Single Cell Sequencing Analysis

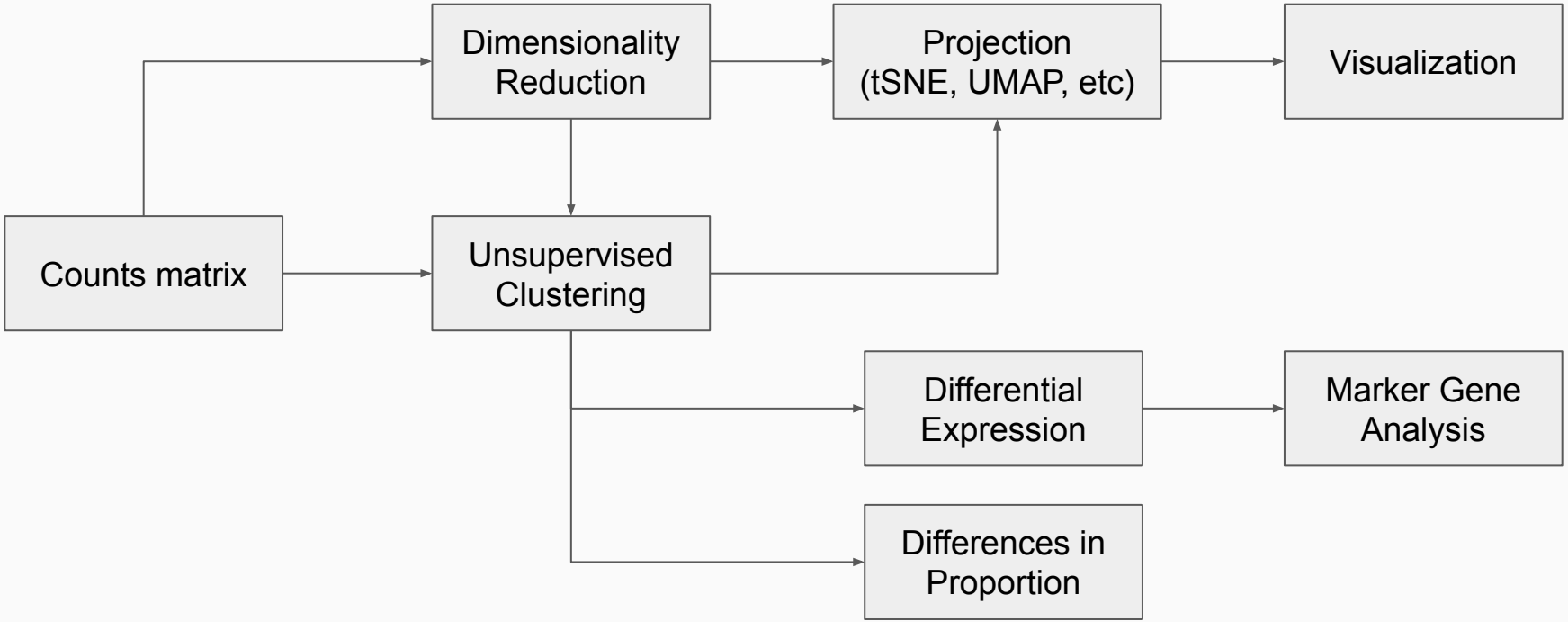
Part 2

The Counts Matrix

- Counts matrix contains either:
  - Read counts or
  - UMI counts if used
- Each cell has:
  - Total number of counts (col. sum, “library size”)
  - Number of non-zero genes
- Each gene has:
  - # of non-zero cells
  - Non-zero mean/variance
- Matrix is *sparse*: many zeros
- Zeros may be:
  - Cell lacks gene
  - A “drop-out”: gene present but was missed by qPCR

	cell1	cell2	cell3	cell4	cell5	cell6	…	cellM
gene1	93	25	0	0	3335	0		82
gene2	5	2	0	3	1252	0		12
gene3	0	0	0	0	0	0		0
gene4	98	21	1	1	5318	0		75
gene5	0	0	513	0	0	325		135
gene6	0	0	113	0	1	497		255
gene7	3	0	0	0	6	0		0
…								
geneN	68	52	0	2	4313			63

Typical Analysis Paths



Unsupervised Clustering

- Wish to identify subpopulations of cells using similarity of transcript abundance
- Clustering methods discover patterns in the data
- *A priori* no knowledge of number of clusters
- Many available methods and metrics:
  - PCA/Spectral analysis
  - Hierarchical or Ward agglomerative clustering
  - K-nearest neighbor clustering
  - Jaccard similarity
  - Louvain community detection
  - K-means
  - Graph based clustering
  - Many, many more...

