Frederick National Laboratory
for Cancer Research

Multi-Sample Analysis and Batch Effect Correction

Vicky Chen
CCR-SF Bioinformatics Group
Advanced Biomedical and Computational Sciences
Biomedical Informatics and Data Science (BIDS) Directorate
Frederick National Laboratory for Cancer Research

Multi-Sample Analysis

- Comparisons between multiple samples are needed to investigate many research questions
- Many of the processing/analysis methods are similar to single sample
- Additional cross-sample comparisons can be applied
- Combining samples increases the potential sources of technical variations

Multi-Sample Analysis

- Combining Samples
- Normalizing and Pre-processing Data
- Feature Selection
- Dimensionality Reduction
- Cluster Analysis
- Visualization
- Differential Expression / Marker Gene Identification
- Cell Type Annotation
- Copy Number Variation Estimate
- Batch Effect Correction

Multi-Sample Analysis

- **Combining Samples**
- **Normalizing and Pre-processing Data**
- **Feature Selection**
- **Dimensionality Reduction**
- **Cluster Analysis**
- **Visualization**
- **Differential Expression / Marker Gene Identification**
- **Cell Type Annotation**
- **Copy Number Variation Estimate**
- **Batch Effect Correction**

- **Most methods are similar to single sample analysis**

Combining Samples

- **10X Cell Ranger can be used to combine samples for 10X captures**
  - Aggregate function
  - Generates UMI expression matrices, basic sample statistics, and interactive analysis platform
  - Generates a summary report and also a loupe file that can be used for additional analysis

- **Many different analysis tools are available that can be used to combine samples**
  - Use of a tool that can handle further downstream analysis is simpler
    - Seurat
    - scran
    - scanpy

Combining Samples

Frederick National Laboratory for Cancer Research

Summary | Analysis

17,213

Estimated Number of Cells

Aggregation

Post-Normalization Total Number of Reads	215,118,612
Pre-Normalization Total Number of Reads	511,189,192
Pre-Normalization Mean Reads per Cell	29,698
Post-Normalization Mean Reads per Cell	12,497

Cells

Estimated Number of Cells	103,278
Fraction Reads in Cells	92.7%
Median Genes per Cell	1,551
Median UMI Counts per Cell	5,759

Sample

Sample ID	AggregatedDatasets

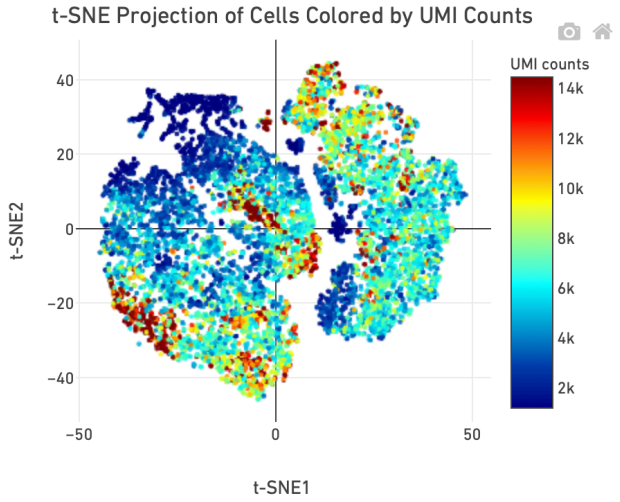
Combining Samples

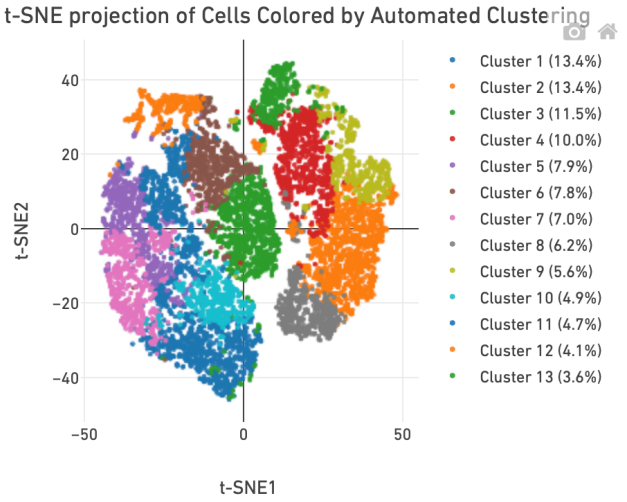
Frederick National Laboratory for Cancer Research

Summary | Analysis

t-SNE Projection

Clustering Type: Graph-based



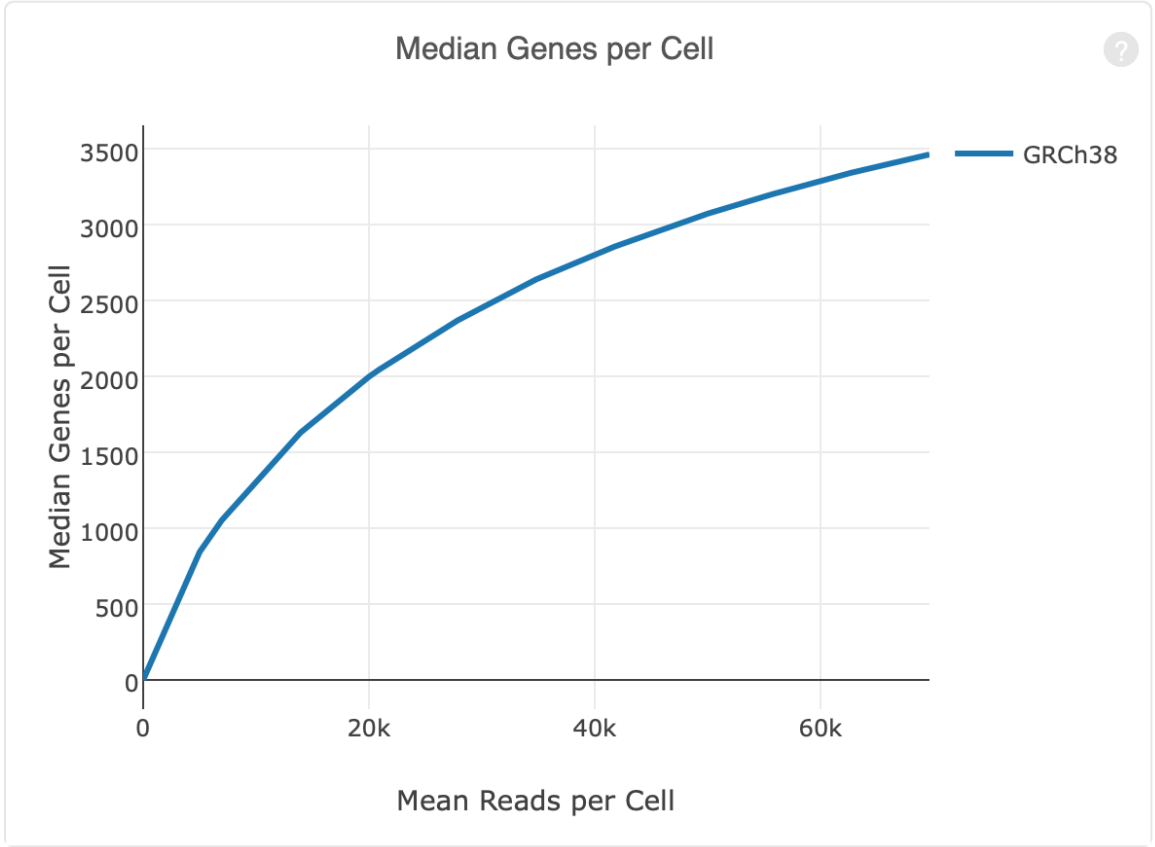


Top Features By Cluster (Log2 fold-change, p-value)

Feature		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		C
ID	Name	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2F
ENSG00000197061	HIST1H4C	1.52	3e-07	-2.45	2e-08	-3.47	3e-16	1.64	2e-06	-1.05	7e-01	-4.99	4e-07	-0
ENSG00000136732	GYPC	1.40	4e-06	-7.76	2e-31	0.45	3e-01	-7.97	7e-25	0.71	7e-01	0.05	1e+00	0
ENSG00000115758	ODC1	1.36	7e-06	-2.77	1e-09	0.67	7e-02	-1.86	4e-04	0.12	1e+00	-0.11	1e+00	0
ENSG00000163950	SLBP	1.32	2e-05	-2.02	6e-06	0.93	6e-03	-0.33	7e-01	-0.57	1e+00	-0.24	1e+00	-0

Normalizing

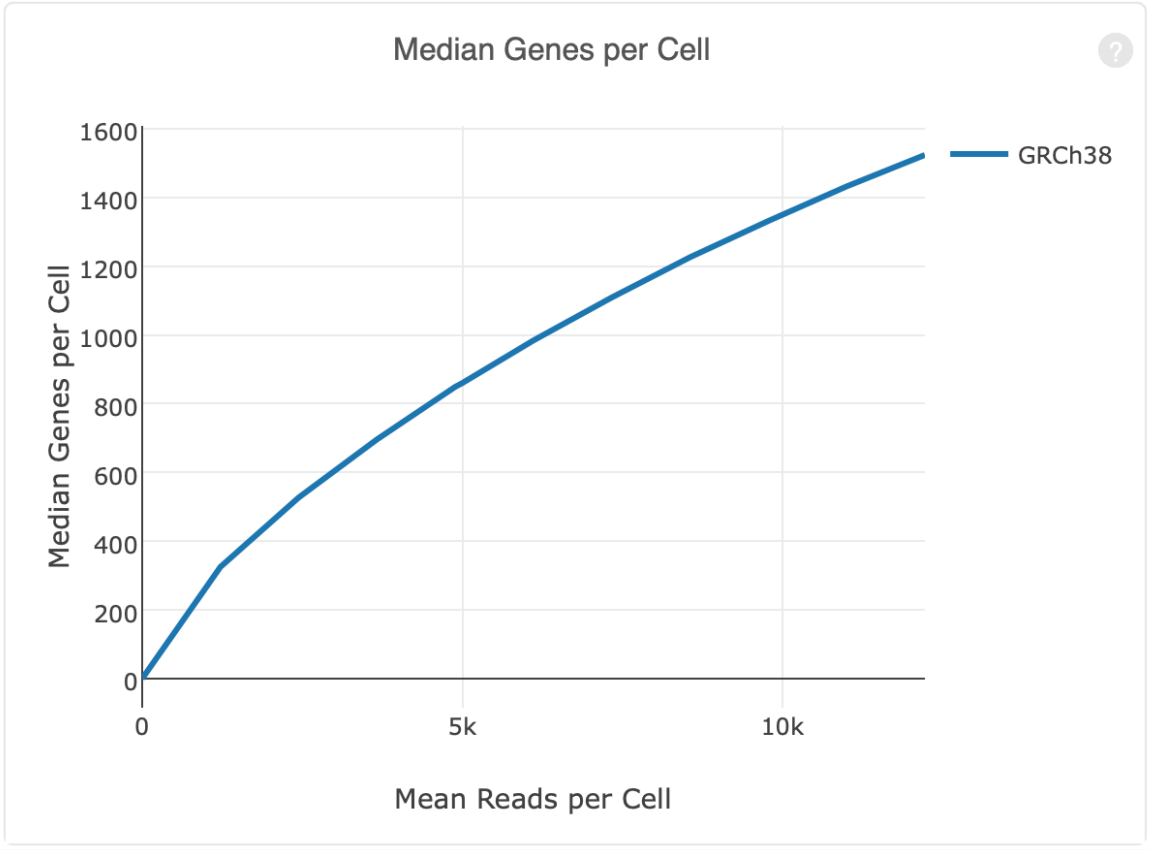
- **Sequencing depth is a factor in detected gene expression**
  - Deeper sequencing results in a more comprehensive picture of the captured transcripts
- **In addition to normalizing by cell can now also normalize sample**
- **Directly comparing samples sequenced at different depths can bias the detection of differences**



