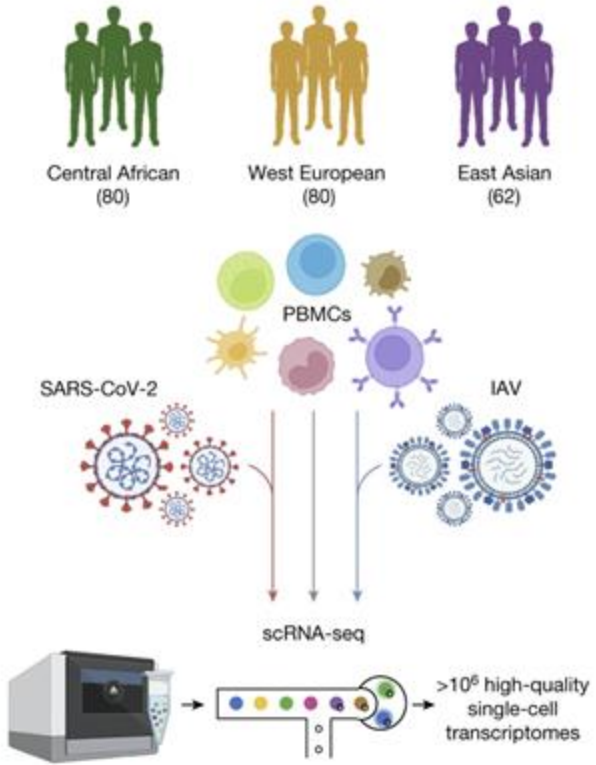
# Batch Effects

Technical, non-biological factors that affect variation in data

Mary O'Neill, Ph.D.
Director, Single Cell Genomics
ONeillMB@uw.edu
@ONeillMB1

Anh Vo, M.S.
Bioinformatician
athuyvo@uw.edu

# Lessons from processing >1M PBMCs



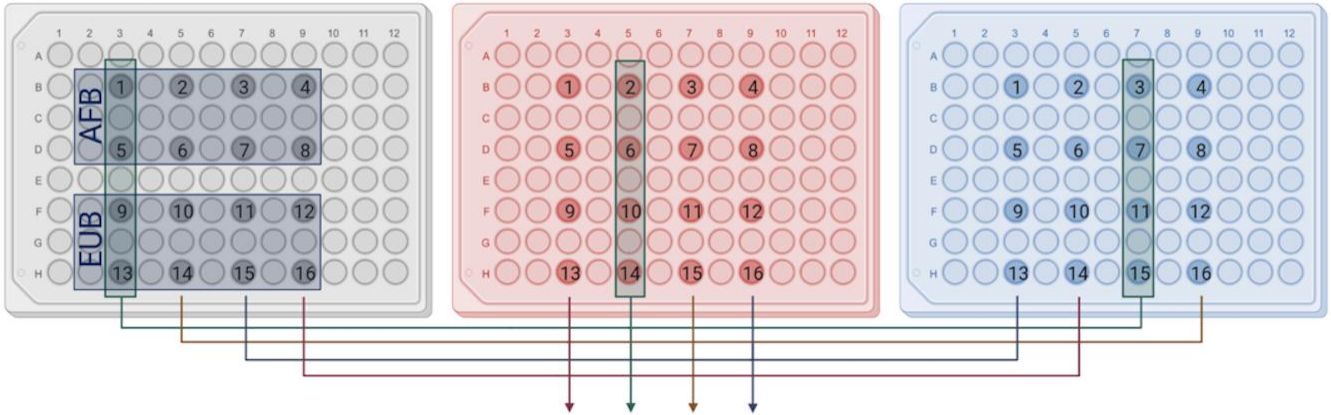

1. Characterize variability of the immune response to SARS-CoV-2 across human populations at single cell resolution (scRNA-seq)

2. Map genetic bases of immune variability in response to SARS-CoV-2 (eQTLs)

3. Uncover natural selection and archaic introgression signals associated to virus-induced immune reponses

Aquino*, Bisiaux*, Li*, O'Neill*, *et al.* 2024 (Nature)

# Library design:




2 runs per week (16 runs total)
16 individuals/8 libraries
per experimental run

Each library contains 12 samples
(4 from each condition)

