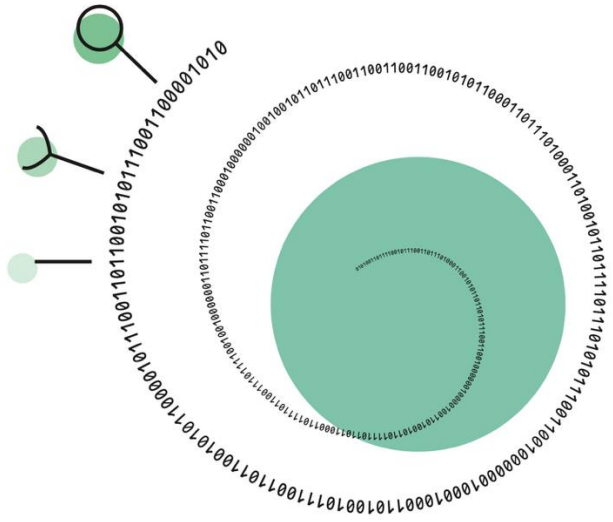
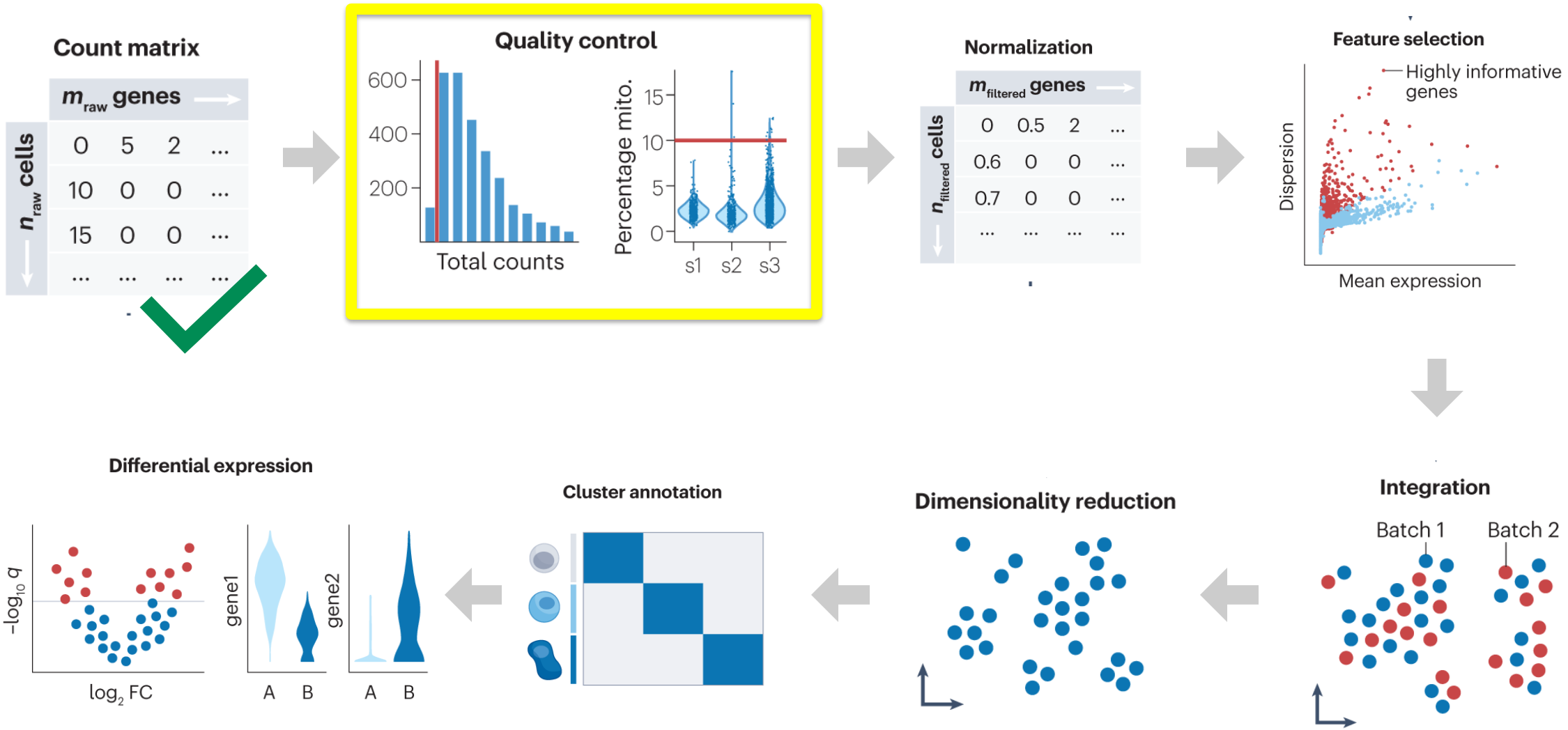


Single Dell Data Analysis Course

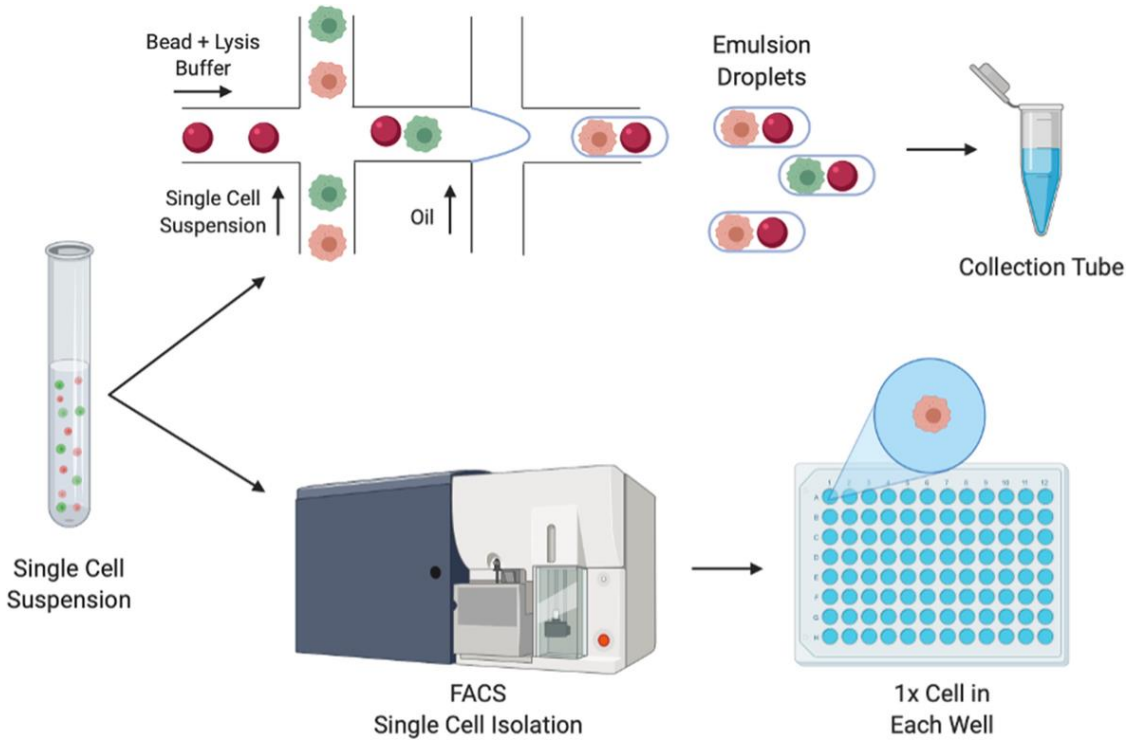
Quality Control

Lisa Buchauer
Professor of Systems Biology of Infectious Diseases
Department of Infectious Diseases and Intensive Care
Charité - Universitätsmedizin Berlin



Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w

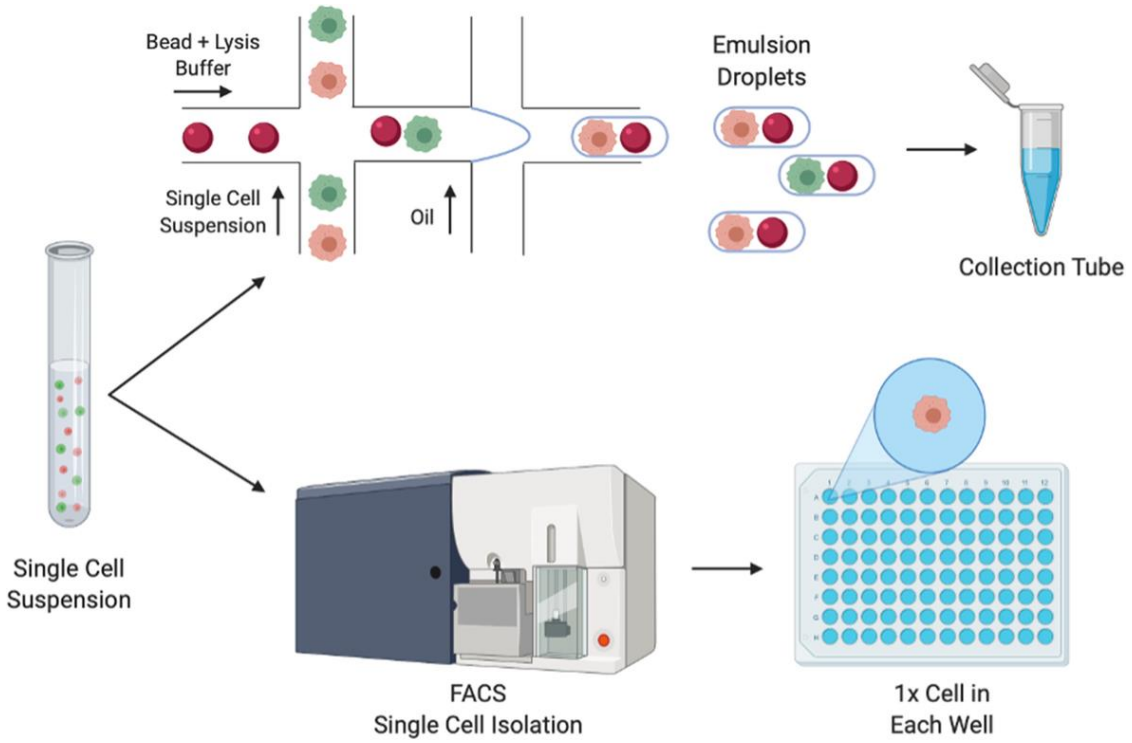
Single cell barcodes are not necessarily single cells



one event (one emulsion droplet/ one well) = one cell barcode

Probst, V., Simonyan, A., Pacheco, F. et al. Benchmarking full-length transcript single cell mRNA sequencing protocols. BMC Genomics 23, 860 (2022). https://doi.org/10.1186/s12864-022-09014-5

Single cell barcodes are not necessarily single cells



one event (one emulsion droplet/ one well) = one cell barcode

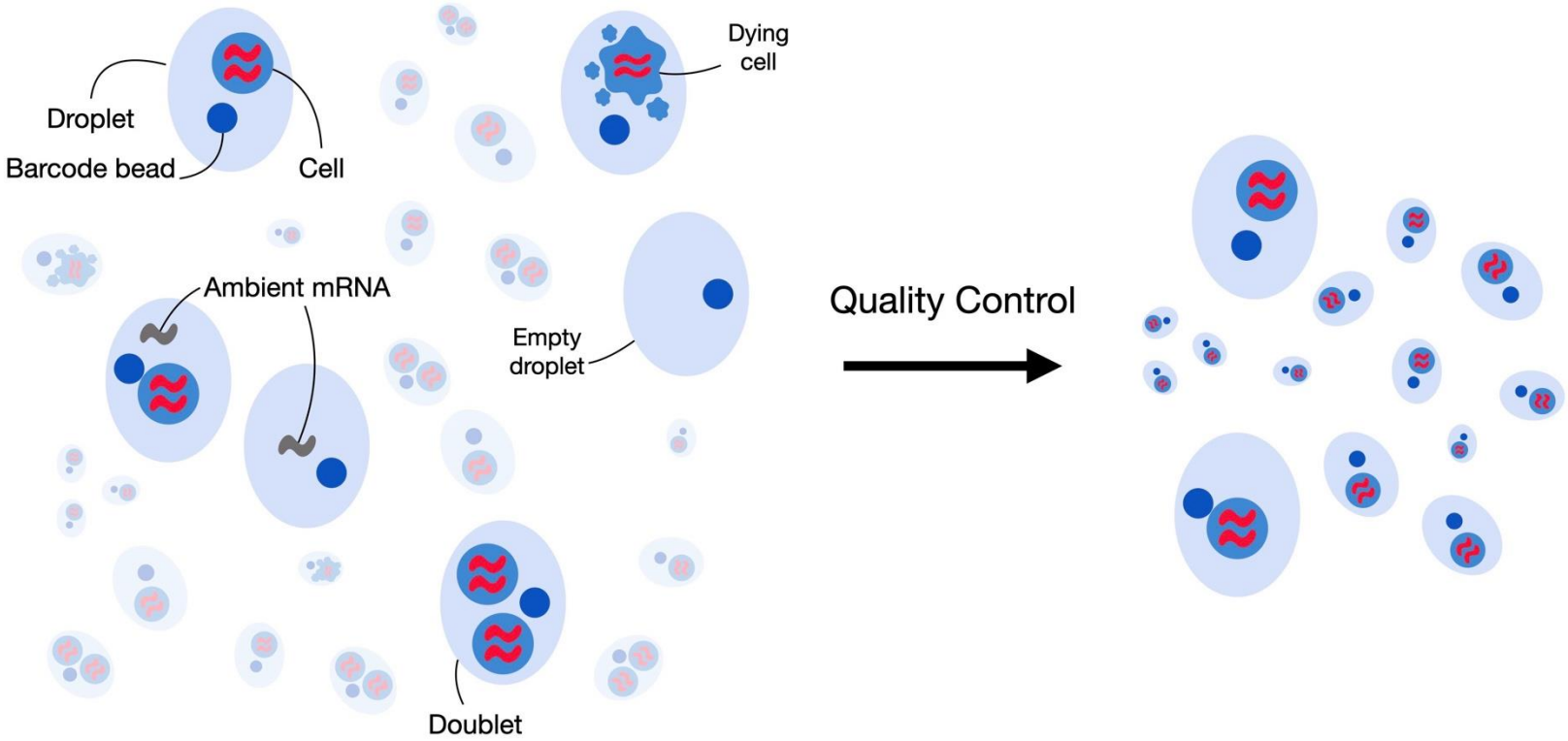
≠

one live cell

could also be
- a dead cell
- no cell
- two cells
- three cells
- …

Probst, V., Simonyan, A., Pacheco, F. et al. Benchmarking full-length transcript single cell mRNA sequencing protocols. BMC Genomics 23, 860 (2022). https://doi.org/10.1186/s12864-022-09014-5

Quality control: removing problematic events and read counts



1) Finding cellular events & basic QC measures

2) Filtering out compromised cells

3) Ambient mRNA removal

4) Doublet removal

https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html

1

Removing empty droplets (cell barcodes without cells)

Single cell alignment software (e.g. STARsolo, cellranger)

raw count matrix

	CB1	CB2	CB3	CB4	CB5
G1	1	5	0	3	0
G2	0	5	1	6	0
G3	0	2	0	4	0

Which events contain cells (one or more)?

https://cf.10xgenomics.com/samples/cell-vdj/5.0.0/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t_web_summary.html

1

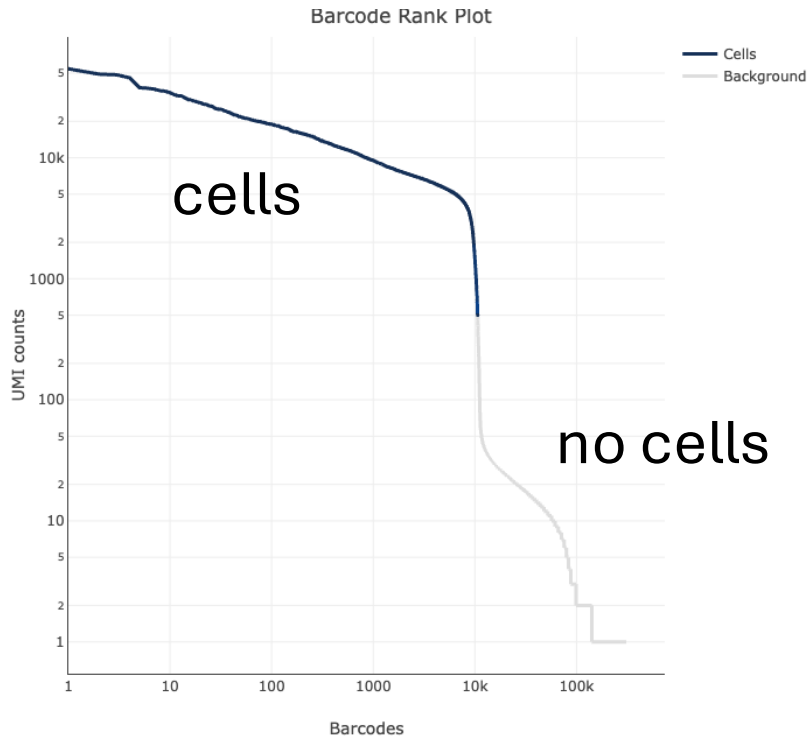
Removing empty droplets (cell barcodes without cells)

Single cell alignment software (e.g. STARsolo, cellranger)

raw count matrix

	CB1	CB2	CB3	CB4	CB5
G1	1	5	0	3	0
G2	0	5	1	6	0
G3	0	2	0	4	0

Which events contain cells (one or more)?

