Single-Cell RNAseq Data

Processing Part

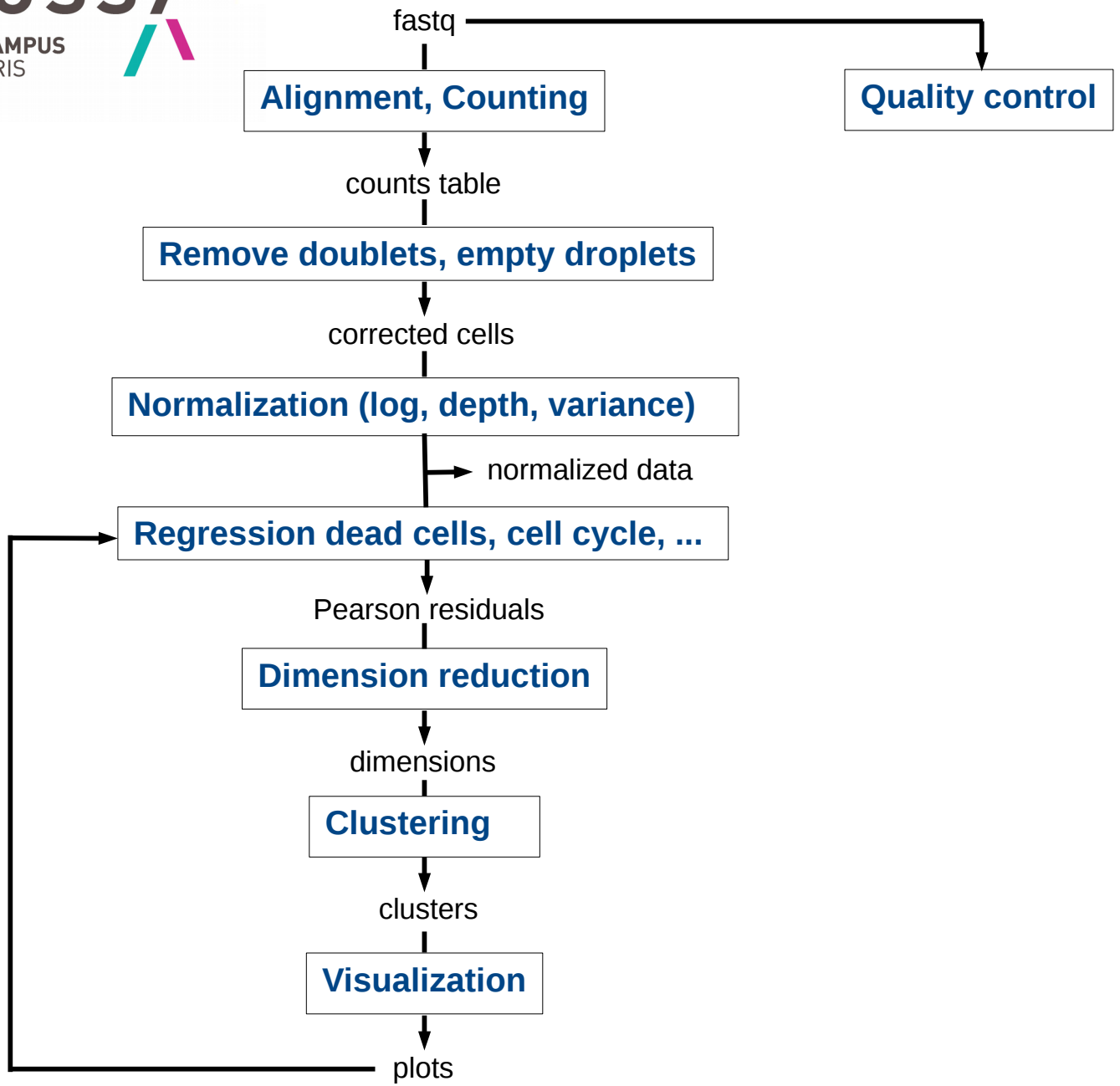
Marine AGLAVE, Bioinformatics Core Facility

GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

Plan

- Reminders on pre-processing
- Annotation of cell types
- Differential expression analysis
- Enrichment and GSEA
- Trajectory Inference
- Integration data

Reminders on pre-processing



Annotation of cell types

fastq

Alignment, Counting

counts table

Remove doublets, empty droplets

corrected cells

Normalization (log, depth, variance)

normalized data

Regression dead cells, cell cycle, ...

Pearson residuals

Dimension reduction

dimensions

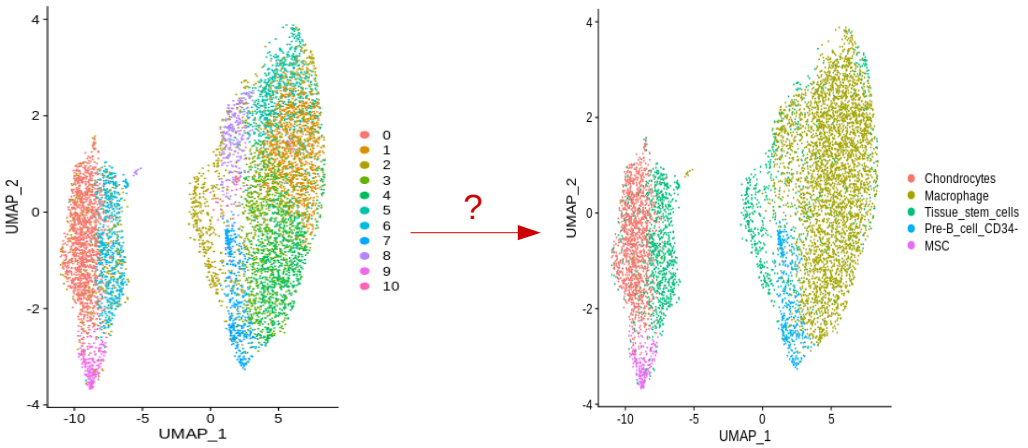
Clustering

clusters

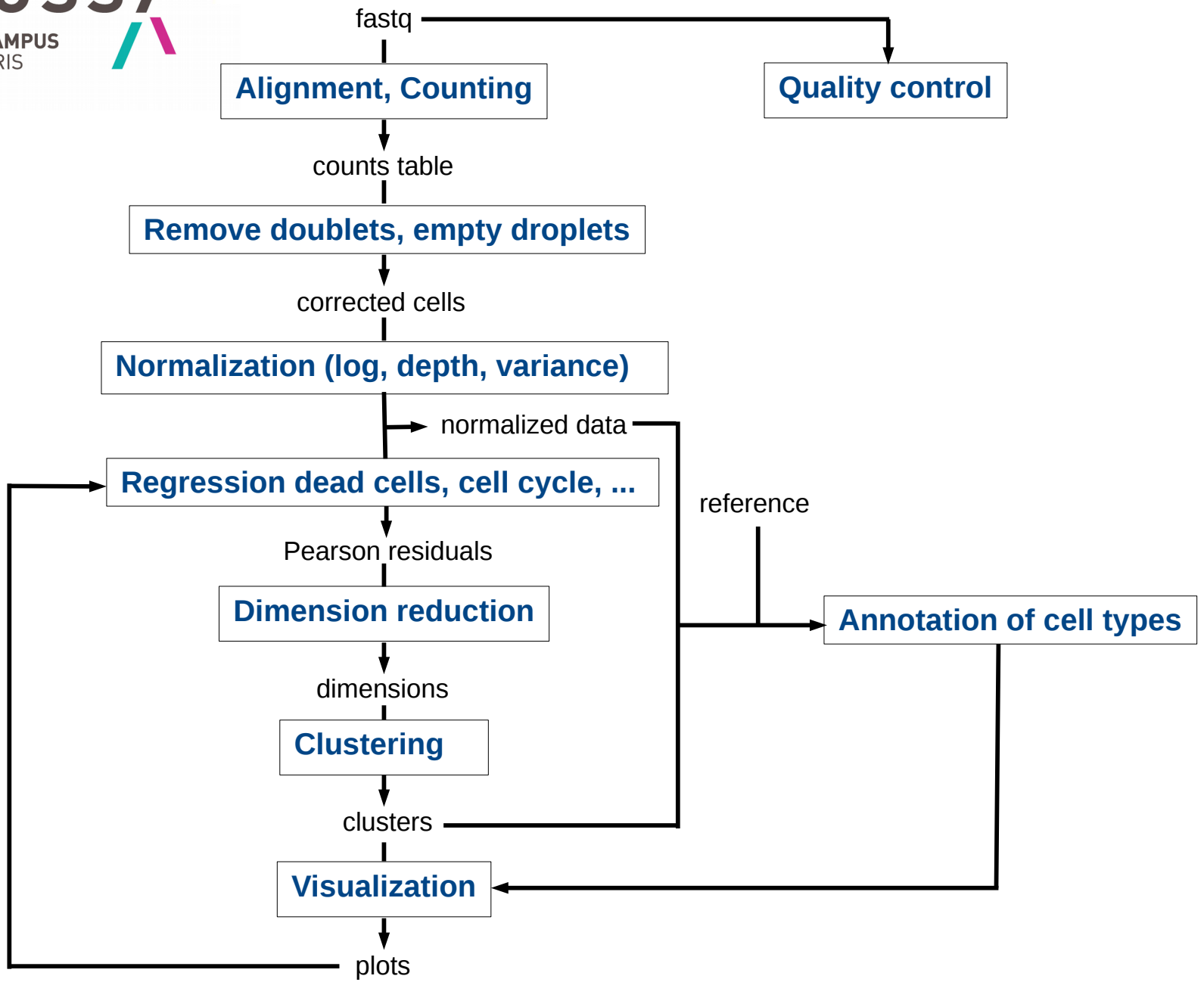
Visualization

plots

Quality control



Annotation of cell types



Annotation of cell types

References

SingleR
clustifyr
scrnaseq

	reference				
	label1		label2	label3	...
	cell a	cell b	cell c	cell d	...
Gene 1					…
Gene 2					...
Gene 3					...
...	...	...	...	...	...

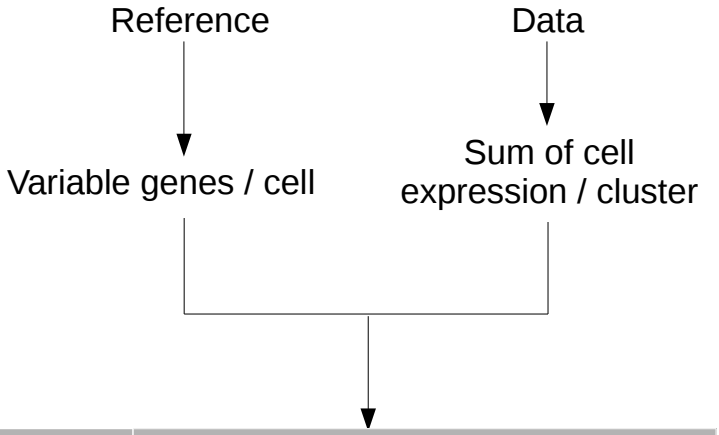
Log Expression

Sources	Organism Origin	Technic	Cell types
Human Primary Cell Atlas (HPCA)	Human	microarray	36
Novershtern Hematopoietic Data			16
Human hematopoietic cell			38
Bluepint epigenomics + ENCODE project		RNA-seq	24
Database of Immune Cell Expression (DICE)			5
Monaco Immune Data			10
Human cortex development		scRNA-seq	47
Human pancreatic cell (inDrop)			14
Human pancreatic cell (SmartSeq2)			12
Human pancreas			14
Muraro Pancreas Data			10
Segerstolpe Pancreas Data			14
Immunologic GenomeProject	Mouse	microarray	20
Mouse Brain		RNA-seq	18
Mouse RNA-seq		scRNA-seq	28
Mouse Cell Atlas			713
Tabula Muris (10X)			112
Tabula Muris (SmartSeq2)			172
Mouse Organogenesis Cell Atlas (MOCA)			37
Mouse pancreas			13
416B Mouse cell line			1
Mouse trophoblasts			1
Shekhar Retina Data			19
Zeisel Brain Data			7
Xenopus tail	Xenopus		46
Personalized reference			

Annotation of cell types

SingleR

Methods



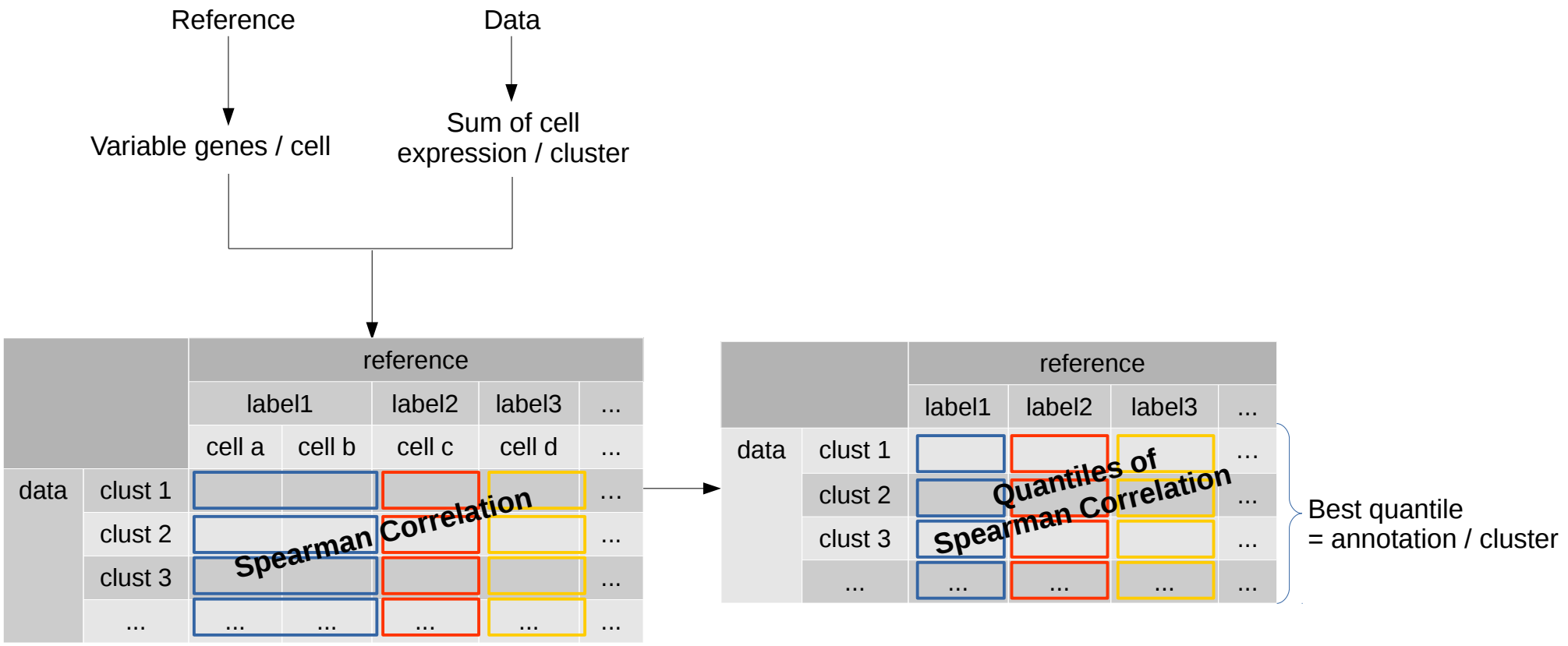
		reference				
		label1		label2	label3	...
		cell a	cell b	cell c	cell d	...
data	clust 1					…
	clust 2					...
	clust 3					...
	...	...	...	...	...	...

