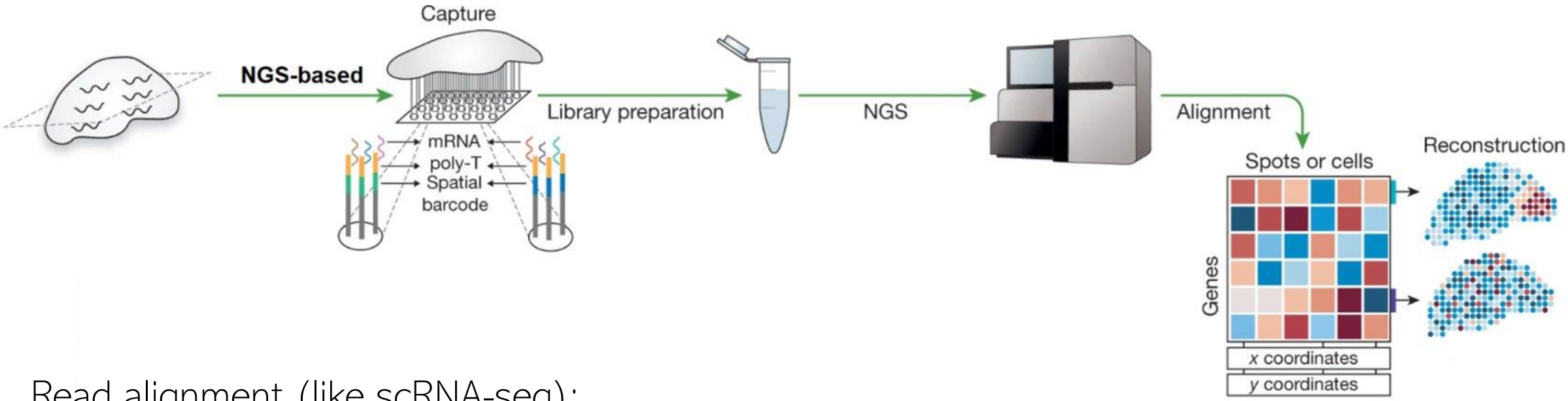
Preprocessing sequencing-based spatially-resolved transcriptomic (SRT) data

Ahmed Mahfouz

Human Genetics, Leiden University Medical Center
Pattern Recognition and Bioinformatics, TU Delft

Sequencing-based SRT

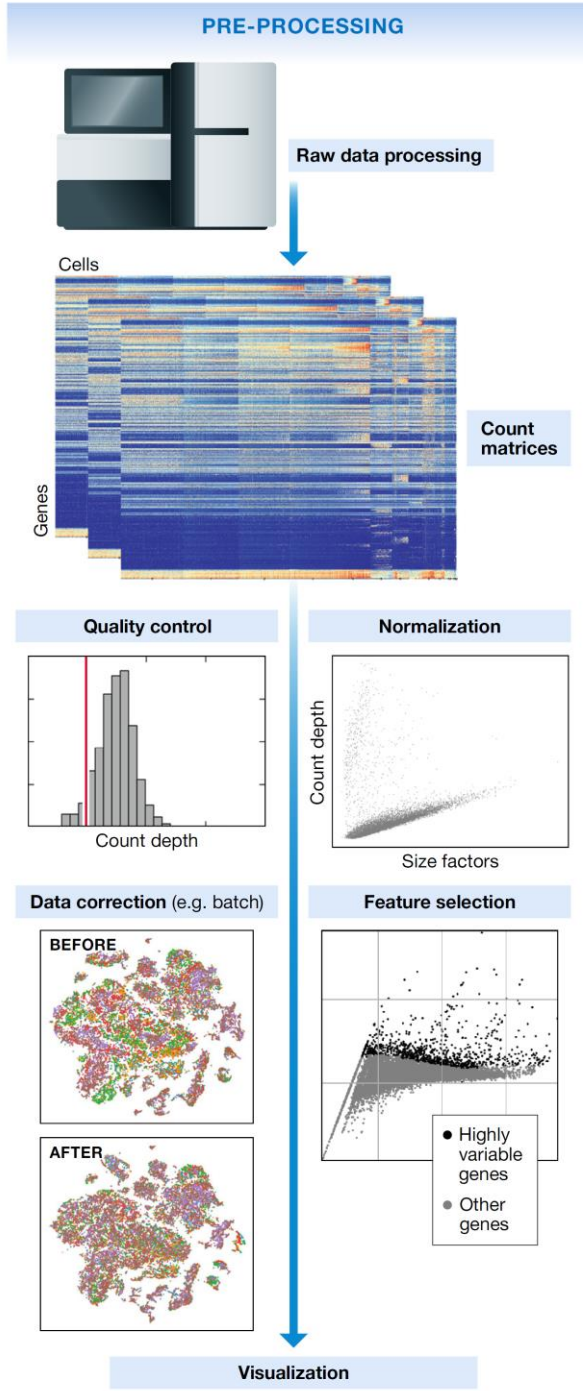


Read alignment (like scRNA-seq):

- SpaceRanger from 10x Genomic
- Stereo-seq Analysis Workflow SAW from STOmics)
- …

Preprocessing SRT data

- Like scRNA-seq data preprocessing:
  - Reads to count matrix
  - Quality control (QC)
  - Normalization
  - Batch correction
  - Feature selection



Luecken and Theis (2019)

Factors that affect data quality in scRNA-seq

Cell dissociation

Cell capture

Cell lysis

Reverse transcription

Preamplification

Library preparation and sequencing

Factors that affect data quality in seq-based SRT

Cell dissociation

Cell capture

Cell lysis/permeabilization

Reverse transcription

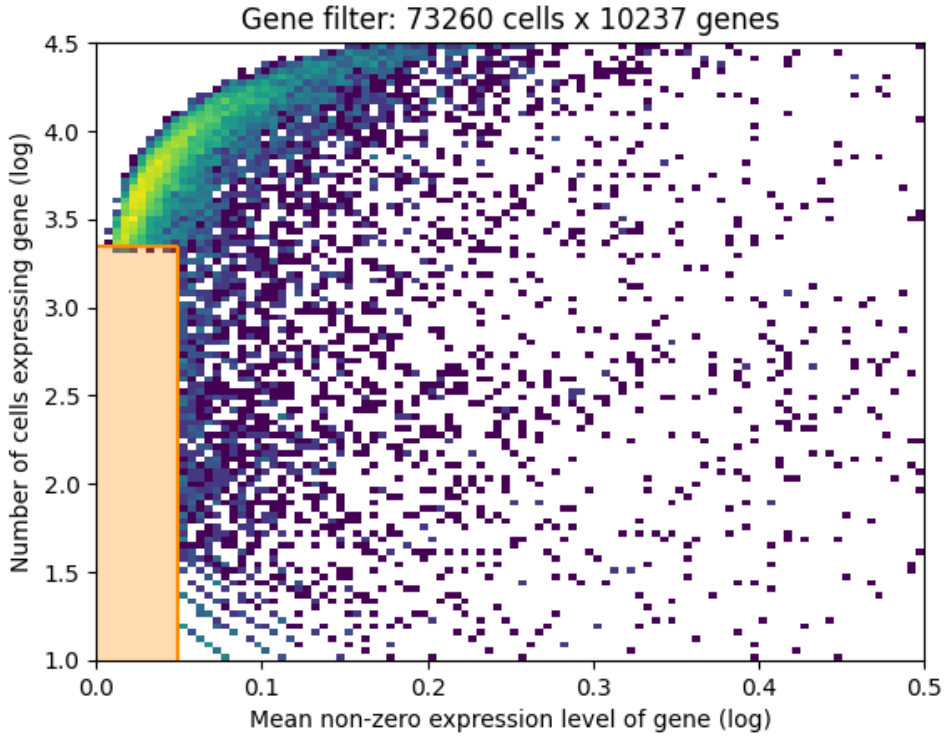
Preamplification

Library preparation and sequencing

Filtering genes

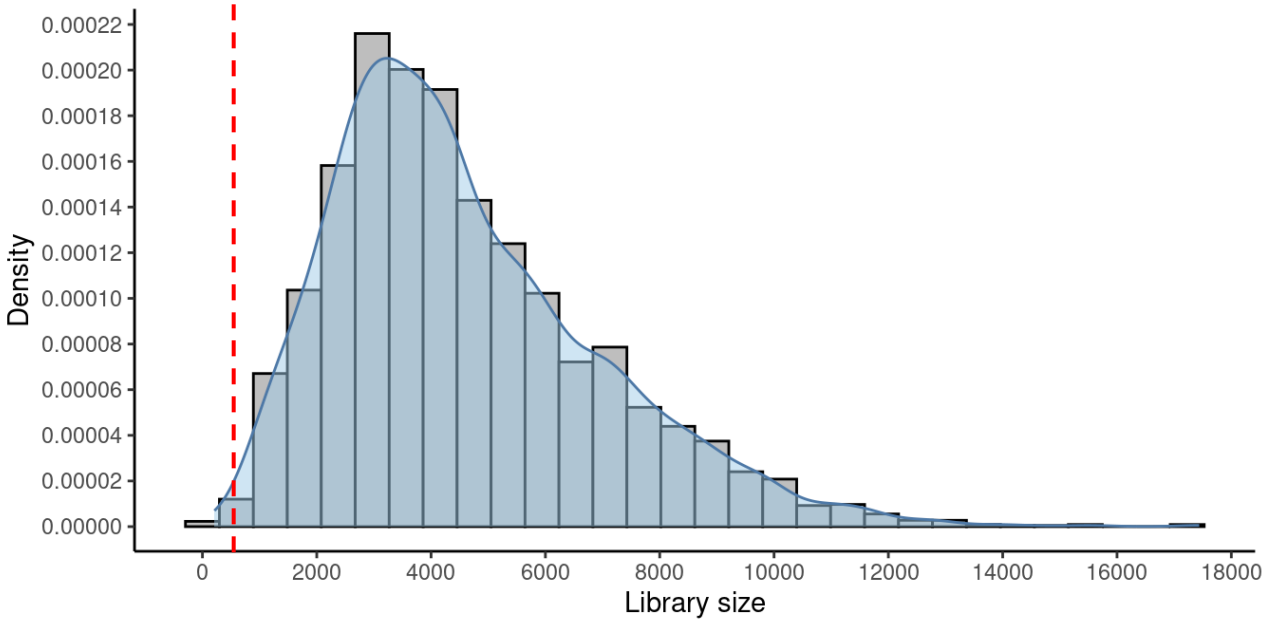
- Remove mitochondrial genes (gene names start with prefix mt- or MT-):
  - They compose 15-40% of mRNA in each location
  - Mostly representing technical artifacts

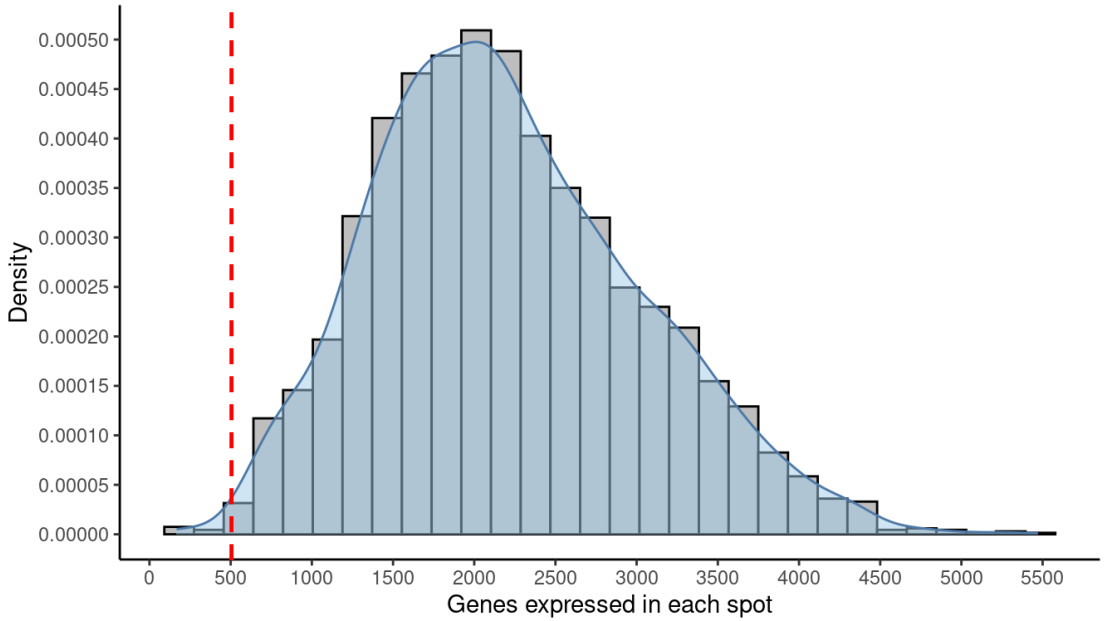
- Filter based on cell count, cell percentage, non-zero mean (aim at 8-16 k genes)



https://cell2location.readthedocs.io/

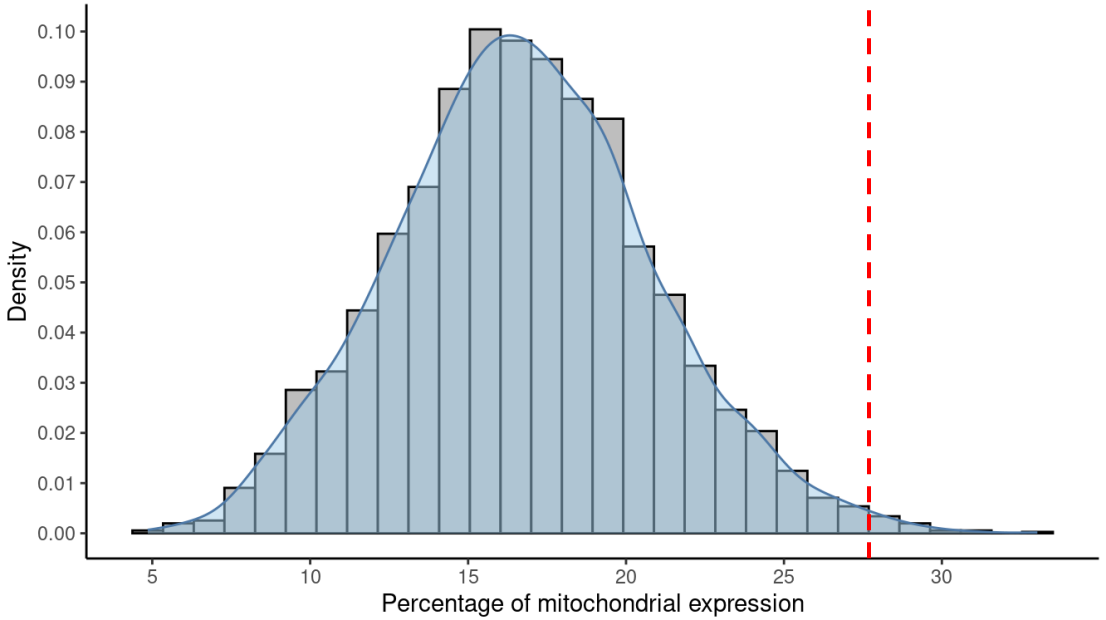
Filtering spots (counts/genes)





