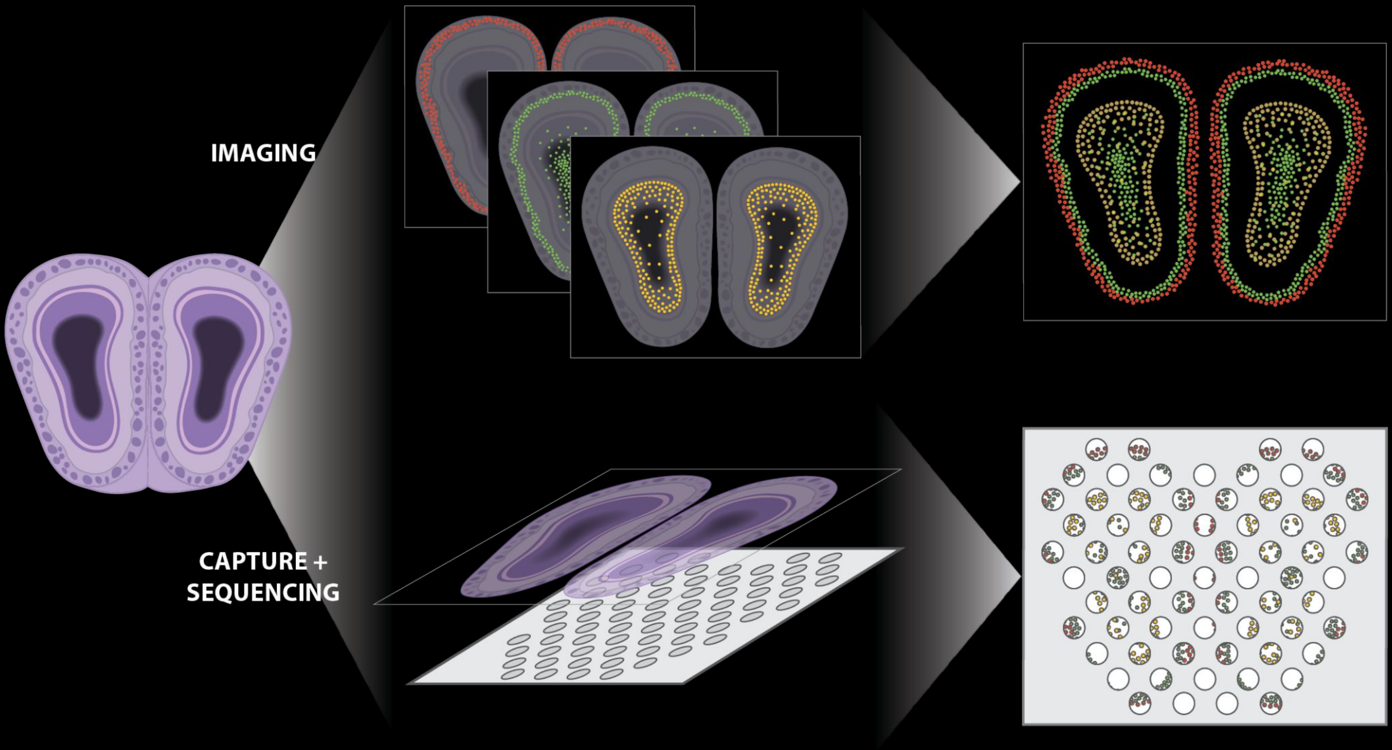
# Gene count normalization in single-cell imaging-based spatially resolved transcriptomics

**Lyla Atta**
MD-PhD Candidate, JEFworks Lab
*Johns Hopkins University*

*@JEFworks @lylaatta*

# Spatially resolved transcriptomics (SRT)



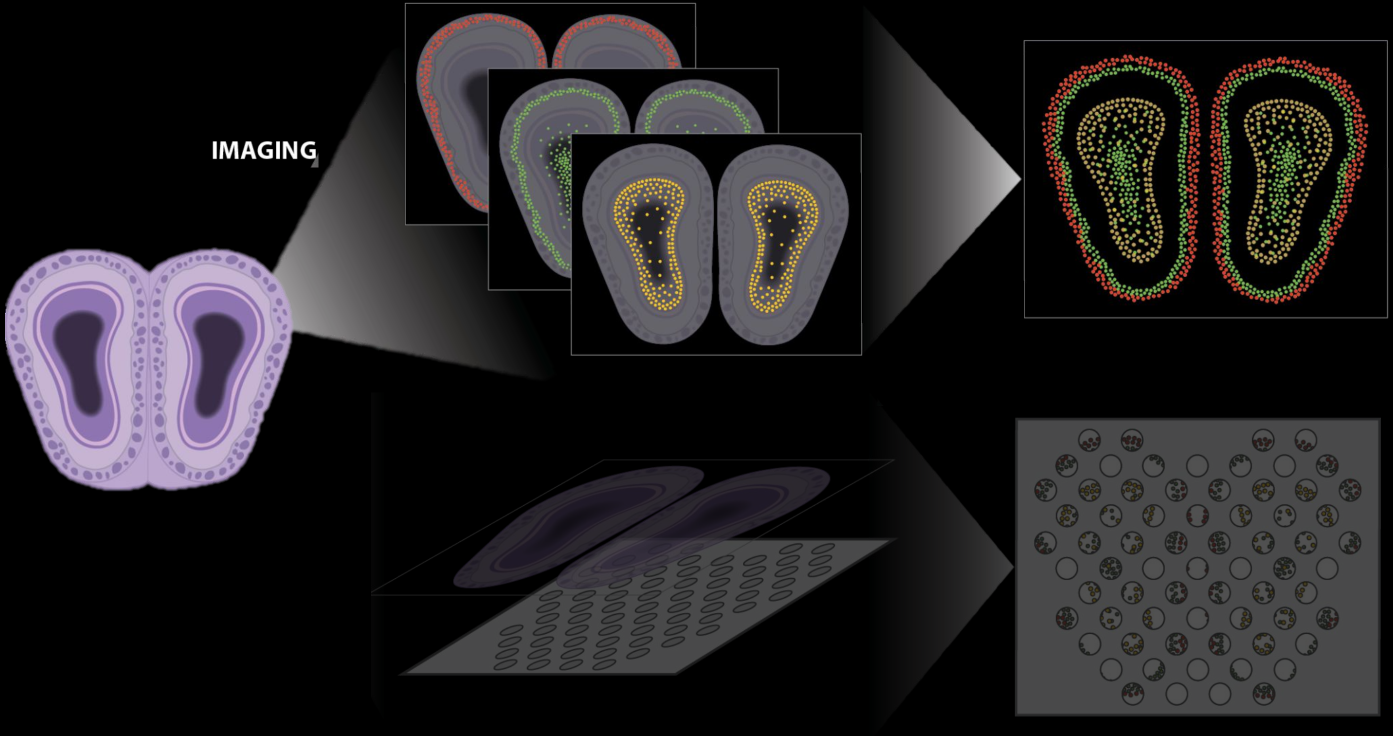
**Imaging-based SRT**

- Single molecule
- Targeted
- ~100-500 genes

**Sequencing-based SRT**

- Spot resolution
- Untargeted
- Whole transcriptome

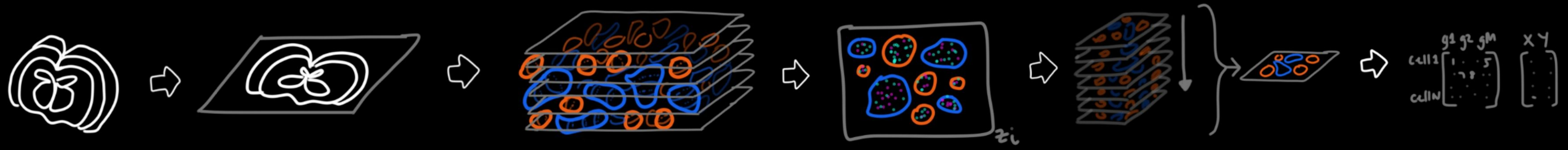
# Spatially resolved transcriptomics (SRT)



**Imaging-based SRT**

- Single molecule
- Targeted
- ~100-500 genes

# imSRT data: generating cell x gene count data



Specimen

Tissue section

Image z-stack

RNA ID

Cell segmentation

Spot assignment + RNA count

# imSRT data: generating cell x gene count data

Specimen

# imSRT data: generating cell x gene count data

Specimen

Tissue section

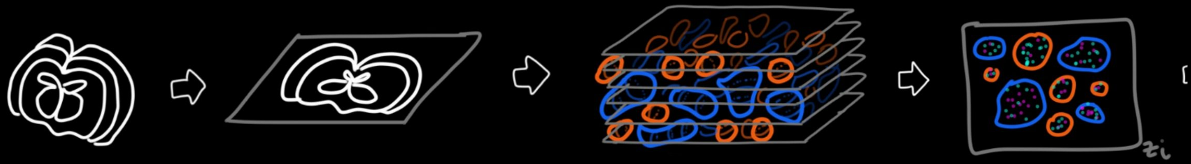
# imSRT data: generating cell x gene count data

Specimen

Tissue section

Image z-stack

# imSRT data: generating cell x gene count data



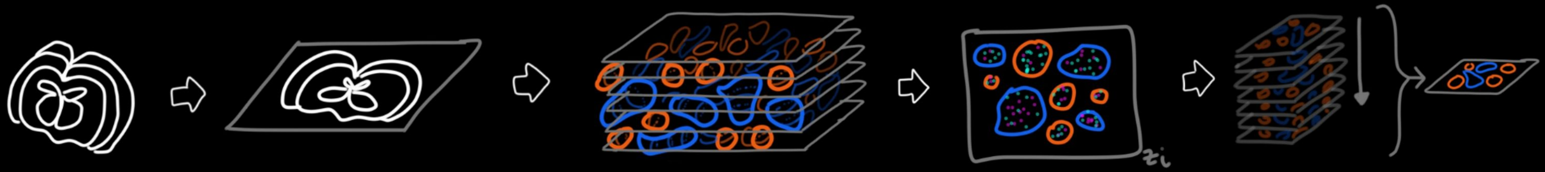
Specimen

Tissue
section

Image
z-stack

RNA ID

# imSRT data: generating cell x gene count data



Specimen

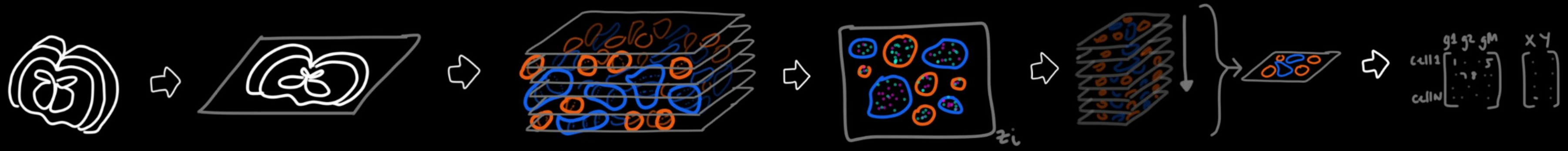
Tissue
section

Image
z-stack

RNA ID

Cell
segmentation

# imSRT data: generating cell x gene count data



Specimen

Tissue section

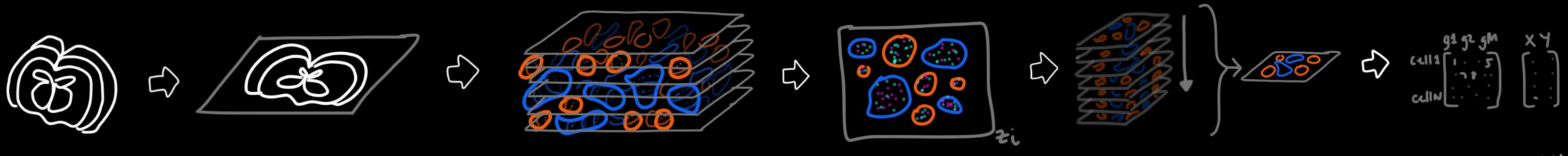
Image z-stack

RNA ID

Cell segmentation

Spot assignment + RNA count

# imSRT data: generating cell x gene count data



Specimen

Tissue section

Image z-stack

RNA ID

Cell segmentation

Spot assignment + RNA count

bioRχiv **Comparative analysis of multiplexed in situ gene expression profiling technologies** ***Hartman, Satija, 2024***

# imSRT: computational methods to generate biological hypotheses

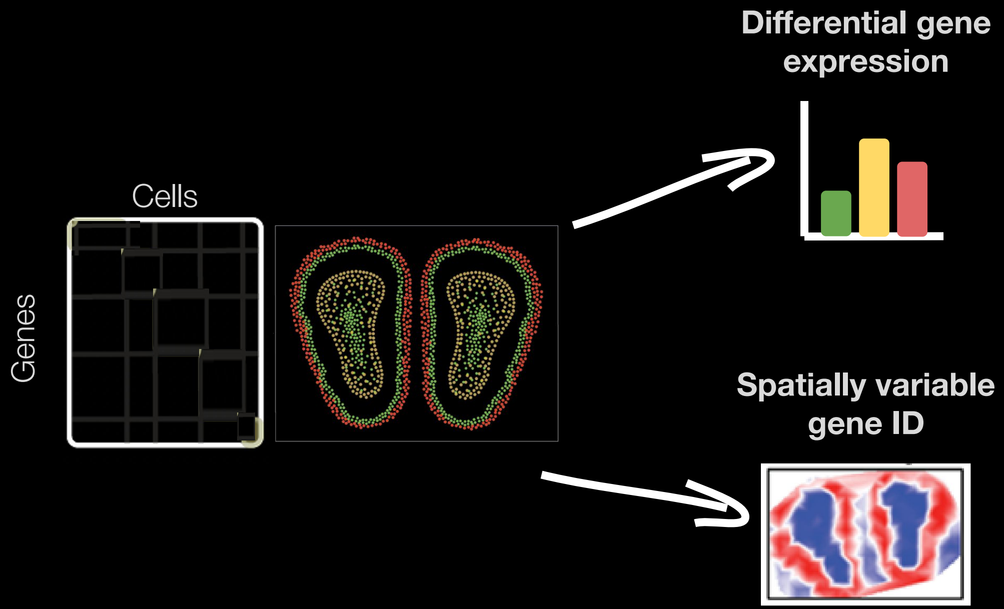
# imSRT: computational methods to generate biological hypotheses