Spearman Correlation

Annotation of cell types

SingleR

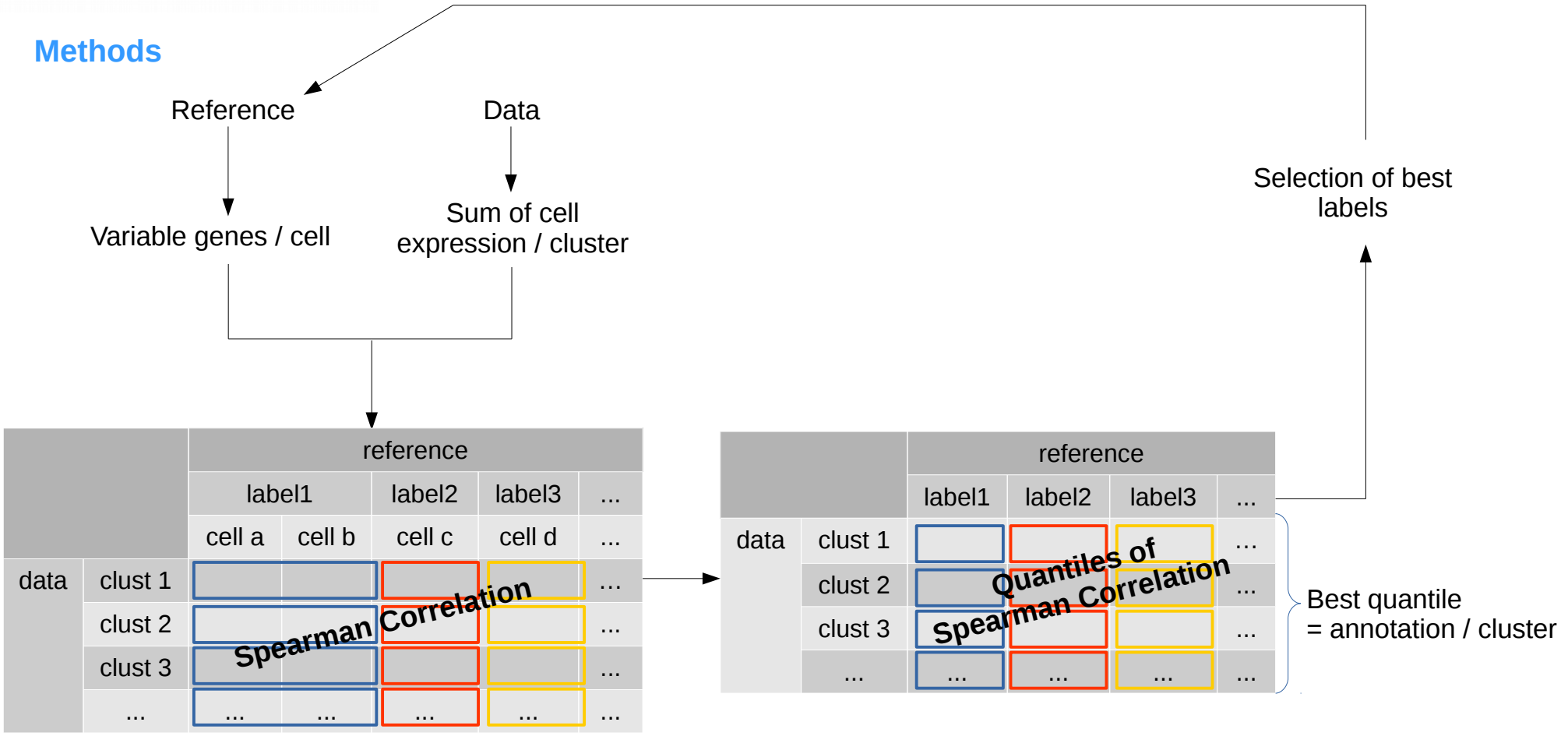
Methods



Annotation of cell types

SingleR

Methods

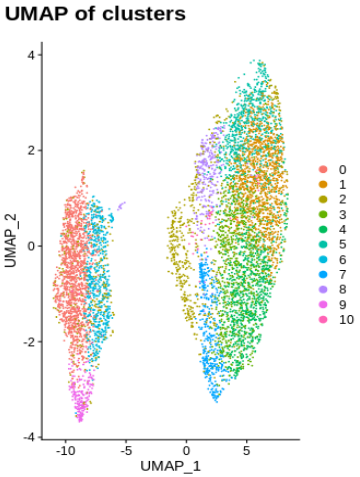


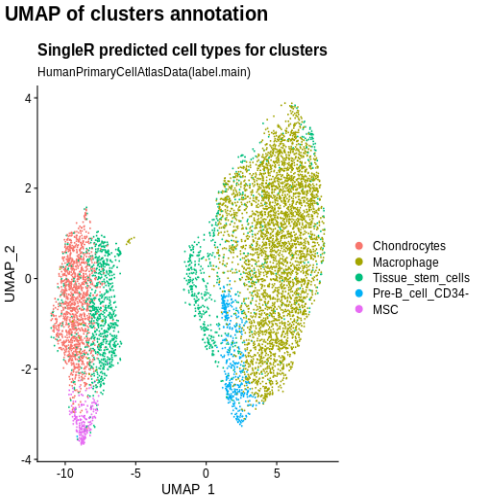
Annotation of cell types

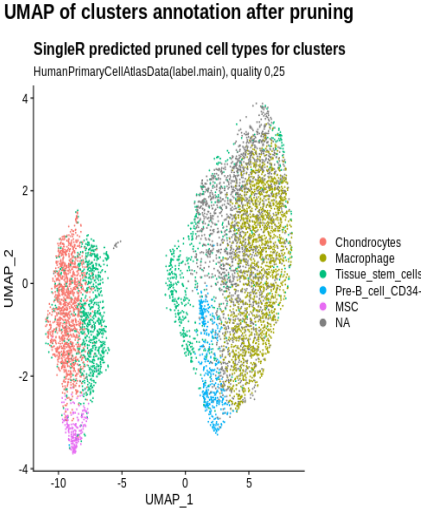
SingleR

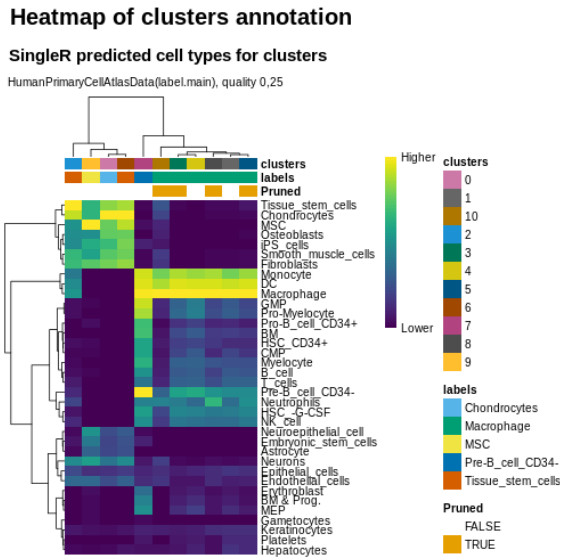
By cluster

Results







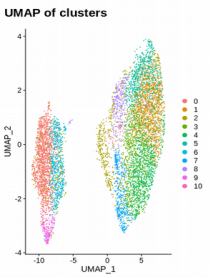


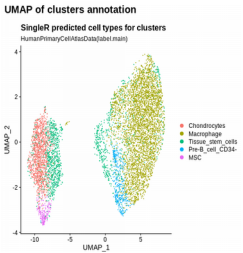
Annotation of cell types

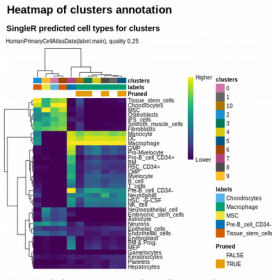
SingleR

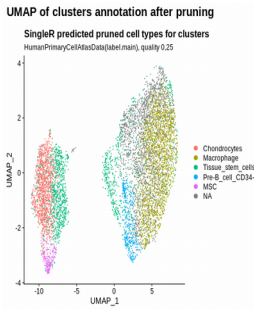
Results

By cluster

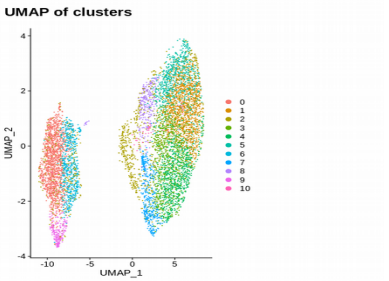


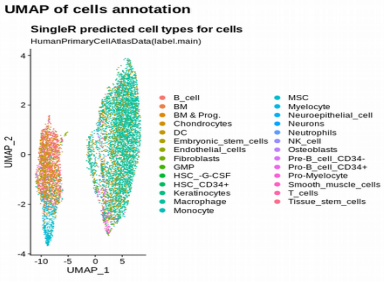


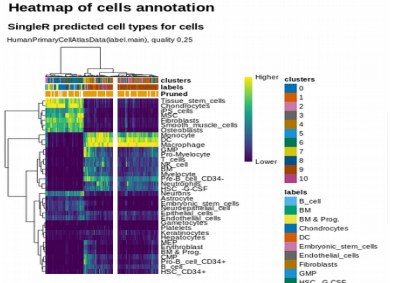


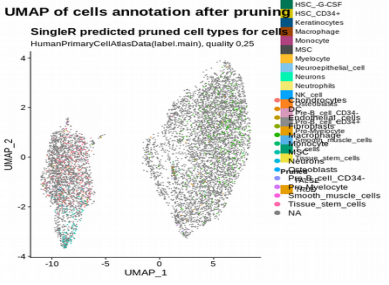


By cells









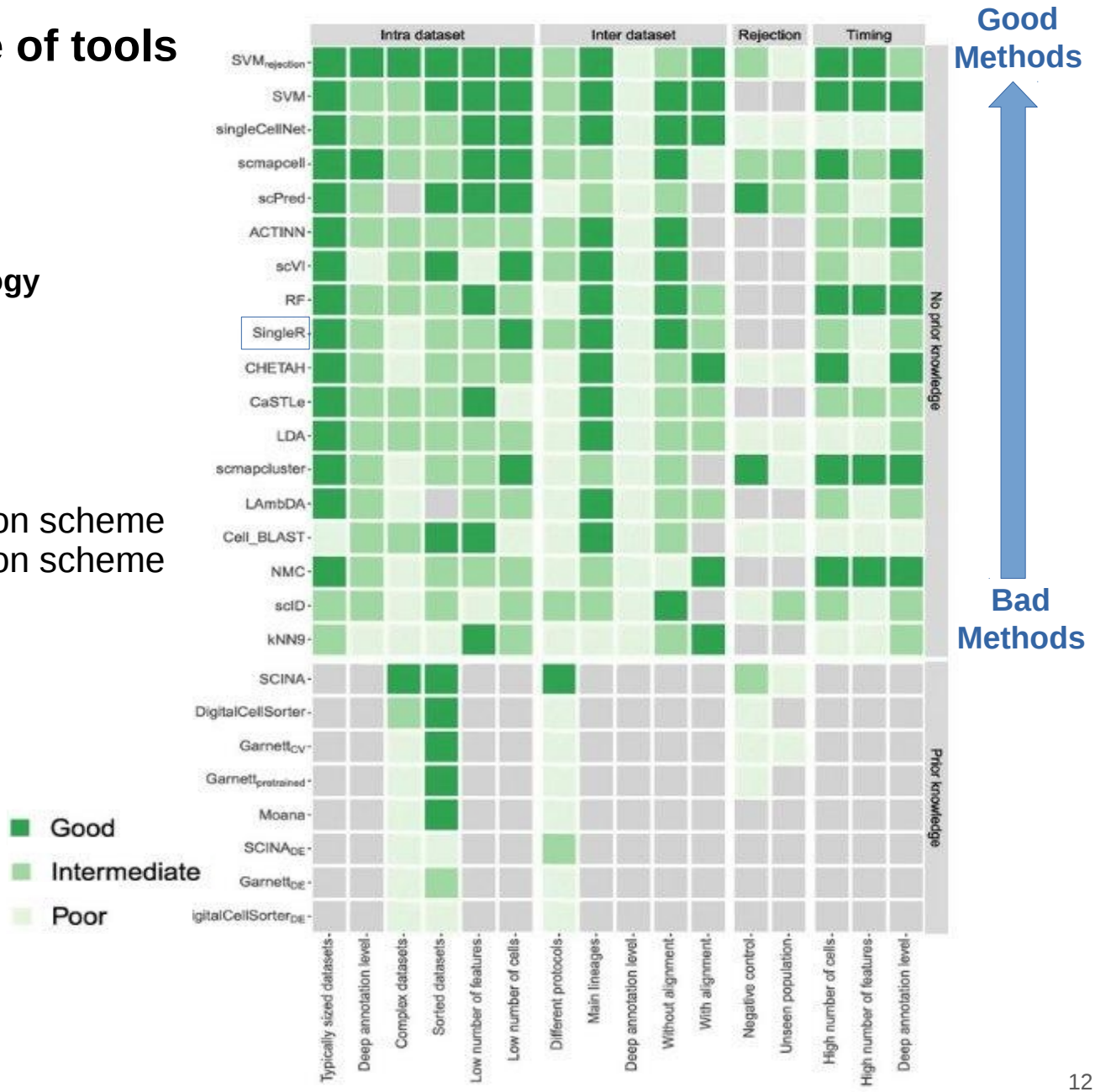
Annotation of cell types

Choice of tools

Abdelaal et al, 2019, Genome Biology

- 22 tools
- 27 public datasets

- Intra-datasets cross-validation scheme
- Inter-datasets cross-validation scheme



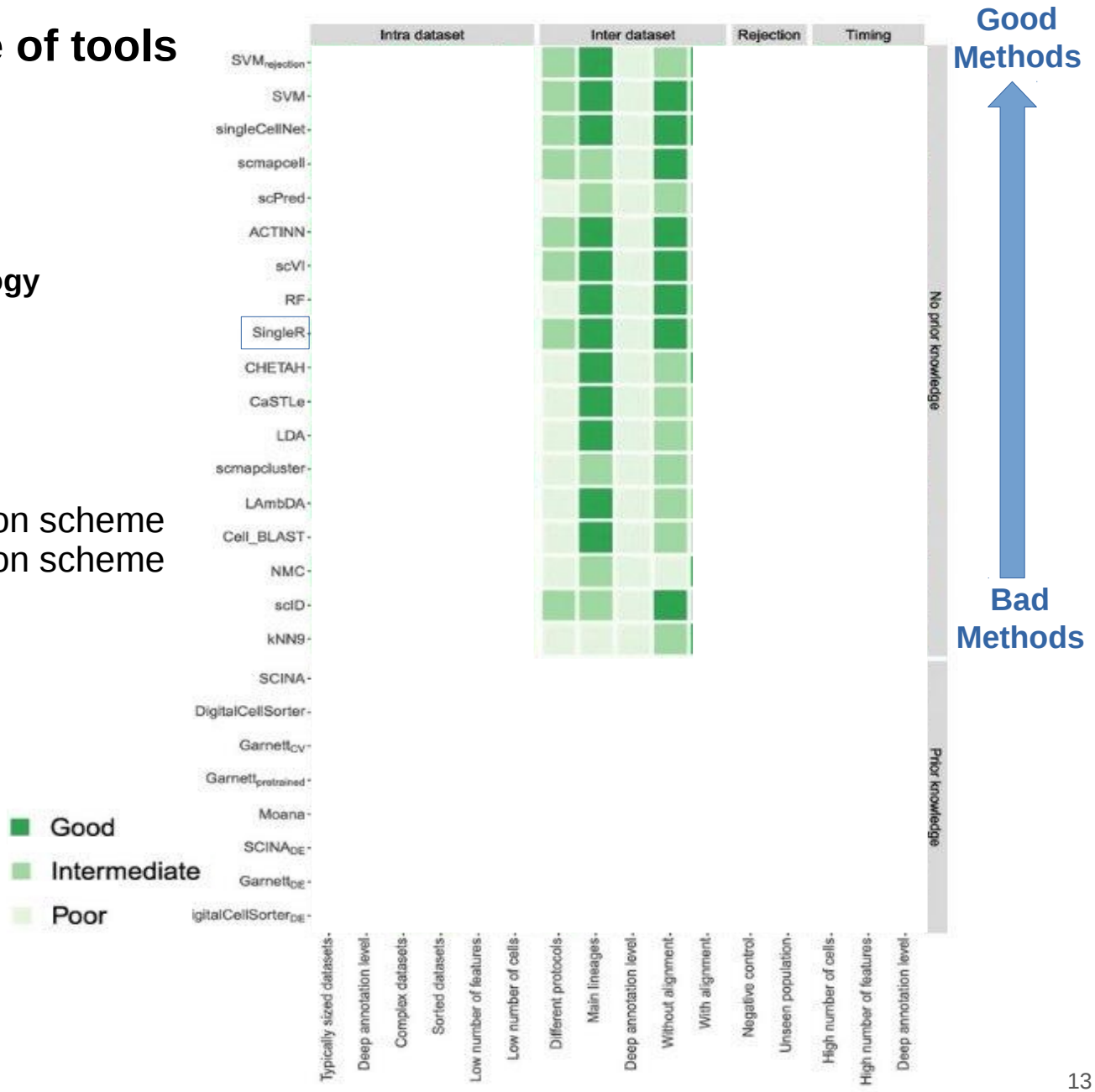
Annotation of cell types

Choice of tools

Abdelaal et al, 2019, Genome Biology

- 22 tools
- 27 public datasets

- Intra-datasets cross-validation scheme
- Inter-datasets cross-validation scheme



Annotation of cell types

Choice of tools

Abdelaal et al, 2019, Genome Biology

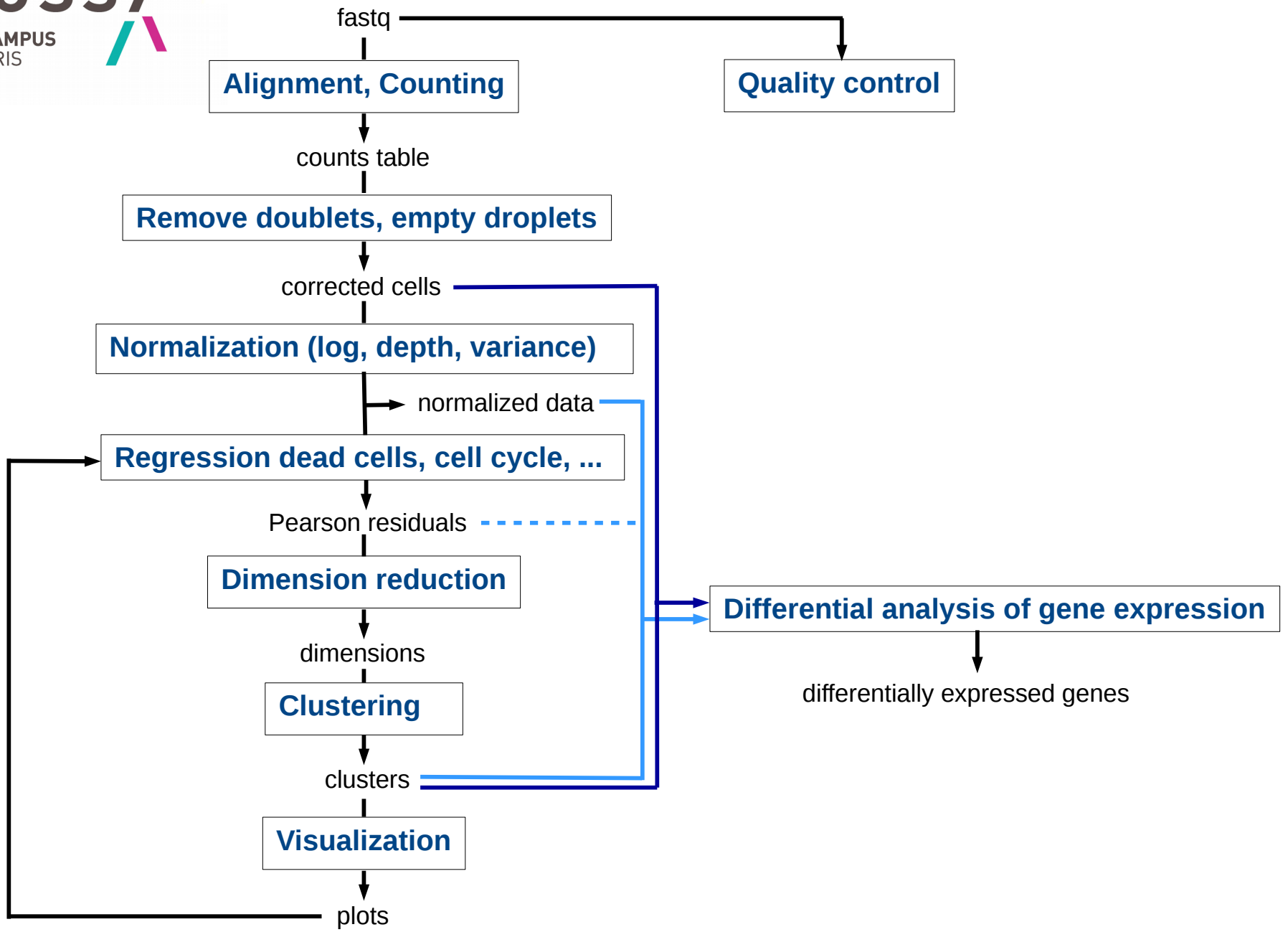
- 22 tools
- 27 public datasets

- Intra-datasets cross-validation scheme
- Inter-datasets cross-validation scheme

Development project:

- clustifyr
- cellassign

Differential expression analysis

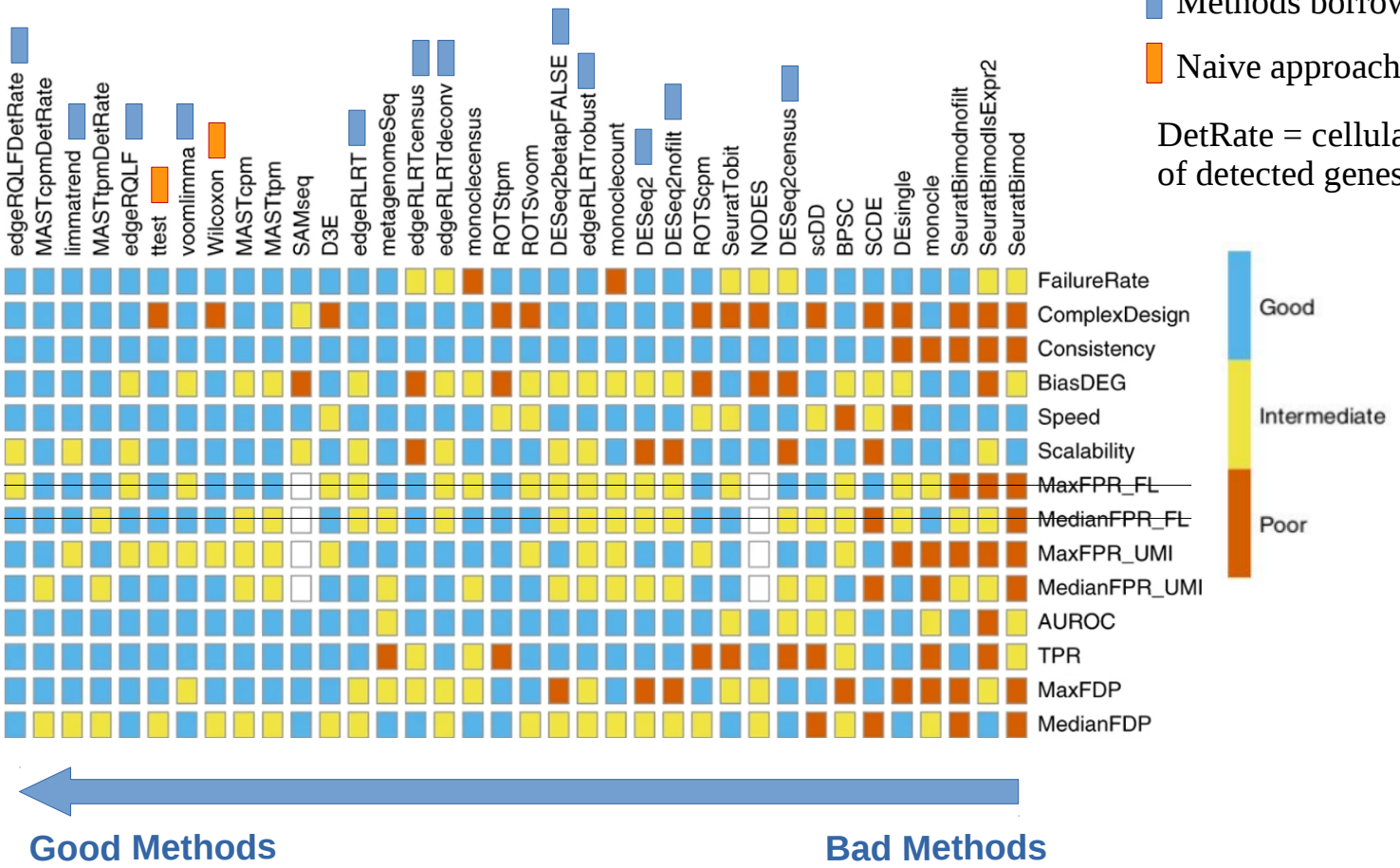


Differential expression analysis

Choice of tools

Soneson & Robinson, 2018, nature methods

- 36 approaches
- 9 datasets (6 full-length + 3 UMI) + simulations



Methods borrowed from bulk-RNA-seq

Naive approaches

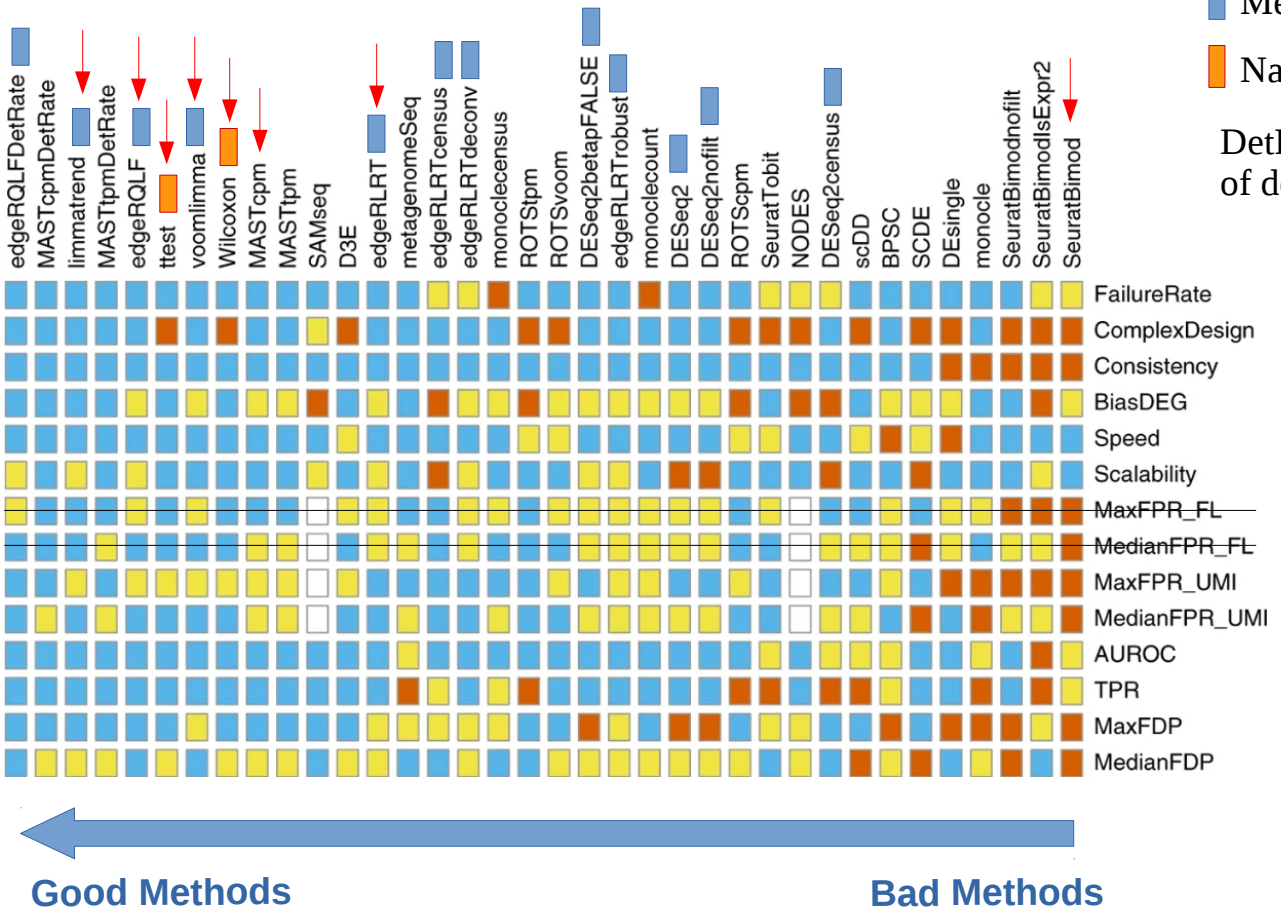
DetRate = cellular detection rate (the fraction of detected genes per cell) as a covariate.

Differential expression analysis

Choice of tools

Soneson & Robinson, 2018, nature methods

- 36 approaches
- 9 datasets (6 full-length + 3 UMI) + simulations



- bulk and naive methods perform well
- edgeR detect a lot FP when expression is weak

Methods borrowed from bulk-RNA-seq

Naive approaches

DetRate = cellular detection rate (the fraction of detected genes per cell) as a covariate.

Being tested:

From Seurat
- wilcoxon,
- bimod,
- t-test,
- poisson,
- negbinomial,
- LogisticRegression,
- MAST

From other:
- edgeRQL
- edgeRLRT
- limma voom
- limma trend

Differential expression analysis

Choice of tools

Methods borrowed from bulk-RNA-seq

Naive approaches

Tian Mou et al, 2020, frontiers in Genetics

- 9 approaches:

Method	Distribution assumption	Test statistic
BPSC	Beta-Poisson	z-test
DEsingle	Zero-Inflated Negative Binomial	Likelihood ratio test
MAST	Normal (Generalized linear hurdle)	Likelihood ratio test
monocle	Normal (Generalized additive model)	Likelihood ratio test
DESeq2	Negative Binomial	Wald test
edgeR	Negative Binomial	Quasi-likelihood F-test
limmatrend	Normal (linear model)	Empirical-Bayes Moderated t-statistics
t-test	Normal	t-test
Wilcoxon	Nonparametric	Wilcoxon

- 3 datasets (full-length) + 1 simulation

- Evaluation by the rediscovery rates (RDR) of top-ranked genes, separately for highly and lowly expressed genes.

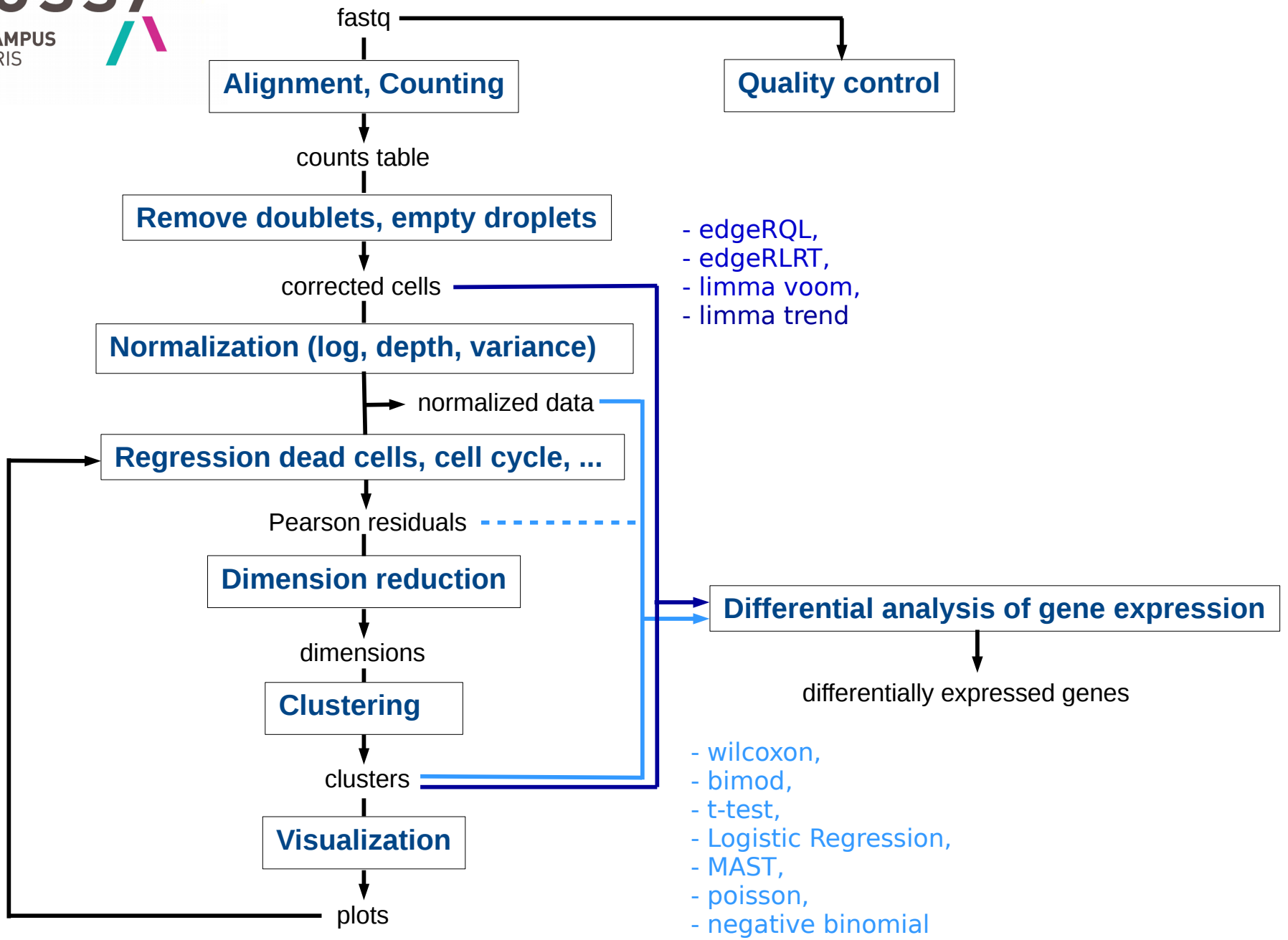
- For the lowly expressed genes: performance varies substantially
- EdgeR and monocle have poor control of false positives
- DESeq2 loses sensitivity
- BPSC, Limma, DEsingle, MAST, t-test and Wilcoxon have similar performances

Being tested:

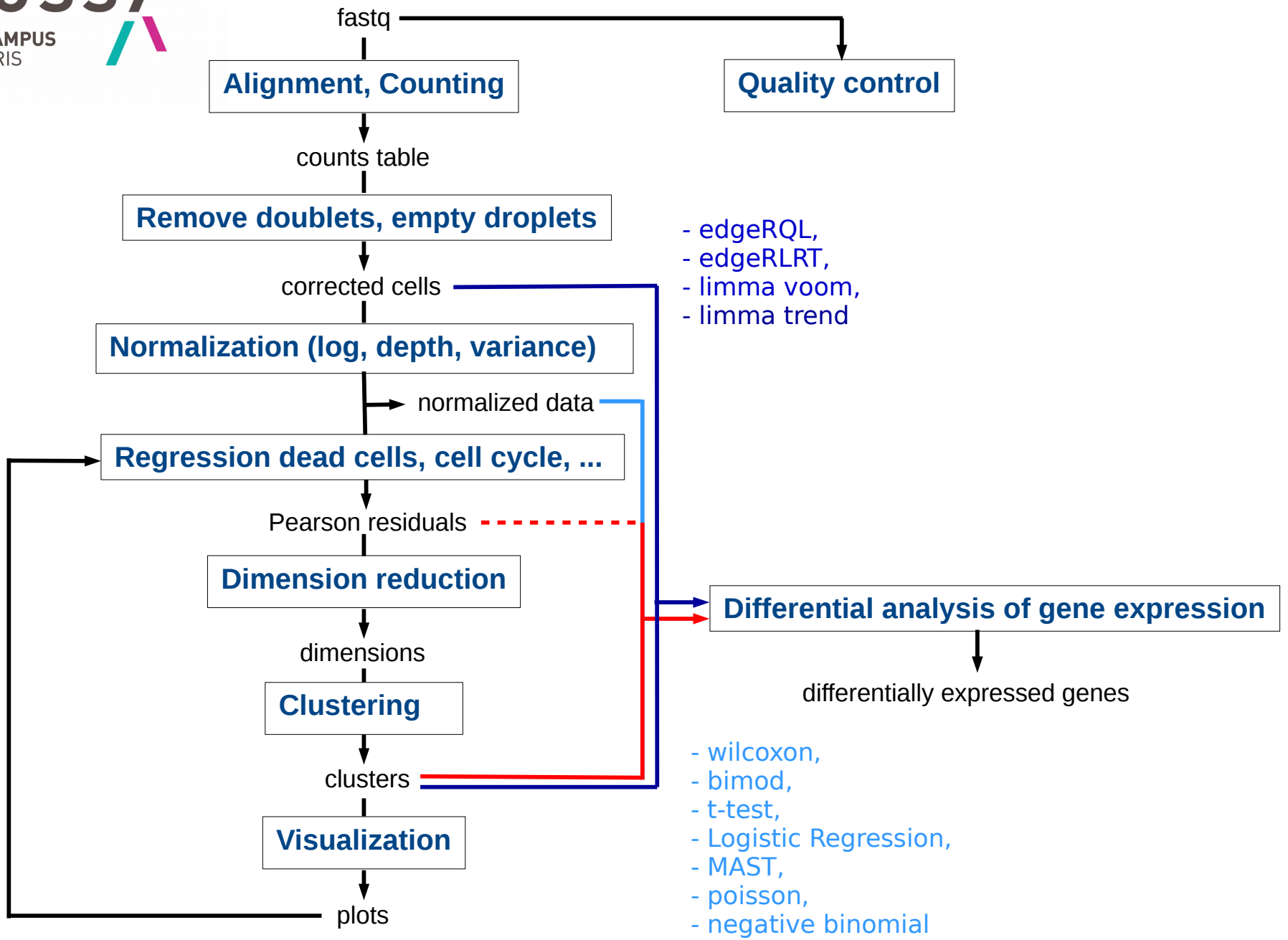
From Seurat
- wilcoxon,
- bimod,
- t-test,
- poisson,
- negbinomial,
- LogisticRegression,
- MAST

