scRNA-seq Workshop Part 2

Applied Bioinformatics for Genomics

JENNIFER A. FOLTZ, PHD

ASSISTANT PROFESSOR, SECTION OF COMPUTATIONAL BIOLOGY

JENNIFER.A.FOLTZ@WUSTL.EDU

With content contributions from Dr. Petti

Review of Plots from Homework (see Rmd file)

Characterizing cells using differential gene expression

Other Differential Expression Tests to Consider:

- AUC/ROC
- DeSeq2
- MAST/Logistic Regression
- Scran pairwiseWilcox() by blocking
- Wilcoxon rank-sum test

Things to Consider:

- Pairwise differential gene expression
- What fraction of cells in each sample express a given gene?
- Of the cells in each sample that express a given gene, does the mean expression in those cells differ?
- Does the distribution of cell types differ between samples?
- Do the samples exhibit cell-type-specific differential gene expression?
- Input into gene ontology algorithms (e.g. do you need a universe/background?)

Characterizing cells using differential gene expression

- Data has zero-inflated negative binomial distribution (lots of zeros, overdispersed) so can't use bulk methods
  - Default in Seurat: Wilcoxon rank-sum test
  - Nonparametric version of t-test
  - For two clusters (A and B), and one gene, rank each cell in each cluster according to expression
  - Determine whether sum-of-ranks for cluster A is significantly different than sum-of-ranks for cluster B
  - Clear explanation of Wilcoxon rank-sum test: http://statweb.stanford.edu/~susan/courses/s141/hononpara.pdf
  - Numerous other tests in Seurat and other packages

    Fold-changes are lower due to noise and low- detection

    Generally accepted to set a minimum detection rate to decrease noise & power

        Most commonly 25% of cells must express the gene for it be detected but can do lower or higher depending on question

Analyzing Multiple Samples

- Merging or batch-correction?
- Avoid batch correction unless absolutely necessary
  - Correct for different technologies (e.g. 3’ and 5’)
  - Correct for different batches
  - Discover conserved biology by finding corresponding cells across different data sets
- Cellranger does faux batch-correction (corrected values are discarded), but batch-corrected tSNE can be visualized in the loupe browser.

What Not to Do

- CRITICAL: Experimental Design Considerations
  - Submission Date
  - What is your hypothesis?
  - How do you envision doing differential expression downstream?
- DO:
  - Treat experimental groups as similar as possible:
    - E.g. if you need to sort one group, sort the other group even if it is technically not needed
  - Include control cell populations that can be used to assess how well a technique is working
- DON'T:
  - Submit control & experimental groups on separate days, technologies, etc.
  - Batch-correct on your experimental question
    - E.g. batch-correct on drug treatment and expect to find clusters that are different with drug treatment