Each sample is done on two separate libraries

2•

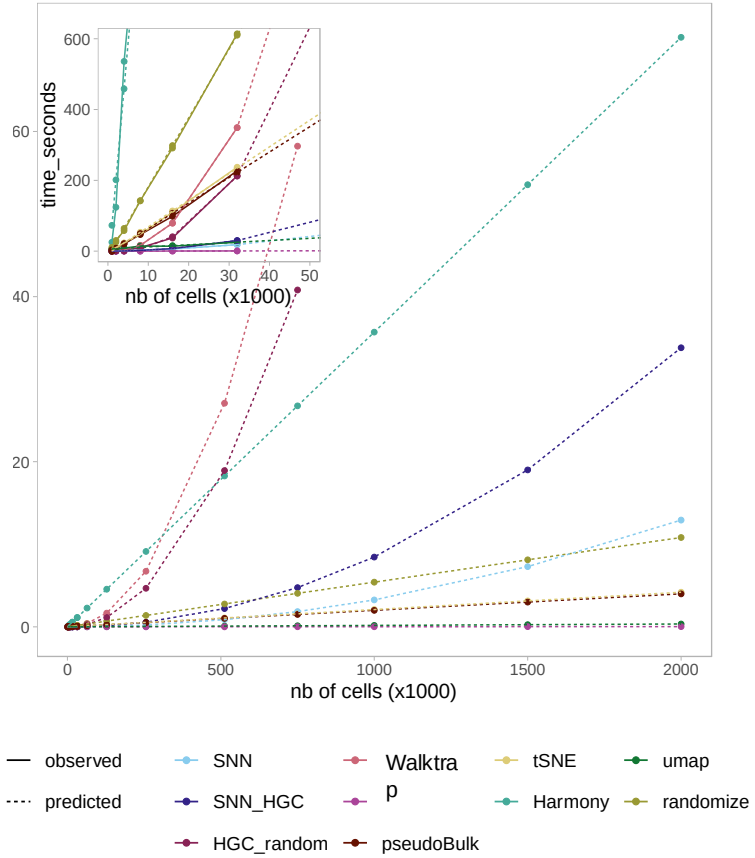
| | L1,L5 | L2,L6 | L3,L7 | L4,L8 |
|---|---|---|---|---|
| NS | 4,8<br>12,16 | 1,5<br>9,13 | 2,6<br>10,14 | 3,7<br>11,15 |
| COV | 1,5<br>9,13 | 2,6<br>10,14 | 3,7<br>11,15 | 4,8<br>12,16 |
| IAV | 2,6<br>10,14 | 3,7<br>11,15 | 4,8<br>12,16 | 1,5<br>9,13 |