From other:
- edgeRQL
- edgeRLRT
- limma voom
- limma trend

Differential expression analysis

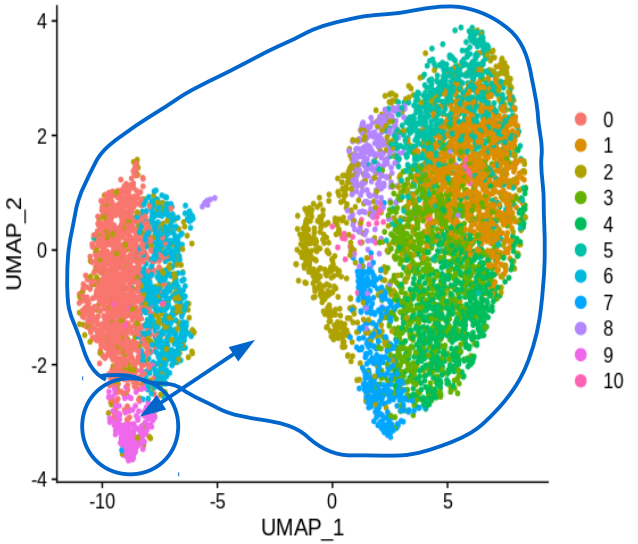


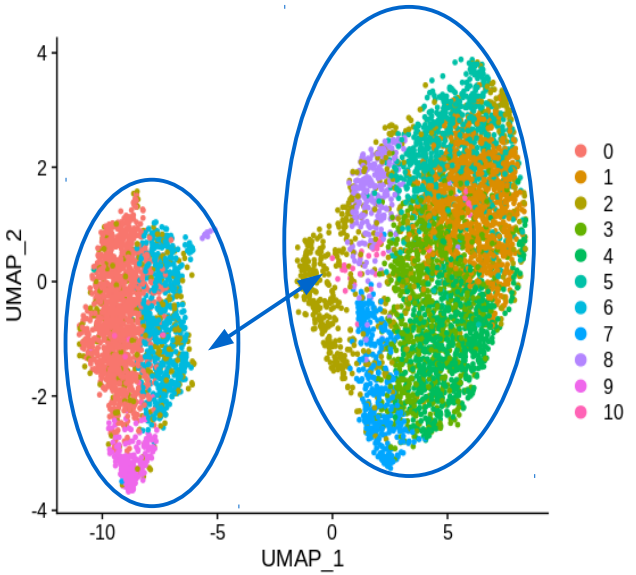
Differential expression analysis

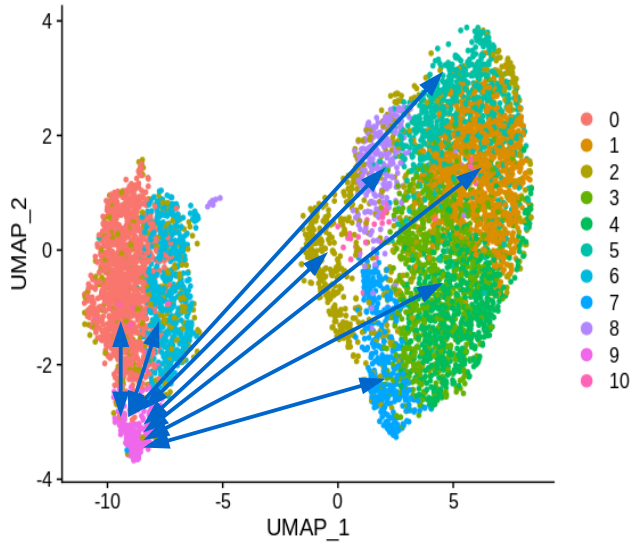


Differential expression analysis

Types of comparisons





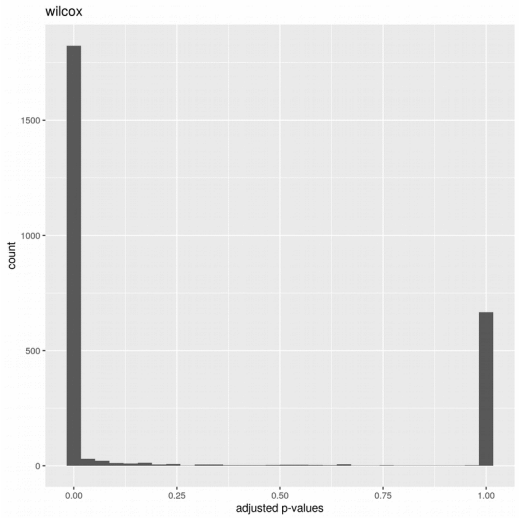


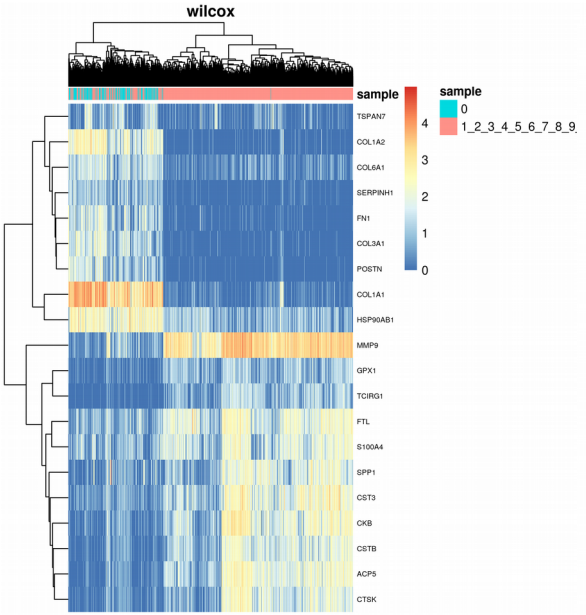
Development project:

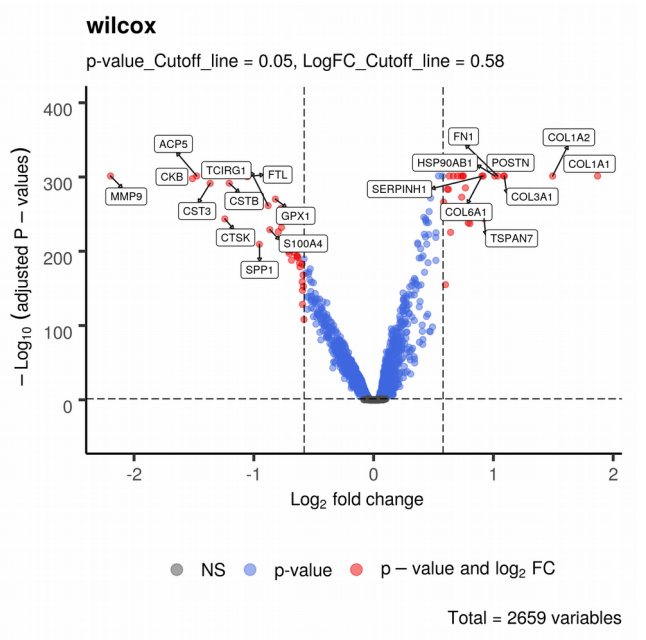
- comparisons between conditions

Differential expression analysis

Results



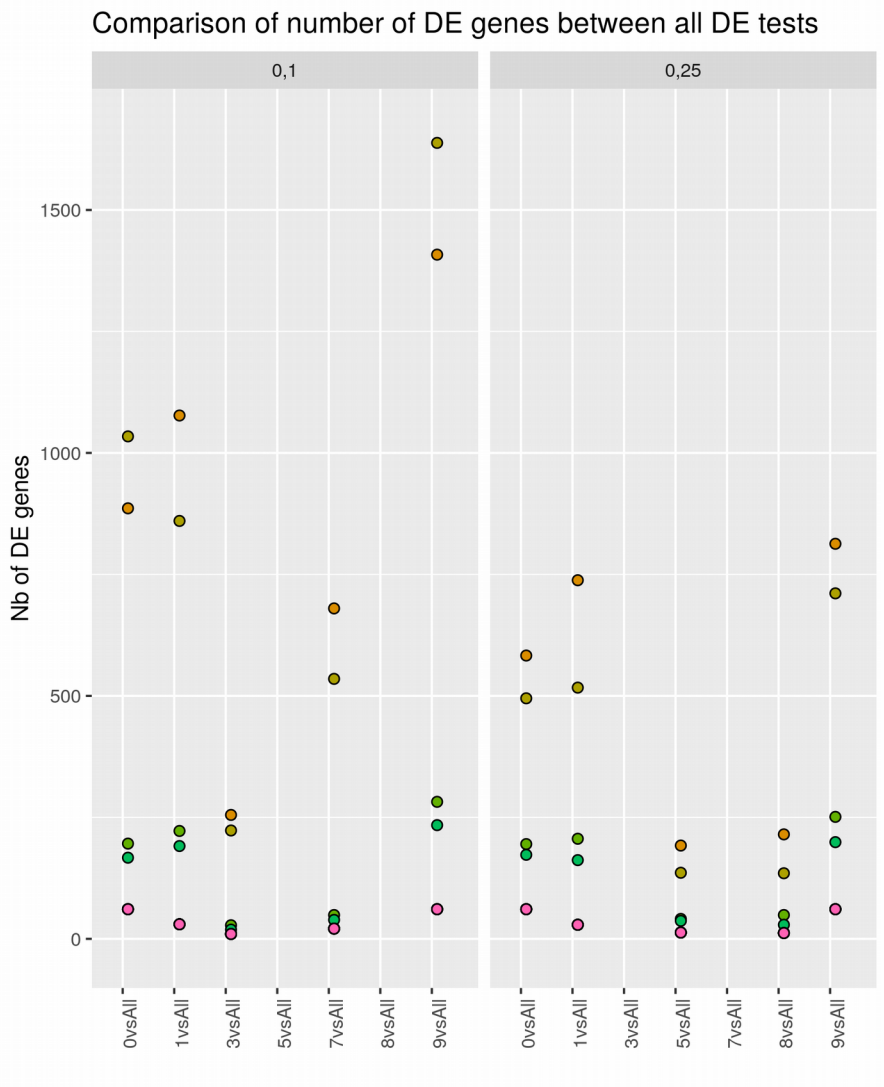


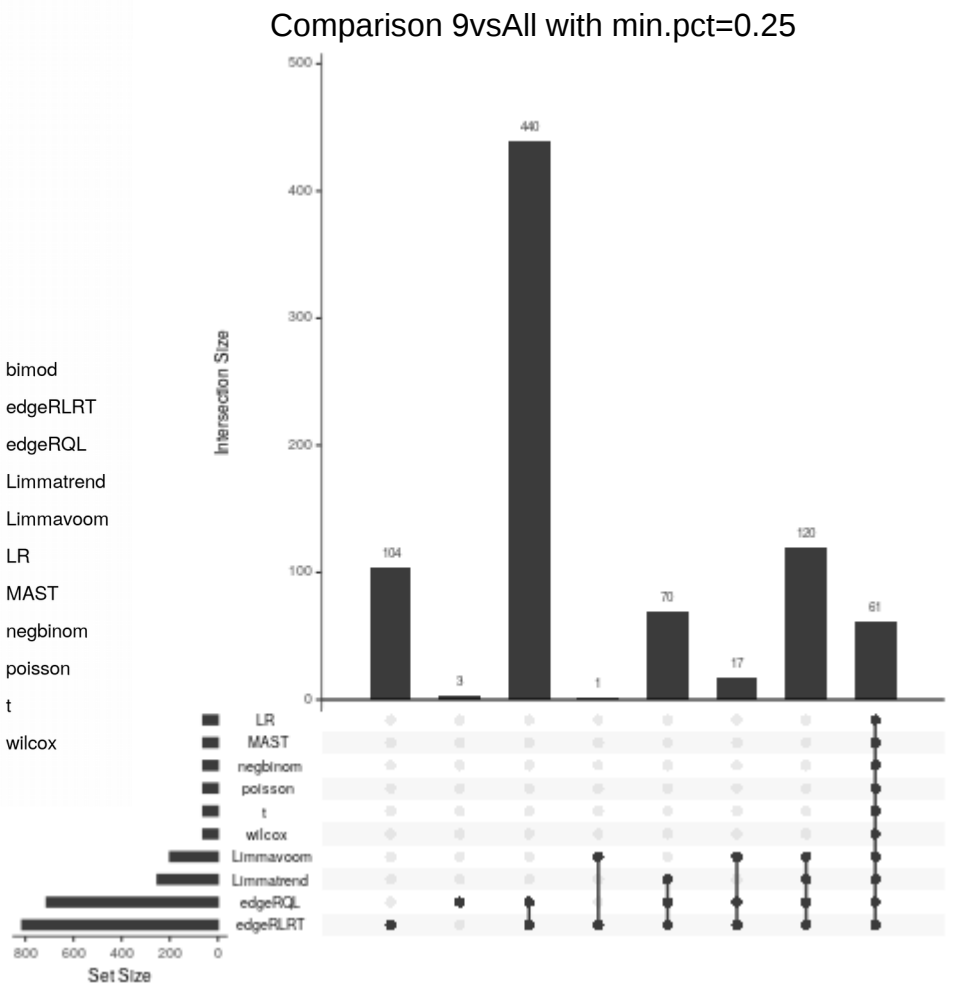


	A	B	C	D	E	F	G	H	I
1	p_val	logFC	pct.1	pct.2	adj.P.Val	tested_cluster	control_cluster	gene	min.pct
2	0	1,8698024958	1	0,403	0	0	All	COL1A1	0,1
3	0	1,4945502039	0,975	0,244	0	0	All	COL1A2	0,1
4	0	1,0910998214	0,87	0,204	0	0	All	COL3A1	0,1
5	0	1,0864930327	0,838	0,187	0	0	All	POSTN	0,1
6	0	1,0360020908	0,852	0,22	0	0	All	FN1	0,1
7	0	1,0117625078	0,993	0,822	0	0	All	HSP90AB1	0,1
8	0	0,9150576426	0,807	0,205	0	0	All	SERPINH1	0,1

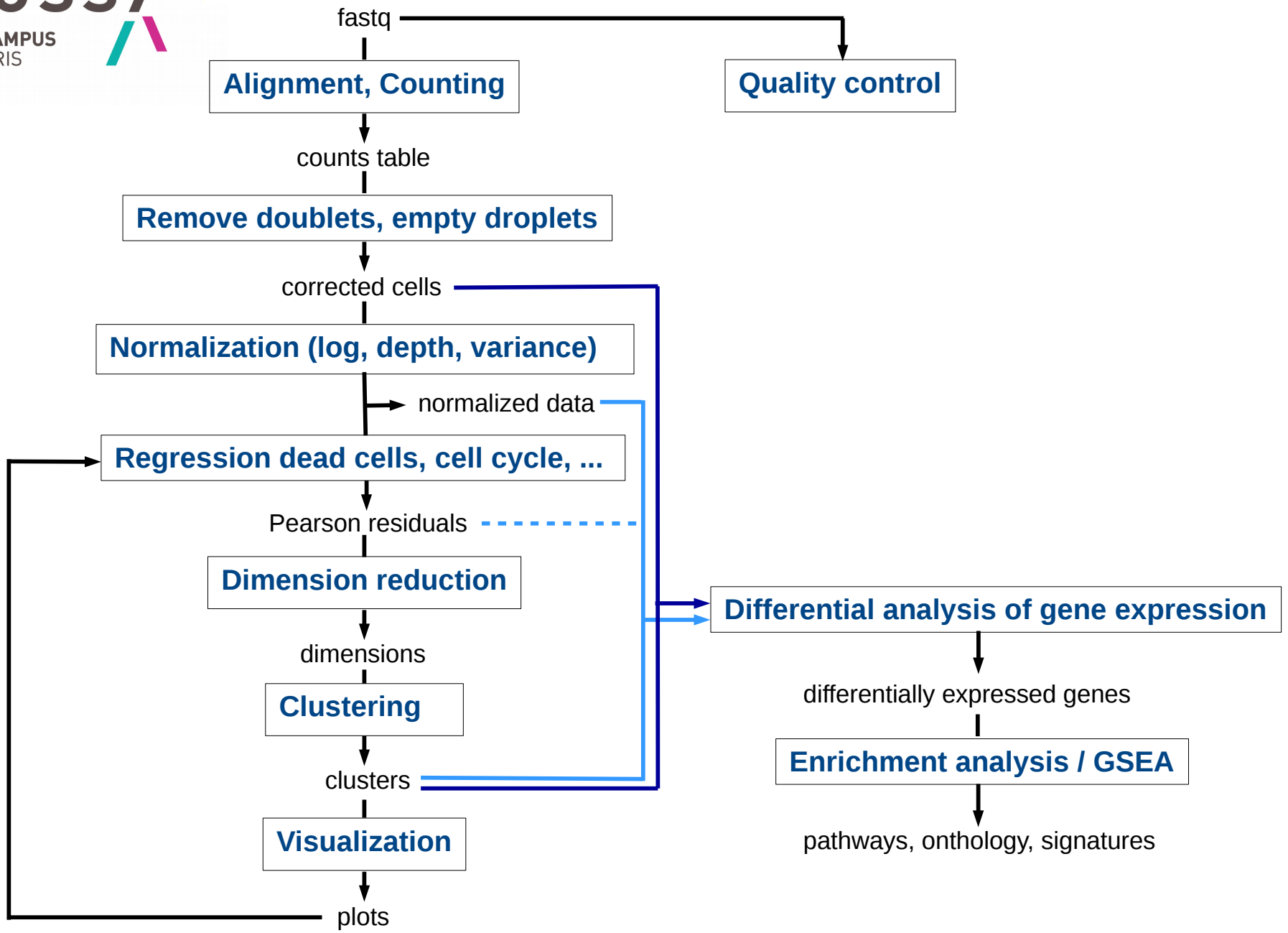
Differential expression analysis

Results





Enrichment and GSEA



Enrichment and GSEA

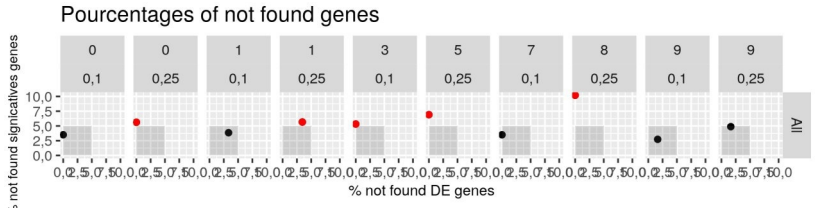
ClusterProfiler

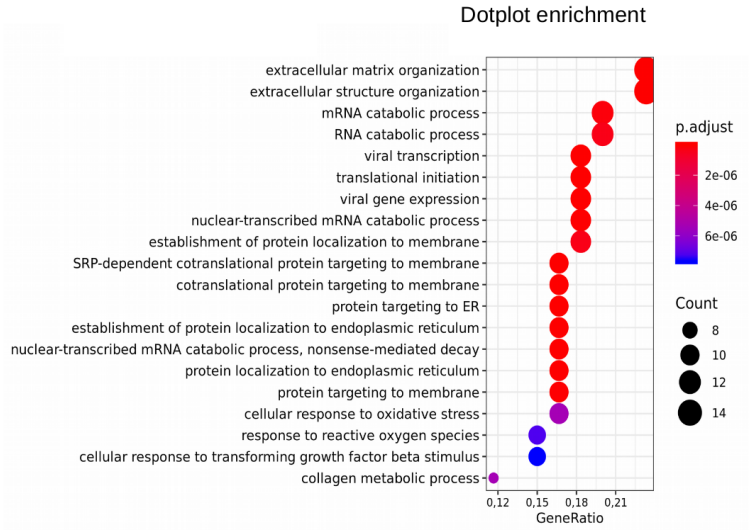
Yu G et al. 2012, OMICS: A Journal of Integrative Biology

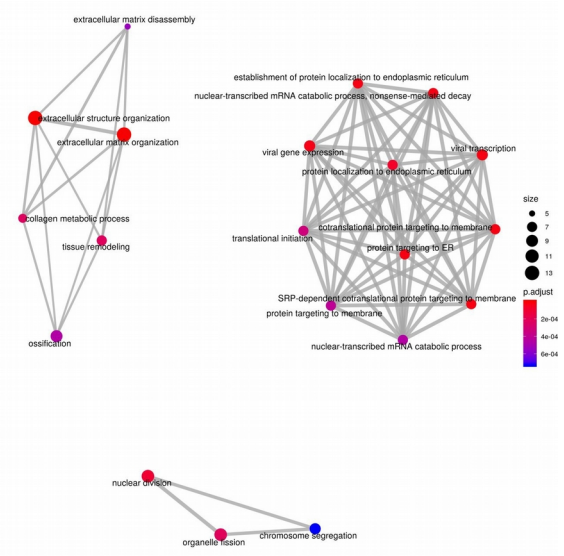
BDD:
- Wikipathways
- MsigDB
- CellMarkers
- Disease Onthology
- GO Terms