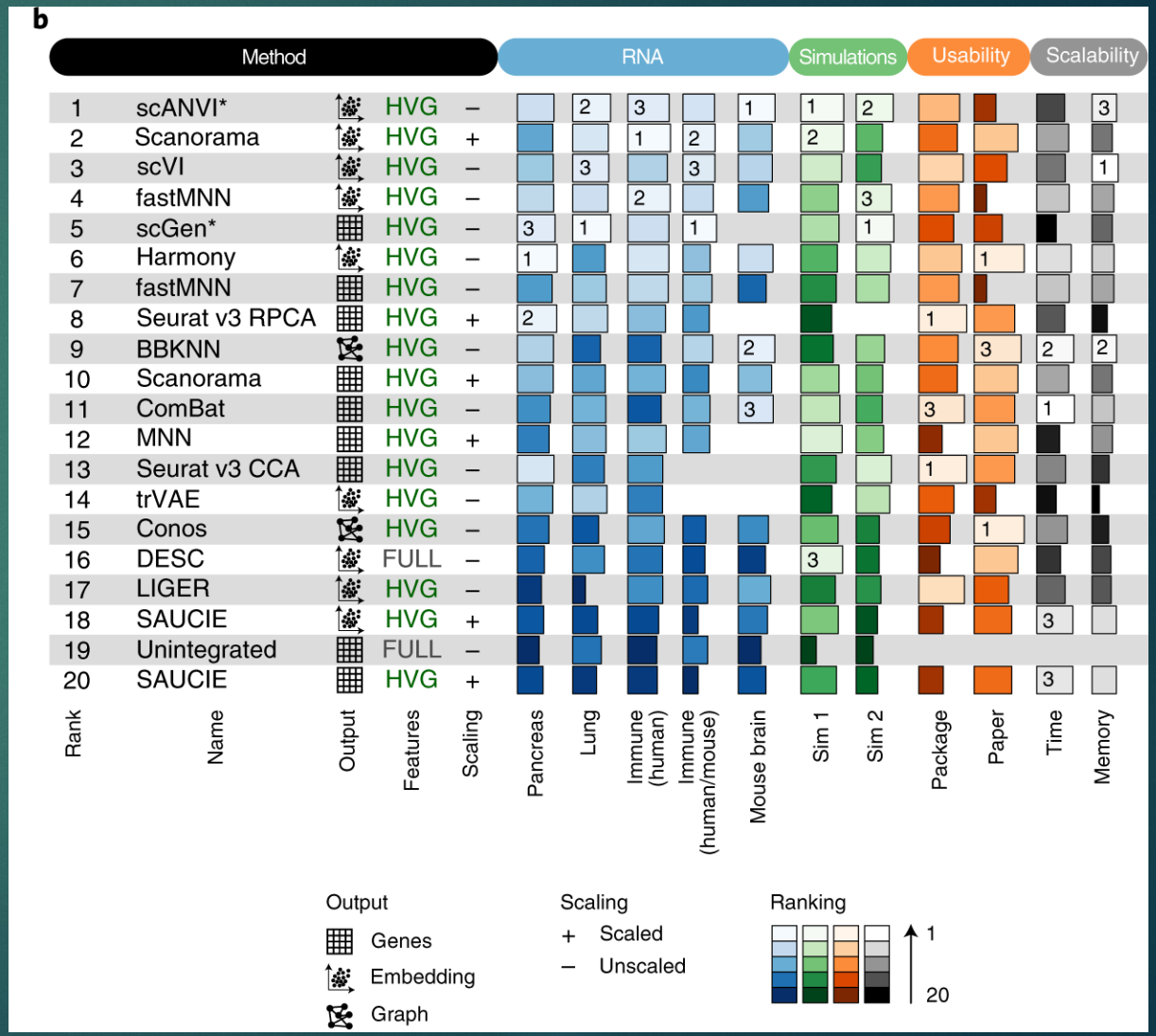
Analyzing Multiple Samples

Luecken, et al., 2022

Table 1 Description of the 14 batch-effect correction methods

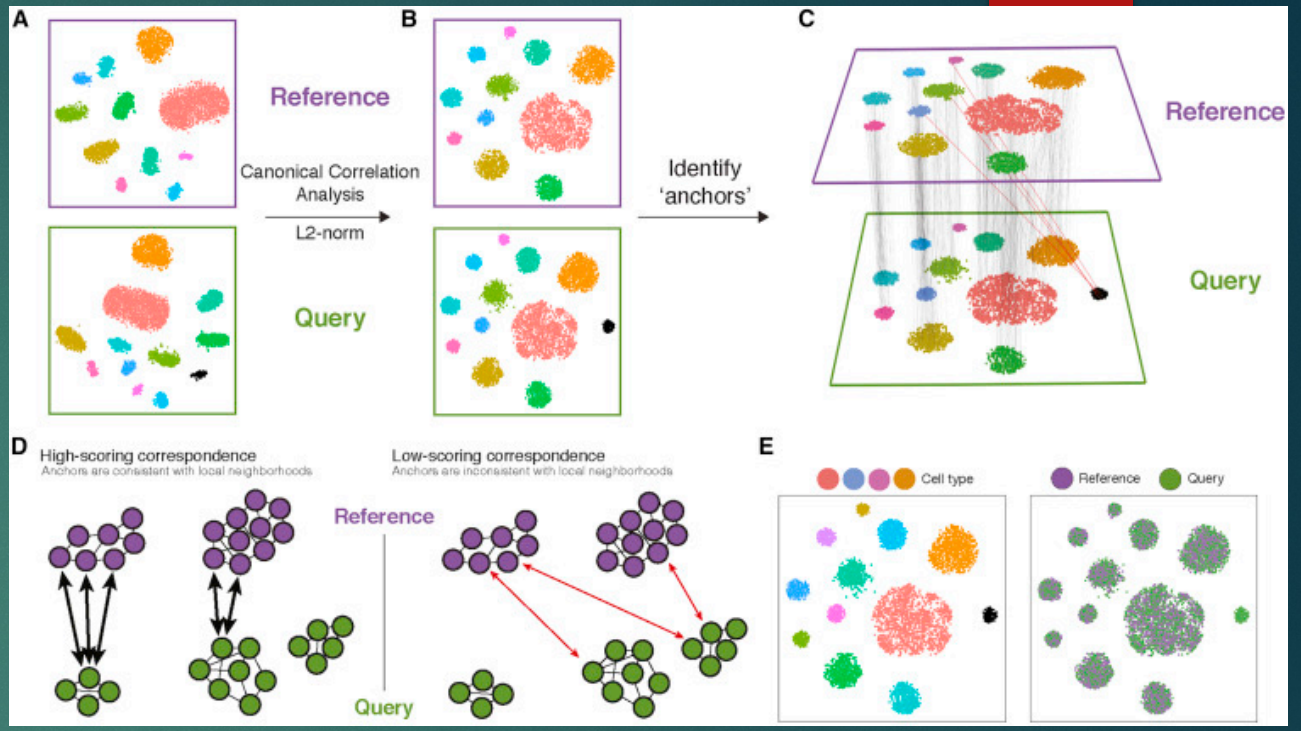
From: A benchmark of batch-effect correction methods for single-cell RNA sequencing data

