Introduction to Single Cell RNA-Seq Data Analysis

Jenny Wu

Director of Bioinformatics

Genomics Research and Technology Hub

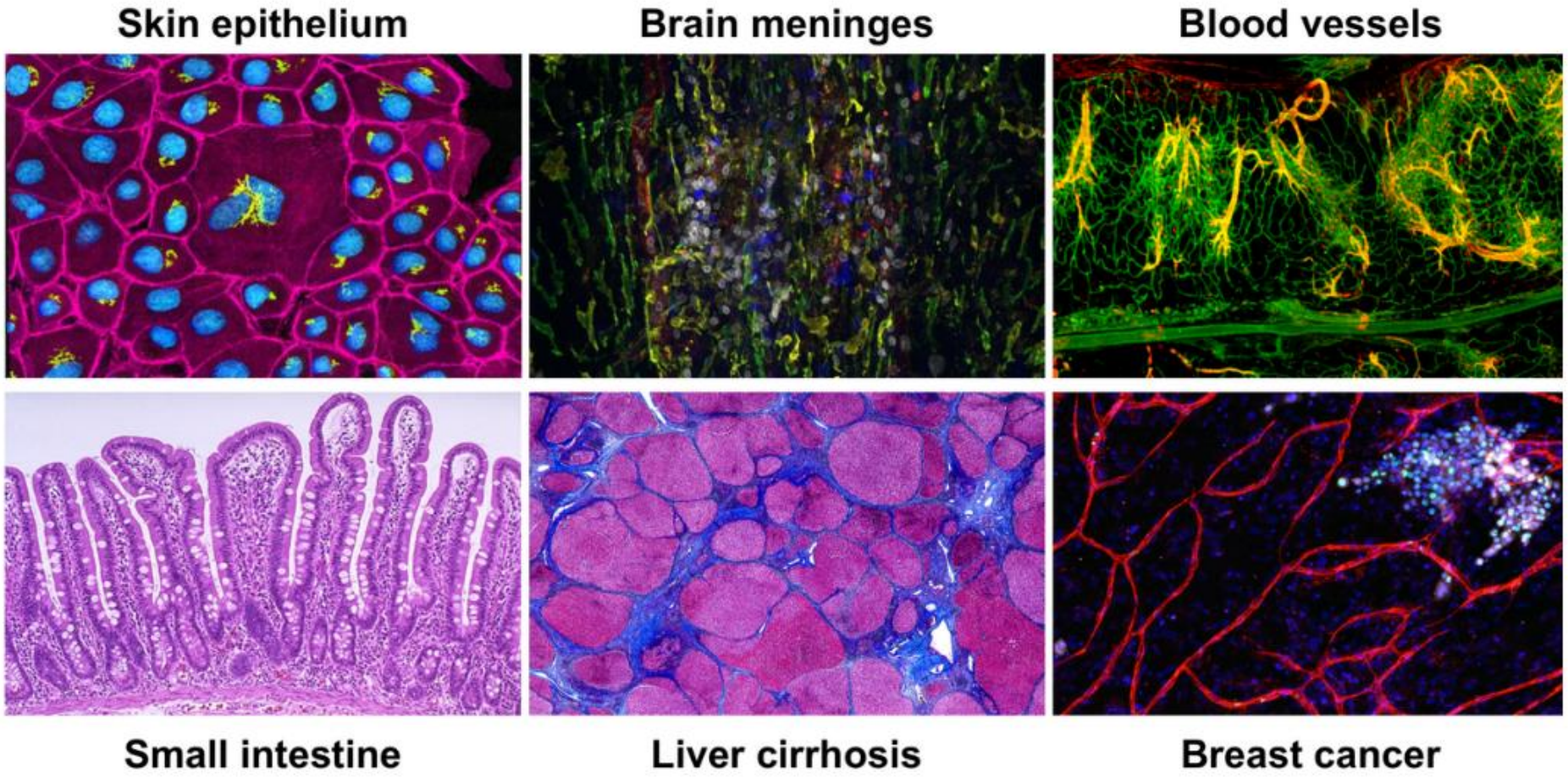
Chao Family Comprehensive Cancer Center

UC Irvine

Outline

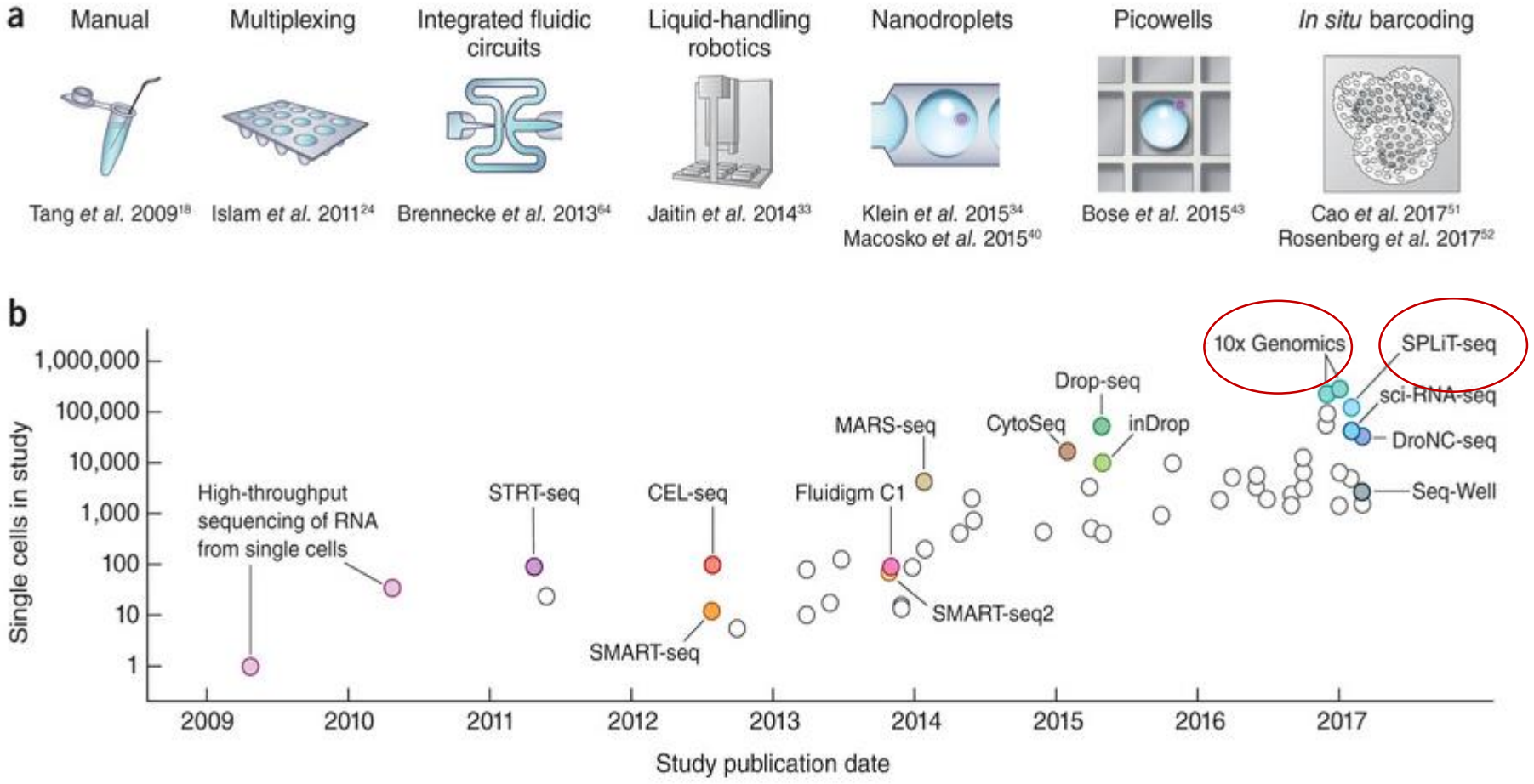
- Why single cell
- Single cell RNA-Seq workflow
- Experimental Design
- Data analysis pipeline and applications
- Single cell RNA-Seq data analysis
  - ✓ Pre-processing
  - ✓ *Downstream Analysis*
- Pre-processing of 10x Chromium data
- Pre-processing of Parsebio data

