Student Computing Club: Dimension reduction algorithms for visualizing single-cell genomic data using R

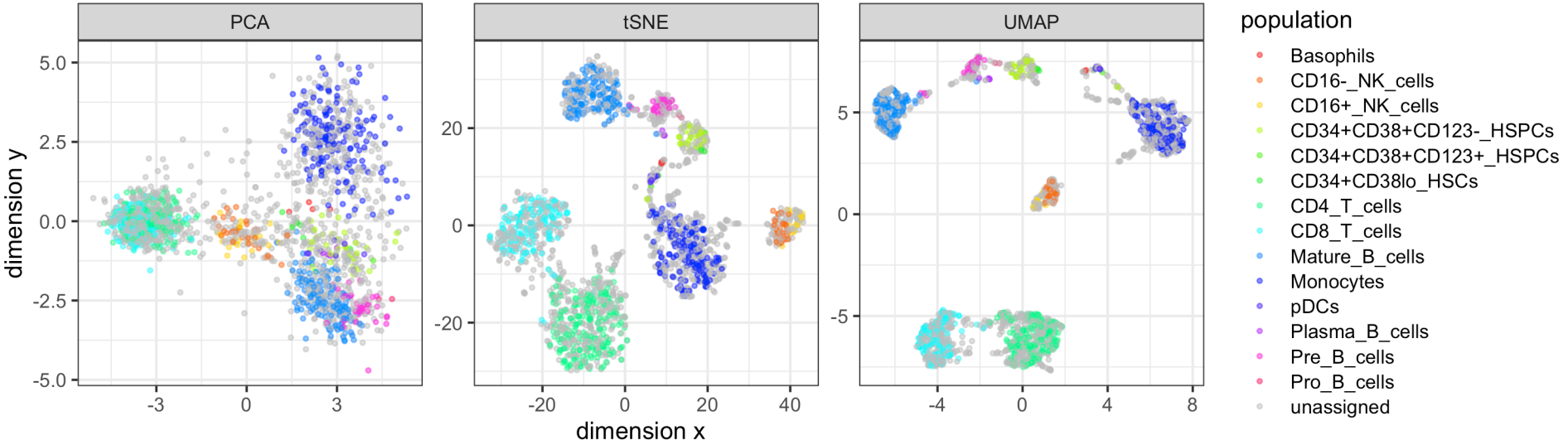
Lukas M. Weber

Hicks Lab
Department of Biostatistics
Bloomberg School of Public Health
Johns Hopkins University