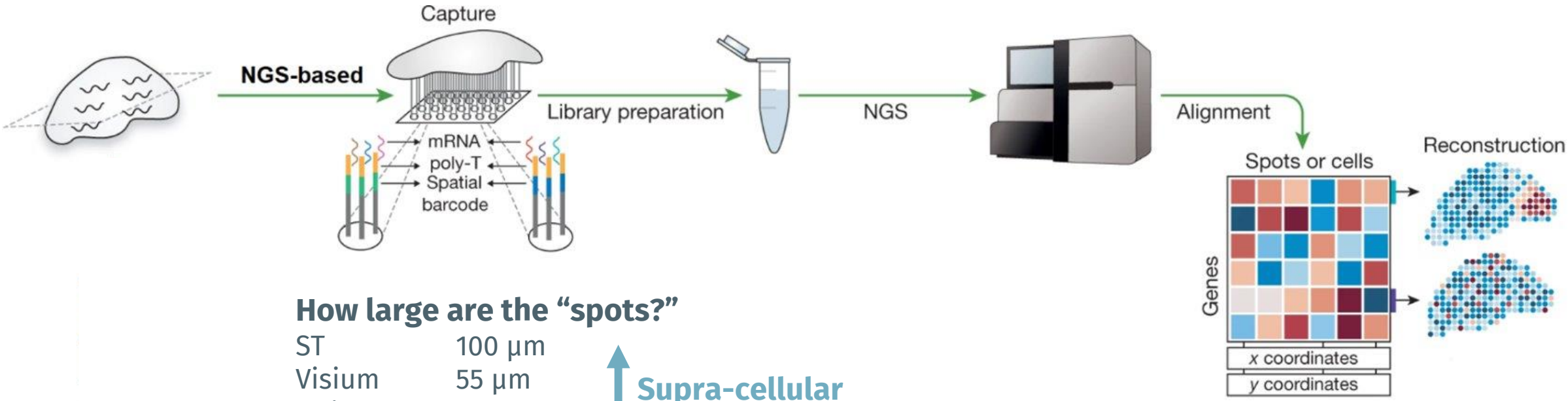
Poor mRNA capture rates due to cell damage and missing mRNAs, or low reaction efficiency

Filtering spots (mitochondrial %)



Partial cell lysis leading to leakage and missing cytoplasmic mRNAs, with the resulting reads therefore concentrated on the remaining mitochondrial mRNAs

Resolution variation in sequencing-based SRT



How large are the "spots?"

ST	100 µm
Visium	55 µm
DBiT-seq	10-50 µm
Slide-Seq	10 µm
HDST	2 µm
PIXEL-seq	1.22 µm
Seq-Scope	0.5-0.8µm
Stereo-seq	0.22 µm

Supra-cellular

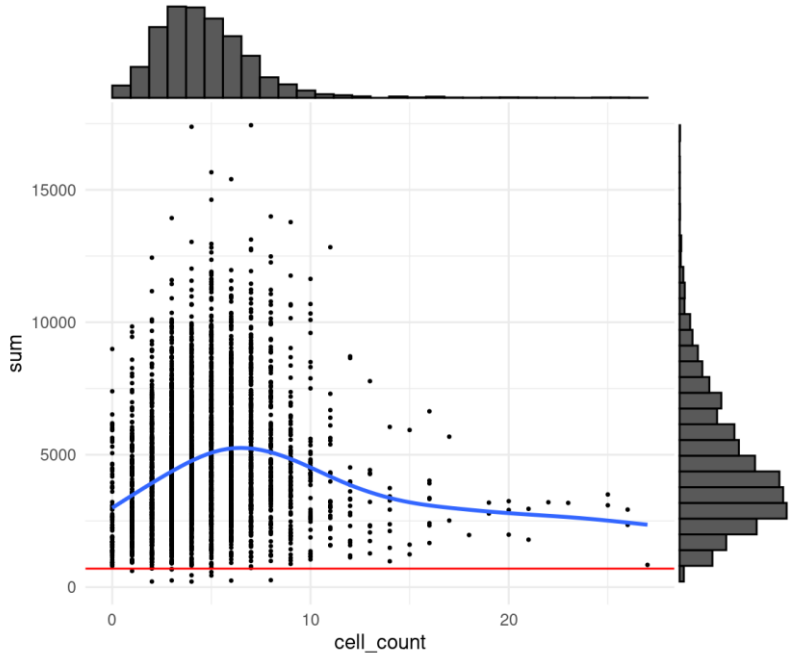
Cellular

Sub-cellular

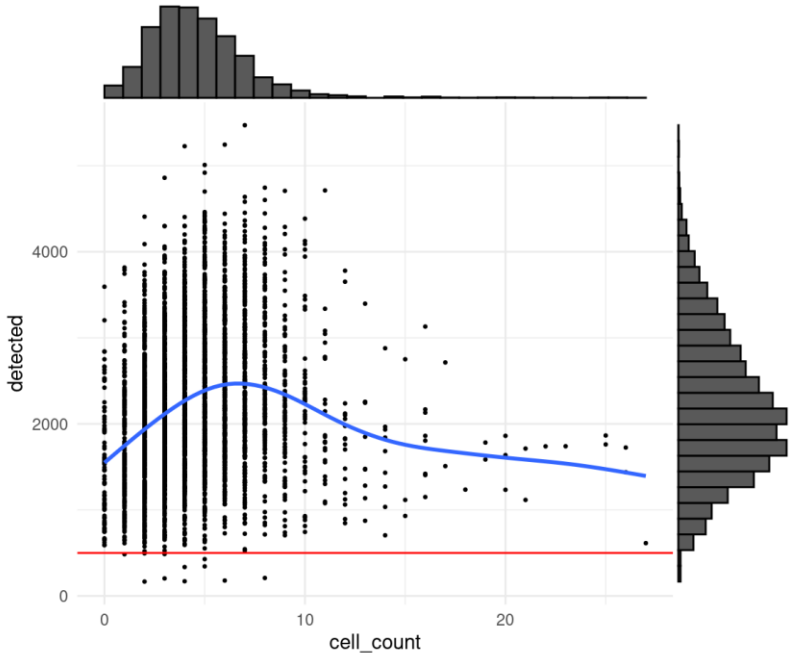
Spots may contain zero, one, or multiple cells

We can plot the library sizes against the number of cells per spot (to make sure we don't remove any spots that may have biological meaning)