Diverse Cell Types, States and Interactions



http://www.cell.com/pictureshow/skin | https://library.med.utah.edu/WebPath/webpath.html

Advances in Single Cell Technology



Svensson *et al*, 2018

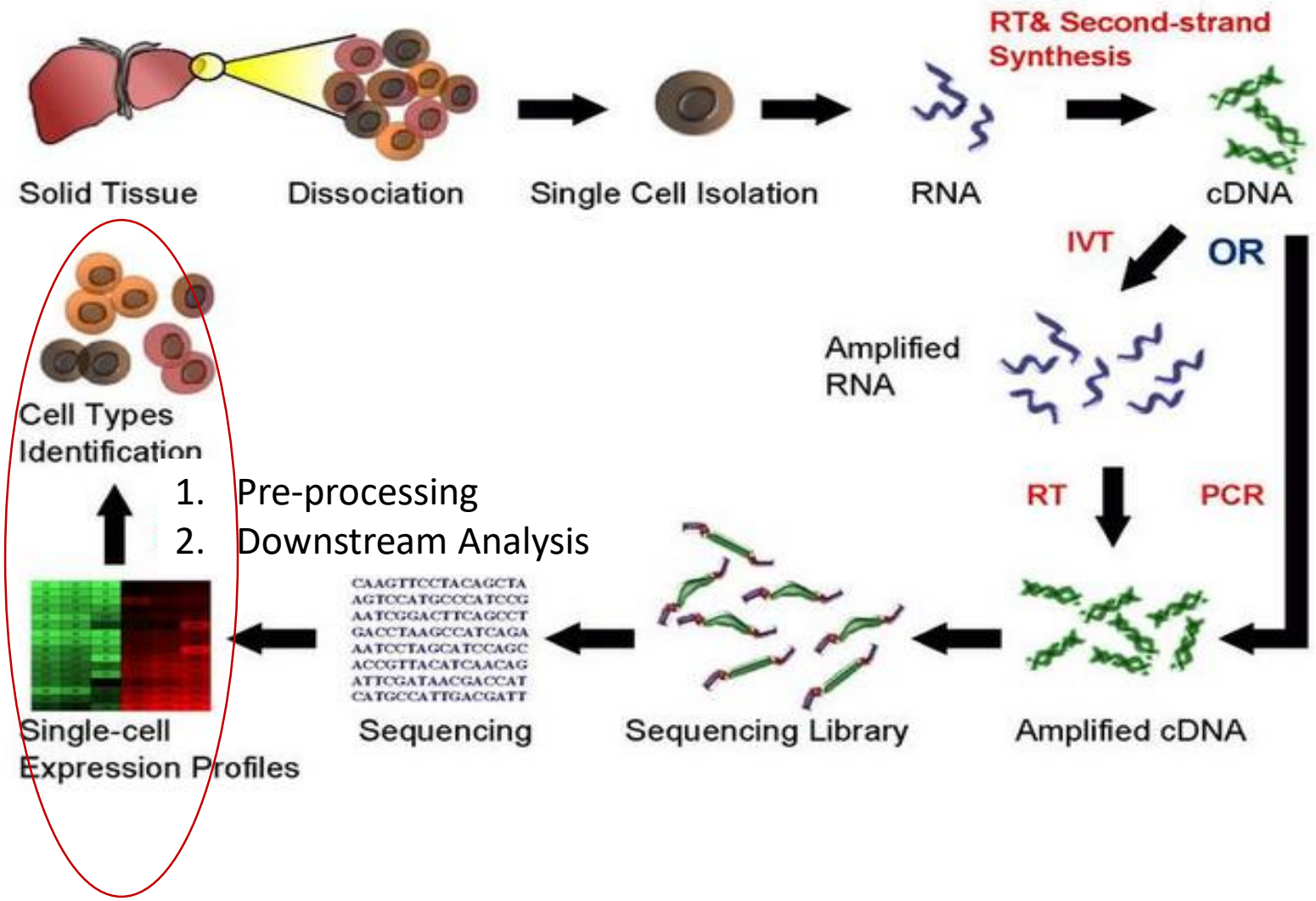
Single Cell RNA-Seq Applications

- Explore which cell types are present in a tissue
- Identify unknown/rare cell types or states
- Elucidate the changes in gene expression during differentiation processes or across time or states
- Identify genes that are differentially expressed in particular cell types between conditions
- Explore changes in expression among a cell type while incorporating spatial, regulatory, and/or protein information