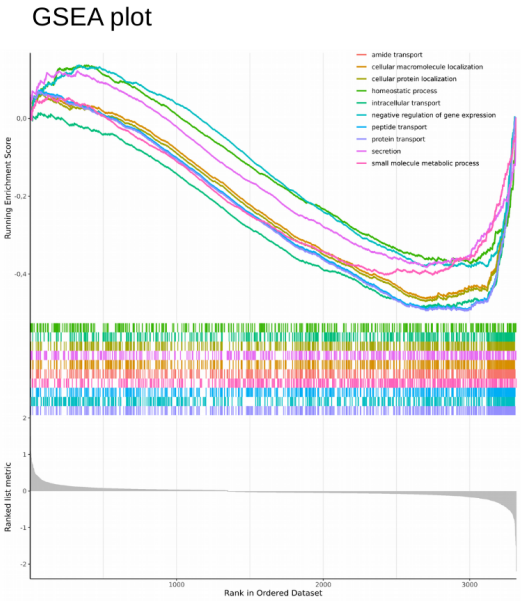
Results

	A	B	C	D	E	F	G	H	I	J	K
1	tested_cluster	control_cluster	min.pct	nb_sig_genes	nb_sig_genes_not_found	pct_sig_not_found	nb_sig_genes_used	nb_DE_genes	nb_DE_genes_not_found	pct_DE_not_found	nb_DE_genes_used
2	0	All	0,1	1855	65	3,5040431267	1790	61	0	0	61
3	1	All	0,1	1946	75	3,8540596095	1871	30	1	3,3333333333	29
4	3	All	0,1	507	27	5,325443787	480	10	0	0	10









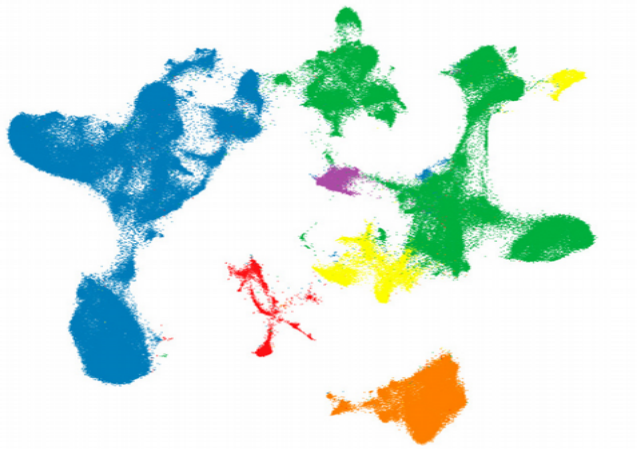
Trajectory Inference

Definition

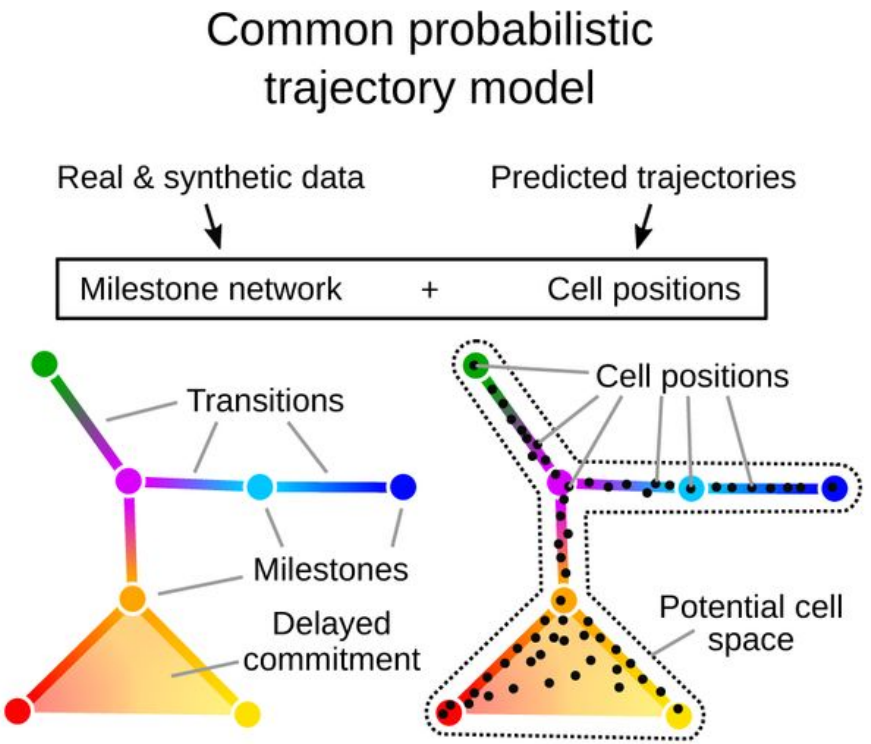
Goal: study cellular dynamic processes (cell cycle, cell differentiation, cell activation...)

How: order cells along a trajectory based on similarities in their expression patterns.

UMAP

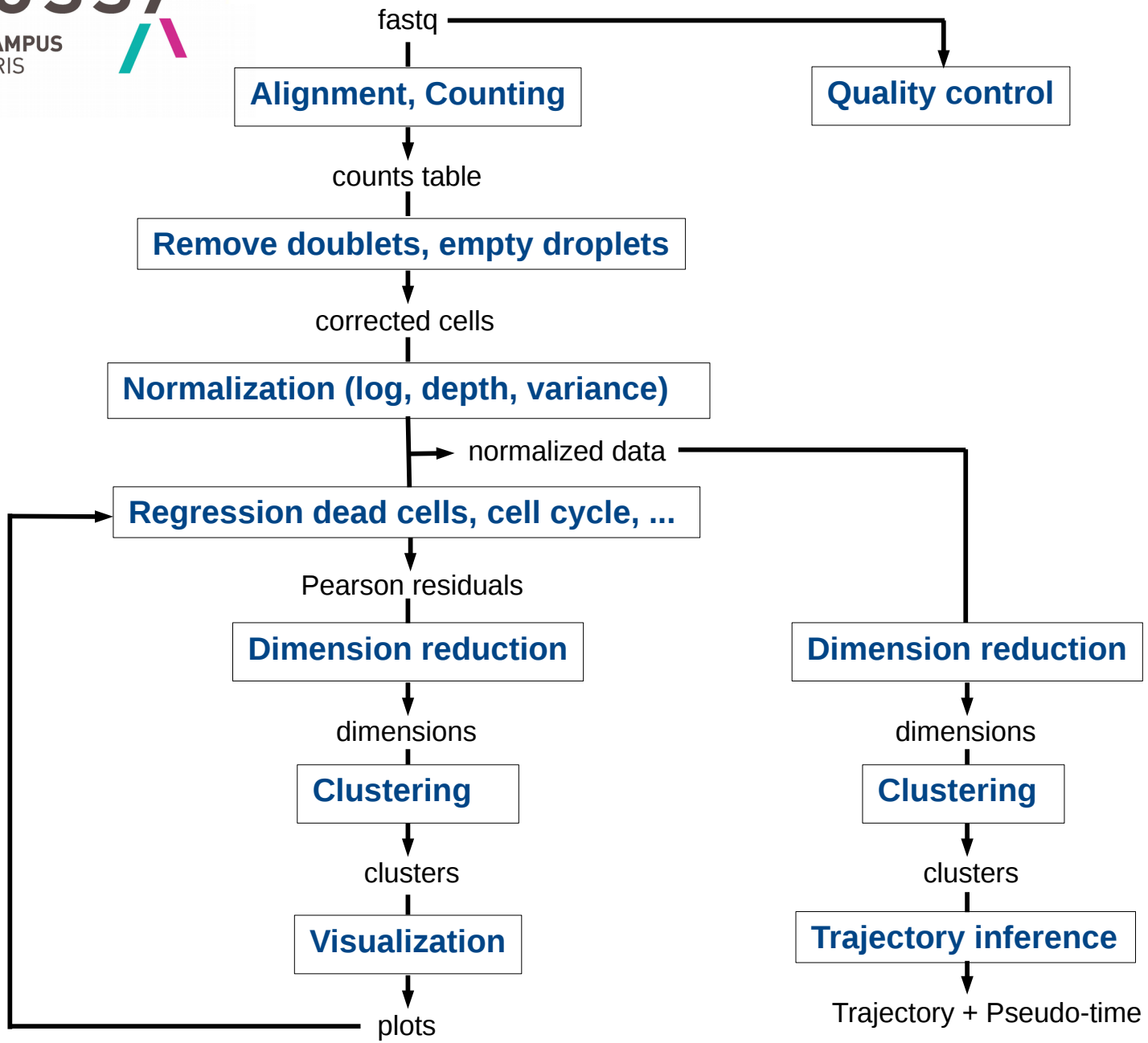


Pseudotimes: a one-dimensional variable representing each cell's transcriptional progression toward the terminal state.



Saelens and al. 2018, bioRχiv

Trajectory Inference



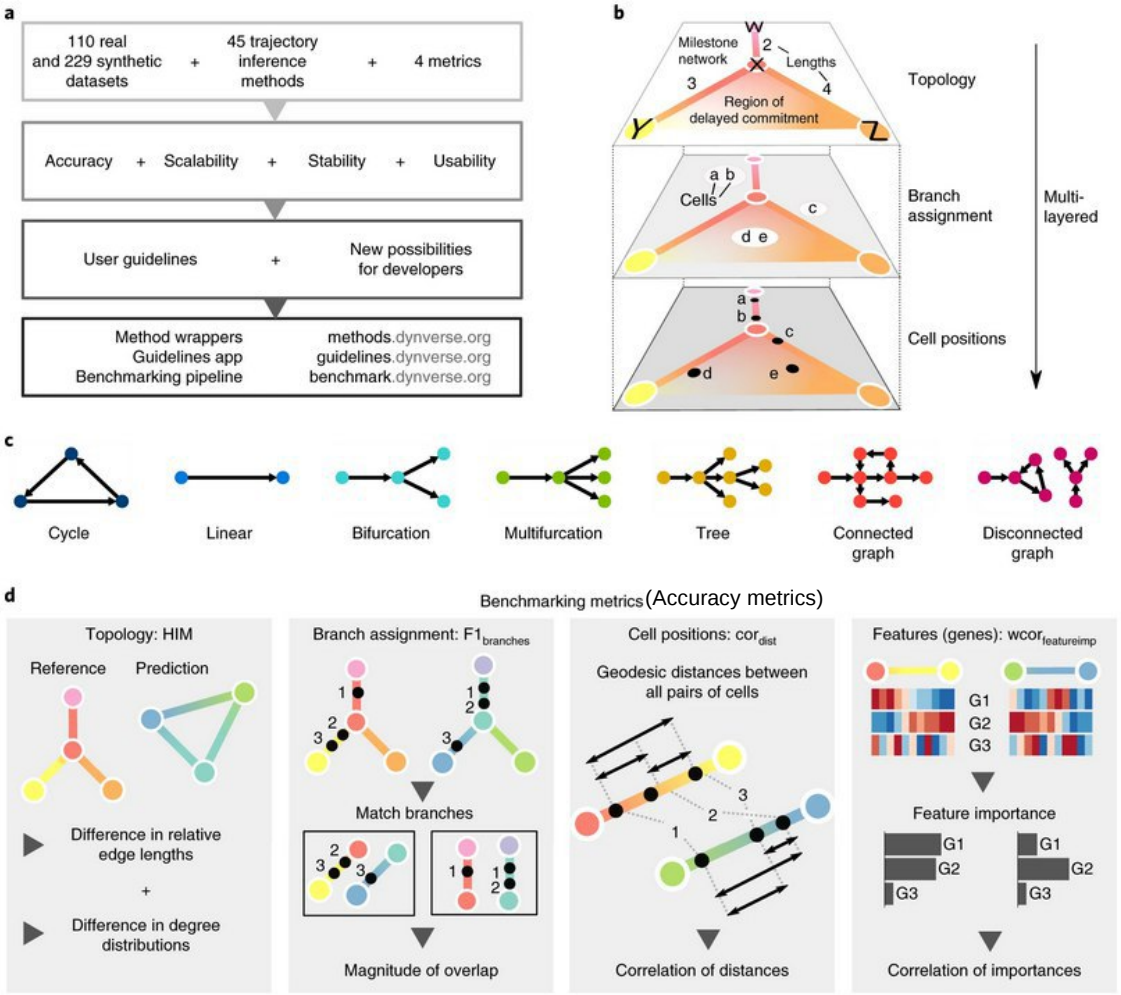
Trajectory Inference

Choice of tools

Saelens and al. 2019, nature biotechnology

- 45 approaches

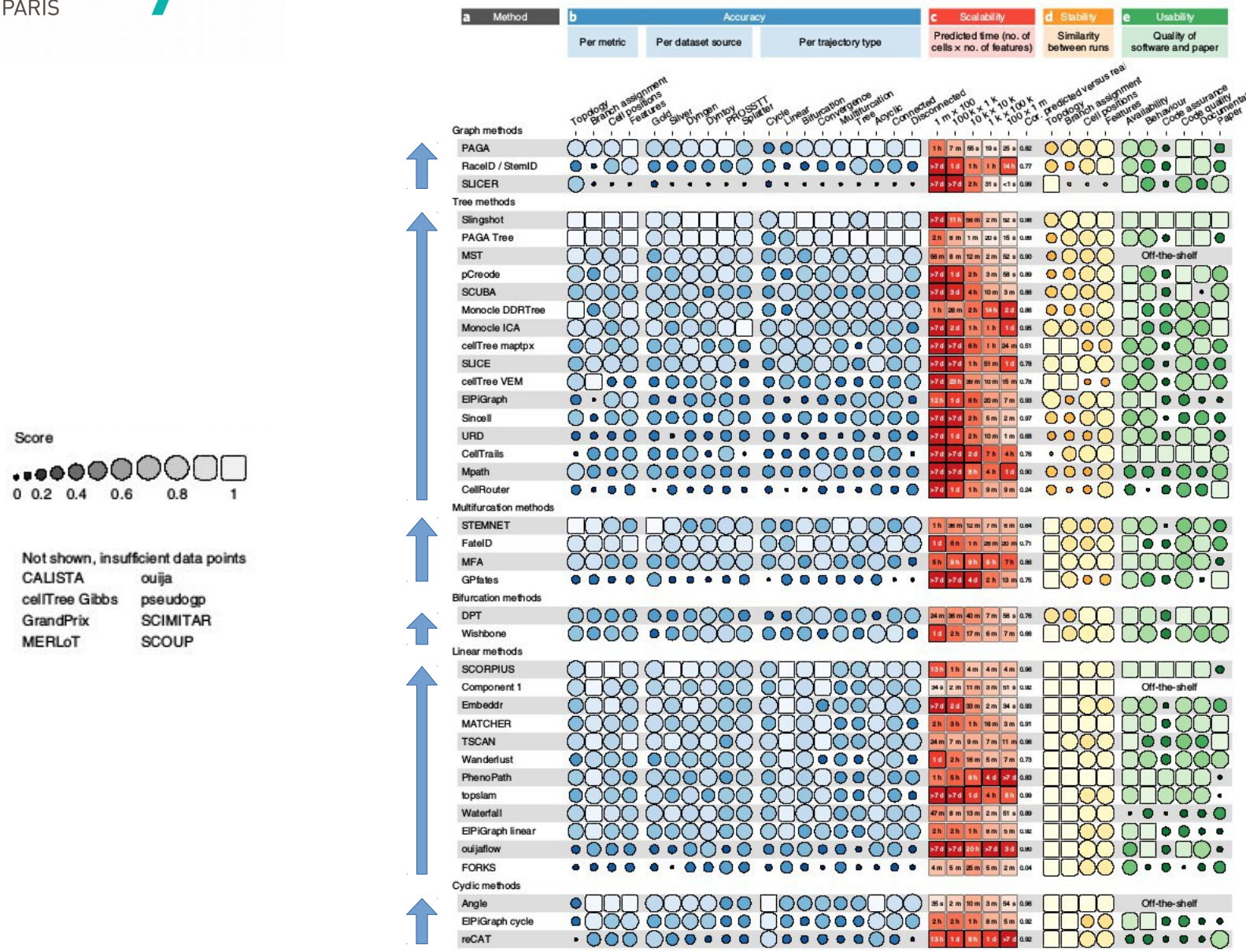
- 110 datasets + 229 synthetic datasets



GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

Trajectory Inference

Choice of tools



Score

0 0.2 0.4 0.6 0.8 1

Not shown, insufficient data points

CALISTA	ouija
cellTree Gibbs	pseudogp
GrandPrix	SCIMITAR
MERLoT	SCOUP

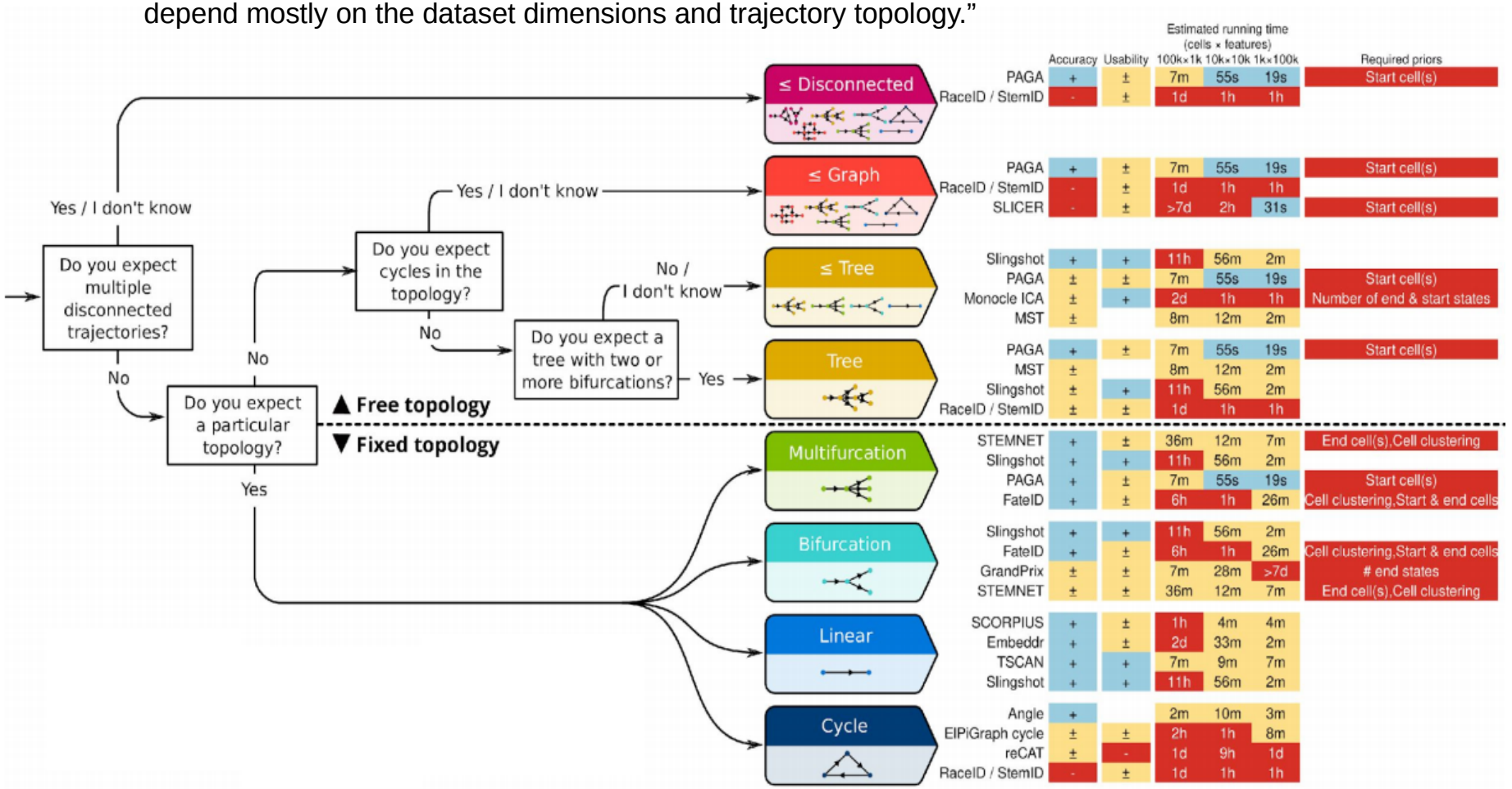
Saelens and al. 2019, nature biotechnology

Trajectory Inference

Choice of tools

Saelens and al. 2019, nature biotechnology

"Our results highlight the complementarity of existing tools, and that the choice of method should depend mostly on the dataset dimensions and trajectory topology."