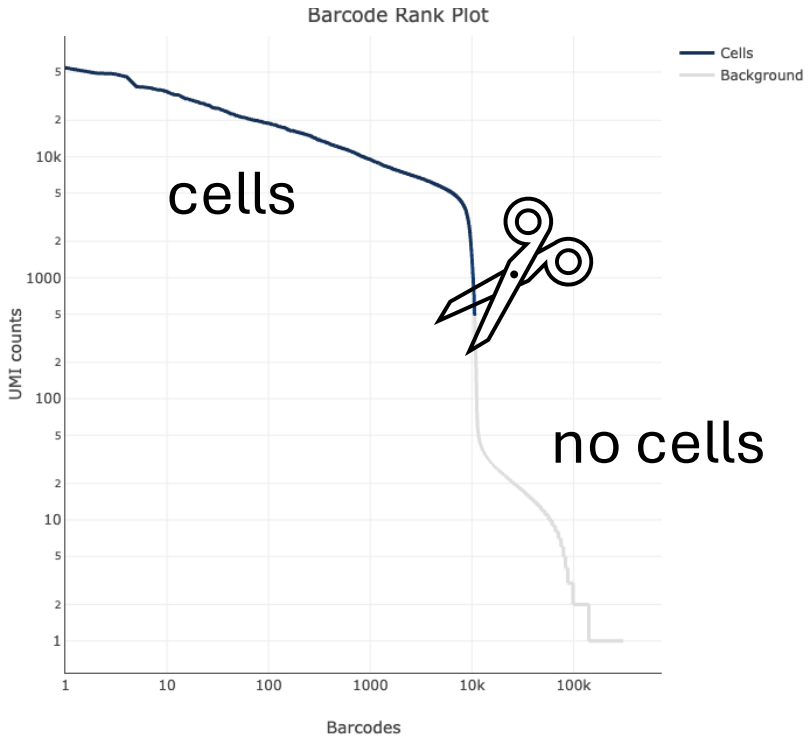


"Knee plot"

https://cf.10xgenomics.com/samples/cell-vdj/5.0.0/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t_web_summary.html

1

Removing empty droplets (cell barcodes without cells)



“Knee plot”

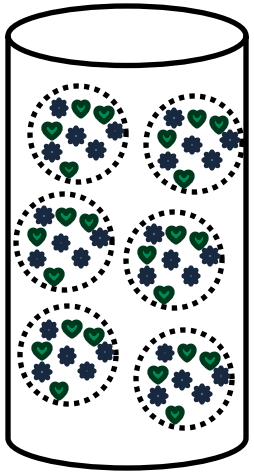
	CB1	CB2	CB3	CB4	CB5
G1	1	5	0	3	0
G2	0	5	1	6	0
G3	0	2	0	4	0

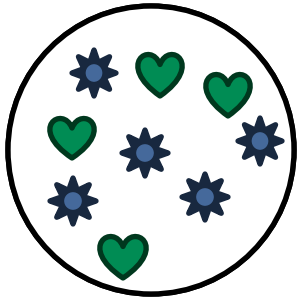
filtered count matrix

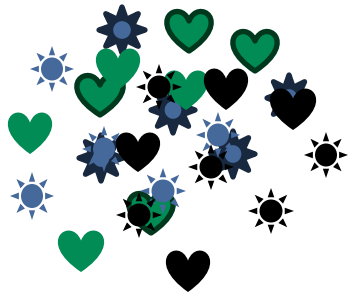
	CB2	CB4
G1	5	3
G2	5	6
G3	2	4

https://cf.10xgenomics.com/samples/cell-vdj/5.0.0/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t_web_summary.html

1 Basic Quality Control (QC) measures for sequencing runs, events and reads



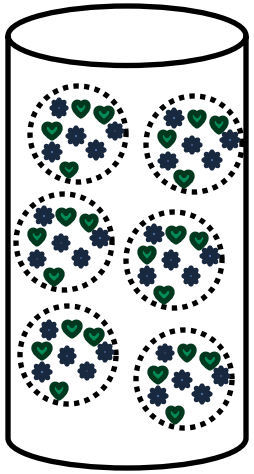


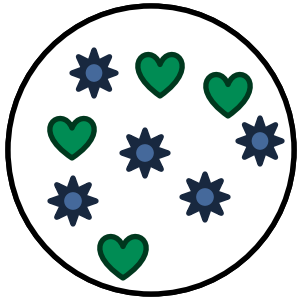


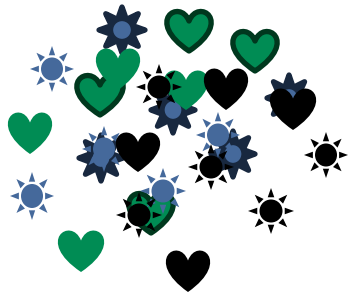
Library/Sequencing level

- Total number of reads
- Fraction of cell barcodes that are valid
- Fraction of UMIs that are valid
- Sequencing saturation

1 Basic Quality Control (QC) measures for sequencing runs, events and reads







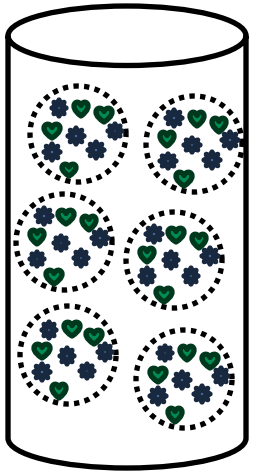
Library/Sequencing level

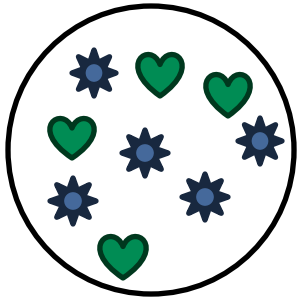
- Total number of reads
- Fraction of cell barcodes that are valid
- Fraction of UMIs that are valid
- Sequencing saturation

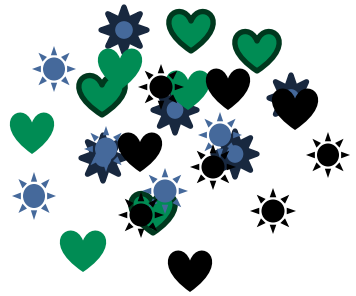
Cellular level

- Estimated number of cellular events
- reads per cell
- genes per cell
- UMIs per cell

1 Basic Quality Control (QC) measures for sequencing runs, events and reads







Library/Sequencing level

- Total number of reads
- Fraction of cell barcodes that are valid
- Fraction of UMIs that are valid
- Sequencing saturation

Cellular level

- Estimated number of cellular events
- reads per cell
- genes per cell
- UMIs per cell

Read/Alignment level

- Fraction of reads mapped to genome
- Fraction of reads mapped to transcriptome
- …introns, exons, intergenic regions

1

Example 10x cellranger count QC report

sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t - Human PBMC 10k (v2)

Count Summary | Count Analysis | VDJ-T Summary | VDJ-T Analysis | VDJ-B Summary | VDJ-B Analysis

10,548
Estimated Number of Cells

60,510
Mean Reads per Cell

1,865
Median Genes per Cell

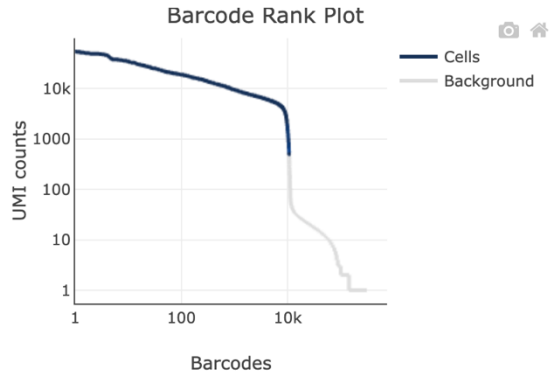
Sequencing

Number of Reads	638,257,832
Number of Short Reads Skipped	0
Valid Barcodes	91.1%
Valid UMIs	99.9%
Sequencing Saturation	83.7%
Q30 Bases in Barcode	95.3%
Q30 Bases in RNA Read	91.3%
Q30 Bases in UMI	95.2%

Mapping

Reads Mapped to Genome	93.4%
Reads Mapped Confidently to Genome	80.3%
Reads Mapped Confidently to Intergenic Regions	4.0%
Reads Mapped Confidently to Intronic Regions	7.7%
Reads Mapped Confidently to Exonic Regions	68.6%
Reads Mapped Confidently to Transcriptome	63.3%
Reads Mapped Antisense to Gene	3.5%

Cells



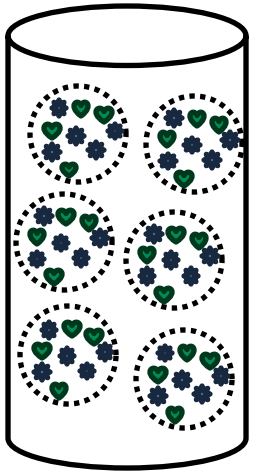
Estimated Number of Cells	10,548
Fraction Reads in Cells	98.1%
Mean Reads per Cell	60,510
Median Genes per Cell	1,865
Total Genes Detected	23,571
Median UMI Counts per Cell	5,504

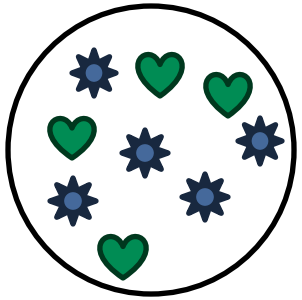
Sample

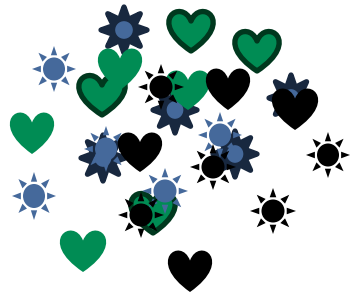
Sample ID	sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t
Sample Description	Human PBMC 10k (v2)
Chemistry	Single Cell 5' R2-only
Include introns	False
Reference Path	…references/refdata-gex-GRCh38-2020-A
Transcriptome	GRCh38-2020-A
Pipeline Version	cellranger-5.0.0

https://cf.10xgenomics.com/samples/cell-vdj/5.0.0/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t_web_summary.html

1 Basic Quality Control (QC) measures for sequencing runs, events and reads







Library/Sequencing level

- Total number of reads
- Fraction of cell barcodes that are valid
- Fraction of UMIs that are valid
- Sequencing saturation

Cellular level

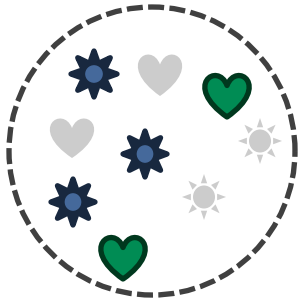
- Estimated number of cellular events
- reads per cell
- genes per cell
- UMIs per cell

Read/Alignment level

- Fraction of reads mapped to genome
- Fraction of reads mapped to transcriptome
- …introns, exons, intergenic regions

2

Filtering low quality cells



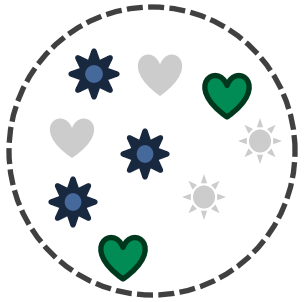
Stressed/dying cell with broken membrane

mRNA degrades → less UMI counts and genes

Mitochondrial RNA is protected by an extra membrane → fraction of mt RNA rises

https://satijalab.org/seurat/articles/pbmc3k_tutorial

Filtering low quality cells



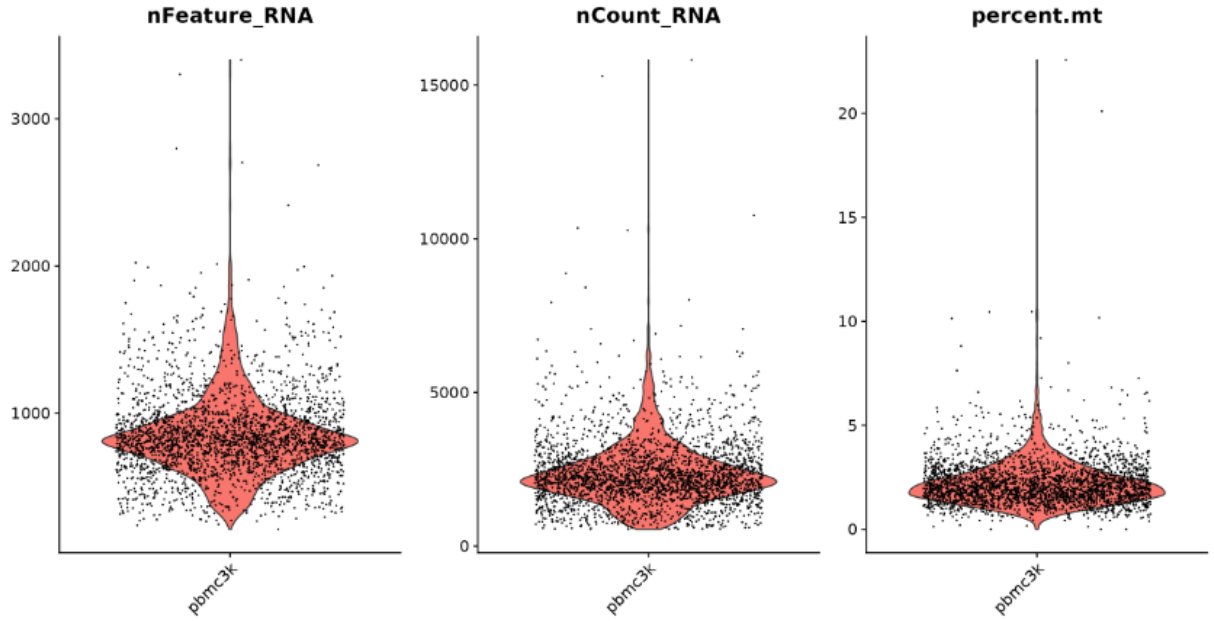
Stressed/dying cell with broken membrane

