Single cell RNA sequencing analysis: from CellRanger outputs to cluster annotation

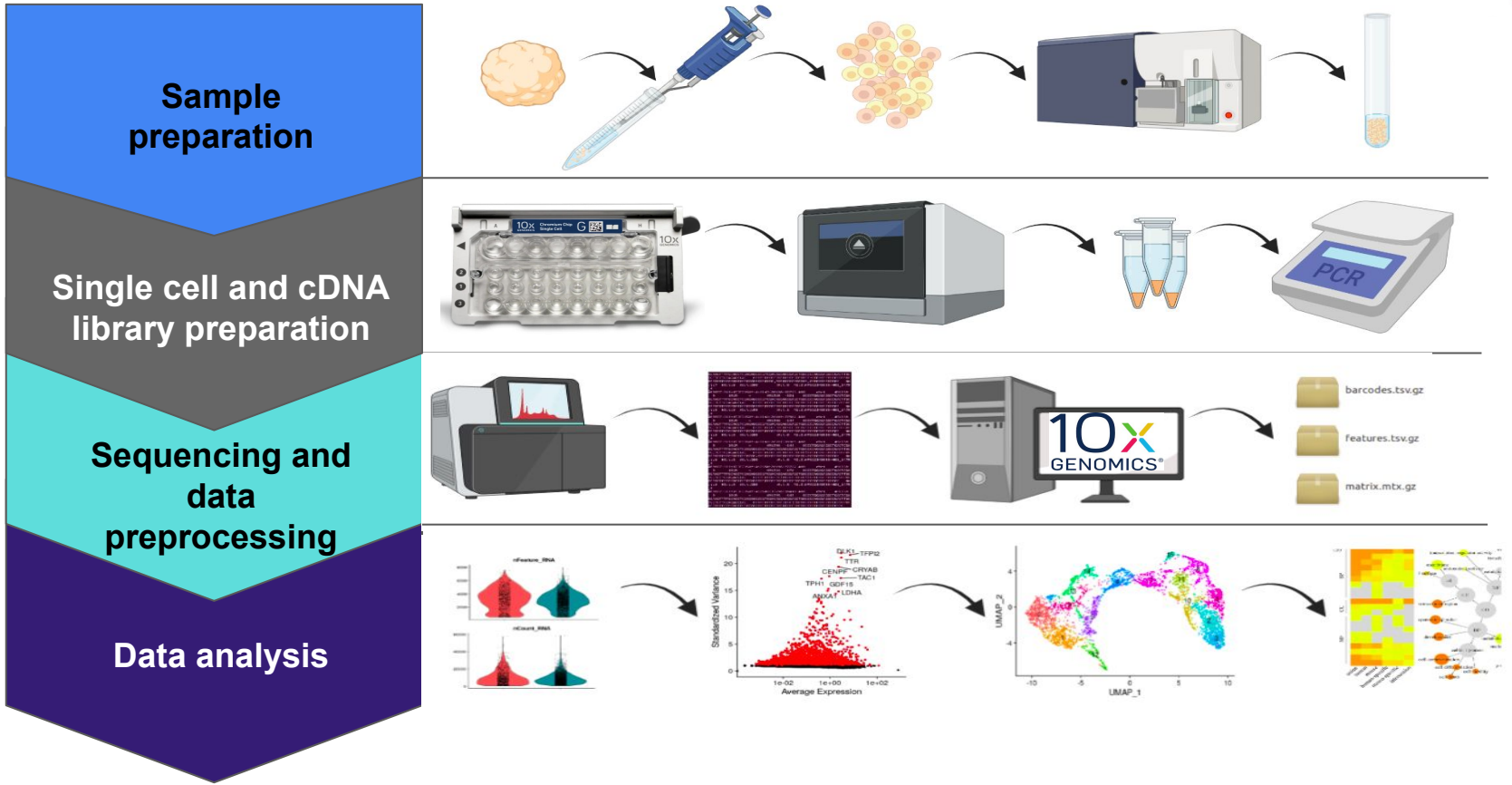
Rhalena Thomas, Moein Yaqubi, Malosree Maitra

Friday June 2nd, 2023

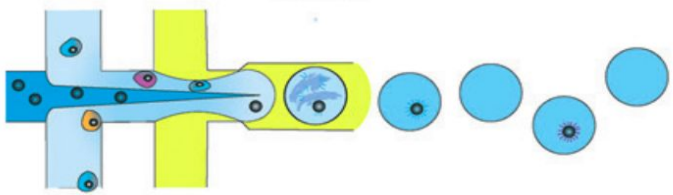
Outline

- Background recap
- Analysis overview
- Alignment & count generation
- Loading data
- Quality Control
- Normalization
- Feature selection
- Dimensionality reduction
- Batch correction
- Clustering
- Annotation
- Comparison to other datasets

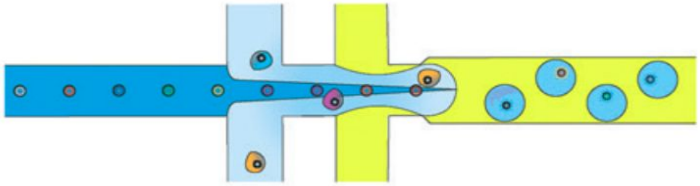
Single cell RNA sequencing (scRNAseq)



scRNAseq: single cell and cDNA library



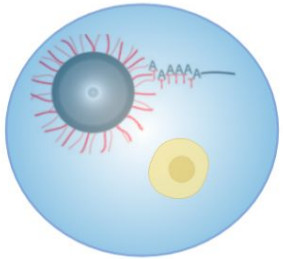
Drop-seq



10X Chromium

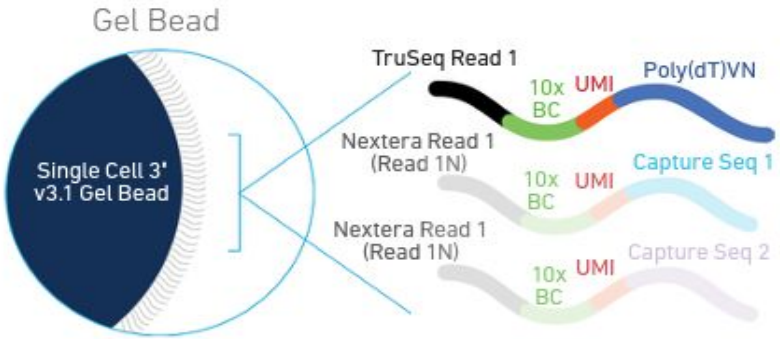
Salomon, Robert, et al. "Droplet-based single cell RNAseq tools: a practical guide." *Lab on a Chip* 19.10 (2019): 1706-1727.

Bead and cell with lysis buffer



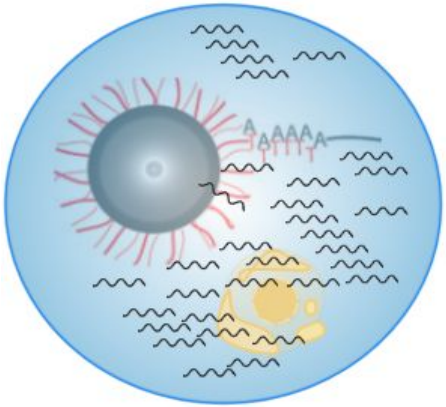
Each bead is covered in primers

- Each primer on one bead contains a bead specific sequence - all the primers have this same sequence
- Each primer contains a different unique modifier sequence (UMI) - each mRNA will have a different sequence attached

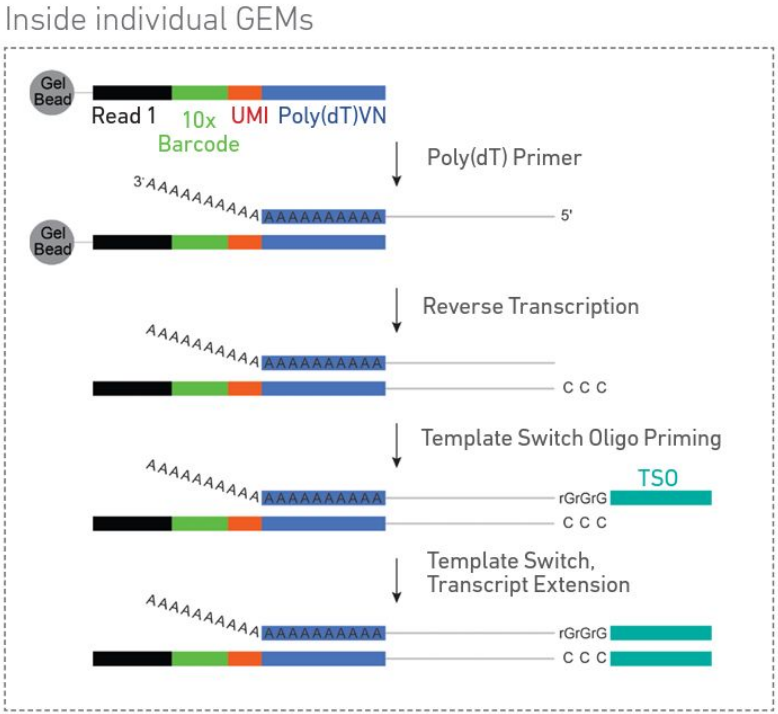


scRNAseq: single cell and cDNA library

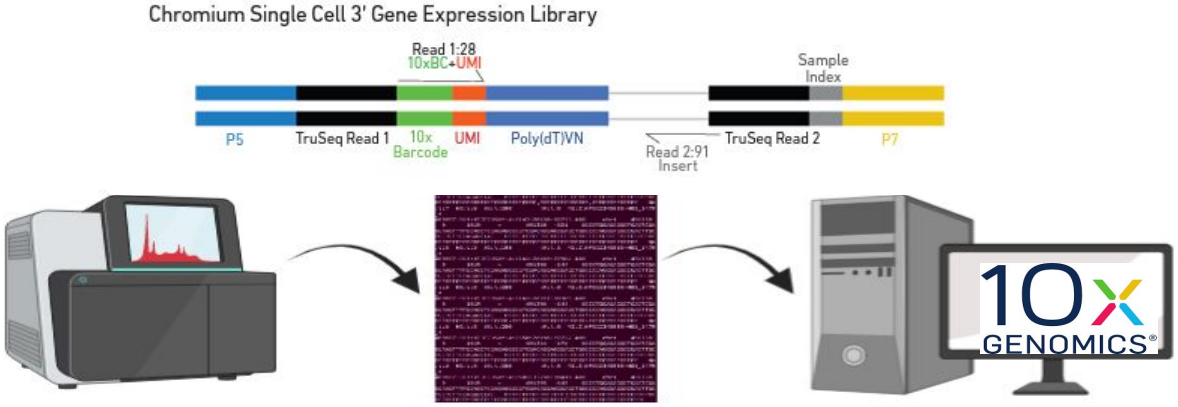
Reverse Transcription Reaction



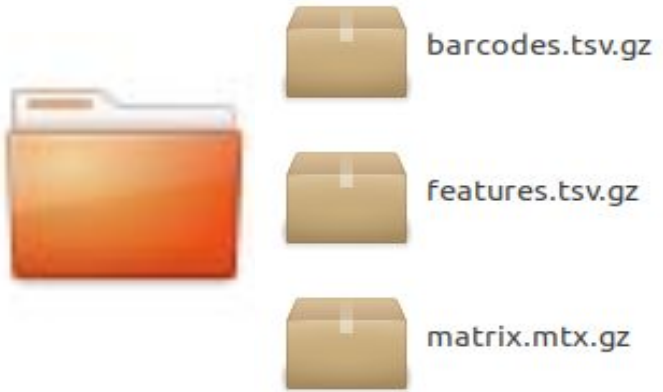
GEM



scRNAseq: next generation sequencing



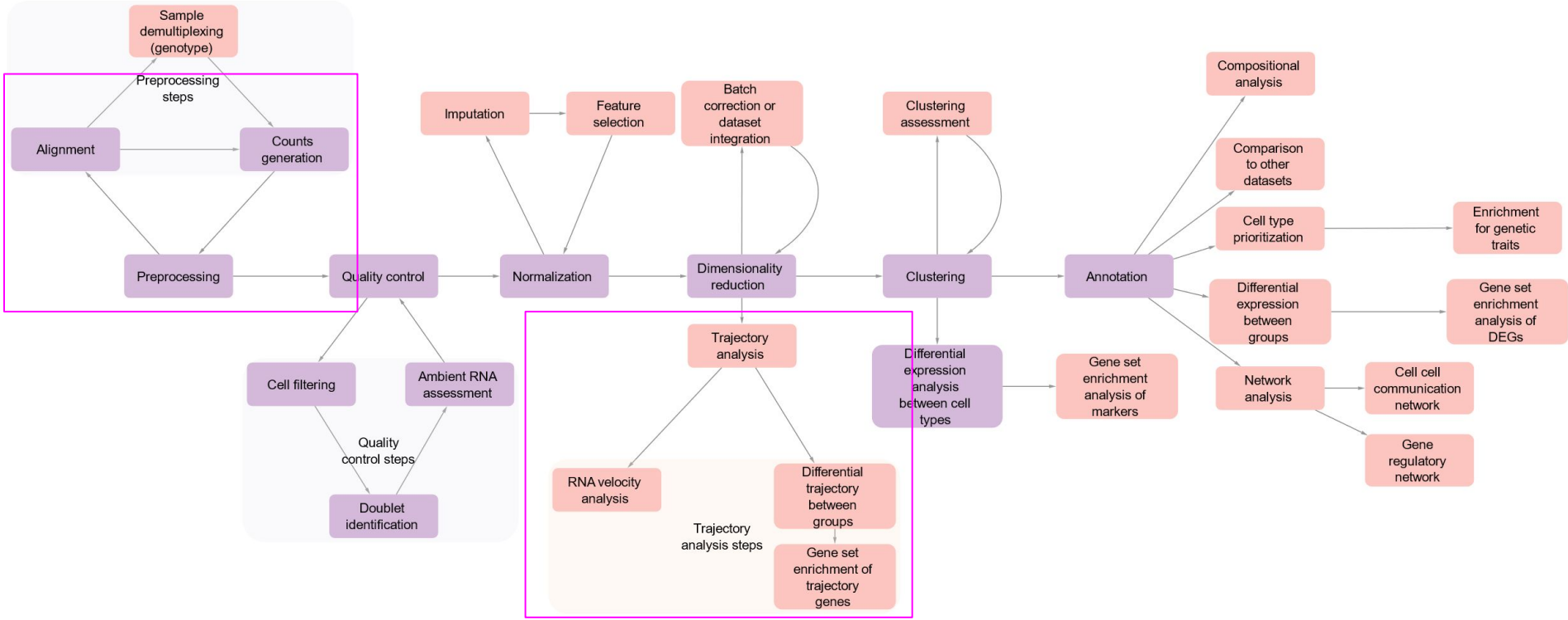
- cDNA is sequenced
- Each UMI is a read of RNA
- Cell barcodes with each RNA fragment
- Needs to be aligned to genome
- Read counts calculated
- Matched to cell barcodes



Analysis overview

Analysis overview

Already done!