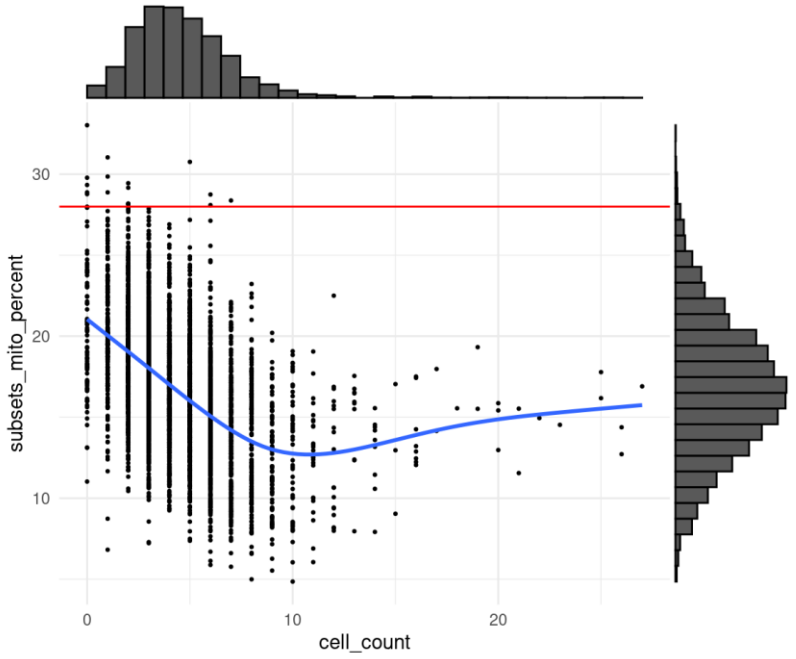
Library size



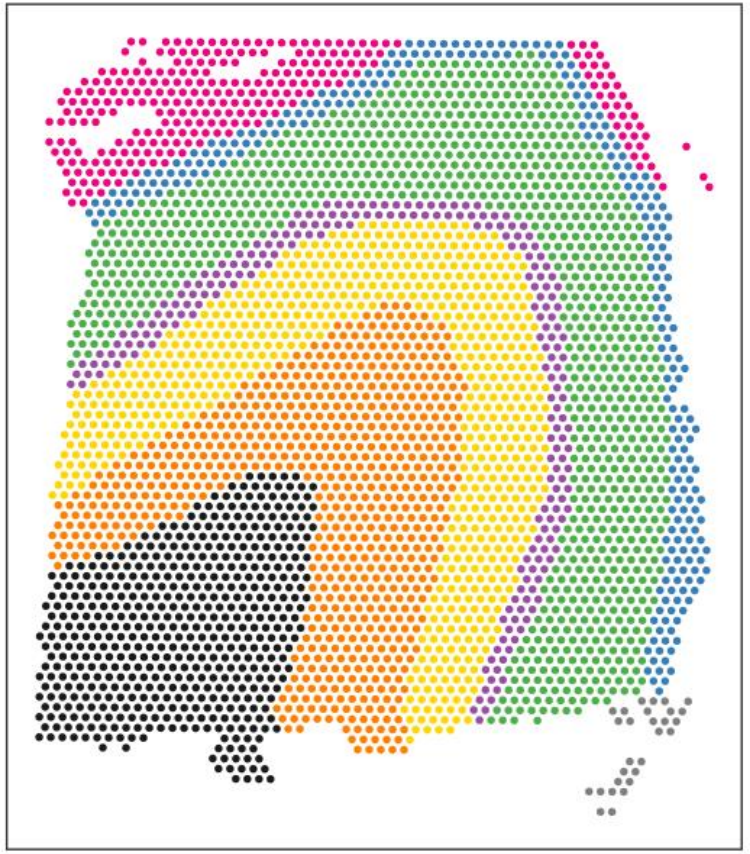
Total Genes

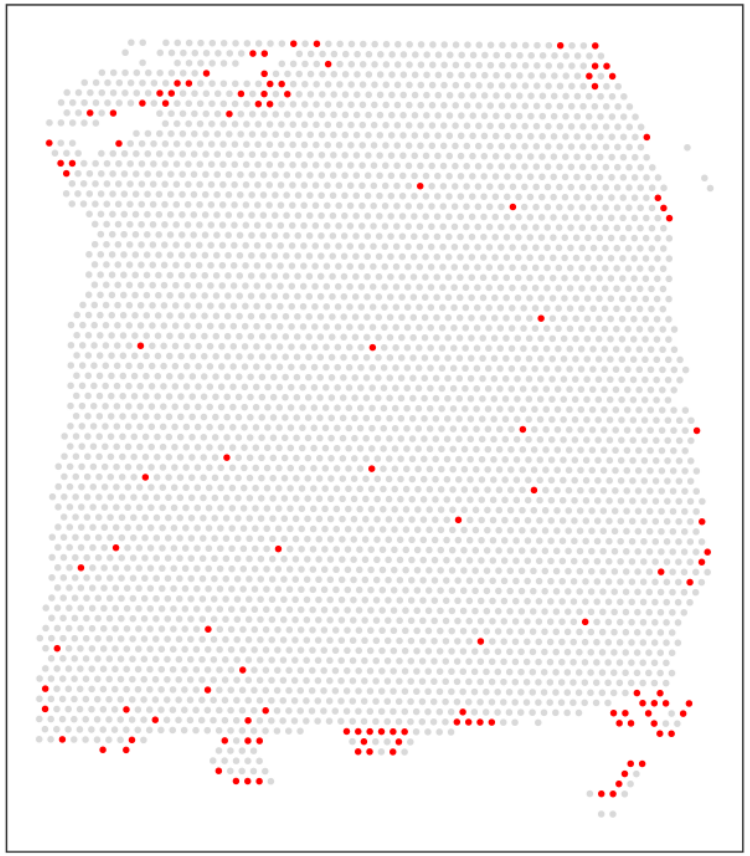


Mitochondrial %

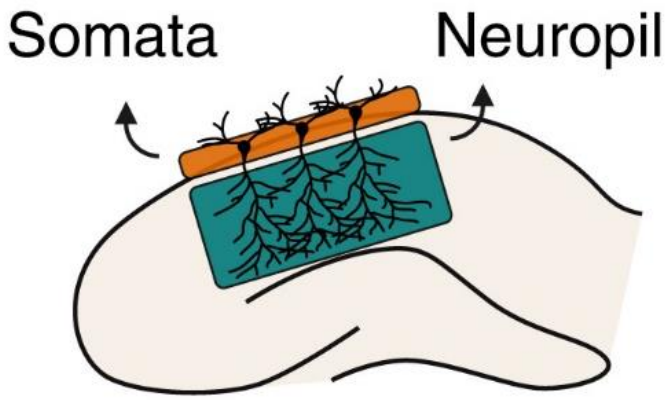


Check spatial patterns





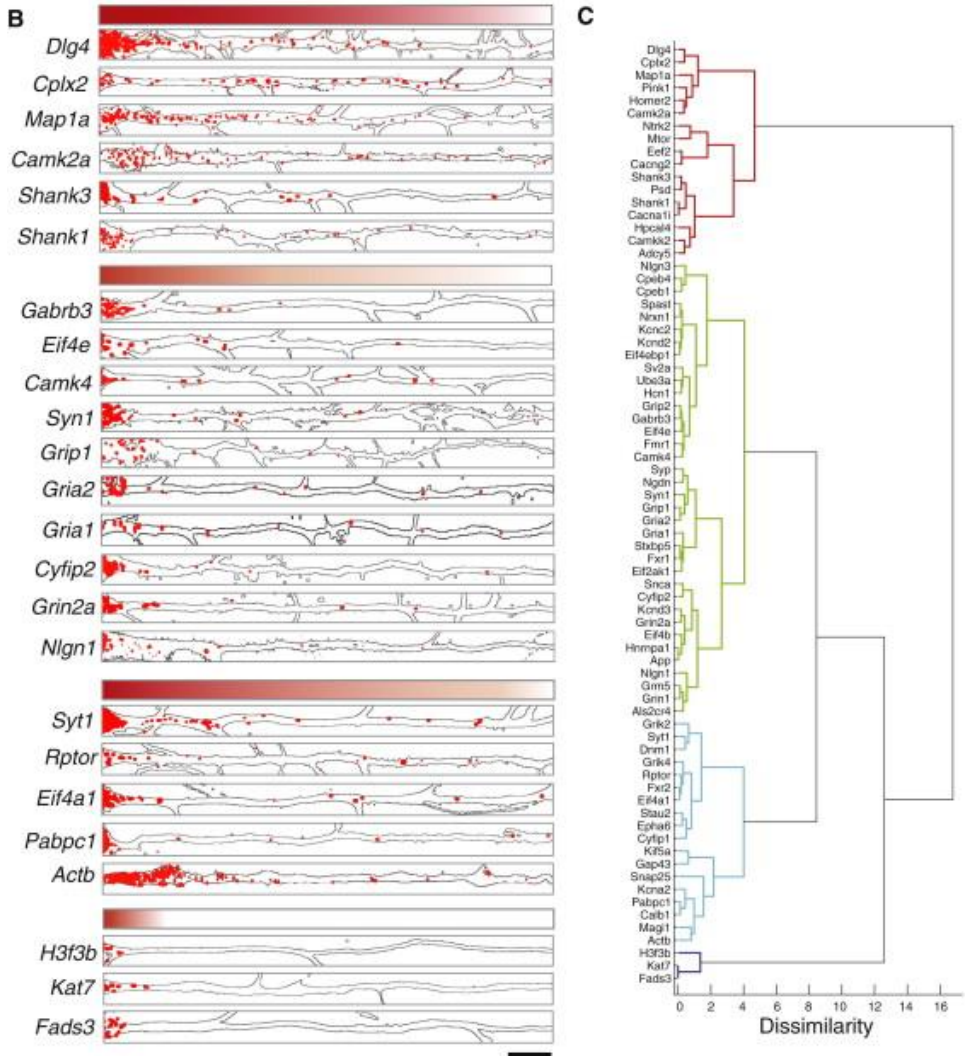
mRNA localization within cells



Neuropil comprises dendrites, axons, glia, and a sparse population of interneurons, but lacks principal neuron cell bodies

SUMMARY

In neurons, dendritic protein synthesis is required for many forms of long-term synaptic plasticity. The population of mRNAs that are localized to dendrites, however, remains sparsely identified. Here, we use deep sequencing to identify the mRNAs resident in the synaptic neuropil in the hippocampus. Analysis of a neuropil data set yielded a list of 8,379 transcripts of which 2,550 are localized in dendrites and/or axons. Using a fluorescent barcode strategy to label individual mRNAs, we show that their relative abundance in the neuropil varies over 3 orders of magnitude. High-resolution in situ hybridization validated the presence of mRNAs in both cultured neurons and hippocampal slices. Among the many mRNAs identified, we observed a large fraction of known synaptic proteins including signaling molecules, scaffolds and receptors. These results reveal a previously unappreciated enormous potential for the local protein synthesis machinery to supply, maintain and modify the dendritic and synaptic proteome.

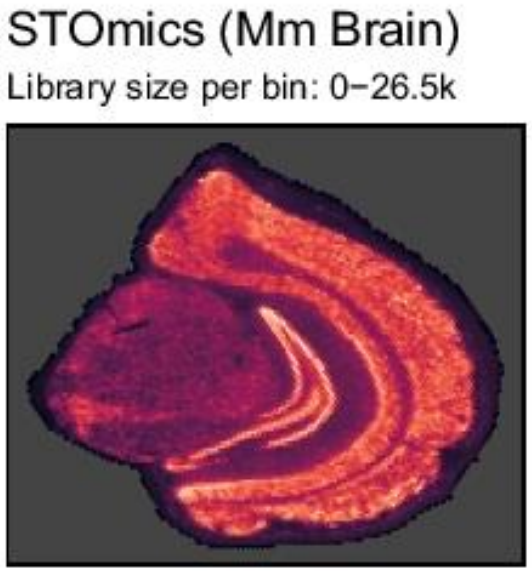


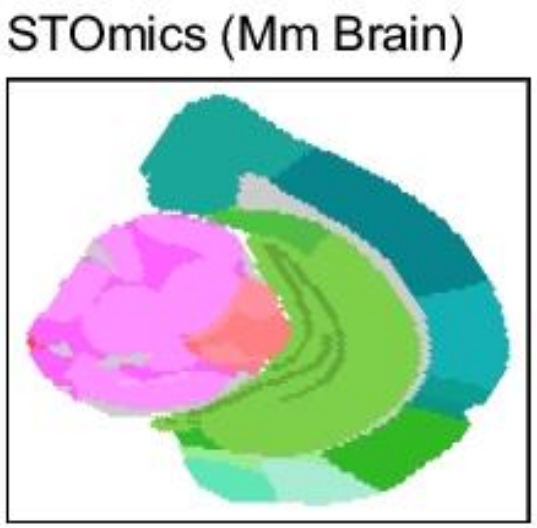
Cajigas*, Tushev*,… (2012)

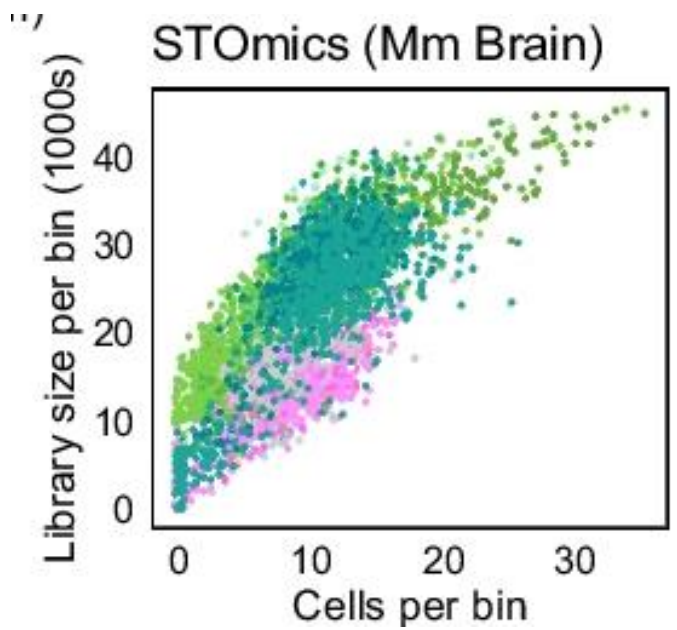
QC in sc/snRNA seq vs SRT

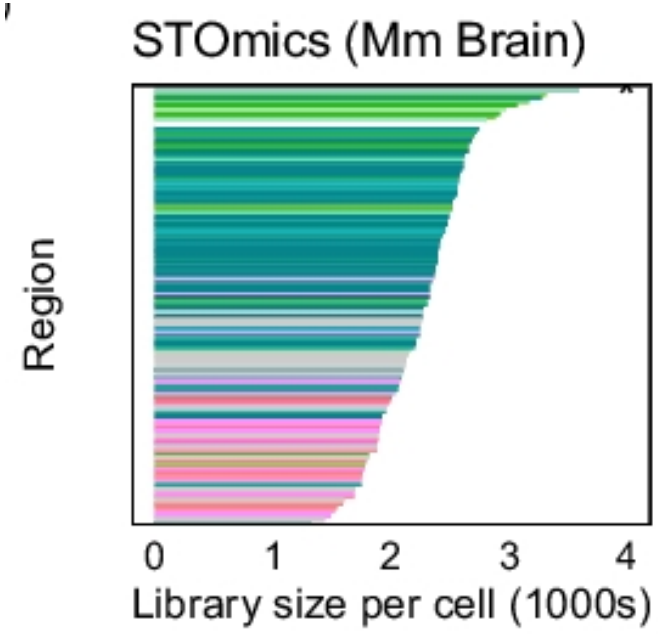
- sc/snRNA-seq: mRNA transcripts from a cell body or nucleus
- SRT: mRNA transcripts from a wide variety of biological domains
- SRT: tissue architecture can result in differences in reagent permeability
  - ➢ This would lead to sampling differences across the tissue and subsequently library size differences

Detection density and total detections/library sizes are associated with biology





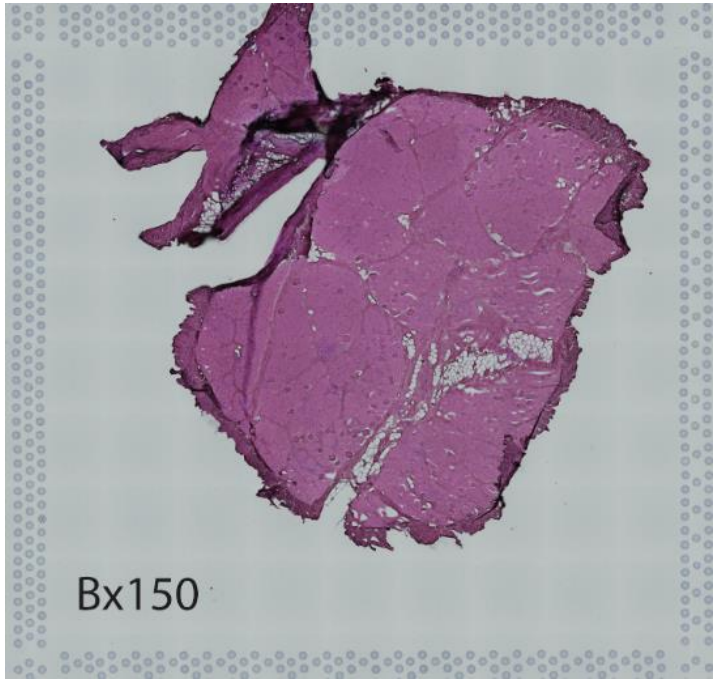




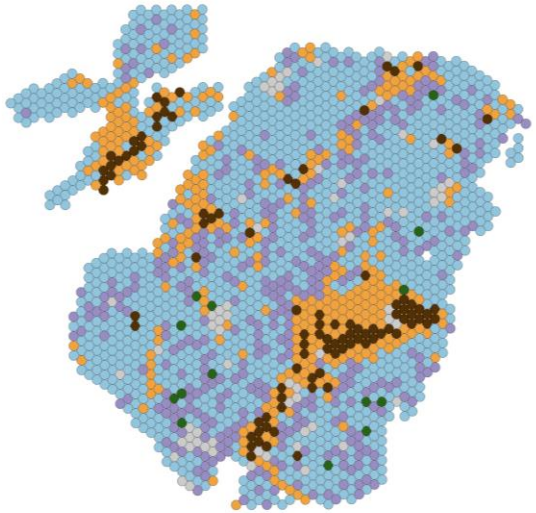
Bhuva,…, Davis (2024)