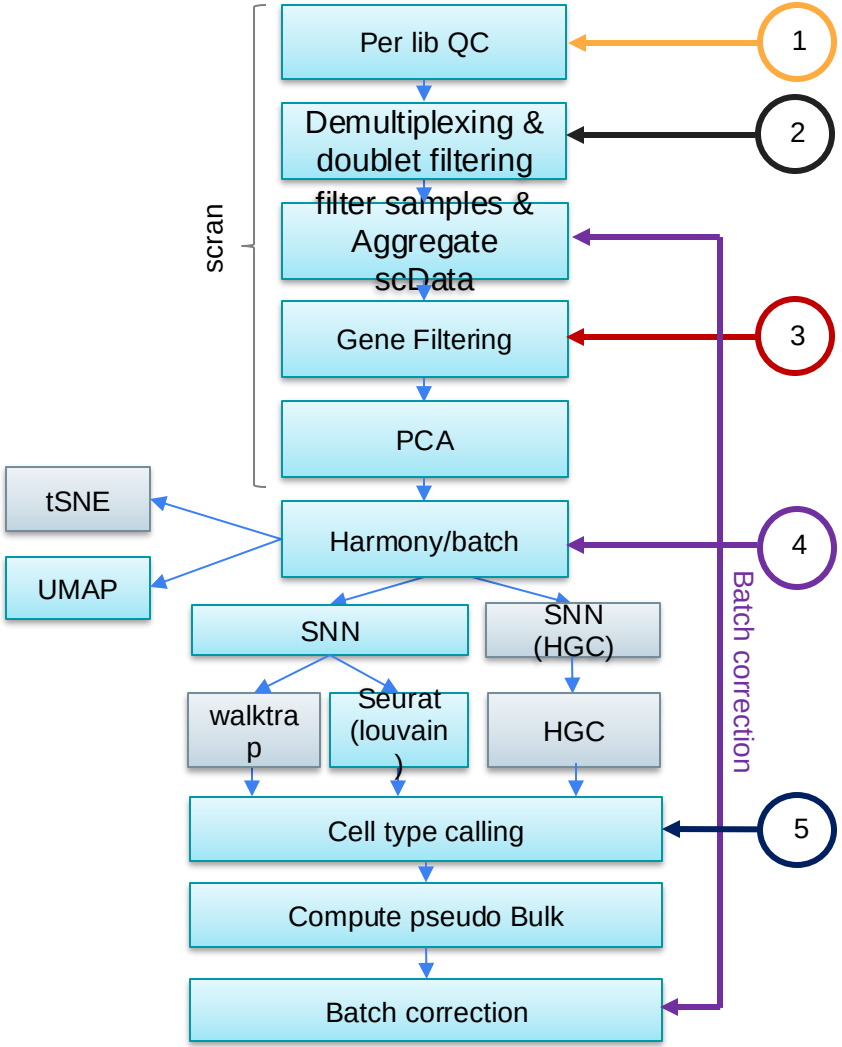
Pooling design

**target:**
1,667 cells per individual/condition

**After QC**:
~1500 cells on average
(median; IQR=558 cells)

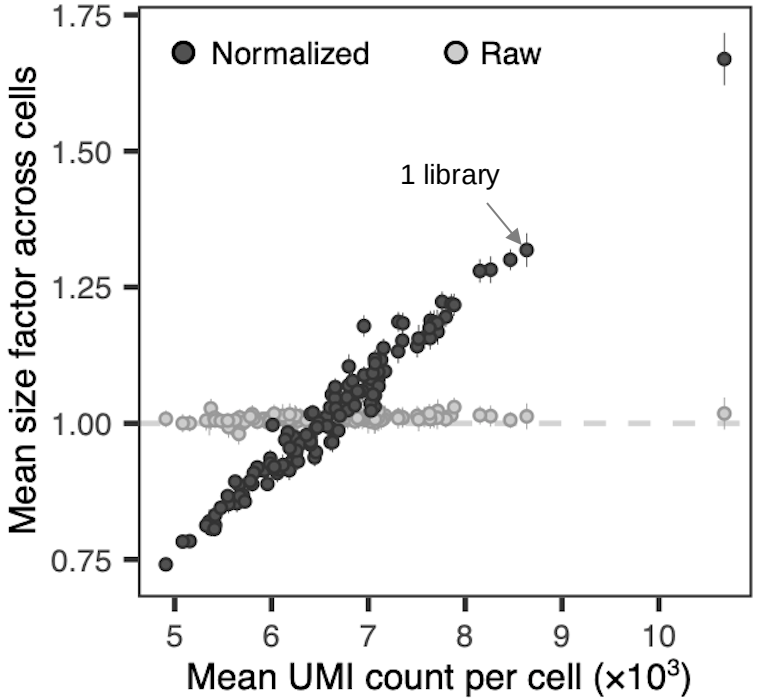
# General pipeline overview



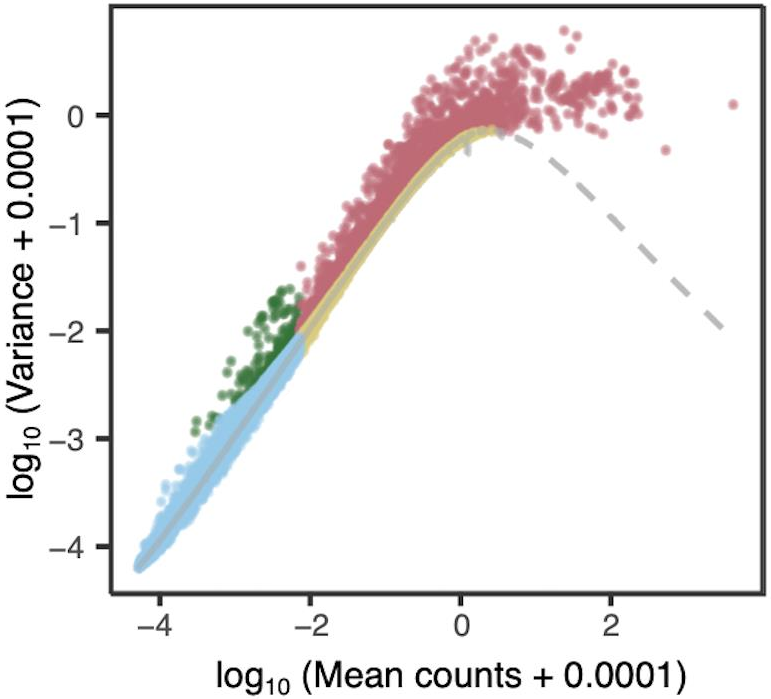
# Size factor normalization & Batch effect removal:

Use `MultibatchNorm` to allow for differences in mean size factor across libraries

