Analysis of scRNA-seq using R

Annamaria Carissimo, Luisa Cutillo
Istituto per le Applicazioni del Calcolo “Mauro Picone”, CNR, Naples, Italy.
School of Mathematics, University of Leeds

Bioinformatics Awareness Day (BAD) – Single Cell RNAseq Hackathon (12th May 2022)

Outline and organization

- Introduction to scRNA-seq
- Overview on scRNA-seq data analysis
- Step-by-step data guided example

Annamaria Carissimo
Luisa Cutillo
Valeria Policastro

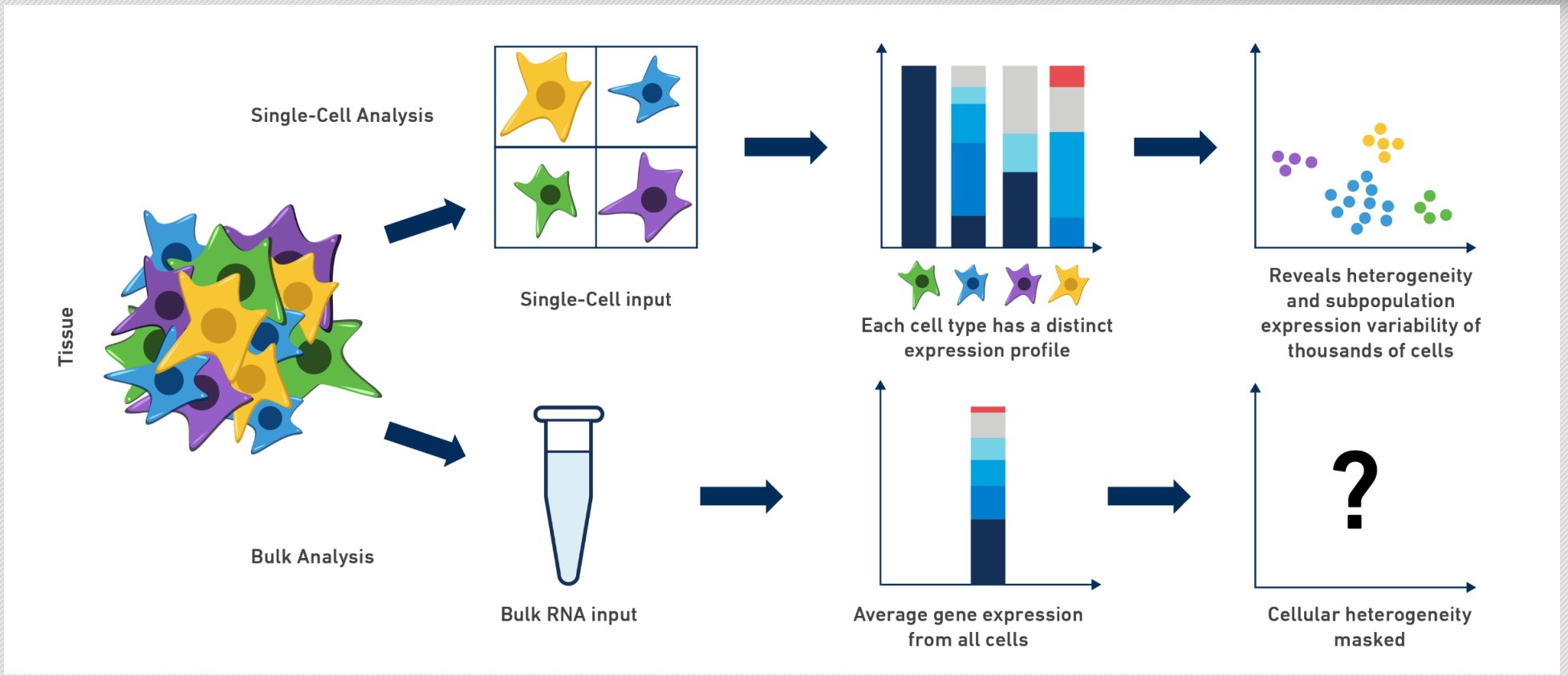
- Models for trajectories and branching identifications
- Monocle pipeline

Davide Cacchiarelli
Francesco Panariello

- Try it yourself
- Present your results
- Let’s have a joint discussion and question time
- Some advanced applications

Introduction to scRNA-seq

Bulk RNA-Seq vs scRNA-Seq



scRNA-seq could play a key role in personalized medicine by facilitating characterization of cells, pathways, and genes associated with human diseases such as cancer. Sci Transl Med. 2017

What can I investigate with scRNA-seq?

Most relevant scRNA-seq **applications** include

- Study cellular **heterogeneity**
- Discover n**ovel** cell populations
- Predict **cell fate** differentiation

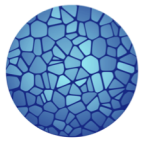
By-cell analysis

- Detect cell-type specific **differentially** expressed genes
- Understand cell-type specific **regulatory** networks

By-gene analysis

Other applications include

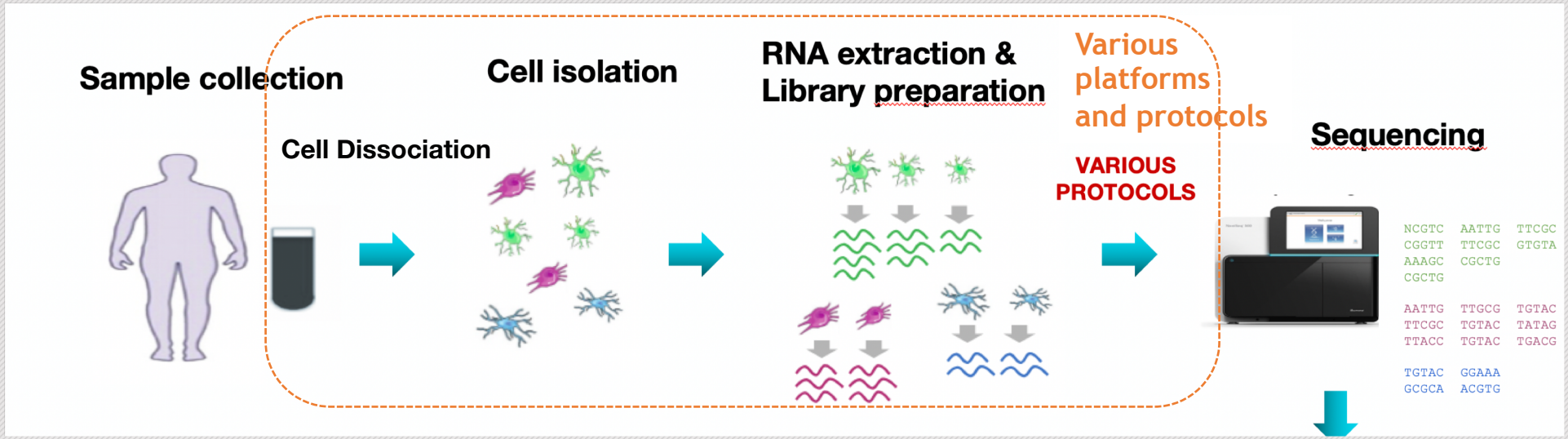
- **Cell Atlas**: Human Cell Atlas, Mouse Cell Atlas,….



HUMAN CELL ATLAS

To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease.