Next session

Step	Tools
Clustering assessment	clustree, scclusteval
Gene set enrichment of trajectory genes	enrichR, fGSEA, ClusterProfiler
RNA velocity analysis	scVelo, velocyto
Differential trajectory between groups	tradeSeq
Counts generation	CellRanger*, alevin-fry, kallisto-bustools, DropSeq tools
Alignment	CellRanger*, alevin-fry, kallisto-bustools, DropSeq tools
Enrichment for genetic traits	EWCE, scDRS
Cell type prioritization	Augur
Trajectory analysis	monocle3*,slingshot
Gene set enrichment analysis of markers	enrichR*, fGSEA, ClusterProfiler
Gene regulatory network	SCENIC, SCENIC+, hdWGCNA
Cell cell communication network	CellChat*, CellPhoneDB, iTALK, LIANA
Network analysis	Various depending on type
Gene set enrichment analysis of DEGs	enrichR, fGSEA, ClusterProfiler
Differential expression between groups	muscat, Libra, edgeR, limma, DESeq2
Compositional analysis	propeller, scCODA
Comparison to other datasets	MetaNeighbor*, FR-Match
Differential expression analysis between cell types	MAST, Wilcoxon tests, presto
Annotation	scClassify*, BRETIGEA, BrainInABlender, Azimuth, scArches
Clustering	Louvain*, Leiden, k-nearest neighbors
Batch correction or dataset integration	Seurat integration*, Harmony, LIGER, Scanorama, ComBat
Dimensionality reduction	PCA*, UMAP*, tSNE
Feature selection	Seurat vst*, marker based
Imputation	MAGIC
Normalization	Log normalization*, scran, sctransform
Quality control	SampleQC
Preprocessing	Various depending on substeps
Sample demultiplexing (genotype)	CellSNP, demuxafy

Resources

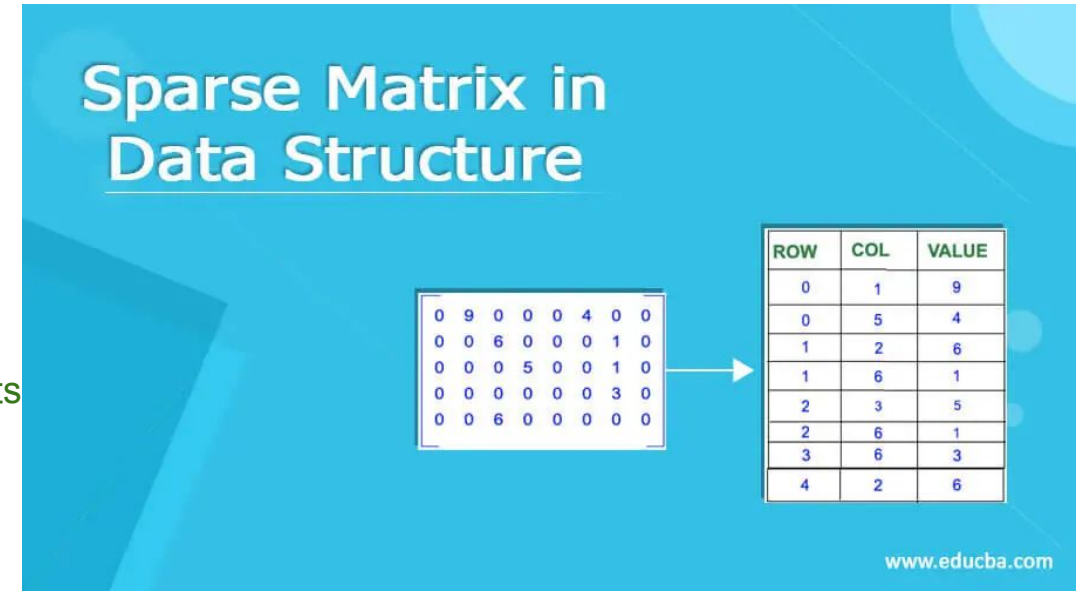
- Growing database of single-cell data analysis tools:
  - https://www.scrna-tools.org/ (Zappia & Theis, 2021)
- Guides for single cell data analysis:
  - Theis lab: https://www.sc-best-practices.org/preamble.html (Luecken & Theis, 2019, Heumos et al., 2023)
  - Hemberg lab: https://www.singlecellcourse.org/ (Andrews et al., 2020)
  - Bioconductor: http://bioconductor.org/books/3.16/OSCA.workflows/
  - 10X Analysis guides presented yesterday: https://www.10xgenomics.com/resources/analysis-guides
- GUI based options (similar to NeuroHub):
  - https://www.genap.ca/p/help/single-cell-in-genap

Tips

- Develop your general programming skills
- Always look at the documentation
- Go to Github or the Bioconductor or Stack Overflow forum for help
- When in doubt, always refer to benchmarking studies:
  - Examples:
    - Doublet detection: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7897250/
    - Differential expression: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9225332/
    - Gene regulatory network: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7098173/

Loading data

Loading data



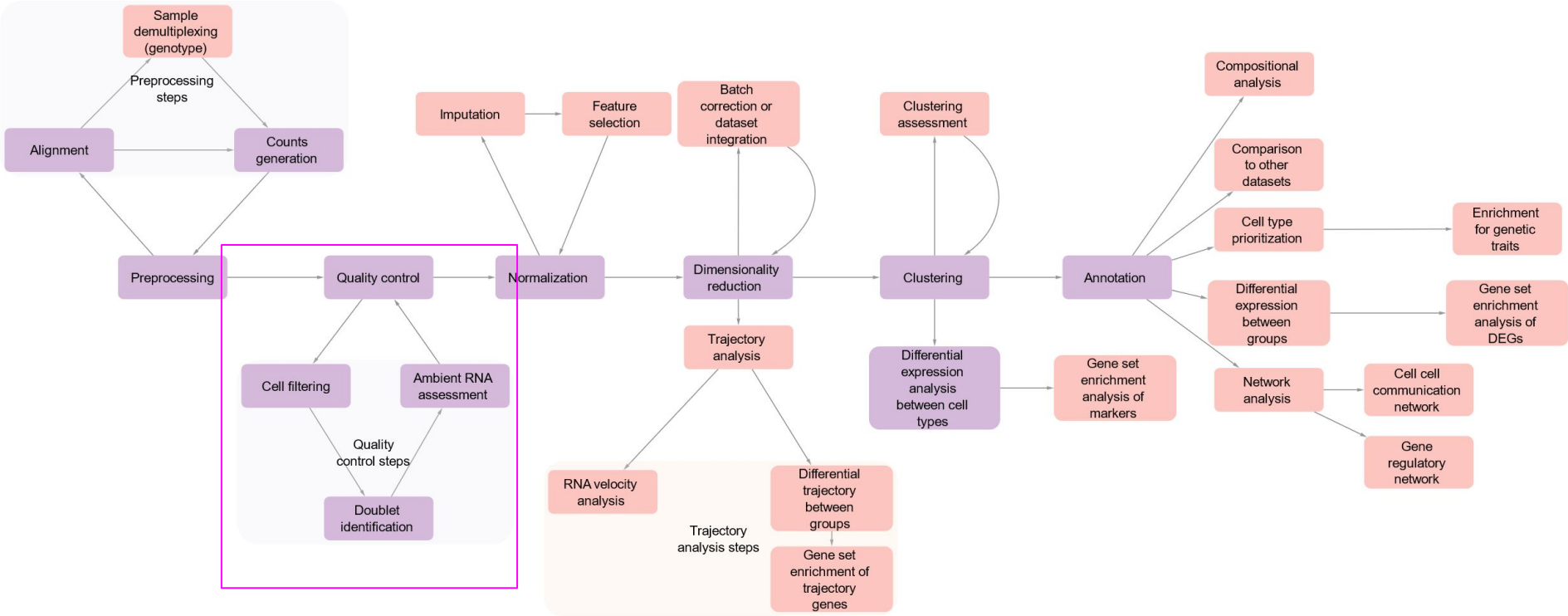
Understanding the R data object

nCount_RNA is the number of molecules of RNA in one cell

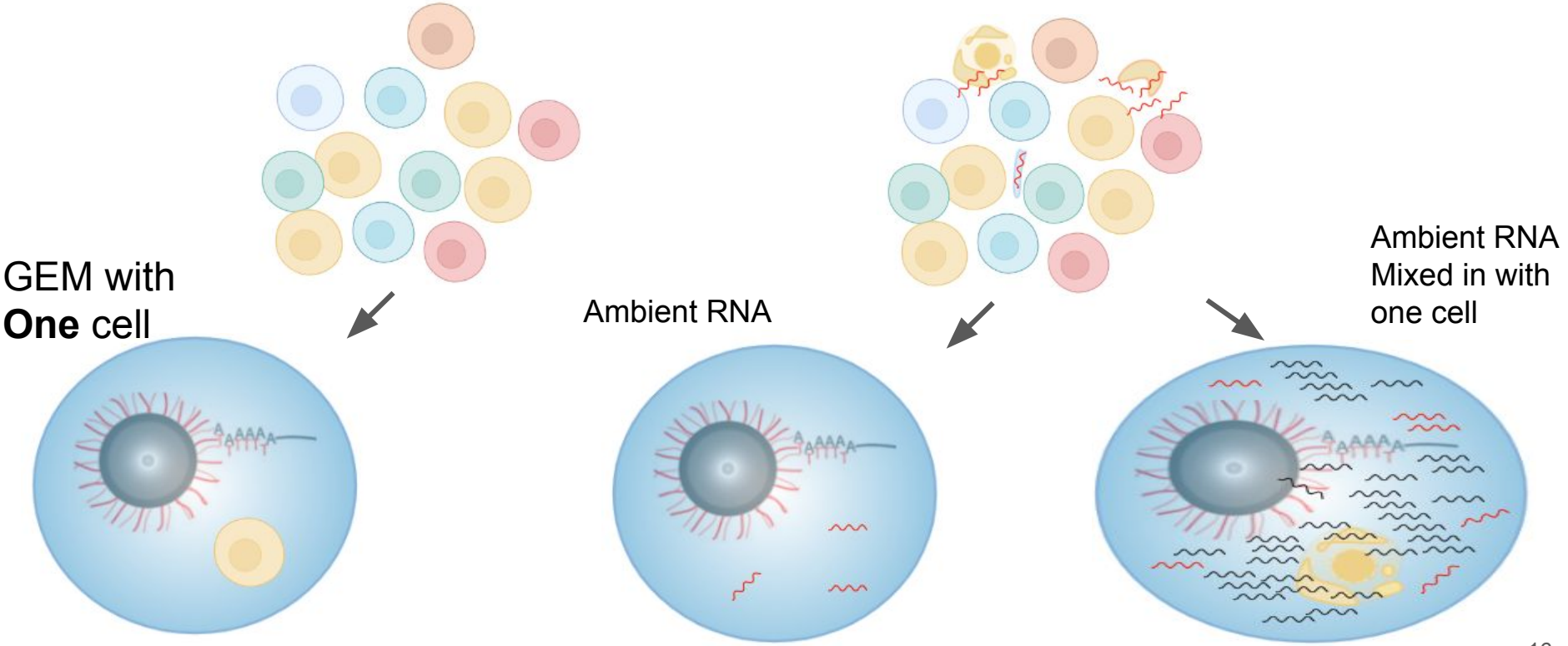
nFeature_RNA is the number of different genes in one cell

	Cell 1	Cell 2	Cell 3
Gene 1	0	0	50
Gene 2	1	1	1
Gene 3	20	0	4
nCount_RNA	21	1	55
nFeature_RNA	2	1	3

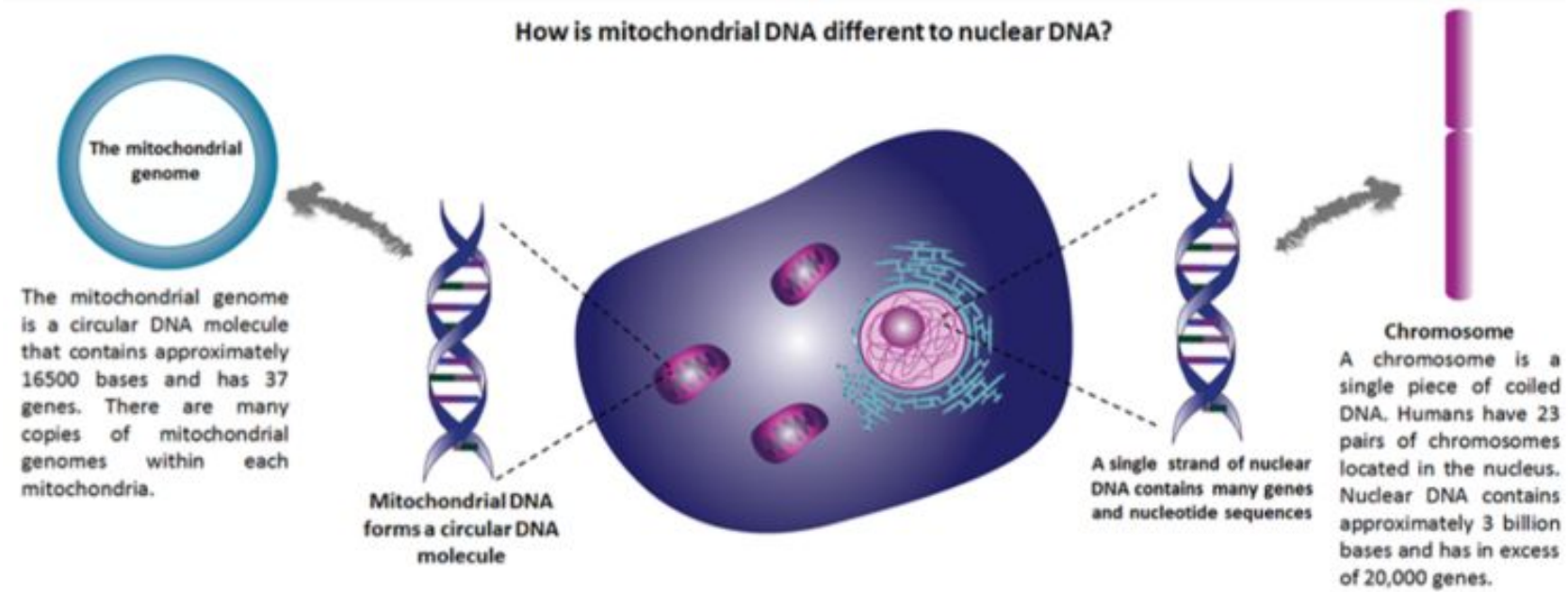
Quality control



Quality control: empty drops and ambient RNA



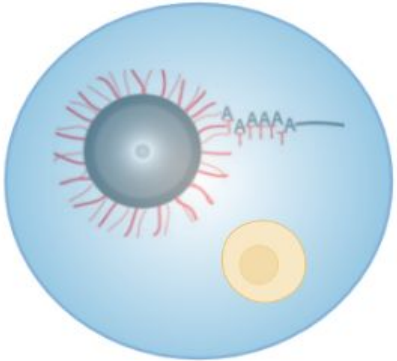
Quality control: mitochondrial reads



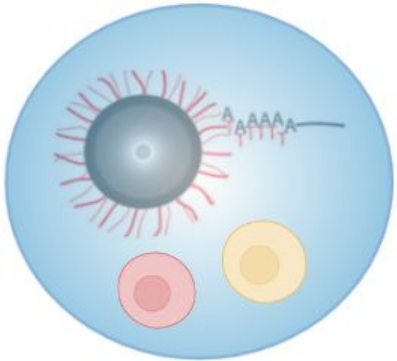
For single-nucleus sequencing - may choose to exclude mitochondrial genes, for single-cell sequencing may filter for high mitochondrial gene percentage.