29 October 2019

Motivating example

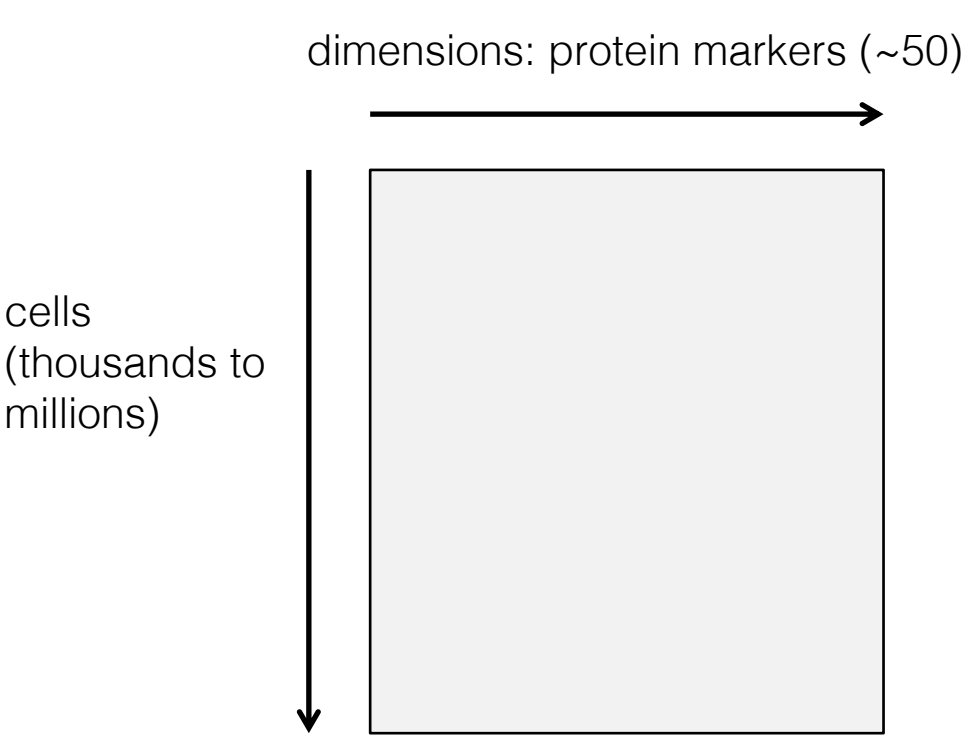
Dimension reduction



Single-cell data

Single-cell data

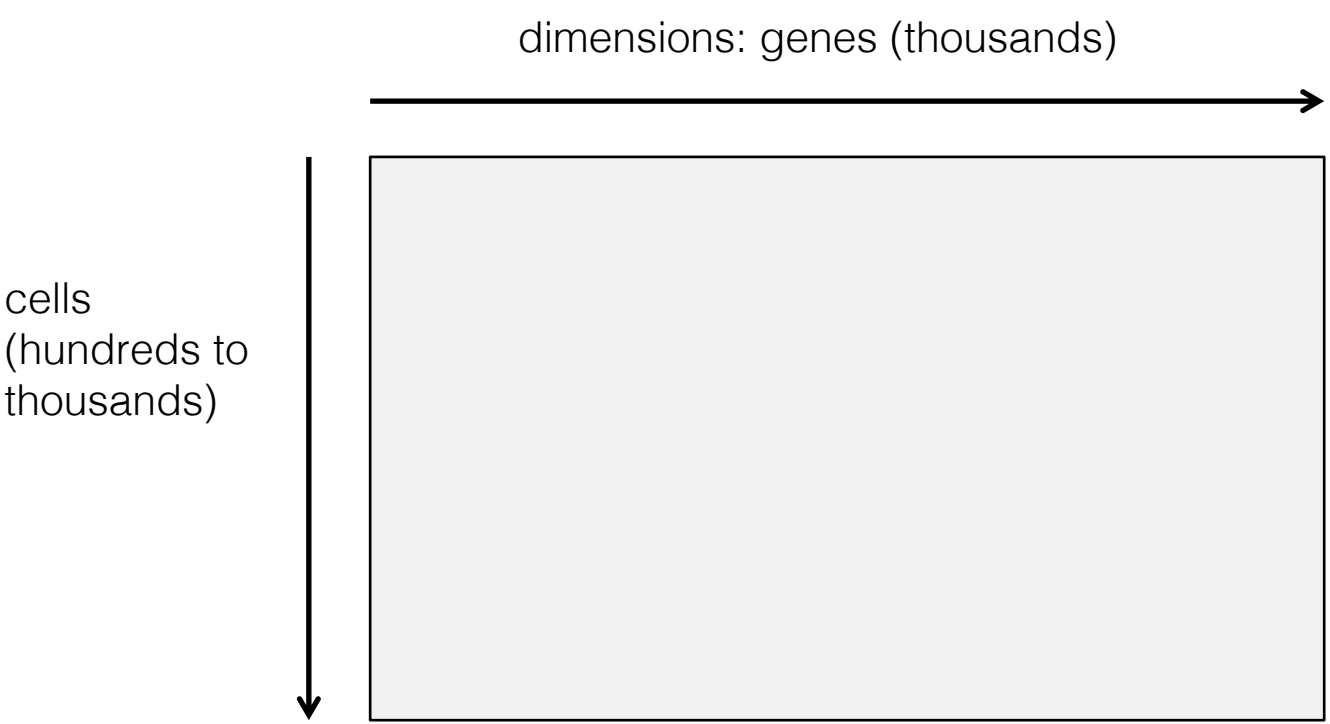
Example: Mass cytometry (CyTOF)



targeted set of proteins of interest; bind to known antibodies

Single-cell data

Example: Single-cell RNA sequencing (scRNA-seq)



untargeted:
measure all
genes

often $p >> n$

(note: usually
represented in
transpose
format)

Dimension reduction

Dimension reduction

Issue: too many dimensions!

How to represent visually?

→ exploratory data analysis; presentation of results (reveal or display patterns of interest, e.g. clusters, trajectories, differential sample features)

How to analyze computationally?