ScRNA-seq overview



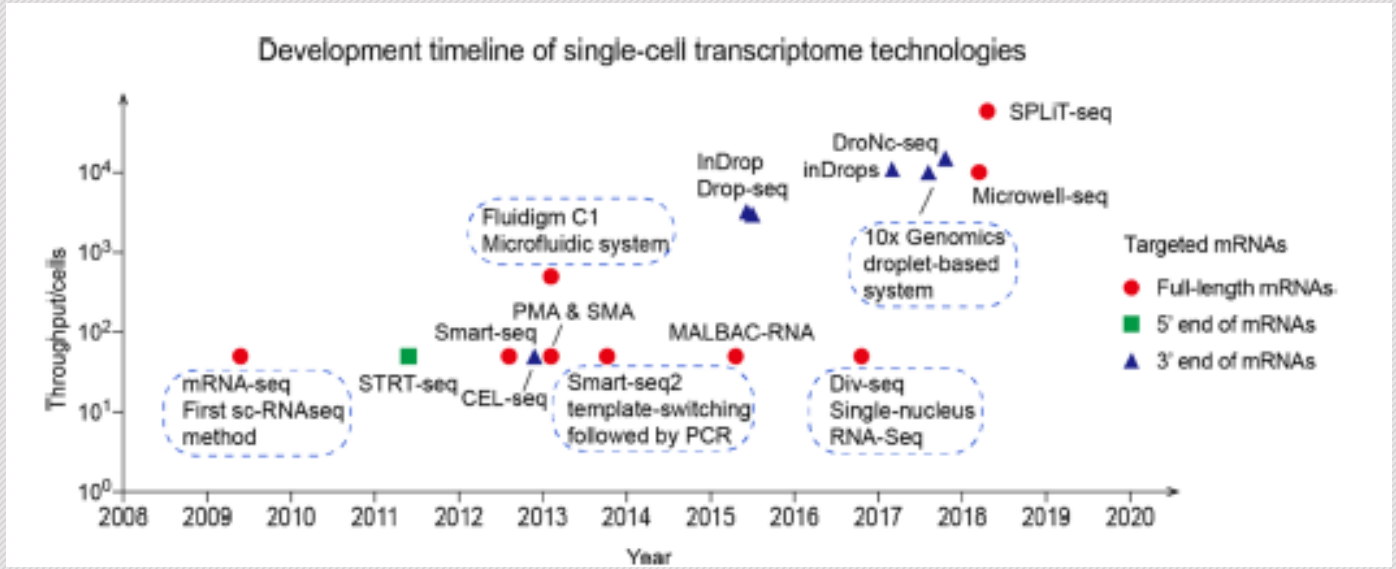
Wet experimental phase

Platforms isolate cells, extract mRNA, and prepare libraries for sequencing

Protocols refer to the specific chemistries used to prepare sequencing libraries

platforms may be protocol-specific (proprietary) or allow multiple protocols (open-source)→**platform+protocols =systems**

ScRNA-seq experimental systems



Zhang et al, 2022

Systems differ with respect to the tecnology and chemistry used to dissociate, isolate cells, extract RNA and prepare the library.

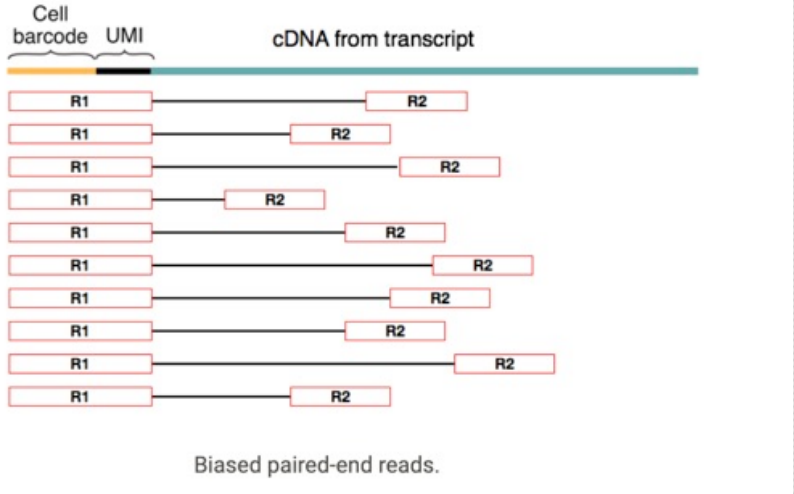
Different systems:

- Low vs high cell throughput
- Read count vs UMI
- Full lenght vs 3'/5' coverage
- Sequencing depth and accuracy
- RNA capturing efficiency
- Multiplexing
- Others….

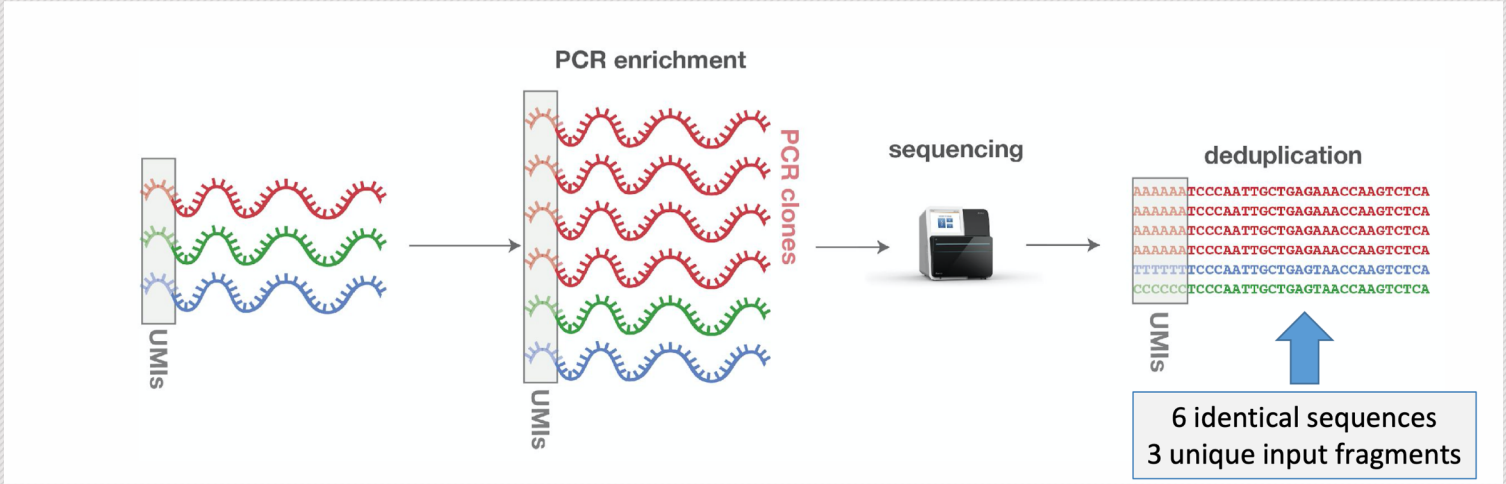
Nowadays, droplets technology allows to process up to tens of thousands of cells together.