Experimental Design

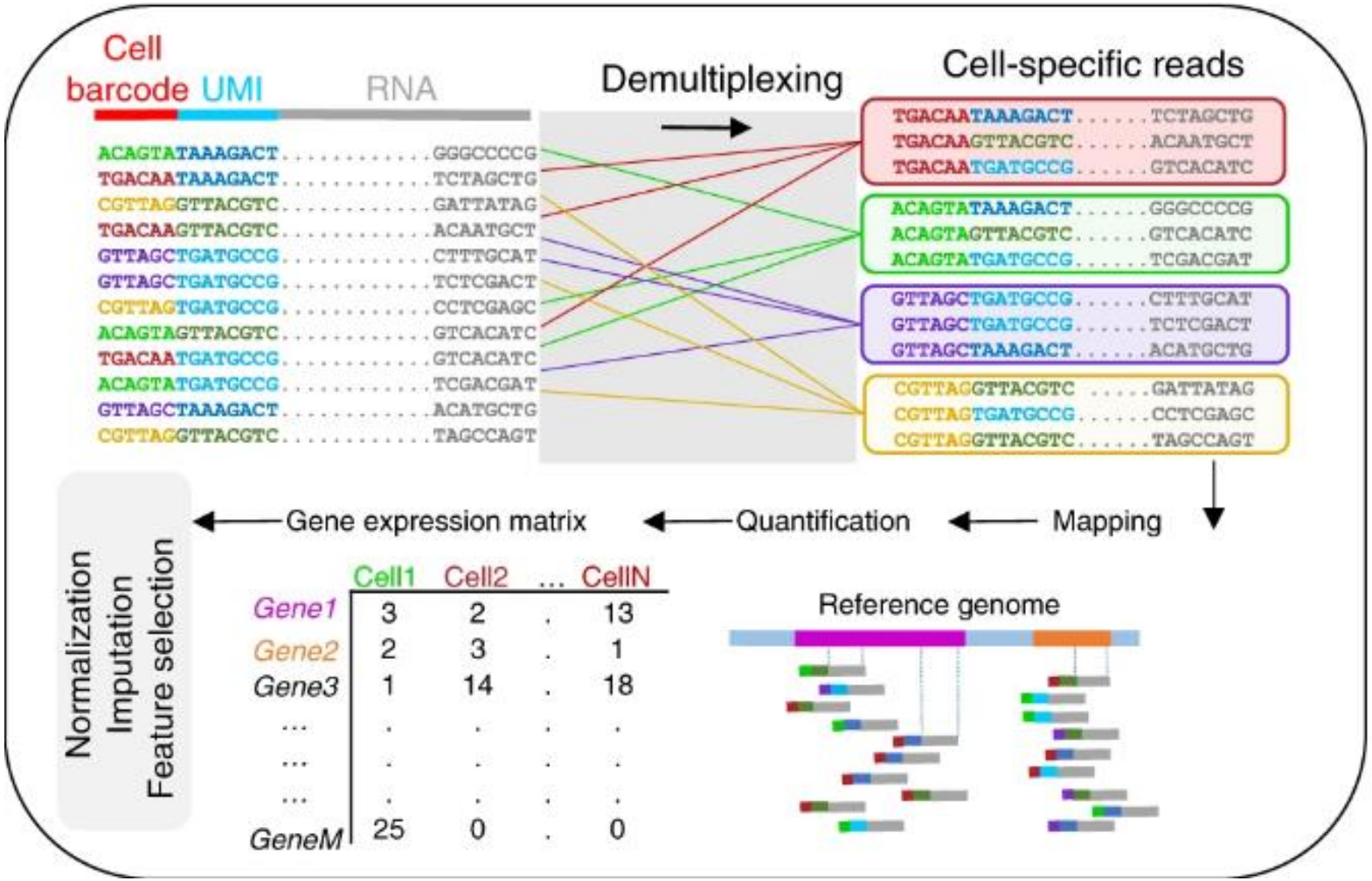
- Choose the platforms/protocols based on the biological question you wish to address
- How many samples and how many cells
  - https://satijalab.org/howmanycells/
- Single cell vs Single nucleus
- Isolate RNA and prepare libraries at same time for all samples or alternate sample groups to avoid batch confounding
- Do not confound sample groups by sex, age, or batch

Single Cell RNA-Seq Workflow



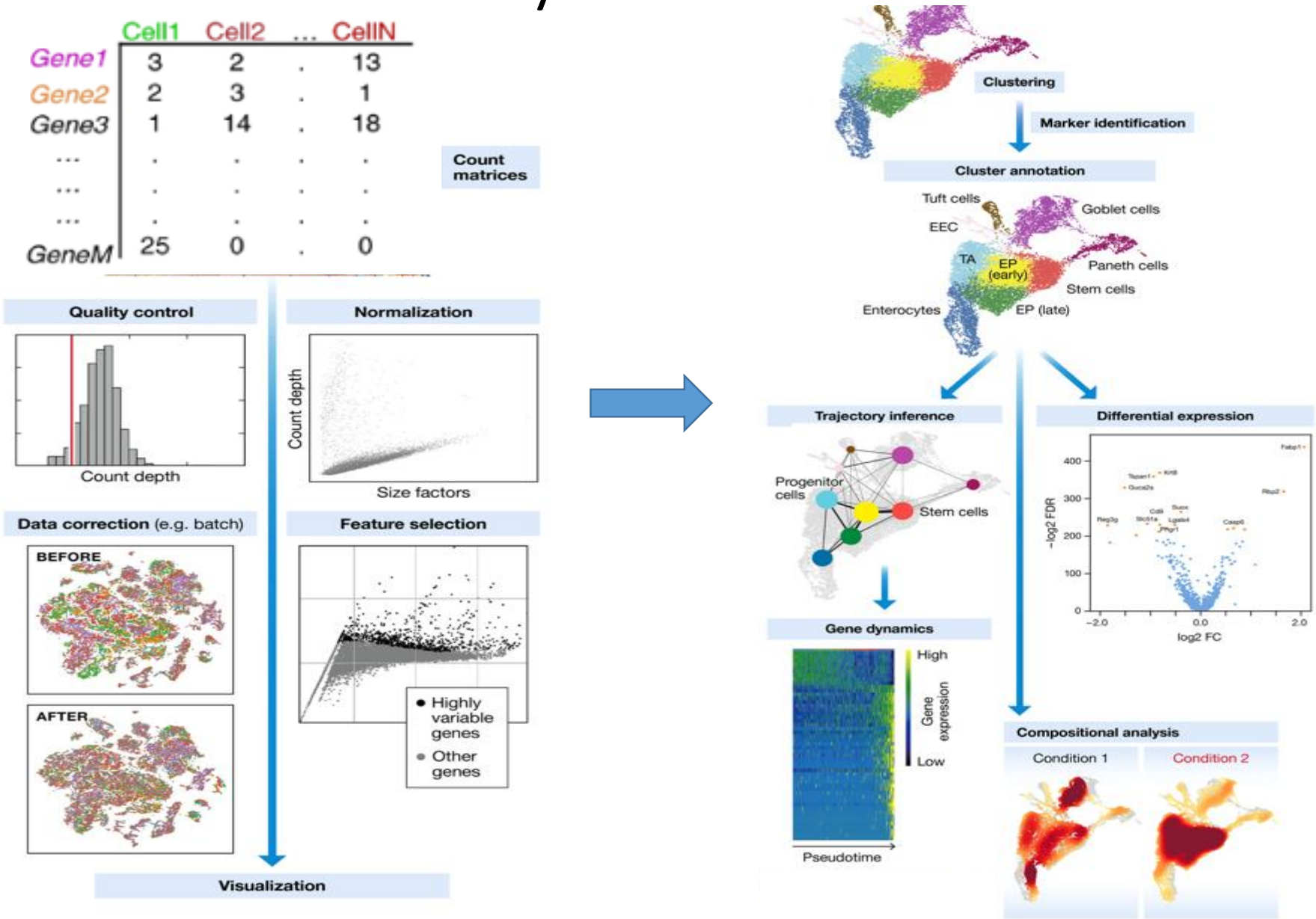
https://en.wikipedia.org/wiki/File:RNA-Seq_workflow-5.pdf

