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

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

How to deal with batch effects

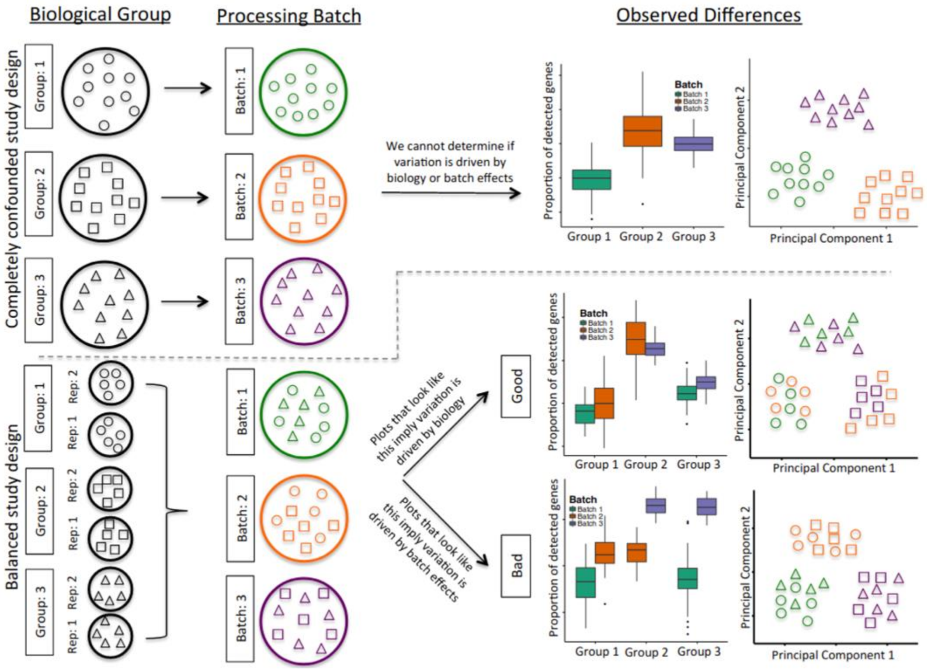
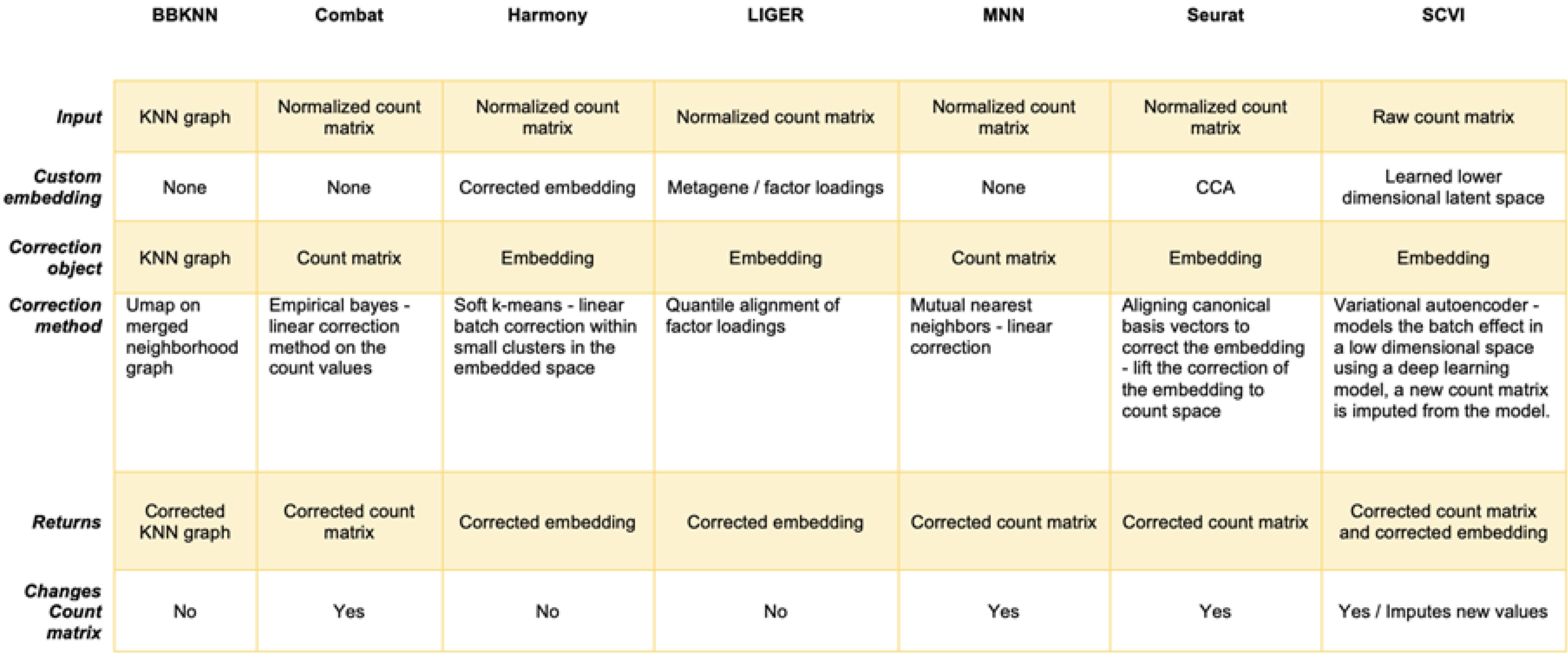