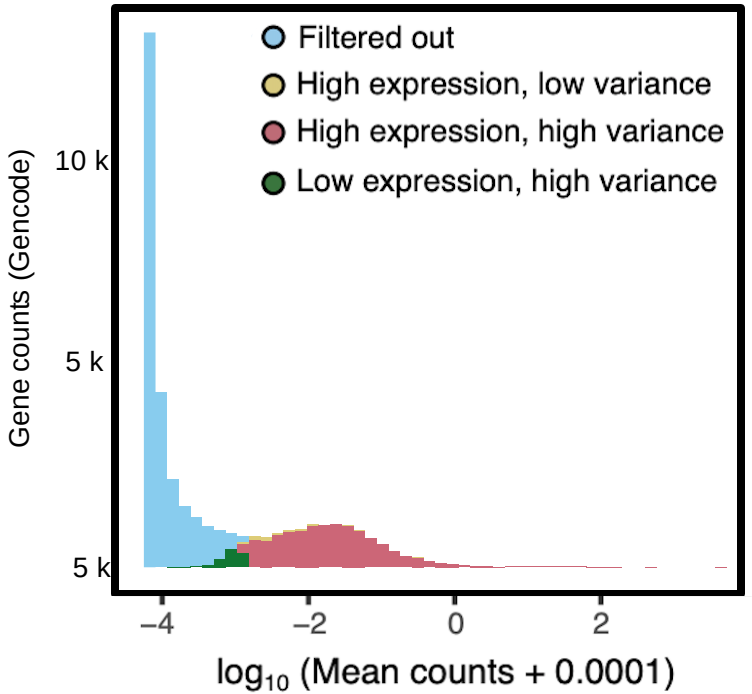


# Gene Filtering :

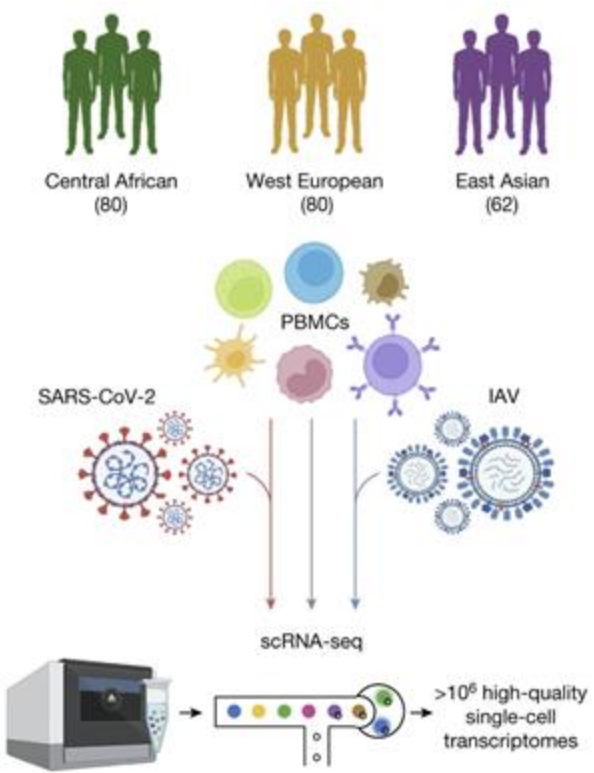
Use `modelGeneVarByPoisson` to estimate biological/technical variance

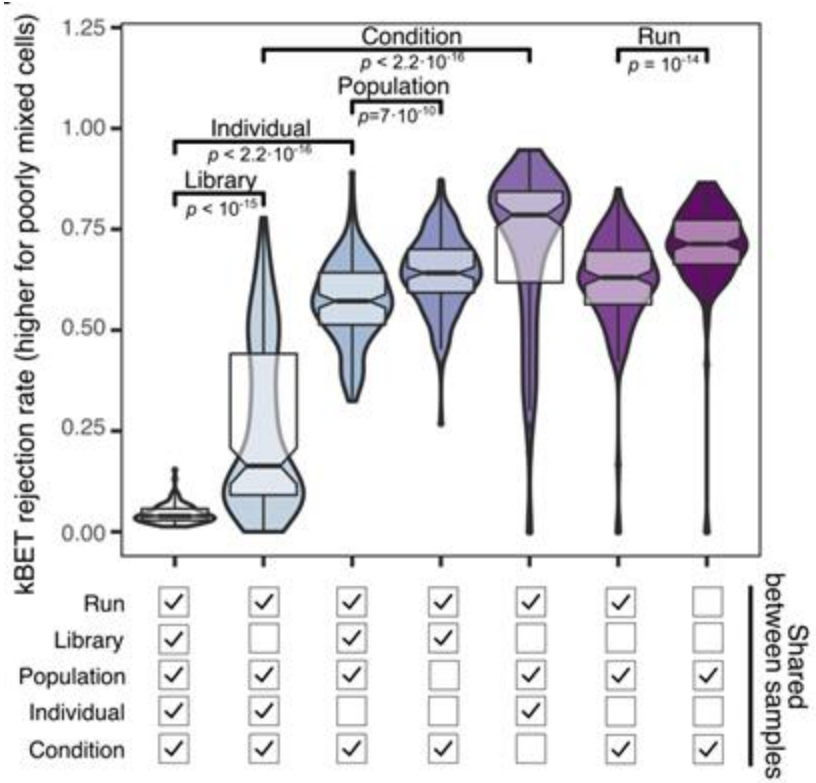



Keep genes with either high expression/high variance



# Sources of variation




Aquino*, Bisiaux*, Li*, O'Neill*, *et al.* 2024 (Nature)

