Introduction to Single-Cell RNA-seq

Thanks to Dennis Wylie for some slides

Why single-cell RNA-seq?

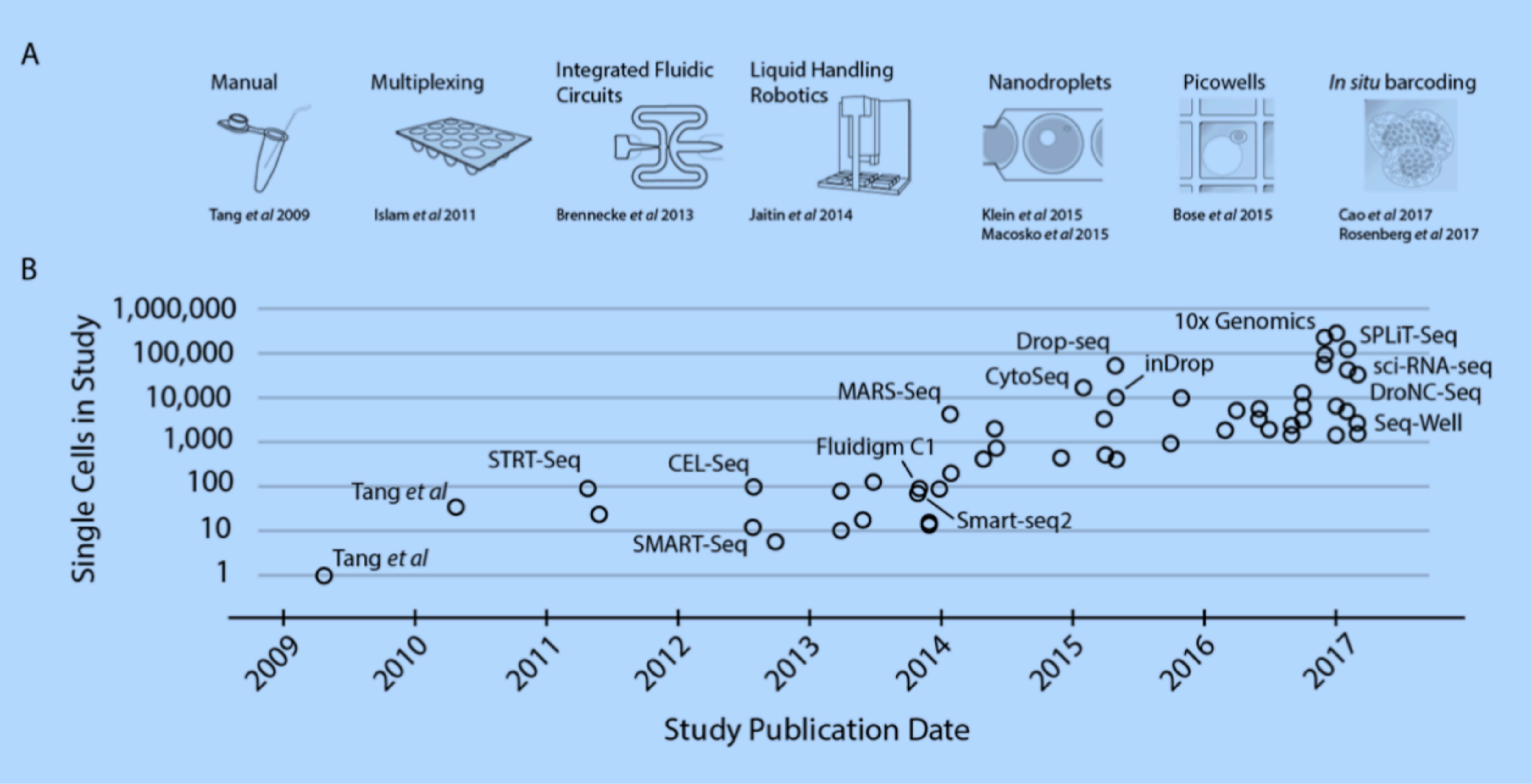
- Allows profiling of gene expression in individual cells.
  - To look at heterogeneity across cell type subpopulations
  - Identify cell to cell variations in alternative splicing.

Unique Challenges with single-cell RNA-seq

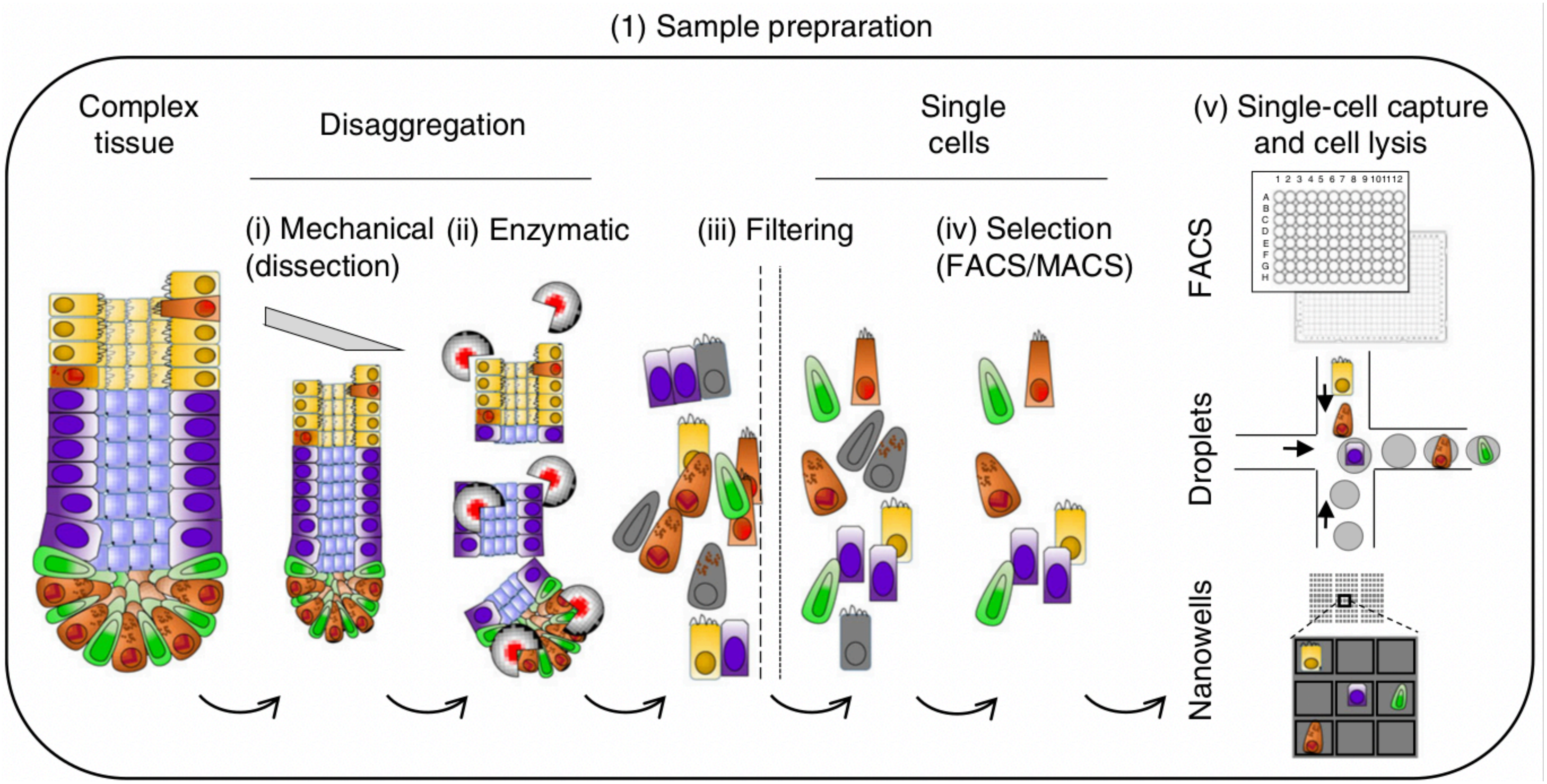
- **Gene dropouts**
    - Due to low amounts of RNA per cell.
    - Some cells are easier to capture than others.
- **Large, but sparse gene expression matrix**
    - Expression values for all genes across 10^2 to 10^5 cells.
    - Many zeros