Biological Group
Processing Batch
Observed Differences
Completely confounded study design
Group: 1
Group: 2
Group: 3
Batch: 1
Batch: 2
Batch: 3
We cannot determine if variation is driven by biology or batch effects
Proportion of detected genes
Batch
Batch 1
Batch 2
Batch 3
Group 1
Group 2
Group 3
Principal Component 2
Principal Component 1
Balanced study design
Rep: 1
Rep: 2
Plots that look like this imply variation is driven by biology
Plots that look like this imply variation is driven by batch effects
Good
Bad

Illustration of a confounded (top panels) and balanced (bottom panels) designs. Shapes denote different sample types (e.g. tissues or patients) and colours processing batches. In the confounded design it's impossible to disentangle biological variation from variation due to the processing batch. In the balanced design, by using tissue replicates and mixing them across batches, it is possible to distinguish between biological and batch-related variation. Figure from Hicks et al..

https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

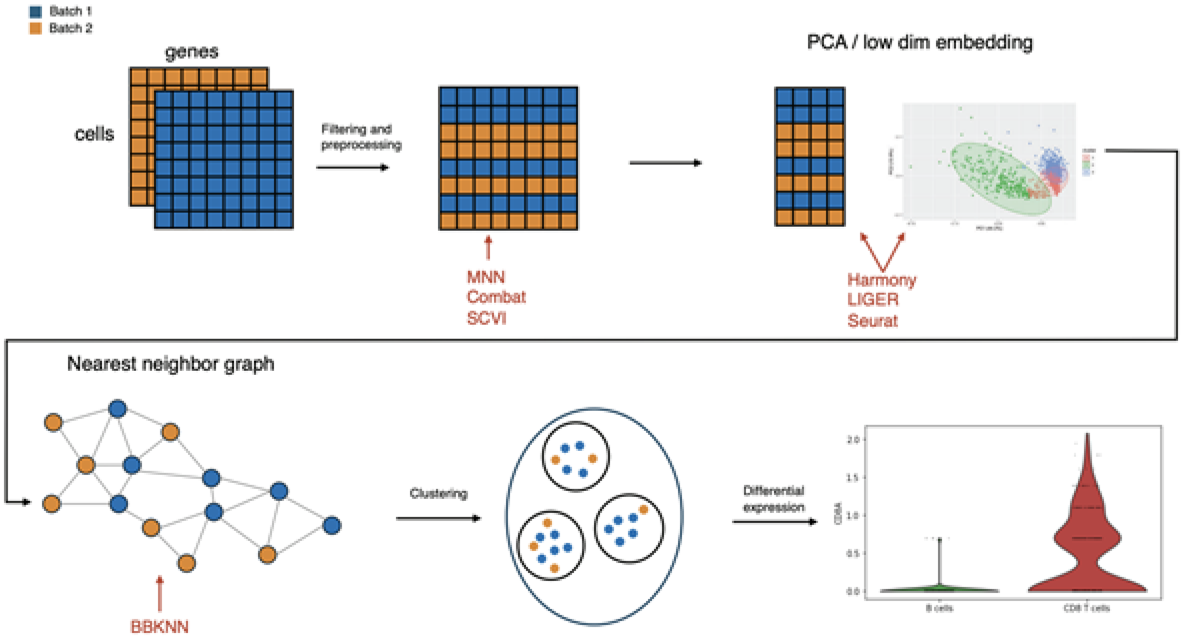
Batch Correction Methods



	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



🟦 Batch 1
🟧 Batch 2

genes

cells

Filtering and preprocessing

MNN
Combat
SCVI

PCA / low dim embedding

Harmony
LIGER
Seurat

Nearest neighbor graph

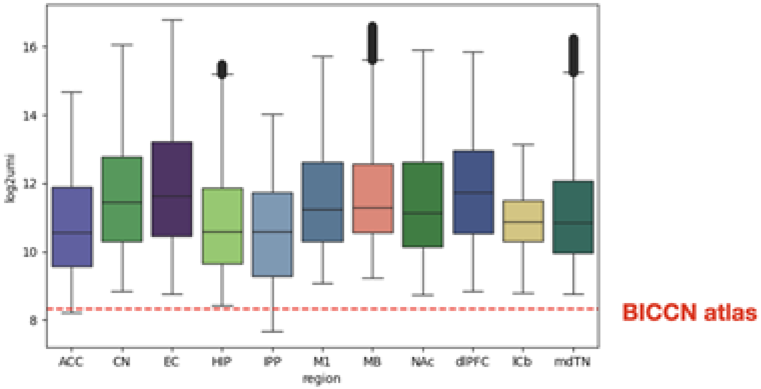
BBKNN

Clustering

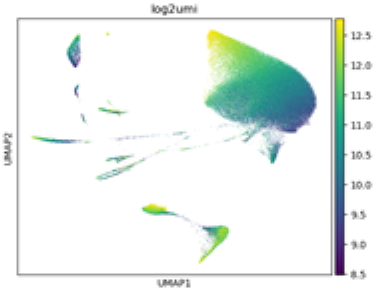
Differential expression

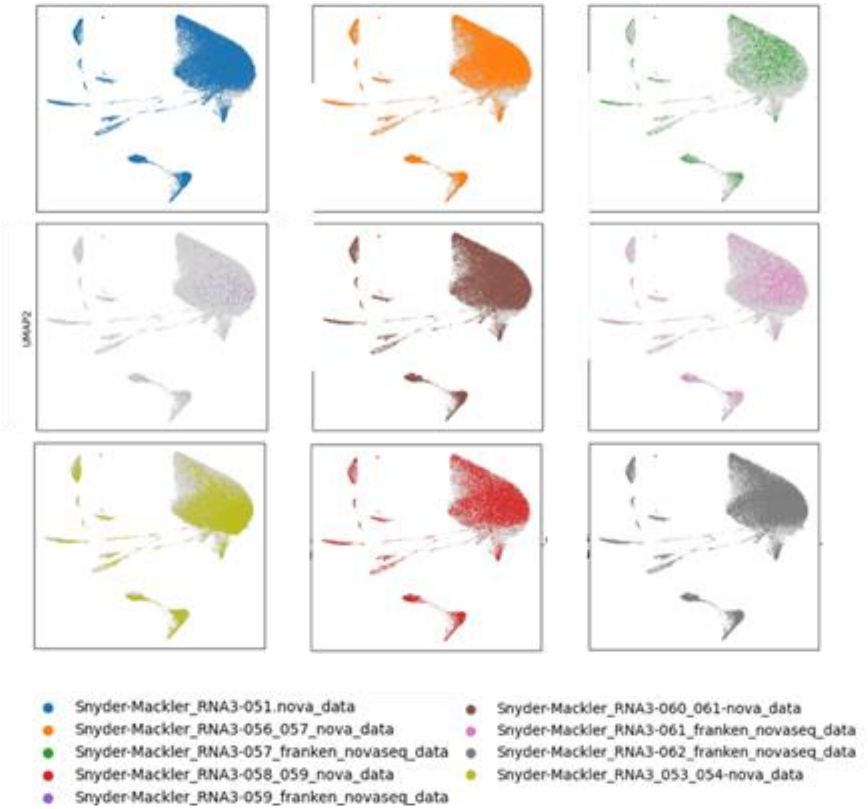
Antonsson & Melsted, 2024 (BioXriv)

Sidenote: we see minimal batch effects in sci-RNA-seq!



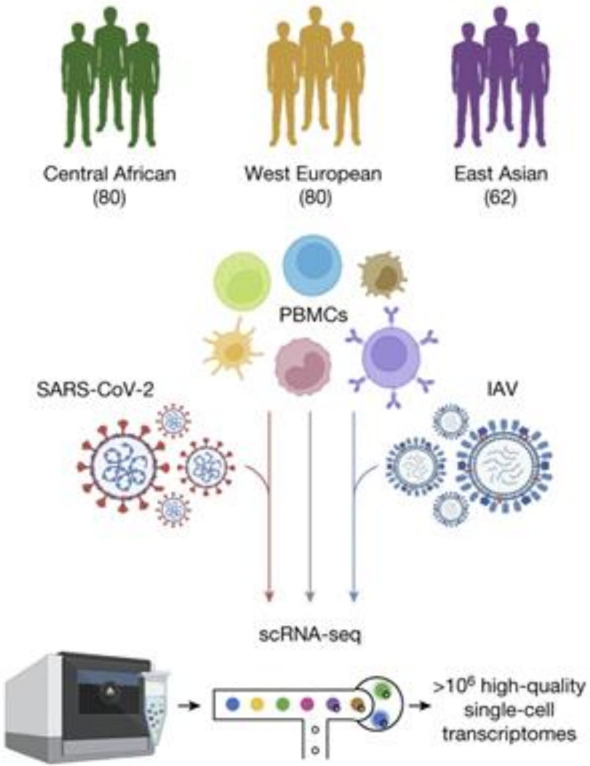
Meidan UMI = 1,918 (~ 6x times higher than BICCN)