Quality control: doublets and multiplets

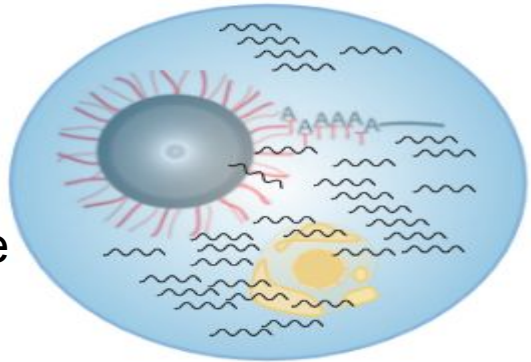
GEM with **One** cell
"Singlet"



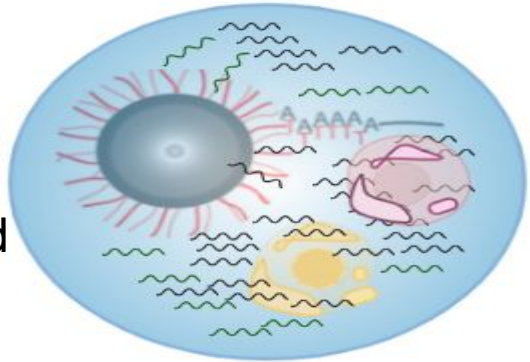
GEM with **Two** cells
"Doublet"



RNA from **One** cell
Desired single cell library



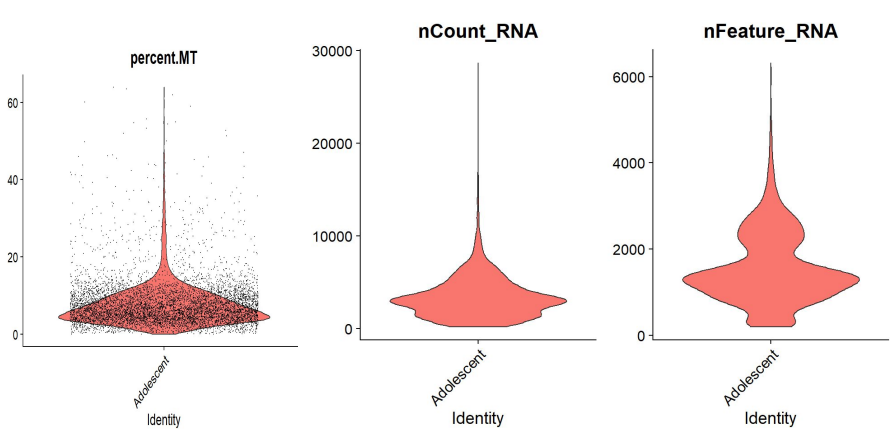
RNA from **Two** cells
Undesired and should be removed



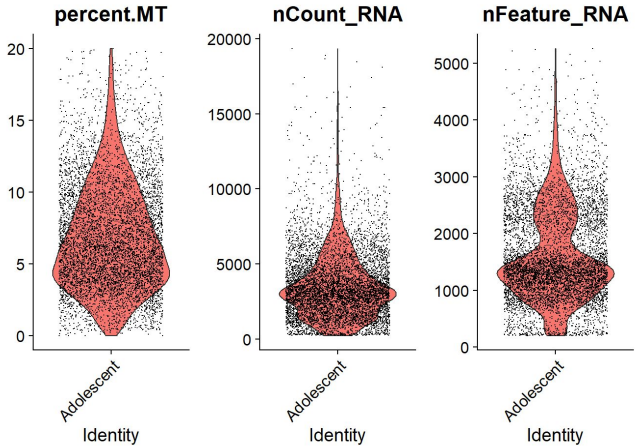
Deciding on filters - what to consider

- You want to remove non-cells or possibly dead/unhealthy cells
- You might want to remove doublets (or try)
  - Keep in mind the methods are estimates
  - You can mark potential doublets like we will do here
- The idea is to reduce noise
- We should apply the same cut off to data that we later want to merge
  - Sometimes the method of sequencing is different and then you might need different filtering parameters
- You might want to look at each sample separately before deciding on cut-offs

Visualize pre and post filter

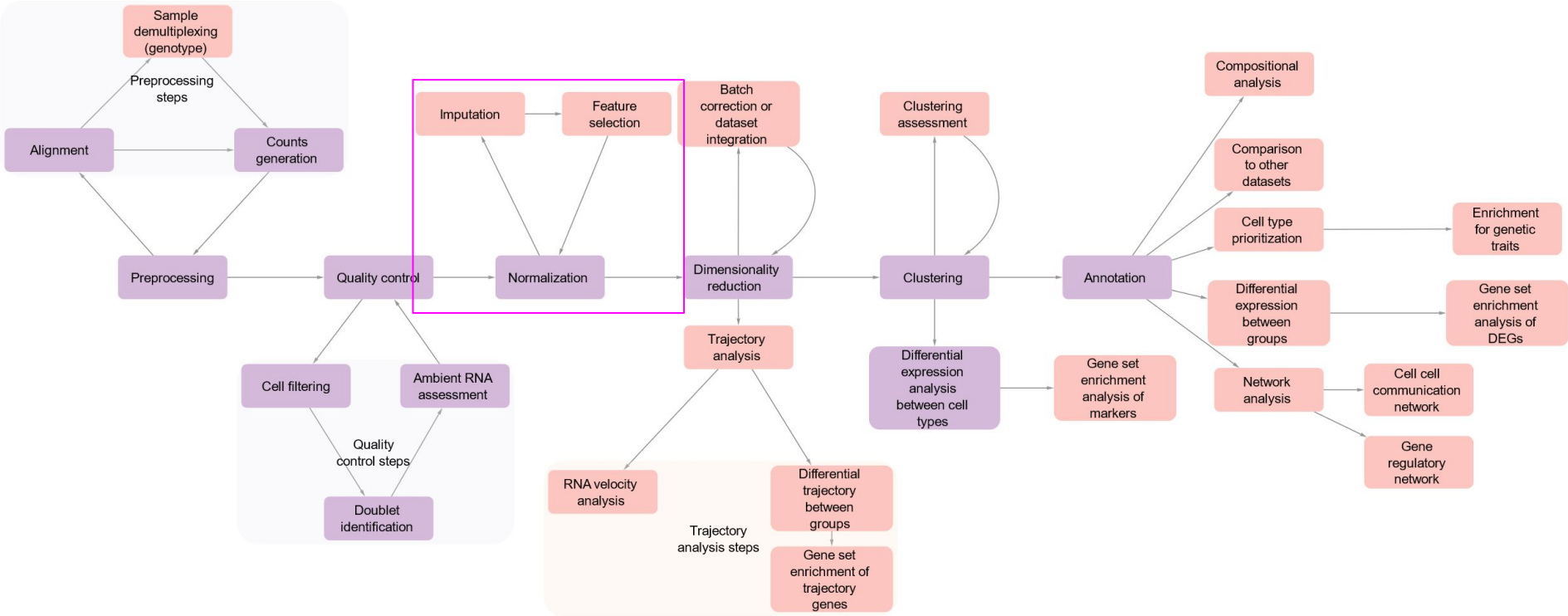


Pre filter



Post filter

Normalization and Feature Selection



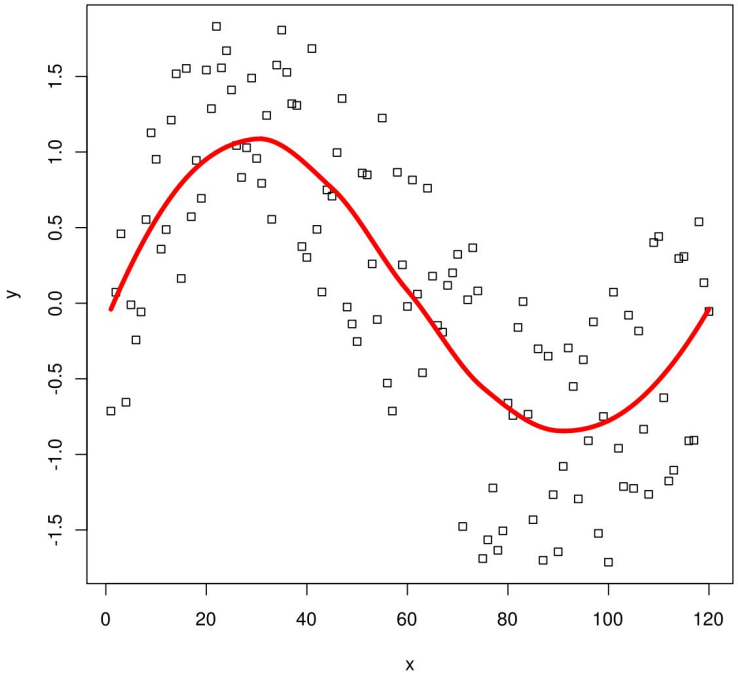
Normalize the counts of genes per cell or nucleus

Method	Author	Year	Spike-ins	Model Description
SAMstrt	Katayama et al.	2013	Yes	Poisson resampling and non-parametric statistics
BASiCS	Vallejos et al.	2015	Yes	Use spike-ins for hierarchical Poisson/ Gamma model for technical variability. Expand model to incorporate biological genes with new Poisson model
GRM	Ding et al.	2015	Yes	Gamma regression model from spike-ins
Simple Norm.	Satija et al.	2015	No	Divide gene counts for cells, then multiply by scale factor and apply a log(x +1) transformation to the result (included in the Seurat package as NormalizeData)
scran	Lun et al.	2016	No	Deconvolution of size factors from constructed linear system. Form pools of cells and normalize using summed expression values
SCnorm	Bacher et al.	2017	Optional	Quantile based model for log sequencing depth.
Linnorm	Yip et al.	2017	Optional	Linear models defined with a normalization strength coefficient to update means. Focuses on stable genes to perform normalization

- Additional methods like sctransform, also implemented in Seurat (Haefmeister & Satija, 2019)
- All methods seek to account for the fact that we have different amounts on information for different cells

Lytal et al., 2020

Feature selection



Local polynomial regression

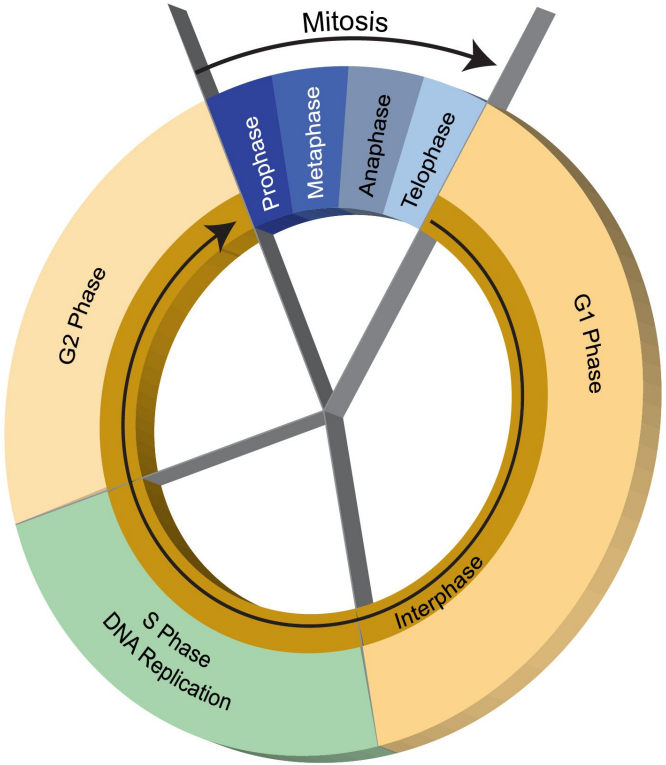
- Can be done manually or automatically
- Consider how many features to use

Using vst algorithm

- fits a line to the relationship of log(variance) and log(mean) using local polynomial regression (loess)
- standardizes the feature values using the observed mean and expected variance (given by the fitted line)
- Feature variance is then calculated on the standardized values after clipping to a maximum
- X most variable features are selected

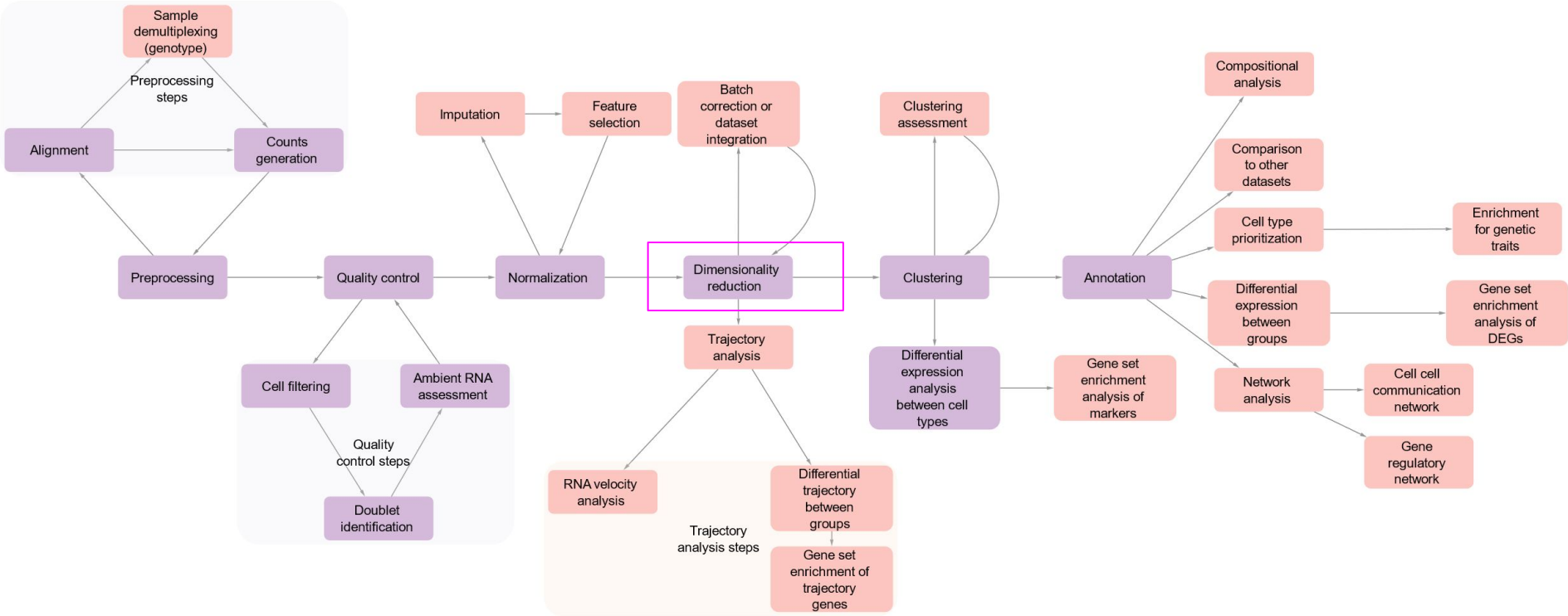
Using a measure of dispersion such as the variance to mean ratio, in log scale

Cell cycle scoring



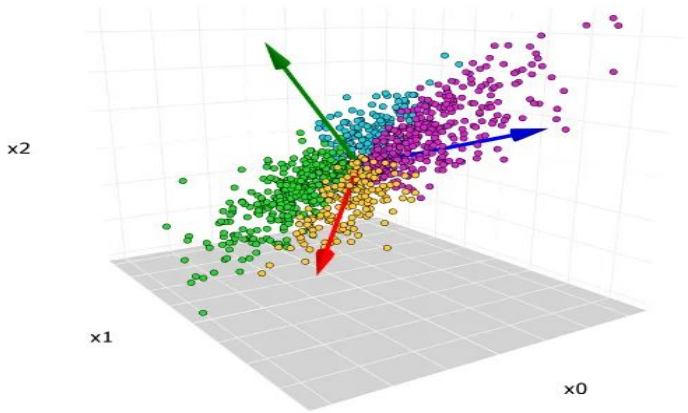
https://www.genome.gov/genetics-glossary/Cell-Cycle