Single-cell RNA-seq Technology Improvements

https://arxiv.org/abs/1704.01379

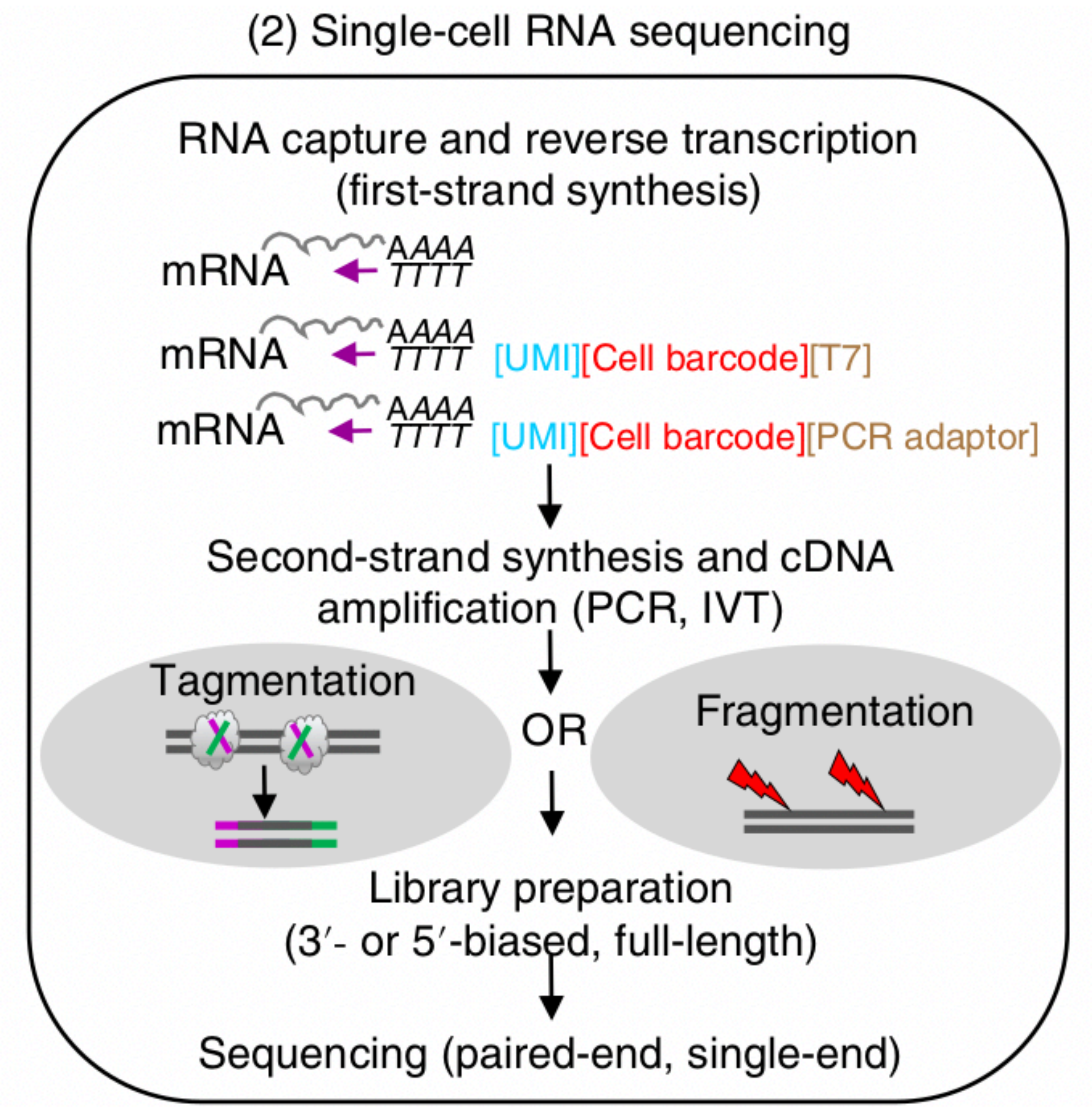


Single-cell RNA-seq Sample Prep



Taken from Lafzi et al. (2018)

Single-cell RNA-Seq Library Prep


(2) Single-cell RNA sequencing
RNA capture and reverse transcription (first-strand synthesis)
mRNA
AAAA
TTTT
mRNA
AAAA
TTTT
[UMI][Cell barcode][T7]
mRNA
AAAA
TTTT
[UMI][Cell barcode][PCR adaptor]
Second-strand synthesis and cDNA amplification (PCR, IVT)
Tagmentation
OR
Fragmentation
Library preparation (3′- or 5′-biased, full-length)
Sequencing (paired-end, single-end)

Taken from Lafzi et al. (2018)

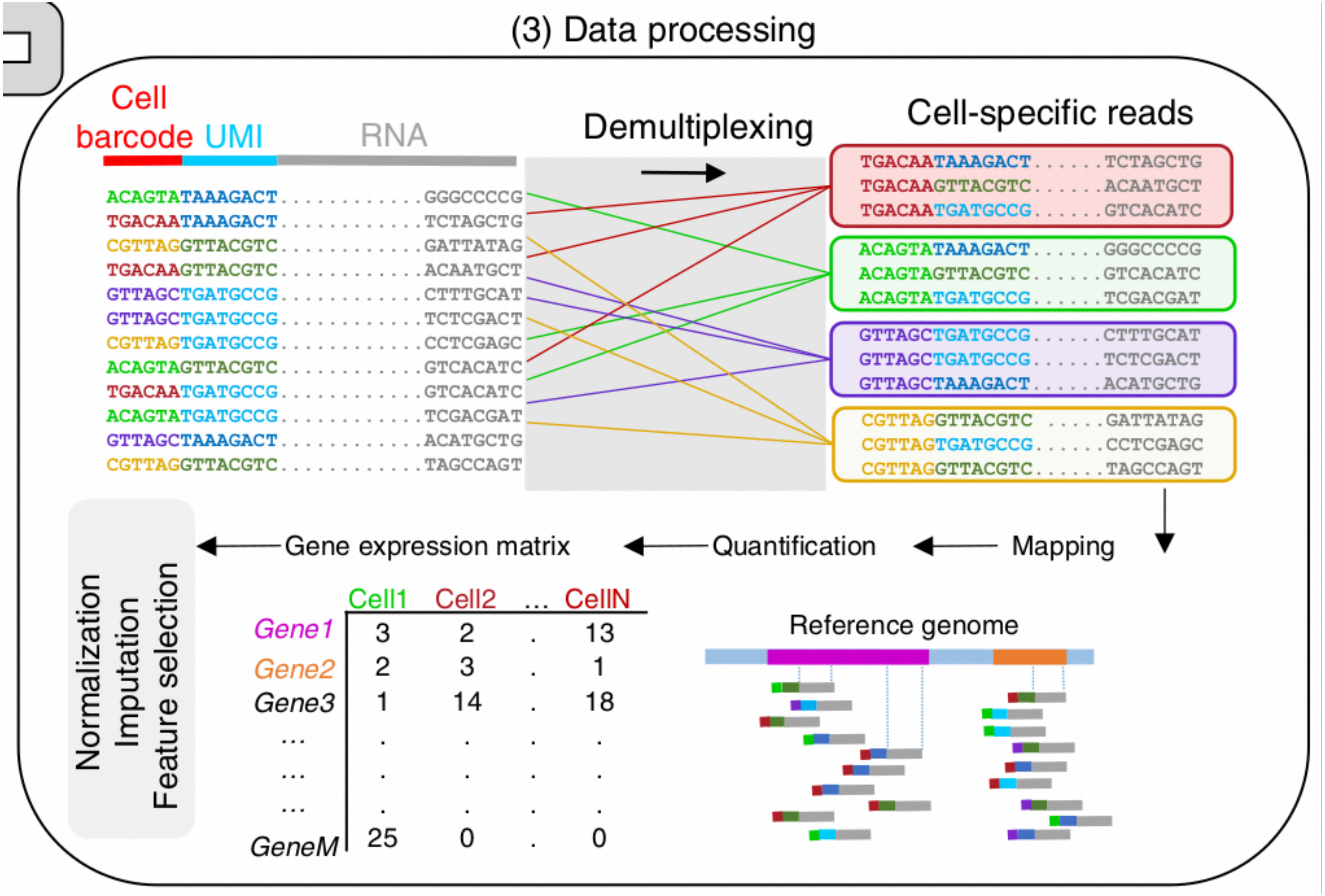
- **UMI (Unique molecular index)**
  - Random 4-20 bp sequences attached to each RNA fragment/template to uniquely identify that RNA fragment/template.
  - One per fragment.
  - For detection of PCR duplicates.
- **Cell barcode**
  - A cell-specific sequence attached to RNA fragments.
  - One per cell
  - For differentiating by cell.
- **Sample barcode/index**
  - One per sample
  - Allows pooling multiple samples on the same sequencing run.

Coverage Recommendations

- How many cells per sample?
    - ~10,000 cells per ‘typical’ sample.
- How many reads per cell?
    - 30,000-50,000 reads per cell for ‘typical’ samples.
- This would take up >1 lane/sample

Coverage decisions should be made based on the purpose/questions of the study.