mRNA degrades → less UMI counts and genes

Mitochondrial RNA is protected by an extra membrane → fraction of mt RNA rises

3 most common QC metrics for cell filtering

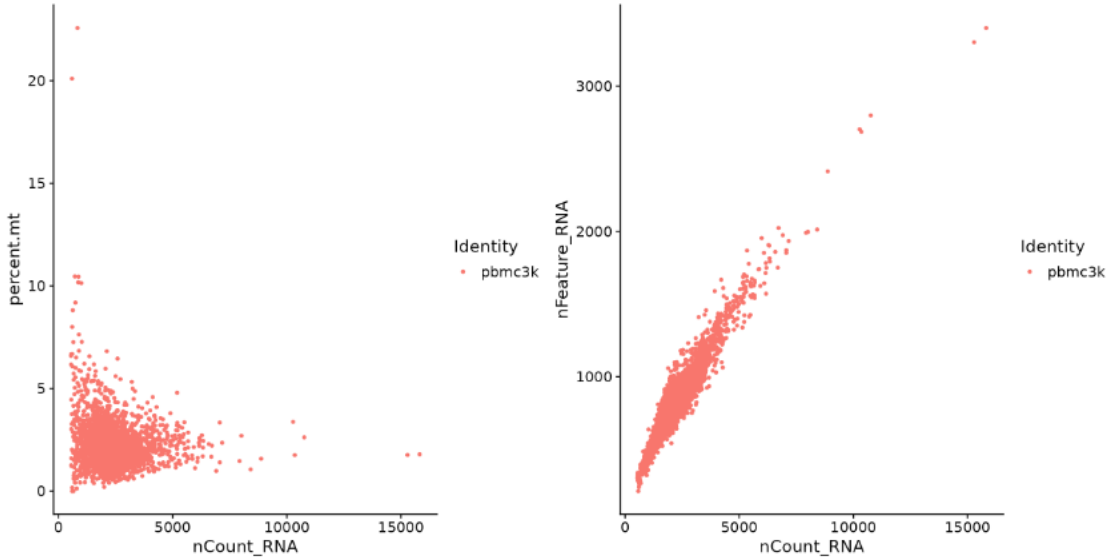
1) Number of detected genes per barcode
2) Number of counts (UMIs) per barcode
3) Fraction of mitochondrial read counts per barcode



https://satijalab.org/seurat/articles/pbmc3k_tutorial

2

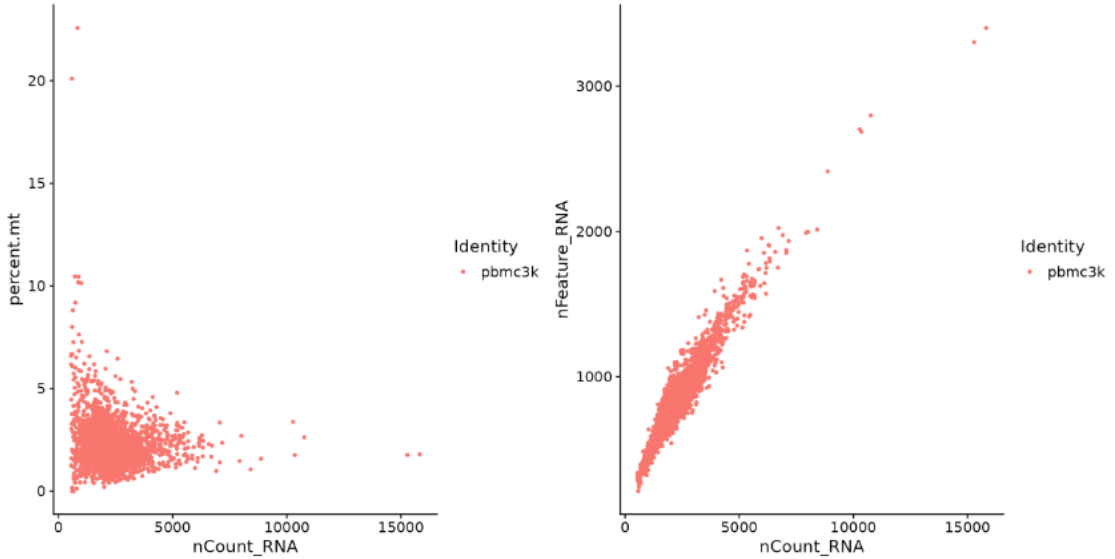
Filtering low quality cells

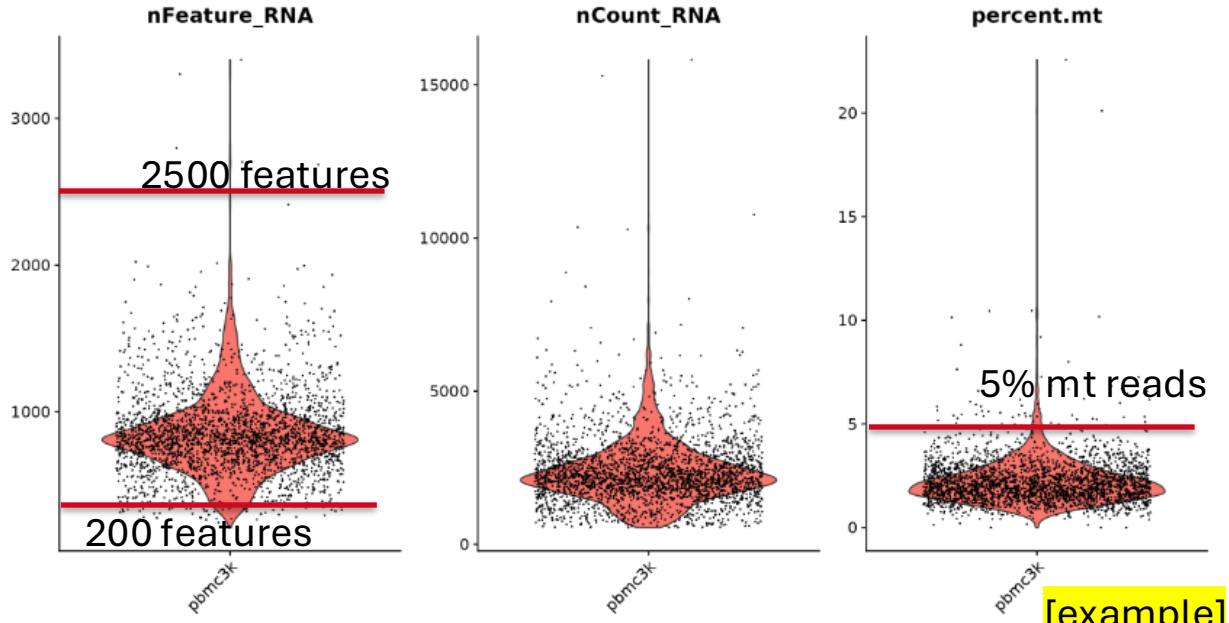


https://satijalab.org/seurat/articles/pbmc3k_tutorial

2

Filtering low quality cells





[example]

Option 1

Filtering with manually chosen cut-offs after visual inspection

https://satijalab.org/seurat/articles/pbmc3k_tutorial

2

Filtering low quality cells

Option 2
Automatic thresholding via median absolute deviations (MAD)

MAD is a measure of statistical dispersion (like standard deviation, but more robust)

$$\mathrm{MAD} = \mathrm{median}(|X_i - \tilde{X}|)$$

$$\tilde{X} = \mathrm{median}(X)$$

https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html

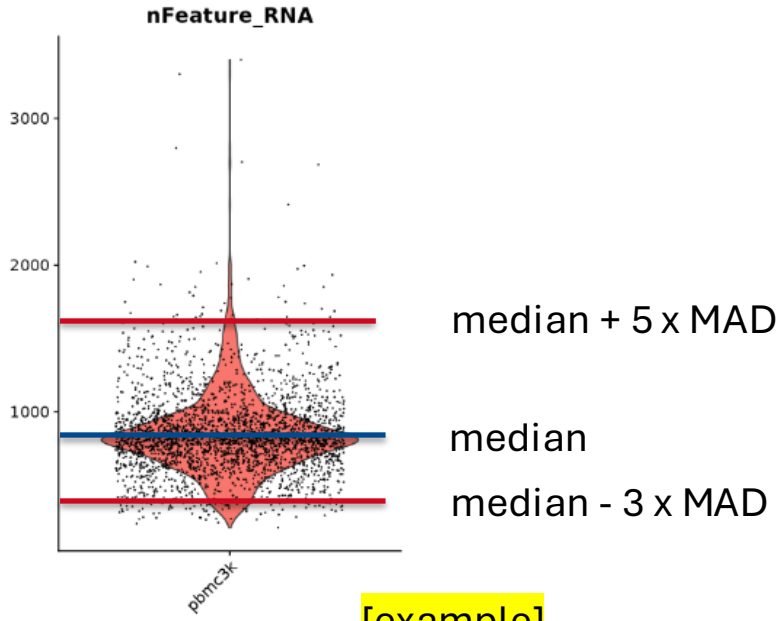
Filtering low quality cells

Option 2
Automatic identification of outliers via median absolute deviations (MAD)

MAD is a measure of statistical dispersion (like standard deviation, but more robust)

$$\mathrm{MAD} = \mathrm{median}(|X_i - \tilde{X}|)$$

$$\tilde{X} = \mathrm{median}(X)$$



median + 5 x MAD

median

median - 3 x MAD

[example]

advantage: can be applied automatically, e.g. if there are many data sets/samples

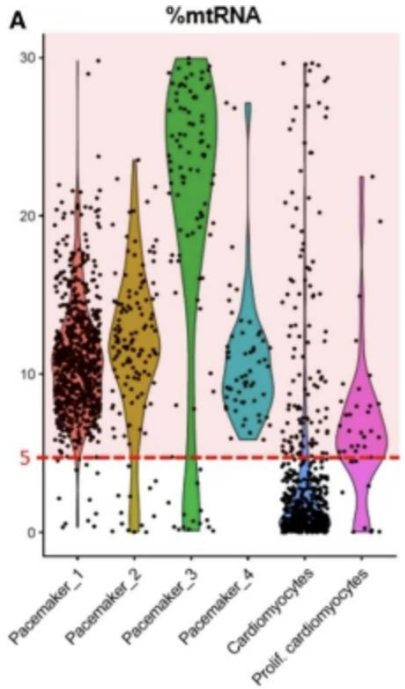
risk: some cell types have higher average RNA content than others, may get filtered out

https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html

2

Filtering low quality cells – general advice

→ Be permissive during initial filtering and revisit later, e.g. remove clusters with high average mitochondrial read fraction.



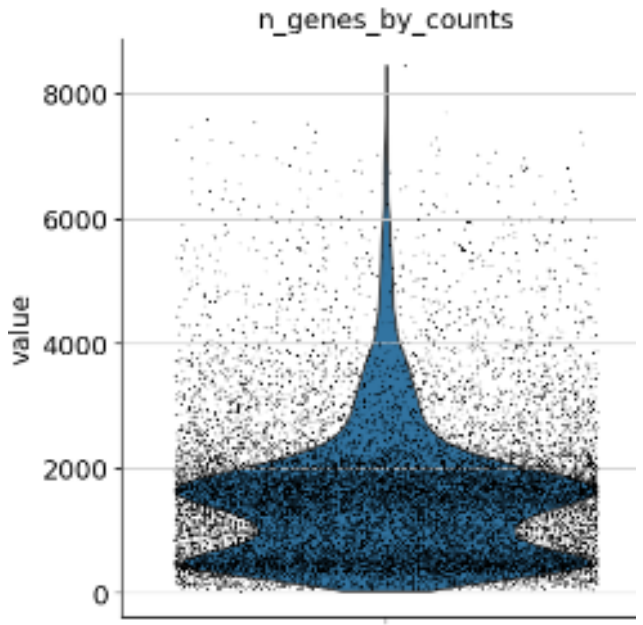
Galow, AM., Kussauer, S., Wolfien, M. et al. Quality control in scRNA-Seq can discriminate pacemaker cells: the mtRNA bias. Cell. Mol. Life Sci. 78, 6585–6592 (2021). https://doi.org/10.1007/s00018-021-03916-5

2

Filtering low quality cells – general advice

→ Be permissive during initial filtering and revisit later, e.g. remove clusters with high average mitochondrial read fraction.

→ Perform filtering per batch / per sample as QC metrics may vary strongly between them.



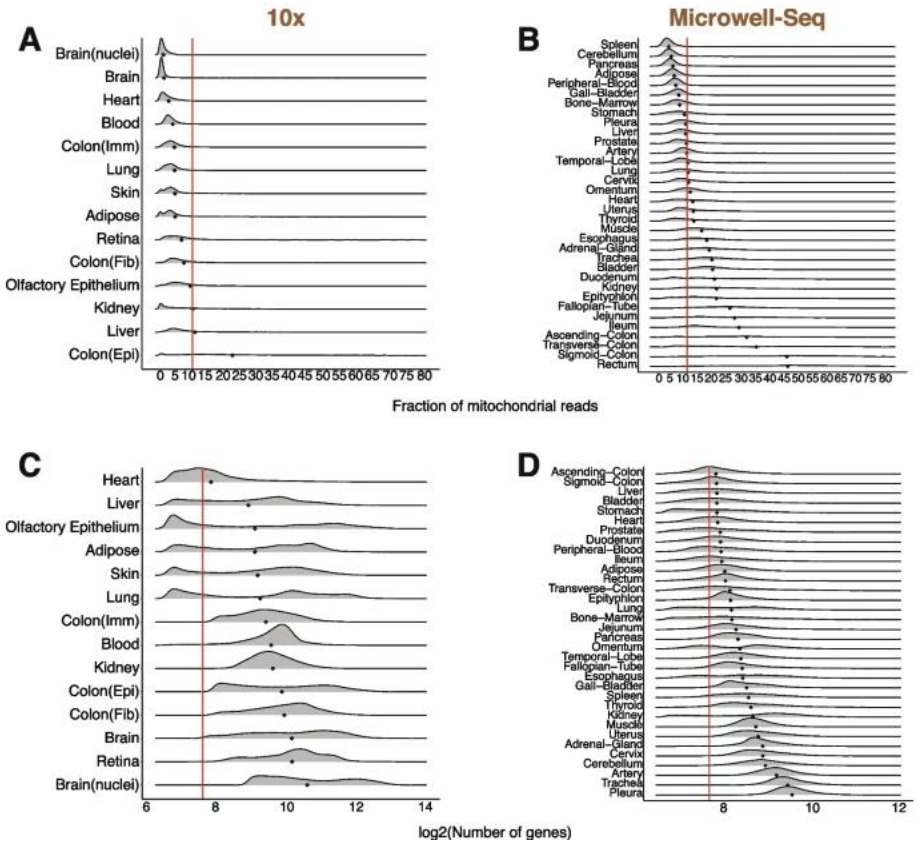
https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html

Filtering low quality cells – general advice

→ Be permissive during initial filtering and revisit later, e.g. remove clusters with high average mitochondrial read fraction.

