

Introduction to Single-Cell RNA-seq

Thanks to Dennis Wylie for some slides

Why single-cell RNA-seq?

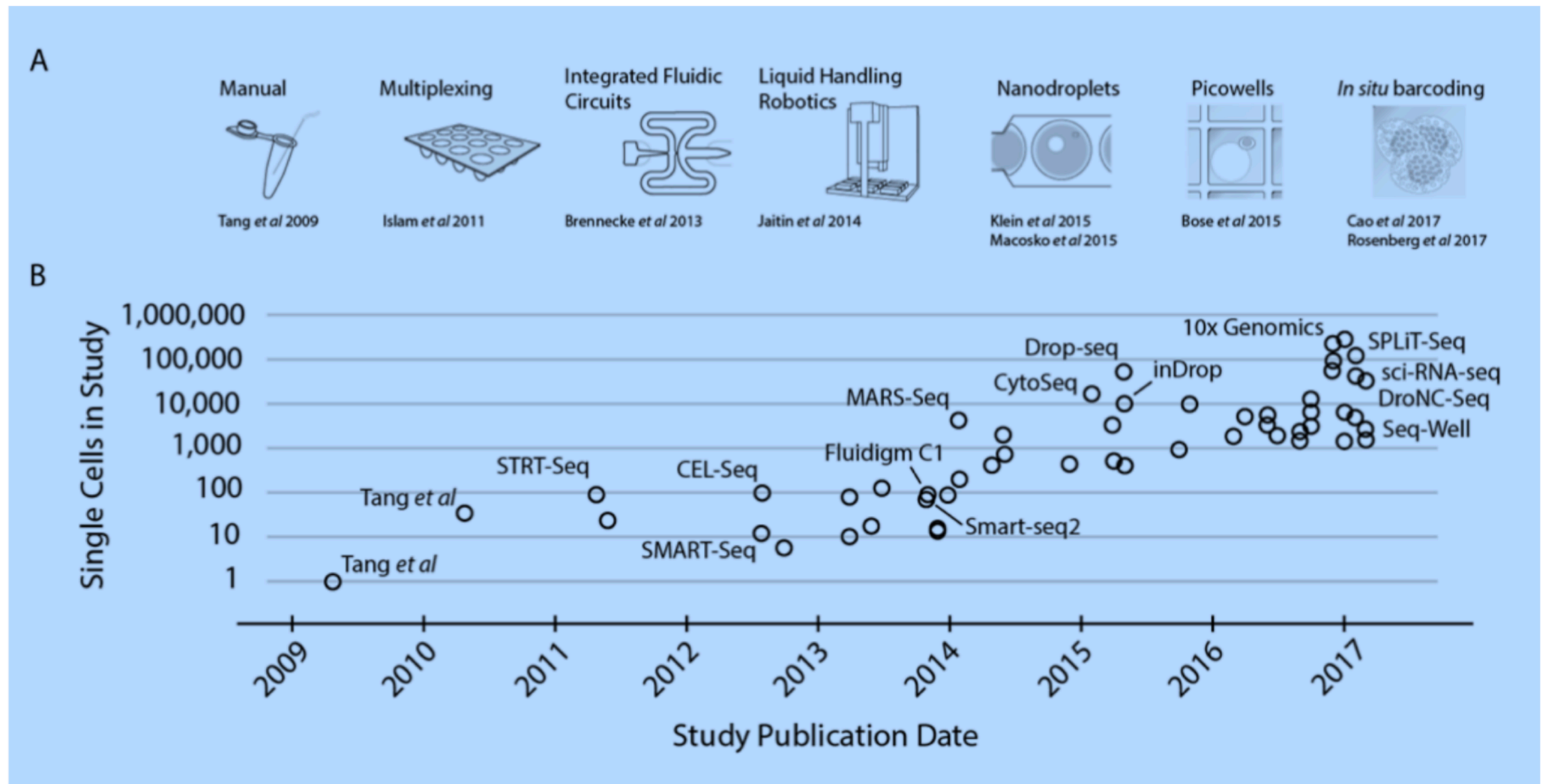
- Allows profiling of gene expression in individual cells.
 - To look at heterogeneity across cell type subpopulations

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

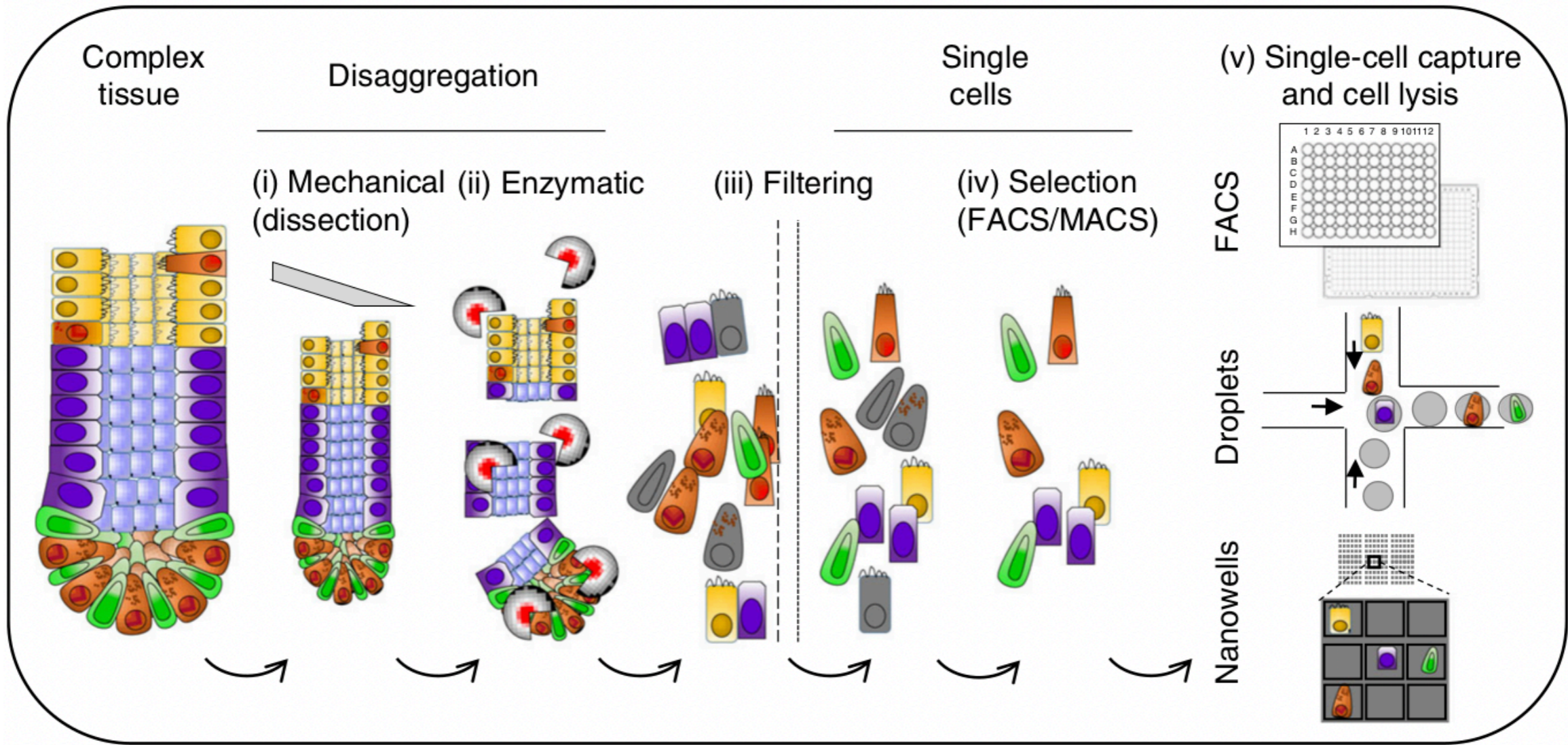
Single-cell RNA-seq Technology Improvements