Differential gene expression

Seurat v5

scanpy

# imSRT: computational methods to generate biological hypotheses



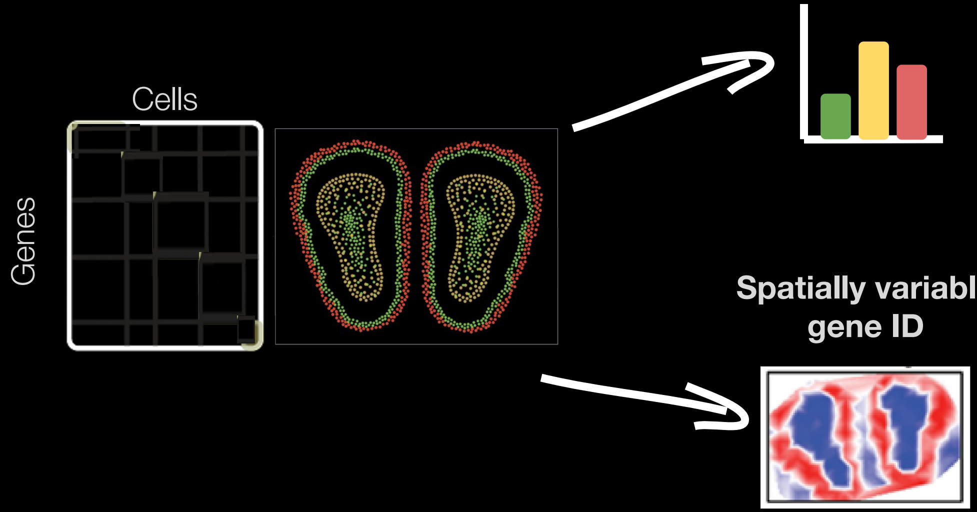
**Differential gene expression**

**Spatially variable gene ID**

Seurat v5

scanpy

# imSRT: computational methods to generate biological hypotheses



**Differential gene expression**

Seurat v5

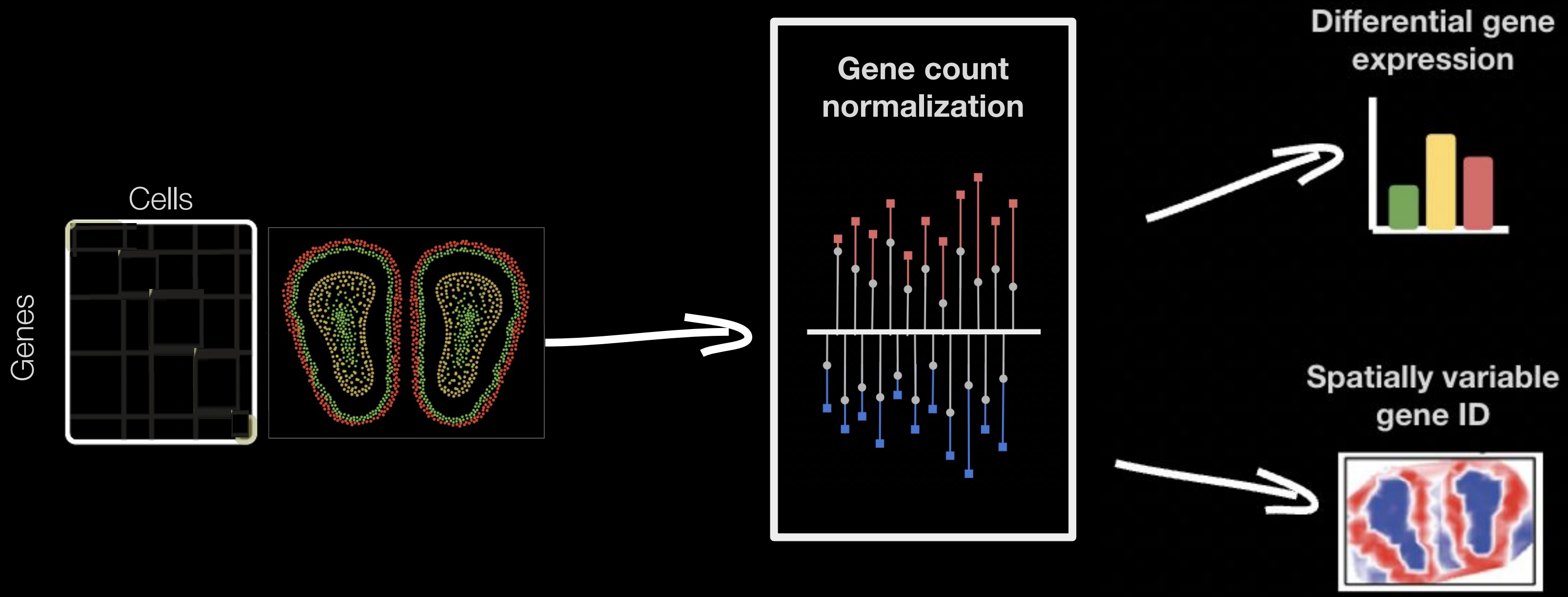
scanpy

**Spatially variable gene ID**

MERINGUE

nnSVG, spatialDE

# imSRT: gene count normalization is the first step to many downstream analyses



Cells

Genes

Gene count normalization

Differential gene expression

Spatially variable gene ID

# Count normalization in imaging-based SRT

☐ Goals of count normalization, current normalization approaches, evaluation criteria

☐ Case study: ventricle gene panel in mouse brain MERFISH imSRT

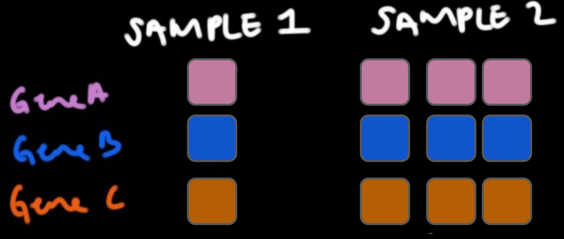
☐ Evaluating normalization across gene panels, technologies and tissues

☐ Recommendations

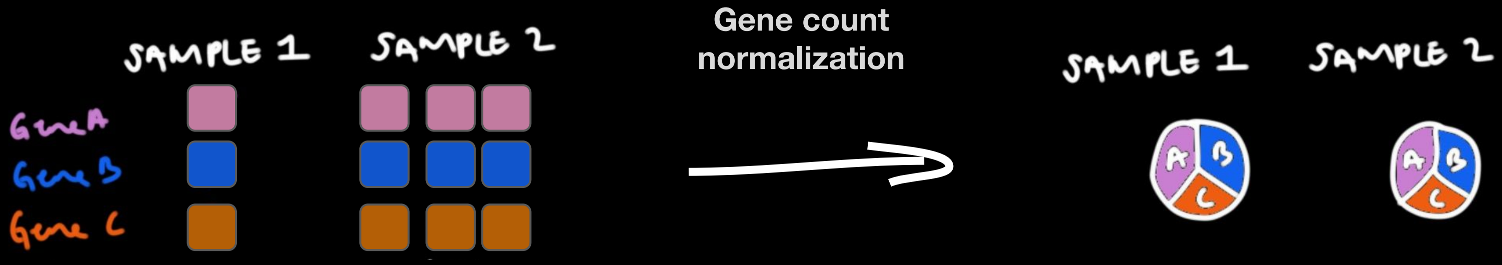
# Count normalization in imaging-based SRT

- ☐ **Goals of count normalization, current normalization approaches, evaluation criteria**
- ☐ Case study: ventricle gene panel in mouse brain MERFISH imSRT
- ☐ Evaluating normalization across gene panels, technologies and tissue types
- ☐ Recommendations

# Gene count normalization: accounting for technical variation

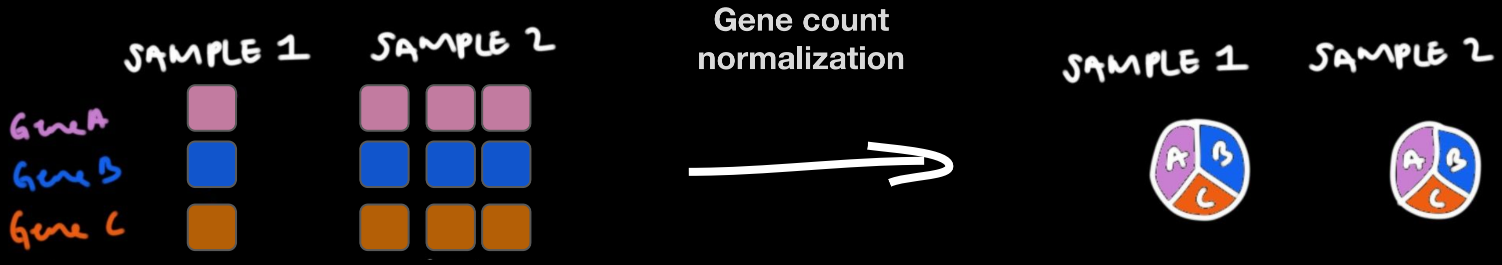


# Gene count normalization: accounting for technical variation



Gene count normalization

# Gene count normalization: accounting for technical variation


Gene count normalization

- **Bulk RNA-seq:** PCR bias, reagent variation
- **scRNA-seq:** stochastic RNA capture
- **imSRT:** partial cell volume imaging

# Gene count normalization: current approaches

| Library size | DESeq2 | TMM (edgeR) | scTransform | Cell volume/ cell area |
|---|---|---|---|---|
| Bulk RNA-seq | Bulk RNA-seq | Bulk RNA-seq | | |
| scRNA-seq | | | scRNA-seq | |
| imSRT | | | | imSRT |

# Gene count normalization: current approaches

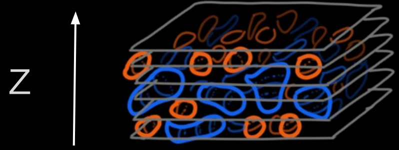
| Library size | DESeq2 | TMM (edgeR) | scTransform | Cell volume/ cell area |
|---|---|---|---|---|
| Bulk RNA-seq | Bulk RNA-seq | Bulk RNA-seq | | |
| scRNA-seq | | | scRNA-seq | |
| imSRT | | | | imSRT |

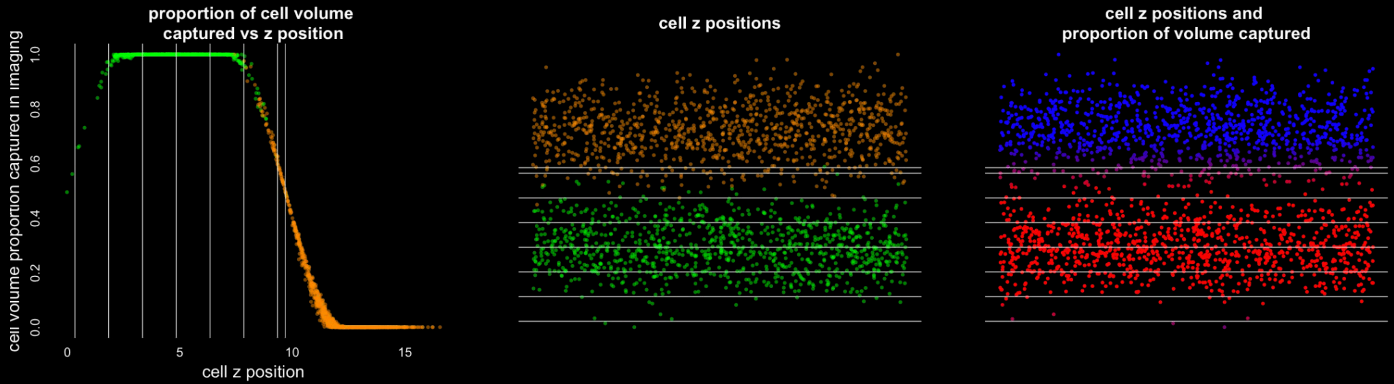
COUNT BASED NORMALIZATION

# Gene count normalization: is it necessary?

## Simulation to evaluate systematic RNA capture biases

One cell type, two z locations

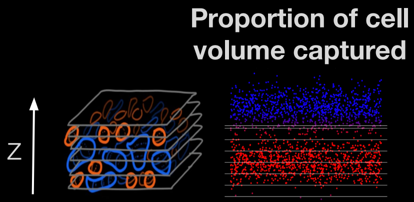


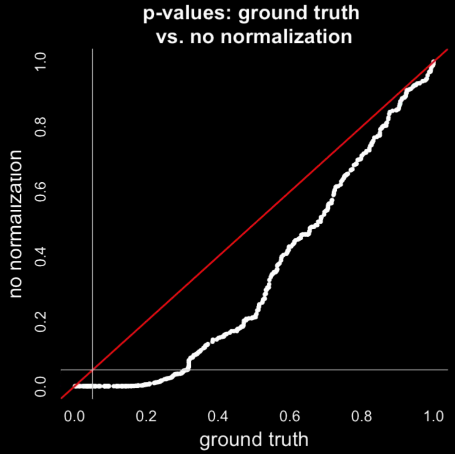


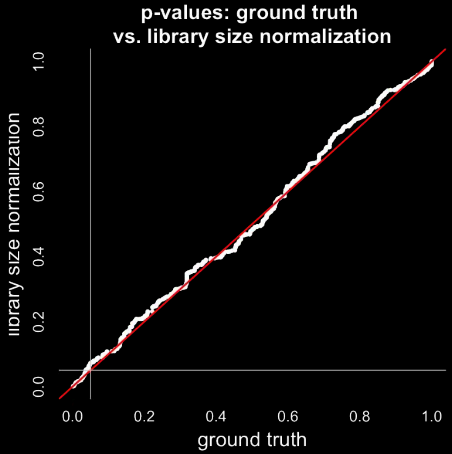
# Gene count normalization: is it necessary?

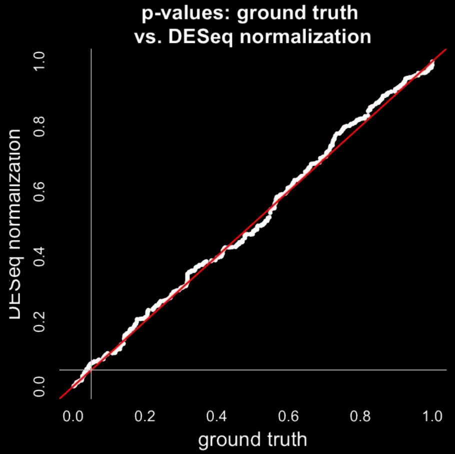
## Simulation to evaluate systematic RNA capture biases

Partial cell capture in imaged volume: inflated Type I error rate without normalization



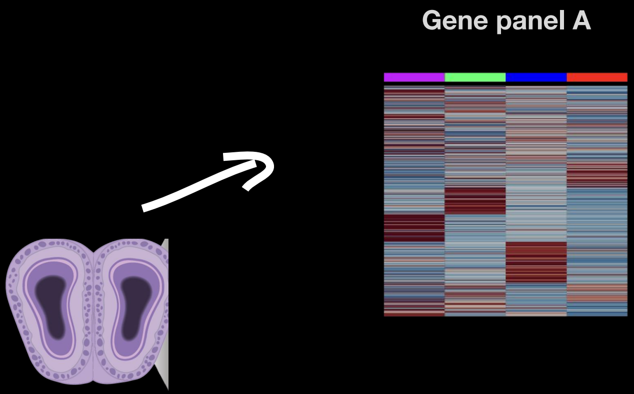

p-values: ground truth vs. no normalization


p-values: ground truth vs. library size normalization


p-values: ground truth vs. DESeq normalization

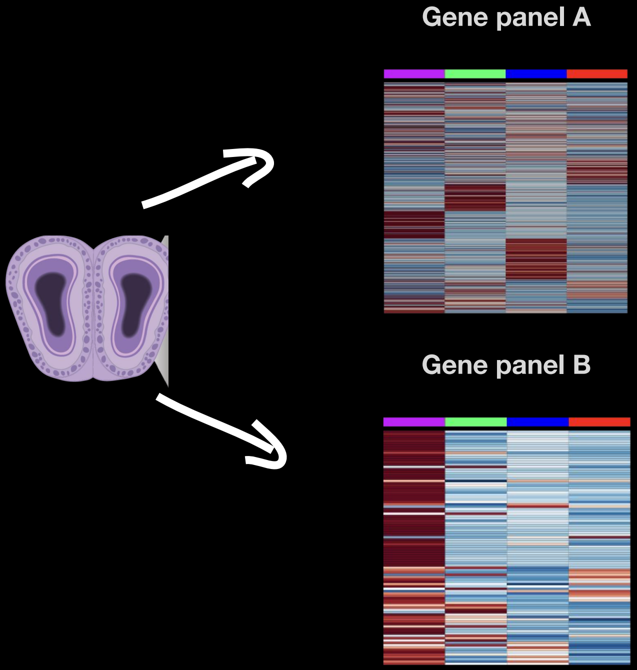
# Gene count normalization: evaluation criteria

## Robustness of downstream analyses with different gene panels



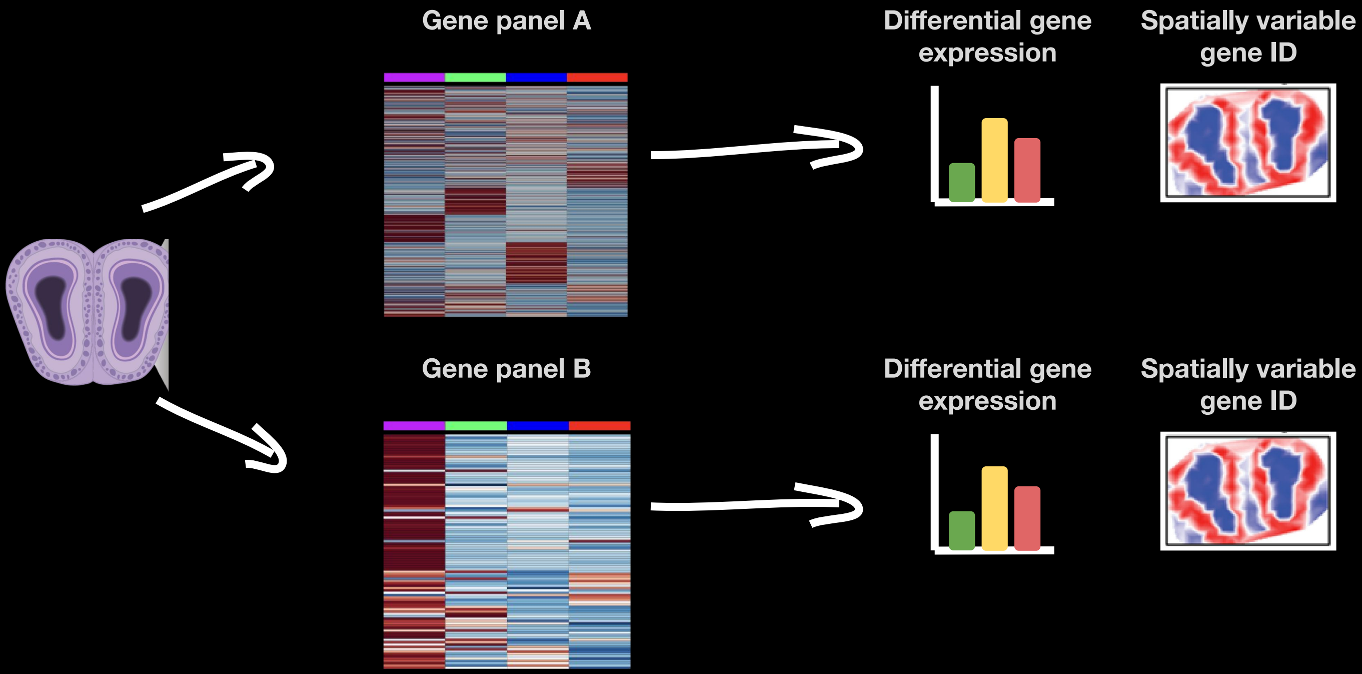
# Gene count normalization: evaluation criteria

