Introduction to the single-cell OPEN* group

24/09/2020

OPEN* : CIT, LBC, I. Curie, Inserm, ...

Goal of the single-cell OPEN group

- Analytical steps for single cell omics with consensual best practices

  >> Share know-how : choose the appropriate method for each objective

- Analytical steps for single cell omics without consensual best practices

  >> Identify needs for new developments/benchmarking

  >> Create task forces to address specific problems

- Format
  - Regular meetings
  - Discussion space
  - Know-how resource : Wiki (?)
  - Dev/benchmark resource : Github (?)

Focus of this 1st meeting : single cell RNA-seq

Agenda :

- Gael Blivet (LBC / CIT)

  Overview of scRNA-seq typical analytical steps, with emphasize on known difficulties

- Andrei Zinovyev (I. Curie)

  Methods for trajectory analysis and related limits

- Jing Liu (I. Curie / CIT)

  Illustration of the difficulty of data integration when mixing tumor samples for cell-class discovery

- Aziz Fouché (I. Curie)

  The data integration problem

Overview of scRNA-seq typical analytical steps with emphasize on known difficulties

Gael Blivet - LBC, ESPCI / CIT, LNCC

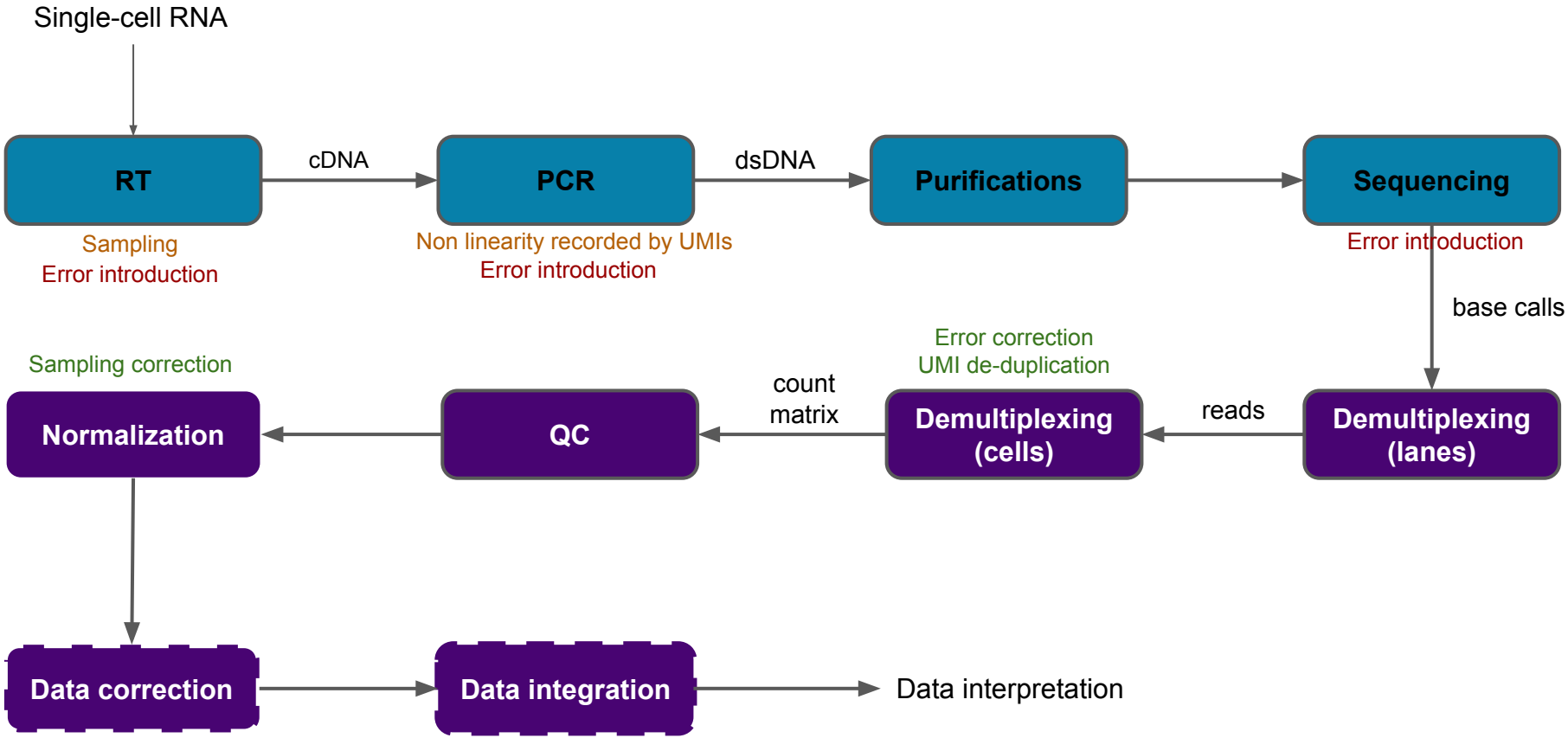
24/09/2020

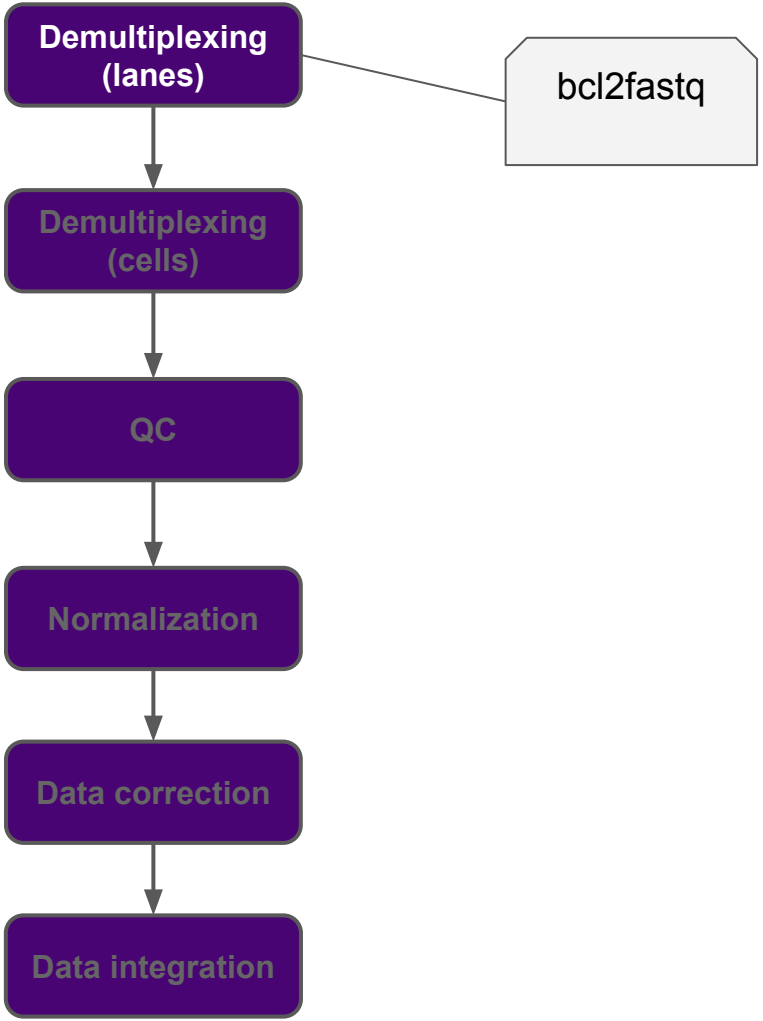
Main resources used

- “Current best practices in single-cell RNA-seq analysis: a tutorial”, Luecken *et al.*, Molecular Systems Biology 2019
- “Orchestrating Single-Cell Analysis with Bioconductor”
  - Focusing on Bioconductor packages (no Seurat)

