1. Interactive scRNA-Seq analysis with the Single Cell Toolkit (SCTK)

2. Decontamination of ambient RNA with DecontX

W. Evan Johnson, Ph.D. and Joshua Campbell, Ph.D.

Boston University School of Medicine

CSBC scRNA-seq working group

9/11/2019

Single Cell RNA-seq

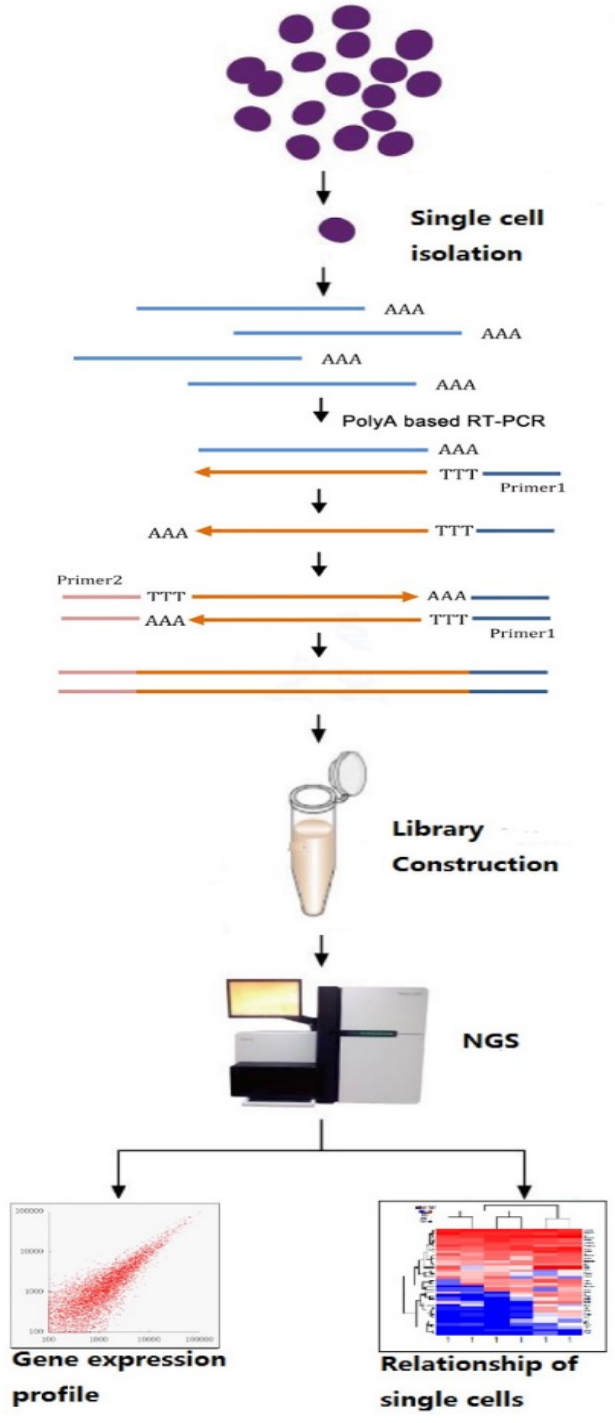
(The Future of RNA-sequencing)

Single Cell Profiling

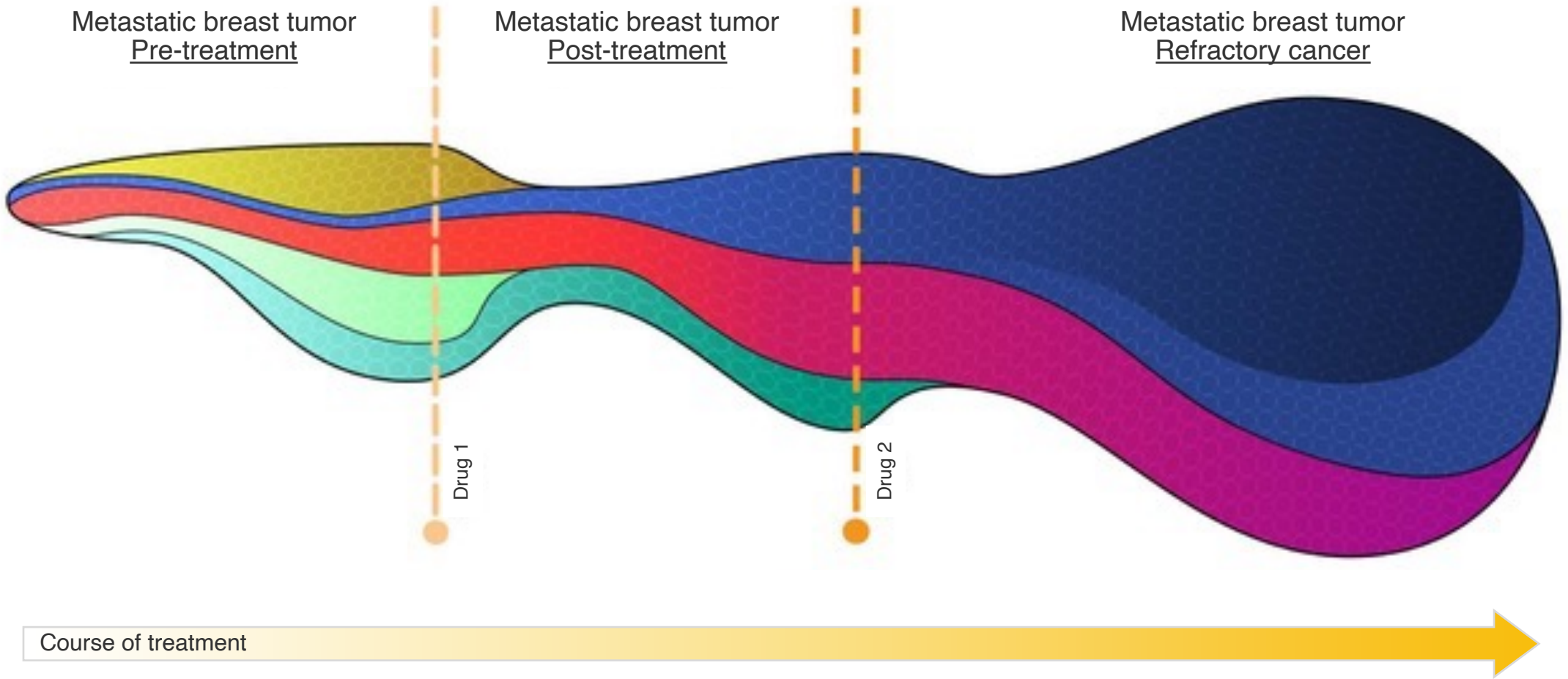
Applications for Single Cell Sequencing:

- **Cancer:** intra-tumoral heterogeneity
- **Development:** characterize every cell in blastocyst
- **Other applications:** novel/rare cell type discovery
- **Infectious Diseases:** combine with metagenomics to explore host response (e.g. TB)
- And more!

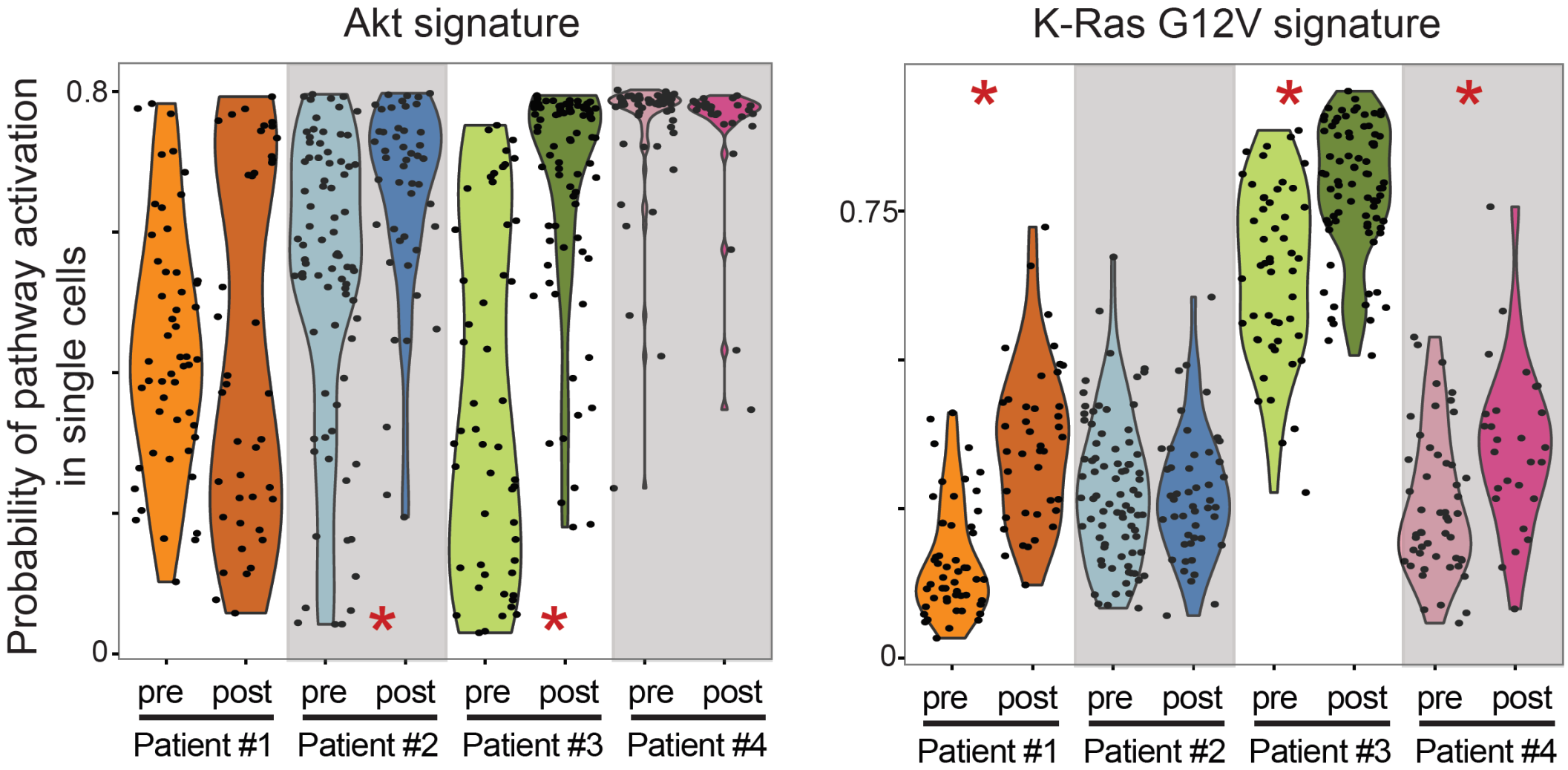
<u>The Future of RNA-sequencing</u>



Tumor Heterogeneity



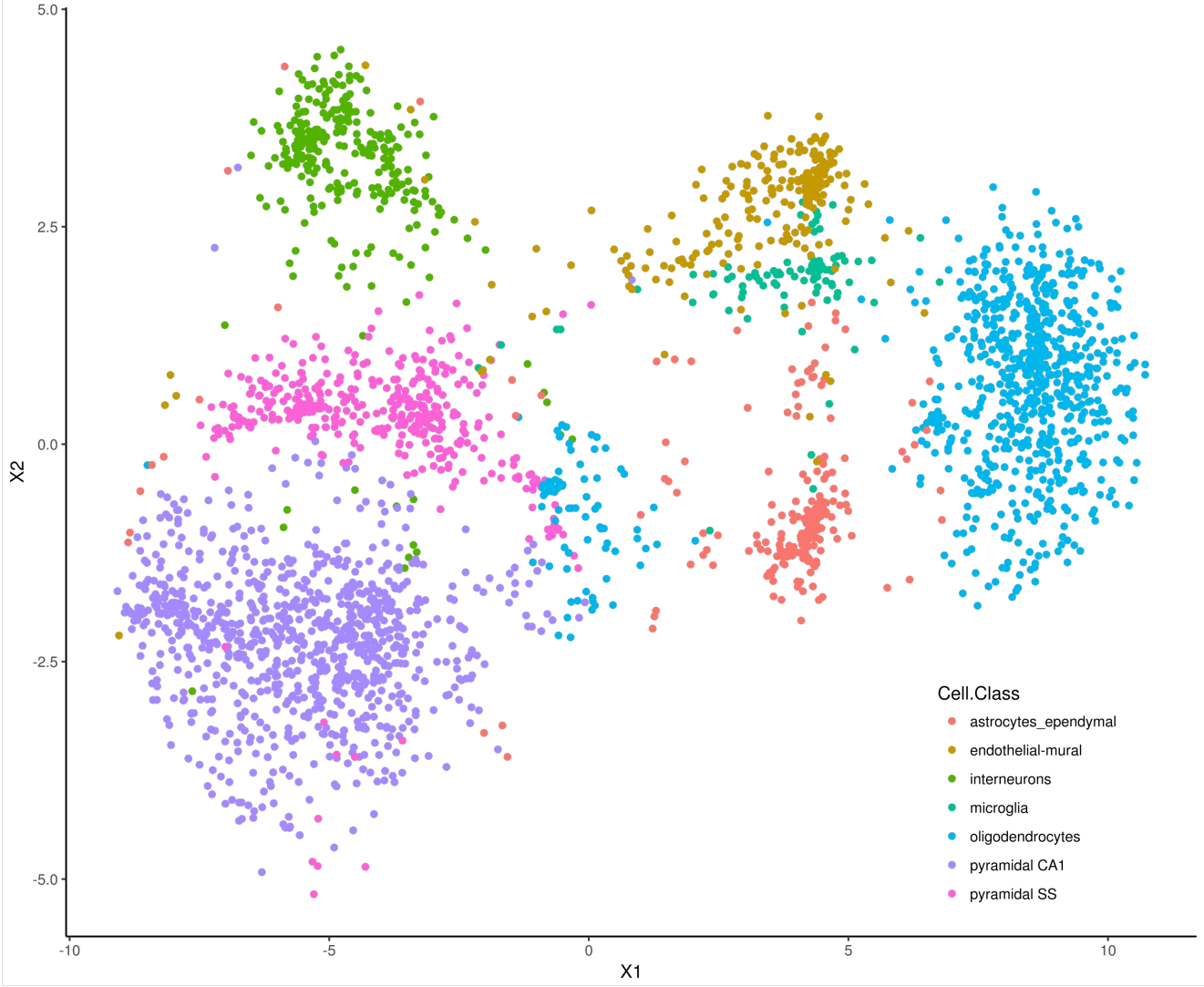
Pathway Activity in Breast Cancer



Brady S, et al. Combating subclonal evolution of resistant cancer phenotypes. *Nat Commun.* 2017

Critical Gaps in scRNA-seq data

- Complex data
  - New analysis challenges
- Interactive, Simple Analysis?
  - Inexperienced Users
  - Optimizing Parameters
  - Filtering Failed Samples
  - Filtering Low expression genes
- Some packages already exist
  - QC
  - Clustering
  - Full Analysis Portal?



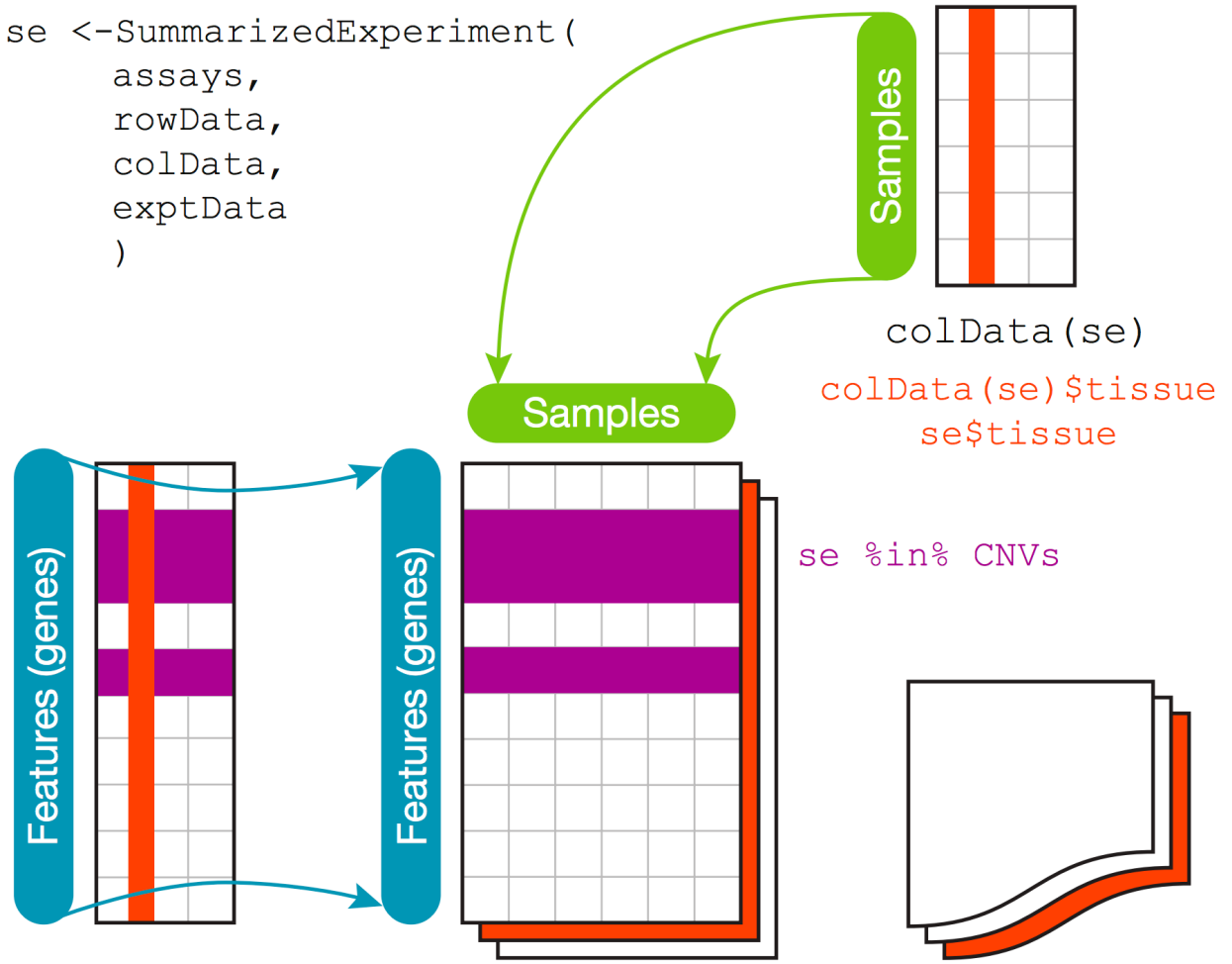
SingleCellTK

Single Cell Toolkit

- Standard R package *with* a Shiny toolkit
- R functions on top of a SingleCellExperiment (SCE) object
  - Can download analysis performed in the toolkit and continue on the command line
- SCE object can be brought in/out of Shiny at any stage
  - Great for common tasks:
    - Interactive clustering/visualization
    - Differential gene/pathway analysis
- (Also works for bulk RNA-seq analysis)

SummarizedExperiment

```
se <-SummarizedExperiment(
     assays,
     rowData,
     colData,
     exptData
     )
```



colData(se)

colData(se)$tissue

se$tissue

se %in% CNVs

rowData(se)

rowData(se)$entrezId

assays(se)

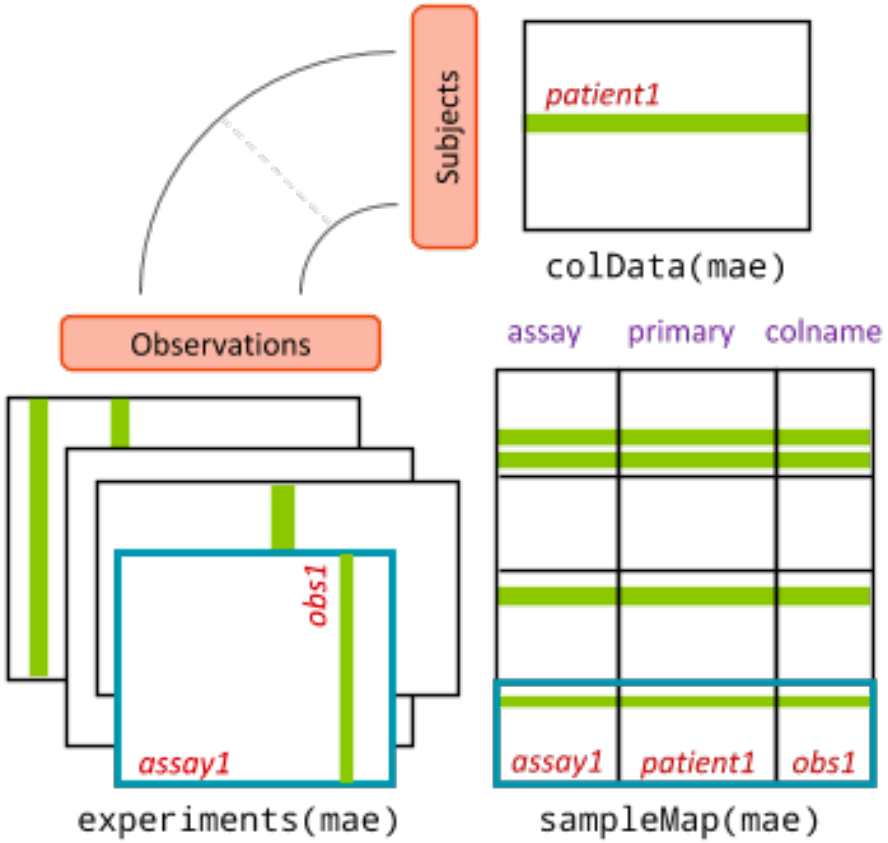
assays(se)$count

exptData(se)

exptData(se)$projectId

MultiAssayExperiment

- Combine multiple data types in similar objects with functionality to query across features and cells.

