Basic Single Cell & Spatial Transcriptomics

Jenny Drnevich, PhD

Assistant Director, HPCBio

Roy J. Carver Biotechnology Center

June 26, 2024

Learning objectives

- To elucidate the differences between bulk, single cell and spatial RNA-Seq.
- To get a brief overview of the various single cell and spatial platforms
- To learn the details and vocabulary of 10x Genomic's sequencing methods for single cell and Visium
- To review the standard steps of quantification and analysis of 10x data, alternative methods and limitations of all methods.
- To present pros/cons of manual vs. automatic cell type calling
- To discuss the coming development of spatial data methods.

What is transcriptomics?

- The study of the "transcriptome" or the transcribed portion of the genome, **RNA**
- Most often this focuses on **mRNAs**, which are translated into proteins
- But also can include all other species of RNAs: **rRNAs, miRNAs, lncRNAs**, etc.

Bulk vs. single cell vs. spatial

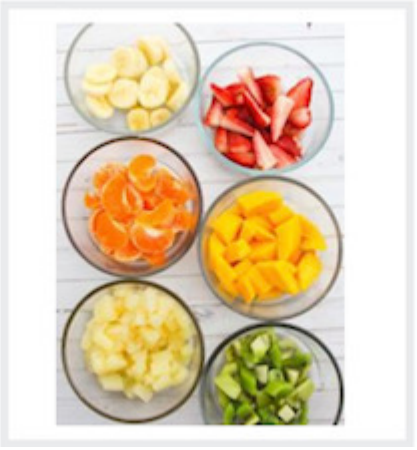
Spatial - the actual arrangement of cells in a tissue

Single cell - dissociated cells that can be separated into types

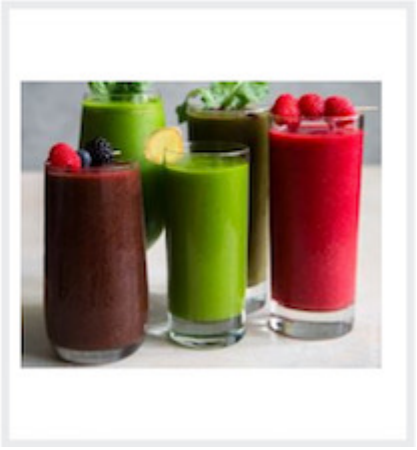
Bulk - whole tissue extractions averaged over all cells

Spatial

Single cell



Bulk



Hernadez, A. DNA Services, Applications [Internet]. Urbana (IL): University of Illinois, Carver Biotechnology Center; [cited 2024 Jun 6]. Available from: https://biotech.illinois.edu/htdna/applications

Single cell methods overview (see review)

Droplet based

- CEL-seq
- Drop-Seq
- Smart-Seq2
- 10x Genomics

Spatial methods overview (see review)

Sequencing based

- 10x Genomics Visium, Visium HD
- Slide-seq, Stereo-seq, Light-seq

Probe based - NanoString GeoMx

Imaging based

- NanoString CosMx
- MERFISH
- STARmap
- 10x Genomics Xenium

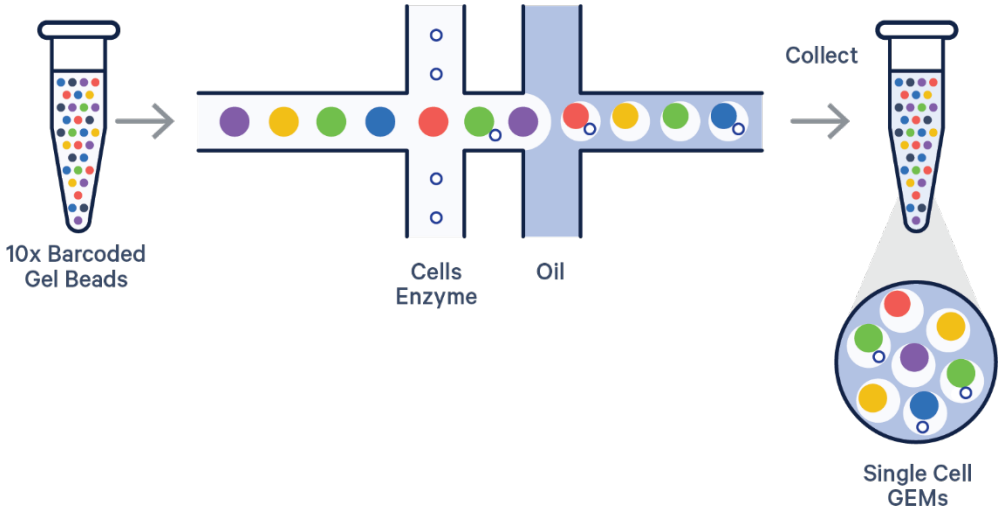
Short reads vs. long reads

Illumina's short reads (<150 nt) allowed all the sequencing-based technologies to be developed and allows excellent gene-level quantification

Pac Bio's HiFi long reads allow full-length (~10 kb) transcript sequencing plus Kinnex kit improves through-put; have worked with 10x Genomics to make the **10x single cell** libraries able to be sequenced by Pac Bio.

10x Genomics Chromium single cell

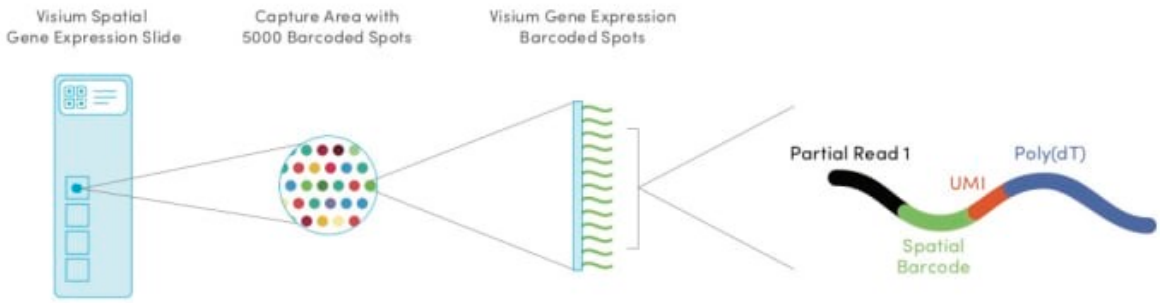
- Special Gel Beads flow in a lane past a channel with cells + enzyme cocktail, then past an oil channel creating nanodroplets.
- Most nanodroplets do not contain any cell; up to ~7% can contain 2+ cells
- Cell lysis and capturing of RNAs occurs within each droplet
- The output of each main lane becomes **one sample**.



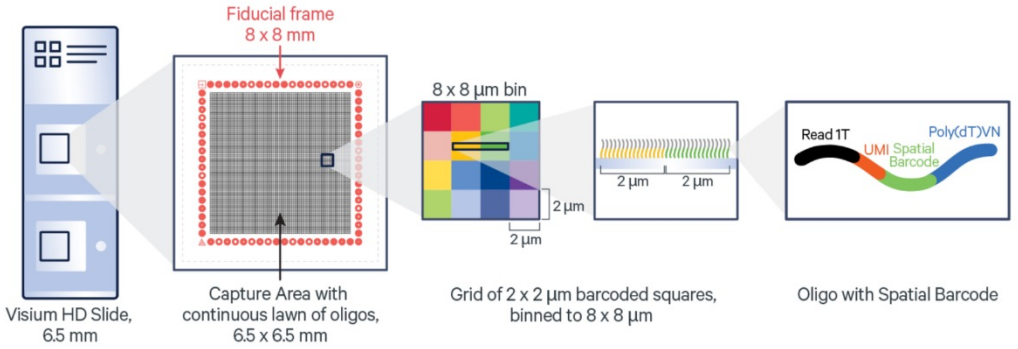
Ortolano, N. The neXt generation of single cell RNA-seq: An introduction to GEM-X technology [Internet]. 10x Genomics, Inc. 2024 Mar 11 [cited 2024 Jun 6]. Available from https://www.10xgenomics.com/blog/the-next-generation-of-single-cell-rna-seq-an-introduction-to-gem-x-technology. Used with permission of 10x Genomics, Inc.

10x Genomics Visium and Visium HD spatial

- Special slides have 2-4 capture areas compose of separated 55 µm "spots" or continuous 2 µm "squares"
- Tissue sections are imaged & placed on capture areas, then permeabilized to release RNAs
- RNAs migrate down to spots/squares below.
- The output of each capture area becomes **one sample**.



Habern, O. Answering Your Questions About the Visium Spatial Gene Expression Solution [Internet]. 10x Genomics, Inc. 2020 Jan 13 [cited 2024 Jun 6]. Available from https://www.10xgenomics.com/blog/answering-your-questions-about-the-visium-spatial-gene-expression-solution. Used with permission of 10x Genomics, Inc.



Habern, O. Your introduction to Visium HD: Spatial biology in high definition [Internet]. 10x Genomics, Inc. 2024 Apr 19 [cited 2024 Jun 6]. Available from https://www.10xgenomics.com/blog/your-introduction-to-visium-hd-spatial-biology-in-high-definition. Used with permission of 10x Genomics, Inc.

10x Genomics terminology (sc and spatial)

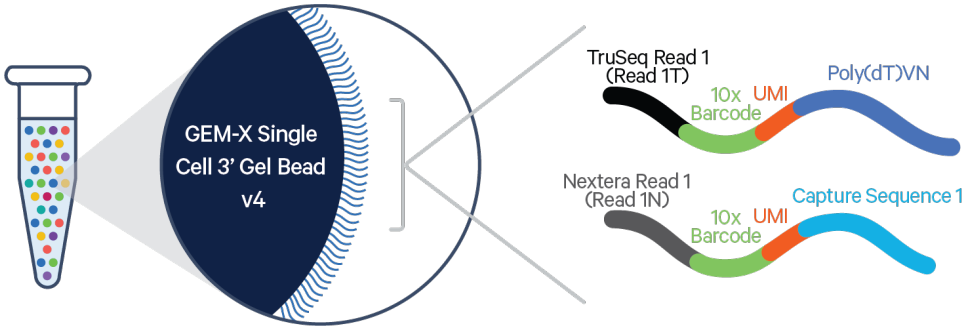
GEMs - Gel Beads-in-emulsions. The gel bead is covered in a lawn of oligonucleotides

Spot/Square - Physical location on a Visium slide covered in a lawn of oligonucleotides

Barcode - Part of the oligo sequence repeated within a GEM bead or slide spot; used to ID which transcripts were in the same GEM or spot/square.

UMI - Unique molecular identifier; UMIs differ between every oligo to uniquely label each original transcript. Used to collapse PCR duplicates.

Capture sequence - Part of the oligo used to capture the species of interest. Usually Poly(dT) for mRNAs.



Ortolano, N. The neXt generation of single cell RNA-seq: An introduction to GEM-X technology [Internet]. 10x Genomics, Inc. 2024 Mar 11 [cited 2024 Jun 6]. Available from https://www.10xgenomics.com/blog/the-next-generation-of-single-cell-rna-seq-an-introduction-to-gem-x-technology. Used with permission of 10x Genomics, Inc.

Anatomy of 10x-Illumina library construct

P5 & P7 - Illumina adapters to attach to the flow cell.

i5 & i7 indexes - Used to mark sequences from the same sample (i.e. each Chromium lane or Visium capture area).

Read 1 - Barcode + UMI

Read 2 - Sequence of the captured transcript

From Chromium Next GEM Single Cell 3' v3.1: Dual Index Libraries Technical Note (CG000325 Rev B), 10x Genomics, Inc. (2021, March 31). Used with permission of 10x Genomics, Inc.

Overview of sequencing outputs

- Multiple samples usually combined together and sequenced in 1+ NovaSeq lane
- Each sequencing lane produces one set of 4 fastq files
  1. R1 - 10x barcode + UMI
  2. R2 - actual transcript sequence
  3. I1 - sample i5 index
  4. I2 - sample i7 index
- Depending on the sequencing center, they may further **demultiplex** into one set of 4 fastq files per sample
  - I1 and I2 no longer needed at this point

Next steps after sequencing - quantification

1. Transcript **read** sequences need to be identified (i.e. which gene each came from)
2. **Reads** grouped by barcode (i.e., GEM/spot/square)
3. **Reads** collapsed to **UMI counts** (discard PCR duplicates)
4. Call cells/spots/squares
    1. SC - Use **UMI counts** to see which GEMs likely contained a cell
    2. Visium - visually find tissue-covered spots/squares
5. Output **UMI counts** per gene per cell/spot/square

Alignment and quantification – 10x pipelines

- Cell Ranger
  - Set of pipelines to process single cell gene expression and all other types of Chromium libraries; aligns to genome using STAR
  - System requirements: 16 cores, 128 GB RAM, 1 TB disk space, 64-bit Linux OS
  - Alternatively, upload to 10x Genomics Cloud Atlas for free*
    - Web page point-and-click interface for most Cell Ranger pipelines
    - *limited number of analyses and downloads; deleted after 6 months
    - Faster run time than most compute clusters (not including up/download)

- Space Ranger
  - Set of pipelines to process Visium and Visium HD libraries; aligns to genome using STAR
  - System requirements: 32 cores, 128 GB RAM, 1 TB disk space, 64-bit Linux OS
  - Support for analyzing Visium HD in 10x Genomics Cloud Atlas coming in 2024
  - Often may need to use Loupe Browser to manually align images and select tissue-covered spots

Both available on IGB's Biocluster

Alignment and quantification – alternative pipelines for single cell data

- Salmon/Alevin
  - Pseudo-align to the transcriptome; runs 30X faster
- Kallisto/bustools
  - Also a pseudo-aligner to transcriptome; runs ~40x faster
  - More focused on downstream trajectory analyses?
- STARsolo
  - Output almost identical to `cellranger count` except no secondary analyses
  - 10X faster than `cellranger count --nosecondary`

Brüning et al. 2022 compare all 3 to Cell Ranger; biggest difference was faster run times

All available on IGB's Biocluster

Output from cellranger count

```
├── analysis/
│   ├── clustering/
│   ├── diffexp/
│   ├── pca/
│   ├── tsne/
│   └── umap/
├── cloupe.cloupe                      <- file for Loupe Browser
├── filtered_feature_bc_matrix/         <- filtered UMI counts to read into R
│   ├── barcodes.tsv.gz
│   ├── features.tsv.gz
│   └── matrix.mtx.gz
├── filtered_feature_bc_matrix.h5      <- Filtered UMI counts compressed to Hdf5
├── metrics_summary.csv
├── molecule_info.h5
├── raw_feature_bc_matrix/              <- raw UMI counts if want to call cells on own
│   ├── barcodes.tsv.gz
│   ├── features.tsv.gz
│   └── matrix.mtx.gz
├── raw_feature_bc_matrix.h5
└── web_summary.html                   <- Main summary file
```

Output from spaceranger count (Visium)

All the same as cellranger count plus additionally:

```
├── deconvolution          <- reference-free spot partitioning
│   ├── deconvolution_k2
│   ├── deconvolution_k3
│   ├── deconvolution_k4
│   ├── deconvolution_k5
│   ├── deconvolution_k6
│   ├── deconvolution_k7
│   ├── dendrogram_k7_distances.png
│   └── dendrogram_k7.png
└── spatial
    ├── aligned_fiducials.jpg
    ├── detected_tissue_image.jpg
    ├── scalefactors_json.json
    ├── spatial_enrichment.csv
    ├── tissue_hires_image.png
    ├── tissue_lowres_image.png
    └── tissue_positions.csv
```

Loupe Browser

- The last part of 10x's "end-to-end" product offerings
- Free, GUI-interface software to explore Chromium and Visium results in cloupe.cloupe files create by *ranger pipelines; many tutorials
- Always do **first QC check** of the cloupe.cloupe and the web_summary.html
  - Alerts or warnings?
  - Number of cells/spots and number of genes detected?
  - % of reads mapped to genome?
  - First look at clustering of cells/spots
- Can do limited additional analyses:
  - Additional cell filtering based on QC metrics
  - Re-do clustering/tSNE/UMAP
  - Differential expression testing between defined groups
  - Can compare > 1 sample if have run *ranger aggr pipeline
- Cons:
  - "Point-and-click" less reproducible
  - Sub-optimal normalization, although can import other results via .csv files or LoupeR.

Why additional downstream analysis is needed

- Within one sample, no normalization done
- When aggregating >1 sample, only normalization is to down-sample to smallest library size
- Cell calling is liberal by design to not miss cells with naturally low RNA abundance
- QC thresholds for dead/dying cells can vary greatly between tissue and cell types so can't be automated
- Proper statistical methods for > 2 samples not available in Loupe
- Loupe figures not high enough quality for publication

Analysis software options (free)

R-based

- Seurat software suite (will do in lab)
- Bioconductor
  - 282 single cell and 50 spatial packages
  - Amezquita et .al's Orchestrating Single-Cell Analysis with Bioconductor book
  - Weber et al.'s Best Practices for Spatial Transcriptomics Analysis with Bioconductor
  - Righelli et .al's ISMB 2023 workshop

Python-based

- scverse
  - 64 total packages (scanpy, Squidpy, scVelo, Squidpy, SpatialData,etc.)
  - 21 tutorials

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)

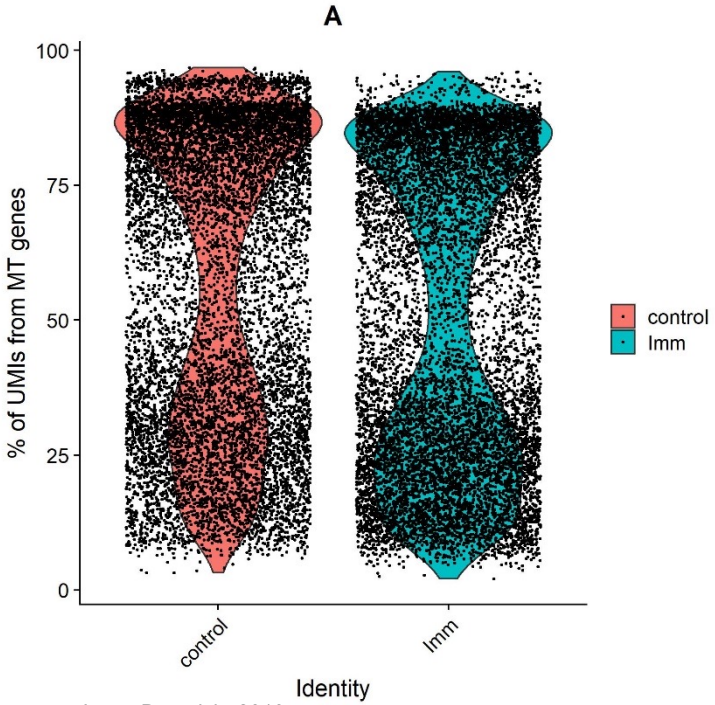
Cell/gene QC filtering

- 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

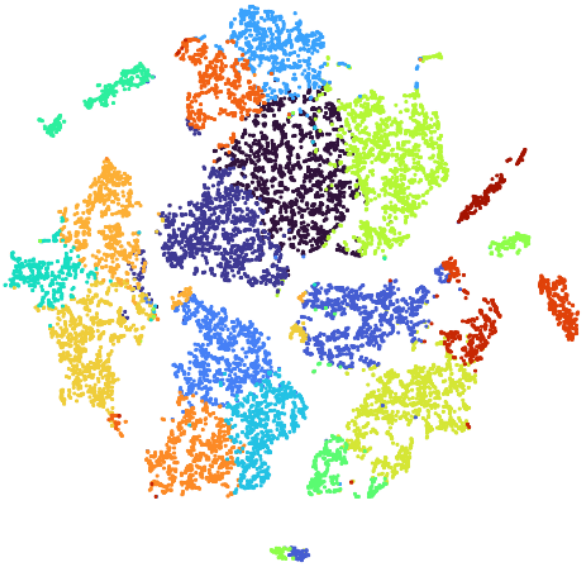
2nd dimension reduction option: tSNE

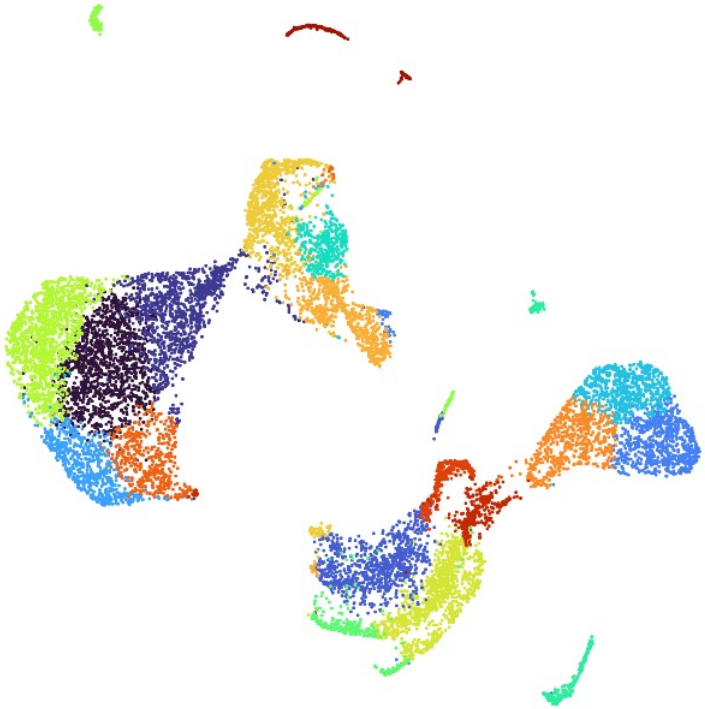
- t-Distributed Stochastic Neighbor Embedding (van der Maaten & Hinton, 2008)
- Non-linear dimension reduction algorithm to visualize high-dimensional data in just 2 dimensions (XY plot)
- StatQuest explanatory video
- First widely used in single cell analysis and still first one shown in Cell/Space 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 place 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.

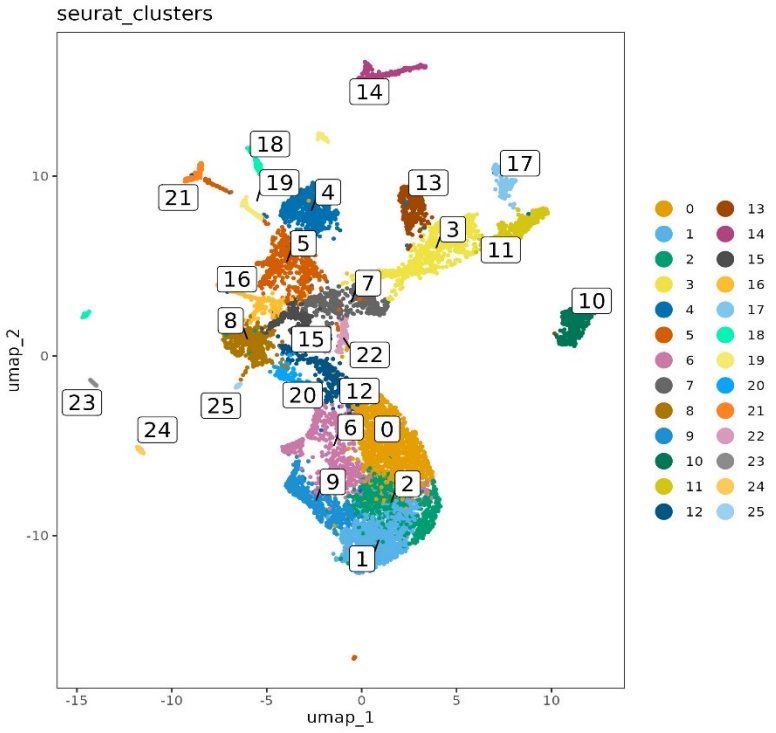
Warning about 2D representations!

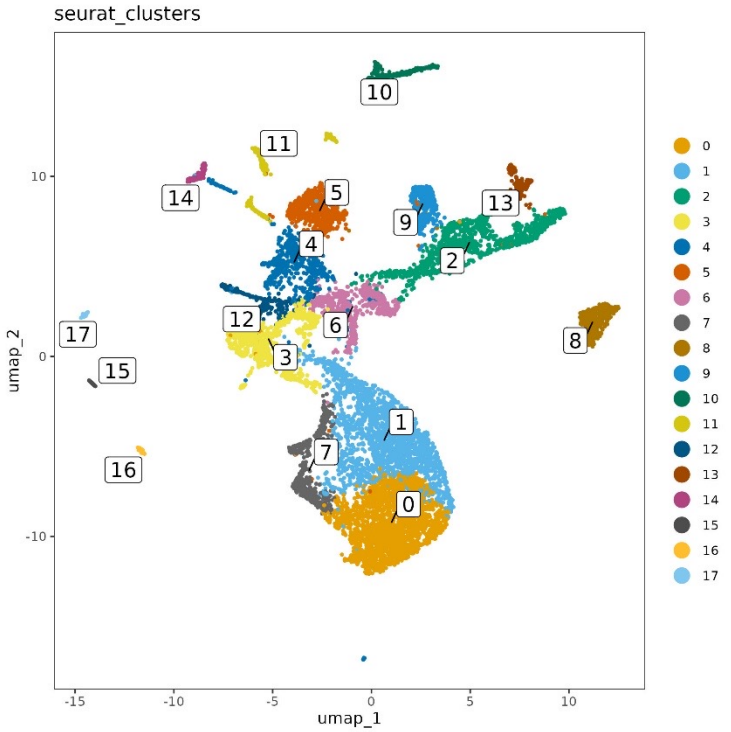
- There is no single "correct" 2D representation of your single cell data.
- Both algorithms are non-deterministic, starting from a random number. A different random number will lead to a slightly different XY graph
- Within t-SNE and UMAP, there are many parameters that could be tweaked and lead to 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"



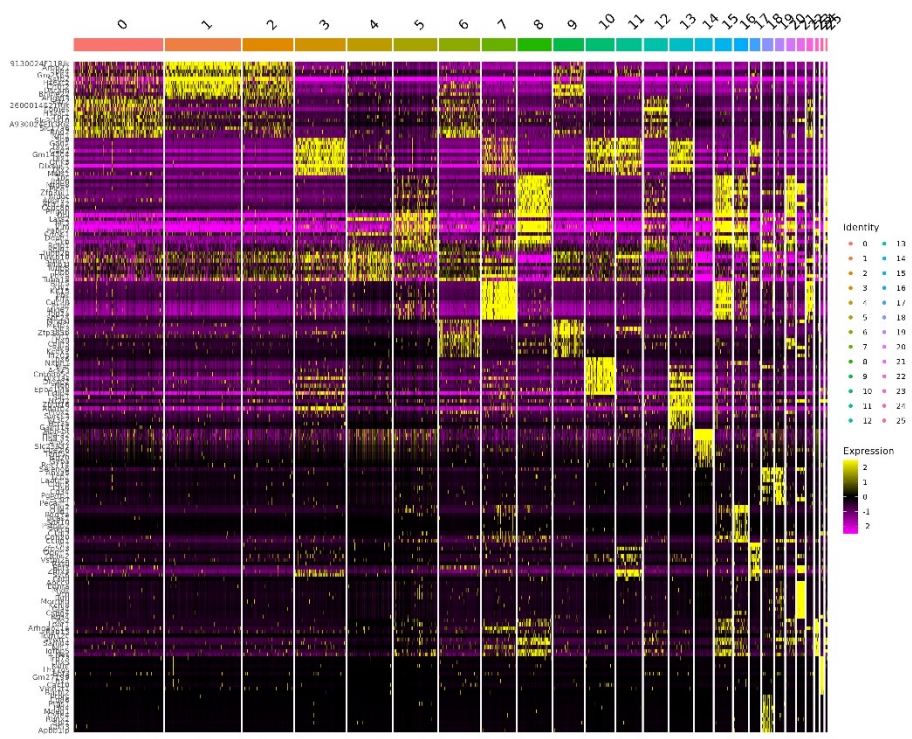


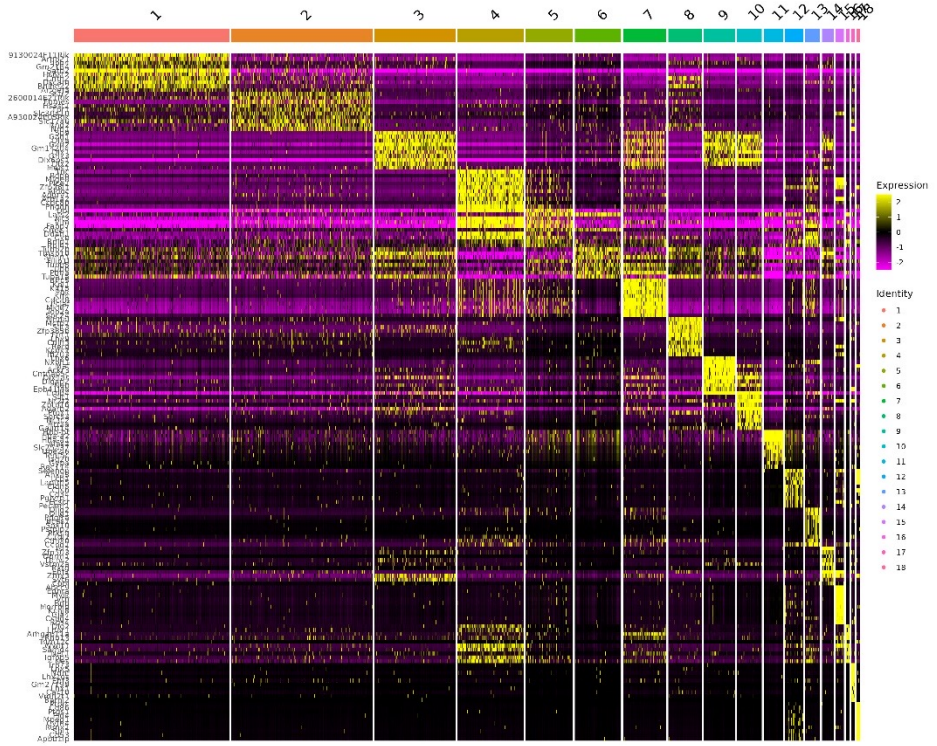
10k Mouse E18 Combined Cortex, Hippocampus and Subventricular Zone Cells, Chromium GEM-X Single Cell 3' Gene Expression Dataset by Cell Ranger 8.0.0, 10x Genomics, Inc, (2024, Apr 15). Used with permission of 10x Genomics, Inc.

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?

Upcoming FREE spatial analysis workshop!

Spatial multi-omics analysis with Giotto Suite

August 5-7, 2024 10 am - 4 pm EST

Register now

"Our Giotto Suite prototype pipeline is generally applicable on various different datasets, such as those created by state-of-the-art spatial technologies, including in situ hybridization (seqFISH+, merFISH, osmFISH, CosMx), sequencing (Slide-seq, Visium, STARmap, Seq-Scope, Stereo-Seq) and imaging-based multiplexing/proteomics (CyCIF, MIBI, CODEX)."

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!