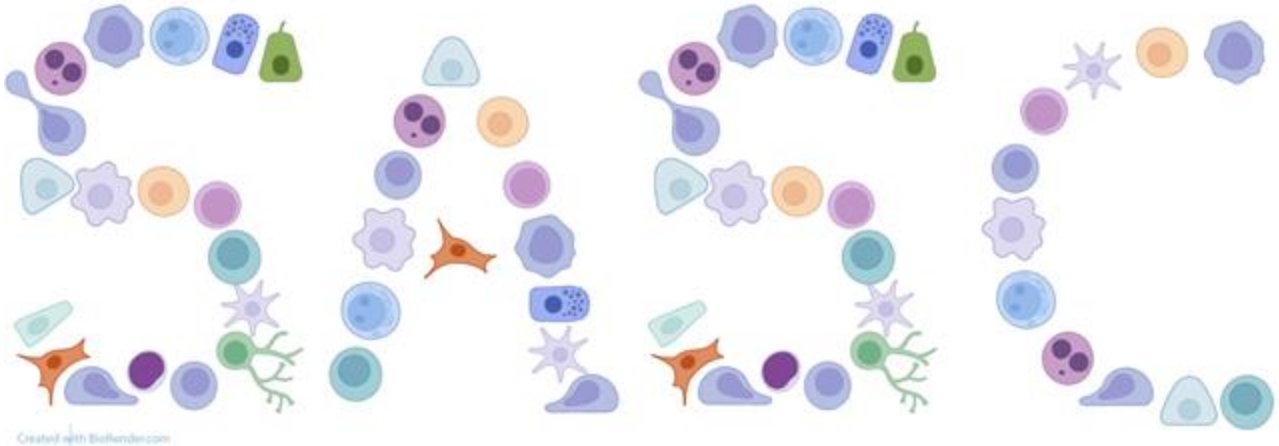
Welcome to



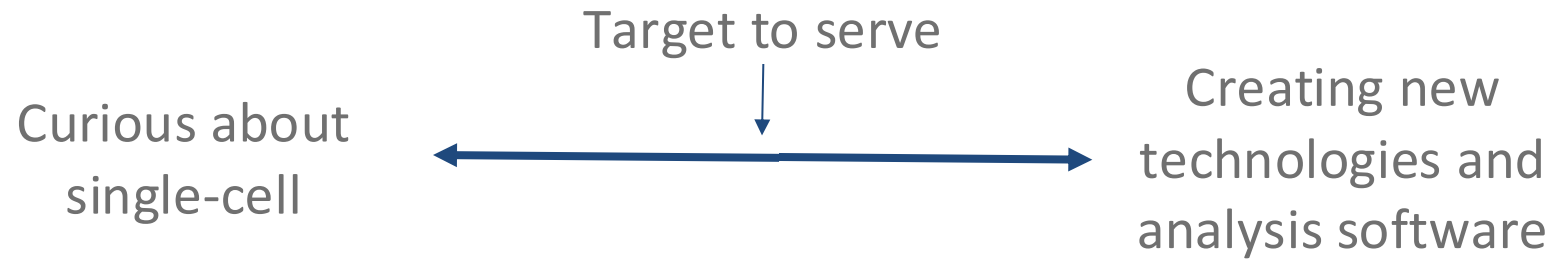
Seattle Area Single Cell (SASC) User Group
March 13, 2025

Our disclosures

- BBI has a fee-for-service single-cell sequencing platform
  - We offer non-commercial combinatorial indexing methods
  - We work closely with the Trapnell Lab who develop Monocle3
- In the past, some of us have worked with 10X Genomics 3’ Gene Expression
  - Performed most analyses in the Bioconductor Single Cell Universe (e.g. Scater, Scran, Scuttle, DropletUtils)

SASC Goals

- We are striving for this user group meeting be technology (and software) agnostic
  - **Focus on common challenges**



Agenda

- Brief introduction to normalization
- Workshop/tutorial
- Networking (remaining time)
    - Refreshments sponsored by Illumina - thank you!

Favorite online resources

https://bioconductor.org/books/release/OSCA/

Orchestrating Single-Cell Analysis with Bioconductor

https://www.sc-best-practices.org/preamble.html

Single-cell best practices

Awesome Single Cell

Community-curated list of software packages and data resources for single-cell, including RNA-seq, ATAC-seq, etc.