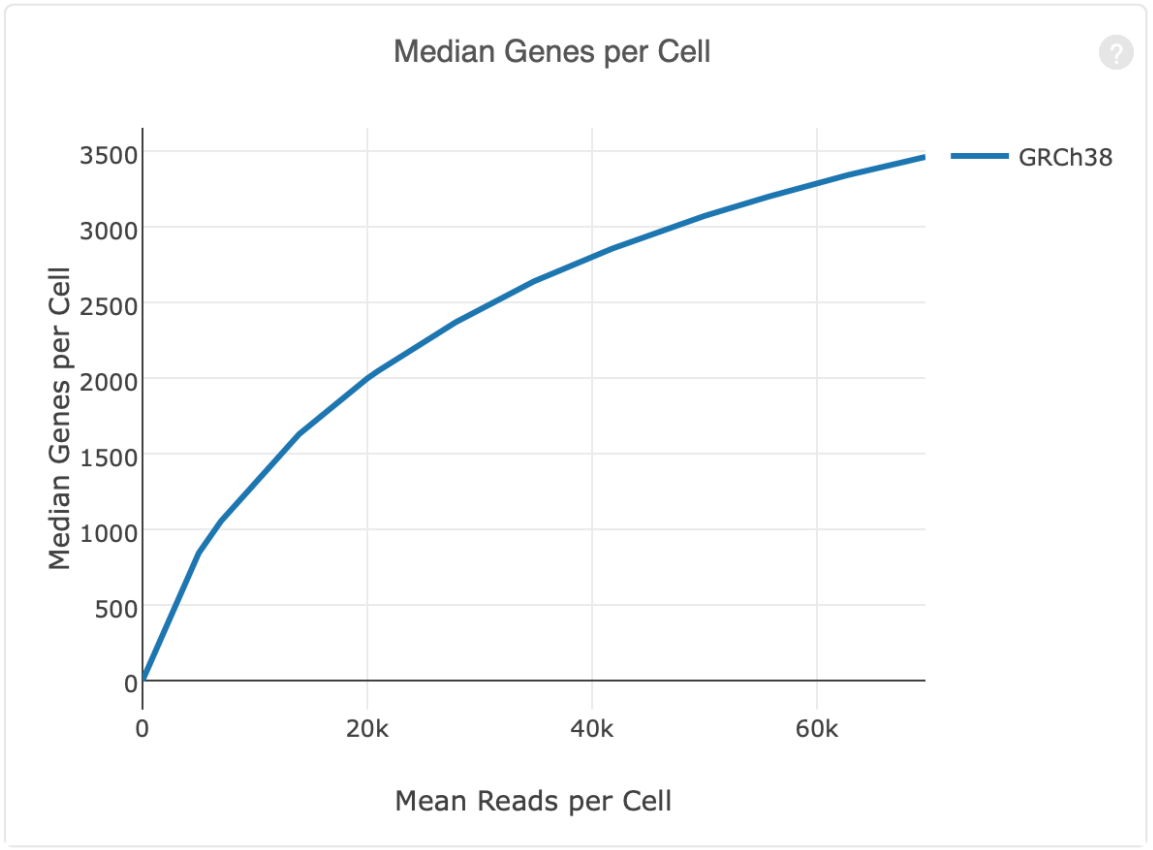
Normalizing

Frederick National Laboratory for Cancer Research

12k Mean Reads per Cell



69k Mean Reads per Cell



Comparing samples at different read depths can bias the detected results, even if samples are normalized by cell

Read Depth Normalization

- **Subsamples reads until all samples have equal number of reads per cell**
  - Data loss with the additional reads essentially “thrown away”
  - 10X Cell Ranger default method
- **Feasibility can depend on amount of data available for each sample**
  - Reaching similar depth for each sample can be costly or not an option

Low Data Loss

Aggregation ⓘ

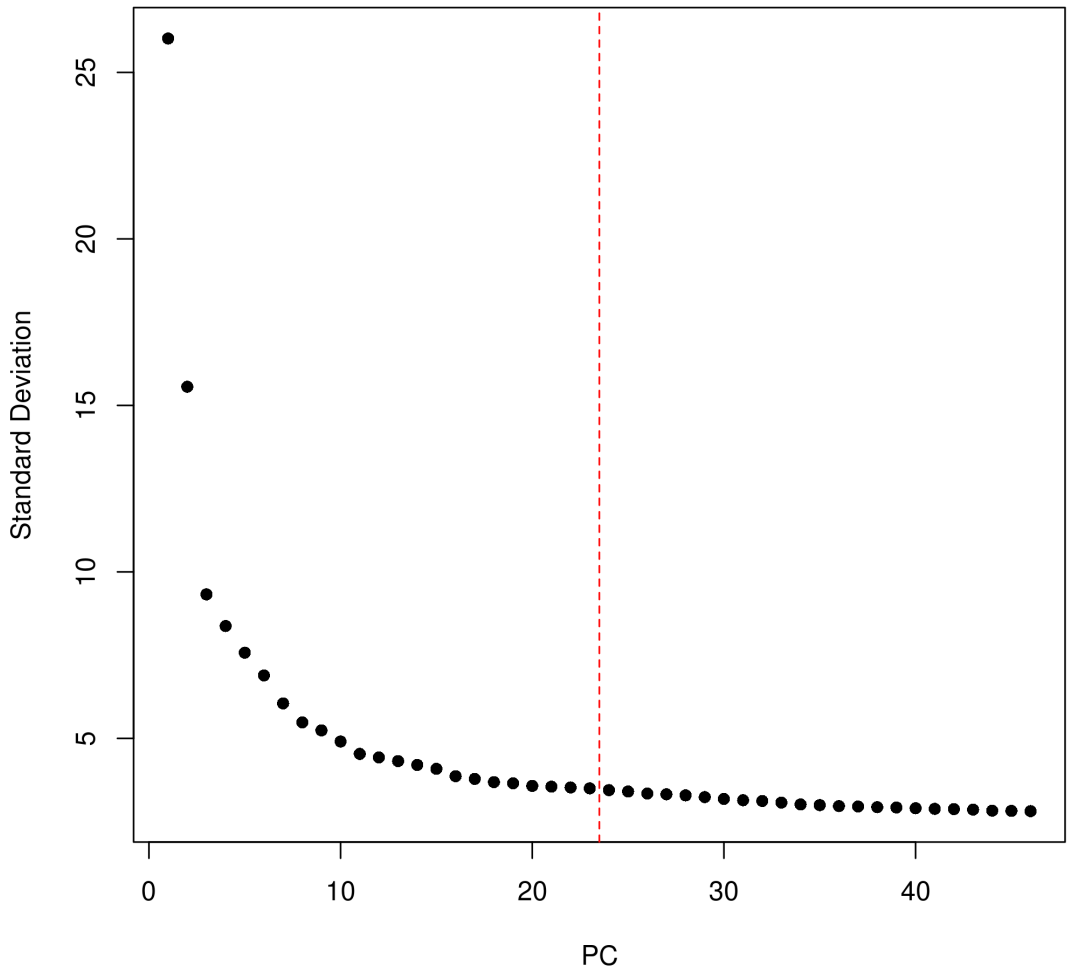
Post-Normalization Total Number of Reads	330,680,726
Pre-Normalization Total Number of Reads	335,415,269
Pre-Normalization Mean Reads per Cell	47,003
Post-Normalization Mean Reads per Cell	46,340

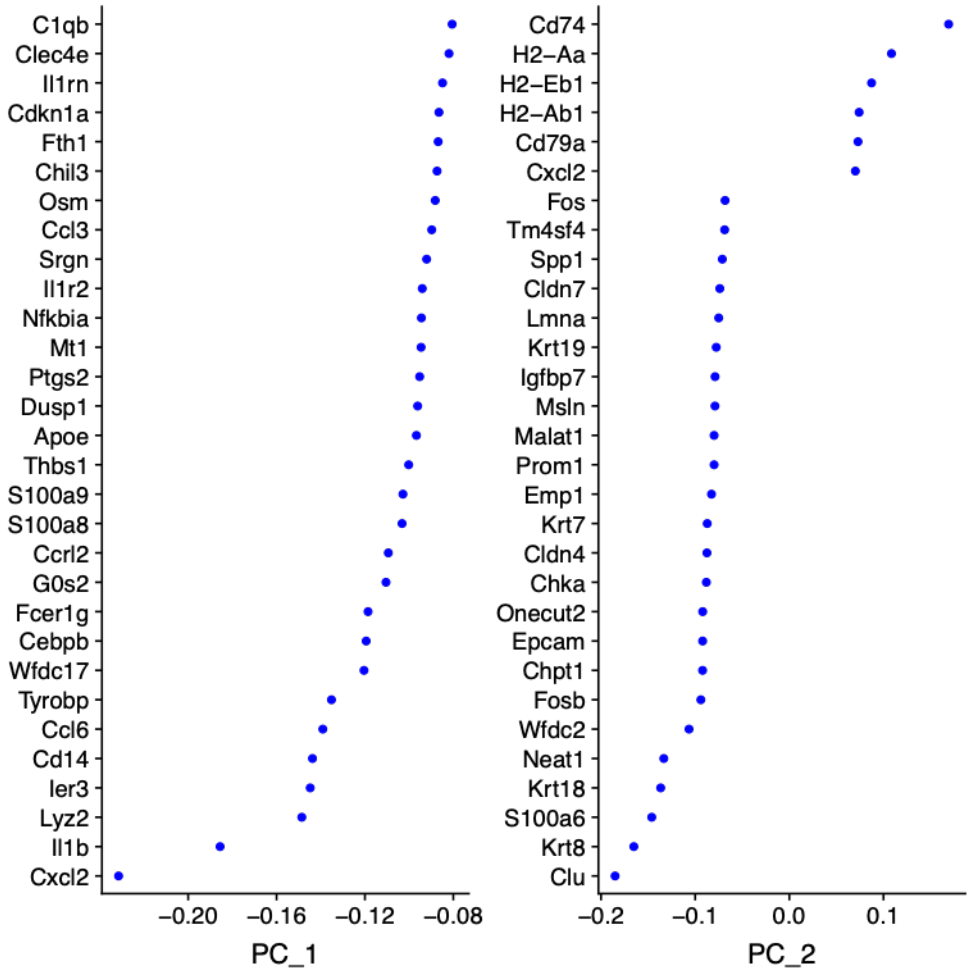
High Data Loss

Aggregation ⓘ

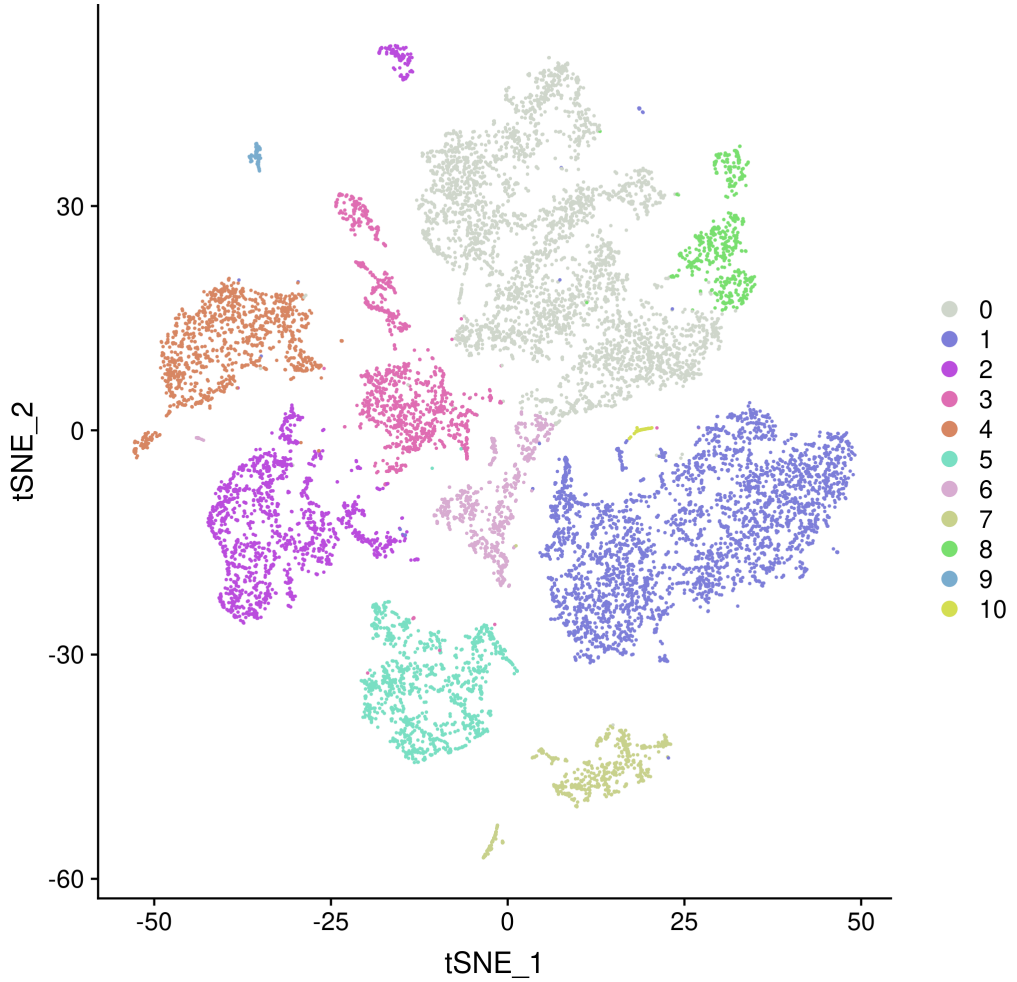
Post-Normalization Total Number of Reads	291,002,098
Pre-Normalization Total Number of Reads	1,251,974,594
Pre-Normalization Mean Reads per Cell	69,931
Post-Normalization Mean Reads per Cell	16,254