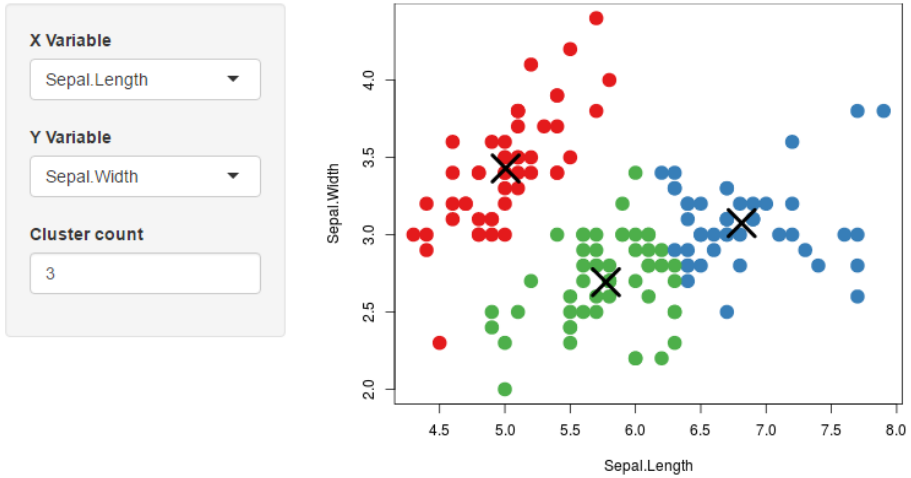


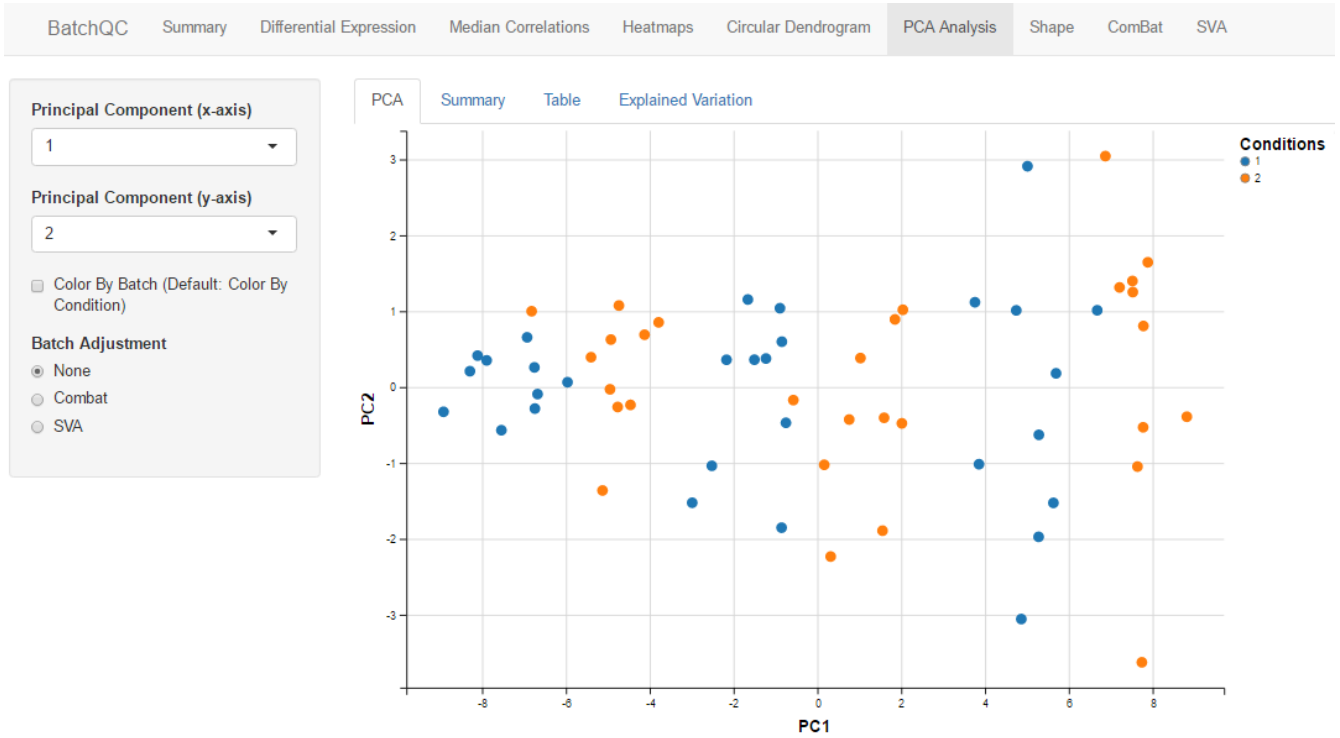
SingleCellExperiment

- https://bioconductor.org/packages/devel/bioc/html/SingleCellExperiment.html
- reducedDims
  - PCA, TSNE, any other sample size matrix of dimensionality reduction data
- isSpike – indicate which probes are spike ins
- sizeFactors – add scaling factors
- Named assays: counts, normcounts, logcounts, cpm, tpm
  - Just a convention
- DelayedArray – run common array functions without loading it into memory

Shiny

- Simple Web Apps Written in R
- Interactive and Reactive
- Iris k-means clustering





SingleCellTK

Data Upload	QC and Filtering	Visualization and Clustering
ComBat Batch Correction	Differential Expression	MAST
Pathway Activity Analysis	Experimental Design	Data Export

singleCellTK

platforms all | rank unknown | posts 0 | in Bioc 1 year
build ok | updated before release

DOI: 10.18129/B9.bioc.singleCellTK

Interactive Analysis of Single Cell RNA-Seq Data

Bioconductor version: Release (3.9)

Run common single cell analysis directly through your browser including differential expression, downsampling analysis, and clustering.

Author: David Jenkins

Maintainer: David Jenkins <dfj at bu.edu>

Citation (from within R, enter `citation("singleCellTK")`):

Jenkins D, Faits T, Khan MM, Briars E, Carrasco Pro S, Johnson WE (2019). *singleCellTK: Interactive Analysis of Single Cell RNA-Seq Data*. R package version 1.4.0, https://compbiomed.github.io/sctk_docs/.

Installation

Documentation »

Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

Single Cell Toolkit

Filter, cluster, and analyze single cell RNA-Seq data

Need help? Read the docs.

Upload

(help)

Choose data source:

- (•) Upload files
- ( ) Upload SCtkExperiment RDS File
- ( ) Use example data

Upload data in tab separated text format:

Example count file:

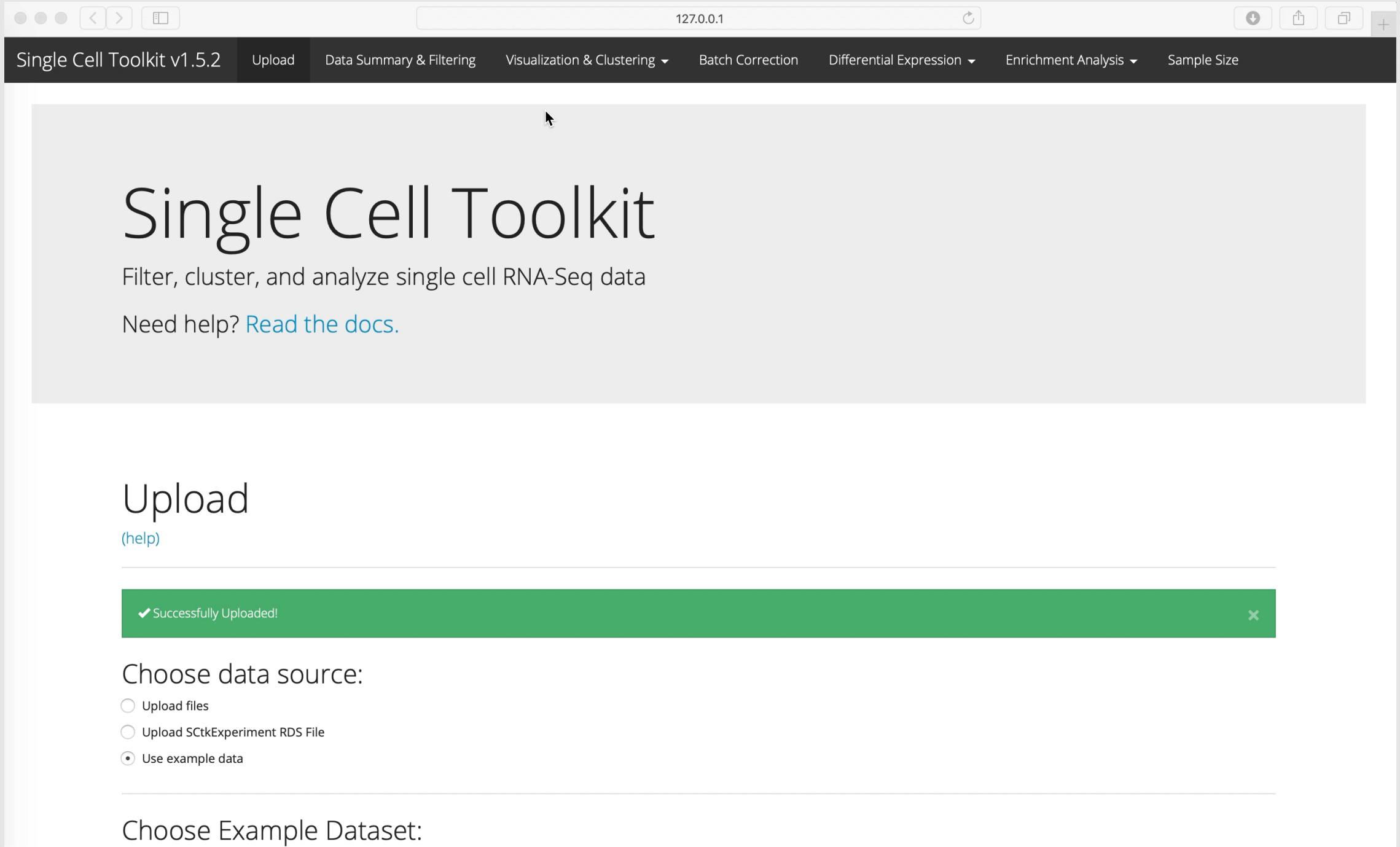
Gene	Cell1	Cell2	...	CellN

Example sample annotation file:

Cell	Annot1	...

Example feature file:

Gene	Annot2	...


127.0.0.1
Single Cell Toolkit v1.5.2
Upload
Data Summary & Filtering
Visualization & Clustering
Batch Correction
Differential Expression
Enrichment Analysis
Sample Size
Single Cell Toolkit
Filter, cluster, and analyze single cell RNA-Seq data
Need help? Read the docs.
Upload
(help)
Successfully Uploaded!
Choose data source:
Upload files
Upload SCtkExperiment RDS File
Use example data
Choose Example Dataset:

Single Cell Toolkit

Filter, cluster, and analyze single cell RNA-Seq data

Need help? Read the docs.

Upload

(help)

✔ Successfully Uploaded!

Choose data source:

- ○ Upload files
- ○ Upload SCtkExperiment RDS File
- ◉ Use example data

Choose Example Dataset:

mouseBrainSubset

SCTK features – *Differential Expression analysis-MAST*

Single Cell Toolkit v0.3.9 | Upload | Data Summary and Filtering | DR & Clustering | Batch Correction | Differential Expression | Pathway Activity Analysis | Sample Size

MAST

Select Assay:

logcounts

Adaptive Thresholding:

Run Thresholding

Hurdle Model:

Use Adaptive Thresholds

Select fold change threshold

0 — 0.6 — 1

Select expression threshold

0 — 0.1 — 1

Select Condition for Hurdle Model

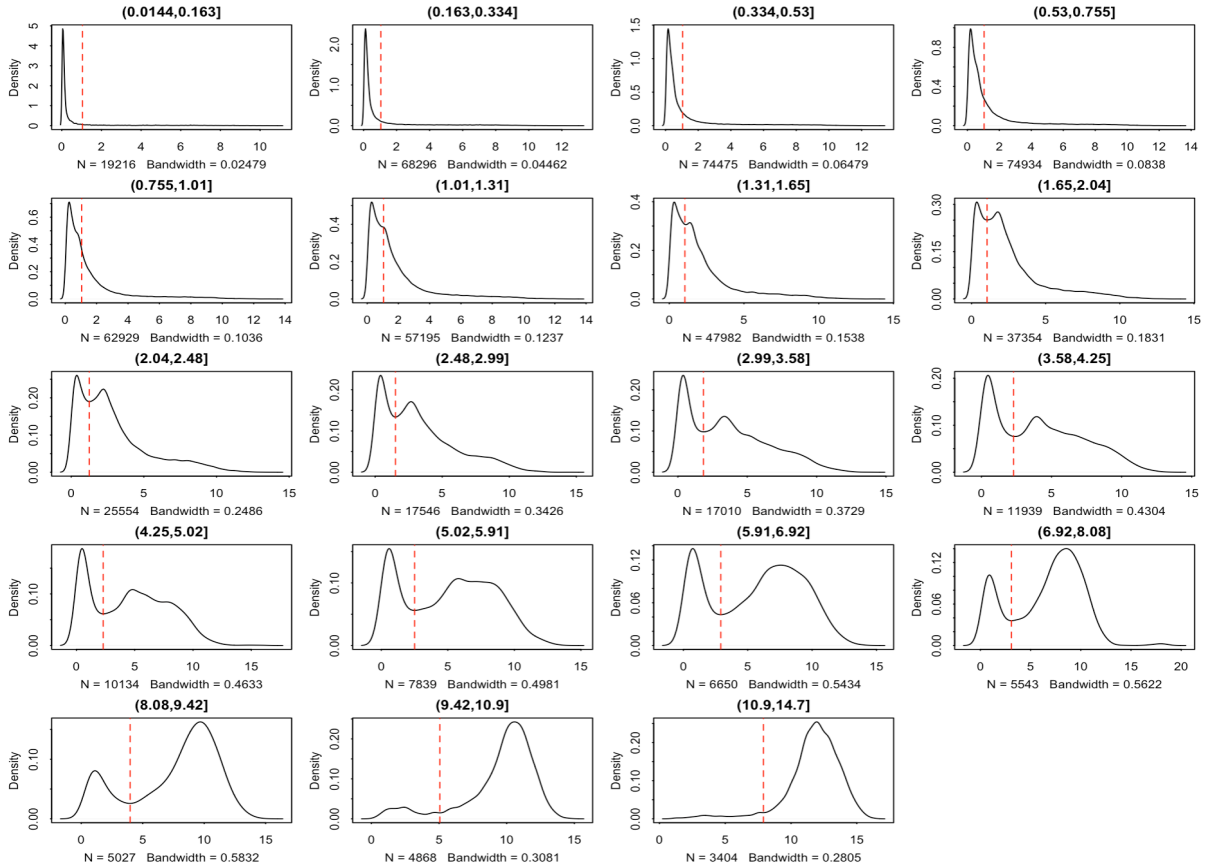
condition

p-value (FDR) cutoff:

0.01 — 0.05 — 0.2

Run DE Using Hurdle

Adaptive thresholding | Results Table | Violin Plot | Linear Model | Heatmap



Single Cell Toolkit v0.3.9 | Upload | Data Summary and Filtering | DR & Clustering | Batch Correction | Differential Expression | Pathway Activity Analysis | Sample Size

MAST

Select Assay:

logcounts

Adaptive Thresholding:

Run Thresholding

Hurdle Model:

Use Adaptive Thresholds

Select fold change threshold

0 — 0.6 — 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Select expression threshold

0 — 0.1 — 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Select Condition for Hurdle Model

condition

p-value (FDR) cutoff:

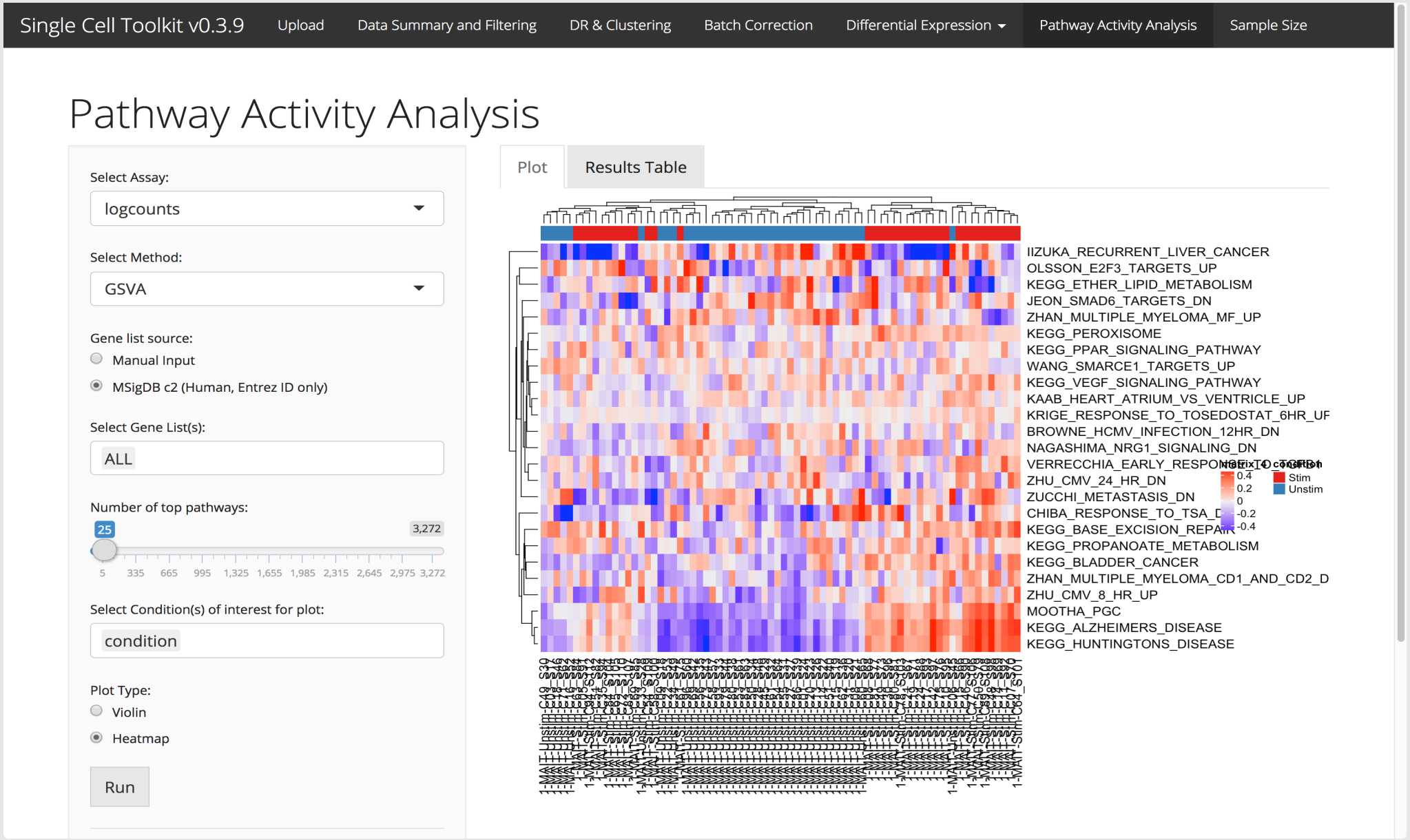
0.01 — 0.05 — 0.2

0.01 0.03 0.05 0.07 0.09 0.11 0.13 0.15 0.17 0.190.2

Run DE Using Hurdle

Adaptive thresholding | Results Table | Violin Plot | Linear Model | Heatmap

SCTK features – *Pathway activity analysis - GSVA*



Enter Top 'N' Genes value:

500

Max genes: 19972

- [ ] Apply p-value Cutoff
- [ ] Apply logFC Cutoff
- [x] Scale Expression values?

Options

General Options

- [x] Row Labels
- [x] Column Labels
- [x] Column Dendrograms
- [x] Row Dendrograms
- [x] Cluster Rows
- [x] Cluster Columns

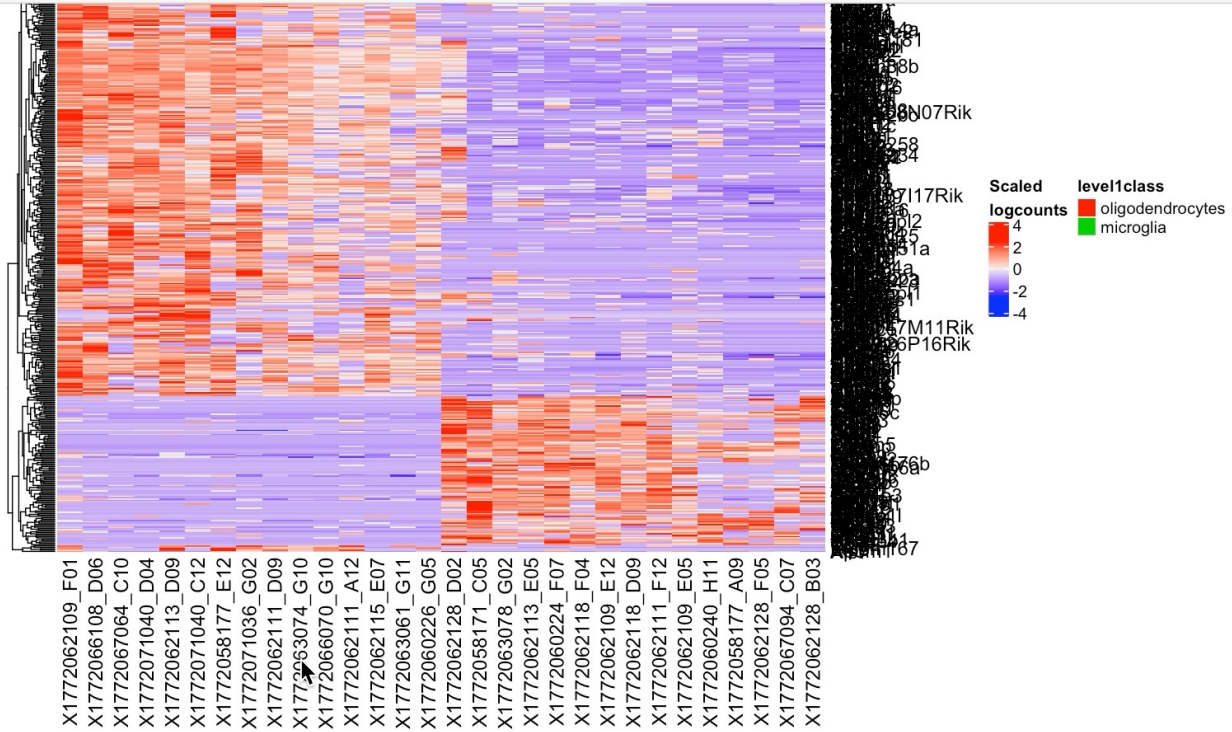
Columns Title

Differential Expression

Colorbar Options

- [x] Display Color Bar

Select Condition



Single Cell Toolkit v0.3.9 | Upload | Data Summary and Filtering | DR & Clustering | Batch Correction | Differential Expression | Pathway Activity Analysis | Sample Size

Sequencing Depth | Number of cells | Snapshot

Minimum readcount to detect gene

10

Minimum number of cells with nonzero expression to detect gene

3

Number of bootstrap iterations.