Single-cell RNA-Seq Data Analysis

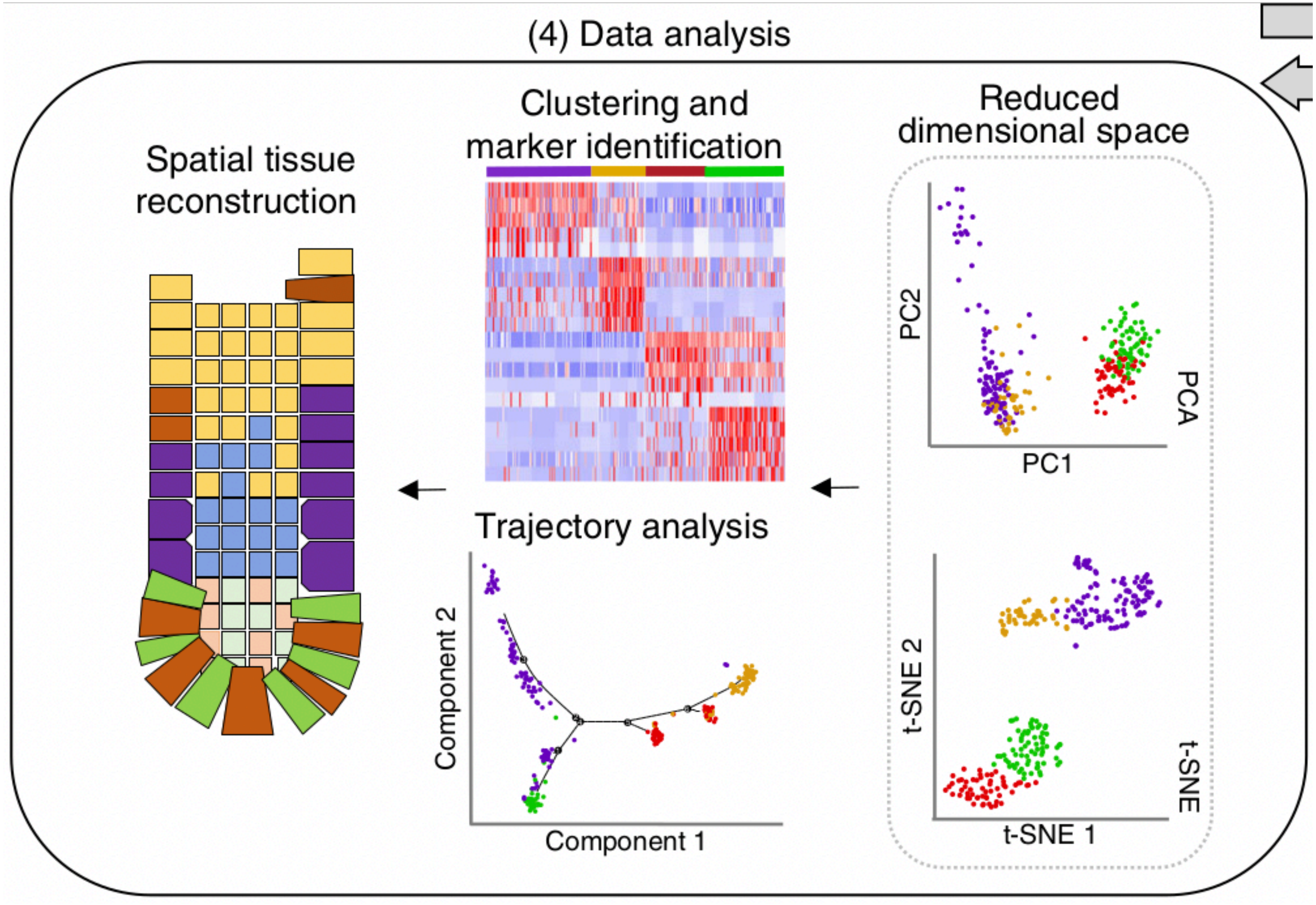


(3) Data processing
Cell barcode
UMI
RNA
ACAGTATAAAGACT............GGGCCCCG
TGACAATAAAGACT............TCTAGCTG
CGTTAGGTTACGTC............GATTATAG
TGACAAGTTACGTC............ACAATGCT
GTTAGCTGATGCCG............CTTTGCAT
GTTAGCTGATGCCG............TCTCGACT
CGTTAGTGATGCCG............CCTCGAGC
ACAGTAGTTACGTC............GTCACATC
TGACAATGATGCCG............GTCACATC
ACAGTATGATGCCG............TCGACGAT
GTTAGCTAAAGACT............ACATGCTG
CGTTAGGTTACGTC............TAGCCAGT
Demultiplexing
Cell-specific reads
TGACAATAAAGACT......TCTAGCTG
TGACAAGTTACGTC......ACAATGCT
TGACAATGATGCCG......GTCACATC
ACAGTATAAAGACT......GGGCCCCG
ACAGTAGTTACGTC......GTCACATC
ACAGTATGATGCCG......TCGACGAT
GTTAGCTGATGCCG......CTTTGCAT
GTTAGCTGATGCCG......TCTCGACT
GTTAGCTAAAGACT......ACATGCTG
CGTTAGGTTACGTC .....GATTATAG
CGTTAGTGATGCCG......CCTCGAGC
CGTTAGGTTACGTC......TAGCCAGT
Mapping
Quantification
Gene expression matrix
Normalization
Imputation
Feature selection
Reference genome

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

Taken from Lafzi et al. (2018)

Single-cell RNA-Seq Data Analysis


(4) Data analysis
Spatial tissue reconstruction
Clustering and marker identification
Reduced dimensional space
PC2
PC1
PCA
Trajectory analysis
Component 2
Component 1
t-SNE 2
t-SNE 1
t-SNE

Taken from Lafzi et al. (2018)

Single-cell RNA-seq Analysis Steps

Many, many tools available: http://www.scRNA-tools.org

- **QC Assessment** (FastQC)
- **Alignment to reference** (STAR, BWA)
  - Pseudoalignment + quantification (kallisto, salmon etc)
- **Quantification** (Cell Ranger, UMI-tools)
  - Error correction of UMI
  - UMI demultiplexing

Single-cell RNA-seq Analysis Steps

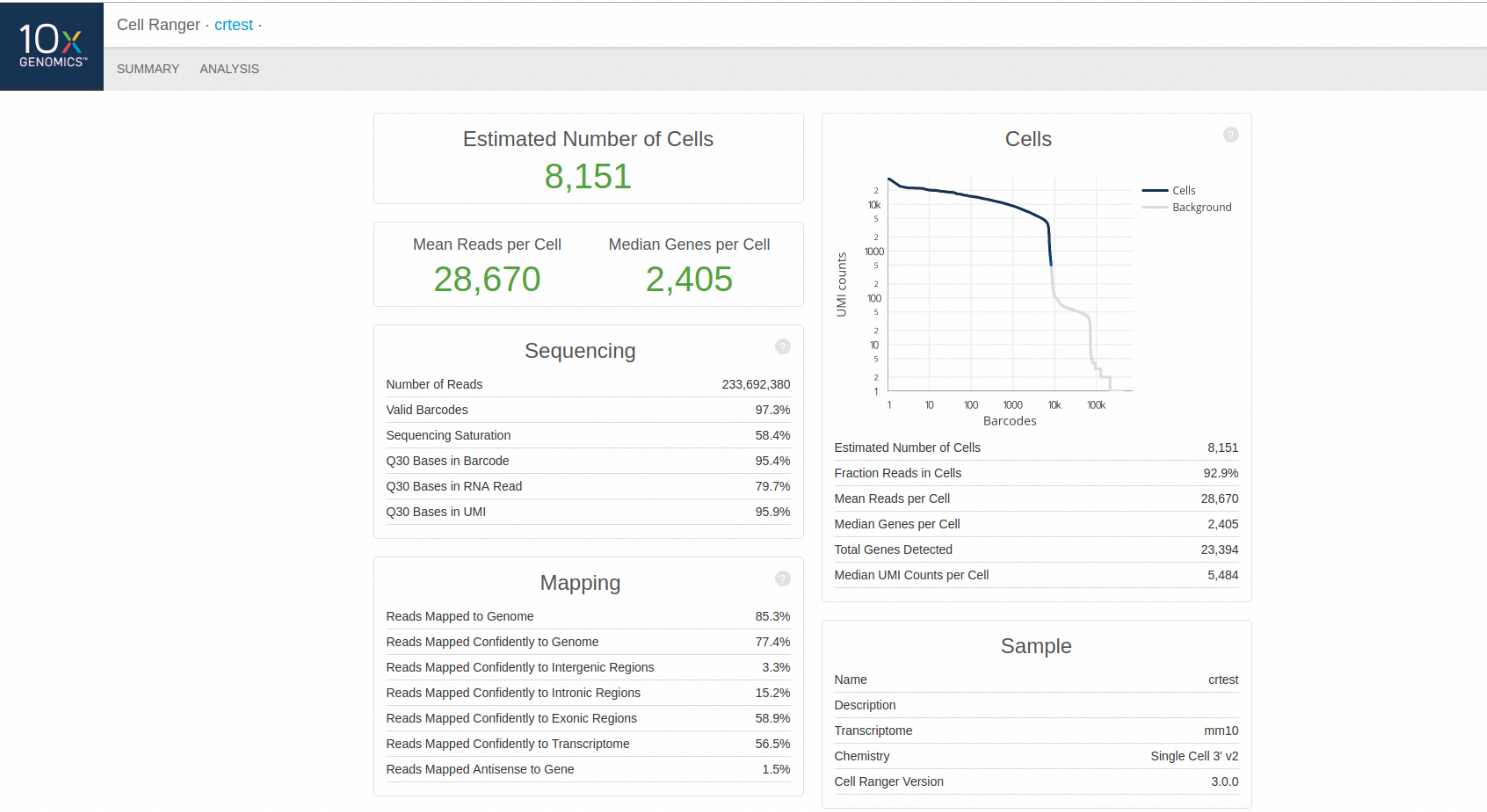
- **Imputation** (SAVER)
- **Normalization** (scran, TPM, CPM, TMM)
- **Dimensionality reduction** (PCA, tSNE, etc)
- **Clustering** (hierarchical, k-means, seurat, etc)
- **Differential expression analysis** (deseq2, edgeR, limma, MAST)

Cell Ranger

- Cell ranger is a set of analysis pipelines that process Chromium (10x) single-cell RNA-Seq data.

1. **Assess quality**
2. **Aligns reads (using star)**
3. **UMI, cell barcode error correction and demultiplexing**
4. **Generates a gene expression matrix after 1 and 2.**
5. Will also do further downstream analysis (normalization, clustering, DE analysis).
6. Analyses provided in a nice interactive report.