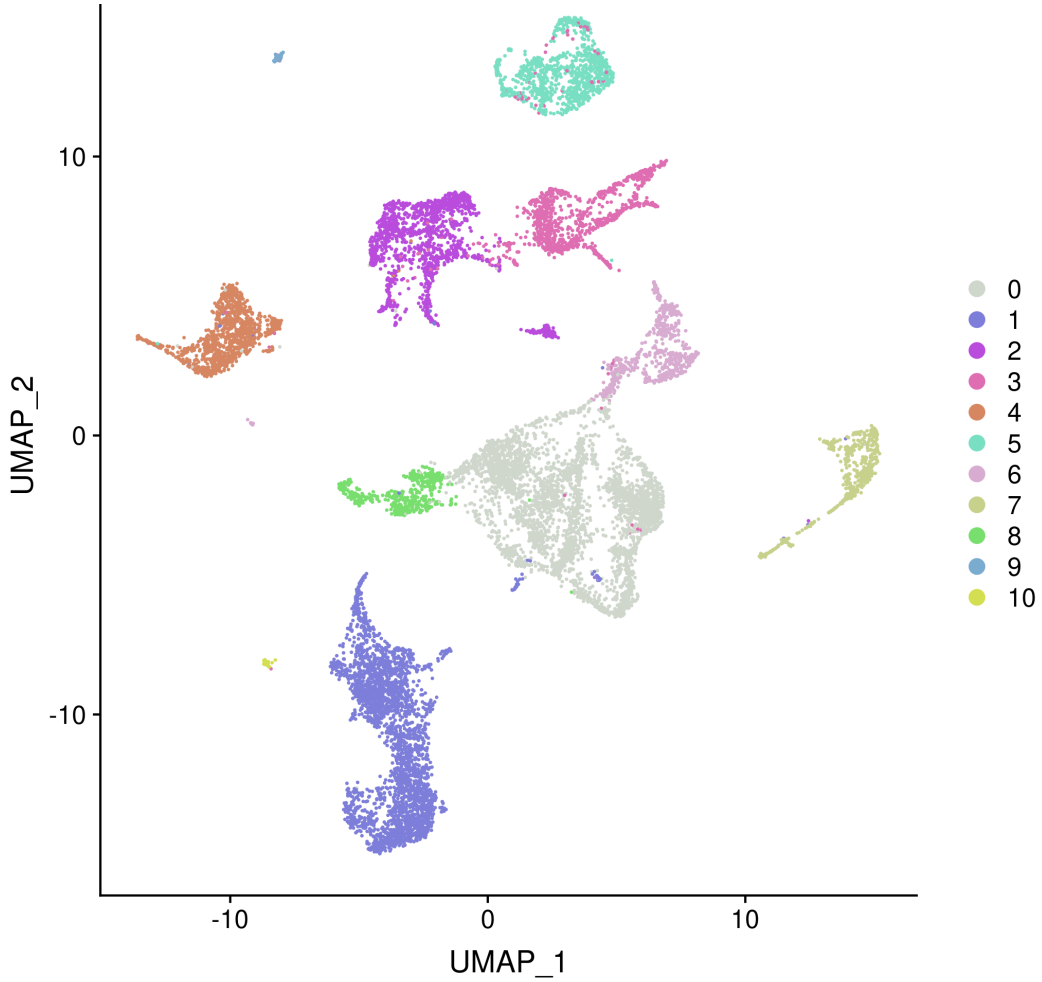
Feature Selection, Dimensionality Reduction