→ curse of dimensionality; computational scalability

Dimension reduction

Summarize data using a lower number of dimensions

Single-cell data: two main applications

- visualization (i.e. plot in 2 or 3 dimensions)
- data preprocessing (curse of dimensionality, remove noise, correlated features, computational scalability)

Dimension reduction algorithms

- select or calculate smaller number of dimensions (features) that capture the underlying patterns of interest in the dataset
- many approaches
- relevant patterns depend on scientific question

Examples

Dataset

Levine_32dim: mass cytometry (CyTOF) dataset from Levine et al. (2015)

- healthy human bone marrow mononuclear cells (BMMCs)
- 32 surface protein markers
- reference cell population (cluster) labels for 14 immune cell populations
- 265,627 cells (104,184 or 39% assigned)
- previously used to benchmark clustering algorithms in our publication (Weber and Robinson, 2016); available as formatted R/Bioconductor objects via HDCytoData package (Weber and Soneson, 2019)

Example: principal component analysis (PCA)

Intuitively: sequentially project data onto rotated orthogonal axes, where each axis captures maximal amount of remaining variance in data

Linear algorithm

– reduced dimensions (principal components) can be interpreted as combinations of original dimensions

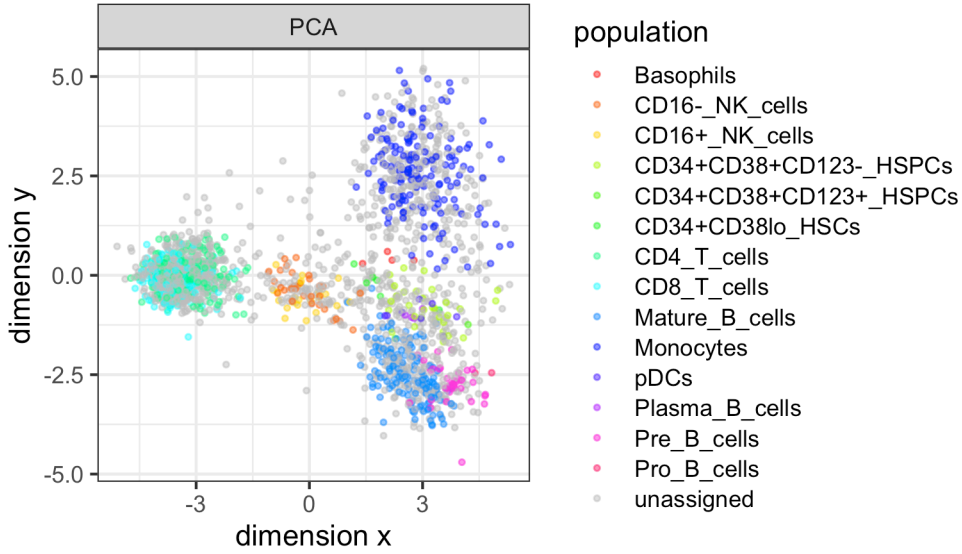
Single-cell data

– PCA commonly used for preprocessing, i.e. reduce dimensionality prior to downstream analysis (e.g. keep top 50 or 100 PCs in scRNA-seq data)

– Often does not work well for visualization, due to nonlinear data structure

Example: principal component analysis (PCA)

Levine_32dim dataset



Example: t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008; van der Maaten 2014)

Developed for visualizing datasets in machine learning; quickly adopted by single-cell biology community