Cell Ranger Web QC Page



Estimated Number of Cells

8,151

Mean Reads per Cell	Median Genes per Cell
28,670	2,405

Sequencing

Number of Reads	233,692,380
Valid Barcodes	97.3%
Sequencing Saturation	58.4%
Q30 Bases in Barcode	95.4%
Q30 Bases in RNA Read	79.7%
Q30 Bases in UMI	95.9%

Mapping

Reads Mapped to Genome	85.3%
Reads Mapped Confidently to Genome	77.4%
Reads Mapped Confidently to Intergenic Regions	3.3%
Reads Mapped Confidently to Intronic Regions	15.2%
Reads Mapped Confidently to Exonic Regions	58.9%
Reads Mapped Confidently to Transcriptome	56.5%
Reads Mapped Antisense to Gene	1.5%

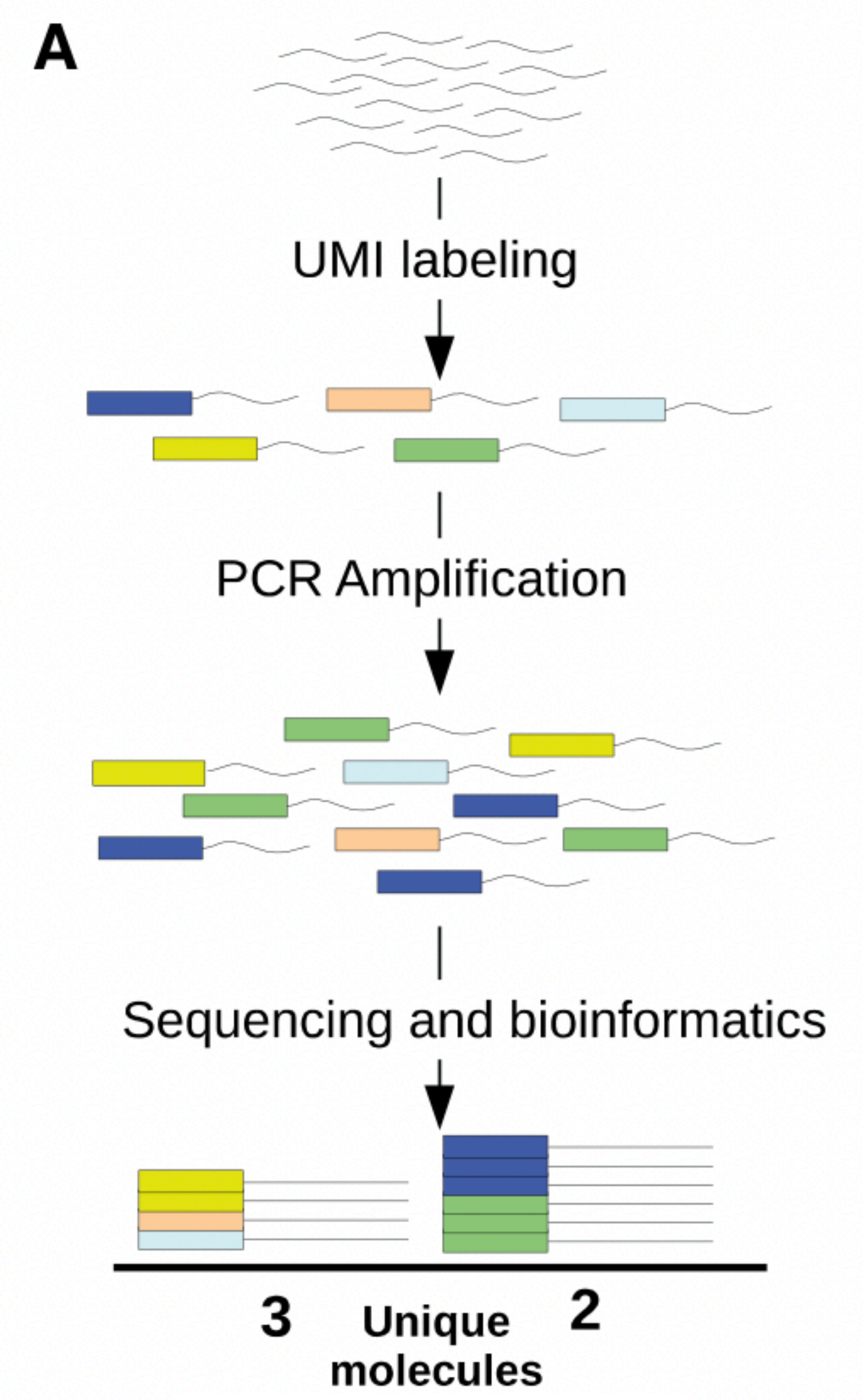
Cells

Estimated Number of Cells	8,151
Fraction Reads in Cells	92.9%
Mean Reads per Cell	28,670
Median Genes per Cell	2,405
Total Genes Detected	23,394
Median UMI Counts per Cell	5,484

Sample

Name	crtest
Description	
Transcriptome	mm10
Chemistry	Single Cell 3' v2
Cell Ranger Version	3.0.0

Single-cell RNA-Seq Data Analysis
UMI demultiplexing and error correction



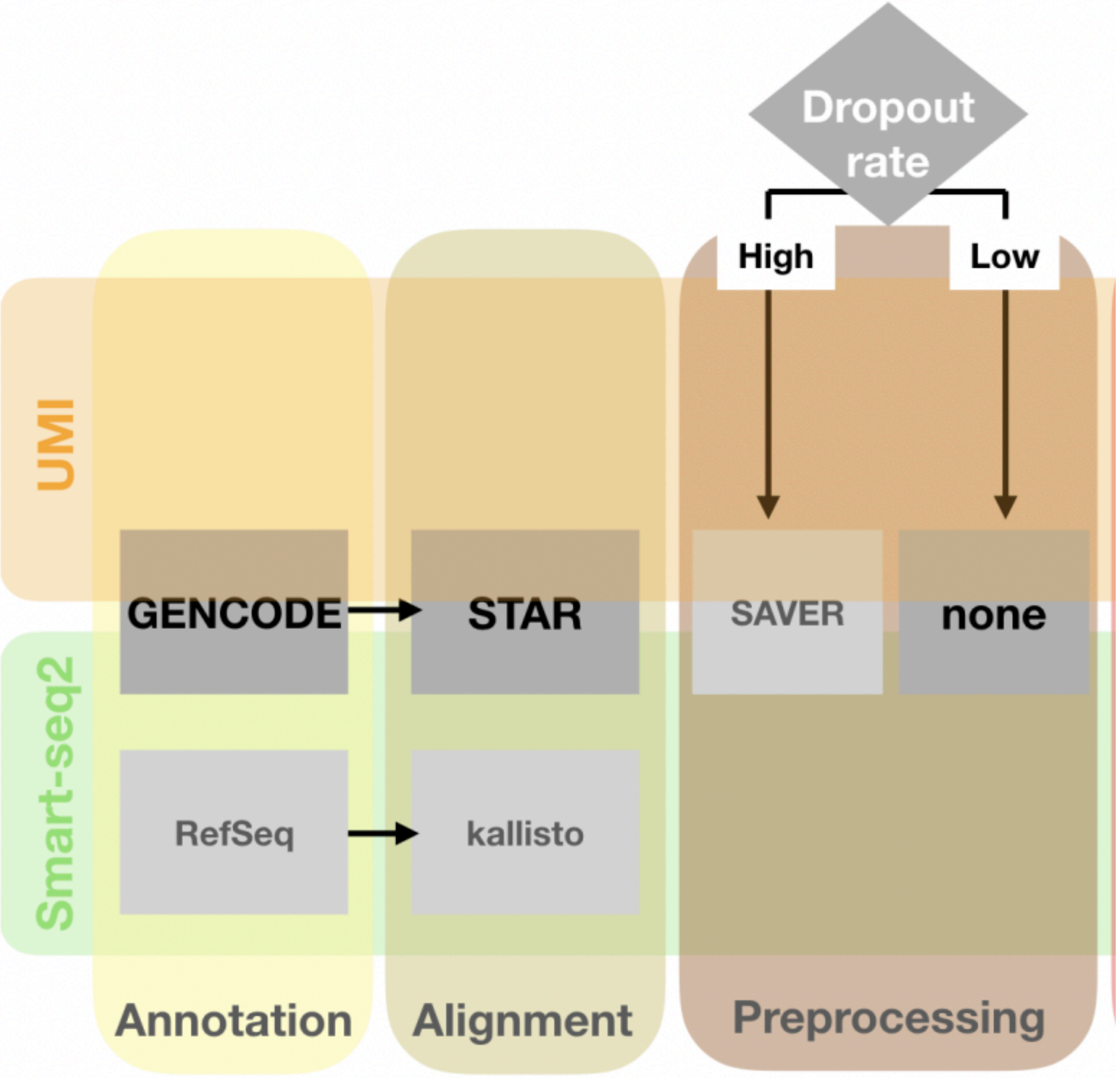
Taken from Smith et al. (2017)

- UMIs can have sequencing errors.
- Some sequences tend to have more than others.
- UMI error correction/filtering
  - No homopolymers
  - No N's
  - No bases with quality lower than 10
  - If a UMI is 1 base pair substitution away from a higher-count UMI, it's corrected to the higher count UMI if they share a cell barcode.

Imputation

- Method to deal with dropouts (genes with zero counts) by borrowing information from other cells.
- For a dropout gene X in cell Y,
    - Impute expression based on expression of gene X in other similar cells.

Single-cell RNA-seq Analyses Benchmarked



Taken from Vieth et al. (2019).

