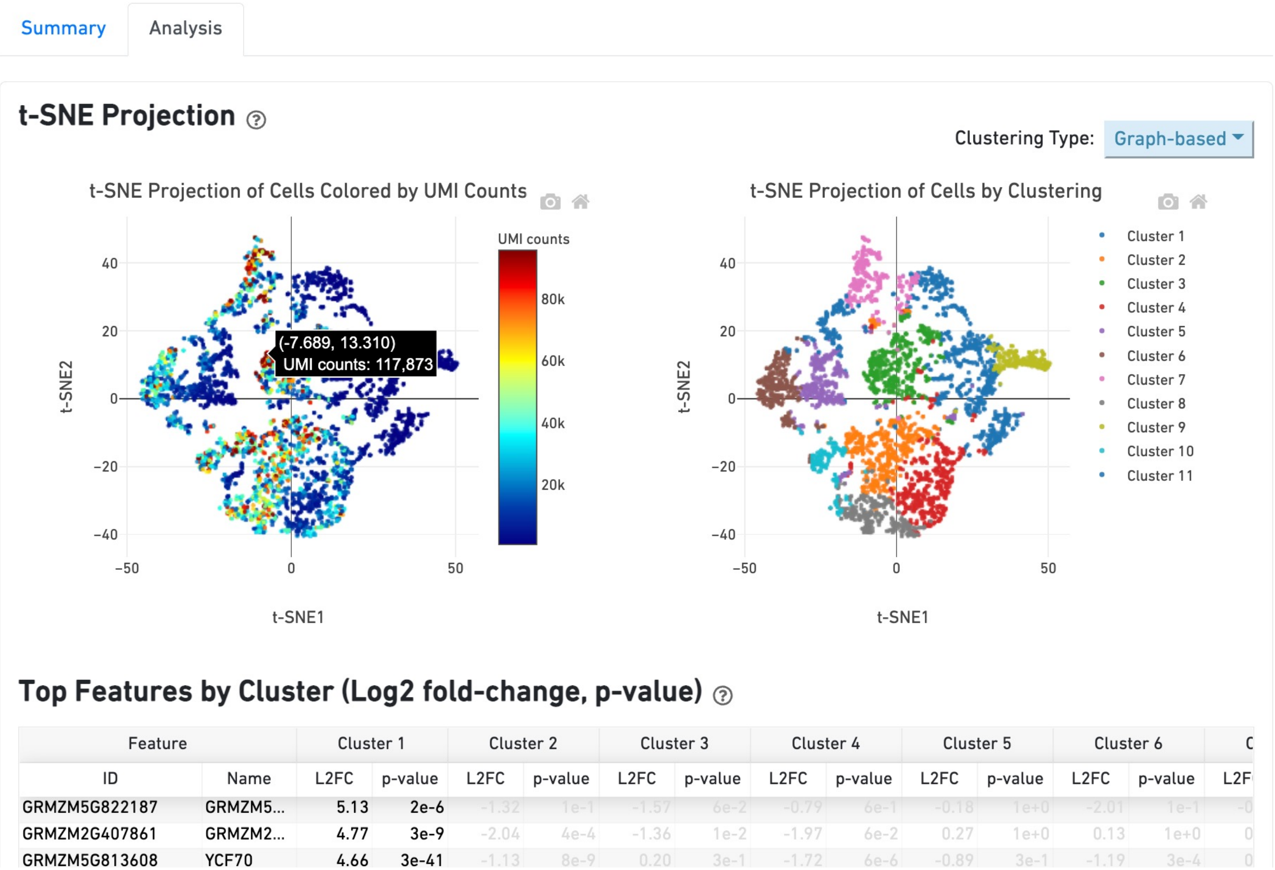
Mapping

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

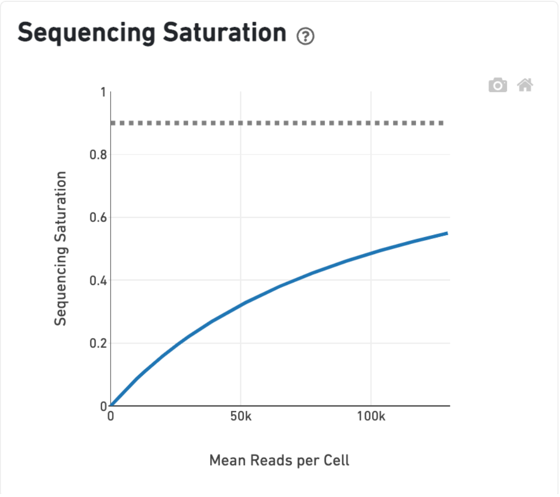
Sample ID	FTPS22_Ctrl
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	False
Reference Path	…anger/references/Zea_Mays_v3_Mar2019
Transcriptome	Zea_Mays_v3_Mar2019-
Pipeline Version	cellranger-6.0.0

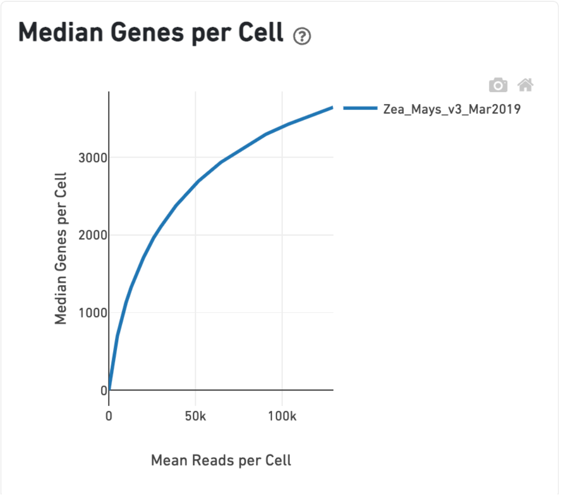
Quality Control: web_summary.html



Top Features by Cluster (Log2 fold-change, p-value)

Feature		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		C
ID	Name	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2F
GRMZM5G822187	GRMZM5...	5.13	2e-6	-1.32	1e-1	-1.57	6e-2	-0.79	6e-1	-0.18	1e+0	-2.01	1e-1	-0
GRMZM2G407861	GRMZM2...	4.77	3e-9	-2.04	4e-4	-1.36	1e-2	-1.97	6e-2	0.27	1e+0	0.13	1e+0	0
GRMZM5G813608	YCF70	4.66	3e-41	-1.13	8e-9	0.20	3e-1	-1.72	6e-6	-0.89	3e-1	-1.19	3e-4	0

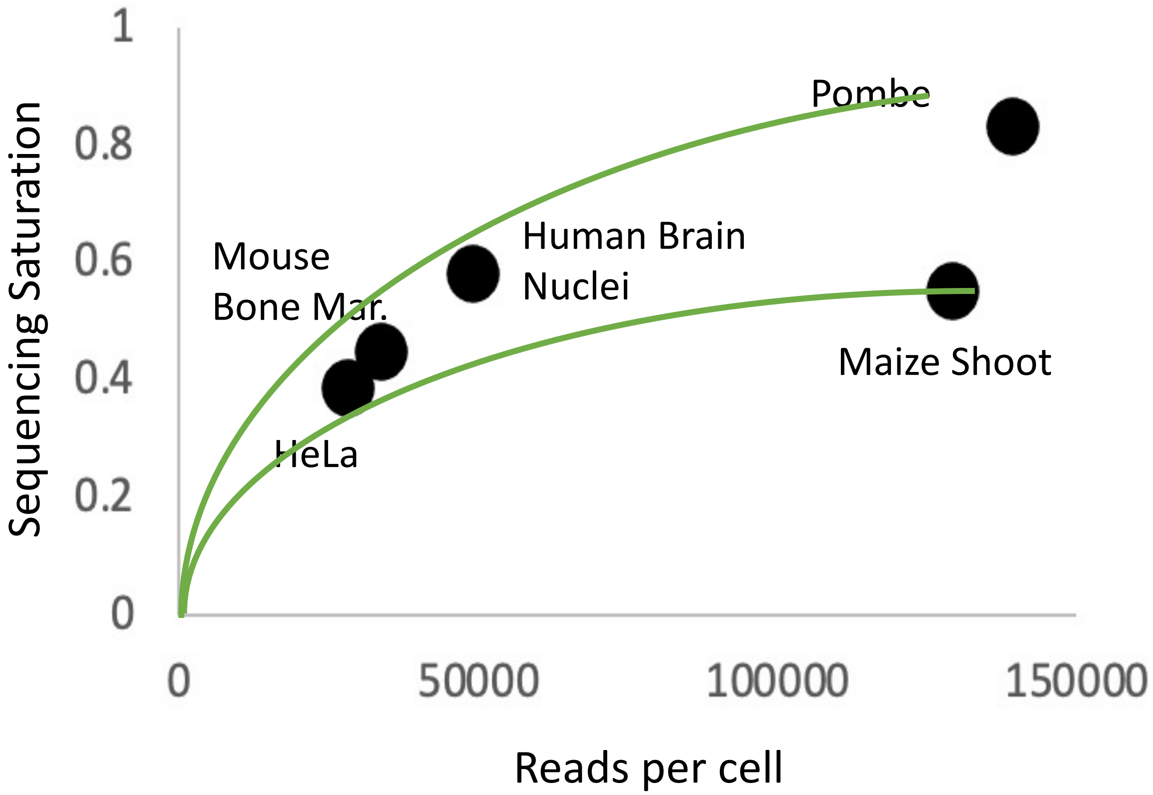




Putting up the numbers

	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:

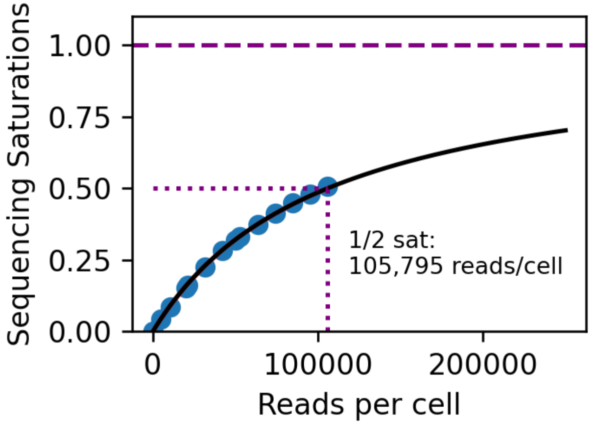
Sample Composition:

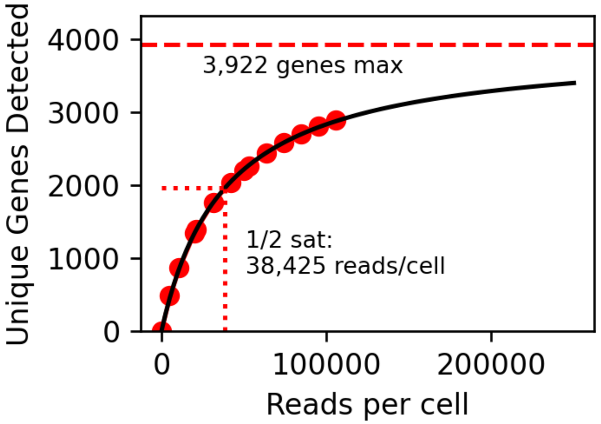
- # of cells
- # of UMIs/cell

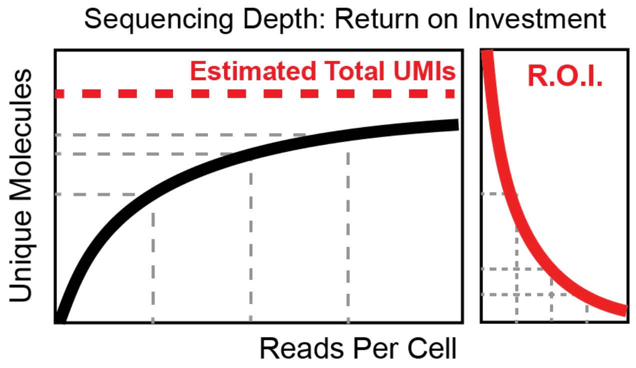
Library Quality:

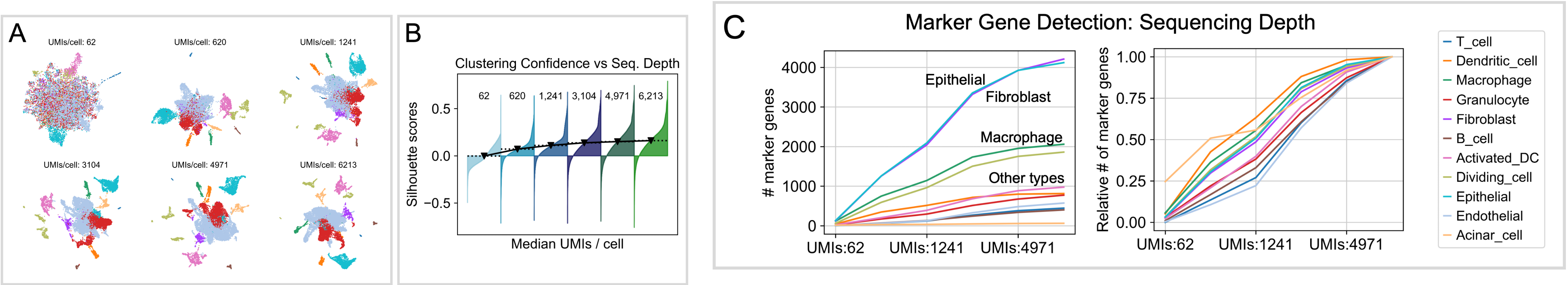
- PCR duplication
- Adapter dimers
- Contaminating DNA (mapping rate)

Sequencing Saturation – How deep?









Regan and Preall, Current Protocols, 2022

Combine samples with cellranger aggr

First, make an aggr.csv file

```
sample_id,molecule_h5
SeqTech22_RNA_10k,/fake/path/SeqTech22/count/SeqTech22_RNA_10k/outs/molecule_info.h5
SeqTech22_RNA_12k,/fake/path/SeqTech22/count/SeqTech22_RNA_12k/outs/molecule_info.h5
```

```
PROJECTDIR=/fake/path/SeqTech22/

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.

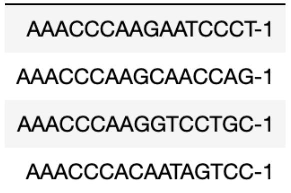
--normalize=none simply concatemerizes the two matrices together, and gives each samples a unique barcode suffix, in the order they appear in the aggr.csv file (-1, -2, -3, etc)

Matrix format: HDF5 vs MTX (MEX)

Market Exchange (MEX) Format

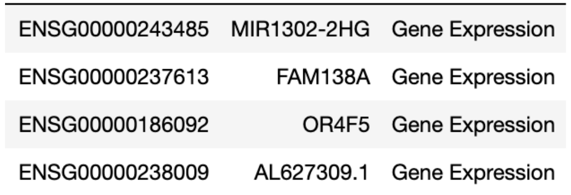
- Simple, deprecation-resistant
- Slow, lazy about metadata

barcodes.tsv



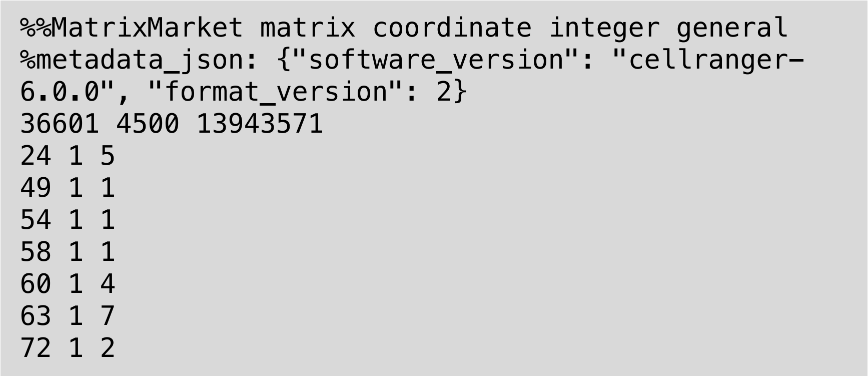
```
AAACCCAAGAATCCCT-1
AAACCCAAGCAACCAG-1
AAACCCAAGGTCCTGC-1
AAACCCACAATAGTCC-1
```

features.tsv



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

matrix.mtx

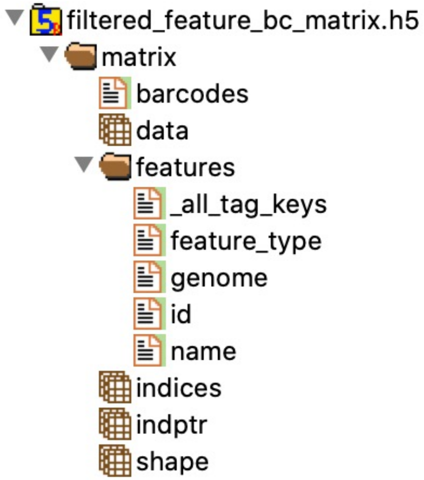