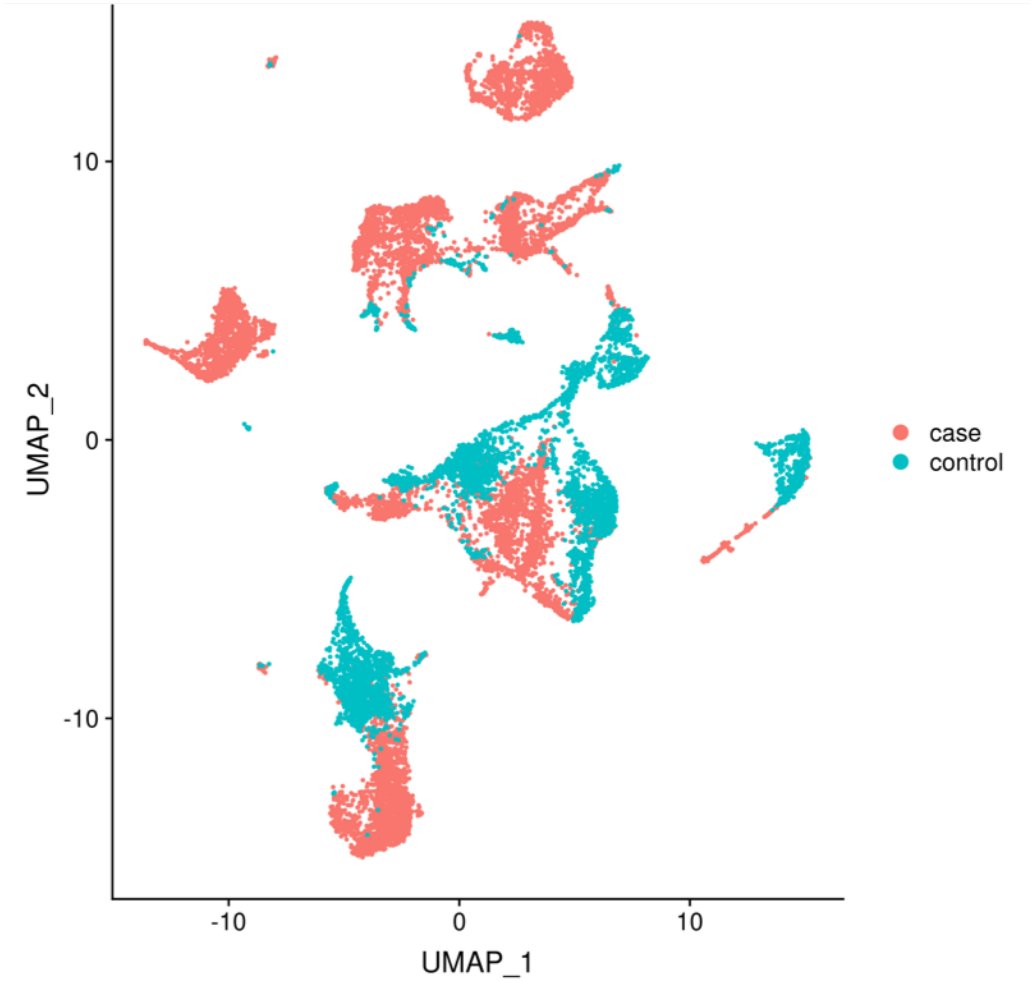
Cluster Analysis and Visualization

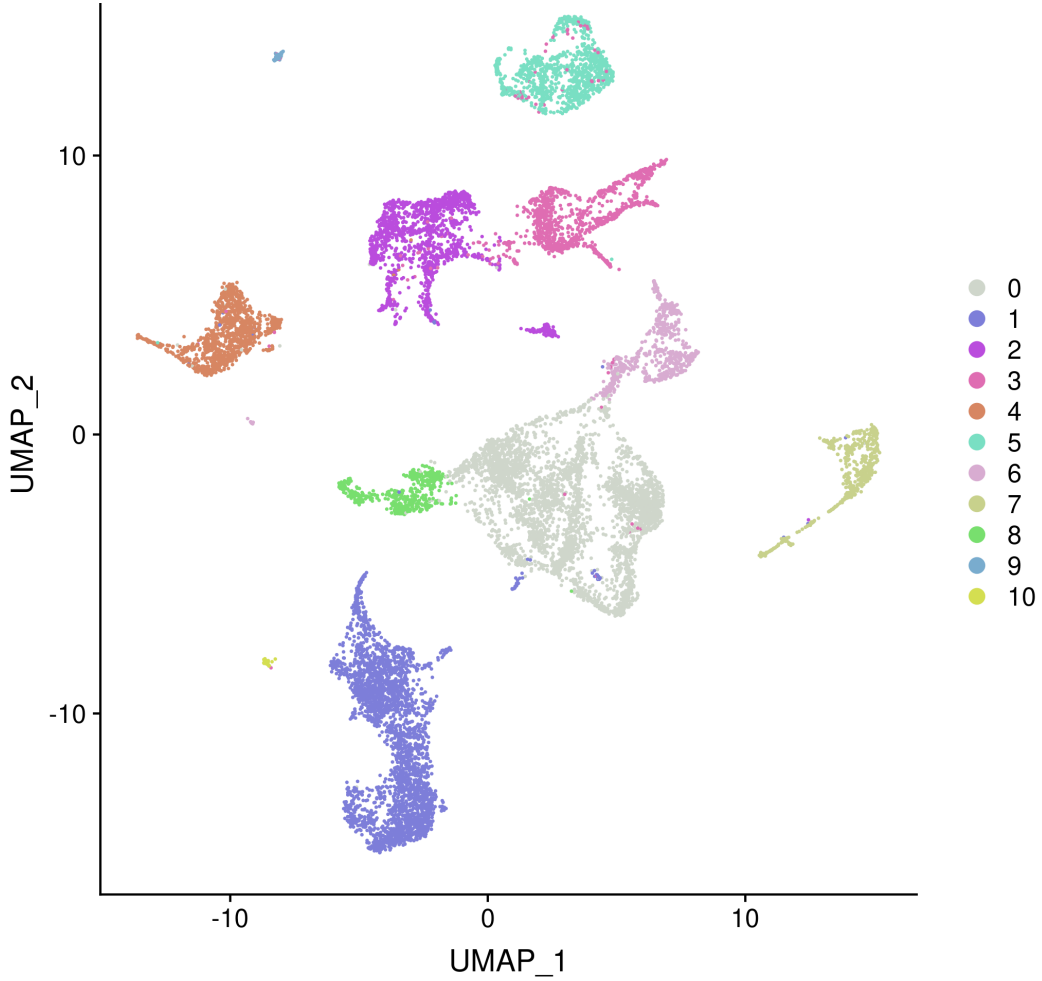




Cluster Analysis and Visualization

Frederick National Laboratory for Cancer Research

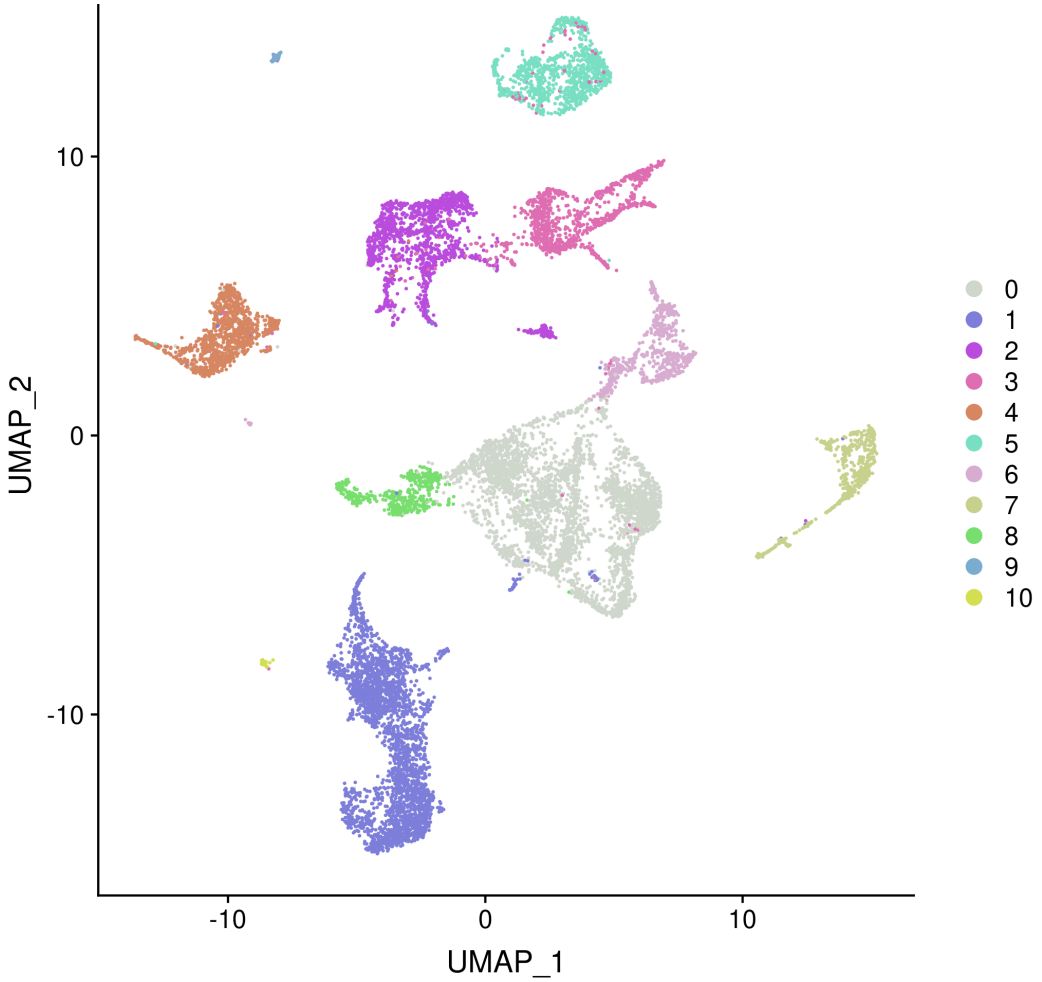


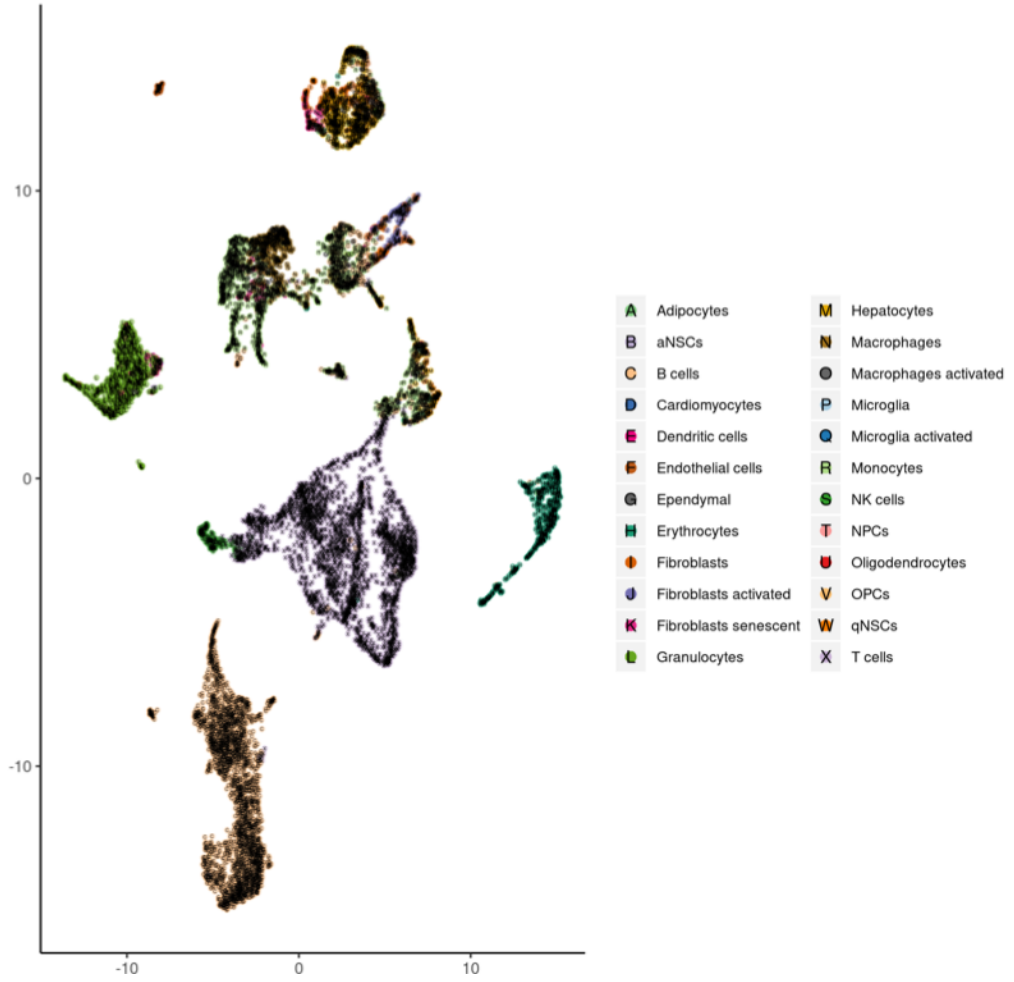


Differential Expression

- **Can now do cross-sample comparison in addition to cluster analysis**
- **It may be useful to identify a set of cells of interest or exclude cells that are not of interested prior to cross-sample comparison**
- **Actual computational methods are the same as those mentioned in single sample analysis**
  - limma-trend, edgeR, MAST, etc

Cell Type Annotation



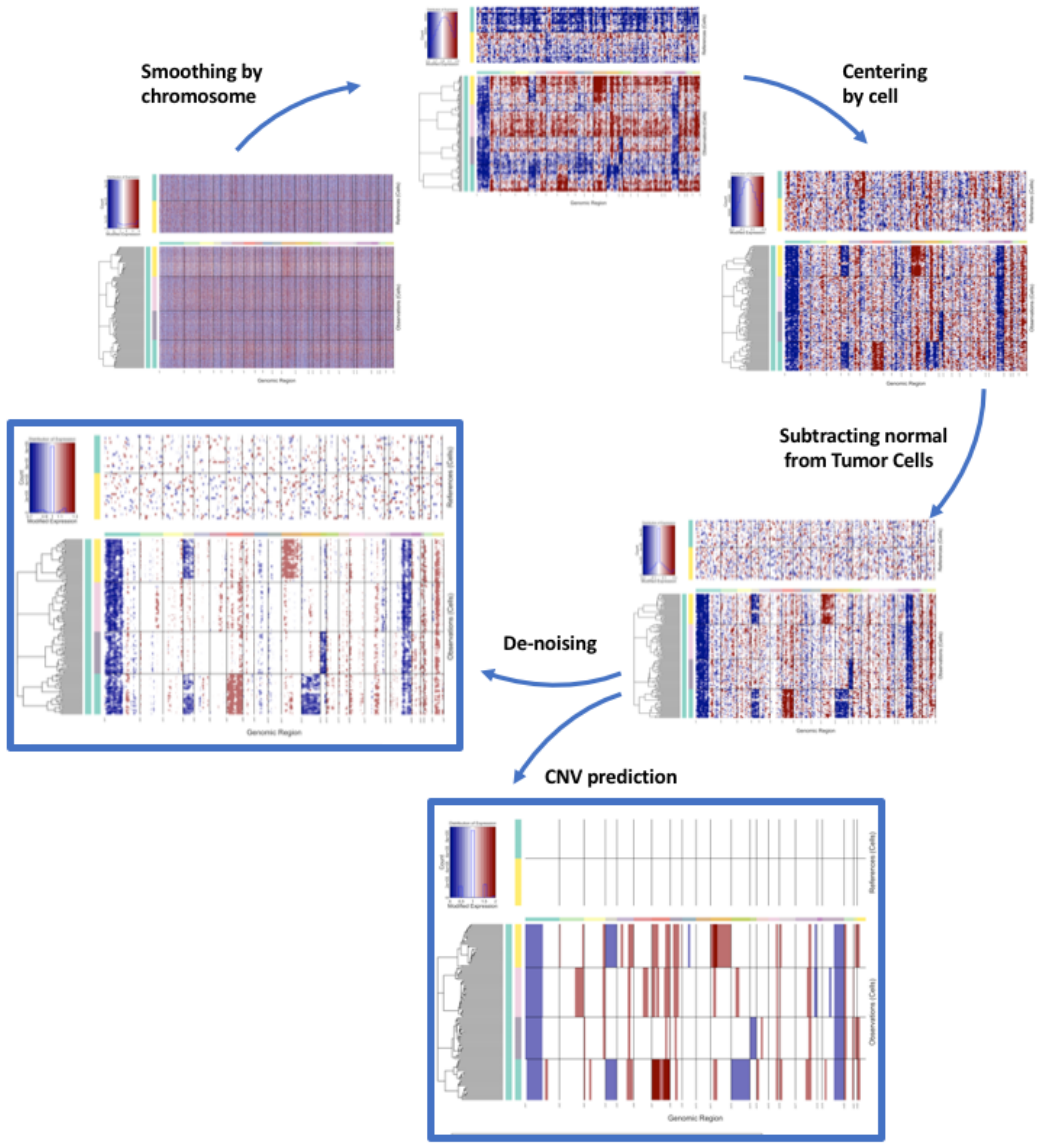


Copy Number Variation Estimate

- **Using detected gene expression intensities to estimate potential copy number variations**
- **Uses either a given reference or estimated normal to detect decreases/increases in expression**
- **Multiple tools available to generate estimates**
  - InferCNV
  - BADGER / HoneyBADGER
  - Conics