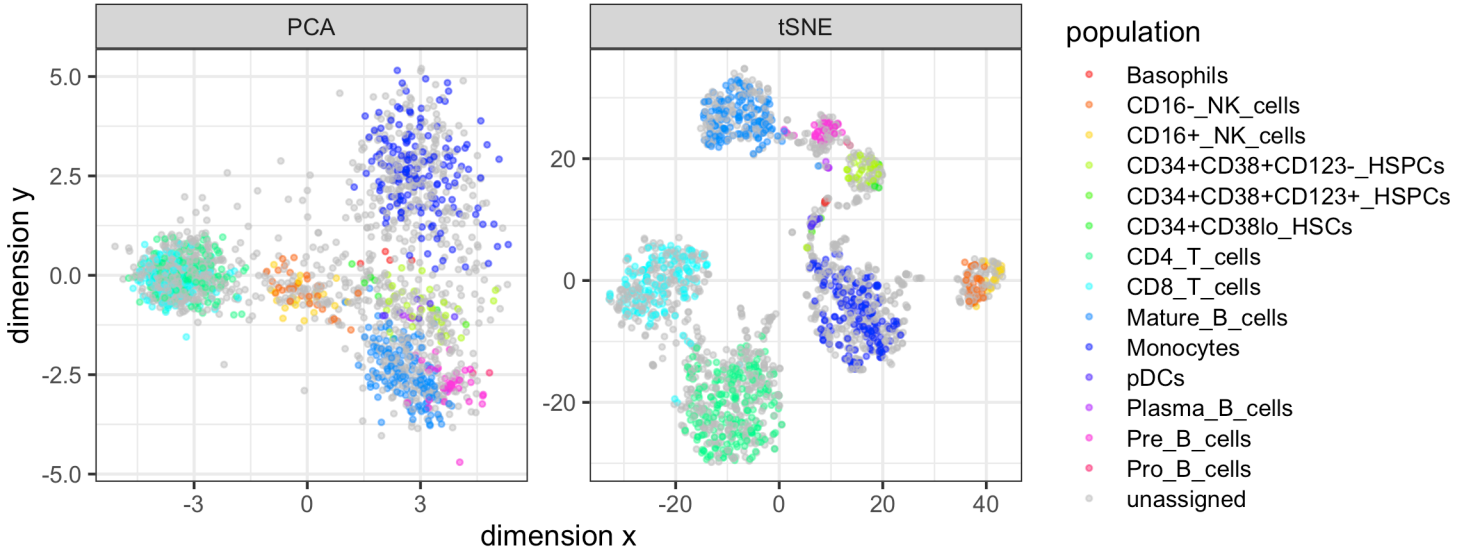
Nonlinear algorithm

Single-cell data

- Advantages: tends to clearly separate clusters (cell populations)
- Disadvantages: reduced dimensions difficult to interpret (especially global distances); can “force” cluster structure; computational scalability

Example: t-SNE

Levine_32dim dataset



Example: UMAP

Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018)

Widely adopted for single-cell data within the last year

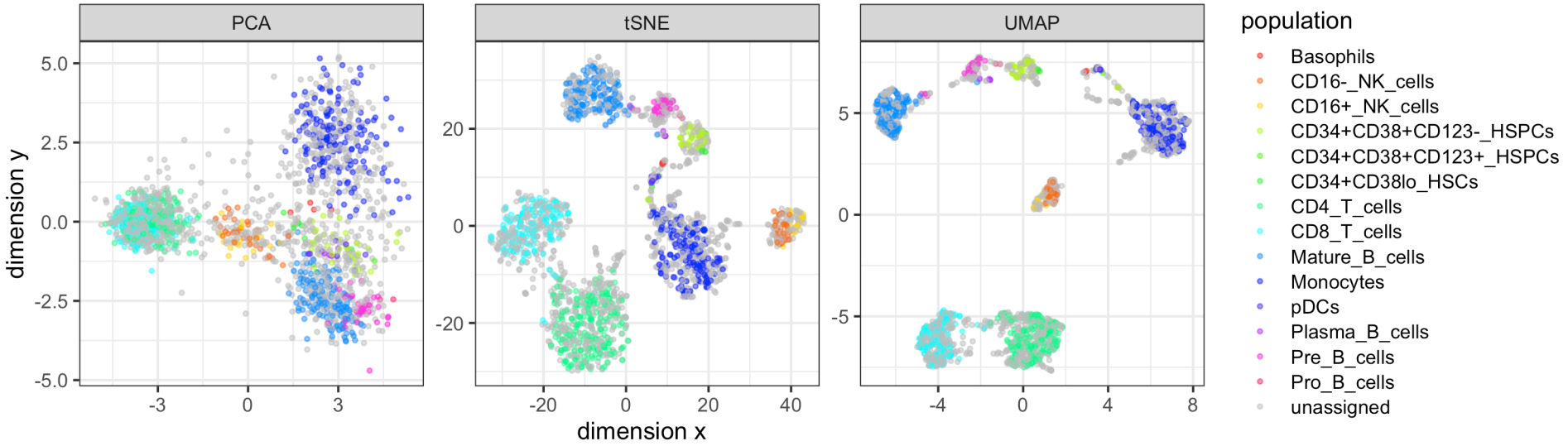
Nonlinear algorithm

Single-cell data

– Advantages: tends to separate clusters as well as t-SNE but preserves global distances more accurately; computationally efficient