```
%%MatrixMarket matrix coordinate integer general
%metadata_json: {"software_version": "cellranger-
6.0.0", "format_version": 2}
36601 4500 13943571
24 1 5
49 1 1
54 1 1
58 1 1
60 1 4
63 1 7
72 1 2
```

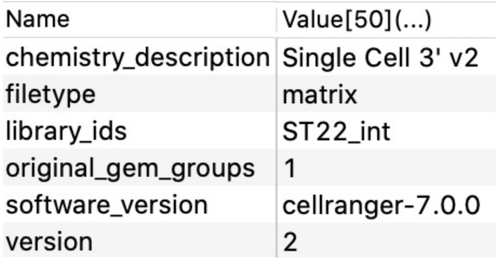
Both store data in SPARSE format

Hierarchical Data Format 5 (HDF5)

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



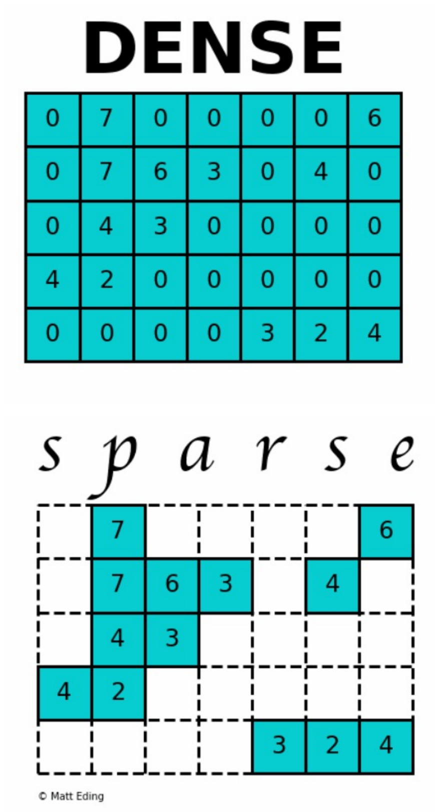
```
▼ filtered_feature_bc_matrix.h5
  ▼ matrix
      barcodes
      data
    ▼ features
        _all_tag_keys
        feature_type
        genome
        id
        name
      indices
      indptr
      shape
```



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



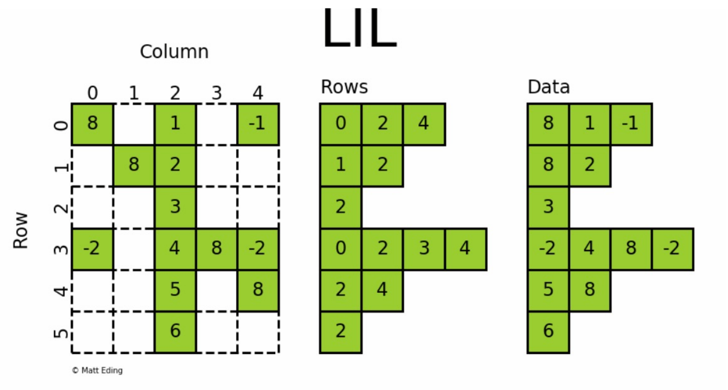
DENSE

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

sparse

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

© Matt Eding

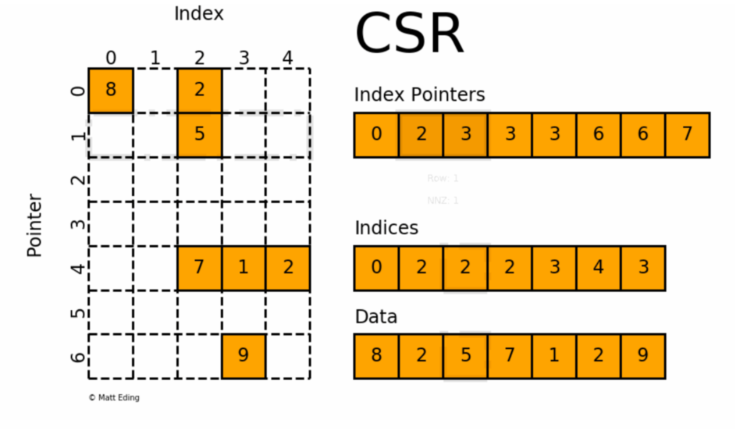


LIL

Row \ Column	0	1	2	3	4
0	8		1		-1
1		8	2		
2			3		
3	-2		4	8	-2
4			5		8
5			6		

Rows				Data			
0	2	4		8	1	-1	
1	2			8	2		
2				3			
0	2	3	4	-2	4	8	-2
2	4			5	8		
2				6			

© Matt Eding



CSR

Pointer \ Index	0	1	2	3	4
0	8		2		
1			5		
2					
3					
4			7	1	2
5					
6				9	

Index Pointers

0	2	3	3	3	6	6	7

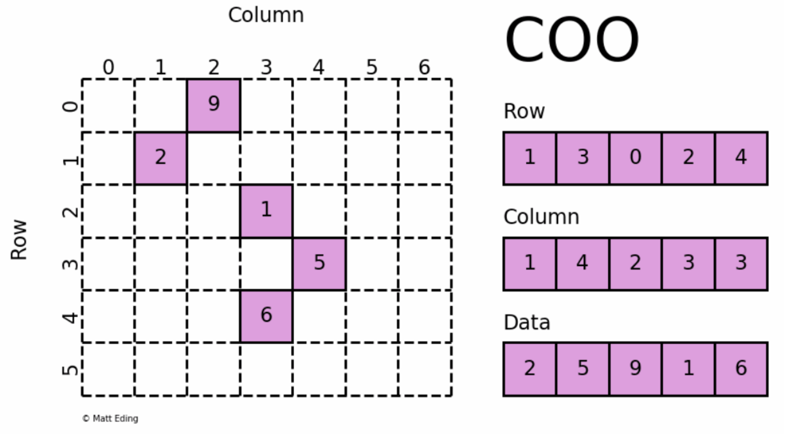
Indices

0	2	2	2	3	4	3

Data

8	2	5	7	1	2	9

© Matt Eding



COO

Row \ Column	0	1	2	3	4	5	6
0			9				
1		2					
2				1			
3					5		
4				6			
5							

Row	1	3	0	2	4
Column	1	4	2	3	3
Data	2	5	9	1	6

© 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

https://matteding.github.io/2019/04/25/sparse-matrices/

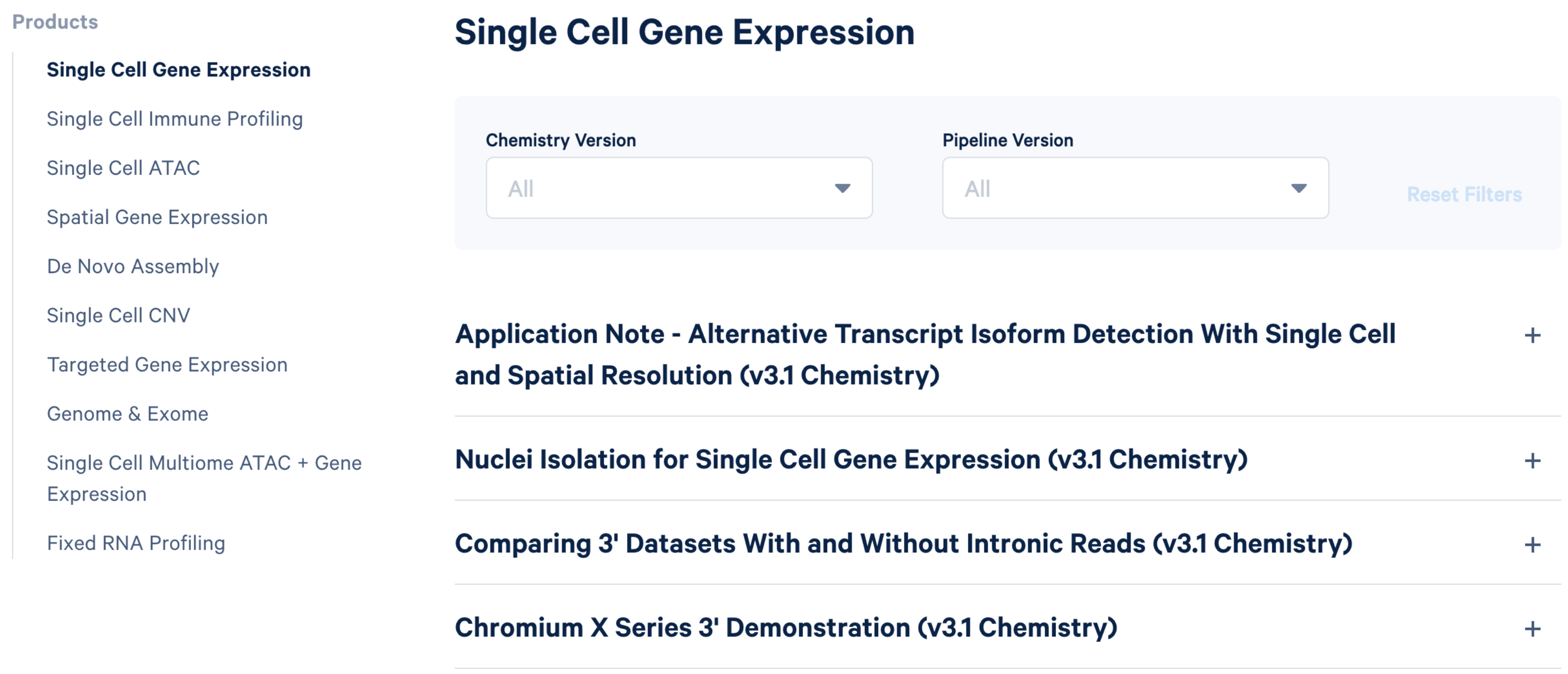
Cell Ranger’s Report Card

Calculate QC metrics	A-
Predict doublets	F
Filter cells (dead, doublets, outliers)	F
Normalize & transform (ie. log)	B
Feature selection (eg. HVG)	D
Regress unwated variation	F
PCA (Linear Dim. Reduc)	A
Neighbor Graph	A
Clustering (Leiden)	B
tSNE / UMAP	A
Differential Expression	B+

Batch Correction	D
Cell Type Identification	F
Sharing Data with Biologist Colleages	A++
Quickly visualizing markers	A++
Quickly finding diffex. genes	A-

Demo Datasets

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



Single Cell Gene Expression

Chemistry Version: All

Pipeline Version: All

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 2017