Wei Yang, Unpublished

Lessons from processing >1M PBMCs



AIMS

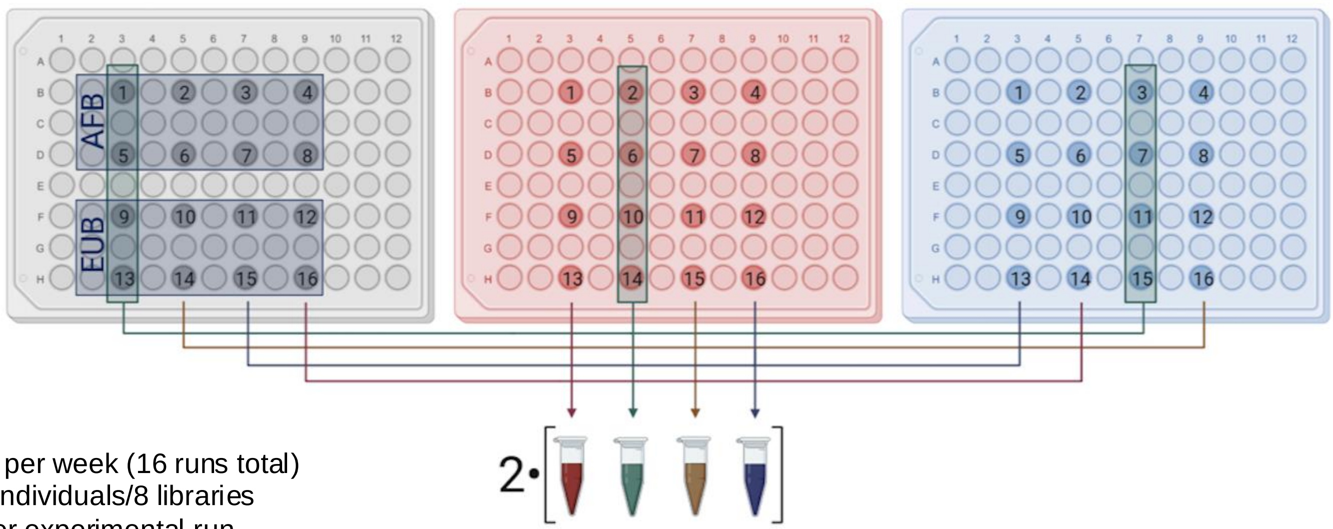
1. Characterize variability of the immune response to SARS-CoV-2 across human populations at single cell resolution (scRNA-seq)

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:



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

Pooling design

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

Each library contains 12 samples
(4 from each condition)

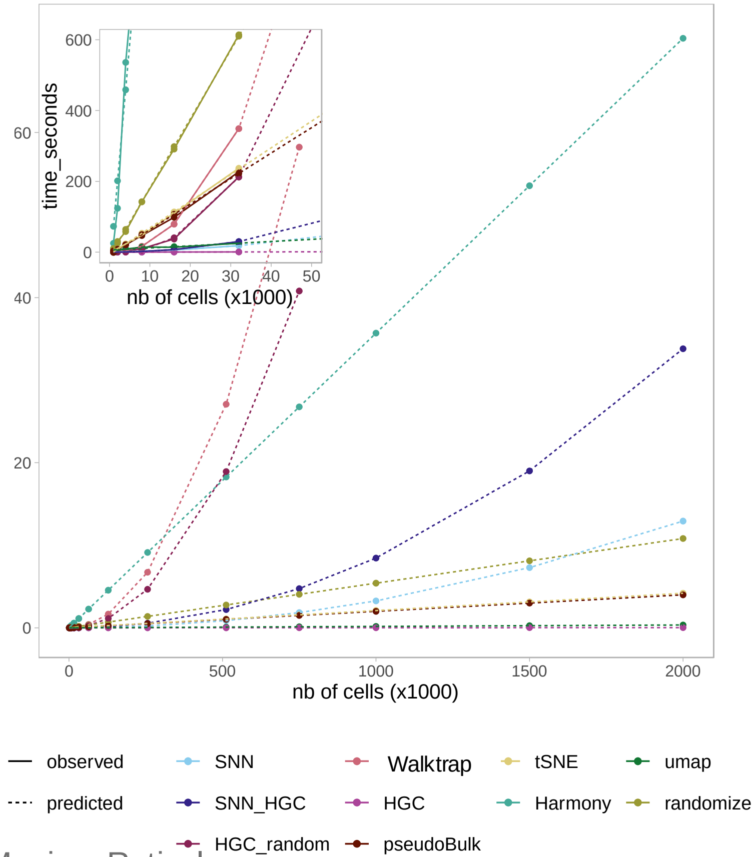
Each sample is done on two separate libraries