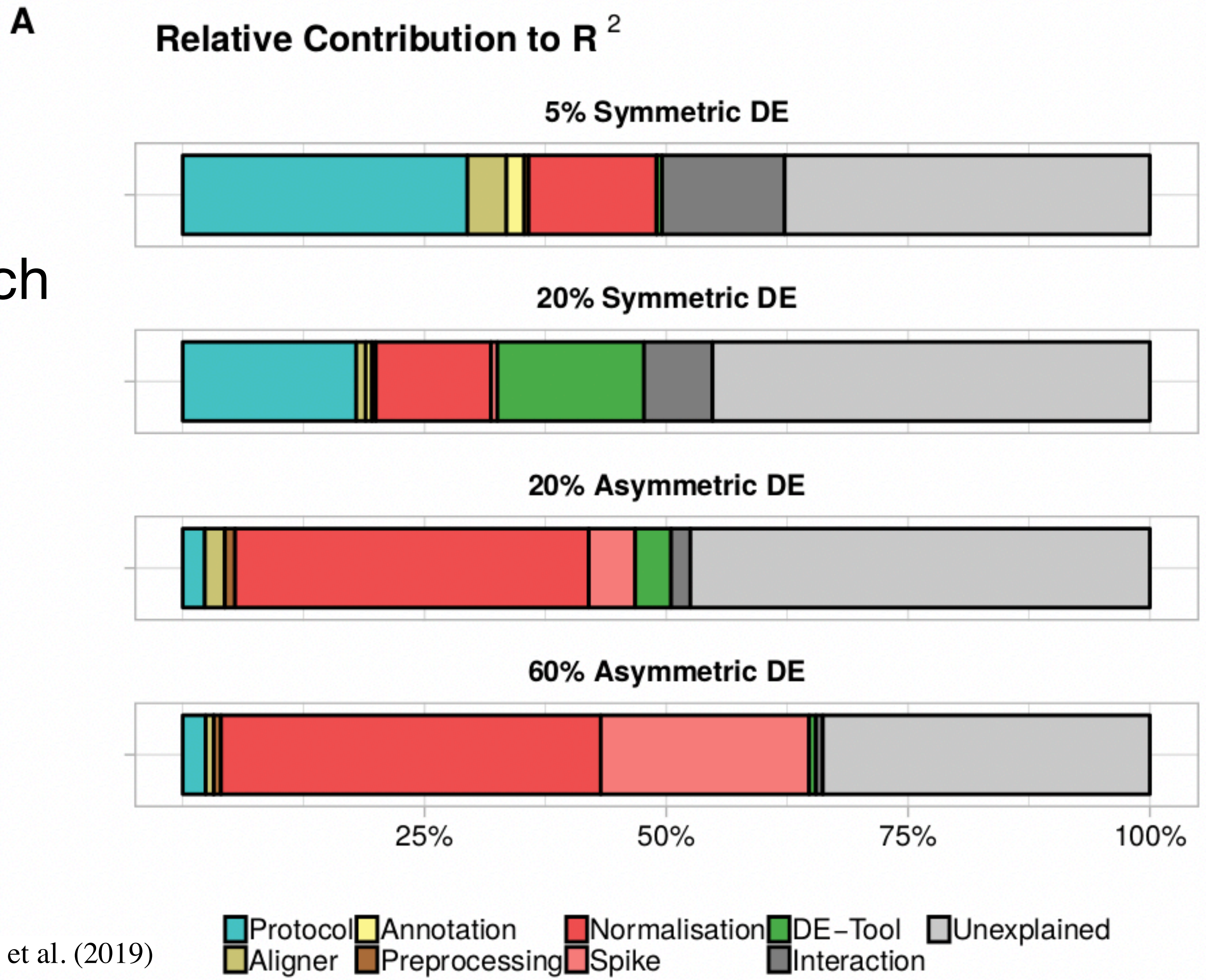
- STAR mapper works well for UMI based/ chromium (10x) scRNA-Seq data
- Imputation only if dropout rate is high

Normalization is important!

- **Remove Technical variations without removing biological variation**
  - dropout events, amplification bias, sequencing depth
  - batch effects
- **Why is it different from normalization of bulk RNA-Seq?**
  - **“**One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up-and down-regulated so that the mean total mRNA content does not differ between groups. This assumption is no longer true when diverse cell types are considered.” - Taken from Vieth et al. (2019).

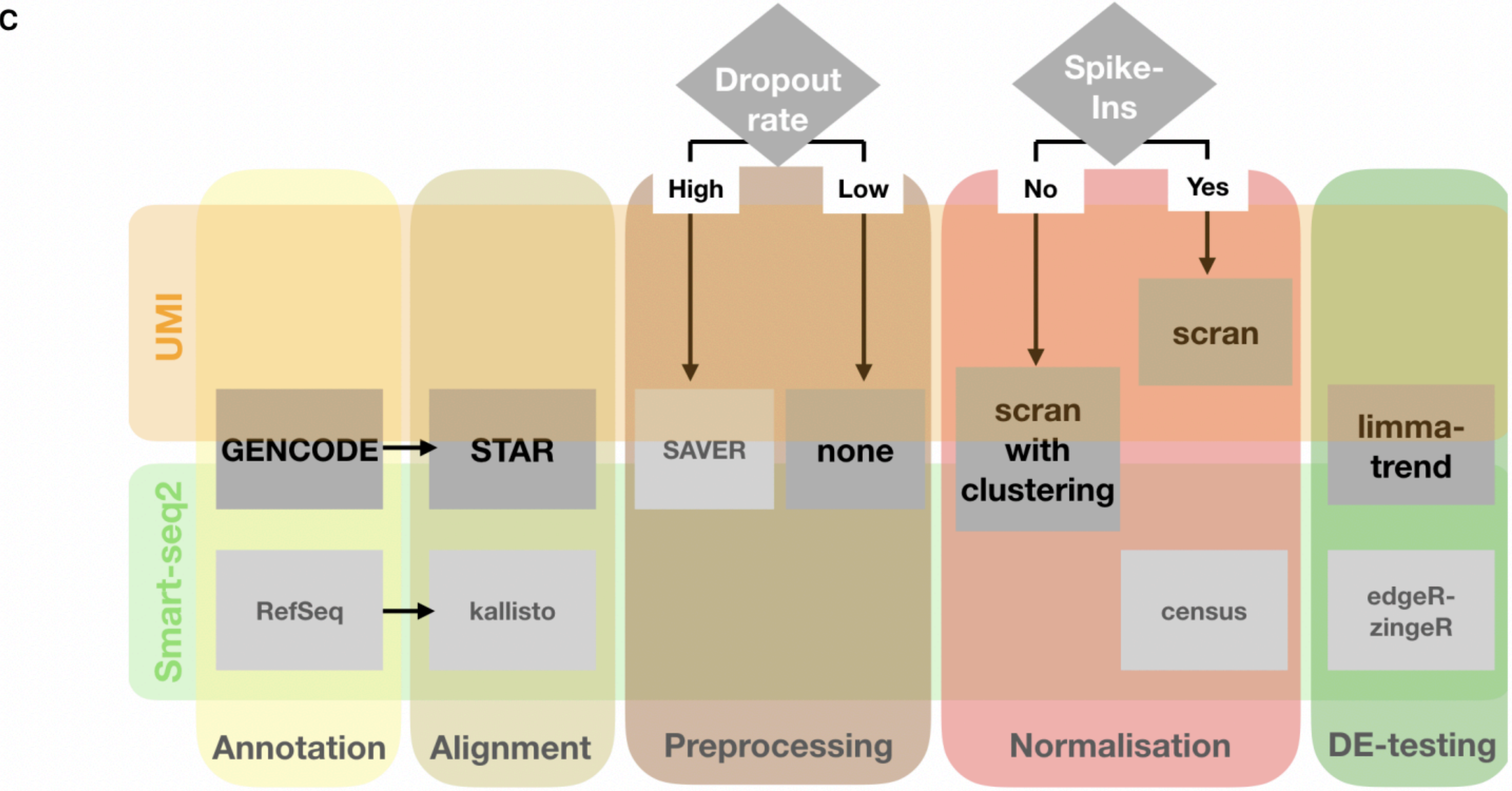
Normalization is important!

Contribution of each step on differential expression (DE) performance



Taken from Vieth et al. (2019)

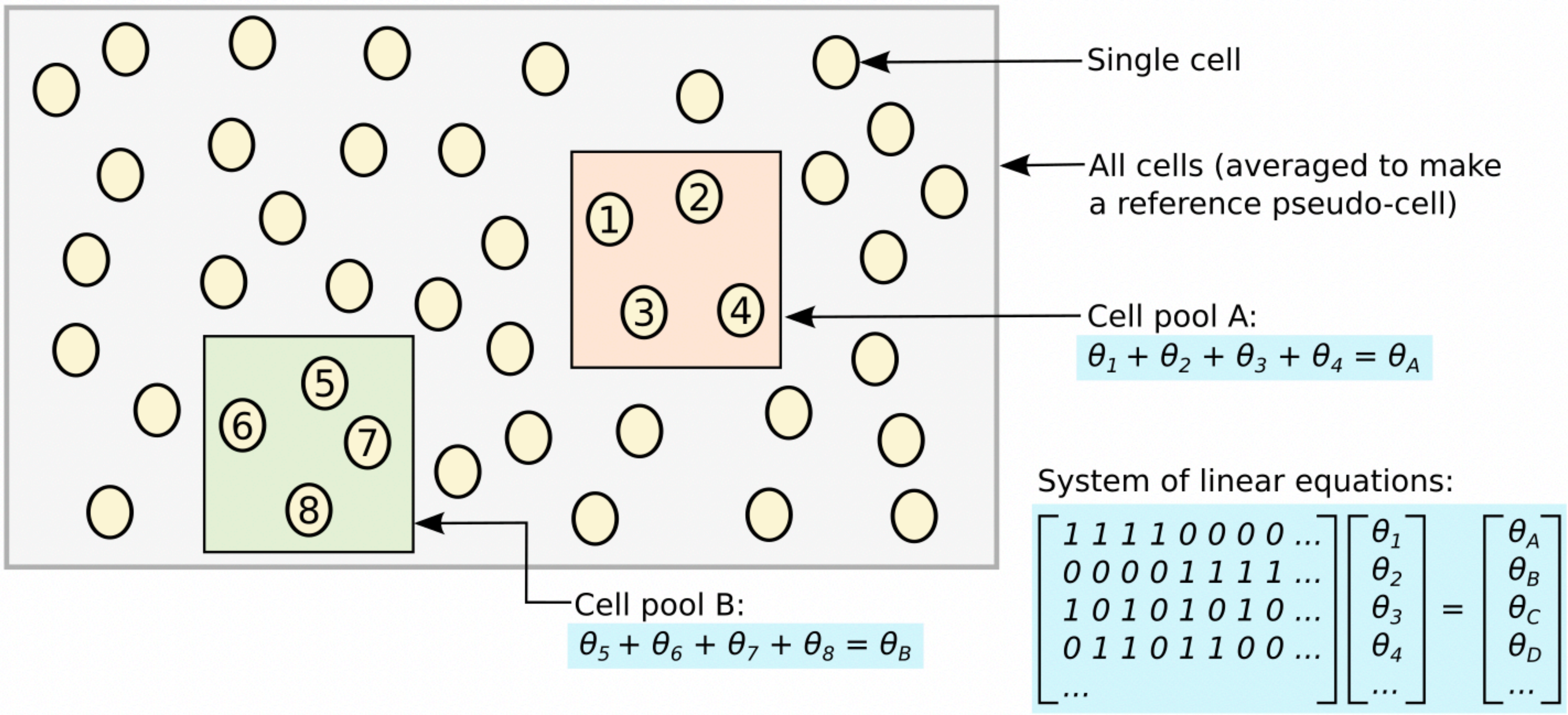
Single-cell RNA-seq Analyses Benchmarked