## Robustness of downstream analyses with different gene panels



# Gene count normalization: evaluation criteria

## Robustness of downstream analyses with different gene panels



Gene panel A

Differential gene expression

Spatially variable gene ID

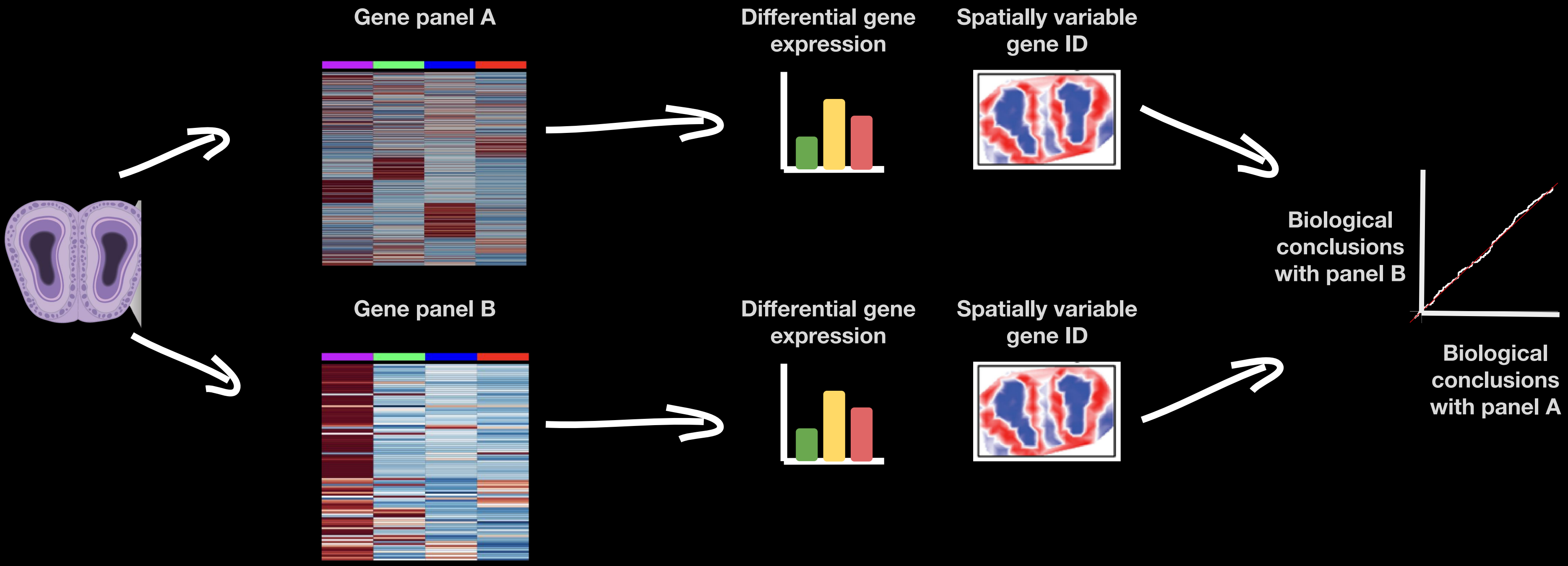
Gene panel B

Differential gene expression

Spatially variable gene ID

# Gene count normalization: evaluation criteria

## Robustness of downstream analyses with different gene panels



Gene panel A

Differential gene expression

Spatially variable gene ID

Biological conclusions with panel B

Gene panel B

Differential gene expression

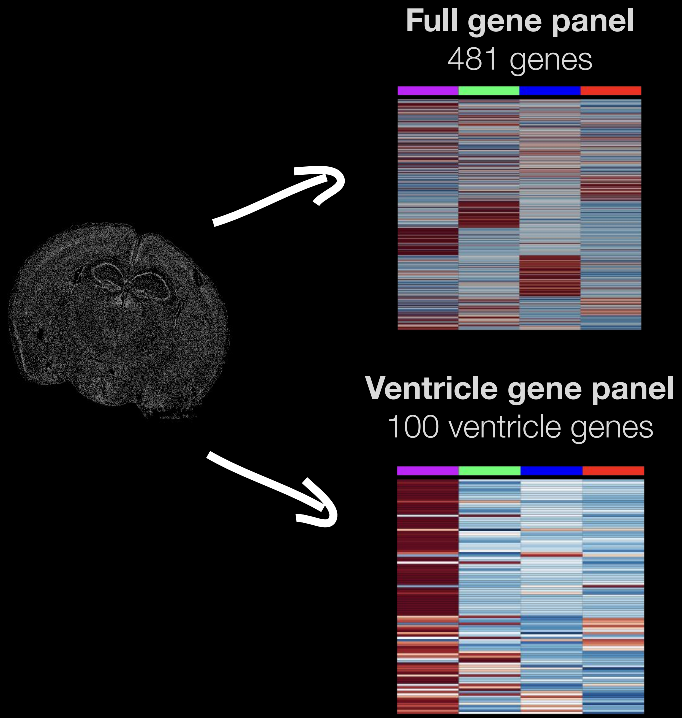
Spatially variable gene ID

Biological conclusions with panel A

# Count normalization in imaging-based SRT

- Goals of count normalization, current normalization approaches, evaluation criteria
- **Case study: ventricle gene panel in mouse brain MERFISH imSRT**
- Evaluating normalization across gene panels, technologies and tissues
- Recommendations

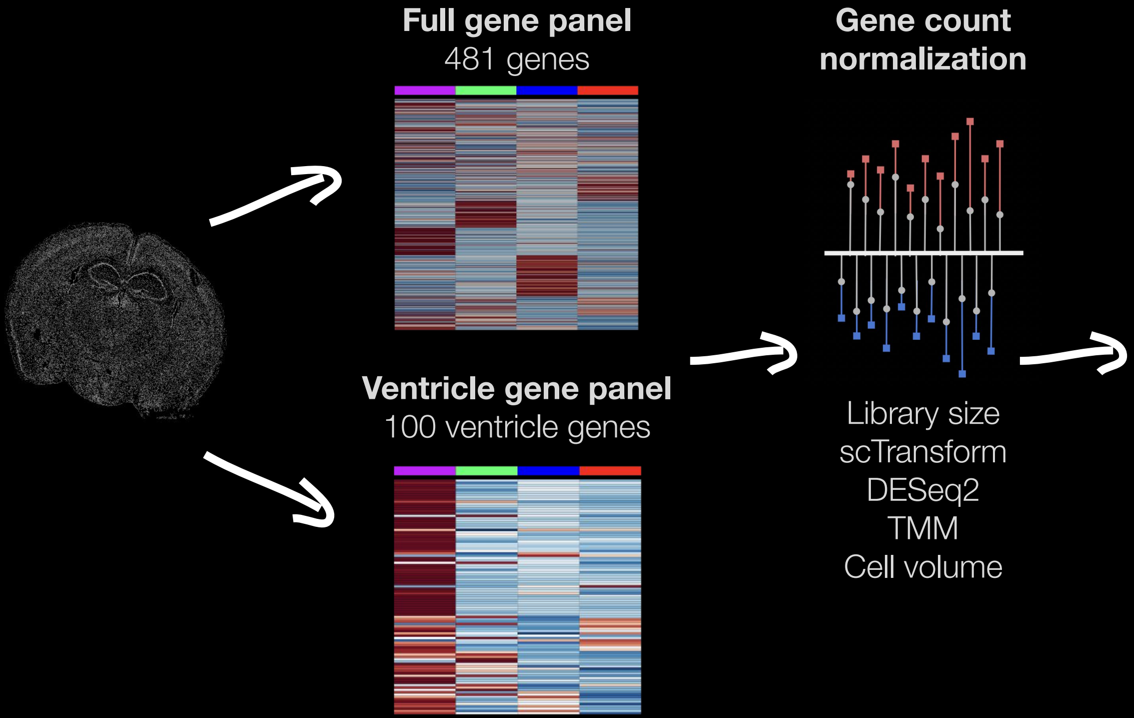
# Comparing normalization methods in mouse brain MERFISH with different gene panels



**Full gene panel**
481 genes

**Ventricle gene panel**
100 ventricle genes

# Comparing normalization methods in mouse brain MERFISH with different gene panels



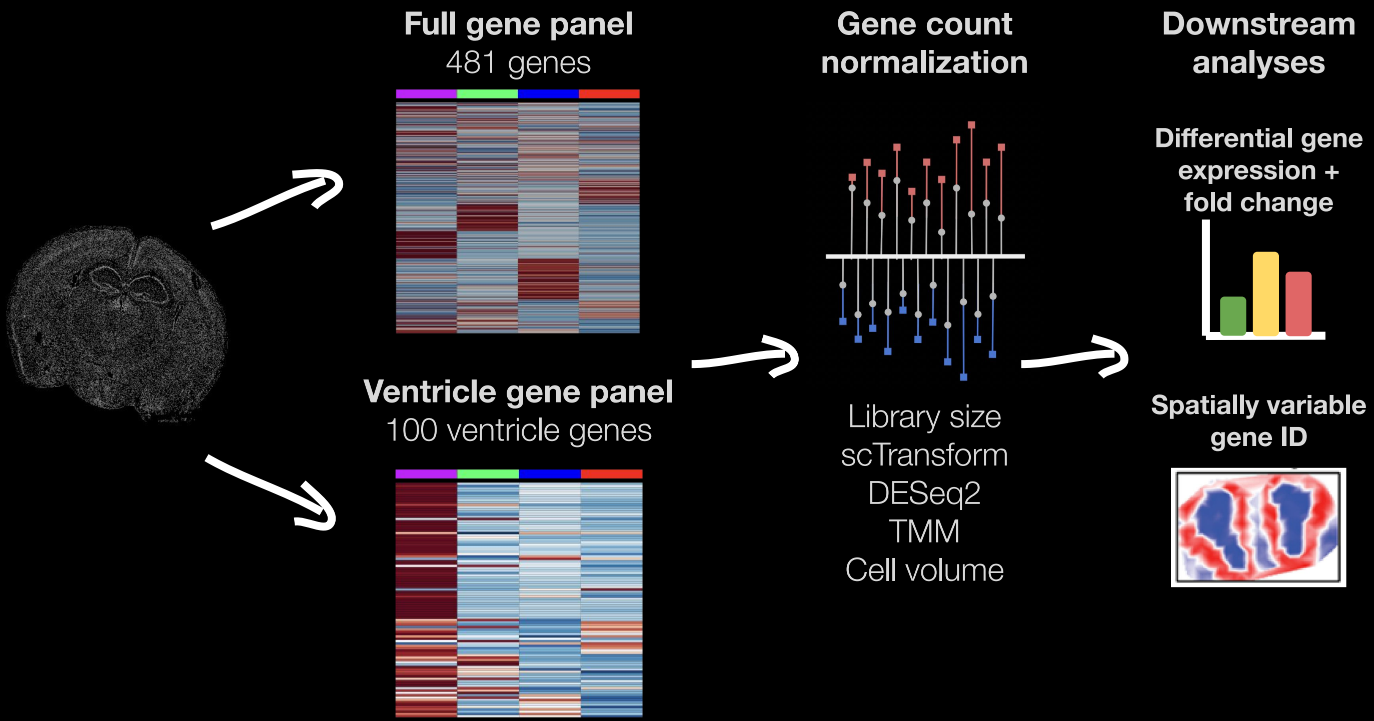
**Full gene panel**
481 genes

**Ventricle gene panel**
100 ventricle genes

**Gene count normalization**

Library size
scTransform
DESeq2
TMM
Cell volume

# Comparing normalization methods in mouse brain MERFISH with different gene panels



**Full gene panel**
481 genes

**Ventricle gene panel**
100 ventricle genes

**Gene count normalization**

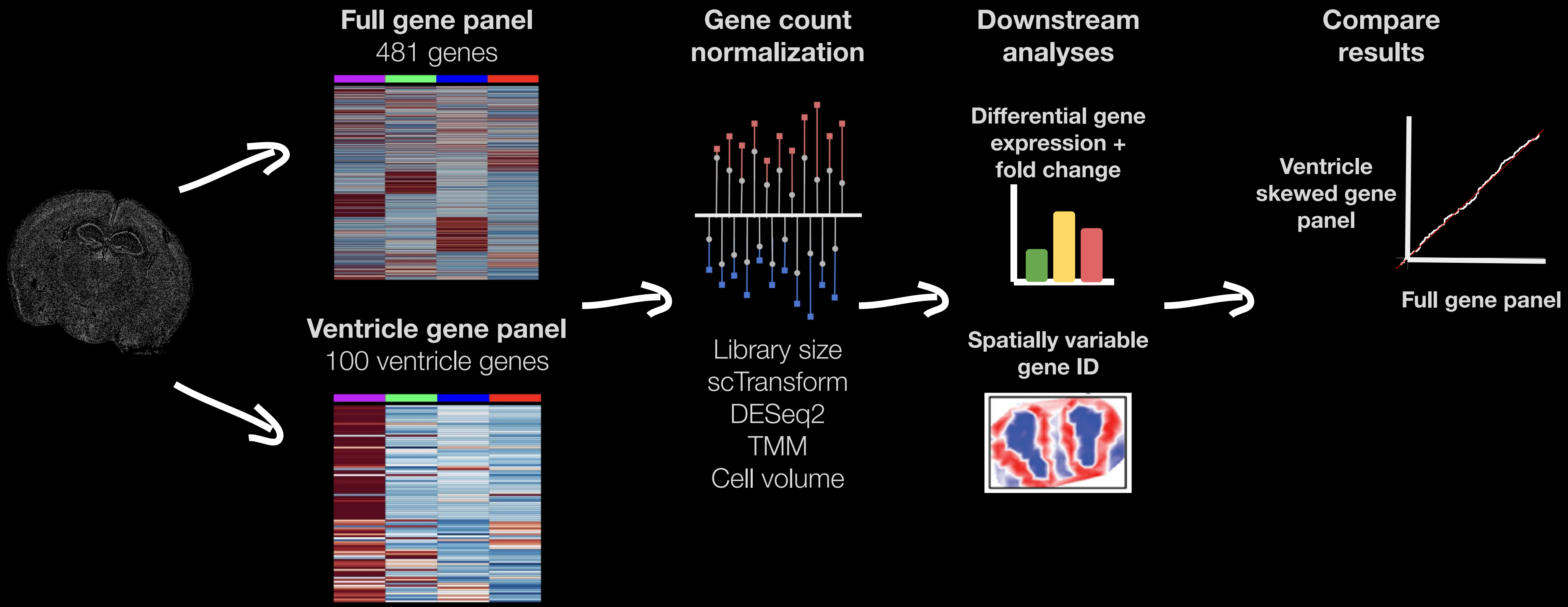
Library size
scTransform
DESeq2
TMM
Cell volume