Example: UMAP

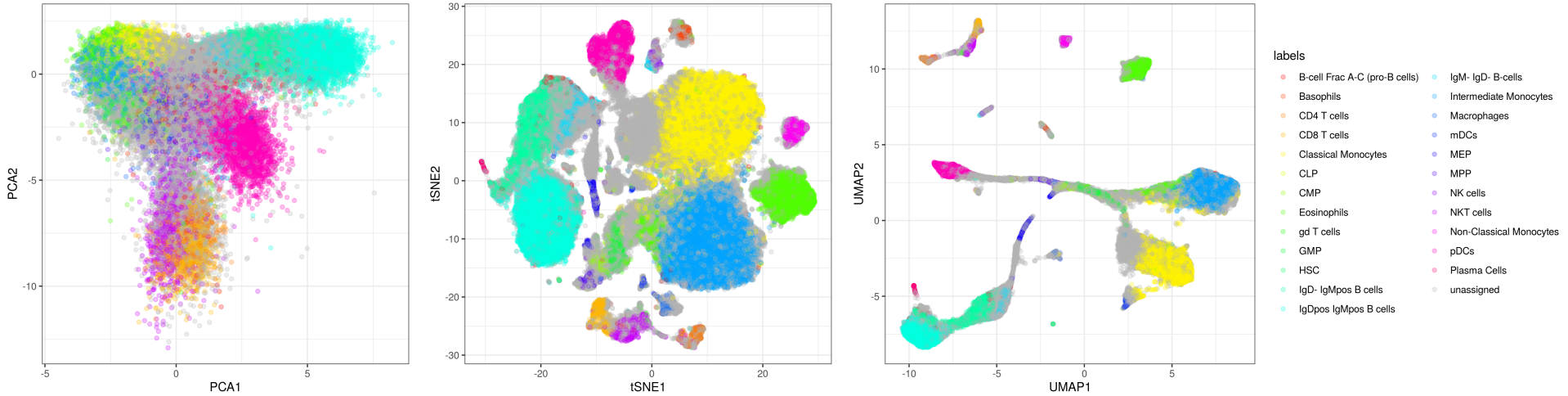
Levine_32dim dataset



More examples

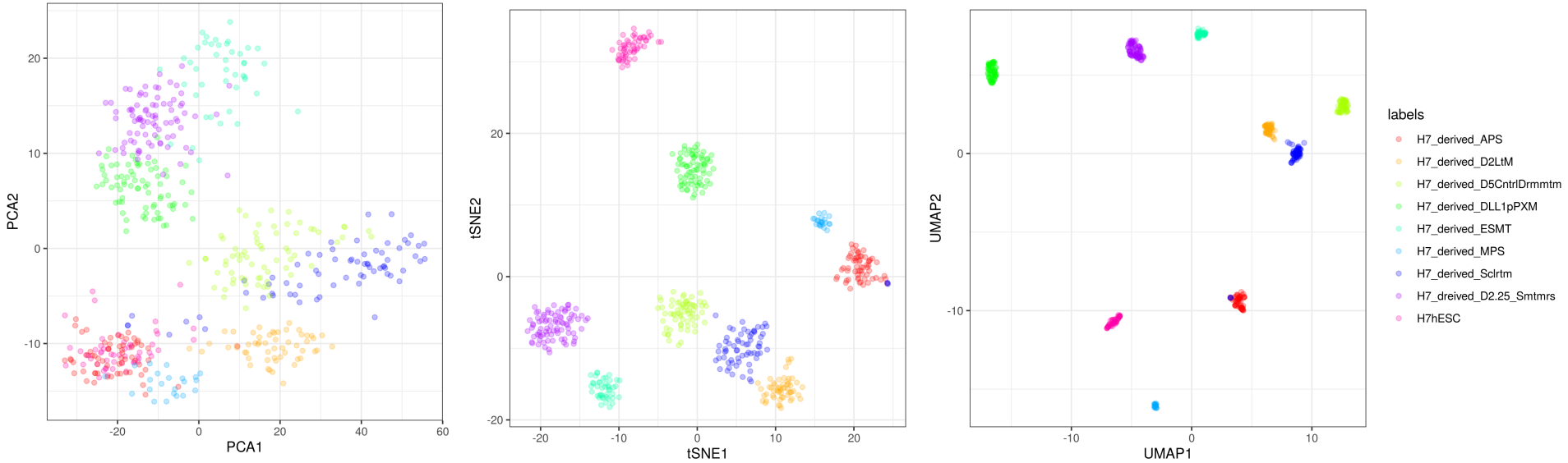
Results

Reduced dimension plots for each method/dataset: Samusik_01 dataset (CyTOF)



