Geteiltes Leid ist halbes Leid

Common difficulties in analysis and interpretation of single-cell RNA-seq data

2025-03-25 ♦ Niko Papadopoulos

Other good resources

Current best practices in single-cell RNA-seq analysis: a tutorial
Luecken & Theis, Molecular Systems Biology, 2019

Single-cell best practices book
(maintained & expanded by the Theis lab and the single-cell community)
https://www.sc-best-practices.org/preamble.html

scRNA-seq analysis: overview*

scRNA-seq analysis: overview*

* most basic example possible

scRNA-seq analysis: overview*

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

* most basic example possible

scRNA-seq analysis: overview*

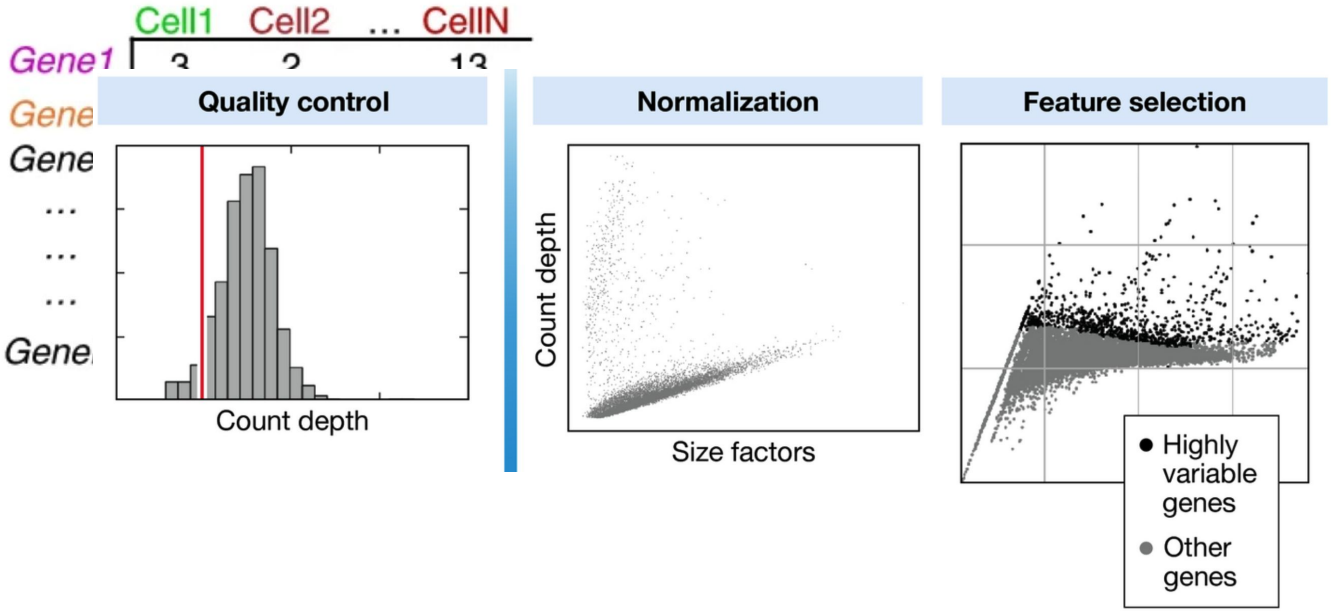
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