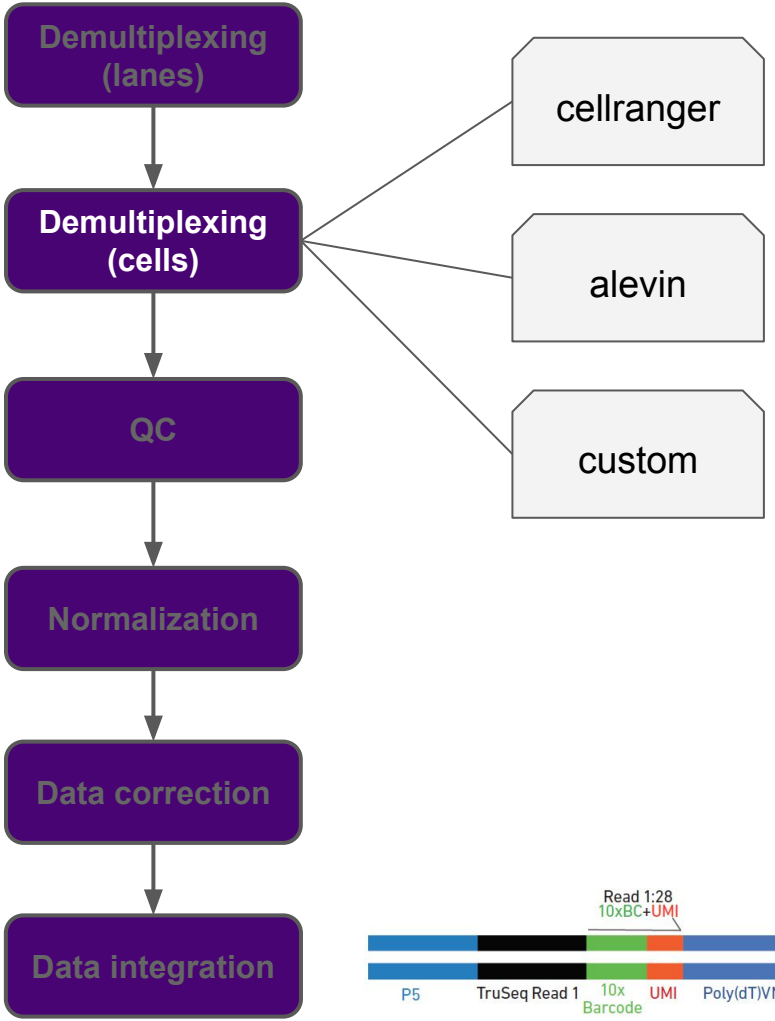
Disclaimer

- Impossible to list all tools
- All recommendations are highly context dependent and have to be rationally examined in the biological context

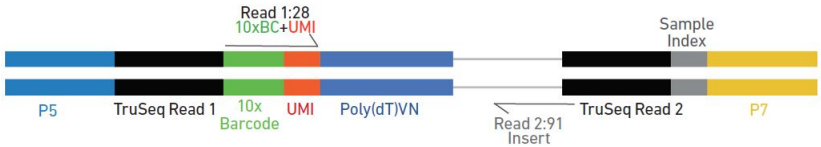

Single-cell RNA
RT
cDNA
PCR
dsDNA
Purifications
Sequencing
Sampling
Error introduction
Non linearity recorded by UMIs
Error introduction
Error introduction
base calls
Sampling correction
Error correction
UMI de-duplication
count
matrix
reads
Normalization
QC
Demultiplexing
(cells)
Demultiplexing
(lanes)
Data correction
Data integration
Data interpretation

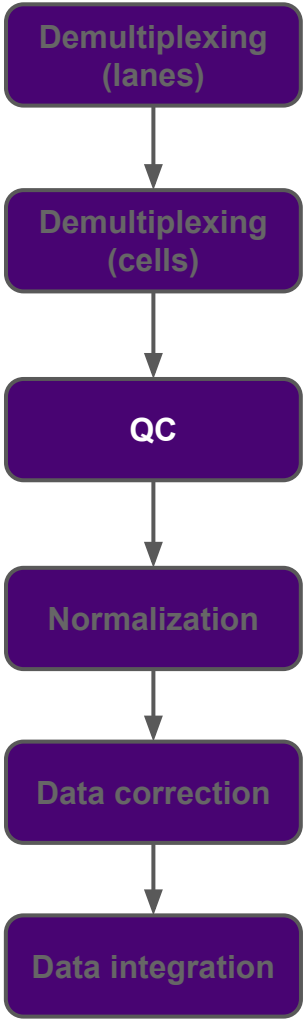


- Task
  - Convert base call files to FASTQ
  - Demultiplex different samples



- Task
  - BC/UMI/insert extraction
  - with error correction
- Platform dedicated
  - [10X] Cellranger (black box)
  - [CEL-seq(2)] scruff
  - [inDrop] inDrop pipeline
- Flexible tools
  - alevin
    - From salmon software (RNA-seq transcript quantification tool)
    - Natively Drop-seq and 10X but can be parametrized for others
  - scater, scPipe, zUMIs, etc.

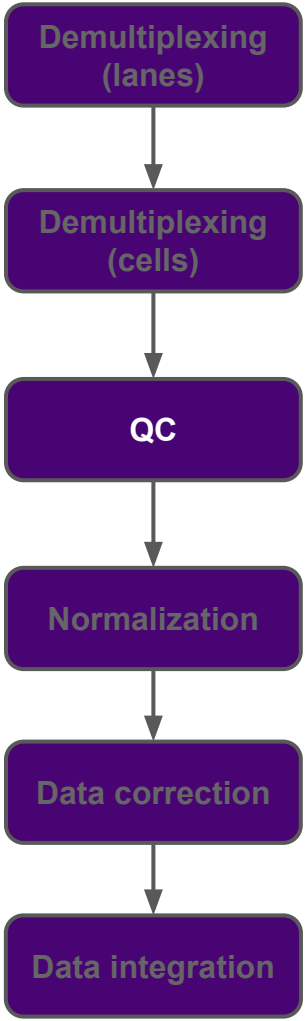




Quality control

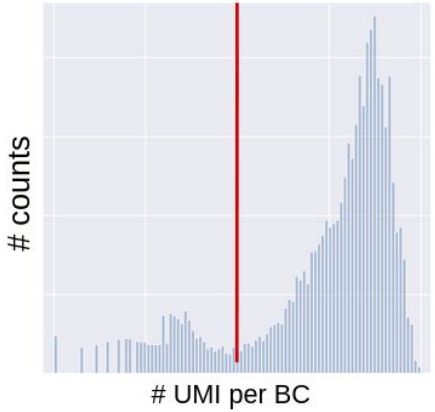
Early stage low-quality data removal to clean downstream analysis

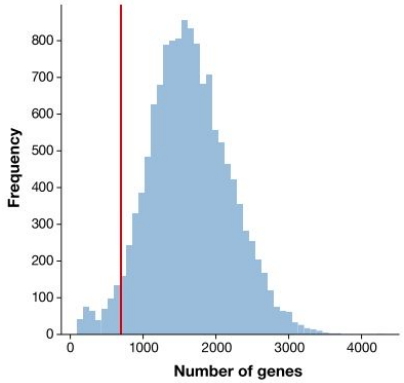
- Warning : “Sufficient” data quality can only be assessed based on downstream analyses …
- Expected: 1 BC = 1 clean cell expression
- To filter out:
  - Empty droplet/well
  - Damaged/perforated/dying cells
  - Failure in lib prep (inefficient RT or PCR)
  - Doublets