10

Maximum log10(number of simulated reads)

3 5 12

3 4 5 6 7 8 9 10 11 12

how many values to simulate

5 10 100

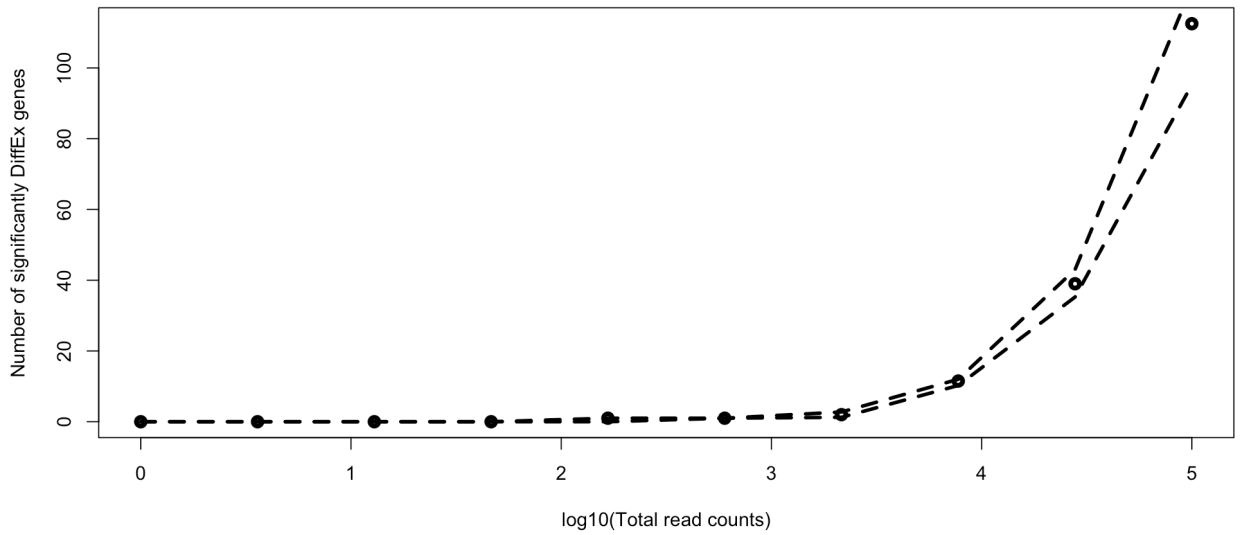
5 15 25 35 45 55 65 75 85 95 100

Condition for diffex

condition

Run subsampler

Genes Detected | Minimum Detectable Effect Size | Number of Diffex Genes



BOSTON UNIVERSITY | c b m

127.0.0.1:3236/#tab-8388-6

Thanks

- Johnson Lab
  - David Jenkins
  - Mohammad Khan
  - Tyler Faits
  - Yuqing Zhang
  - Yue Zhao
- Additional Toolkit Contributors
  - Emma Briars
  - Sebastian Carrasco Pro
  - Steve Cunningham
  - Sean Corbett
- Bild Lab (University of Utah)
  - Andrea Bild
  - Mumtahena Rahman
  - Shelley Macneil
  - Sam Brady
- Single Cell Workgroup
  - Josh Campbell
  - Masanao Yajima

Common Questions

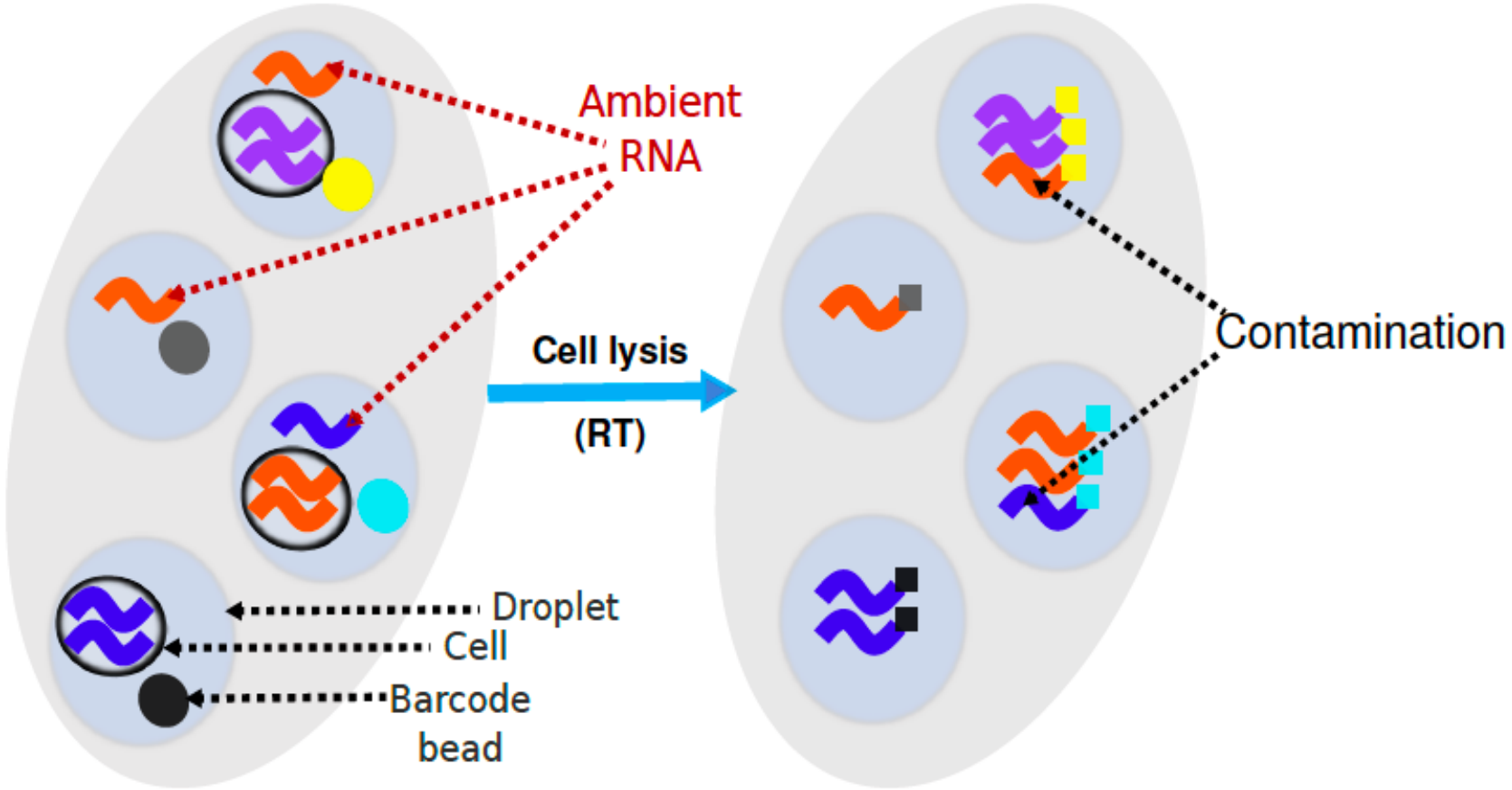
- How is this different from other tools/packages?
  - (Is this the right question?)
- Can we import data to/from other tools?
  - CellRanger object
  - To/from Seurat
- Is it scalable to larger datasets?

Decontamination of ambient RNA with DecontX

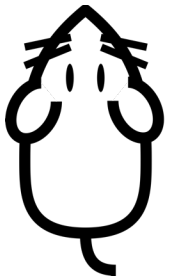
https://www.biorxiv.org/content/10.1101/704015v1

https://github.com/campbio/celda

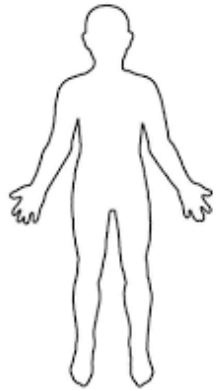
Contamination from ambient RNA in scRNA-seq data



Example of contamination in a mouse-human cell mixture dataset from 10X.

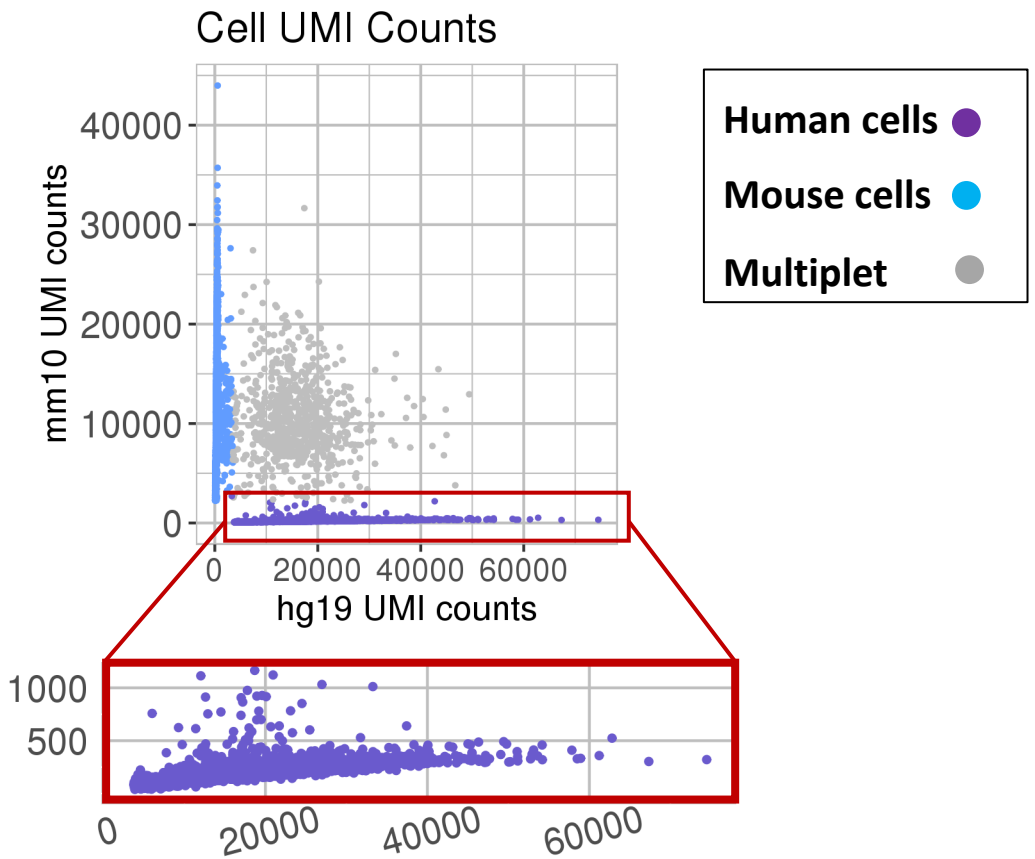


Fresh frozen mouse (NIH3T3) cells

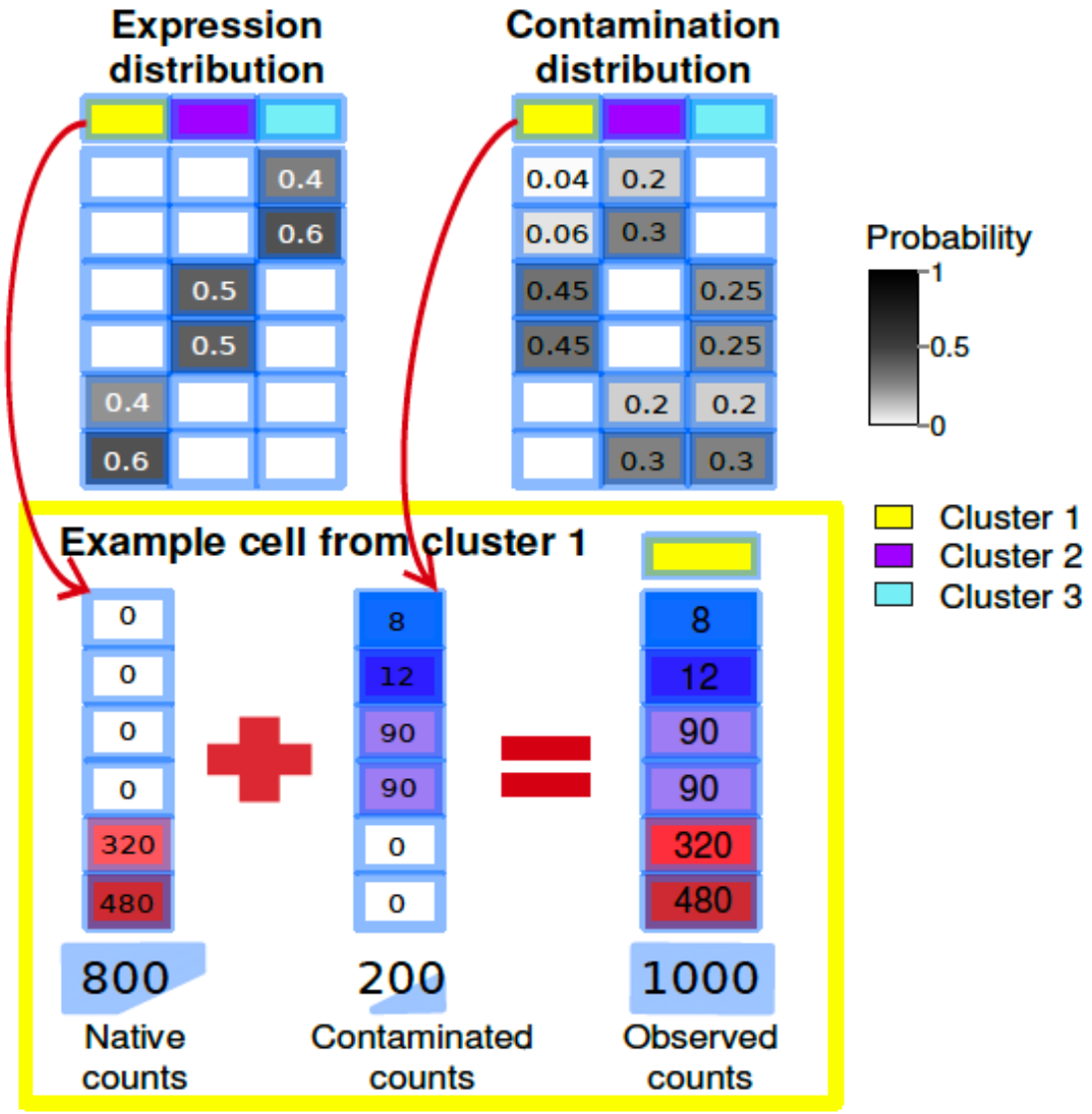


Fresh frozen human (HEK293T) cells

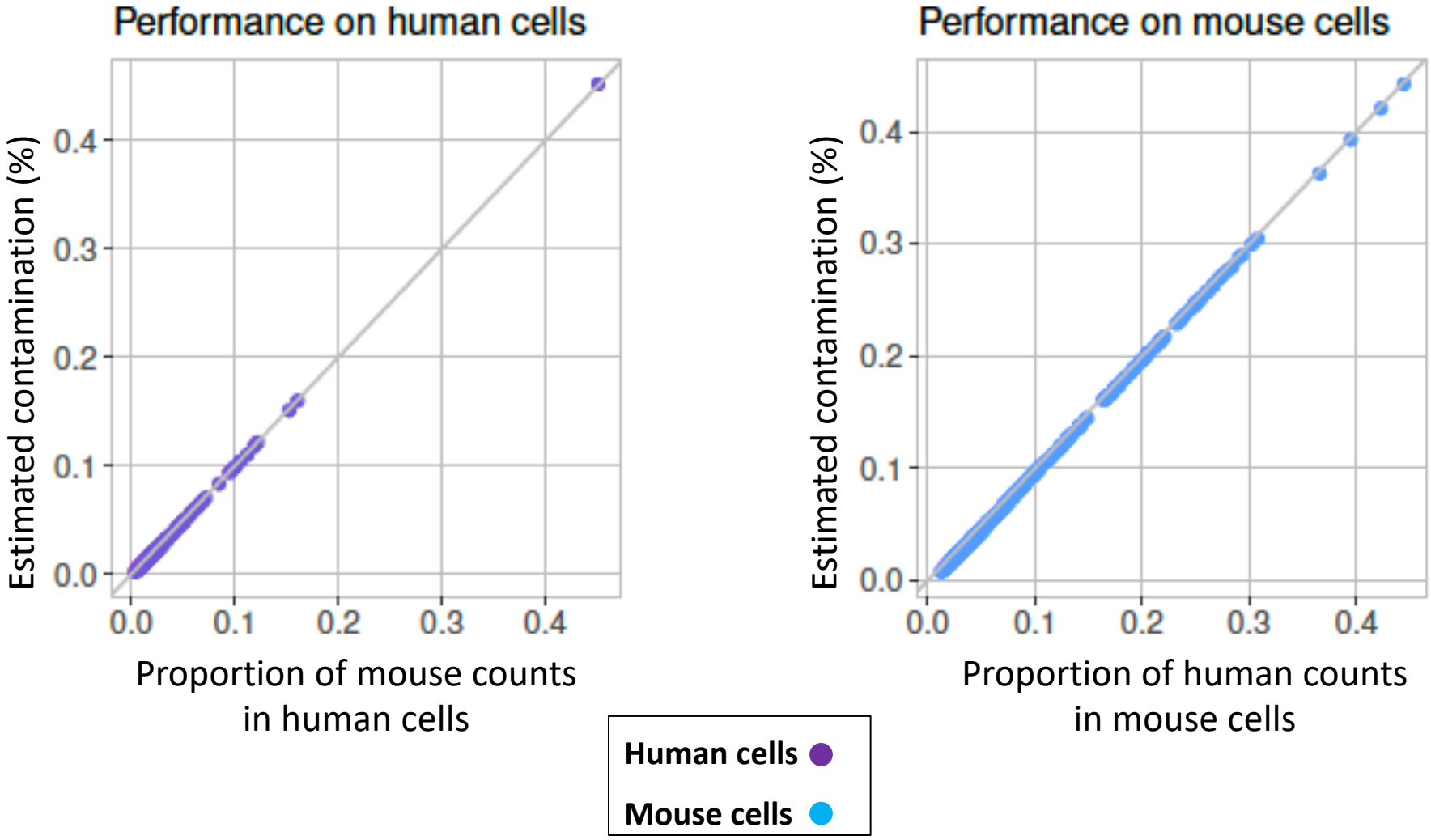
1:1 mixture of 12k human-mouse cells
12,820 cells detected



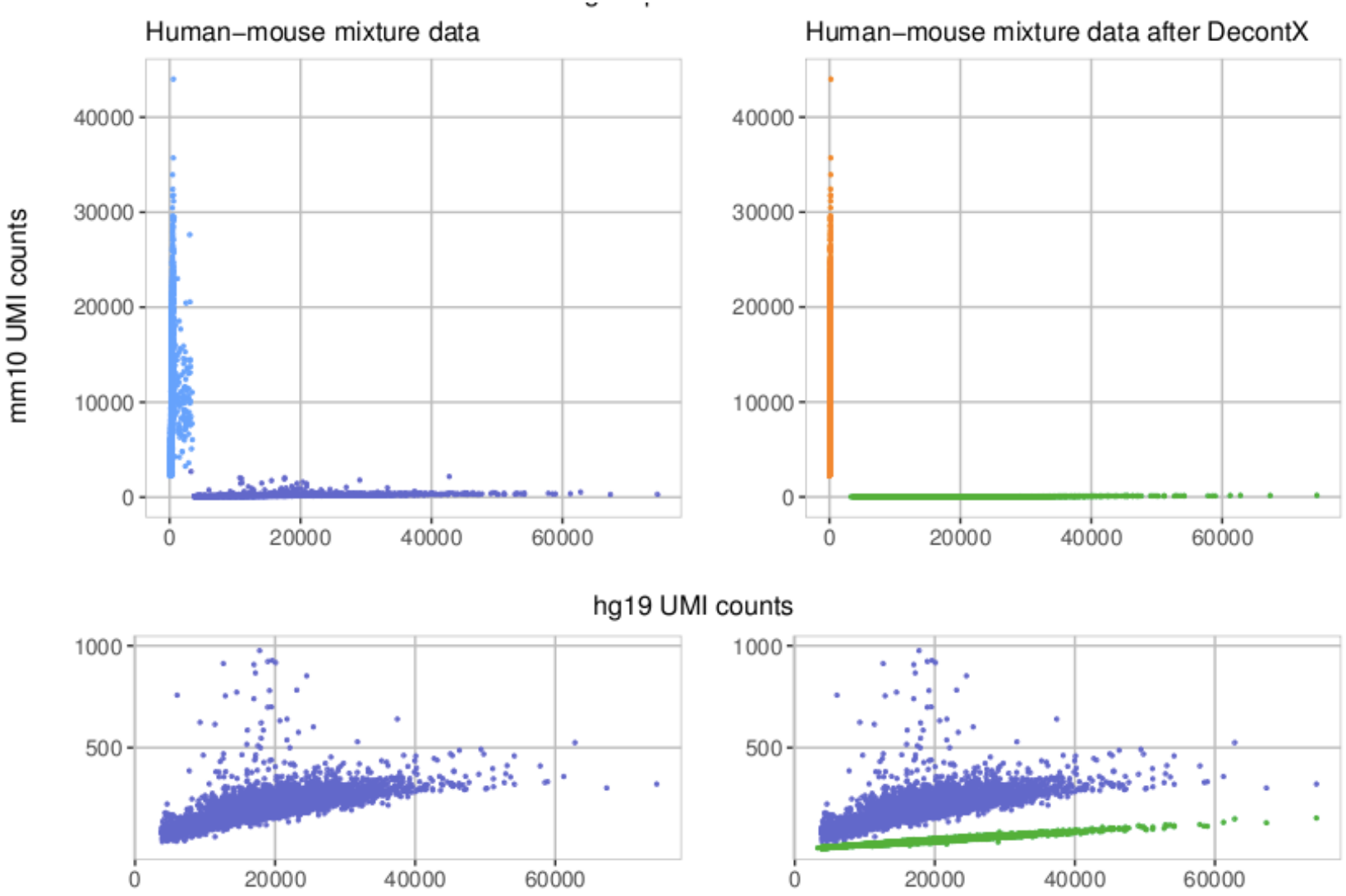
Each cell is a mixture of transcripts from its native cell type and a weighted combination of all other cell types.



