Swiss Institute of Bioinformatics

INTRODUCTION TO SEQUENCING-BASED SPATIAL TRANSCRIPTOMICS DATA ANALYSIS

QC and Normalization

Deepak Tanwar

December 9-10, 2025

elixir SWITZERLAND

Which QC do you use for single-cell data?

QC for Visium HD is conceptually the same as single-cell QC – but computed *per High-Definition spot* rather than per cell.

Cell → Spot

Why QC of spatial data?

High-density spots can capture technical artifacts (folds, bubbles, staining issues).

Why QC of spatial data?

High-density spots can capture technical artifacts (folds, bubbles, staining issues).

Artifacts are often **spatially coherent** – affecting contiguous regions, not just individual spots.

Why QC of spatial data?

High-density spots can capture technical artifacts (folds, bubbles, staining issues).

Artifacts are often **spatially coherent** – affecting contiguous regions, not just individual spots.

Poor-quality regions can bias downstream analyses (clustering, spatial DE, trajectory analysis).

QC in Visium HD

Same metrics as single-cell RNA-seq:

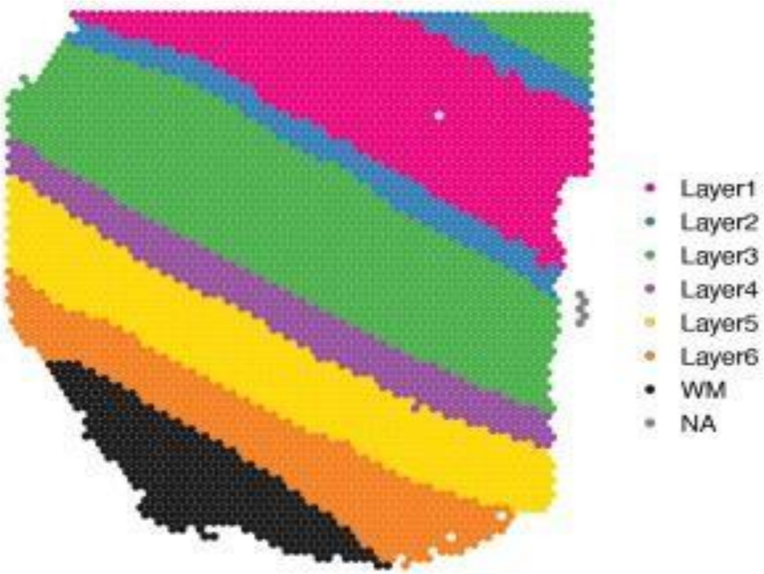
- Total UMIs per spot
- Number of genes per spot
- Mitochondrial fraction

Key difference: unit is HD spot, not single cell.

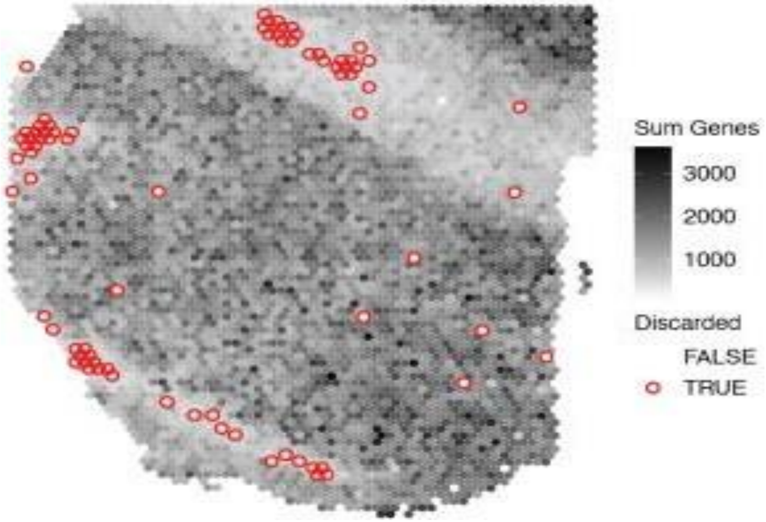
Neighboring spots share microenvironment → spatial artifacts can emerge