Dimensionality Reduction

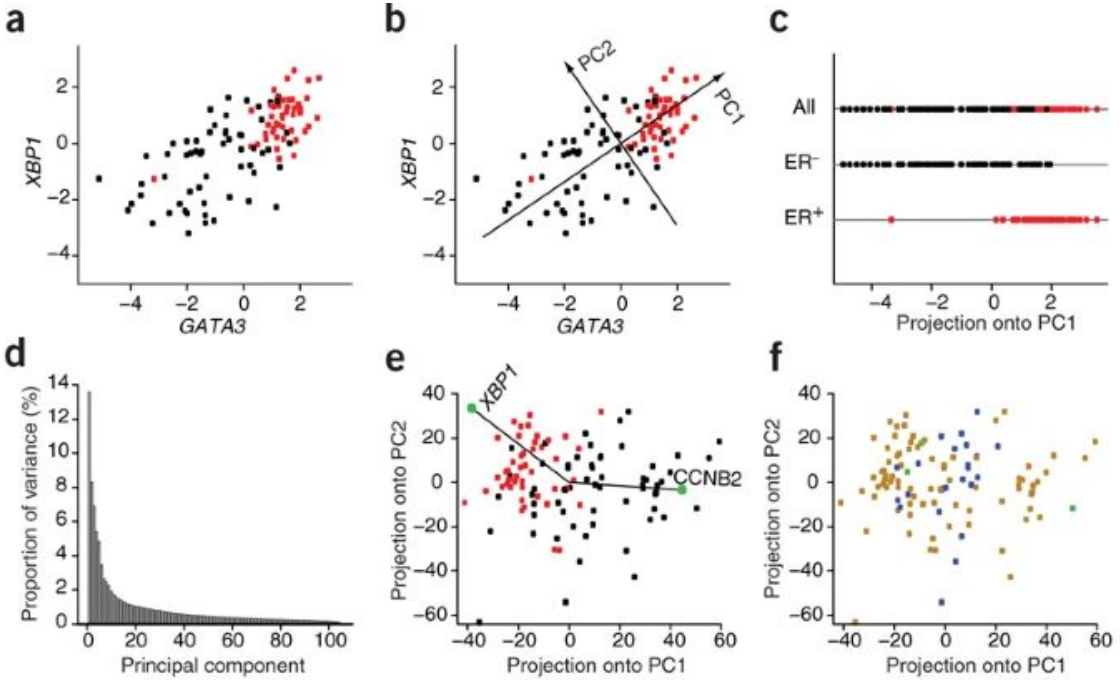


Principal component analysis

Multidimensional data



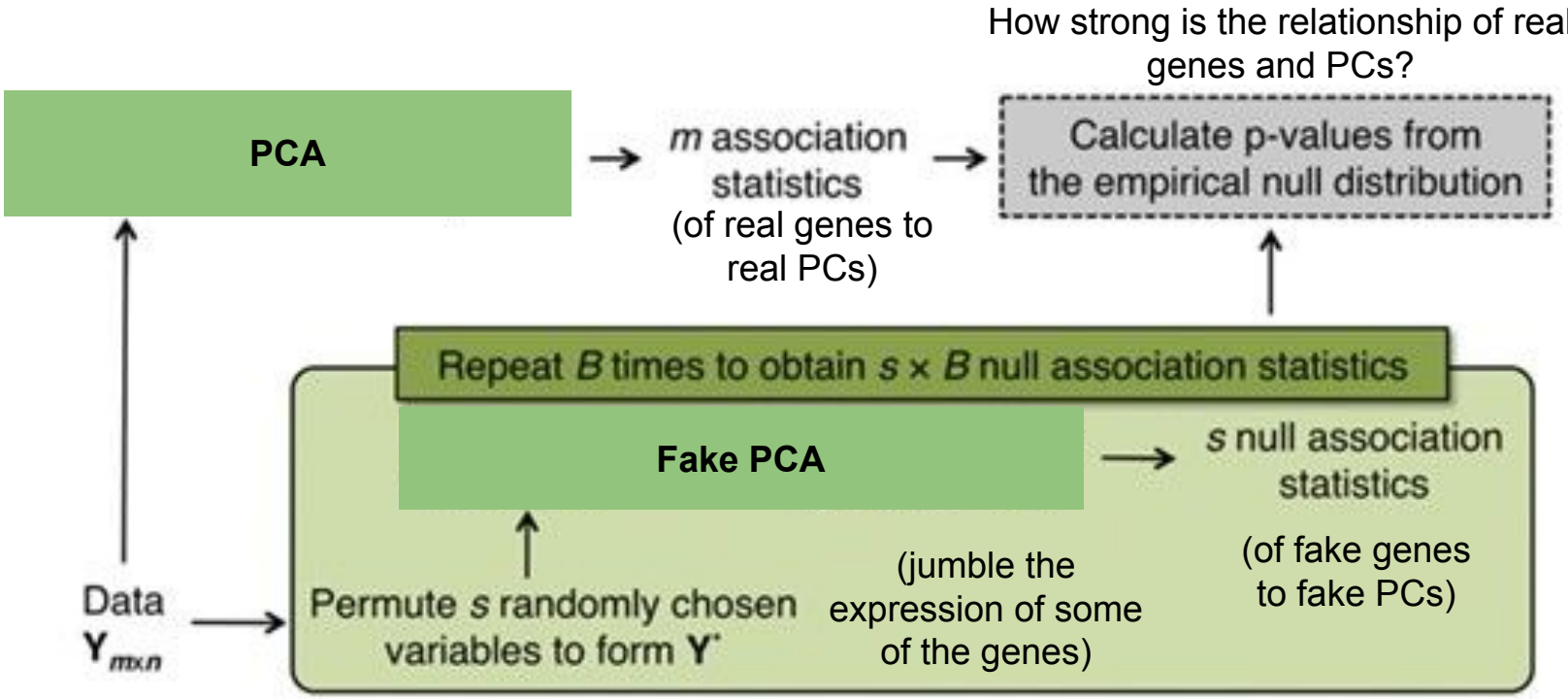
Calculate the variance of gene expression across cells and combine the variance to create a PC



Ringnér, 2008 (Nat Biotech)

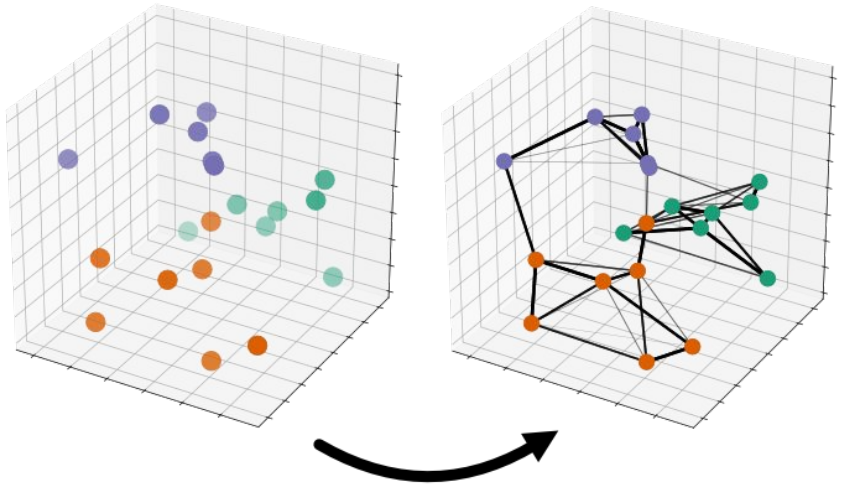
https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

Principal component analysis choosing number of PCs

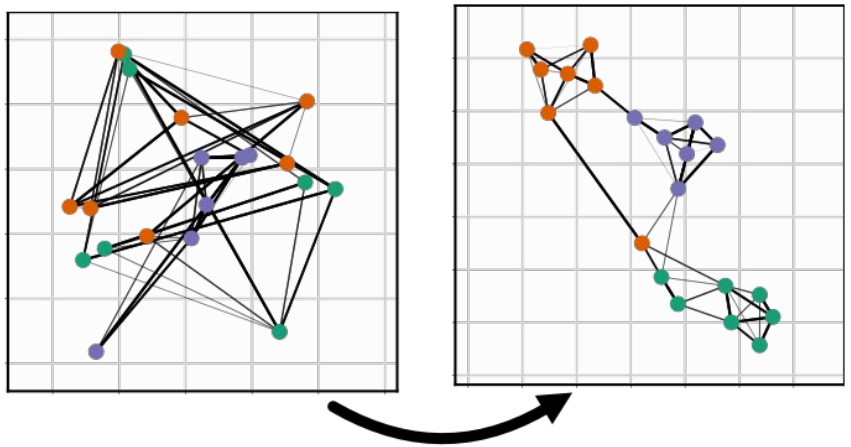


JackStraw: [Chung & Storey, 2014]

UMAP visualization



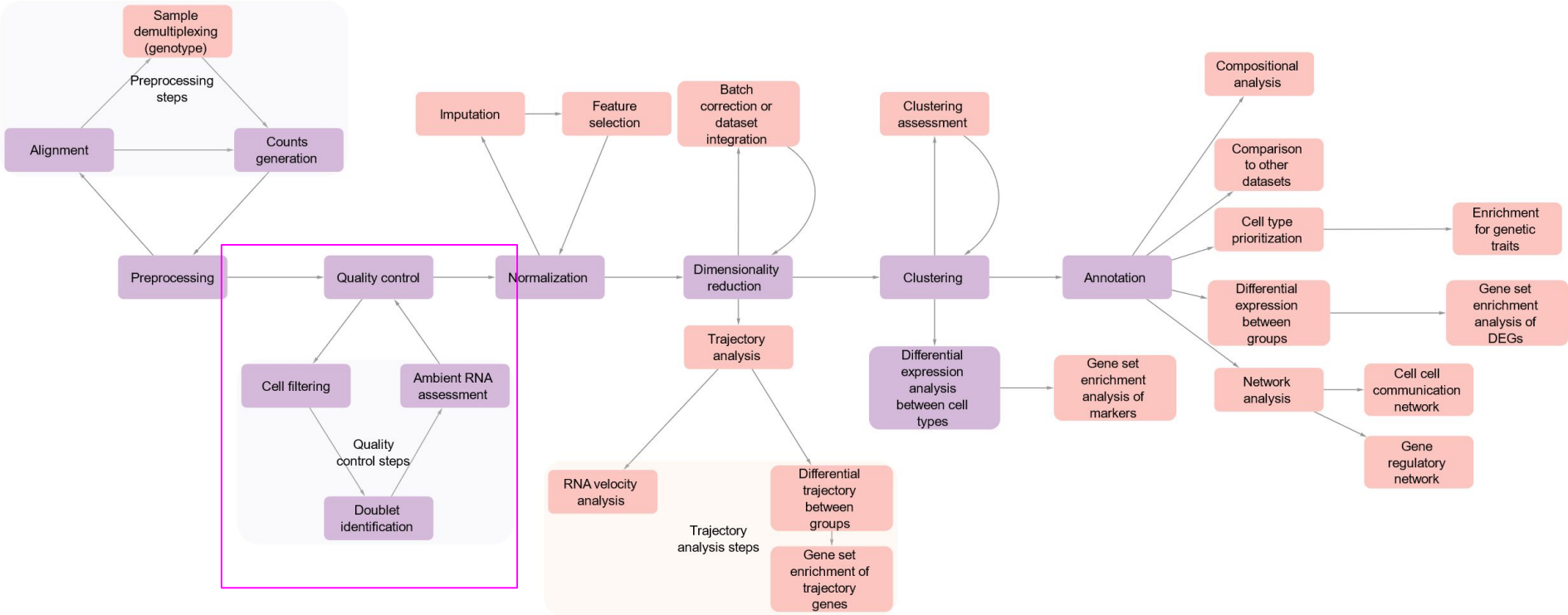
Compute a graphical representation
of the dataset



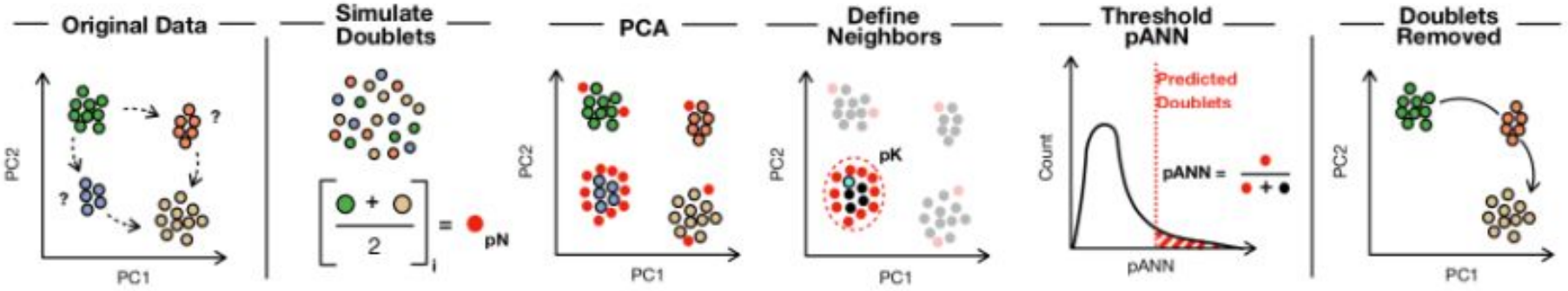
Learn an embedding that preserves
the structure of the graph

https://umap-learn.readthedocs.io/en/latest/parametric_umap.html

Quality control



Quality Control - DoubletFinder (McGinnis et al., 2019)

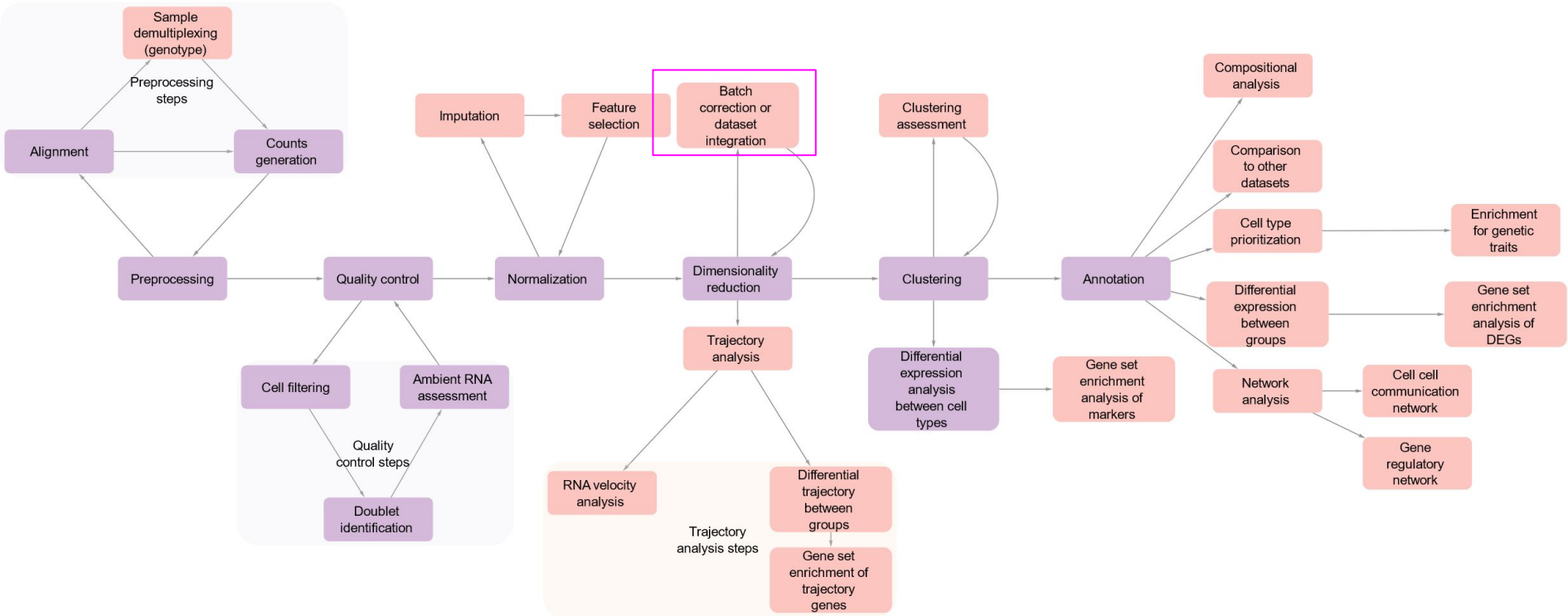


https://github.com/chris-mcginnis-ucsf/DoubletFinder

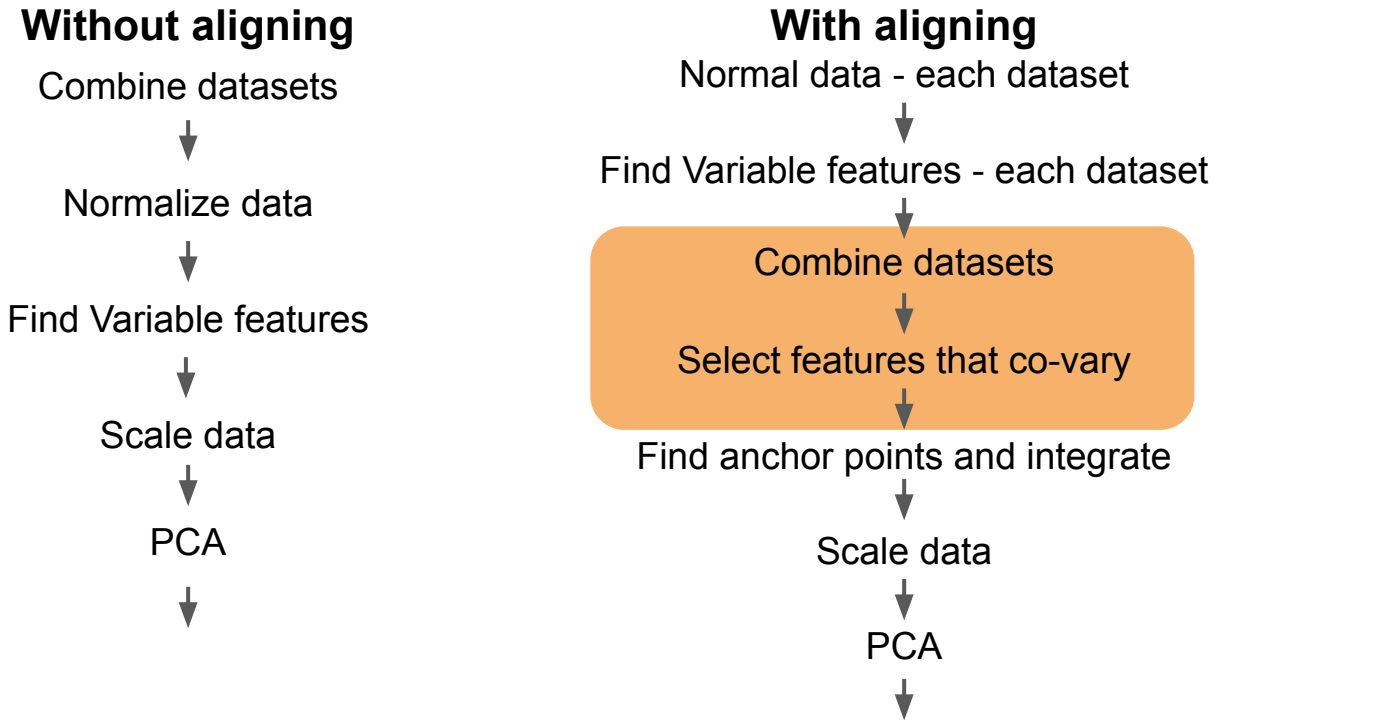
Types of doublets:

- Heterotypic (of two different cell types) or Homotypic (of the same cell type)
- Ground truth (when different subjects are multiplexed)
  - Barcode corresponding to two different genotypes is a doublet
  - Using sample “tags”, barcodes with tags from different samples are doublets

Batch correction



Prepare for principal component analysis and clustering



Merging samples and aligning is important to remove technical variation
BUT it can falsely remove biological variation

Combine data - integrate and align

1321	1325	7806	7343
3242	476	658	8313
5478	5454	7697	679
427	5747	136	7271

Data set 1

+

1321	1325	7806	7343
3242	476	658	8313
5478	5454	7697	679
427	5747	136	7271

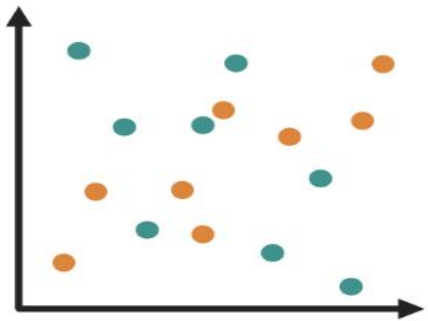
Data set 2

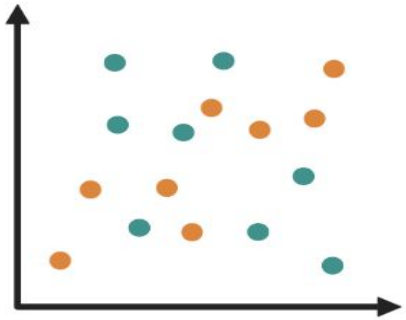
=

1321	1325	7806	7343
3242	476	658	8313
5478	5454	7697	679
427	5747	136	7271
1321	1325	7806	7343
3242	476	658	8313
5478	5454	7697	679
427	5747	136	7271

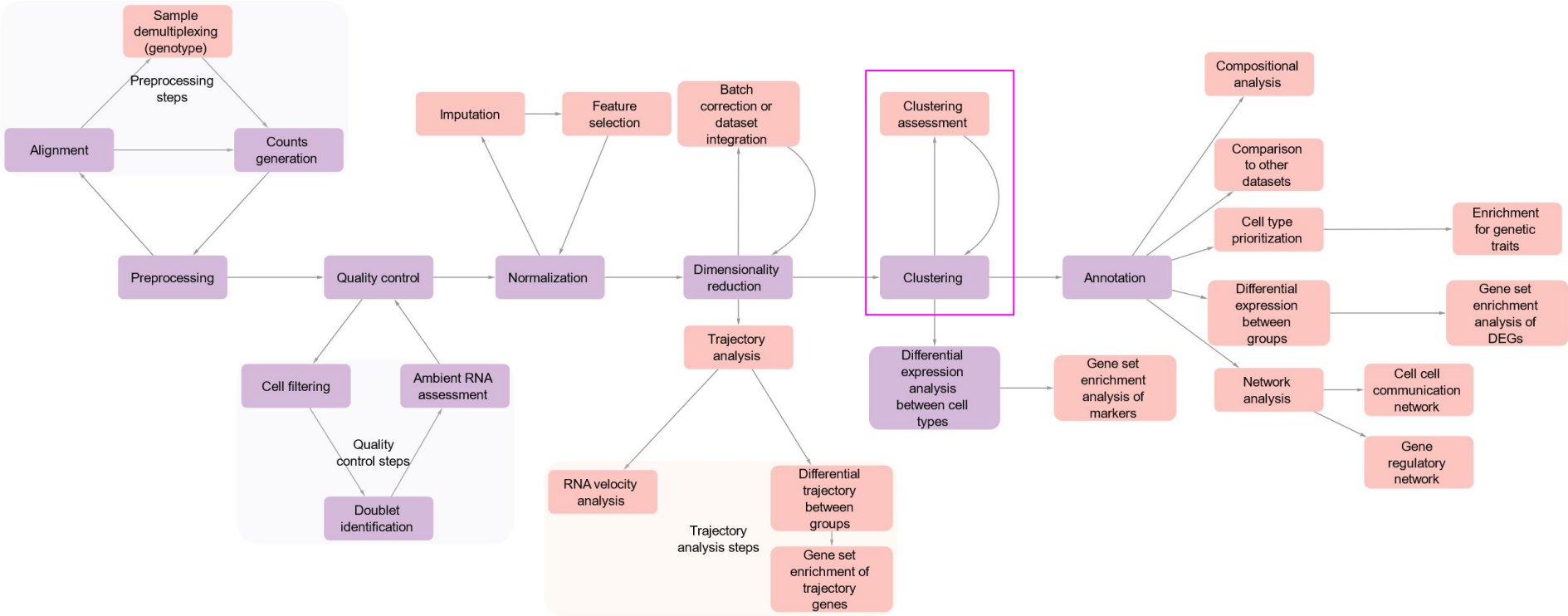
Normalize and find variable features

Integrate the data (shifts values) and finds variable features that vary in both datasets