DecontX-estimated % contamination is highly correlated with the proportion of counts from the other organism.

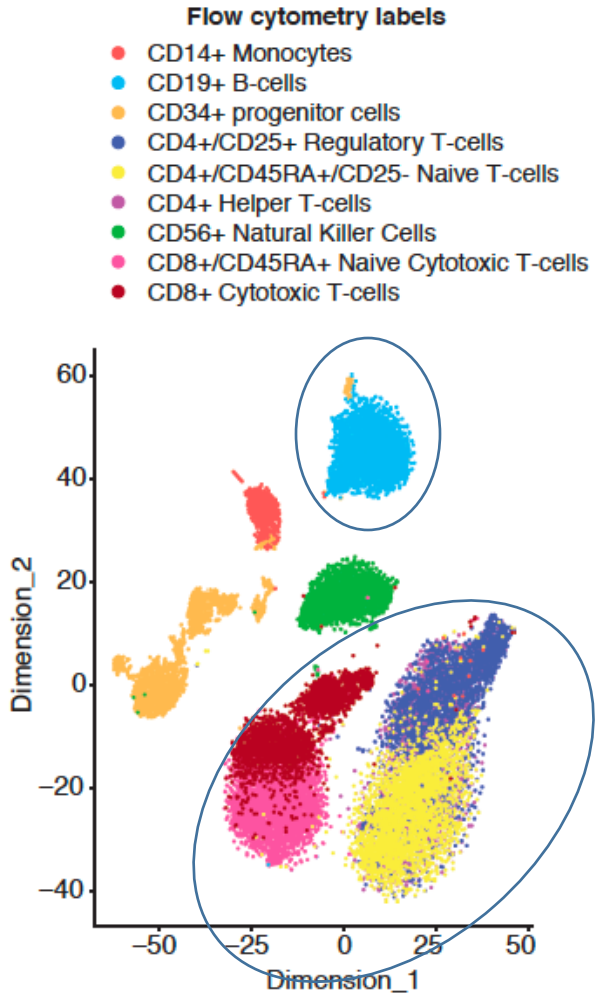


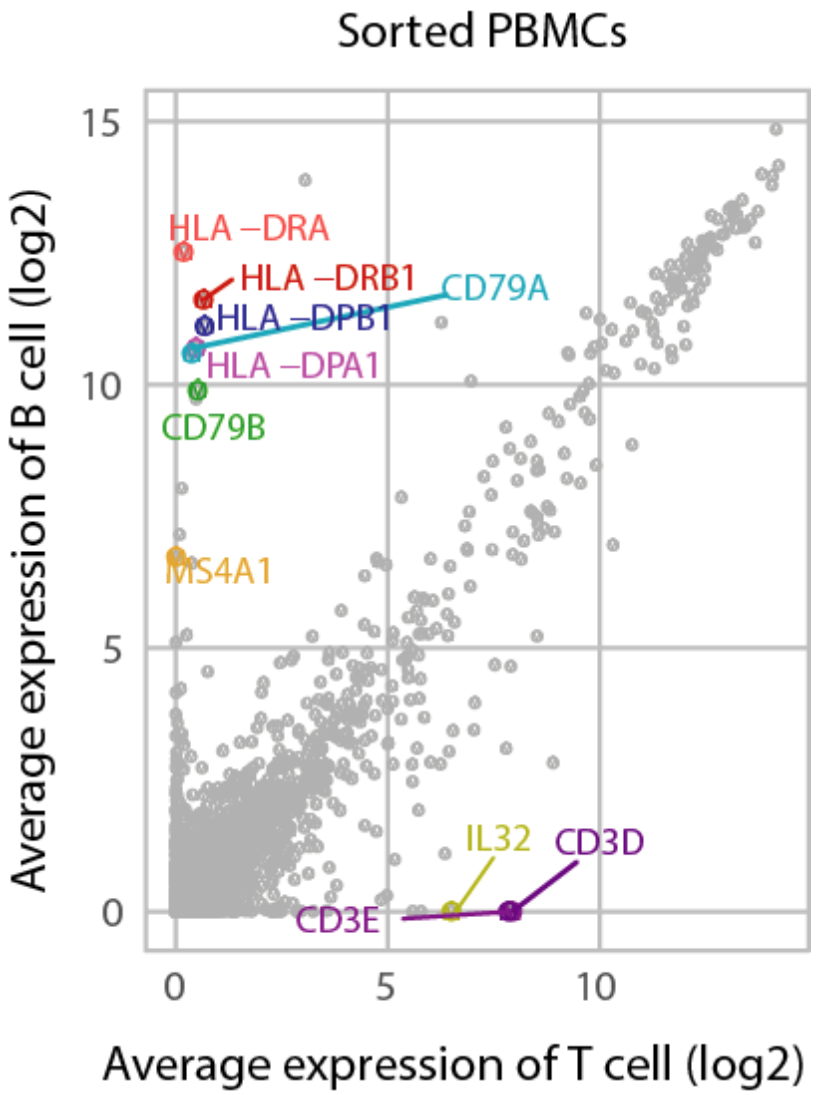
DecontX largely removed the contamination counts from ambient RNA of the other organism.



PBMCs sorted and profiled in separate channels of 10X showed little cross contamination.

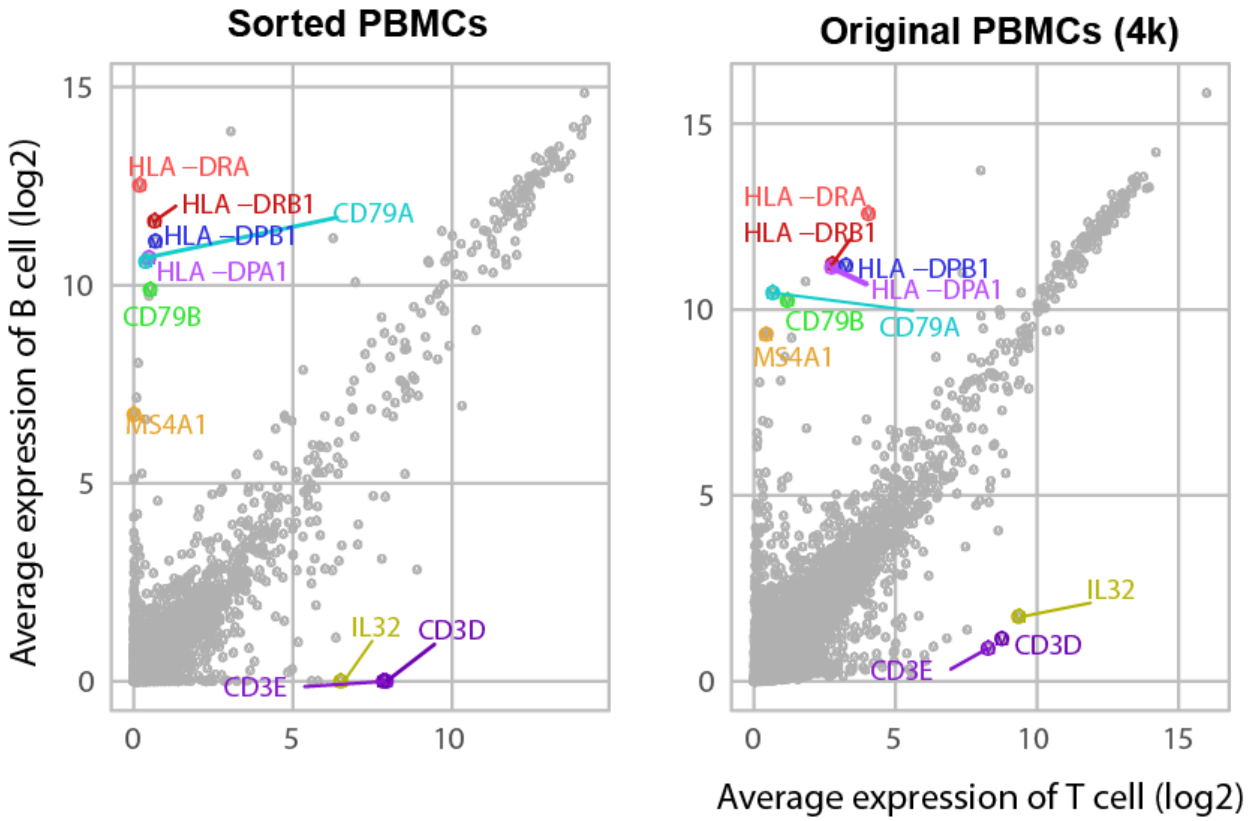
Number of cells: 84,431



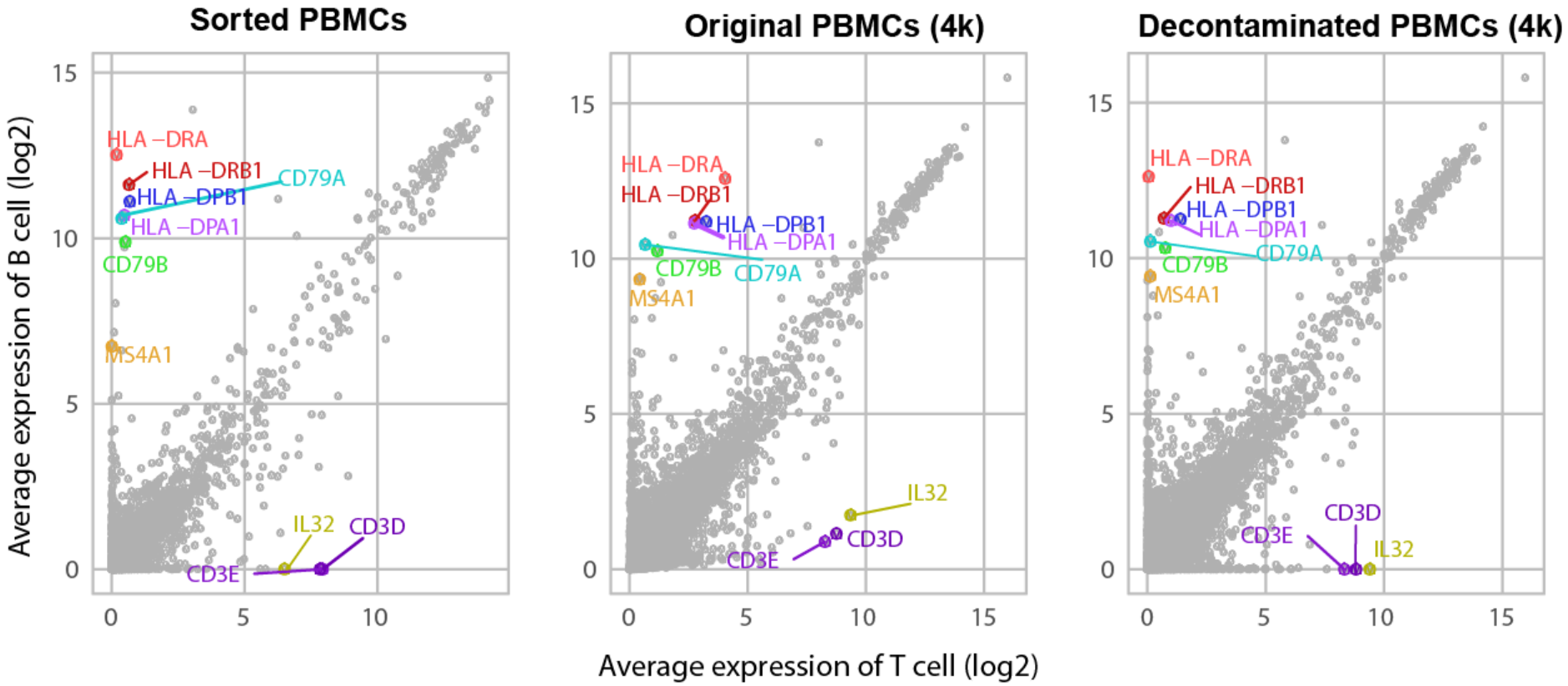


Zheng et al, Nat Commun, 2017

PBMCs that were profiled in the same channel of 10X showed higher levels of marker gene cross contamination.



DecontX removed the majority of marker genes cross contamination from T- and B-cell populations.



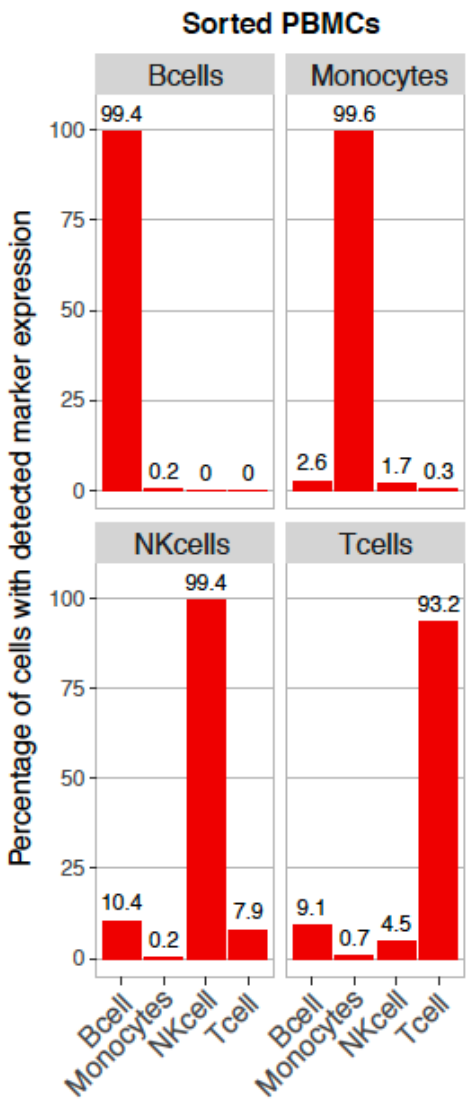
When cell types are profiled in different channels, cell-type specific genes are lowly detectable in other populations.

B-cells
CD79A
CD79B
MS4A1

T-cells
CD3E
CD3D

NK-cells
GNLY

Monocytes
S100A8
S100A9
VCAN



Marker Genes

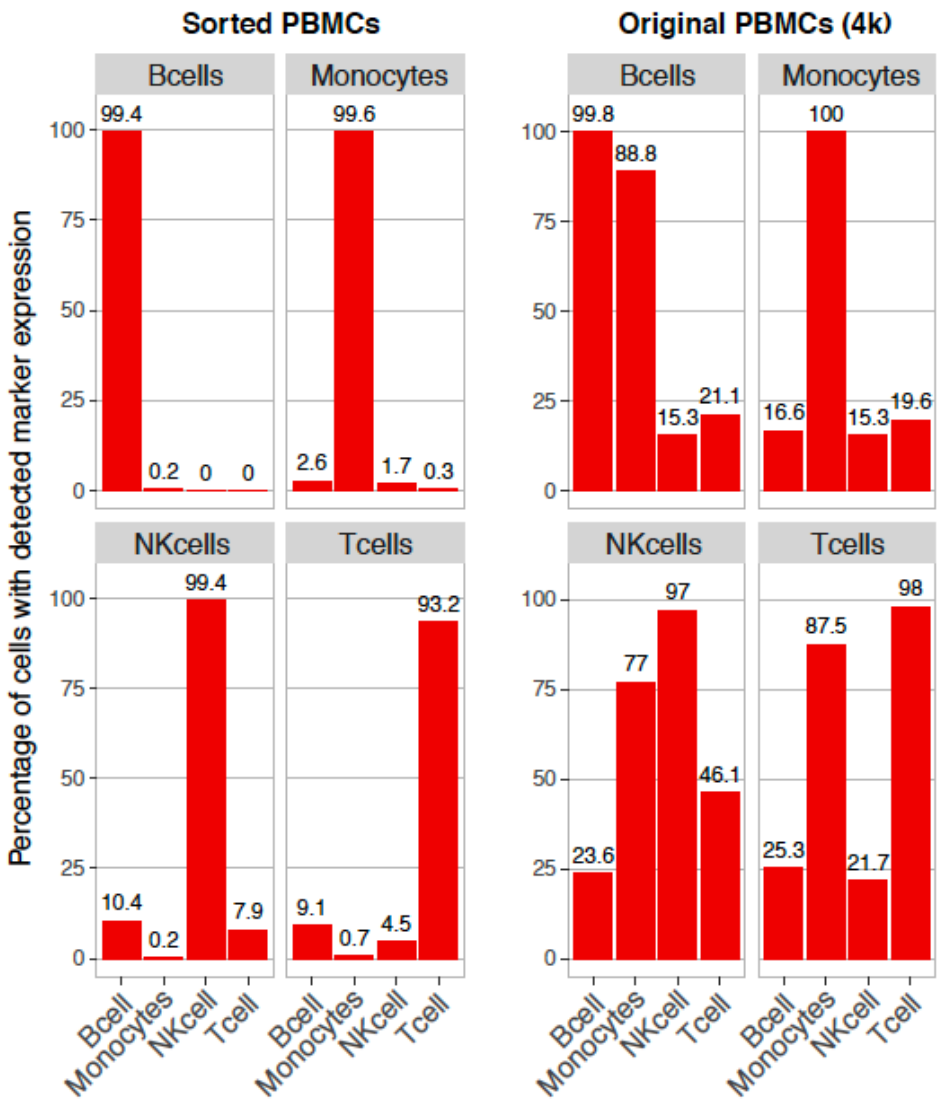
When cell types are profiled in the same channel, cell-type specific genes are detected at high levels in other populations.

B-cells
CD79A
CD79B
MS4A1

T-cells
CD3E
CD3D

NK-cells
GNLY

Monocytes
S100A8
S100A9
VCAN



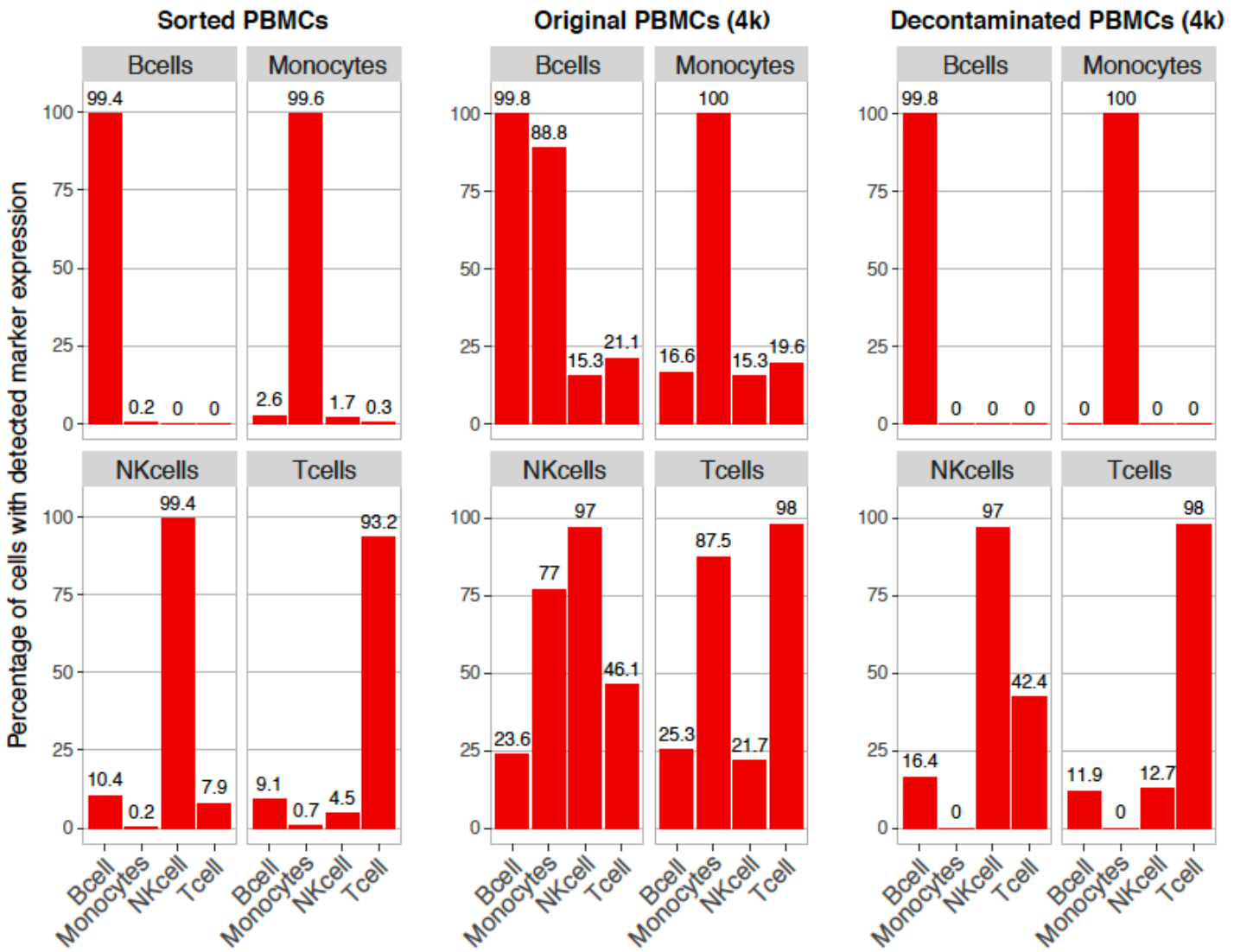
DecontX removed the majority of marker gene cross contamination from all populations.

