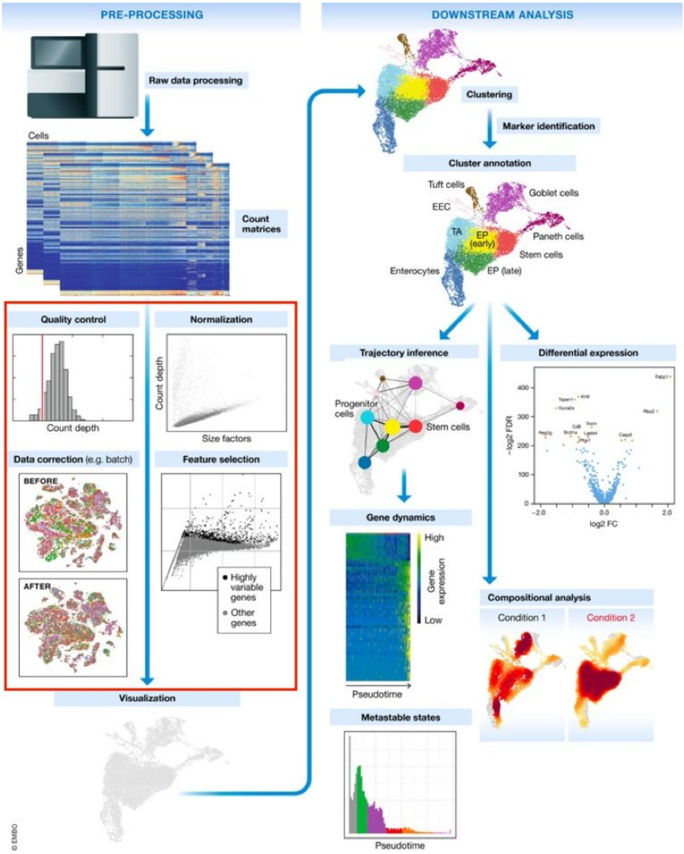
The Many Steps of scRNAseq Processing

Tara Chari
03/27/24

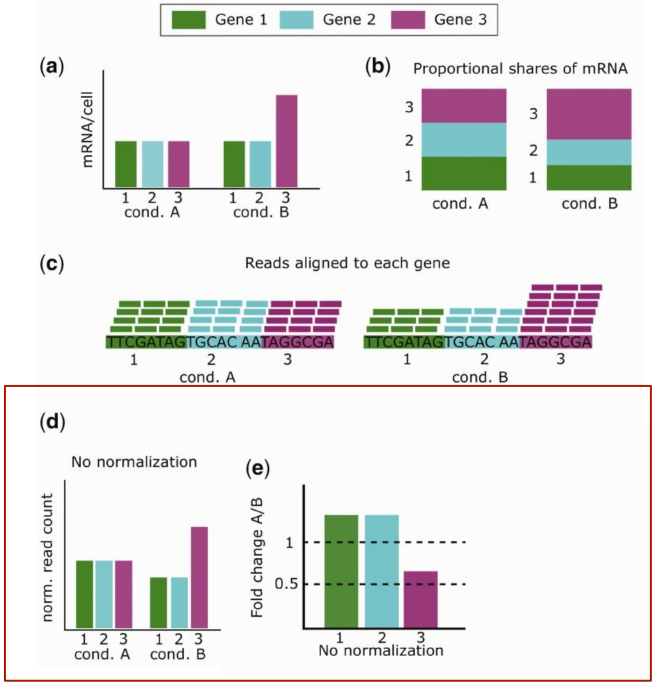
QC and Pre-processing of scRNAseq Counts



Biological question underlying normalization

Which genes are 'differentially-expressed', or where is difference in expression significant

- Relies on measure of difference in magnitude or fold-change



Evans et al. 2018

Why do we do **read-depth normalization**?