Tools	Programming language	Batch-effect-corrected output	Methods	Reference package version
Seurat 2 (CCA, MultiCCA)	R	Normalized canonical components (CCs)	Canonical correlation analysis and dynamic time warping	Butler et al. [4], Seurat package version 2.3.4
Seurat 3 (Integration)	R	Normalized gene expression matrix	Canonical correlation analysis and mutual nearest neighbors-anchors	Stuart et al. [12], Seurat package version 3.0.1
Harmony	R	Normalized feature reduction vectors (Harmony)	Iterative clustering in dimensionally reduced space	Korsunsky et al. [13], Harmony version 0.99.9
MNN Correct	R	Normalized gene expression matrix	Mutual nearest neighbor in gene expression space	Haghverdi et al. [5], Scran package version 1.12.0
fastMNN	R	Normalized principal components	Mutual nearest neighbor in dimensionally reduced space	Haghverdi et al. [5], Lun ATL [7], Scran package version 1.12.0
ComBat	R	Normalized gene expression matrix	Adjusts for known batches using an empirical Bayesian framework	Johnson et al. [1]
limma	R	Normalized gene expression matrix	Linear model/empirical Bayes model	Smyth et al. [2], limma version 3.38.3
scGen	Python	Normalized gene expression matrix	Variational auto-encoders neural network model and latent space	Lotfollahi et al. [16], 2019, scGen version 1.0.0
Scanorama	Python/R	Normalized gene expression matrix	Mutual nearest neighbor and panoramic stitching	Hie et al. [9], Scanorama version 1.4.
MND-ResNet	Python	Normalized principal components	Residual neural network for calibration	Shaham et al. [15] updated code to Python 3
ZINB-WaVE	R	Normalized feature reduction vectors (ZINB-WaVE)/normalized gene expression matrix	Zero-inflated negative binomial model, extension of RUV model	Risso et al. [6], ZINB-WaVE version 1.6.0
scMerge	R	Normalized gene expression matrix	Stably expressed genes (scSEGs) and RUVIII model	Lin et al. [18], scMerge version 1.1.3
LIGER	R	Normalized feature reduction vectors (LIGER)	Integrative non-negative matrix factorization (iNMF) and joint clustering + quantile alignment	Welch et al. [14], liger version 1.0
BBKNN	Python/R	Connectivity graph and normalized dimension reduction vectors (UMAP)	Batch balanced *k*-nearest neighbors	Polański et al. [10], bioRxiv. BBKNN version 1.3.2