Be aware, variation in QC measure can be biological

H&E



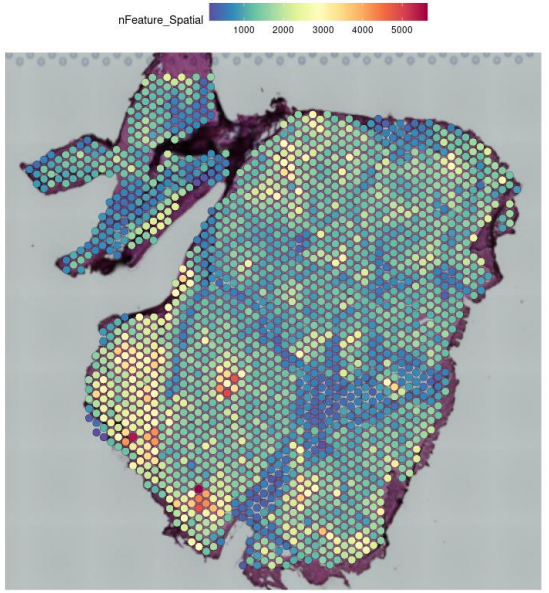
Annotation



- Type 1
- Type 2A
- Type 2X
- Connective tissue
- Fat
- Non muscle fibers

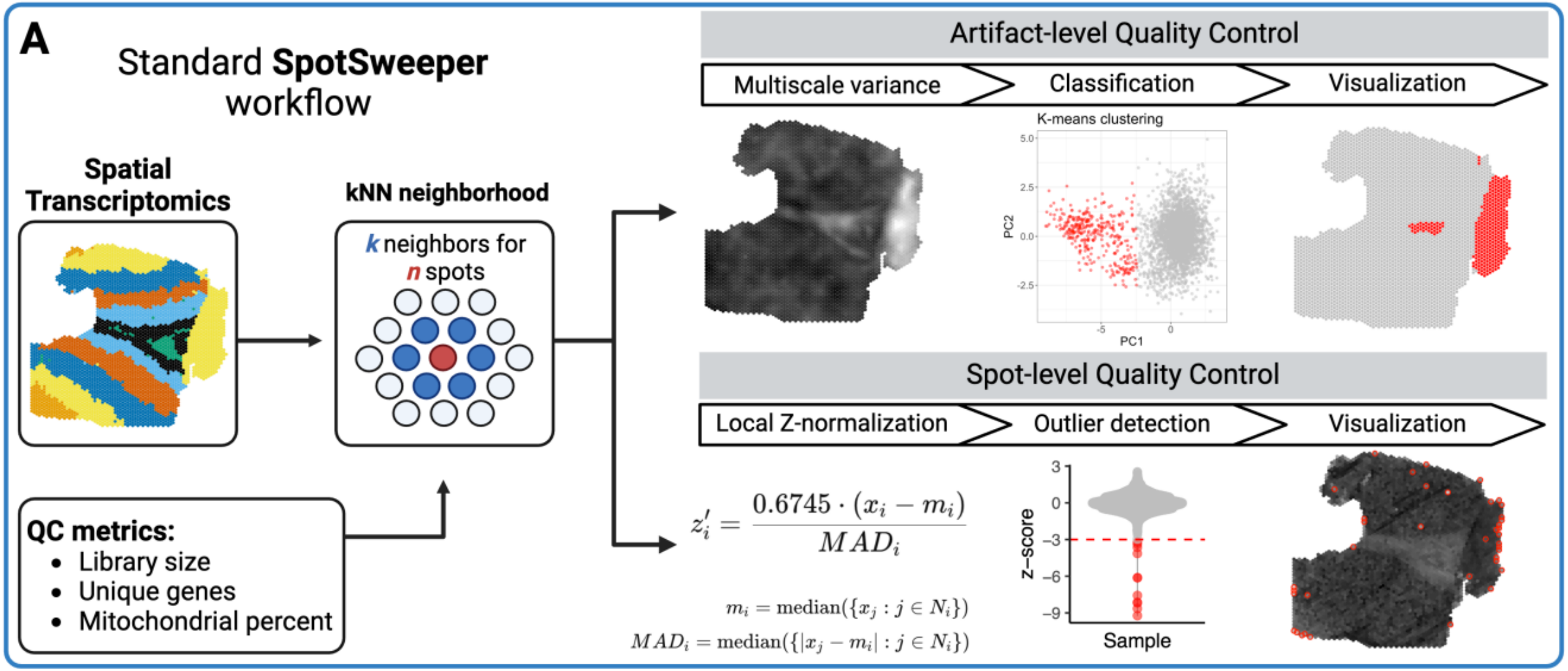
Library size

Total Genes



Heezen*, Mao*…, Mahfouz*, Spitali*(submitted)

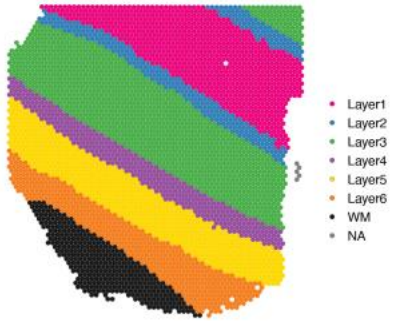
Spatially-aware QC (SpotSweeper)



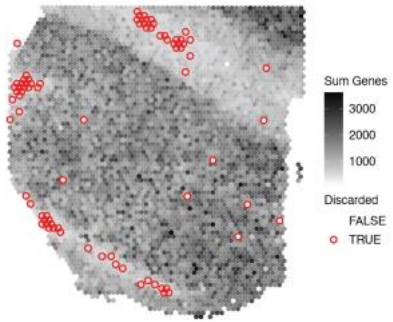
Totty, Hicks, Guo (2024)

SpotSweeper (spot-level artifacts)

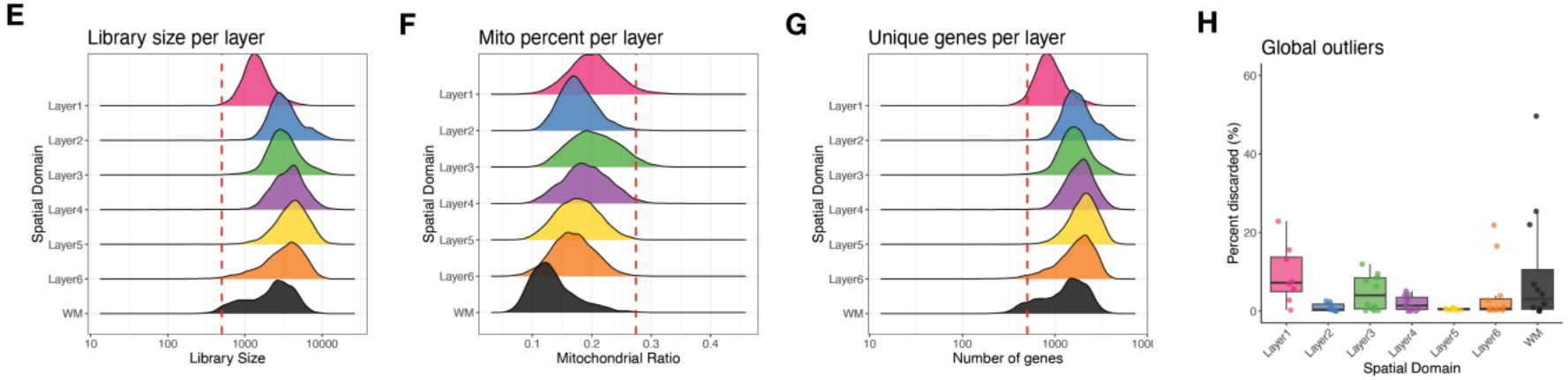
Annotation



Global

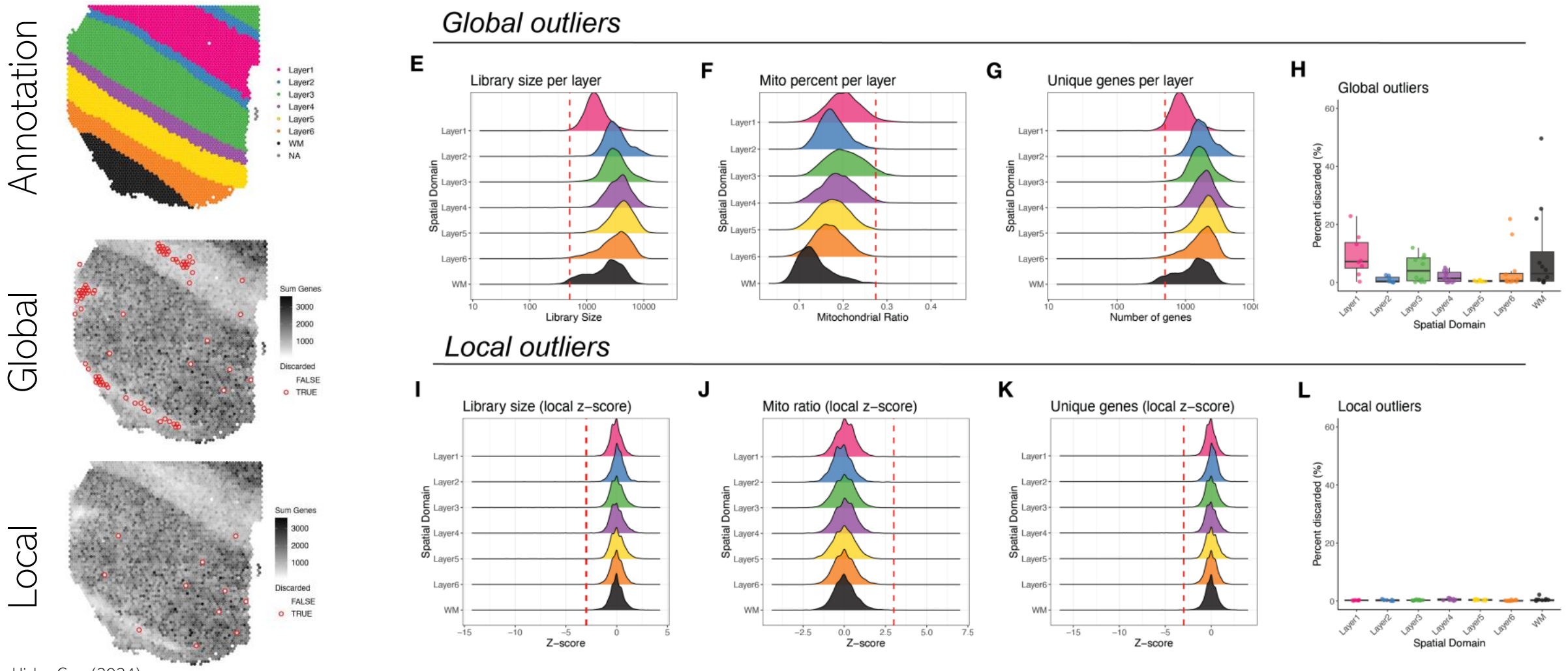


Global outliers



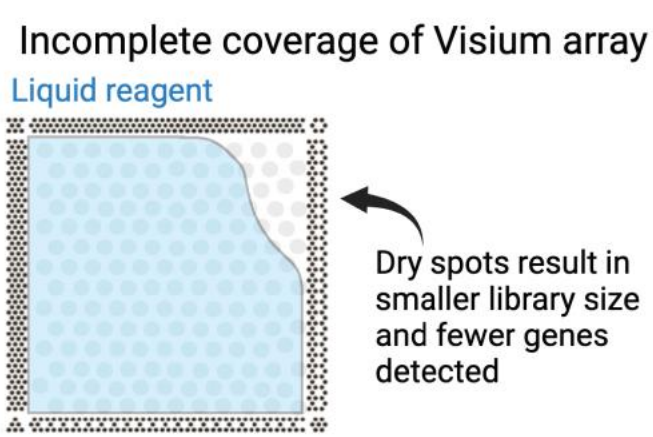
Totty, Hicks, Guo (2024)

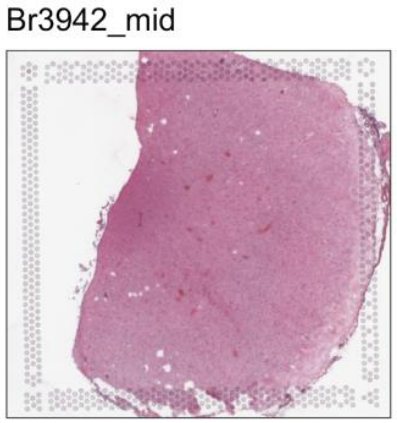
SpotSweeper (spot-level artifacts)

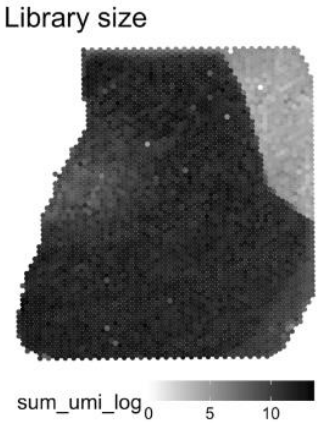


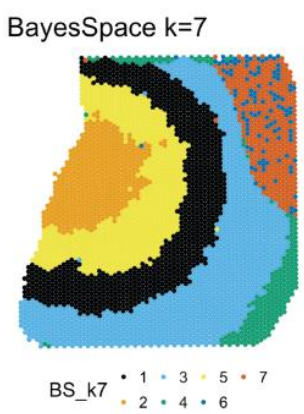
Totty, Hicks, Guo (2024)

SpotSweeper (region-level artifacts)



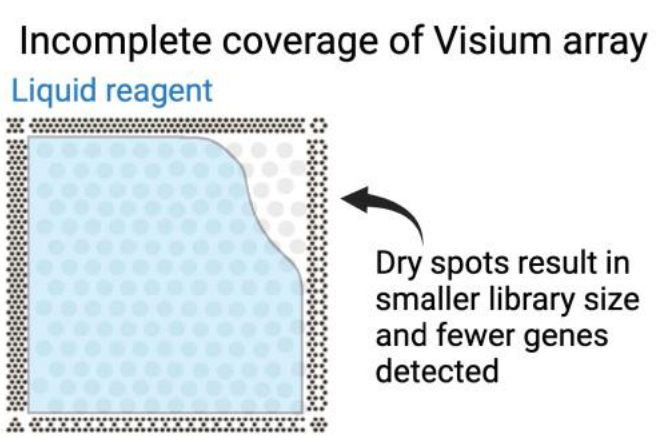


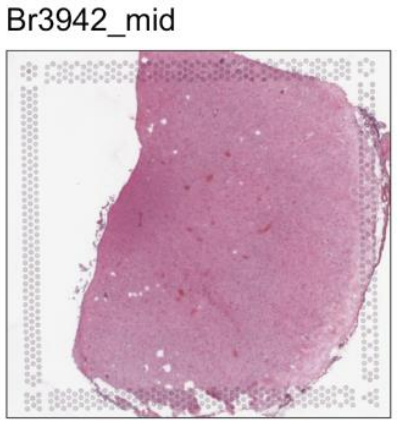


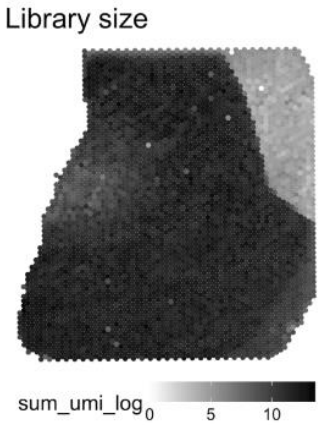


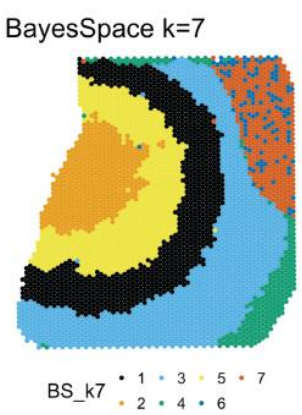
Totty, Hicks, Guo (2024)