- https://www.dropbox.com/sh/qksaunln69yrqd1/AAAKLZ4E-yyfhb5-eYSnvnnZa?dl=0

EXERCISE 2

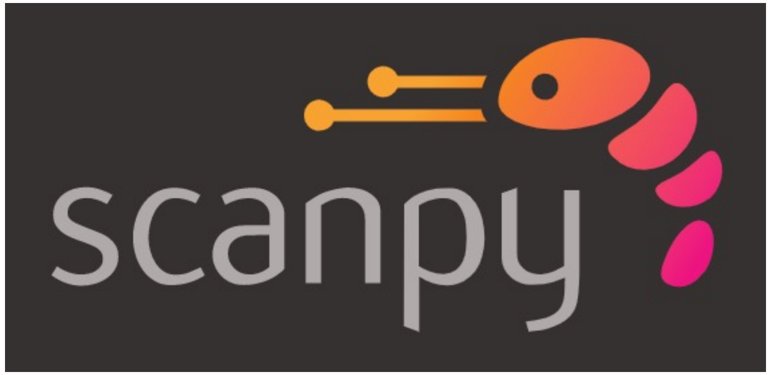
- Step 1: Export the category 'Protospacer Per Tag' using Loupe Browser

- Step 2: Modify the CSV file to collapse sgRNAs by gene target.
- eg. merge ACE2-1 and ACE2-2 into a single category, 'ACE2'

- Step 3: Import back into Loupe and run differential expression on sgRNAs

Secondary Analysis

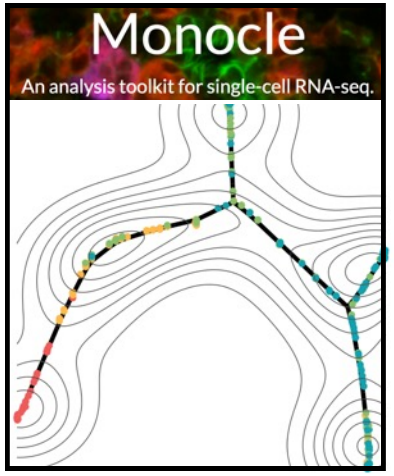
Fabian Theis - München



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

Python

R



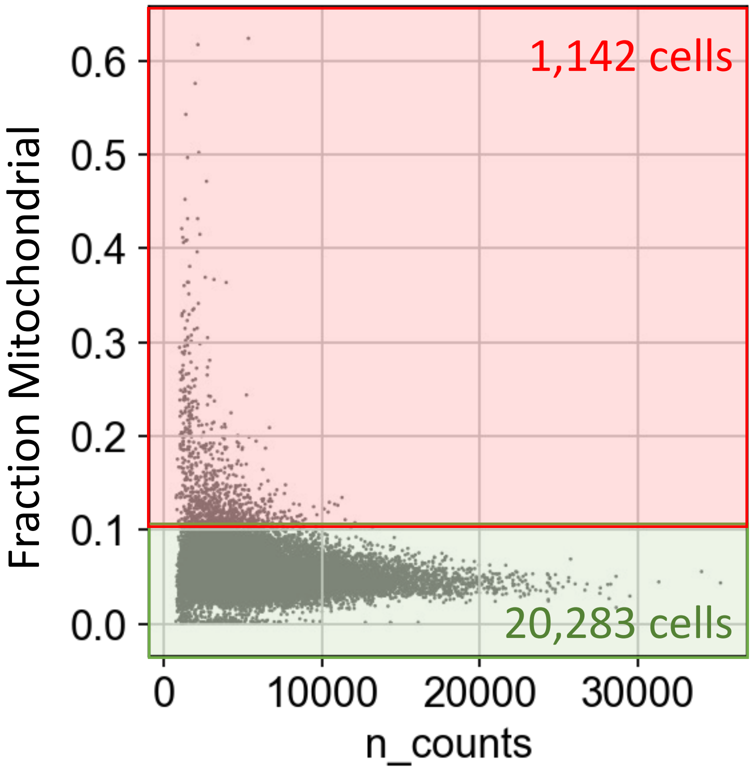
Cole Trapnell –WashU

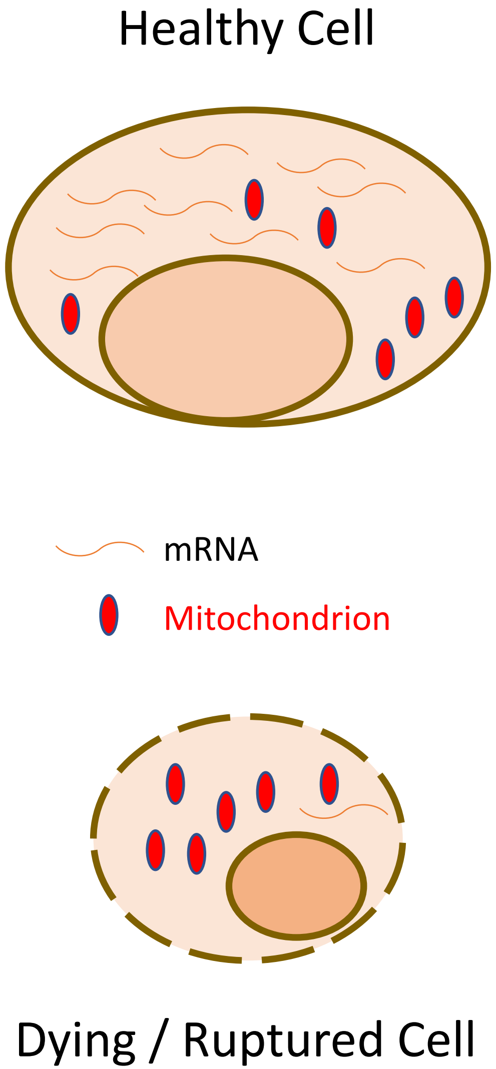
Liger **R**

Macosko lab

QC metrics – Cell Death

Filter cells by Fraction Mitochondrial Reads



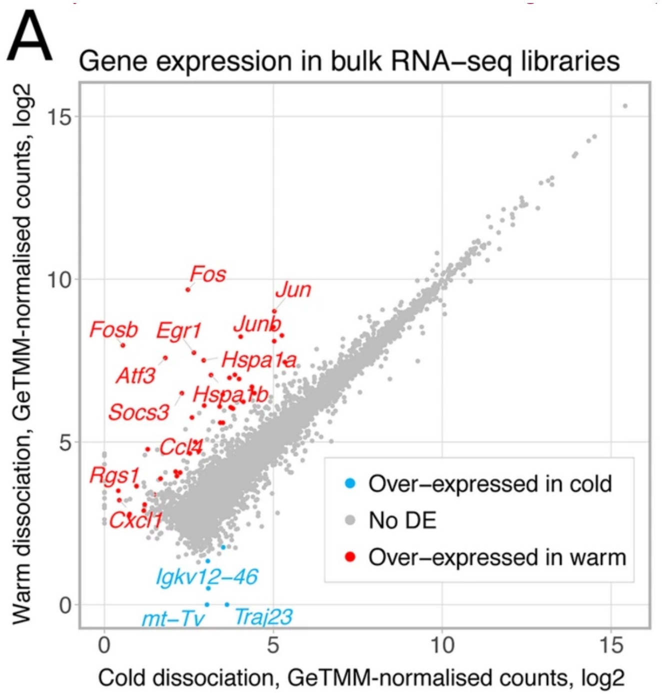

Healthy Cell

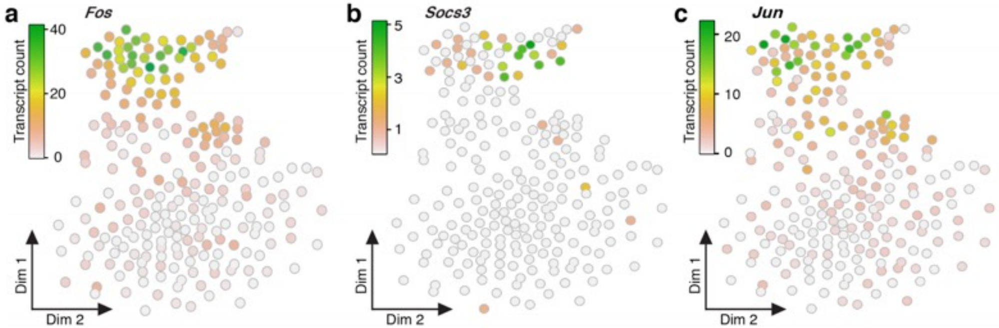
Dying / Ruptured Cell

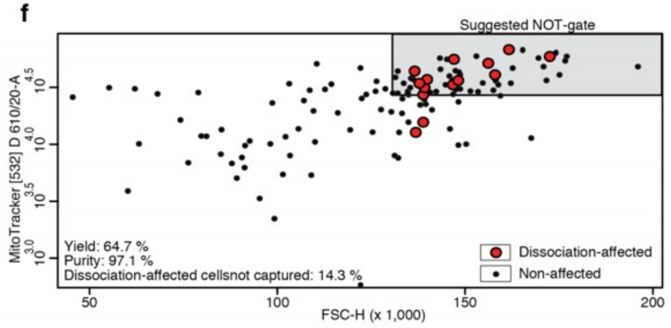
QC metrics – Dissociation Stress

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



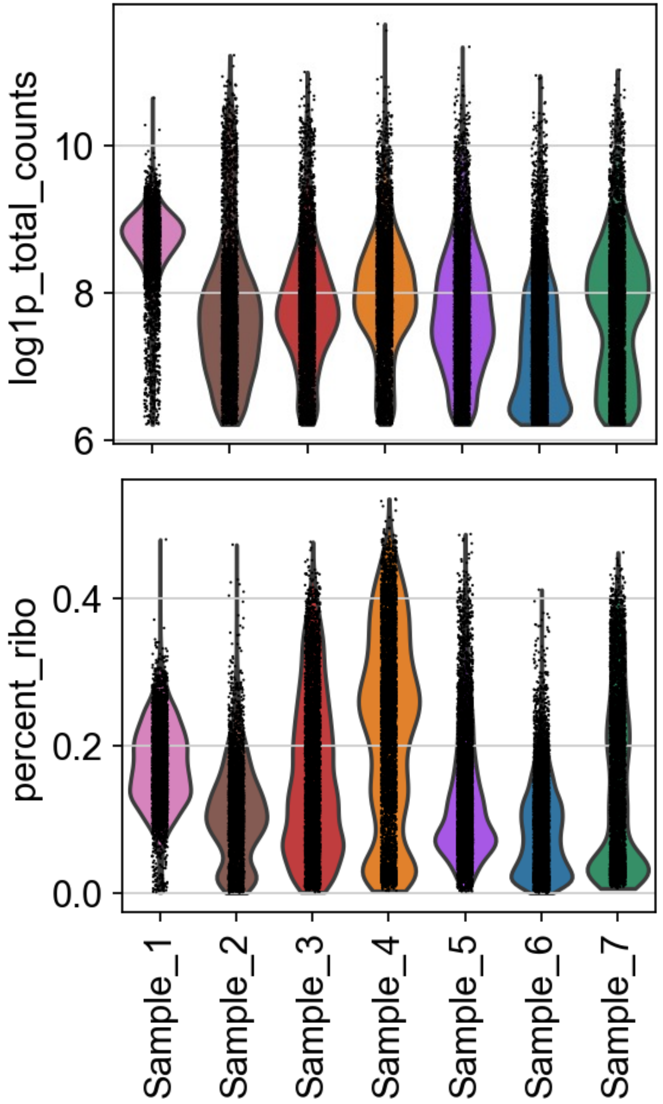


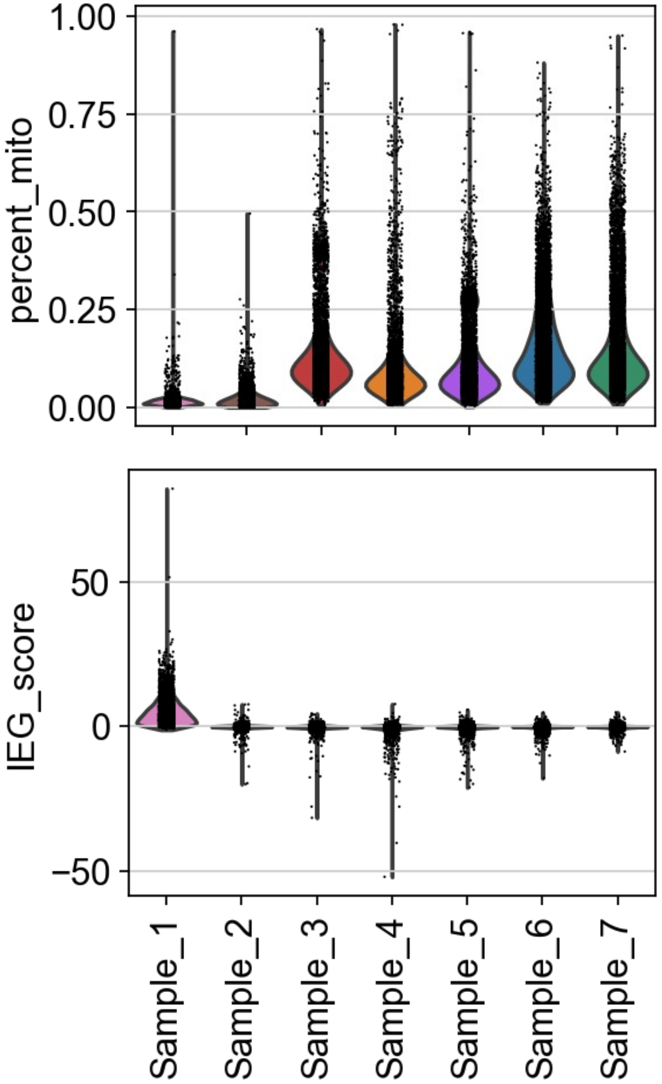


QC metrics don’t always correspond to artifacts

UMI counts:
Cell types
Viability
Sequencing depth
Library handling

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

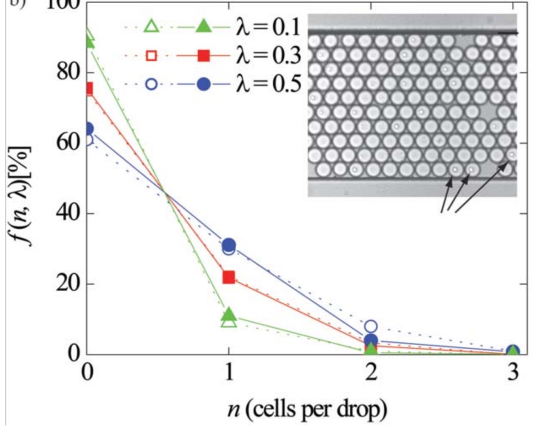




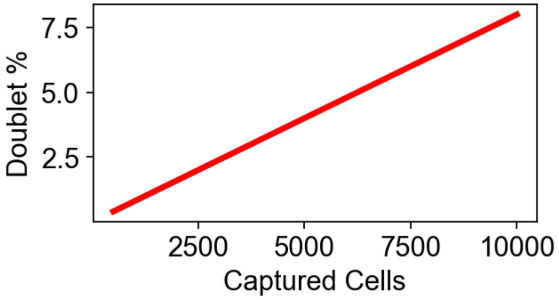
Mitochondrial %
Dying cells
Respiration activity!
(some cells have > 50% but are alive!)

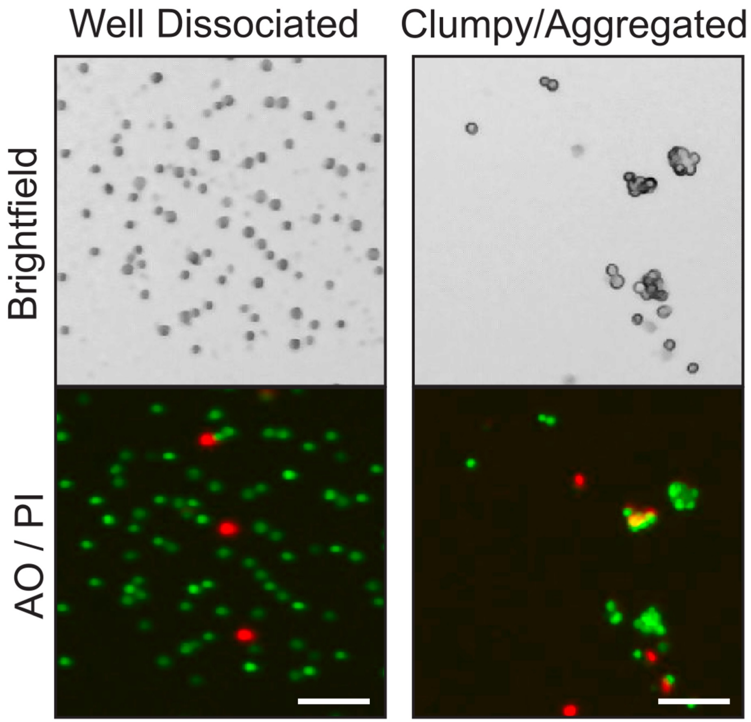
Immediate-Early Genes
Post-dissoc. stress
In vivo acute activation

Doublets / Multiplets



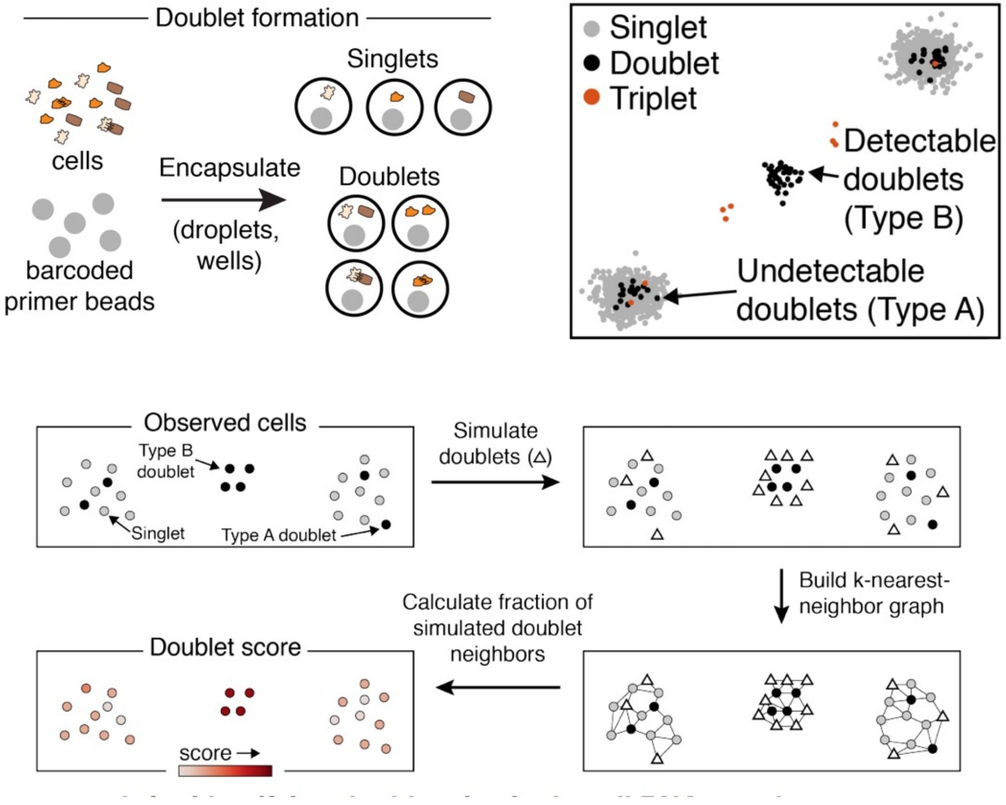
Köster…Weitz. *Lab on a Chip* (2008)





Doublet Filtering

Scrublet



Wolock et al. (2018) bioRxiv

- •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

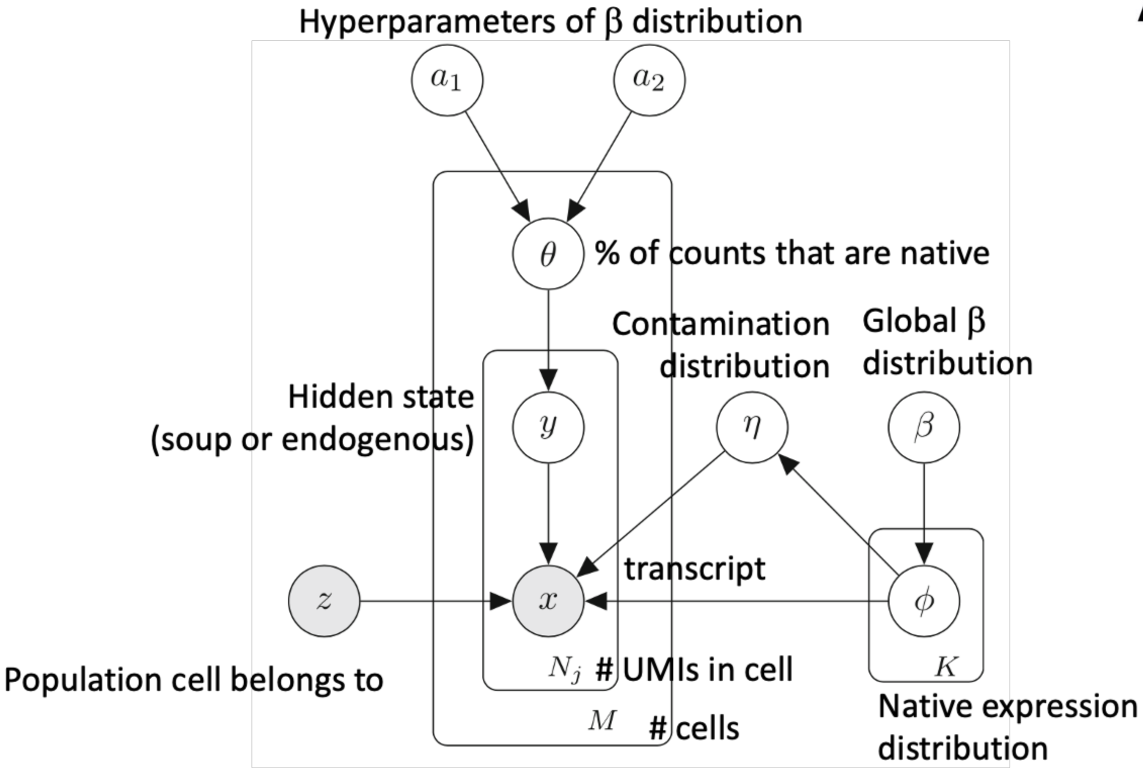
METHOD Open Access

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]*

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

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



Hyperparameters of β distribution

% of counts that are native

Contamination distribution

Global β distribution

Hidden state (soup or endogenous)

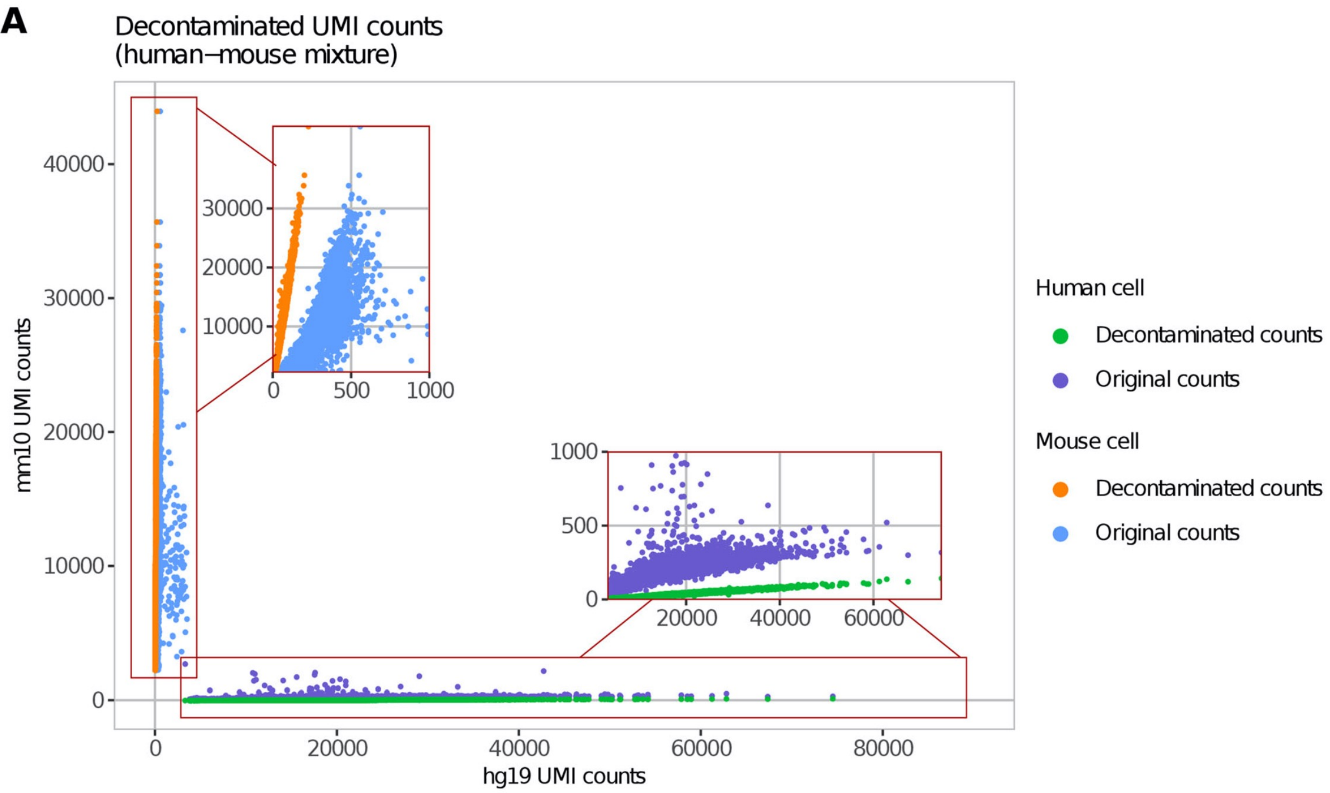
transcript

Population cell belongs to

UMIs in cell

cells

Native expression distribution



Normalization & log transformation