scripting

programming

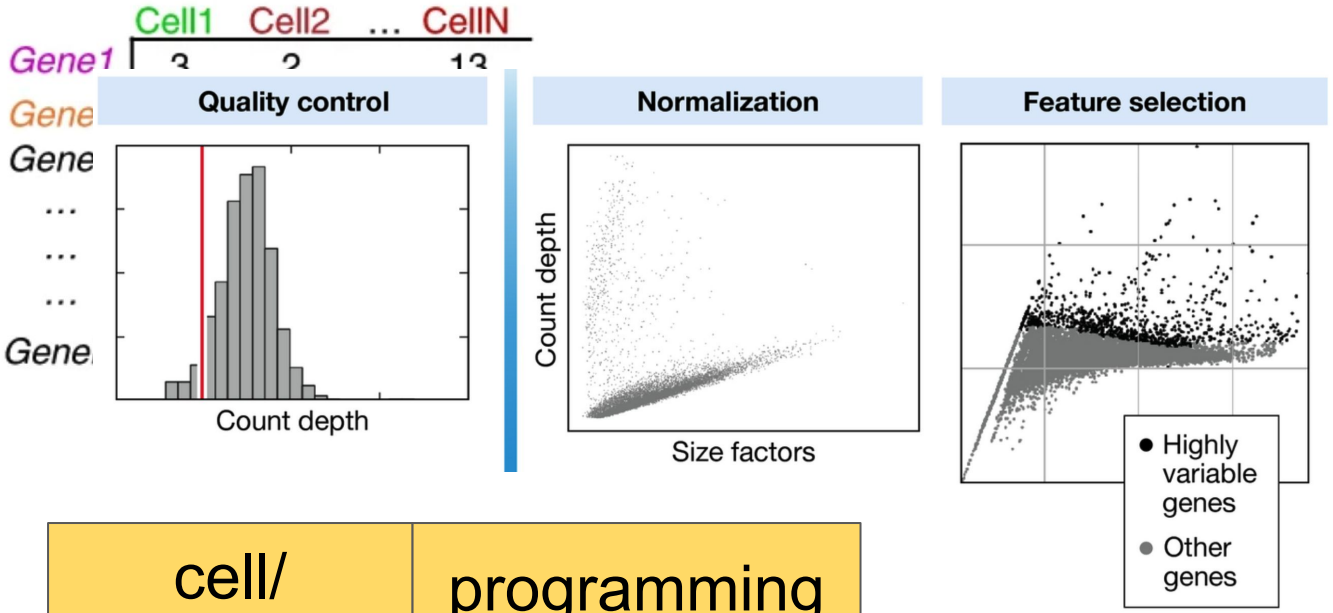
* most basic example possible

scRNA-seq analysis: overview*



* most basic example possible

scRNA-seq analysis: overview*



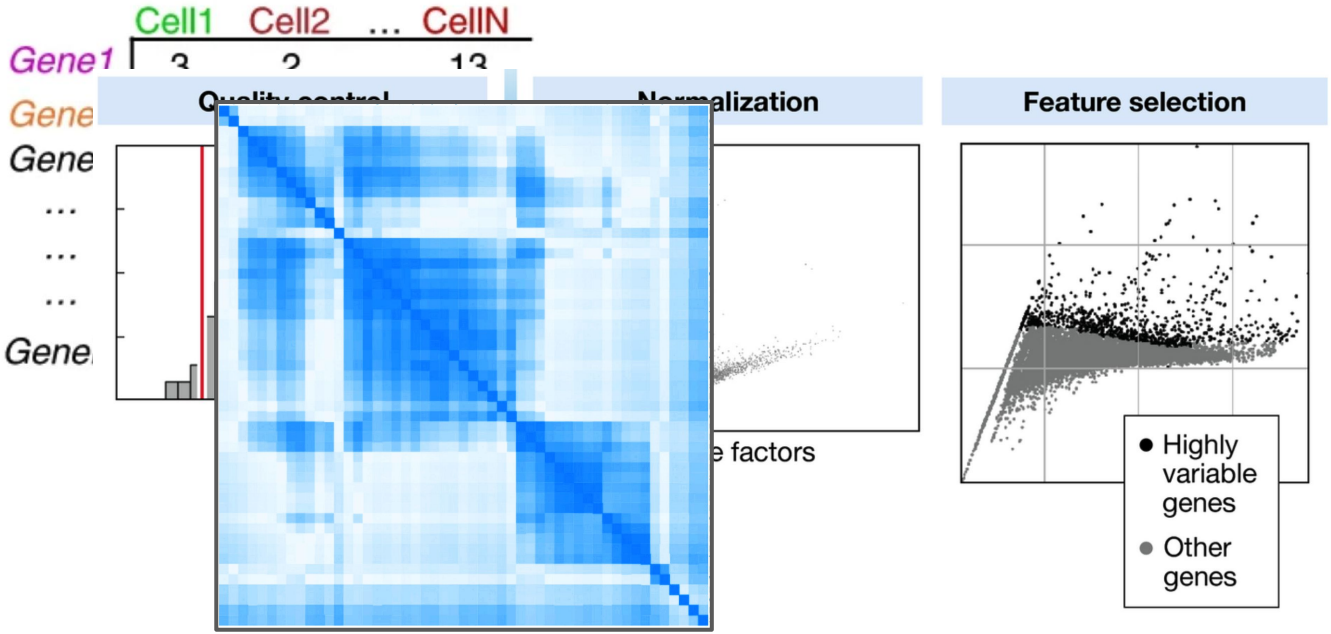
cell/ molecular biology

programming

statistics

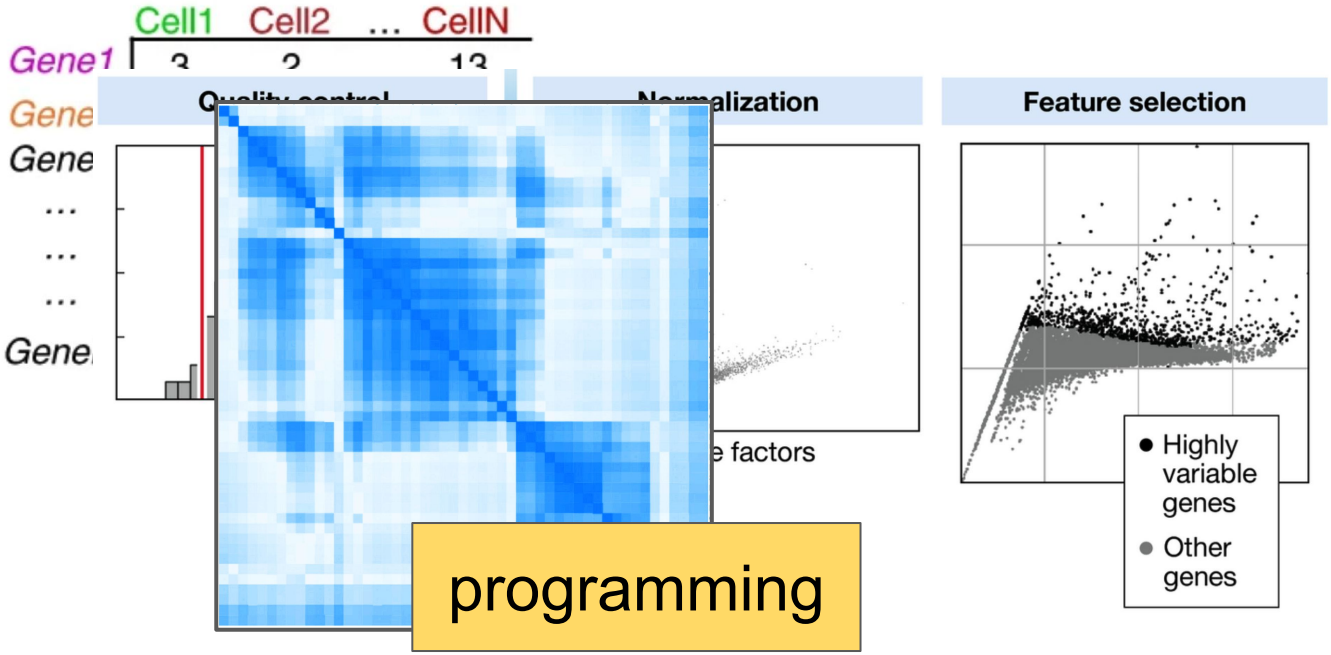
* most basic example possible

scRNA-seq analysis: overview*



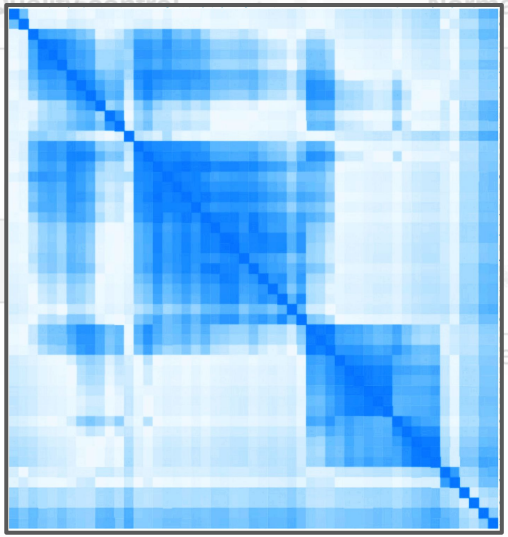
* most basic example possible

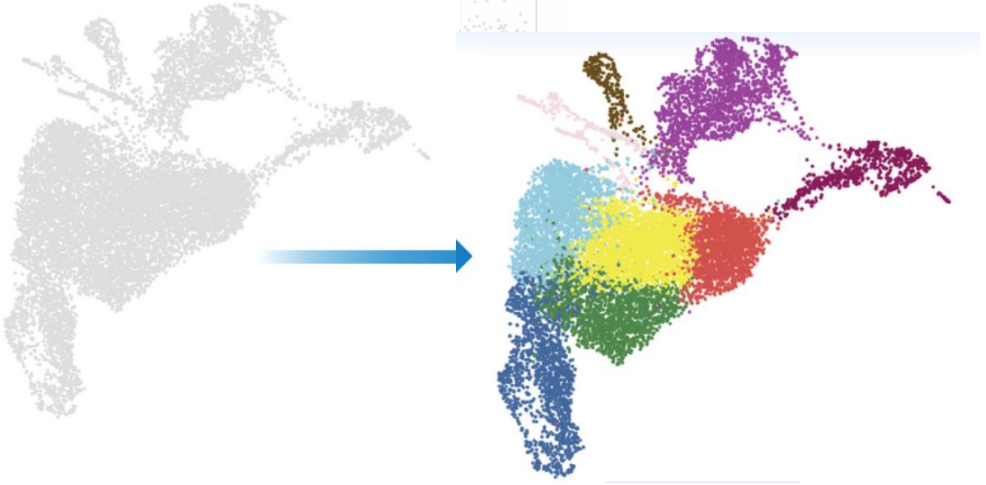
scRNA-seq analysis: overview*



* most basic example possible

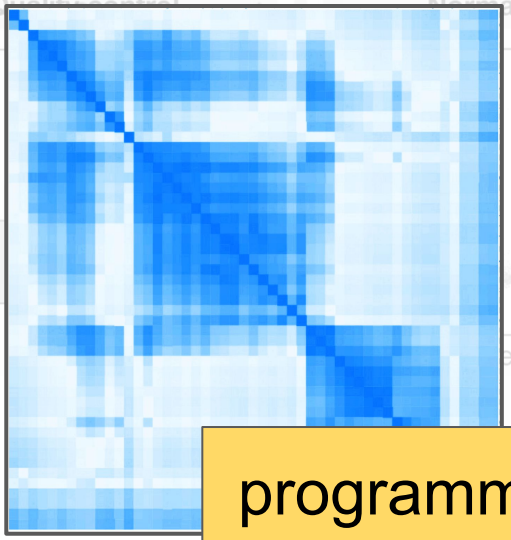
scRNA-seq analysis: overview*





* most basic example possible

scRNA-seq analysis: overview*

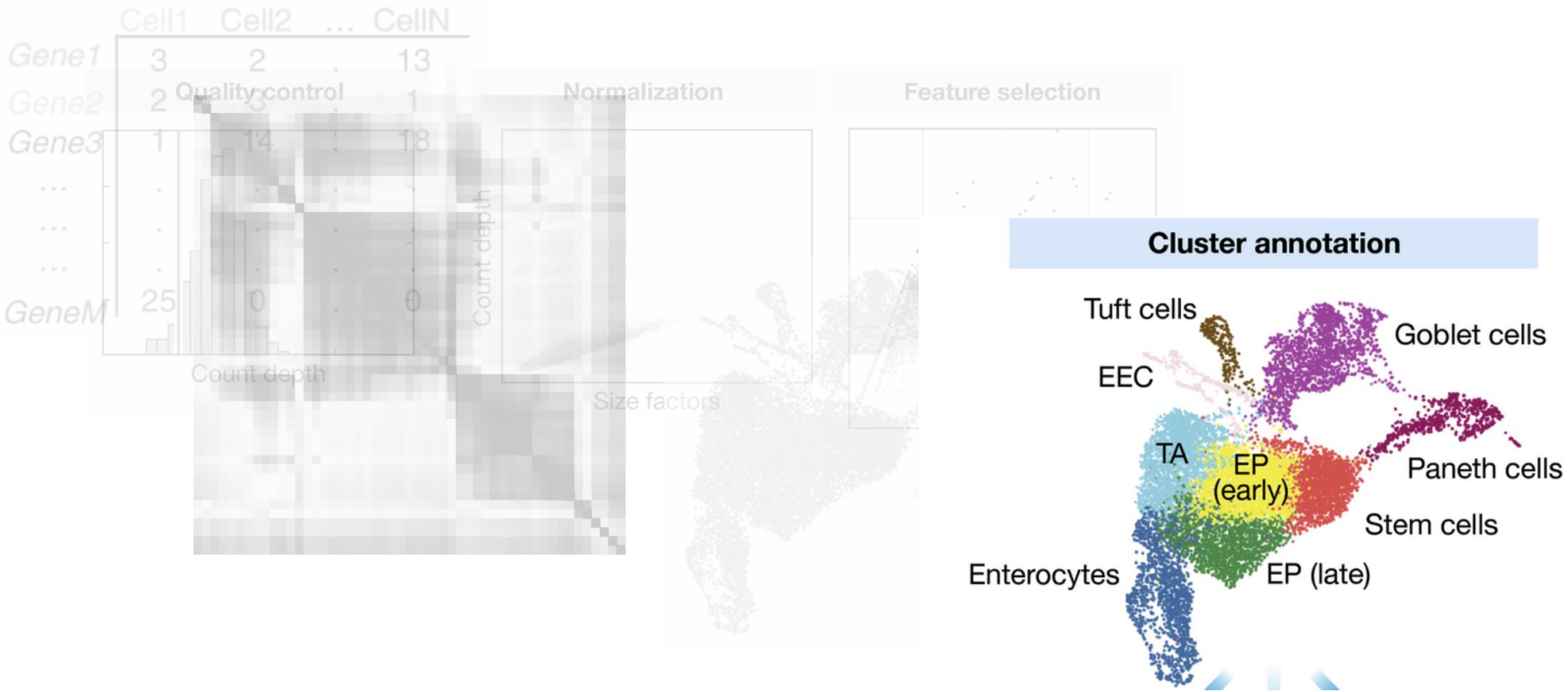


programming

<math>

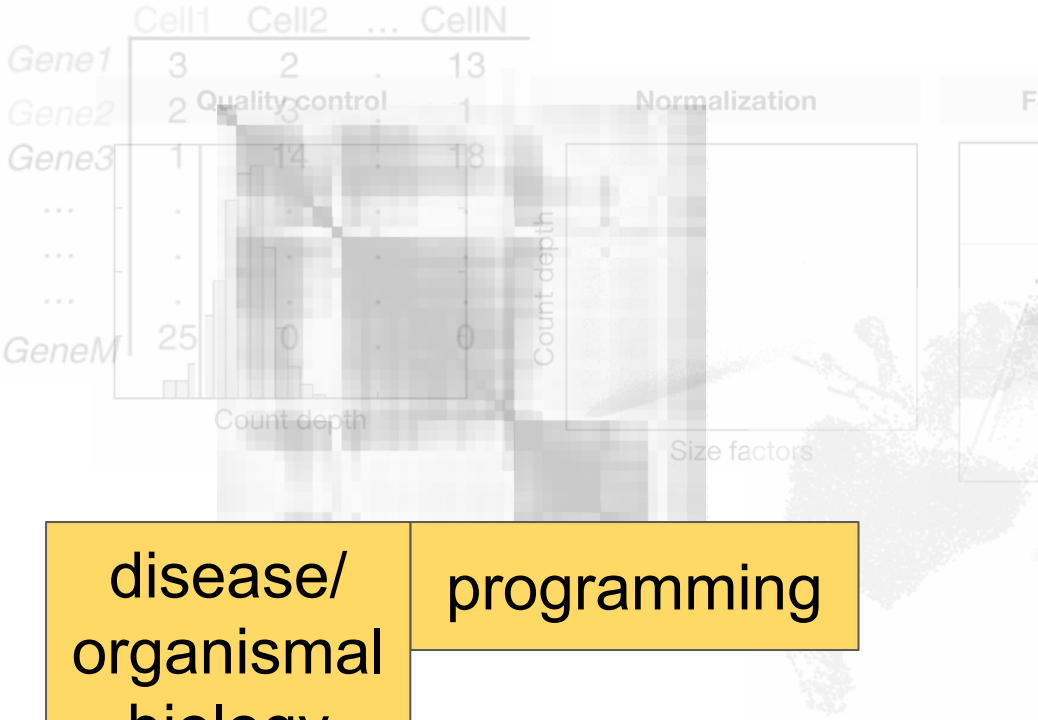
* most basic example possible

scRNA-seq analysis: overview*



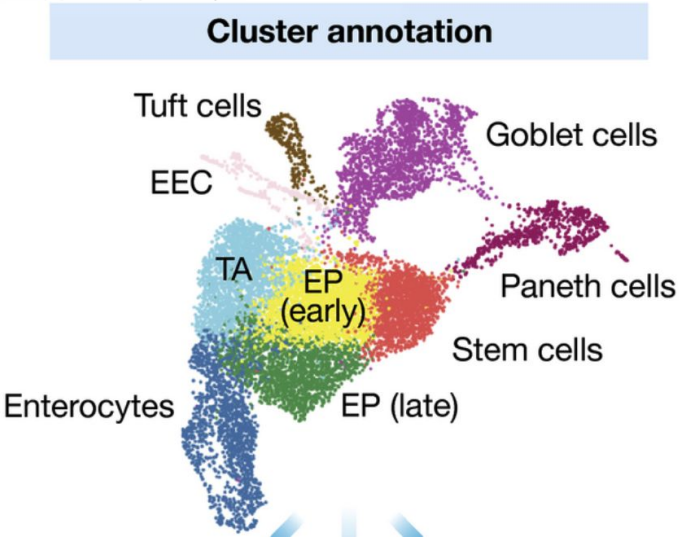
* most basic example possible

scRNA-seq analysis: overview*



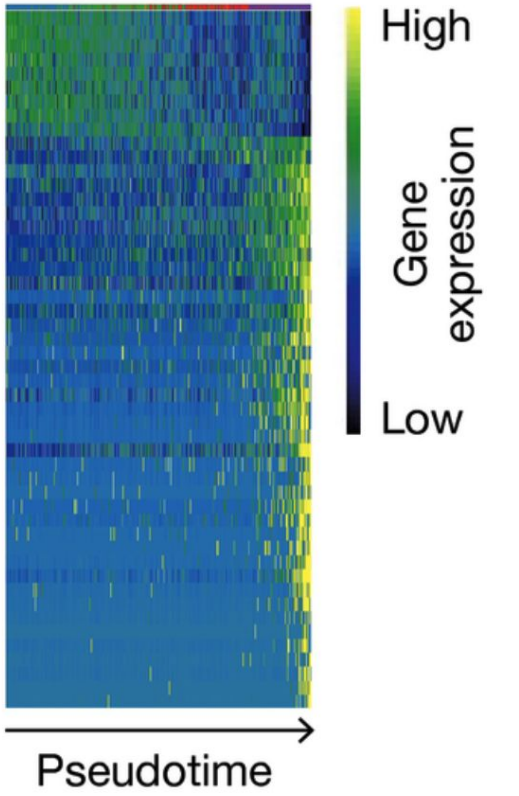
disease/ organismal biology

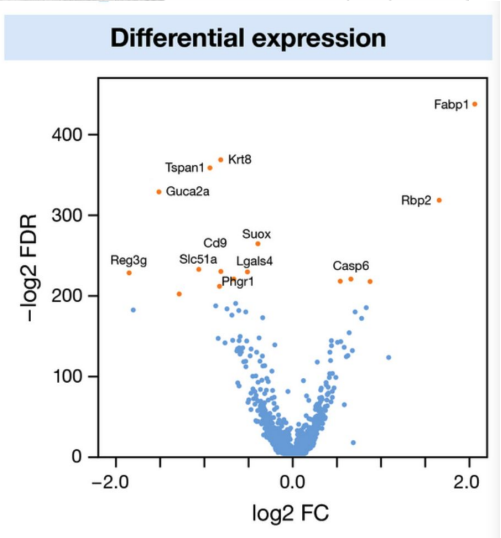
programming

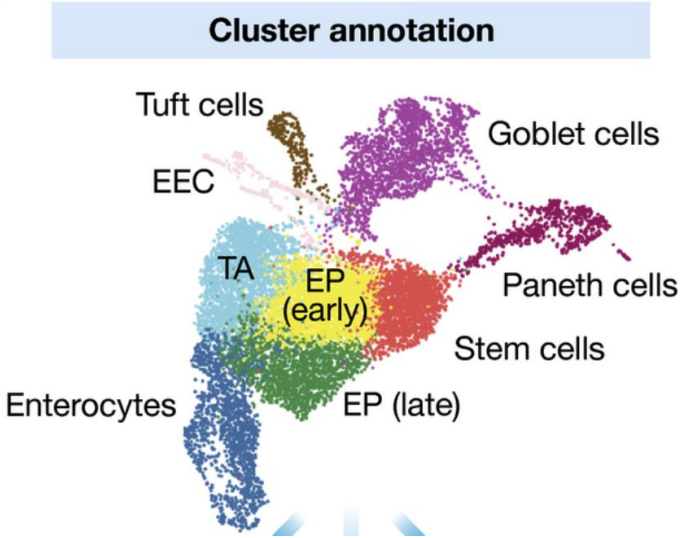


* most basic example possible

scRNA-seq analysis: overview*

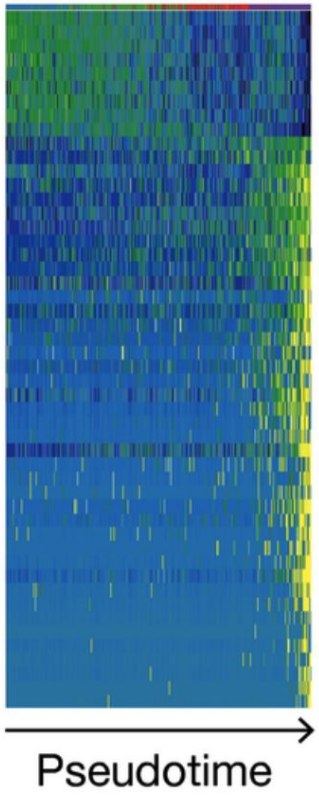






* most basic example possible

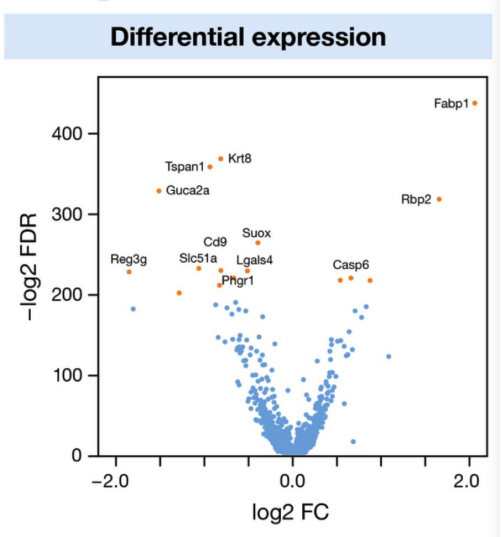
scRNA-seq analysis: overview*

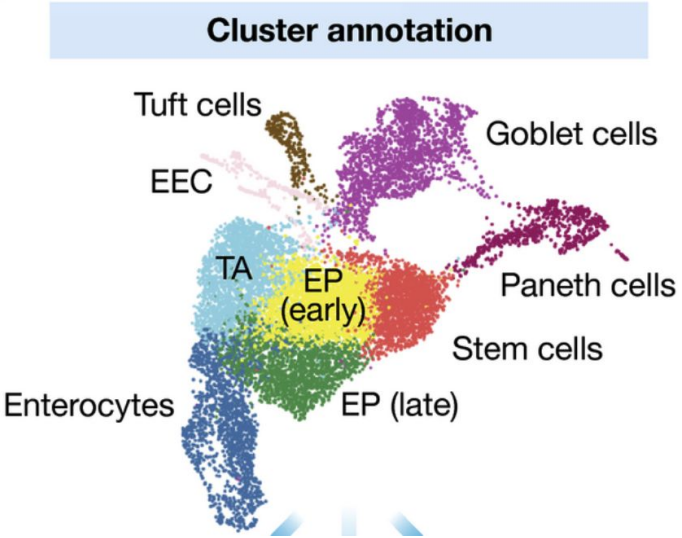


disease/ organismal biology

programming

statistics





* most basic example possible

Why is scRNA-seq analysis so hard?

- To get the most out of the data you (probably) need non-standard analysis
- Not clear what standard analysis is anyway
- Too many knobs to turn → analysis paralysis
- Often: biological questions unclear → hard to formulate testable hypotheses
- “Best” scRNA-seq skillset:
  - Scripting
  - Programming
  - Statistics
  - (cellular/molecular) biology
  - (organismal/tissue/disease) biology

→ does not exist

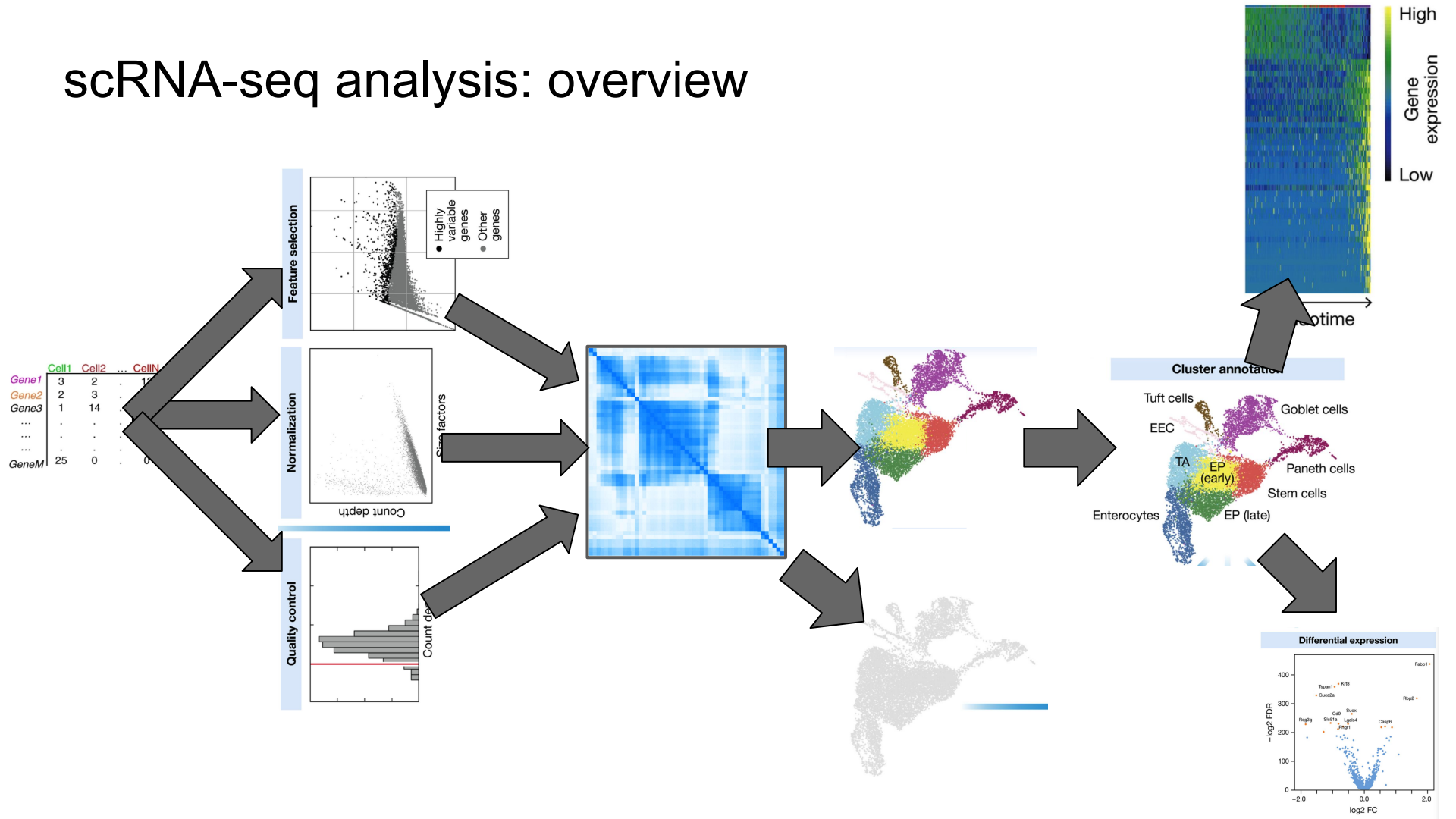
Why is scRNA-seq analysis so hard?