1. Want to be able to compare gene counts between two samples (cells).

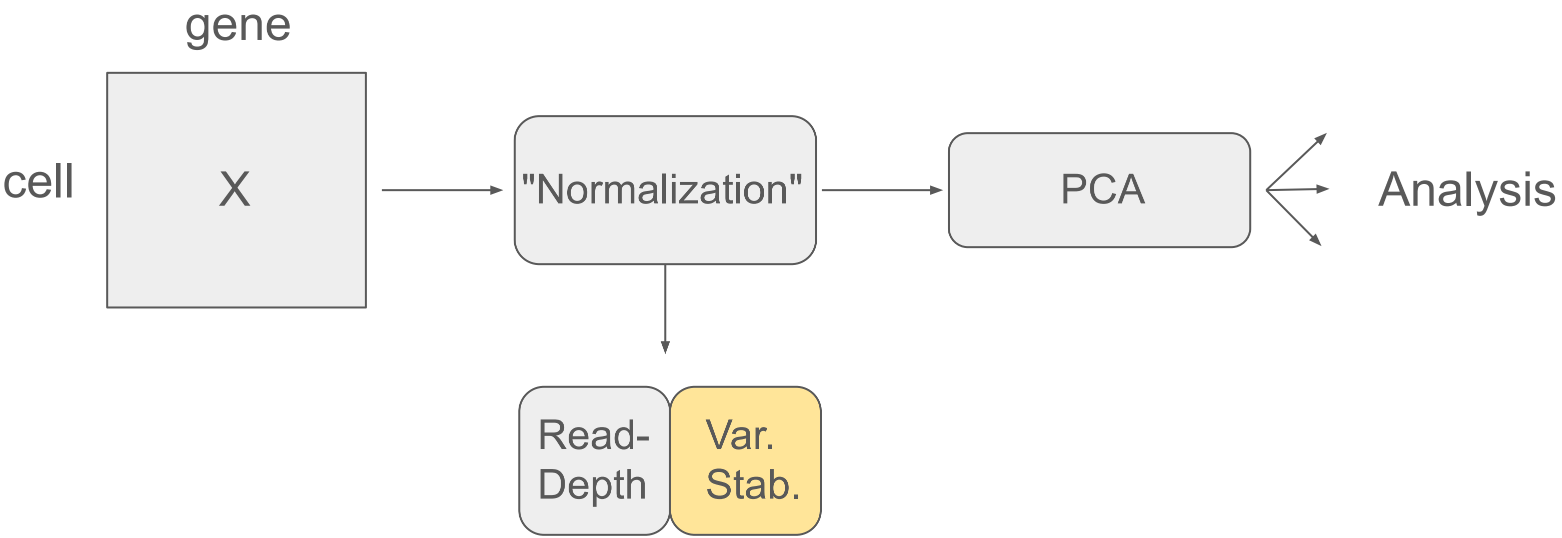
2. The process of sequencing has many possible steps for technical bias in molecule capture:
    a. Biased capture of particular molecules
    b. PCR of molecules
    c. Binding to flow cells…

3. Assumes we want to undo these *technical* effects

4. It is a *scaling* of the data to make counts between groups comparable

    Cell counts * New Total/Total Cell counts

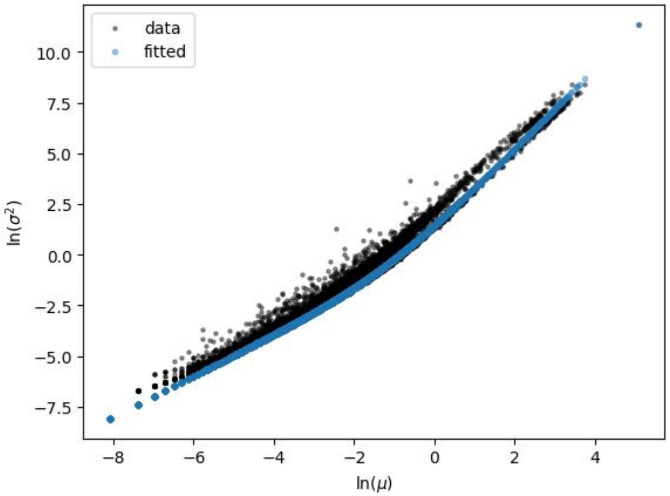
Variance stabilization often motivated by use of PCA



High Expression Genes are Very Variable

'Overdispersion'