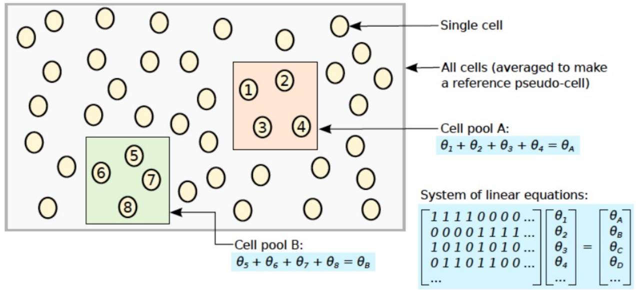
Basic Method: $$\frac{\text{Gene(i)Counts}}{\text{All Gene Counts}} * 10{,}000$$ (ie. transcripts per 10k, “tp10k”)

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.

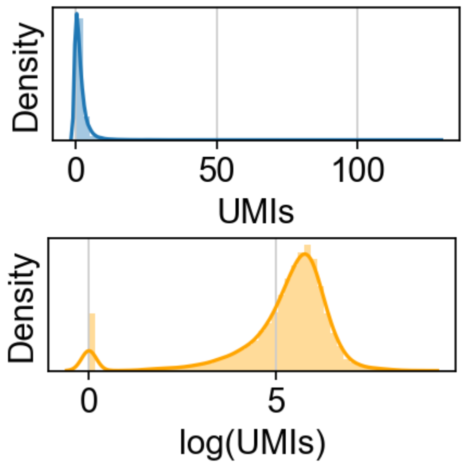


Single cell
All cells (averaged to make a reference pseudo-cell)
1
2
3
4
5
6
7
8
Cell pool A:
$\theta_1 + \theta_2 + \theta_3 + \theta_4 = \theta_A$
Cell pool B:
$\theta_5 + \theta_6 + \theta_7 + \theta_8 = \theta_B$
System of linear equations:
$\begin{bmatrix} 1\,1\,1\,1\,0\,0\,0\,0 \dots \\ 0\,0\,0\,0\,1\,1\,1\,1 \dots \\ 1\,0\,1\,0\,1\,0\,1\,0 \dots \\ 0\,1\,1\,0\,1\,1\,0\,0 \dots \\ \dots \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \dots \end{bmatrix} = \begin{bmatrix} \theta_A \\ \theta_B \\ \theta_C \\ \theta_D \\ \dots \end{bmatrix}$

Lun *et al.*, Genome Biology (2016)

1. Cluster cells together
2. Pool cells to increase counts, reduce 0’s
3. Robust estimate of each pool size factor
4. Wash & repeat for multiple pools
5. Solve the linear system of equations to obtain *per-cell* size factors

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

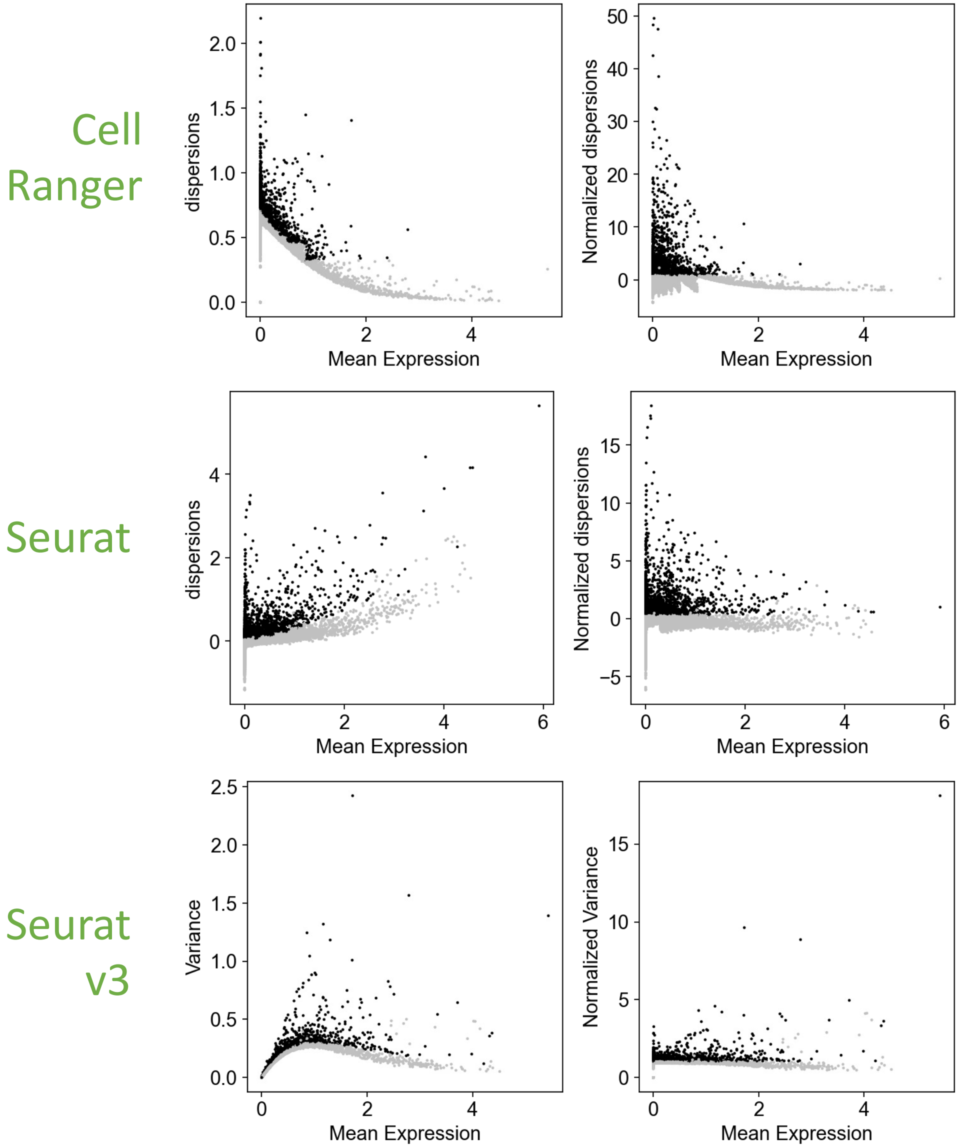


Add 1

Logarithmize

Basic variance stabilizing transformation
Other options:
Arcsin(h)
SCTransform (Seurat)
Pearson residuals

Feature Selection: Highly Variable Genes (HVGs)

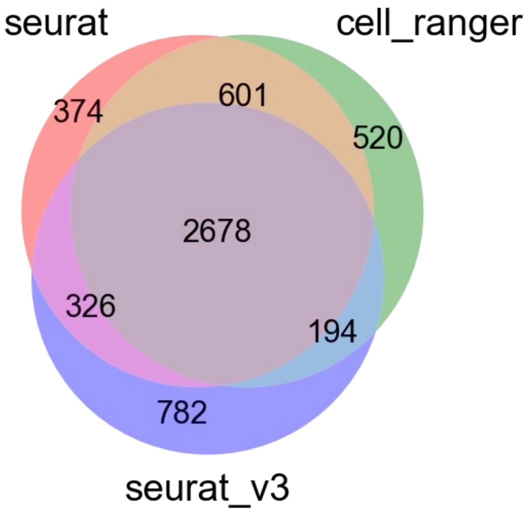


Assumption: genes that are interesting in the data will have higher variance

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

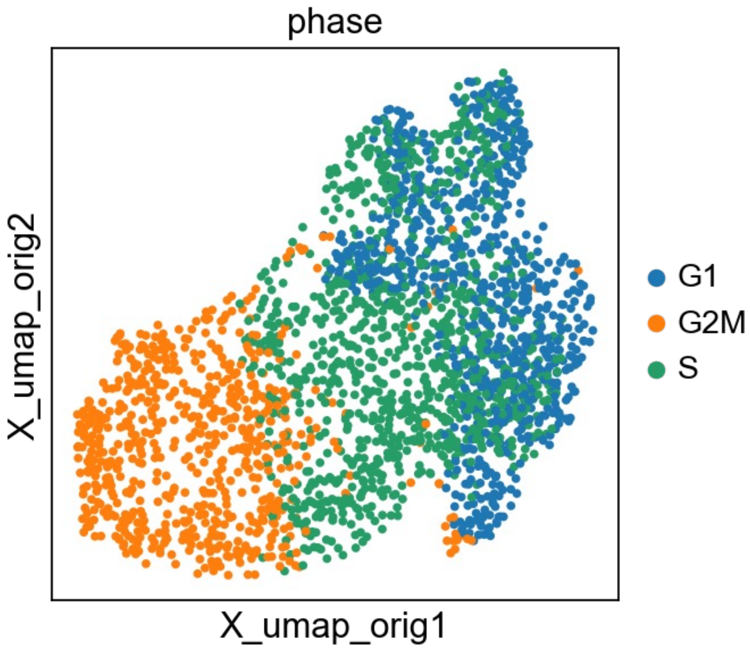
Thus, different approaches to select HVGs exist to select relevant HVGs to account for their expression ‘bin’

Alternatively: do PCA on ALL genes (computationally intensive)



Feature Selection: Supervised gene sets

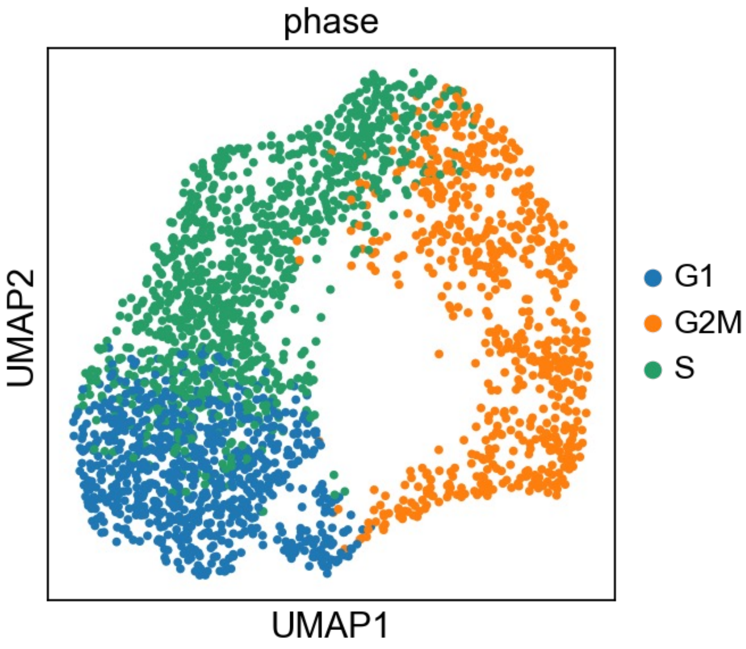
Siha cervical cancer cell line



529 curated cell-cycle genes

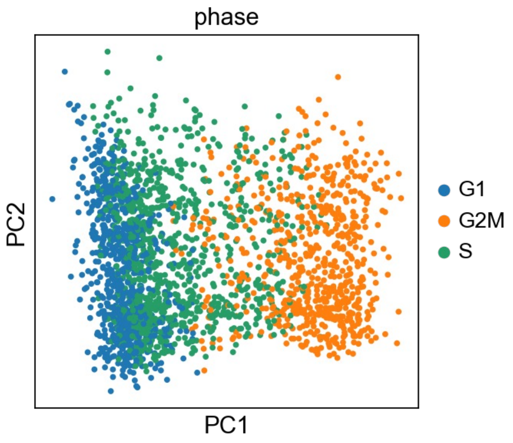
Macosko et al. (Cell) 2015 May 21; 161(5): 1202–1214.

PCA, UMAP



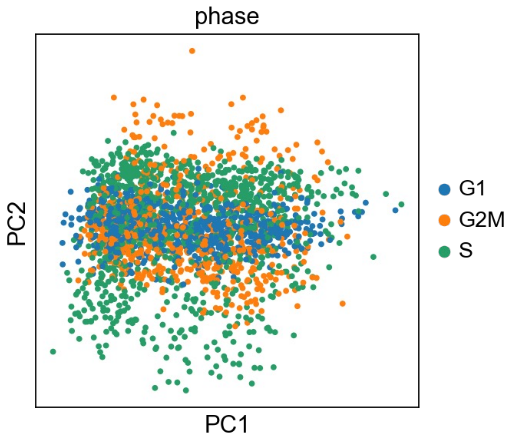
~1D circular manifold!

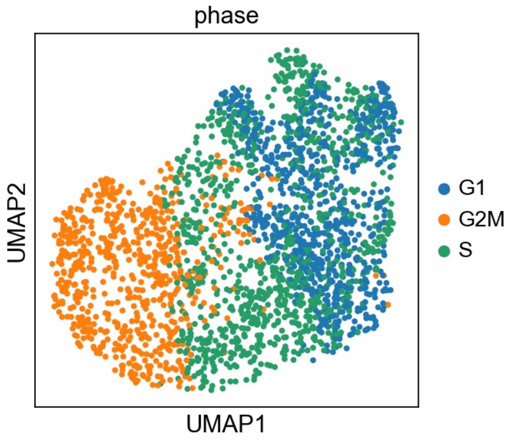
Linear Regression of Unwanted Variation

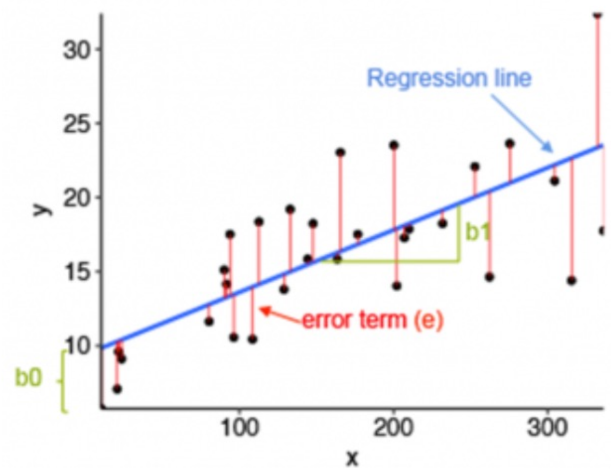


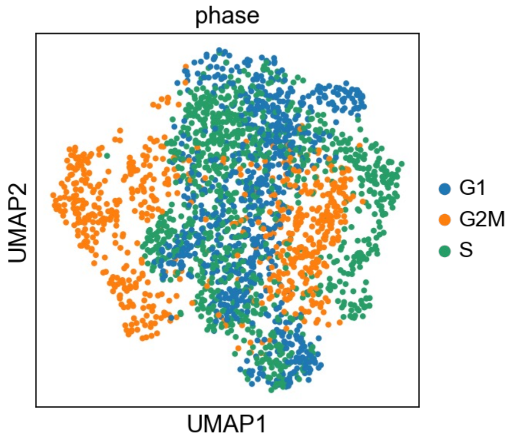
- Fit a linear model of each gene's expression vs confounder variable