→ Perform filtering per batch / per sample as QC metrics may vary strongly between them.

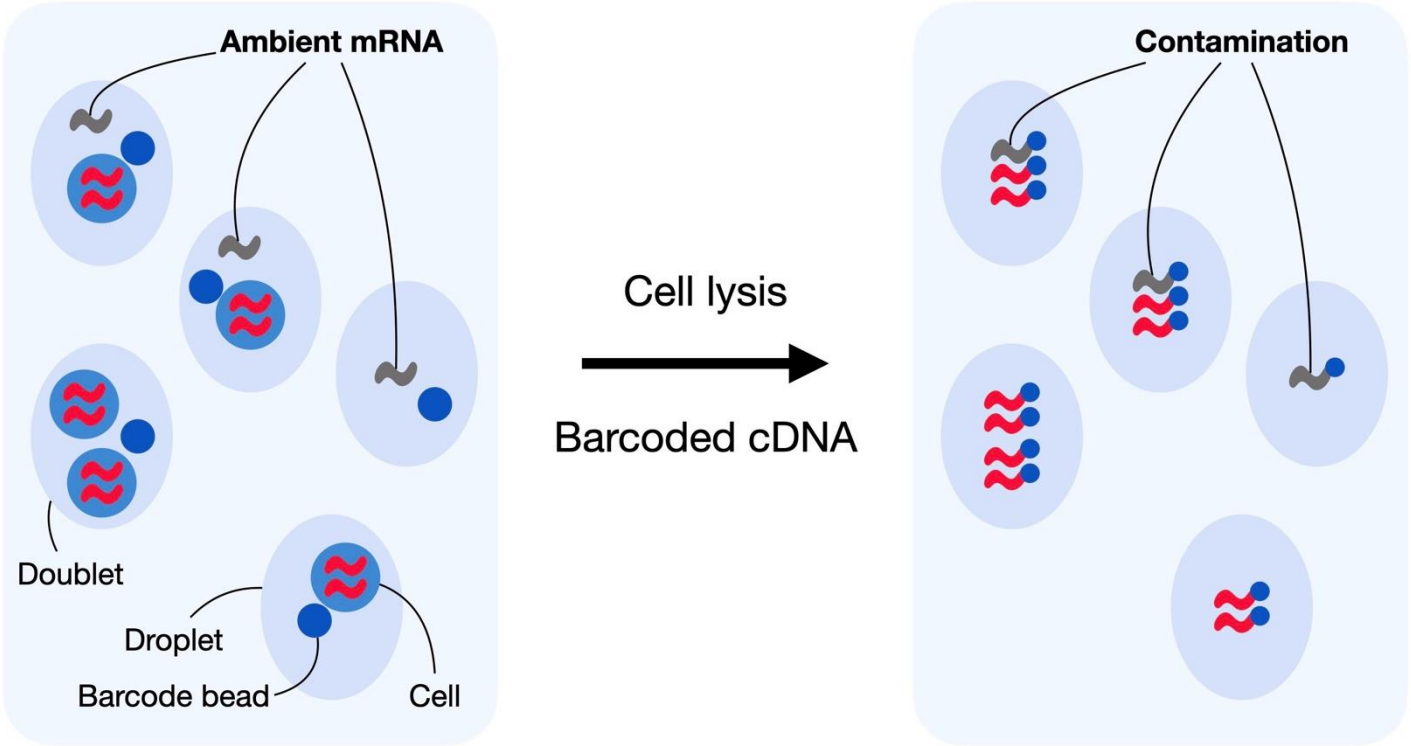
→ QC metrics vary by tissue and protocol, don't freak out if your values are different from tutorials.



Subramanian, A., Alperovich, M., Yang, Y. et al. Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. Genome Biol 23, 267 (2022). https://doi.org/10.1186/s13059-022-02820-w

3

Ambient mRNA / mRNA “soup”: cell-free mRNA from burst cells enters reaction volumes (droplets, wells)

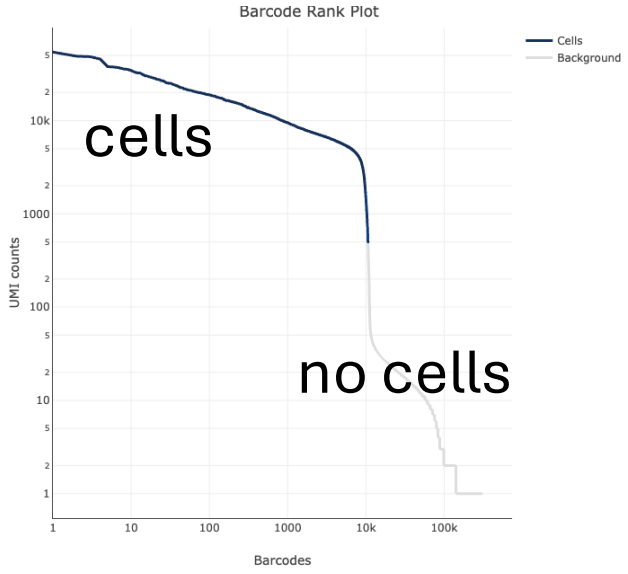


https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html

3

Removal of ambient mRNA / mRNA “soup”: basic idea

Total mRNA counts for gene x in cell y	=	True counts from cell y	+	Extra counts from the soup
measure				estimate from empty cells



https://cf.10xgenomics.com/samples/cell-vdj/5.0.0/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t_web_summary.html

3

Removal of ambient mRNA / mRNA “soup”: basic idea

True counts from cell y	=	Total mRNA counts for gene x in cell y	-	Extra counts from the soup

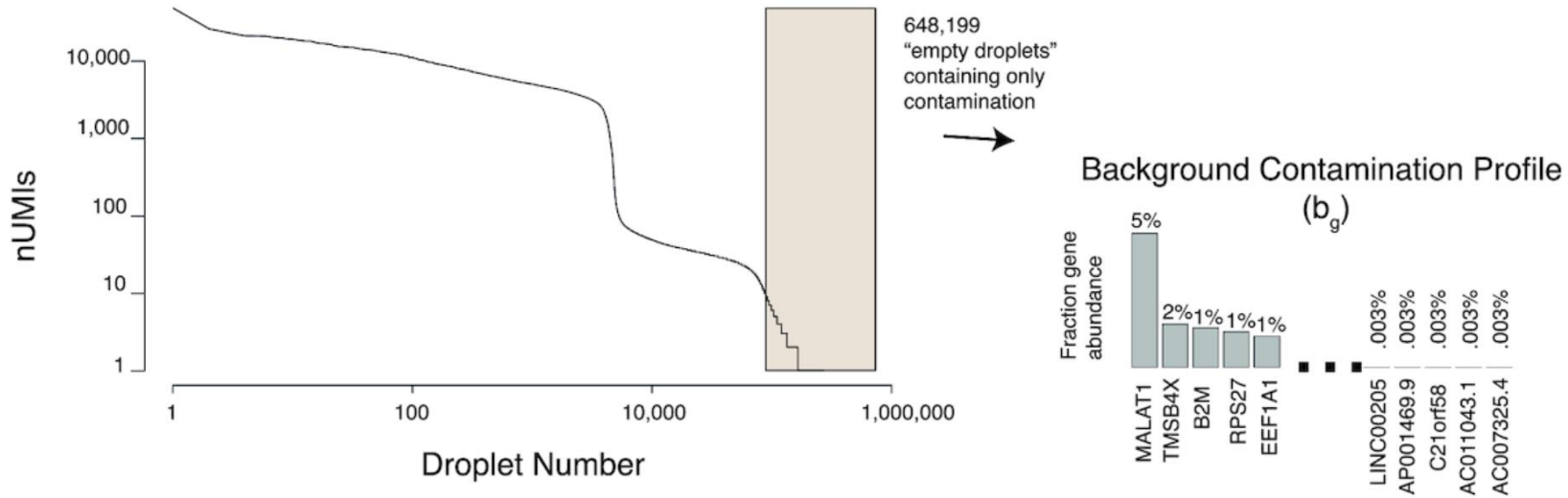
Potential problems with naïve approach:

- After substraction, you may end up with negative counts / non-integers
- Individual cells have different library sizes (count sums), so a one-size-fits-all substraction of contamination may not be appropriate

https://cf.10xgenomics.com/samples/cell-vdj/5.0.0/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t_web_summary.html

Removal of ambient mRNA, example method: SoupX

1) Determine **contamination profile** from empty droplets



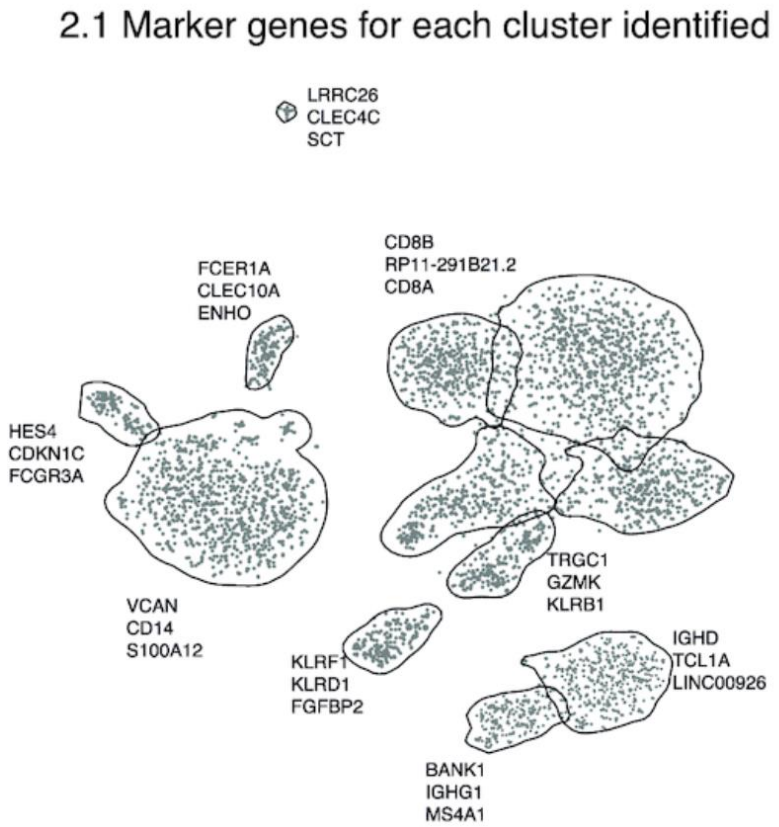
Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w

https://cf.10xgenomics.com/samples/cell-vdj/5.0.0/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t/sc5p_v2_hs_PBMC_10k_multi_5gex_5fb_b_t_web_summary.html

Removal of ambient mRNA, example method: SoupX

1) Determine **contamination profile** from empty droplets

2) Estimate the **contamination rate** cell-containing droplets via highly specific marker genes



Assumption: strong marker gene (e.g. CD8A) of one cluster is not expressed in the other clusters → occurrence in another cluster is contamination

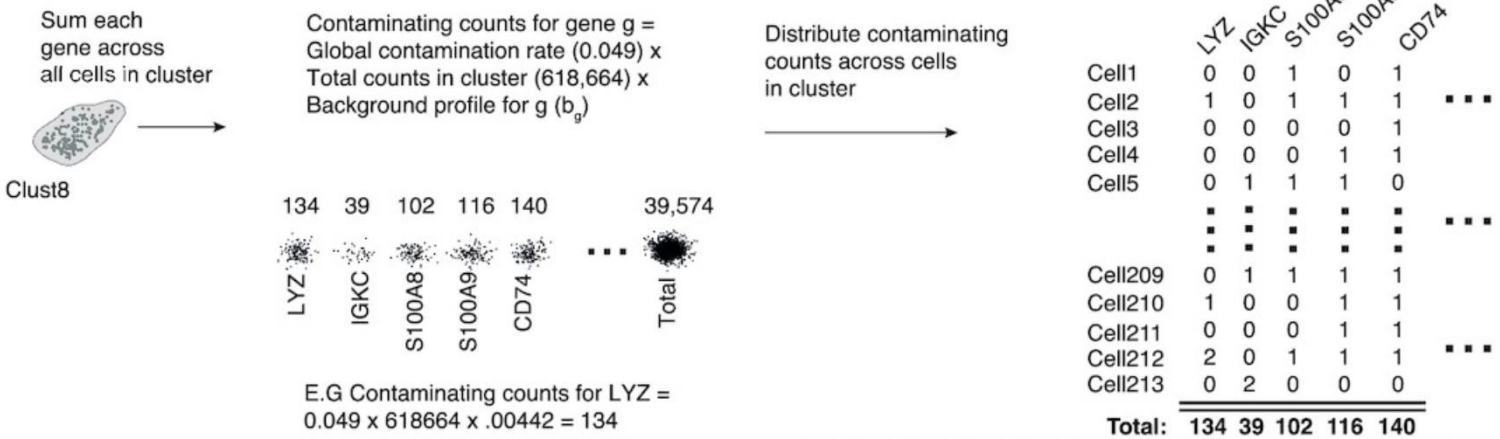
Matthew D Young, Sam Behjati, SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data, *GigaScience*, Volume 9, Issue 12, December 2020, giaa151, https://doi.org/10.1093/gigascience/giaa151

Removal of ambient mRNA, example method: SoupX

1) Determine **contamination profile** from empty droplets

2) Estimate the **contamination rate** cell-containing droplets via highly specific marker genes

3) Draw **corrected counts** from a multinomial model (positive integer output) using contamination profiles, contamination rate and measured counts as input



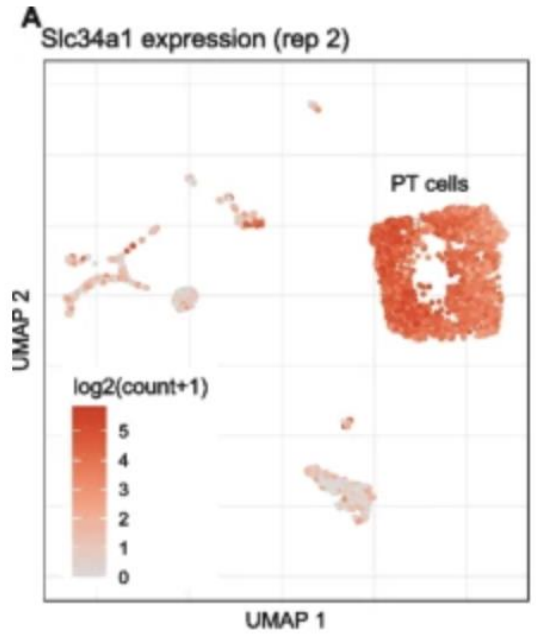
Matthew D Young, Sam Behjati, SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data, *GigaScience*, Volume 9, Issue 12, December 2020, giaa151, https://doi.org/10.1093/gigascience/giaa151

3

Ambient RNA removal has positive effects on downstream analysis

The impact of contamination on marker gene analyses

The ability to distinguish hitherto unknown cell types and states is one of the greatest achievements made possible by single cell transcriptome analyses. To this end, marker genes are commonly used to annotate cell clusters for which available classifications appear insufficient. An ideal marker gene would be expressed in all cells of one type but in none of the other present cell types. Thus, when comparing expression levels of one cell type versus all others, we expect high log2-fold changes, the higher the change the more reliable the marker. However, such a reliance on marker genes also makes this type of analysis vulnerable to background noise. Our whole kidney data can illustrate this problem well, because with the