B-cells
CD79A
CD79B
MS4A1

T-cells
CD3E
CD3D

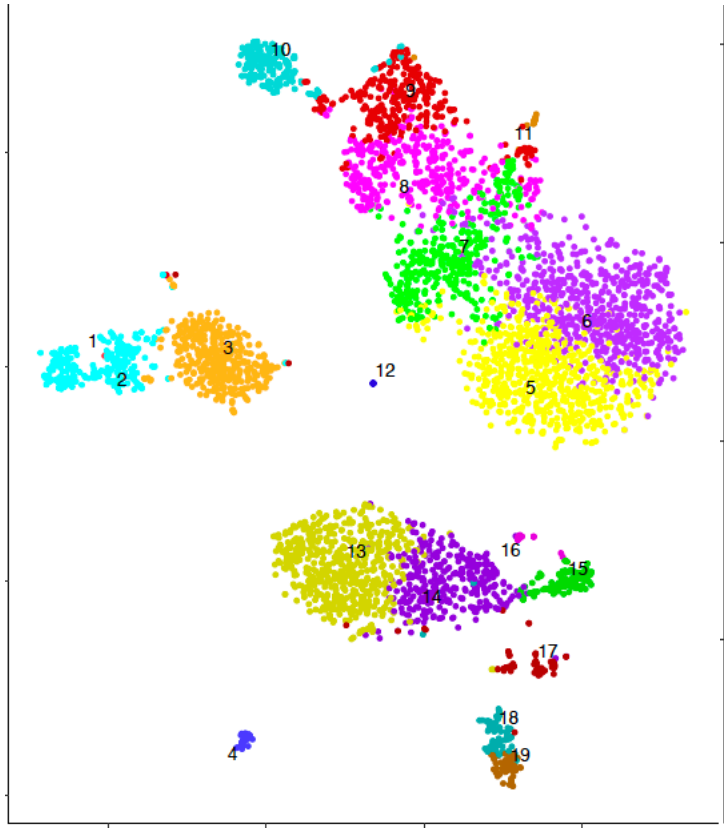
NK-cells
GNLY

Monocytes
S100A8
S100A9
VCAN

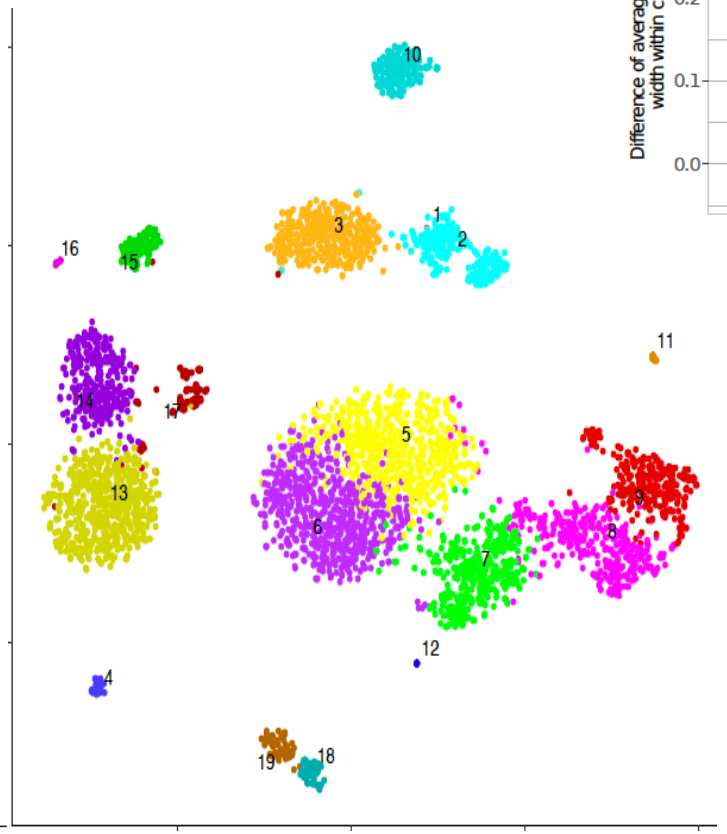


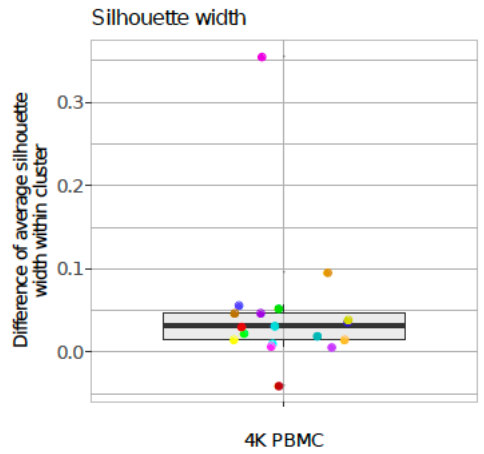
Decontamination improved overall cell cluster separation.

Original Counts
PBMC 4K



Decontaminated Counts
PBMC 4K





Average contamination level in PBMC 4K was 6.9%

Summary

- Estimation of contamination from ambient RNA can be useful to quality control experimental pipelines.
- Decontaminated counts can improve downstream clustering and visualization.
- DecontX is available in the Celda R package on Bioconductor and Github:

https://github.com/campbio/celda

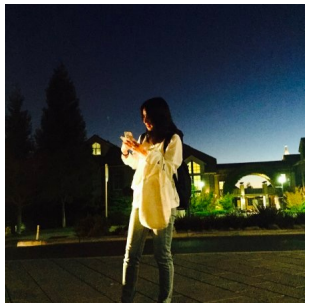
DecontX Acknowledgements

Boston University School of Medicine
Computational Biomedicine

Masanao Yajima

Evan Johnson

Shiyi (Iris) Yang



Zhe Wang

Yusuke Koga

Sean Corbett

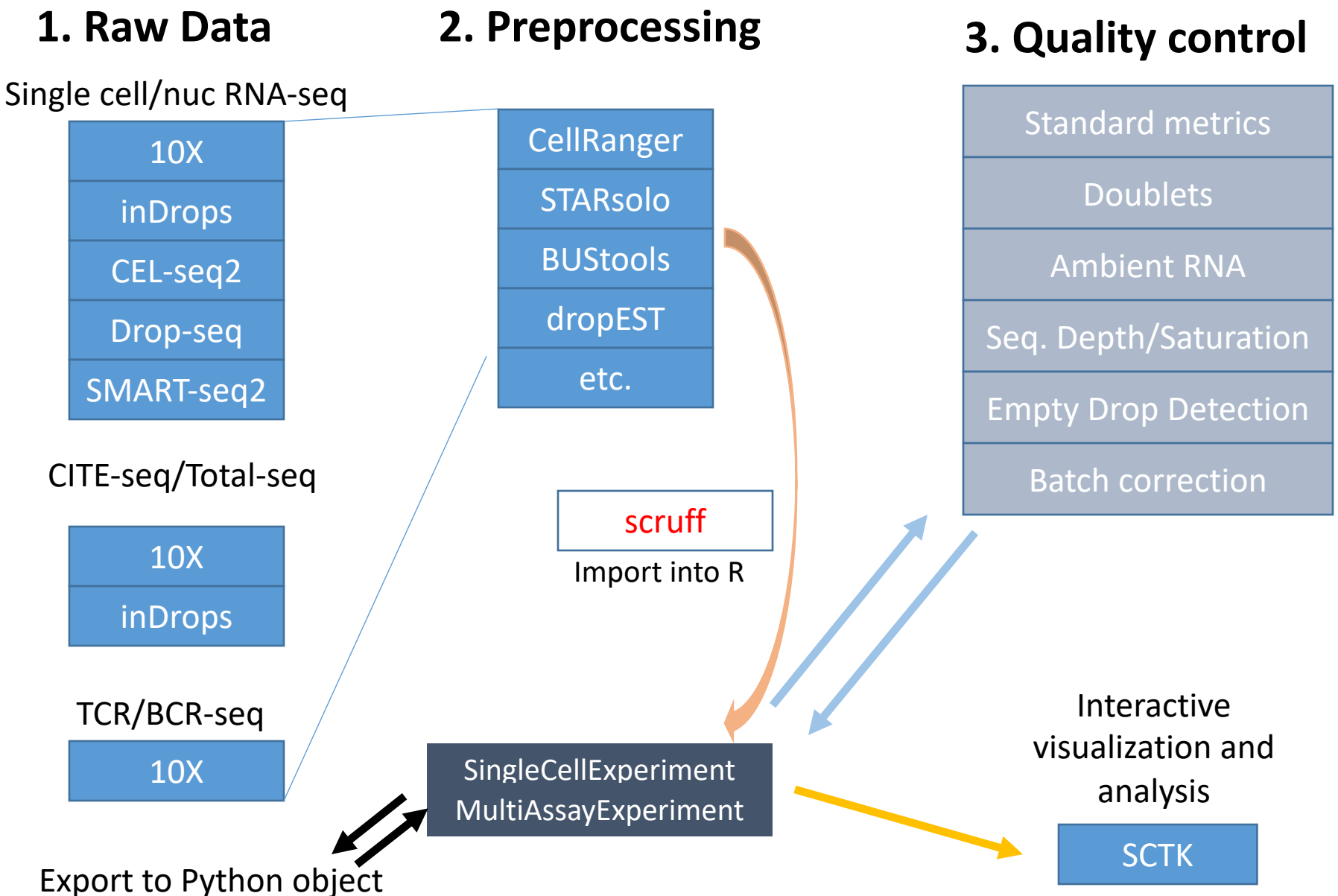
Eric Reed
Ahmed Youssef
Jing Zhang

Salam Alabdullatif
Ali Lashkaripour
Zhaorong Li
Timo Hu

https://github.com/campbio/celda

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

QC pipeline development



Summary of QC metrics

Standard metrics
- Total UMI per cell
- Total Genes per cell
- % Mitochondria per cell
- % counts in top X features
  - 50, 100, 200, 500

Doublet Detection
Scrublet
doubletCells (Scran)
DoubletFinder
DoubletDecon
DoubletDetection

Ambient RNA estimation
DecontX

Additional items?

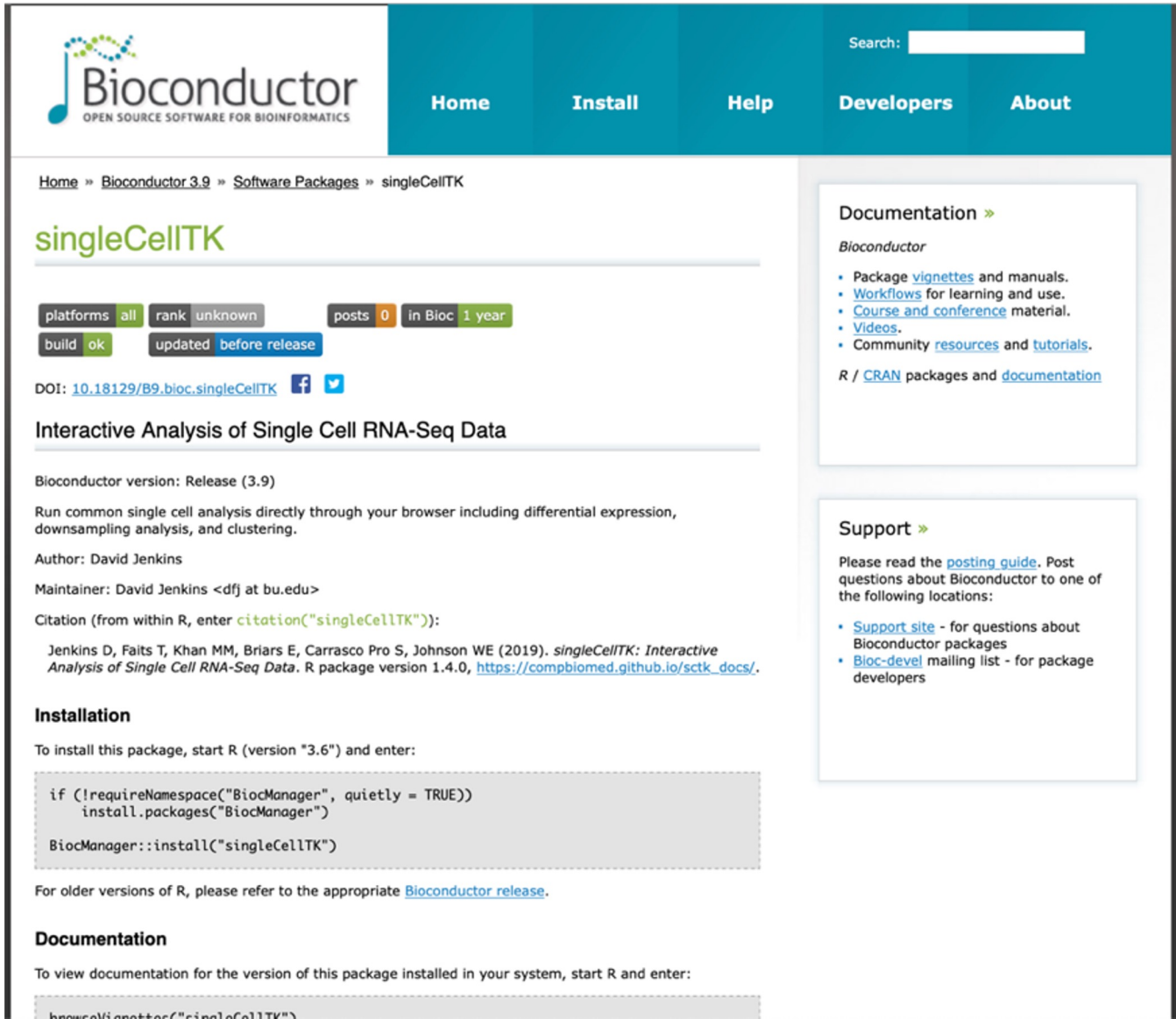
Interactive scRNA-Seq analysis with the Single Cell Toolkit (SCTK)

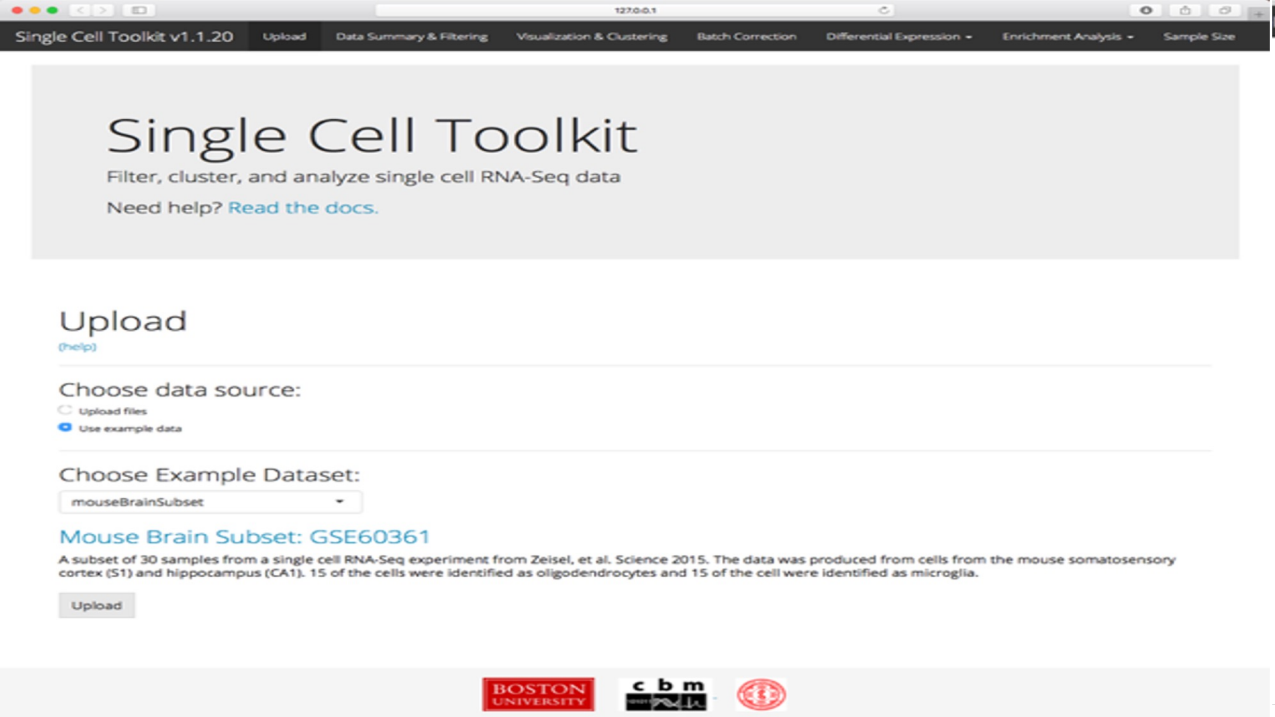
Evan Johnson and Josh Campbell
Division of Computational Biomedicine
Boston University School of Medicine

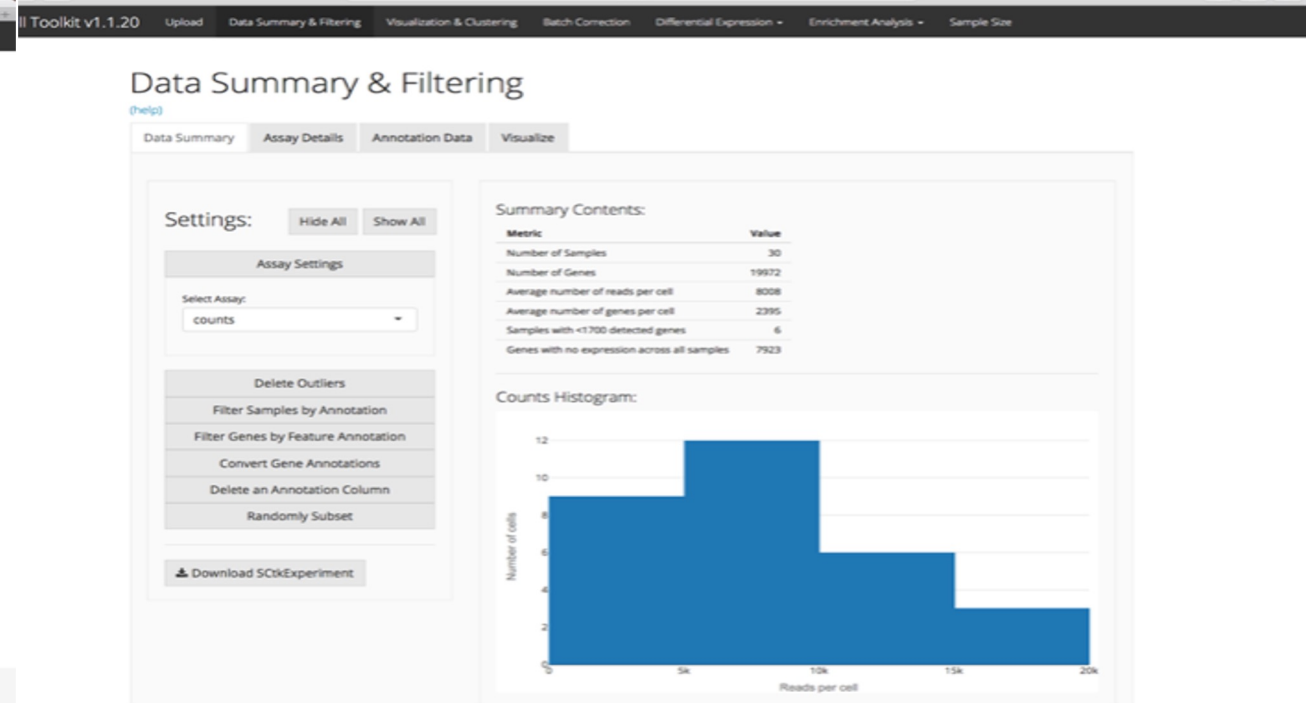
May 1, 2020

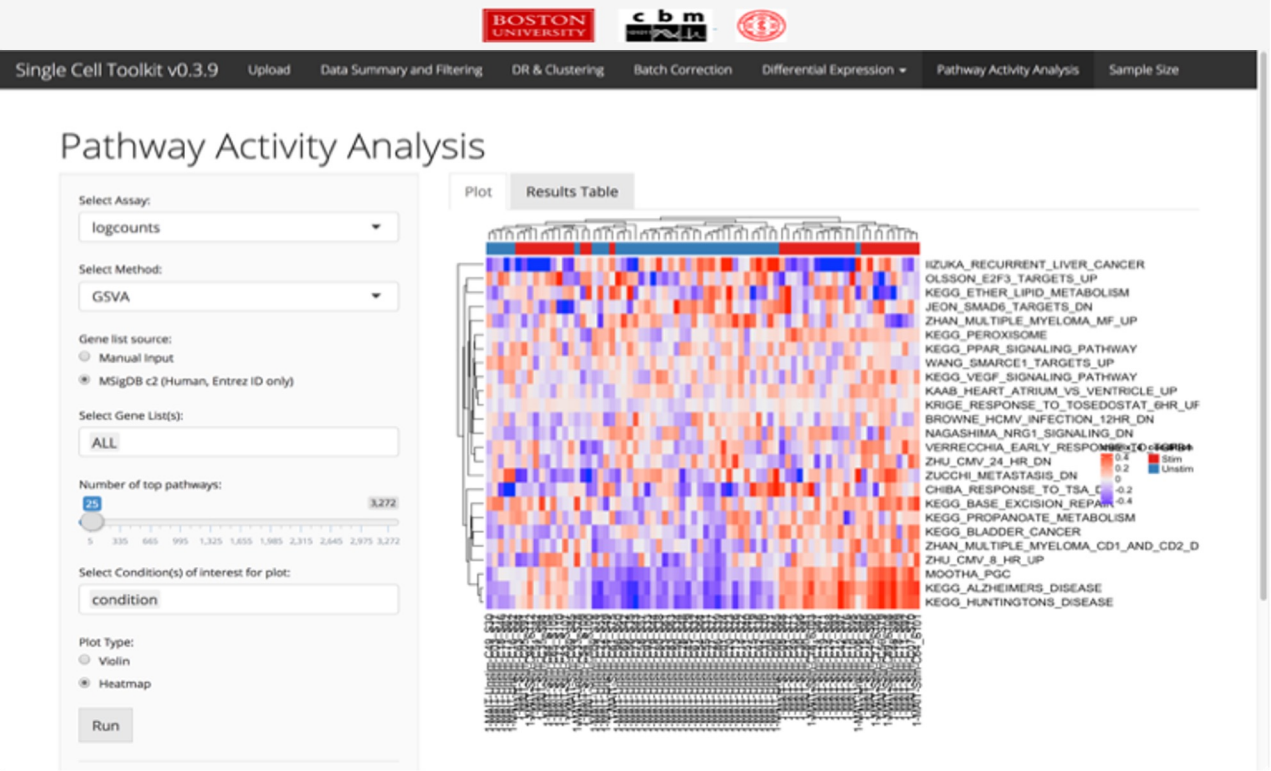
Single Cell Toolkit (SCTK)