- To get the most out of the data you (probably) need non-standard analysis
- Not clear what standard analysis is anyway
- Too many knobs to turn → analysis paralysis
- Often: biological questions unclear → hard to formulate testable hypotheses
- “Best” scRNA-seq skillset:
  - Scripting
  - Programming
  - Statistics
  - (cellular/molecular) biology
  - (organismal/tissue/disease) biology

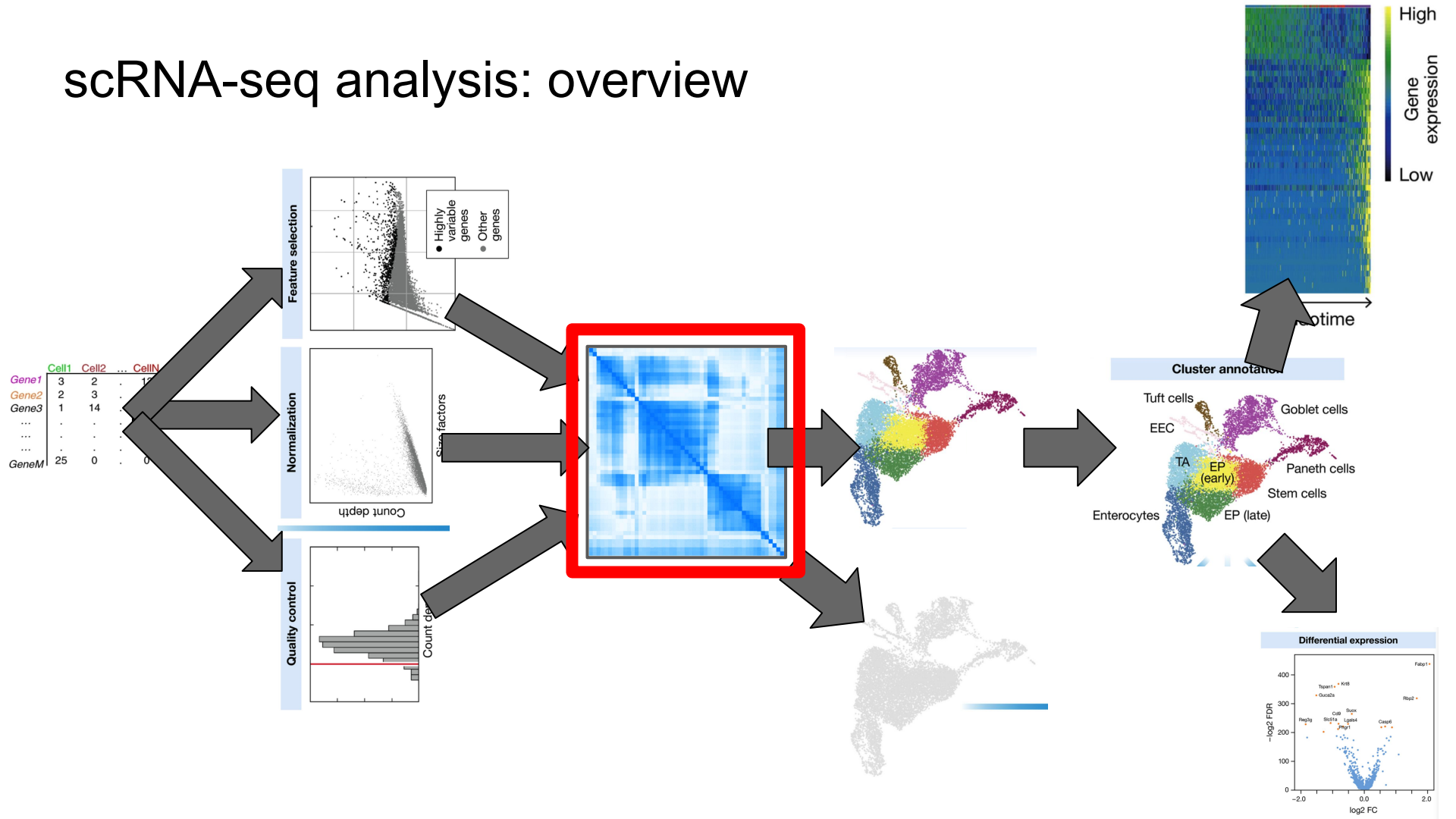
→ does not exist

- Recognize gaps
- Prioritize
- Develop “working understanding”
- Focus on biology/questions
- Ask for help

scRNA-seq analysis: overview

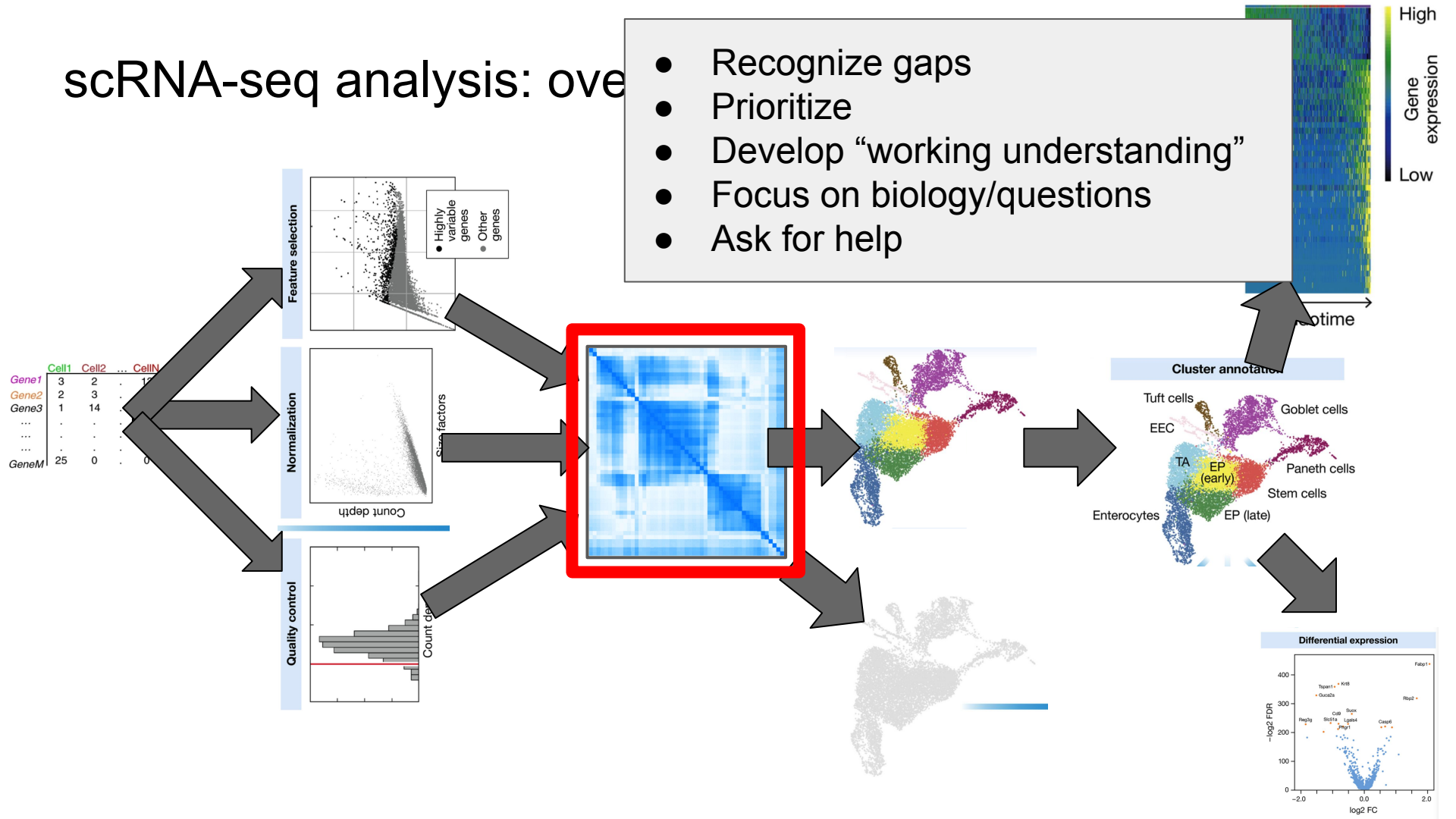


scRNA-seq analysis: overview



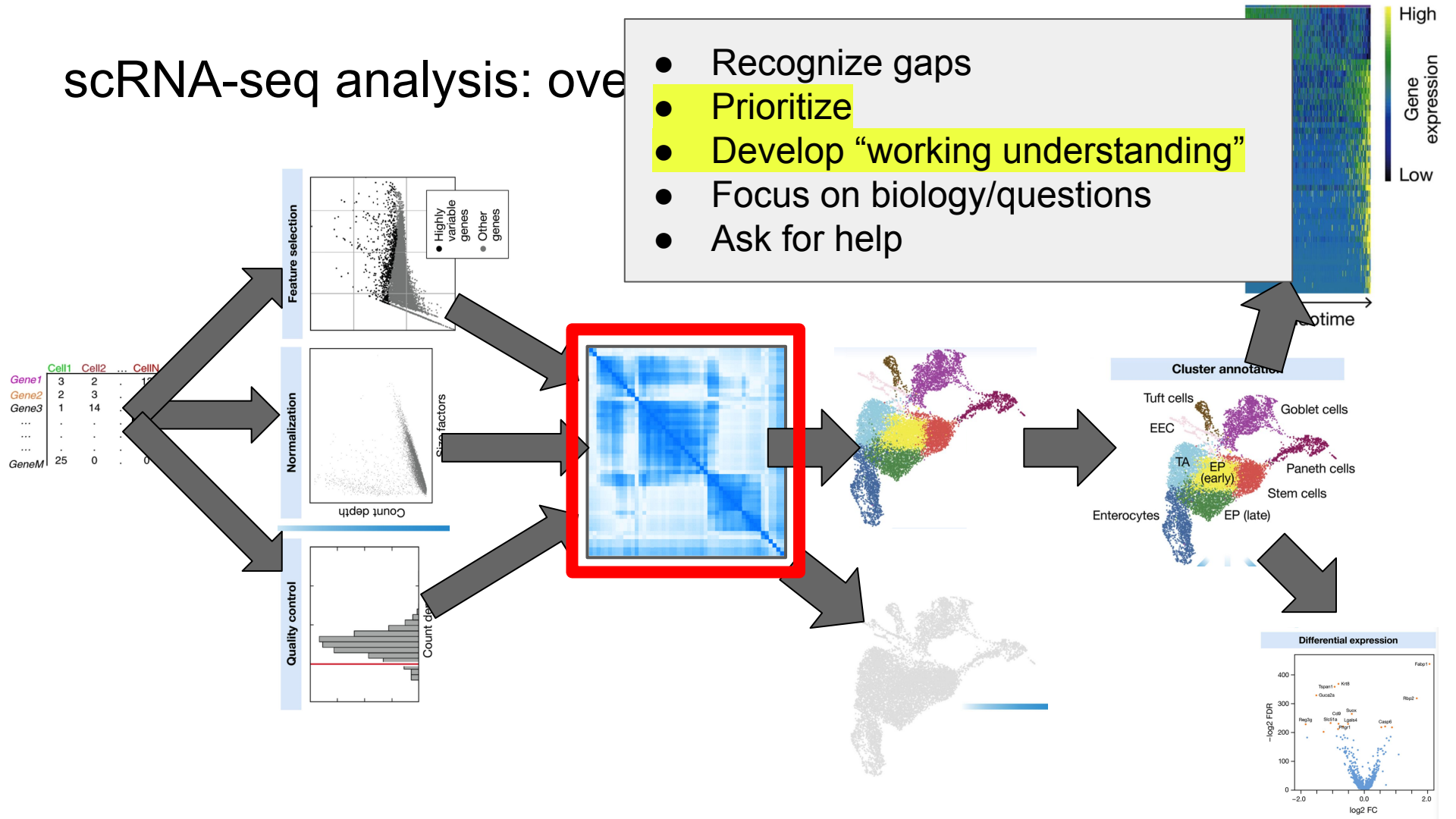
scRNA-seq analysis: ove

- Recognize gaps
- Prioritize
- Develop “working understanding”
- Focus on biology/questions
- Ask for help



scRNA-seq analysis: ove

- Recognize gaps
- Prioritize
- Develop “working understanding”
- Focus on biology/questions
- Ask for help