Option 1: Integration

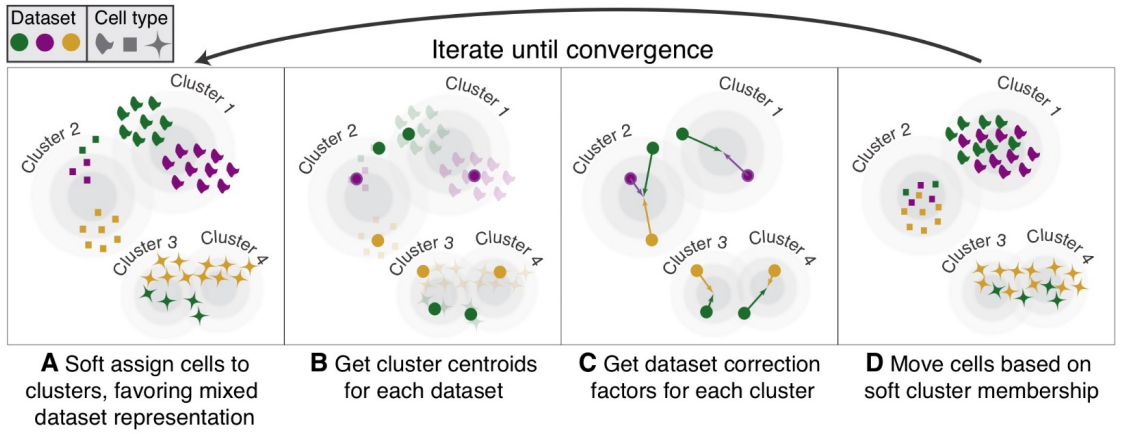
```
anchors <- FindIntegrationAnchors(object.list = scrna.list, di
ms = 1:30) # find anchors
scrna.int <- IntegrateData(anchorset = anchors, dims = 1:30
) # Integrate data
```