<https://arxiv.org/abs/1704.01379>



Single-cell RNA-seq Sample Prep

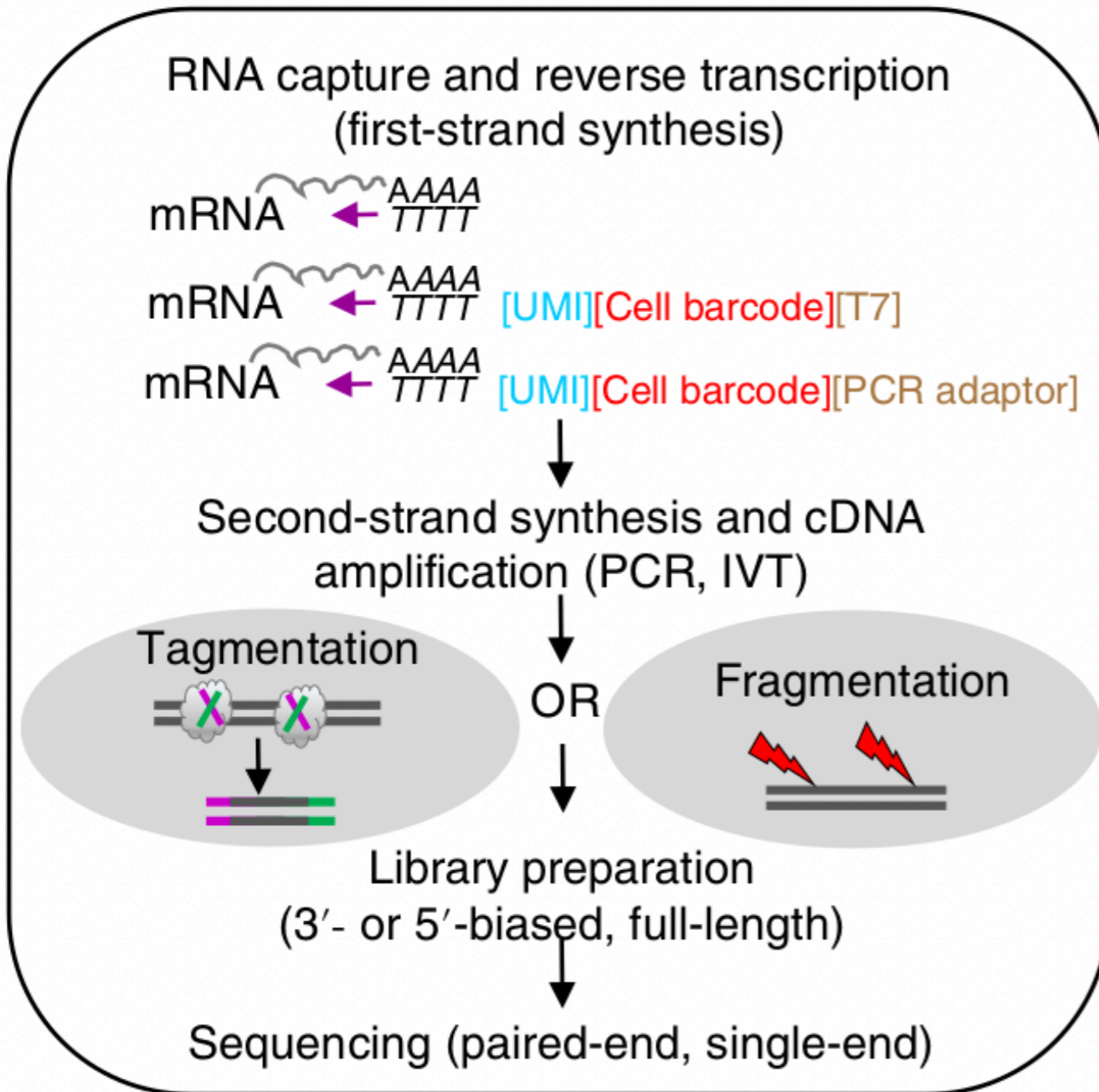
(1) Sample preparation



Taken from Lafzi et al. (2018)

Single-cell RNA-Seq Library Prep

(2) Single-cell RNA sequencing



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

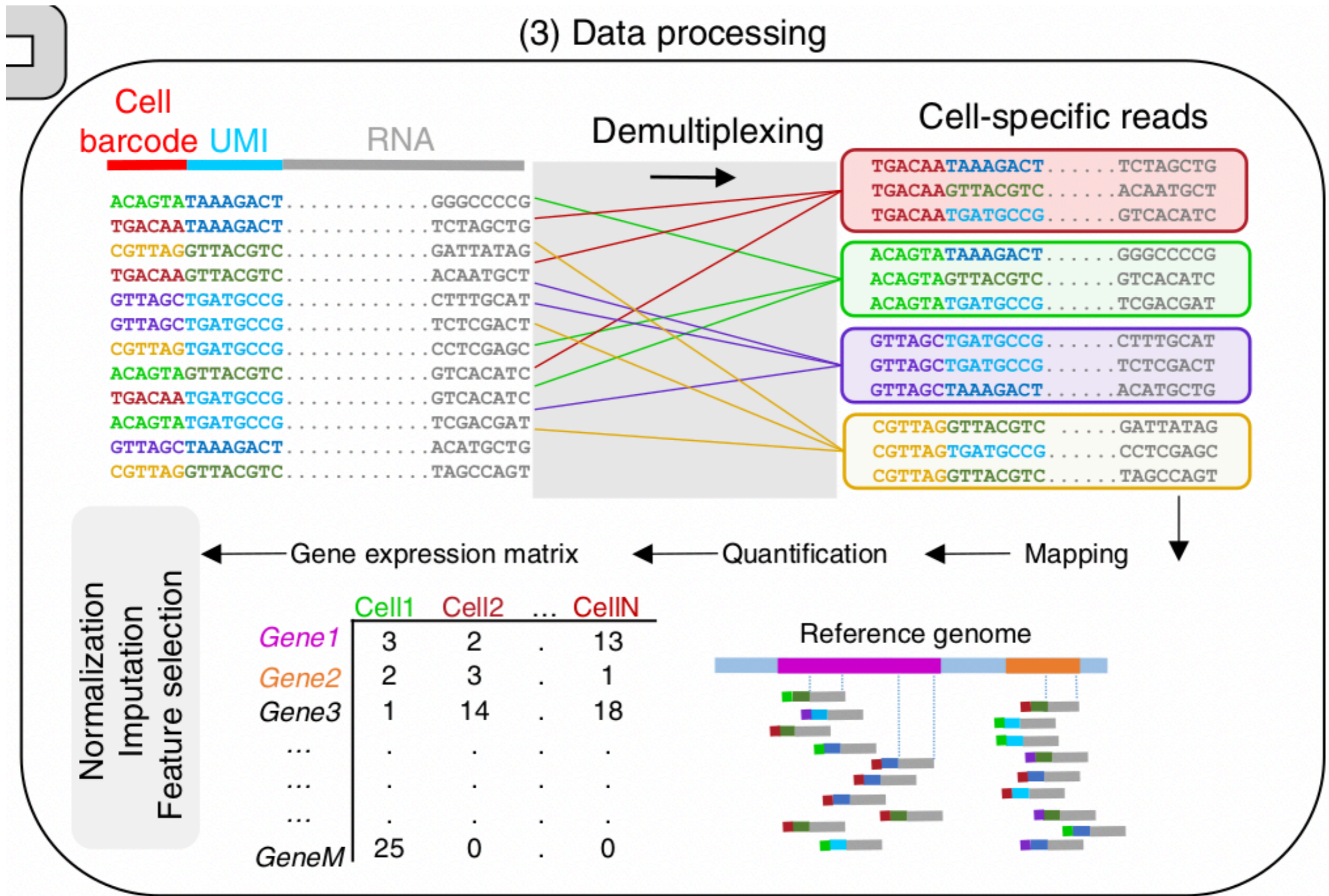
- How many cells per sample?
 - ~10,000 cells per 'typical' sample.
- How many reads per cell?
 - 30,000-50,000 reads per cell for 'typical' samples.
- This would take up >1 lane/sample