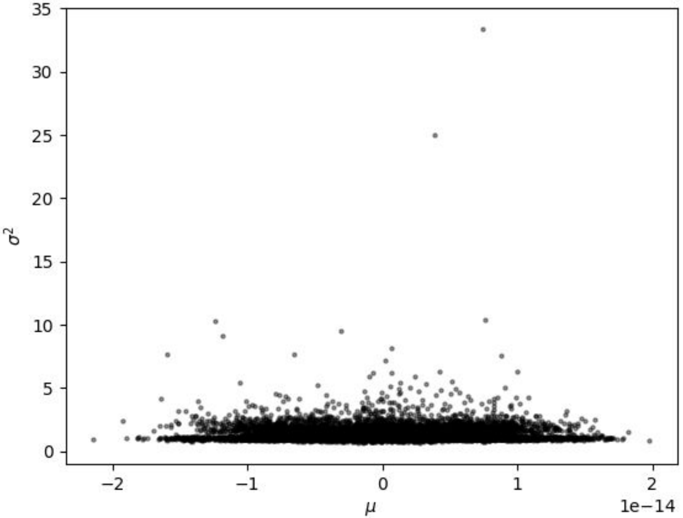
Var Expr



Mean Expr

Variance-Stabilized Genes

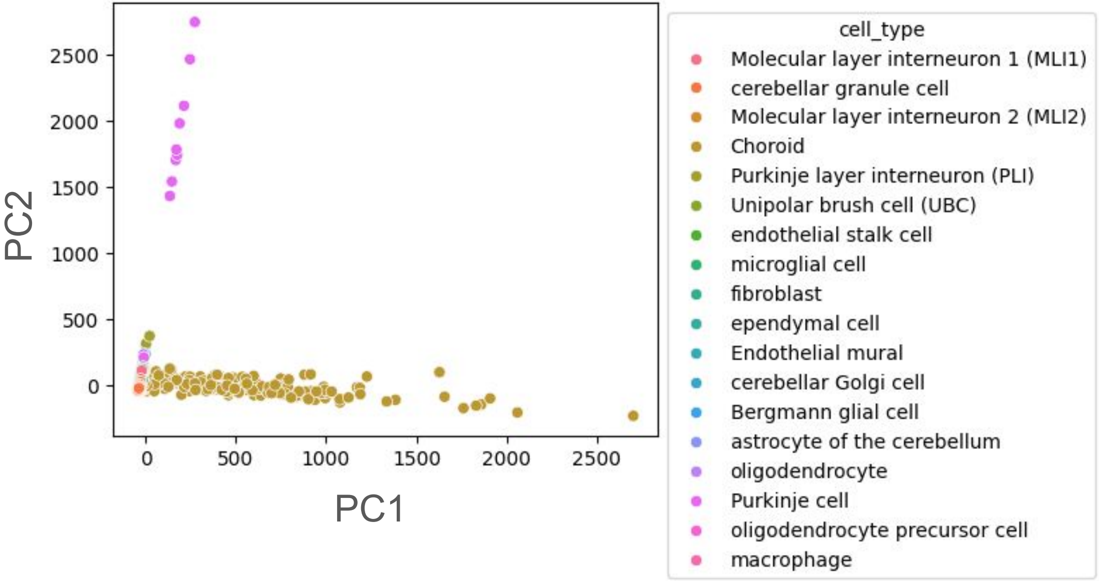
Var Expr

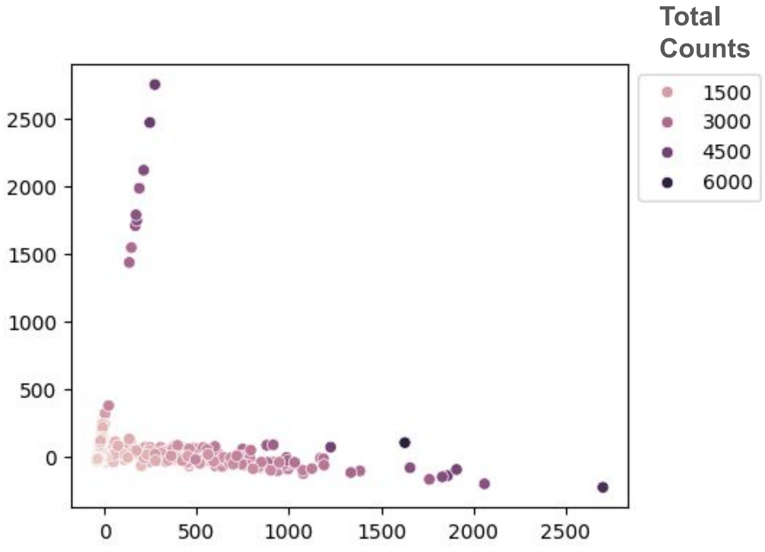


Mean Expr

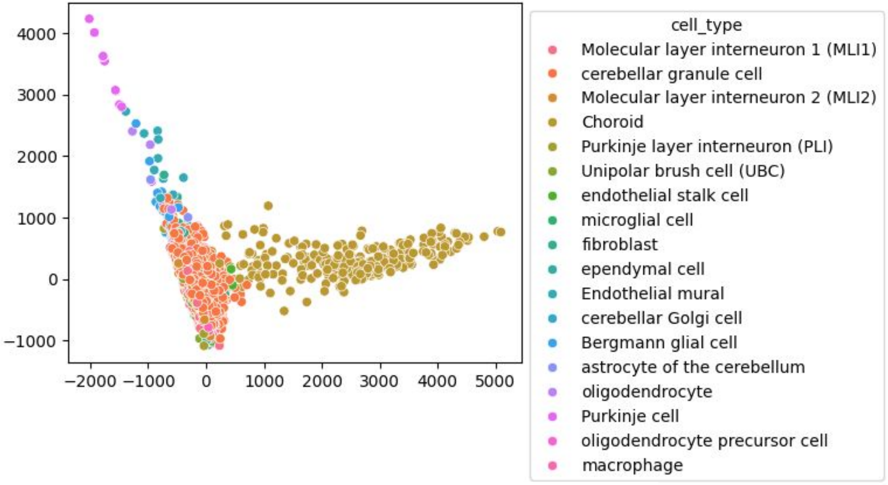
Why are normalization and stabilization useful?