Analysis: Differential Expression

- **Goal:** identify gene expression differences between cell types (i.e. clusters)
- Simple solution: DESeq2 of each cluster vs all the others
- Significant genes drove the clustering
- Examine for marker genes

Marker Gene Analysis

- **Goal:** Label each cluster to known cell type
- Biological domain experts know which genes are expressed by each cell type
- Some clusters may be difficult to label (novel cell types?)
- **NB:** cells of one cell type may cluster by *state*, e.g. cell cycle phase G1

Dimensionality Reduction

- Counts matrix may have many dimensions
  - 1000s of genes x 1000s of cells
- Reduces # dimensions while preserving variance
- May be necessary for large datasets (>1M cells) to make downstream analysis algorithms tractable
- Many methods available, including:
  - PCA
  - Multidimensional Scaling (MDS)
  - Downsampling

Projection + Visualization

- **Goal:** accurately visualize cell clusters in two dimensions
- Projection: embed samples of high dimensional space into lower dimensional space, retaining structure of data
  - e.g. map from (1000s genes x 1000s cells) to 2 (i.e. <x,y>)
- Ideally preserve both local and global structure
- **NB:** this is challenging to do efficiently+accurately!
- Available methods:
  - t-SNE: t-statistic Stochastic Neighbor Embedding
  - UMAP: Uniform Manifold Approx. and Mapping
  - PCA

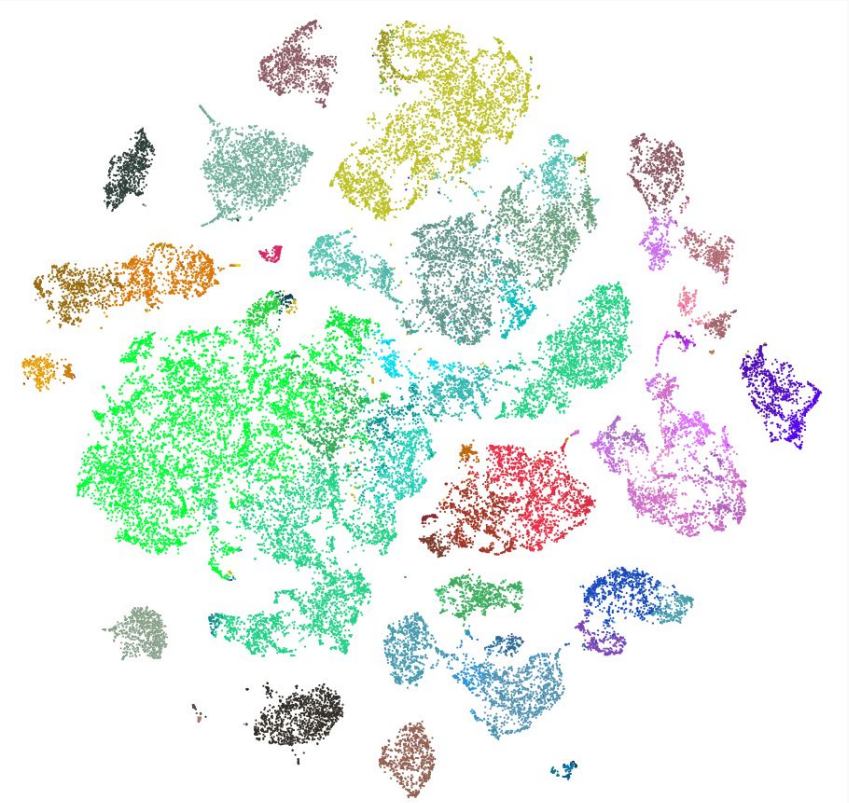
t-SNE, UMAP, et cetera

Conceptual idea:

1. Compute distance between all pairs of cells in high dimensional (i.e. all genes) space
2. Find a function that maps samples into 2D space s.t. cells that are close in original space are also close in embedded space

Can compute locally accurate mapping quickly:

- Cells near each other are similar
- **BUT** cells far away from each other are *not necessarily proportionately far* from each other!
- **Local** structure is preserved, **global** structure is not!