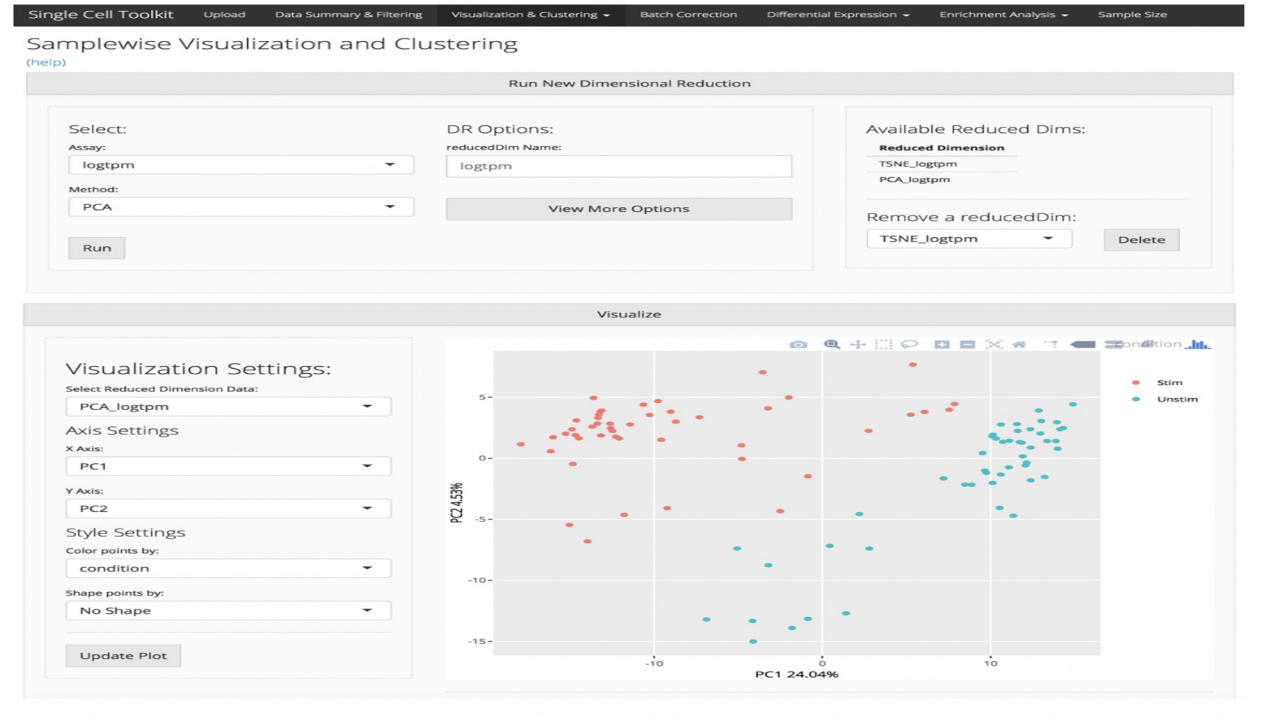
- Standard R package w/Shiny toolkit
- Interacts and operates on an SCE (SingleCellExperiment) object
- SCE object in/out of Shiny any stage
- Great for common tasks:
  - Interactive clustering/visualization
  - Differential gene/pathway analysis
- (Also works for bulk RNA-seq analysis)




Single Cell Toolkit v1.1.20
Upload
Data Summary & Filtering
Visualization & Clustering
Batch Correction
Differential Expression
Enrichment Analysis
Sample Size
Single Cell Toolkit
Filter, cluster, and analyze single cell RNA-Seq data
Need help? Read the docs.
Upload
(help)
Choose data source:
Upload files
Use example data
Choose Example Dataset:
mouseBrainSubset
Mouse Brain Subset: GSE60361
A subset of 30 samples from a single cell RNA-Seq experiment from Zeisel, et al. Science 2015. The data was produced from cells from the mouse somatosensory cortex (S1) and hippocampus (CA1). 15 of the cells were identified as oligodendrocytes and 15 of the cell were identified as microglia.
Upload


Toolkit v1.1.20
Upload
Data Summary & Filtering
Visualization & Clustering
Batch Correction
Differential Expression
Enrichment Analysis
Sample Size
Data Summary & Filtering
(help)
Data Summary
Assay Details
Annotation Data
Visualize
Settings:
Hide All
Show All
Assay Settings
Select Assay:
counts
Delete Outliers
Filter Samples by Annotation
Filter Genes by Feature Annotation
Convert Gene Annotations
Delete an Annotation Column
Randomly Subset
Download SCtkExperiment
Summary Contents:
Metric
Value
Number of Samples 30
Number of Genes 19972
Average number of reads per cell 8008
Average number of genes per cell 2395
Samples with <1700 detected genes 6
Genes with no expression across all samples 7923
Counts Histogram:
Number of cells
Reads per cell


BOSTON UNIVERSITY
cbm
Single Cell Toolkit v0.3.9
Upload
Data Summary and Filtering
DR & Clustering
Batch Correction
Differential Expression
Pathway Activity Analysis
Sample Size
Pathway Activity Analysis
Select Assay:
logcounts
Select Method:
GSVA
Gene list source:
Manual Input
MSigDB c2 (Human, Entrez ID only)
Select Gene List(s):
ALL
Number of top pathways:
25
3,272
Select Condition(s) of interest for plot:
condition
Plot Type:
Violin
Heatmap
Run
Plot
Results Table
IIZUKA_RECURRENT_LIVER_CANCER
OLSSON_E2F3_TARGETS_UP
KEGG_ETHER_LIPID_METABOLISM
JEON_SMAD6_TARGETS_DN
ZHAN_MULTIPLE_MYELOMA_MF_UP
KEGG_PEROXISOME
KEGG_PPAR_SIGNALING_PATHWAY
WANG_SMARCE1_TARGETS_UP
KEGG_VEGF_SIGNALING_PATHWAY
KAAB_HEART_ATRIUM_VS_VENTRICLE_UP
KRIGE_RESPONSE_TO_TOSEDOSTAT_6HR_UP
BROWNE_HCMV_INFECTION_12HR_DN
NAGASHIMA_NRG1_SIGNALING_DN
ZHU_CMV_24_HR_DN
ZUCCHI_METASTASIS_DN
KEGG_BASE_EXCISION_REPAIR
KEGG_PROPANOATE_METABOLISM
KEGG_BLADDER_CANCER
ZHAN_MULTIPLE_MYELOMA_CD1_AND_CD2_D
ZHU_CMV_8_HR_UP
MOOTHA_PGC
KEGG_ALZHEIMERS_DISEASE
KEGG_HUNTINGTONS_DISEASE
Stim
Unstim


Single Cell Toolkit
Upload
Data Summary & Filtering
Visualization & Clustering
Batch Correction
Differential Expression
Enrichment Analysis
Sample Size
Samplewise Visualization and Clustering
(help)
Run New Dimensional Reduction
Select:
Assay:
logtpm
Method:
PCA
Run
DR Options:
reducedDim Name:
logtpm
View More Options
Available Reduced Dims:
Reduced Dimension
TSNE_logtpm
PCA_logtpm
Remove a reducedDim:
TSNE_logtpm
Delete
Visualize
Visualization Settings:
Select Reduced Dimension Data:
PCA_logtpm
Axis Settings
X Axis:
PC1
Y Axis:
PC2
Style Settings
Color points by:
condition
Shape points by:
No Shape
Update Plot
Stim
Unstim
PC2 4.53%
PC1 24.04%

”inDrops” is custom microfluidics device that can process large numbers of cells with high capture rates.

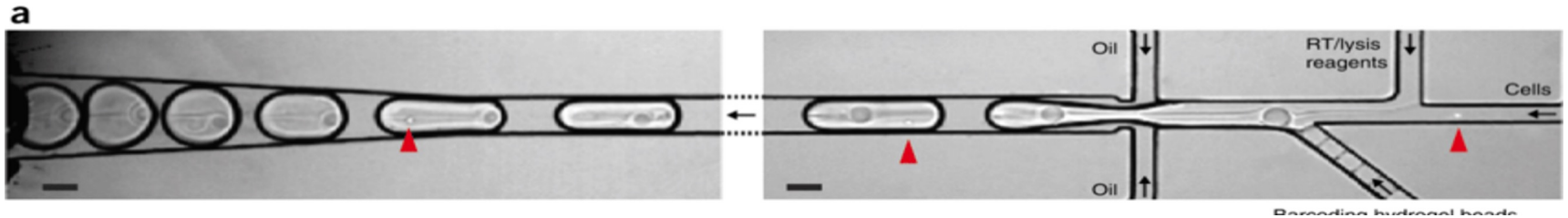
Allon Klein

Harvard Medical School

PROTOCOL

Single-cell barcoding and sequencing using droplet microfluidics

Rapolas Zilionis[1,2], Juozas Nainys[1], Adrian Veres[2–4], Virginia Savova[2], David Zemmour[5], Allon M Klein[2] & Linas Mazutis[1]



Goals of parent R33:

- Aim 1. Optimize inDrops for microscopic samples and fixed cells.
- Aim 2. Optimize inDrops for low-cost, high-throughput, high sensitivity targeted transcriptomics.
- Aim 3. Integrate single cell genomics with histopathology.

Aims of ITCR-IMAT collaboration

IMAT (HMS - Klein)

Aim 1: Develop a Total-seq protocol for the inDrop system.

Aim 2: Develop single cell ATAC-seq for inDrop.

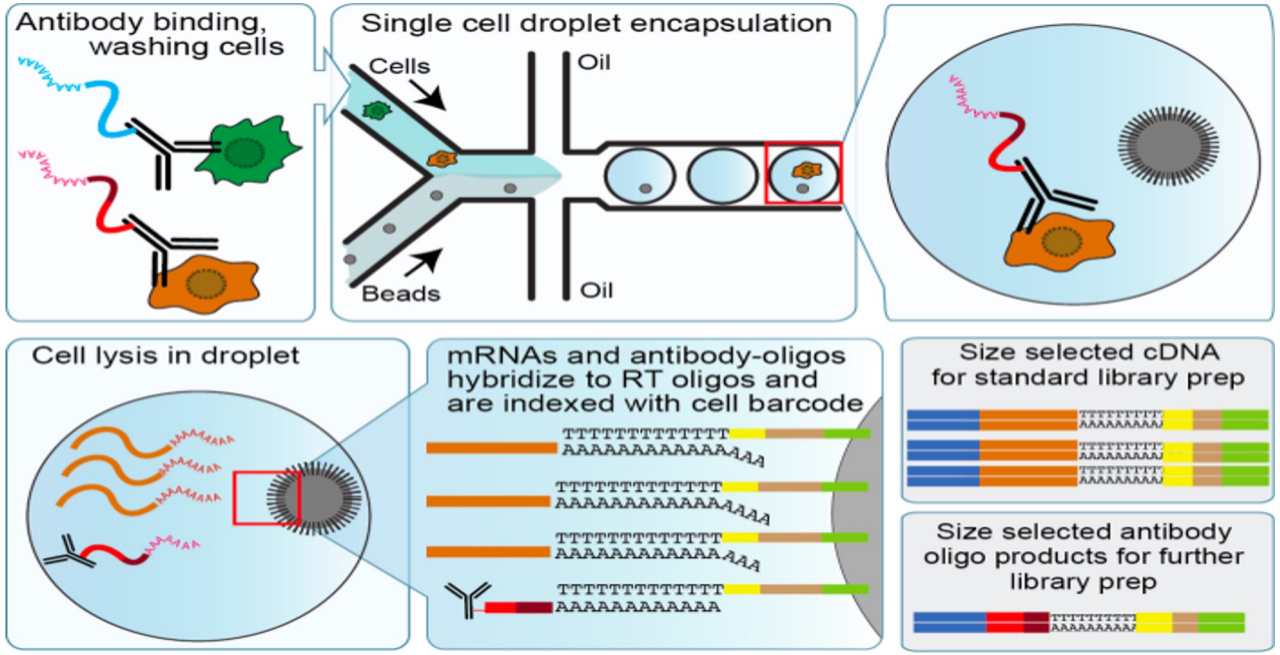
ITCR (BU – Johnson/Campbell)

Aim 1: Develop computational modules for analysis and display of single-cell Total/CITE-seq and ATAC-seq data generated from tumor specimens.

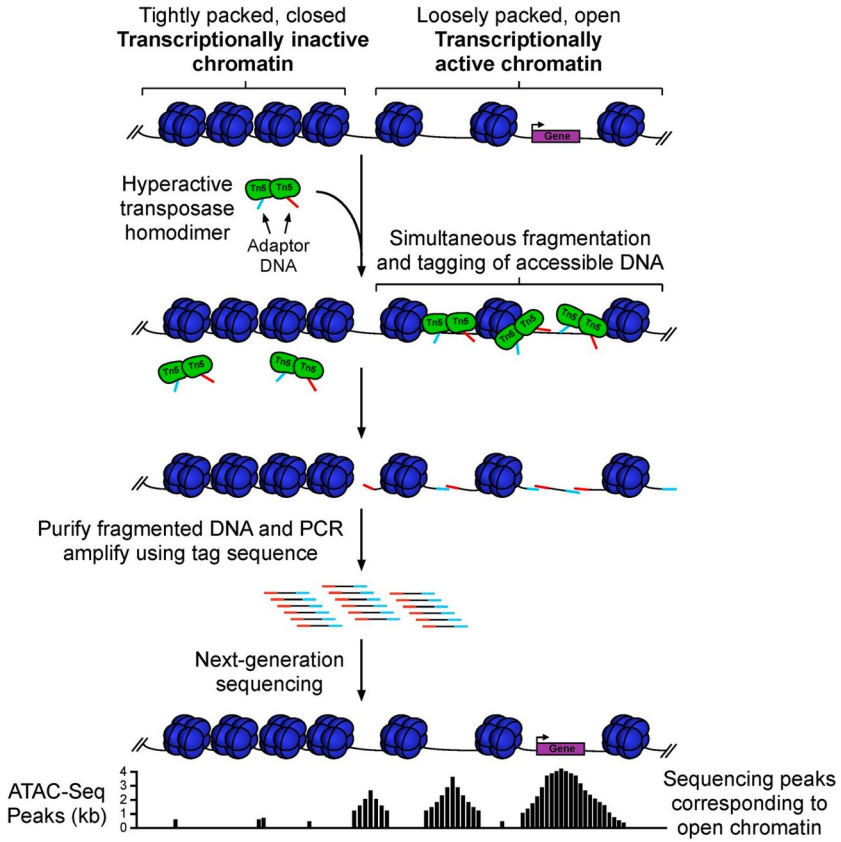
Aim 2: Develop computational modules for comprehensive assessment and correction of batch effects or sample-specific effects across tumor specimens.

Aim 1: Develop computational modules for analysis and display of single-cell CITE-seq and ATAC-seq data generated from tumors.

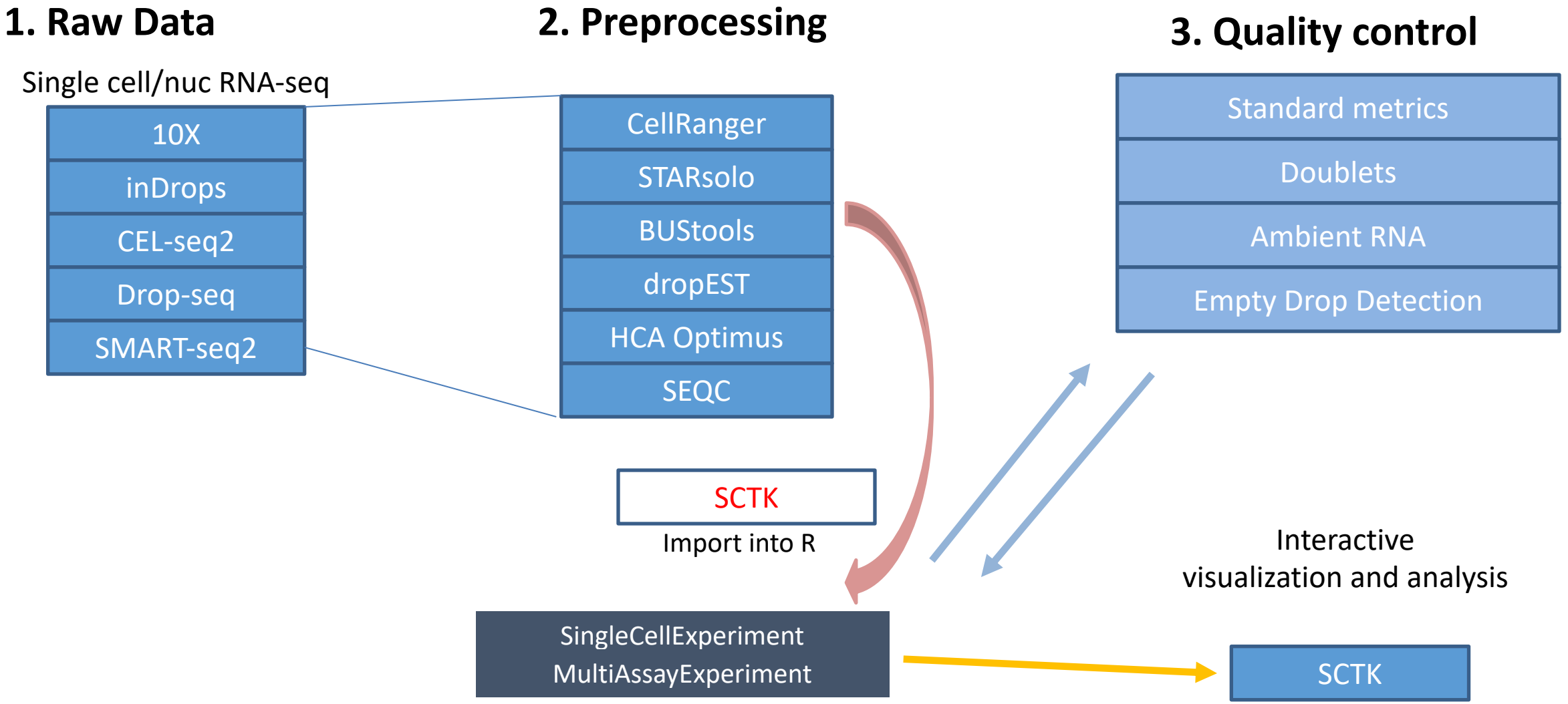
Total-Seq/CITE-seq antibody derived tags (ADTs) can be used to measure protein levels at single cell resolution



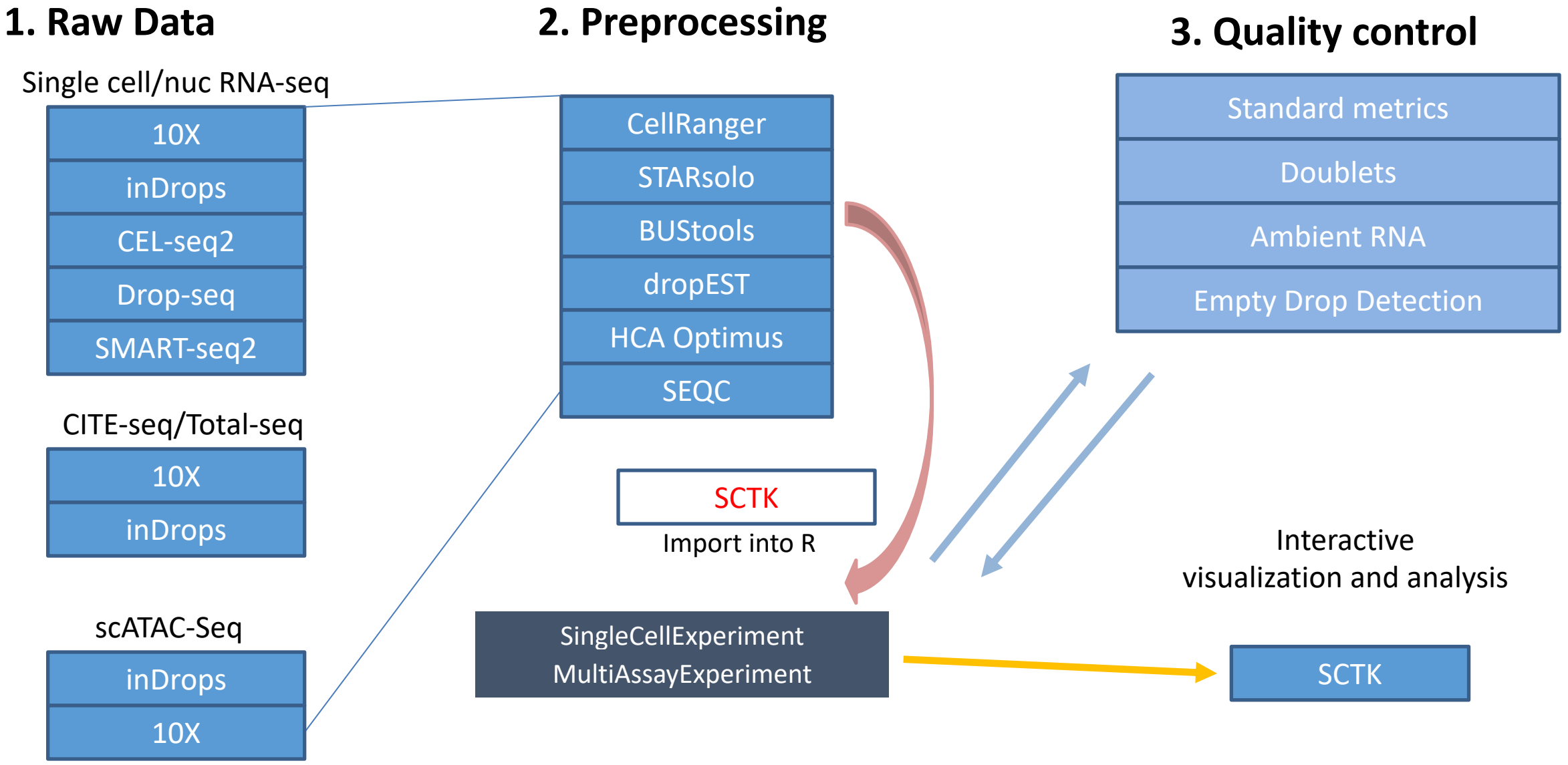
ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a technique used in molecular biology to assess genome-wide chromatin accessibility.