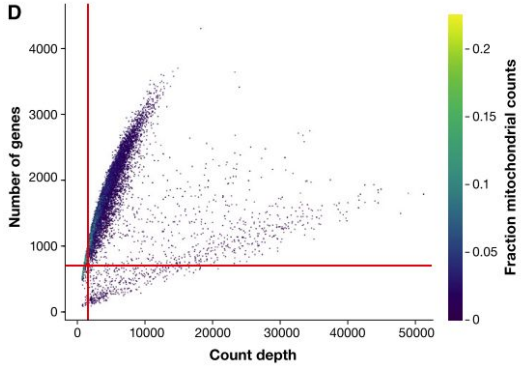


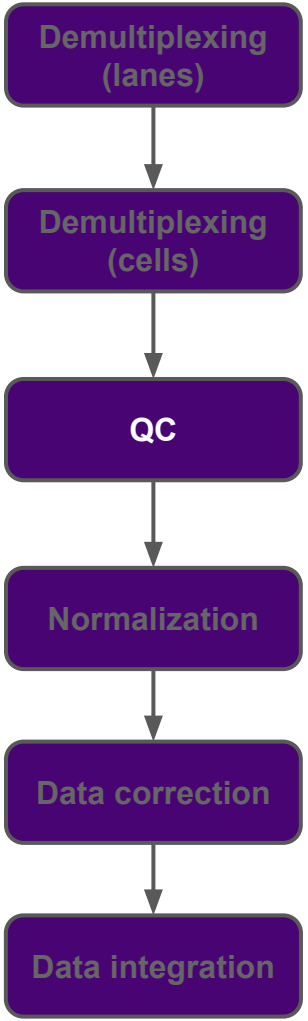
Quality control >> Usual metrics

- Read/UMI count per BC
- Gene count per BC
- Fraction of mitoRNA / ERCC / spike-in RNA



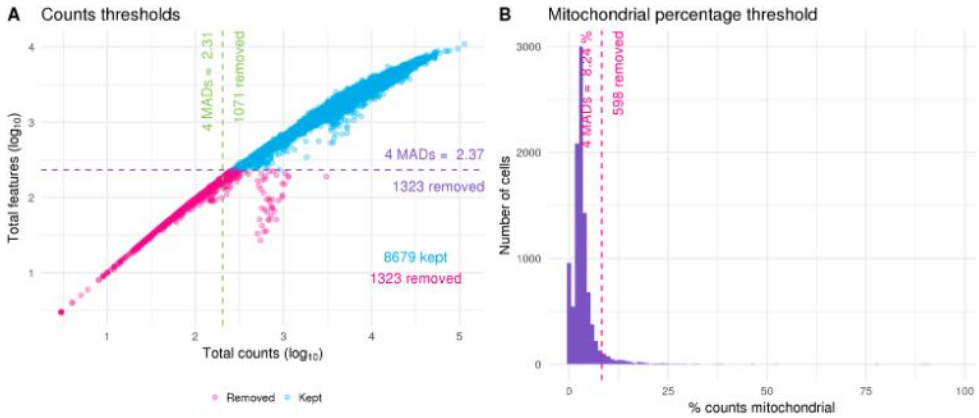


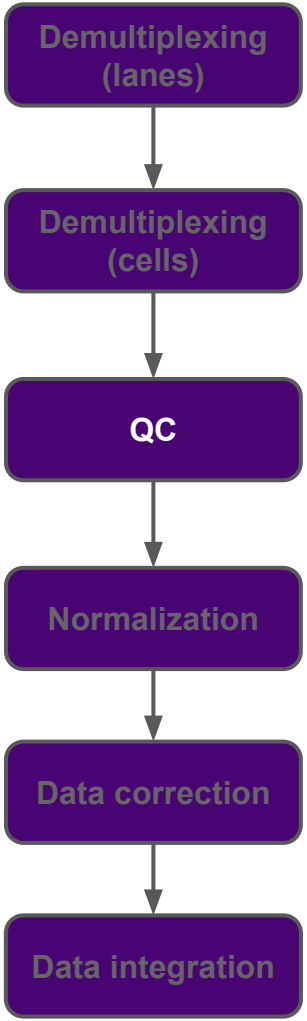




Quality control >> Methods of threshold selection *(whatever the metrics)*

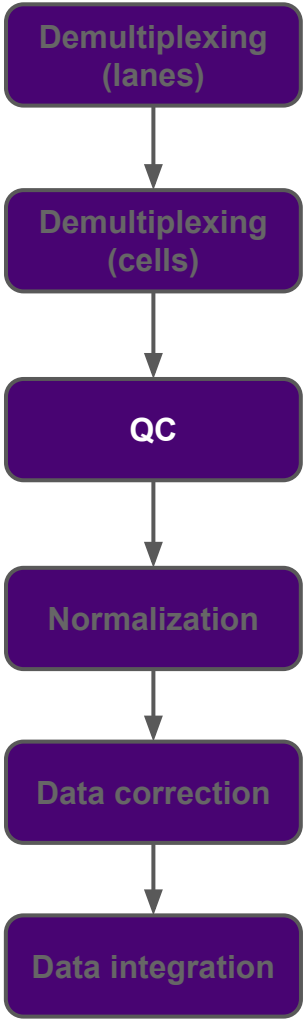
- Visual inspection
- A priori knowledge
- Data-driven: MADs (median absolute deviations)





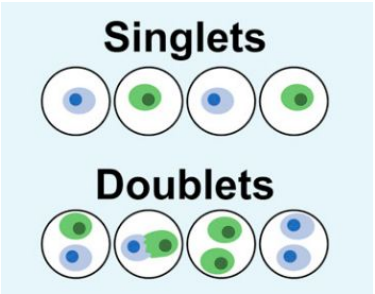
Quality control >> Methods of threshold selection *(whatever the metrics)*

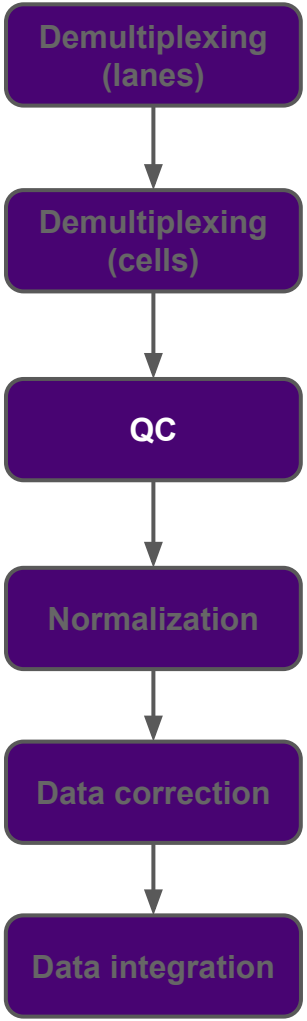
- Other advanced methods less obvious to interpret
  - PCA-based approaches?
  - Support vector machines?



Quality control >> Doublets detection

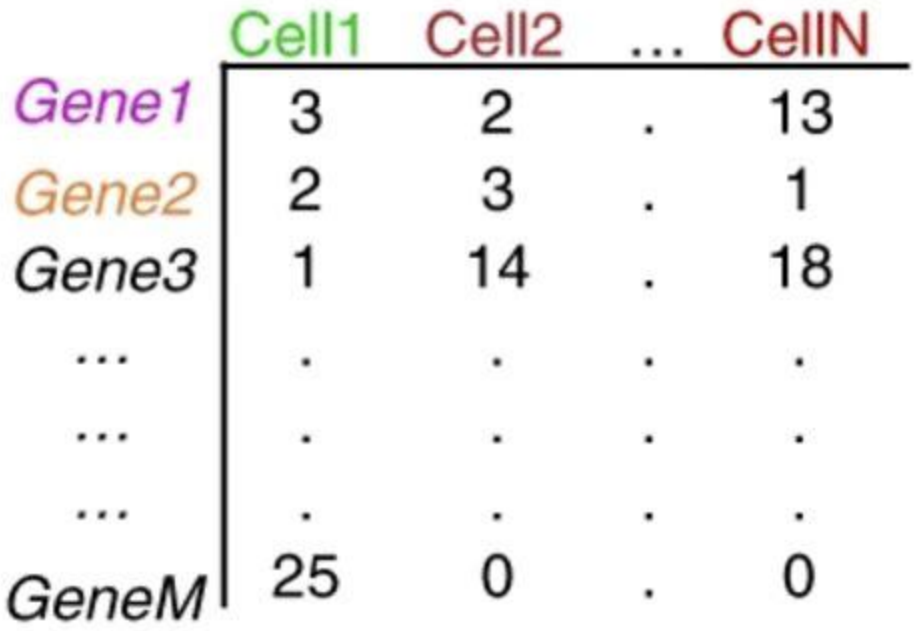
- DoubletDecon: combination of deconvolution analyses
- Scrublet: simulation of multiplets from the data and building a nearest neighbor classifier
- Doublet Finder: actual cell data comparison to artificial pair cell average