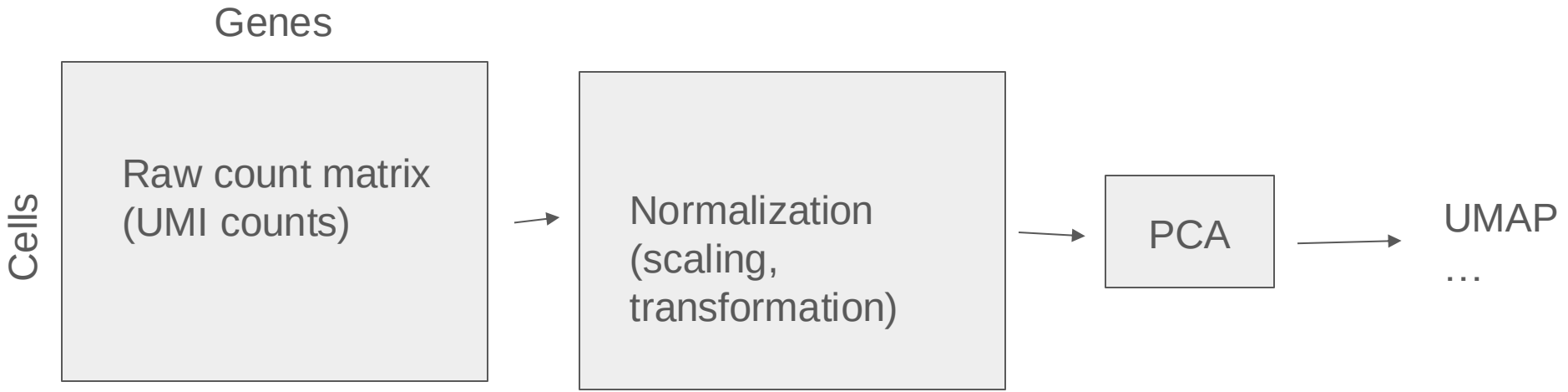
https://github.com/seandavi/awesome-single-cell?tab=readme-ov-file

Normalization

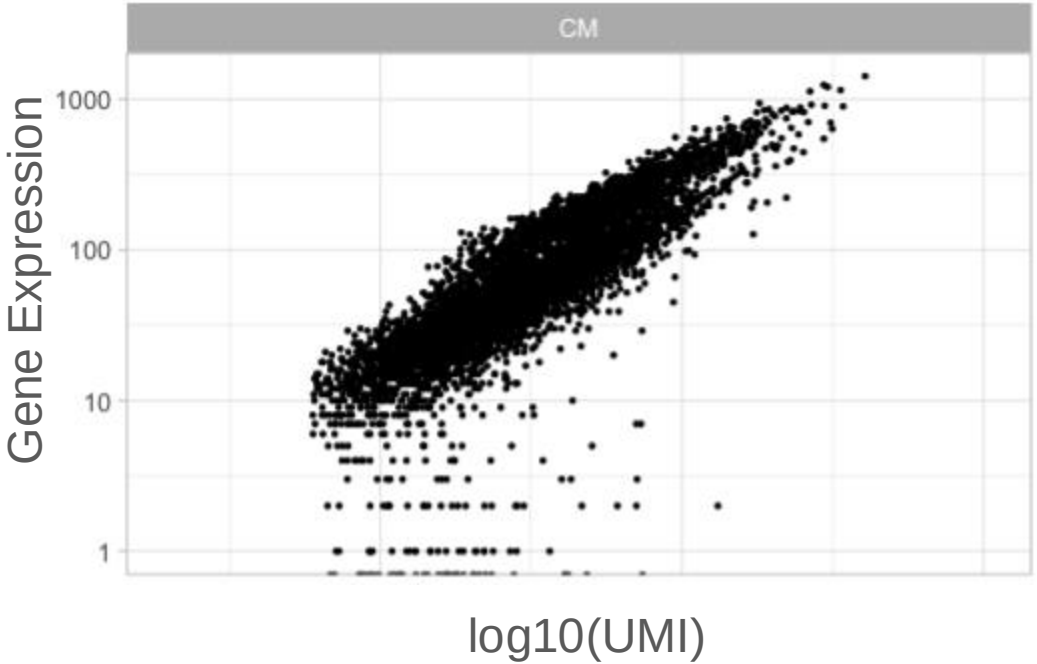
Great video on youtube: https://www.youtube.com/watch?v=huxkc2GH4lk

Many slides shamelessly lifted and condensed from Florian Wagner's video above!