# How to deal with batch effects

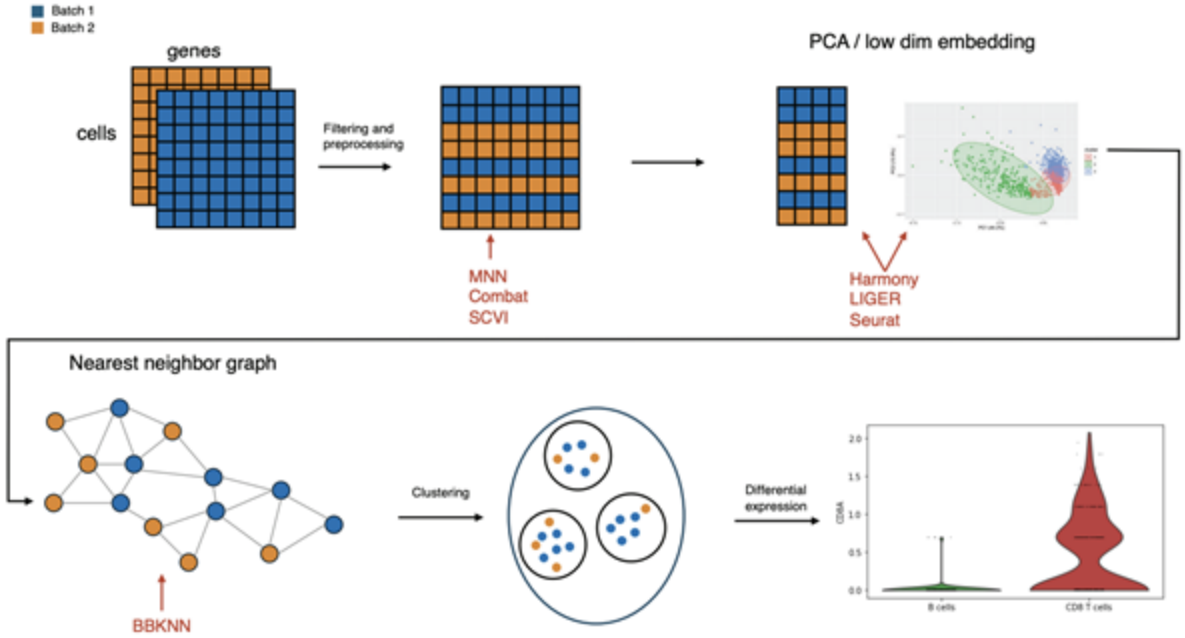
- Best way to avoid batch effects is not to introduce them in the first place!
- If unavoidable, it is very important that study design does not confound batch and other variables
  - Nothing can salvage poor study design
- Make sure you have a batch effect
  - Sometimes (often?) batch correction introduces more artifacts than they alleviate
- Apply methods thoughtfully
  - Don’t blindly trust methods
  - Know what they are doing, what to use them for, and where they can lead you astray

# Batch Correction Methods

| | BBKNN | Combat | Harmony | LIGER | MNN | Seurat | SCVI |
|---|---|---|---|---|---|---|---|
| ***Input*** | KNN graph | Normalized count matrix | Normalized count matrix | Normalized count matrix | Normalized count matrix | Normalized count matrix | Raw count matrix |
| ***Custom embedding*** | None | None | Corrected embedding | Metagene / factor loadings | None | CCA | Learned lower dimensional latent space |
| ***Correction object*** | KNN graph | Count matrix | Embedding | Embedding | Count matrix | Embedding | Embedding |
| ***Correction method*** | Umap on merged neighborhood graph | Empirical bayes - linear correction method on the count values | Soft k-means - linear batch correction within small clusters in the embedded space | Quantile alignment of factor loadings | Mutual nearest neighbors - linear correction | Aligning canonical basis vectors to correct the embedding - lift the correction of the embedding to count space | Variational autoencoder - models the batch effect in a low dimensional space using a deep learning model, a new count matrix is imputed from the model. |
| ***Returns*** | Corrected KNN graph | Corrected count matrix | Corrected embedding | Corrected embedding | Corrected count matrix | Corrected count matrix | Corrected count matrix and corrected embedding |
| ***Changes Count matrix*** | No | Yes | No | No | Yes | Yes | Yes / Imputes new values |