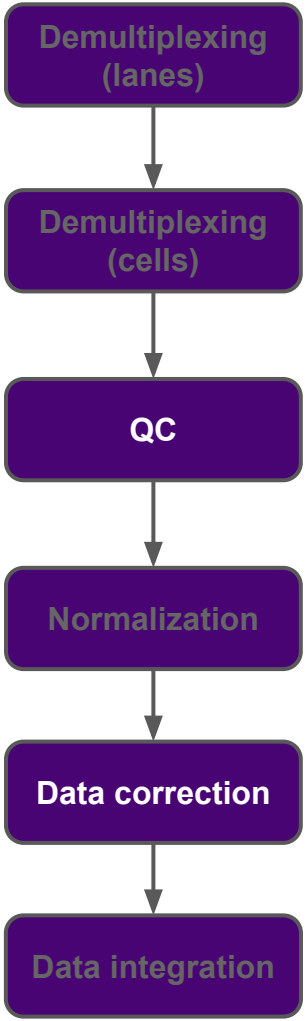




Quality control >> Pitfalls & recommendations

- Consider QC metrics jointly instead of separately.
- Be permissive on QC thresholding and revisit according to downstream clustering interpretability.
- Determine QC thresholds separately for each sample.

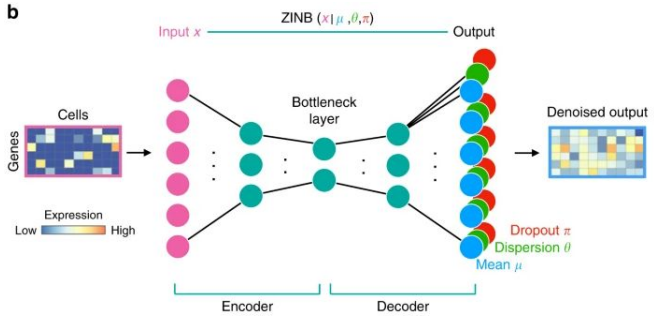

Cell1 Cell2 … CellN
Gene1 3 2 . 13
Gene2 2 3 . 1
Gene3 1 14 . 18
… . . . .
… . . . .
… . . . .
GeneM 25 0 . 0

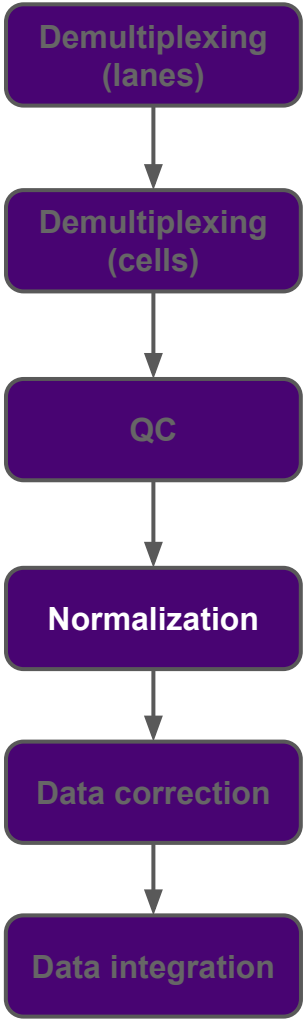


Expression recovery (*vs* filtering)

Fighting dropout through data imputation

- Improves the estimation of gene-gene correlations
- Optional for exploratory analysis
- Approaches:
  - Pool-based size factors (MAGIC, scImpute)
  - Negative binomial model parameter estimations (SAVER)
  - Neural network (DCA, scVI)

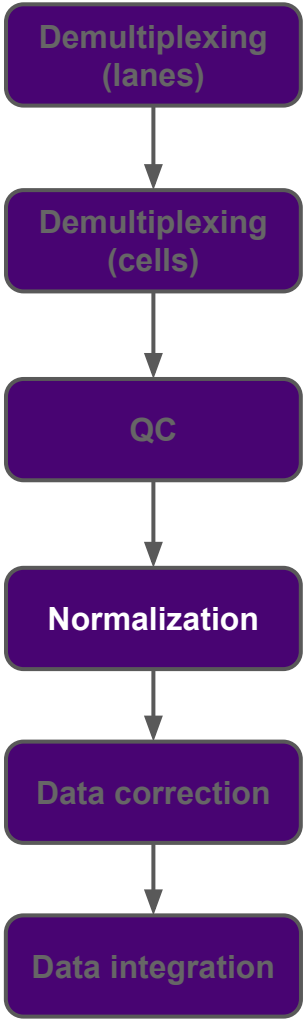




Normalization

Scaling counts to eliminate sampling effect and get relevant relative gene expression between cells

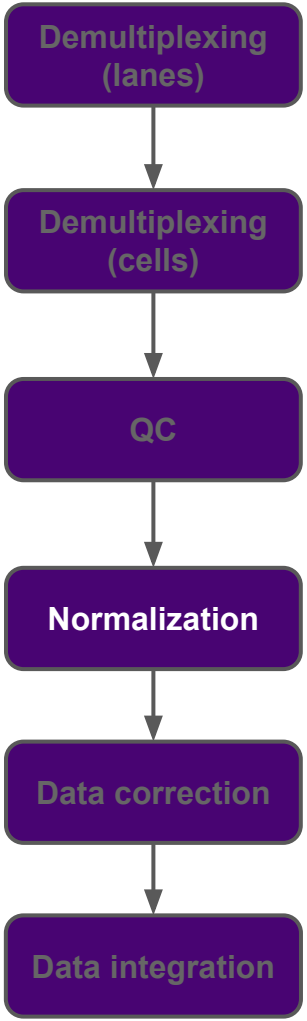
- Normalization ≠ batch correction
- “Normalization is overall the most influential step” (Vieth et al. 2019)
- The best normalization method is dataset dependent
  - scone tool assess efficacy of various normalization methods



Normalization

Methods

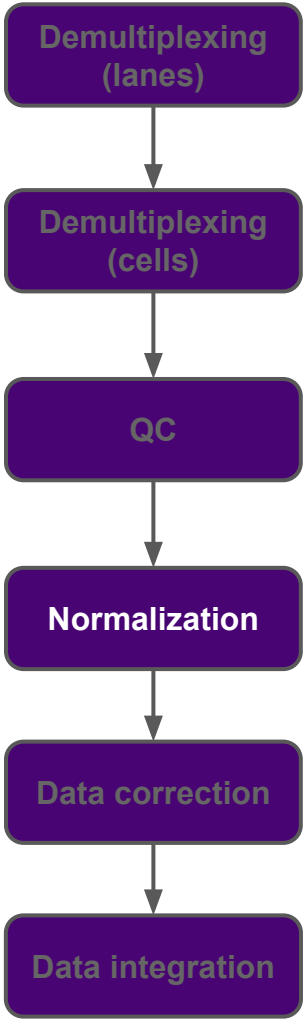
- Identify a library (= 1 cell) scaling factor
  - Library size normalization
  - Spike-ins (preserves cell total RNA differences)
- Downsampling
  - To overcome effect and library size strong correlation