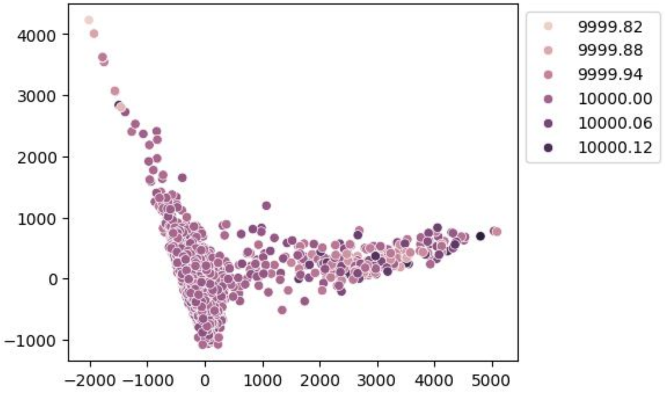
Results of PCA on Raw Counts:

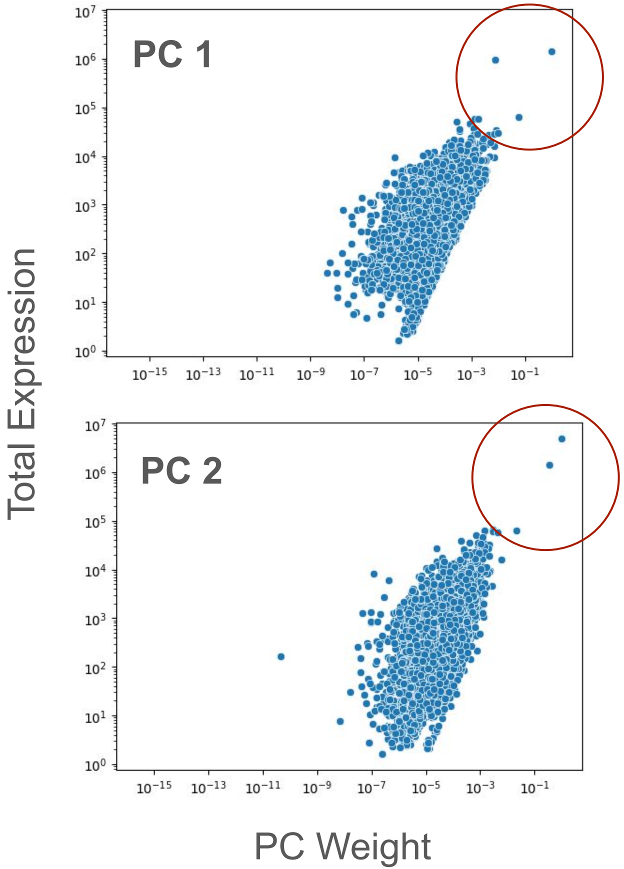




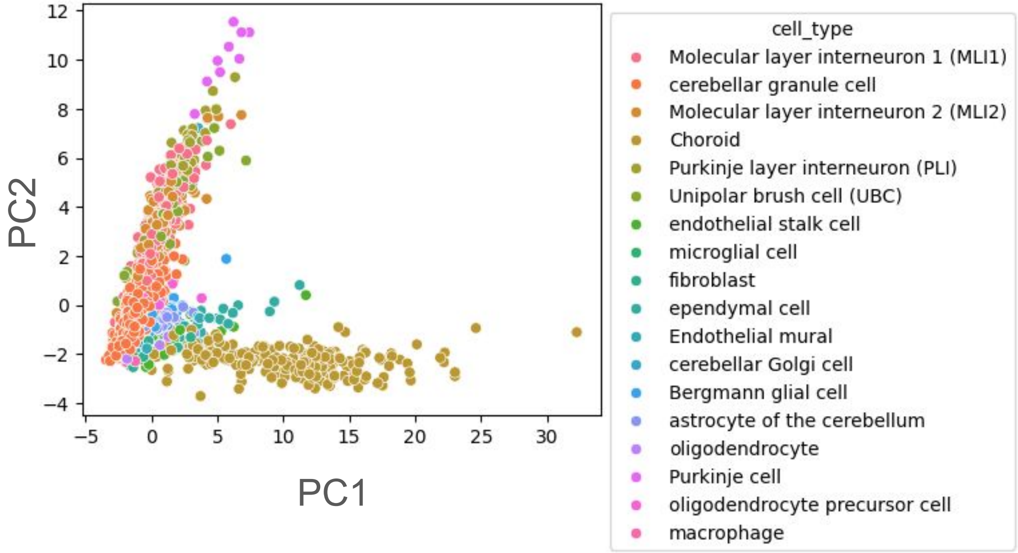
Results of only read-depth correction

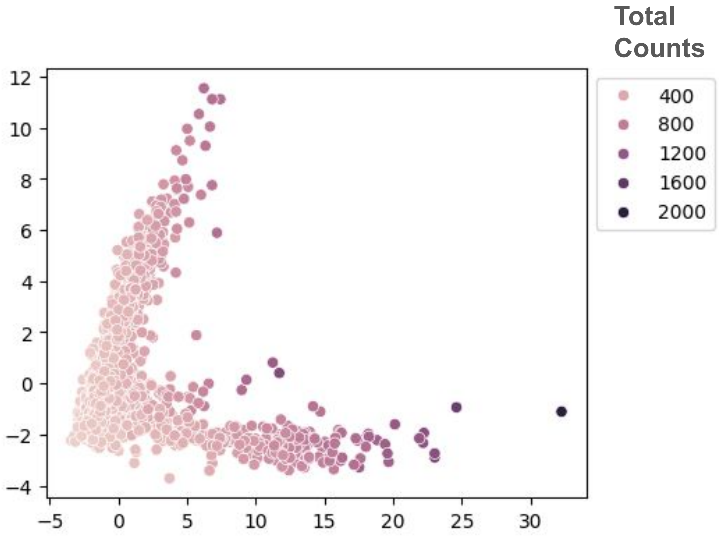




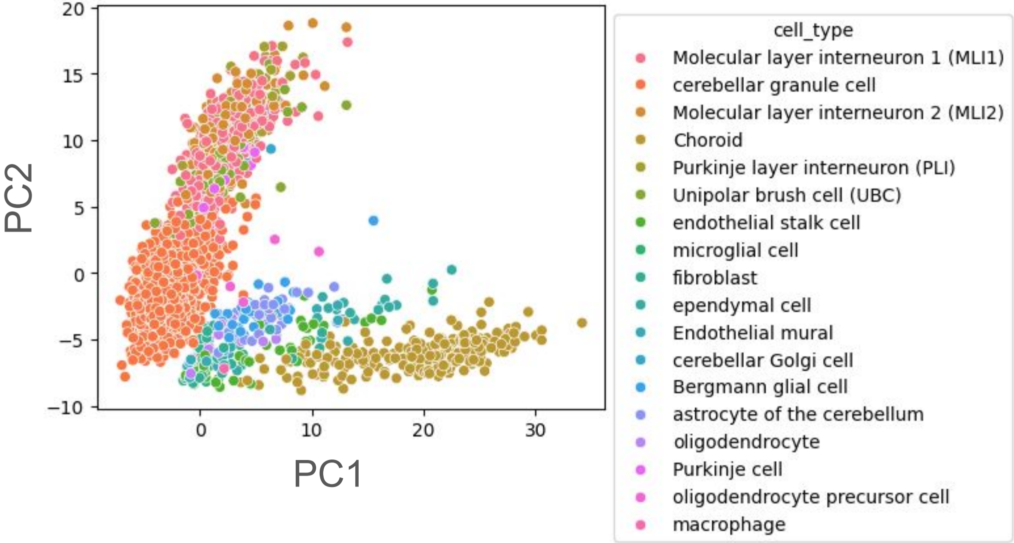


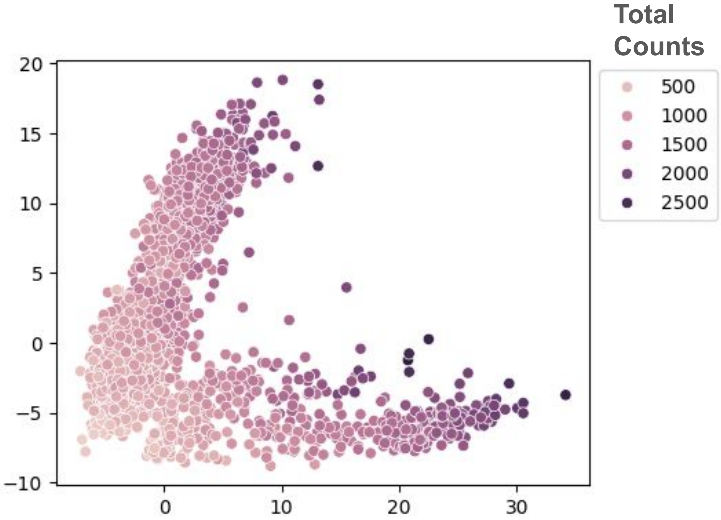
Results of only log-normalization





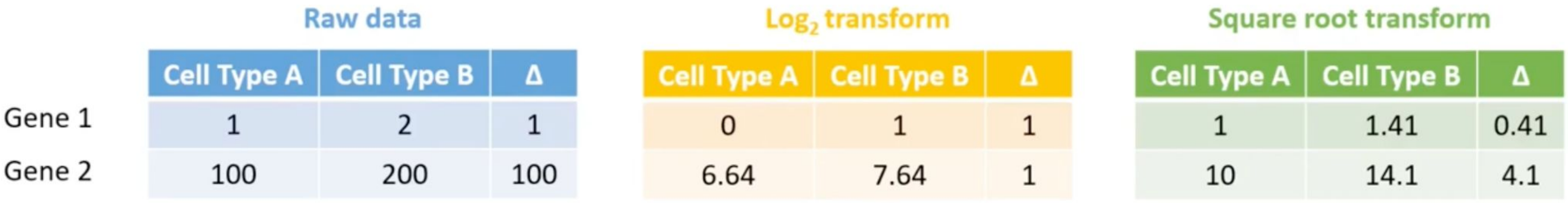
Results with both transformations





Transformations change impact of high expression genes

~log1p



Raw data

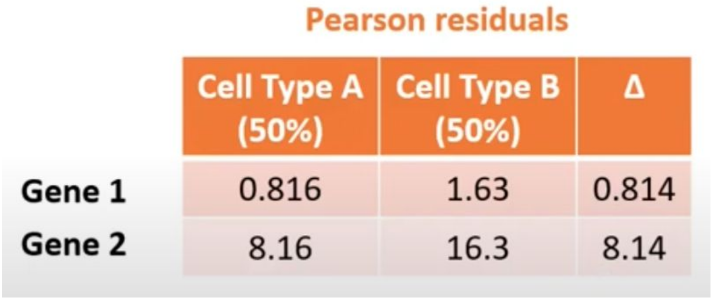
	Cell Type A	Cell Type B	Δ
Gene 1	1	2	1
Gene 2	100	200	100

Log_2 transform

	Cell Type A	Cell Type B	Δ
Gene 1	0	1	1
Gene 2	6.64	7.64	1