**Downstream analyses**

**Differential gene expression + fold change**

**Spatially variable gene ID**

# Comparing normalization methods in mouse brain MERFISH with different gene panels



**Full gene panel**
481 genes

**Ventricle gene panel**
100 ventricle genes

**Gene count normalization**

Library size
scTransform
DESeq2
TMM
Cell volume

**Downstream analyses**

**Differential gene expression + fold change**

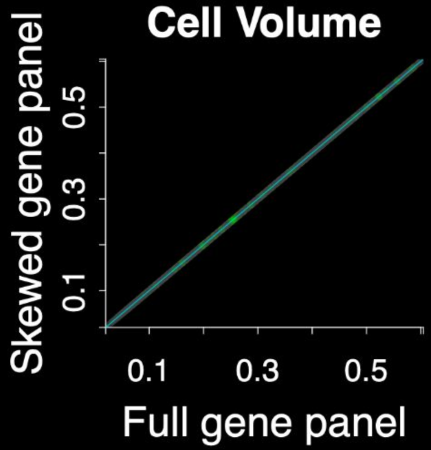
**Spatially variable gene ID**

**Compare results**

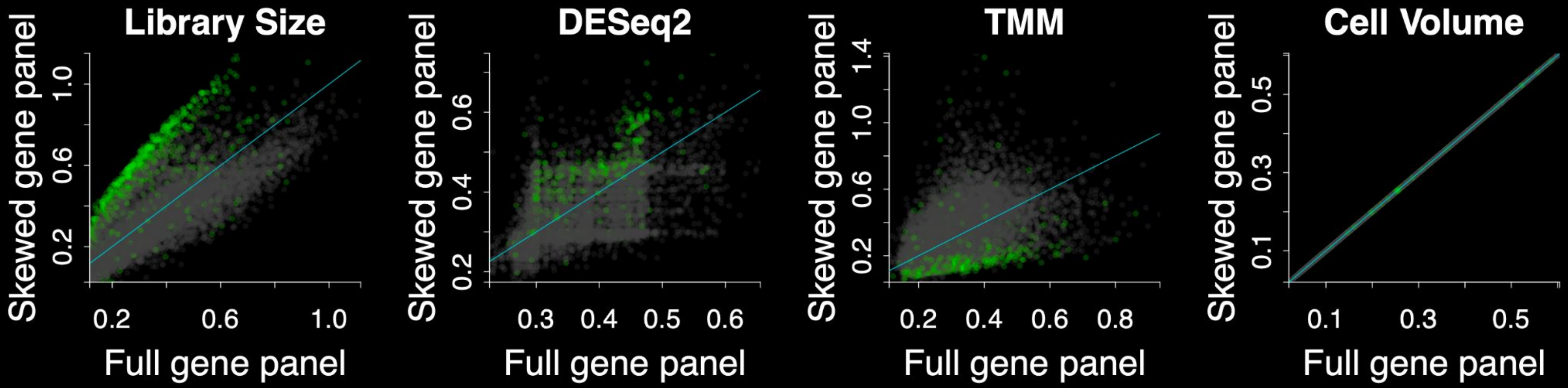
**Ventricle skewed gene panel**

**Full gene panel**

# Normalizing scaling factors: tissue region specific bias with ventricle gene panel

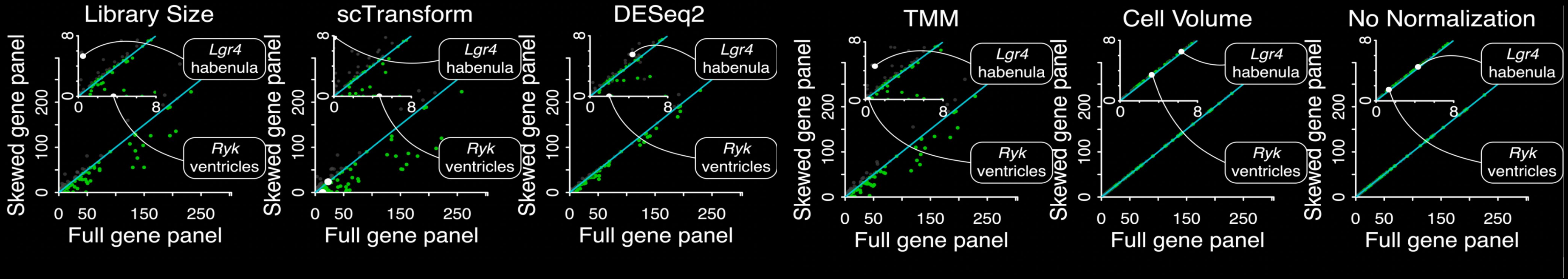


# Normalizing scaling factors: tissue region specific bias with ventricle gene panel



COUNT BASED NORMALIZATION

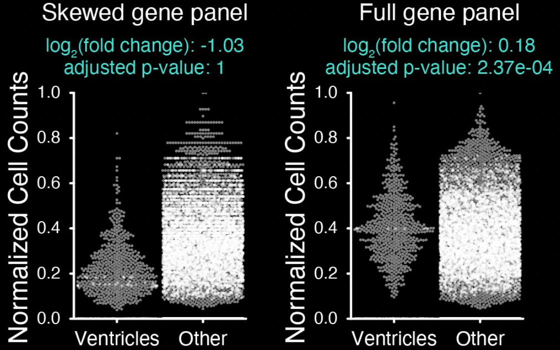
# Differential gene expression: tissue region specific bias with ventricle gene panel



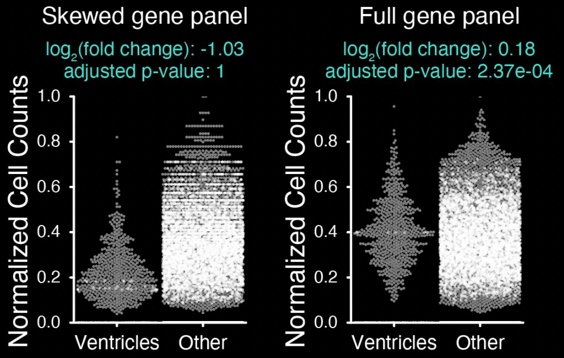
COUNT BASED NORMALIZATION

# Differential gene expression: tissue region specific bias with ventricle gene panel results in <u>false positives and negatives</u>

***Ryk*: DE false negative**



# Differential gene expression: tissue region specific bias with ventricle gene panel results in false positives and negatives



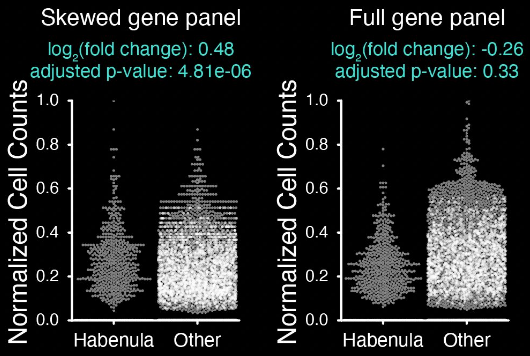
***Ryk*: DE false negative**

Skewed gene panel

log$_2$(fold change): -1.03
adjusted p-value: 1

Full gene panel

log$_2$(fold change): 0.18
adjusted p-value: 2.37e-04



***Lgr4*: DE false positive**

Skewed gene panel

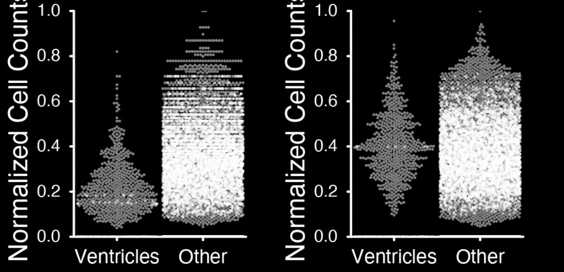
log$_2$(fold change): 0.48
adjusted p-value: 4.81e-06

Full gene panel

log$_2$(fold change): -0.26
adjusted p-value: 0.33

# Differential gene expression: tissue region specific bias with ventricle gene panel results in false positives and negatives

## *Ryk*: DE false negative

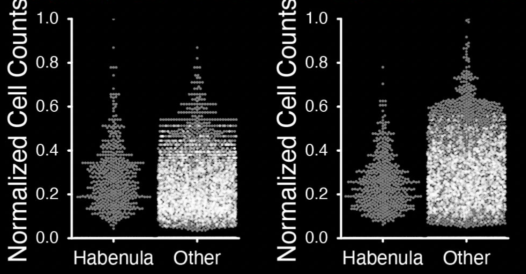

Skewed gene panel

log$_2$(fold change): -1.03
adjusted p-value: 1

Full gene panel

log$_2$(fold change): 0.18
adjusted p-value: 2.37e-04

## *Lgr4*: DE false positive

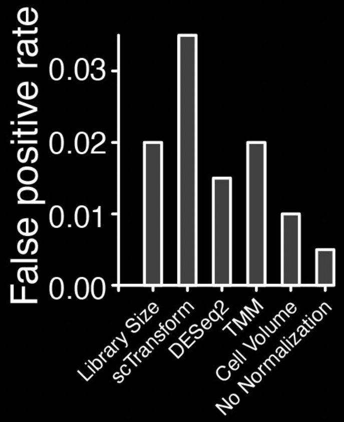

Skewed gene panel

log$_2$(fold change): 0.48
adjusted p-value: 4.81e-06

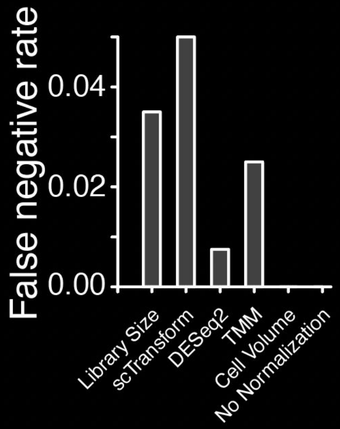
Full gene panel

log$_2$(fold change): -0.26
adjusted p-value: 0.33

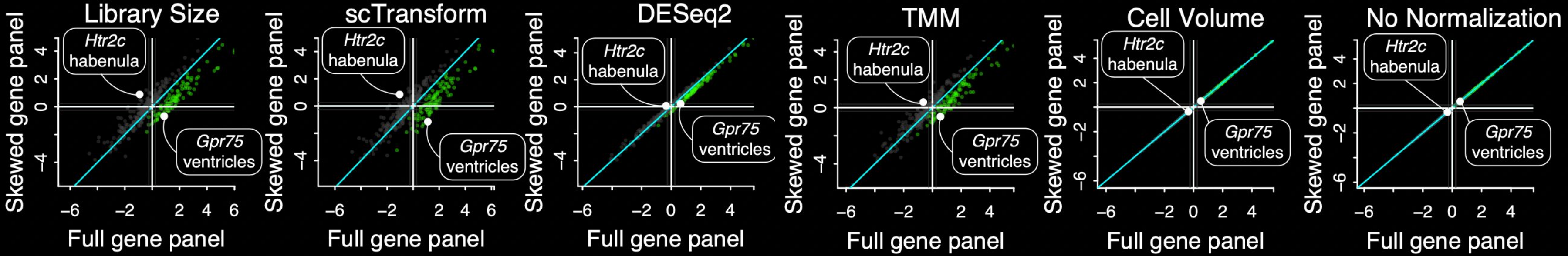
## DE false negative rate



## DE false positive rate



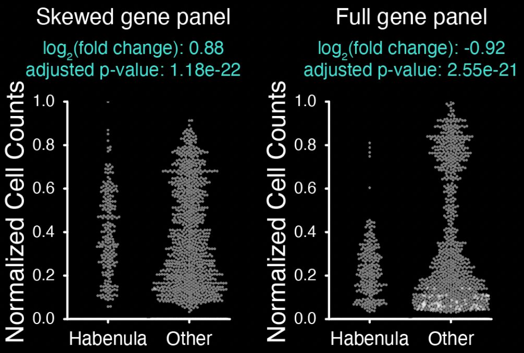
# Differential gene expression: tissue region specific bias with ventricle gene panel



COUNT BASED NORMALIZATION

# Differential gene expression: tissue region specific bias with ventricle gene panel results in <u>switched gene fold changes</u>

***Htr2c*: FC switched positive**



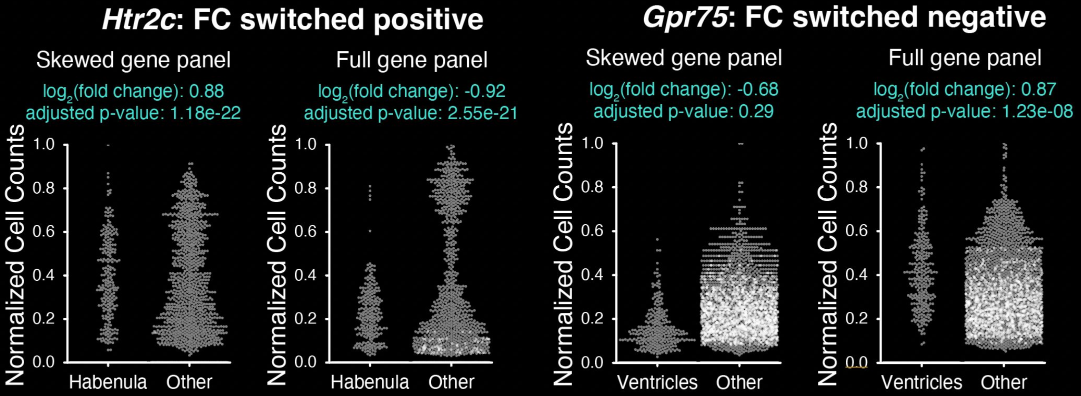
Skewed gene panel

$\log_2$(fold change): 0.88
adjusted p-value: 1.18e-22

Full gene panel

$\log_2$(fold change): -0.92
adjusted p-value: 2.55e-21

# Differential gene expression: tissue region specific bias with ventricle gene panel results in <u>switched gene fold changes</u>



***Htr2c*: FC switched positive**

Skewed gene panel
$\log_2$(fold change): 0.88
adjusted p-value: 1.18e-22

Full gene panel
$\log_2$(fold change): -0.92
adjusted p-value: 2.55e-21

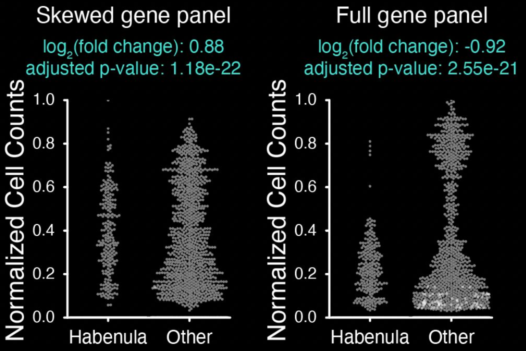
***Gpr75*: FC switched negative**

Skewed gene panel
$\log_2$(fold change): -0.68
adjusted p-value: 0.29

Full gene panel
$\log_2$(fold change): 0.87
adjusted p-value: 1.23e-08

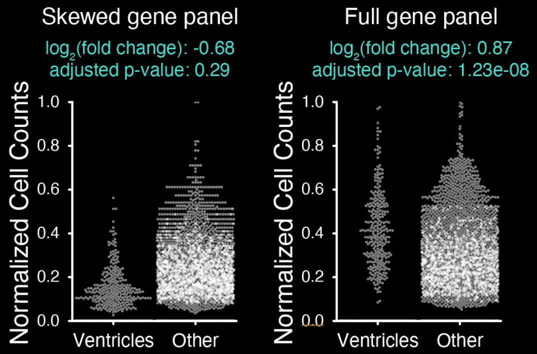
# Differential gene expression: tissue region specific bias with ventricle gene panel results in switched gene fold changes

**_Htr2c_: FC switched positive**


Skewed gene panel
$\log_2$(fold change): 0.88
adjusted p-value: 1.18e-22

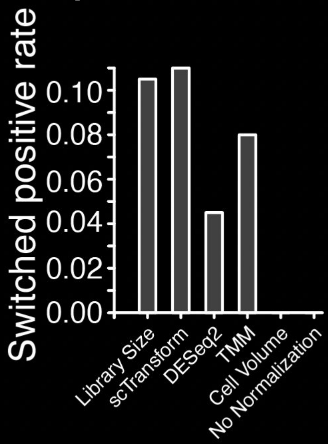
Full gene panel
$\log_2$(fold change): -0.92
adjusted p-value: 2.55e-21