Unique Molecular Identifiers (UMIs)

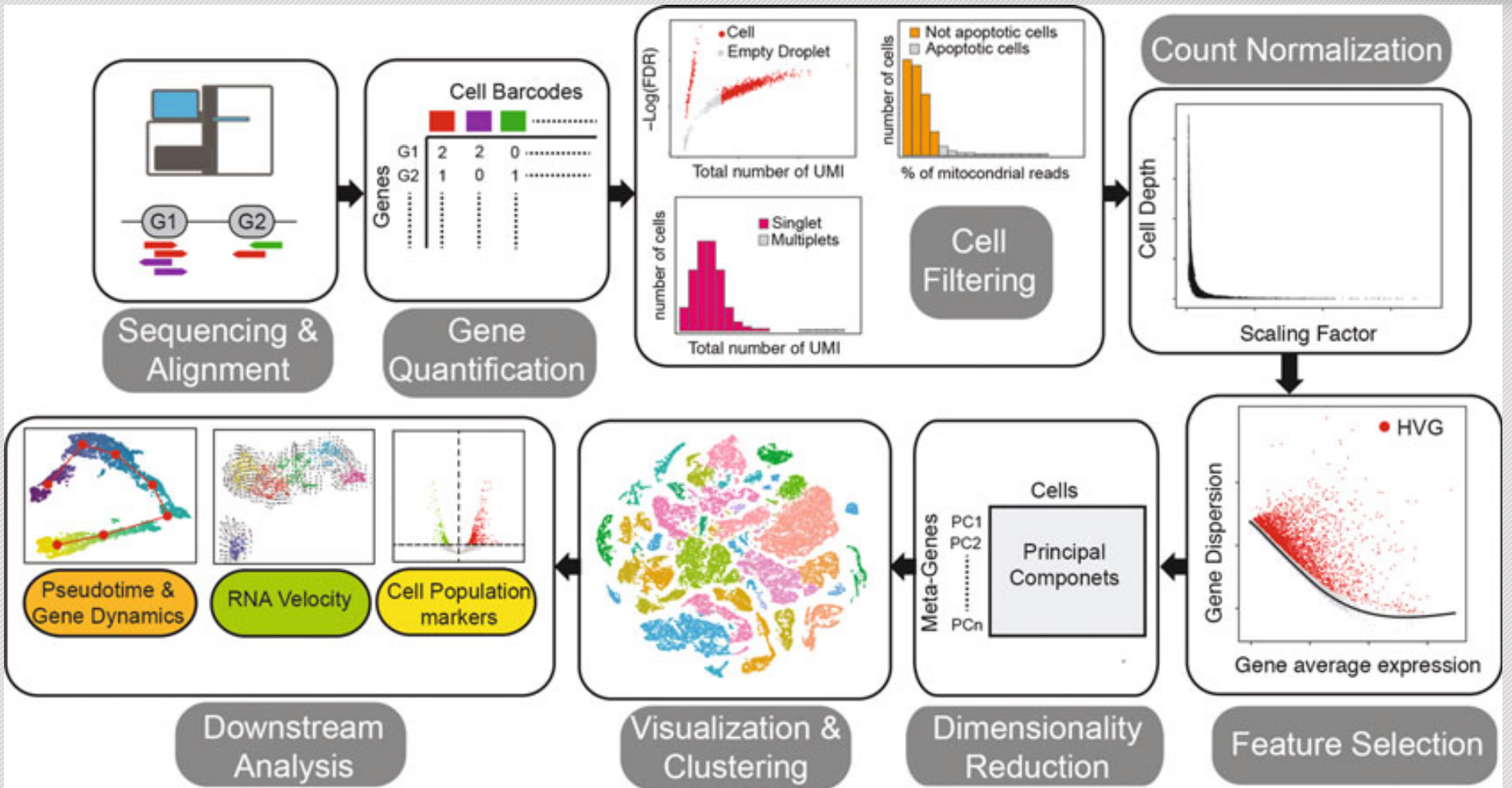
- PCR introduces **nonlinear** amplification bias
- **UMIs** are a way to tag each unique molecule in the sequencing library (before PCR)
- Afterward, sum up only distinct UMIs (collapse reads)



Biased paired-end reads.



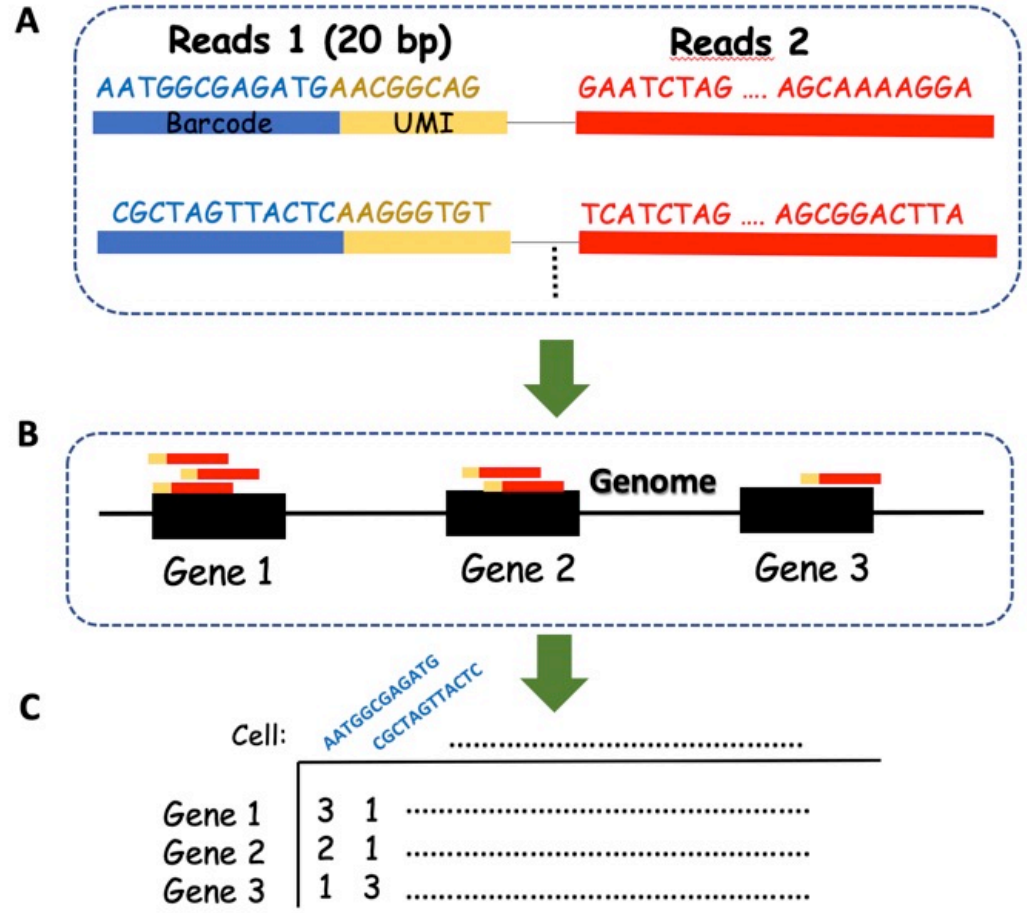
scRNA-seq data analysis overview



Slovin et al, 2021

From sequences to raw counts

- Inspect sequence quality (**FastQC**)
- Remove or trim low quality sequence (**TrimGalore, cutadapt**)
- Demultiplexing → Group reads by cell barcode
- Alignment - (**STAR**), or pseudo-aligner (**Kallisto** or **Salmon**)
- Quantify reads per transcript/gene with tools widely used for bulk (**Rsubread, RSEM, HTSeq**) or specialized tools (collapse UMIs and demultiplex)
- If using UMI → collapse UMIs to count on gene/transcripts (**Umi-tools**)



The '*count*' data: Reads counts vs UMI counts

Usually we only retain **uniquely mapped reads**, thus reads mapping on multiple areas of the genome are discarded. Multiple mapping reads can be "*rescued*"

Read counts

	cell 1	cell 2	cell 3	...	cell M
gene 1	0	0	0		0
gene 2	20	22	1		5
gene 3	90	26	10		10
...					
gene N	5	5	1		5

Regardless the count type, scRNA-seq data are **extremely sparse**, i.e., there is a high proportion of **zero read counts**.
This '***zero-inflation***' arises for both **biological** reasons and **technical** reasons.