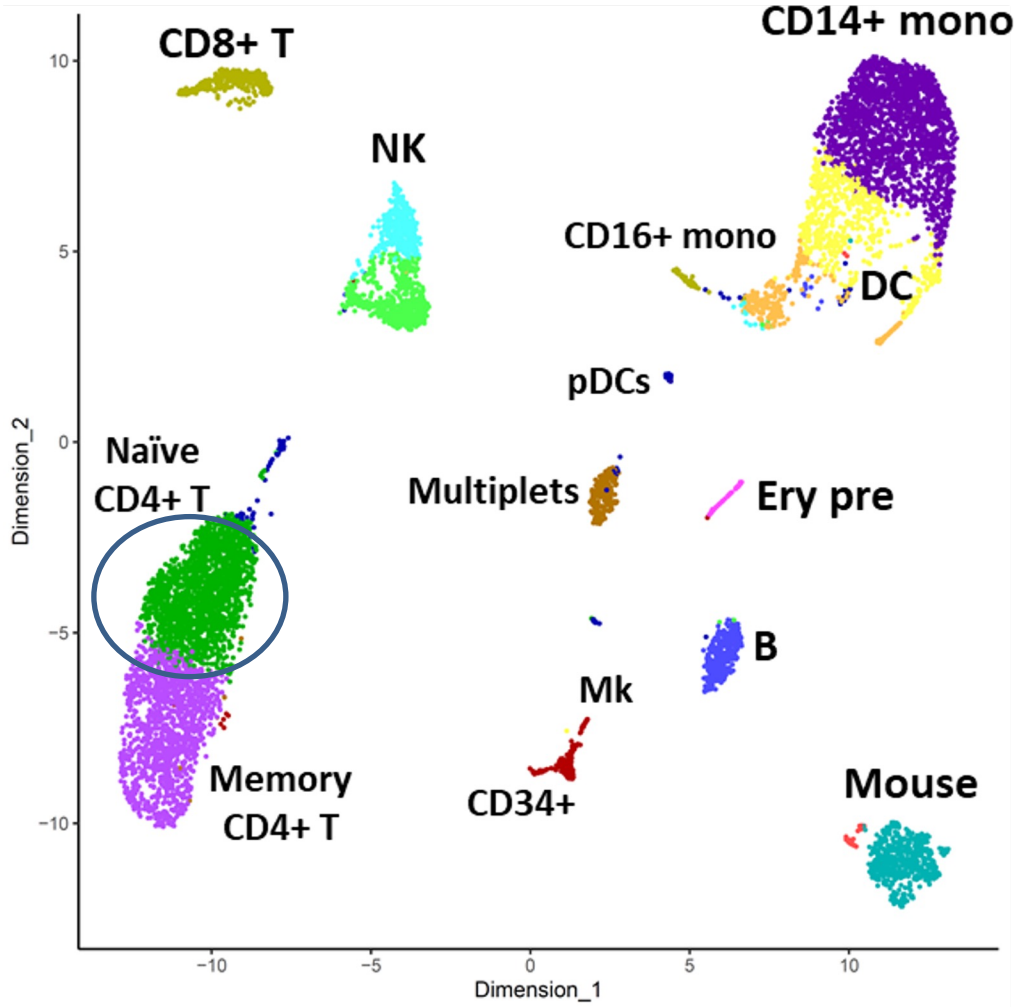
Comprehensive scRNA-seq QC pipeline for the Human Tumor Atlas Network (HTAN)



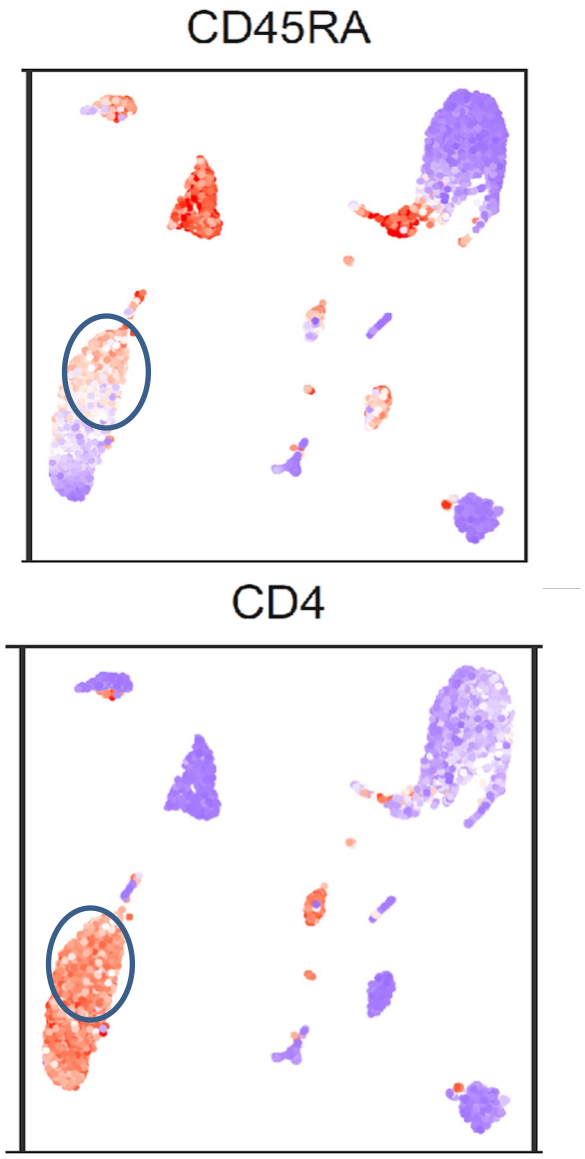
Expansion to include import functions for Total-seq/CITE-seq or scATAC-seq data in R data containers.



Developing novel statistical approaches for joint clustering of ADT (protein) and scRNA-seq data.



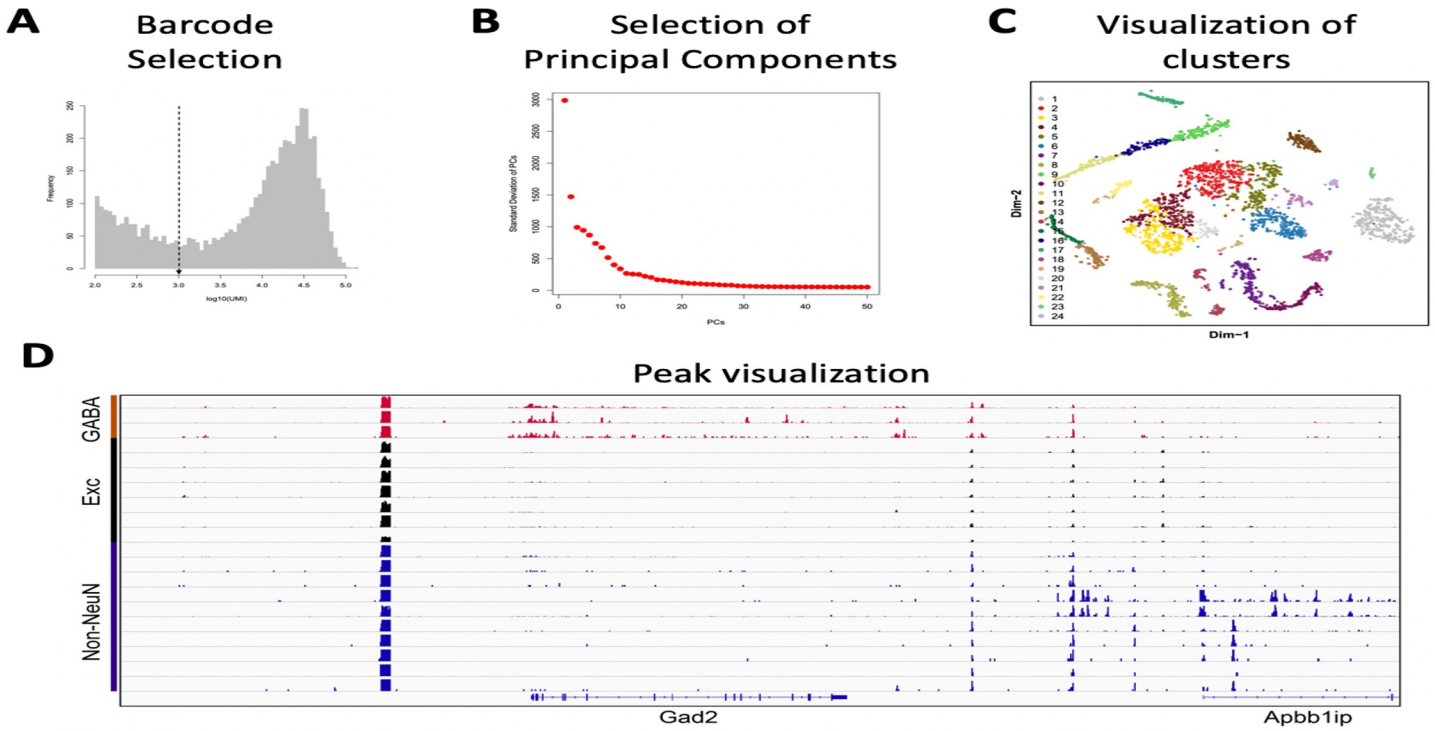
https://github.com/campbio/celda/



Develop computational modules for analysis and display of single-cell ATAC-seq data generated from tumors .

Fast and Accurate Clustering of Single Cell Epigenomes Reveals *Cis*-Regulatory Elements in Rare Cell Types

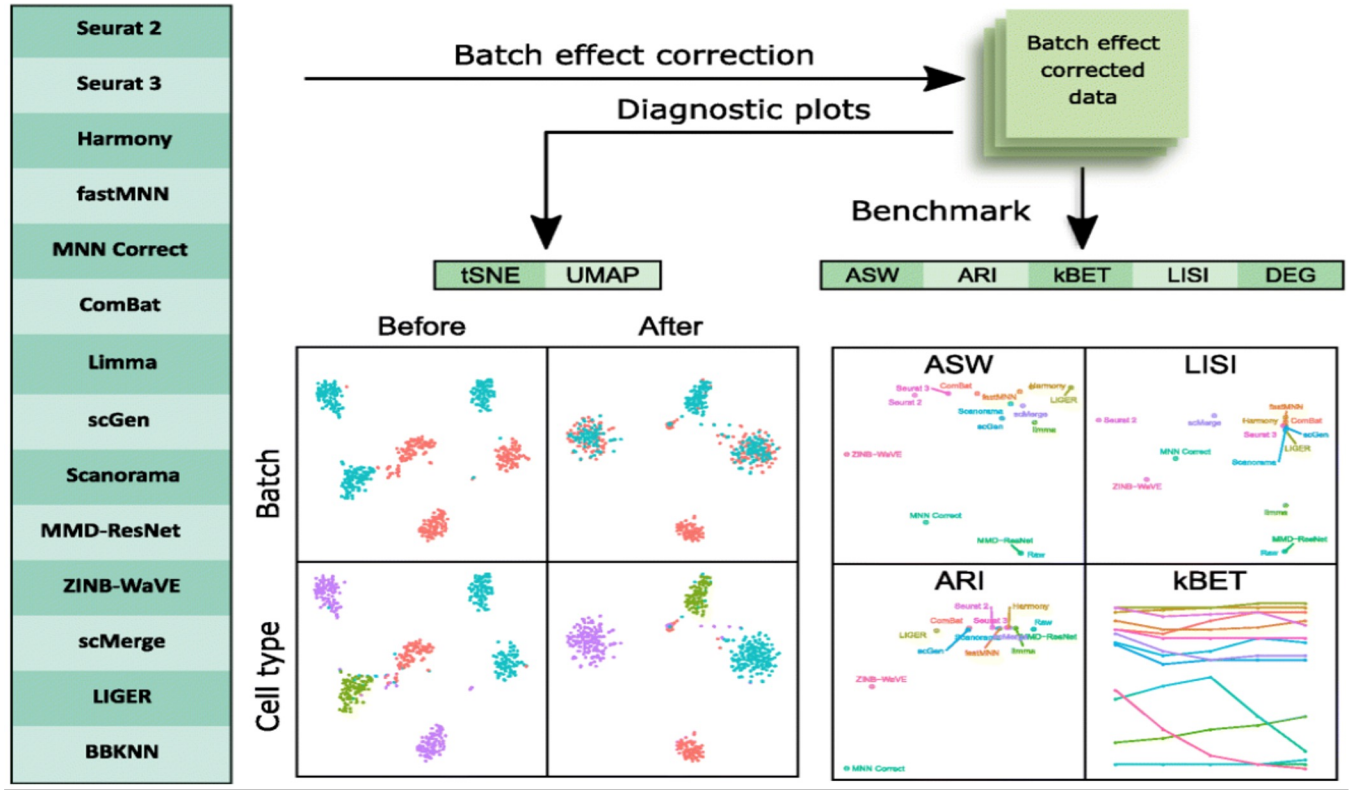
Rongxin Fang, Sebastian Preissl, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, Eran A. Mukamel, Yanxiao Zhang, M. Margarita Behrens, Joseph Ecker, Bing Ren



snapATAC

Aim 2: Develop computational modules for comprehensive assessment and correction of batch effects or sample-specific effects across tumors.

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



Tran et al, Genome Biology, 2020

Batch correction tools implemented in SCTK:

1) ComBat/ComBat-Seq
2) Seurat3 Integration
3) Harmony
4) scMerge
5) FastMNN
6) MNNcorrect
7) BBKNN
8) LIGER
9) scGEN
10) Scanorama
11) ZinB-wave

ComBat-Seq: Appropriate assumptions for count data

Mean-variance dependence in RNA-seq counts:

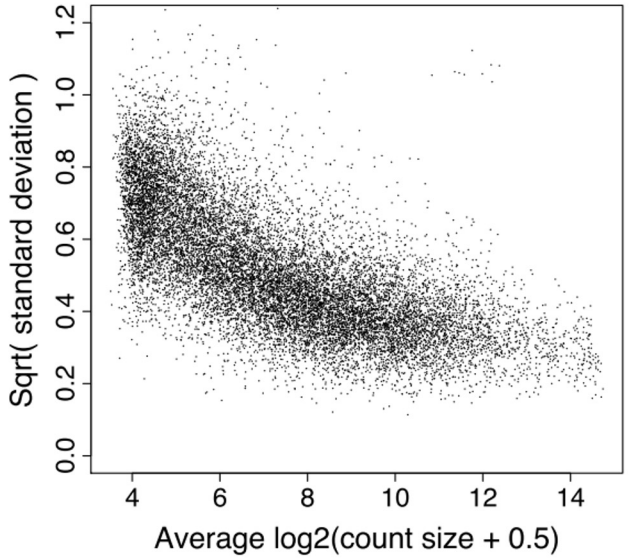
- Over-dispersion (variance > mean)
- Genes with smaller counts tend to have larger variance

Negative Binomial (NB):

$$y \sim NB(\mu, \phi)$$

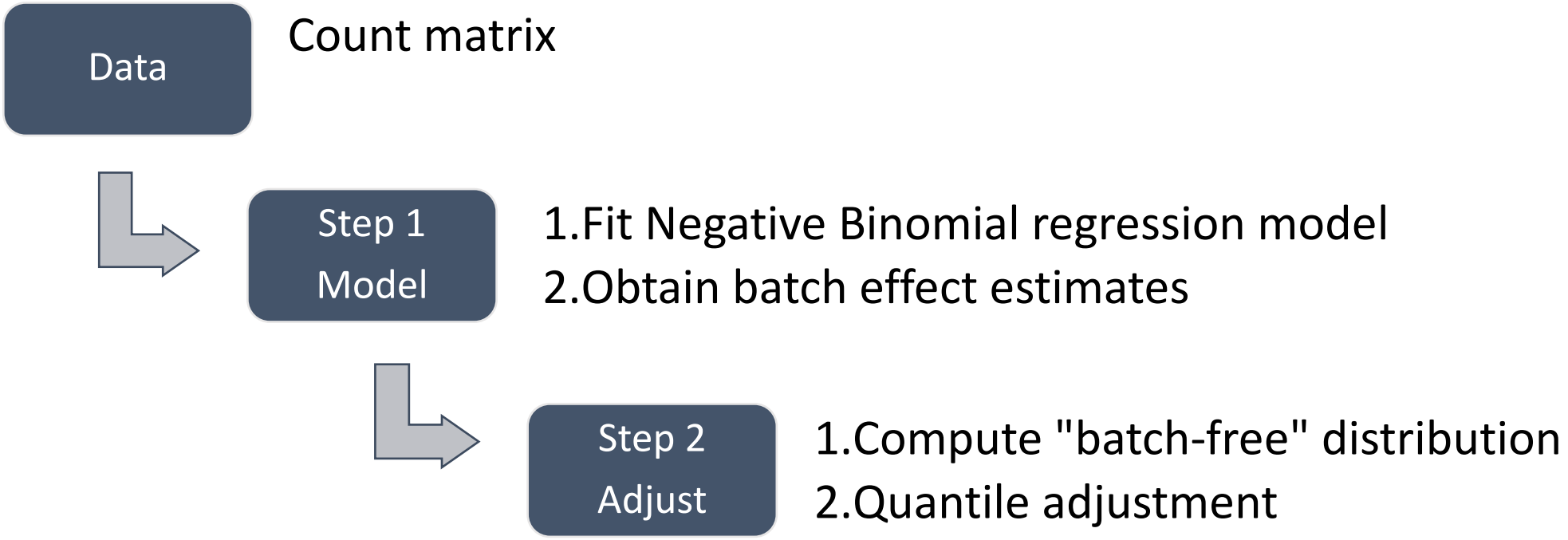
- Negative Binomial regression used in edgeR & DESeq2
- Variance is a function of mean

$$var(y) = \mu + \phi\mu^2$$



Law, Charity W., et al. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." Genome biology 2014

ComBat-Seq algorithm



ComBat-Seq algorithm: Model

Negative Binomial regression

Gene-wise model: for a certain gene g, count in sample j from batch i $\quad y_{gij} \sim NB(\mu_{gij}, \phi_{gi})$

$$\log \mu_{gij} = \alpha_g + X_j \beta_g + \gamma_{gi} + \log N_j \qquad Var(y_{gij}) = \mu_{gij} + \phi_{gi}\mu_{gij}^2$$

Decompose scaled counts into 3 components

- α_g — Average level for gene g (in "negative" samples)
- $X_j \beta_g$ — Biological condition of sample j
- γ_{gi} — Mean batch effect