it’s all about
cell-cell distances

cell filtering

normalisation

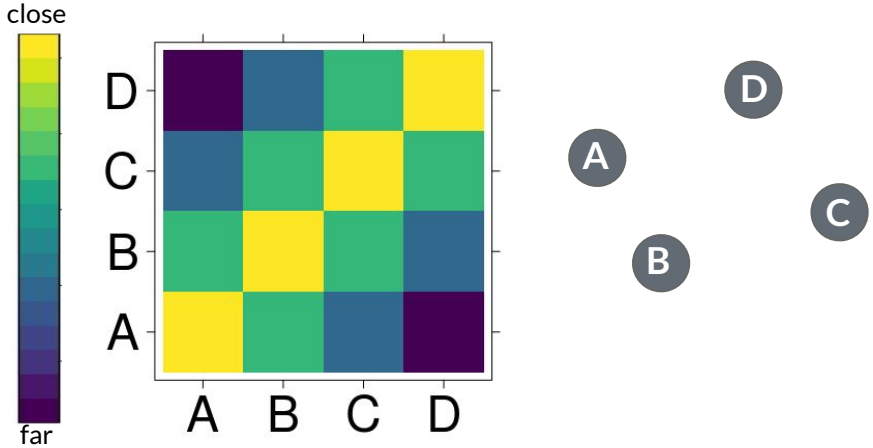
variance stabilisation

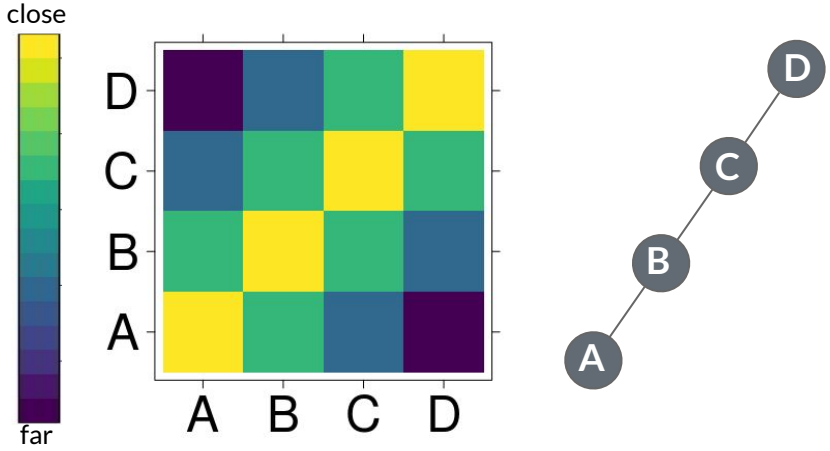
it’s all about cell-cell distances

dimensionality reduction

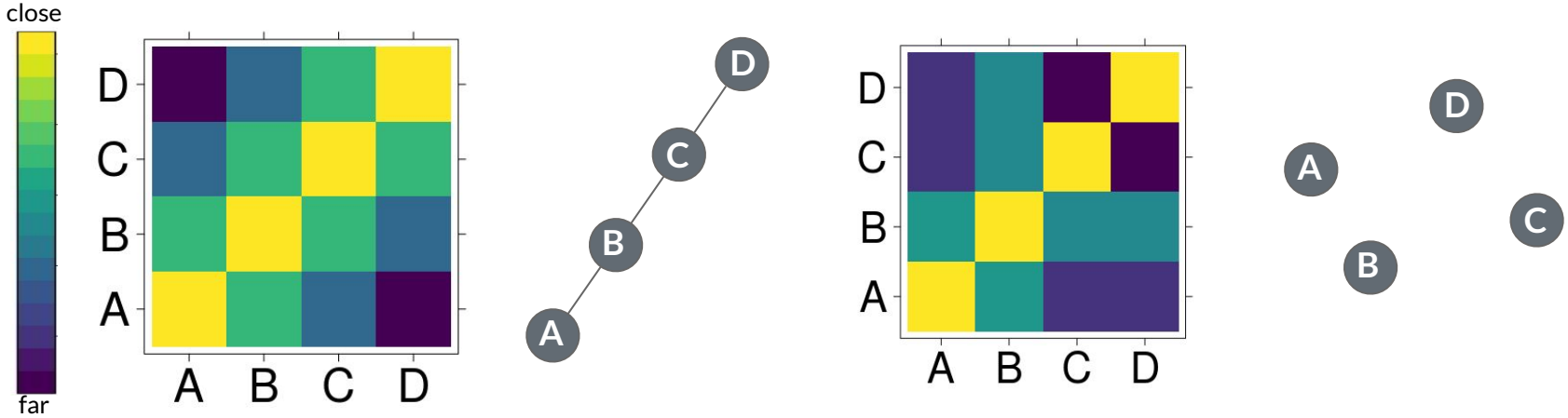
distance measures

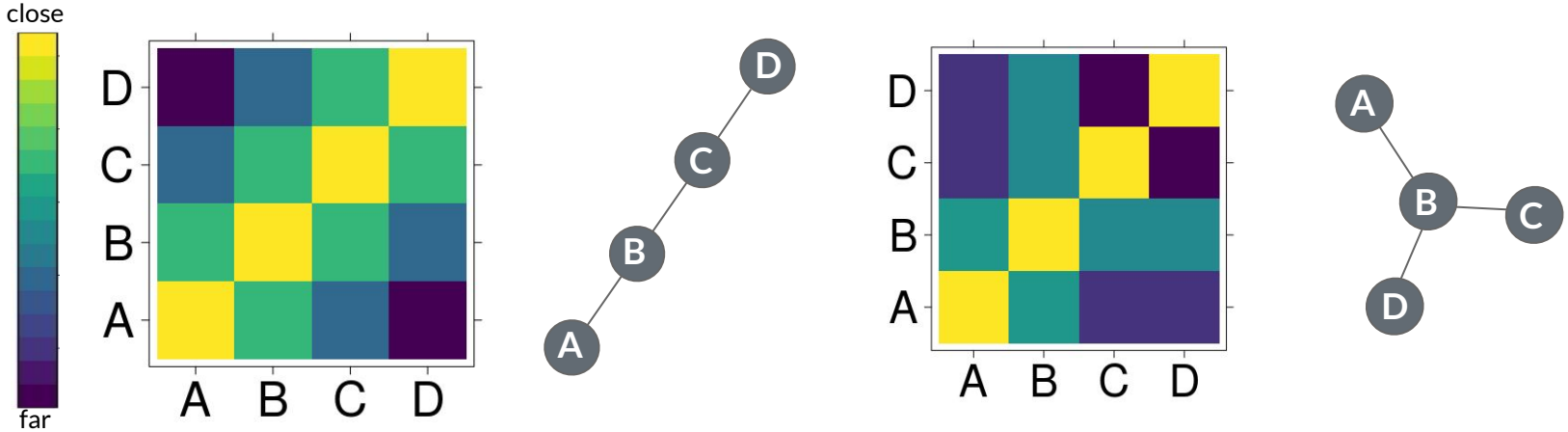
informative genes


close
far
D
C
B
A
A
B
C
D
D
A
C
B

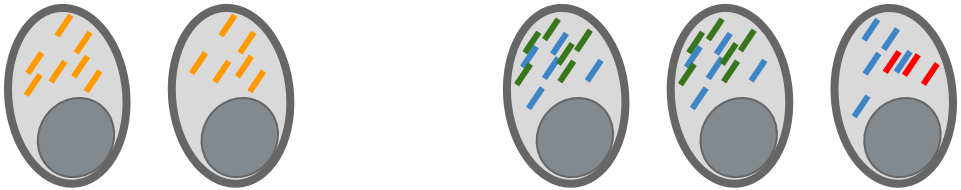

close
far
D
C
B
A
A
B
C
D

D
C
B
A

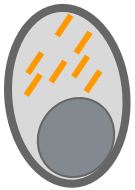
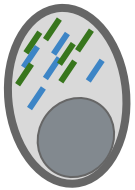
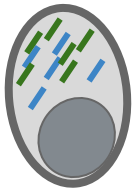
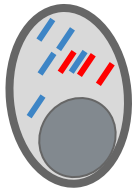

close
far
D
C
B
A
A B C D
D
C
B
A
D
C
B
A
A B C D
D
A
C
B


close
far
D
C
B
A
A B C D
D
C
B
A
D
C
B
A
A B C D
A
B
C
D

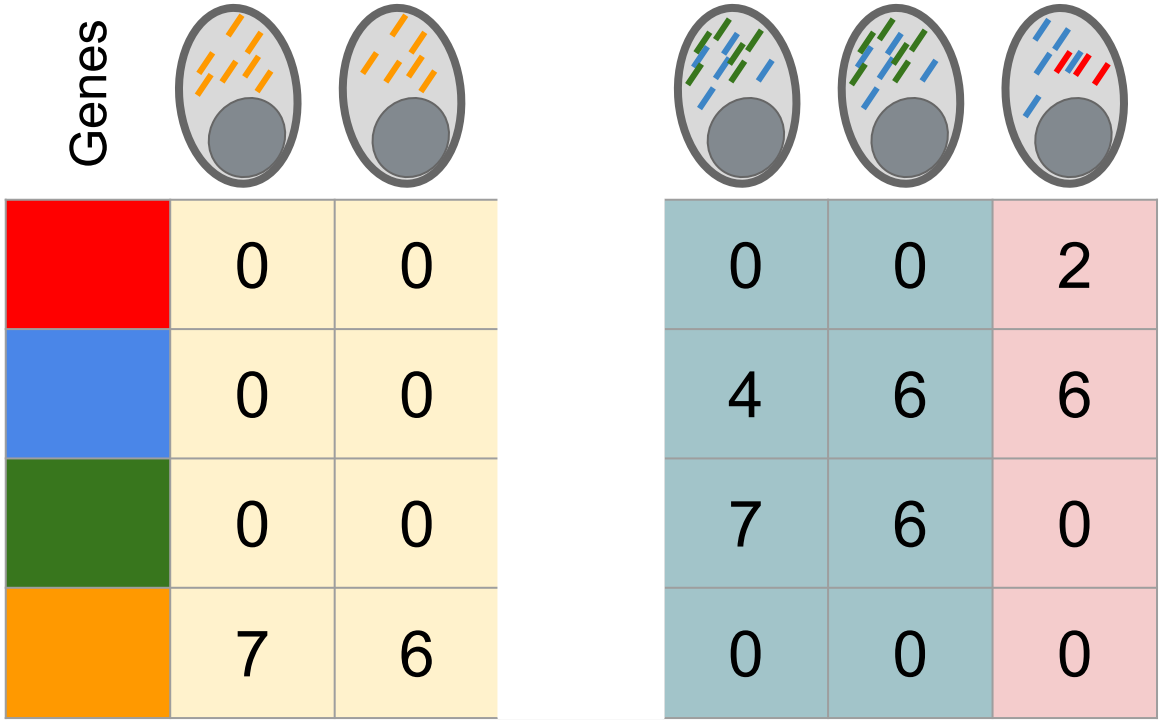
All about distances (1) - doublets



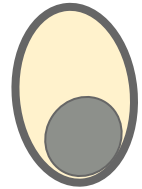
All about distances (1) - doublets

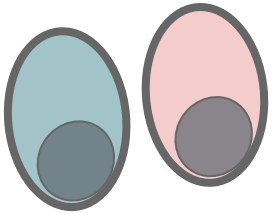
Genes					
	0	0	0	0	2
	0	0	4	6	6
	0	0	7	6	0
	7	6	0	0	0

All about distances (1) - doublets



All about distances (1) - doublets

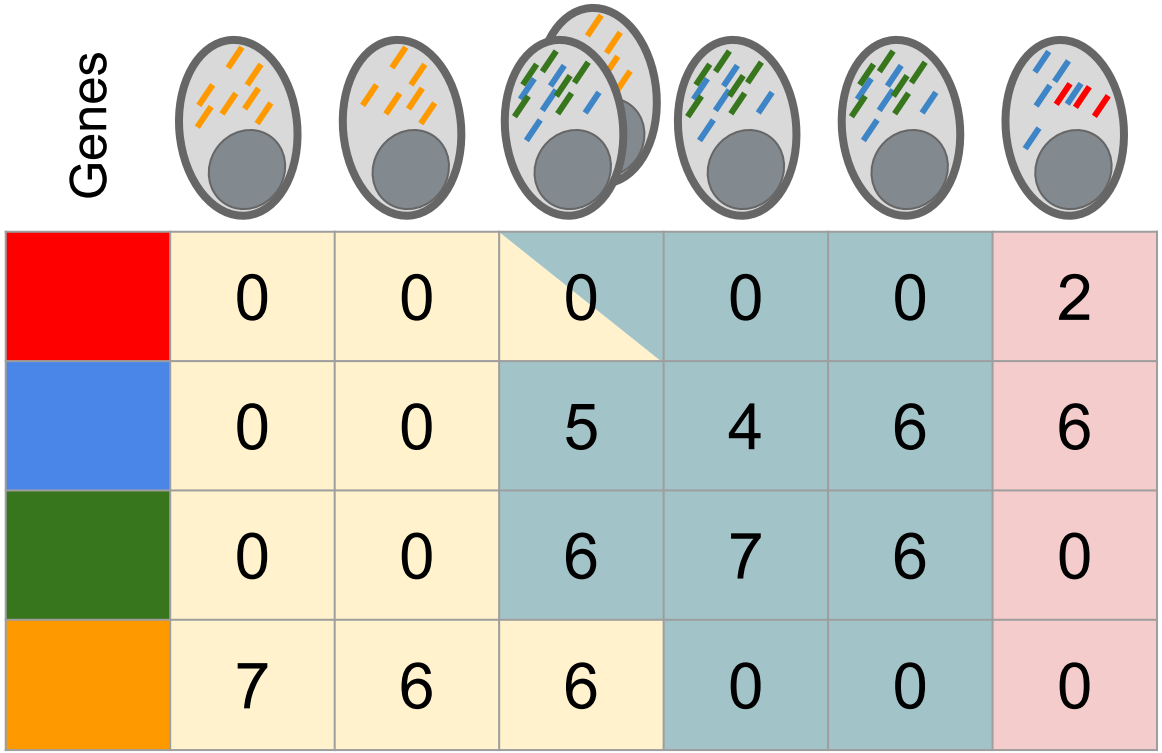




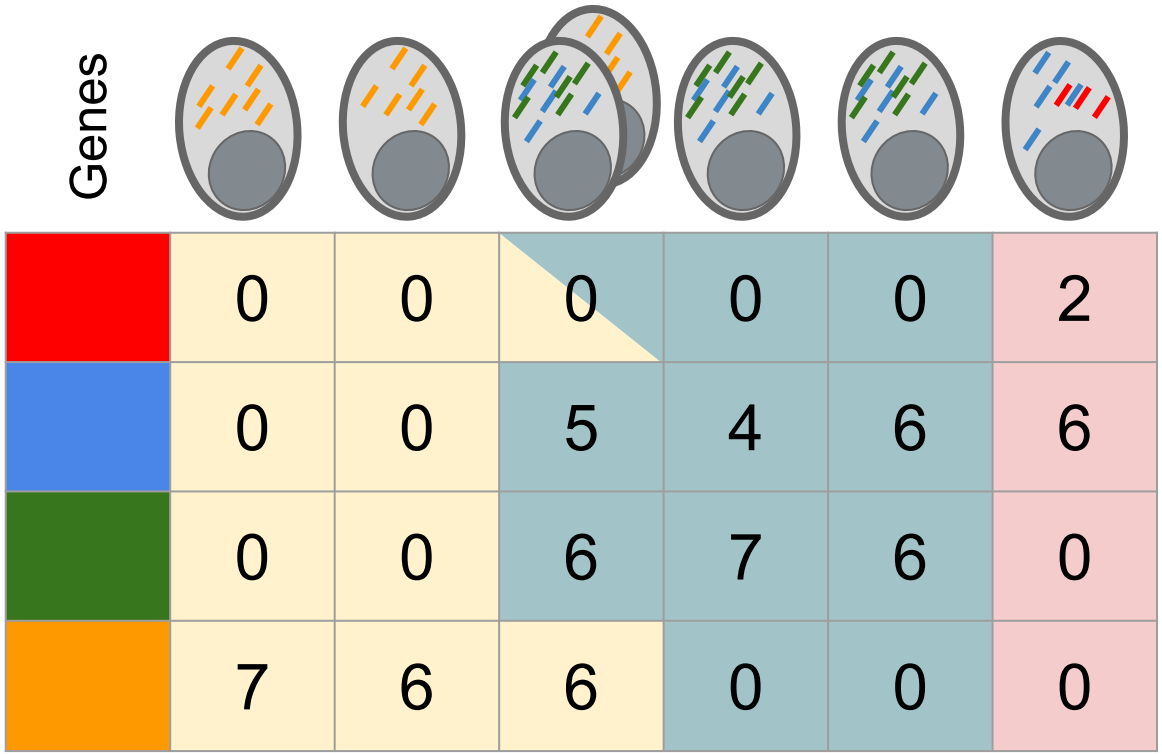
Genes		
	0	0
	0	0
	0	0
	7	6

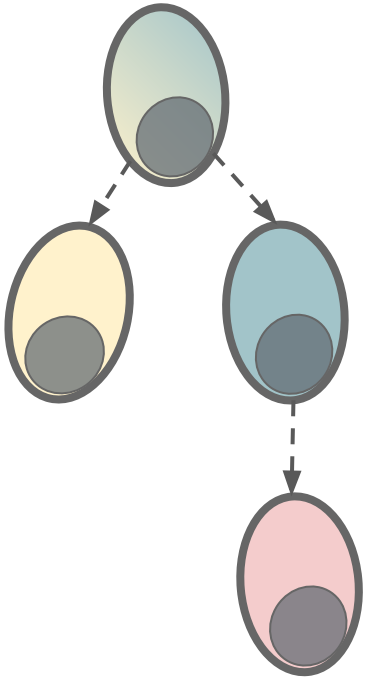
0	0	2
4	6	6
7	6	0
0	0	0

All about distances (1) - doublets



All about distances (1) - doublets





All about distances (2) - mitochondrial content

Tangent - mitochondrial content

Mitochondrial genes are expressed in most cells, and their expression level is cell type-specific.

High expression levels of mitochondrial genes could be an indicator of:

1. Poor sample quality, leading to a high fraction of apoptotic or lysing cells.
2. Biology of the particular sample, for example, tumor biopsies, may have increased mitochondrial gene expression due to metabolic activity and/or necrosis.

- 10X Genomics, 2017

Tangent - mitochondrial content

- dead cells% ~ MT%?

Tangent - mitochondrial content

- dead cells% ~ MT%?

  → try different combinations of happy/unhappy/dying cells* & sequence

Tangent - mitochondrial content

- dead cells% ~ MT%?

  → try different combinations of happy/unhappy/dying cells* & sequence

* human PBMCs from a healthy donor

Tangent - mitochondrial content

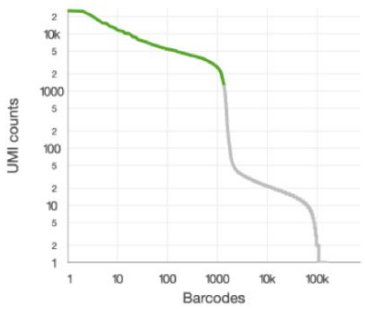
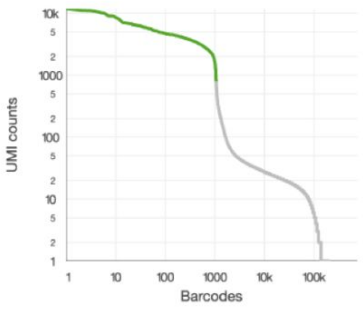
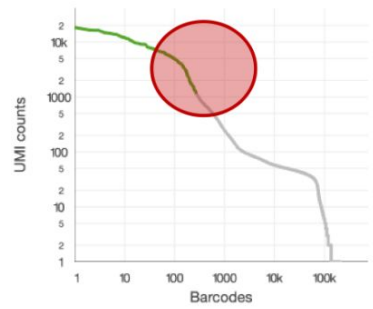
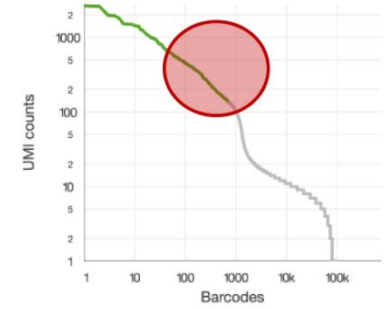
- dead cells% ~ MT%?