**_Gpr75_: FC switched negative**

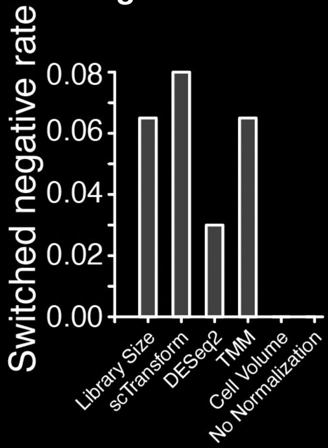

Skewed gene panel
$\log_2$(fold change): -0.68
adjusted p-value: 0.29

Full gene panel
$\log_2$(fold change): 0.87
adjusted p-value: 1.23e-08

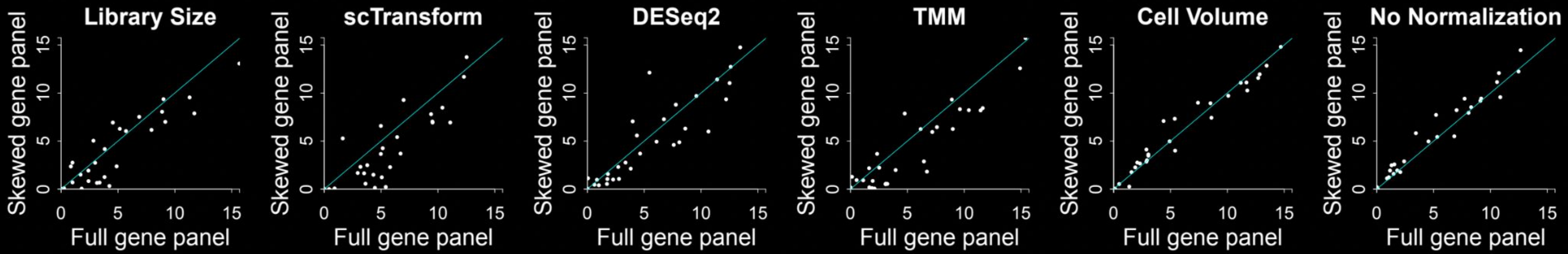
**FC switched positive rate**



**FC switched negative rate**



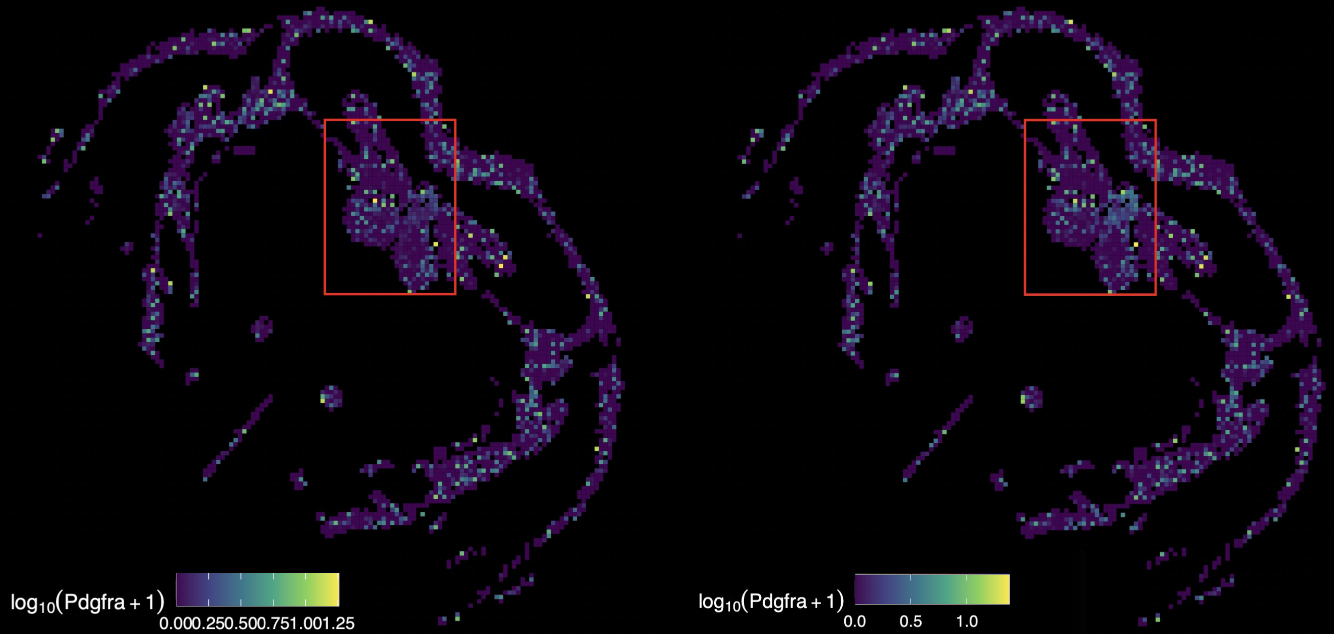
# Spatially variable gene identification with ventricle gene panel results in <u>false positives and negatives</u>



COUNT BASED NORMALIZATION

# Spatially variable gene identification with ventricle gene panel results in <u>false positives and negatives</u>

**Pdgfra, library size normalization**



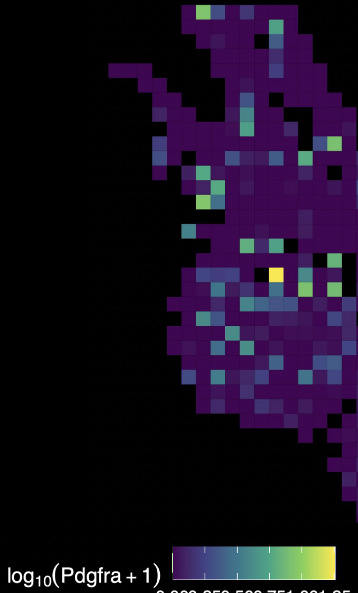
**Ventricle gene panel**

**Full gene panel**

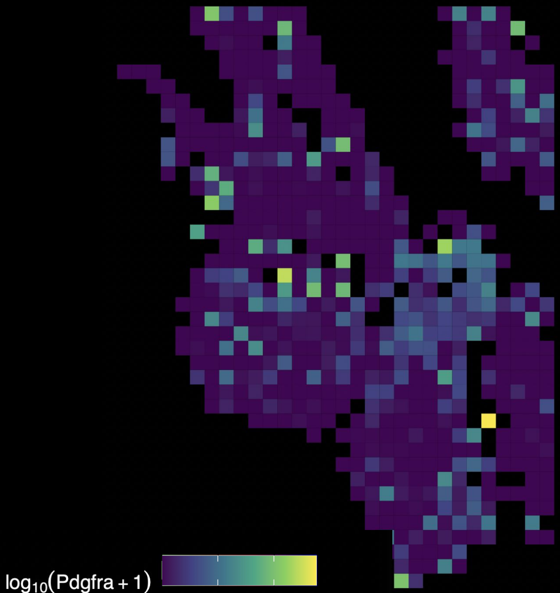
# Spatially variable gene identification with ventricle gene panel results in <u>false positives and negatives</u>

**Pdgfra, library size normalization**

**Ventricle gene panel**



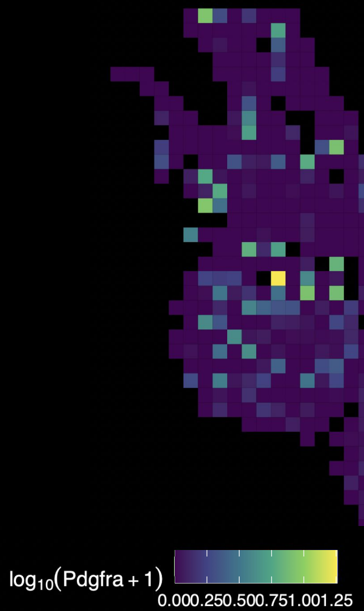
**Full gene panel**



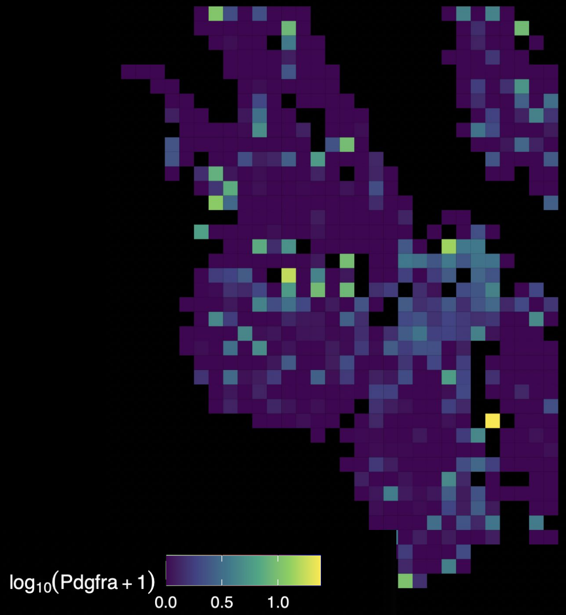
# Spatially variable gene identification with ventricle gene panel results in <u>false positives and negatives</u>

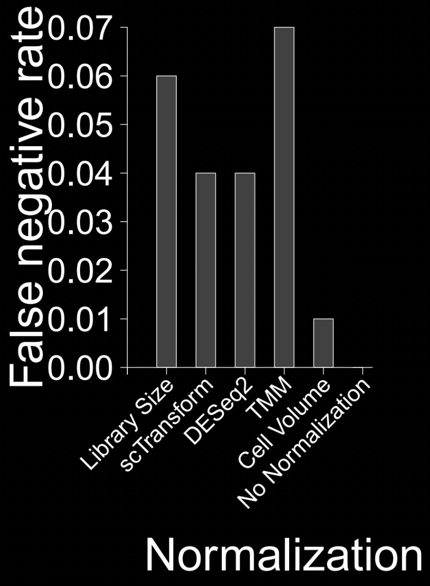
**Pdgfra, library size normalization**

**Ventricle gene panel**



**Full gene panel**

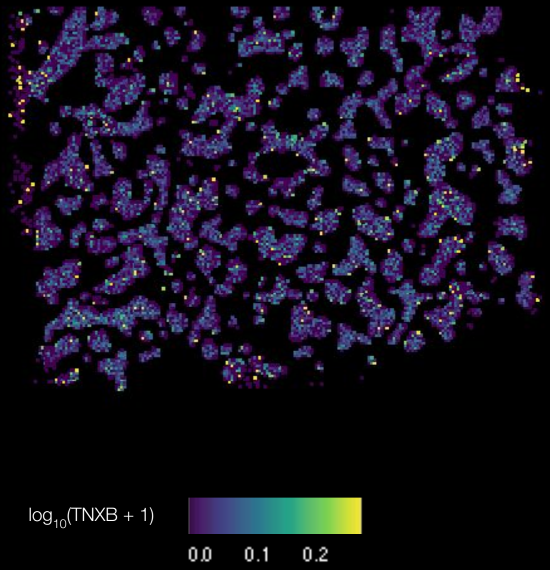




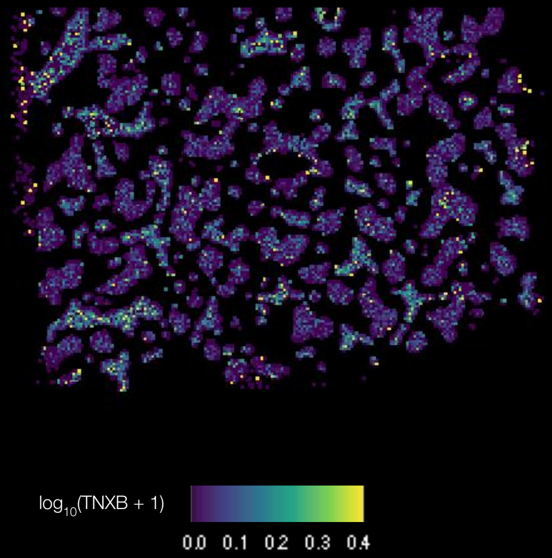
# Spatially variable gene identification with ventricle gene panel results in <u>false negatives</u>

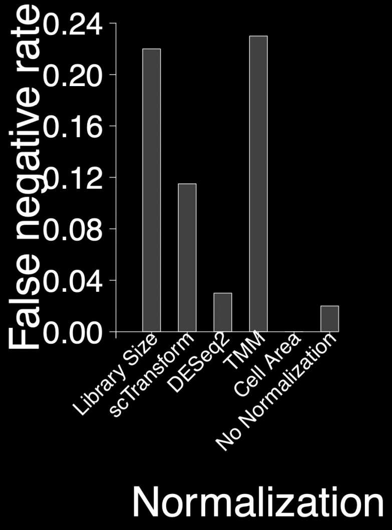
**CosMx human liver: TNXB, library size normalization**

**Zone 1 gene panel**

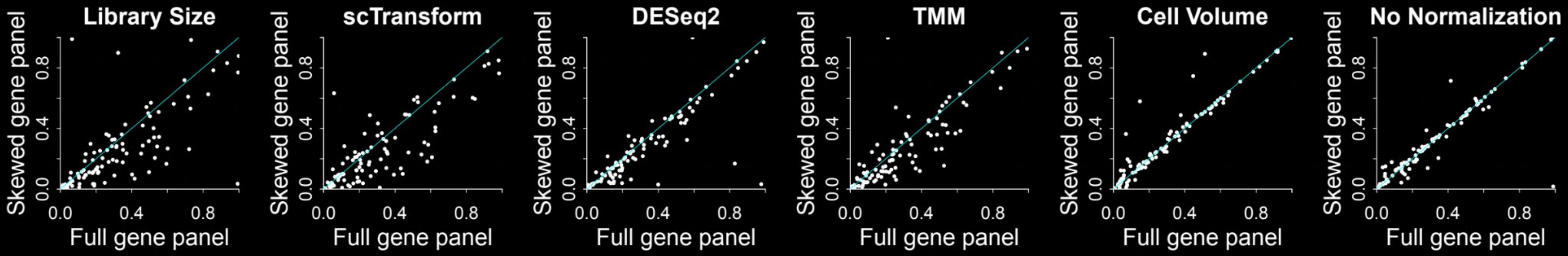


**Full gene panel**





# Spatially variable gene identification with ventricle gene panel results in <u>mis-estimates of spatial contribution to variance</u>

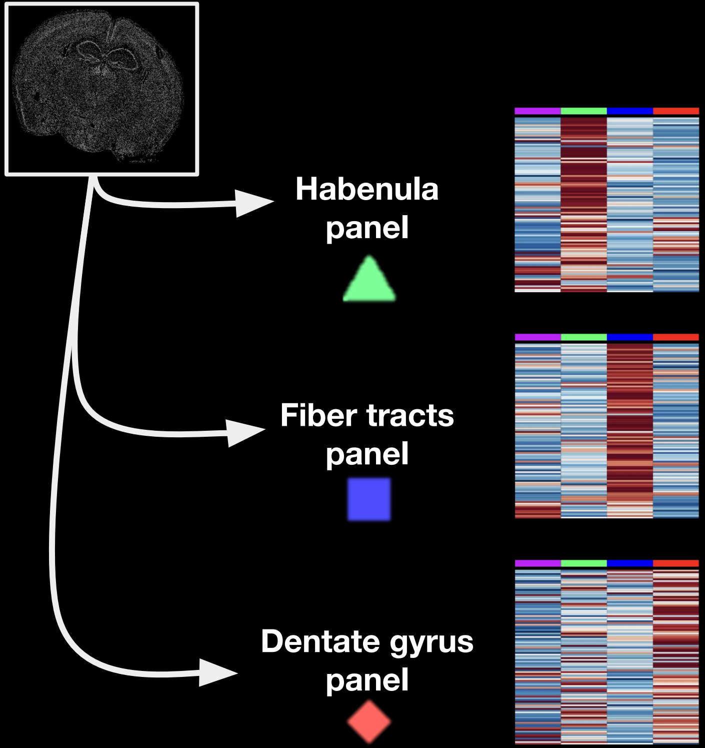


**COUNT BASED NORMALIZATION**

# Count normalization in imaging-based SRT

- ☐ Goals of count normalization, current normalization approaches, evaluation criteria
- ☐ Case study: ventricle gene panel in mouse brain MERFISH imSRT
- ☐ Evaluating normalization across gene panels, technologies and tissues
- ☐ Recommendations