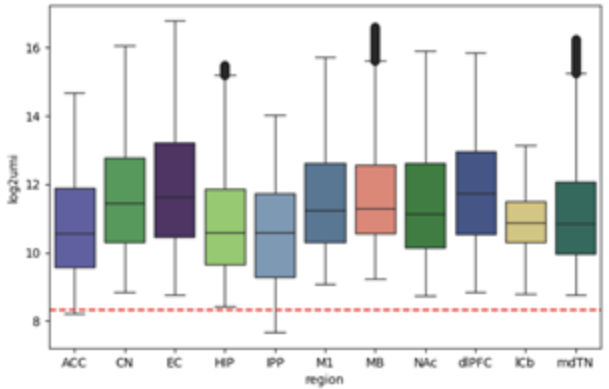
Antonsson & Melsted, 2024 (BioXriv)

# Batch Correction Methods



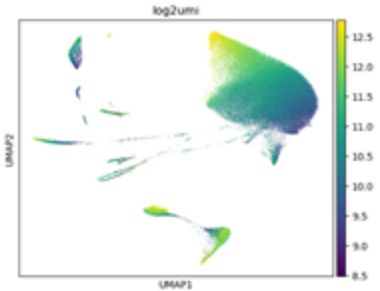

Antonsson & Melsted, 2024 (BioXriv)

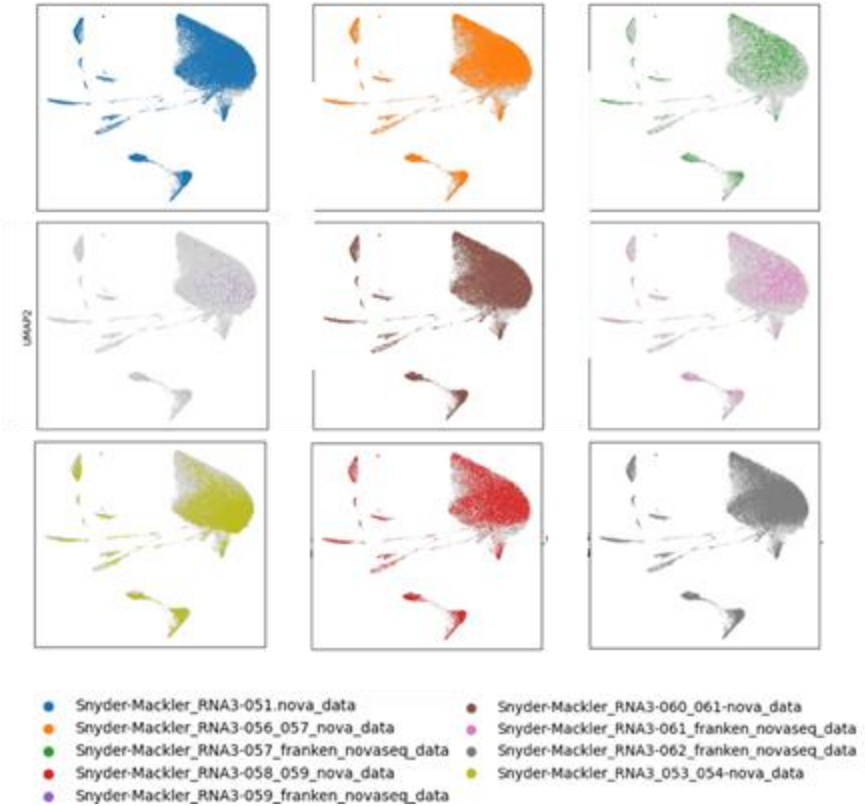
# Minimal batch effect in sci-RNA-seq!