Defining “normalization”



Scaling → removing correlation with sequencing depth



‘Size Factor’

e.g. sum of UMI counts

Transformation → shift the weight

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
0	1	1
6.64	7.64	1

- Simple log transformation → applied to each measurement independently
- Log transformation weights fold change, not expression level

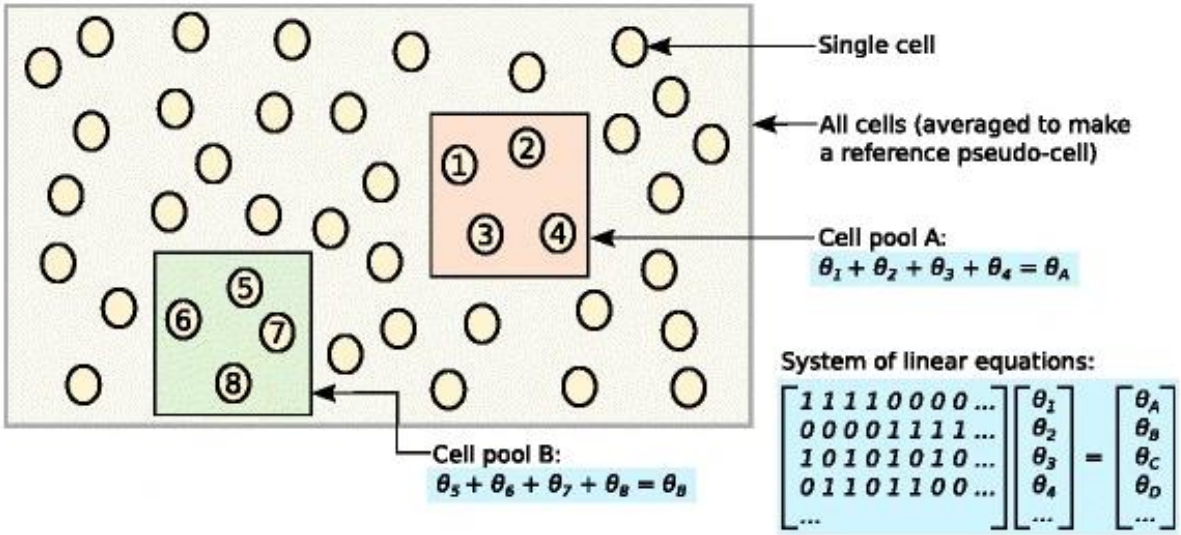
https://www.youtube.com/watch?v=huxkc2GH4Ik

Shifted logarithm, aka library size normalization

$$f(y) = \log\left(\frac{y}{s} + y_0\right)$$

- Y = raw counts
- S = size factor
- y_0= pseudo-count

Deconvolution, aka pooling-based size factor



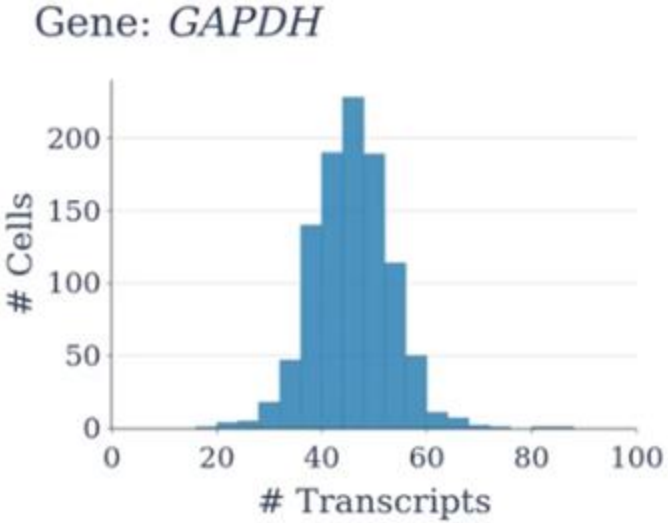
- Estimates size factors based on a linear regression over genes for pools of cells
- Aims to better account for differences in count depths/composition biases across all cells

Lun, Bach & Marioni (2016)

Pearson residuals (e.g. SCTransform)

- Each gene contributes to the analysis according to how much evidence there is that it is non-uniformly expressed
- Favors genes that are expressed in only a small fraction of cells
- Lowly expressed genes with very cell-type-specific expression patterns can contribute more to the analysis than highly expressed genes with broader expression patterns
- In principle, very useful for identification of rare cell types
- In practice, really depends on the dataset
    - subject to overcorrect especially on sparse data
    - computationally expensive

Noisy data!