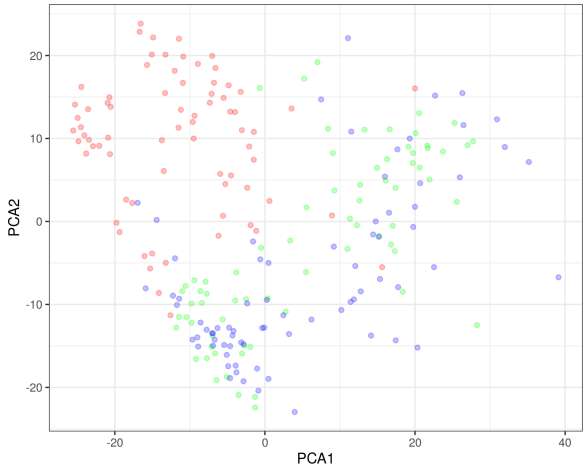
Results

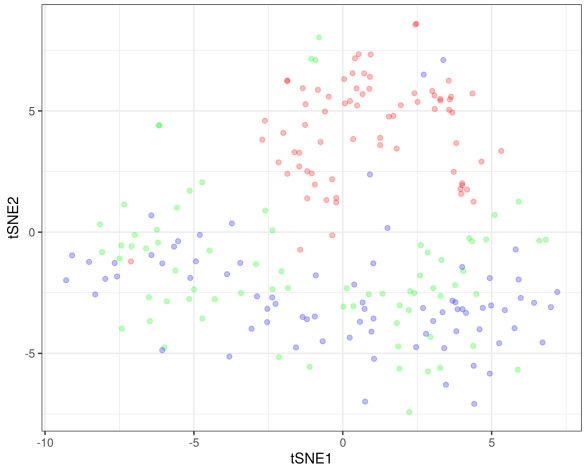
Reduced dimension plots for each method/dataset: Koh dataset (scRNA-seq)

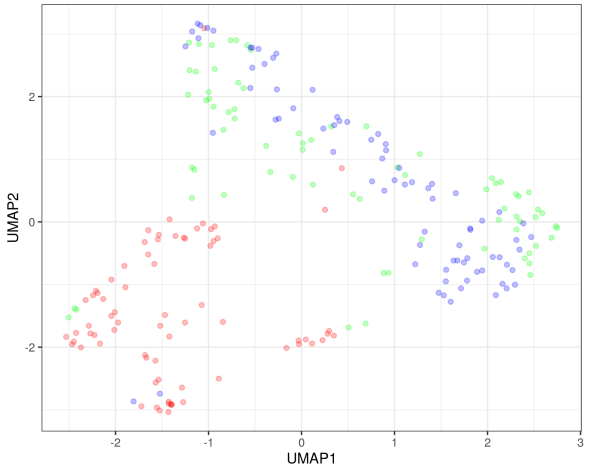


Results

Reduced dimension plots for each method/dataset: Trapnell dataset (scRNA-seq)







Interactive demo

Interactive demo

See RStudio

Thank you!

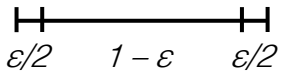
Additional slides

Curse of dimensionality

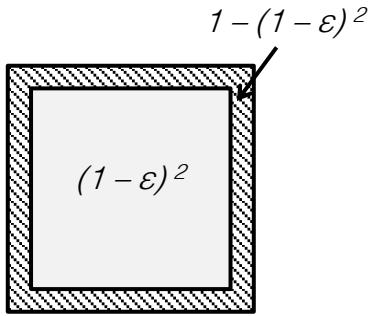
Standard (e.g. Euclidean) distances become largely meaningless in very high-dimensional spaces

→ all points reside in thin “shell” of high-dimensional sphere or cube, with ~zero interior volume; all points are approximately the same distance apart

1-dimensional



2-dimensional



…

ε = 0.01

p	*(1 – ε) p*	*1 – (1 – ε) p*
1	0.99	0.01
2	0.9801	0.0199
3	0.9703	0.0297
…		
10	0.9044	0.0956
100	0.3660	0.6340
1000	4.32e-05	~1.0
10000	2.25e-44	~1.0