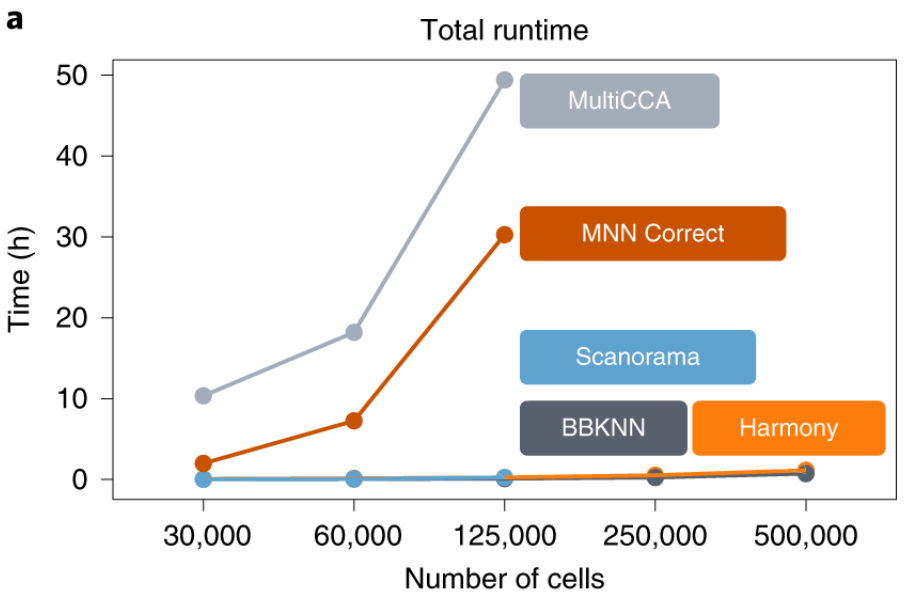
Stuart et al. 2019

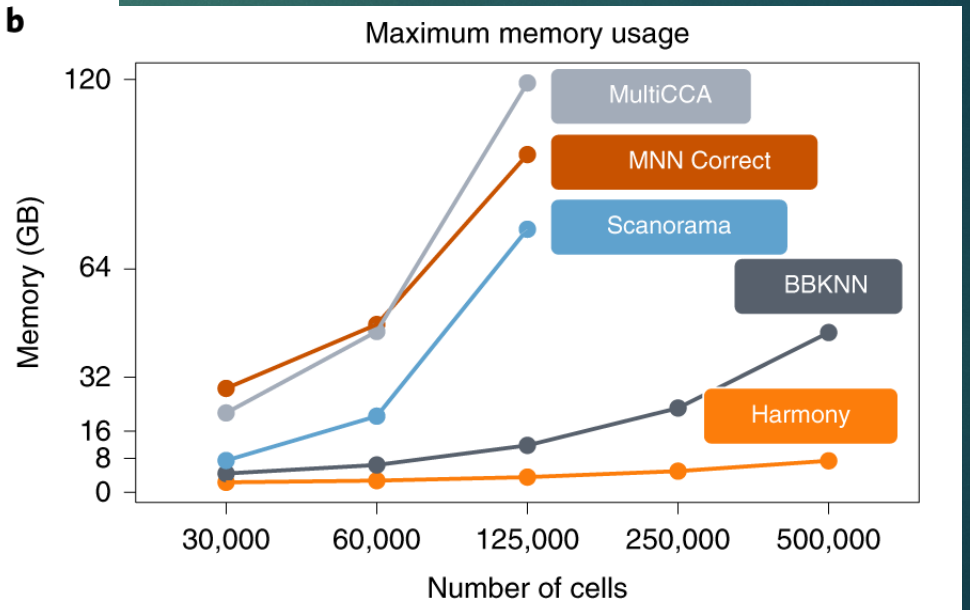
- Integrated values not intended for use with differential expression calculations.
  - We recommend running your differential expression tests on the “unintegrated” data. By default this is stored in the “RNA” Assay. There are several reasons for this.
  - The integration procedure inherently introduces dependencies between data points. This violates the assumptions of the statistical tests used for differential expression.
- TransferData function uses data integration to classify cells based on a reference data set.

Option 2: Harmony Batch Correction



- Ability to batch correct on multiple classes
- Minimal increase in the saved object file size

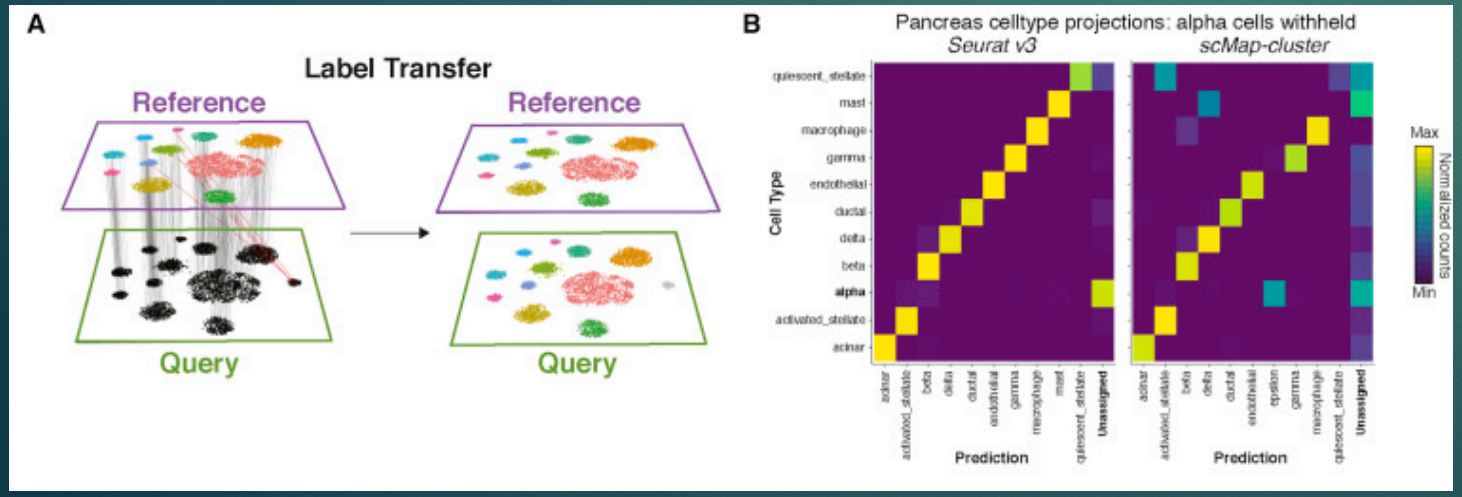




Korsunsky et al., 2019

Label Harmonization/Transfer Approaches

- In lieu of batch correction, can “integrate” data by aligning the sample annotations
  - e.g. look for gene changing between the same cell populations across samples
- Pros: No batch correction required, smaller datasets easier to work with computationally
- Cons: No increase in power by combining samples, annotations may not be 100% consistent