# Region-specific biases in normalized gene expression generalize across <u>gene panels</u>



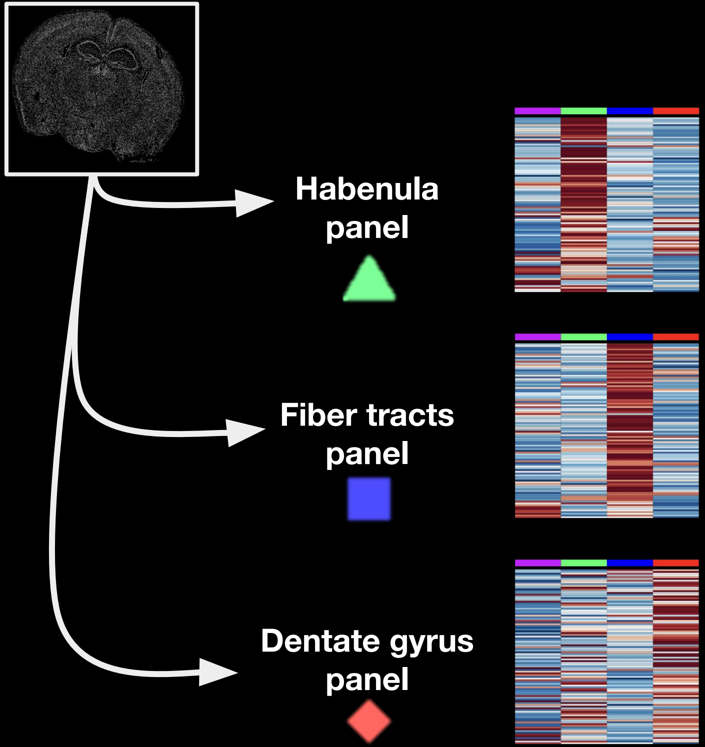
Habenula panel ▲

Fiber tracts panel ■

Dentate gyrus panel ◆

● Ventricle panel

# Region-specific biases in normalized gene expression generalize across <u>gene panels</u>



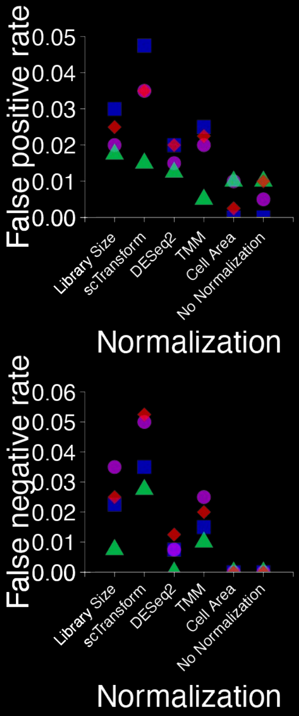
**Habenula panel** ▲

**Fiber tracts panel** ■

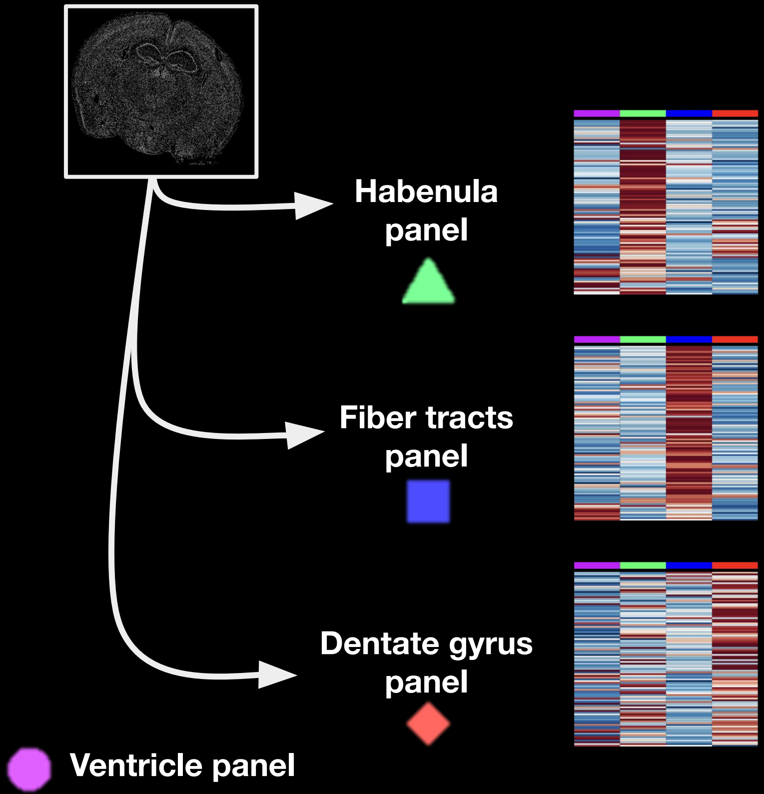
**Dentate gyrus panel** ◆

● **Ventricle panel**

## DE error rates



# Region-specific biases in normalized gene expression generalize across <u>gene panels</u>

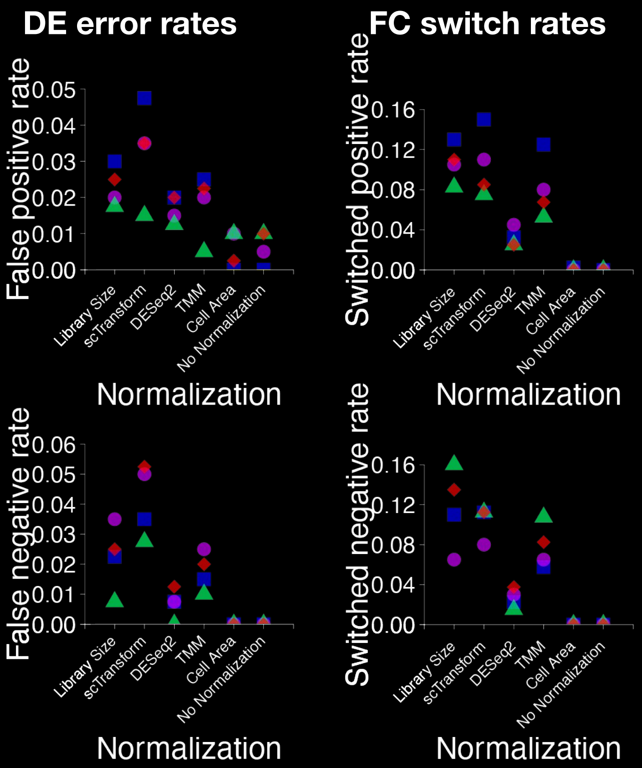


**Habenula panel** ▲

**Fiber tracts panel** ■

**Dentate gyrus panel** ◆

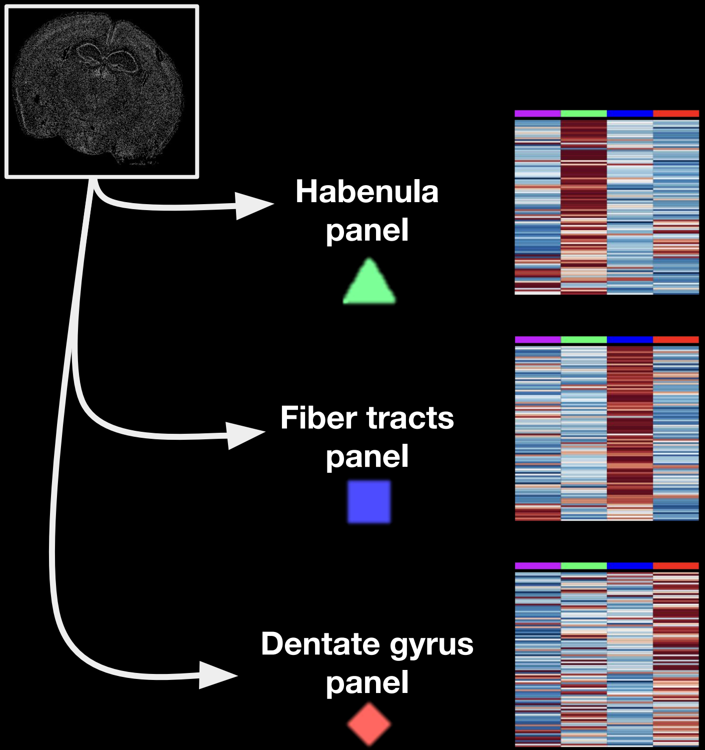
● **Ventricle panel**



DE error rates

FC switch rates

# Region-specific biases in normalized gene expression generalize across <u>gene panels</u>

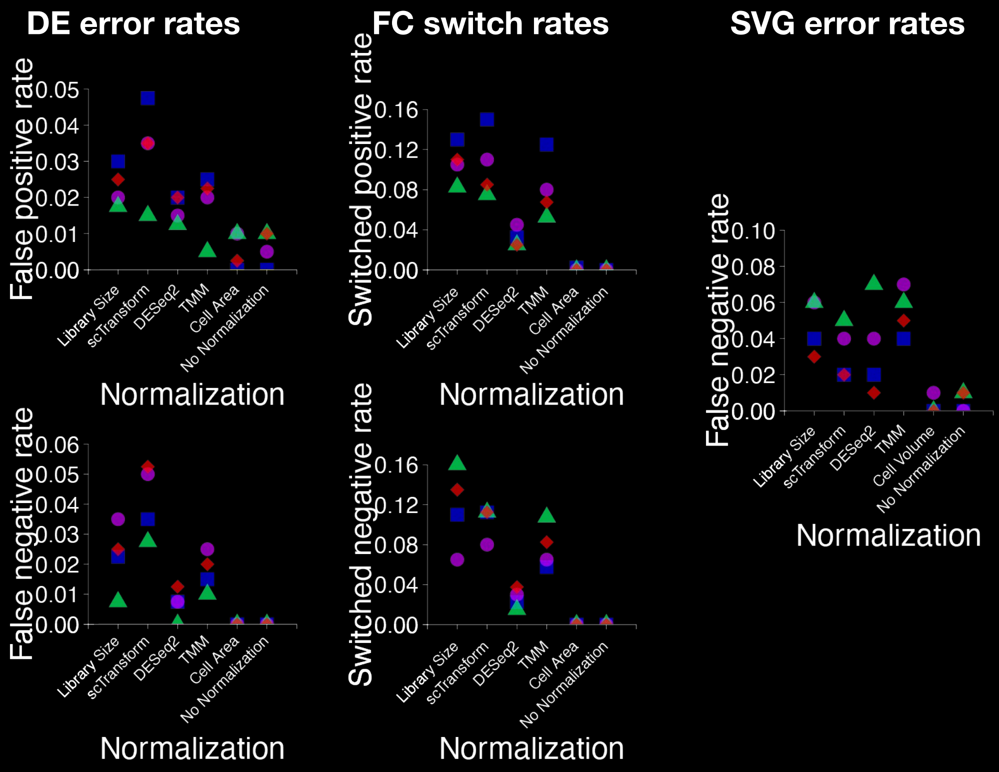


**Habenula panel** ▲

**Fiber tracts panel** ■

**Dentate gyrus panel** ◆

● **Ventricle panel**



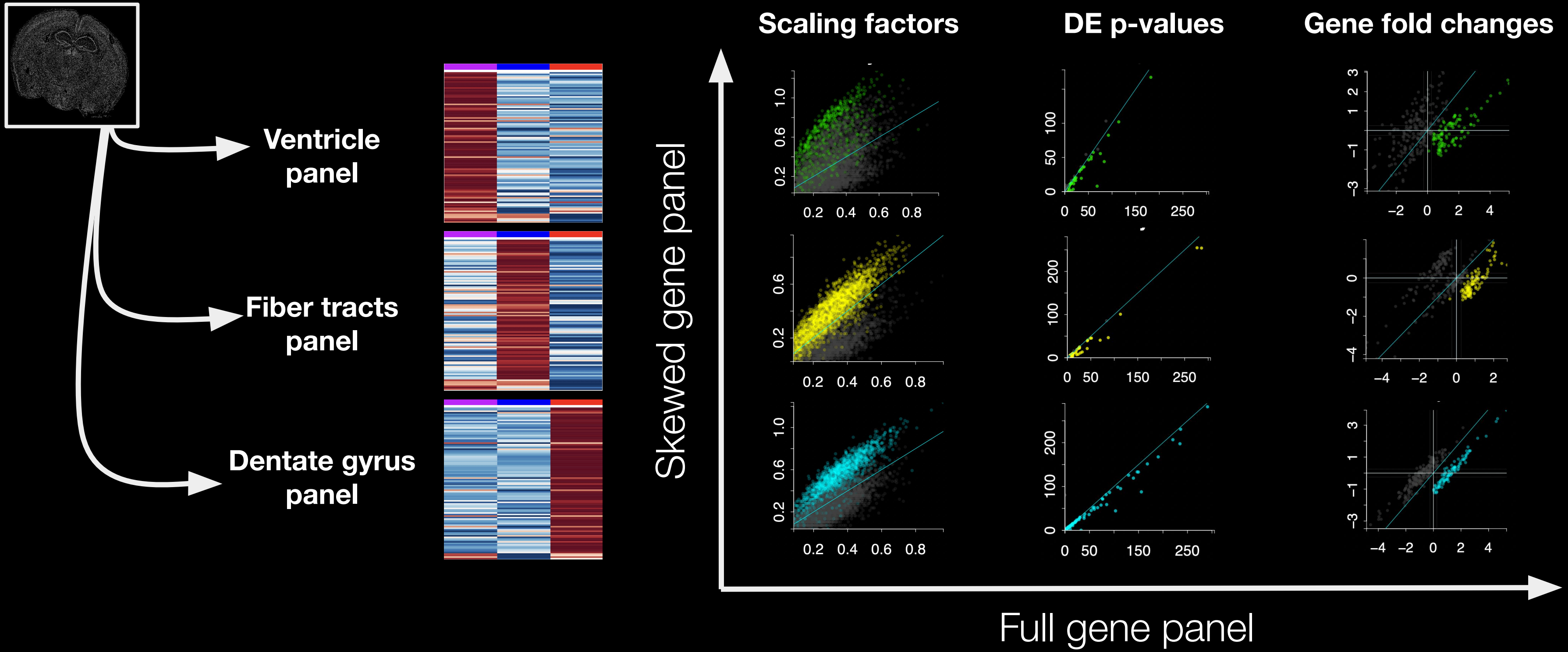
**DE error rates**

**FC switch rates**

**SVG error rates**

# Region-specific biases in normalized gene expression generalize across imSRT technologies

## Library size normalization with STARmapPLUS in mouse brain



**Ventricle panel**

**Fiber tracts panel**

**Dentate gyrus panel**

**Scaling factors**

**DE p-values**

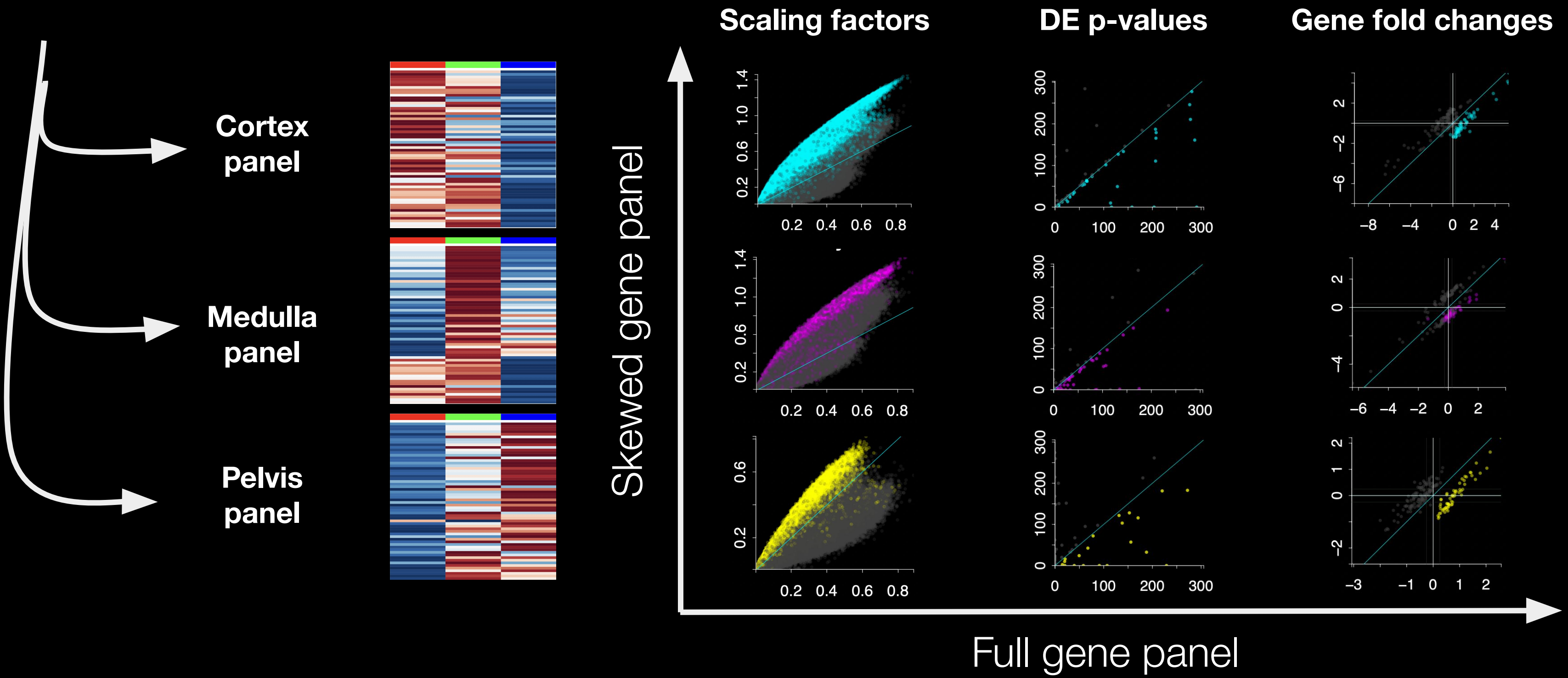
**Gene fold changes**

Skewed gene panel

Full gene panel

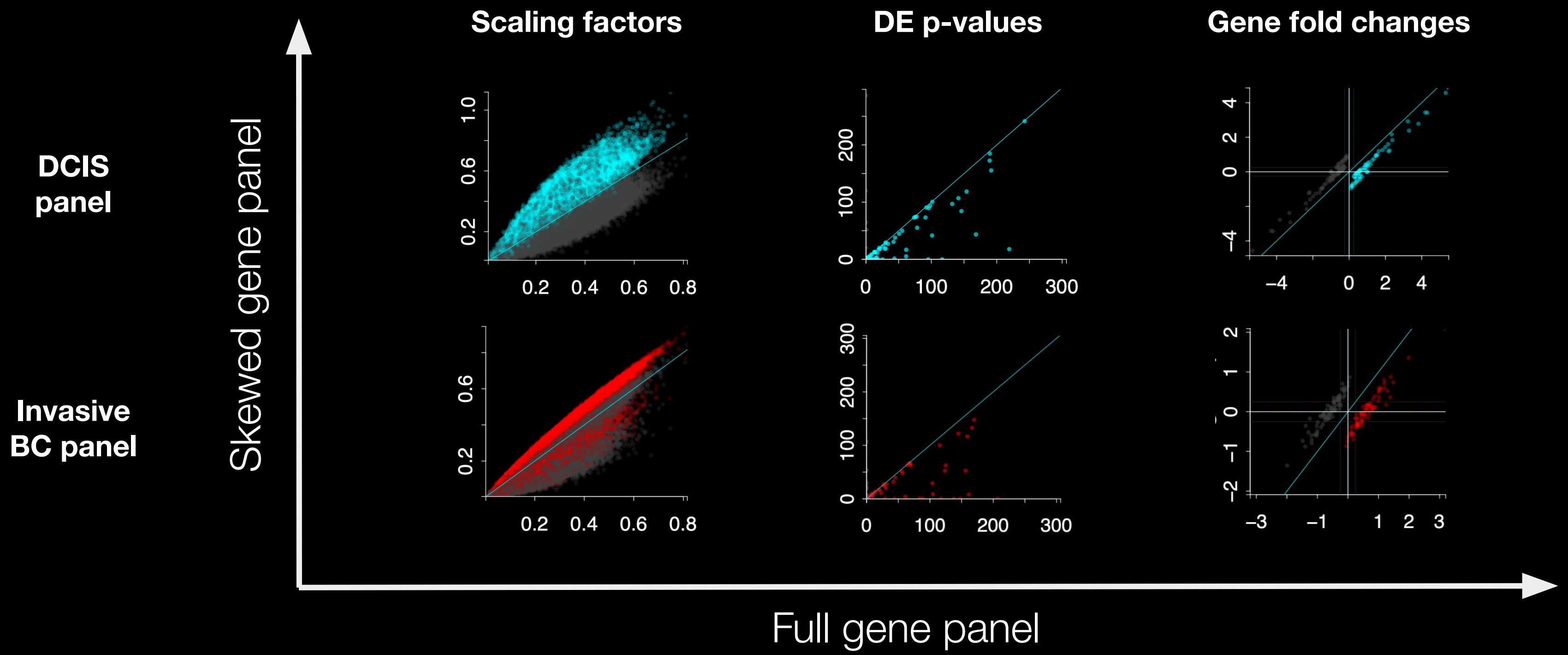
# Region-specific biases in normalized gene expression generalize across <u>tissue types</u>