UMI counts

	cell 1	cell 2	cell 3	...	cell M
gene 1	0	0	0		0
gene 2	10	5	1		2
gene 3	27	10	3		3
...					
gene N	3	2	1		0

- Larger counts
- Include potential PCR artefacts
- Large extra Poisson variability
- Both full-length and tag based (3'-5')

- Smaller counts
- UMIs reduce extra Poisson variability
- No PCR artefacts
- Only tag based (3'-5')

Quality control and filtering

Quality control at **cell level**

- Removing dying cells or with broken membrane
- Removing doublets, empty drops/wells

Quality control at **gene/transcript level**

- Removing not expressed genes/transcripts when dealing with dropouts
- Removing genes detected in a small number of cells

Some QC-metrics

- % of uniquely mapped reads
- Total counts per cells/barcode
- Number of expressed genes per cells/barcode
- Fraction of counts from mitocondrial genes per cells/barcode
- Spike-in detection – ratio, if availble

- Count per gene
- Some specific gene category that do not contribute to cell variation

FILTERING
low quality cells

	cell1	cell2	cell3
gene A	18	28	3
gene B	6	140	0
gene C	180	35	0
gene D	0	0	2

FILTERING
lowly expressed genes

	cell1	cell2	cell3
gene A	18	28	3
gene B	6	140	0
gene C	180	35	0
gene D	0	0	2

Tutorial

Dropouts: Impute or Not?

scRNA-seq data are **extremely sparse** due to **dropouts**

The **zeros** arise for different reasons:

- The gene was not expressed in the cell → there are no transcripts to sequence. **It is a true zero.**
- The gene was expressed, but for some reason (trancriptional bursting or capturing efficiency) the transcripts were lost somewhere prior to sequencing. **It is a dropouts zero.**
- The gene was expressed and transcripts were captured, but the sequencing depth was not sufficient to produce any reads. **It is a dropouts zero.**

Data imputation aims to replace zero-abundance values with expected values under a drop-out model

Imputation is a difficult challenge and prone to creating false-positive results in downstream analysis.

→ In alternative, one can use implicit approaches that model counts with **zero-inflated distributions**

There are many different imputation methods available ALRA (Linderman et al. 2022) MAGIC (Dijk et al. 2017), DrImpute and scImpute (Li and Li 2017).

Normalization

Aims: Removing systematic non-biological variation (i.e., biases) and making **count distributions comparable**

Biases in scRNAseq data are due to several factors

- Low mRNA amount per cell
- Variable mRNA capture efficiency
- Variable sequencing depth
- Technical batches
- etc

Normalizing single-cell RNA sequencing data: challenges and opportunities

Catalina A Vallejos[1–4,10], Davide Risso[5,9,10], Antonio Scialdone[2,10], Sandrine Dudoit[5,6] & John C Marioni[2,7,8]

Scaling approaches for normalization

- **Global scaling:** It attemps scaling expression measures within each cells by a constant factor
- **Non-linear scaling:** It attemps scaling expression measures depending on the expression level

Main approaches

1. (Adapt) bulk RNA-seq **size factors**
2. Pooling and deconvolution approaches

- **scran**: *Lun et al. 2016* proposed a pooling and deconvolution approach. Cells are pooled together and normalized against a global pseudoreference, then deconvolved by solving a system of linear equations

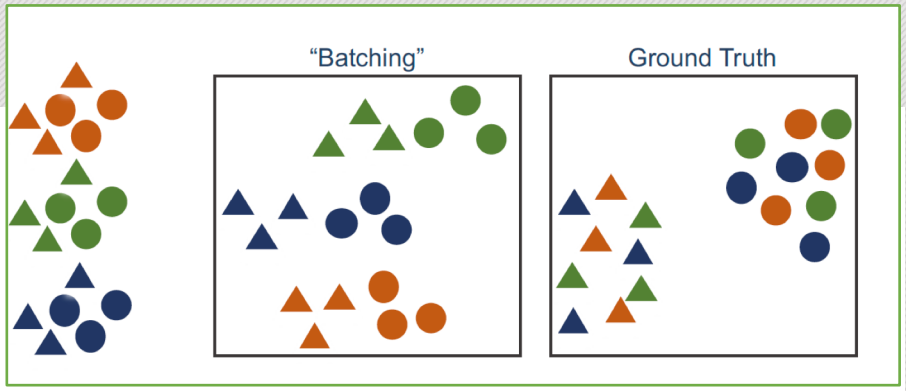
3. Regression based approaches

- **SCnorm**: *Bacher et al 2017* use two step quantile regression

Classical bulk RNA-seq normalization methods (CPM, TMM, upper quartile, etc) might not work well for scRNA-seq data since data are **zero-inflated** due to dropouts and transcriptional bursting

Tutorial

What is a Batch-effect ?

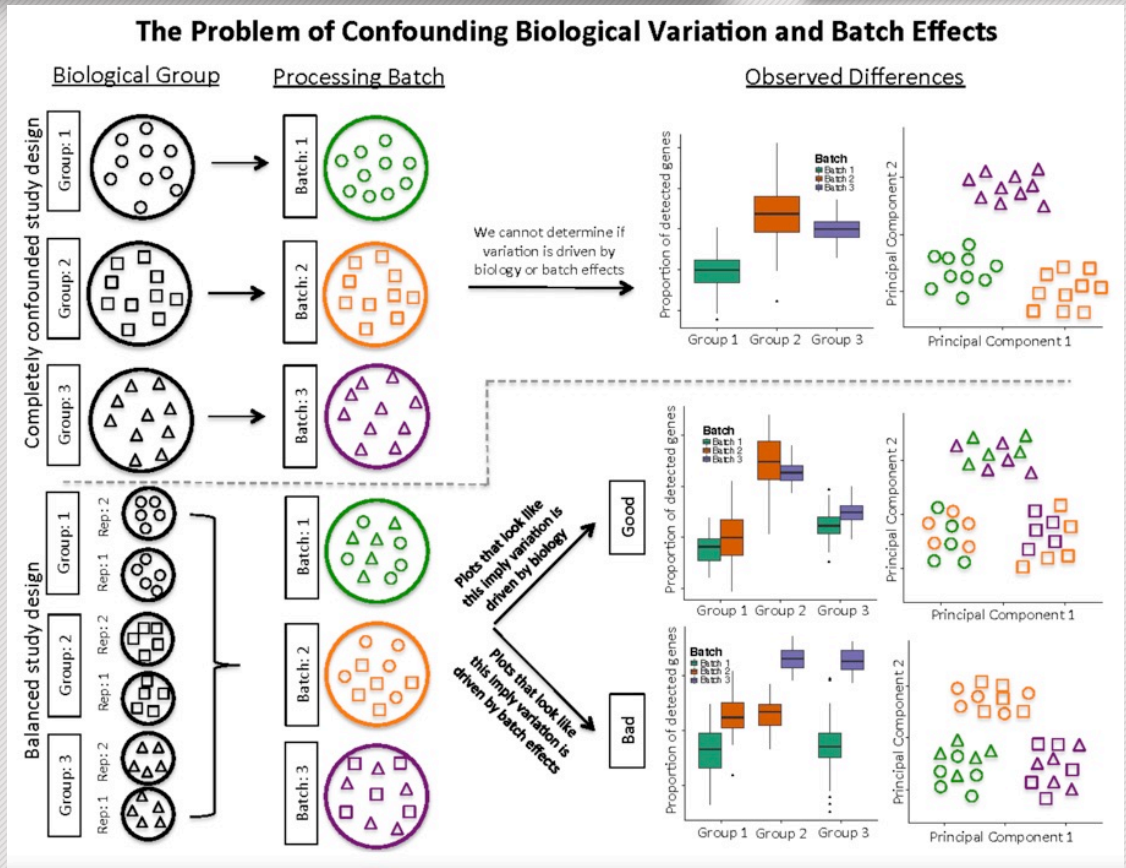


- Data may have been be generated in multiple laboratories, at different times, using different cell dissociation and handling protocols, library-preparation technologies and/or sequencing platforms.