Example of spots removal

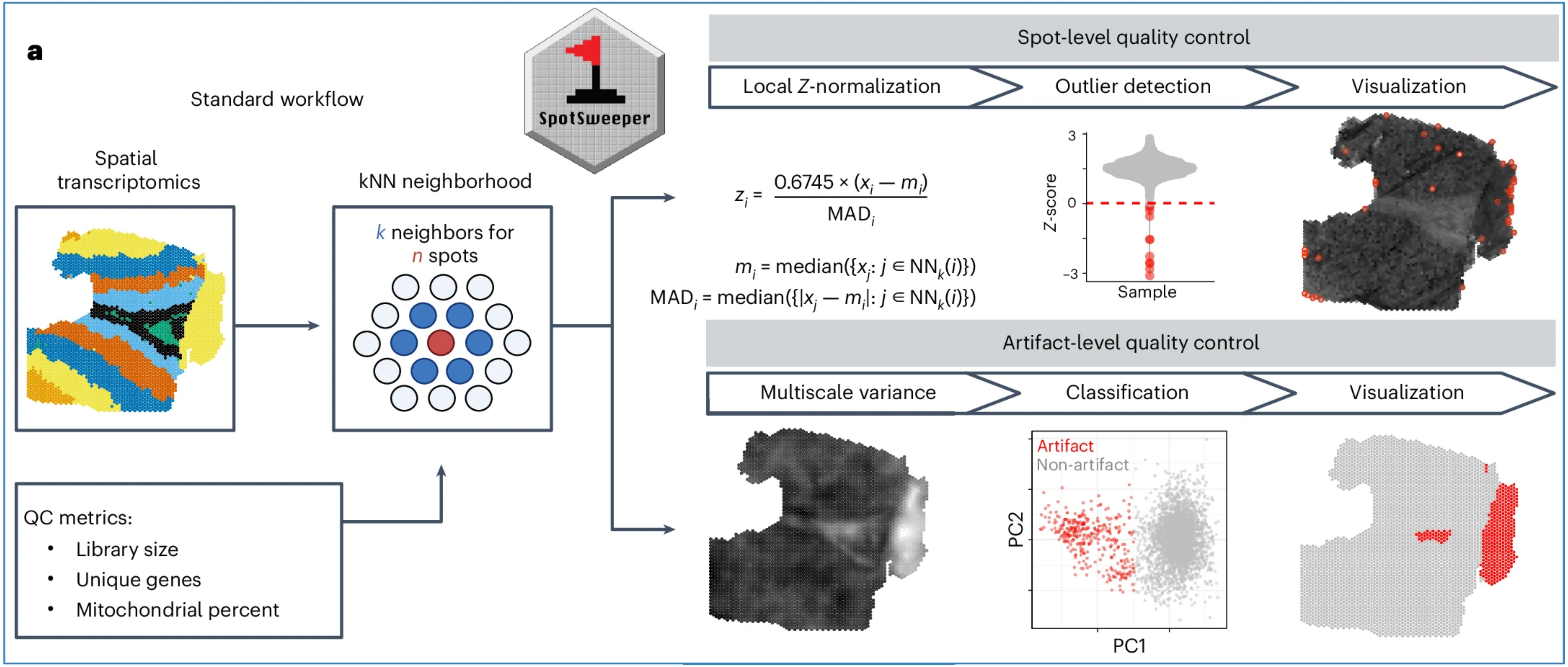
Annotation



Global



SpotSweeper: spatially aware QC



Standard workflow

Spatial transcriptomics

kNN neighborhood

k neighbors for *n* spots

QC metrics:

- Library size
- Unique genes
- Mitochondrial percent

Spot-level quality control

Local *Z*-normalization → Outlier detection → Visualization

$$z_i = \frac{0.6745 \times (x_i - m_i)}{\mathrm{MAD}_i}$$

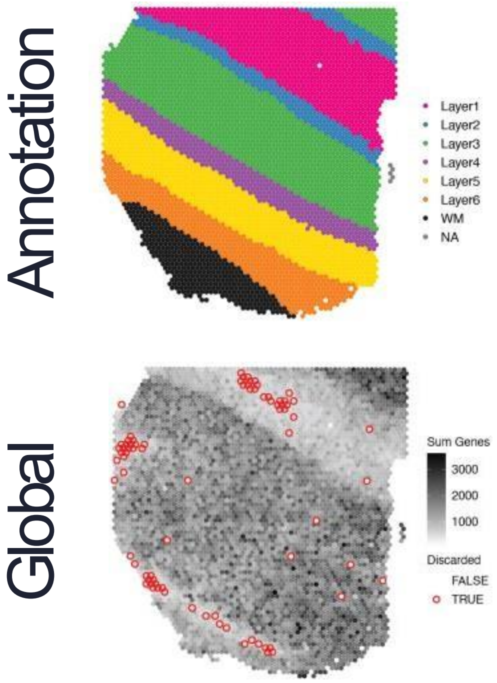
$$m_i = \mathrm{median}(\{x_j : j \in \mathrm{NN}_k(i)\})$$