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

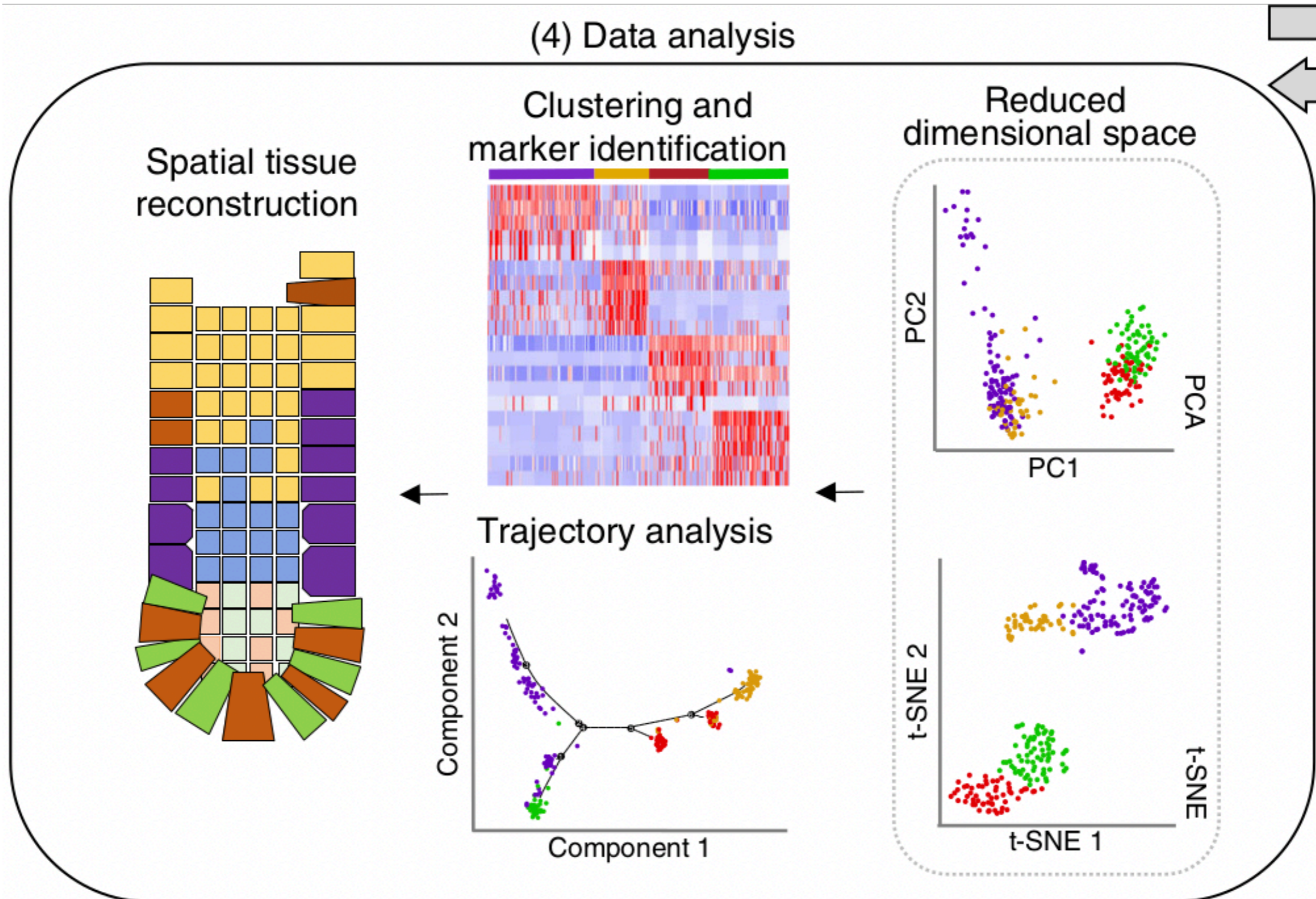
Coverage Recommendations

- Interested in measuring genes expressed at low levels?
 - Increase number of reads per cell.
- Detection of novel transcripts, or quantification of isoforms
 - Need full-length libraries (not 10x)
 - Increase number of reads per cell.
- Interested in finding rare cell types?
 - Increase number of cells
 - Increase number of reads per cell.

Single-cell RNA-Seq Data Analysis



Single-cell RNA-Seq Data Analysis



Single-cell RNA-seq Analysis Steps

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

- **QC Assessment** (FastQC)
- **Alignment to reference** (STAR within Cell Ranger)
- **Quantification** (within Cell Ranger, UMI-tools)

Single-cell RNA-seq Analysis Steps

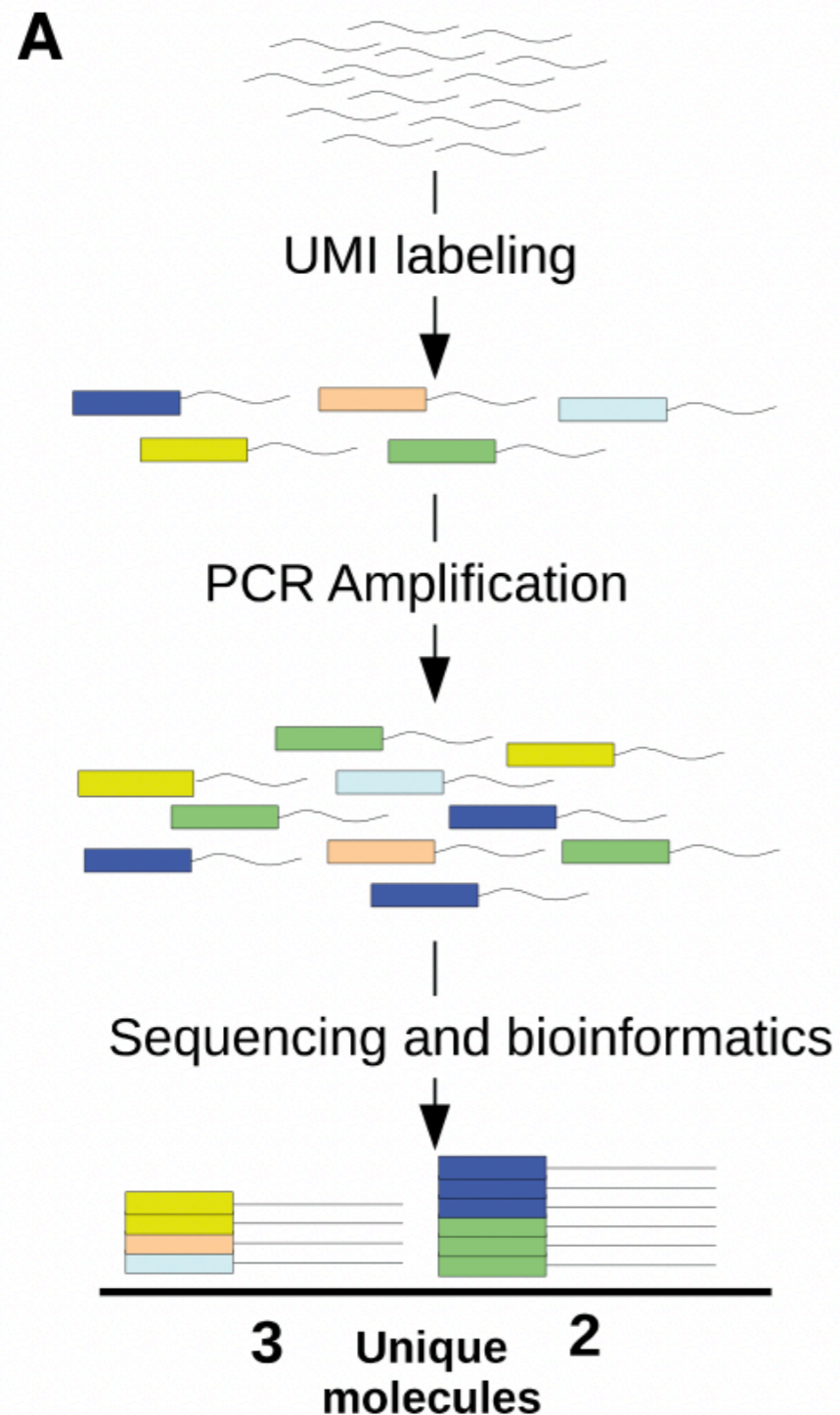
- Imputation (OPTIONAL) (SAVER)
- **Normalization** (scran or SCTransform)
- **Dimensionality reduction** (tSNE, UMAP, PCA, etc)
- **Clustering** (hierarchical, k-means, Seurat, etc)
- **Differential expression analysis** (DESeq2, edgeR, limma, MAST)

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.