    - Eg. cell cycle gene score, % mito, dissociation-stress response, library size

- Take model residuals as new scaled values
- This often doesn't work well.







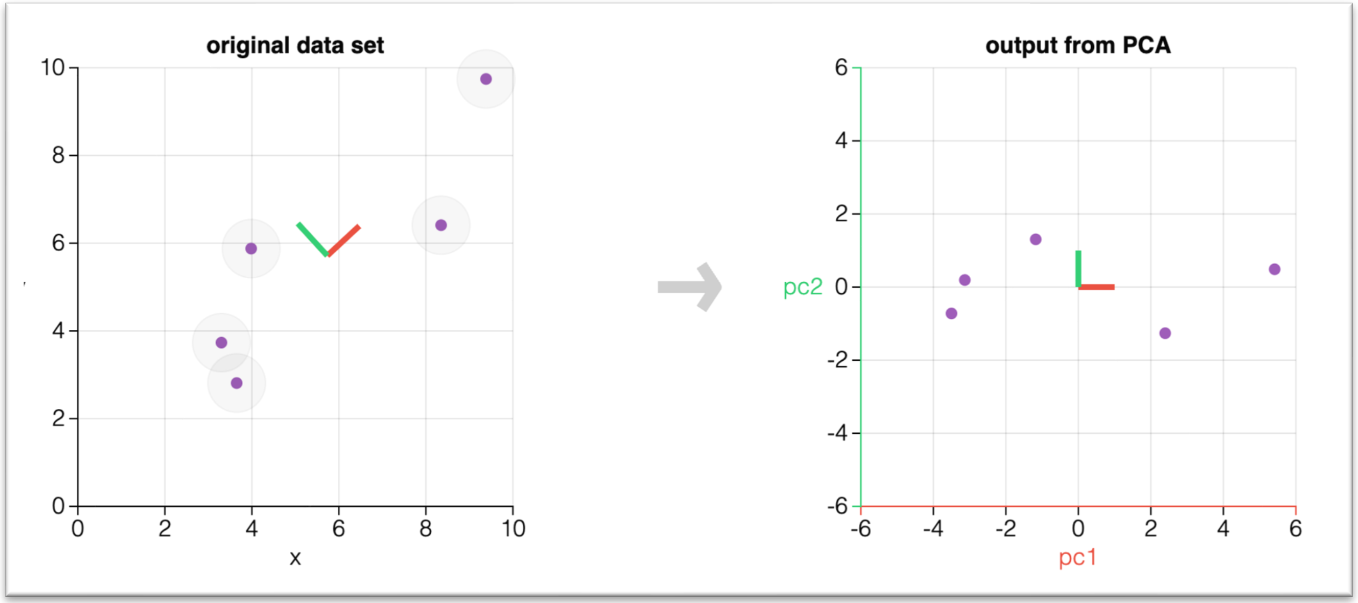


Principal Component Analysis

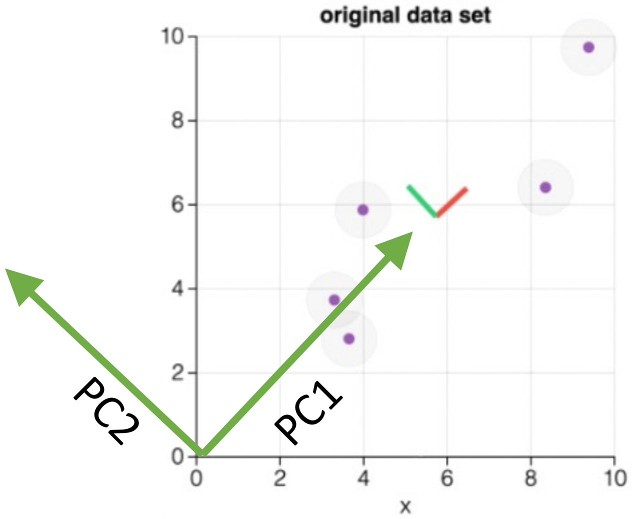
NOT a dimensionality reduction method, *per se.* It is a ROTATION.

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:

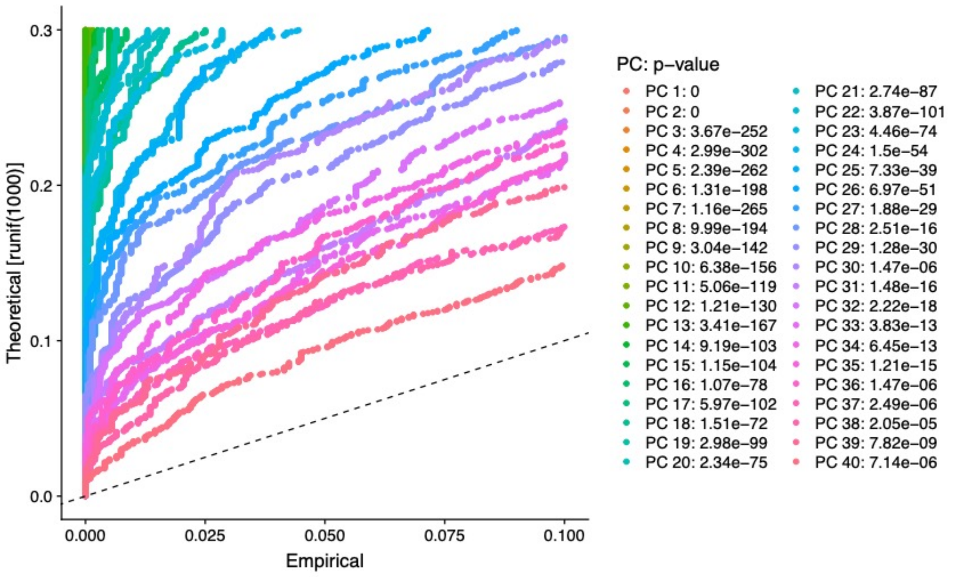
Jackstraw analysis: Determine “significant” PCs

Randomly permute data & recalculate PCs

Compare “Real” PCs with random noise.

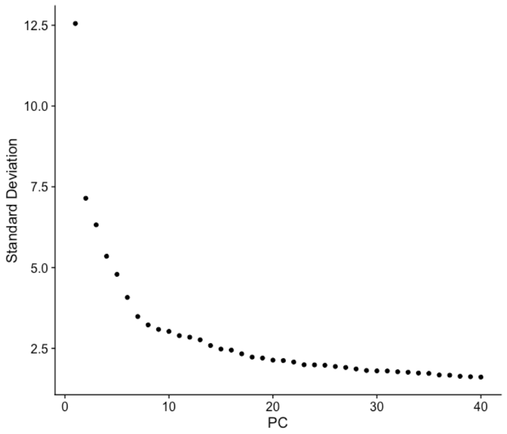
-slow

-still subject to weakness of PCA



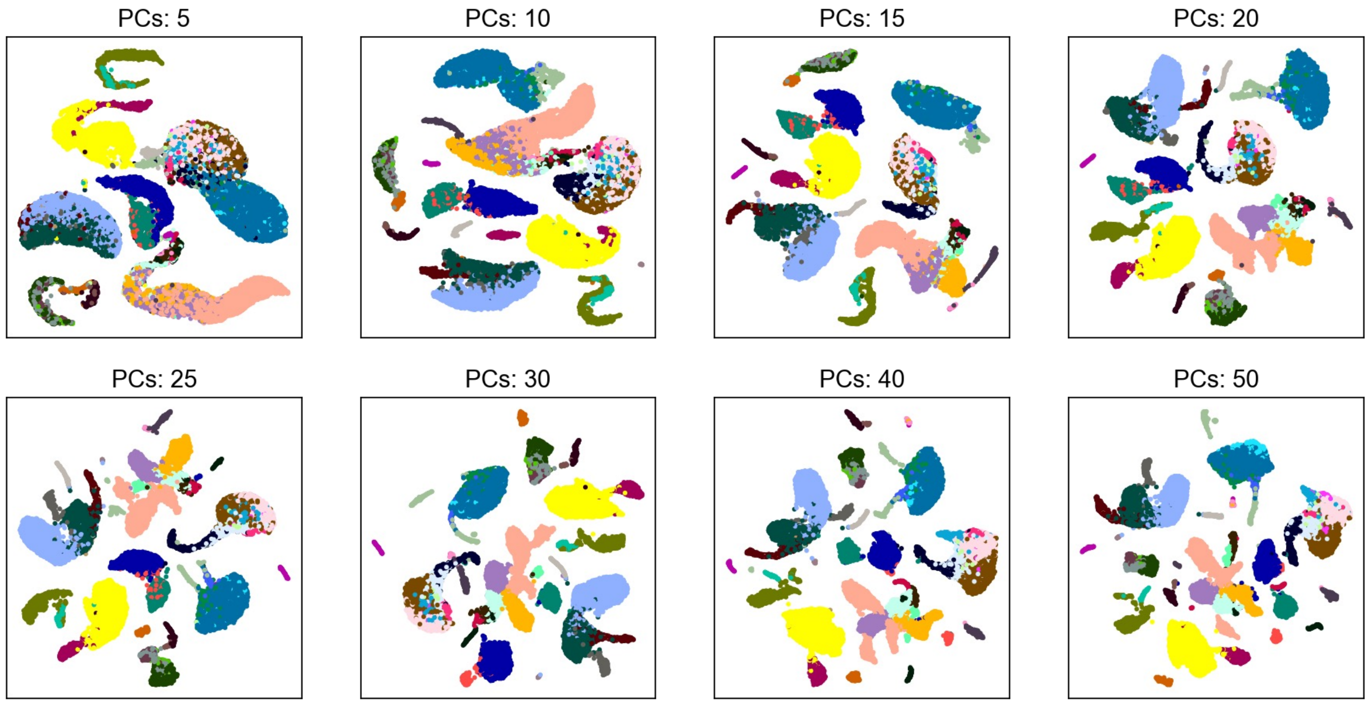
Option 2:

Elbow Plot



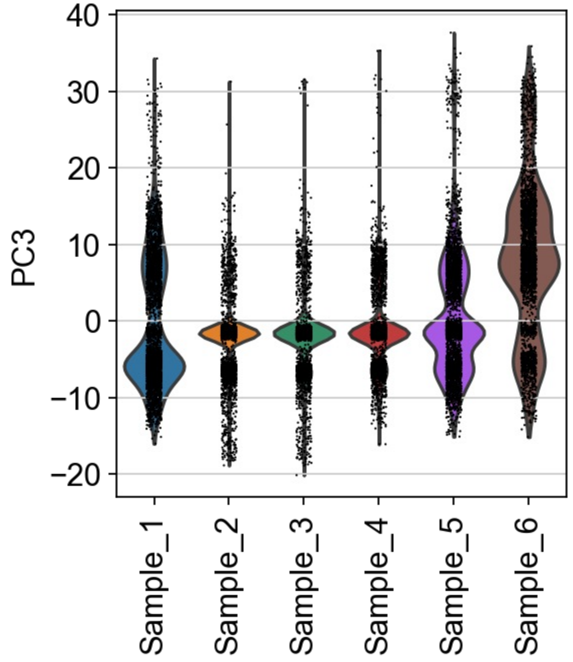
Option 3:

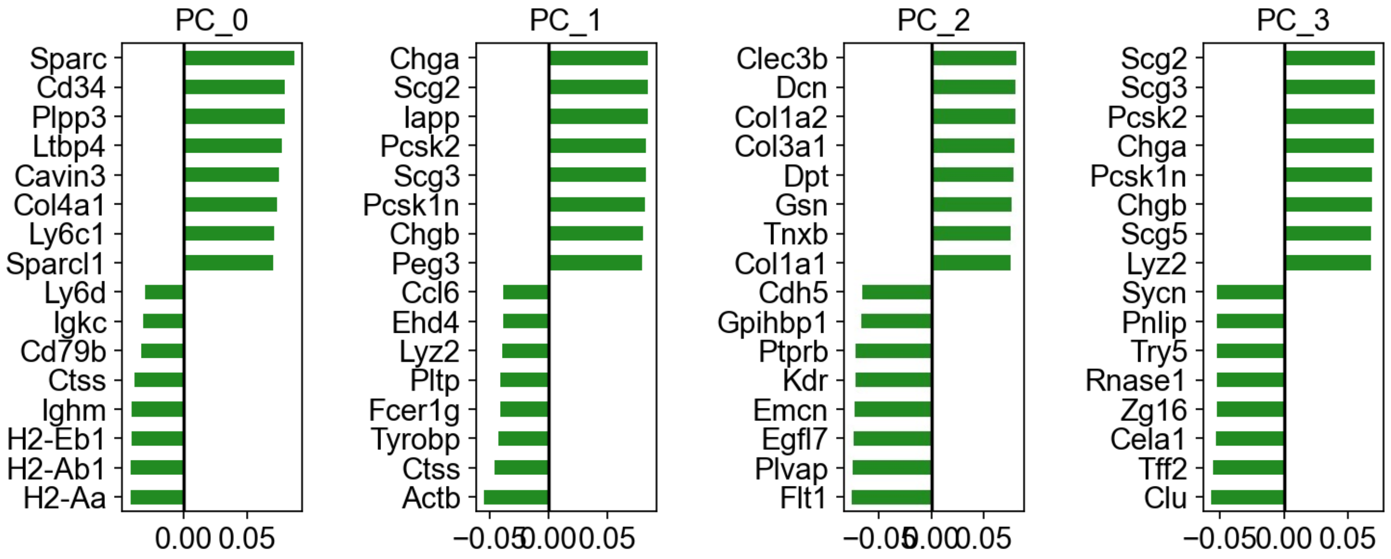
Pick a number that gives you results that you can interpret and defend.

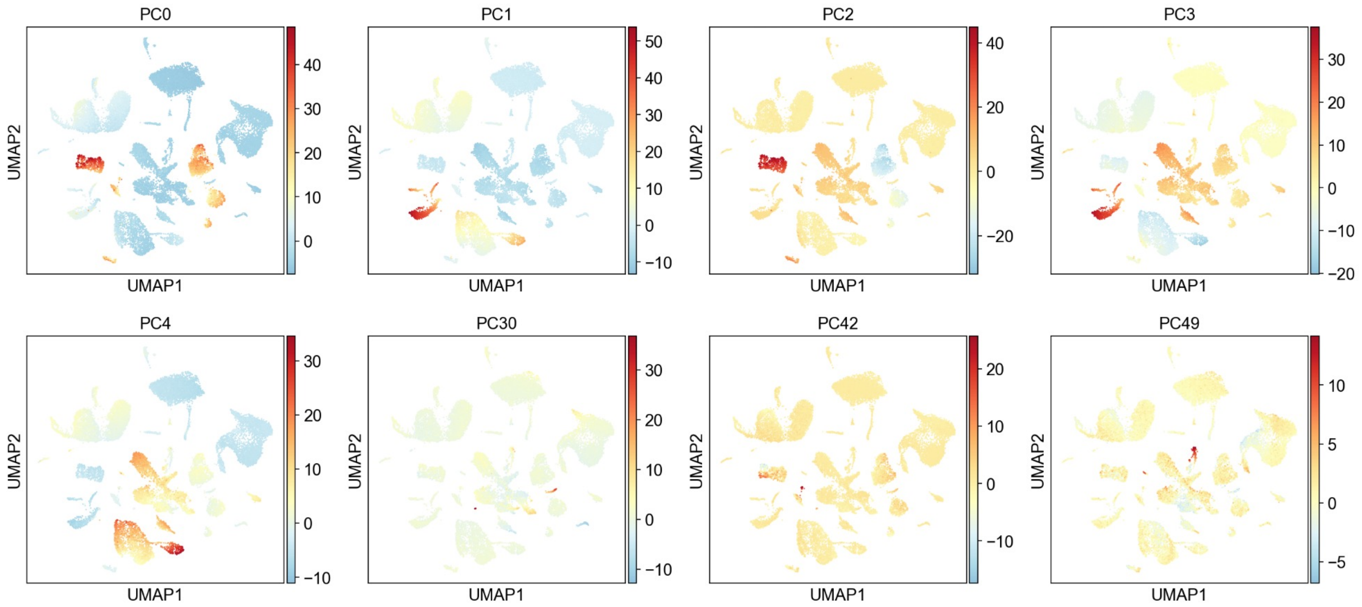


PC “Loadings”

- 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

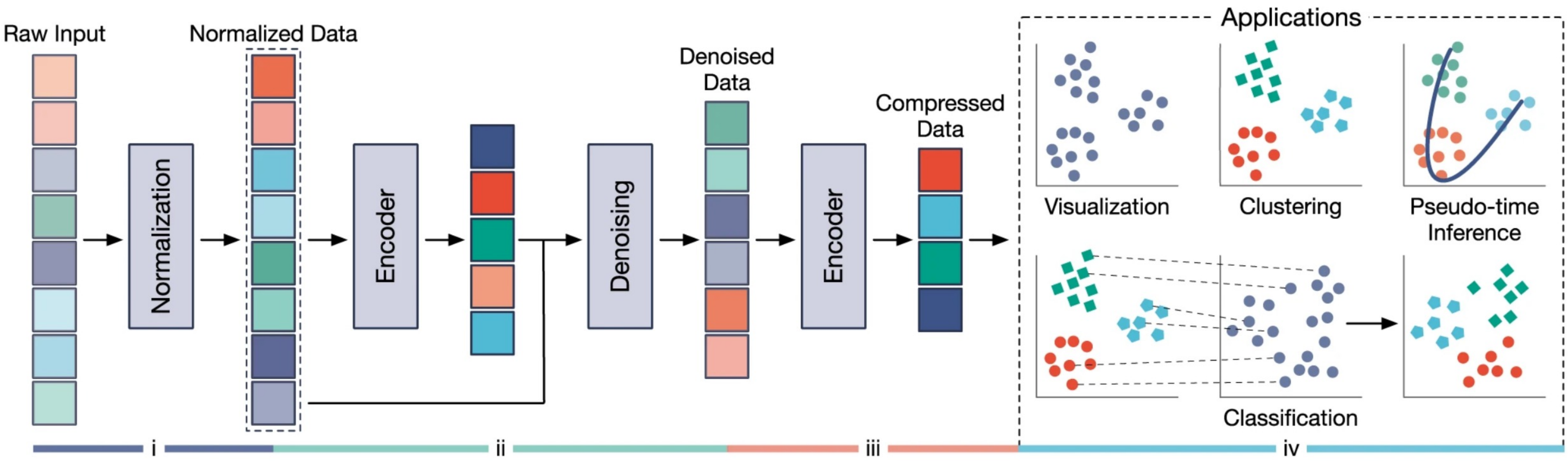
Deep Learning

- scVI
- scVAE
- SAUCIE
- VASC
- DCA
- scvis
- scSemiGAN

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

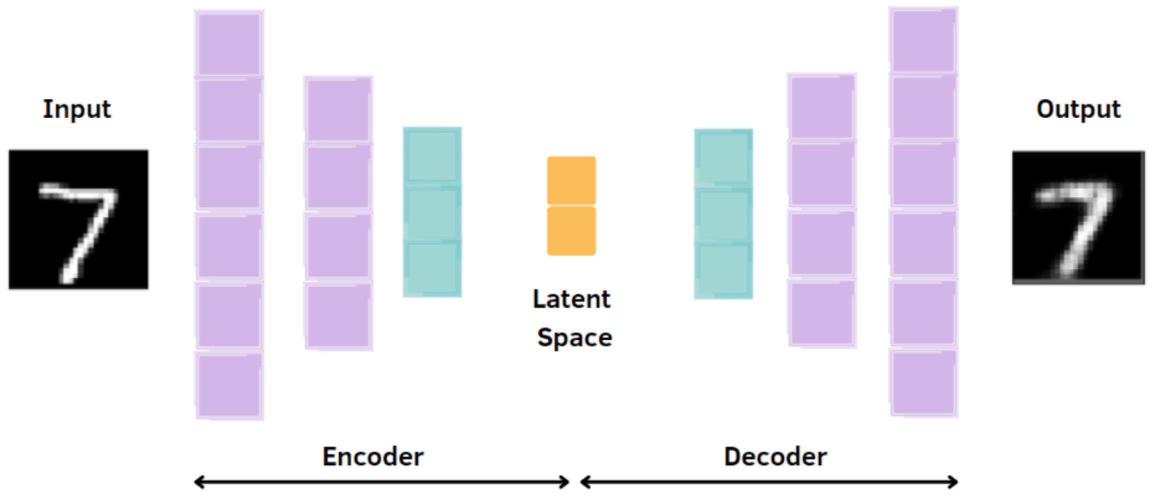
Adapted from Xiang e al. Front Genet. doi: 10.3389/fgene.2021.646936.

Deep Learning methods: Autoencoders



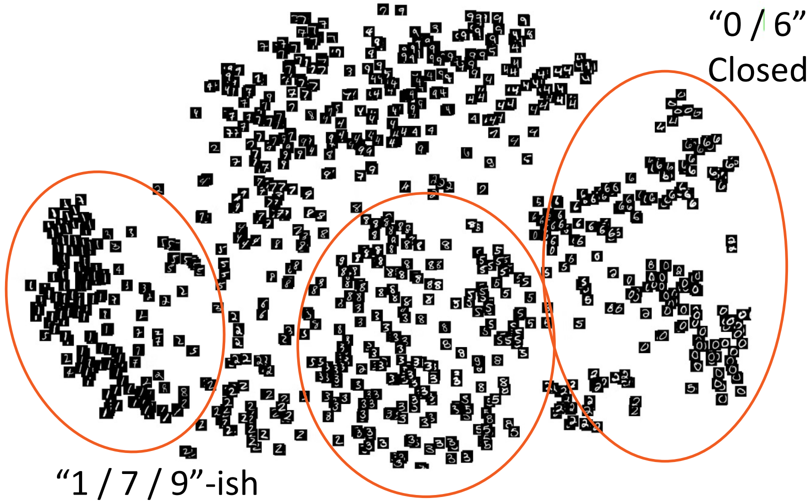
Batch correction
Denoising
Non-linear mapping of latent space (gene programs)

Tran et al. Nat Commun. 2021 Feb 15;12(1):1029. doi: 10.1038/s41467-021-21312-2.



Training set of handwritten numbers

2D Latent-space representation



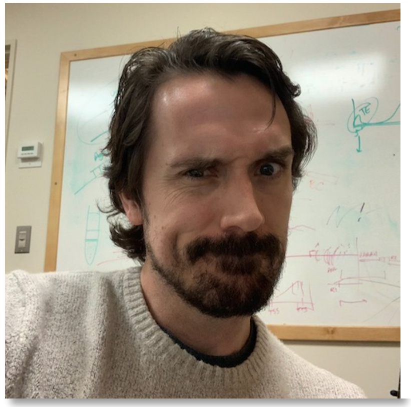
“0 / 6”-ish
Closed circles

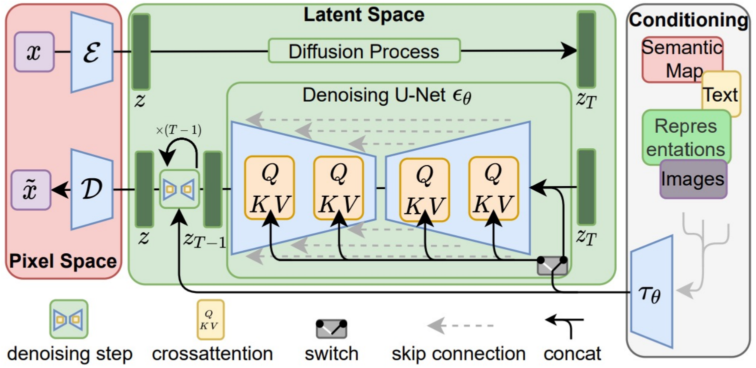
“1 / 7 / 9”-ish
Straight lines

“3 / 5 / 8”-ish
”bumpy”

Variational Autoencoders

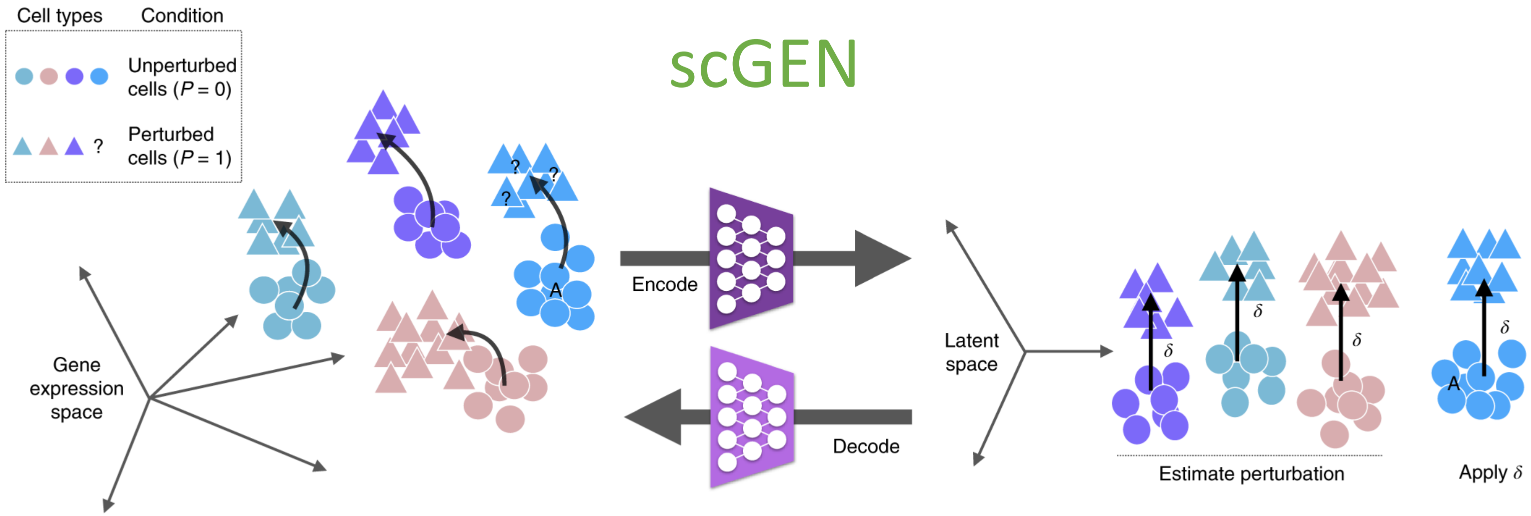
Stable Diffusion





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

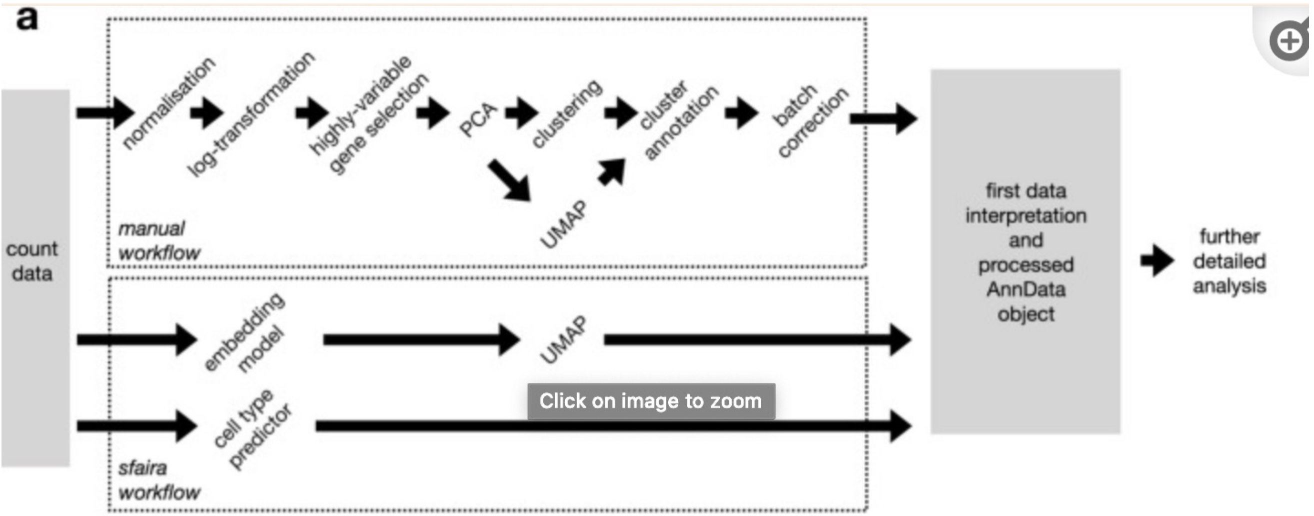
scGEN



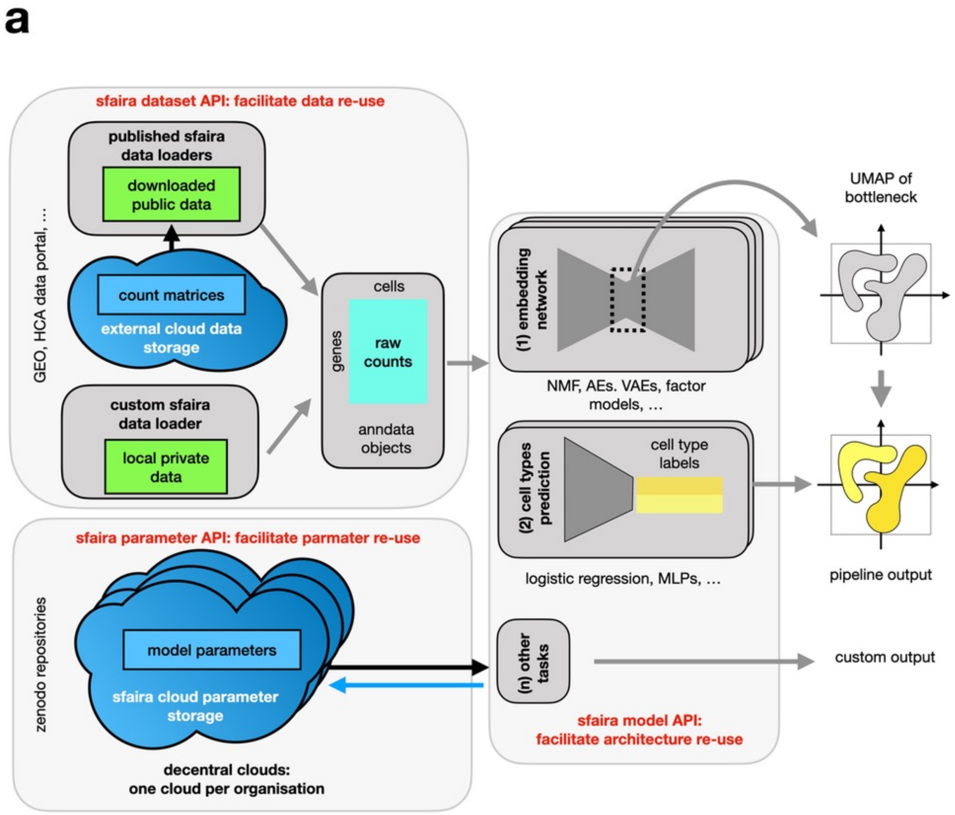
Towards sharable and reproducible embeddings

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
sfaira dataset API: facilitate data re-use
GEO, HCA data portal, ...
published sfaira data loaders
downloaded public data
count matrices
external cloud data storage
custom sfaira data loader