## Library size normalization with seqFISH in mouse kidney



# Region-specific biases in normalized gene expression generalize across <u>tissue types</u>

## Library size normalization with 10X Xenium in human breast cancer



Scaling factors

DE p-values

Gene fold changes

**DCIS panel**

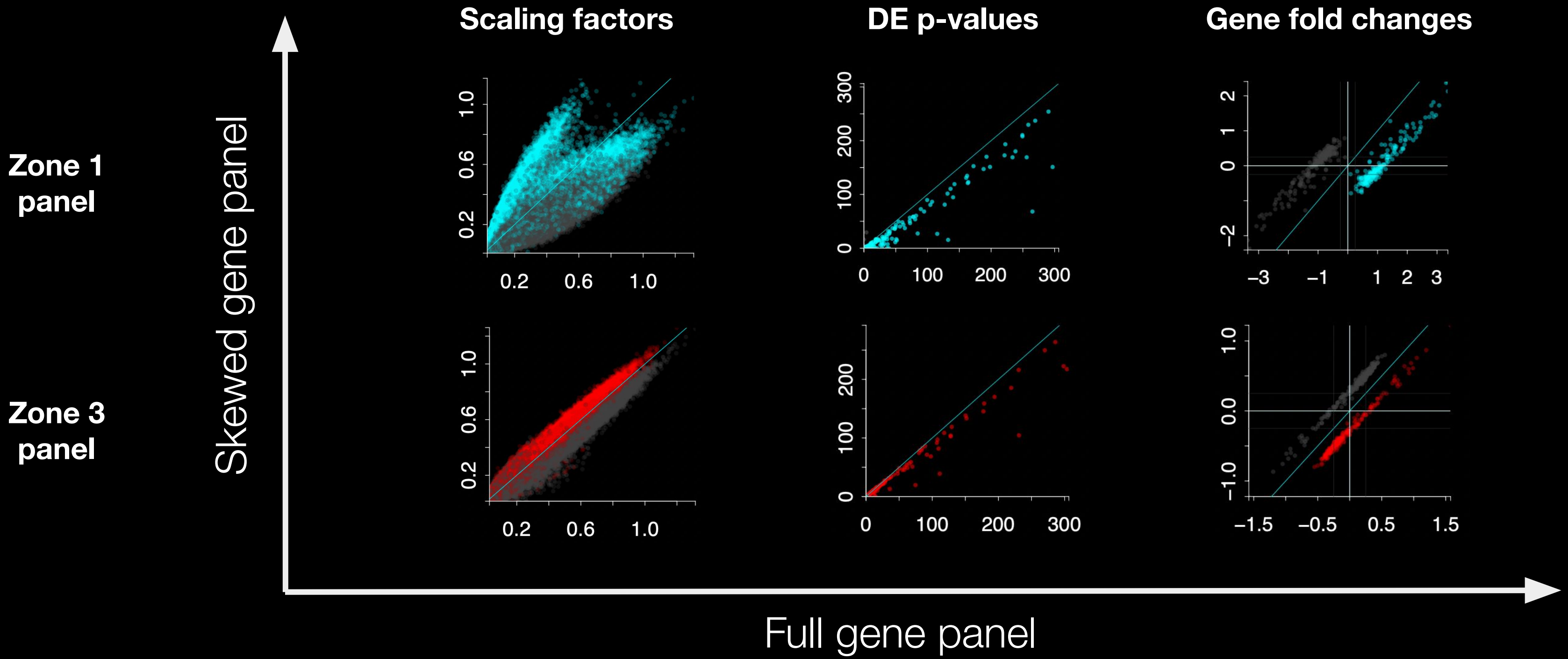
**Invasive BC panel**

Skewed gene panel

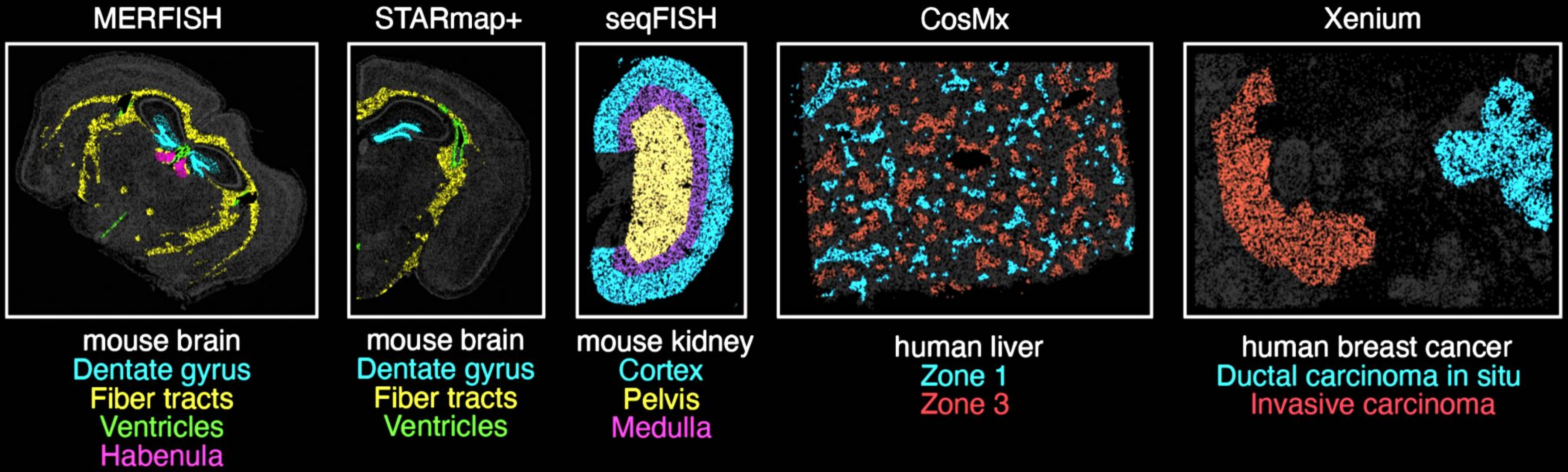
Full gene panel

# Region-specific biases in normalized gene expression generalize across tissue types

Library size normalization with CosMx in human liver

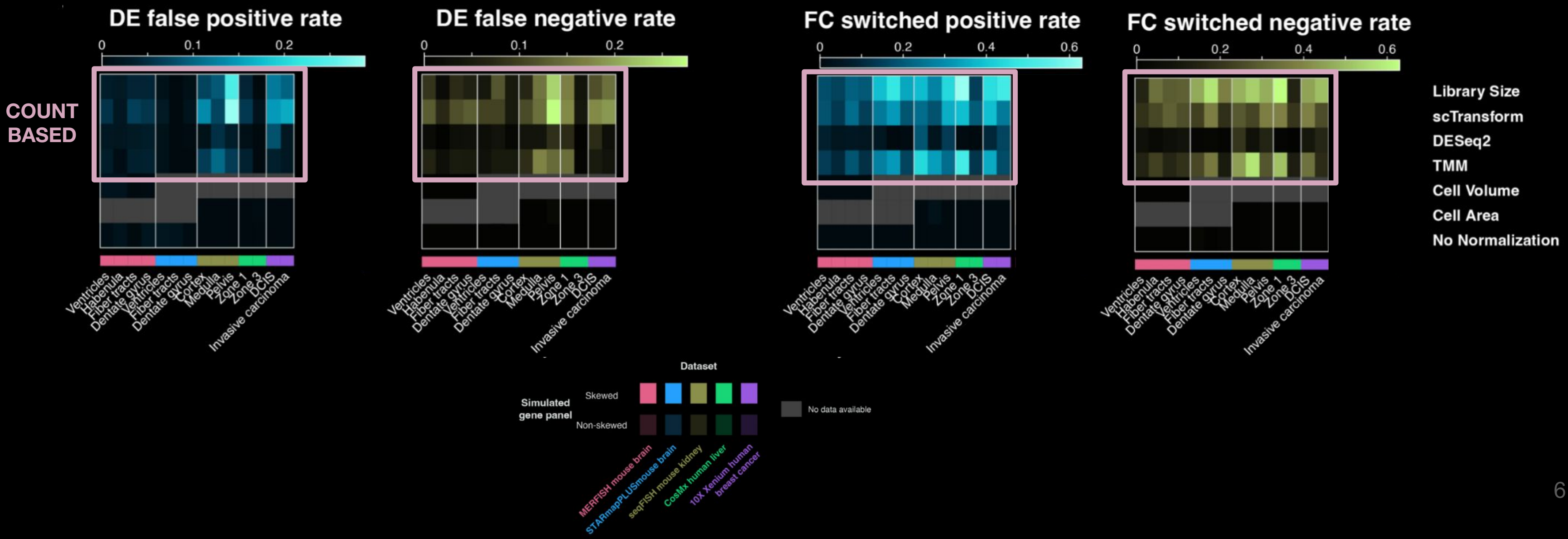


# Region-specific biases in gene differential expression and fold change generalize across <u>gene panels</u>, <u>technologies</u>, and <u>tissues</u>

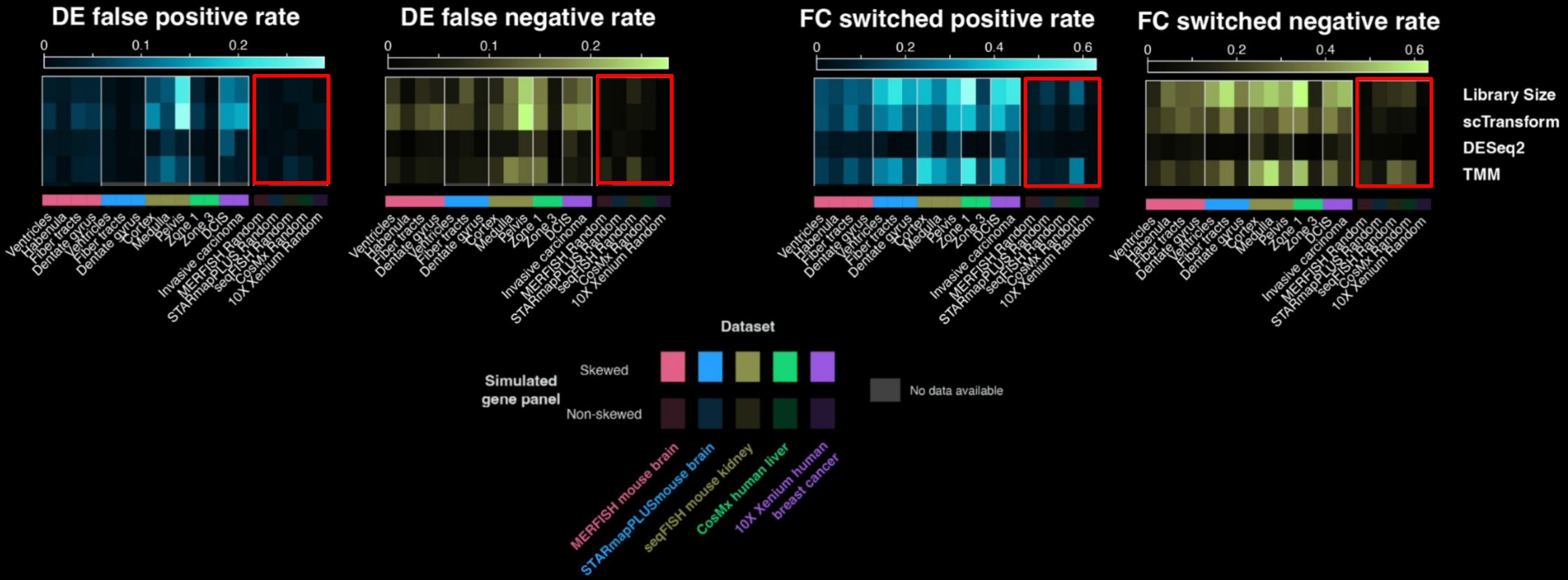


# Region-specific biases in gene differential expression and fold change generalize across gene panels, technologies, and tissues

**Differential expression error rates** **Fold change switch rates**



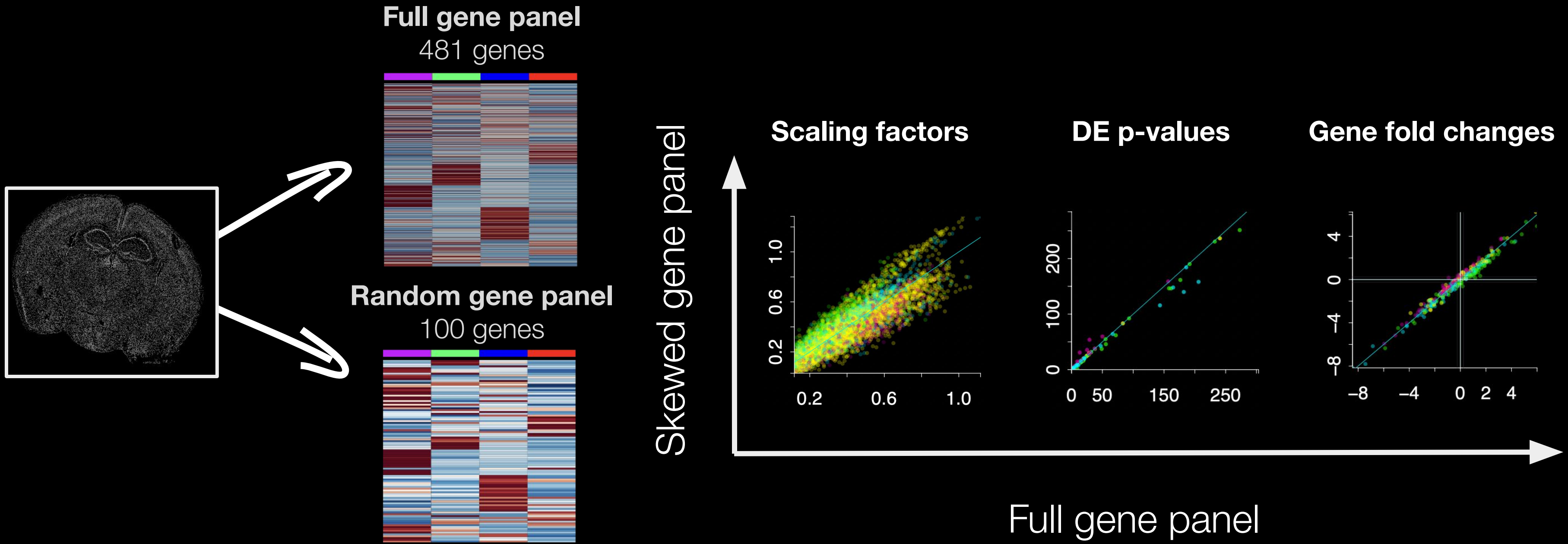
# Region-specific biases can be mitigated with more representative gene panels