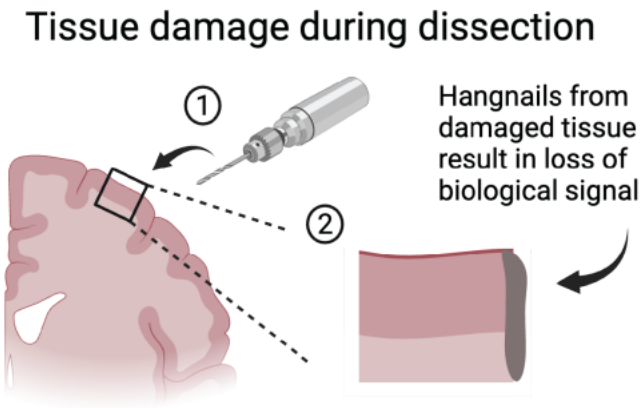
SpotSweeper (region-level artifacts)

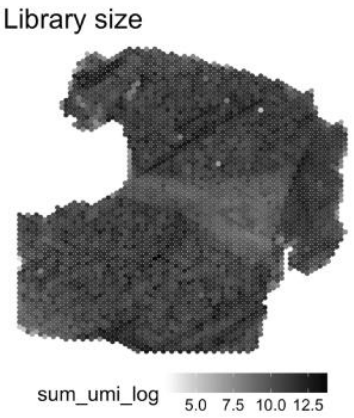


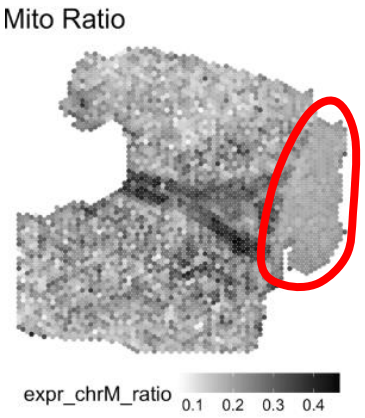








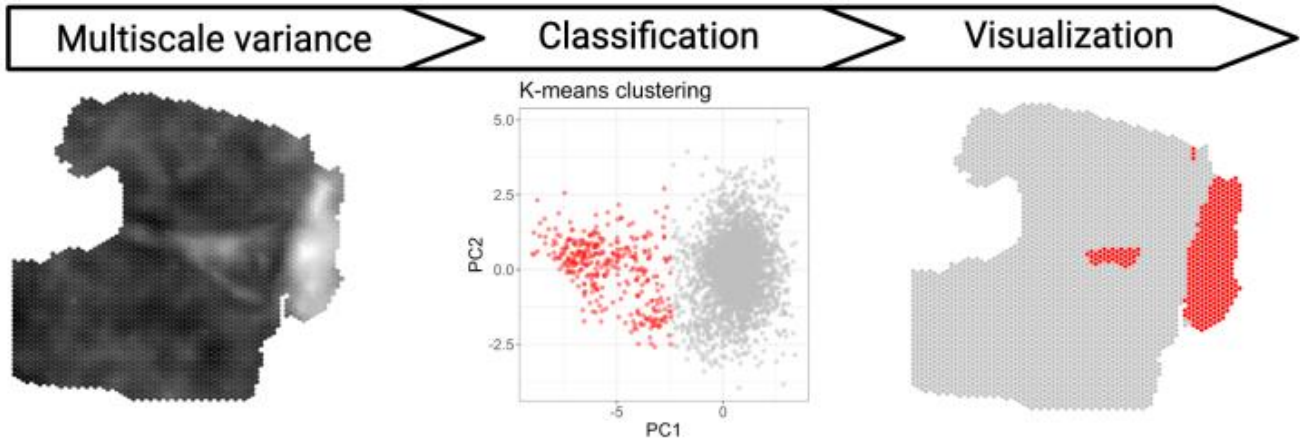




Totty, Hicks, Guo (2024)

SpotSweeper (region-level artifacts)

1. The k-NN for each spot are identified based on the spatial coordinates
2. For each neighborhood size (i.e., scale), local variance of the mitochondrial ratio is calculated and adjusted for a mean-variance relationship using linear regression
3. Mean-corrected local variance = the residuals of the linear regression
4. Perform PCA on the mean-corrected local variances of all neighborhood sizes
5. Apply k-means clustering (k=2) in the first two PCs to identify regional artifacts compared to high-quality tissue



Totty, Hicks, Guo (2024)

Normalization

Sources of variation

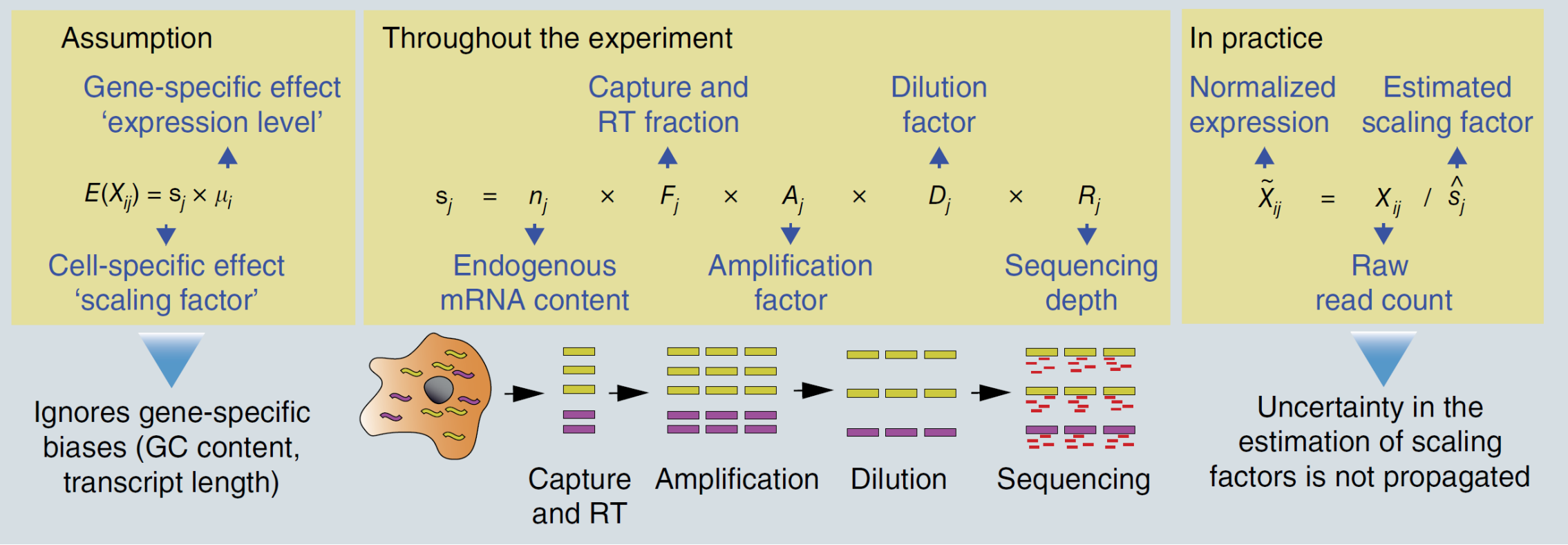
Biological

- Cell type/state
- Cell cycle
- Cell size
- Sex, Age, …
- …

Technical

- Cell quality
- Library prep efficiency
- Batch effects
- …

Normalization (1)



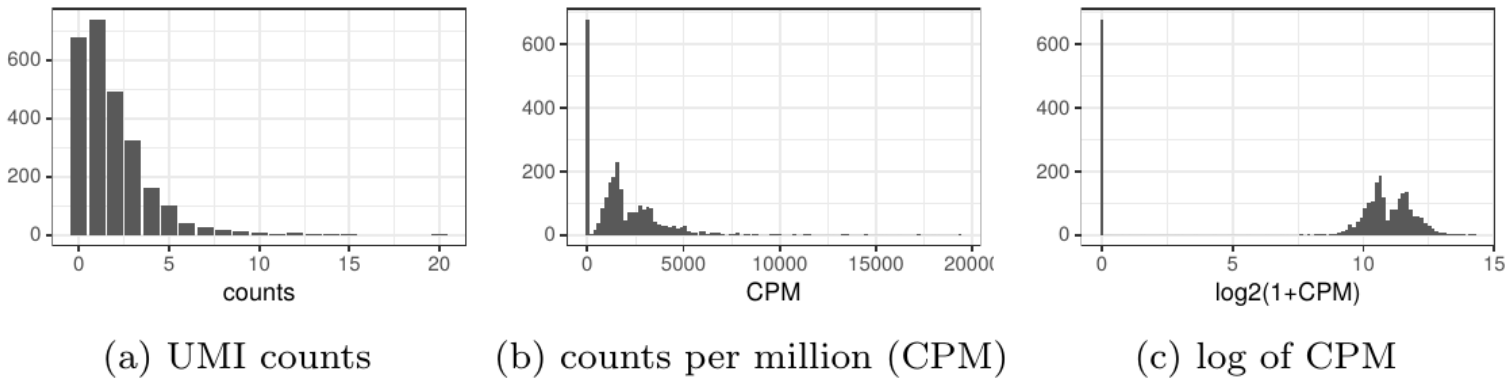
Vallejos,…, Dudoit, Marioni (2017)

Log-normalization

$$Y_{ij} = log_e((\frac{X_{ij}}{\sum_i X_{ij}} \times 10{,}000) + 1)$$

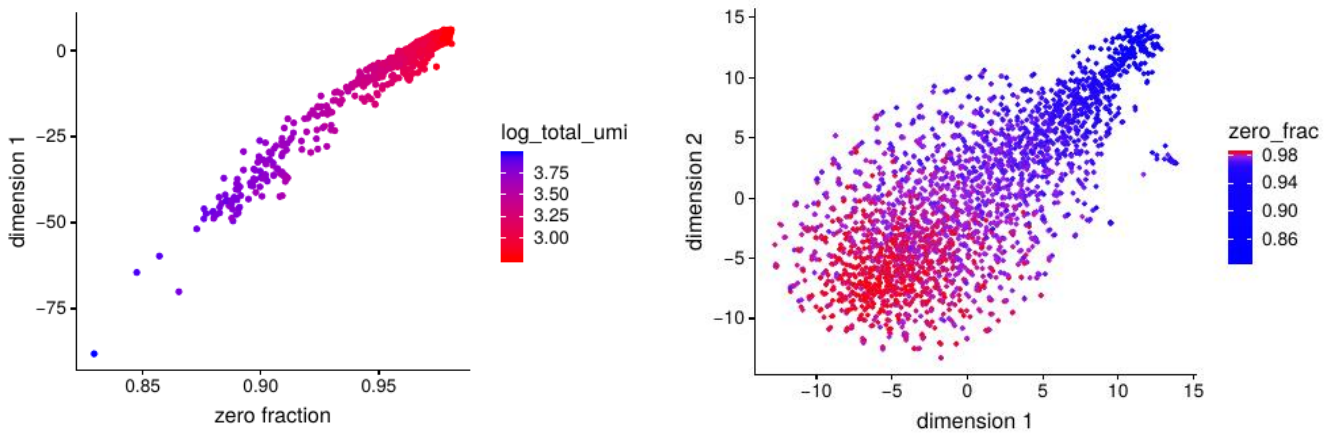
- Simplest and most commonly-used normalization strategy
- Divide all counts for each cell by a cell-specific scaling factor (i.e. size factor)
- Assumes that any cell-specific bias (e.g., in capture or amplification efficiency) affects all genes equally via scaling of the expected mean count for that cell
- A modified version of CPM normalization

Effect of dropouts on normalization



(a) UMI counts (b) counts per million (CPM) (c) log of CPM

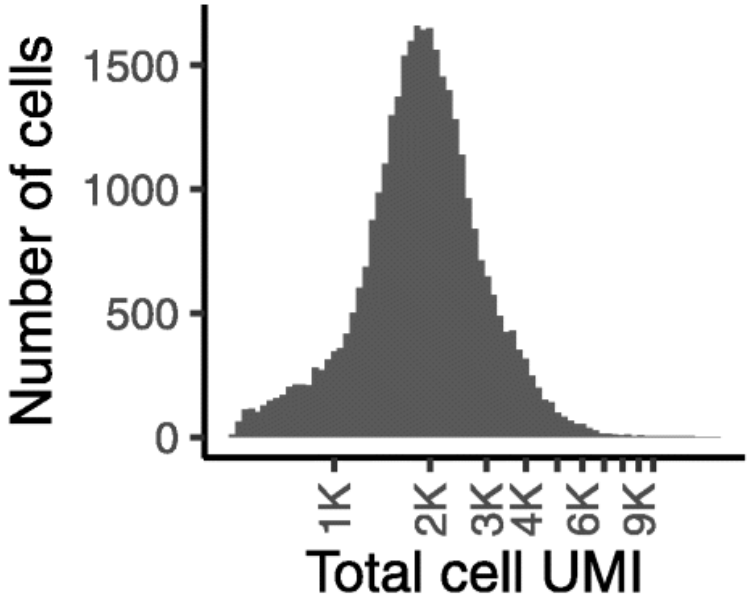
Fraction of zeros become main source of variability