  → try different combinations of happy/unhappy/dying cells* & sequence

Sample	Name	What happened
1	control	sample prep into seq
2	Room temperature (24h)	Clearly unhappy cells
3	Digitonin Low	1:1 mix of dying (digitonin) and happy cells
4	Digitonin High	5:1 mix of dying/happy

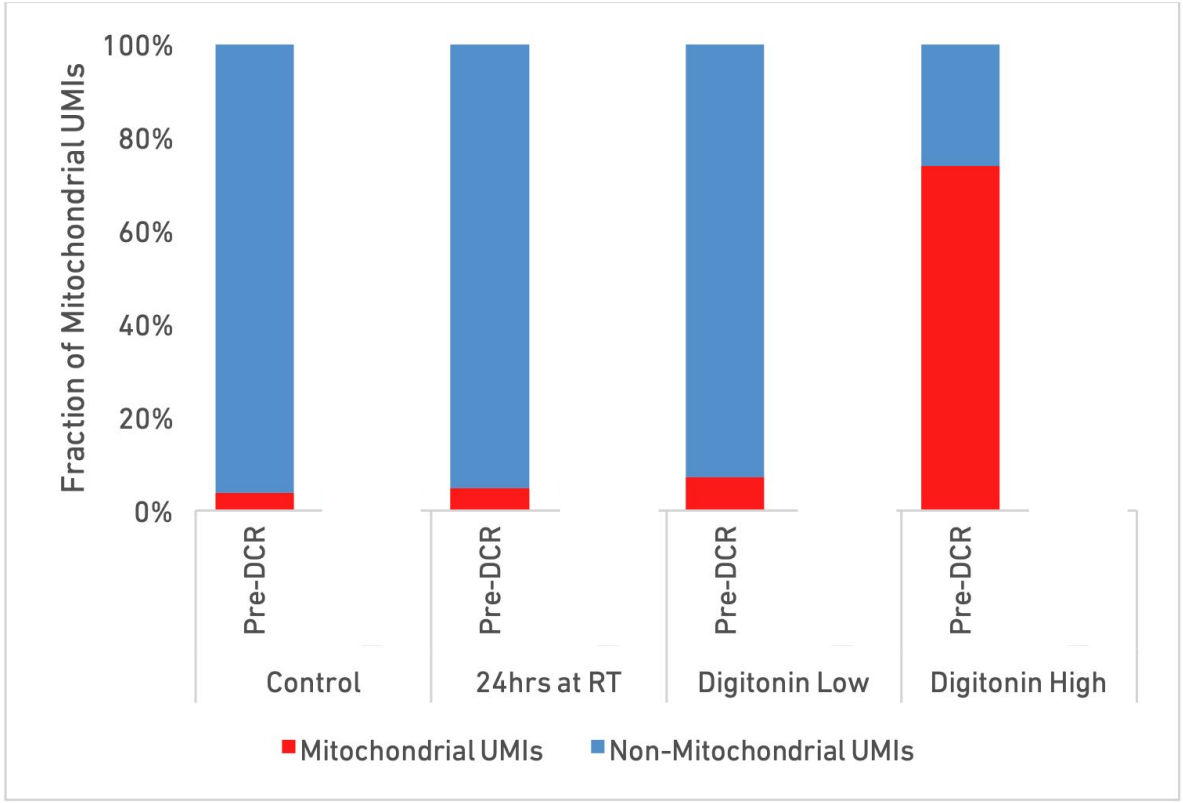
* human PBMCs from a healthy donor

Tangent - mitochondrial content

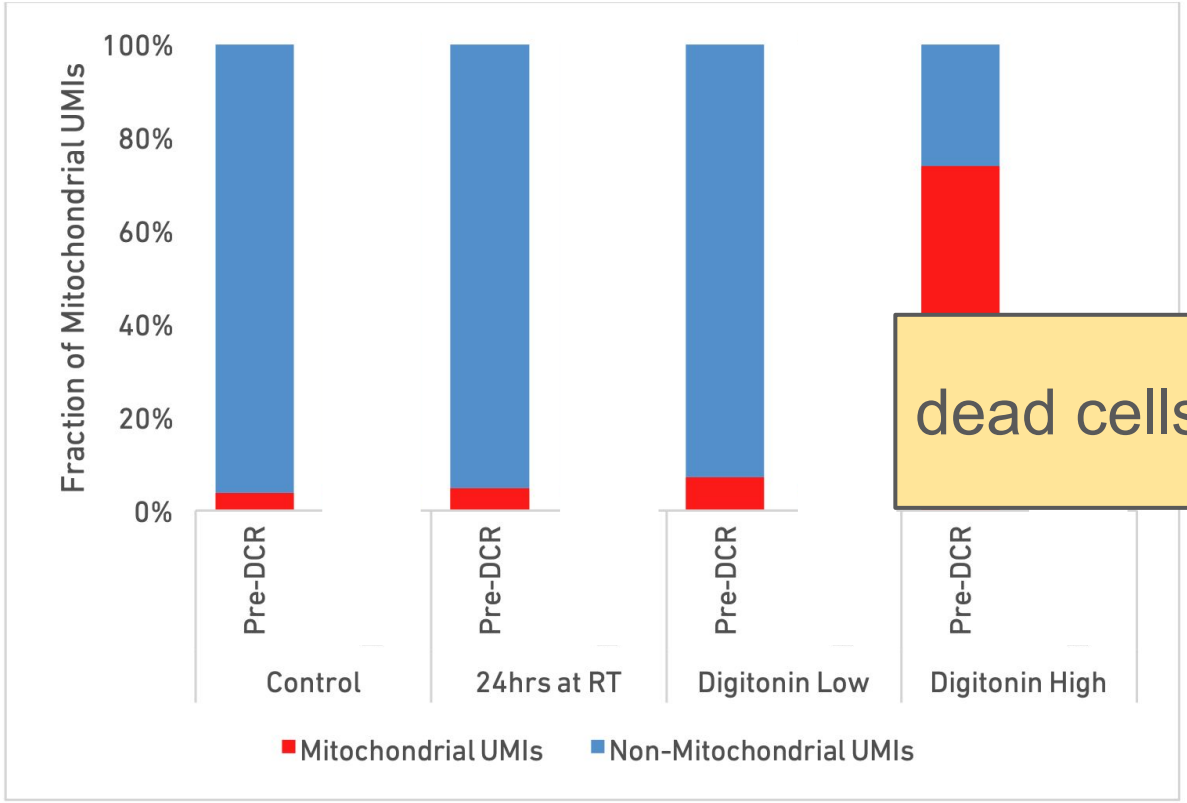
Treatment	Control	24 h at RT	Digitonin Low	Digitonin High
Pre-Dead Cell Removal				

https://support.10xgenomics.com/single-cell-gene-expression/index/doc/technical-note-removal-of-dead-cells-from-single-cell-suspensions-improves-performance-for-10x-genomics-single-cell-applications.

Tangent - mitochondrial content

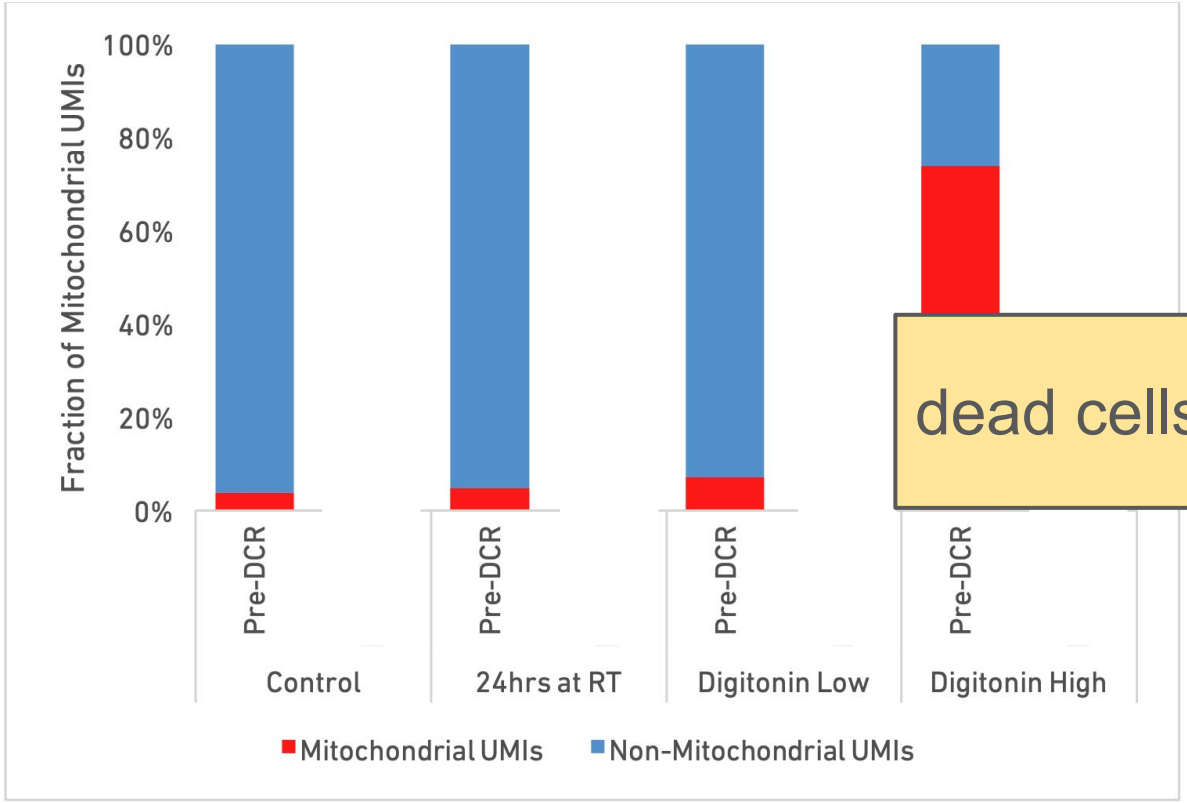


Tangent - mitochondrial content



dead cells% ~ MT% ✅

Tangent - mitochondrial content



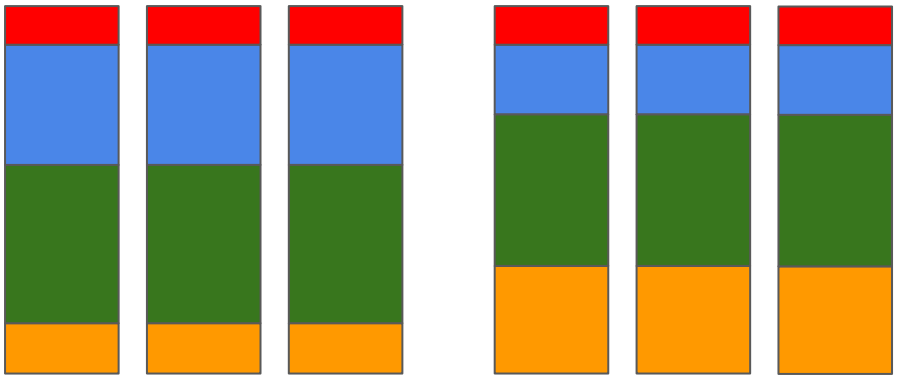
dead cells% ~ MT% ✅

*

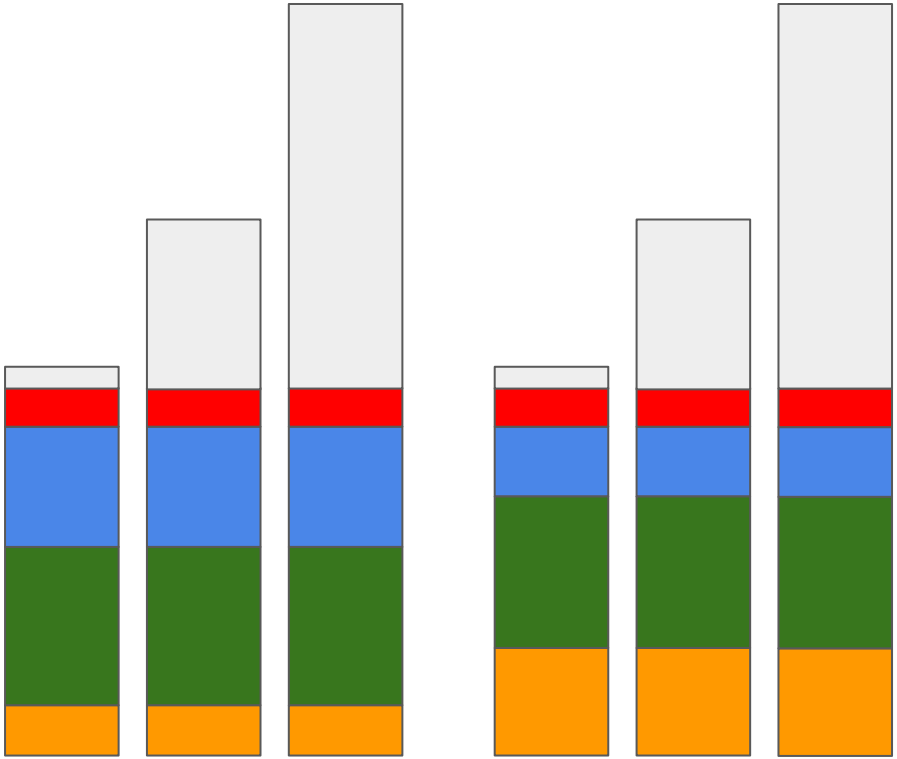
* in human PBMCs

All about distances (2) - mitochondrial content

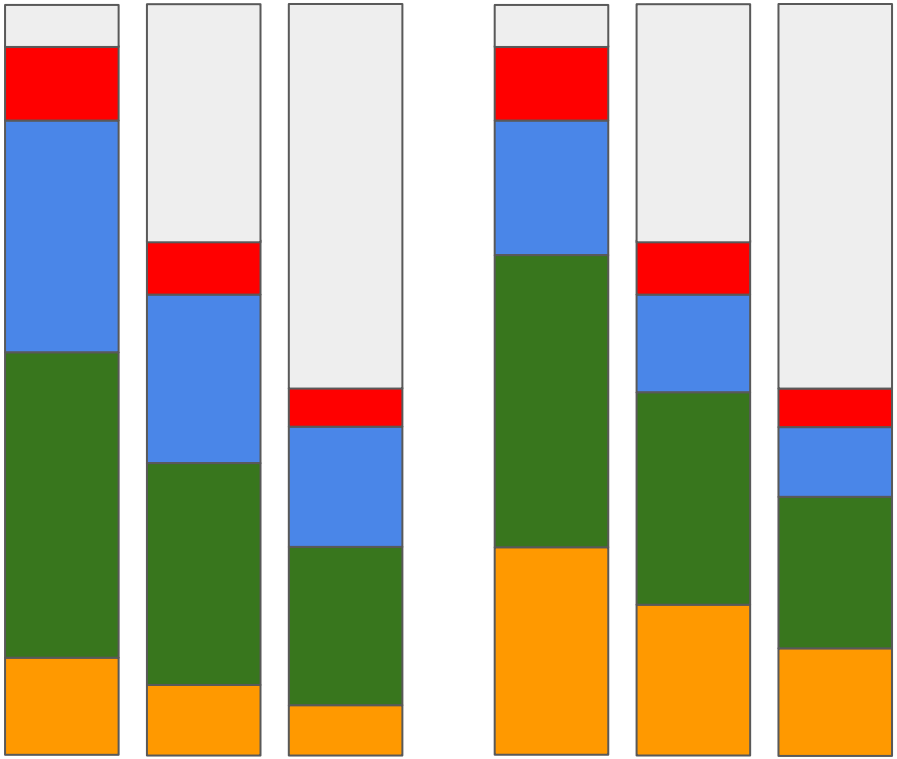
All about distances (2) - mitochondrial content

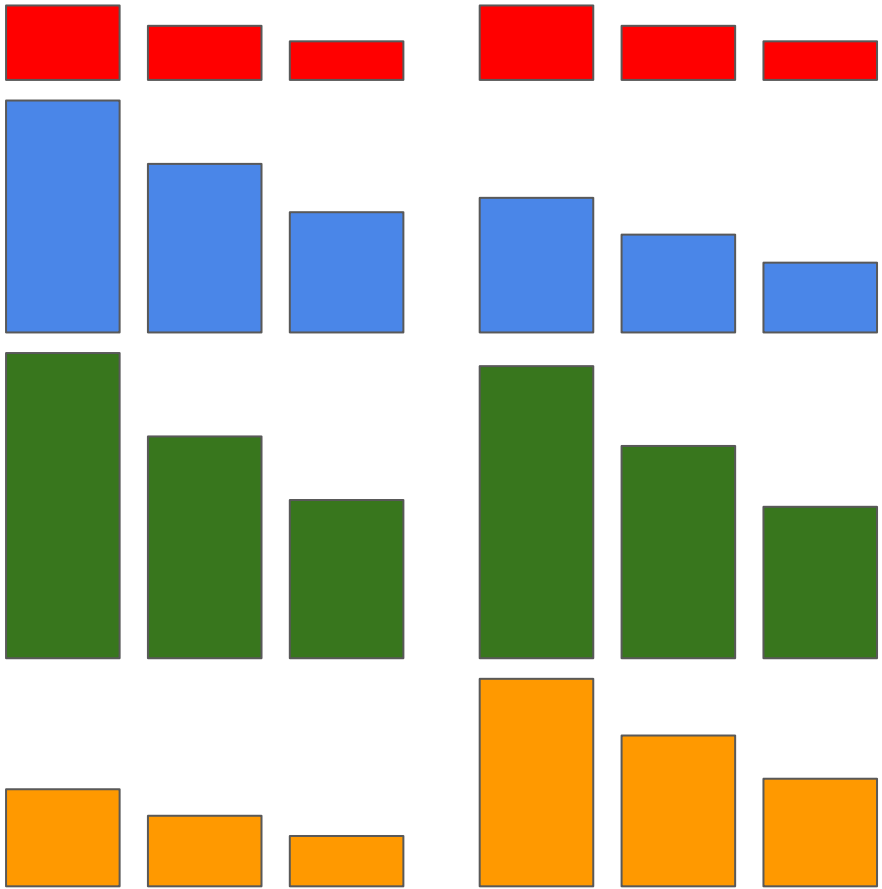


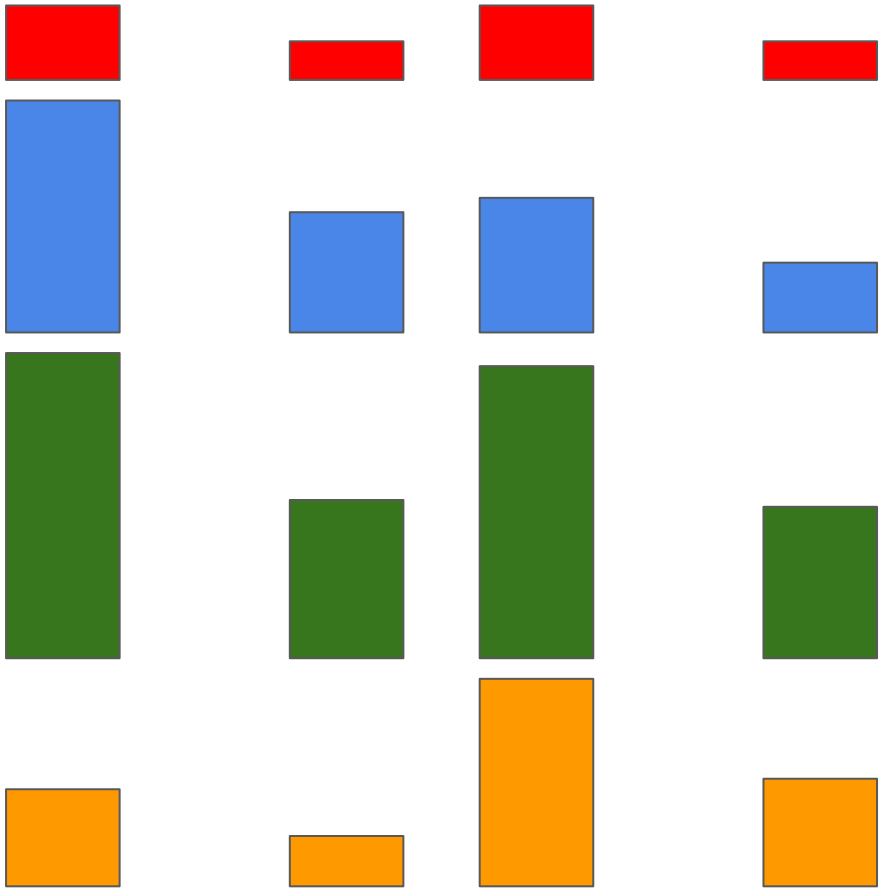
All about distances (2) - mitochondrial content



All about distances (2) - mitochondrial content







QC heuristics

QC heuristics

- Doublets & low-quality cells (usually) increase noise in data*

* technically possible to redeem, under very specific circumstances. It very probably doesn’t apply to your project.