Pre-Processing of Single Cell Raw Data



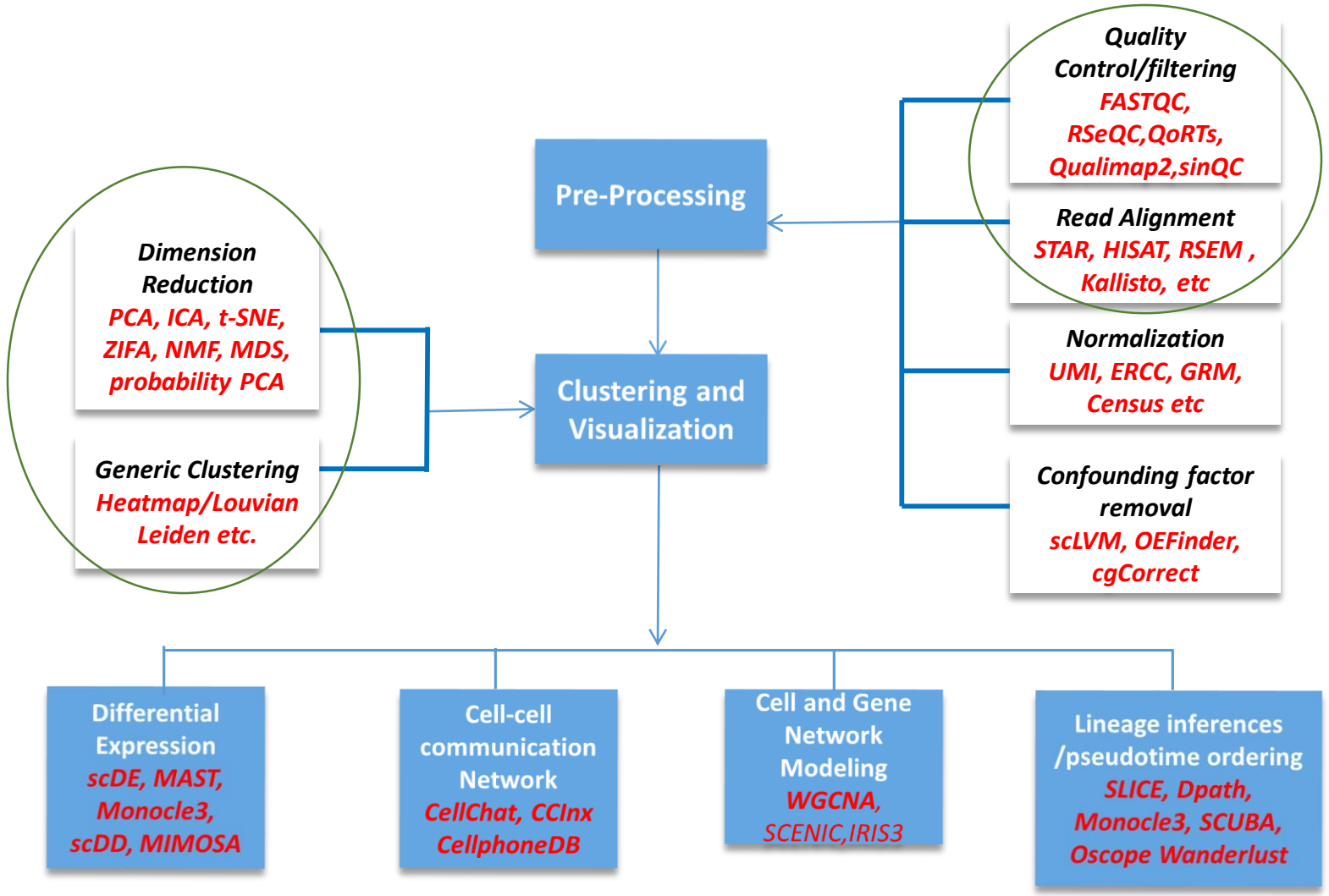
Lafzi *et al*, 2018

Downstream Analysis



Luecken *et al*, 2019

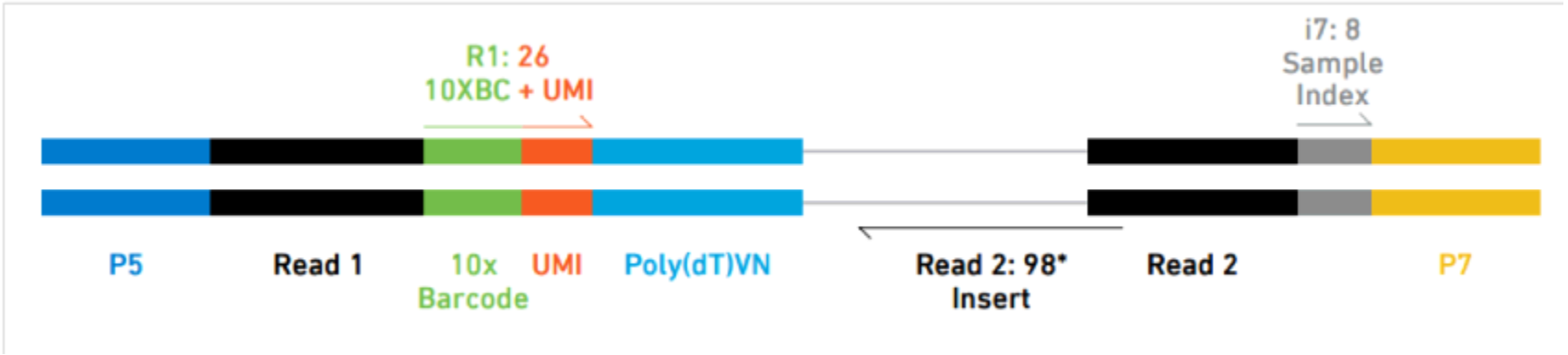
scRNA-seq Data Analysis Tools



CellRanger, bustools, Parsebio pipeline

R packages: Seurat4, Monocle3, Pagoda, scater, scran
Integrated analysis: Scell, fast Project

10X Chromium Data and CellRanger




```
$ cd /home/jdoe/runs
$ cellranger count --id=sample345 \
                   --transcriptome=/opt/refdata-cellranger-GRCh38-1.2.0 \
                   --fastqs=/home/jdoe/runs/HAWT7ADXX/outs/fastq_path \
                   --sample=mysample \
                   --expect-cells=1000
```