local private data
cells
genes
raw counts
anndata objects
sfaira parameter API: facilitate parmater re-use
zenodo repositories
model parameters
sfaira cloud parameter storage
decentral clouds: one cloud per organisation
(1) embedding network
NMF, AEs. VAEs, factor models, ...
(2) cell types prediction
cell type labels
logistic regression, MLPs, ...
(n) other tasks
sfaira model API: facilitate architecture re-use
UMAP of bottleneck
pipeline output
custom output

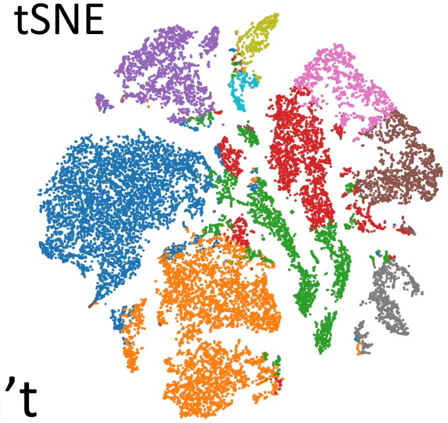
Fischer, D.S., Dony, L., König, M. *et al.* Sfaira accelerates data and model reuse in single cell genomics. *Genome Biol* **22**, 248 (2021). https://doi.org/10.1186/s13059-021-02452-6

Visualization with tSNE/UMAP

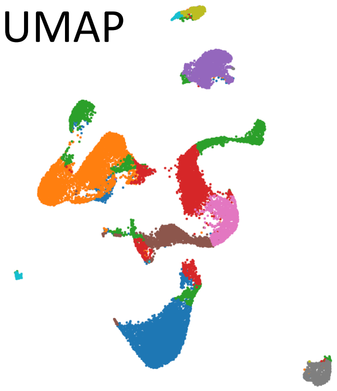
T-distributed stochastic neighbor embedding
Uniform Manifold Approximation

- 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

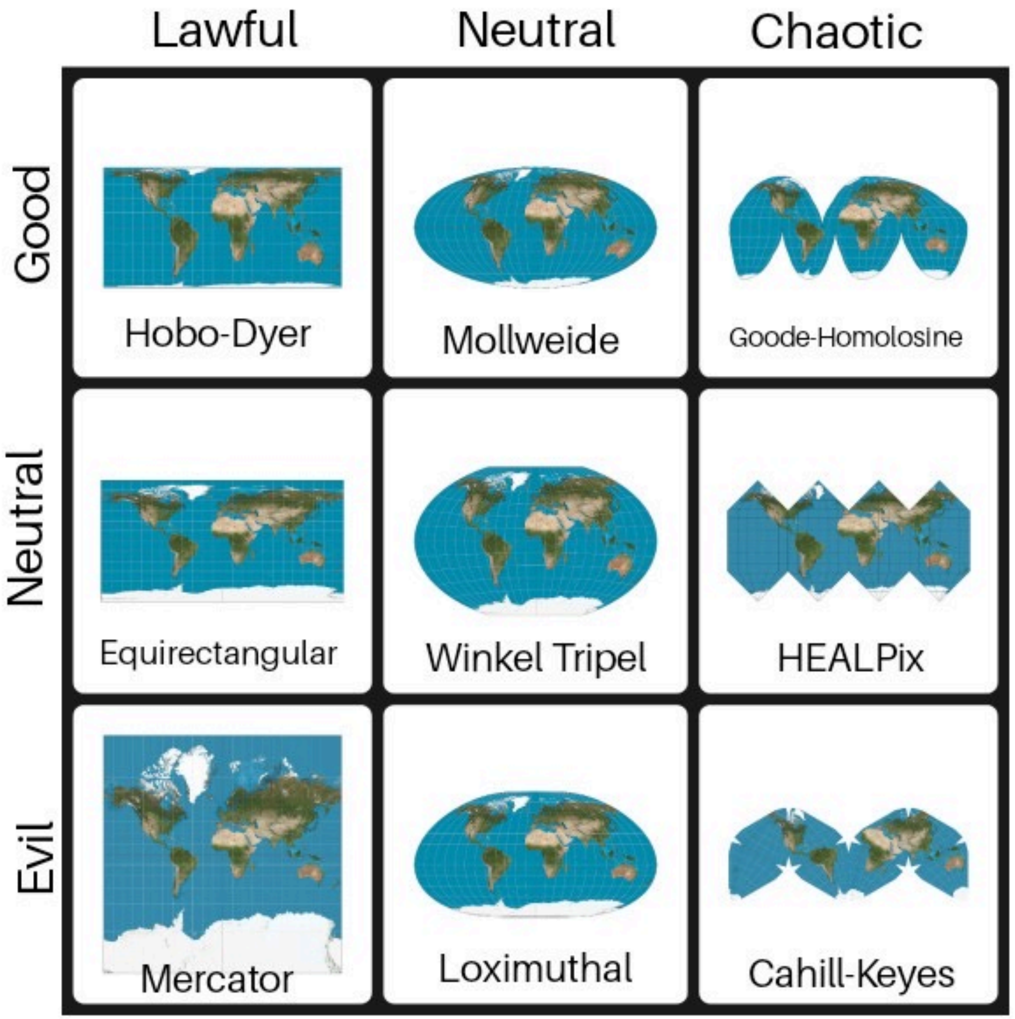
tSNE



UMAP



No one projection is “correct”



	Lawful	Neutral	Chaotic
Good	Hobo-Dyer	Mollweide	Goode-Homolosine
Neutral	Equirectangular	Winkel Tripel	HEALPix
Evil	Mercator	Loximuthal	Cahill-Keyes

Clustering

Generally run on latent space (eg. PC space)

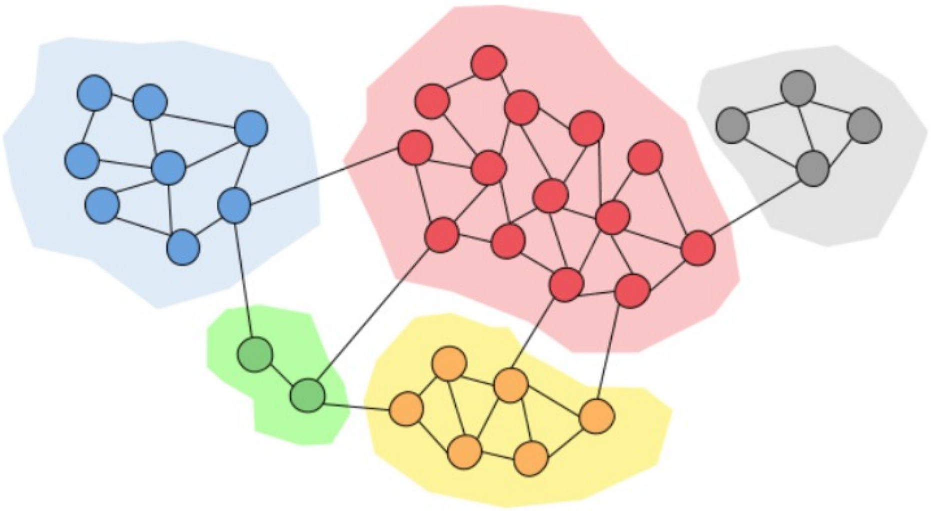
Built into Cell Ranger, most 3rd party tools:

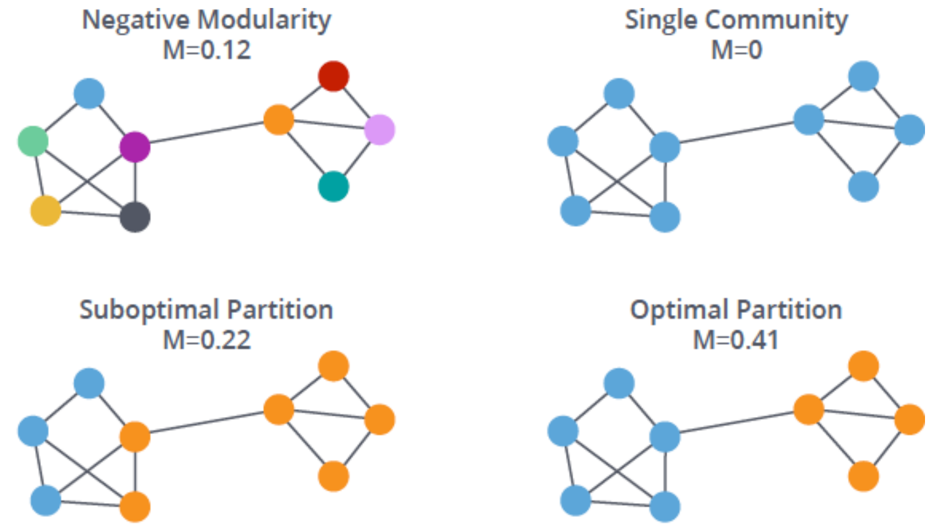
K-means:

- 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



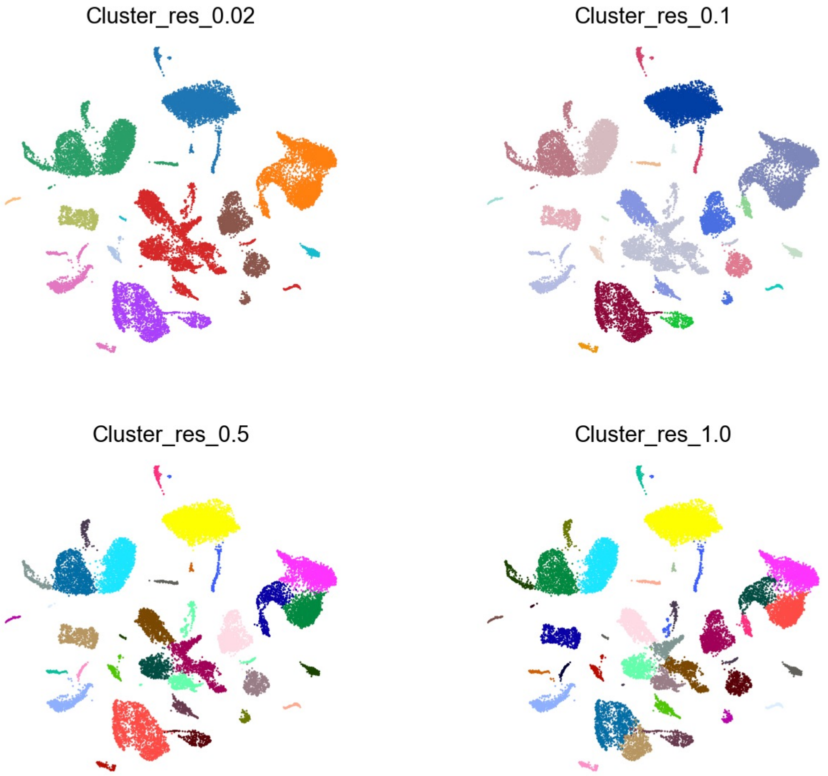


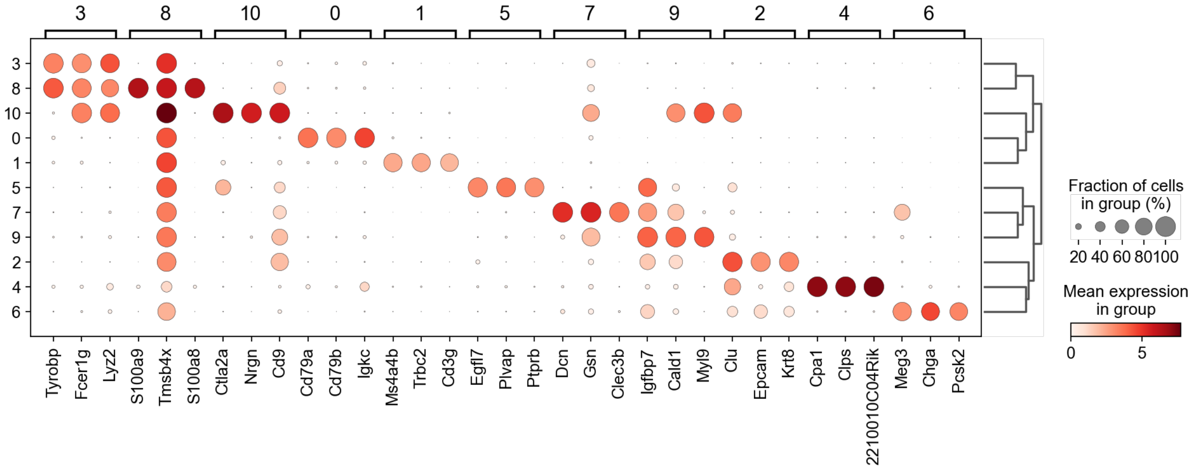
Modularity

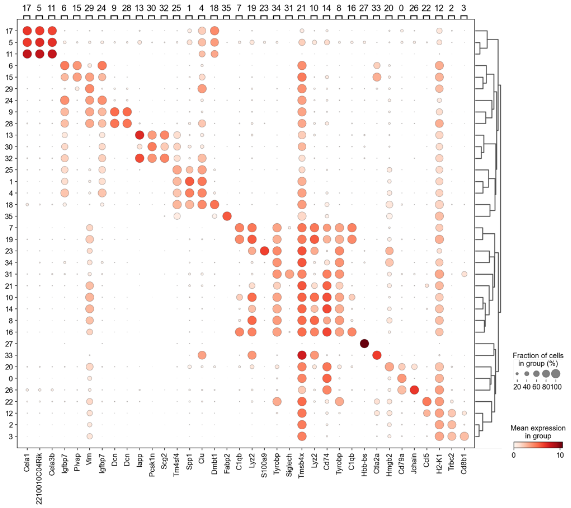
*(Completely unrelated to K in K-means!)

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!







Too Low: miss subtypes, phenotypic states

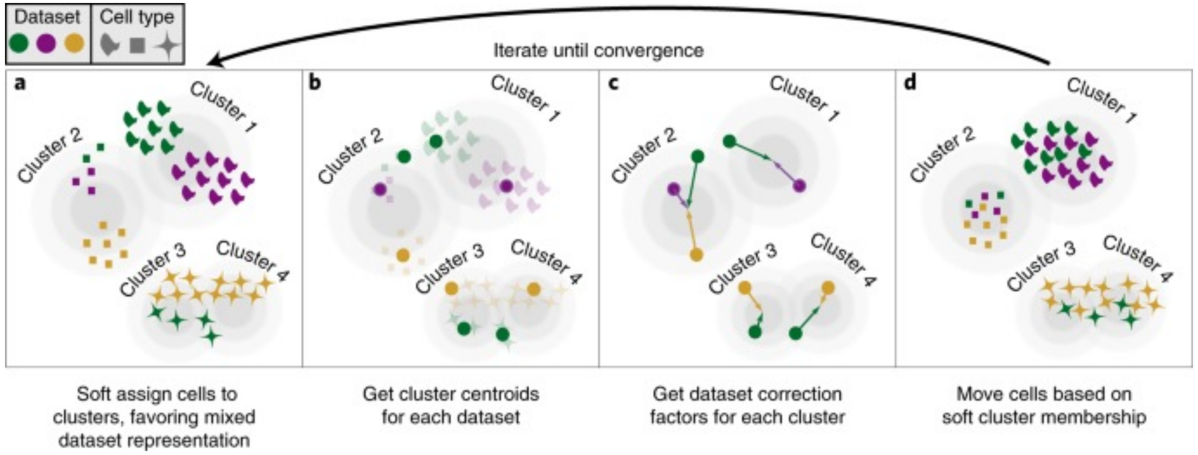
Too High: over-clustered, no meaningful distinguishing marker genes

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.

Harmony



Soft assign cells to clusters, favoring mixed dataset representation

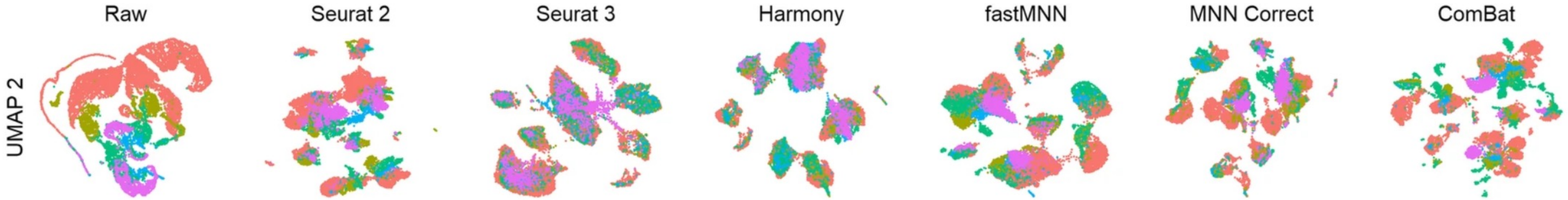
Get cluster centroids for each dataset

Get dataset correction factors for each cluster

Move cells based on soft cluster membership

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 not directly modify data matrix
ie. often can’t be used for differential expression

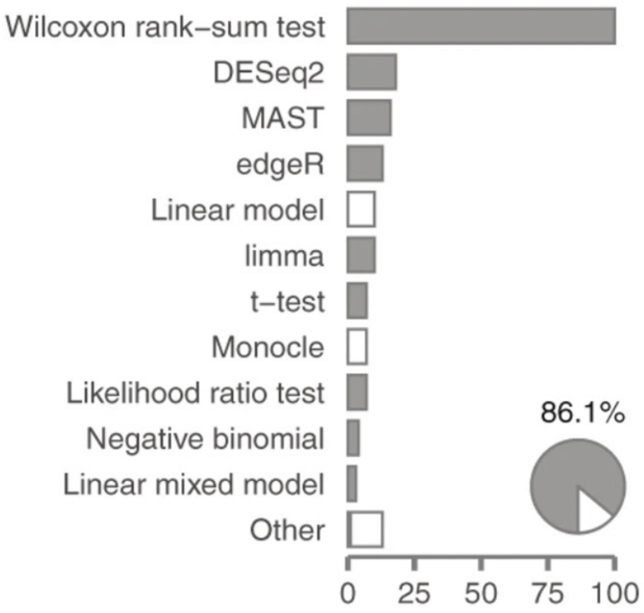


“The best batch correction happens at the bench” –J Preall