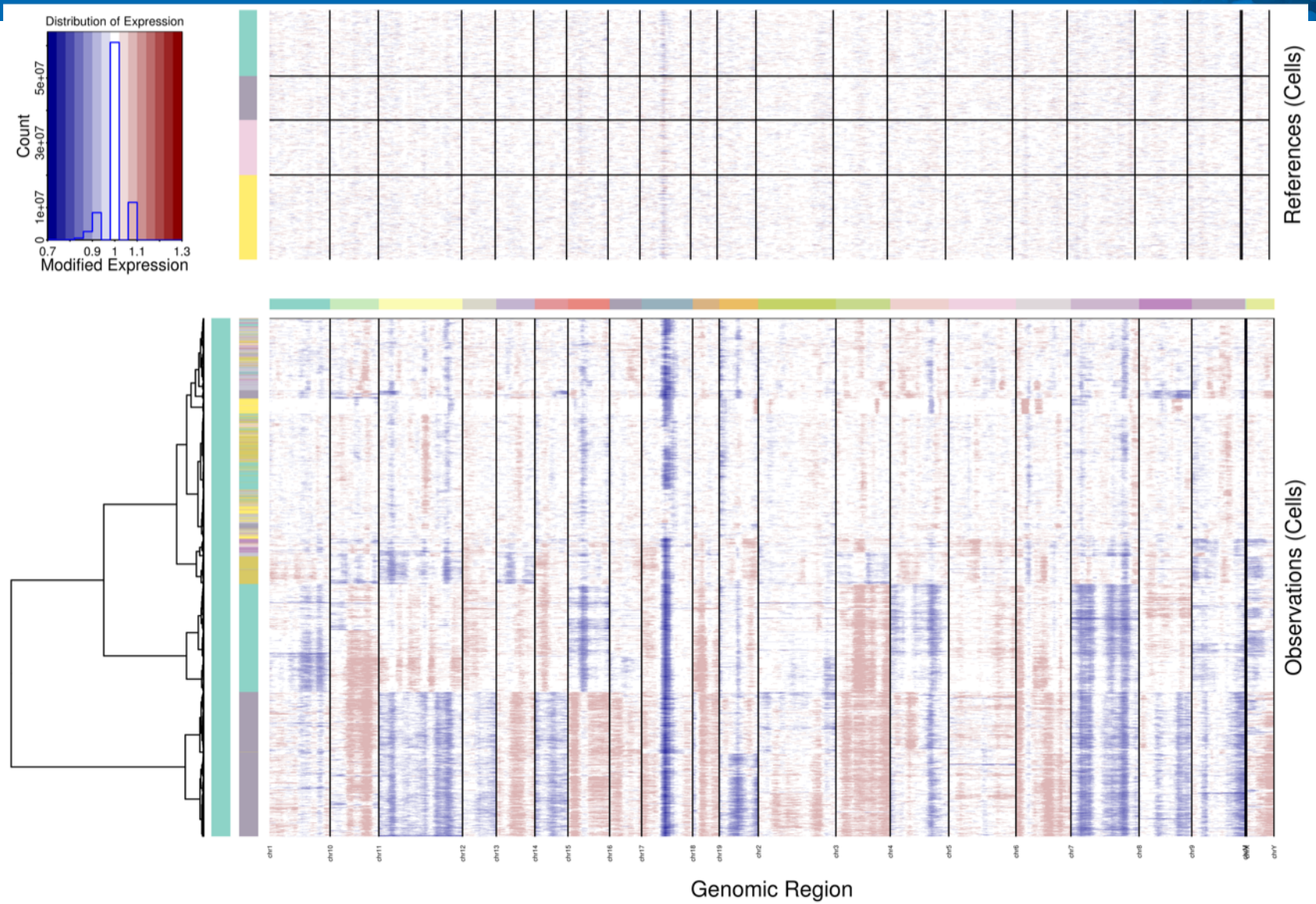
inferCNV

- **Expression intensity of genes along each chromosome are smoothed**
- **Expression intensity of each cell is centered at zero**
- **The mean of the normal cells are subtracted from tumor cells**

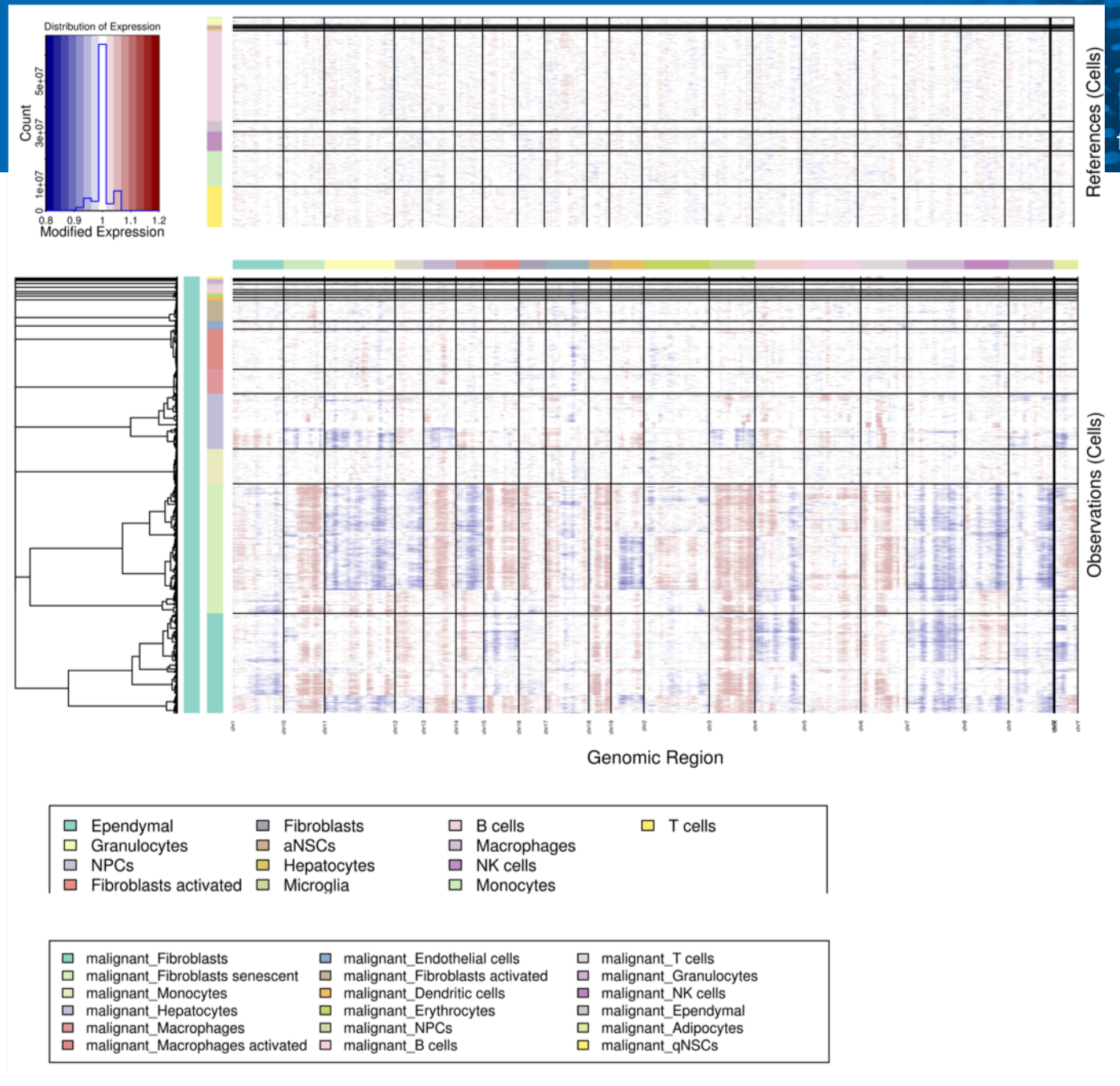


https://github.com/broadinstitute/inferCNV

inferCNV – Normal vs Tumor



inferCNV – Cell Type



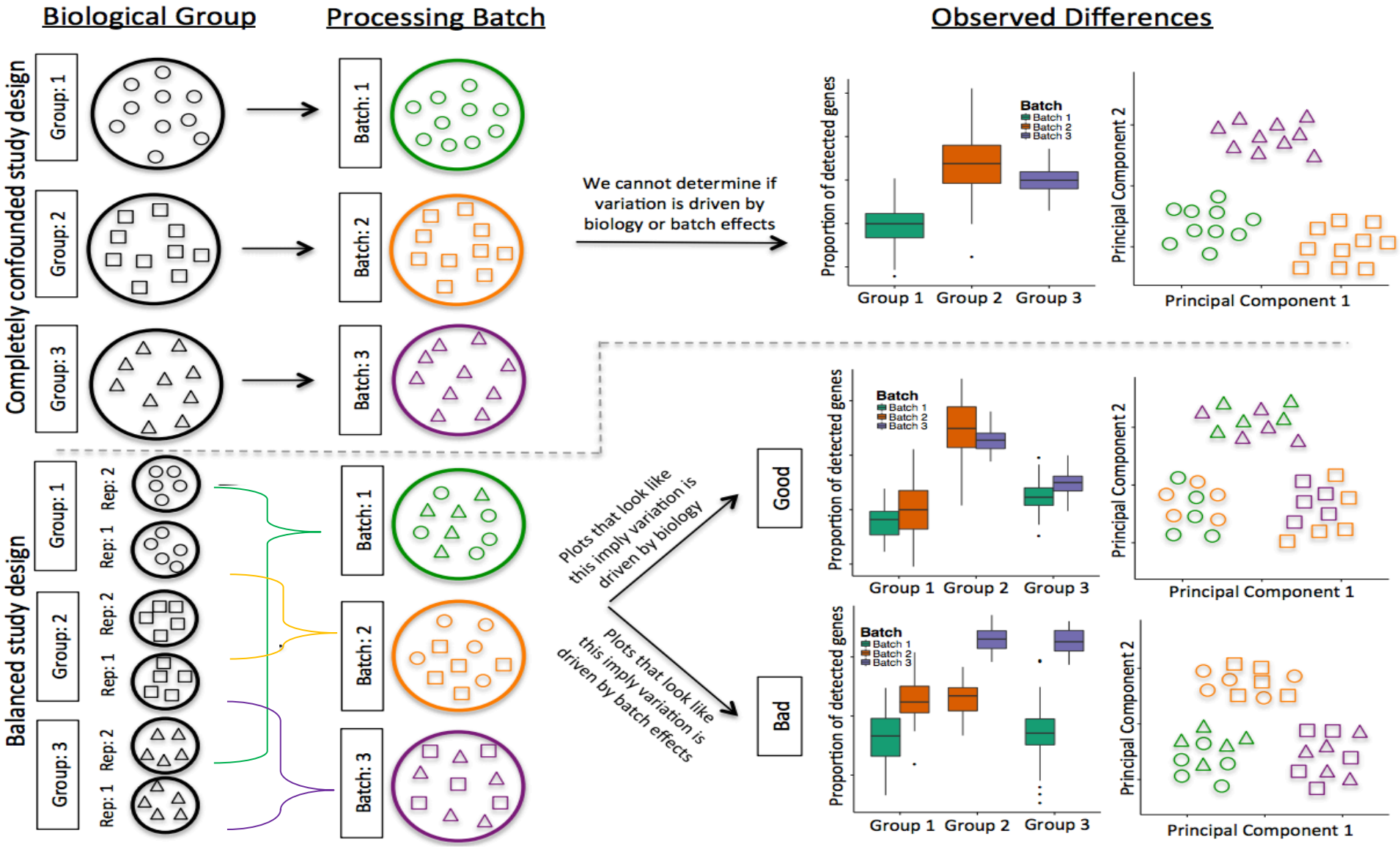
Batch Effect

- **Technical sources of variation that exist between samples**
  - Captures performed at different times (not different time points)
  - Generated in different laboratories/protocols
  - Different sequencing platforms, etc.

- **Obscures the actual underlying biology of the samples**
  - Can result in misleading conclusions