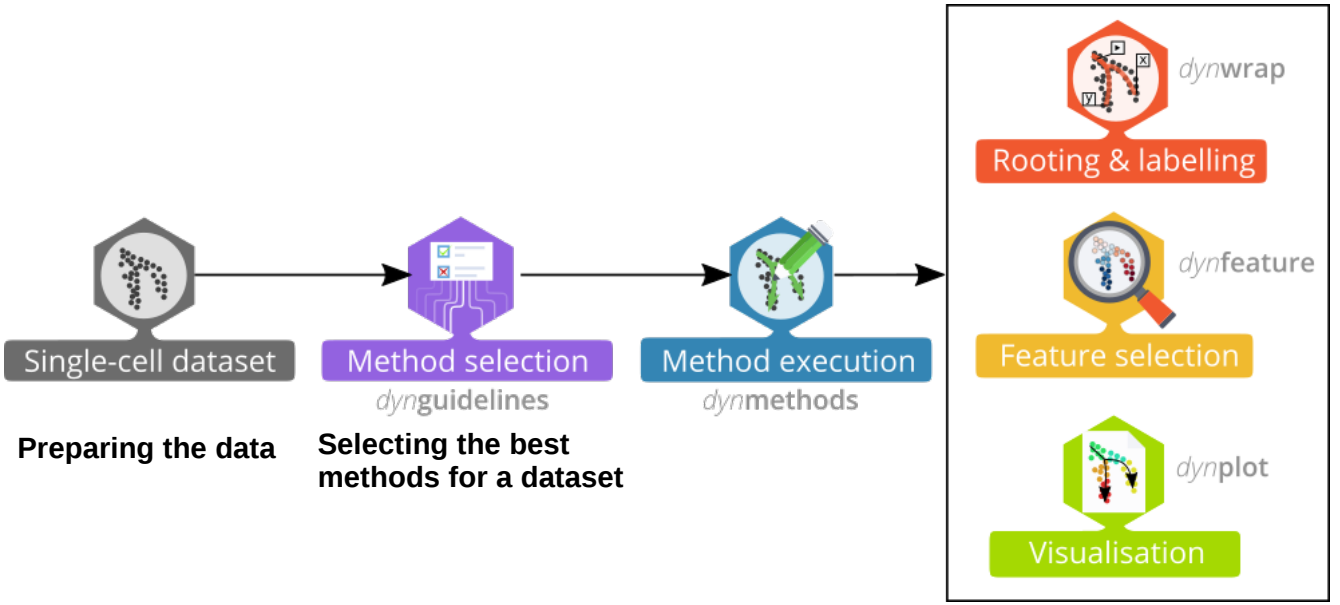
Trajectory Inference

Dynverse

Saelens and al. 2019, nature biotechnology

- 50+ methods

- end-users who want to apply TI on their dataset of interest



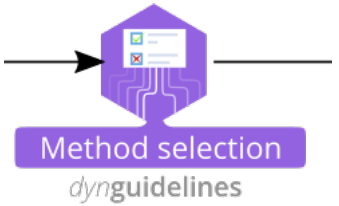
Preparing the data

Selecting the best methods for a dataset

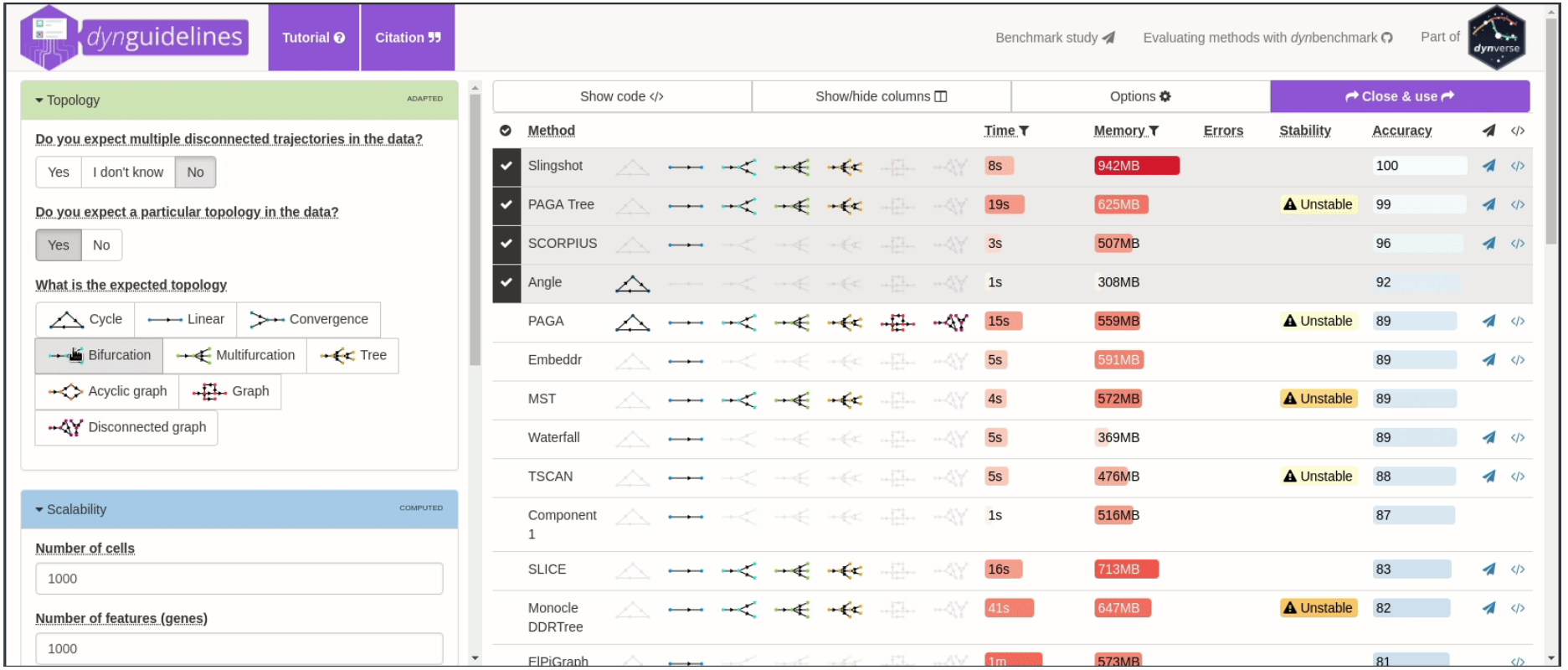
Trajectory Inference

Dynverse

Saelens and al. 2019, nature biotechnology



Selecting the best methods for a dataset



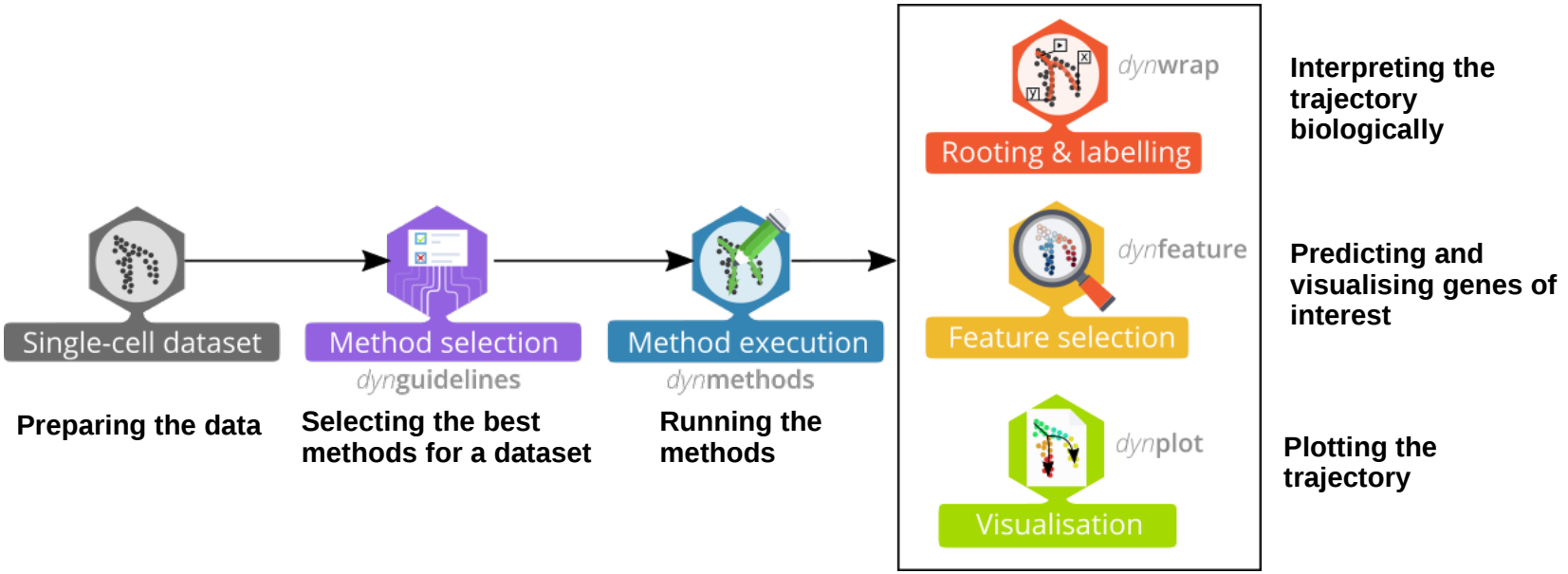
Trajectory Inference

Dynverse

Saelens and al. 2019, nature biotechnology

- 50+ methods

- end-users who want to apply TI on their dataset of interest



Preparing the data

Selecting the best methods for a dataset

Running the methods

Interpreting the trajectory biologically

Predicting and visualising genes of interest

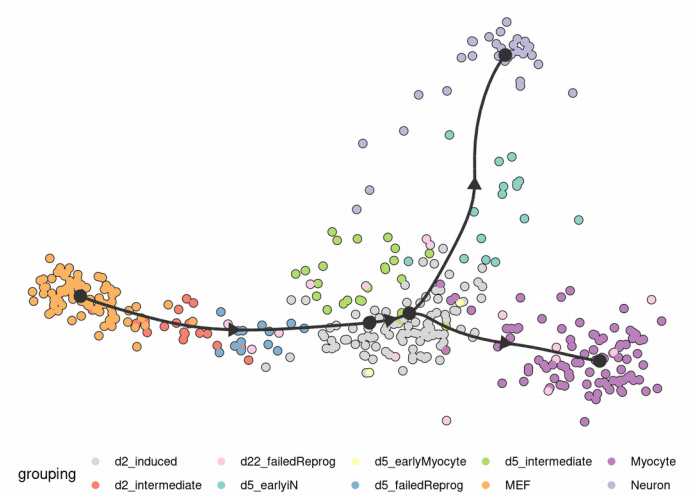
Plotting the trajectory

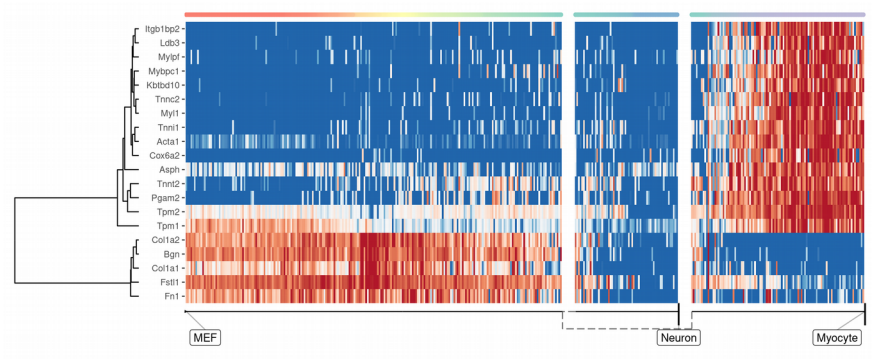
- developers who seek to easily quantify the performance of their TI method and compare it to other TI methods