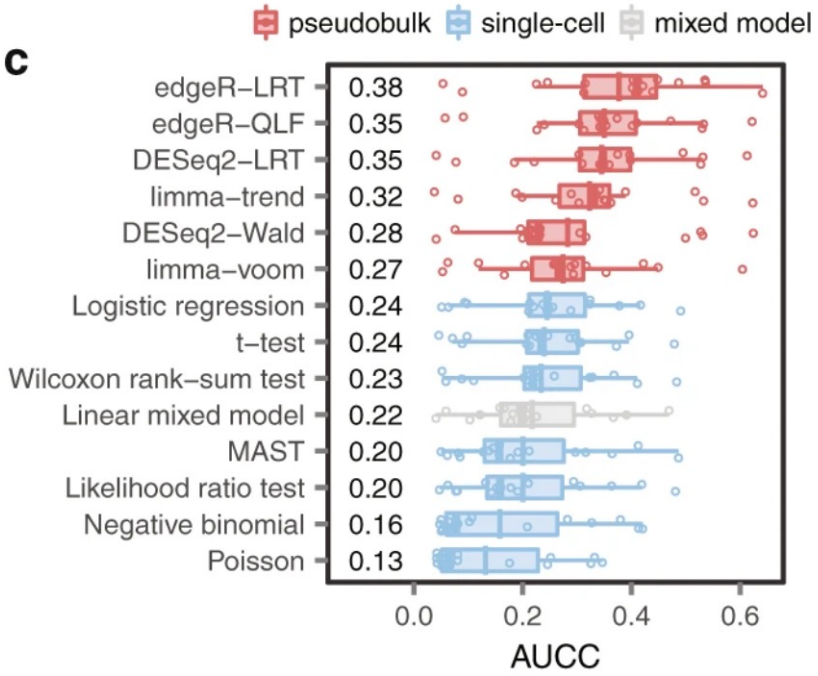
Harmony: Korsunsky, I., Millard, N., Fan, J. et al. Nat Methods 16, 1289–1296 (2019).
Benchmarking: Tran, H.T.N., Ang, K.S., Chevrier, M. et al. Genome Biol 21, 12 (2020). https://doi.org/10.1186/s13059-019-1850-9

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)

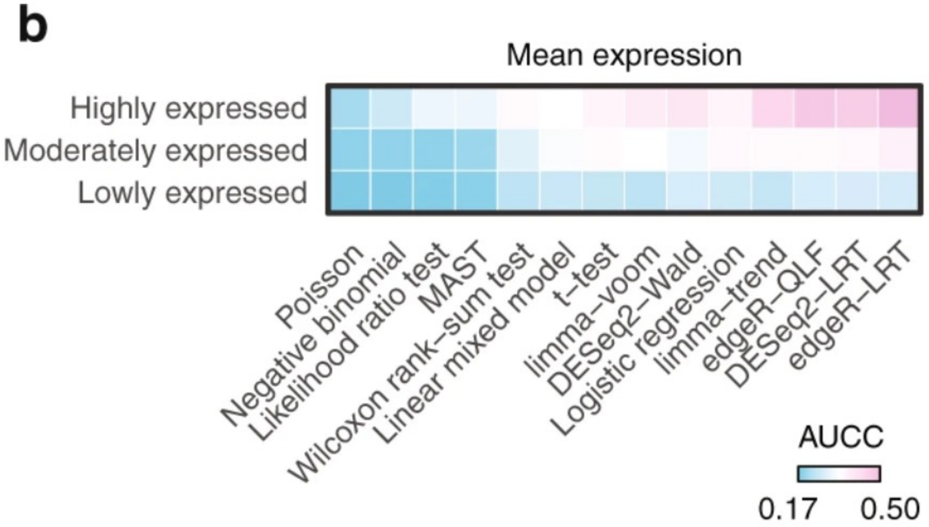
Squair, J.W., Gautier, M., Kathe, C. et al. Confronting false discoveries in single-cell differential expression. Nat Commun 12, 5692 (2021). https://doi.org/10.1038/s41467-021-25960-2

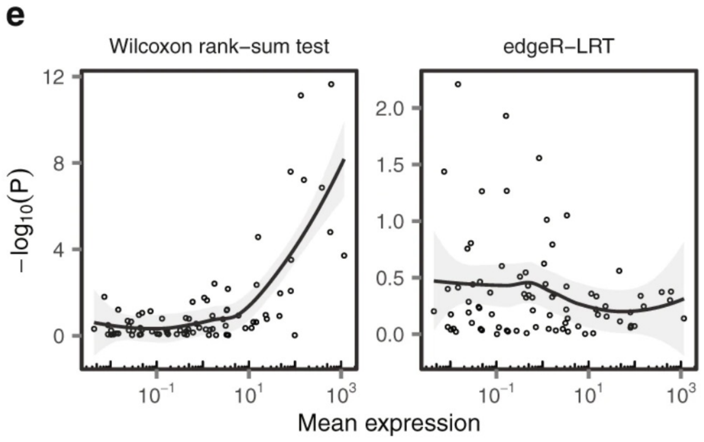
Differential Gene Expression

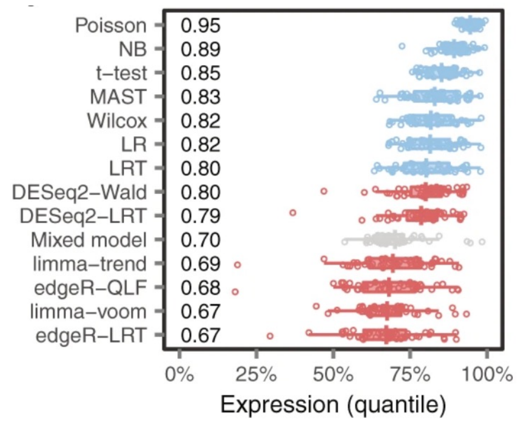
Single-cell methods biased towards highly expressed genes

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)



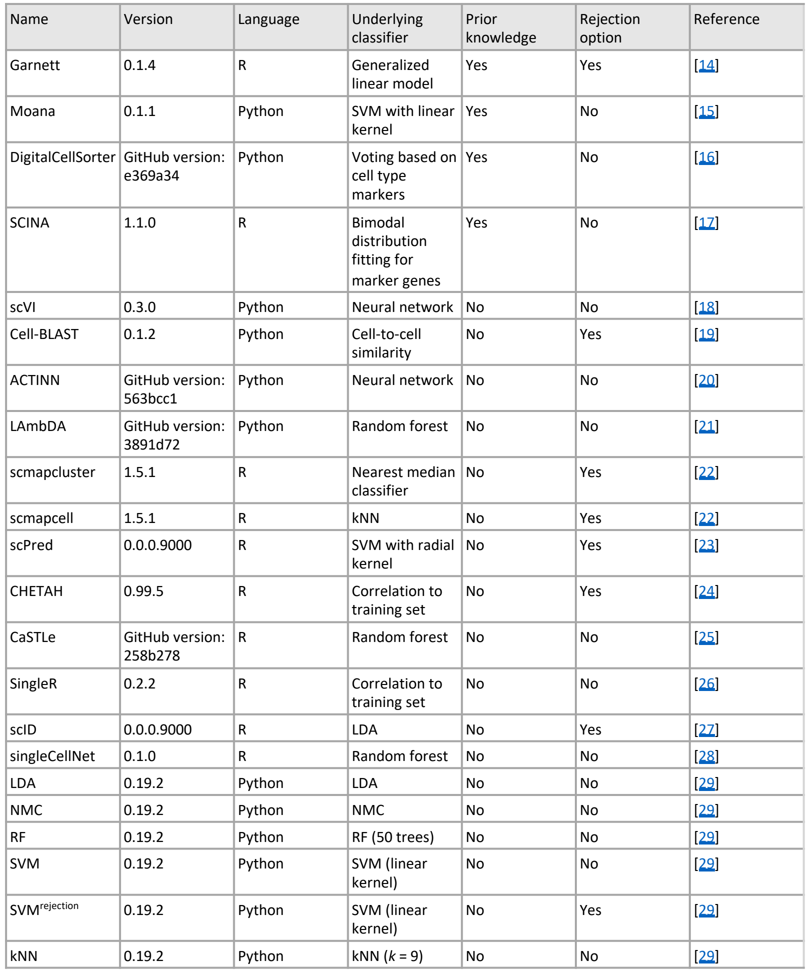




Squair, J.W., Gautier, M., Kathe, C. et al. Confronting false discoveries in single-cell differential expression. Nat Commun 12, 5692 (2021). https://doi.org/10.1038/s41467-021-25960-2

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

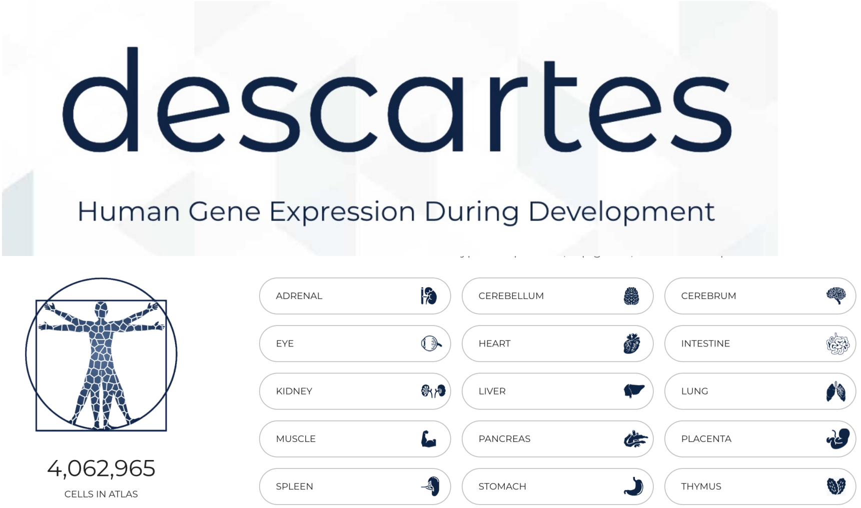
scRNA-tools

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

Atlas Efforts building Cell Type Classifiers



descartes

Human Gene Expression During Development

ADRENAL CEREBELLUM CEREBRUM
EYE HEART INTESTINE
KIDNEY LIVER LUNG
MUSCLE PANCREAS PLACENTA
SPLEEN STOMACH THYMUS

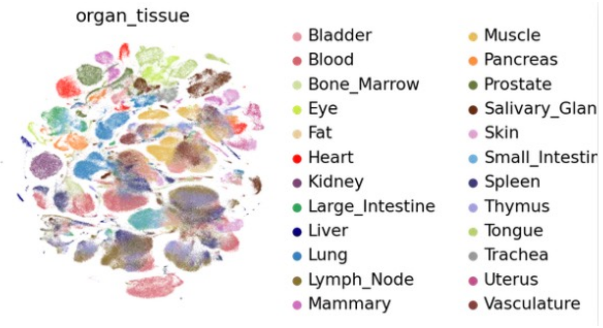
4,062,965
CELLS IN ATLAS

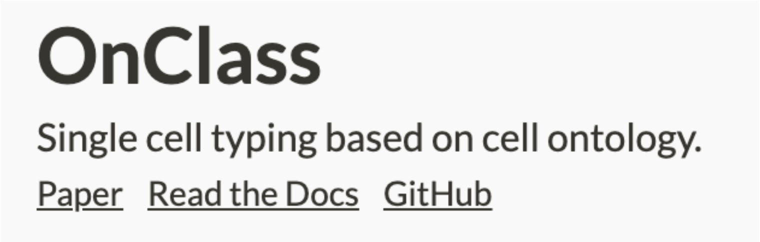
Garnett

Automated cell type classification

Tabula Sapiens

Human transcriptome reference at single cell resolution





OnClass

Single cell typing based on cell ontology.

Paper Read the Docs GitHub

Going Beyond Gene Expression

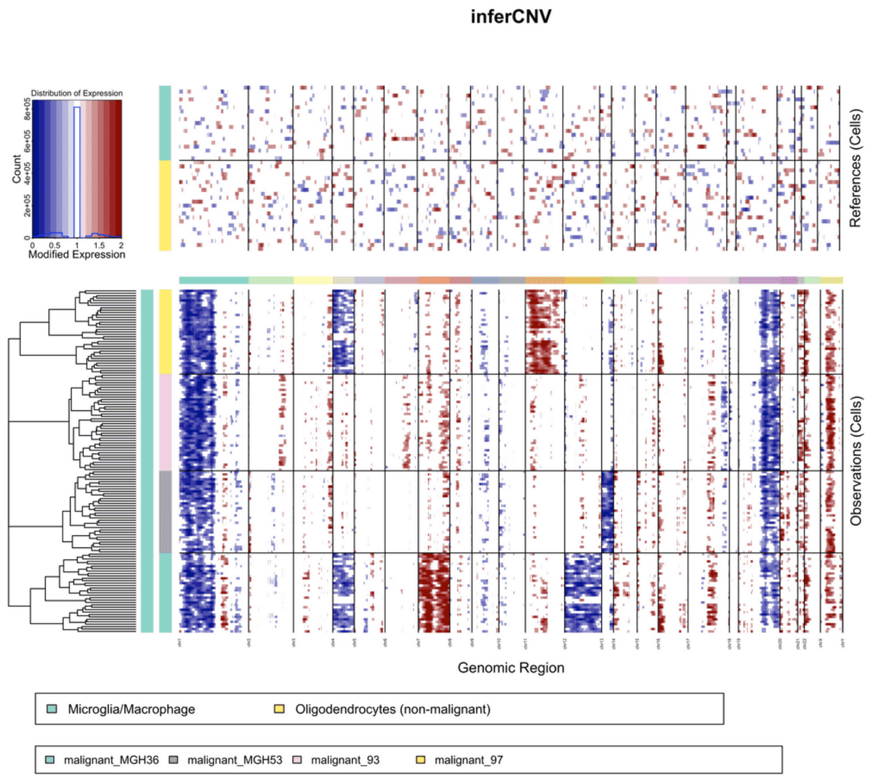
- Copy Number Inference:
  - InferCNV: Tirosh et al (2016) Science doi: 10.1126/science.aad0501
  - KopyKAT: Gao R et al. (2021) Nat Biotechnol. doi: 10.1038/s41587-020-00795-2
- Alternative TSS Utilization:
  - scRCAT-seq: Hu Y et al (2020) Nat Commmun. 10.1038/s41467-020-18976-7
- Alternative Polyadenylation: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

CNV inference

Eg. inferCNV, CaSpER, CopyKAT, SCYN

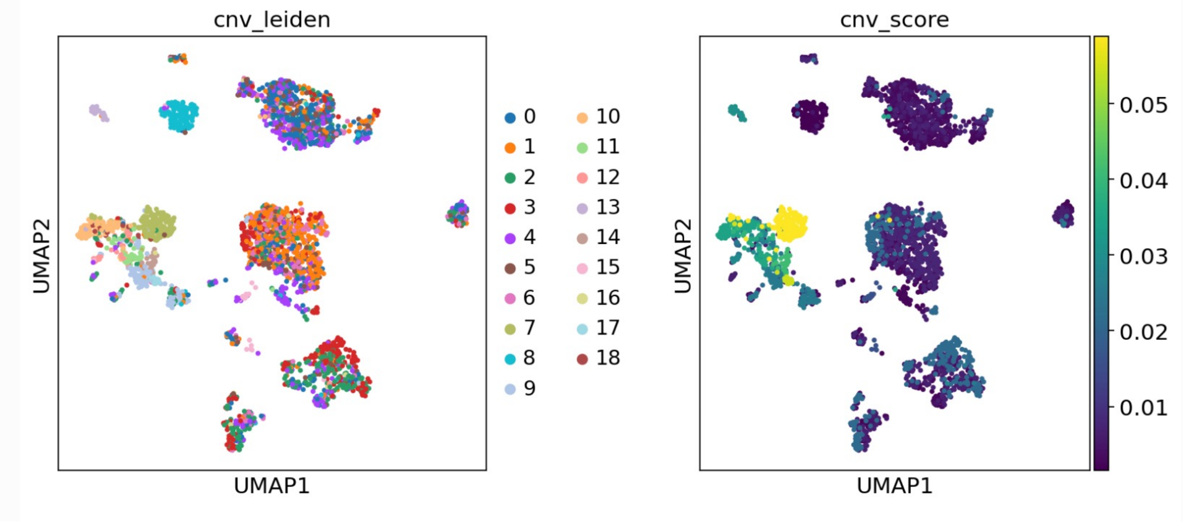
- Counts binned expression data across chromosomes