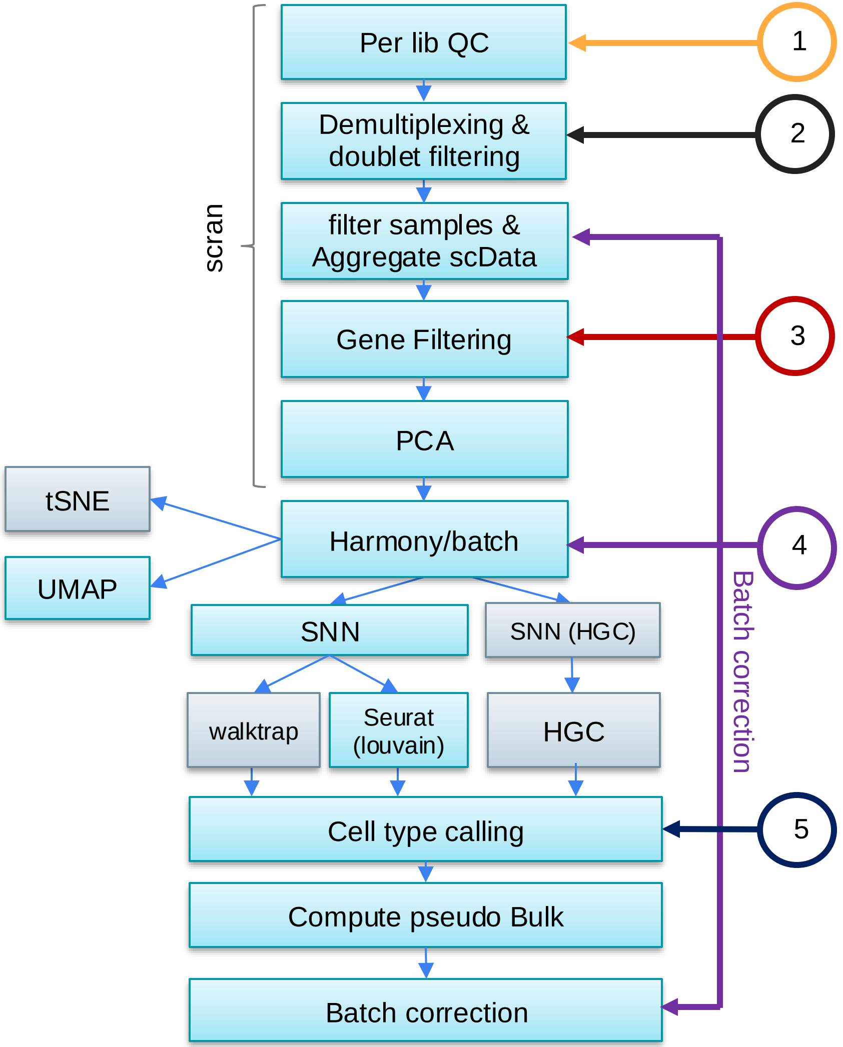
target:
1,667 cells per
individual/condition

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

Slide from Maxime Rotival

General pipeline overview





scran

Per lib QC

Demultiplexing & doublet filtering

filter samples & Aggregate scData

Gene Filtering

PCA

Harmony/batch

tSNE

UMAP

SNN

SNN (HGC)

walktrap

Seurat (louvain)

HGC

Cell type calling

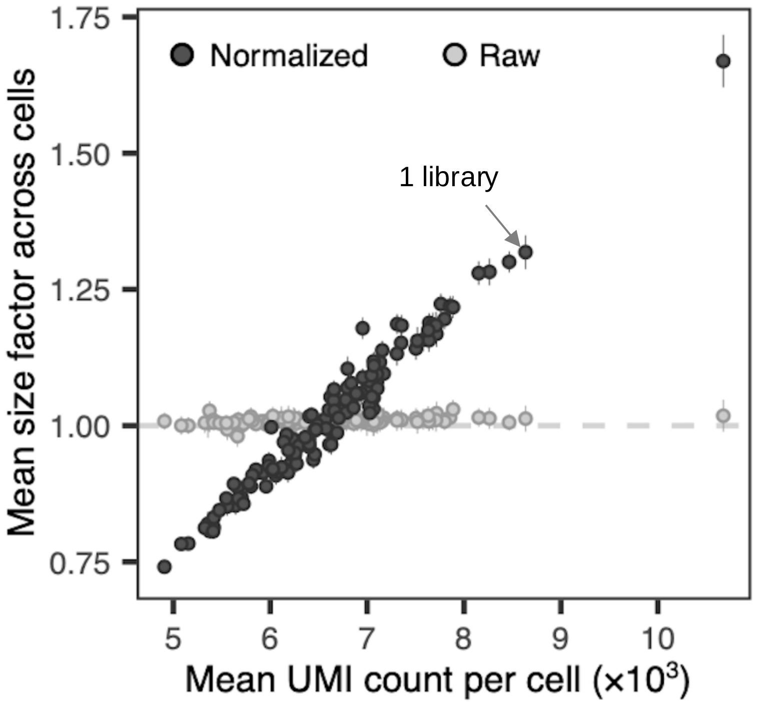
Compute pseudo Bulk

Batch correction

Slide from Maxime Rotival

Size factor normalization & Batch effect removal:

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



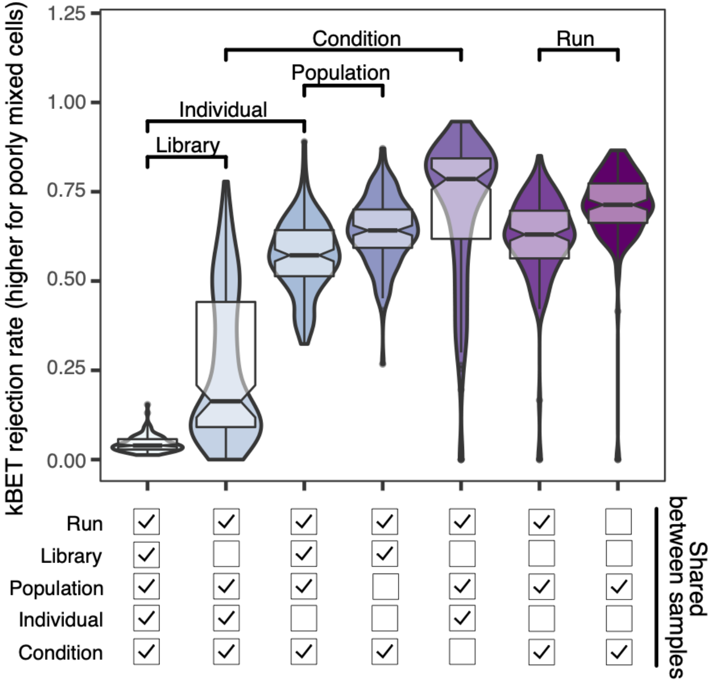
Slide from Maxime Rotival

Size factor normalization & Batch effect removal:

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

Use `kBET` to estimate batch effects on cell mixing…

…and `Harmony` to correct for batch effects across experimental runs (for cell clustering purposes)



Run	☑	☑	☑	☑	☑	☑	□
Library	☑	□	☑	☑	□	□	□
Population	☑	☑	☑	□	☑	☑	☑
Individual	☑	☑	□	□	☑	□	□
Condition	☑	☑	☑	☑	□	☑	☑

Slide from Maxime Rotival

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

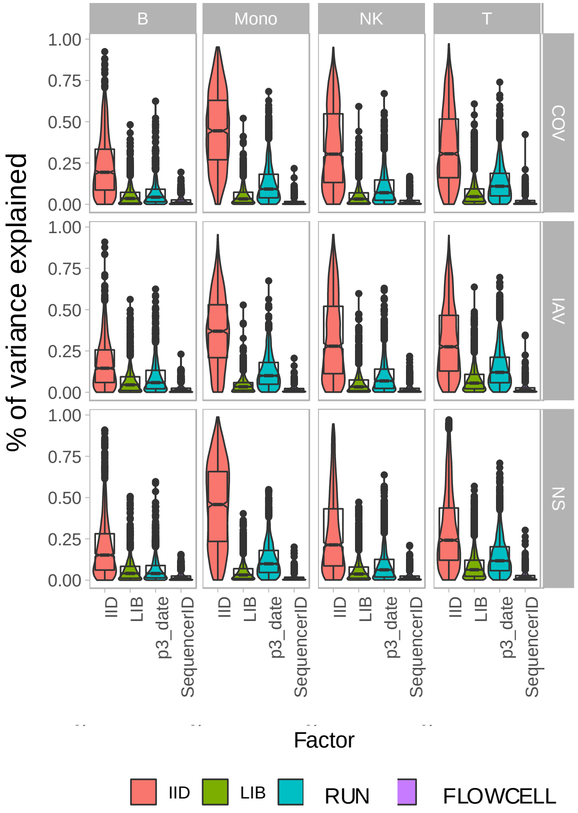
Size factor normalization & Batch effect removal:

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

Use kBET to estimate batch effects on cell mixing…

…and Harmony to correct for batch effects across experimental runs (for cell clustering purposes)

Use linear mixed models (lme4) to estimate & correct for batch effects at pseudobulk level
(differential expression/ eQTL mapping)