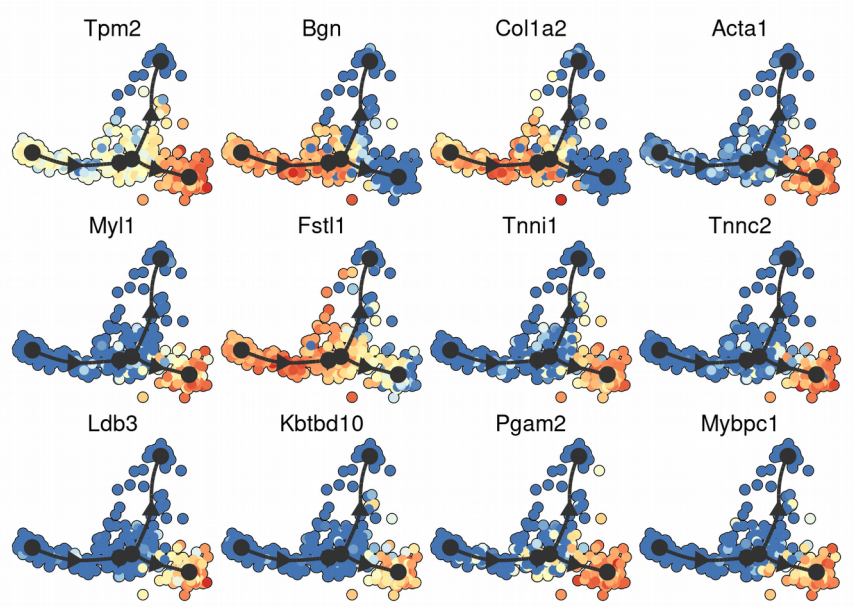
Trajectory Inference

Dynverse

Saelens and al. 2019, nature biotechnology



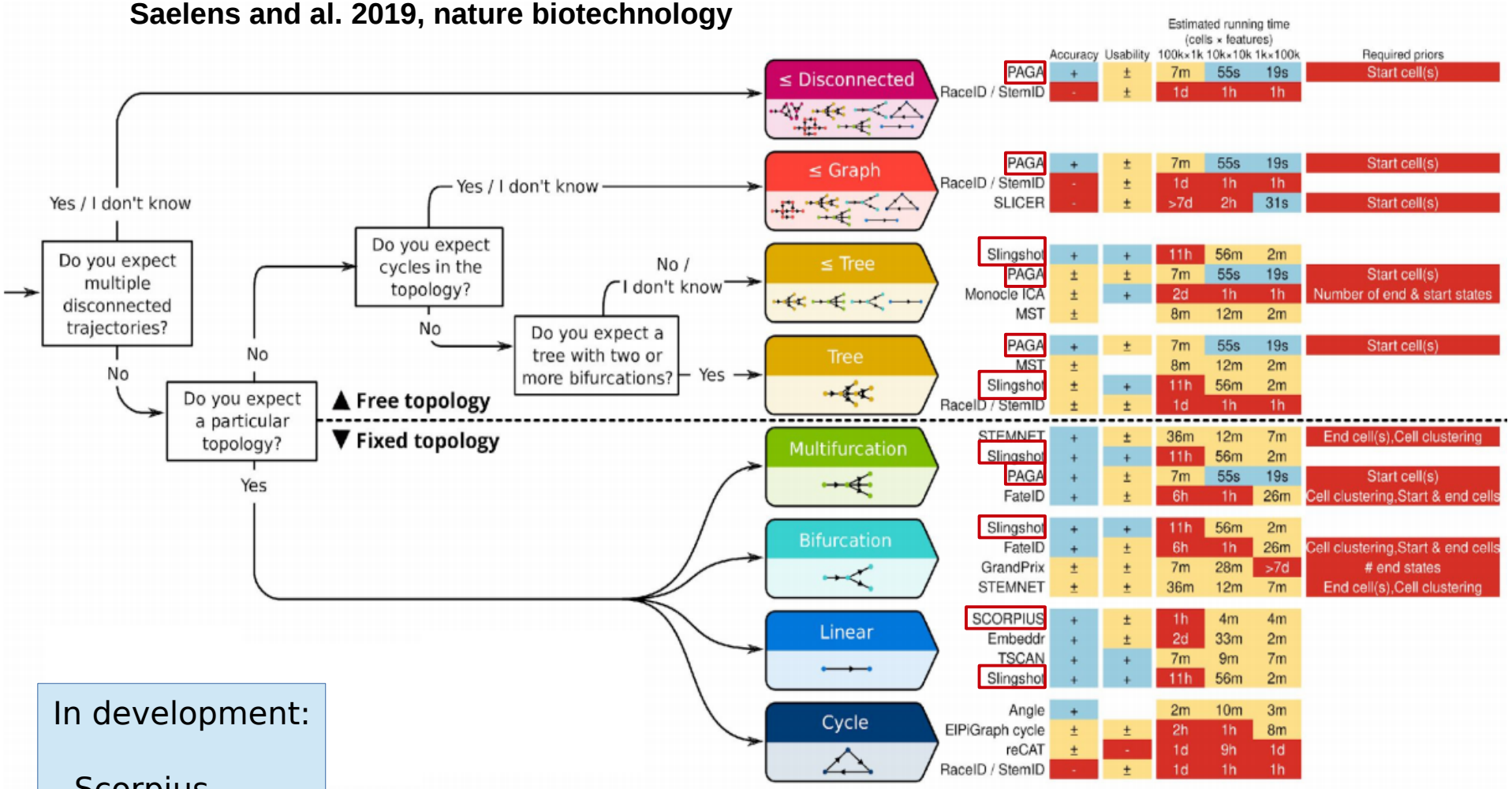




Trajectory Inference

Choice of tools

Saelens and al. 2019, nature biotechnology



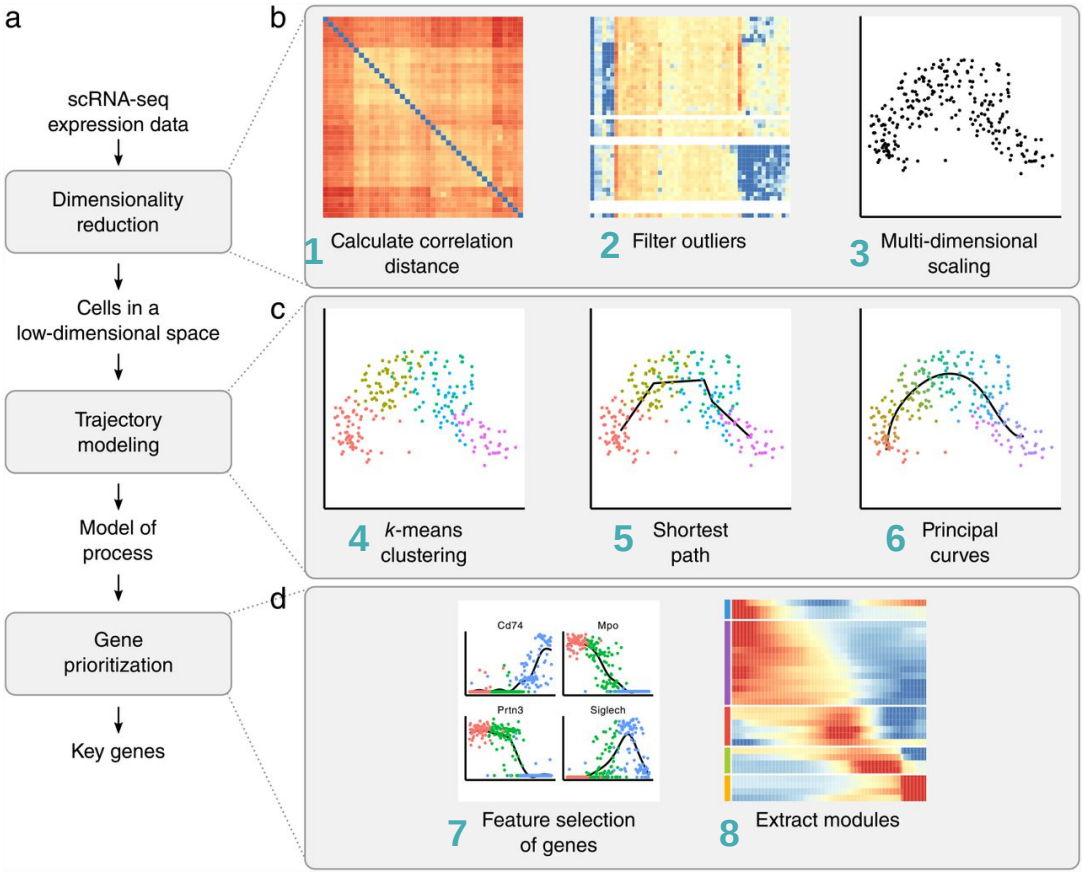
In development:

- Scorpius
- Slingshot
- PAGA

Trajectory Inference

Scorpius

Robrecht Cannoodt et al. 2016, bioRχiv



1) Spearman correlation: the more robust there is noise.

2) Deletion of abberent values.

3) Dimension reduction (MDS: multi-dimensional scaling)⇒ main data structure.

4) K-means clustering.

5) Find the shortest path between the cluster centers.

6) Iterative refinement using the principal curves algorithm, then ordering of cells by projection.

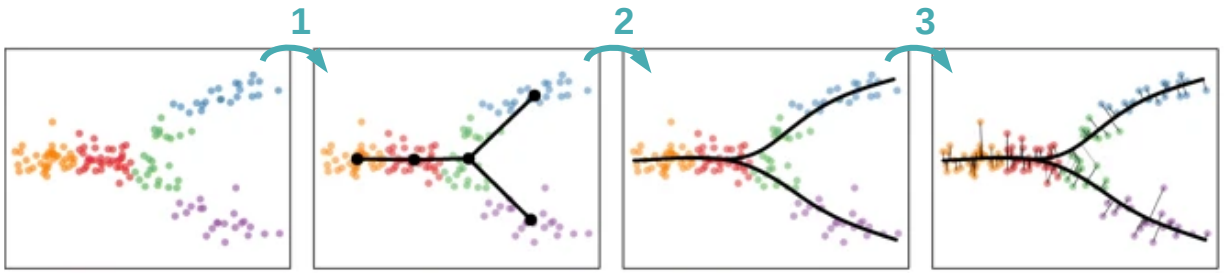
7) Classification of genes according to their ability to predict the order of cells from expression data (algo Random Forest).

8) Grouping of genes into coherent modules + visualization.

Trajectory Inference

Slingshot

Kelly Street et al. 2018, BMC Genomics



Start with reduced (zinbwave) and clustered (RSEC) data.

Identification of lineages:

1) Inference of the global structure: minimum spanning tree (MST) based on clusters => identification of the number of lineages and milestones.

Identification of pseudotimes:

2) Fitting of curves by simultaneous principal curves.

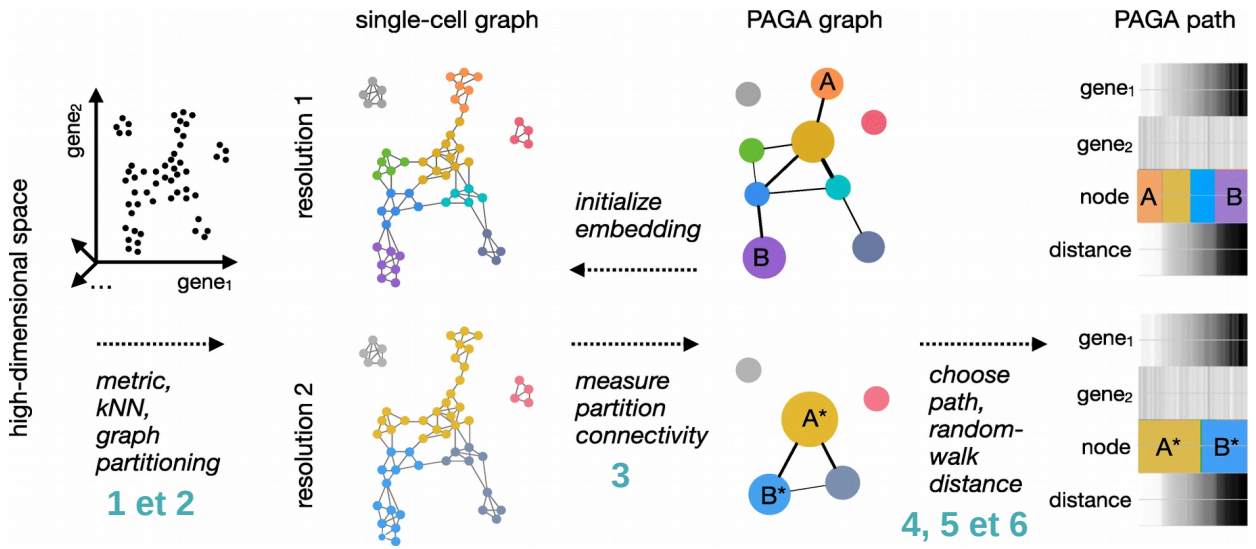
3) Pseudotime values are obtained by orthogonal projection onto the curves.

Trajectory Inference

PAGA (Partition-based Graph Abstraction)

F. Alexander Wolf et al. 2019, Genome Biology

Groups as connected if their number of inter-edges exceeds a fraction of the number of inter-edges expected under random assignment.



1) Dimension reduction (PCA) + Euclidean distance + representation of the data in the form of a kNN graph.

2) Partitioning the graph at a desired resolution (partitions = groups of connected cells = clusters) thanks to the Louvain algorithm

3) PAGA graph: 1 node / partition + connection of the nodes to each other with a connectivity weight (sort of probability of connection)

4) Removal of connections if weak connectivity (noise)

5) Order cells in each partition (pseudo-time) = distance measurement based on random walk on the single-cell graph compared to a starting cell + to the remaining connections in the PAGA graph.

6) Followed changes in gene expression along trajectories, averaging all of the single-cell paths through the corresponding groups of cells.

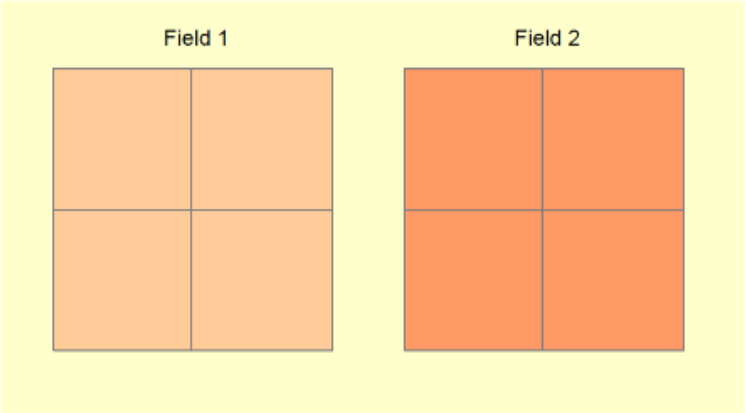
Definitions

integration: combination of multiple single-cell RNA-seq datasets into one.

experimental design

We wish to compare two groups of plants,
before and after treatment.

There are two fields and 4 zones per field.

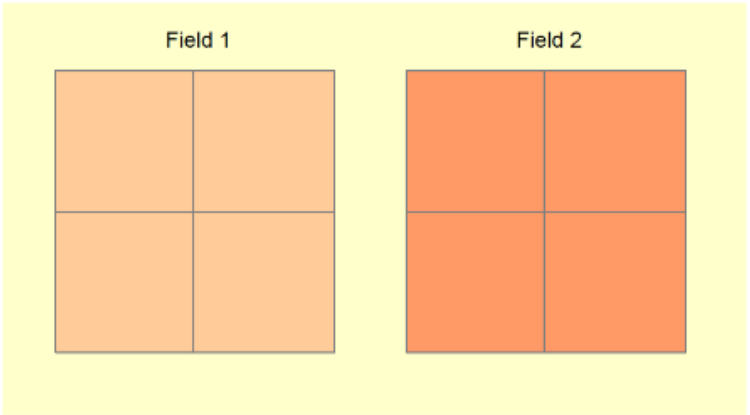


Definition

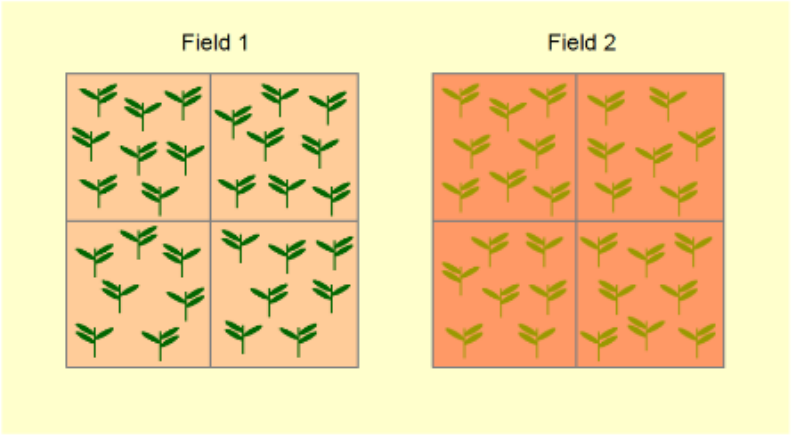
experimental design

We wish to compare two groups of plants, before and after treatment.

There are two fields and 4 zones per field.



4 replicates / group



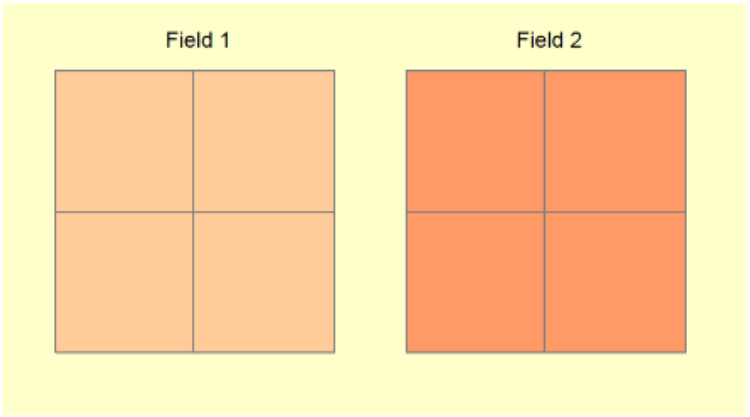
Is this experimental design good?

Definition

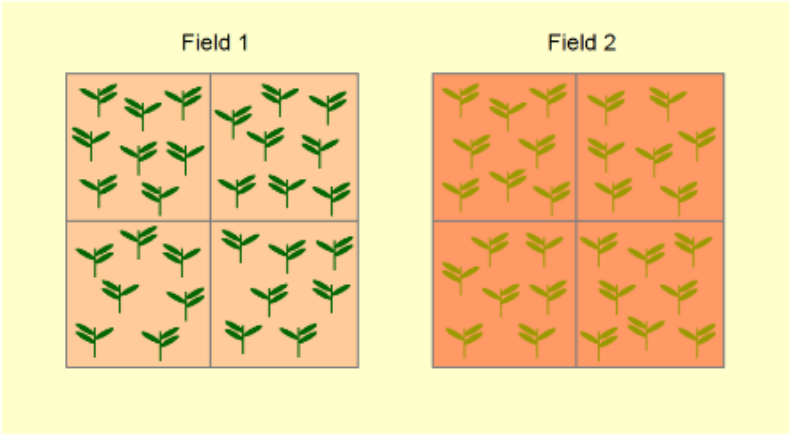
experimental design

We wish to compare two groups of plants, before and after treatment.

There are two fields and 4 zones per field.



4 replicates / group



Is this experimental design good?

No, confounding factors:

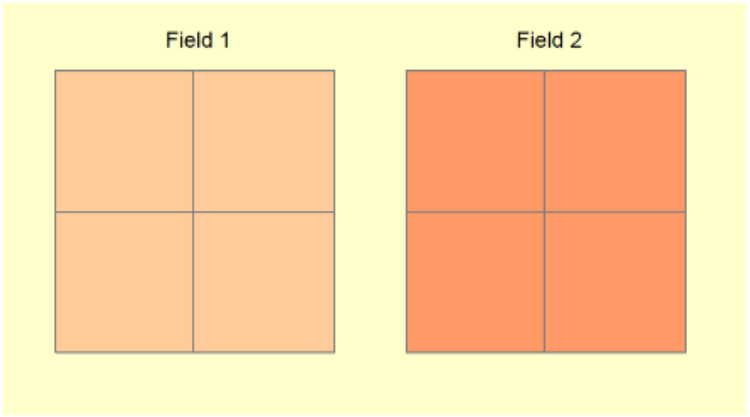
- "Experimental condition" factor: control or treatment
- "Environment" factor: Field 1 or Field 2

Definition

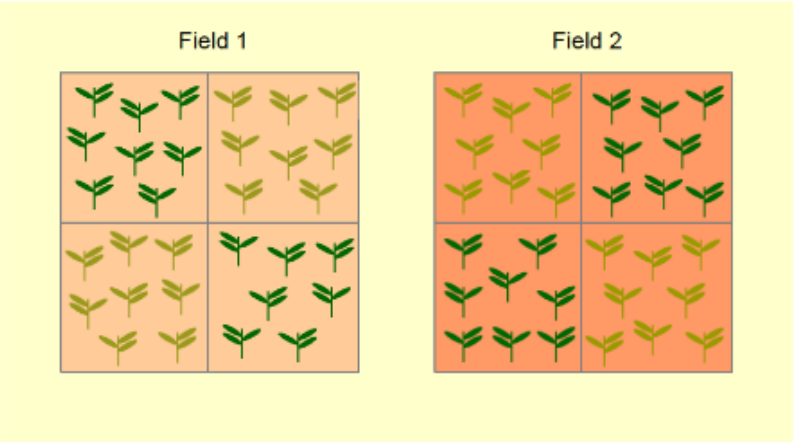
experimental design

We wish to compare two groups of plants, before and after treatment.

There are two fields and 4 zones per field.



4 replicates / group



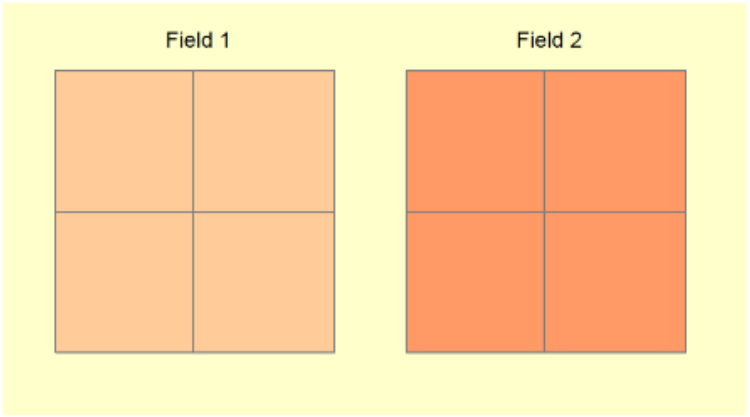
Is this experimental design good?

Definition

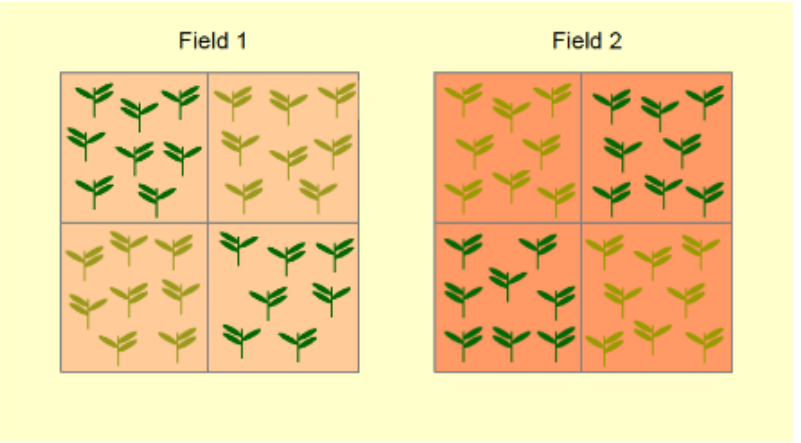
experimental design

We wish to compare two groups of plants, before and after treatment.