Single-cell RNA-Seq Data Analysis

UMI demultiplexing and error correction



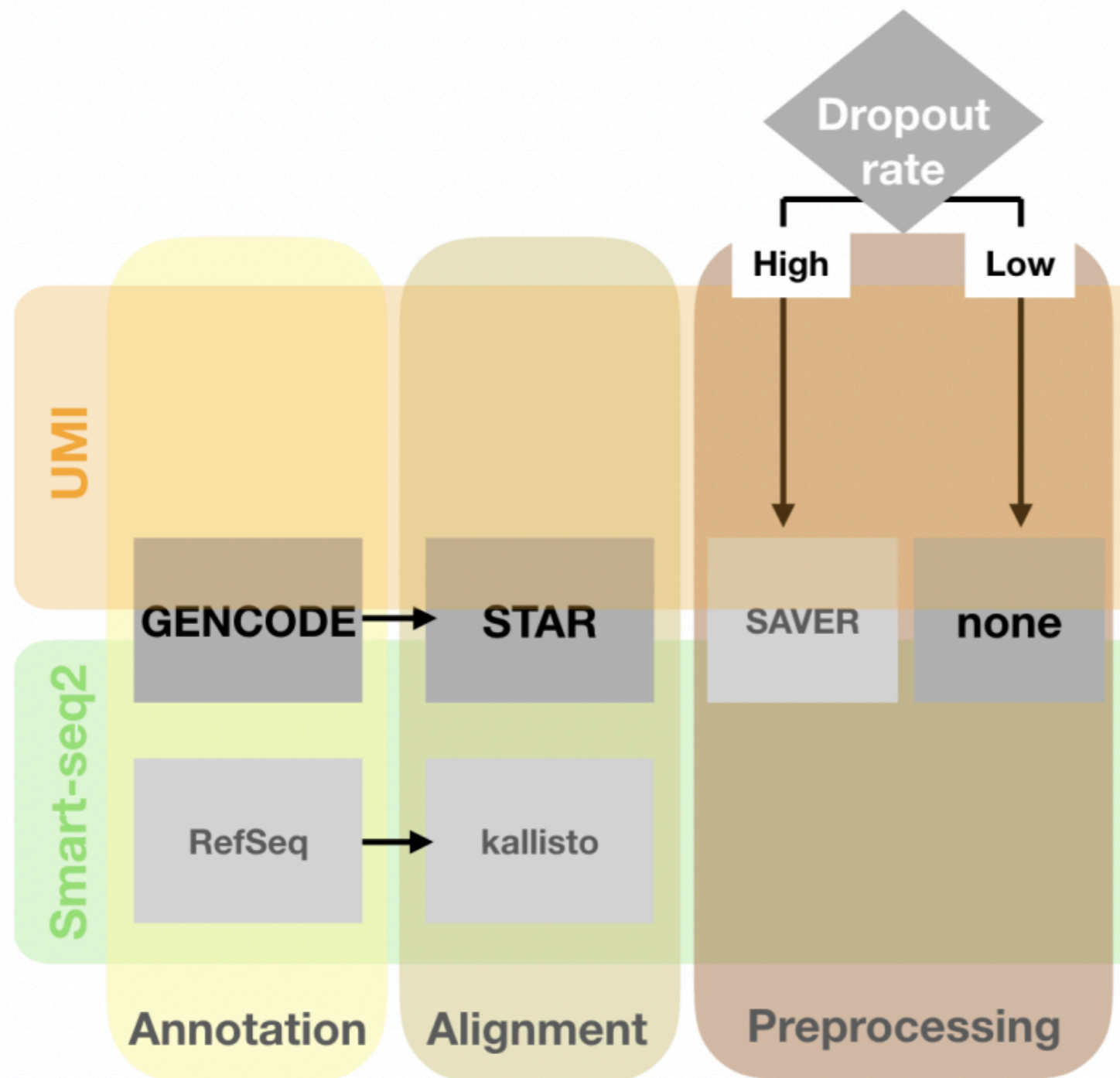
Taken from Smith et al. (2017)

- 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



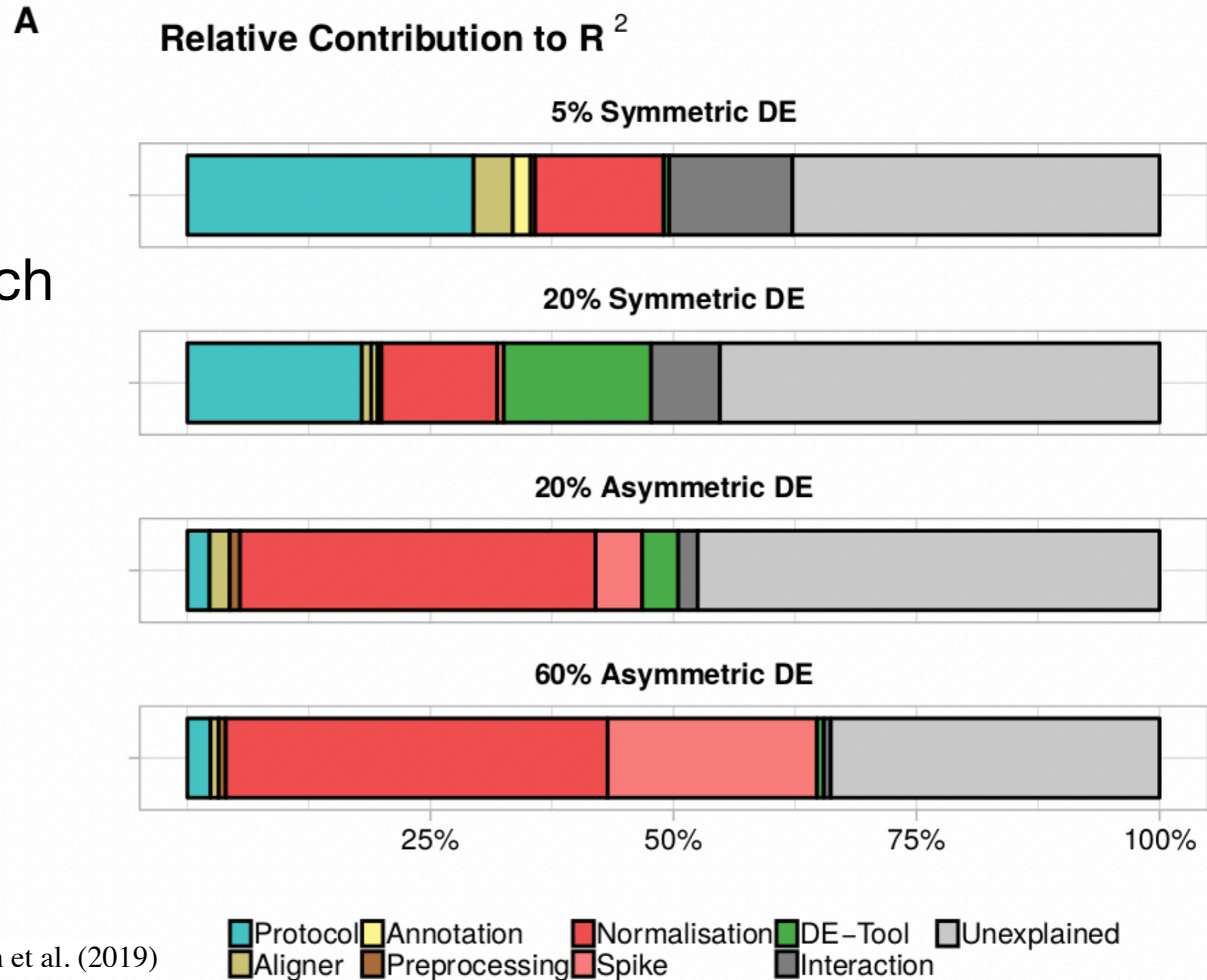
- STAR mapper works well for UMI based/ chromium (10x) scRNA-Seq data
- Imputation only if dropout rate is high

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).

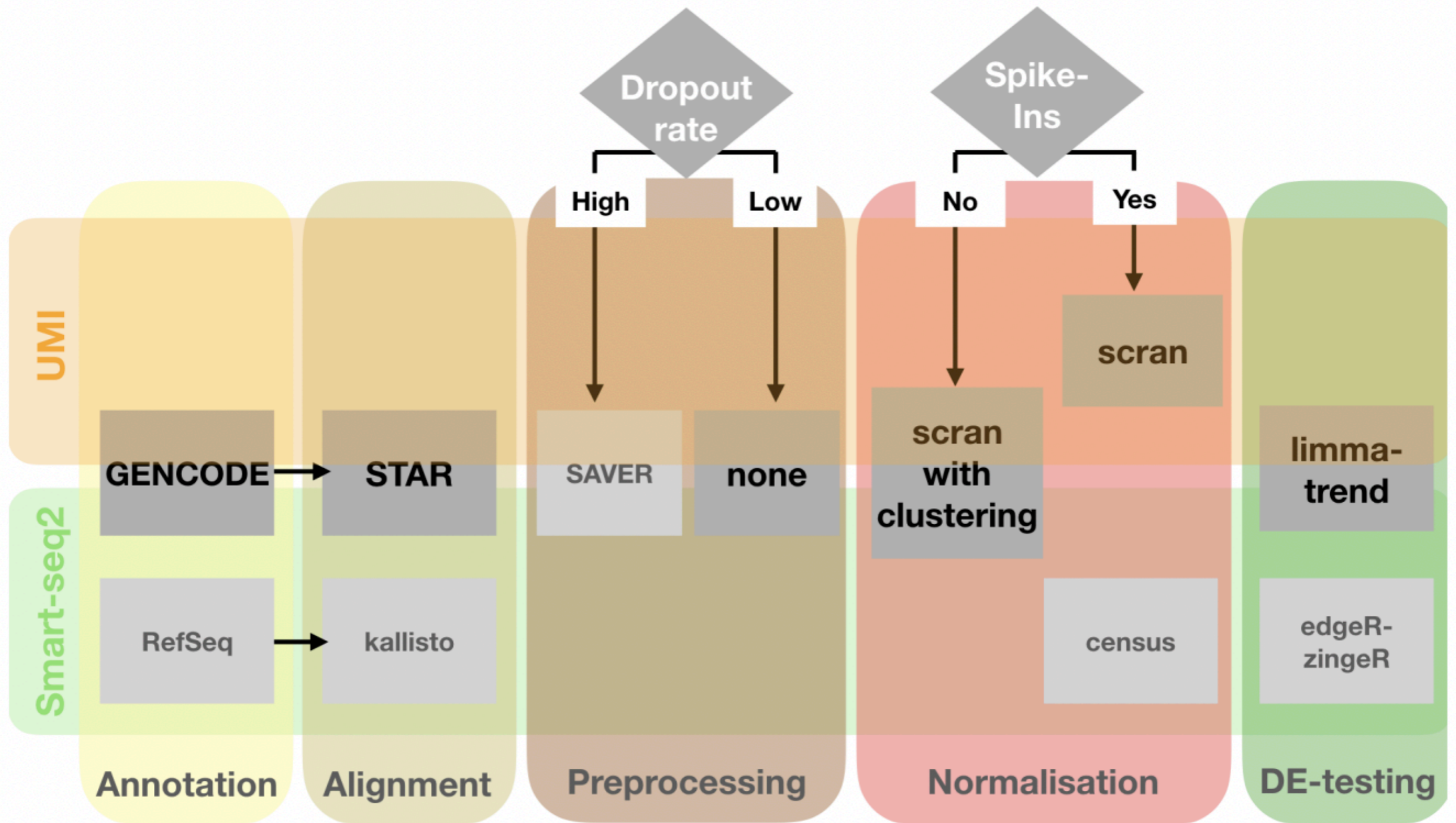
Normalization is important!