QC heuristics

- Doublets & low-quality cells (usually) increase noise in data*
- Mammalian cut-offs probably don’t apply

* technically possible to redeem, under very specific circumstances. It very probably doesn’t apply to your project.

QC heuristics

- Doublets & low-quality cells (usually) increase noise in data*
- Mammalian cut-offs probably don’t apply
- No silver bullet :(

* technically possible to redeem, under very specific circumstances. It very probably doesn’t apply to your project.

QC heuristics

- Doublets & low-quality cells (usually) increase noise in data*
- Mammalian cut-offs probably don’t apply
- No silver bullet :(
- Compare to other indicators of cell quality *in context*

* technically possible to redeem, under very specific circumstances. It very probably doesn’t apply to your project.

QC heuristics

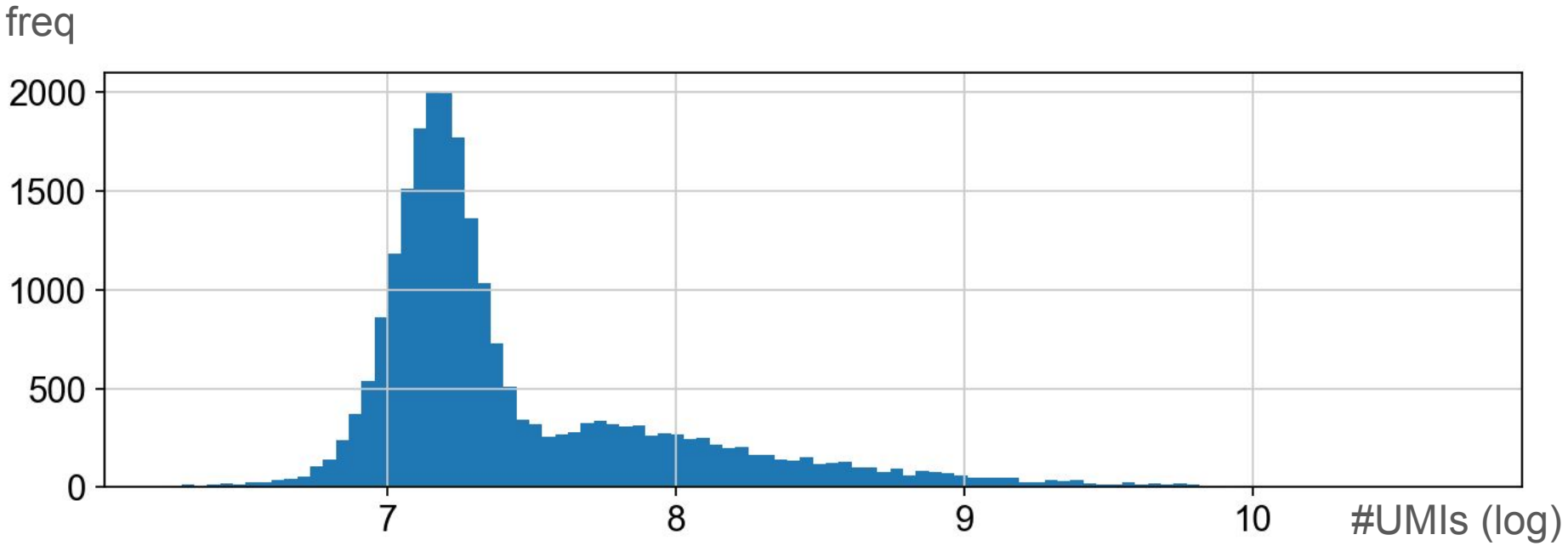
- Doublets & low-quality cells (usually) increase noise in data*
- Mammalian cut-offs probably don’t apply
- No silver bullet :(
- Compare to other indicators of cell quality *in context*
  - #genes detected
  - #total reads
  - %ribosomal content
  - %highest expressed genes
  - Compare to housekeeping genes**
  - Ratio of useful reads/mito+housekeeping

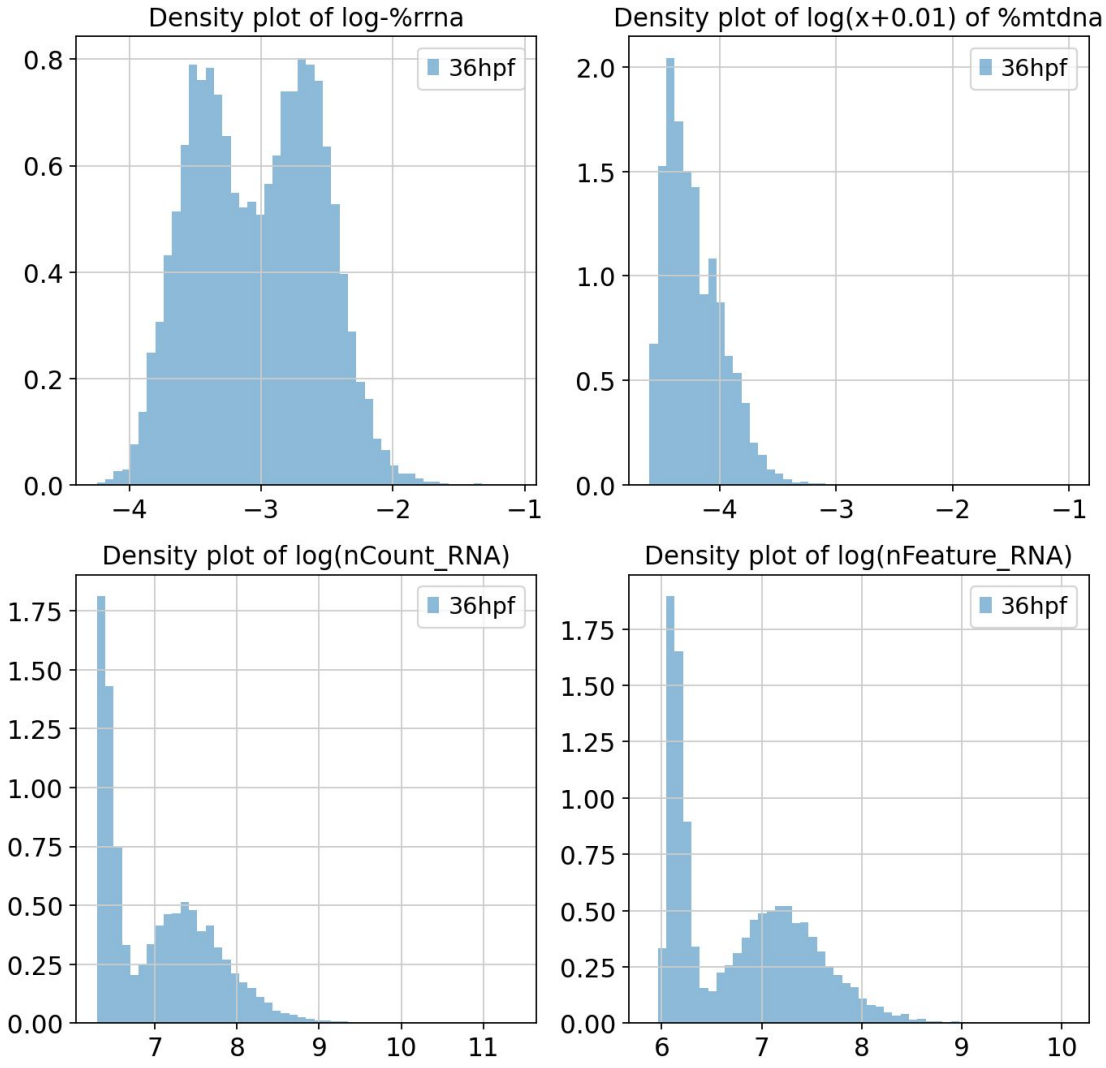
* technically possible to redeem, under very specific circumstances. It very probably doesn’t apply to your project.
** if you can get a list of them for your species

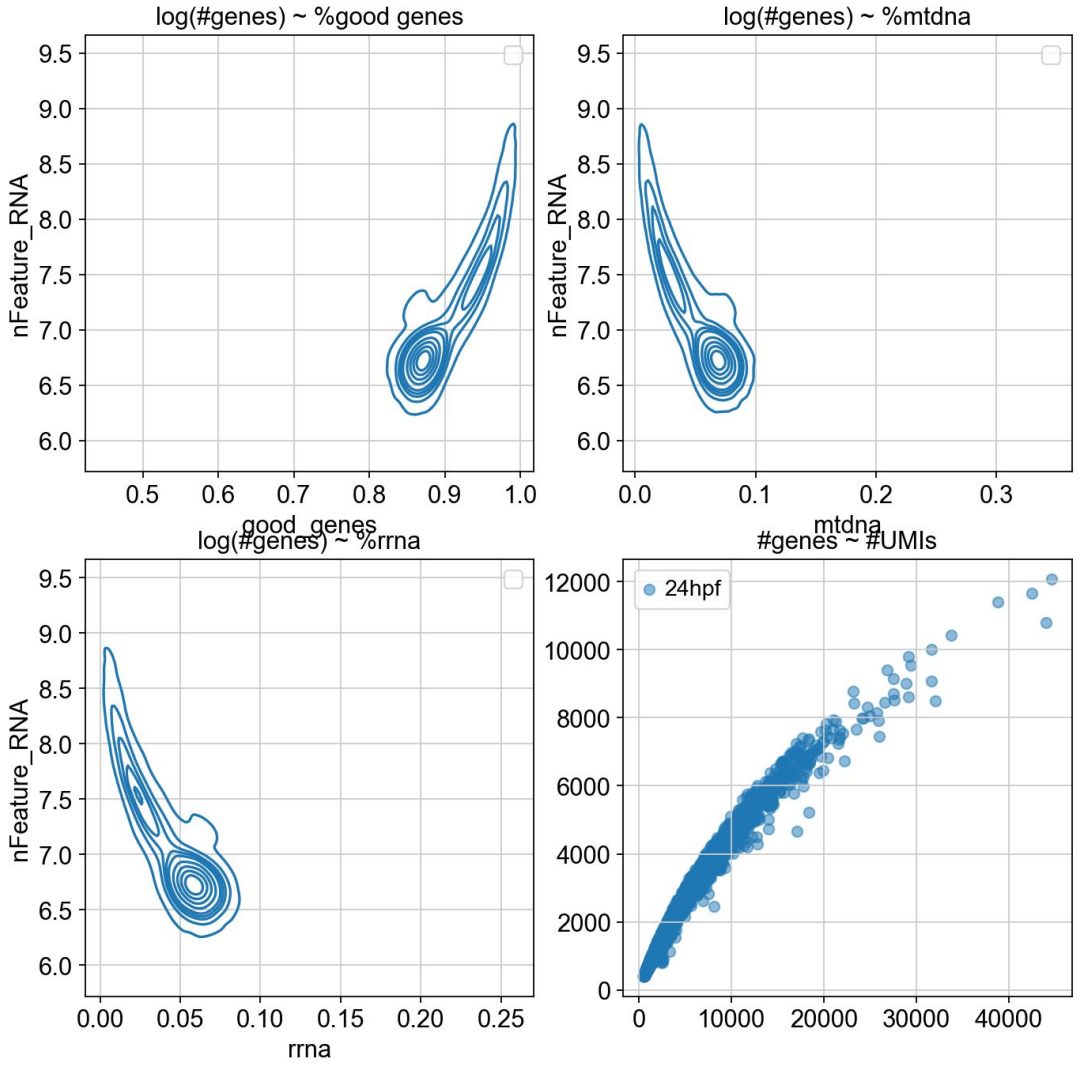
QC heuristics

- Doublets & low-quality cells (usually) increase noise in data*
- Mammalian cut-offs probably don’t apply
- No silver bullet :(
- Compare to other indicators of cell quality *in context*
  - #genes detected
  - #total reads
  - %ribosomal content
  - %highest expressed genes
  - Compare to housekeeping genes**
  - Ratio of useful reads/mito+housekeeping

* technically possible to redeem, under very specific circumstances. It very probably doesn’t apply to your project.
** if you can get a list of them for your species


freq
2000
1500
1000
500
0
7
8
9
10
#UMIs (log)


Density plot of log-%rrna
36hpf
0.0
0.2
0.4
0.6
0.8
−4
−3
−2
−1
Density plot of log(x+0.01) of %mtdna
36hpf
0.0
0.5
1.0
1.5
2.0
−4
−3
−2
−1
Density plot of log(nCount_RNA)
36hpf
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
7
8
9
10
11
Density plot of log(nFeature_RNA)
36hpf
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
6
7
8
9
10


log(#genes) ~ %good genes
nFeature_RNA
good_genes
log(#genes) ~ %mtdna
mtdna
log(#genes) ~ %rrna
rrna
#genes ~ #UMIs
24hpf

QC heuristics

- Doublets & low-quality cells (usually) increase along in data*
- Mammalian cut-offs probably don't apply
- No silver bullet :(
- Compare to other indicators of cell quality *in context*
  - #genes detected
  - #total reads
  - %ribosomal content
  - %highest expressed genes
  - Compare to housekeeping genes**
  - Ratio of useful reads/mito/housekeeping

The biology is right

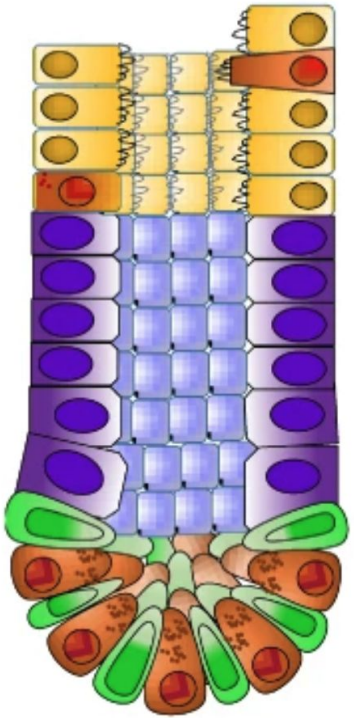
* technically possible to redeem, under very specific circumstances. It very probably doesn't apply to your project.
** if you can get a list of them for your species

Overall summary

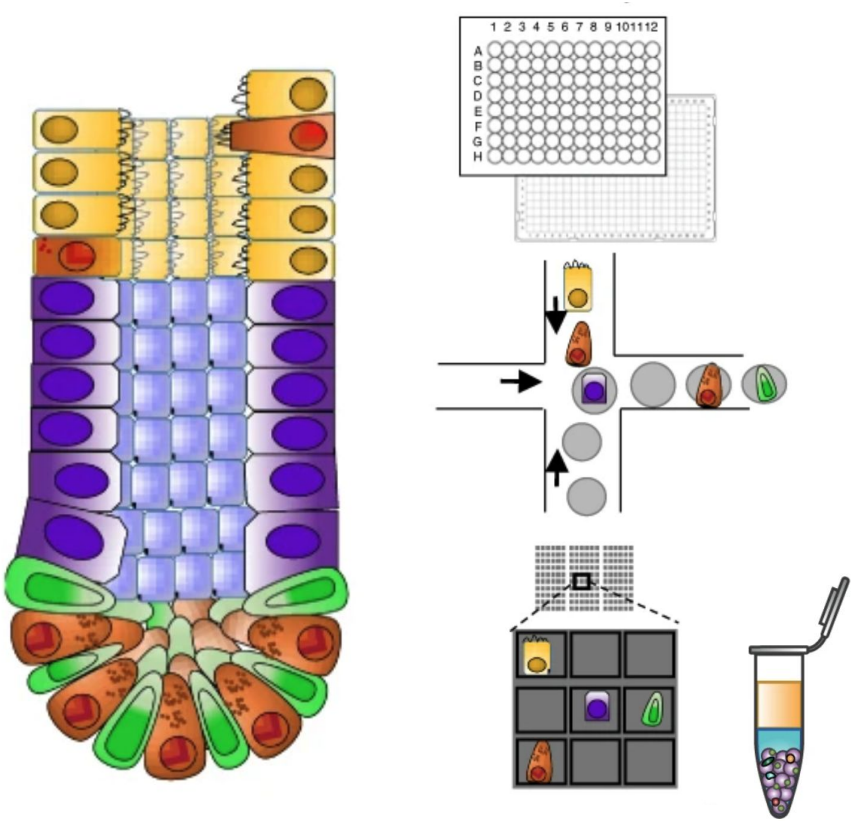
- scRNA-seq analysis has many moving parts from different areas of expertise → easy to have blind spots
- Cell-to-cell distances at the core of (most) analysis
- How is <insert choice> affecting the (pattern of) distances between cells?
- QC: try to combine “orthogonal” indicators
- QC: use common sense + knowledge of system
- QC: When in doubt…
  - Is this crucial for answering my question?
  - Can I validate this with other data?