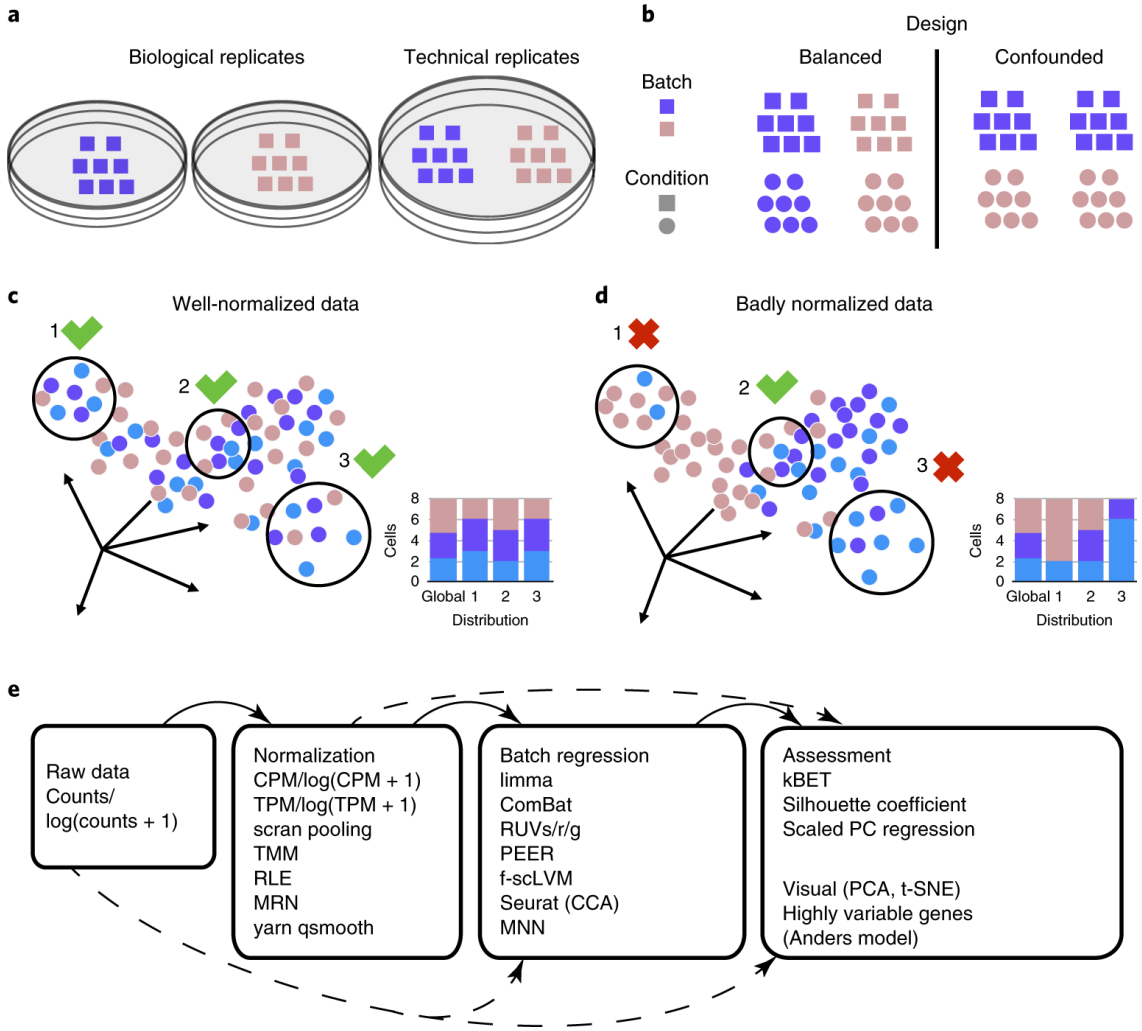
- **Different ways of experimental design can help avoid batch effect**

The Problem of Confounding Biological Variation and Batch Effects



Quantifying Batch Effect

- **Some metrics are available to estimate the amount of batch effect**
- **kBET measures local batch label distribution and compares to global**
  - If similar then does not reject null hypothesis that batches are well-mixed
- **Biological knowledge of data is still useful for accurately judging metrics**



Batch Effect Correction

- **Some batch effect correction methods work on samples with low complexity**
  - Samples all share the same cell types or states, i.e. technical replicates
- **Other methods have been designed to handle integrating varying sets of cell types across different samples**
- **Needs to be careful to avoid over-correcting**
  - Samples that do not contain overlapping cell types

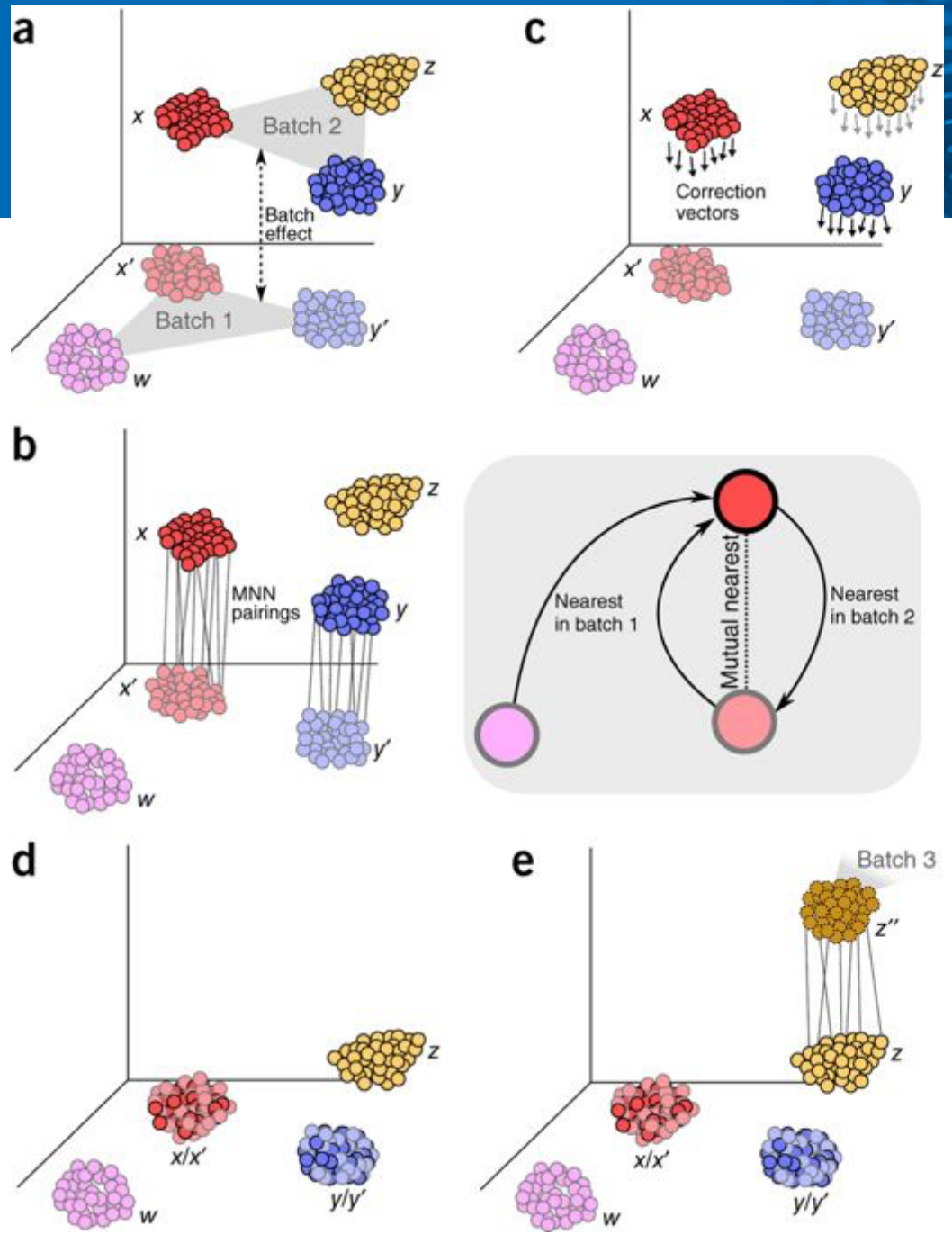
Batch Effect Tools

- **Many different tools available**
  - fastMNN
  - Scanorama
  - Harmony
  - Conos
  - CCA
  - BBKNN
  - ComBat
  - Limma
  - scMerge
  - scAlign
  - LIGER, etc…

fastMNN/mnnCorrect

- **fastMNN applies the mutual nearest neighbors algorithm on the principal components of a dataset**
  - mnnCorrect applied the algorithm on the gene expression data