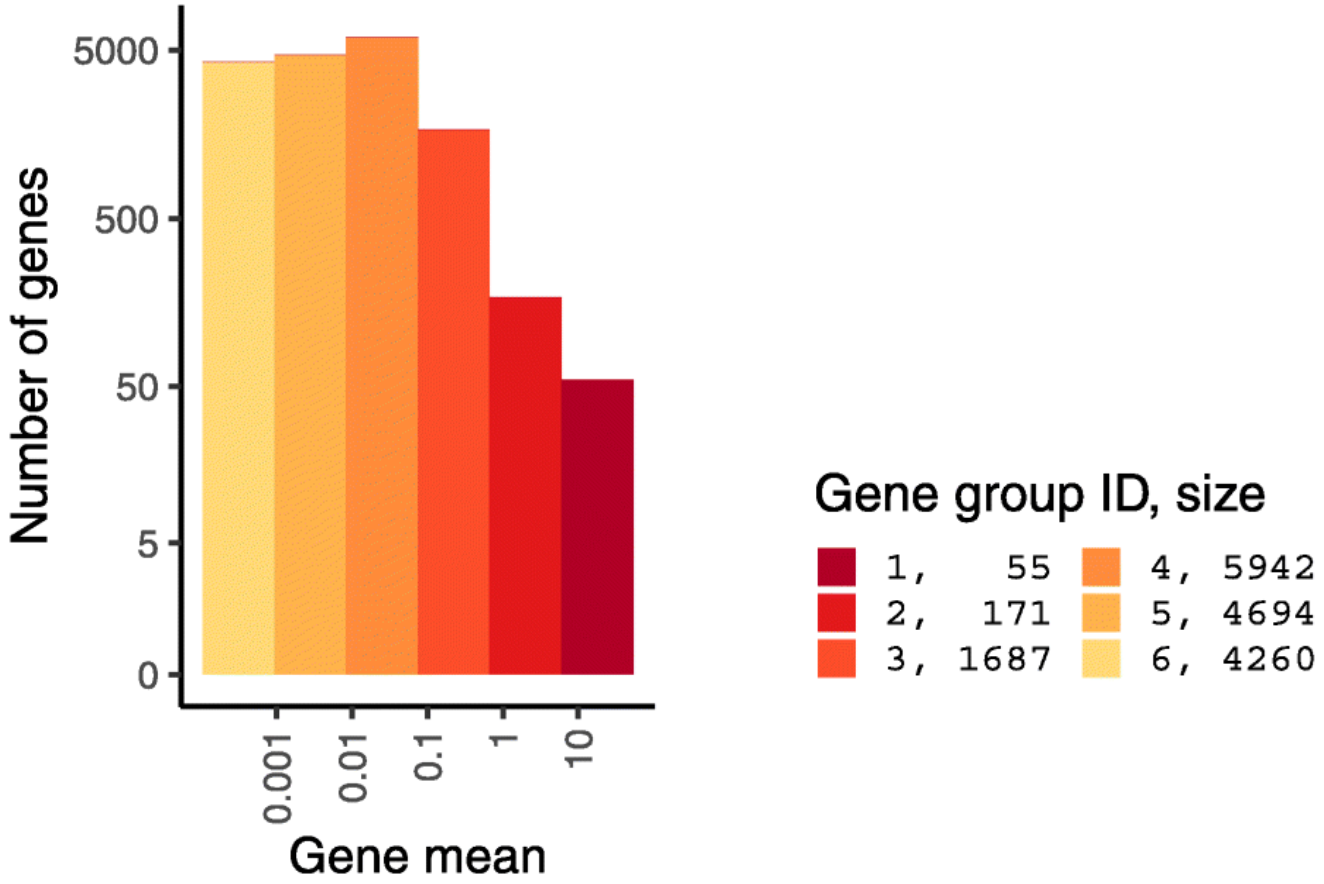


Townes,…, Irizarry (2019)

Does log-normalization (scaling) work?

33,148 PBMCs, 10x Genomics
16,809 genes detected ≥ 5 cells

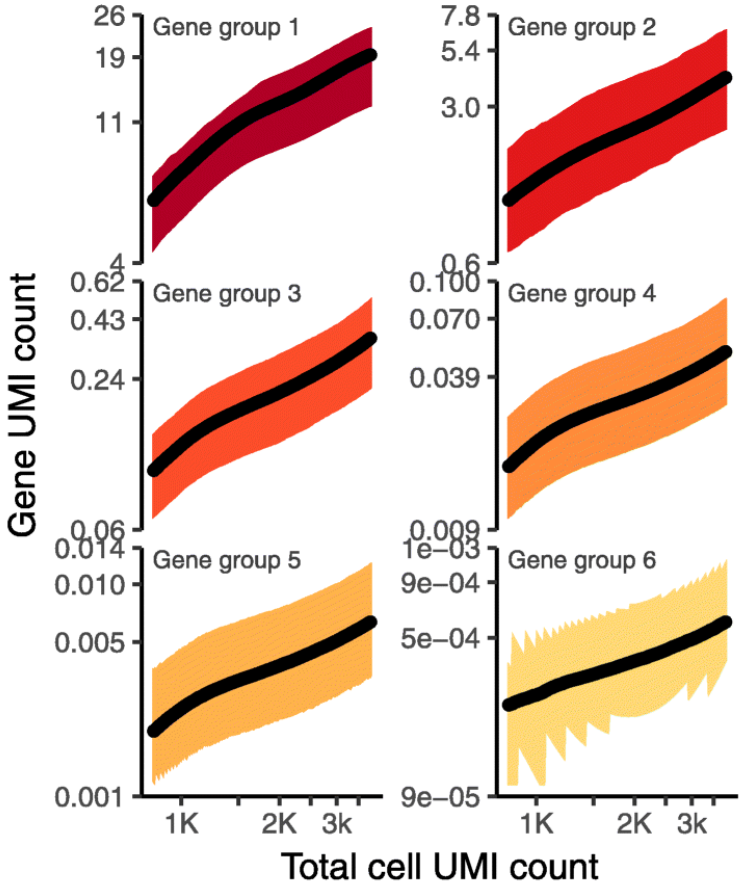




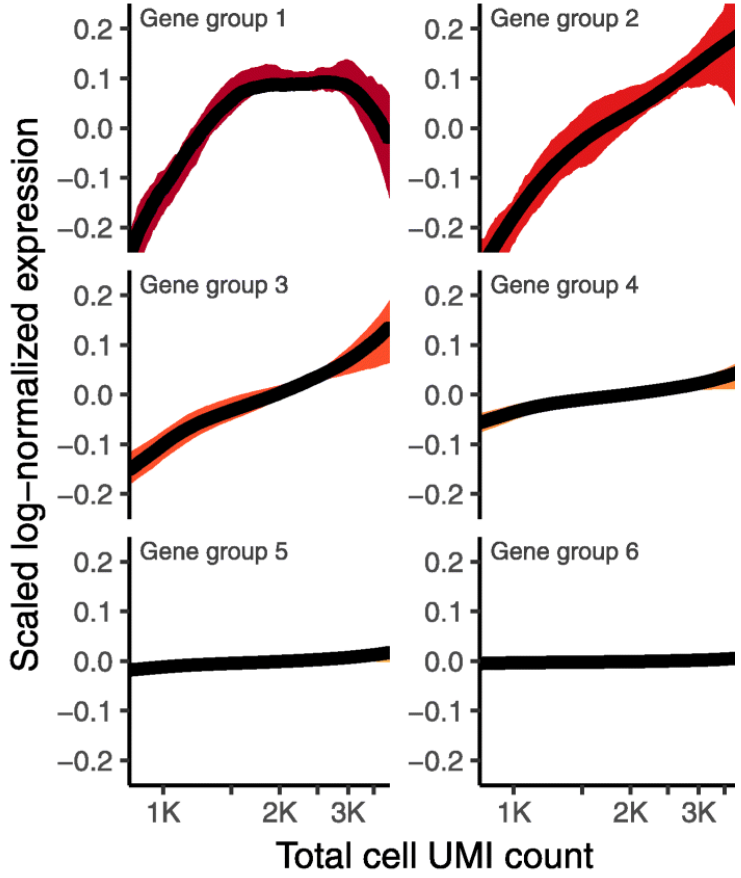
Hafemeister & Satija (2019)

Does log-normalization (scaling) work?

Before normalization



After normalization



Hafemeister & Satija (2019)

Modeling scRNAseq data

- Model the UMI counts for a given gene using a generalized linear model

$$\log\big(\mathbb{E}(x_i)\big) = \beta_0 + \beta_1 \log_{10} m + e_i$$

Allows variation in total library size rather than enforcing it as a constant metric.

x_i: vector of UMI counts assigned to gene i

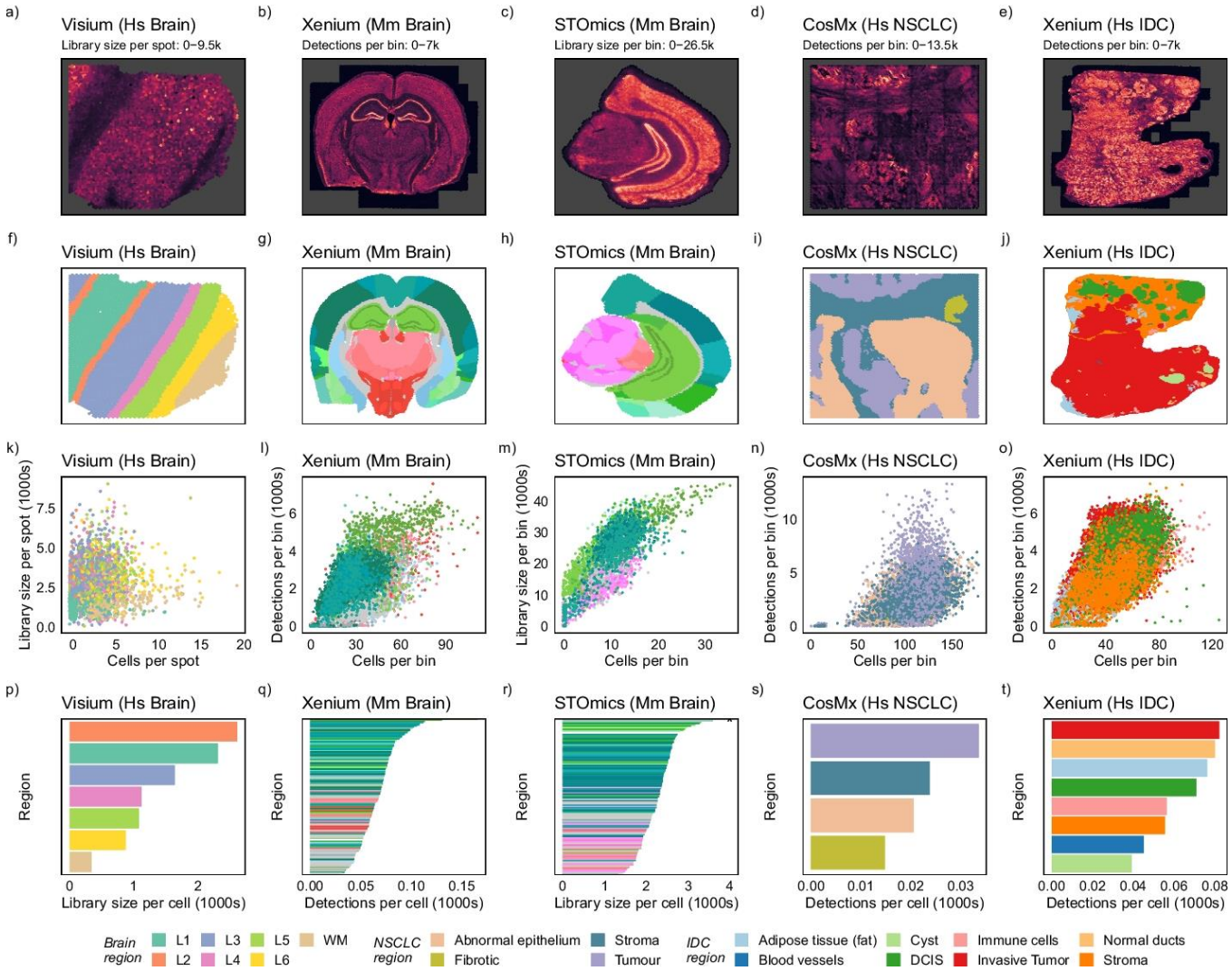
m: vector of molecules assigned to the cells, i.e., $m_j = \sum_i x_{ij}$

e_i: negative binomial (NB) error distribution, parameterized with mean μ and variance $\mu + \frac{\mu^2}{\sigma}$

Hafemeister & Satija (2019)

Library size ≈ biological variation

Library size normalization assumes that all regions of the tissue have the same underlying mRNA abundance



Bhuva,…, Davis (2024)

Which normalization method to use?

Bhuva *et al. Genome Biology* (2024) 25:99
https://doi.org/10.1186/s13059-024-03241-7

Genome Biology

SHORT REPORT **Open Access**

Library size confounds biology in spatial transcriptomics data

Dharmesh D. Bhuva[1,2,3*], Chin Wee Tan[2,3,4], Agus Salim[2,5], Claire Marceaux[3,6], Marie A. Pickering[7], Jinjin Chen[2,3], Malvika Kharbanda[1,2,3], Xinyi Jin[2,3], Ning Liu[1,2,3], Kristen Feher[1,2,3], Givanna Putri[2,3], Wayne D. Tilley[7], Theresa E. Hickey[7], Marie-Liesse Asselin-Labat[3,6], Belinda Phipson[2,3†] and Melissa J. Davis[1,2,3,4,8†]

- Tested the effects of normalization on spatial domain identification

Though sctransform removes library size effects effectively, their confounding with biology results in removal of biological effects as well.