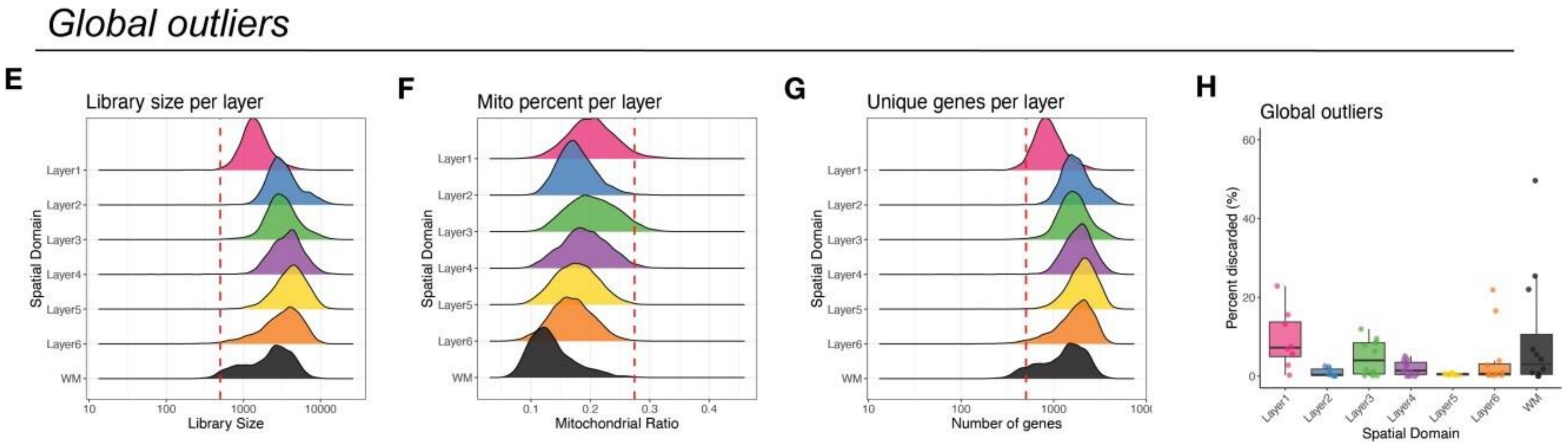
$$\mathrm{MAD}_i = \mathrm{median}(\{|x_j - m_i| : j \in \mathrm{NN}_k(i)\})$$

Artifact-level quality control

Multiscale variance → Classification → Visualization

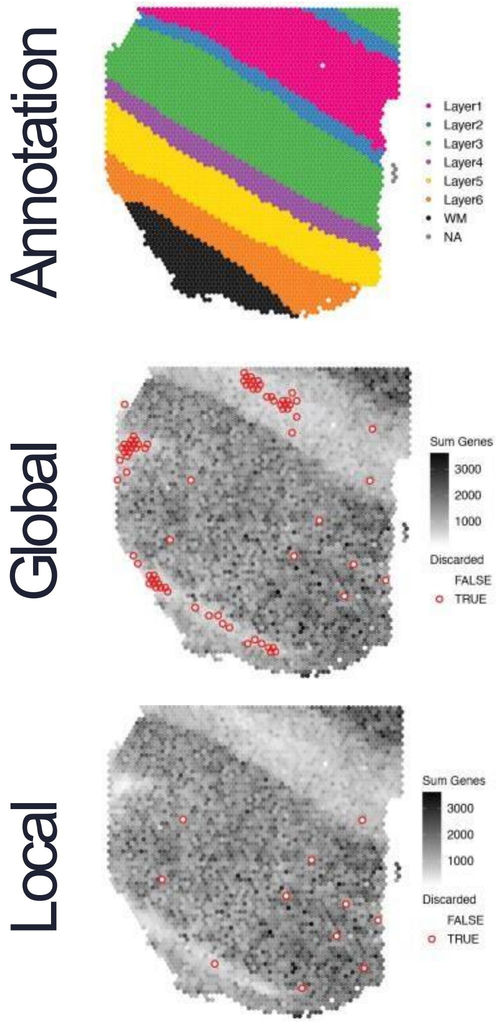
SpotSweeper (spot-level artifacts)