All of these factors result in **technical batch** effects, in which the expression of genes in one batch differs systematically from that in another batch.

Such differences can **mask underlying biology** or introduce **spurious structures** in the data → to avoid misleading conclusions, they must be corrected before further analysis.

→ **We distinguish between known and unknown factors**



Note there could be **biological batches** we are not interested in (cell size, age, gender, donor …)

Feature selection and Dimension reduction

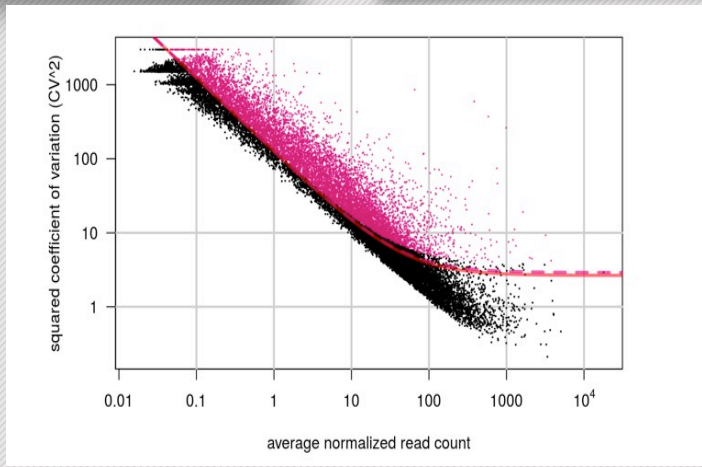
ScRNA-seq are **high dimensional** data → dimensionality has to be reduced before proceeding into the **analysis for reducing** both **technical noise** and **computational burden**

- **Feature selection** is aimed at selecting a subset of the most '*biologically associated*' genes, say for example the **HGV** (Highly variable genes)
- **Dimension reduction** consists of either linear or non linear embeddings that project data into a lower dimensional space that keep most of the biological signal (i.e., cell heterogeneity)

The main assumption is that only a portion of genes will show a response to the biological condition of interest, e.g. differences in cell-type, drivers of differentiation and so on…

Feature selection: Highly Variable Genes (HVG)

- **Assumption:** large differences in expression across cells are due to biological differences between the cells rather than technical noise.
- Due to the nature of count data, there is a positive relationship between the **mean expression** of a gene and **the variance** in the read counts across cells.
- **HGV** consists in **identifying genes that exhibit high cell-to-cell variation in the dataset**



*[Brennecke et al., 2013] implemented in **M3Drop** package*

There are several others similar ideas

1. *normalize data*
2. *calculate the mean and the square coefficient of variation (CV^2) for each gene.* $CV^2 = \frac{\delta^2}{\mu^2}$
3. *Fits a quadratic model (gamma generalized linear model) to the relationship between mean expression and the CV^2*
4. *Use a chi-square test to find genes significantly above the curve*

Tutorial

Dimensionality Reduction for scRNA-seq

Principal component analysis (PCA):

When PC components correlate with cells heterogeneity, **PCs loadings** can be used to identify gene contribution to cell populations

Reduce dimensionality for downstream analysis or for visualizing the data → The number of components to retains depends on the aim

- There are plenty of other methods both based on ***matrix factorization*** such as ICA, NMF, or on ***graph-theory*** (tSNE, UMAP, Isomap, Diffusion maps)
- Moreover, there are sc-specific approaches, such as zero inflated factor analysis (**ZIFA**) or zero-inflated negative binomial wanted variation extraction (**ZINBWaVE**)

Tutorial

Dimensionality Reduction: Other Approaches

tSNE (*t-distributed Stochastic Neighbor Embedding algorithm*)

- It is a non linear approach aimed at preserving local distances (i.e., distance between nearest neighbor cells) → **local embedding,** it does not preserve a global data structure
- It is an **iterative** and **stochastic** method that depends on an inizialization seed and several tuning parameters (i.e., perplexity, number of iterations, many others)
- Note, for large datasets use a novel more efficient implementation based on FFT

UMAP (*Uniform Manifold Approximation and Projection*)

- It is a non linear approach based on **topological structures**
- It is **both local** and **global embeddings**
- Not completely stochastic
- As tSNE, it has many tuning parameters (minimum distance, number of neighbors)
- It is quite fast

tSNE and **UMAP** are implemented in
- Seurat3
- Scater
- Monocole3

Etc

UMAP is relatively novel approach that is becoming extremely popular

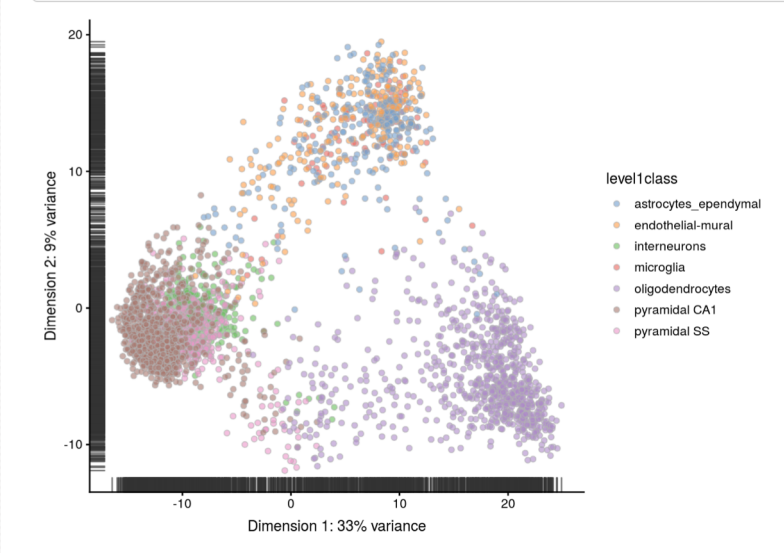
Some R tools for quality control, normalization and dimension reduction