Stuart et al. 2019

Reclustering of Data

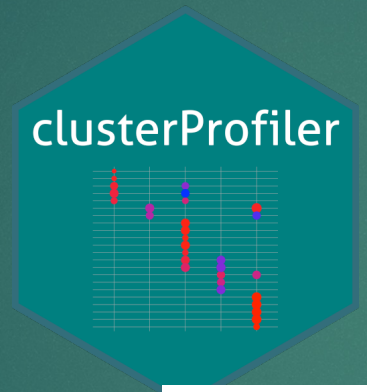
- Need to rerun FindVariableFeatures (in order to increase chance of finding additional heterogeneity or cell populations)
- Rerun ScaleData
  - I also rerun CellCycleScoring although likely not necessary
- Rerun dimensionality reduction tests, and batch correction (if using Harmony)

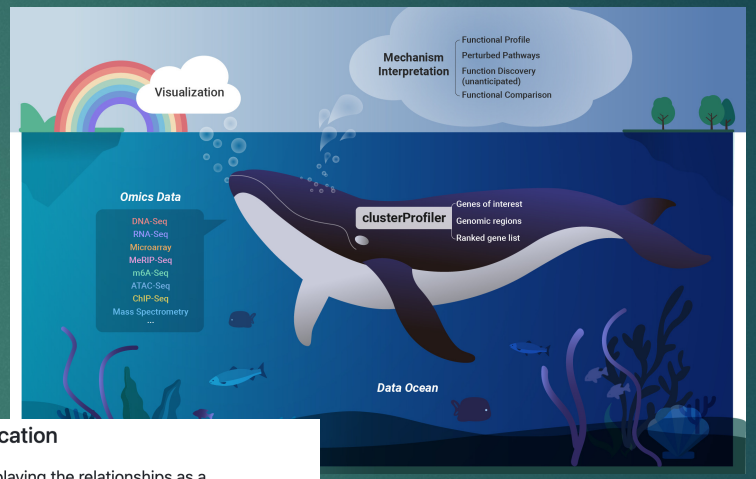
Pathway & Gene Set Analysis

https://toppgene.cchmc.org/enrichment.jsp

ToppFun

Clusterprofiler
https://yulab-smu.top/biomedical-knowledge-mining-book/





15.4 **Heatmap-like functional classification**

The `heatplot` is similar to `cnetplot`, while displaying the relationships as a heatmap. The gene-concept network may become too complicated if user want to show a large number significant terms. The `heatplot` can simplify the result and more easy to identify expression patterns.

```
p1 <- heatplot(edox, showCategory=5)
p2 <- heatplot(edox, foldChange=geneList, showCategory=5)
cowplot::plot_grid(p1, p2, ncol=1, labels=LETTERS[1:2])
```