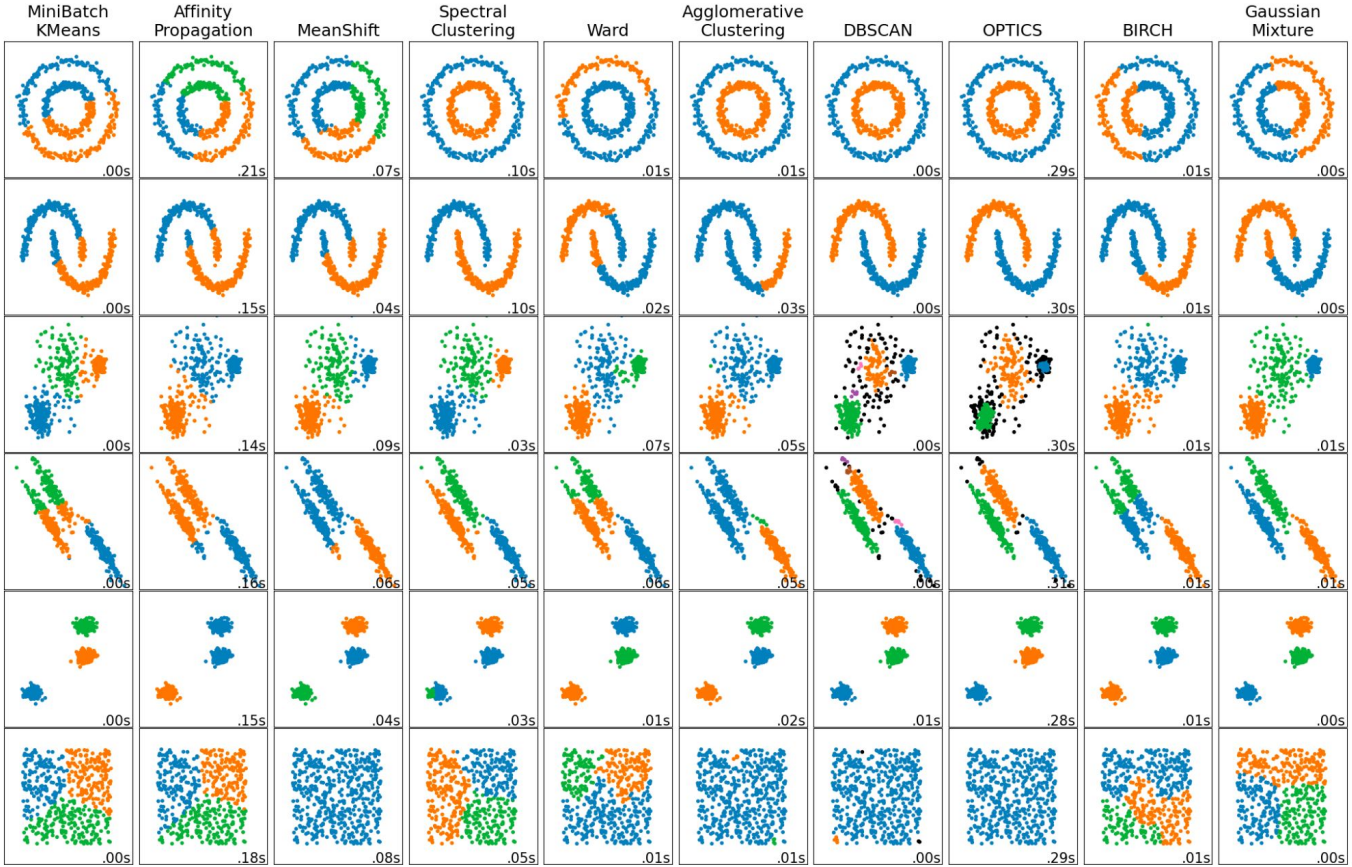




Clustering and clustering assessment



Clustering: similar things are grouped together



https://scikit-learn.org/stable/modules/clustering.html

Unsupervised Clustering of scRNAseq data: Network Detection

Combined data object: Integrated, Normalized, Scaled

↓

Dimensional reductions : PCA

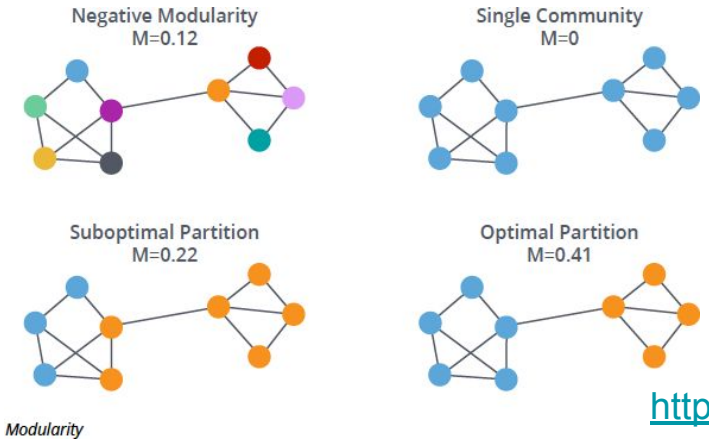
↓

Create nearest neighbour network : KNN or SNN

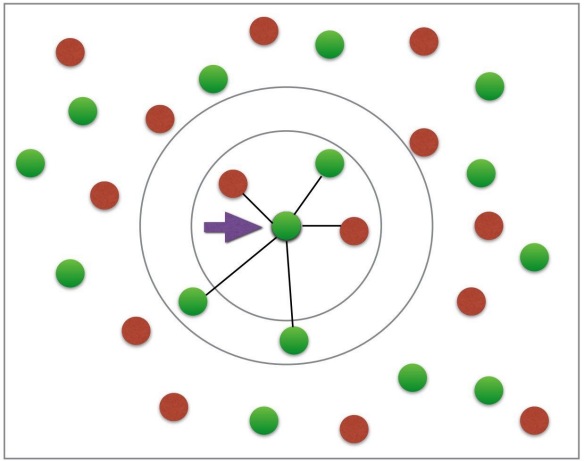
↓ **Set K**

Find clusters using network detection : Louvain

Set resolution



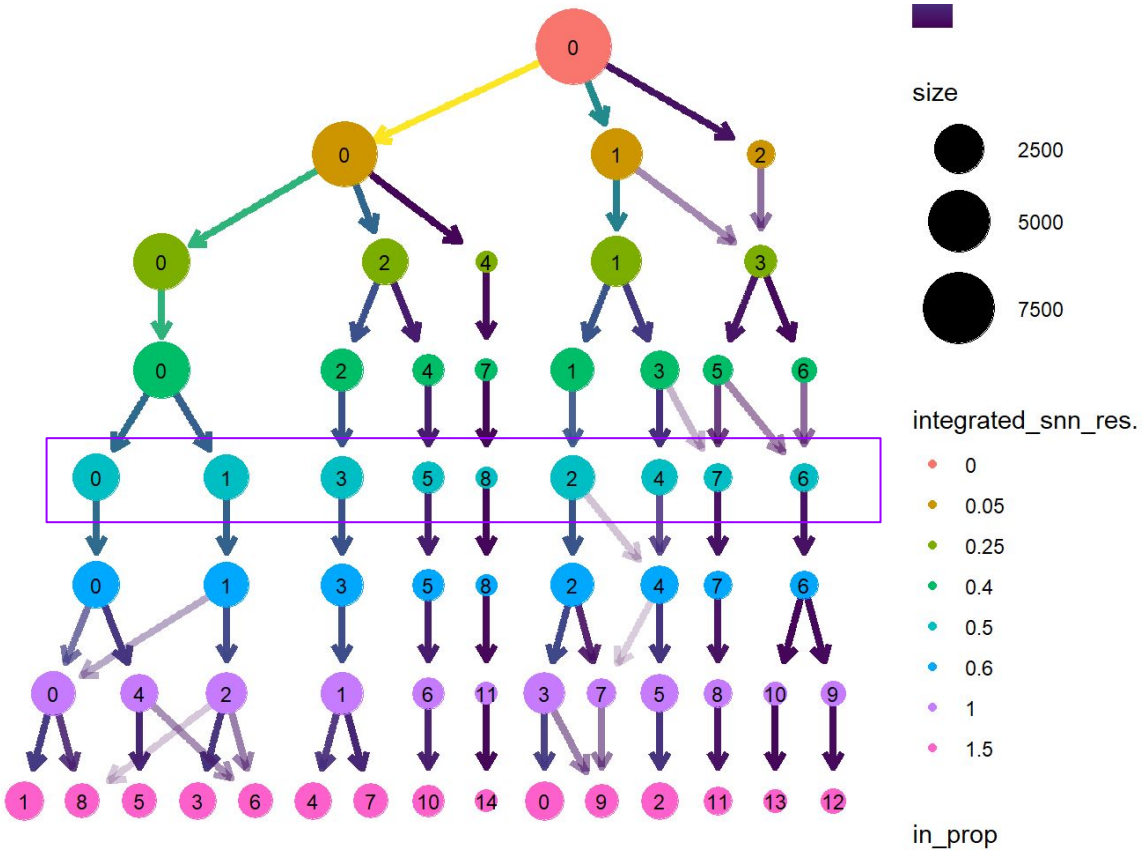
K- Nearest Neighbor



https://neo4j.com/blog/graph-algorithms-neo4j-louvain-modularity/

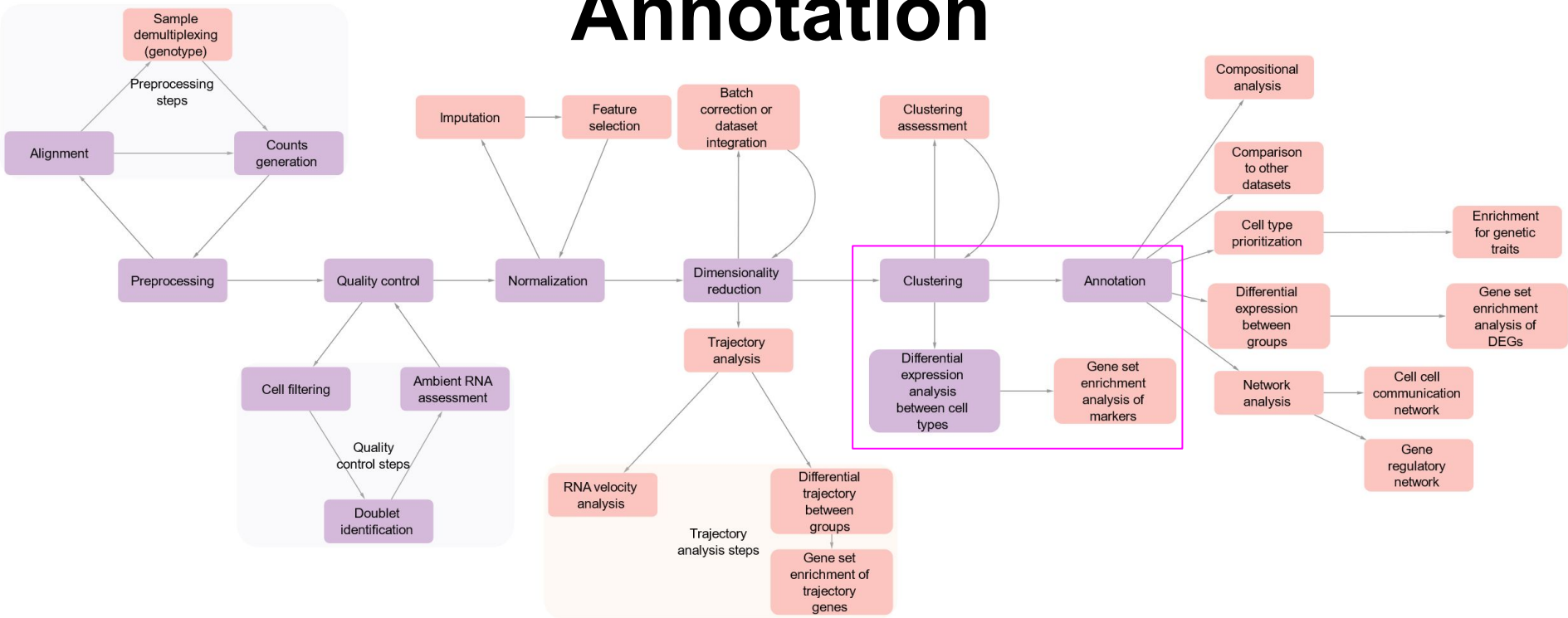
https://towardsdatascience.com/building-a-k-nearest-neighbors-k-nn-model-with-scikit-learn-51209555453a

Unsupervised Clustering at different resolutions



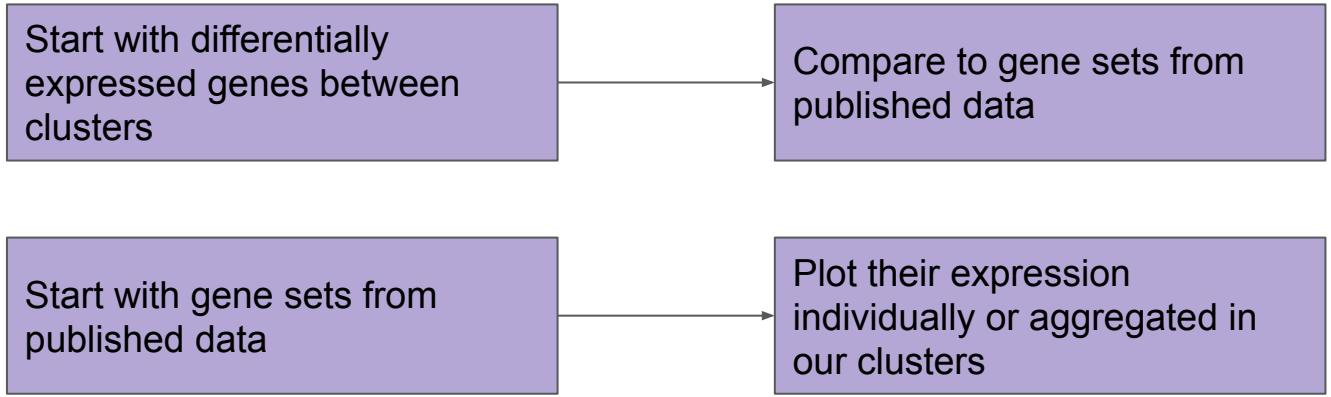
Clustree package

Annotation


Sample demultiplexing (genotype)
Preprocessing steps
Alignment
Counts generation
Preprocessing
Quality control
Cell filtering
Ambient RNA assessment
Quality control steps
Doublet identification
Normalization
Imputation
Feature selection
Batch correction or dataset integration
Dimensionality reduction
Trajectory analysis
RNA velocity analysis
Trajectory analysis steps
Differential trajectory between groups
Gene set enrichment of trajectory genes
Clustering
Clustering assessment
Differential expression analysis between cell types
Gene set enrichment analysis of markers
Annotation
Compositional analysis
Comparison to other datasets
Cell type prioritization
Enrichment for genetic traits
Differential expression between groups
Gene set enrichment analysis of DEGs
Network analysis
Cell cell communication network
Gene regulatory network

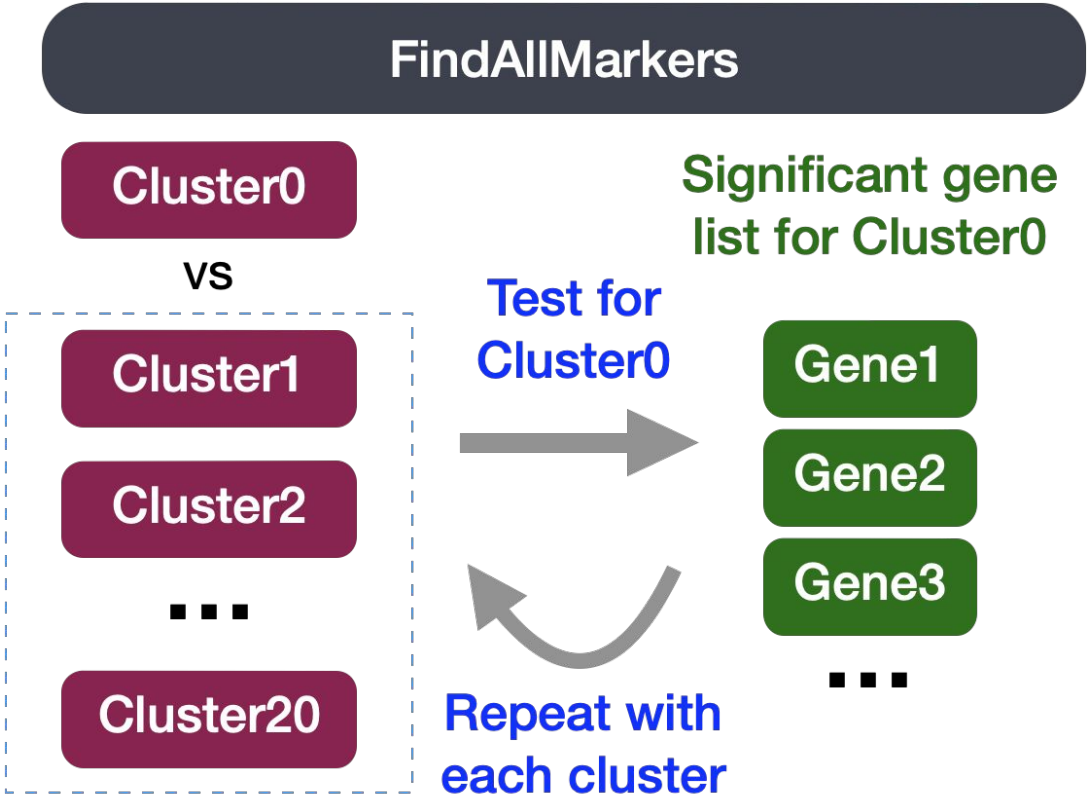
Cluster annotation approaches

- We need to decide what resolution of clusters to select for annotation
- We can take two annotation approaches

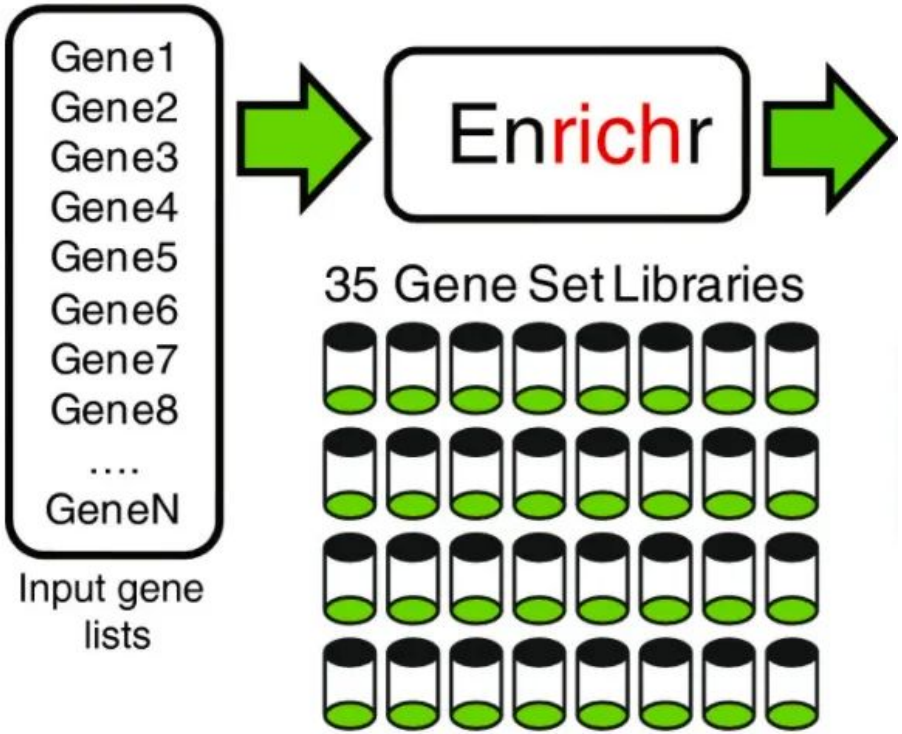


Find cluster markers

- Each cluster (X) is compared to all the other clusters
- Wilcoxon test is the default method (non-parametric t-tests)
- There is a correction for multiple comparisons
- For the DEG n= number of cells in a cluster



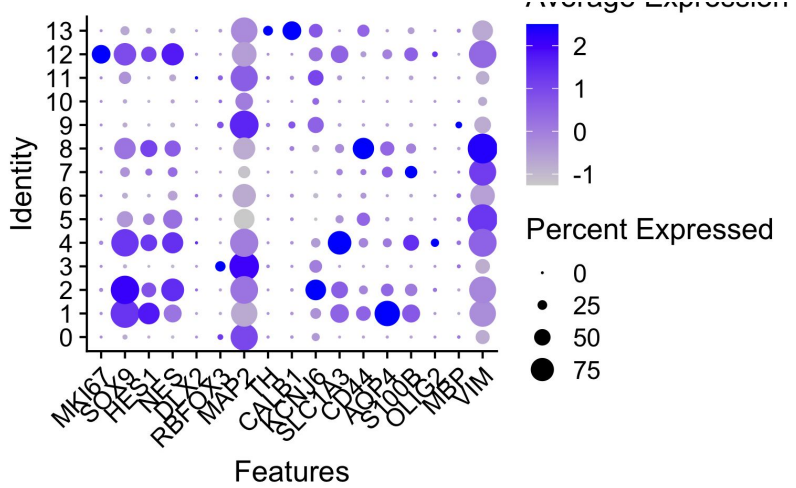
Look up cluster markers using EnrichR

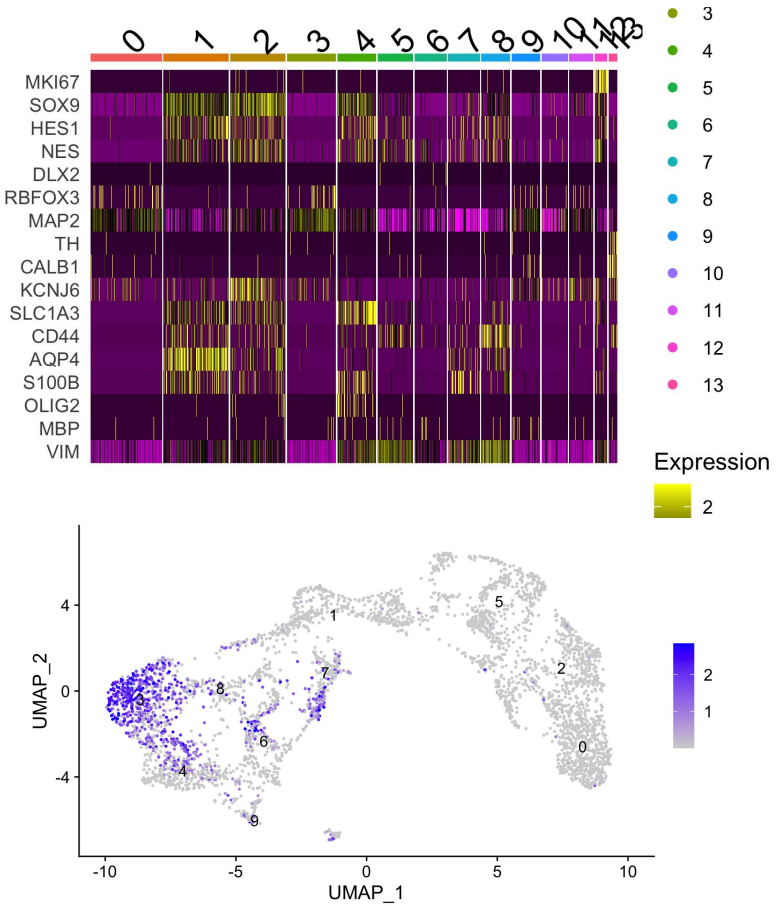


Cell types prediction for each reference library

Visualize marker expression

- Requires a custom list of cell type markers (from literature)
- Dot plot
- Heatmap
- Feature plot