CellRanger Output: Summary

SUMMARY ANALYSIS

Estimated Number of Cells

8,627

Mean Reads per Cell	Median Genes per Cell
16,889	2,136

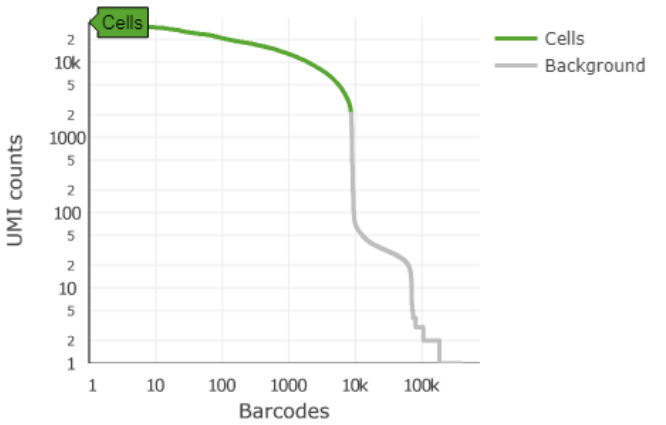
Sequencing

Number of Reads	145,709,844
Valid Barcodes	98.5%
Sequencing Saturation	33.3%
Q30 Bases in Barcode	96.8%
Q30 Bases in RNA Read	76.4%
Q30 Bases in UMI	96.9%

Mapping

Reads Mapped to Genome	91.6%
Reads Mapped Confidently to Genome	87.1%
Reads Mapped Confidently to Intergenic Regions	2.6%
Reads Mapped Confidently to Intronic Regions	11.6%
Reads Mapped Confidently to Exonic Regions	72.9%
Reads Mapped Confidently to Transcriptome	69.6%
Reads Mapped Antisense to Gene	1.1%

Cells



Estimated Number of Cells	8,627
Fraction Reads in Cells	96.1%
Mean Reads per Cell	16,889
Median Genes per Cell	2,136
Total Genes Detected	17,069
Median UMI Counts per Cell	6,306

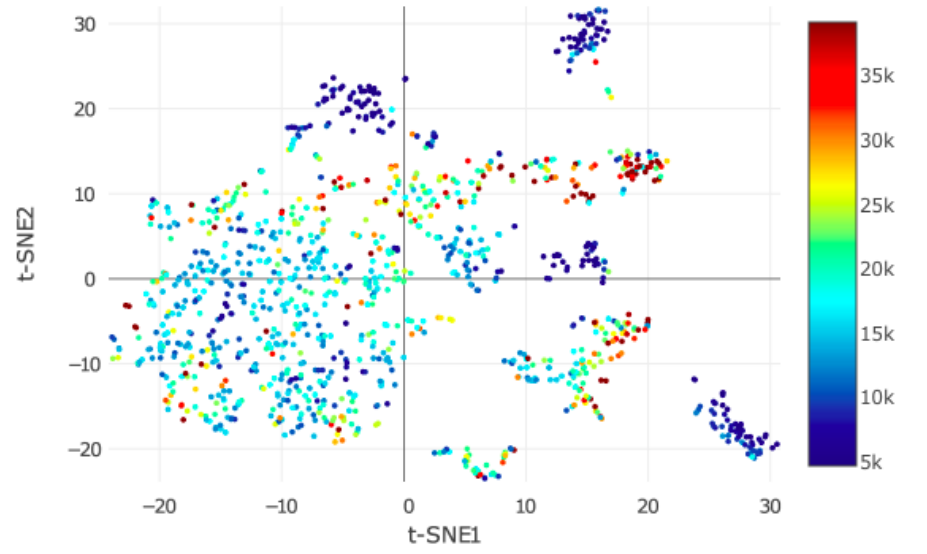
Sample

Name	plus_Jan9
Description	
Transcriptome	mm10_eGFP_rick
Chemistry	Single Cell 3' v2
Cell Ranger Version	2.1.0