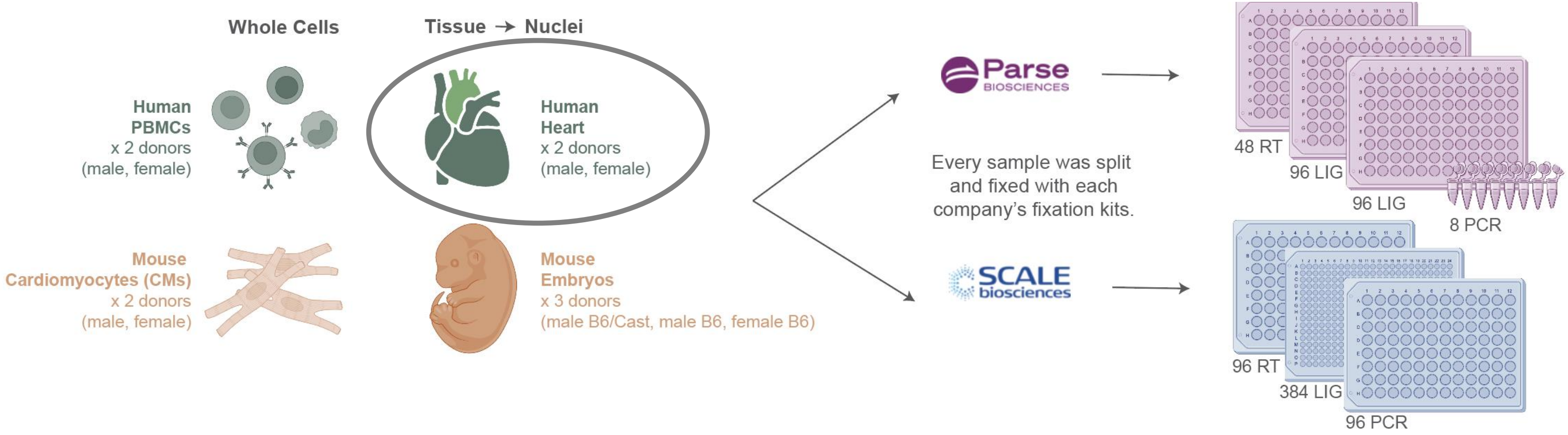


Slide from Maxime Rotival

Tutorial

Background on the dataset

Combinatorial-indexing benchmarking (generated *in-house*)



Whole Cells

Tissue → Nuclei

Human PBMCs
x 2 donors
(male, female)

Human Heart
x 2 donors
(male, female)

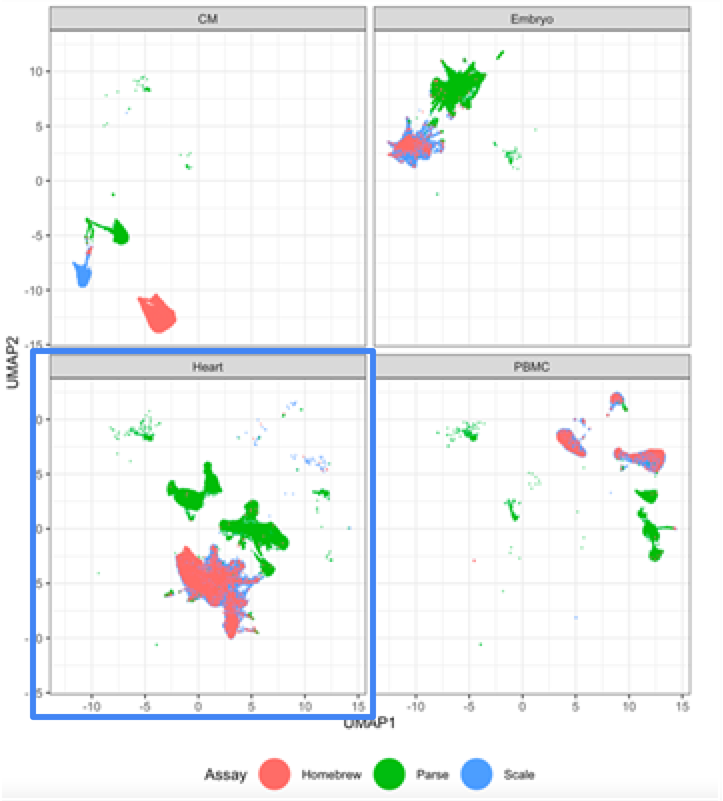
Mouse Cardiomyocytes (CMs)
x 2 donors
(male, female)

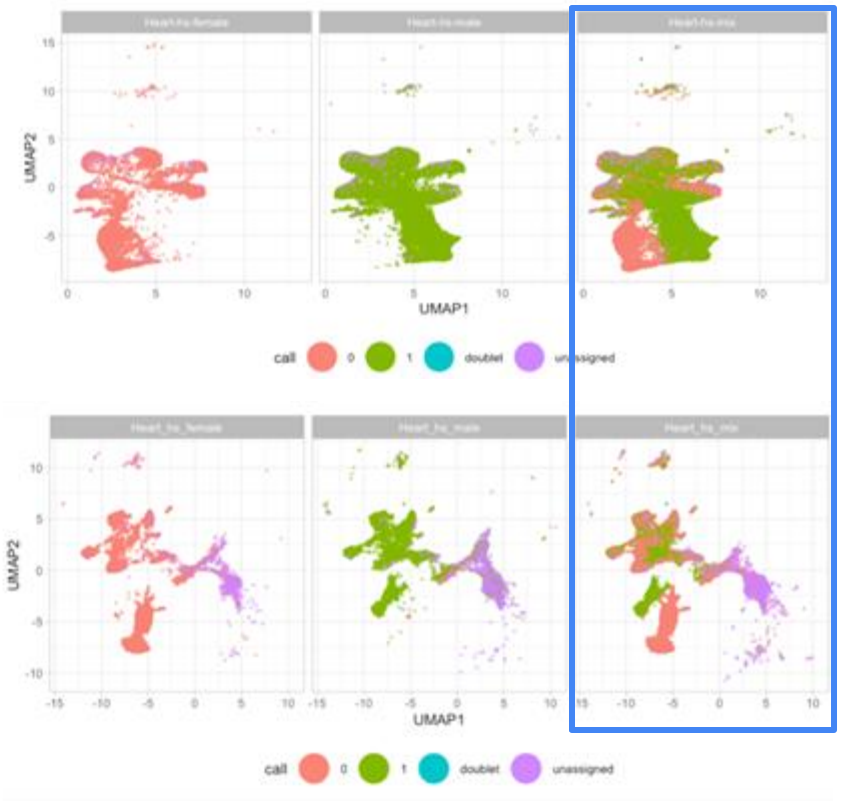
Mouse Embryos
x 3 donors
(male B6/Cast, male B6, female B6)