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.

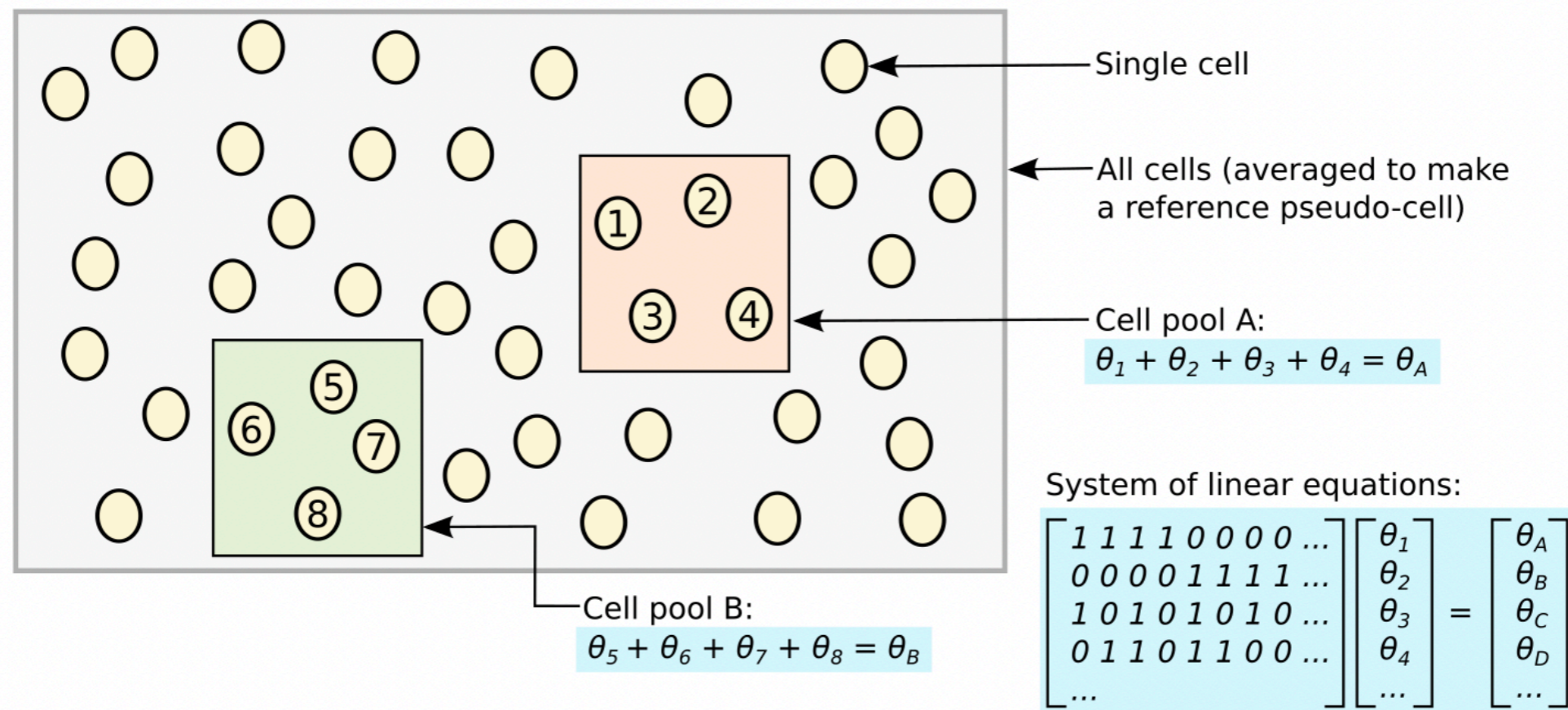


Fig. 3 Schematic of the deconvolution method. All cells in the data set are averaged to make a reference pseudo-cell. Expression values for cells in pool A are summed together and normalized against the reference to yield a pool-based size factor θ_A . This is equal to the sum of the cell-based factors 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.

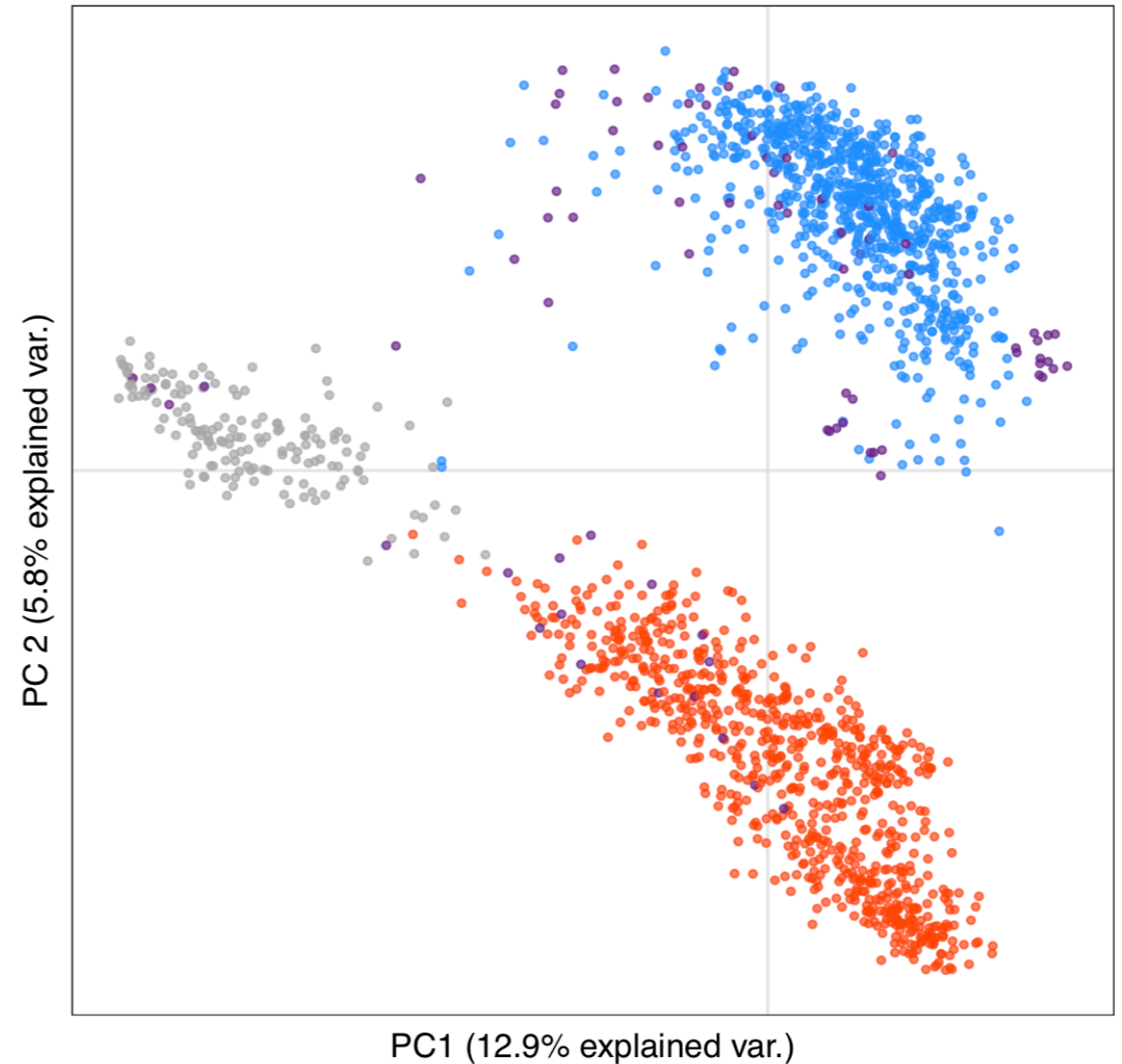
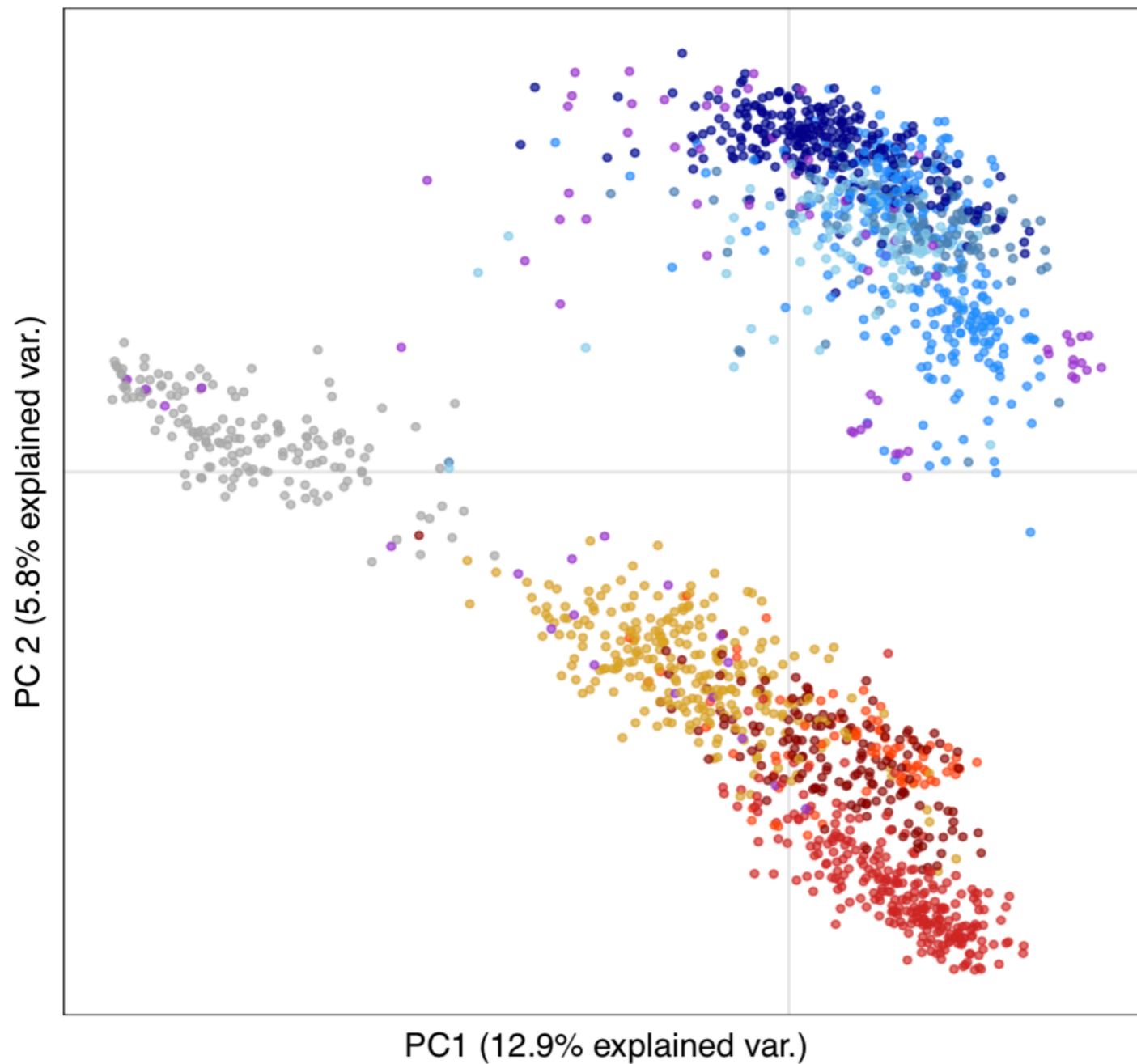