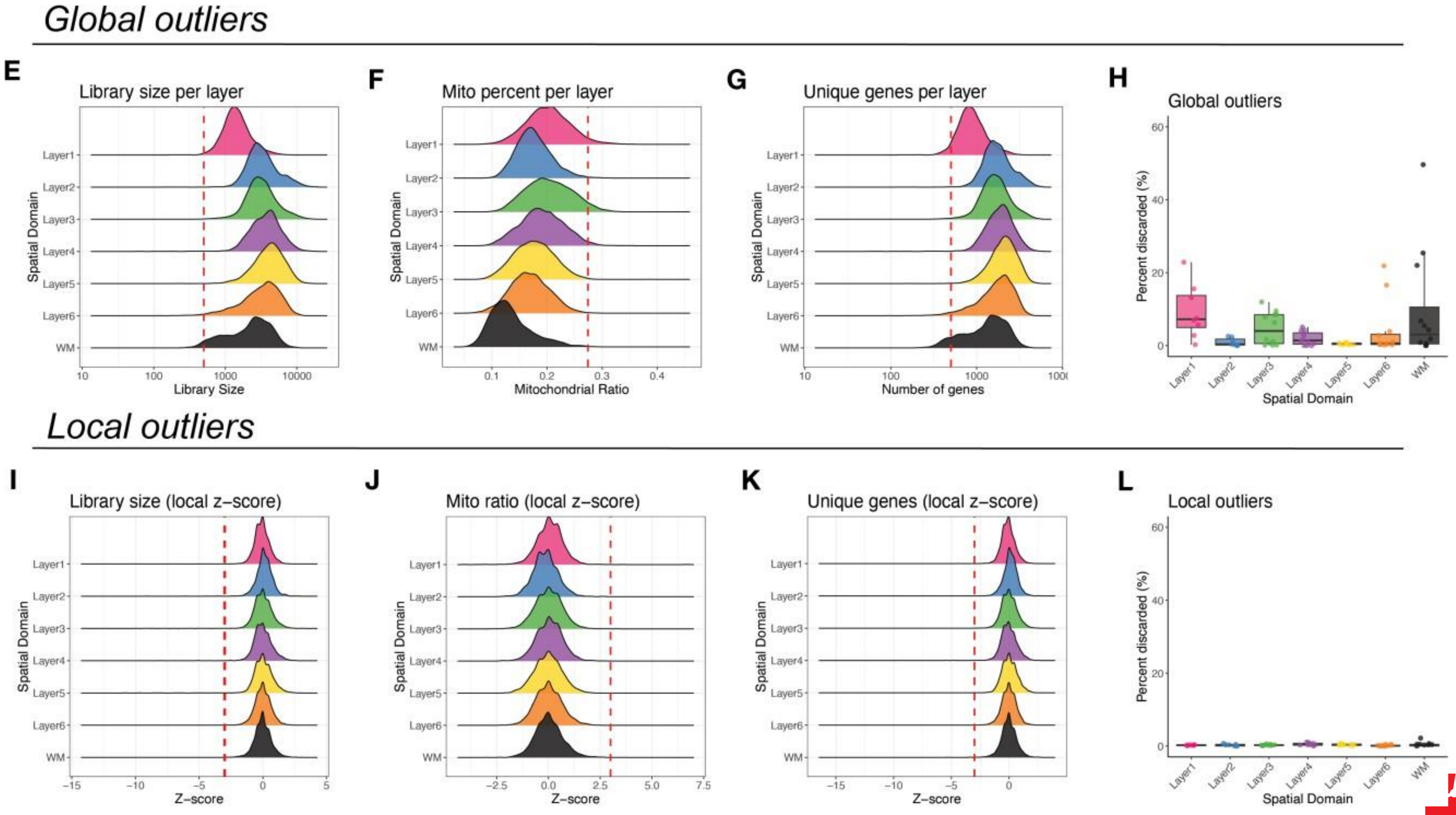




Totty, Hicks, Guo (2024)

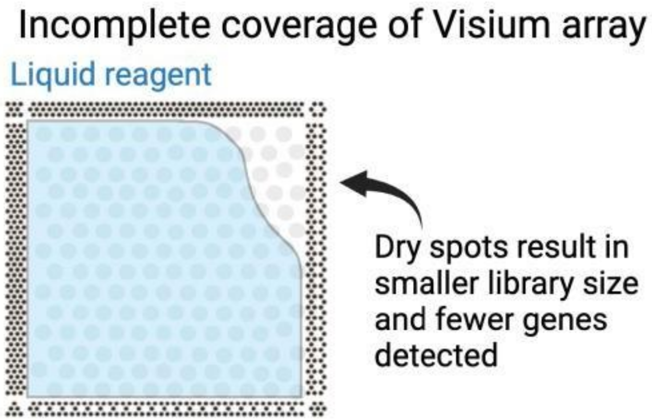
SpotSweeper (spot-level artifacts)