ϕ_{gi} Variance batch effect

ComBat-Seq algorithm: Adjust

Adjust the data

- Calculate parameters for “batch-free” distribution: $y^*_{gj} \sim NB(\mu^*_{gj}, \phi^*_g)$

$$\log \mu^*_{gj} = \log \hat{\mu}_{gij} - \hat{\gamma}_{gi}$$

$$\phi^*_g = \frac{1}{N_{batch}} \sum_i \hat{\phi}_{gi}$$

- Map quantiles from empirical distribution to the batch-free distribution

ComBat-Seq algorithm: Adjust

Adjust the data

Original count matrix

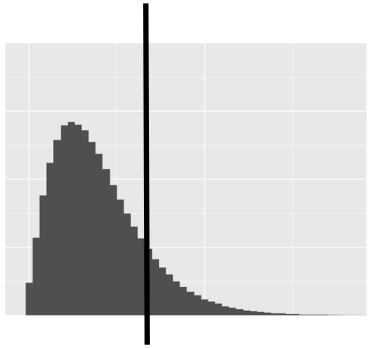
	S 1	S 2
G 1	14	12
G 2	0	0
G 3	112	11
…		

Adjusted count matrix

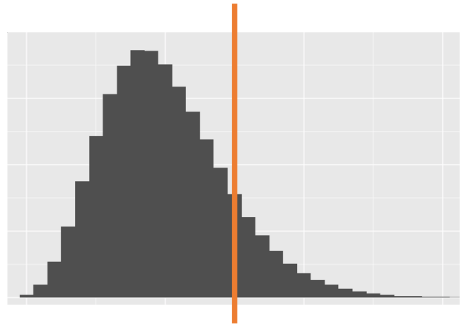
	S 1	S 2
G 1	9	8
G 2	0	0
G 3	60	47
…		

Empirical distribution of original counts:

$$y_{gij} \sim NB(\hat{\mu}_{gij}, \hat{\phi}_{gi})$$



12



8

Batch-free distribution for adjusted counts:

$$y^*_{gj} \sim NB(\mu^*_{gj}, \phi^*_g)$$

ComBat Methods for scRNA-seq Analysis

ComBat-Cell-Seq

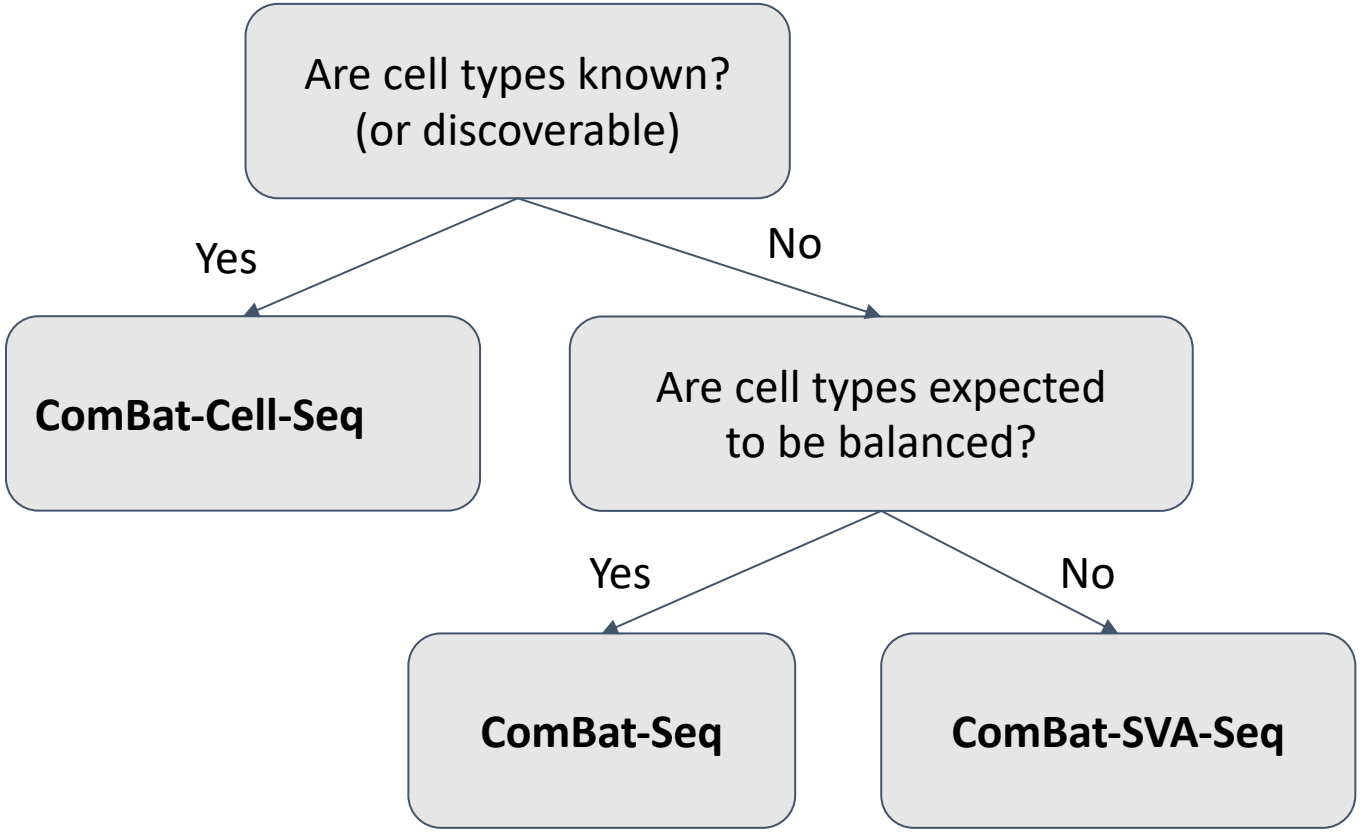
- Include cell types in ComBat design model

ComBat-Seq

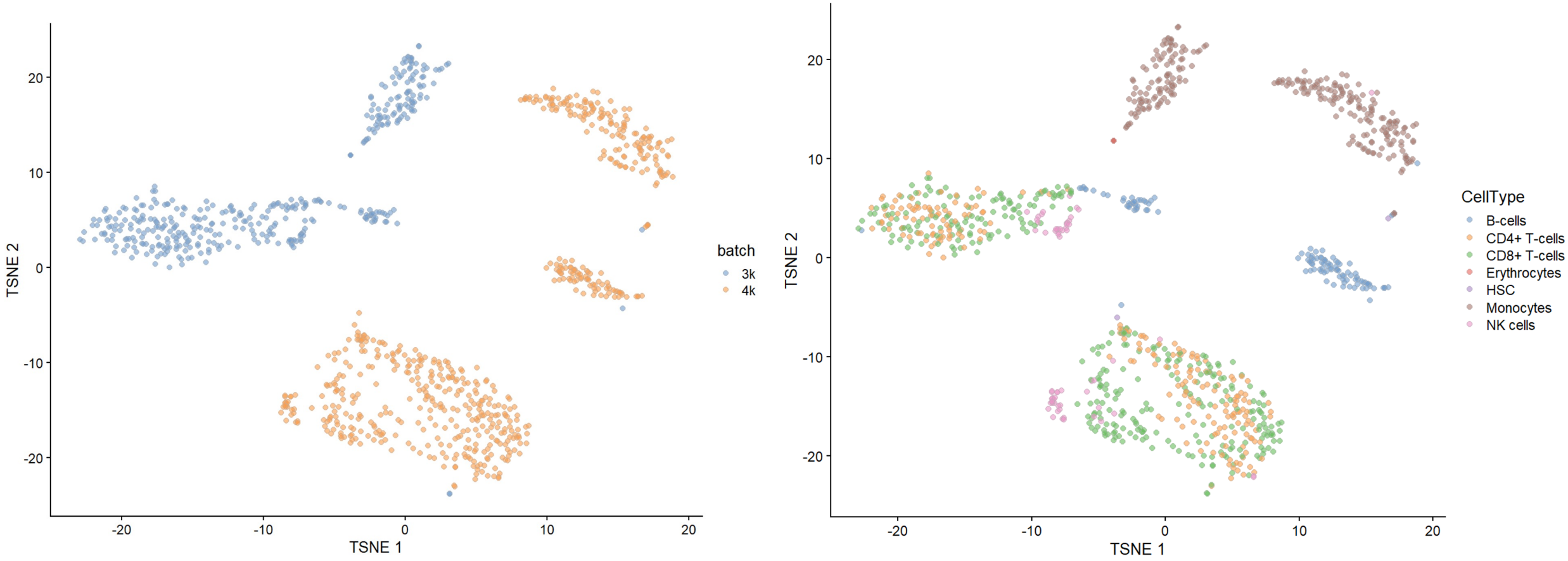
- Okay if combining batches with ‘balanced’ cell types

ComBat-SVA-Seq

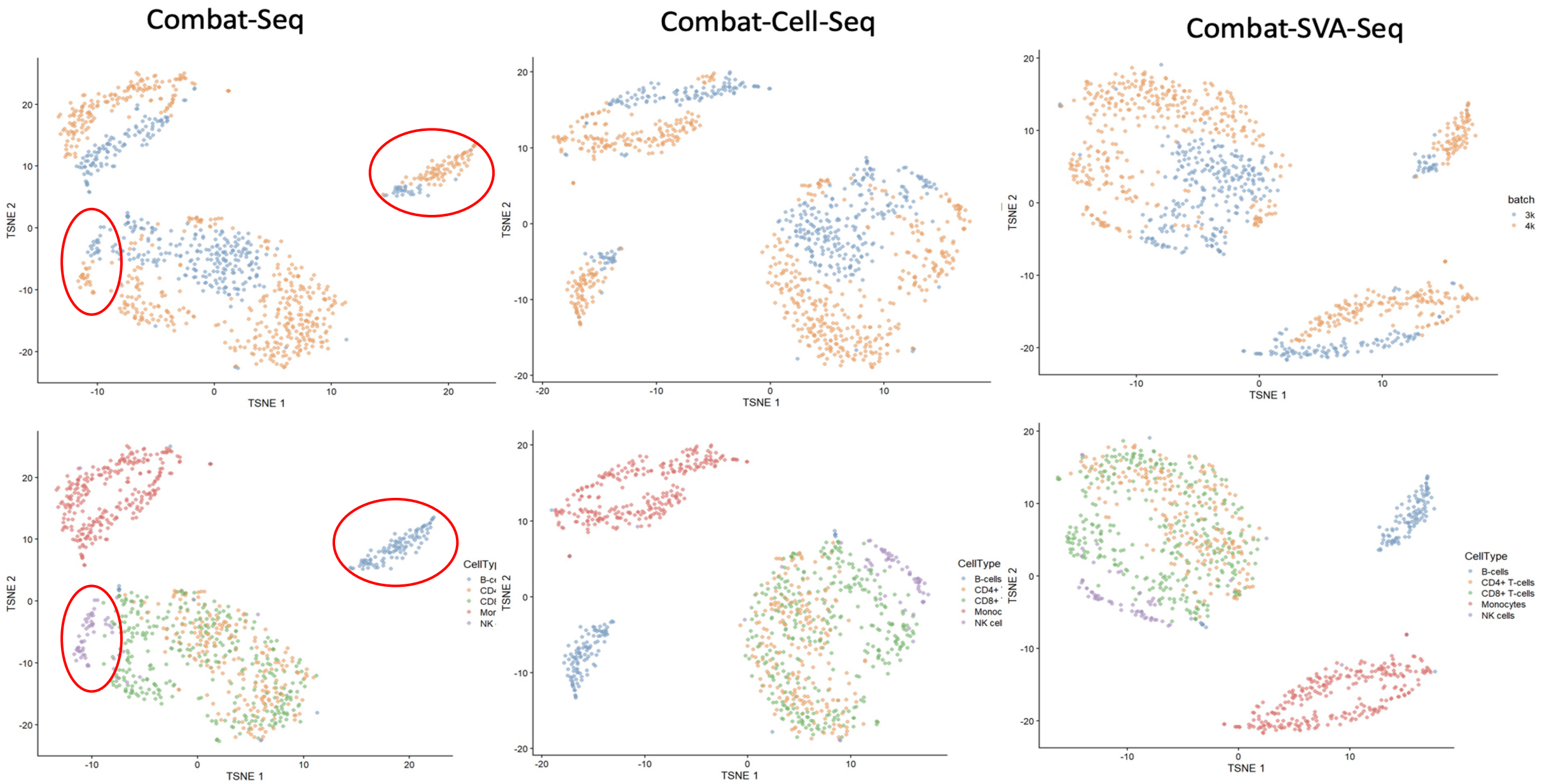
- Use SVA to identify surrogate cell-type variability



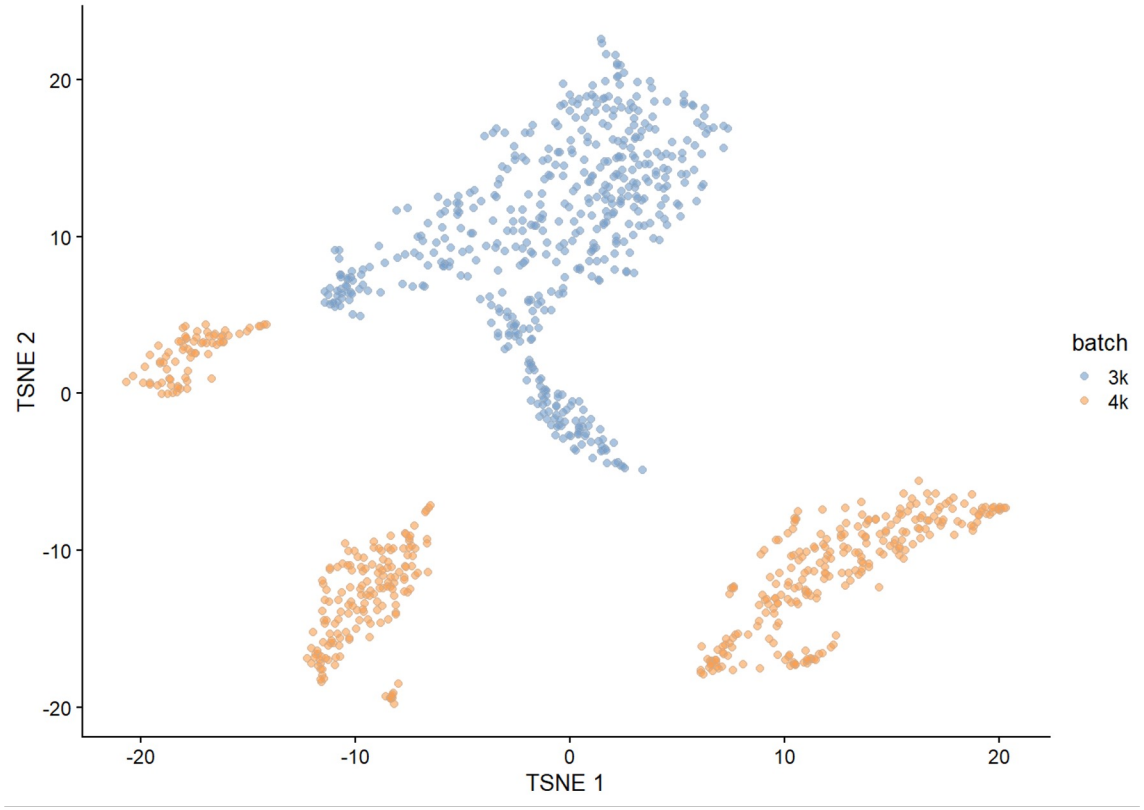
ComBat-Seq for Balanced Designs (ComBat-**Cell**-Seq, ComBat-Seq)

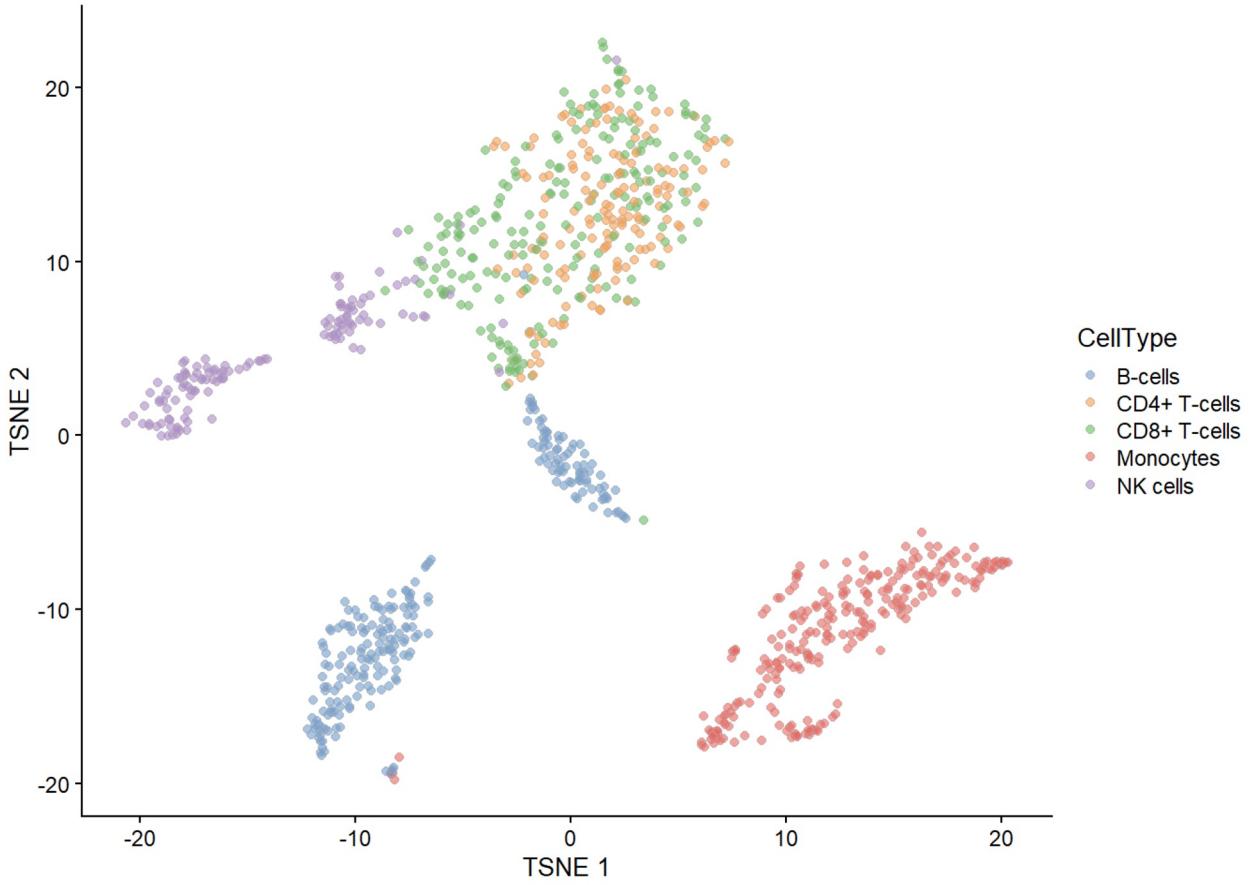


ComBat-Seq for Balanced Designs (ComBat-**Cell**-Seq, ComBat-Seq)

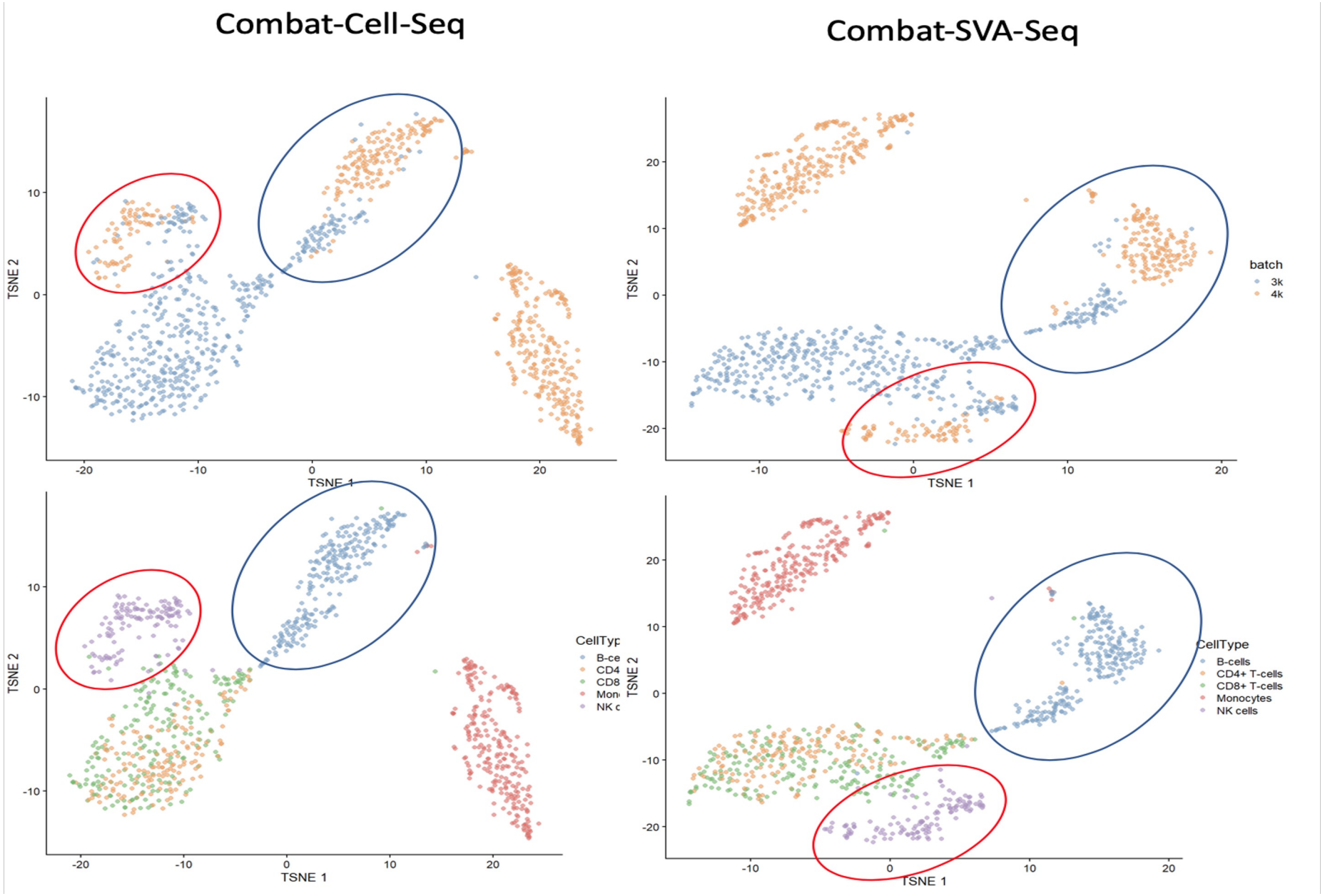


ComBat-Seq for Unbalanced Designs (ComBat-**CeII**-Seq, ComBat-**SVA**-Seq)





ComBat-Seq for Unbalanced Designs (ComBat-**Cell**-Seq, ComBat-**SVA**-Seq)



ComBat-seq Summary

For balanced designs:

- **ComBat-Seq**, **ComBat-Cell-Seq**, **ComBat-SVA-Seq** all work well!

For unbalanced designs:

- **ComBat-Seq:** May remove cell-type specific variation
- **Combat-Cell-Seq:** Performs extremely well
- **Combat-SVA-Seq**: Not quite as good as Combat-Cell-Seq, but performs well in both simulated and real-data examples