# Count normalization in imaging-based SRT

- ☐ Goals of count normalization, current normalization approaches, evaluation criteria
- ☐ Case study: ventricle gene panel in mouse brain MERFISH imSRT
- ☐ Evaluating normalization across gene panels, technologies and tissues
- ☐ **Recommendations**

# Region-specific biases can be mitigated with more representative gene panels



**Full gene panel**
481 genes

**Random gene panel**
100 genes

**Scaling factors**

**DE p-values**

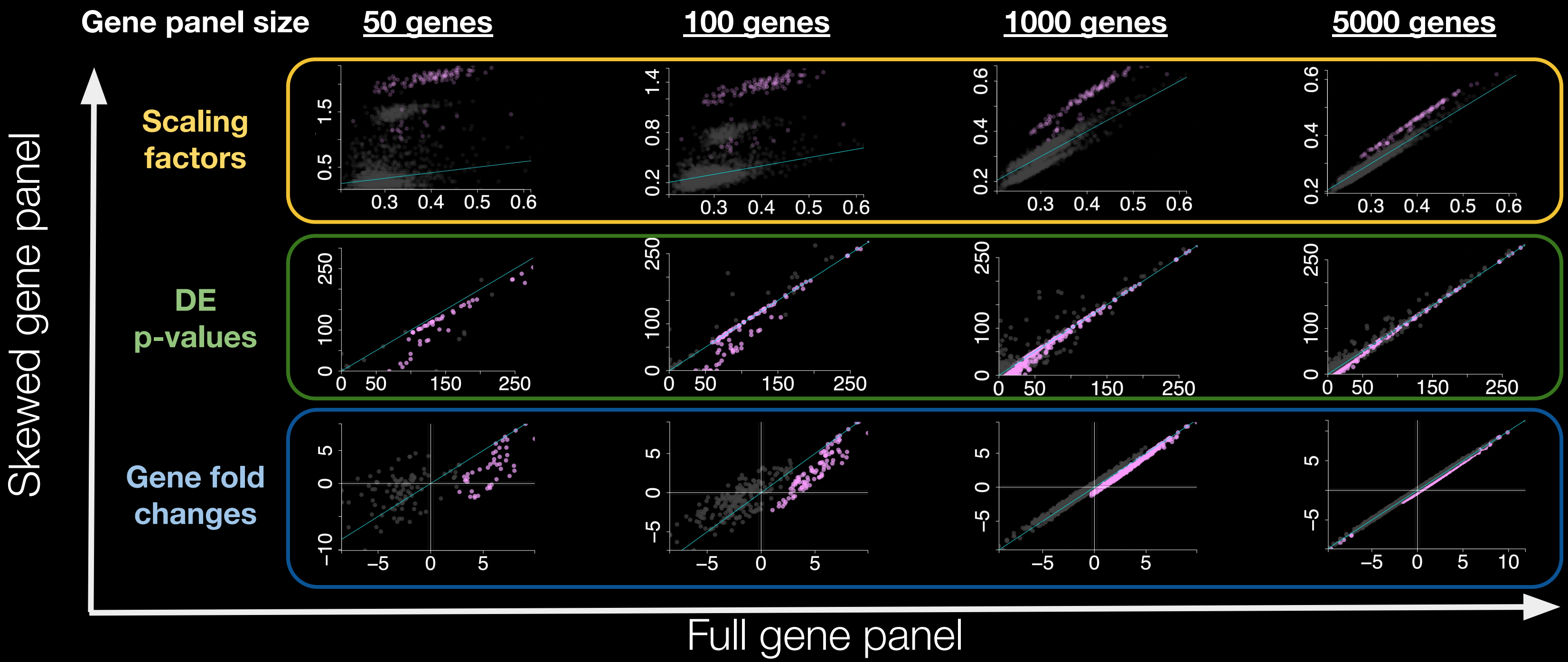
**Gene fold changes**

Skewed gene panel

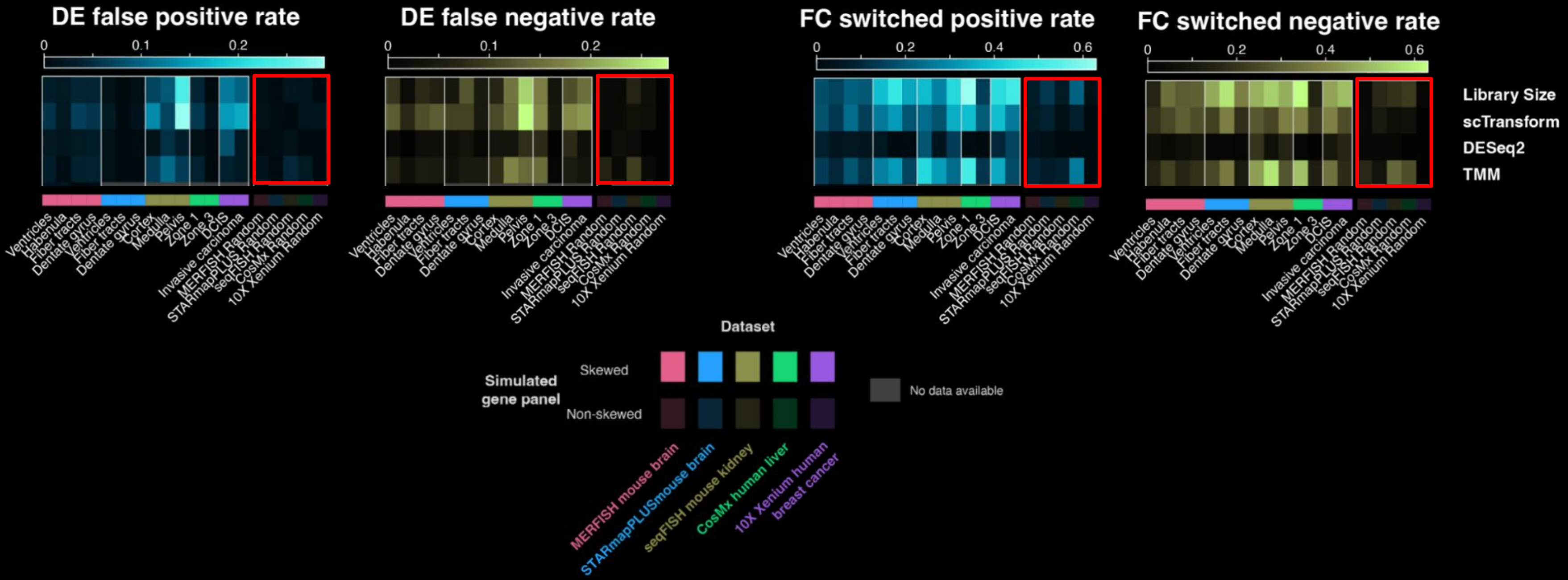
Full gene panel

# Region-specific biases can be mitigated with larger gene panels

Library size normalization with monocyte-skewed gene panels simulated from sorted PBMS scRNA-seq



# Region-specific biases can be mitigated with more representative gene panels



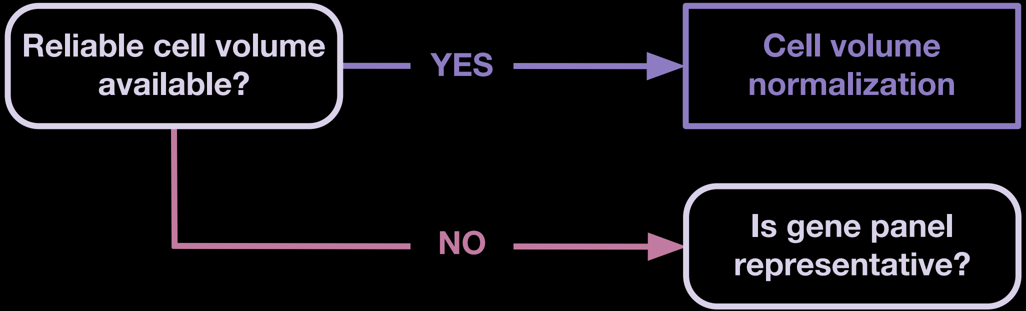
# Recommendations

Reliable cell volume available?

YES

Cell volume normalization

# Recommendations



Reliable cell volume available?

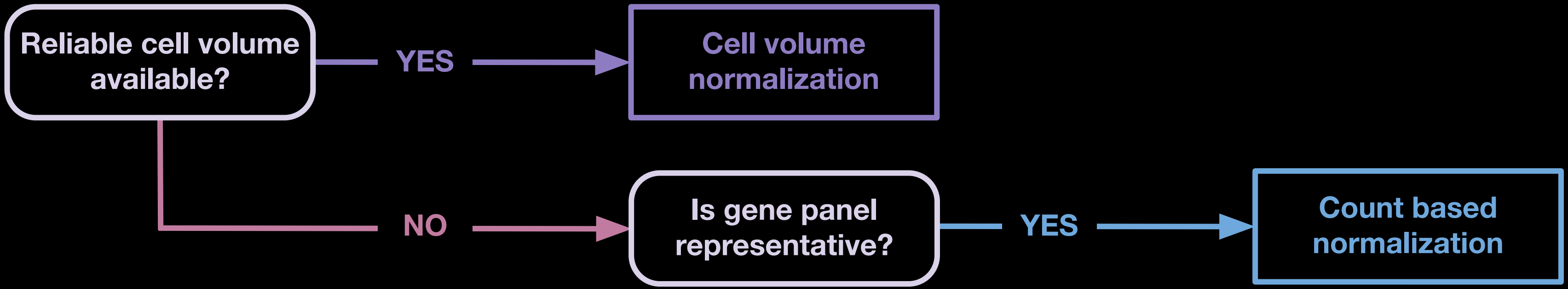
YES

Cell volume normalization

NO

Is gene panel representative?

# Recommendations



Reliable cell volume available?

YES

Cell volume normalization

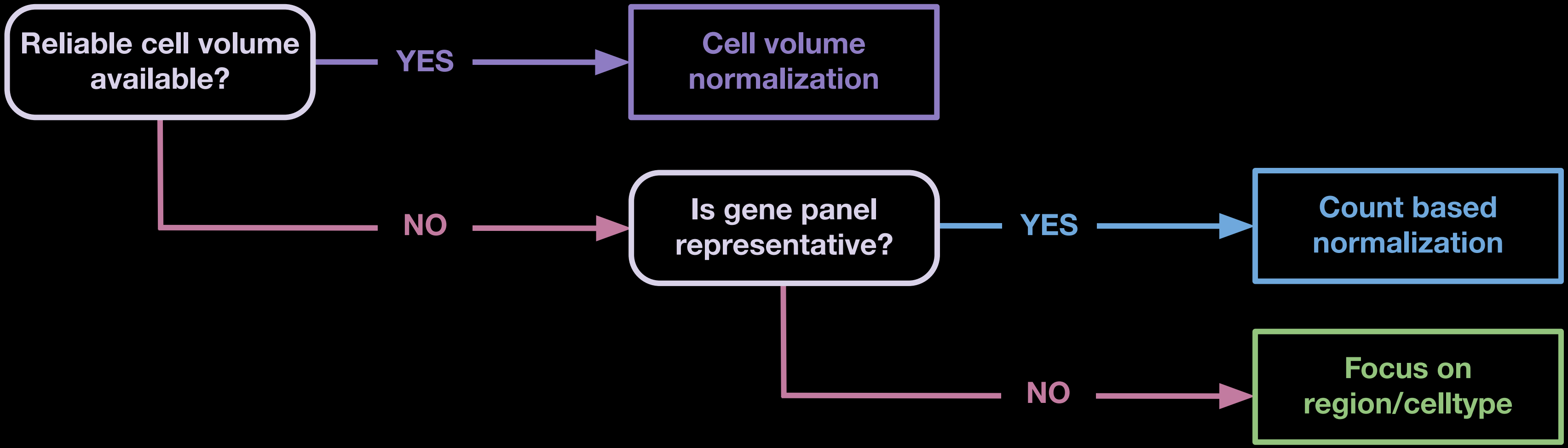
NO

Is gene panel representative?

YES

Count based normalization

# Recommendations



Reliable cell volume available?

YES

Cell volume normalization

NO

Is gene panel representative?

YES

Count based normalization

NO

Focus on region/celltype

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

**JEFworks Lab**
**Jean Fan, PhD**
**Kalen Clifton** **– STalign**
**Manjari Anant**
**Gohta Aihara** **– SEraster**
Srujan Singh
Rafael dos Santos Piexoto
Mayling Chen
Dee Velazquez
Vivien Jiang

# TPM

- Divide by (gene length then) total counts

| | SAMPLE 1 | SAMPLE 2 |
|---|---|---|
| Gene A | 5 | 15 |
| Gene B | 5 | 15 |
| Gene C | 5 | 15 |
| Total | 15 | 45 |

↓ TPM

| | SAMPLE 1 | SAMPLE 2 |
|---|---|---|
| Gene A | 5/15 = 1/3 | 15/45 = 1/3 |
| gene B | 1/3 | 1/3 |
| Gene C | 1/3 | 1/3 |
| Total | 1 | 1 |

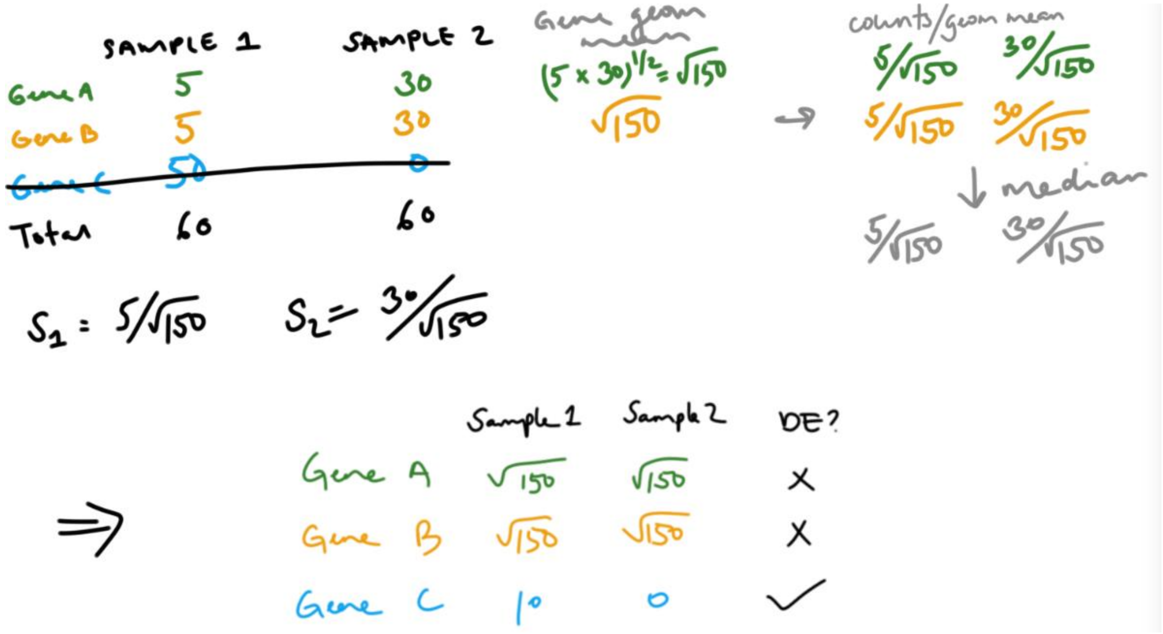
# DESeq normalization: accounting for compositional differences

- Filter out genes not expressed in all samples
- Sample scaling factor:

$$\hat{s}_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{1/m}}.$$

i:genes, j:samples

k:counts



| | SAMPLE 1 | SAMPLE 2 | Gene geom mean |
|---|---|---|---|
| Gene A | 5 | 30 | $(5 \times 30)^{1/2} = \sqrt{150}$ |
| Gene B | 5 | 30 | $\sqrt{150}$ |
| ~~Gene C~~ | ~~50~~ | ~~0~~ | |
| Total | 60 | 60 | |

→ counts/geom mean

| | |
|---|---|
| $5/\sqrt{150}$ | $30/\sqrt{150}$ |
| $5/\sqrt{150}$ | $30/\sqrt{150}$ |

↓ median

| | |
|---|---|
| $5/\sqrt{150}$ | $30/\sqrt{150}$ |

$S_1 = 5/\sqrt{150}$ $\quad S_2 = 30/\sqrt{150}$

⇒

| | Sample 1 | Sample 2 | DE? |
|---|---|---|---|
| Gene A | $\sqrt{150}$ | $\sqrt{150}$ | ✗ |
| Gene B | $\sqrt{150}$ | $\sqrt{150}$ | ✗ |
| Gene C | 10 | 0 | ✓ |