- **Part of R packages scran and newly developed batchelor**

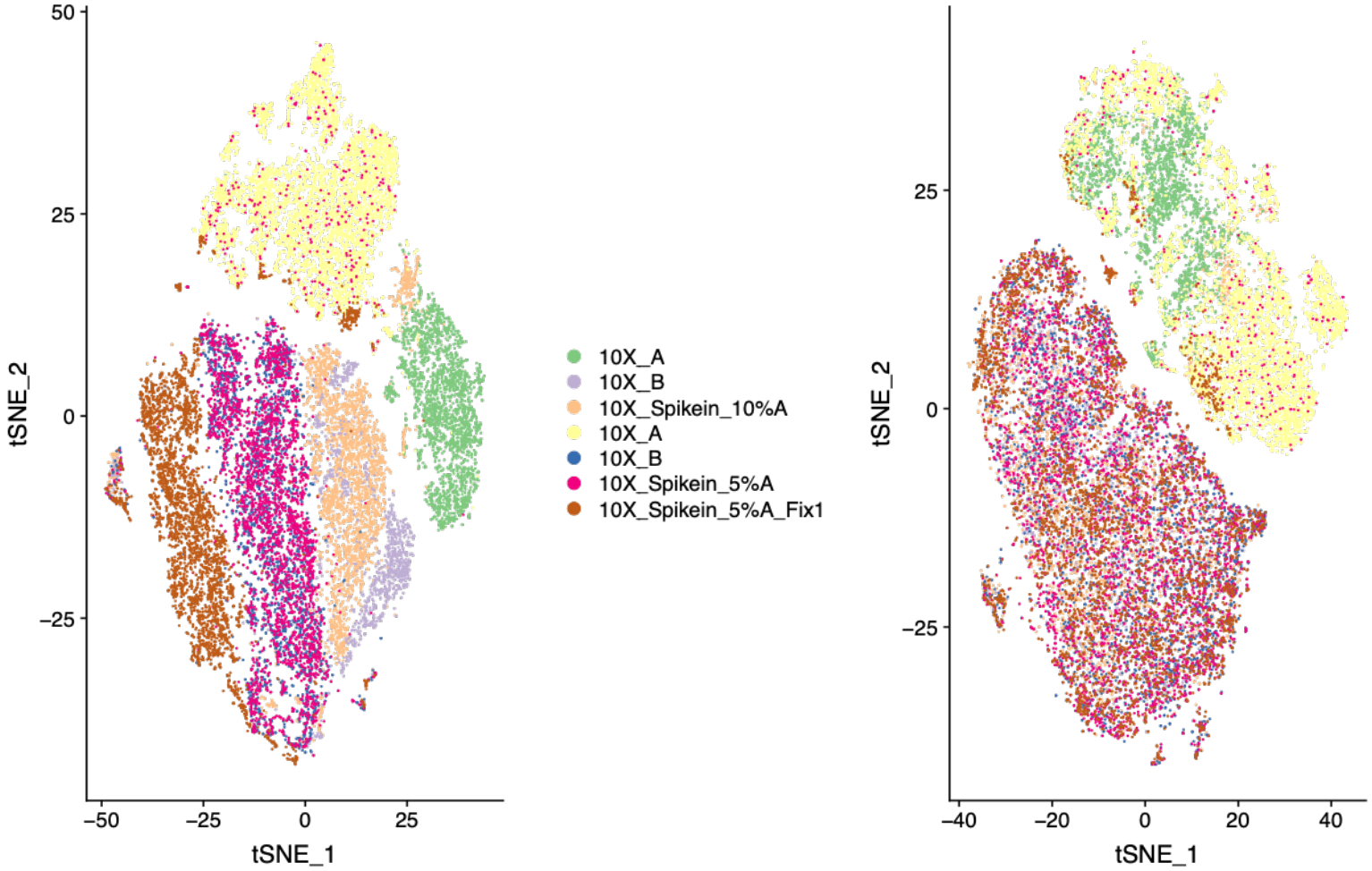


https://www.nature.com/articles/nbt.4091

Frederick National Laboratory for Cancer Research

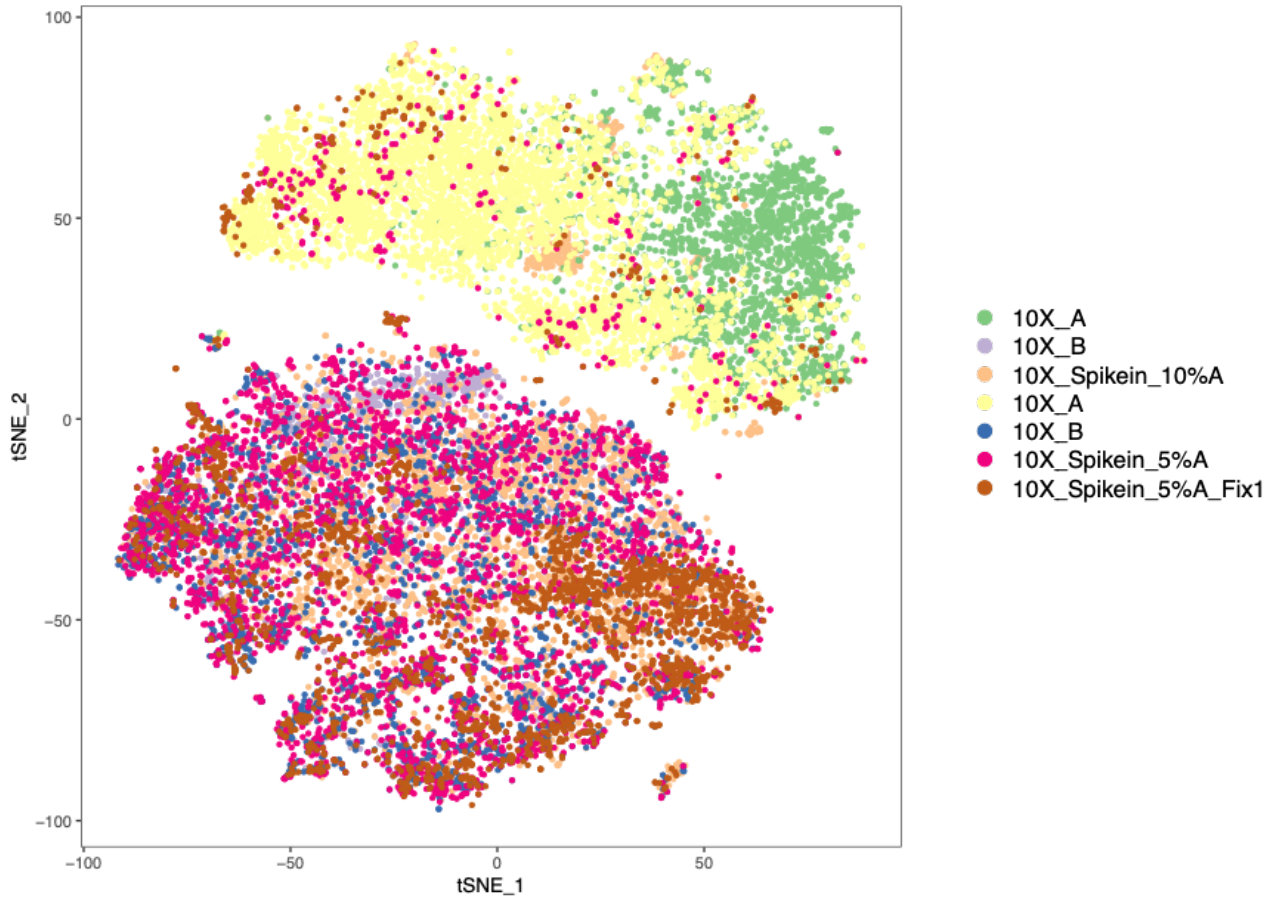
fastMNN/mnnCorrect

- **fastMNN applies the mutual nearest neighbors algorithm on the principal components of a dataset**
  - mnnCorrect applied the algorithm on the gene expression data
- **Part of R packages scran and newly developed batchelor**



Scanorama

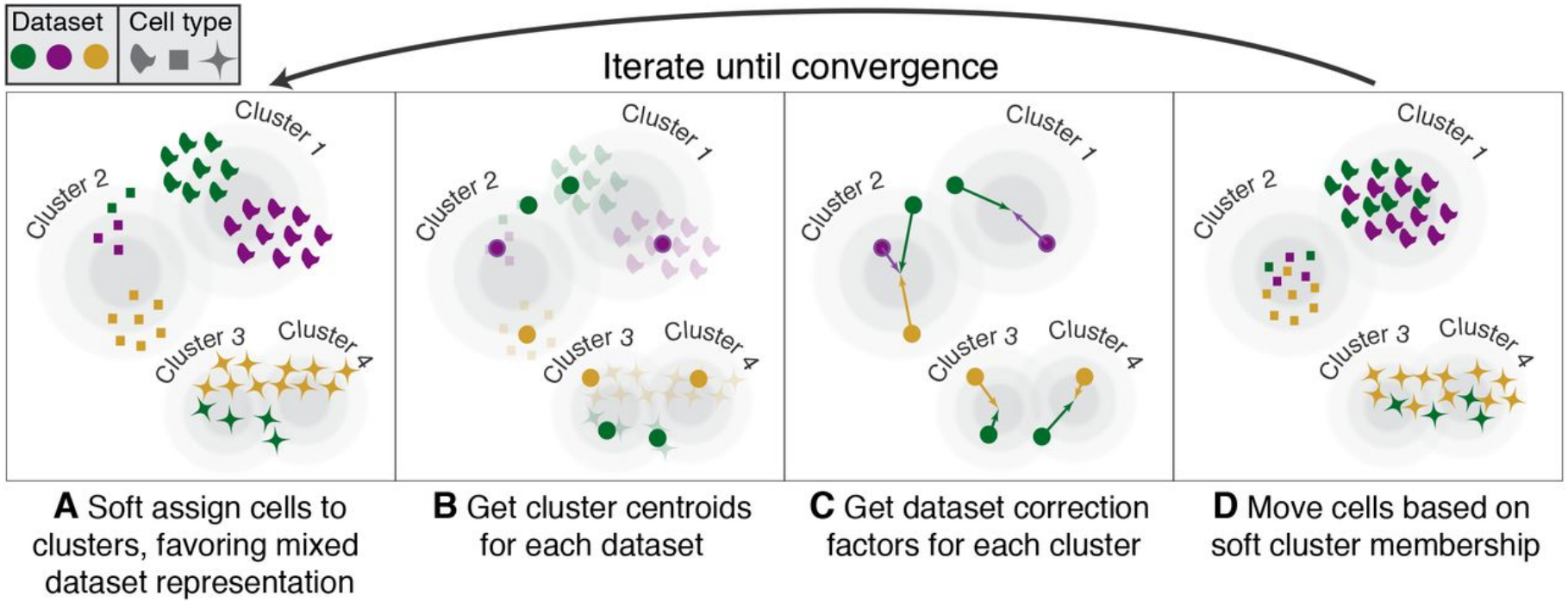
- **Uses a variation of the mutual nearest neighbor algorithm**
- **Less likely to overcorrect when no overlapping cell types exist in dataset**
- **Implemented in python**
  - Can interact with python scanpy object or be called from R



Harmony

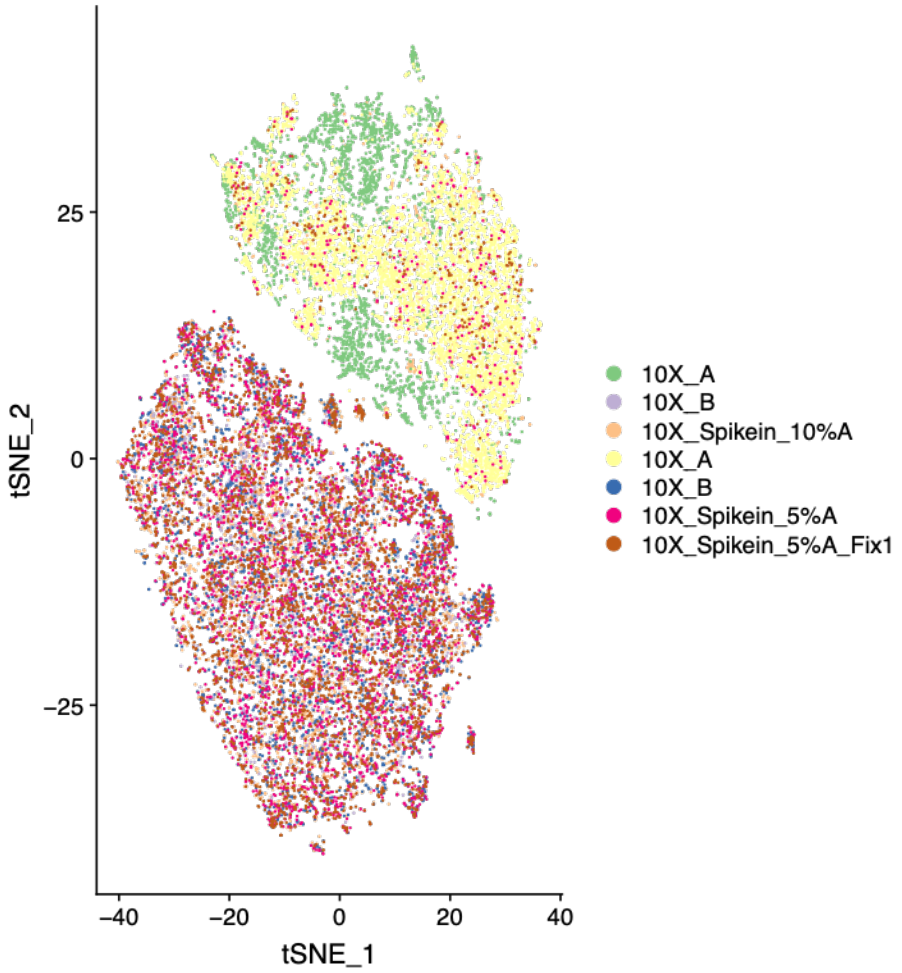
- **Iteratively learns a linear correction function**

- **Adjusts low dimensional cell embeddings**
  - Generates adjusted embeddings that can be used downstream



Harmony

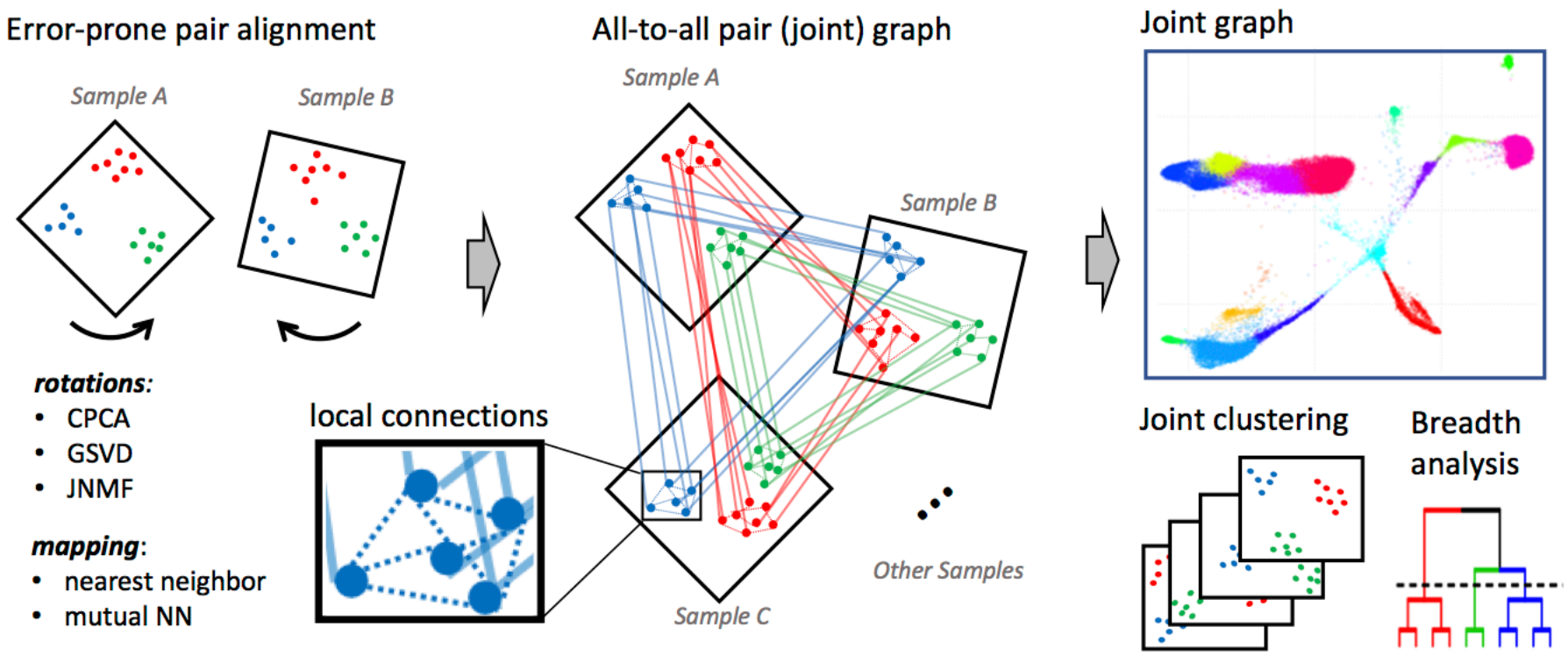
- **Iteratively learns a linear correction function**
- **Adjusts low dimensional cell embeddings**
  - Generates adjusted embeddings that can be used downstream
- **Can be used to integrate data across different technologies**
- **Implemented in R**