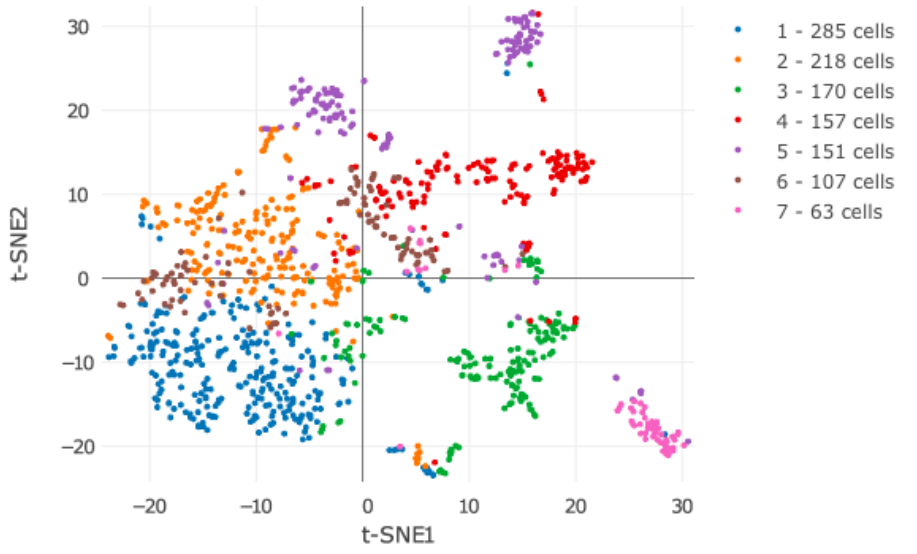
CellRanger Output: Analysis

SUMMARY ANALYSIS

t-SNE Projection of Cells Colored by UMI Counts



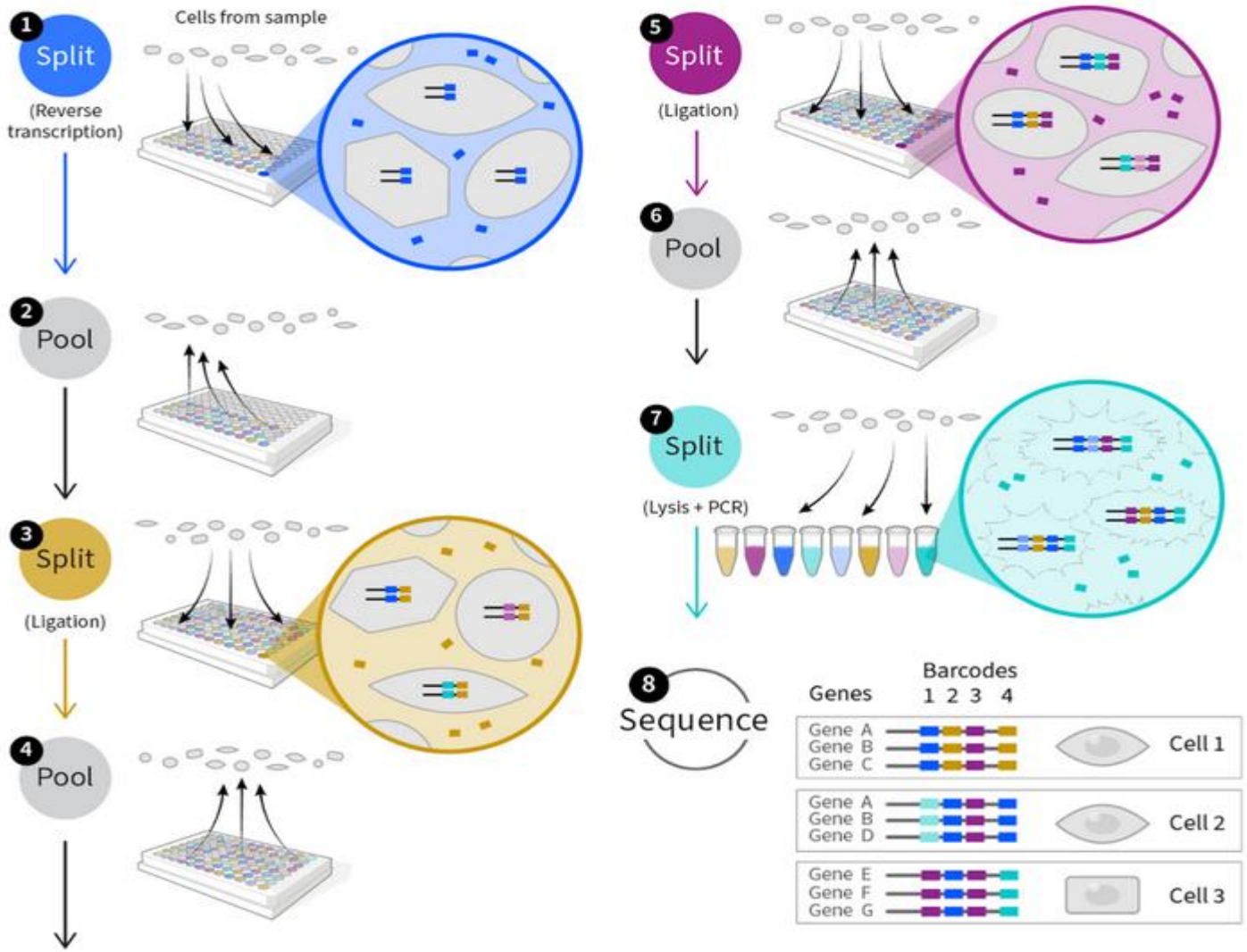
t-SNE projection of Cells Colored by Automated Clustering



Top Genes By Cluster (Log2 fold-change, p-value)

		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Cluster 7	
Gene ID	Gene name	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value
ENSG00000163536	SERPINI1	2.49	1e-13	-0.93	3e-01	-2.19	8e-05	-2.57	6e-08	-0.79	1e+00	0.35	1e+00	-0.43	1e+00
ENSG00000173714	WFIKKN2	2.37	4e-04	-1.81	4e-01	-4.09	3e-03	-3.84	3e-04	-1.70	1e+00	1.85	4e-01	-2.65	1e+00

Parsebio (SPLit-Seq)



Parsebio Raw Reads

Read 1 - transcript sequence

Illumina index sequence

transcript sequence

@NB551368:195:HCVJVBGXJ:1:11104:5453:1061 1:N:0:ACTTGA

AAGCCNTGGTATCAACGCAGAGTGAATNGGAGATGTCACATNCCTATGACCCTAATGTACATAAAATTGACAGG

+

A/AAA#AEE/EEA/EAEEEEEEEEEA#EE///6/EEE/6A#//<<///E//<EAE//<E/AA</A//E//<</

Read 2 - barcodes and degenerate sequence

@NB551368:195:HCVJVBGXJ:1:11104:13360:1088 2:N:0:ACTTGA

TCAGAGGTATAGGCTAACGTGGCCGATGTTTCGCATCGGCGTACGACTGTCTGTCAATCCACGTGCTTGAGACTGTGGGGGCGATG

+

AAAAAEEEEEEEEEEEEAA/</EAEEEEEA<AEEEEE</AE6<<AA/<AAEE</AAEEE<<A<<AAAEAEEEEAE<EA/<E<A<<

Round 3 barcode

30bp Linker

Round 2 barcode

22bp Linker

Round 1 barcode

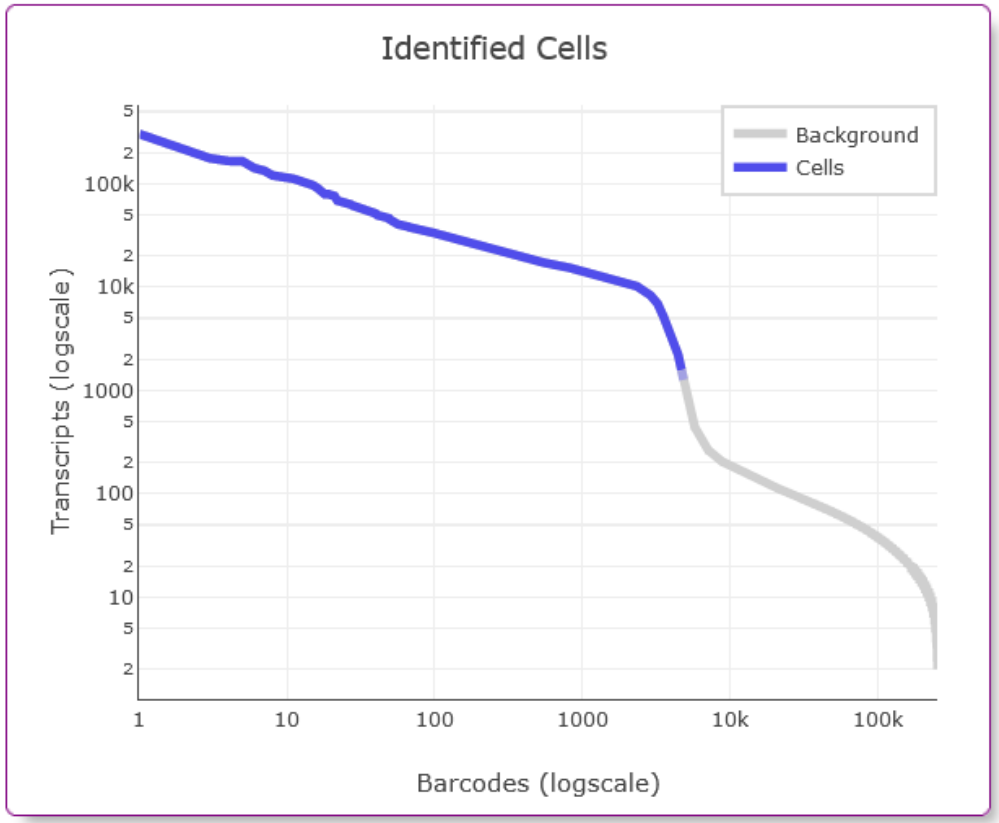
```
singularity exec /dfs3a/singularity_containers/rcic/parsebio.simg split-pipe \
--mode all \
--kit WT \
--genome_dir ${GPATH} \
--fq1 ${IPATH}/Sublibrary1_S1_R1_001.fastq.gz \
--fq2 ${IPATH}/Sublibrary1_S1_R2_001.fastq.gz \
--output_dir ${OPATH} \
--samp_list ${SPATH}
```