BICCN atlas

**Meidan UMI = 1,918 (~ 6x times higher than BICCN)**



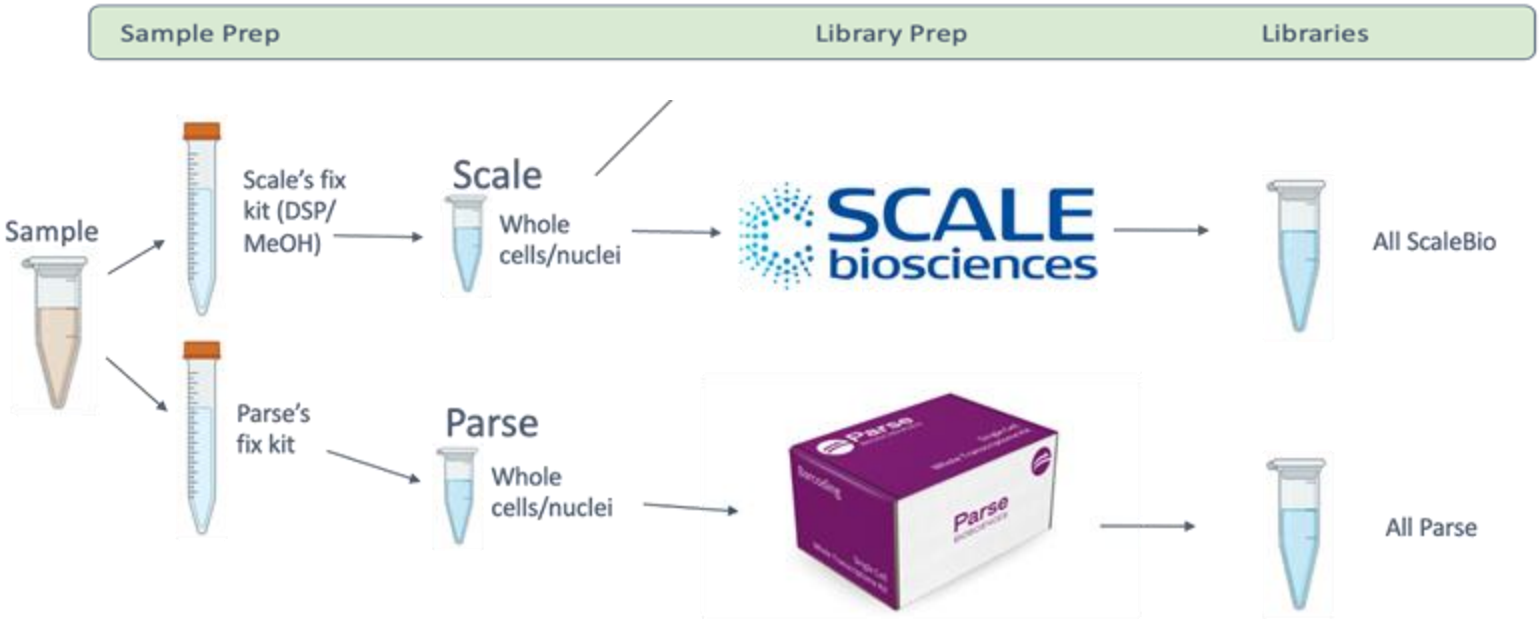




Wei Yang, Unpublished

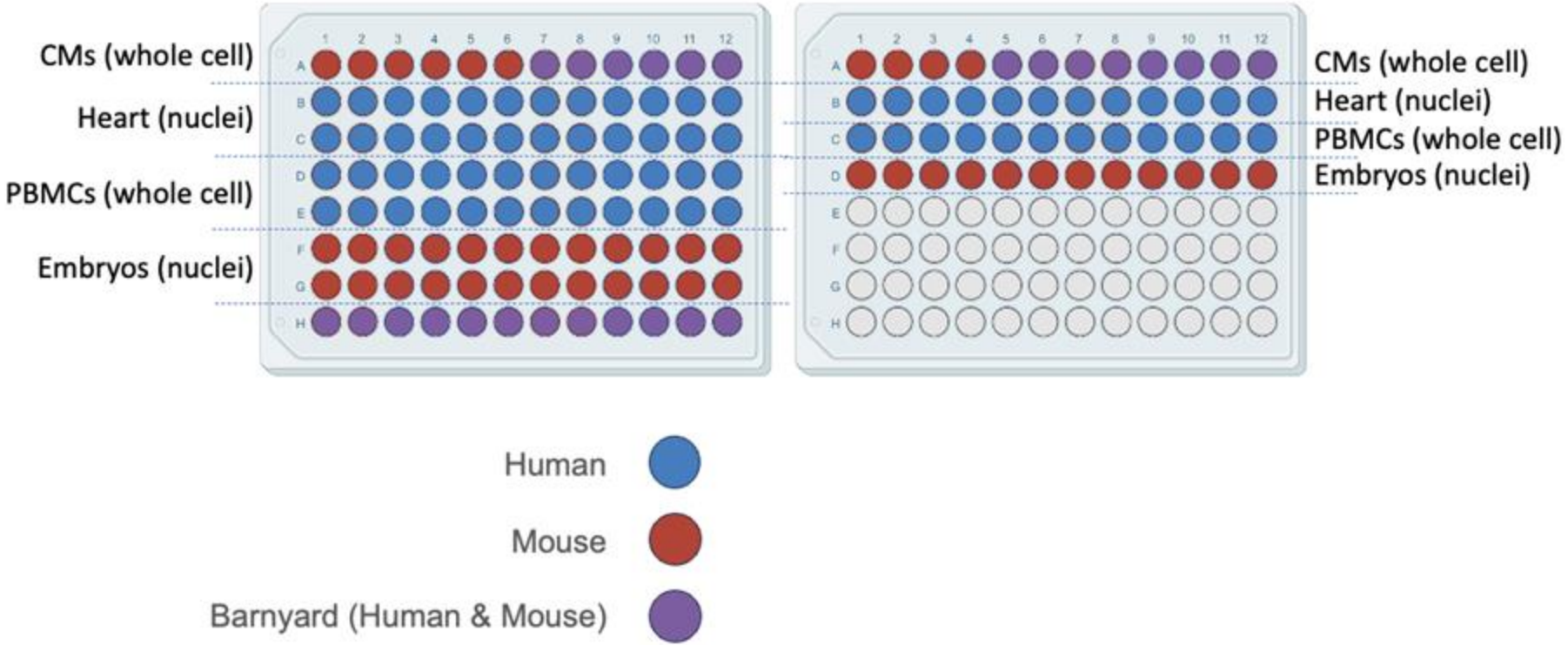
# Tutorial

Background on the dataset

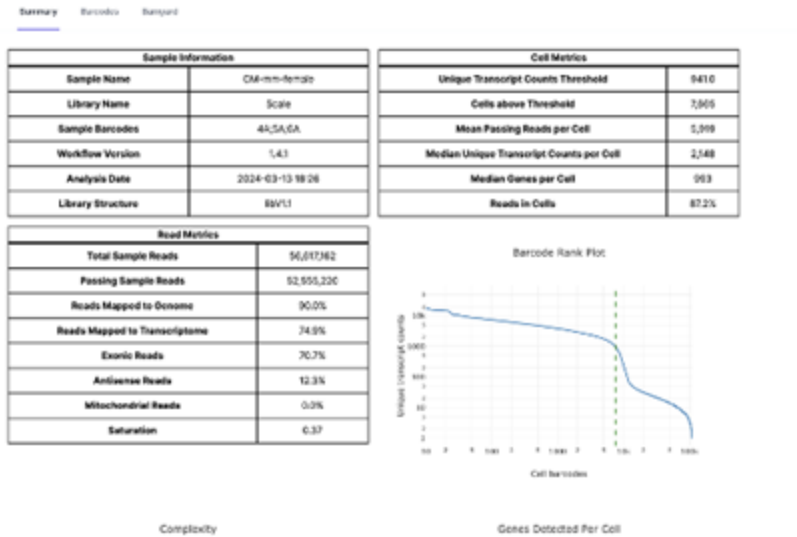
# Study Design



# Study Design

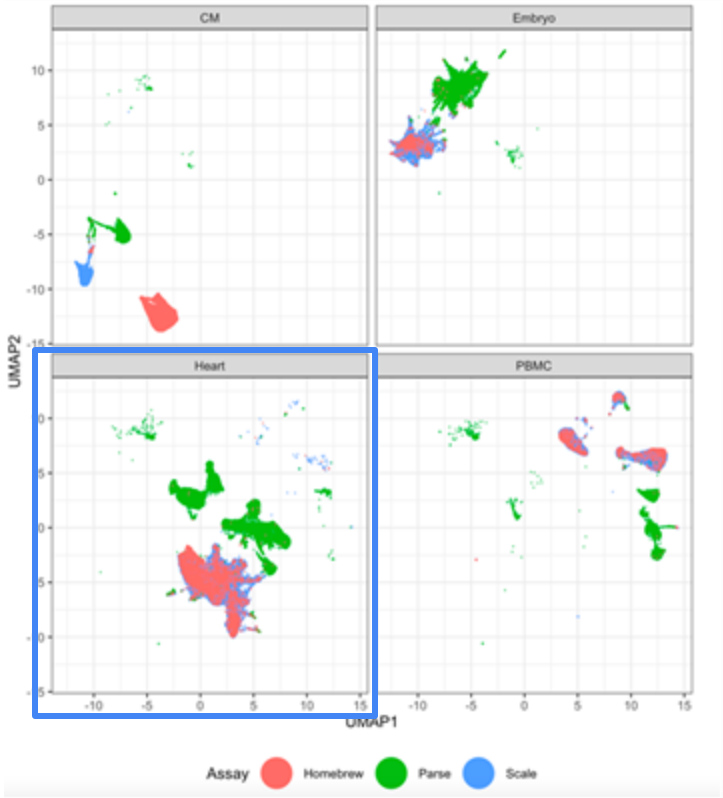


# Each company has a pipeline


- ***Aggressively*** call cells – inflate UMIs?
- Both pipelines cut off almost an entire cell population (e.g. PBMCs) in barnyard samples
  - So, we start with every barcode >= 100 UMIs
- Eventually, will try to use a near universal pipeline. For now, using respective pipelines.

# Will play with human heart data today




# Thank you!