See also for imageing-based SRT…

Atta *et al. Genome Biology* (2024) 25:153
https://doi.org/10.1186/s13059-024-03303-w

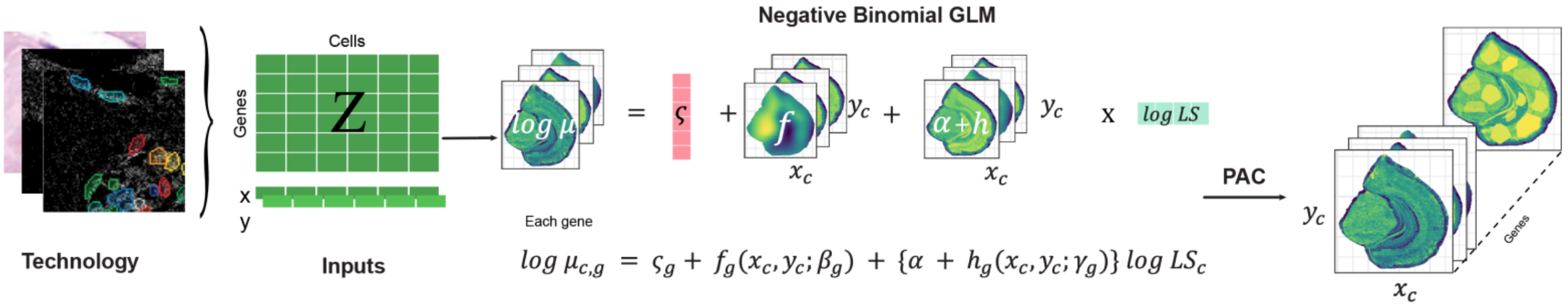
Genome Biology

RESEARCH **Open Access**

Gene count normalization in single-cell imaging-based spatially resolved transcriptomics

Lyla Atta[1,2], Kalen Clifton[1,2], Manjari Anant[2,3], Gohta Aihara[1,2] and Jean Fan[1,2*]

SpaNorm



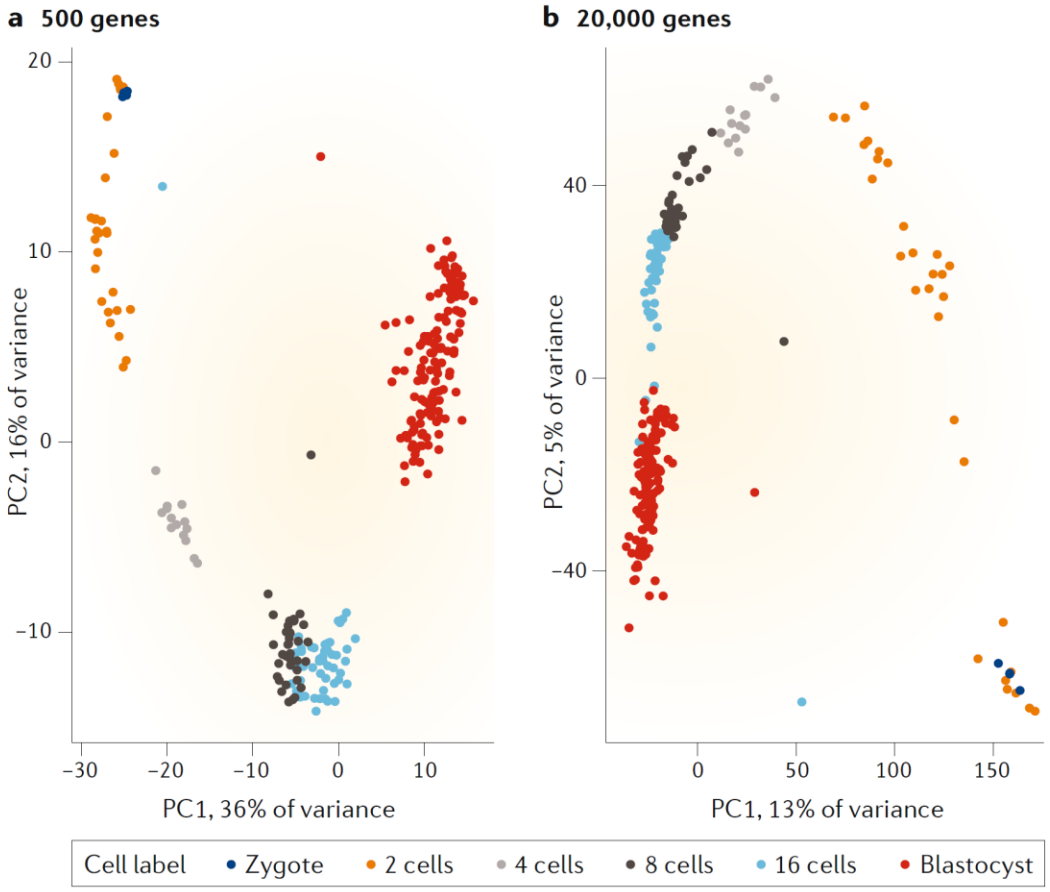
Salim*, Bhuva*,…, Davis (2024)

Feature (Gene) selection

Feature selection

- Curse of dimensionality
  - More features (genes) -> noise dominates distances between samples (cells), effectively all cells get 'same' distance

- Remove genes which only exhibit technical noise
  - Increase the signal:noise ratio
  - Reduce the computational complexity



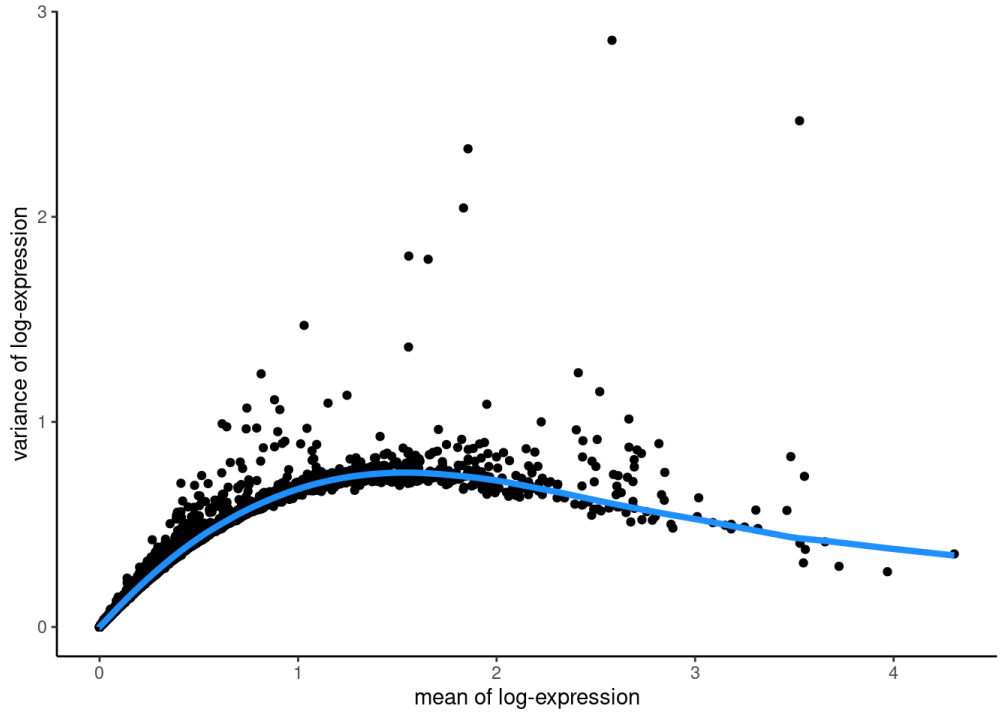
Kiselev,... (https://doi.org/10.1038/s41576-018-0088-9)

Feature selection

Highly Variable Genes (HVG)

- The simplest approach to quantifying per-gene variation is to compute the variance of the log-normalized expression values across all cells

- We define the **biological component** for each gene as the difference between its total variance and the technical component.
  - Used as the metric for HVG selection (e.g. top 2000 HVGs)

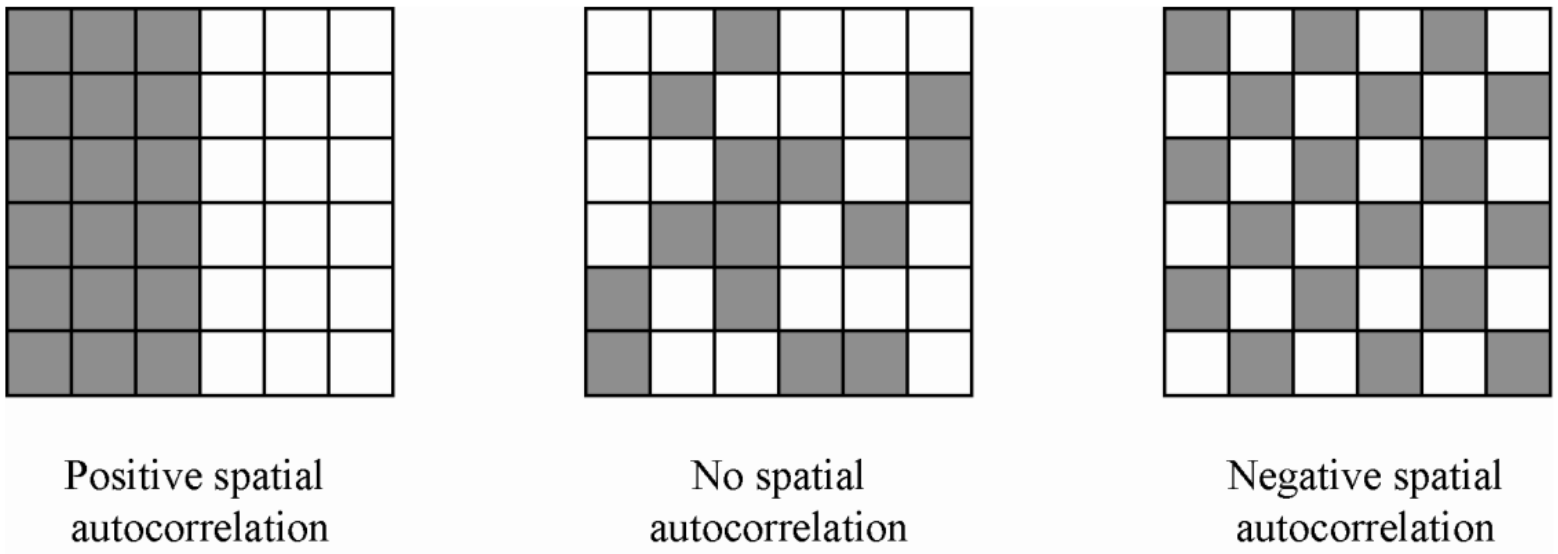


Orchestrating Single-Cell Analysis with Bioconductor; Lun, McCarthy, Marioni (2016)

Feature selection

Spatially Variable Genes (SVG)

"The first law of geography: Everything is related to everything else, but near things are more related than distant things." Waldo R. Tobler (Tobler 1970)



Boraks,…, Amend (2021)

Feature selection

Spatially Variable Genes (SVG)

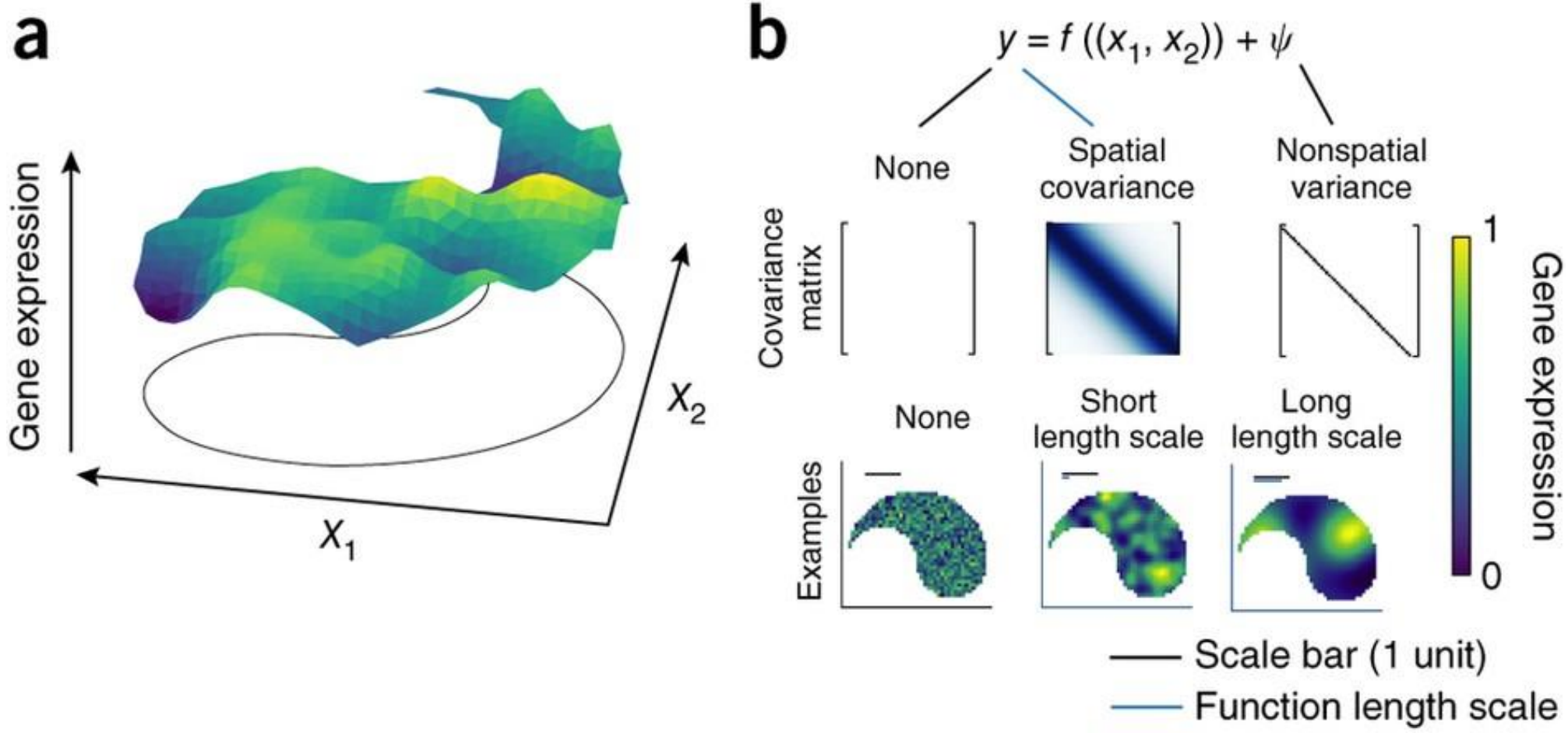
- HVGs selection methods designed for sc/snRNA-seq data ignores spatial coordinates