Parsebio pre-processing Output: Summary

Parse Experimental Report Summary Clustering

Sample: S4 **Wells:** B1-B4

Estimated Number of Cells	4,795
mm10 Number of Cells Detected	4,795
mm10 Median Transcripts/Cell	10,058
mm10 Median Genes/Cell	3,377
Mean Reads/Cell	41,891
Number of Reads	200,868,545
Sequencing Saturation	0.504
BC1 (RT) >Q30	0.902
BC2 >Q30	0.927
BC3 >Q30	0.939
cDNA >Q30	0.927



Parsebio pre-processing Output : Analysis

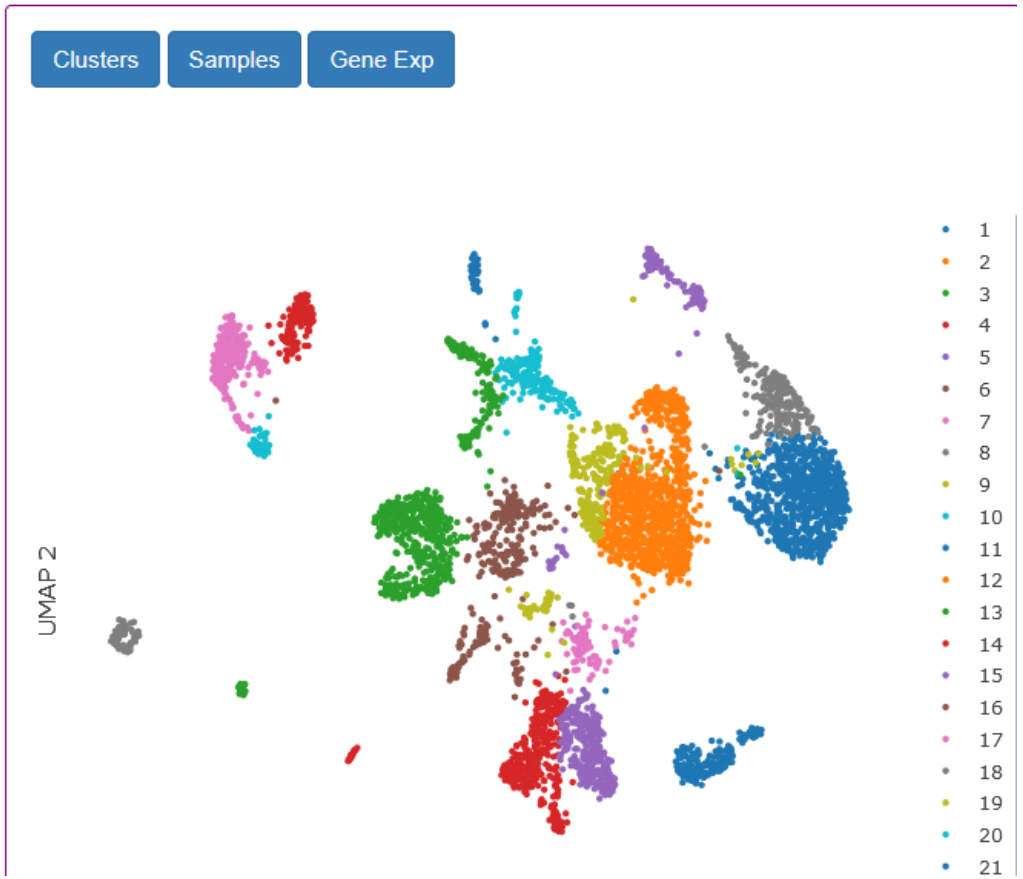
Parse Experimental Report Summary Clustering

Show clusters

- all
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

Differentially Expressed Genes

Cluster	Gene name	Score	Log2 FC	Pval adj
1	A830012C17Rik	33.2	3.4	1.5E-27
1	Adora2a	26.4	3.1	8.1E-116
1	Gpr6	23.8	2.5	8.1E-02
1	Drd2	21.0	3.4	0.0E+00
1	Penk	19.2	2.9	3.3E-117
1	Oprd1	18.9	2.9	1.1E-56
1	Gm30313	14.1	2.6	3.4E-11
1	Nt5e	13.7	2.5	1.5E-35
1	Gm39043	12.1	2.4	6.0E-05
1	Gm37711	10.3	2.1	4.2E-01
1	Gm19410	10.0	2.3	2.0E-29
1	Nell1	8.9	2.3	3.4E-236
1	B230110G15Rik	7.8	2.1	1.8E-15



Alternative Tools for pre-processing

ARTICLES
https://doi.org/10.1038/s41592-022-01408-3

nature methods

Check for updates

Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data

Dongze He[1], Mohsen Zakeri[2], Hirak Sarkar[3], Charlotte Sones and Rob Patro[2]

The rapid growth of high-throughput single-cell and single-nucleus RNA-sequencin has produced a wealth of data over the past few years. The size, volume and distinc the development of new computational methods to accurately and efficiently quar that constitute the input to downstream analyses. We introduce the alevin-fry fra In addition to being faster and more memory frugal than other accurate quantifica memory scalability and false-positive expression issues that are exhibited by ot alevin-fry can be effectively used to quantify sc/snRNA-seq data, and also how th tion required as input for RNA velocity analyses can be seamlessly extracted from t normal gene expression count matrices.