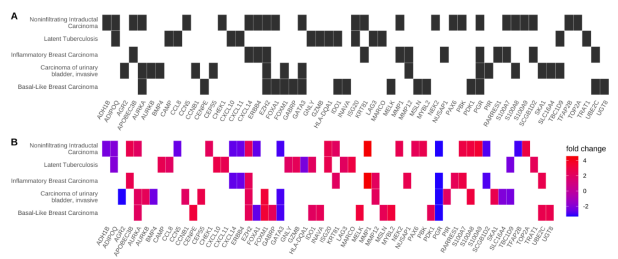
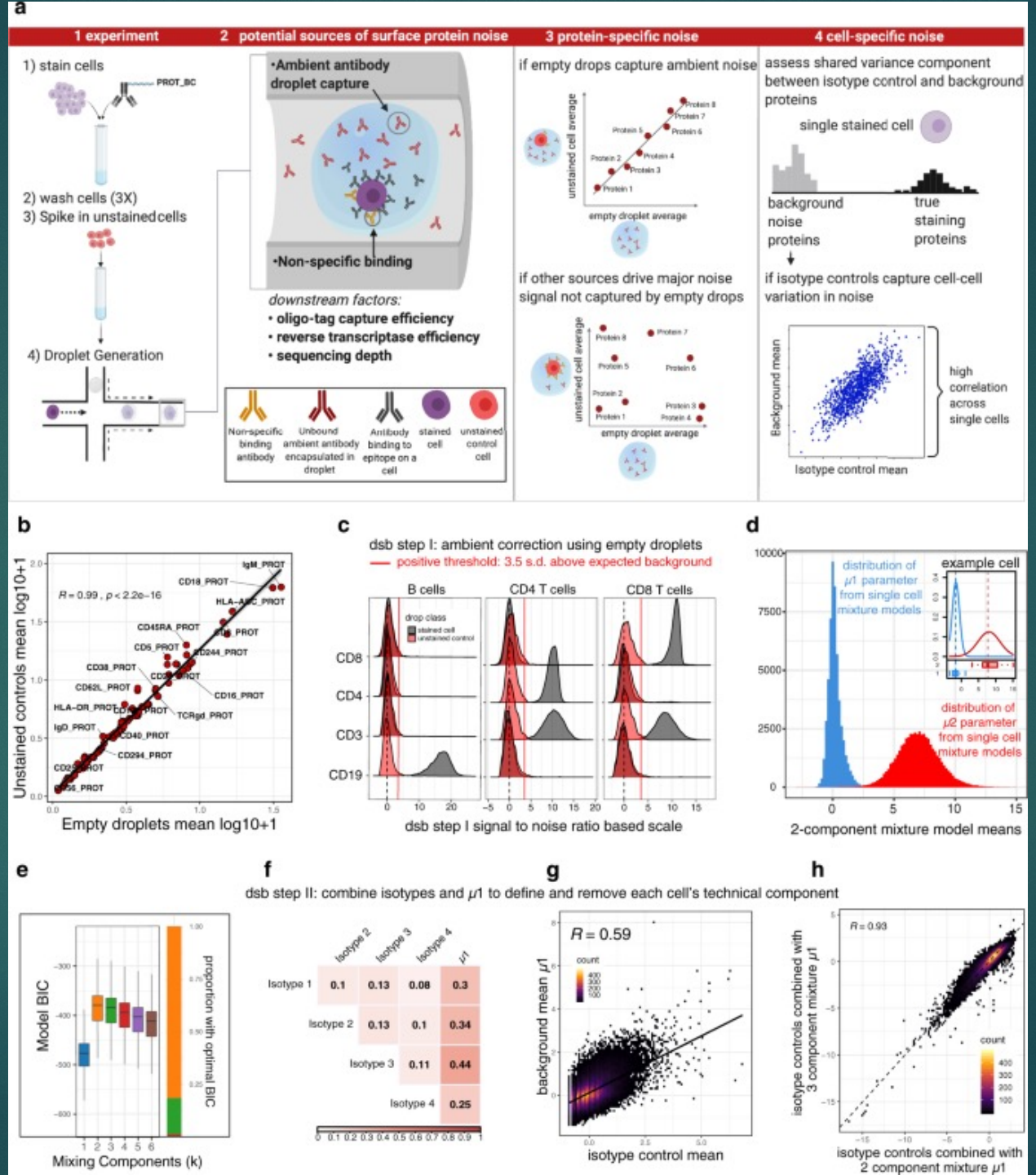


Figure 15.6: **Heatmap plot of enriched terms.** default (A), `foldChange=geneList` (B)

CITE-seq Data

- 3 main types of normalization:
  - 1. CLR (centered log ratio) in Seurat, with margin=1, plots low to high for each features
  - CLR in Seurat, with margin 2, requires the assumption that each cell is stained with roughly the same amount of antibodies- normalizes per cell
  - Denoised & Scaled by Background (dsb) (separate package)
    - This requires the raw output with empty droplets from 10x to specify a background
    - Recommended to have isotype controls as well