Image generated by Dennis Wylie

Dimensionality Reduction

PCA

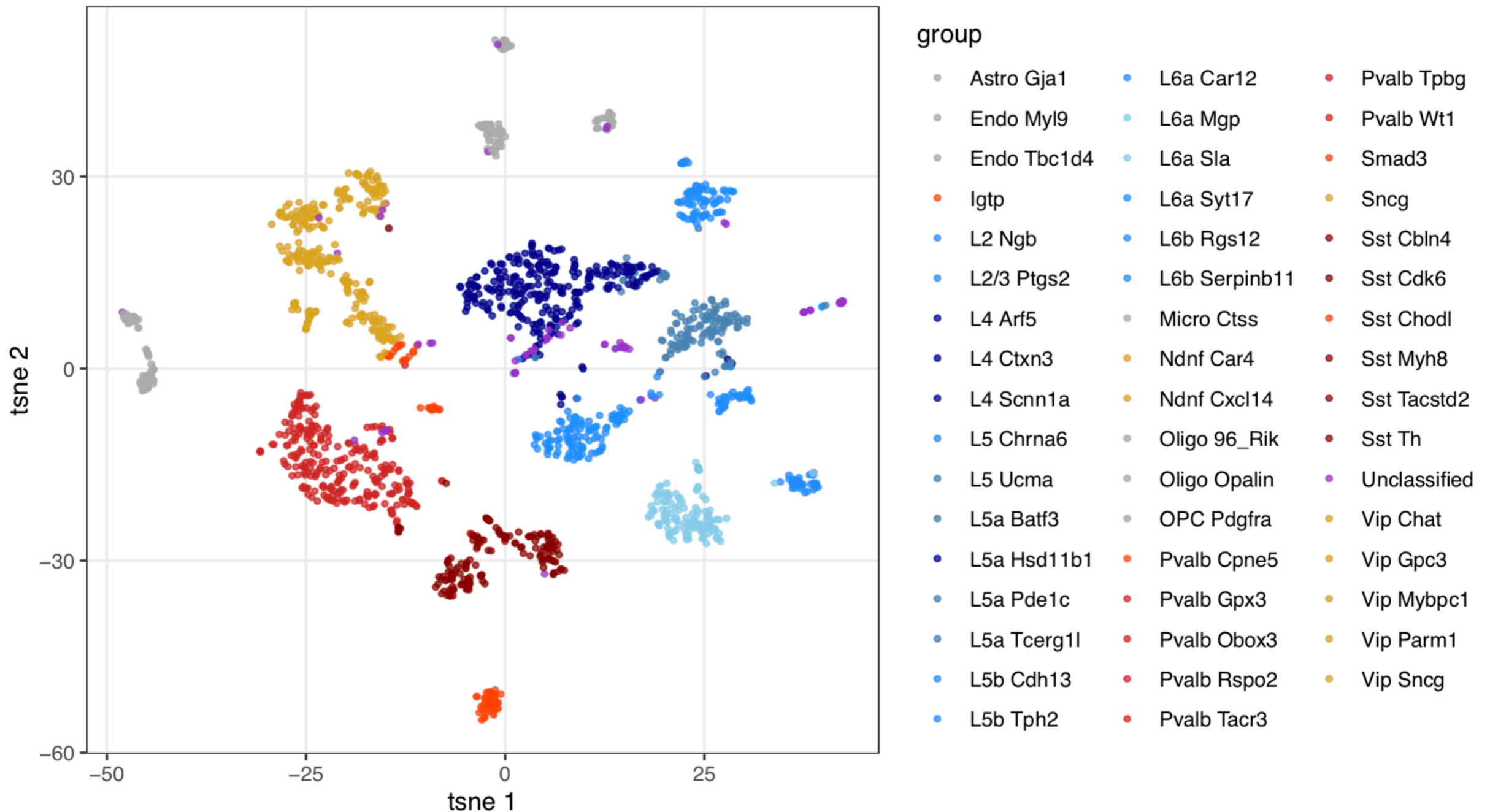


- | | | |
|---------------|-----------------|----------------|
| ● Astro Gja1 | ● L6a Car12 | ● Pvalb Tpbg |
| ● Endo Myl9 | ● L6a Mgp | ● Pvalb Wt1 |
| ● Endo Tbc1d4 | ● L6a Sla | ● Smad3 |
| ● Igtp | ● L6a Syt17 | ● Sncg |
| ● L2 Ngb | ● L6b Rgs12 | ● Sst Cbln4 |
| ● L2/3 Ptgs2 | ● L6b Serpinb11 | ● Sst Cdk6 |
| ● L4 Arf5 | ● Micro Ctss | ● Sst Chodl |
| ● L4 Ctxn3 | ● Ndnf Car4 | ● Sst Myh8 |
| ● L4 Scnn1a | ● Ndnf Cxcl14 | ● Sst Tacstd2 |
| ● L5 Chrna6 | ● Oligo 96_Rik | ● Sst Th |
| ● L5 Ucma | ● Oligo Opalin | ● Unclassified |
| ● L5a Batf3 | ● OPC Pdgfra | ● Vip Chat |
| ● L5a Hsd11b1 | ● Pvalb Cpne5 | ● Vip Gpc3 |
| ● L5a Pde1c | ● Pvalb Gpx3 | ● Vip Mybpc1 |
| ● L5a Tcerg1l | ● Pvalb Obox3 | ● Vip Parm1 |
| ● L5b Cdh13 | ● Pvalb Rspo2 | ● Vip Sncg |
| ● L5b Tph2 | ● Pvalb Tacr3 | |