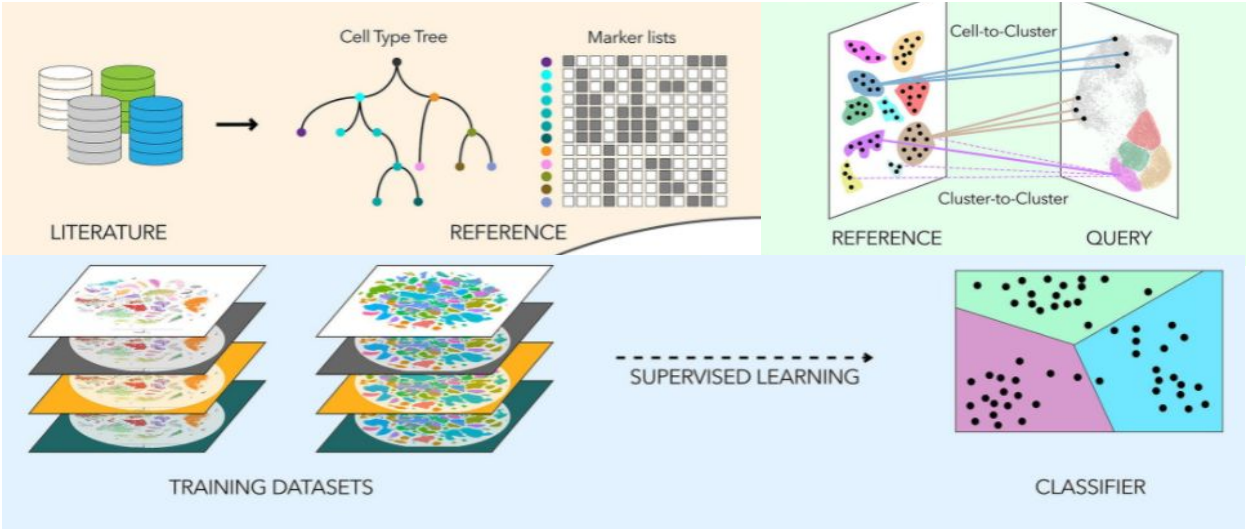




Automated annotation

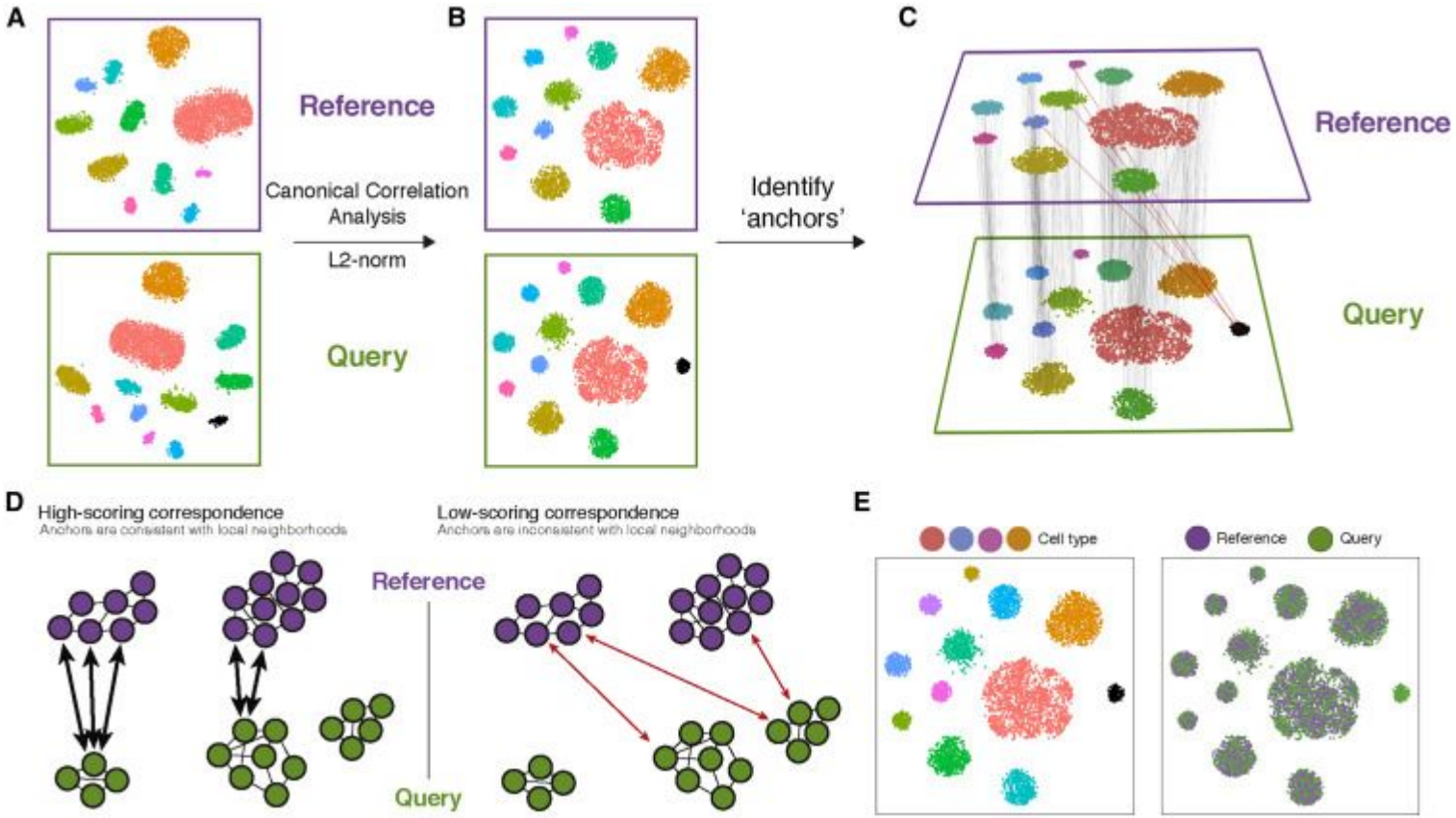
Some tools annotate individual cells and some annotate clusters

- Marker Gene Database-based: Requires public databases and ontologies describing cell type specific markers for the expected cell types
- Correlation-based: Requires labelled scRNAseq reference datasets
- Supervised Classification-based: Requires labelled scRNAseq reference datasets



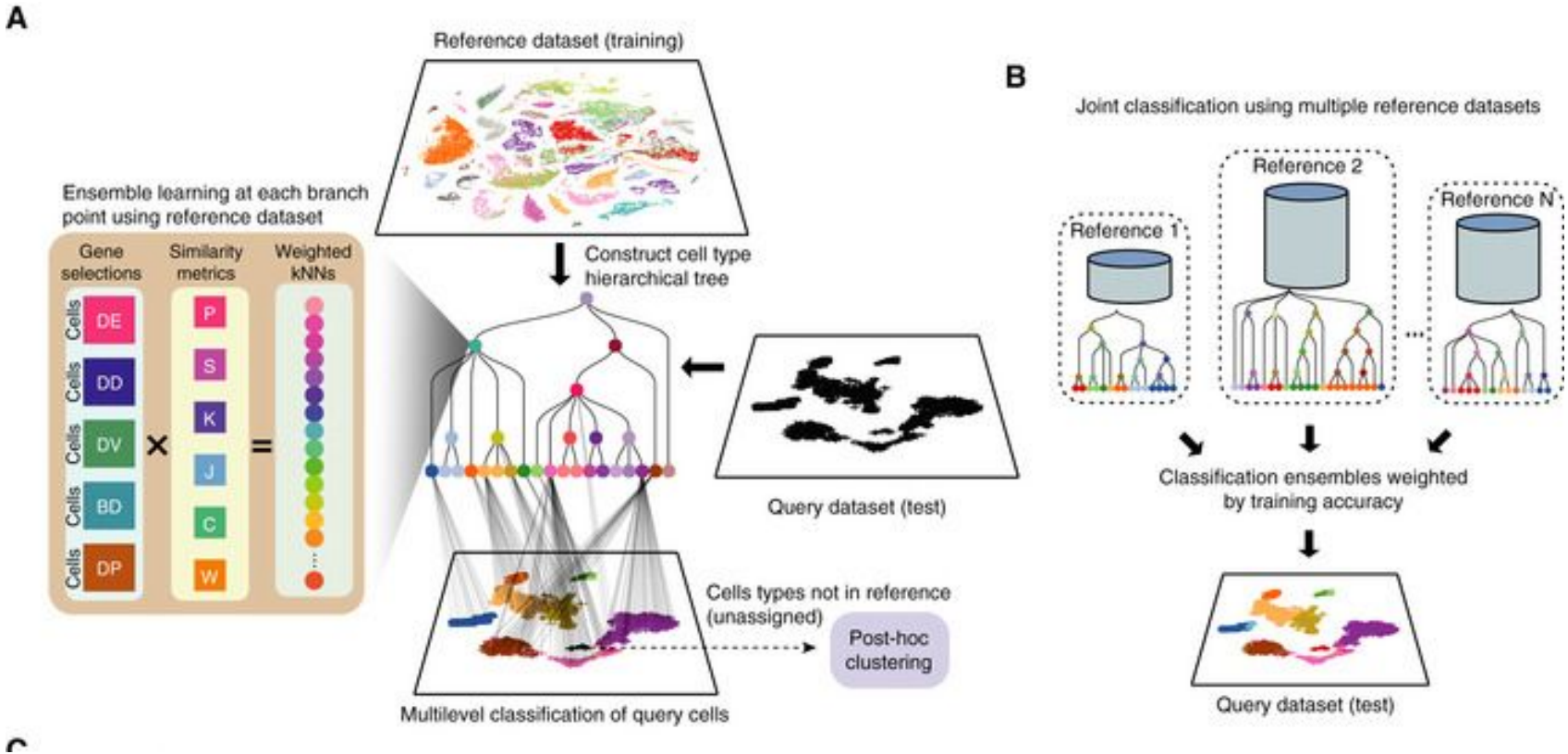
Pasquini, Giovanni, et al. "Automated methods for cell type annotation on scRNA-seq data." *Computational and Structural Biotechnology Journal* (2021).

Automated cluster annotation with Seurat



Stuart, Tim, et al. "Comprehensive integration of single-cell data." *Cell* 177.7 (2019): 1888-1902.

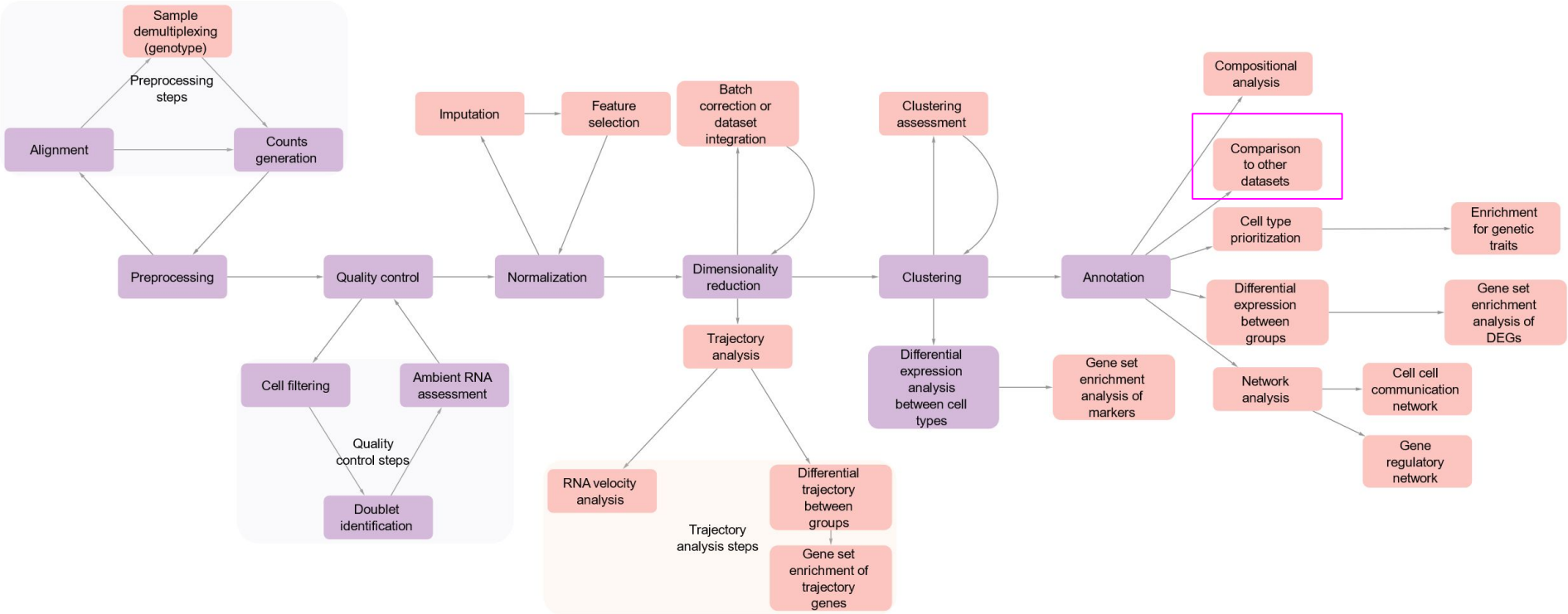
Automated cluster annotation with scClassify



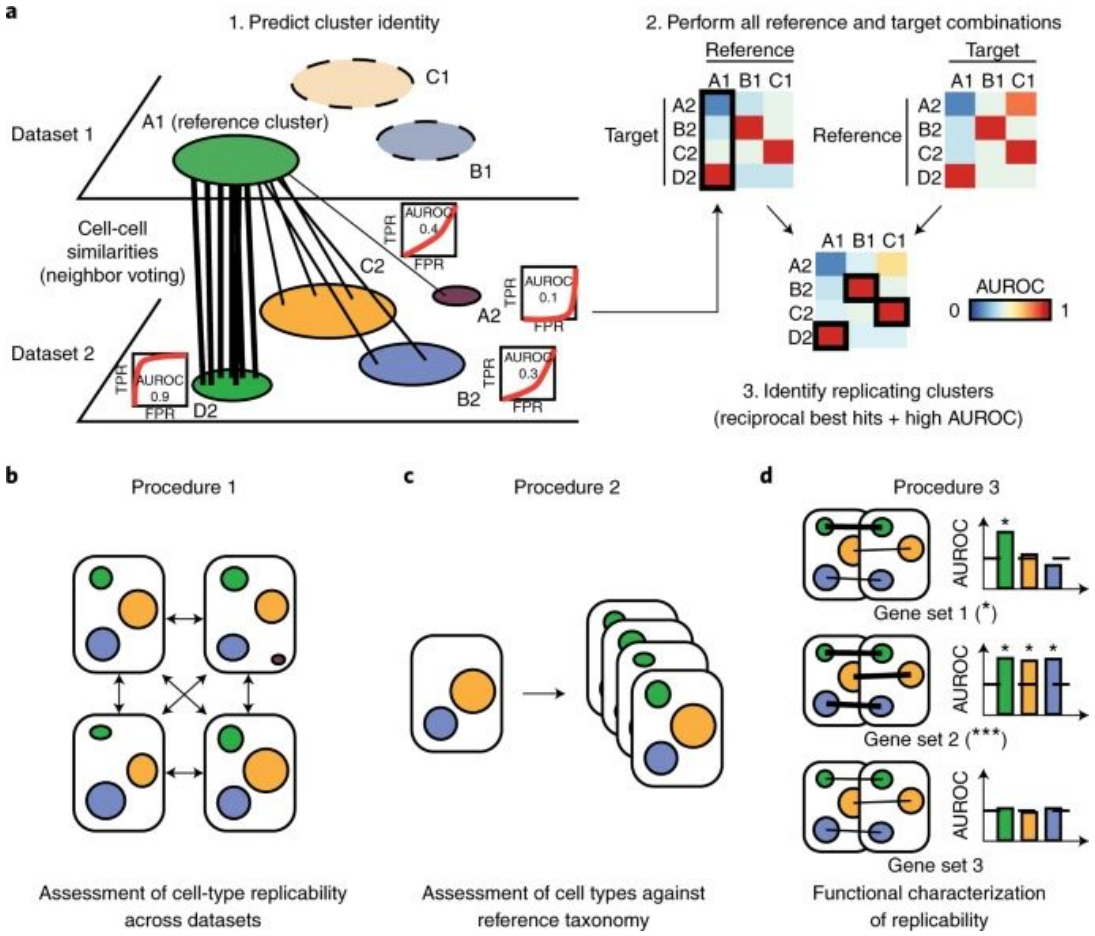
Lin, Yingxin, et al. "scClassify: sample size estimation and multiscale classification of cells using single and multiple reference." *Molecular systems biology* 16.6 (2020): e9389.