- SVGs are genes with a highly spatially correlated pattern of expression, which varies along with the spatial distribution of a tissue structure of interest

- Statistical measures of spatial autocorrelation:
  - Moran's I
  - Geary's C

  …

Statistical models of SRT

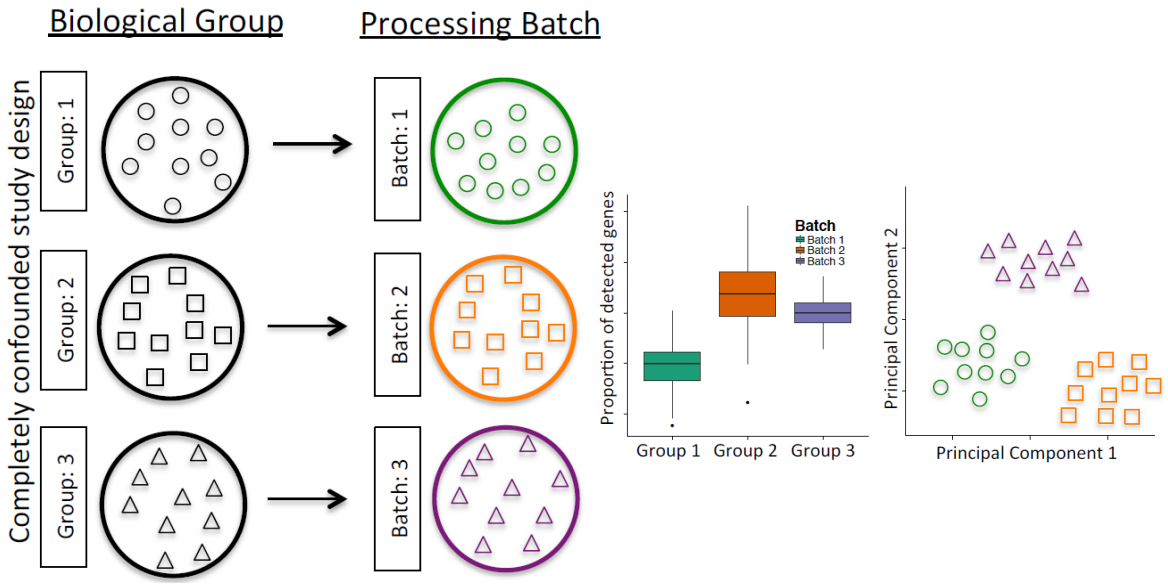


SpatialDE: Svensson et al. Nature Methods 2018
SPARK: Sun et al. Nature Methods 2020
SPARK-X: Zhu et al. genome Biology 2021

Batch Correction

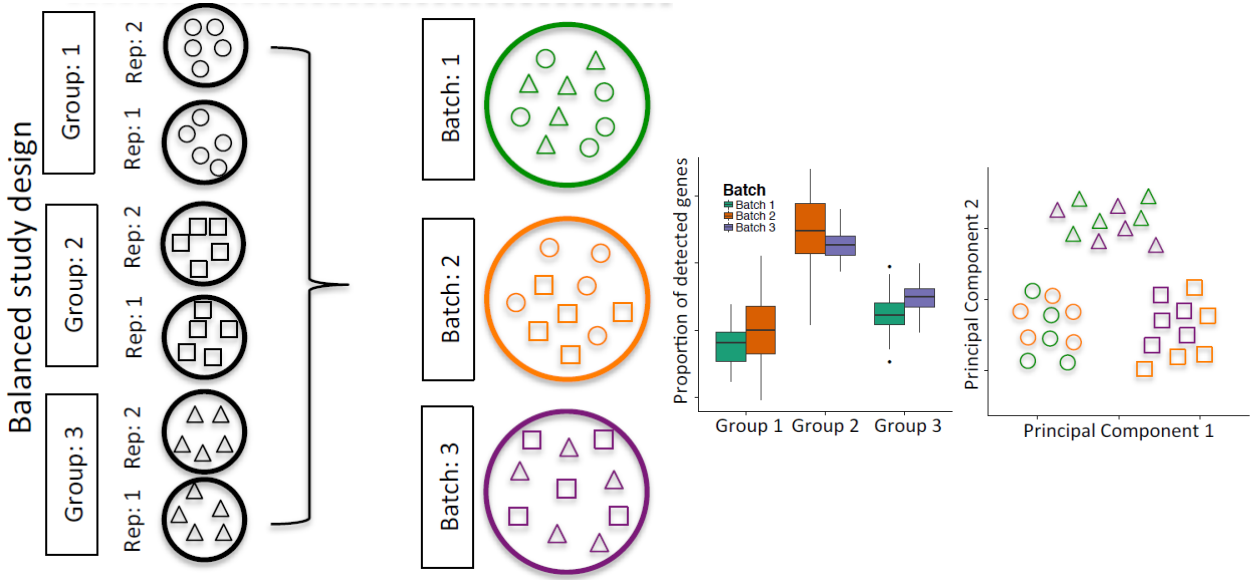
Confounders and batch effects

Confounded design



Don’t design your experiment like this!!!

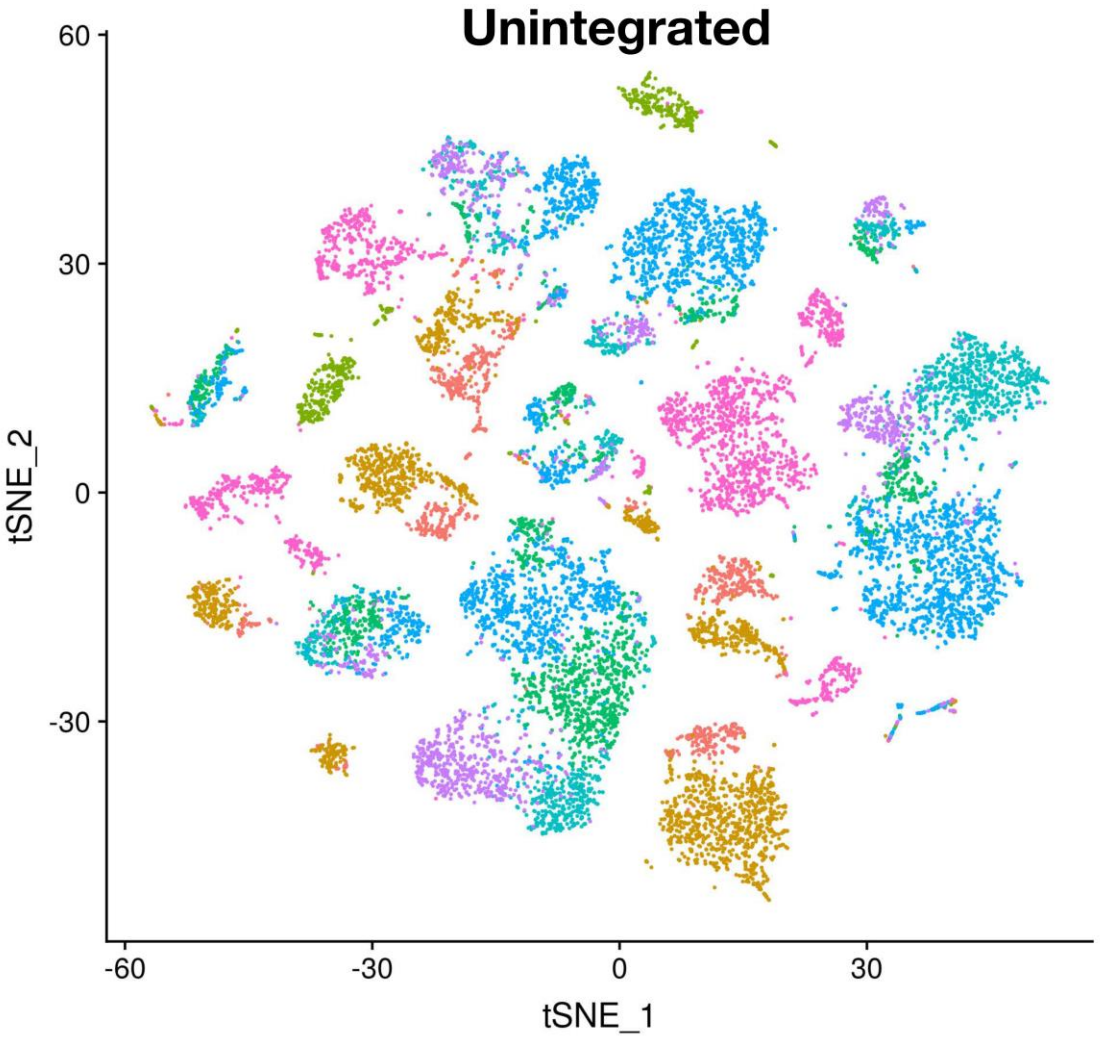
Not confounded design

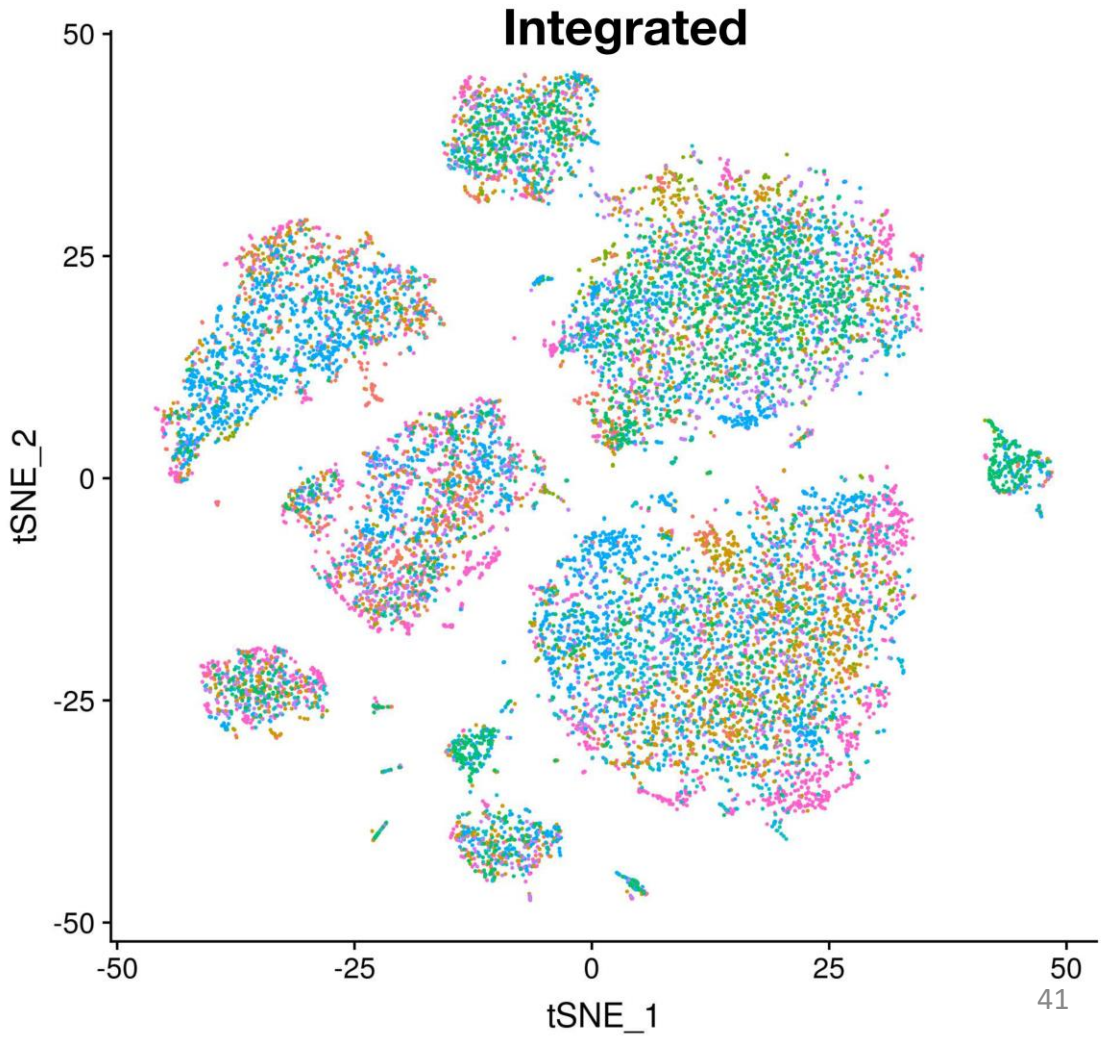


Good experimental design *does not remove batch effects*, it prevents them from biasing your results.

Hicks et al. (https://doi.org/10.1093/biostatistics/kxx053)

Batch correction





Batch correction methods

- scVI (https://doi.org/10.1038/s41592-018-0229-2)
- MNNcorrect (https://doi.org/10.1038/nbt.4091)
- CCA + anchors (Seurat v3) (https://doi.org/10.1101/460147)
- CCA + dynamic time warping (Seurat v2) (https://doi.org/10.1038/nbt.4096)
- LIGER (https://doi.org/10.1101/459891)
- Harmony (https://doi.org/10.1101/461954)
- Conos (https://doi.org/10.1101/460246)
- Scanorama (https://doi.org/10.1101/371179)
- scMerge (https://doi.org/10.1073/pnas.1820006116)
- …

Spatially-unaware!

Summary

- Quality control (QC)
- Normalization
- Feature selection
- Batch correction

Thank You!

a.mahfouz@lumc.nl
mahfouzlab.org