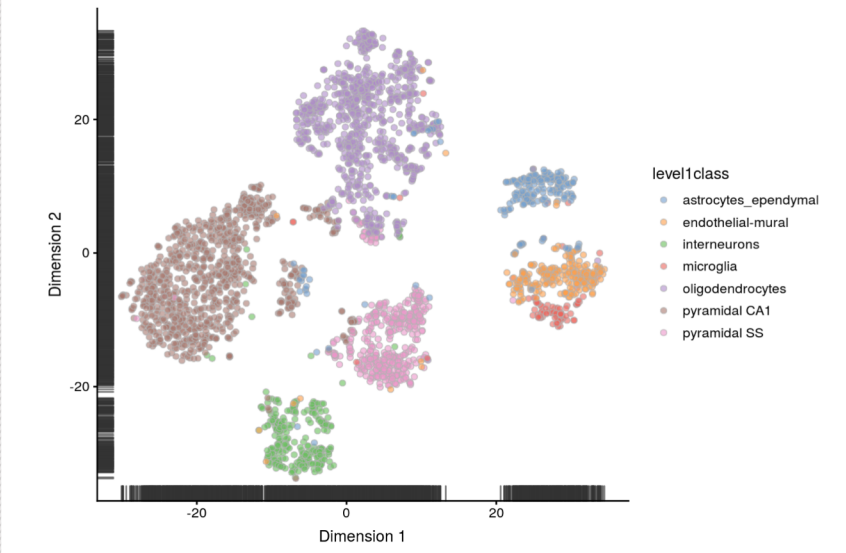
- **scater:** quality control, dimension reduction, visualization (Bioconductor)
- **scran:** normalization, doublet detection, batch effect correction, detection of HGV, dimension reduction, clustering (Bioconductor)
- **SCnorm:** normalization (Bioconductor)
- **scone:** normalization, batch effect correction, quality filtering (Bioconductor)
- **Seurat:** general purpose tool, including normalization and batch removal (CRAN)

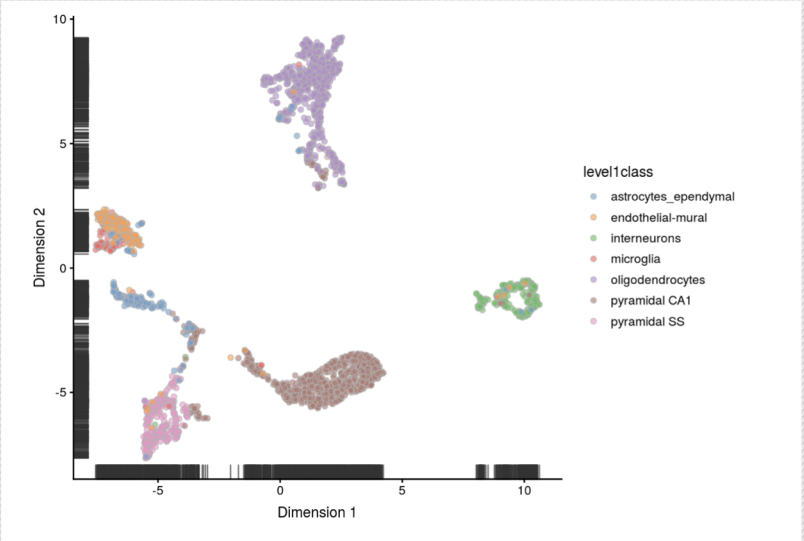
Others

- **scruff:** UMI tools, quality control (Bioconductor)
- **Cellity:** quality control(Bioconductor)
- **simpleSingleCell:** quality control, normalization (Bioconductor)
- **ZINB-WaVE**: dimension reduction (Bioconductor)
- **sctransform:** normalization (CRAN)
- **DropletUtils:** removal of empty droplets (Bioconductor)

Visualization: PCA, tSNE, UMAP







PCA tSNE UMAP

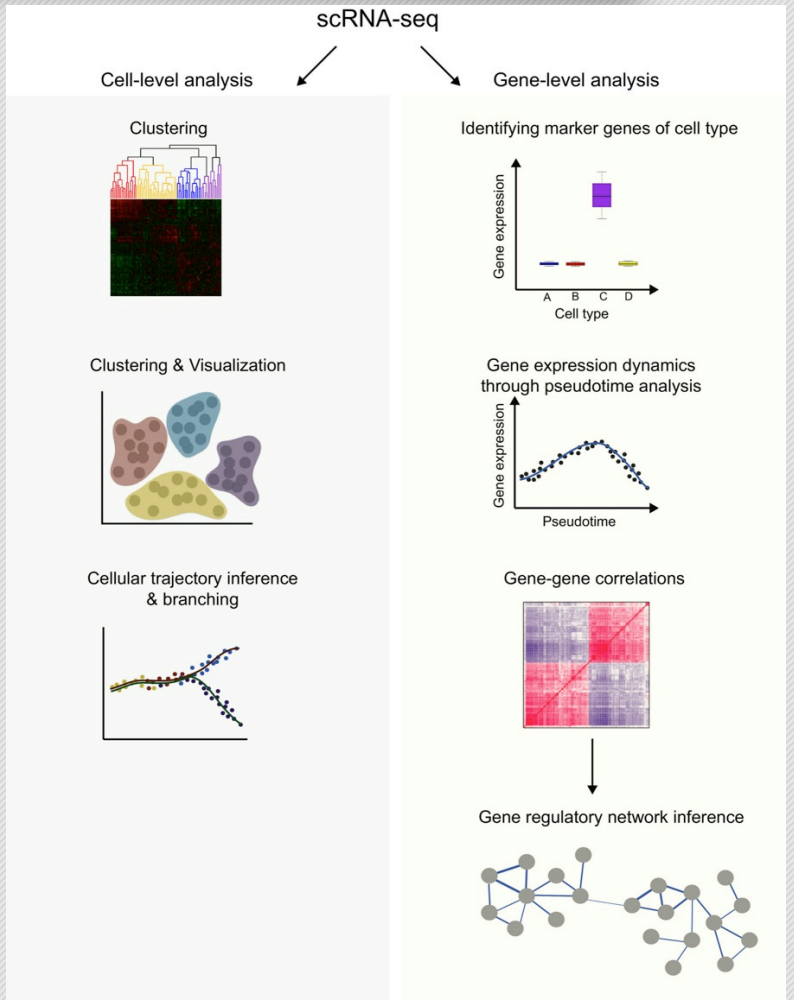
Other visualization methods such as **Diffusion Maps** might be more appropriate for cell differentiation.
Diffusion map is a non linear method for dimension reduction where the distance between the points are measured as probability from going from one to another.

Tutorial

Introduction to downstream analysis

- Clustering
- Cluster annotation
- Trajectory inference and branching
- Cell marker identification
- Differential expression
- Gene regulatory networks
- Others… continuosly emerging

Each question would require a specific course



Clustering

Clustering is one of the most popular scRNA-seq applications, since it allows to identify subpopulations of cells

The most popular clustering approaches are:

- **Hierarchical clustering**: build and cut the dendrogram based on pairwise distance matrix
- **K-means clustering**: iterative approach that reassign cells to the nearest cluster center
- **Graph-based clustering**: Cells are first organized in a network using a KNN approach, then cluster are identified as network modules (i.e, regions of highly connected nodes)

Challenges in scRNA-seq

- What is the number of clusters?
- What about dropouts and other scRNA-seq technical issues ?
- Scalability: in scRNA-Seq experiments the number of cells could be millions, tools developed for single-cell data don't scale well.