Add annotations

Comparison to other datasets

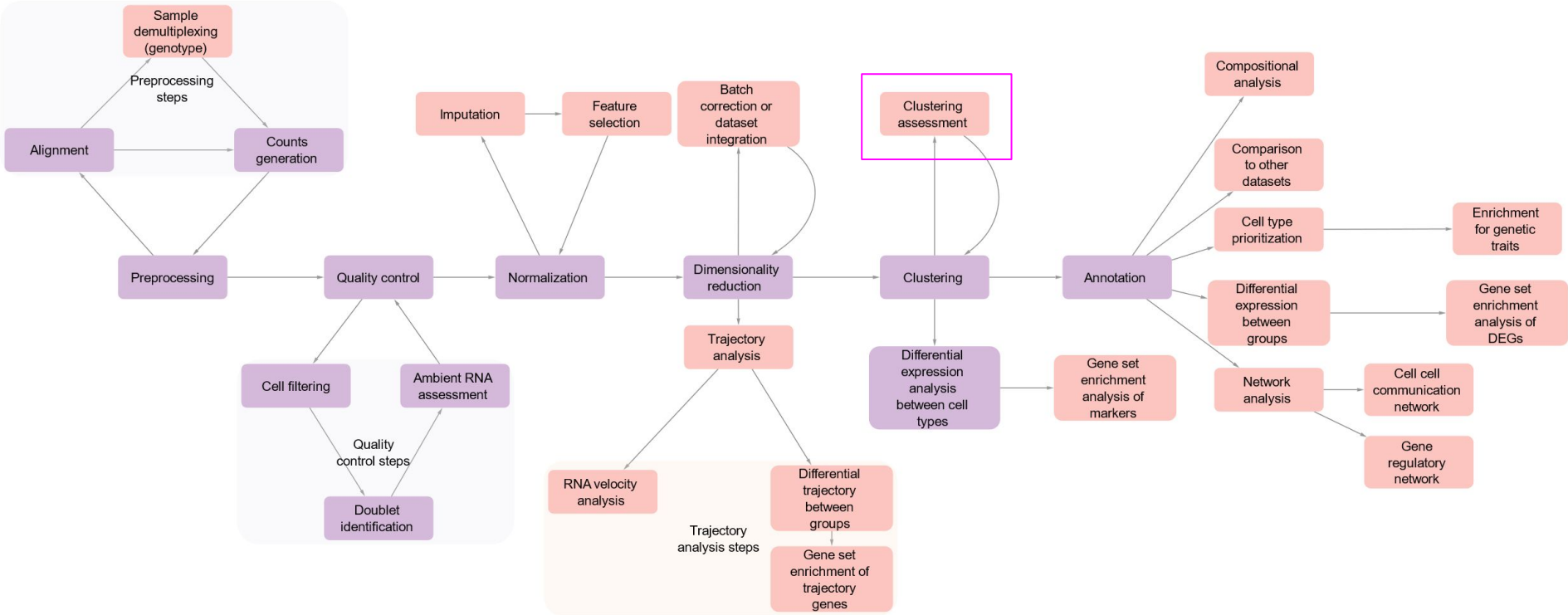


MetaNeighbor (Fishcer et al., 2021)



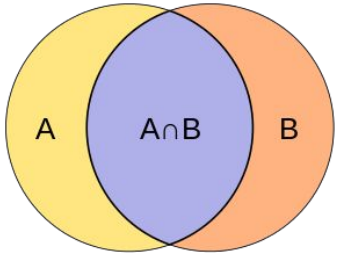
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8826496/

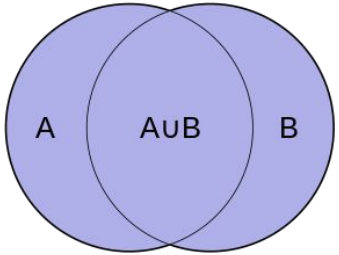
Clustering assessment - revisited



Jaccard Indices and Adjusted Random Index

https://blog.paperspace.com/ml-evaluation-metrics-part-2/





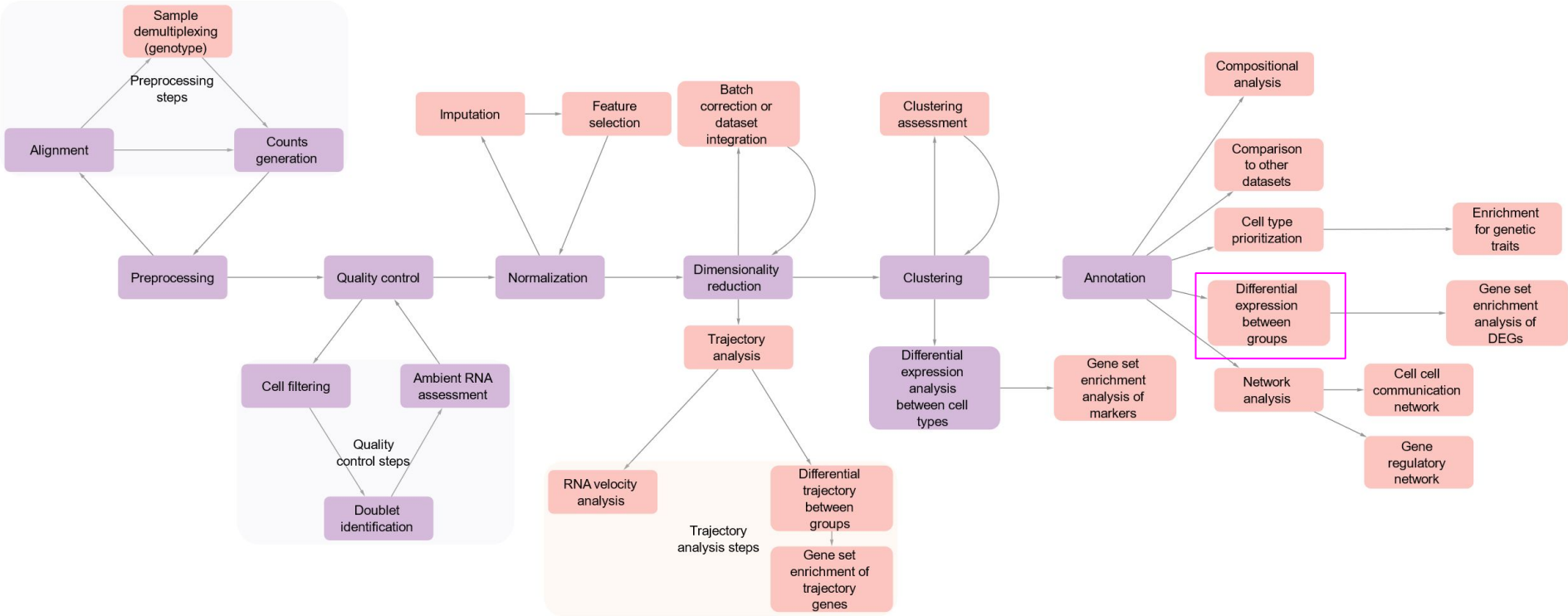
- Jaccard Index compares two clusters from two different clustering labels for the same data
- Random Index and Adjusted Random Index compare all the clusters from two different clustering labels of the same data

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

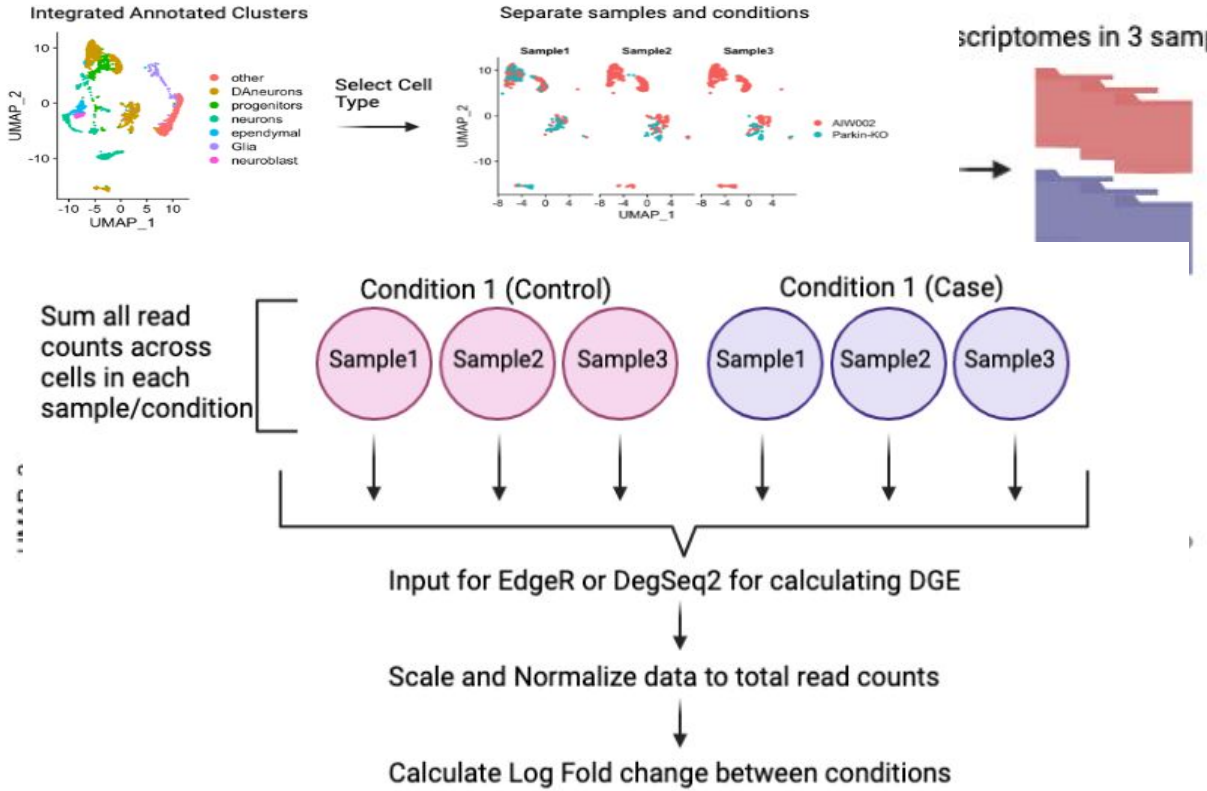
https://en.wikipedia.org/wiki/Jaccard_index

https://en.wikipedia.org/wiki/Rand_index

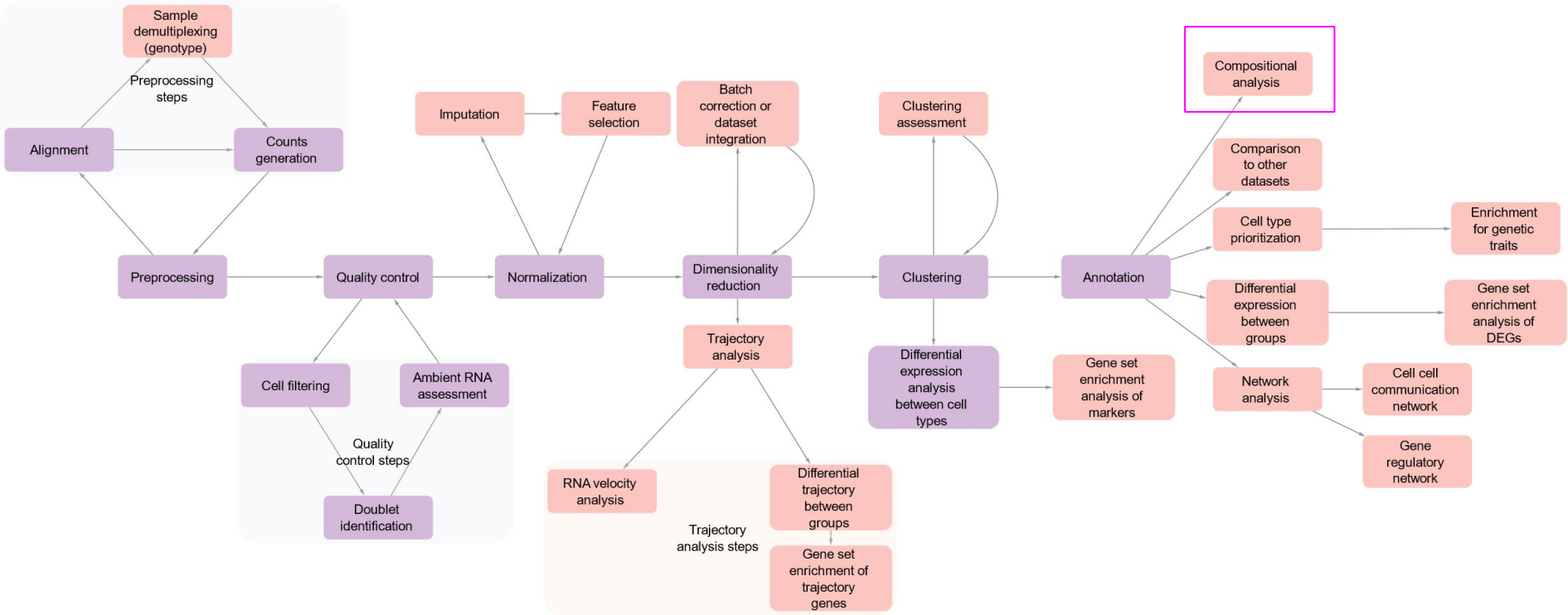
Differential expression between groups (overview)



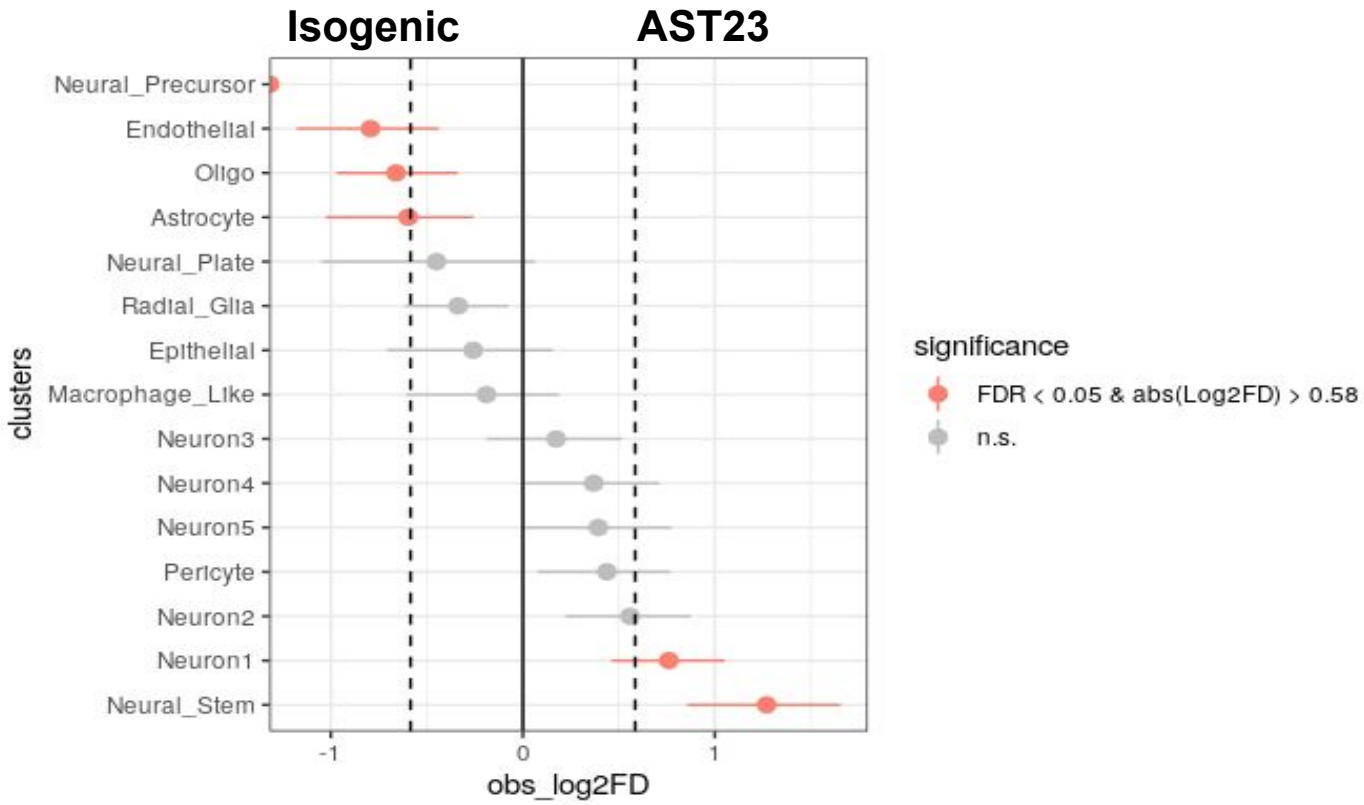
Testing Differential Gene Expression between Groups



Comparison of proportions



Testing proportions of cell types between Groups



Thank you for attending!

Any questions?