Normalization

Normalization: log(x+1) transform

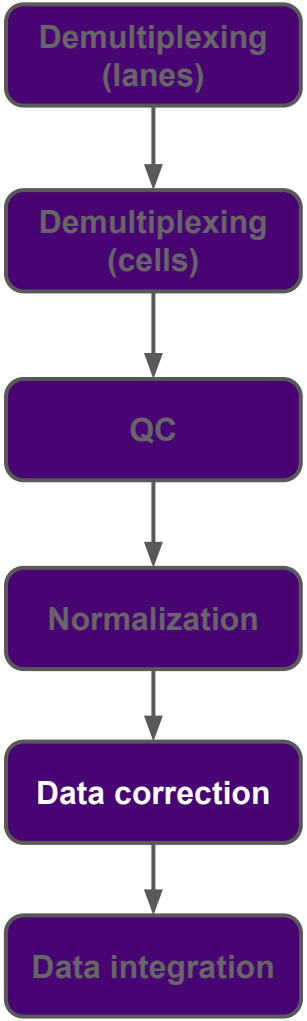
- Distances between expression values are log fold changes (“10 vs 50” > “1000 vs 1100”)
- Mitigates the mean–variance relationship
- Reduces data skewness



Normalization

Gene normalization

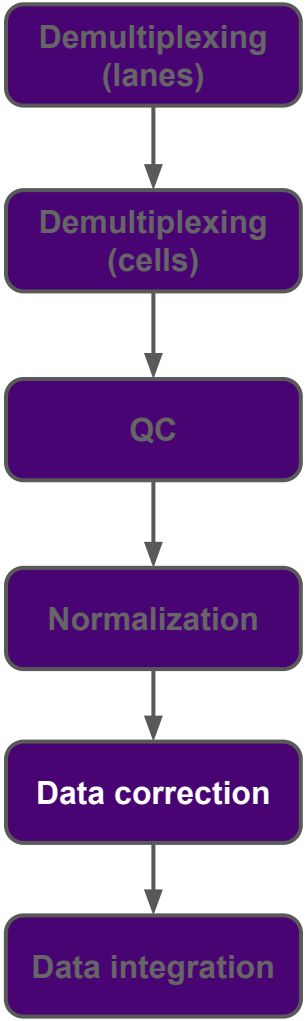
- Scaling gene counts to have zero mean and unit variance
- No consensus on this question
- When all genes should be weighted equally for downstream analysis and the magnitude of expression is not of interest



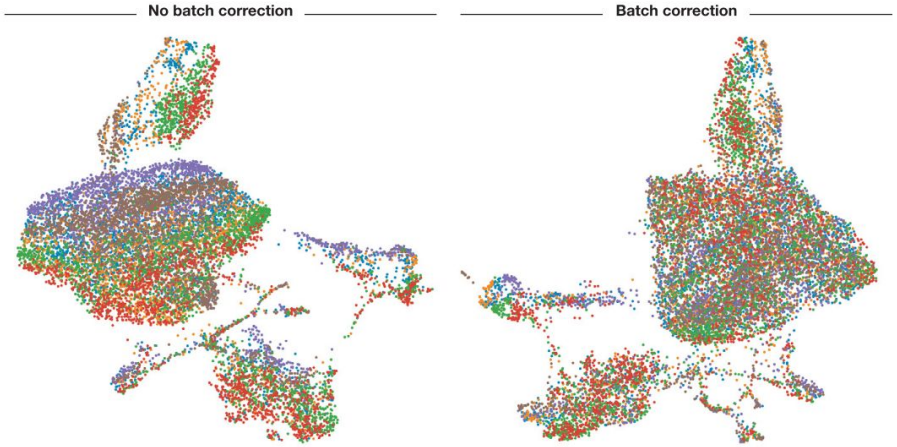
Data correction

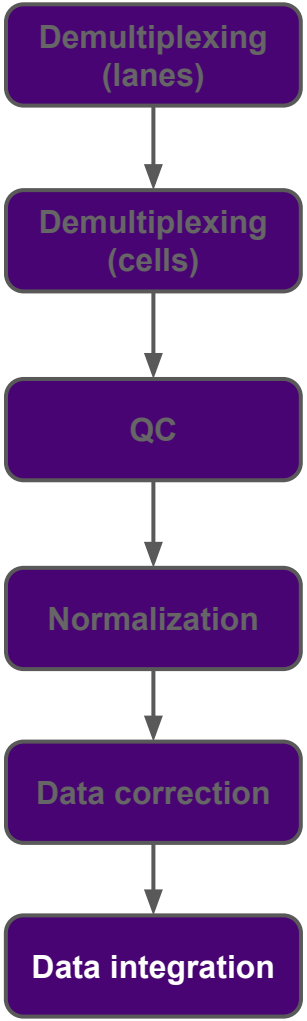
Remove biological and technical effects

- The effects of the cell cycle (recommended for trajectory inference)
  - Simple linear regression against a cell cycle score (Scanpy, Seurat, etc.)
  - More complex mixture model (f-scLVM)
  - Warning: variation in cell size accounts for the transcriptomics effects generally attributed to the cell cycle thus normalization can partially correct this



- Batch correction between samples/cells in a same experiment
- Typically solved by linear approaches
- Tool: ComBat takes into account both mean and variance of the data

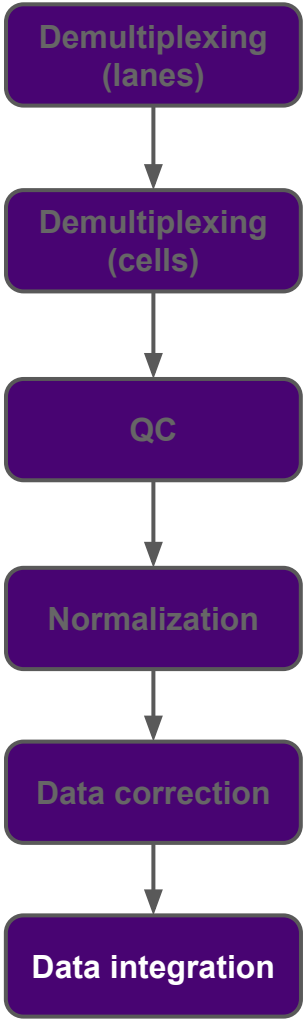




Data integration

Combine datasets from different origins

- Non-linear approaches
- Tools:
  - CCA (Butler *et al.* 2018)
  - MNN (Haghverdi *et al.* 2018)
  - Scanorama (Hie *et al.* 2018)
  - RISC (Liu *et al.* 2018)
  - etc.
- Next talks topic