Square root transform

	Cell Type A	Cell Type B	Δ
Gene 1	1	1.41	0.41
Gene 2	10	14.1	4.1



Pearson residuals

	Cell Type A (50%)	Cell Type B (50%)	Δ
Gene 1	0.816	1.63	0.814
Gene 2	8.16	16.3	8.14

Seurat Transformations/Normalizations

`"LogNormalize`": Feature counts for each cell are divided by the total counts for that cell and multiplied by the `scale.factor`. This is then natural-log transformed using `log1p`

"`RC`": Relative counts. Feature counts for each cell are divided by the total counts for that cell and multiplied by the `scale.factor`. No log-transformation is applied

CLR: "Seurat CLR removes 0 counts first by `x[x>0]` and then `log1p` transform the raw counts, sum them up, calculate the average of the log counts, `exp` it back, and then divided the raw counts by this average and finally `log1p` again"

Workflow in Hasel et al . 2021

1. **Normalization/stabilization**
2. **Integration**
3. **PCA Reduction**
4. Clustering
5. DE (Differential Expression) Analysis
6. Spatial Analysis (?)

Discussion Question 1

How would you assess what type of normalization or variance stabilization procedure is 'best'?

Discussion Question 2

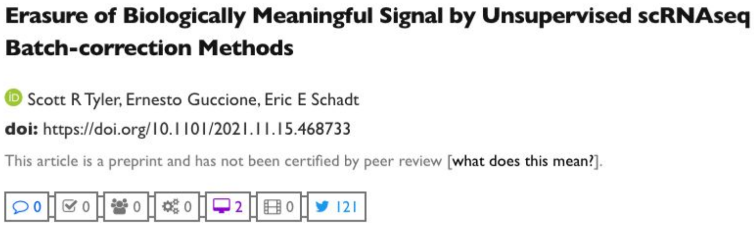
How can you determine if the PCA reduction captures biological variability?

Discussion Question 3

How could you assess the effects integration is having on the downstream reduction and, ultimately, clustering/annotations?