Taken from Vieth et al. (2019).

Normalization with SCRAN

- Cluster cells into cell pools by similarity first.
- Perform normalization within each cluster/cell pool.


1
2
3
4
5
6
7
8
Single cell
All cells (averaged to make a reference pseudo-cell)
Cell pool A:
Cell pool B:
System of linear equations:

$$\theta_1 + \theta_2 + \theta_3 + \theta_4 = \theta_A$$

$$\theta_5 + \theta_6 + \theta_7 + \theta_8 = \theta_B$$

$$\begin{bmatrix} 1\,1\,1\,1\,0\,0\,0\,0\ldots \\ 0\,0\,0\,0\,1\,1\,1\,1\ldots \\ 1\,0\,1\,0\,1\,0\,1\,0\ldots \\ 0\,1\,1\,0\,1\,1\,0\,0\ldots \\ \ldots \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \ldots \end{bmatrix} = \begin{bmatrix} \theta_A \\ \theta_B \\ \theta_C \\ \theta_D \\ \ldots \end{bmatrix}$$

Fig. 3 Schematic of the deconvolution method. All cells in the data set are averaged to make a reference pseudo-cell. Expression values for cells in pool A are summed together and normalized against the reference to yield a pool-based size factor θ_A. This is equal to the sum of the cell-based factors θ_j for cells $j = 1$–4 and can be used to formulate a linear equation. (For simplicity, the t_j term is assumed to be unity here.) Repeating this for multiple pools (e.g., pool B) leads to the construction of a linear system that can be solved to estimate θ_j for each cell j

Dimensionality Reduction

- Why?
  - Reduce the number of dimensions in a high dimensional data for visualization.
  - To prepare the dataset for subsequent clustering.