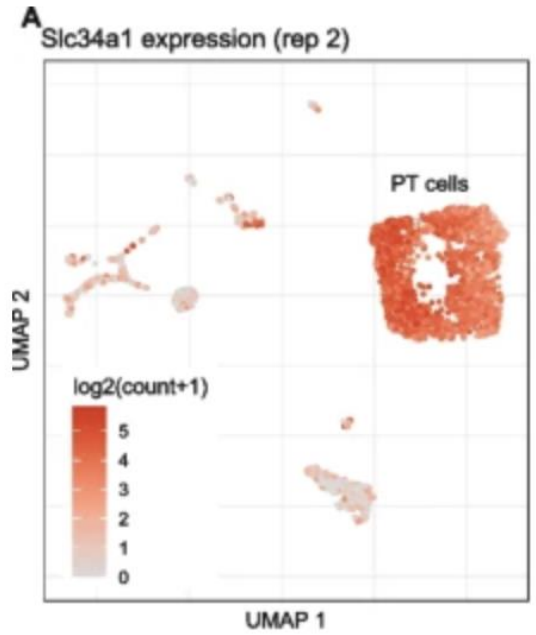


Janssen, P., Kliesmete, Z., Vieth, B. *et al.* The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biol* **24**, 140 (2023). https://doi.org/10.1186/s13059-023-02978-x

Ambient RNA removal has positive effects on downstream analysis

The impact of contamination on marker gene analyses

The ability to distinguish hitherto unknown cell types and states is one of the greatest achievements made possible by single cell transcriptome analyses. To this end, marker genes are commonly used to annotate cell clusters for which available classifications appear insufficient. An ideal marker gene would be expressed in all cells of one type but in none of the other present cell types. Thus, when comparing expression levels of one cell type versus all others, we expect high log2-fold changes, the higher the change the more reliable the marker. However, such a reliance on marker genes also makes this type of analysis vulnerable to background noise. Our whole kidney data can illustrate this problem well, because with the



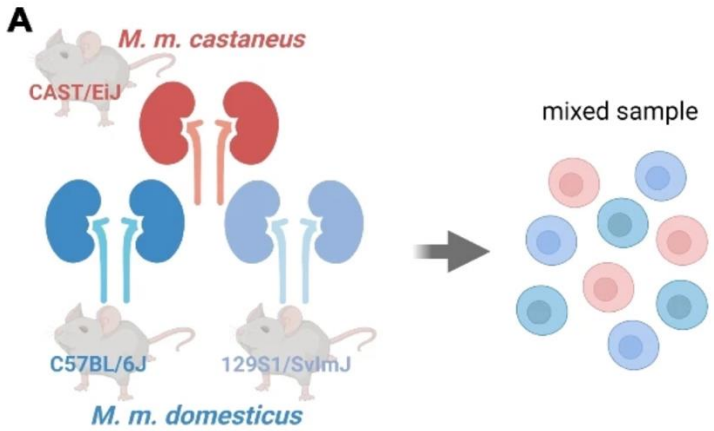
Clean data →

- Clearer clusters
- More meaningful marker genes
- Better results in differential gene expression analyses

Janssen, P., Kliesmete, Z., Vieth, B. *et al.* The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biol* **24**, 140 (2023). https://doi.org/10.1186/s13059-023-02978-x

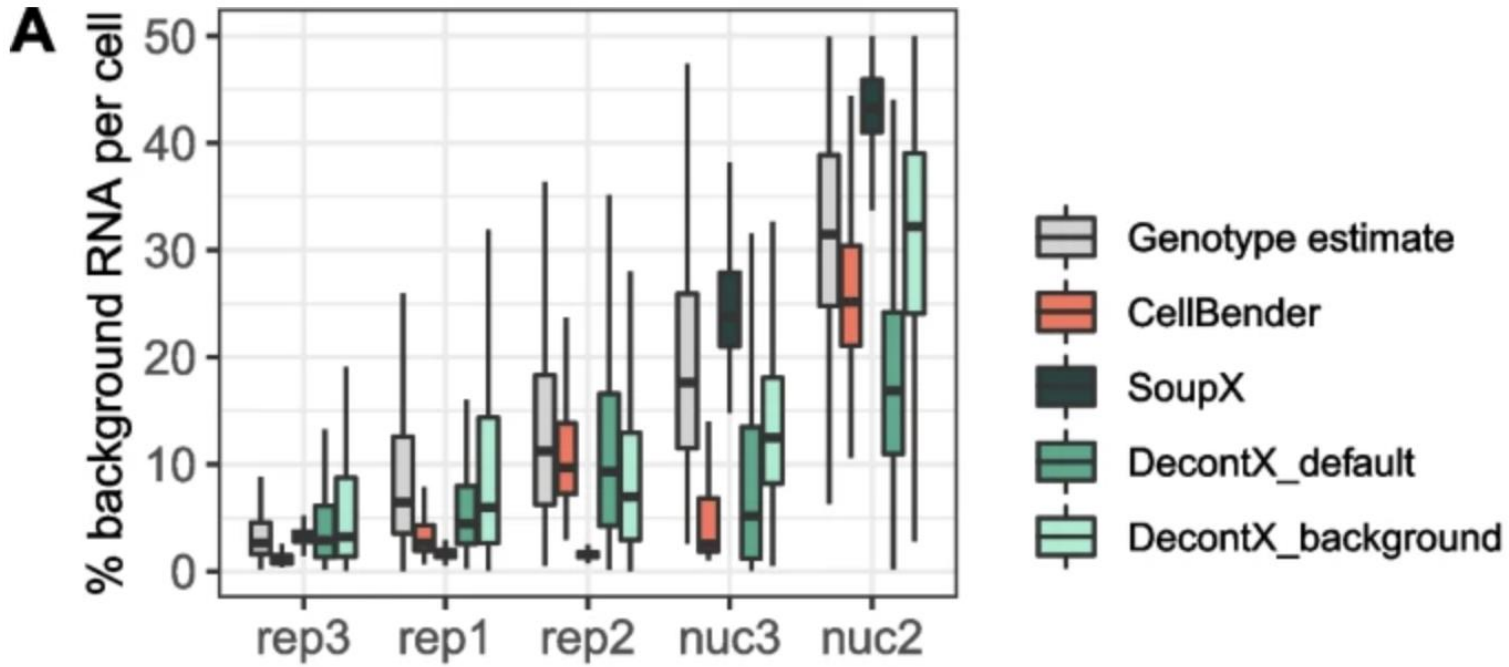
3

Benchmarking ambient RNA removal tools: Cellbender performs slightly better than other methods



experimental setup with 3 mouse strains allows to access ground truth ambient RNA fractions

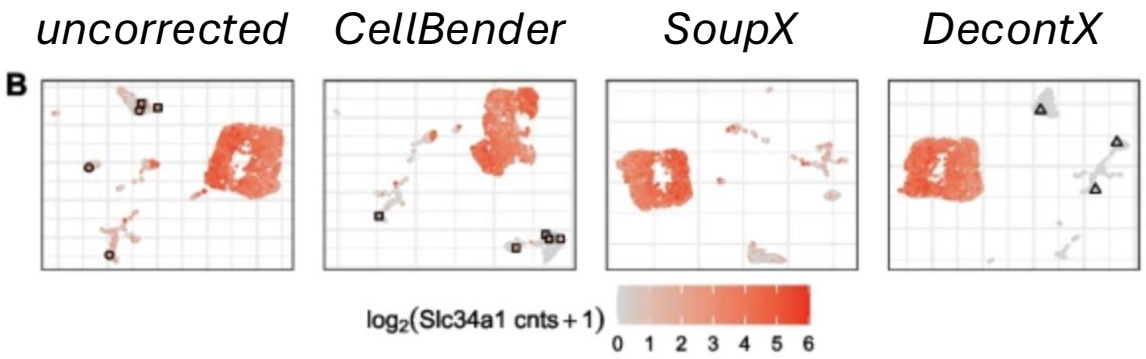
Fig. 5

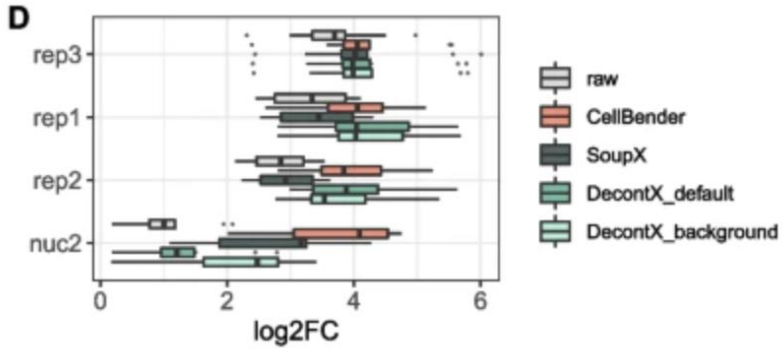


Janssen, P., Kliesmete, Z., Vieth, B. *et al.* The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biol* **24**, 140 (2023). https://doi.org/10.1186/s13059-023-02978-x

3

Benchmarking ambient RNA removal tools: recommendations



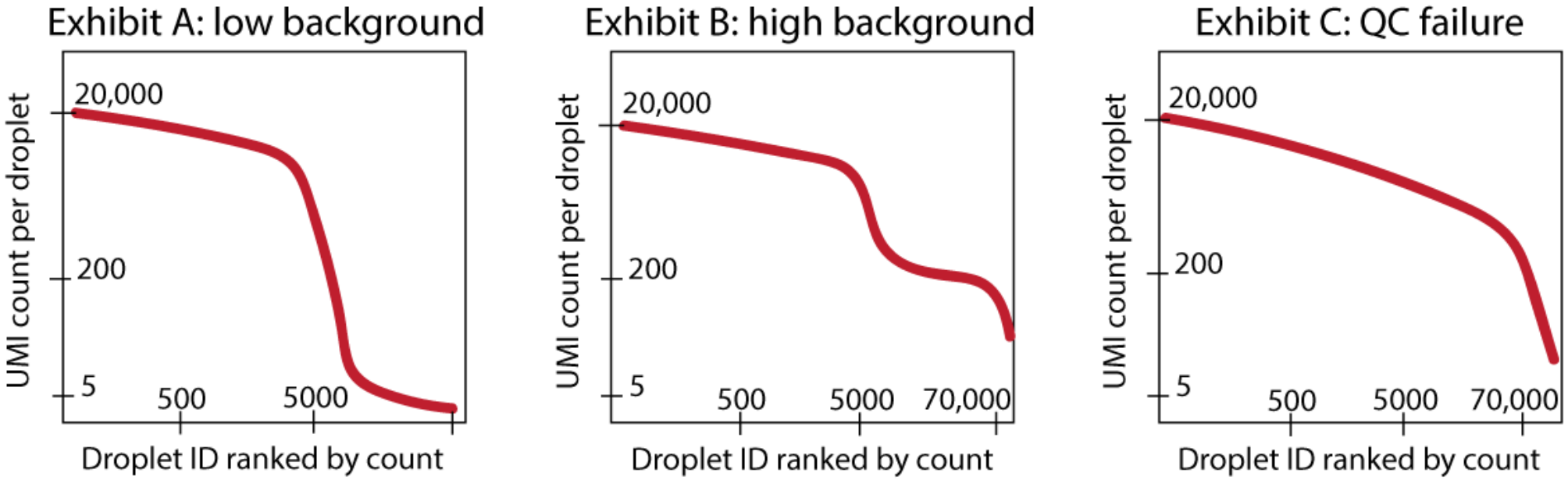


- Ambient RNA removal should always be performed if the goal is marker gene identification
- Classification, clustering and pseudotime analyses are generally robust enough to not require ambient RNA removal → for these analyses, only correct if background RNA levels are high

Janssen, P., Kliesmete, Z., Vieth, B. *et al.* The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biol* **24**, 140 (2023). https://doi.org/10.1186/s13059-023-02978-x

3

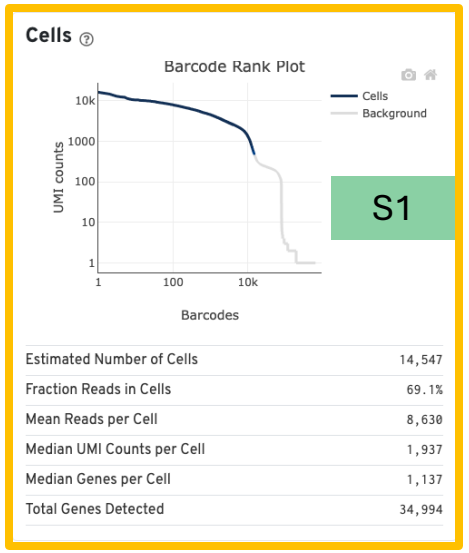
Knee plots help determine the level of ambient RNA

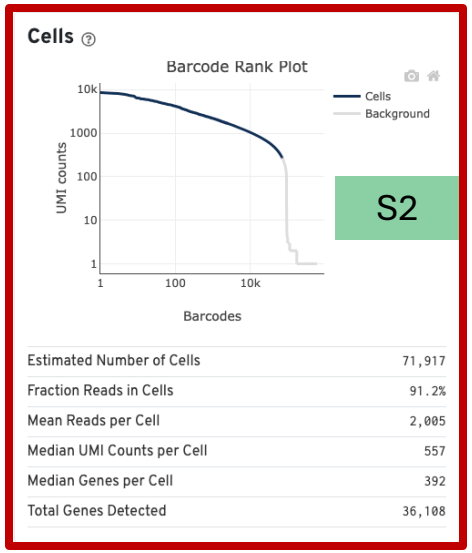


https://cellbender.readthedocs.io/en/latest/_images/UMI_curve_tropes.png

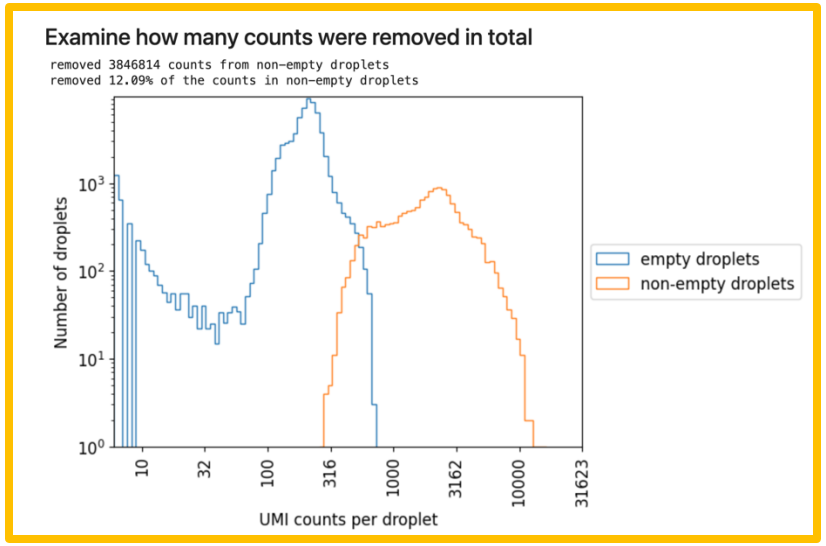
3

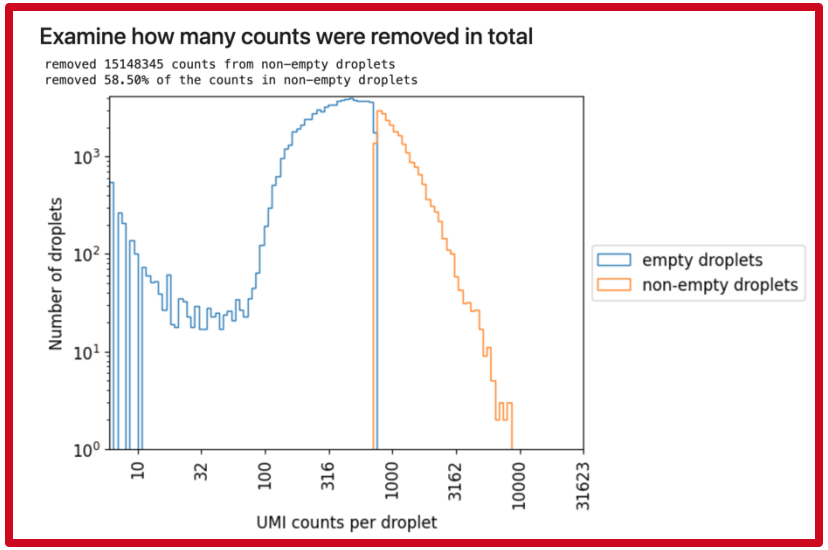
Not every experimental failure can be cleaned up