Clustering scRNA-seq

Plenty of methods that can be used in different modes and with differently pre-processed data

Table 1 | **Clustering methods for scRNA-seq**

Name	Year	Method type	Strengths	Limitations
scanpy[4]	2018	PCA + graph-based	Very scalable	May not be accurate for small data sets
Seurat (latest)[3]	2016			
PhenoGraph[32]	2015			
SC3[22]	2017	PCA + *k*-means	High accuracy through consensus, provides estimation of *k*	High complexity, not scalable
SIMLR[24]	2017	Data-driven dimensionality reduction + *k*-means	Concurrent training of the distance metric improves sensitivity in noisy data sets	Adjusting the distance metric to make cells fit the clusters may artificially inflate quality measures
CIDR[25]	2017	PCA + hierarchical	Implicitly imputes dropouts when calculating distances	
GiniClust[75]	2016	DBSCAN	Sensitive to rare cell types	Not effective for the detection of large clusters
pcaReduce[27]	2016	PCA + *k*-means + hierarchical	Provides hierarchy of solutions	Very stochastic, does not provide a stable result
Tasic et al.[28]	2016	PCA + hierarchical	Cross validation used to perform fuzzy clustering	High complexity, no software package available
TSCAN[41]	2016	PCA + Gaussian mixture model	Combines clustering and pseudotime analysis	Assumes clusters follow multivariate normal distribution
mpath[45]	2016	Hierarchical	Combines clustering and pseudotime analysis	Uses empirically defined thresholds and a priori knowledge
BackSPIN[26]	2015	Biclustering (hierarchical)	Multiple rounds of feature selection improve clustering resolution	Tends to over-partition the data
RaceID[23], RaceID2[115], RaceID3	2015	*k*-Means	Detects rare cell types, provides estimation of *k*	Performs poorly when there are no rare cell types
SINCERA[5]	2015	Hierarchical	Method is intuitively easy to understand	Simple hierarchical clustering is used, may not be appropriate for very noisy data
SNN-Cliq[80]	2015	Graph-based	Provides estimation of *k*	High complexity, not scalable

DBSCAN, density-based spatial clustering of applications with noise; PCA, principal component analysis; scRNA-seq, single-cell RNA sequencing.

Identifying cell populations with scRNASeq

Tallulah S. Andrews, Martin Hemberg*

Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK

Challenges in unsupervised clustering of single-cell RNA-seq data

Vladimir Yu Kiselev, Tallulah S. Andrews & Martin Hemberg

Benchmark and parameter sensitivity analysis of scRNAseq clustering methods.

Monika Krzak[1*], Yordan Raykov[2], Alexis Boukouvalas[3], Luisa Cutillo[4], Claudia Angelini[1]

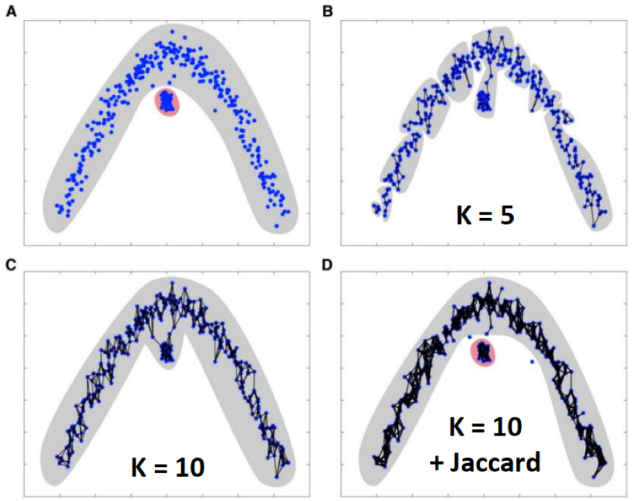
- There is still a significant space for improvments
- Scalability is becoming a serious issue for several methods

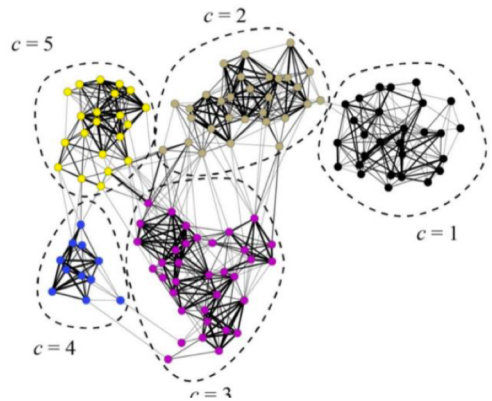
Seurat Clustering

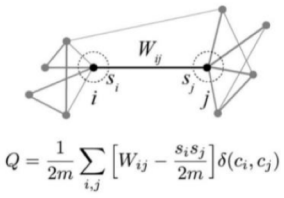
1. Reduce dimensionality using PCA
2. Construct KNN (*k-nearest neighbor*) graph based on the Euclidean distance in PCA space using some of the first components
3. Refine the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard distance, i.e. how many shared edges).
4. Cluster cells by optimizing for modularity (*Louvain algorithm*)