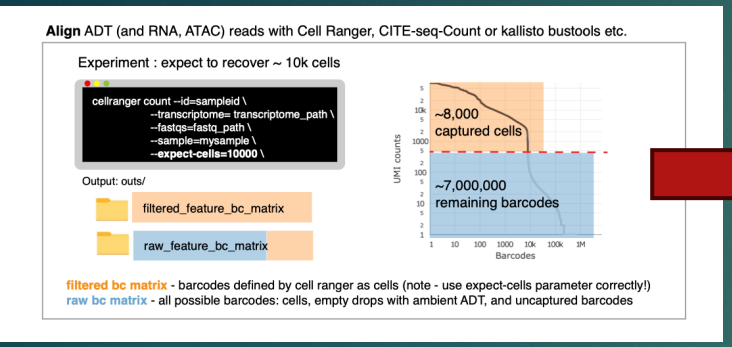
DSB

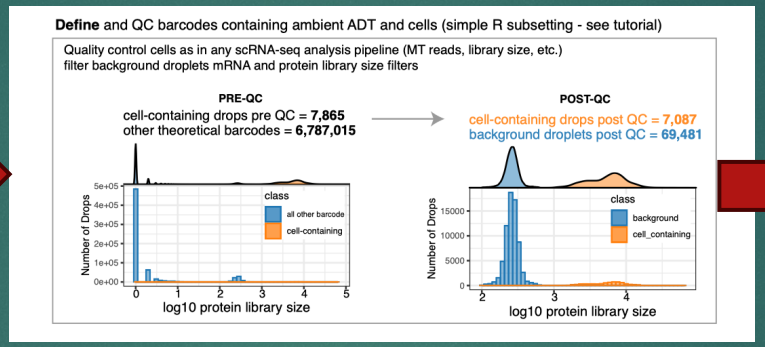


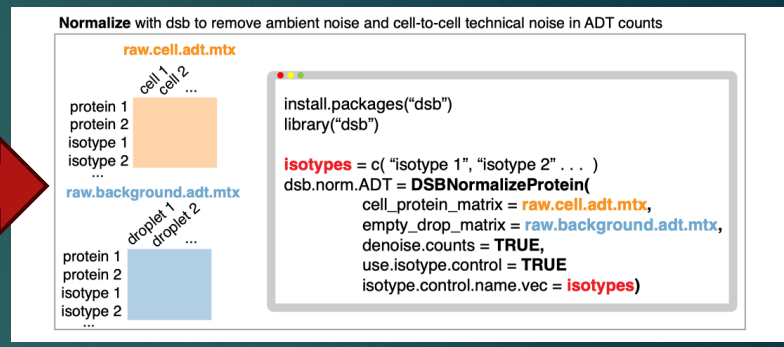
https://github.com/niaid/dsb

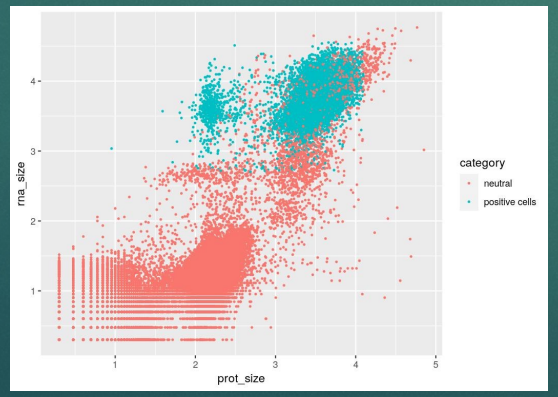
Mule et al., 2022

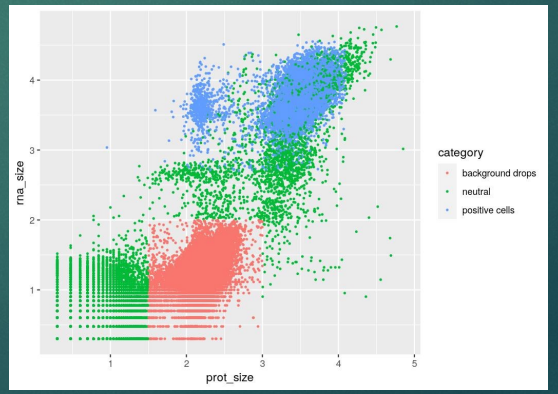
DSB Workflow











Final Thoughts

- Don’t be afraid to not know
- Everyone started at ground zero sometime
- Think of coding as learning how to use a computer- each software/package has its’ own set of quirks
- Not everything published works in real life