Application of cellbender





Article | Published: 07 August 2023

Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender

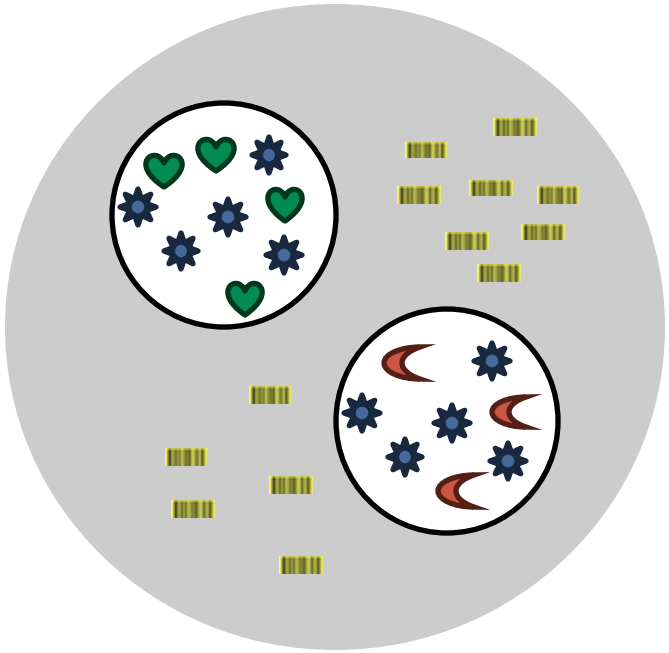
Stephen J. Fleming, Mark D. Chaffin, Alessandro Arduini, Amer-Denis Akkad, Eric Banks, John C. Marioni, Anthony A. Philippakis, Patrick T. Ellinor & Mehrtash Babadi

Nature Methods **20**, 1323–1335 (2023) | Cite this article

https://cellbender.readthedocs.io/en/latest/_images/UMI_curve_tropes.png

4

Doublet detection

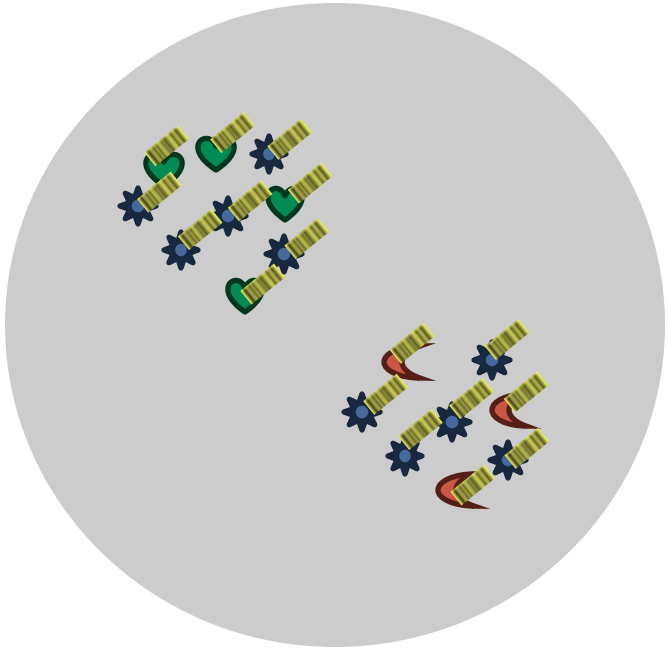


droplet / well

two or more cells enter the same droplet or well

4

Doublet detection

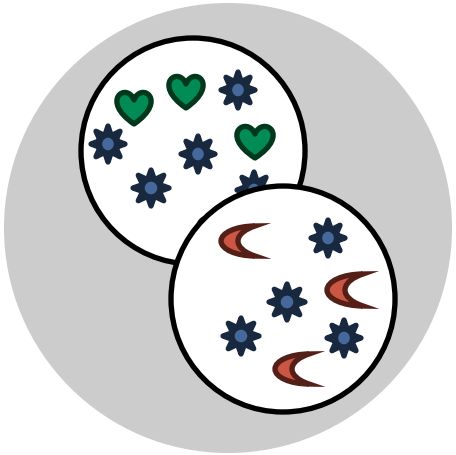


droplet / microwell

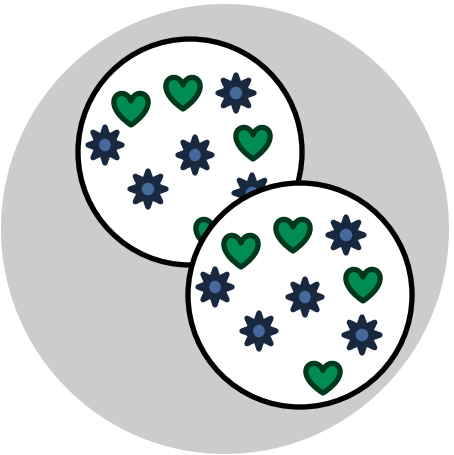
all mRNA molecules get labelled with the same cell barcode

In the count matrix, they become one row (column)

Heterotypic doublet



Homotypic doublet



4

From doublet avoidance to doublet acceptance with multiplexing

Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~825	~500
~0.8%	~1,650	~1,000
~1.6%	~3,300	~2,000
~2.4%	~4,950	~3,000
~3.2%	~6,600	~4,000
~4.0%	~8,250	~5,000
~4.8%	~9,900	~6,000
~5.6%	~11,550	~7,000
~6.4%	~13,200	~8,000
~7.2%	~14,850	~9,000
~8.0%	~16,500	~10,000

Example: doublet rates for 10x Chromium 3' v3.1

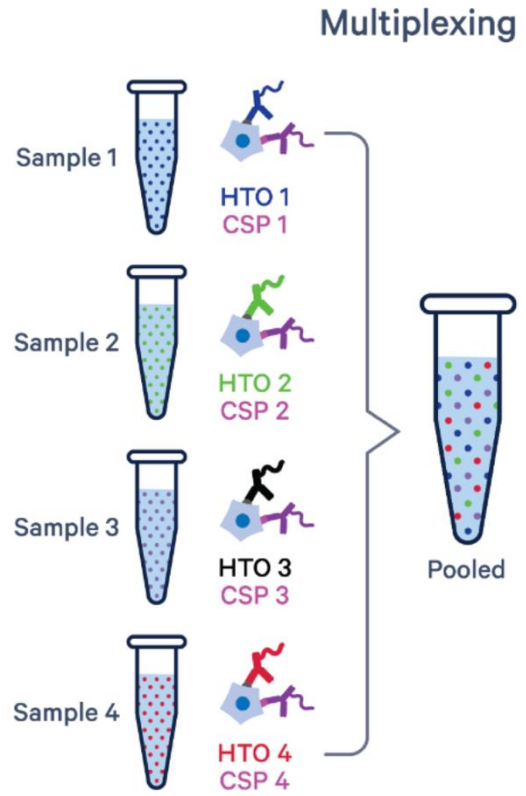
https://kb.10xgenomics.com/hc/en-us/articles/360001378811-What-is-the-maximum-number-of-cells-that-can-be-profiled

4

From doublet avoidance to doublet acceptance with multiplexing

Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~825	~500
~0.8%	~1,650	~1,000
~1.6%	~3,300	~2,000
~2.4%	~4,950	~3,000
~3.2%	~6,600	~4,000
~4.0%	~8,250	~5,000
~4.8%	~9,900	~6,000
~5.6%	~11,550	~7,000
~6.4%	~13,200	~8,000
~7.2%	~14,850	~9,000
~8.0%	~16,500	~10,000

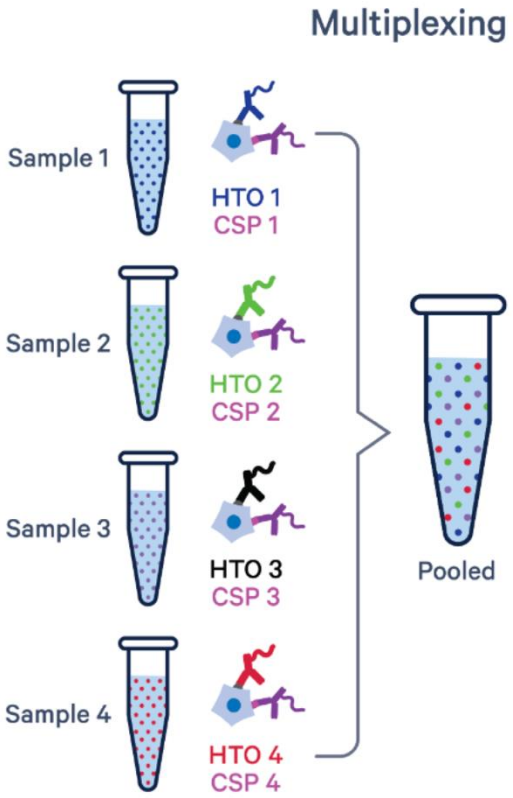
Example: doublet rates for 10x Chromium 3' v3.1



https://kb.10xgenomics.com/hc/en-us/articles/360001378811-What-is-the-maximum-number-of-cells-that-can-be-profiled

4

From doublet avoidance to doublet acceptance with multiplexing

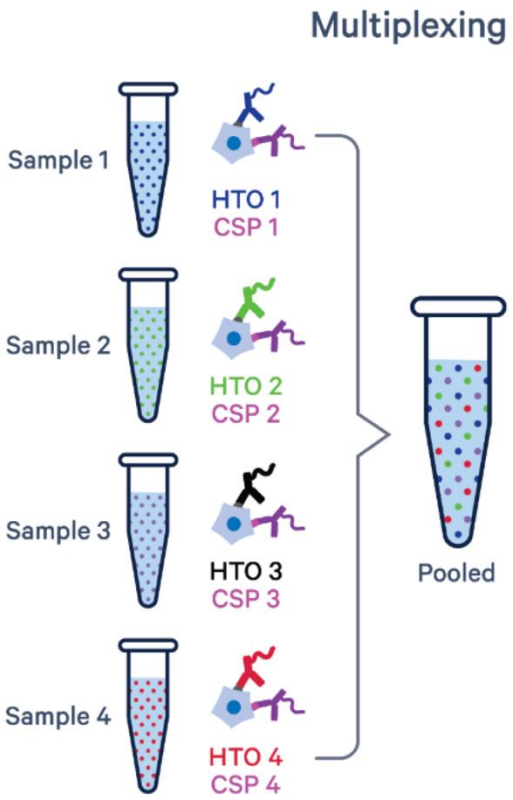


Most doublets will identifiable because they contain more than one hashtag oligonucleotide sequence!

https://kb.10xgenomics.com/hc/en-us/articles/360056584872-How-many-cell-multiplets-will-remain-undetected-in-my-final-data-when-using-the-3-CellPlex-Kit-for-Cell-Multiplexing

4

From doublet avoidance to doublet acceptance with multiplexing



Most doublets will identifiable because they contain more than one hashtag oligonucleotide sequence!

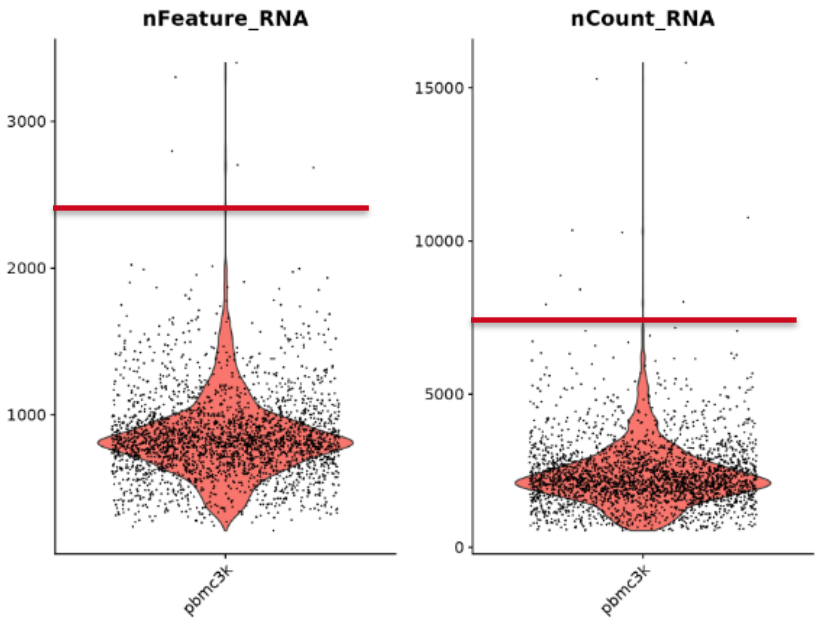
		Targeted Cell Recovery			
		5,000	10,000	20,000	30,000
Cell Barcodes Detected		4,800	9,200	16,900	23,400
Singlets		4,600	8,400	14,100	17,700
Multiplets		210	780	2,800	5,600
Multiplet Rate		~4%	~8%	~16%	~24%
Expected number of multiplets after Cell Ranger filtering					
2 tags	Dectected multiplets	105	390	1,400	2,800
	Undetected multiplets	105	390	1,400	2,800
4 tags	Dectected multiplets	158	580	2,100	4,200
	Undetected multiplets	52	200	700	1,400
8 tags	Dectected multiplets	185	680	2,460	4,930
	Undetected multiplets	25	100	340	670
12 tags	Dectected multiplets	193	720	2,580	5,150
	Undetected multiplets	17	60	220	450

Example: doublet rates for 10x Chromium 3' v3.1 with multiplexing

https://kb.10xgenomics.com/hc/en-us/articles/360056584872-How-many-cell-multiplets-will-remain-undetected-in-my-final-data-when-using-the-3-CellPlex-Kit-for-Cell-Multiplexing

4

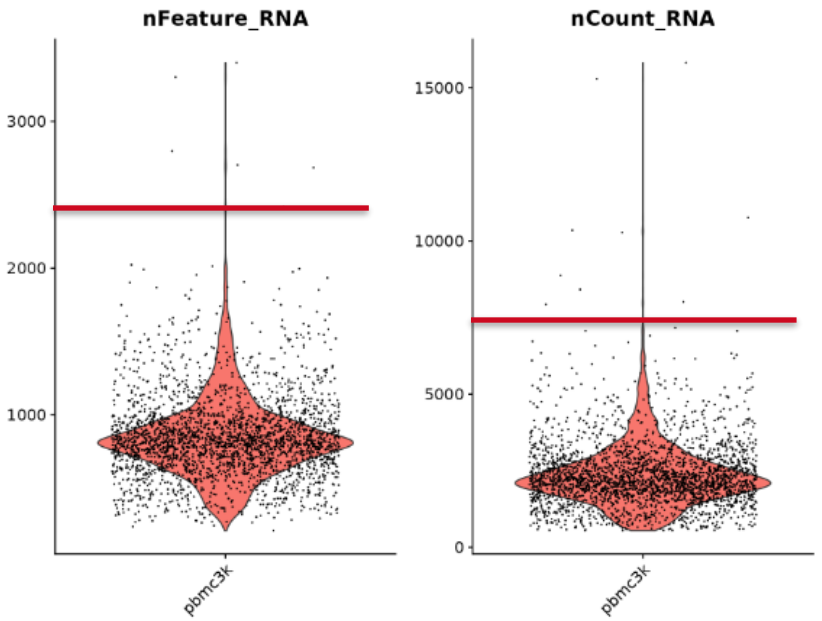
Doublet detection with computational methods basic ideas