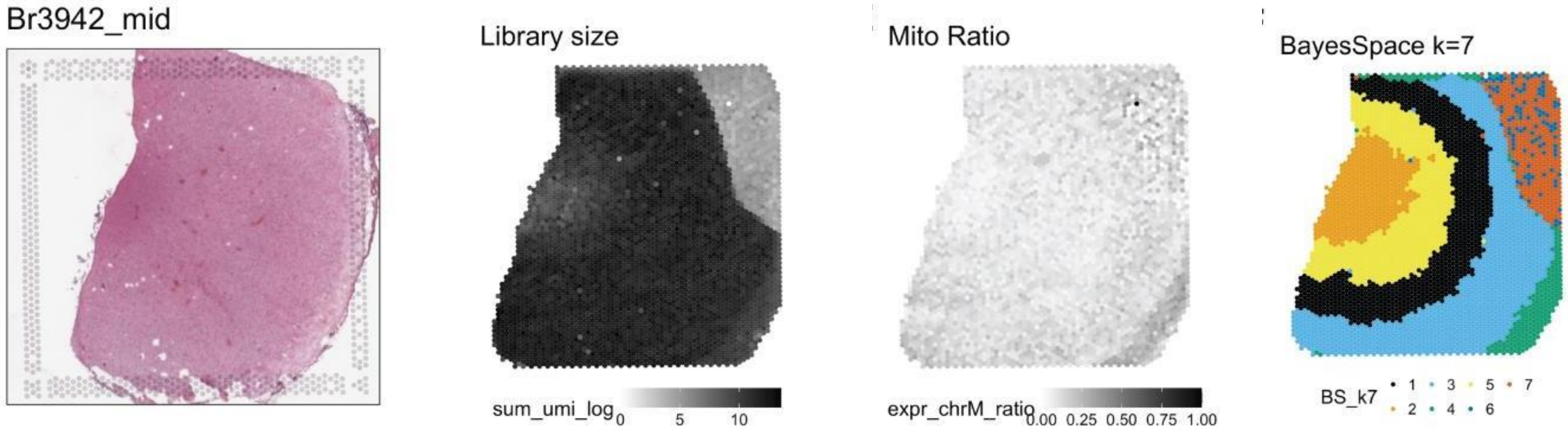




Totty, Hicks, Guo (2024)

SpotSweeper (region-level artifacts)

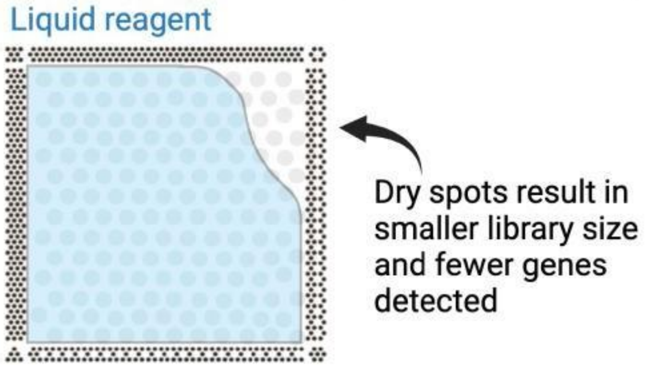




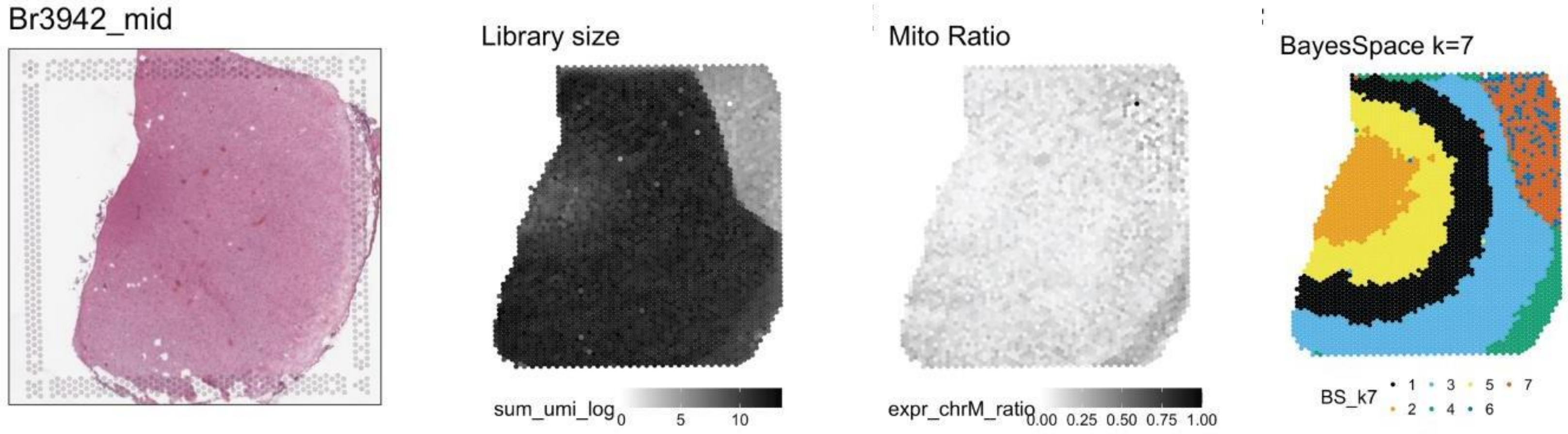
Totty, Hicks, Guo (2024)