Note that it does not require to explicitly choose the number of clusters

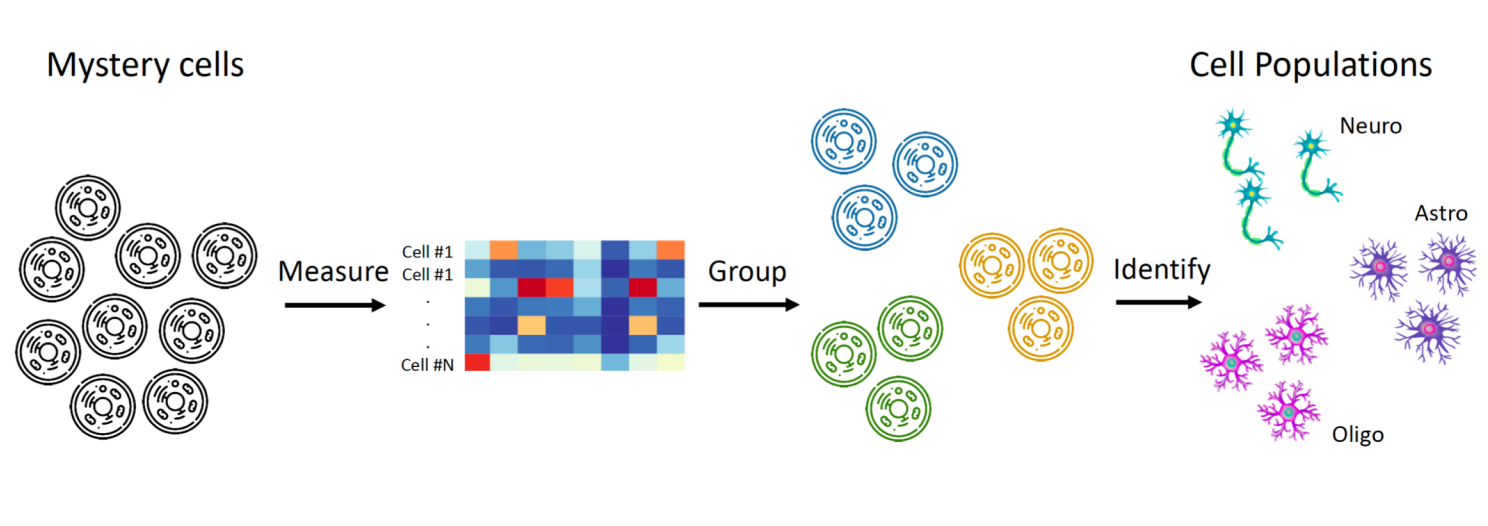
Tutorial







Cluster annotation - Cell identity



Making sense of the data

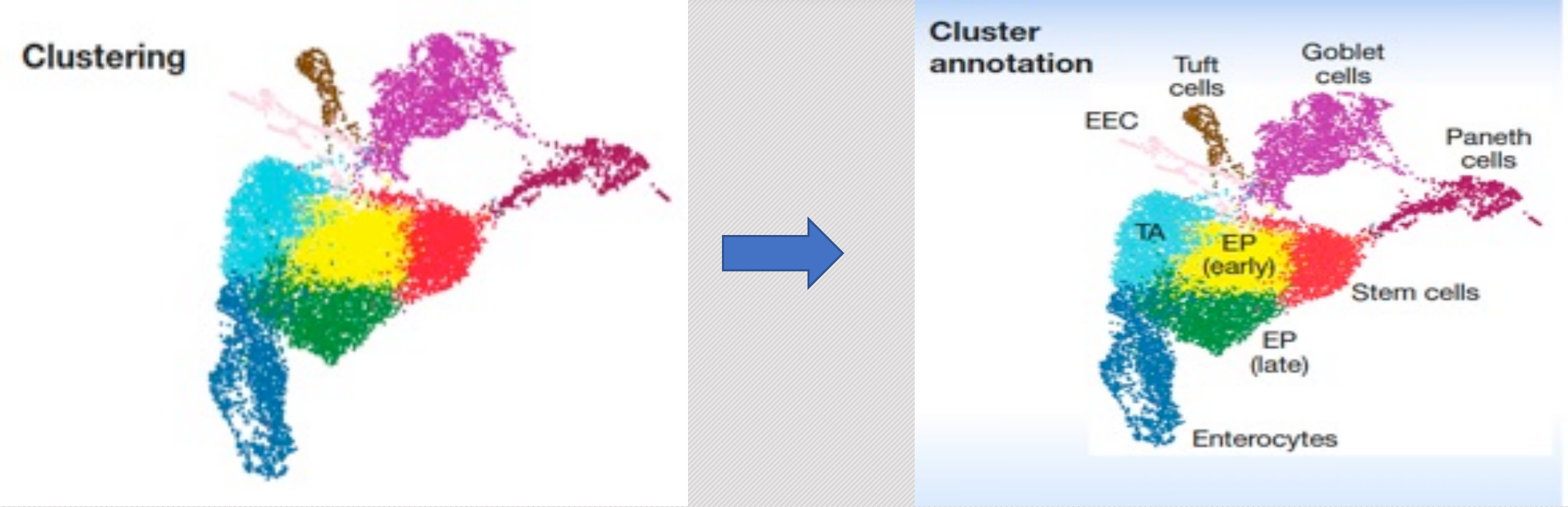
- Samples are heterogeneous, either in cell composition and abundance
- Important in particular for **immunology** and **cancer research**
- Discovering novel cell types

A comparison of automatic cell identification methods for single-cell RNA sequencing data

Tamim Abdelaal[1,2†], Lieke Michielsen[1,2†], Davy Cats[3], Dylan Hoogduin[3], Hailiang Mei[3], Marcel J. T. Reinders[1,2] and Ahmed Mahfouz[1,2*]

- **Manual approaches**: Use cell surface markers or other gene markers to identify known cell population
- Databases with cell type gene signatures (**PanglaoDB, CellMarker**, **Tabula Muris, Cell Atlas**..)
- **Automatic approaches**: Classification/mapping based approaches (**scmapm SingleR**, ..)
- Tumor cells are much more challenging
- Did you identify novel cell sub-population?..... is it a true functionally characterized populations?

Cluster annotation - Cell identity



Note

- It is not always clear what constitutes a '*cell type*' (i.e, at what resolution types are different)
- Cells of the same cell type in different states may be detected in separate clusters

→ it is better to use the term "cell identities" rather than "cell types"

The iteration of clustering, cluster annotation, re- or subclustering and re-annotation can be time-consuming

Differential Expression

Comparison of cell types (often within a single sample) to find **"marker genes"** or cell-type specific **pathways**

However cell types is usually unknown

- **Step 1**: Get the cell populations using clustering approaches
- **Step 2**: Compare expression levels between populations

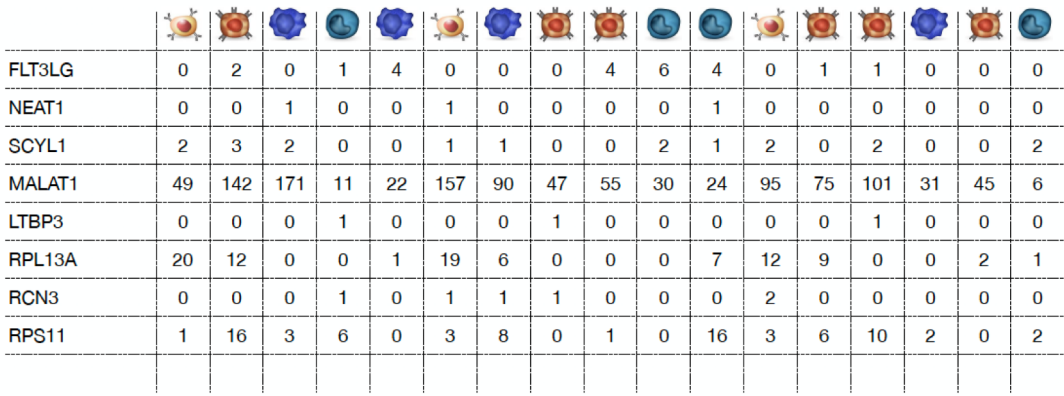
Bulk RNA-Seq methods:

- edgeR, DESeq2, etc

→ scRNAseq data are much more sparse, with higher variability, many more cells than the typical number samples of bulk RNA-seq

Single-cell methods:

- Single Cell Differential Expression (**SCDE**)
- Model-based Analysis of Single-cell Transcriptomics (**MAST**)
- Many others



FLT3LG	0	2	0	1	4	0	0	0	4	6	4	0	1	1	0	0	0
NEAT1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0
SCYL1	2	3	2	0	0	1	1	0	0	2	1	2	0	2	0	0	2
MALAT1	49	142	171	11	22	157	90	47	55	30	24	95	75	101	31	45	6
LTBP3	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
RPL13A	20	12	0	0	1	19	6	0	0	0	7	12	9	0	0	2	1
RCN3	0	0	0	1	0	1	1	1	0	0	0	2	0	0	0	0	0
RPS11	1	16	3	6	0	3	8	0	1	0	16	3	6	10	2	0	2

For a given cluster, are we interested in "*marker genes*" that are:

- DE compared to all cells outside of the cluster
- DE compared to at least one other cluster
- DE compared to each of the other clusters
- DE compared to "most" of the other clusters

`scran::findMarkers` [Tutorial]

Some R tools for Downstream analysis

- SC3: Clustering (Bioconductor)
- Seurat: QC and pre-processing, detection of HGV, dimension reduction, clustering, DE (CRAN)
- Monocole3: Clustering, differential expression, trajectories, identification of marker genes (Bioconductor and CRAN)

Other applications

- Slingshot : trajectories (Bioconductor)
- MAST: DE (Bioconductor)
- SCDE: DE (Bioconductor)
- SCENIC : gene regulatory networks (GitHub)
- SCODE : gene regulatory networks (GitHub)
- ISEE: Shiny-based graphical user interface for exploring data (Bioconductor)

More tools at

- https://www.scrna-tools.org
- https://github.com/seandavi/awesome-single-cell

Conclusions

- **scRNA-seq** allows to investigate cell heterogeneity and development at an unprecedenent level of resolution and throughtput
- **scRNA-seq** data are by far **noisier** and **sparser** than bulk RNA-seq data, → novel computational methods are required
- While their preprocessing (i-e. Data cleaning, normalization, etc) is quite general, the downstream analysis strongly depends on the biological question of interest
- More than **400 methods** have been developed→ **no 'golden standard'**
- scRNA-seq **data size** is rapidally increasing→ novel challenge to **scalability** of methods in terms of **memory** and **run time**
- **Novel applications** of scRNA-seq are continuolsy proposed

Thank you for the attention