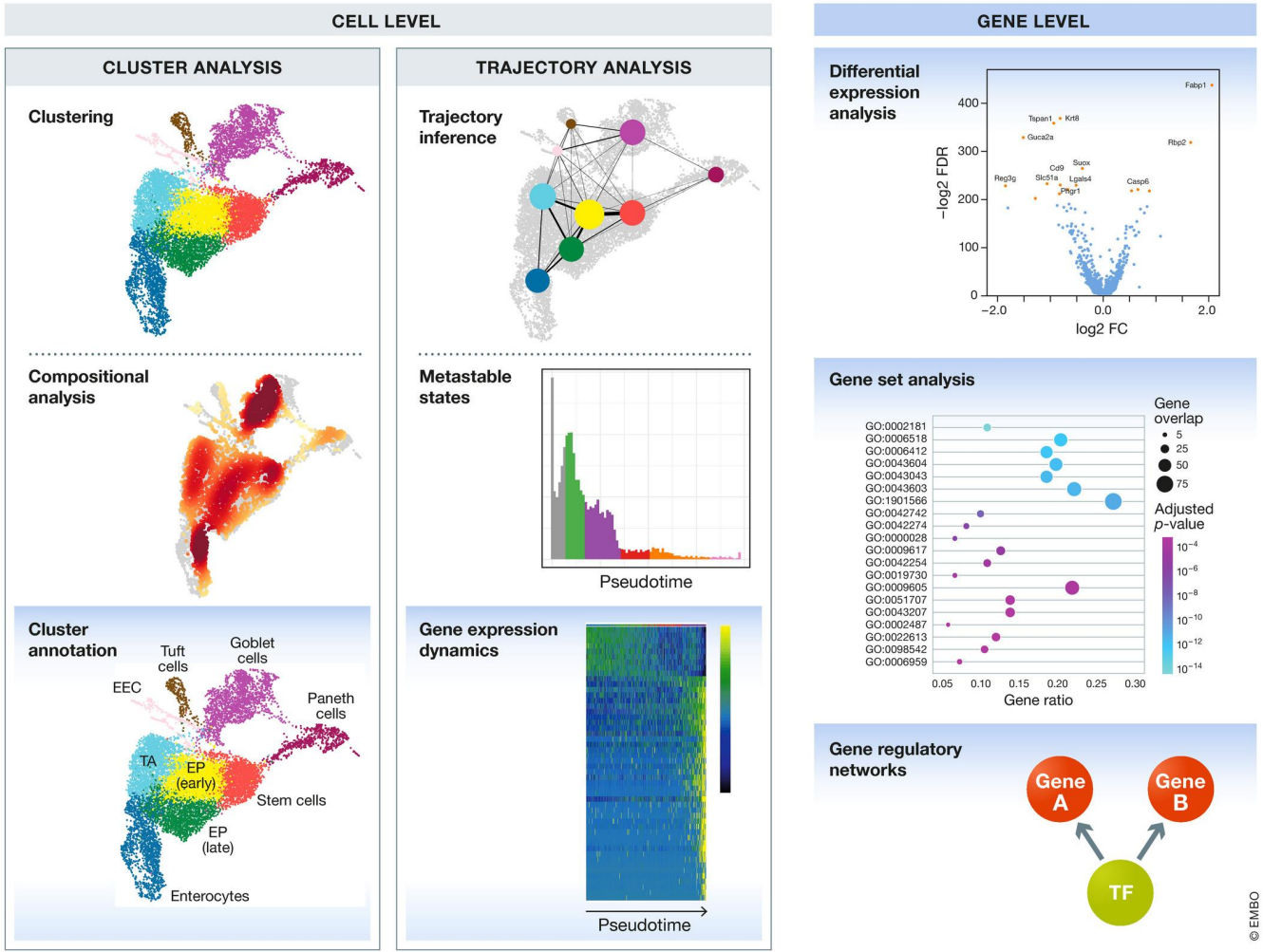
Data correction and integration

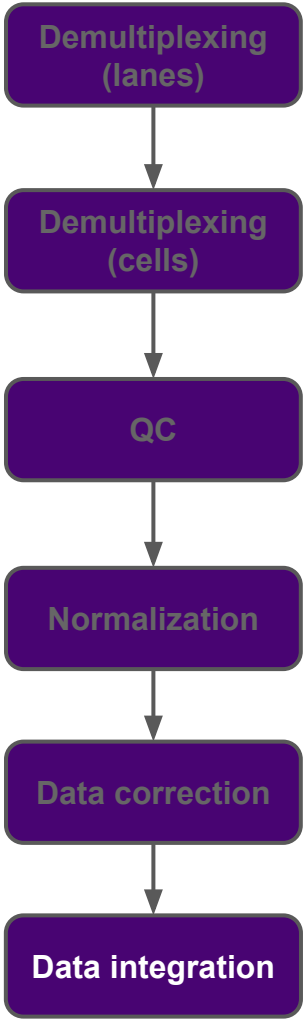
Pitfalls & recommendations

- Regress out both biological and technical jointly
- Check expected input data (normalized vs raw)
- Batch correction via ComBat if cell type and state compositions between batches are consistent
- Correct biological effect for trajectory inference

- Data integration and batch correction should be performed by different methods.
- Data integration may overcorrect batch effects

(More) Specific steps

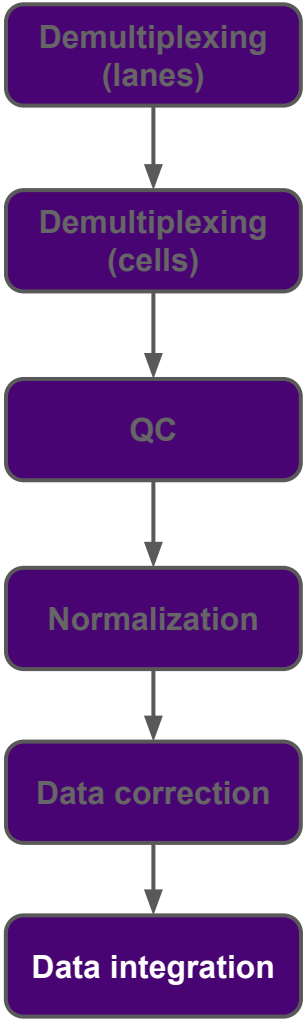

CELL LEVEL
CLUSTER ANALYSIS
Clustering
Compositional analysis
Cluster annotation
Tuft cells
Goblet cells
EEC
Paneth cells
TA
EP (early)
Stem cells
EP (late)
Enterocytes
TRAJECTORY ANALYSIS
Trajectory inference
Metastable states
Pseudotime
Gene expression dynamics
Pseudotime
GENE LEVEL
Differential expression analysis
−log2 FDR
log2 FC
Fabp1
Tspan1
Krt8
Guca2a
Rbp2
Cd9
Sucx
Reg3g
Slc51a
Lgals4
Casp6
Pfkpr1
Gene set analysis
GO:0002181
GO:0006518
GO:0006412
GO:0043604
GO:0043043
GO:0043603
GO:1901566
GO:0042742
GO:0042274
GO:0000028
GO:0009617
GO:0042254
GO:0019730
GO:0009605
GO:0051707
GO:0043207
GO:0002487
GO:0022613
GO:0098542
GO:0006959
Gene ratio
Gene overlap
Adjusted p-value
Gene regulatory networks
Gene A
Gene B
TF



Feature selection

Select relevant genes and ease computation

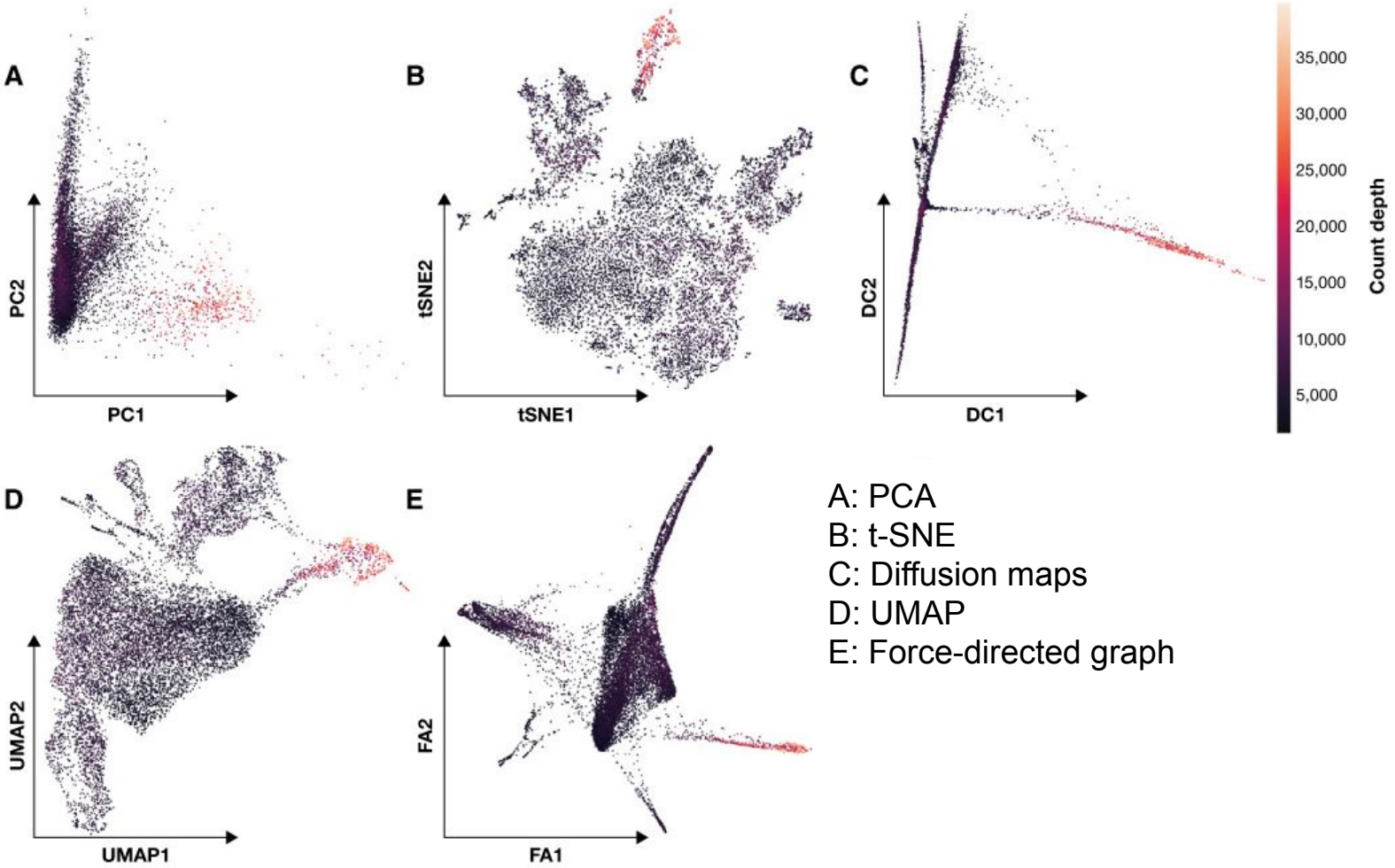
- Assumption: expression profile of most gene is dominated by technical noise
- Methods
  - Highly Variable Genes (HGVs, n = 1-5k)
  - *a priori* genes of interest
  - “All above the trend”
- Methods that use gene expression means and variances cannot be used when gene expression values have been normalized to zero mean and unit variance, or when residuals from model fitting are used as normalized expression values.