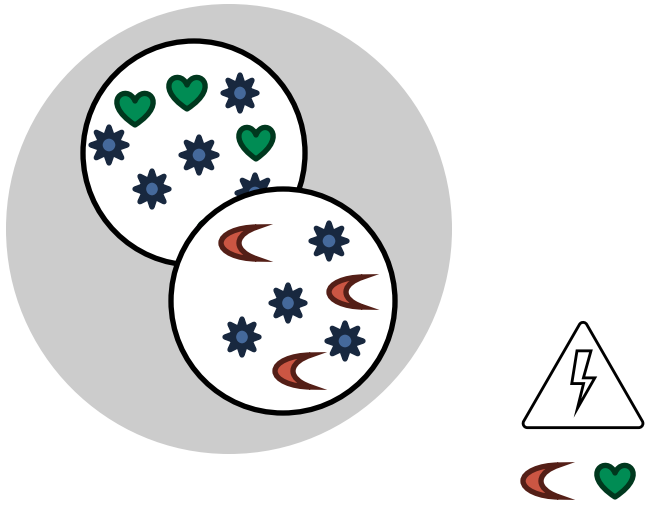


Events with many counts / many genes are suspicious
→ use upper threshold

4

Doublet detection with computational methods basic ideas



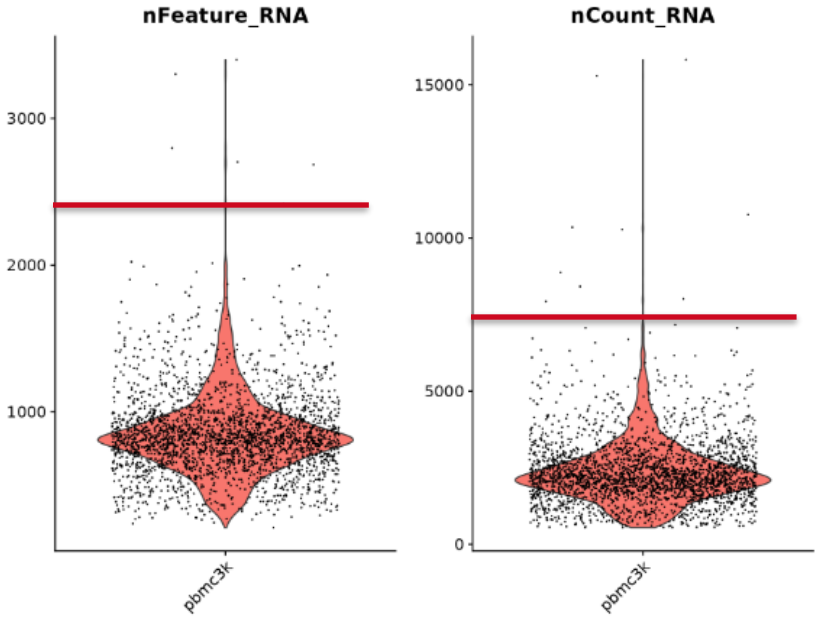


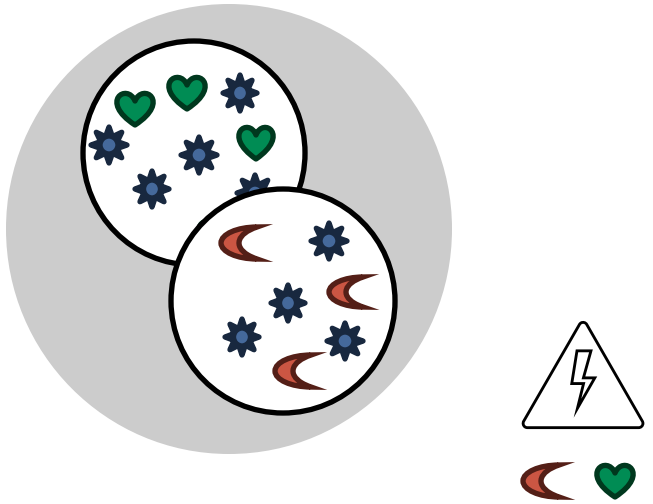
Events with many counts / many genes are suspicious
→ use upper threshold

Use prior knowledge to identify combinations of marker genes which are not thought to exist

4

Doublet detection with computational methods basic ideas





Events with many counts / many genes are suspicious → use upper threshold

Use prior knowledge to identify combinations of marker genes which are not thought to exist

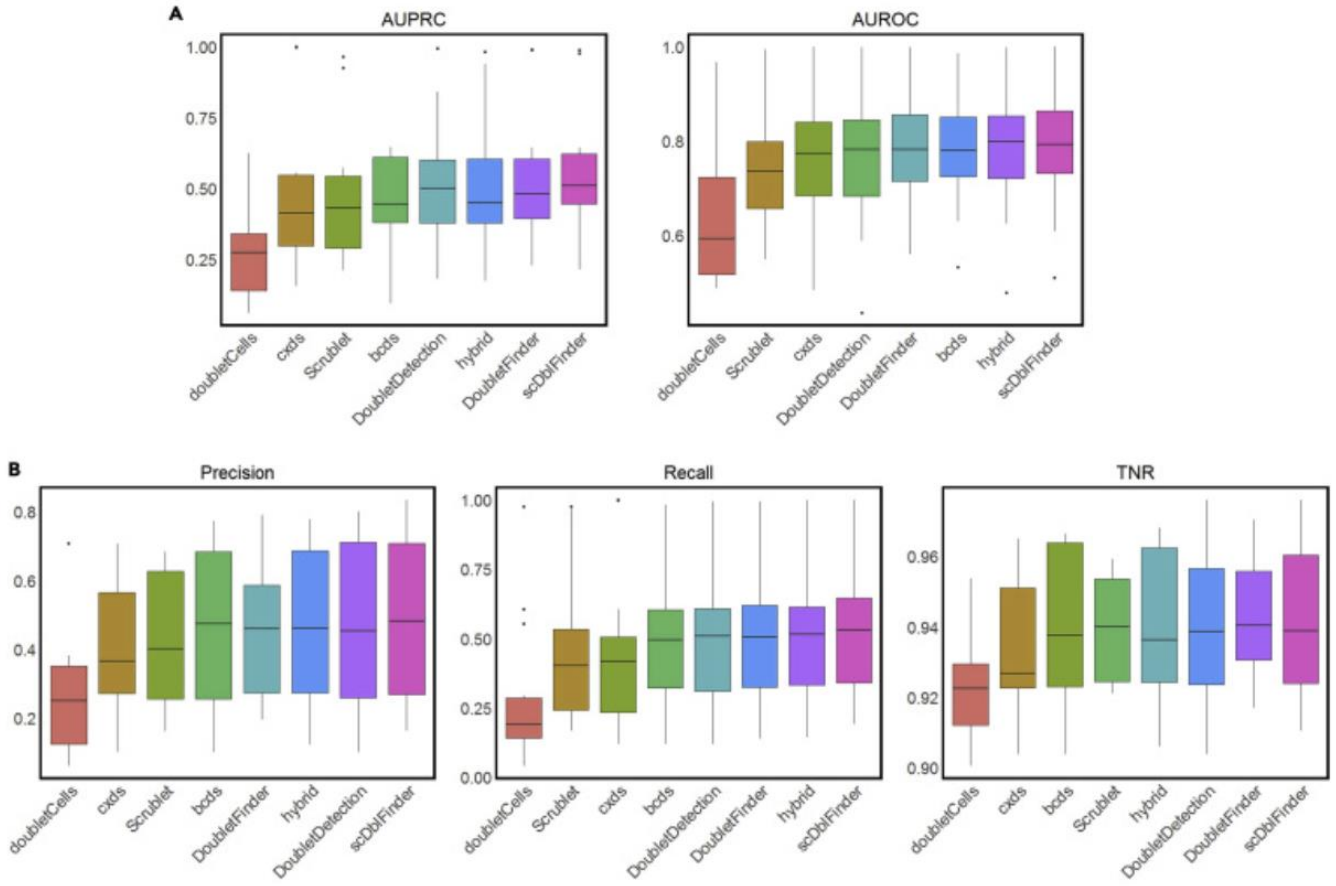
Some cell types are larger / have more mRNA

Right cell type resolution to consider? Novel discoveries?

Doublet detection with computational methods – many options exist

Data for benchmarking doublet detection algorithms:

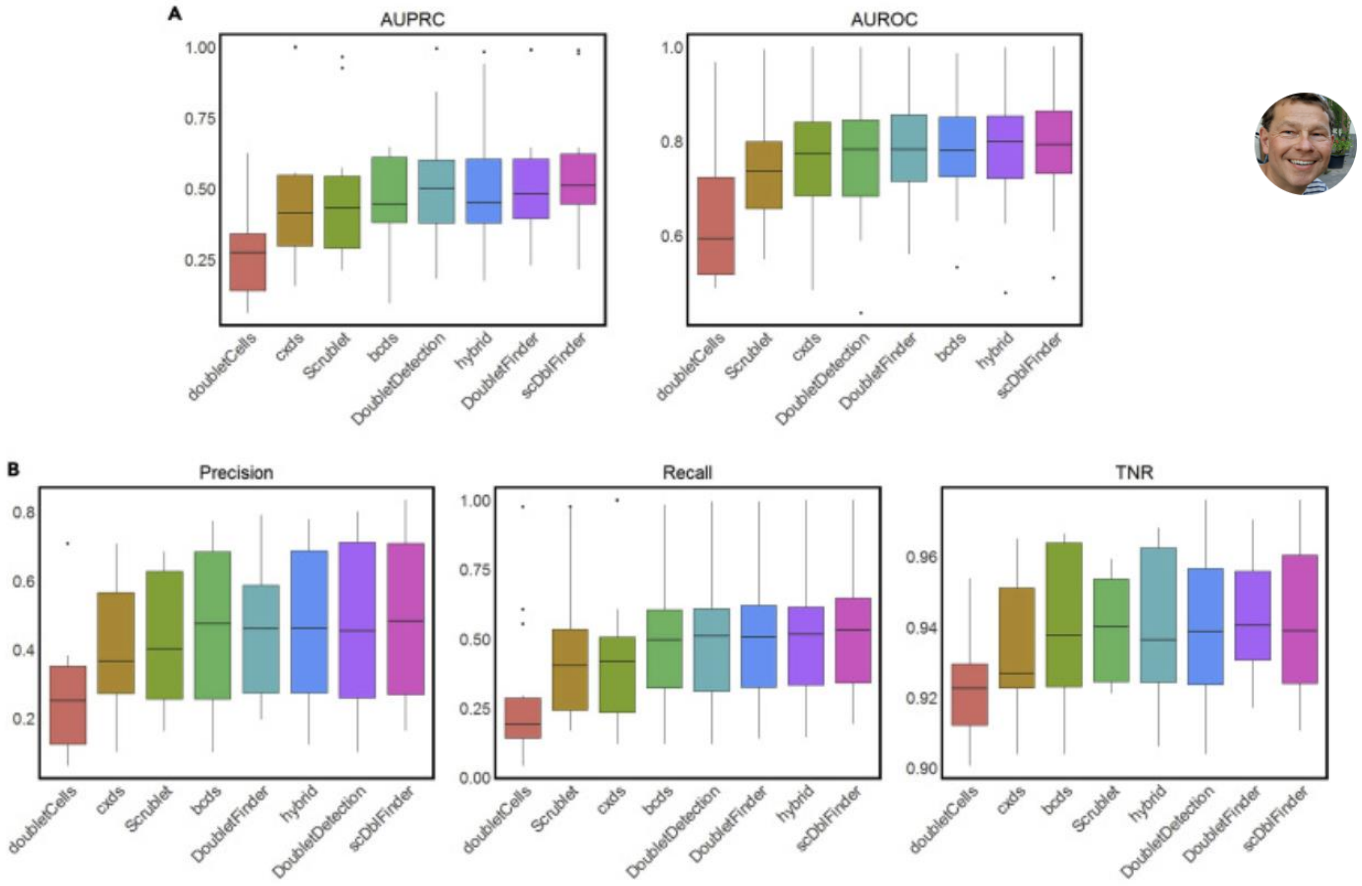
1. Simulated datasets
2. Experimental doublet datasets:
   a. Human/mouse mixture
   b. 2 genotype mixture
   c. cell hashing (multiplexing)



Nan Miles Xi, Jingyi Jessica Li (2021): Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis, STAR Protocols, Volume 2, Issue 3, https://doi.org/10.1016/j.xpro.2021.100699.

4

Doublet detection with computational methods – many options exist



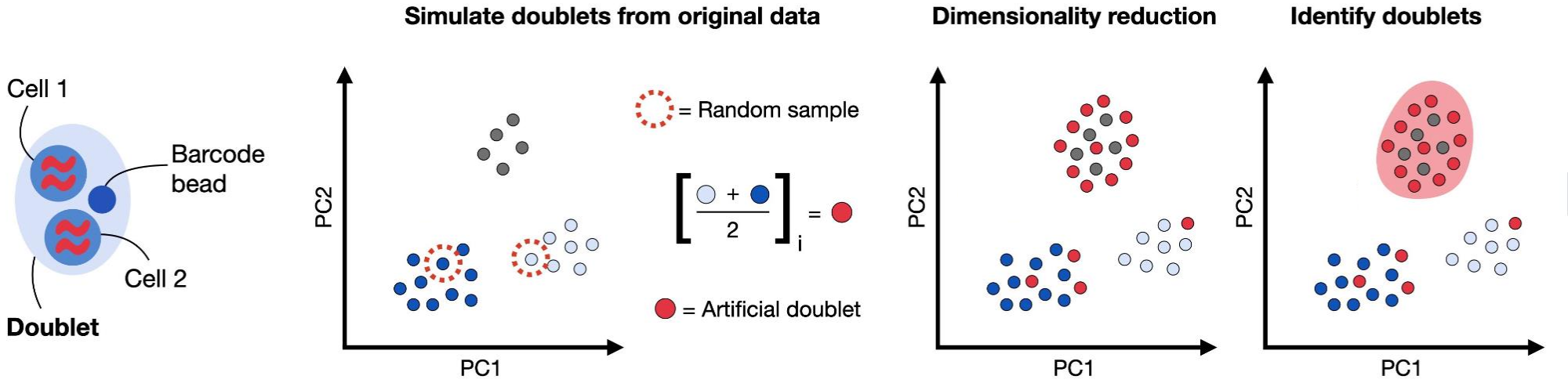
Wolfgang Huber 🇺🇦 @wkhuber.bsky.social · 5mo

Bioinformatics wisdom of the day:
When there are too many methods for the same task, you know that none of them is very good.

3 20

Nan Miles Xi, Jingyi Jessica Li (2021): Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis, STAR Protocols, Volume 2, Issue 3, https://doi.org/10.1016/j.xpro.2021.100699.