SpotSweeper (region-level artifacts)

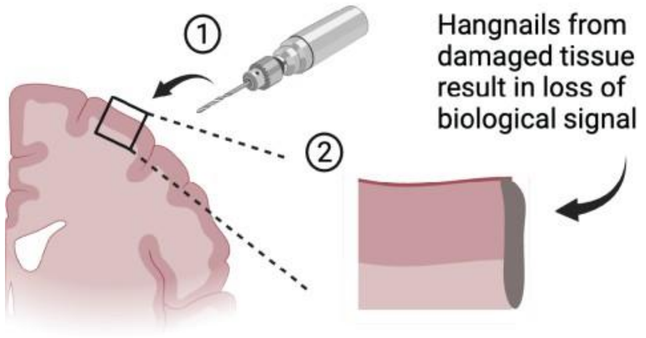
Incomplete coverage of Visium array



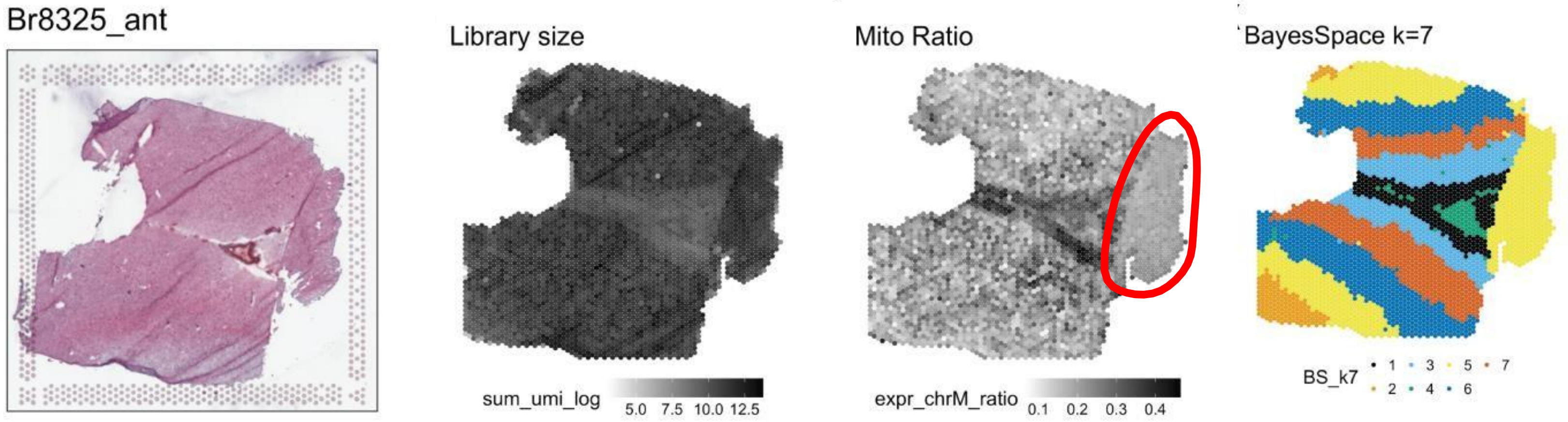
Dry spots result in smaller library size and fewer genes detected