- Builds a background model based on provided “Normal” reference
- Identifies regions with higher than expected expression across entire window

- 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



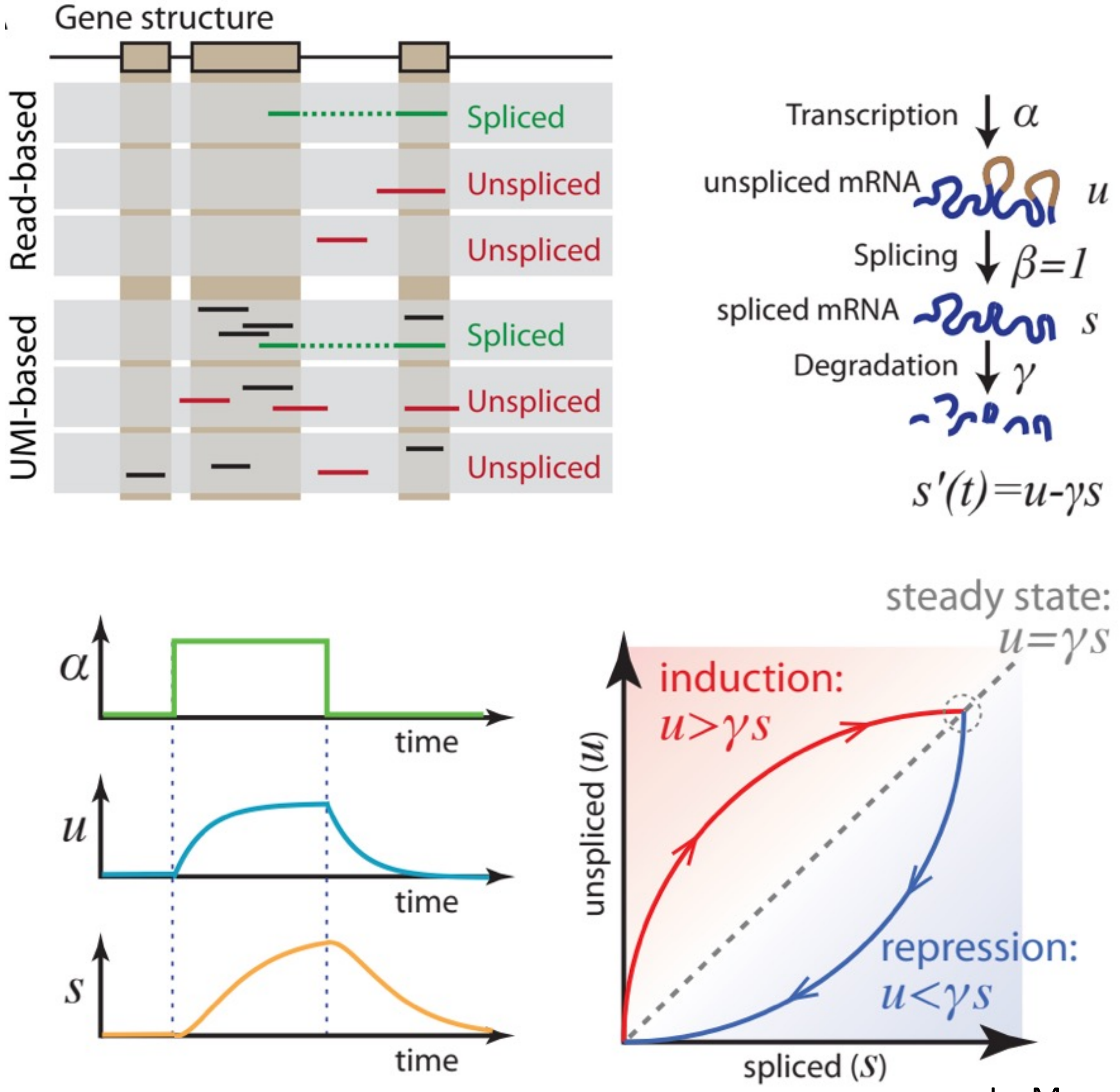
Transcriptional Dynamics: RNA Velocity

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

Quantifies spliced / unspliced

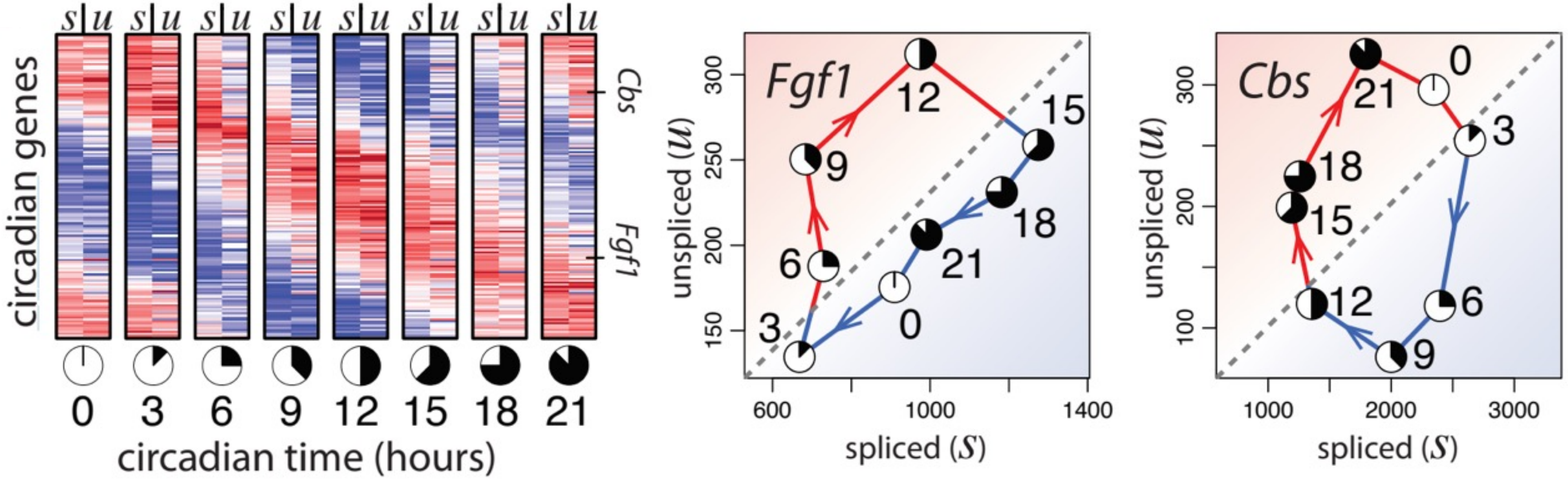
Models dynamics

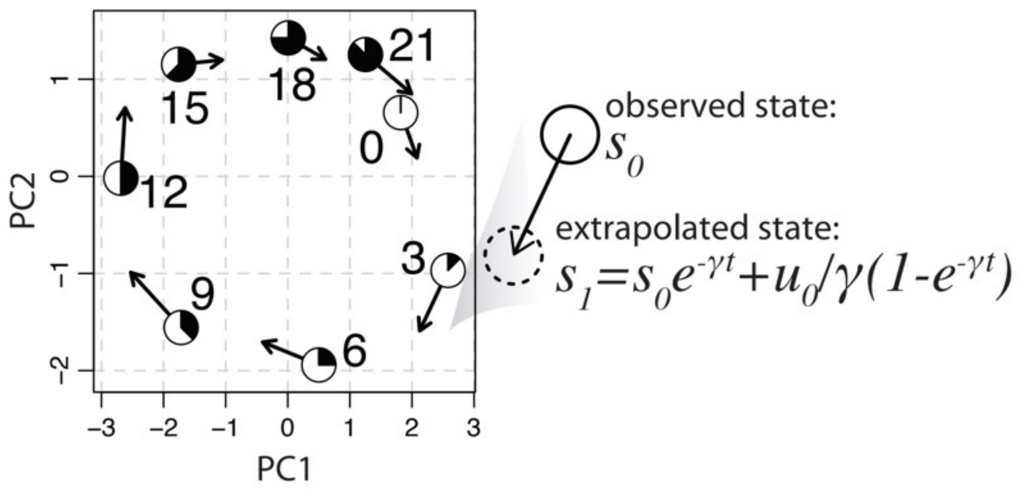
CAVEATS: Gene annotations
Cryptic exons
unannotated intronic genes
repetitive elements



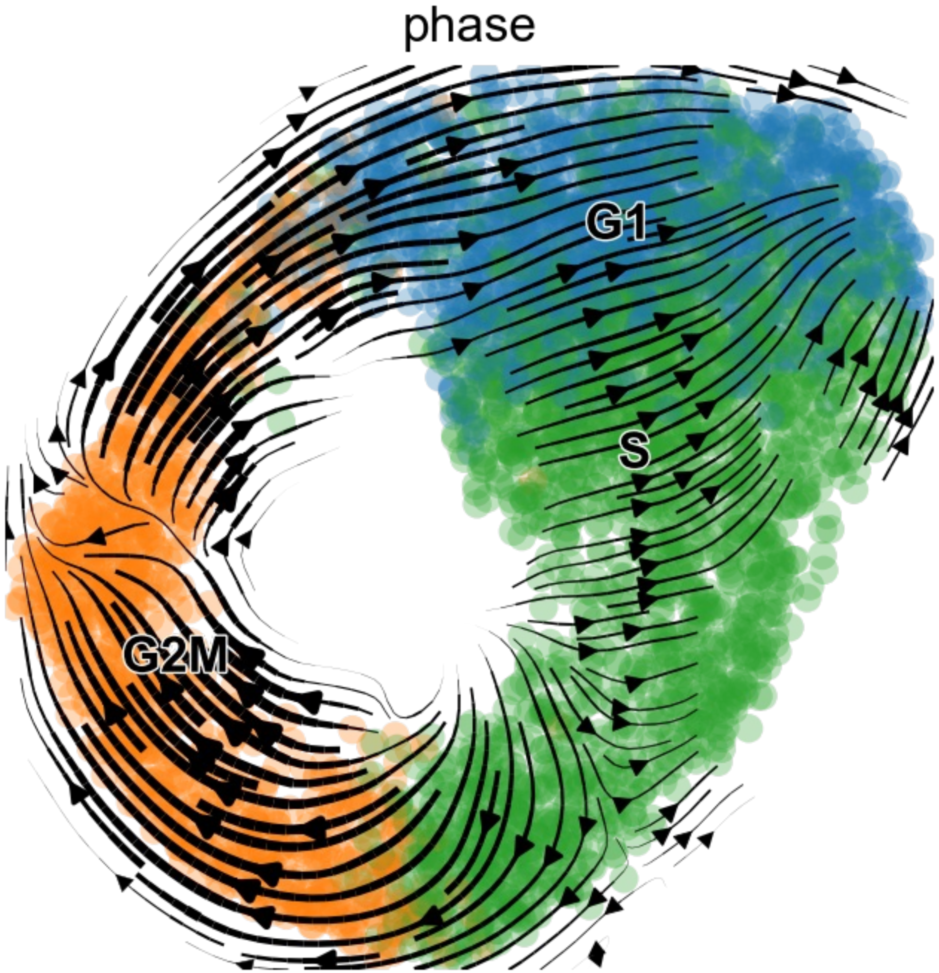
Le Manno et al. (2018) *Nature*

Bulk RNAseq from mouse circadian rhythm data



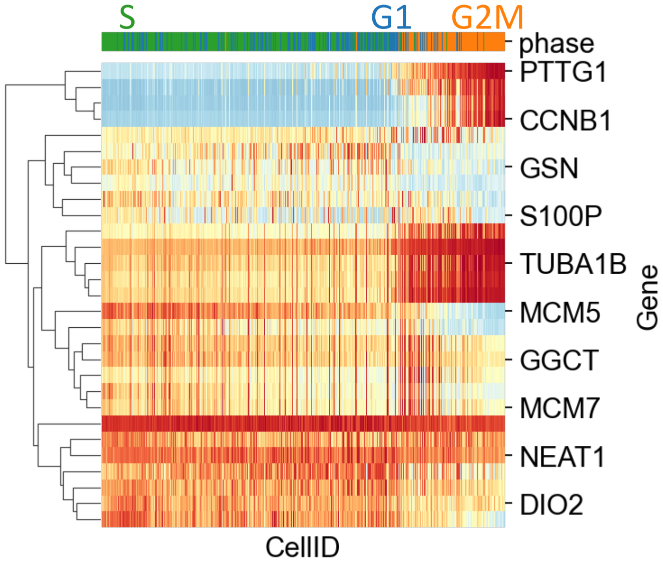


Cultured Cells – RNA Velocity



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

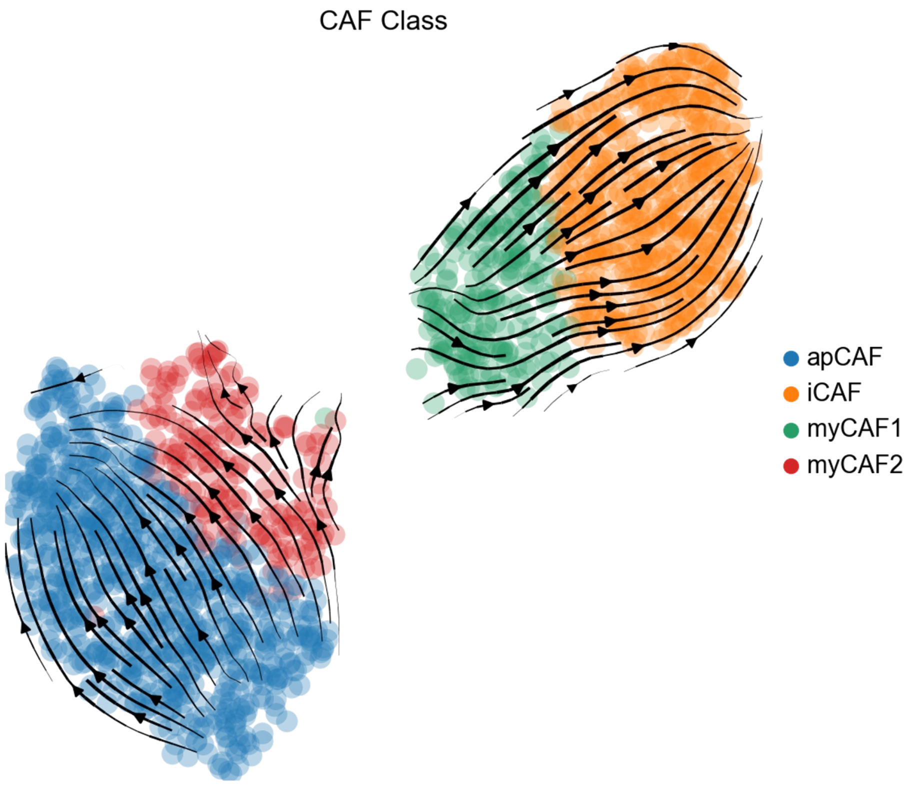
RNA Velocity is relevant over very short time scales, when transition states are abundant in the population



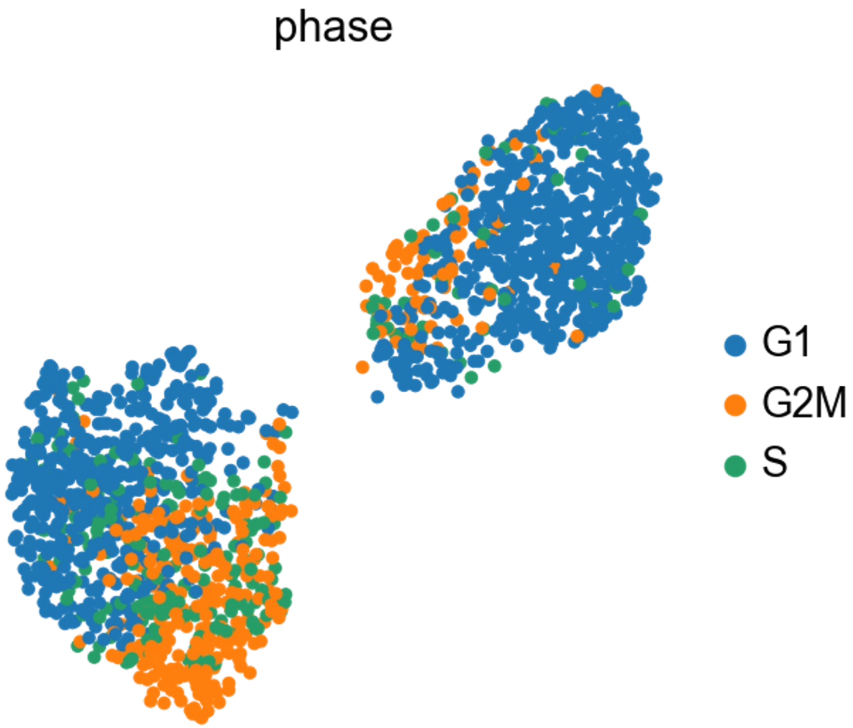
Latent time →

Caveat:

Don’t Confuse Developmental Trajectories with Cell Cycle Kinetics



CAF Class



phase

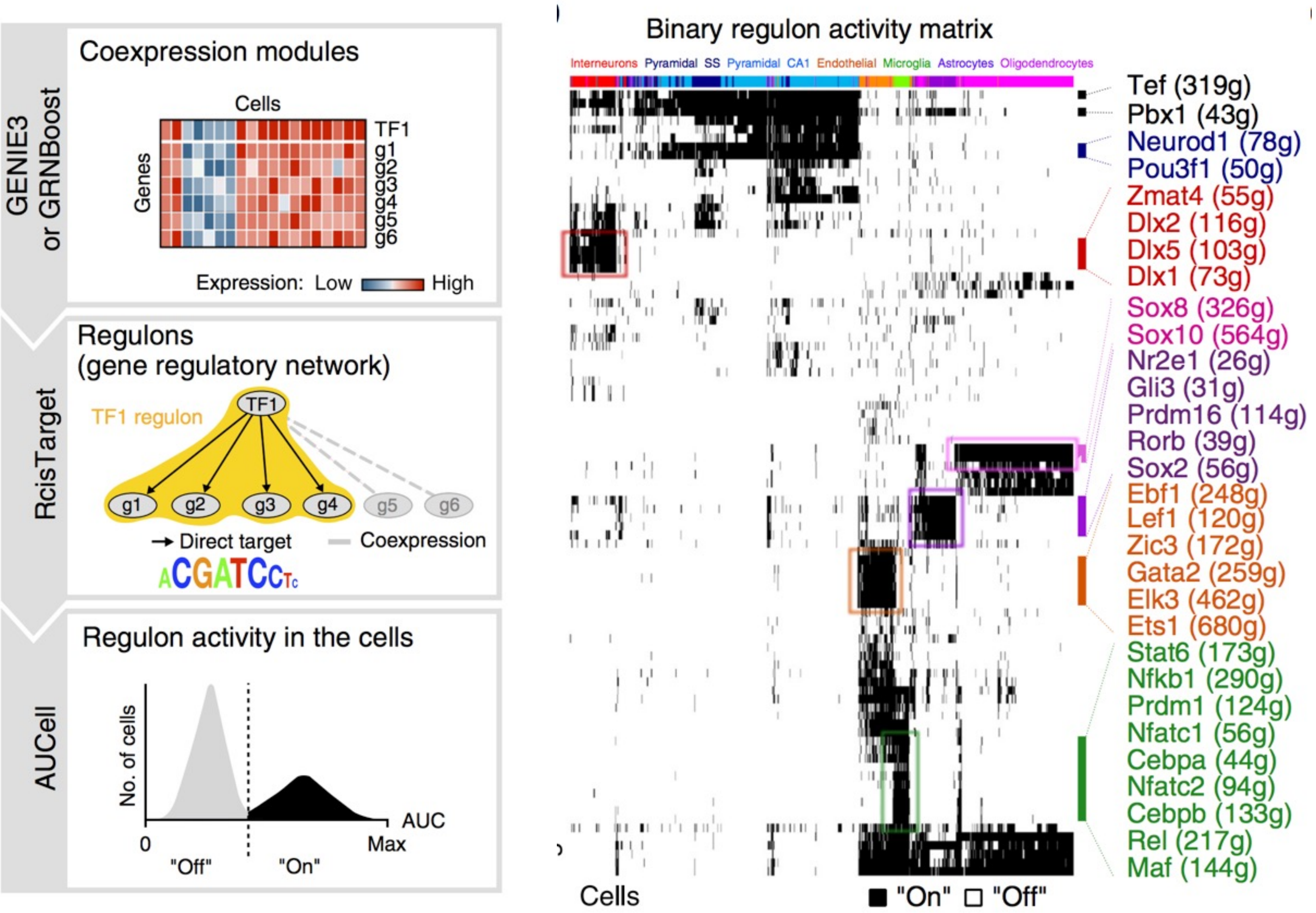
SCENIC single-cell regulatory network inference and clustering

Infers Transcription Factor Activity

Gene Co-expression network

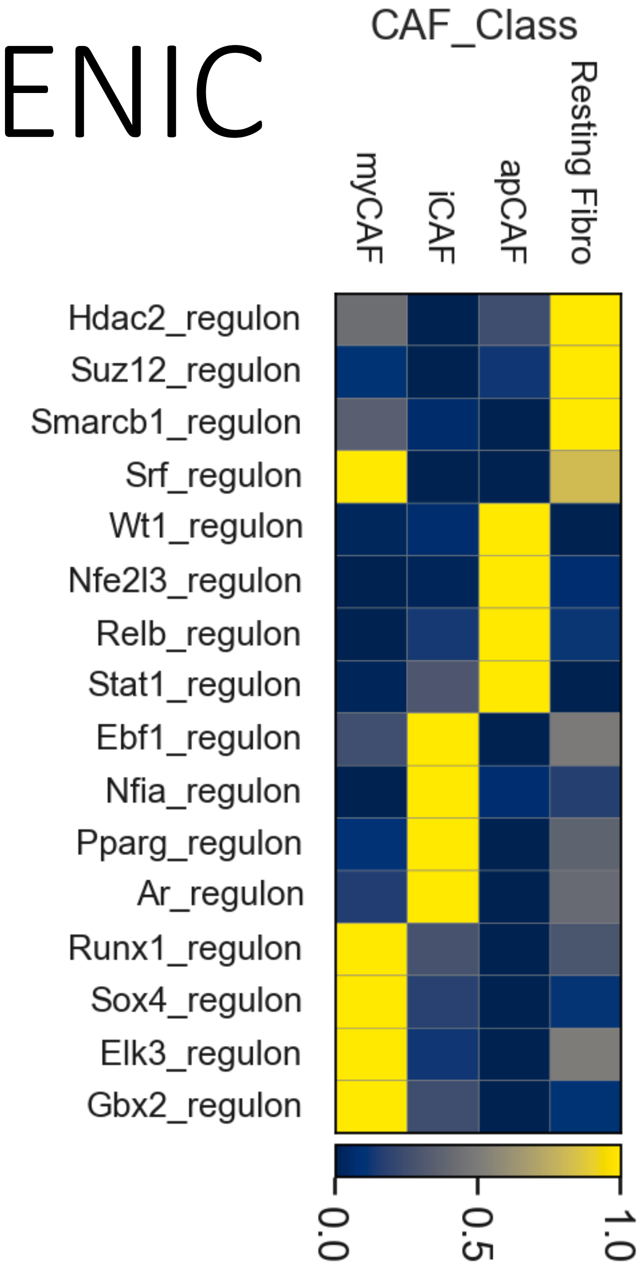
Motif search

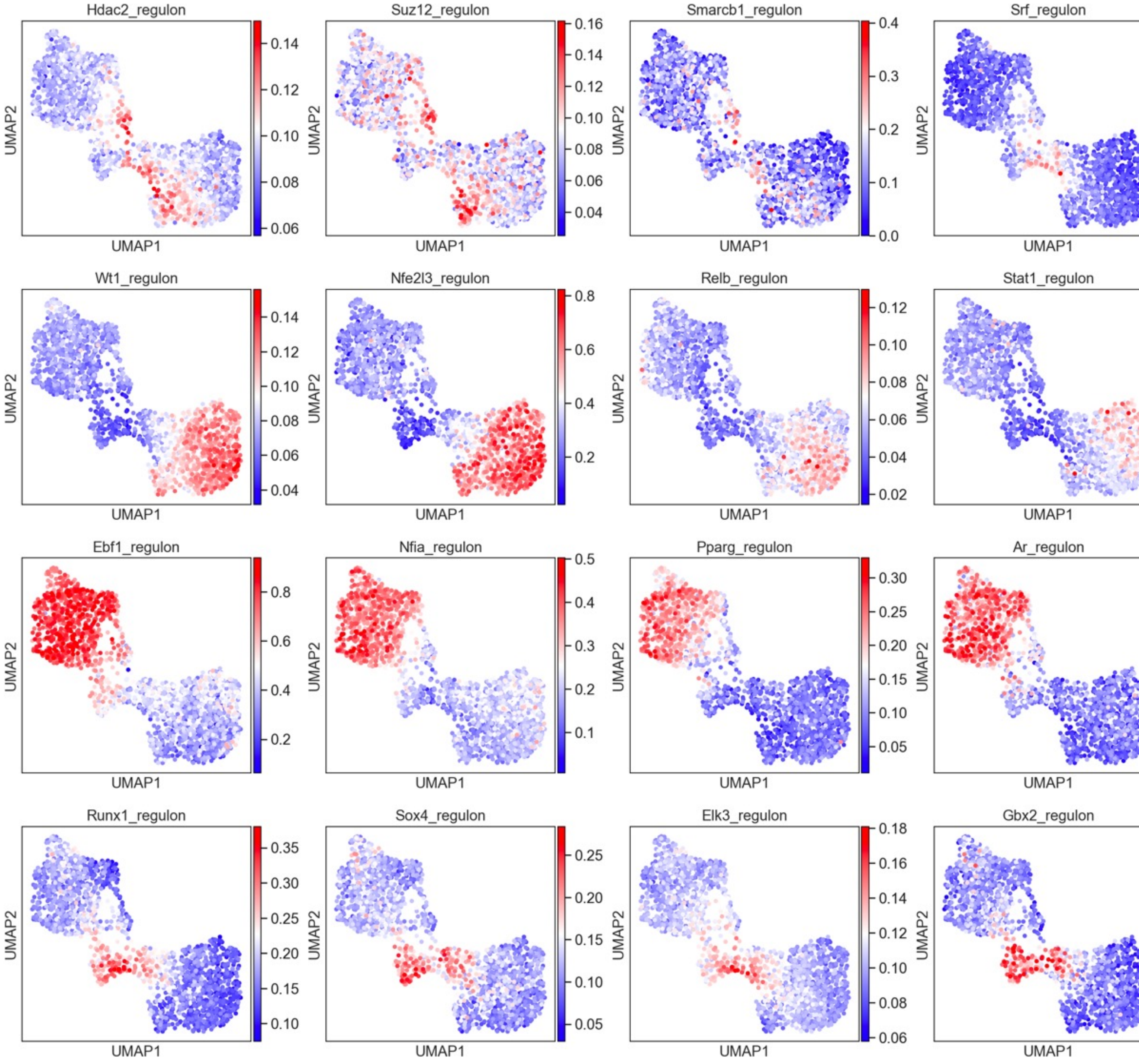
Regulon activity



Aibar et al. (2017) *Nature Methods.* SCENIC: single-cell regulatory network inference and clustering

SCENIC


CAF_Class
myCAF
iCAF
apCAF
Resting Fibro
Hdac2_regulon
Suz12_regulon
Smarcb1_regulon
Srf_regulon
Wt1_regulon
Nfe2l3_regulon
Relb_regulon
Stat1_regulon
Ebf1_regulon
Nfia_regulon
Pparg_regulon
Ar_regulon
Runx1_regulon
Sox4_regulon
Elk3_regulon
Gbx2_regulon
0.0
0.5
1.0


Hdac2_regulon
0.14
0.12
0.10
0.08
0.06
Suz12_regulon
0.16
0.14
0.12
0.10
0.08
0.06
0.04
Smarcb1_regulon
0.4
0.3
0.2
0.1
0.0
Srf_regulon
Wt1_regulon
0.14
0.12
0.10
0.08
0.06
0.04
Nfe2l3_regulon
0.8
0.6
0.4
0.2
Relb_regulon
0.12
0.10
0.08
0.06
0.04
0.02
Stat1_regulon
Ebf1_regulon
0.8
0.6
0.4
0.2
Nfia_regulon
0.5
0.4
0.3
0.2
0.1
Pparg_regulon
0.30
0.25
0.20
0.15
0.10
0.05
Ar_regulon
Runx1_regulon
0.35
0.30
0.25
0.20
0.15
0.10
Sox4_regulon
0.25
0.20
0.15
0.10
0.05
Elk3_regulon
0.18
0.16
0.14
0.12
0.10
0.08
0.06
Gbx2_regulon
UMAP1
UMAP2