Every sample was split and fixed with each company's fixation kits.

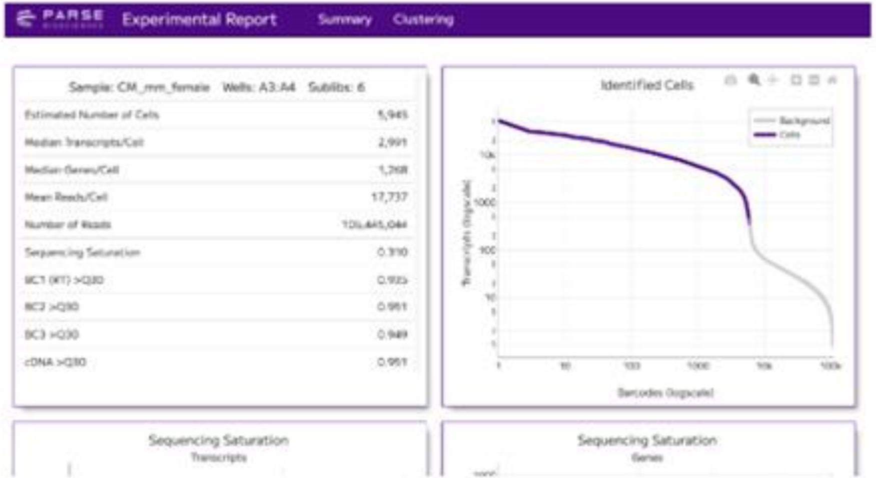
Tutorial: random subsample of mixed sample of nuclei from two human hearts, processed with two different technologies

Will play with human heart data today



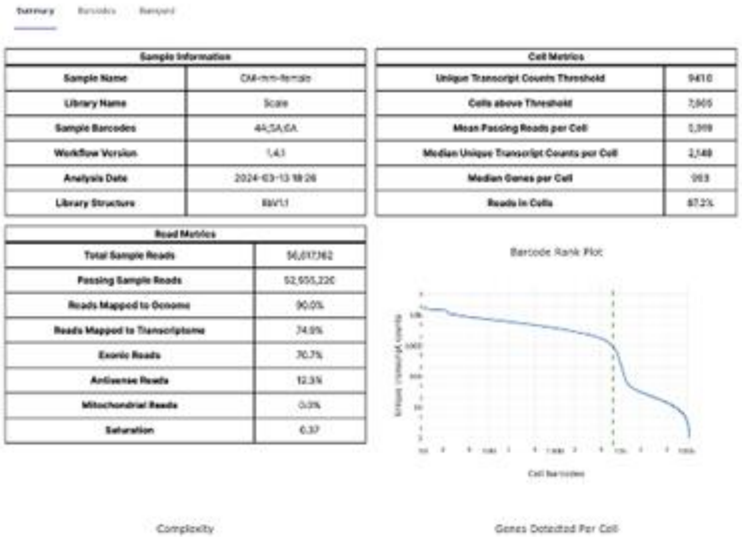


Each company has a pipeline



Sample: CM_mm_female Wells: A3:A4 Sublibs: 6

Estimated Number of Cells	5,945
Median Transcripts/Cell	2,991
Median Genes/Cell	1,268
Mean Reads/Cell	17,737
Number of Reads	105,445,044
Sequencing Saturation	0.310
BC1 (RT) >Q30	0.935
BC2 >Q30	0.951
BC3 >Q30	0.949
cDNA >Q30	0.951



Sample Information	
Sample Name	CM-mm-female
Library Name	Scale
Sample Barcodes	4A,5A,6A
Workflow Version	1.4.1
Analysis Date	2024-03-13 18:26
Library Structure	libV1.1

Cell Metrics	
Unique Transcript Counts Threshold	9410
Cells above Threshold	7,905
Mean Passing Reads per Cell	3,999
Median Unique Transcript Counts per Cell	2,148
Median Genes per Cell	993
Reads in Cells	67.2%

Read Metrics	
Total Sample Reads	56,017,162
Passing Sample Reads	52,655,220
Reads Mapped to Genome	90.0%
Reads Mapped to Transcriptome	74.9%
Exonic Reads	70.7%
Antisense Reads	12.3%
Mitochondrial Reads	0.3%
Saturation	0.37

- ***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