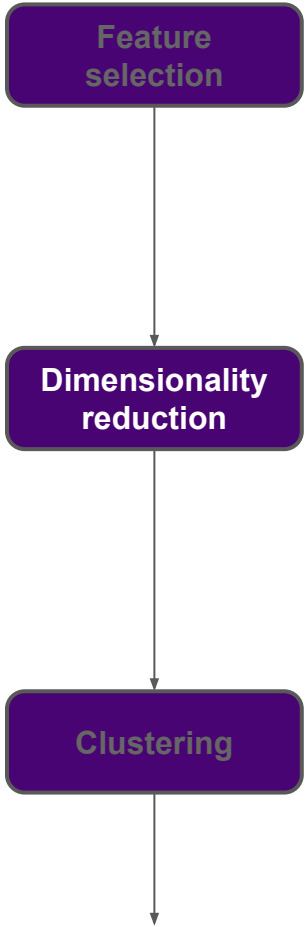
Dimensionality reduction

Describe expression profiles with few dimensions and ease computation

- Target
  - Visualisation (2-3 components)
  - Summarization
- Method choice, what matters?
  - Distances (visual interpretability)
  - Local similarity vs. global structure
  - Computing time



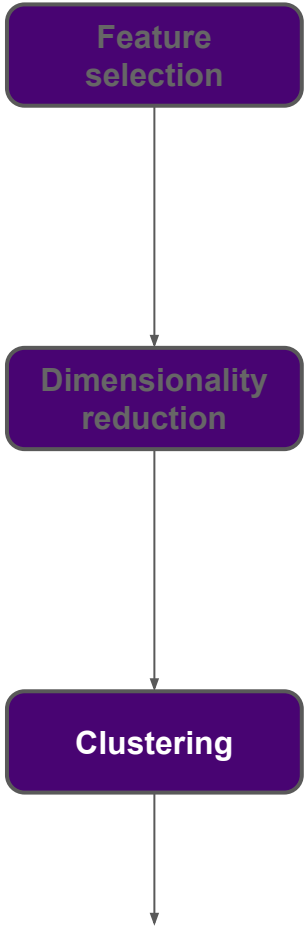
A: PCA
B: t-SNE
C: Diffusion maps
D: UMAP
E: Force-directed graph



Dimensionality reduction

Pitfalls & recommendations

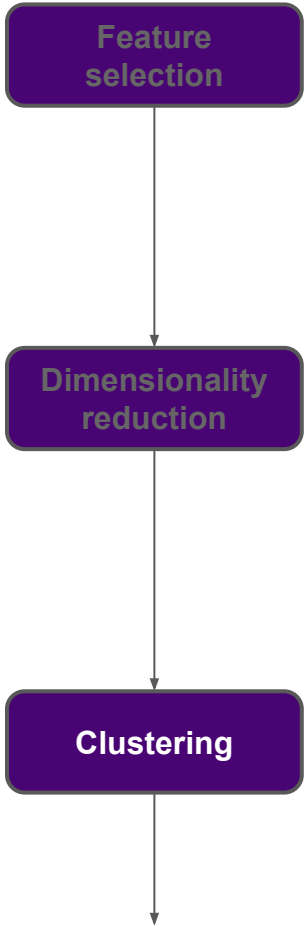
- Dimensionality reduction methods should be considered separately for summarization and visualization.
- UMAP for exploratory visualization; PCA for general purpose summarization; and diffusion maps as an alternative to PCA for trajectory inference summarization.



Clustering

Define discrete groups of cells with similar expression profiles to enable interpretation

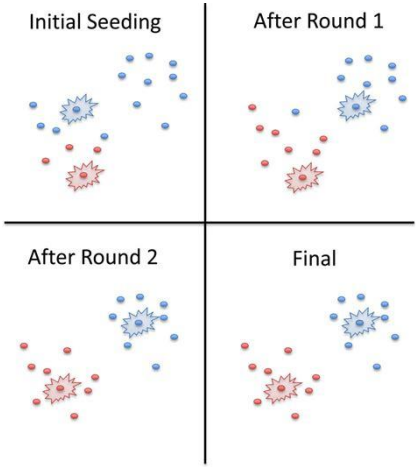
- ~~What is the true number of clusters?~~
- How well do the cluster approximate the cell types/states?
- 3 approaches
  - Community detection / Graph-based
  - Clustering algorithms
  - Mixed



Clustering

Clustering algorithms

- Classical unsupervised learning problem based on a distance matrix
- Most famous: k-means (recommended with correlation-based distances)