How to Assess Effects of Major Processing Steps

1. **Normalization/stabilization:**
   - **a.** Assess mean/var relationship before/after
   - b. Are cell size effects technical or biological?
     - i. Spike-ins, control genes, etc
   - c. Determine if downstream task requires transformation



Erasure of Biologically Meaningful Signal by Unsupervised scRNAseq Batch-correction Methods

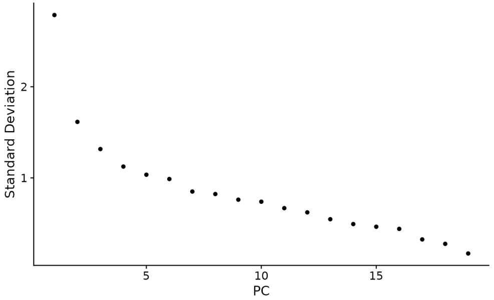
Scott R Tyler, Ernesto Guccione, Eric E Schadt

doi: https://doi.org/10.1101/2021.11.15.468733

This article is a preprint and has not been certified by peer review [what does this mean?].

2. **Integration**
   - a. Use quantitative metrics
     - i. How mixed cells are (or not)
     - ii. If biological variation retained across samples
   - b. Test multiple methods, including something simple like downsampling

3. **PCA Reduction**
   - a. Look at highly weighted genes in Principal Components
     - i. Plot PCs
   - b. Use ElbowPlot() to see how much data variance PCs capture



Expanding the Spatial Possibilities

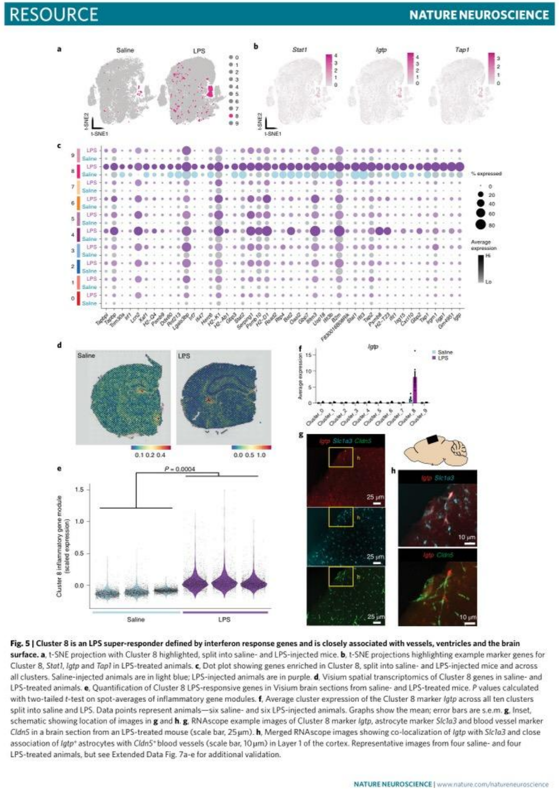


Fig. 5 | Cluster 8 is an LPS super-responder defined by interferon response genes and is closely associated with vessels, ventricles and the brain surface. a, t-SNE projection with Cluster 8 highlighted, split into saline- and LPS-injected mice. **b**, t-SNE projections highlighting example marker genes for Cluster 8, *Stat1*, *Igtp* and *Tap1* in LPS-treated animals. **c**, Dot plot showing genes enriched in Cluster 8, split into saline- and LPS-injected mice and across all clusters. Saline-injected animals are in light blue; LPS-injected animals are in purple. **d**, Visium spatial transcriptomics of Cluster 8 genes in saline- and LPS-treated animals. **e**, Quantification of Cluster 8 LPS-responsive genes in Visium brain sections from saline- and LPS-treated mice. *P* values calculated with two-tailed *t*-test on spot-averages of inflammatory gene modules. **f**, Average cluster expression of the Cluster 8 marker *Igtp* across all ten clusters split into saline and LPS. Data points represent animals—six saline- and six LPS-injected animals. Graphs show the mean; error bars are s.e.m. **g**, Inset, schematic showing location of images in **g** and **h**. **g**, RNAscope example images of Cluster 8 marker *Igtp*, astrocyte marker *Slc1a3* and blood vessel marker *Cldn5* in a brain section from an LPS-treated mouse (scale bar, 25 µm). **h**, Merged RNAscope images showing co-localization of *Igtp* with *Slc1a3* and close association of *Igtp*⁺ astrocytes with *Cldn5*⁺ blood vessels (scale bar, 10 µm) in Layer 1 of the cortex. Representative images from four saline- and four LPS-treated animals, but see Extended Data Fig. 7a-e for additional validation.