Image generated by Dennis Wylie

Dimensionality Reduction

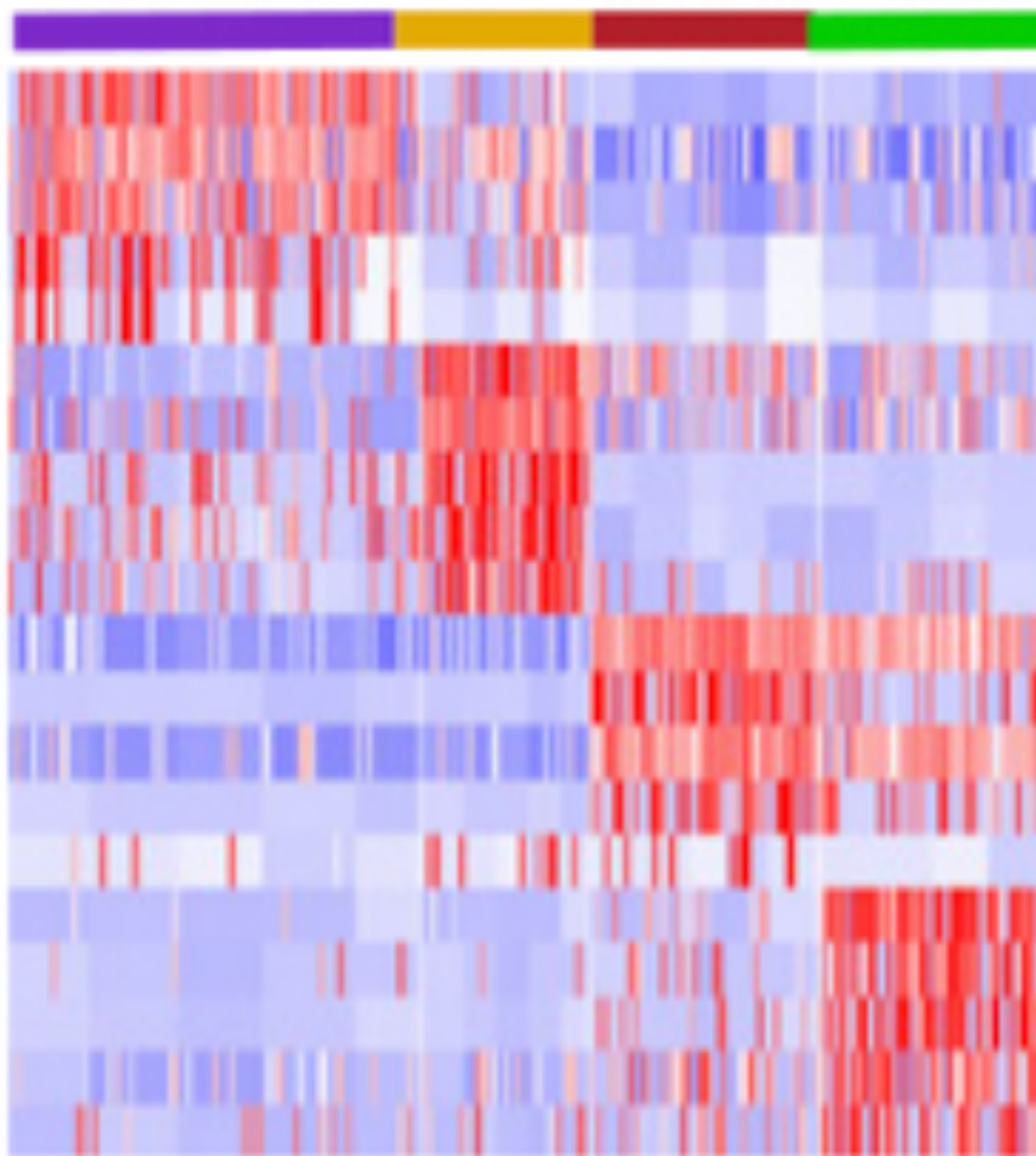
tSNE (t-Distributed Stochastic Neighbor Embedding)



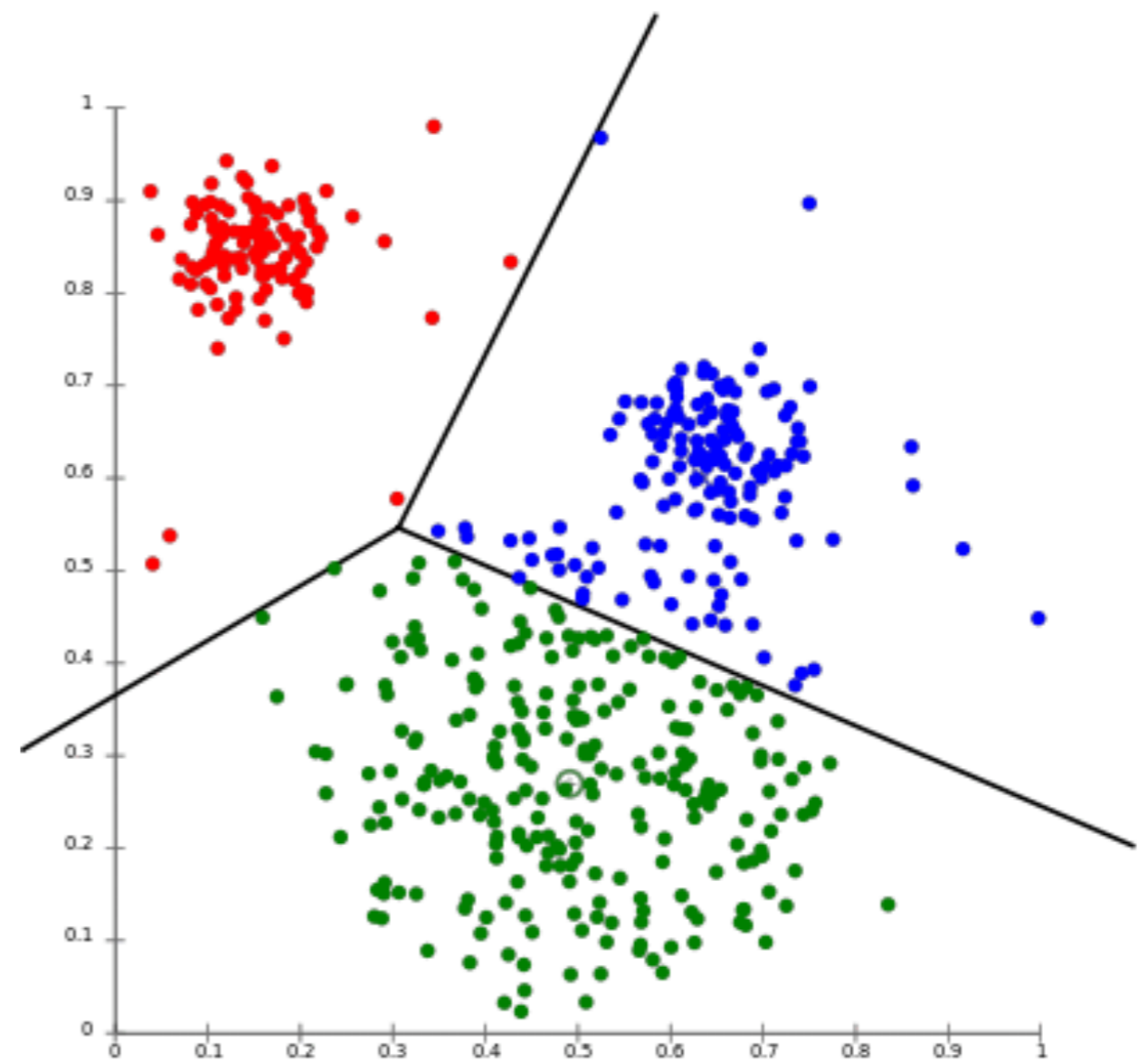
Clustering

Clustering to identify cell subpopulations

Heirarchical clustering



K-means clustering



Differential Expression Analysis

Important to distinguish:

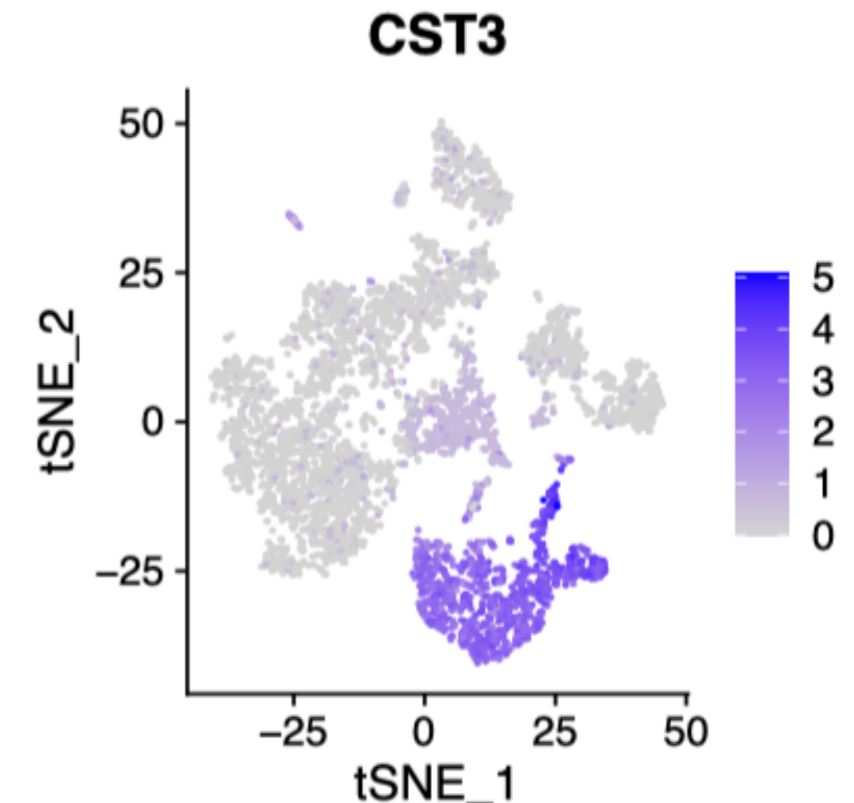
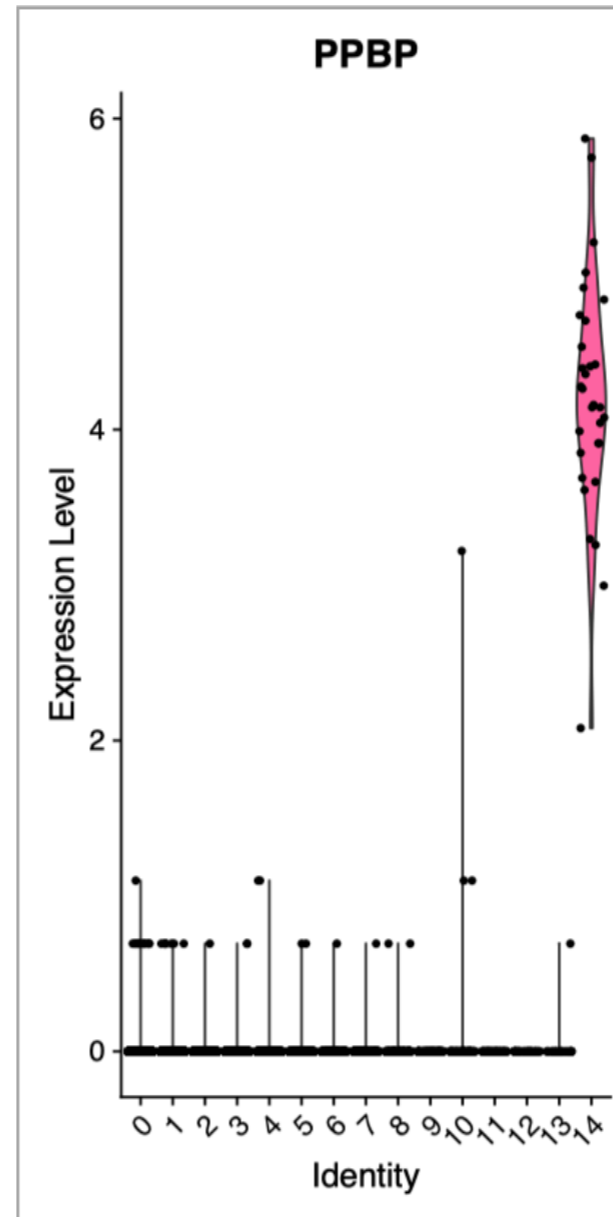
1. DE between predefined cell populations across different samples.
2. DE between clustering-defined subpopulations in 1 sample
 - applying standard statistical tests to clusters learned from same data set will result in very biased p values!
 - Use fold changes(effect sizes) to identify driver genes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

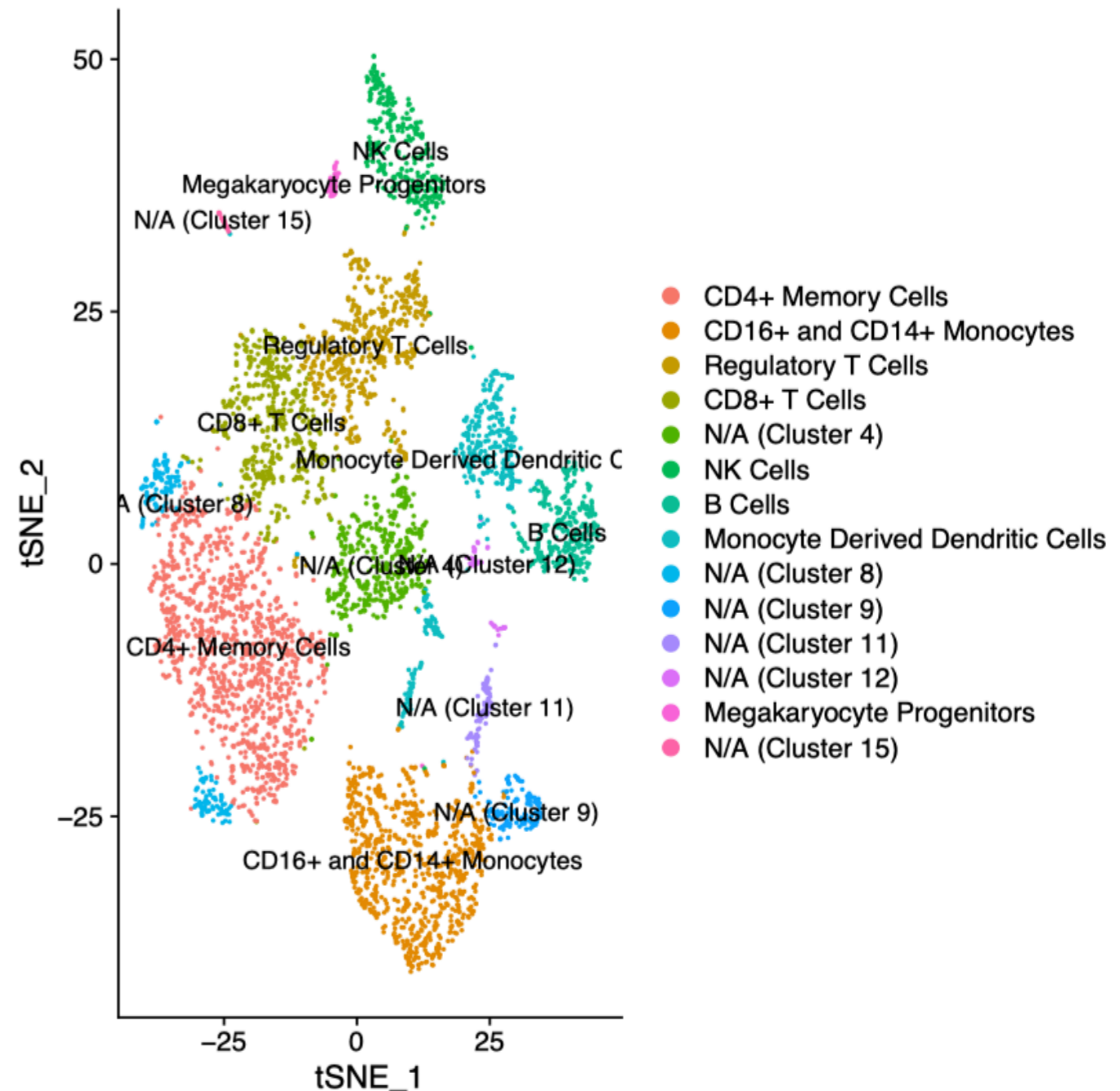
Identifying Marker Genes

- Differential expression analysis across clusters can identify top marker genes for each cluster
 - DEGs for each cluster
- Visualize using violin plots and tSNEs.



Label Clusters/Identify Cell Types

- Label cell clusters based on known markers.
 - <https://panglaodb.se>
- Use tools that automate this process:
 - [Cibersort](#)
 - [scMatch](#)



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. Scraper normalization works very well. Seurat's new scTransform is a good option as well.
- **Clustering** using clustering methods to identify group for doing DE analysis.
- **DE analysis** using standard bulk RNA-Seq methods works fine.
- **Identifying cell types among clusters** requires some prior knowledge about markers expected for cell types.