Tissue damage during dissection



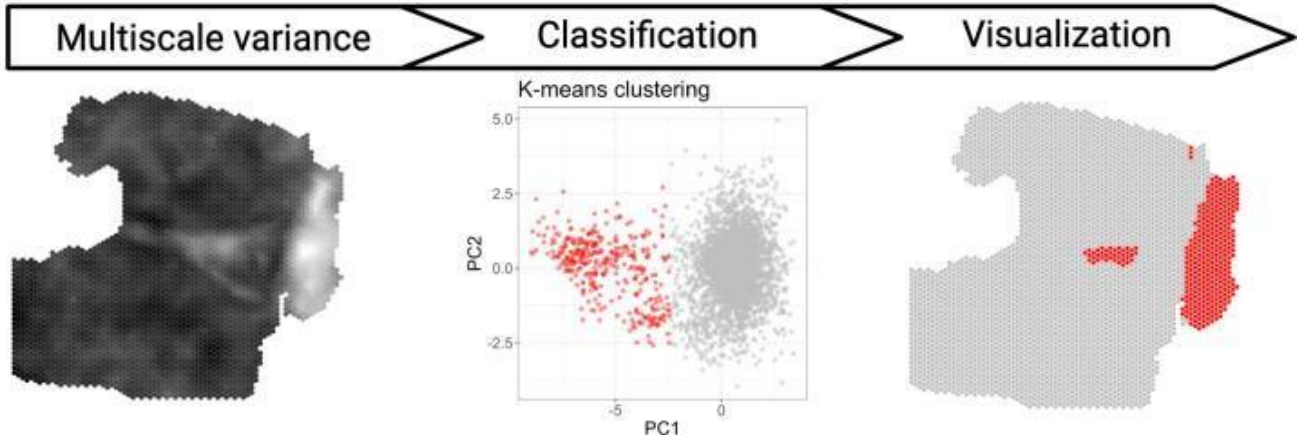
Hangnails from damaged tissue result in loss of biological signal



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)

Preserve biological variability while removing technical artifacts: SpotSweeper

Detects **spatially coherent low-quality regions** in Visium HD slides.

Uses **per-spot QC metrics** (mainly mitochondrial fraction) and **spatial neighborhood info**.

Flags artifacts for downstream masking/removal.

Works across **multiple neighborhood scales** to detect both small and large defects.

How do you normalize your single-cell data?

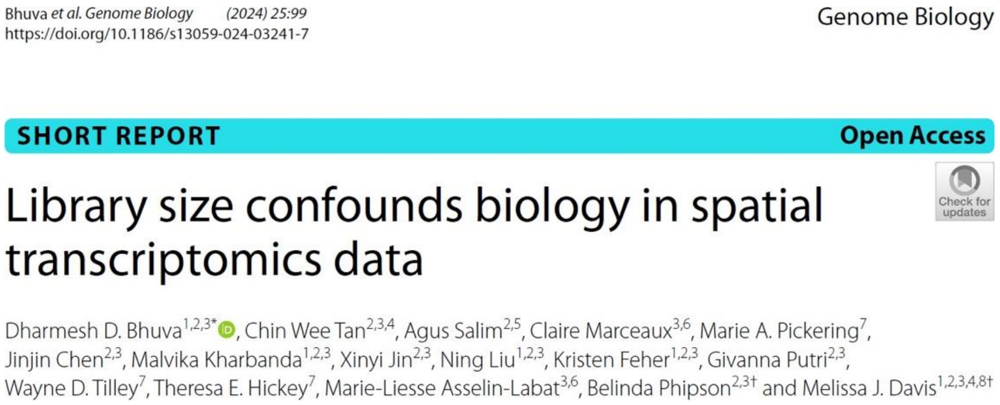
Log-normalization

$$Y_{ij} = log_e((\frac{X_{ij}}{σ_i X_{ij}}) + 1)$$

- Simplest and most commonly-used normalization strategy
- Divide all counts for each spot by a spot-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

How about any other method?

Can we use scTranscform?