Doublet detection with computational methods – example: scDblFinder

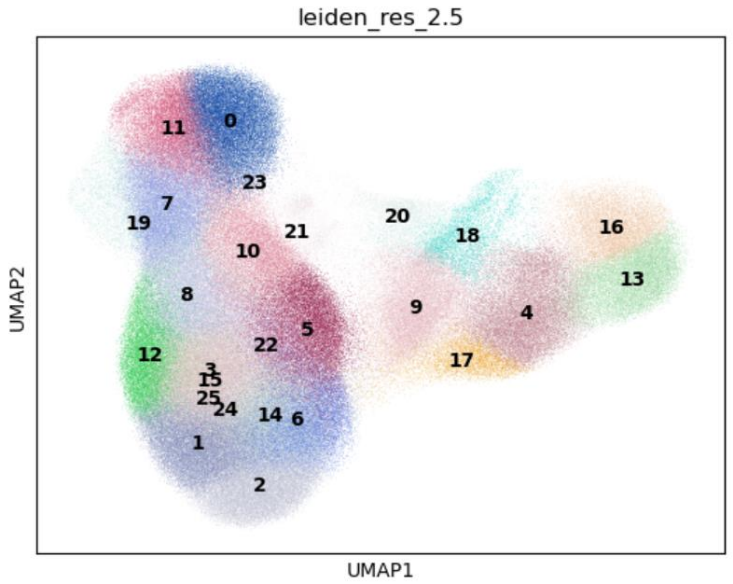


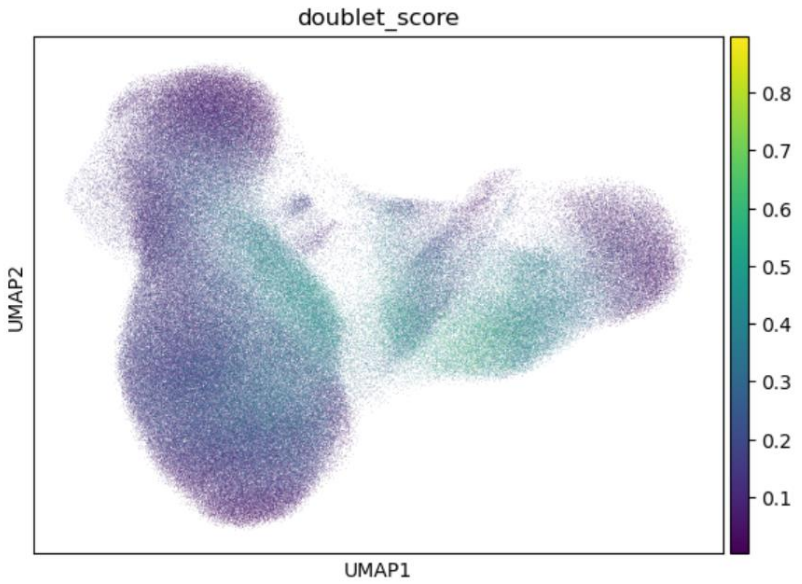
Nearest neighbor classification – if most nearest neighbors are simulated doublets, the cell is probably a doublet, too

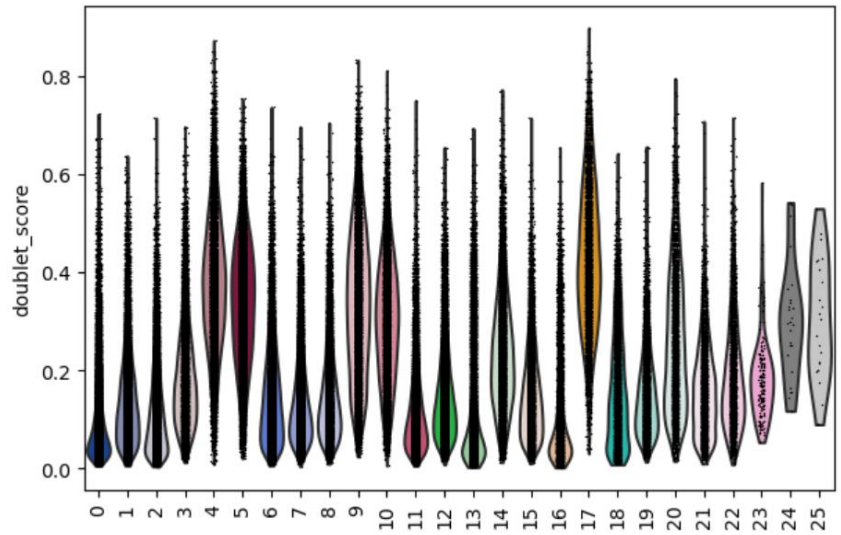
https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html#doublet-detection

4

Strategy: Cluster at high resolution, compare doublet scores at the cluster level rather than the single cell level



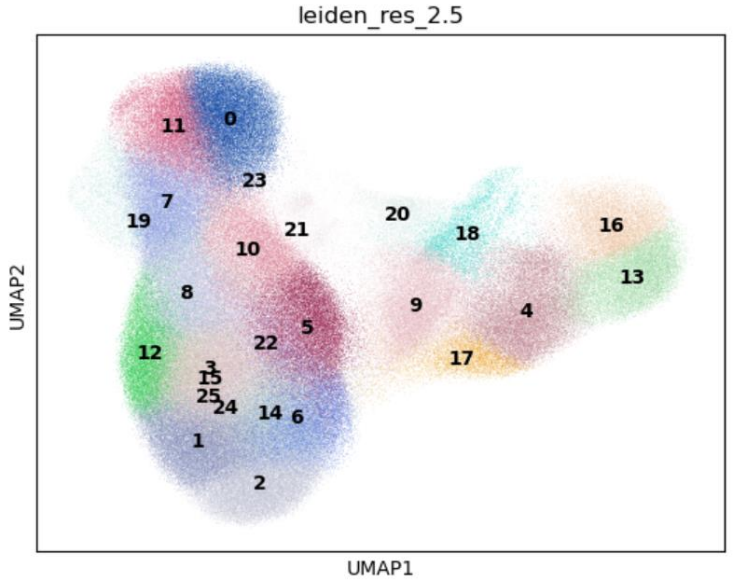




Here: doublet scores from scrublet via scanpy

4

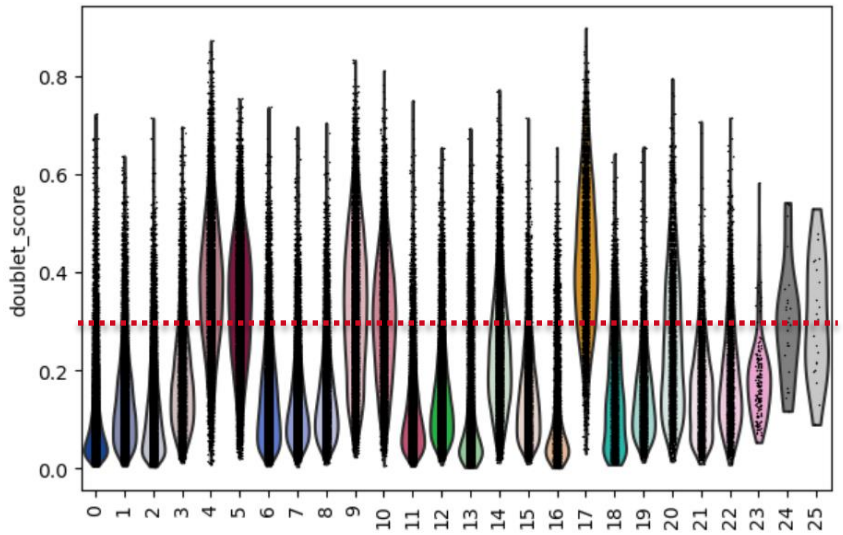
Strategy: Cluster at high resolution, compare doublet scores at the cluster level rather than the single cell level



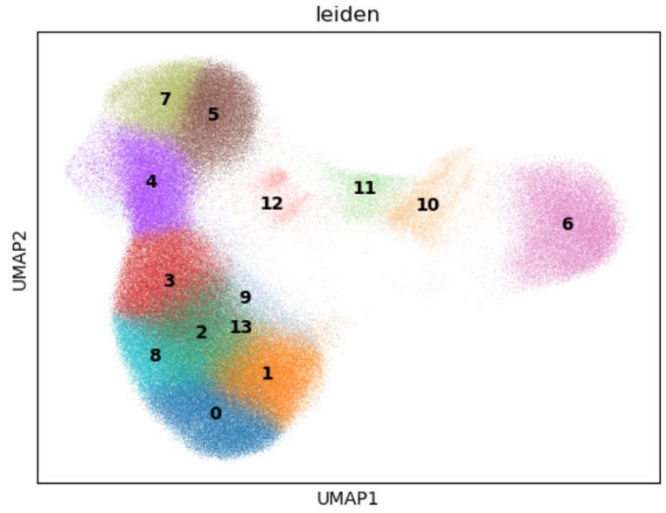
Removed clusters with mean doublet score > 0.3, meaning clusters 4, 5, 9, 10, 17, 24, 25.

Before doublet removal: 338,465 events
After doublet removal: 255,170 events

→ ~24% removed, agrees well with theoretically expected number of doublets

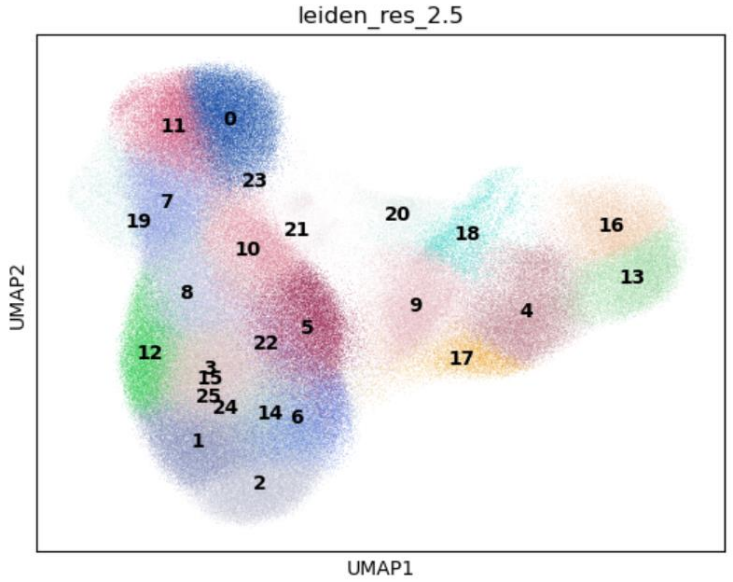


After removal of clusters with high doublet scores, recluster



4

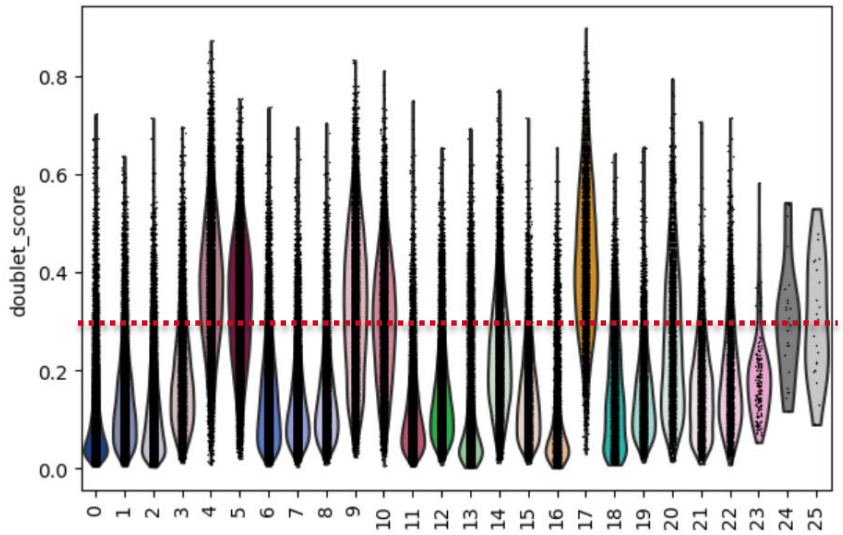
Strategy: Cluster at high resolution, compare doublet scores at the cluster level rather than the single cell level



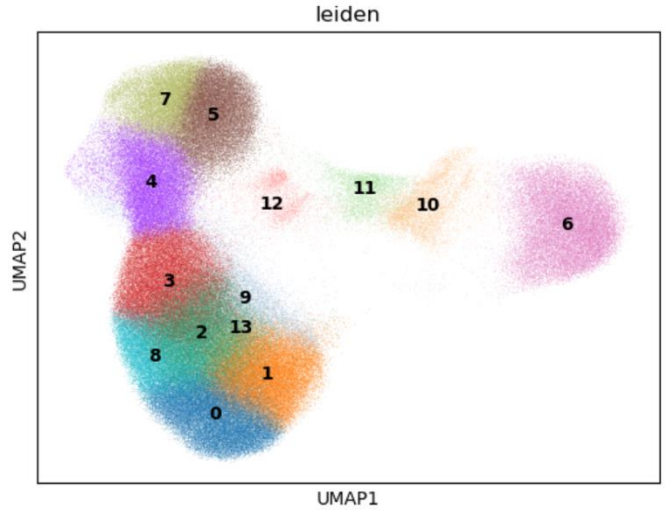
Removed clusters with mean doublet score > 0.3, meaning clusters 4, 5, 9, 10, 17, 24, 25.

Before doublet removal: 338,465 events
After doublet removal: 255,170 events

→ ~24% removed, agrees well with theoretically expected number of doublets



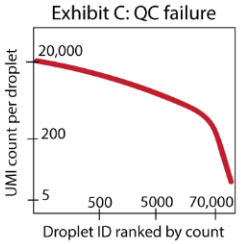
After removal of clusters with high doublet scores, recluster



Quality Control: A Suggested Workflow (but you are allowed to think for yourselves!)

Alignment (e.g. cellranger, STARsolo)

Inspect knee plot and library level QC metrics to identify failures



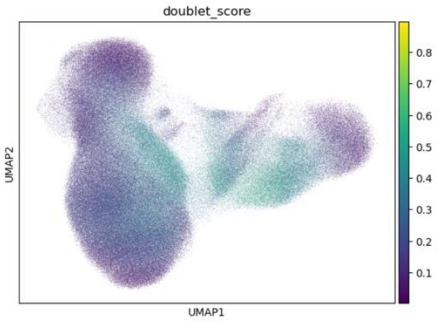
If required, run ambient RNA removal (e.g. cellbender), proceed with corrected count matrix

1st permissive filtering on
- n_counts
- n_genes
- % mt genes

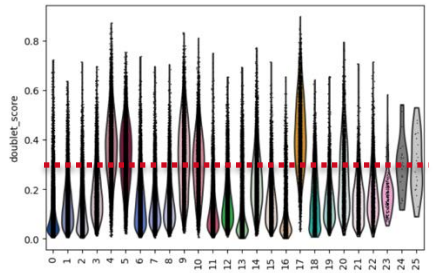
per sample

[If several libraries/datasets/runs/samples are being combined, combine here]

Calculate doublet scores (e.g. scrublet, scDblFinder)



Run a fine clustering (i.e. deliberately overcluster), inspect QC metrics at the cluster level, e.g. as violin plots (doublet score, n_counts, n_genes, % mt genes, marker gene scores if available)



Determine cut-offs and remove outlier clusters

Clean data ☺

Follow-up material

- In-depth discussion of knee plots including how they may be used to detect experimental failures and problematic samples https://www.10xgenomics.com/support/software/cell-ranger/latest/advanced/cr-barcode-rank-plot