Feature selection

Dimensionality reduction

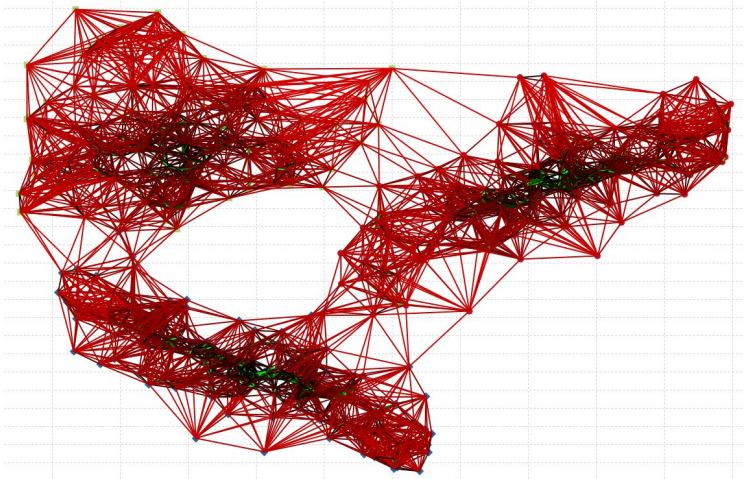
Clustering

Advanced analyses

Clustering

Community detection / Graph-based

- Graph-partitioning method
- Most famous: KNN



Feature selection

Dimensionality reduction

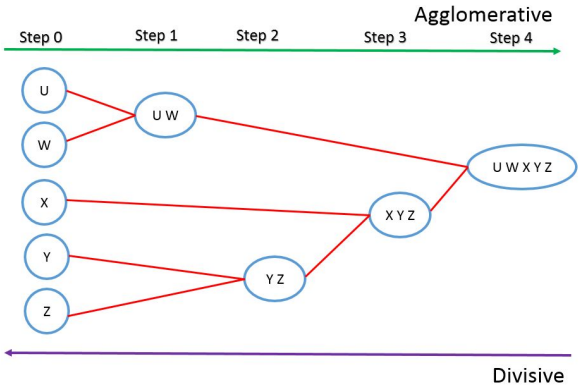
Clustering

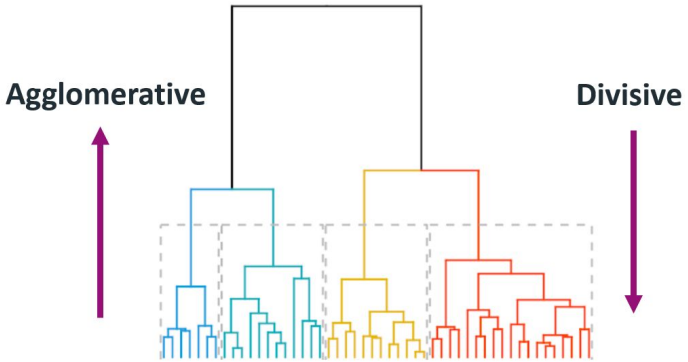
Advanced analyses

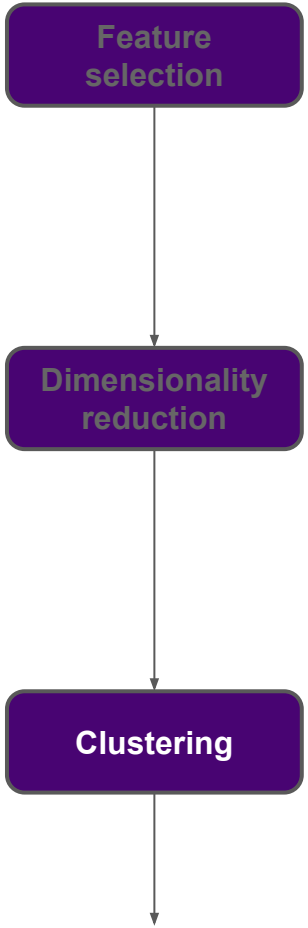
Clustering

Mixed strategy

1. K-means with an inflated *k*
2. Hierarchical clustering on small clusters



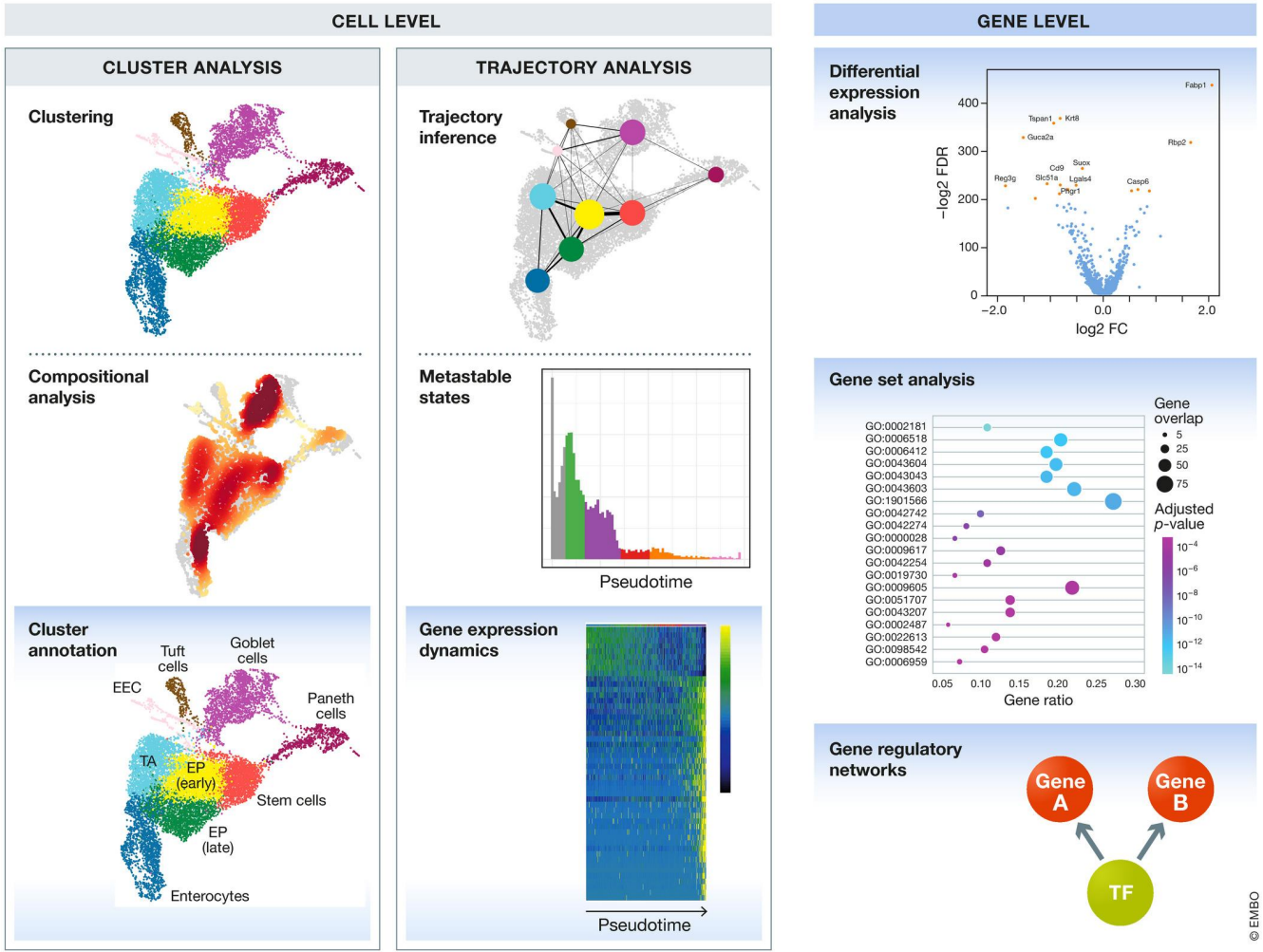




Dimensionality reduction

Pitfalls & recommendations

- Louvain algorithm on single-cell KNN graph (default method in Scanpy and Seurat)
- Can be performed at different resolutions to focus on particular substructures
- Clustering stability: small upstream changes should have little impact on conclusions


CELL LEVEL
CLUSTER ANALYSIS
Clustering
Compositional analysis
Cluster annotation
Tuft cells
Goblet cells
EEC
Paneth cells
TA
EP (early)
Stem cells
EP (late)
Enterocytes
TRAJECTORY ANALYSIS
Trajectory inference
Metastable states
Pseudotime
Gene expression dynamics
Pseudotime
GENE LEVEL
Differential expression analysis
−log2 FDR
log2 FC
Fabp1
Tspan1
Krt8
Guca2a
Rbp2
Cd9
Sucx
Reg3g
Slc51a
Lgals4
Casp6
Pfdr1
Gene set analysis
GO:0002181
GO:0006518
GO:0006412
GO:0043604
GO:0043043
GO:0043603
GO:1901566
GO:0042742
GO:0042274
GO:0000028
GO:0009617
GO:0042254
GO:0019730
GO:0009605
GO:0051707
GO:0043207
GO:0002487
GO:0022613
GO:0098542
GO:0006959
Gene ratio
Gene overlap
Adjusted p-value
Gene regulatory networks
Gene A
Gene B
TF

Thanks for your attention