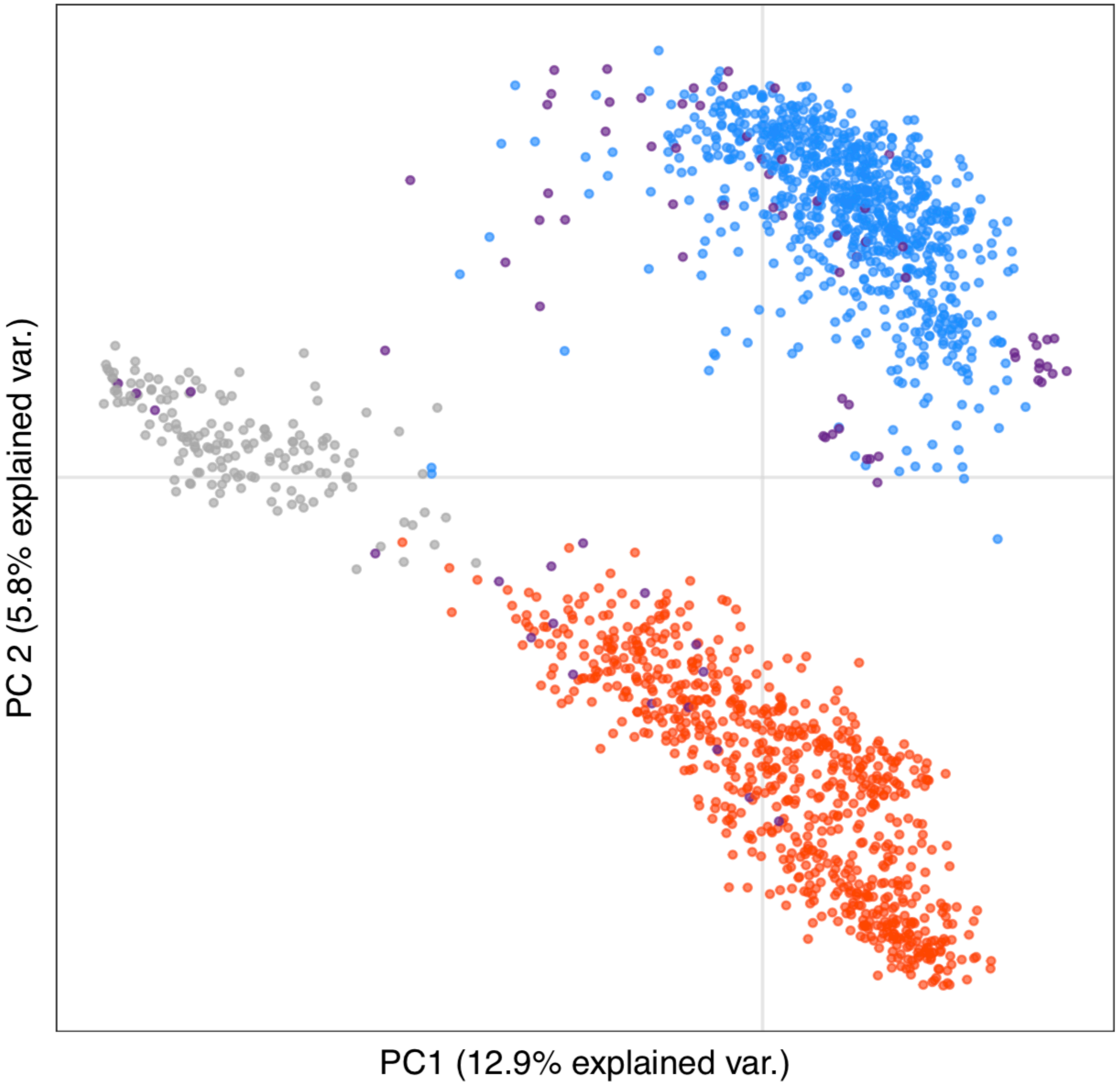


Image generated by Dennis Wylie

Dimensionality Reduction

PCA

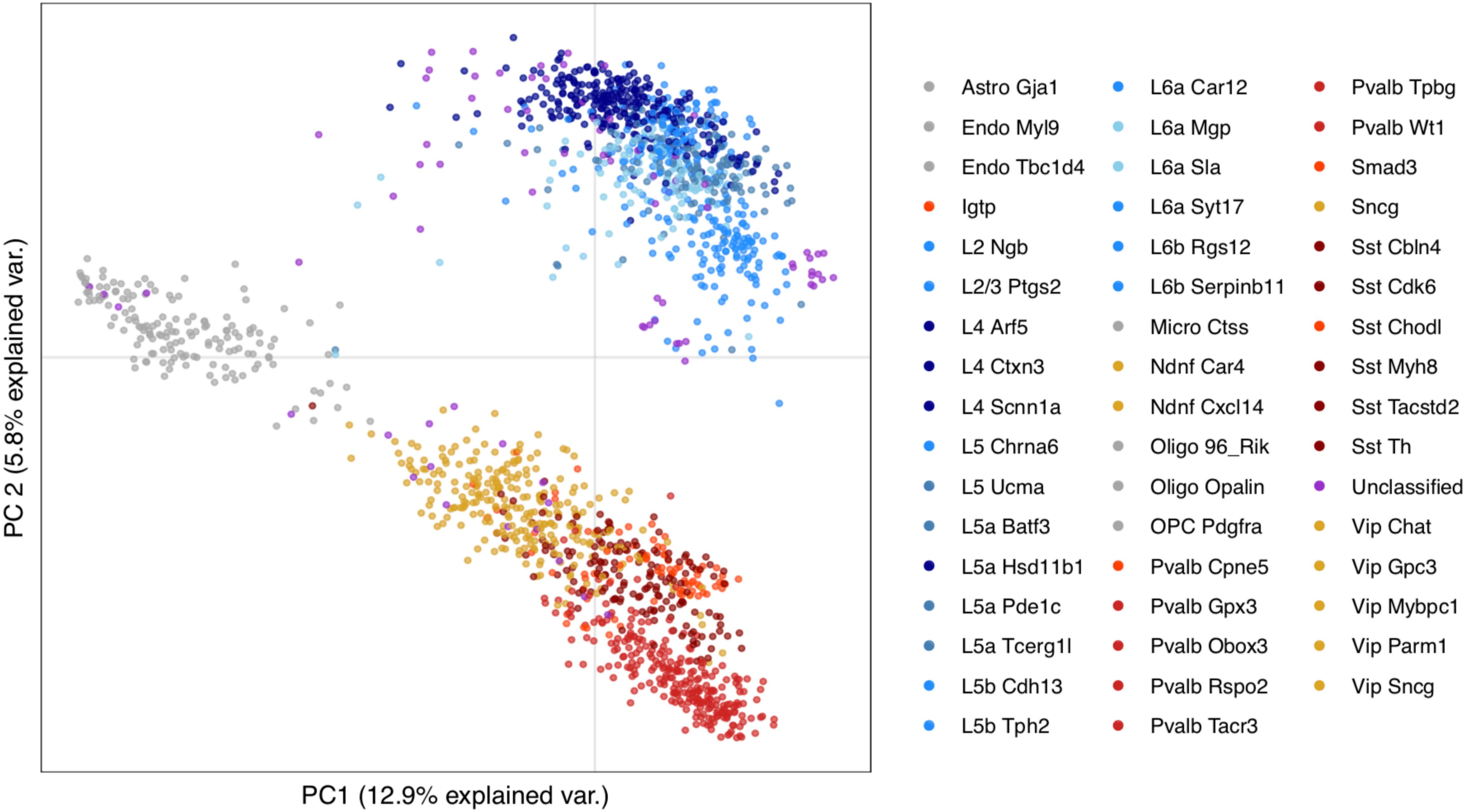
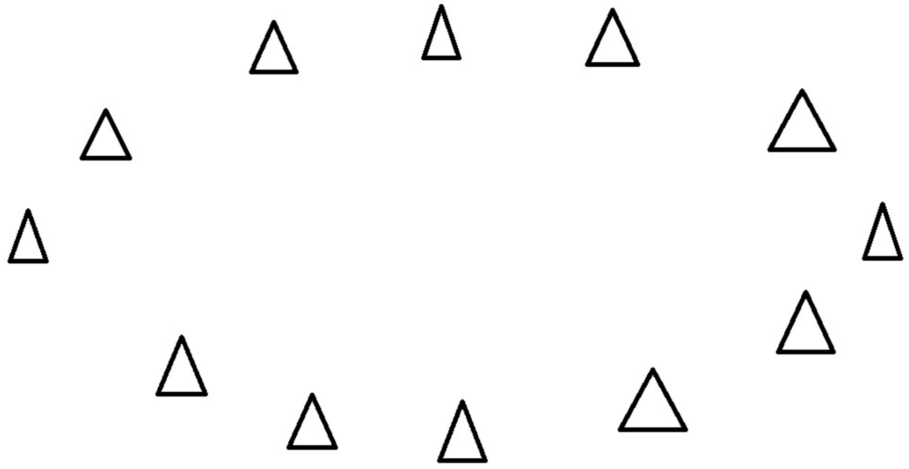
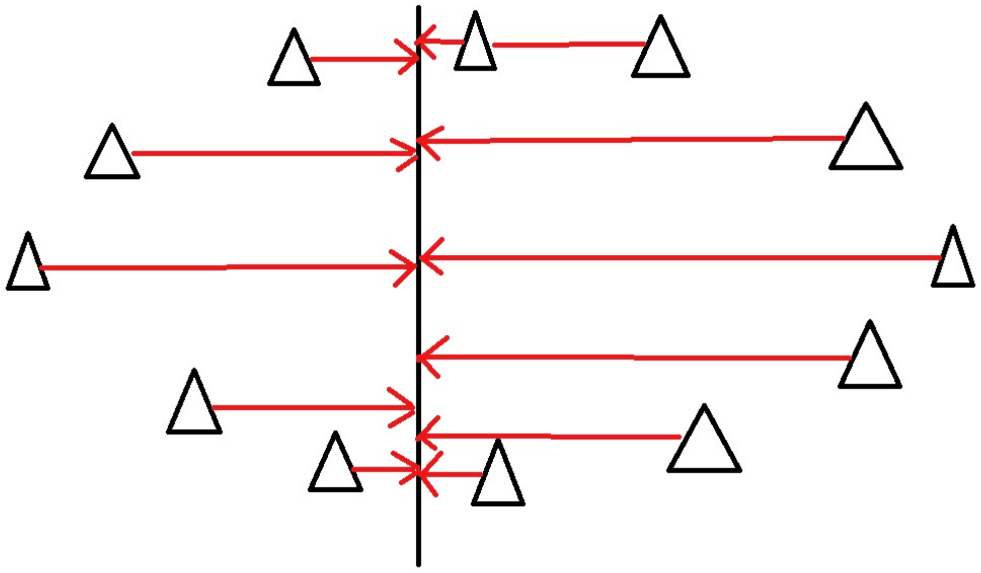


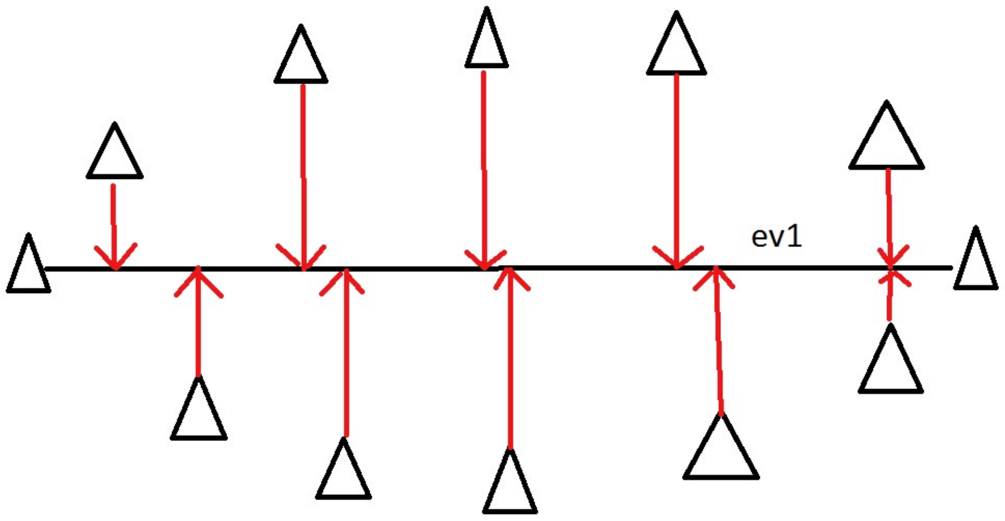
Image generated by Dennis Wylie

Principal Component Analysis



- What are the Principal components of this data?
  - Directions where there is most variance
  - When data is projected onto a straight line, the data is most spread out.





https://georgemdallas.wordpress.com/

Dimensionality Reduction

tSNE (t-Distributed Stochastic Neighbor Embedding)