Ideas graveyard

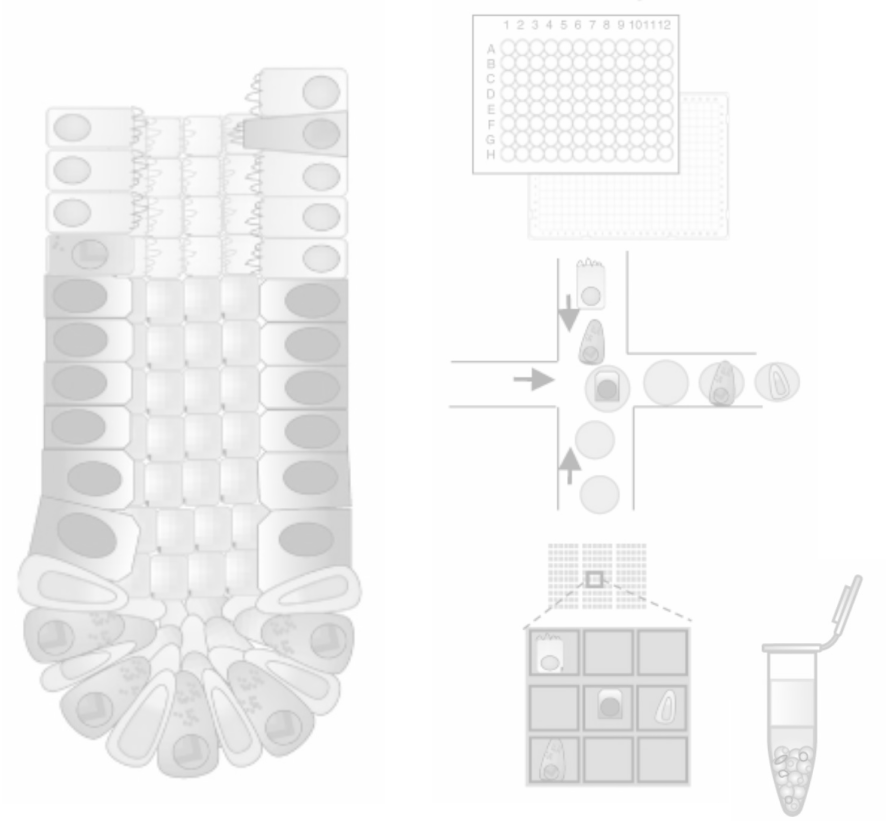
Single-cell RNA sequencing: a refresher

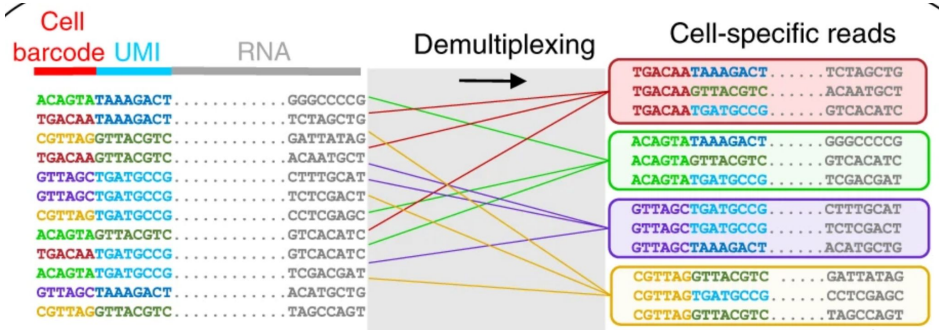


Single-cell RNA sequencing: a refresher

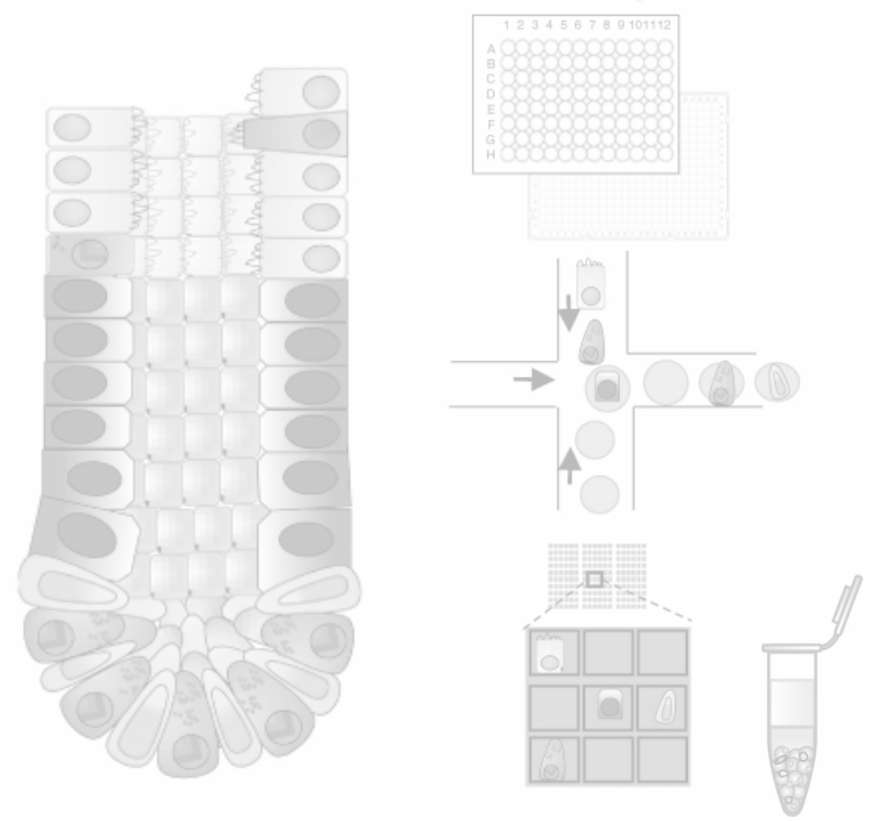


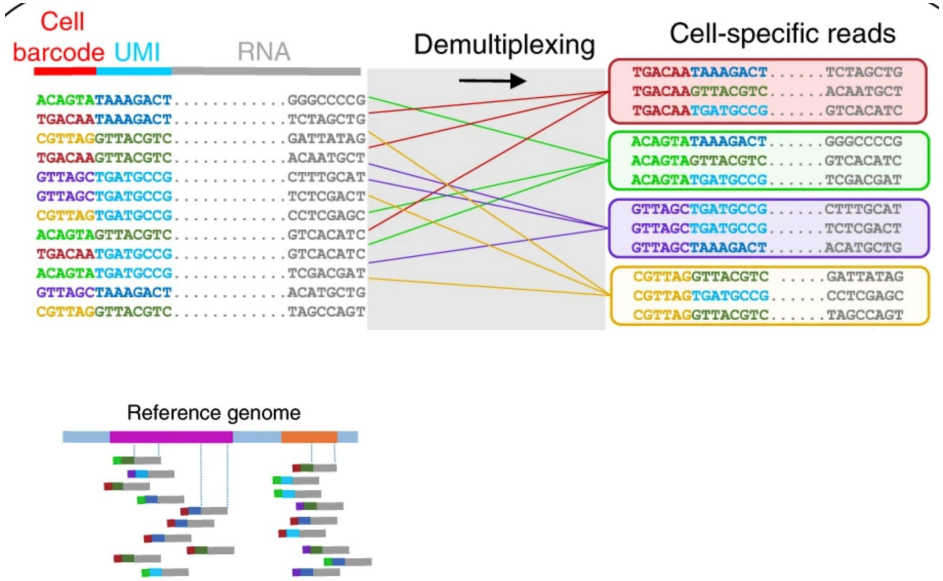
Single-cell RNA sequencing: a refresher



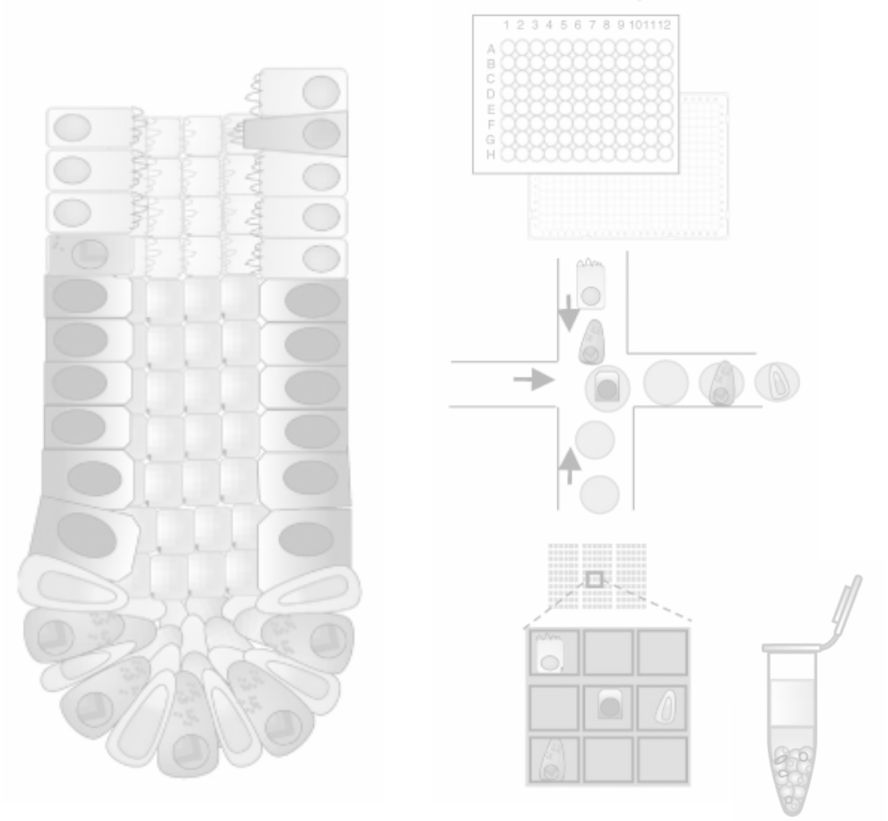


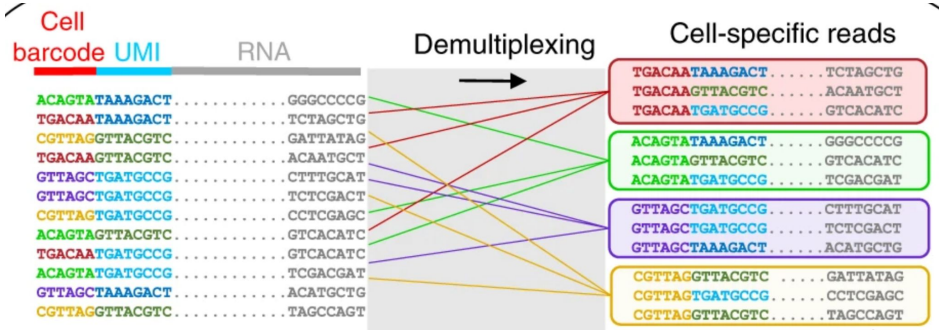
Single-cell RNA sequencing: a refresher

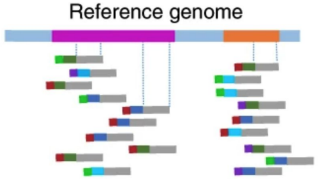




Single-cell RNA sequencing: a refresher

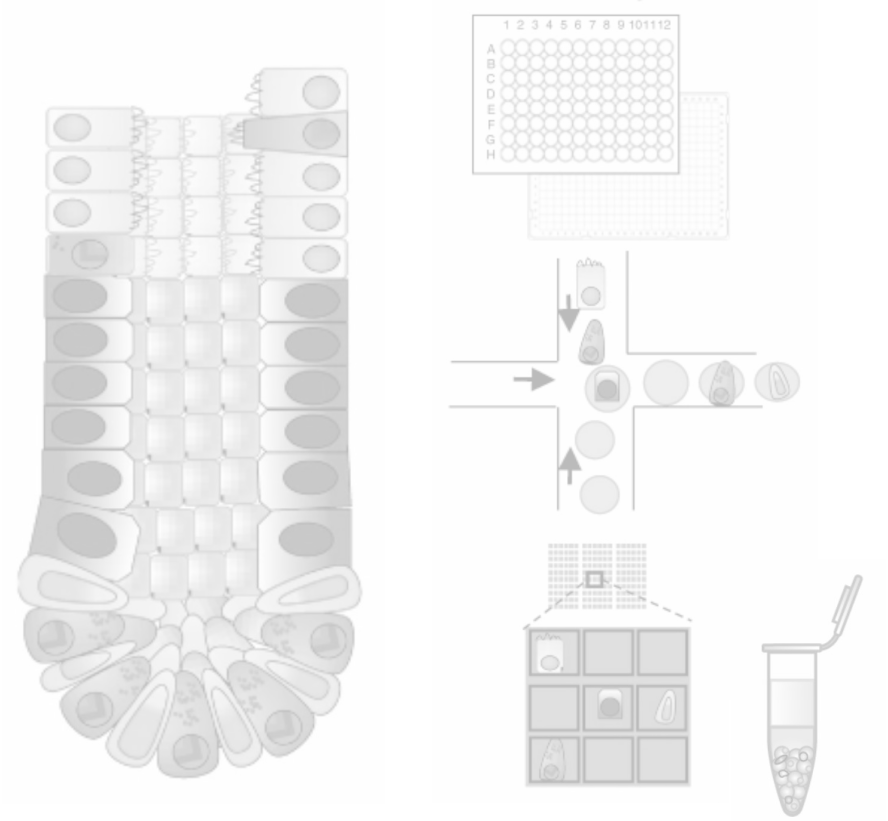


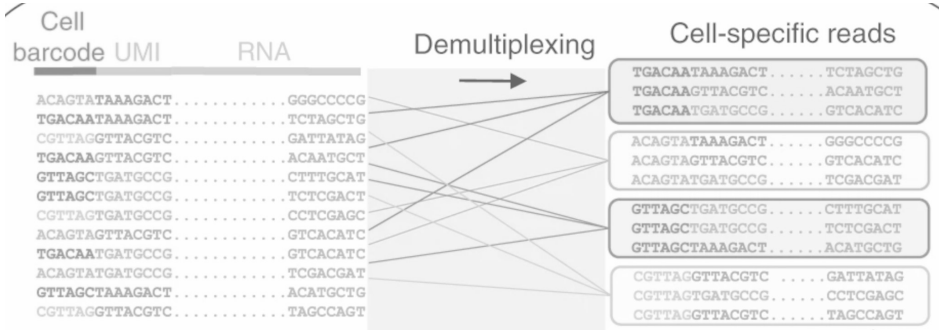


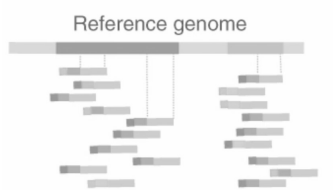


	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

Single-cell RNA sequencing: a refresher



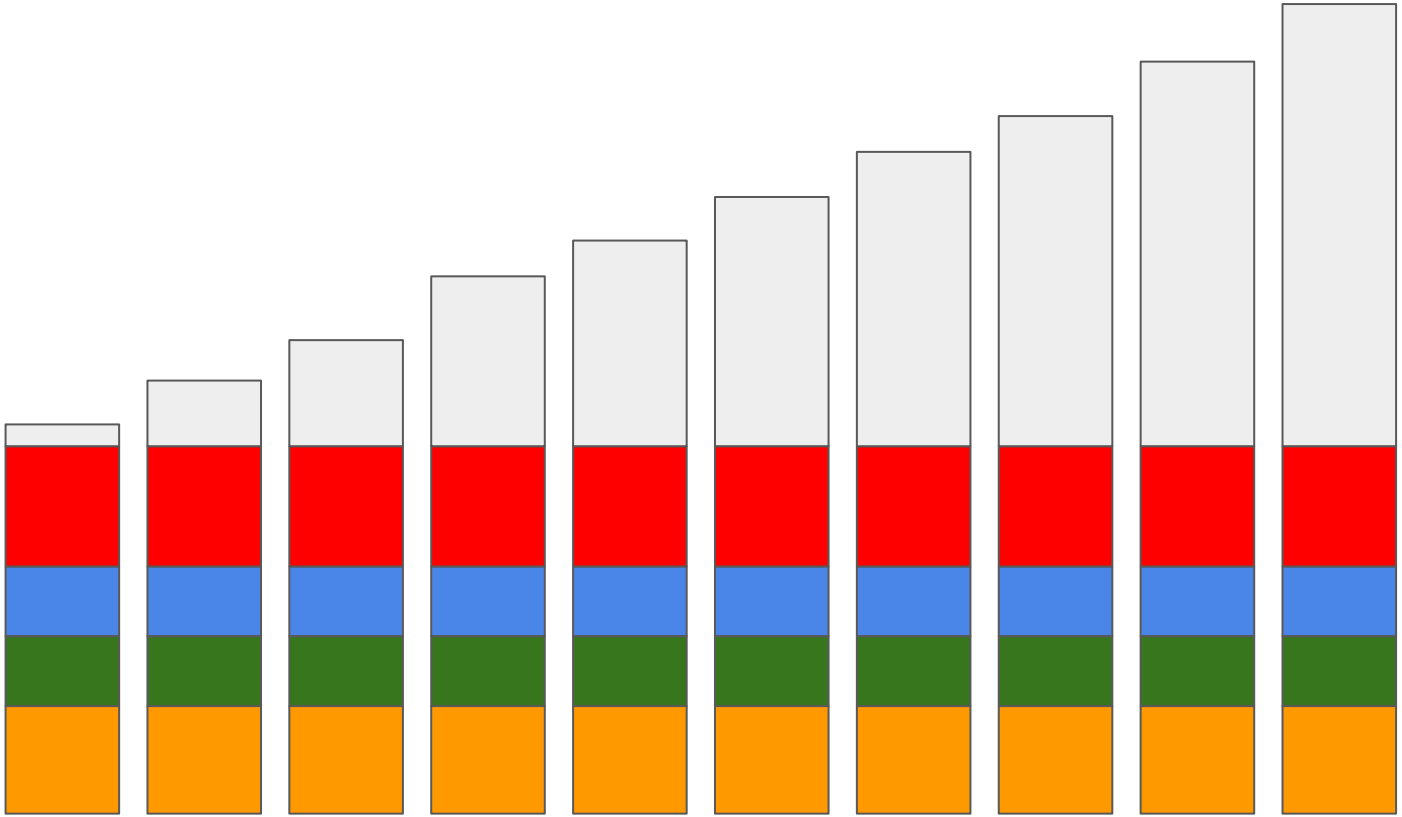




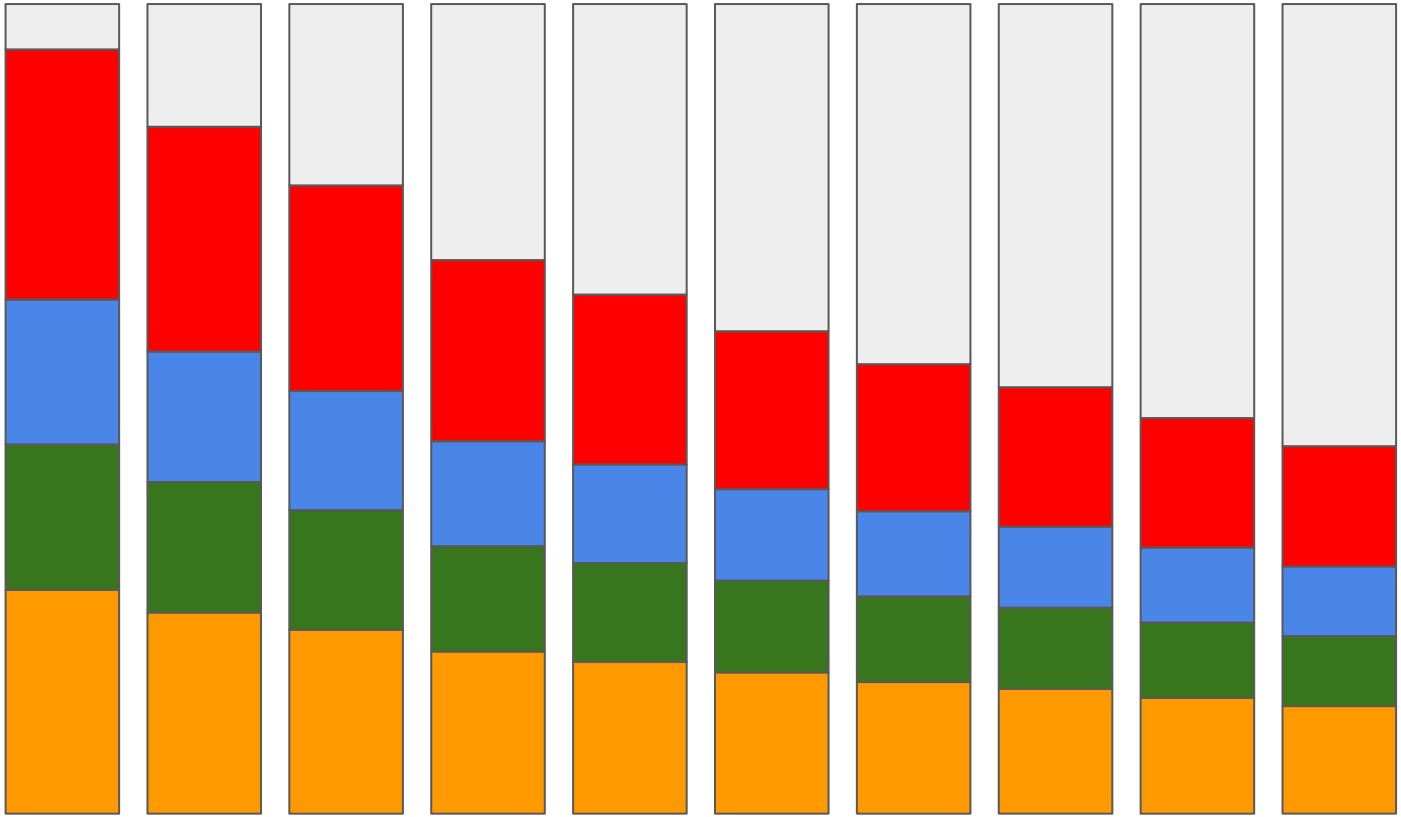
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

All about distances (2) - mitochondrial content

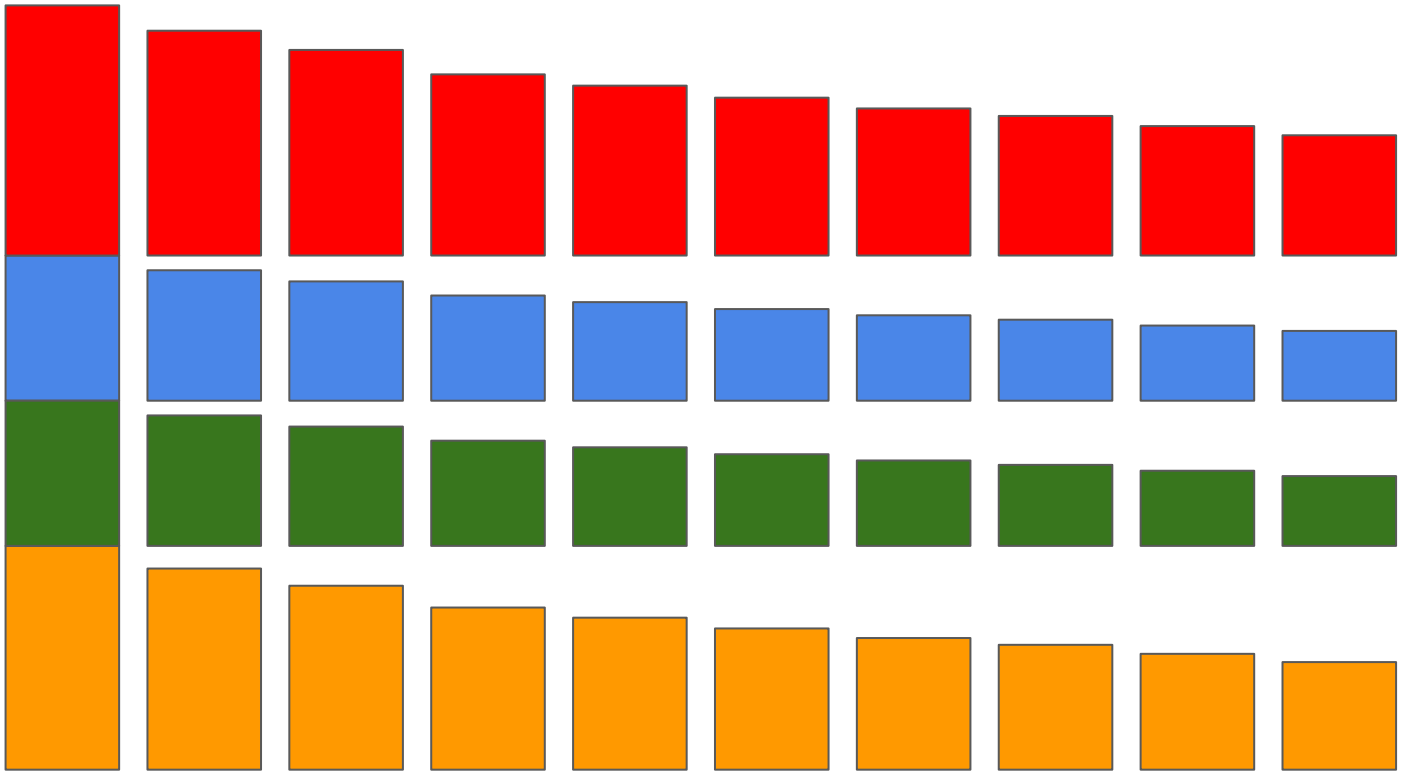
All about distances (2) - mitochondrial content



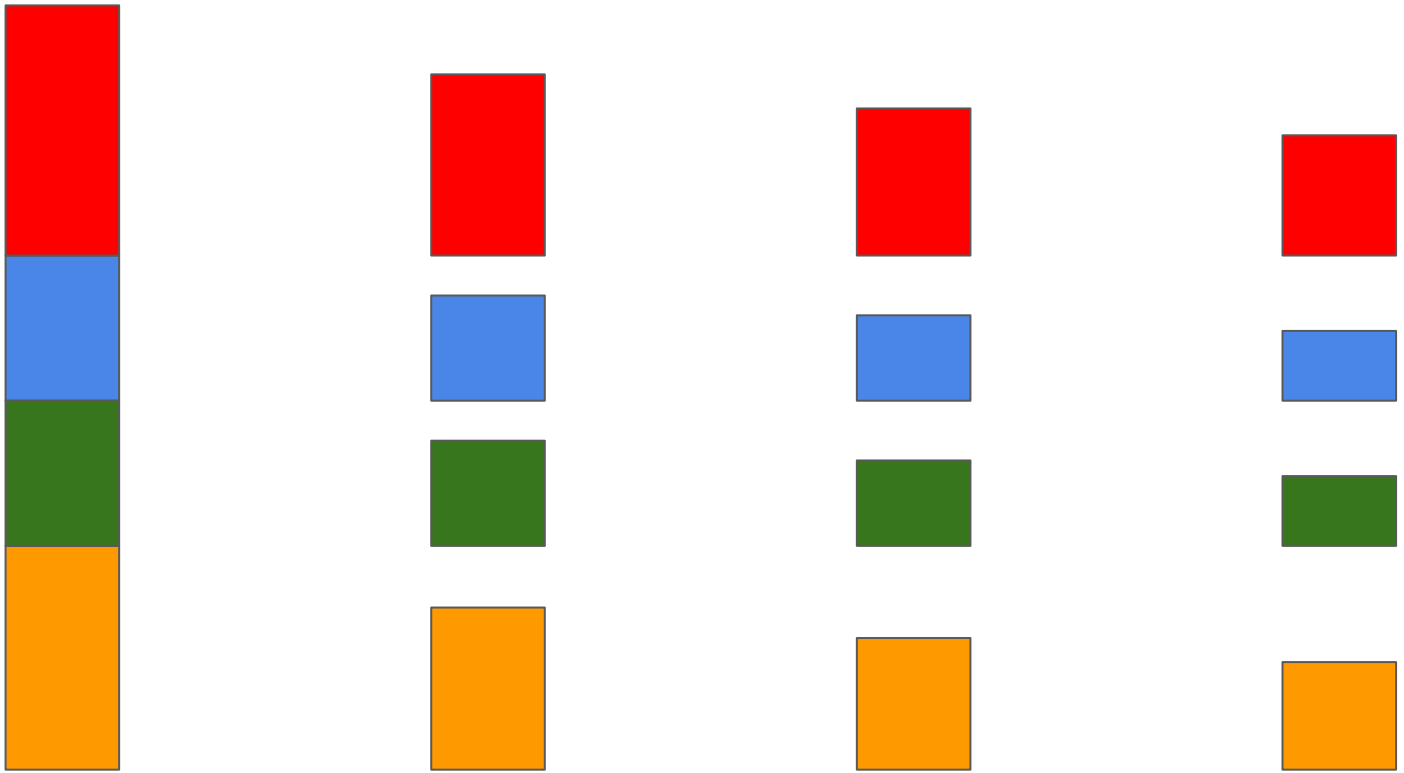
All about distances (2) - mitochondrial content



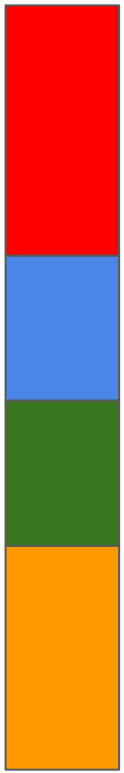
All about distances (2) - mitochondrial content

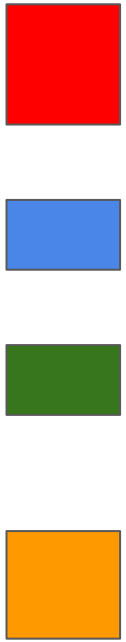