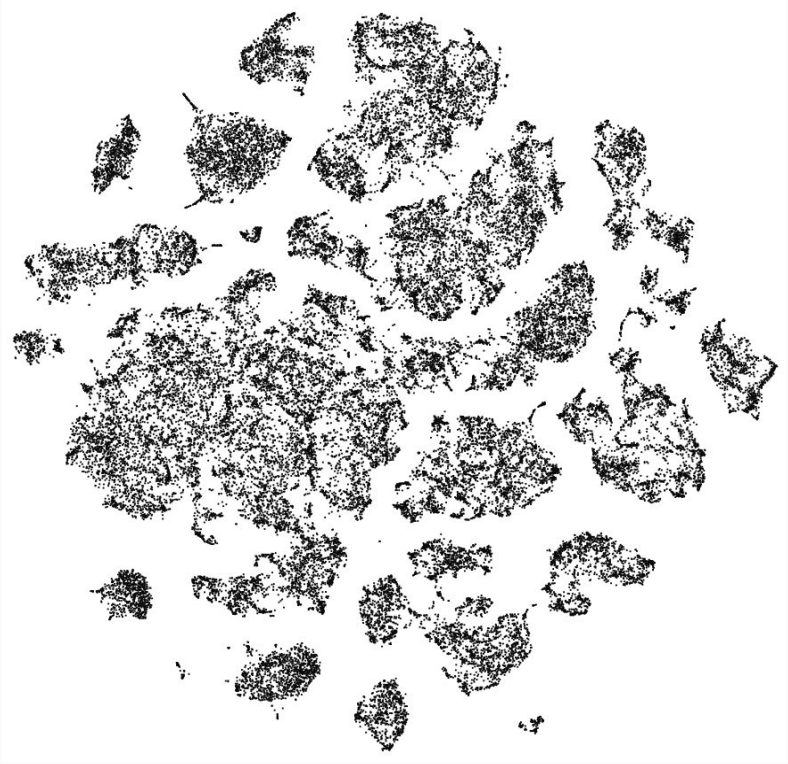


t-SNE projections from human cortex single nuclear RNA-Seq
https://portal.brain-map.org/atlases-and-data/rnaseq

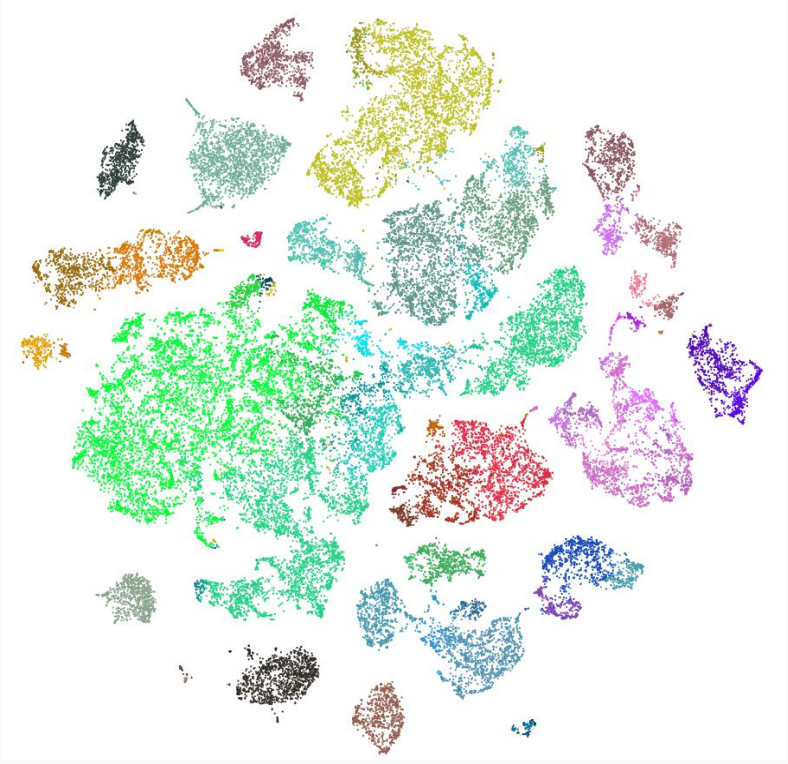
Visualizing Projections

- Wish to use clustering to interpret data
- Projection methods produce an embedding
- **Strategy:** map cell metadata onto embedding and visualize
- Following slides:
  - Single nuclear RNA-Seq from human cortex
  - 6 regions sampled
  - Source: Brain-Map Cell Type Database
    https://portal.brain-map.org/atlases-and-data/rnaseq