There are two fields and 4 zones per field.



4 replicates / group



Is this experimental design good?

Yes, the "environment" factor and the "condition" factor are no longer confused!

We can then estimate the effect of the treatment.

Definition

experimental design

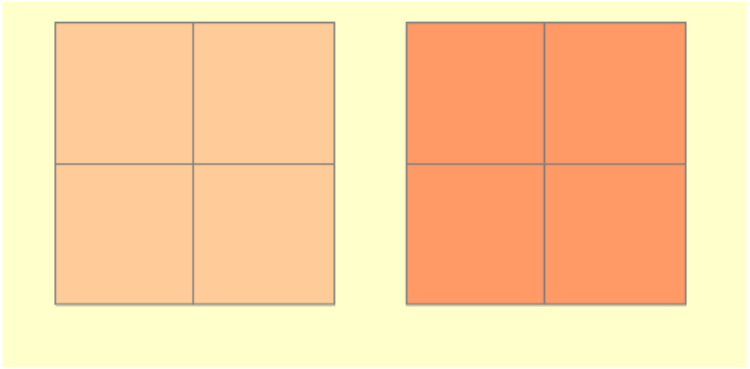
We have 2 patients, same tissue

Patient 1

Patient 2

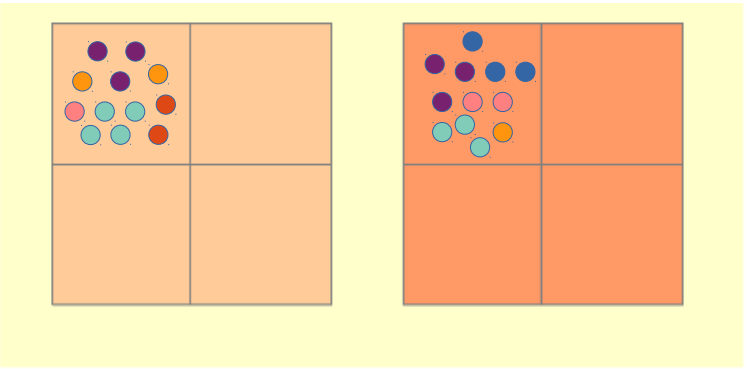
There are two plates and 4 zones per plate.

Flow cell 1 Flow cell 2



1 replicates / group

Flow cell 1 Flow cell 2



Is this experimental design good?

Integration

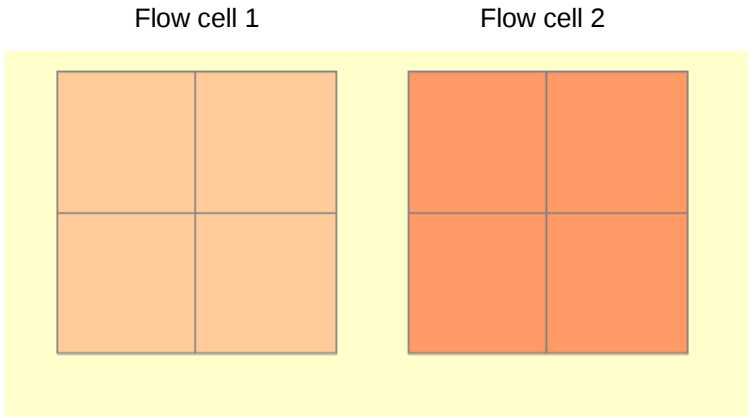
Definition

Most batch correction algorithms require **at least one identical cell type to be shared between any pair of data batches**, to guide the data alignment.

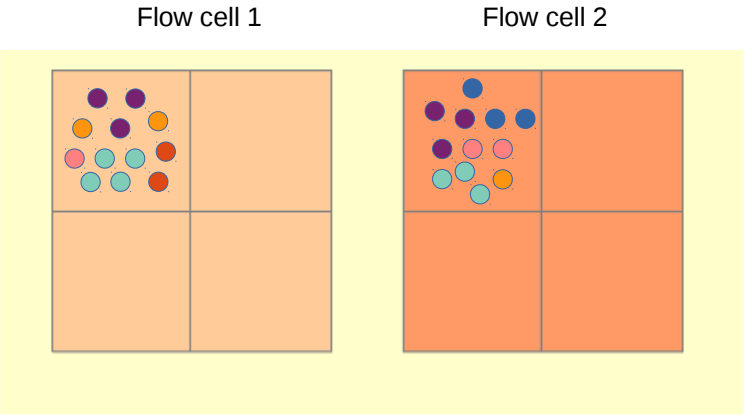
experimental design

We have 2 patients, same tissue

There are two plates and 4 zones per plate.



1 replicates / group



Is this experimental design good?

Yes, if:

-there is at least one cell population that is present in both groups

- batch effect variation is much smaller than the biological-effect variation between different cell types

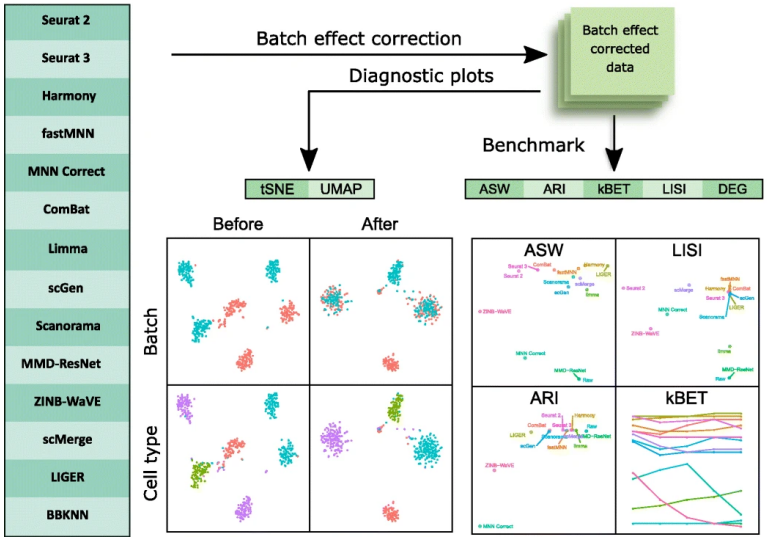
Integration

Choice of tools

Tran et al. 2020, Genome Biology

- 14 methods
- 10 datasets representing 5 integration scenarios
  - Identical cell types but with different technologies
  - Mix of cell types (min 1 type is shared)
  - Multiple batch effect
  - Big data (<500 000 cells)
  - Simulations for DE analysis
- Eval
  - Computational runtime
  - Ability to handle large datasets
  - Batch-effect correction efficacy

A



B

Dataset	Description	Number of batches	Total cell number	Technologies
1	Human Dendritic Cells	2	576	Smart-Seq2
2	Mouse Cell Atlas	2	6,954	Microwell-Seq
				Smart-Seq2
3	Simulation	Refer to Simulation table		
4	Human Pancreas	5	14,767	inDrop
				CEL-Seq2
				Smart-Seq2
				SMARTer
				SMARTer
5	Human Peripheral Blood Mononuclear Cell	2	15,476	10x 3'
				10x 5'
6	Cell line	3	9,530	10x
7	Mouse Retina	2	71,638	Drop-seq
8	Mouse Brain	2	833,206	Drop-seq
				SPLiT-seq
9	Human Cell Atlas	2	621,466	10x
10	Mouse Haematopoietic Stem and Progenitor Cells	2	4,649	MARS-seq
				Smart-Seq2

Integration

Choice of tools

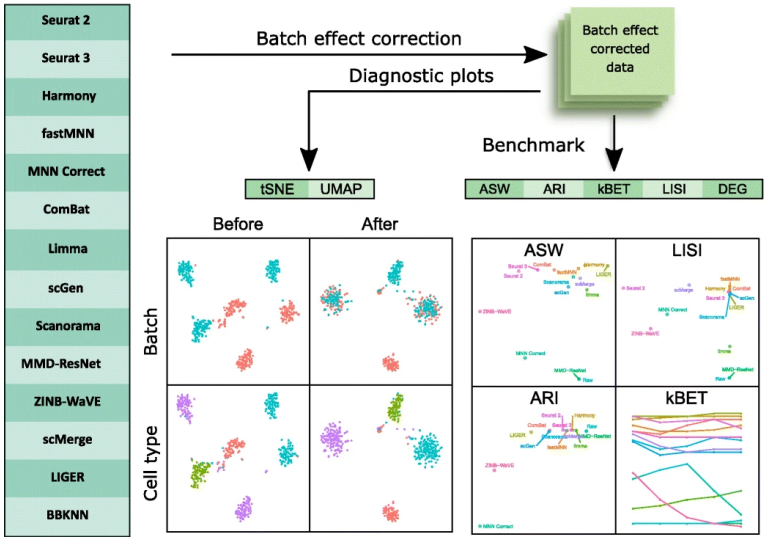
Tran et al. 2020, Genome Biology

- 14 methods
- 10 datasets representing 5 integration scenarios
  - Identical cell types but with different technologies
  - Mix of cell types (min 1 type is shared)
  - Multiple batch effect
  - Big data (<500 000 cells)
  - Simulations for DE analysis
- Eval
  - Computational runtime
  - Ability to handle large datasets
  - Batch-effect correction efficacy

In development:

- Seurat 3
- Harmony
- LIGER

A

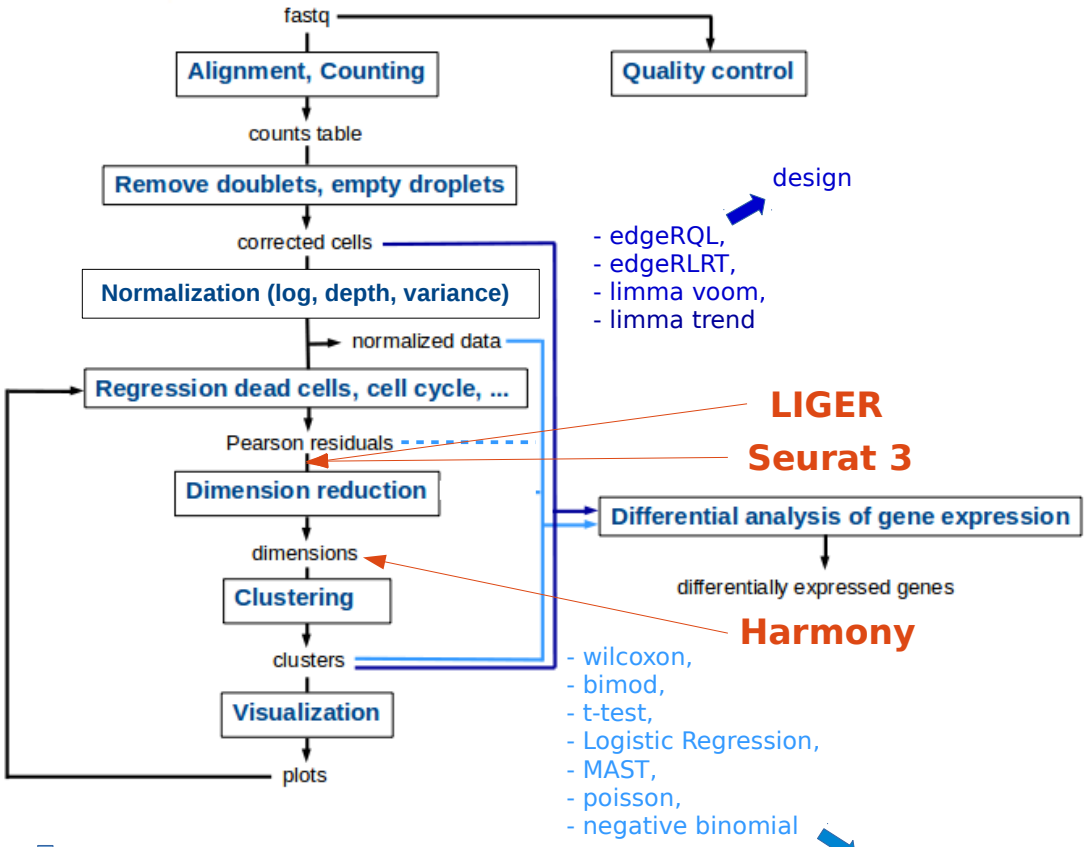


B

Dataset	Description	Number of batches	Total cell number	Technologies
1	Human Dendritic Cells	2	576	Smart-Seq2
2	Mouse Cell Atlas	2	6,954	Microwell-Seq
				Smart-Seq2
3	Simulation	Refer to Simulation table		
4	Human Pancreas	5	14,767	inDrop
				CEL-Seq2
				Smart-Seq2
				SMARTer
				SMARTer
5	Human Peripheral Blood Mononuclear Cell	2	15,476	10x 3'
				10x 5'
6	Cell line	3	9,530	10x
7	Mouse Retina	2	71,638	Drop-seq
8	Mouse Brain	2	833,206	Drop-seq
				SPLiT-seq
9	Human Cell Atlas	2	621,466	10x
10	Mouse Haematopoietic Stem and Progenitor Cells	2	4,649	MARS-seq
				Smart-Seq2

Impact on previous stages

Differential expression analysis

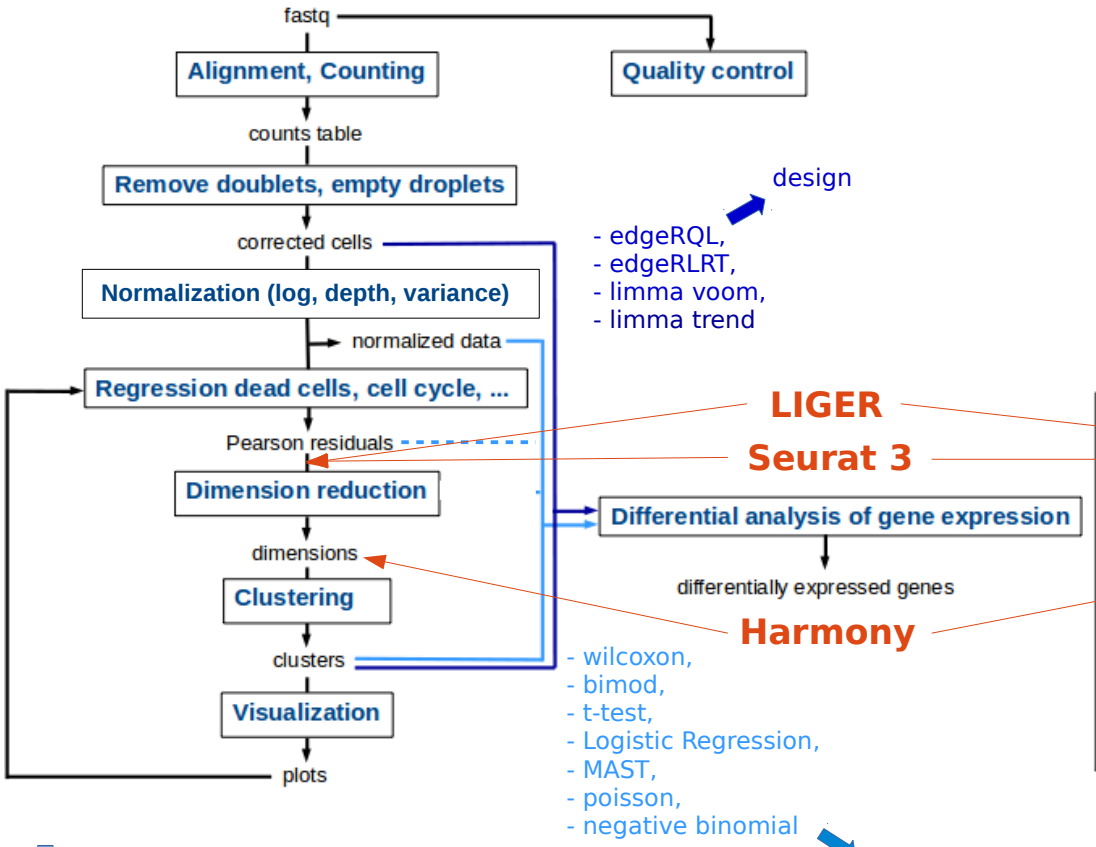


Influences the identification of clusters, but DE is performed on non-integrated data

Integration

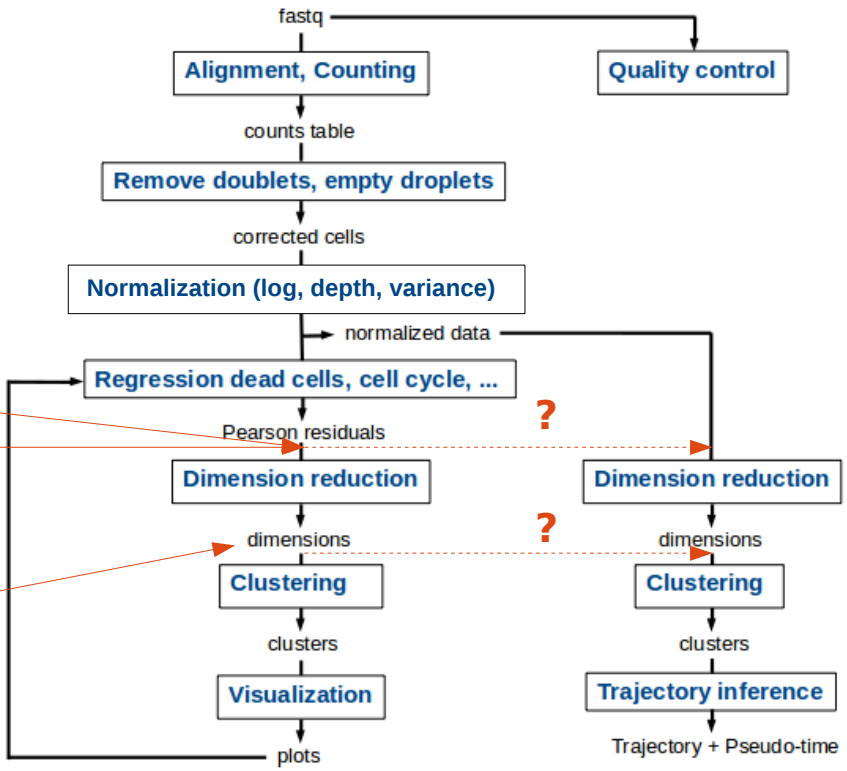
Impact on previous stages

Differential expression analysis



Influences the identification of clusters, but DE is performed on non-integrated data

Trajectory inference



Impact in investigation

Thank you for your attention