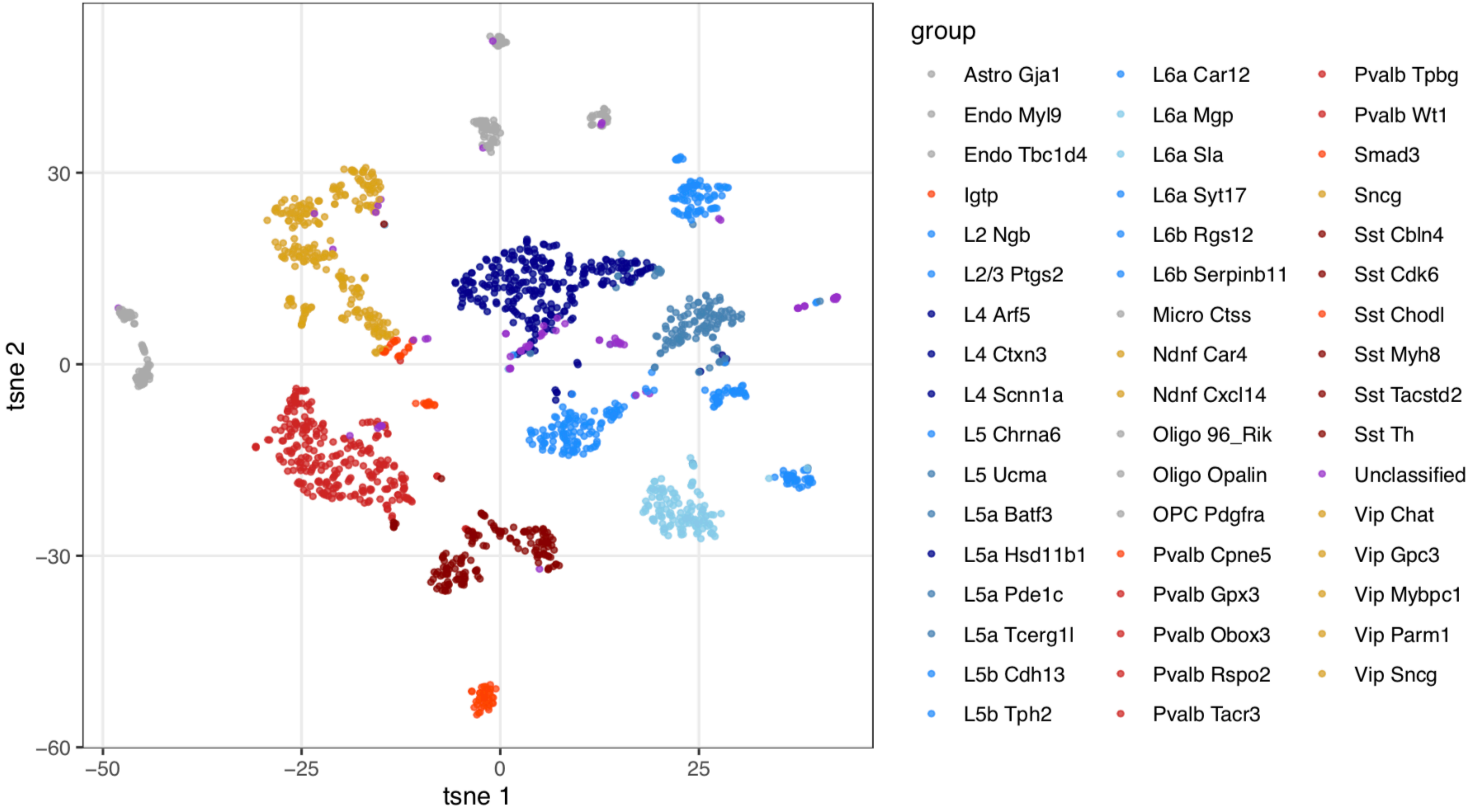
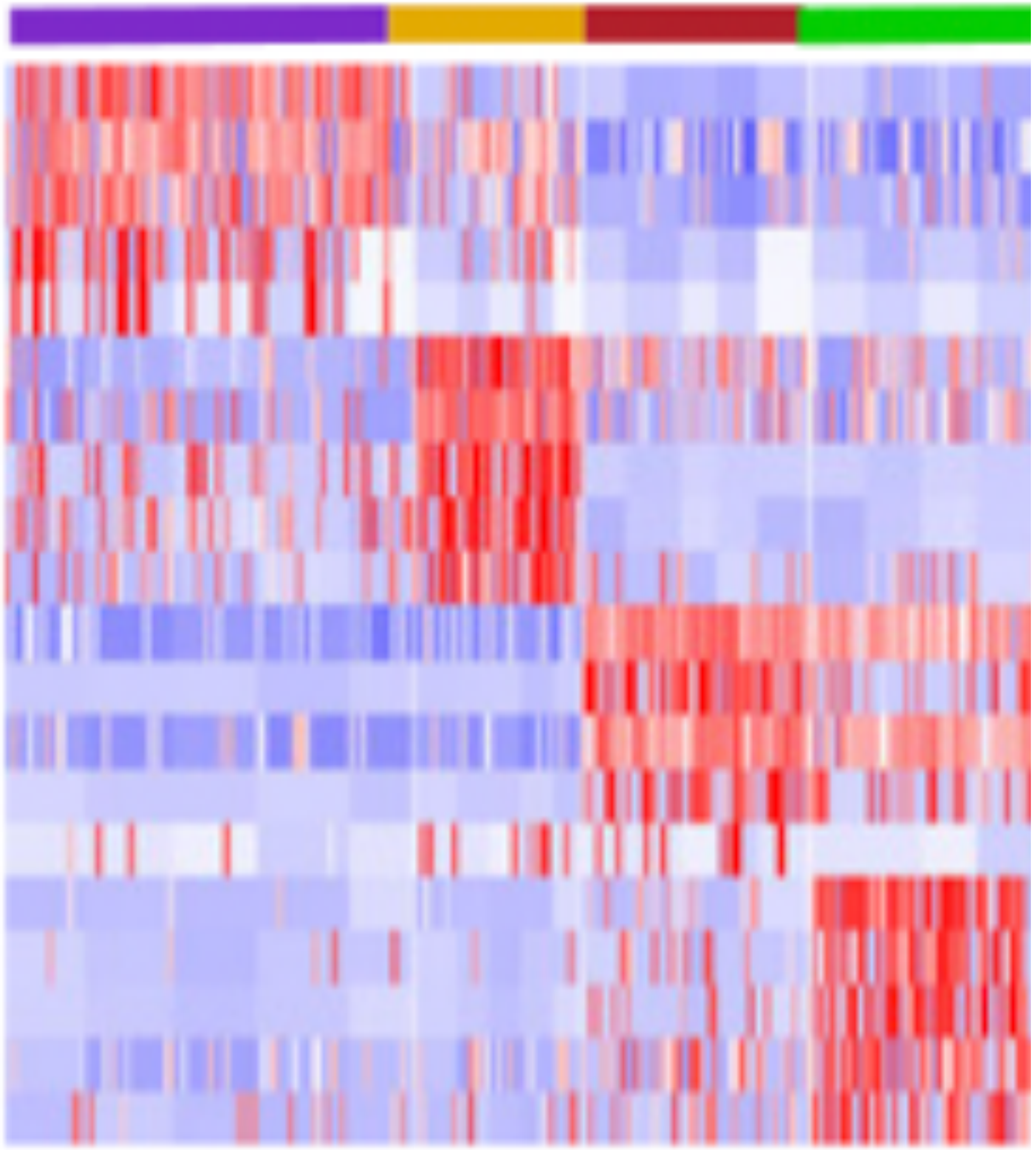


Image generated by Dennis Wylie

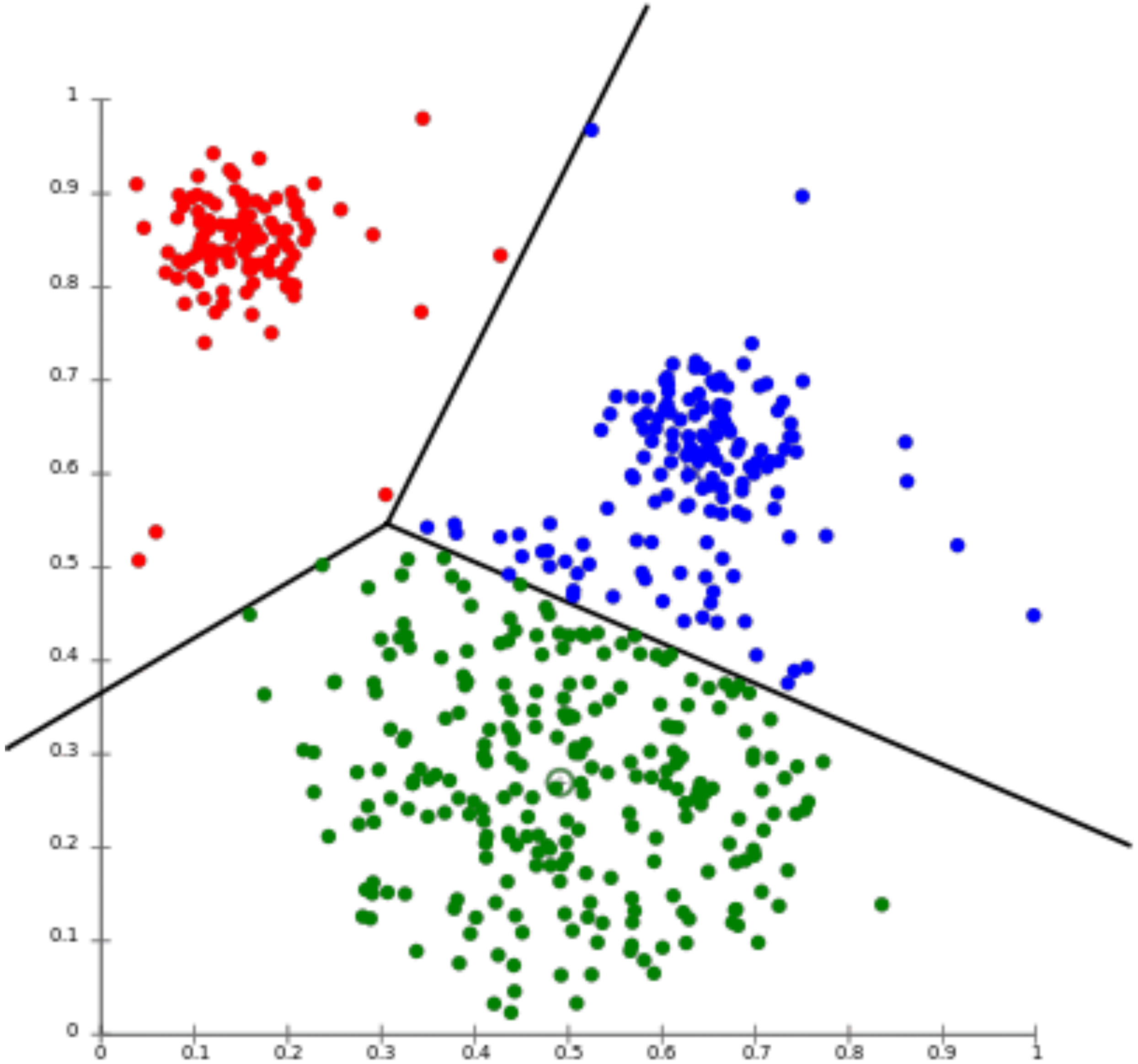
Clustering

Clustering to identify cell subpopulations

Heirarchical clustering



K-means clustering



Differential Expression Analysis

Important to distinguish:

1. DE between predefined cell populations across different samples.
2. DE between clustering-defined subpopulations in 1 sample
   - applying standard statistical tests to clusters learned from same data set will result in very biased p values!
   - Use fold changes(effect sizes) to identify driver genes for each cluster compared to every other cluster.

Differential Expression Analysis

- Bulk RNA-Seq DE analyses methods:
    - DESeq2
    - edgeR

- Specialized scRNA-Seq DE analysis methods
    - Single Cell Differential Expression (SCDE)
    - Model-based Analysis of Single-cell Transcriptomics (MAST)

- Soneson & Robinson (2018) evaluated 36 DE approaches:
    - “bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq.”

Conclusions

- Lots of tools available for scRNA-Seq. Tools are actively being developed, updated and benchmarked.

- **Mapping** using bulk RNA-Seq methods is just fine.

- When **quantifying** genes, UMIs should be taken into account and error corrected.

- **Normalization** is one of the most important steps in scRNA-seq data analysis and needs to be treated differently from bulk RNA-Seq datasets. Scran normalization works very well.

- **Clustering** using clustering methods to identify group for doing DE analysis.

- **DE analysis** using standard bulk RNA-Seq methods works fine.