nature biotechnology

LETTERS
https://doi.org/10.1038/s41587-021-00870-2

Check for updates

Modular, efficient and constant-memory single-cell RNA-seq preprocessing

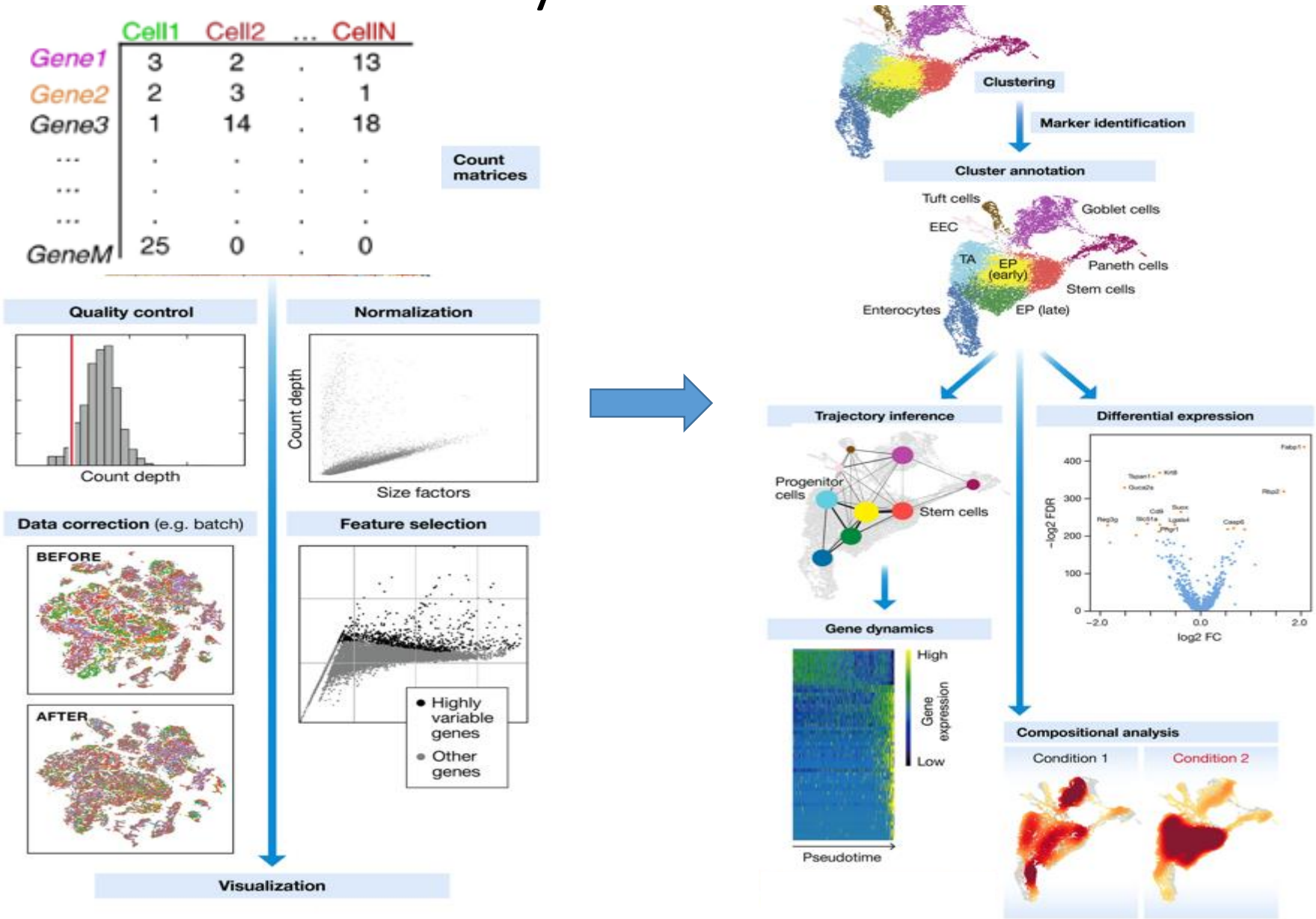
Páll Melsted[1,8], A. Sina Booeshaghi[2,8], Lauren Liu[3], Fan Gao[4,5], Lambda Lu[4], Kyung Hoi (Joseph) Min[6], Eduardo da Veiga Beltrame[4], Kristján Eldjárn Hjörleifsson[3], Jase Gehring[7] and Lior Pachter[3,4]

We describe a workflow for preprocessing of single-cell RNA-sequencing data that balances efficiency and accuracy. Our workflow is based on the kallisto and bustools programs, and is near optimal in speed with a constant memory requirement providing scalability for arbitrarily large datasets. The workflow is modular, and we demonstrate its flexibility by showing how it can be used for RNA velocity analyses.

that is untenable given the pace of improvement in technology and the corresponding increase in data volume.

In recent work, we introduced a format for scRNA-seq data that makes possible the development of efficient workflows by virtue of decoupling the computationally demanding step of associating reads to transcripts and genes (alignment) from the other steps required for scRNA-seq preprocessing[10]. This format, called BUS

Downstream Analysis



Luecken *et al*, 2019

MatrixMarket Format for Gene Count Matrix

Sparse Gene Count Matrix

```
%%MatrixMarket matrix coordinate integer general
%Rows=cells (6907), Cols=genes (29552)
6907 29552 5550212
1 1 43
1 2 6
1 3 29
1 4 108
1 5 11
1 7 3
1 8 3
1 9 5
1 11 21
1 12 3
1 15 6
1 16 7
1 17 26
1 18 231
1 19 30
1 21 28
1 22 42
1 26 1
1 29 4
1 30 12
```

Gene Info

```
gene_id,gene_name,genome
ENSFCAG00000000001,INTS6L,felis
ENSFCAG00000000007,HMGCR,felis
ENSFCAG00000000015,CEP192,felis
ENSFCAG00000000022,RASGRP1,felis
ENSFCAG00000000023,GPR39,felis
ENSFCAG00000000024,LYPD1,felis
ENSFCAG00000000028,RCN3,felis
ENSFCAG00000000029,APOO,felis
ENSFCAG00000000030,CXHXorf58,felis
ENSFCAG00000000031,CB1H4orf19,felis
ENSFCAG00000000032,RELL1,felis
ENSFCAG00000000034,DRC1,felis
ENSFCAG00000000035,OTOF,felis
ENSFCAG00000000036,CA3H2orf70,felis
ENSFCAG00000000038,KAT5,felis
ENSFCAG00000000043,PPARGC1A,felis
ENSFCAG00000000044,LRP6,felis
ENSFCAG00000000049,SYNE1,felis
ENSFCAG00000000050,TARDBP,felis
ENSFCAG00000000051,MASP2,felis
```

Cell Info

```
AAACCCAAGGTAGCAC-1
AAACCCAGTCACTCTC-1
AAACGAAAGATGAATC-1
AAACGAAAGCACCGTC-1
AAACGAATCCCTCATG-1
AAACGAATCTGAATGC-1
AAACGCTAGCACCTGC-1
AAACGCTAGCTTAGTC-1
AAACGCTCAAGATTGA-1
AAACGCTCACTTCAGA-1
AAACGCTTCTGGTGCG-1
AAAGAACAGAAGTCCG-1
AAAGAACAGACGAGCT-1
AAAGAACAGAGAGAAC-1
AAAGAACAGCATCCCG-1
AAAGAACGTTTCACAG-1
AAAGGATAGATGCAGC-1
AAAGGATAGGTATCTC-1
AAAGGATCAGGACGAT-1
AAAGGATGTAGCTAAA-1
AAAGGGCCATACACCA-1
```

Limitations of Single Cell Data

- Low capture rate, low depth of sequencing per cell, 3’ only
- High biological and technical variability across cells/samples
- Missing spatial information

Useful Links

- CellRanger manual

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

- Biojhub3 on HPC3:

  https://hpc3.rcic.uci.edu/biojhub3/hub/login

- Workshop data:

  /dfs6/pub/ucightf/workshop/

- Seurat manual

  https://satijalab.org/seurat/

- 10x Cloud

  https://www.10xgenomics.com/products/cloud-analysis