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

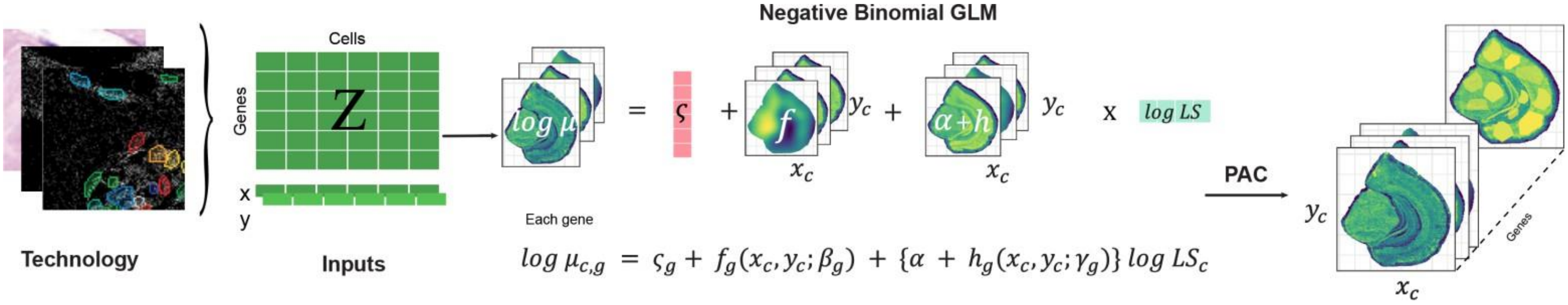
Genome Biology

RESEARCH **Open Access**

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

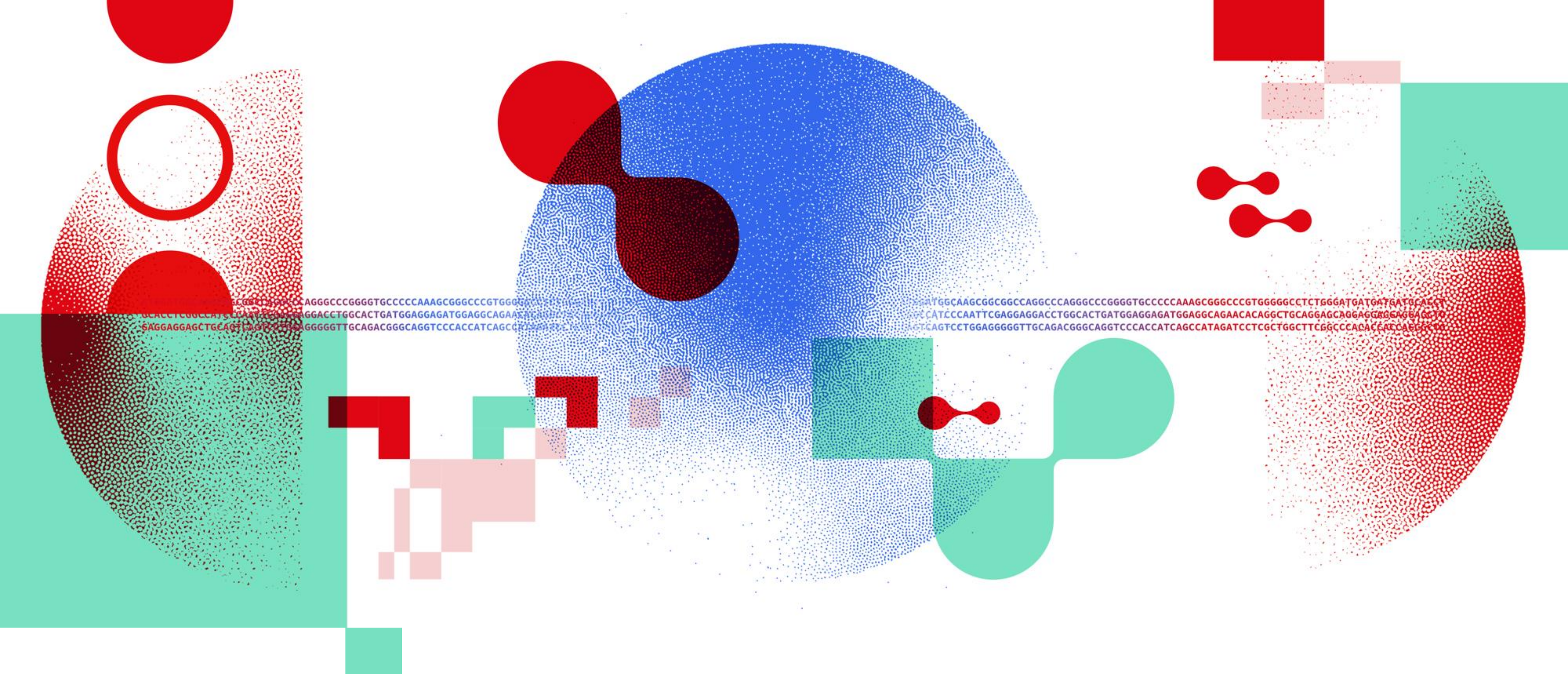
Lyla Atta[1,2], Kalen Clifton[1,2], Manjari Anant[2,3], Gohta Aihara[1,2] and Jean Fan[1,2*]

SpaNorm



$$\log \mu_{c,g} = \varsigma_g + f_g(x_c, y_c; \beta_g) + \{\alpha + h_g(x_c, y_c; \gamma_g)\} \log LS_c$$

Salim*, Bhuva*,…, Davis (2024)



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss