Break!

Steps in "standard" Seurat analysis

1. Cell/gene QC filtering
2. Normalization
3. Dimension reduction (PCA/UMAP/Cluster calling)
4. Marker gene detection
5. Cell type annotation (marker genes and using reference data set)

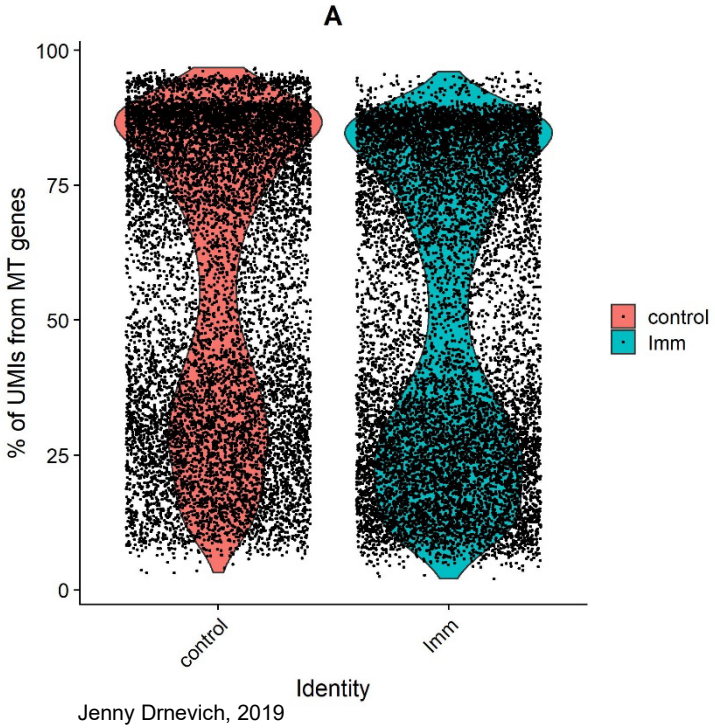
Cell/gene QC filtering

- Remove cells with too few genes/UMIs
- Recall ~7% expected "doublets" (>1 cell)
  - Remove cells with too many genes/UMIs (easiest)
  - Computational double detection (harder, worth ROI?)
- Percentage of UMIs that map to mitochondrial genes; high proportions assumed to indicate dead or dying cells
- Also see OSCA book's extended QC discussion
- Genes only detected in <20 cells likely not interesting but can make up a substantial proportion of genes. Remove to save computational effort and reduce FDR correction

What thresholds to use?

Must decide specifically for each data set or even each sample!

- If relatively uniform distribution of QC metrics across cells, can pick by "eye" or use +/-3 Median Absolute Deviations
- If different samples have very different mean UMI/genes, do per sample
- If bi- or tri-modal distributions, cannot use +/- 3 MAD



Jenny Drnevich, 2019

Normalization

- Within a sample, differences in total UMIs per cell somewhat due to uneven sequencing effort, so need to normalize
- First method similar to bulk RNA-Seq, scaled to total UMI counts per cell then log-transformed
- Seurat's SCTransform normalization better accounts for sequencing depth differences between cells.

How to quantify similarity/differences in expression of thousands of genes across thousands of cells?

- Using all genes computationally expensive and many genes have correlated expression patterns, so using a **reduced data set** give just as good results
- The first step is to pick the top 2000-3000 **most variable genes** across all* cells as these have the most information on cell-to-cell differences.
- Then they are put through **multiple dimension reduction algorithms** to arrive at a final* representation of the cells' relatedness

* You may want to re-process subsets of cells as the most discriminatory genes for a subset is likely very different than across all major cell types, leading to a different representation of relationships among the subset

1st dimension reduction: PCA

- Principal components analysis is run on the top 3000 variable genes to linearly reduce down to dozens of PCs
- Need to select what number of PCs capture most of the variation in the top genes:
  - Too few can fail to capture important variation
  - Way too many is computationally inefficient
  - Usually 30-50 is sufficient for most 10x data sets

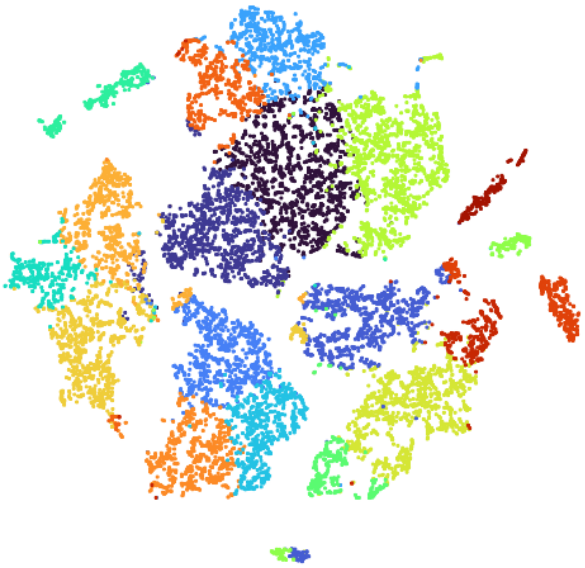
2nd dimension reduction option: tSNE

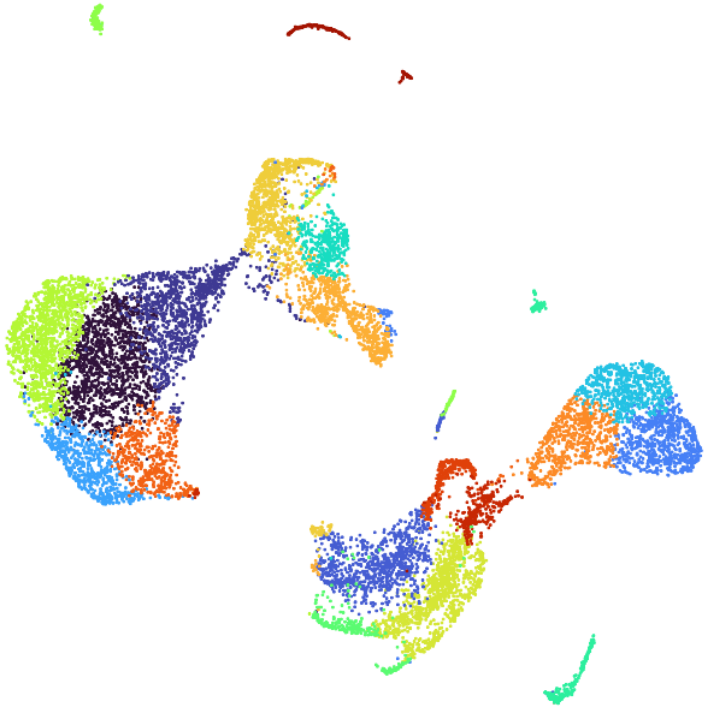
- t-Distributed Stochastic Neighbor Embedding (van der Maaten & Hinton, 2008)
- Non-linear dimension reduction algorithm to visualize high-dimensional data in just 2 dimensions (XY plot)
- StatQuest explanatory video
- First widely used in single cell analysis and still first one shown in Cell/Space Ranger outputs
- Does not scale well to large data sets due to computation time and memory requirements
- Does not preserve "global" data structure (i.e., clusters of cells placed equidistant from other clusters

2nd dimension reduction option: UMAP

- Uniform Manifold Approximation and Projection (McInnes & Healy, 2018)
- Also a non-linear dimension reduction algorithm to visualize high-dimensional data in just 2 dimensions (XY plot)
- UMAP scales better than t-SNE and uses less memory
- Clusters of cells not placed equidistant from other clusters, but possibly still does not preserve true global data structure and is just "art".
- Good, interactive comparison of UMAP and t-SNE

t-SNE vs. UMAP





10k Mouse E18 Combined Cortex, Hippocampus and Subventricular Zone Cells, Chromium GEM-X Single Cell 3' Gene Expression Dataset by Cell Ranger 8.0.0, 10x Genomics, Inc, (2024, Apr 15). Used with permission of 10x Genomics, Inc.

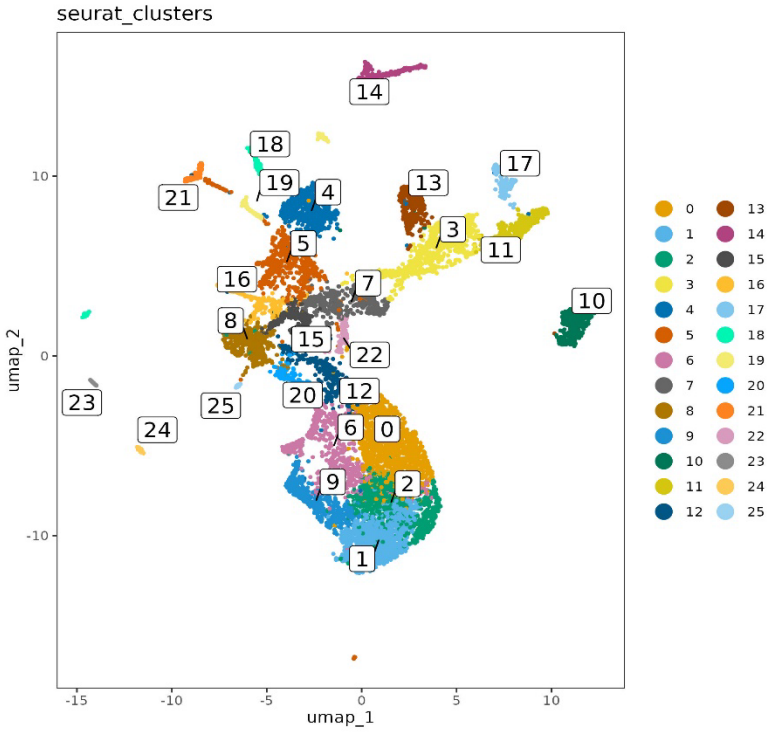
Warning about 2D representations!

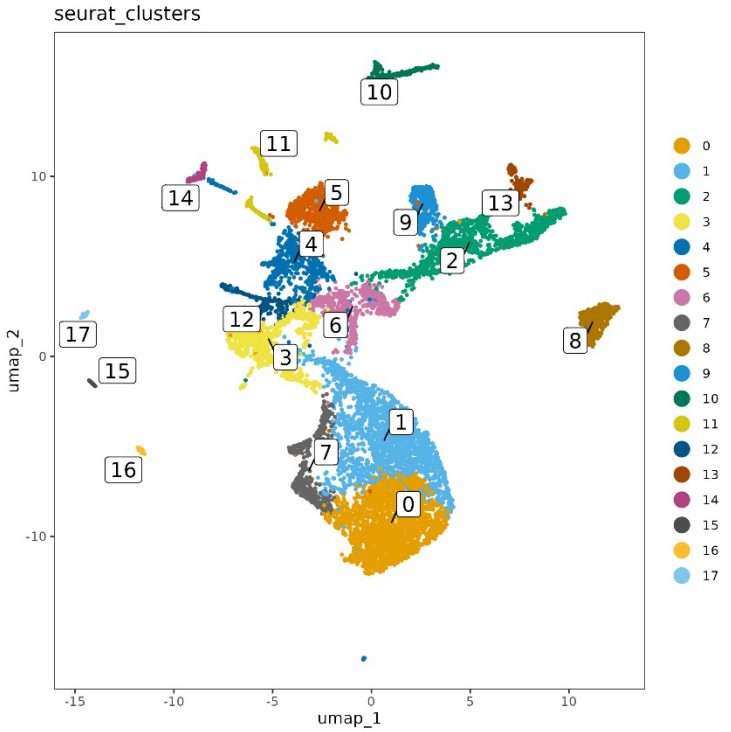
- There is no single "correct" 2D representation of your single cell data.
- Both algorithms are non-deterministic, starting from a random number. A different random number will lead to a slightly different XY graph
- Within t-SNE and UMAP, there are many parameters that could be tweaked and lead to slightly or greatly different XY graphs
- Remember how many reductions have been done (subset genes, PCA then t-SNE/UMAP) to get to this point.
- Consider them only to be "useful" to recognize patterns in your data set.

Cluster calling

- A separate algorithm is run to partition cells into "clusters"; most often shared nearest neighbor graphs
- This is run on the ~40 PC values, not the 2 t-SNE/UMAP values so **cells in a cluster do not co-locate perfectly**!
- Among other parameters, changing the resolution can lead to more or fewer clusters.
- How to determine "correct" number of clusters? At a broad clustering level, I recommend match to major UMAP/t-SNE groupings.

Too many clusters vs. "about right"



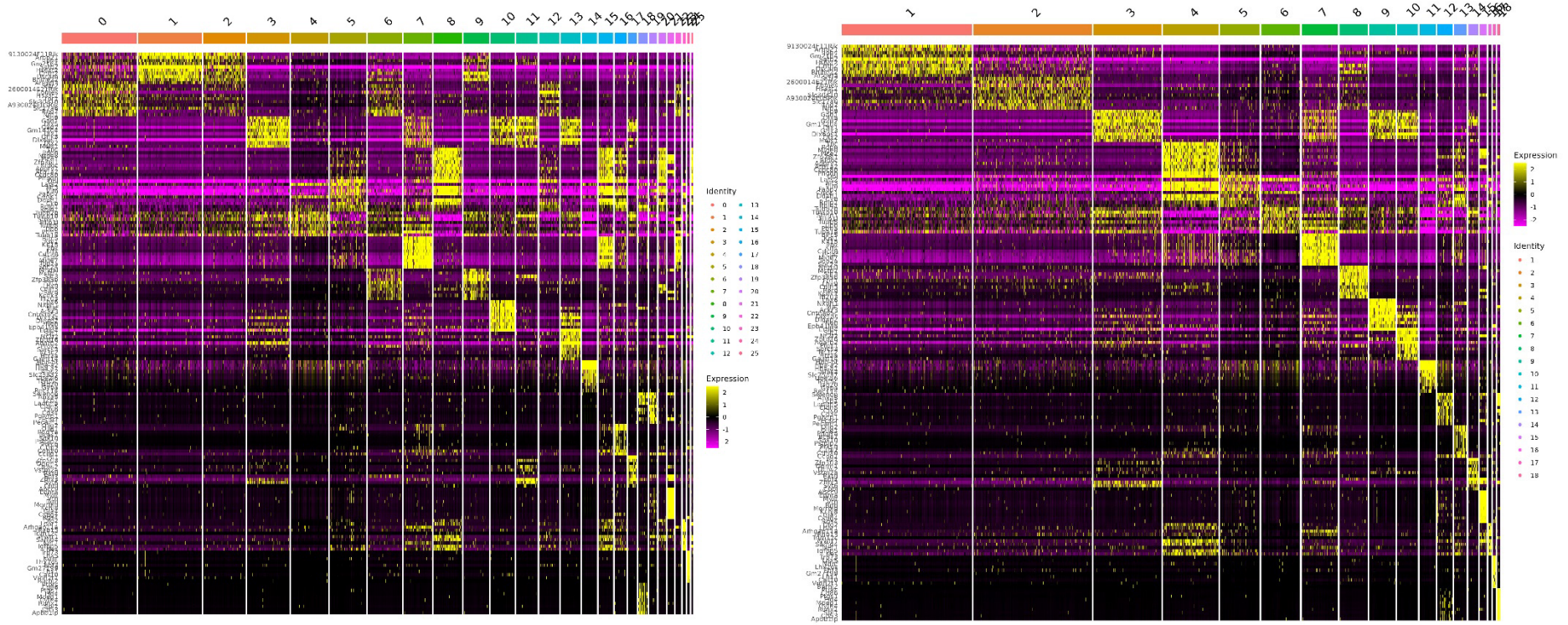


10k Mouse E18 Combined Cortex, Hippocampus and Subventricular Zone Cells, Chromium GEM-X Single Cell 3' Gene Expression Dataset by Cell Ranger 8.0.0, 10x Genomics, Inc, (2024, Apr 15). Used with permission of 10x Genomics, Inc.

Cell type calling: marker gene detection

- Which genes are most responsible for defining a cluster? How to find them?
- Differential expression testing of all cells of one cluster vs. cells of all other clusters combined
- 1-vs-rest may not find good markers if have too many clusters very similar to each other; works best at major groupings
- 1-vs-1 better for defining markers between subtypes (or re-do all steps separately for major clusters!)

Too many clusters vs. "about right"



10k Mouse E18 Combined Cortex, Hippocampus and Subventricular Zone Cells, Chromium GEM-X Single Cell 3' Gene Expression Dataset by Cell Ranger 8.0.0, 10x Genomics, Inc, (2024, Apr 15). Used with permission of 10x Genomics, Inc.

Cell type calling: marker gene detection

Use the marker genes to manually decide cell type:

- Look up genes in annotated databases (PanglaoDB, CellMarker 2.0 or other 10x recommendations)
- Pros:
  - Specific for your data set
- Cons:
  - "Canonical" markers may not be expressed in your data
  - Databases can show more than one cell type for particular genes
  - Tedious to do for many clusters

Cell type calling: computational from reference

Find a reference data set with cell type calls and compare expression values to find which of your cells have similar expression to the annotated cell types.

Within R: can use any reference that has cell type calls and expression values.

Web resources: various sites have set references that are easy to use:

- **Azimuth**
- **Tabula Sapiens**
- **SciBet**

Limitations of computational cell type calling

- You are trusting the cell type calls of others
- If your **data has cell types not in the reference**, they cannot be called correctly
- The reference could have been made with a very different method than your data that can greatly affect expression (e.g. cell-sorting + bulk RNA-Seq vs. non-sorted single cell)
- Even minor differences in experimental methodology can affect gene expression
- Automatic, 100% accurate cell type calling likely un-obtainable!

Comparing different treatments

- Single cell and spatial have quickly moved beyond simple with-sample cell type identification to comparing expression within a cell type between treatments.
- Due to cost limitations, first experiments only compared 1 replicate of each treatment. P-values of 1000s vs. 1000s of cells can get ridiculously low (1e-301)
- However, it is still a 1 vs. 1 comparison in terms of independent biological replicates; **differences could be due to batch effects, not treatment differences**
- Like bulk RNA-Seq, ideally you would have at least 3 biological replicates per treatment.

Challenges in spatial analyses

- Currently, most analyses of spatial data treat it just like single cell and ignore spatial information. Clusters just overlaid on spatial images at end.
- MERFISH, Xenium and other high-resolution methods often analyzed manually.
- New methods of incorporating spatial information during the UMAP/cluster calling are being developed.
- AI may be needed to scan large sections to find all areas of interest.
- How to differential expression between treatments if treatment changes the physical structure?

That's it for the Basic Single Cell & Spatial lecture!

- The lab will go over how to use the standard Seurat pipeline to analyze that single mouse brain sample.
- The HPCBio group provides experimental design advice, general consulting, cell/space ranger processing and/or downstream data analysis. Get in touch with us at hpcbio@biotech.Illinois.edu if you have any questions!
- THANK YOU!!