Visualizing Projections

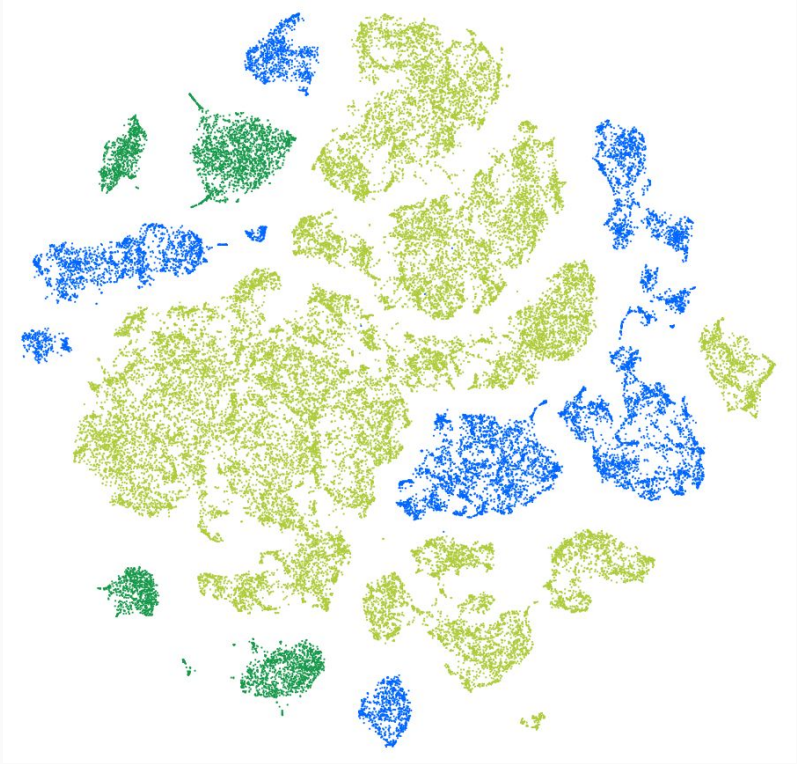


No color



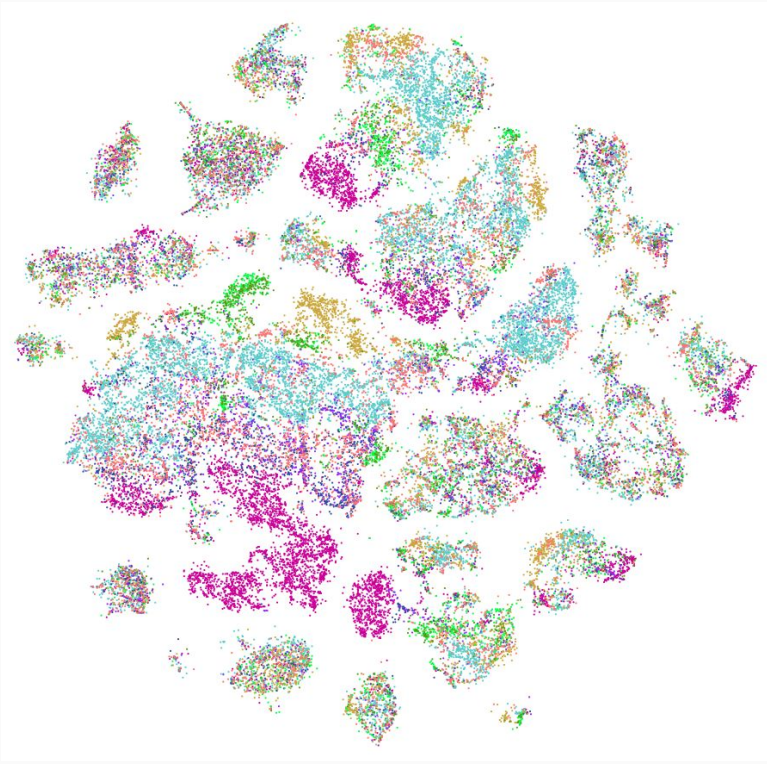
Colored by unsupervised clustering

Visualizing Projections



Colored by cell class, labeled by known marker genes

Non-neuronal (glia)
Glutamatergic neurons
GABAergic neurons



Colored by brain region

The Many Subjects We Didn't Cover

- Supervised clustering of cells by known markers/cell states (e.g. cell cycle)
- Comparing different samples
- Pseudotime Analysis
  - Inferring cellular development/change over time
- Imputation
  - Infer expression values for "dropout" genes
- Many more...

Current Software

- Bioconductor
  - Seurat - One of the first analysis software packages
  - SingleCellExperiment - official Bioconductor class
  - scater - Single Cell Analysis Toolkit
- scanpy - single cell analysis in python
- Many others now
- Millions of others soon...