All about distances (2) - mitochondrial content

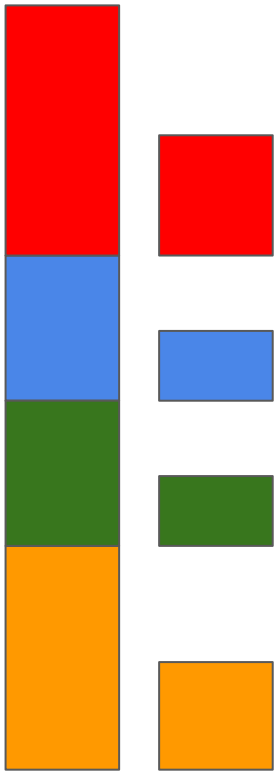


All about distances (2) - mitochondrial content





All about distances (2) - mitochondrial content



Why methods matter/hidden assumptions

- scRNA-seq == distances between cells; the perpetual struggle between noise and signal
- Anything that distorts the distances can change our analysis!
  - Normalisation: changes the values
    → changes distances!
  - Variance stabilisation: changes values
    → changes distances!
  - Mitochondrial/ribosomal counts: keep/remove? before/after normalisation?
    → changes distances!
  - Doublets: keep/remove?
    → changes distance *patterns*

Confusing methods for biology

- Clustering == cell types
- Highly variable genes == informative genes

Data hygiene

- Keep track of (and upload!) mapping reference (transcriptome/genome/proteome)
  - What choices were made (e.g. only protein-coding genes?)
  - State of annotation/reference: e.g. UTRs often extremely important for mapping rate
  - Include gene IDs in sc file/plots (ScanPy: use as .var index)
  - Keep track of annotation (BLAST/emapper/whatever tables)
- Bare minimum:
  - Raw count matrix
  - Filtering results (genes/cells)
  - Clustering results (corresponds to paper)