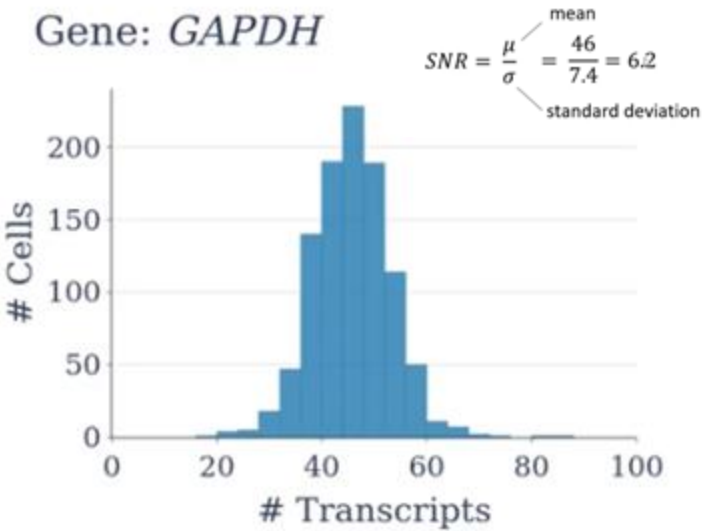
Data: *Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells* (Klein et al., Cell 2015)

- Synthetic transcriptomes
- All variation is technical

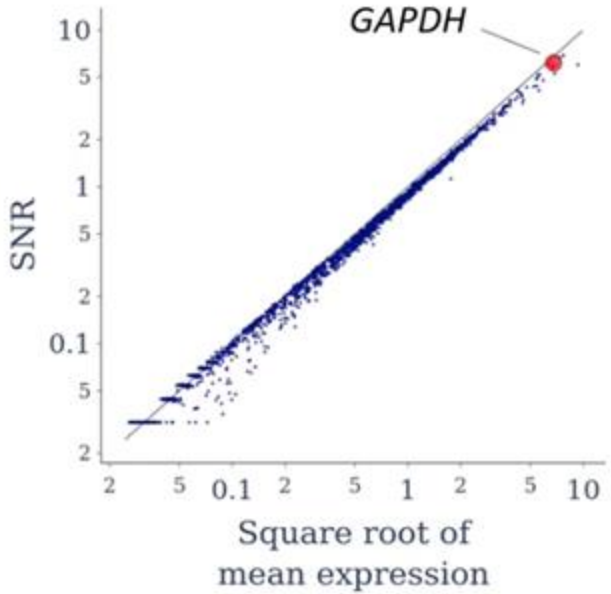
True biological differences or technical noise?

https://www.youtube.com/watch?v=huxkc2GH4lk

The more highly expressed a gene, the more accurately it can be measured



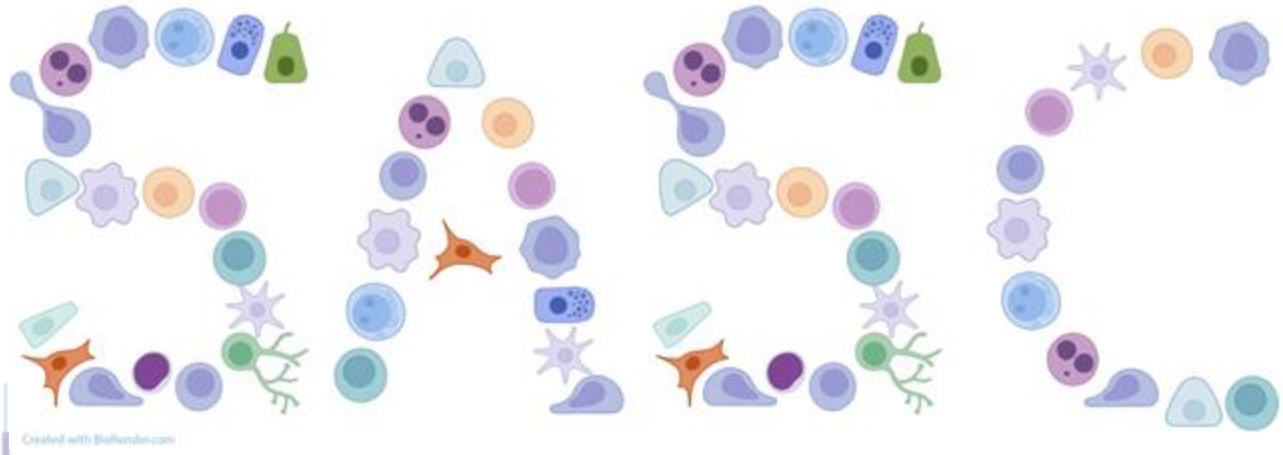
Data: *Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells* (Klein et al., Cell 2015)



https://www.youtube.com/watch?v=huxkc2GH4lk

Tutorial

The next



Kevin Lin, assistant professor at UW Biostatistics, has volunteered to lead the next session on multi-ome (RNA+ATAC)! Stay tuned for the date/time/place.