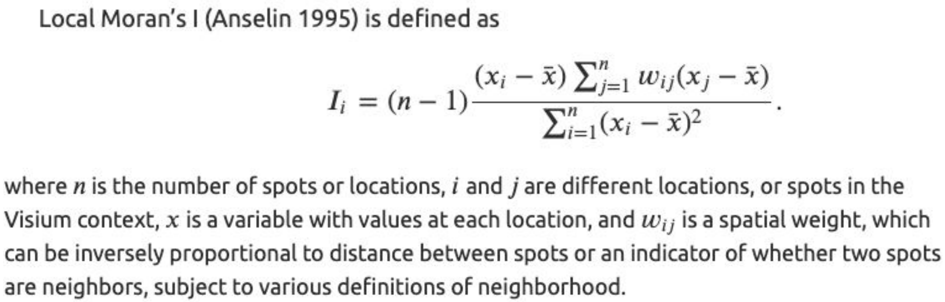
NATURE NEUROSCIENCE | www.nature.com/natureneuroscience

How to quantify spatially distinct patterns?

How to find interesting spatial behavior that isn't just RNAseq cluster markers?

Investigation with Local Moran's I

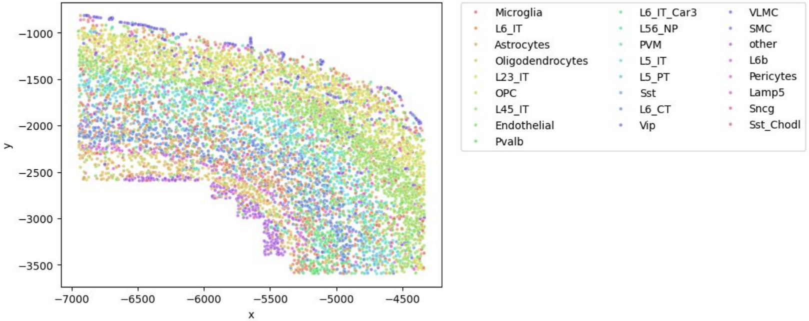
https://pachterlab.github.io/voyager/articles/localmoran_landing.html

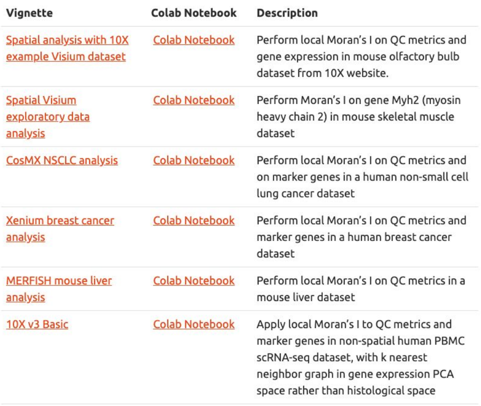


Local Moran's I (Anselin 1995) is defined as

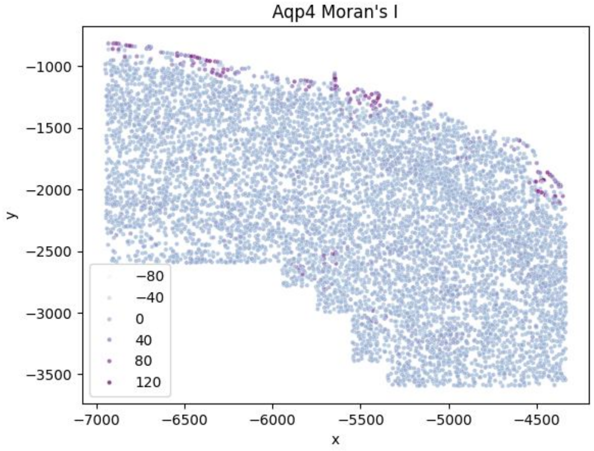
$$I_i = (n-1)\frac{(x_i - \bar{x})\sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

where n is the number of spots or locations, i and j are different locations, or spots in the Visium context, x is a variable with values at each location, and w_{ij} is a spatial weight, which can be inversely proportional to distance between spots or an indicator of whether two spots are neighbors, subject to various definitions of neighborhood.





Vignette	Colab Notebook	Description
Spatial analysis with 10X example Visium dataset	Colab Notebook	Perform local Moran's I on QC metrics and gene expression in mouse olfactory bulb dataset from 10X website.
Spatial Visium exploratory data analysis	Colab Notebook	Perform Moran's I on gene Myh2 (myosin heavy chain 2) in mouse skeletal muscle dataset
CosMX NSCLC analysis	Colab Notebook	Perform local Moran's I on QC metrics and on marker genes in a human non-small cell lung cancer dataset
Xenium breast cancer analysis	Colab Notebook	Perform local Moran's I on QC metrics and marker genes in a human breast cancer dataset
MERFISH mouse liver analysis	Colab Notebook	Perform local Moran's I on QC metrics in a mouse liver dataset
10X v3 Basic	Colab Notebook	Apply local Moran's I to QC metrics and marker genes in non-spatial human PBMC scRNA-seq dataset, with k nearest neighbor graph in gene expression PCA space rather than histological space


Aqp4 Moran's I
y
x