Conos

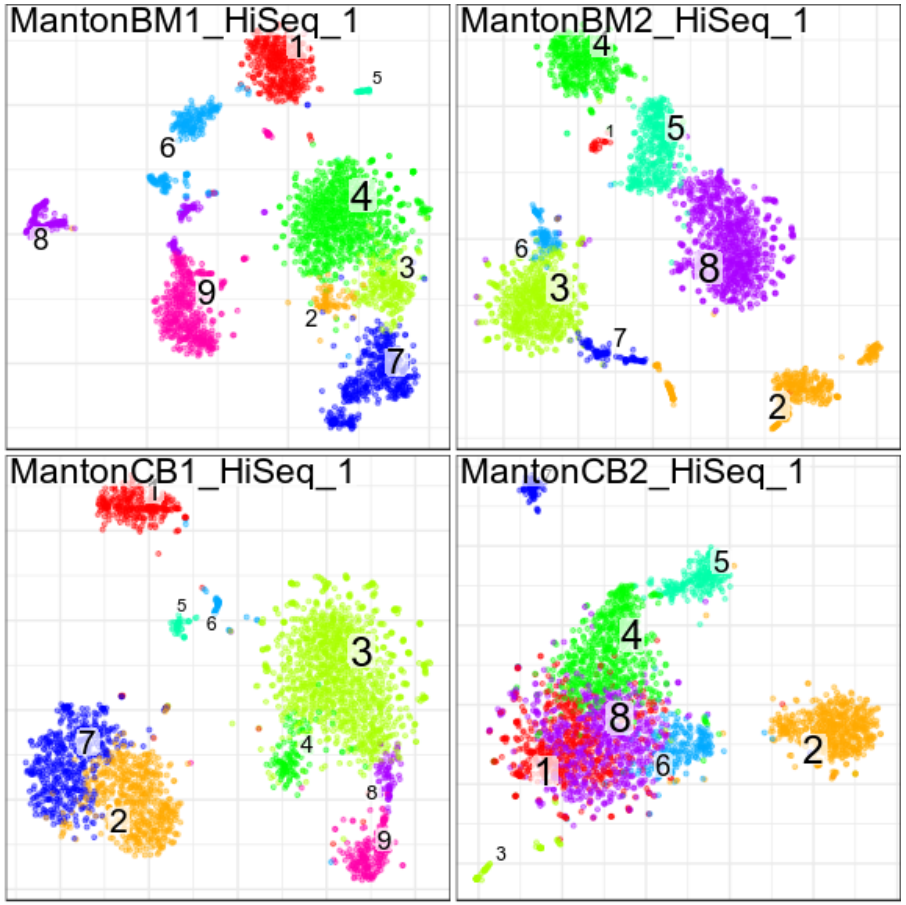
- **Works well with different types and large amounts of data**
  - Resolution increases as the number of samples increases
- **Performs pair-wise alignments between samples to identify subpopulations consistently mapped together**
- **These subpopulations are closer together in the joint graph and can be identified as similar cells**
- **Implemented in R and the results can be saved and uploaded to scanpy**

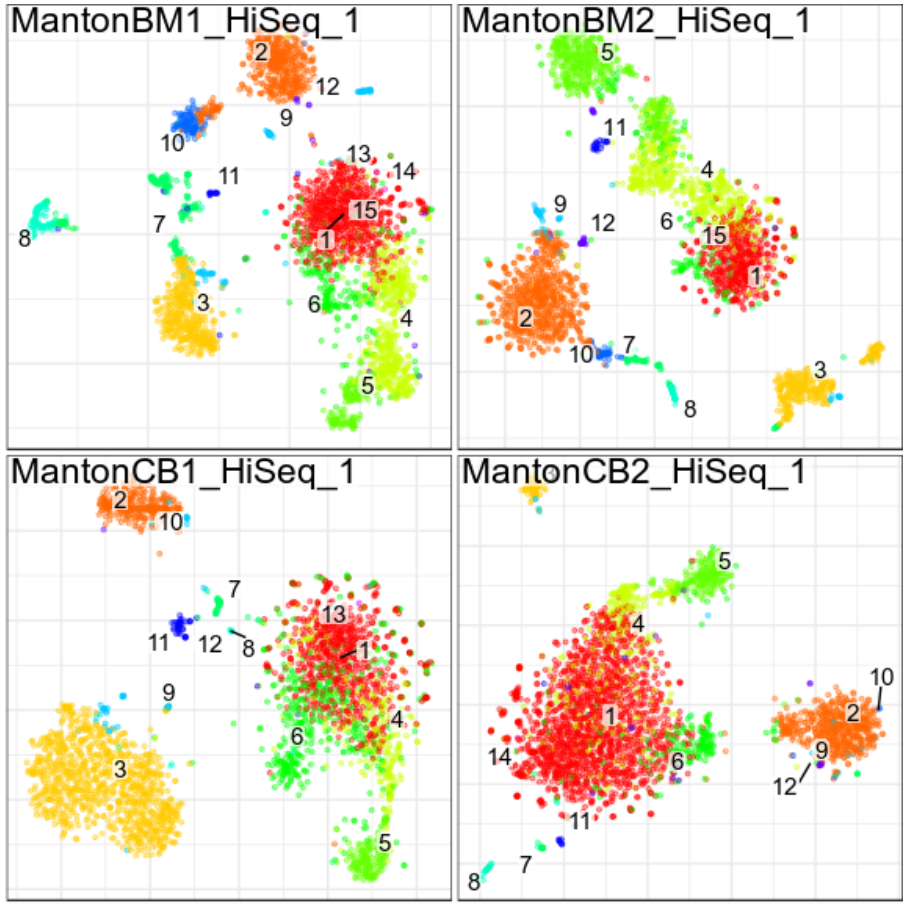
Conos



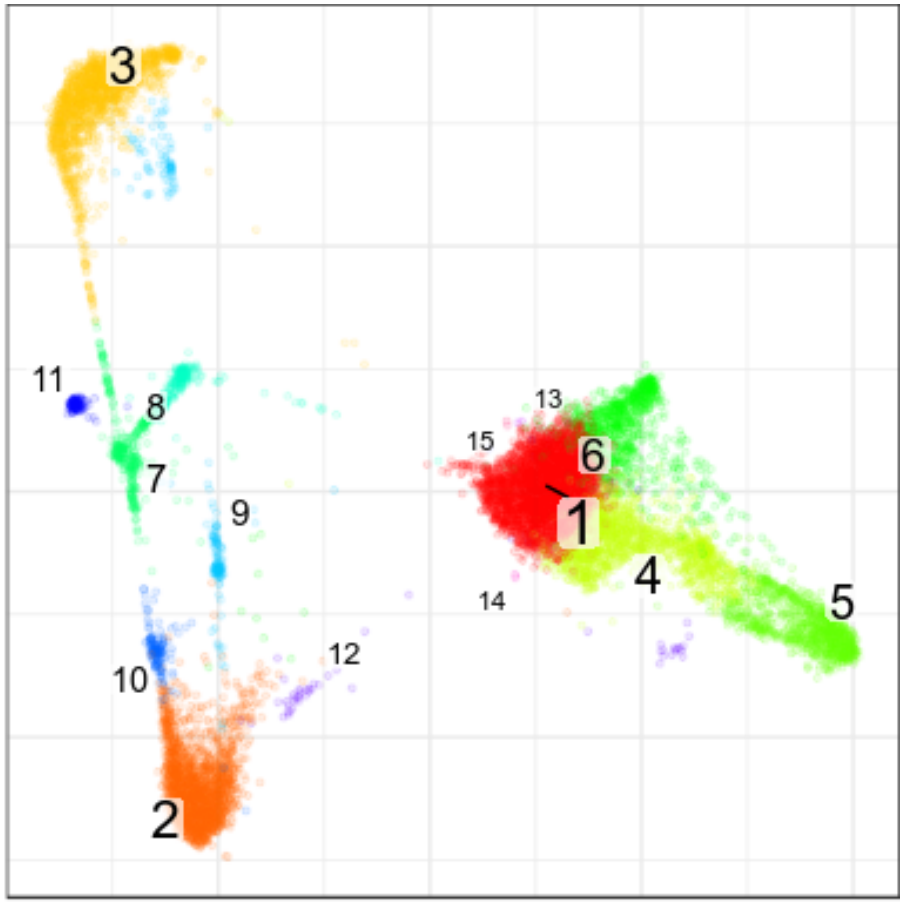
https://github.com/hms-dbmi/conos

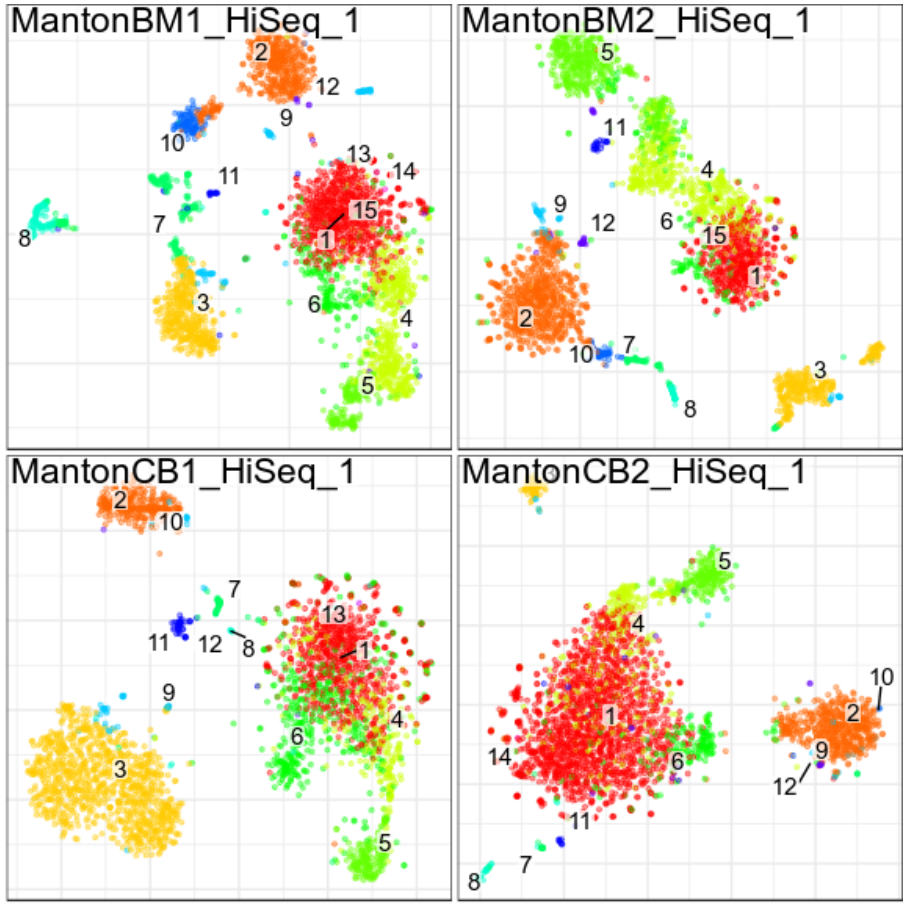
Conos


MantonBM1_HiSeq_1
MantonBM2_HiSeq_1
MantonCB1_HiSeq_1
MantonCB2_HiSeq_1


MantonBM1_HiSeq_1
MantonBM2_HiSeq_1
MantonCB1_HiSeq_1
MantonCB2_HiSeq_1

Conos


3
11
8
7
9
10
12
2
13
15
6
1
4
14
5


MantonBM1_HiSeq_1
MantonBM2_HiSeq_1
MantonCB1_HiSeq_1
MantonCB2_HiSeq_1

Summary

- **Many tools are available for multiple sample analysis**
- **Single sample analysis tools can also be applied to multi-sample analysis**
- **Additional normalization considerations exist when dealing with multiple samples**
- **Batch effects can exist due to technical effects or when combining multiple datasets**
  - May be able to design projects to avoid these technical effects
- **Need to be careful of overcorrecting when correcting for batch effects**