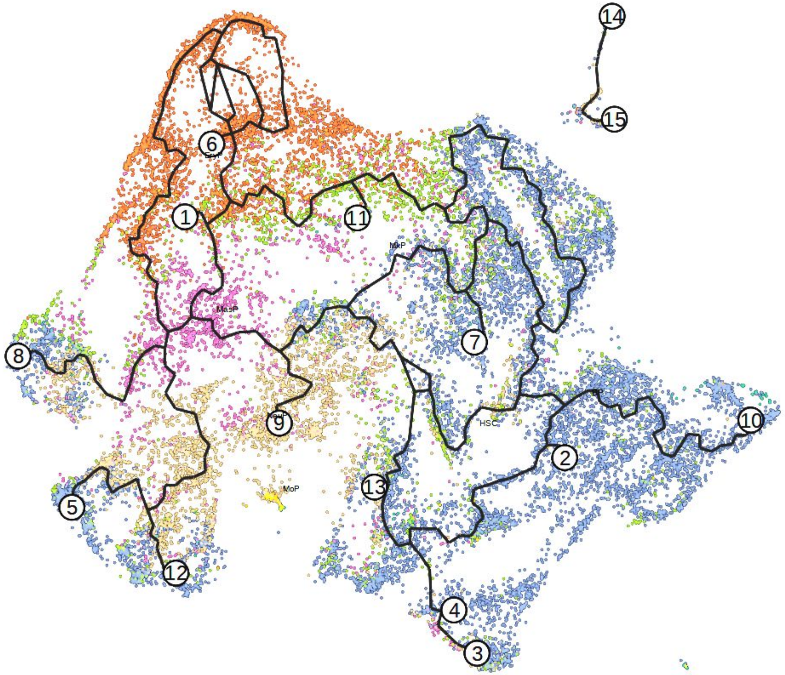
Exploring Single Cell Transcriptomes

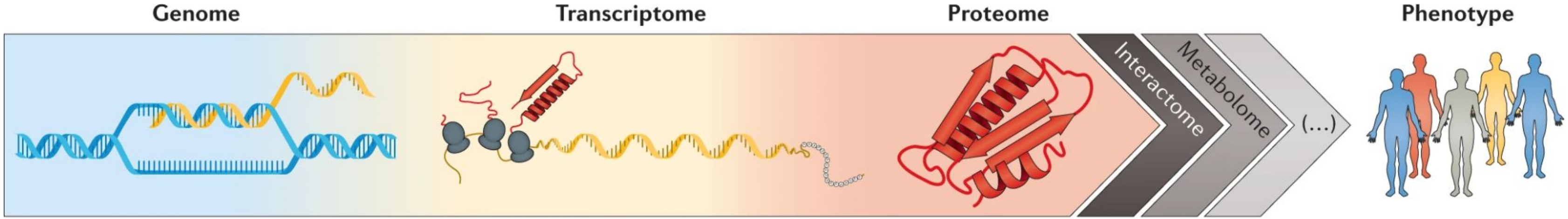


Brett Kiyota
Kieran Maheden

CPSC 547
221214

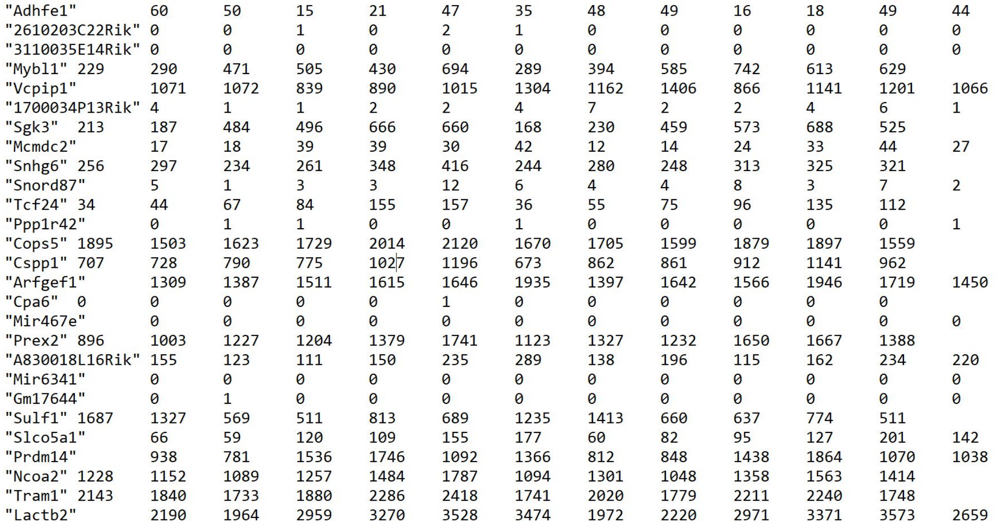
Background

- All cells have the same set of DNA and genes
- How do different cell types arise from the same set of blueprints?
- Not all genes are *expressed* (turned on) in every cell
- Expression entails *Transcription* (DNA to RNA) and *Translation* (RNA to protein)
- RNA is typically considered an inert intermediate between long term information storage (DNA) and the functional form of the information (protein)



Buccitelli et al., Nat Rev Gen, 2020.

How do we learn about gene expression?

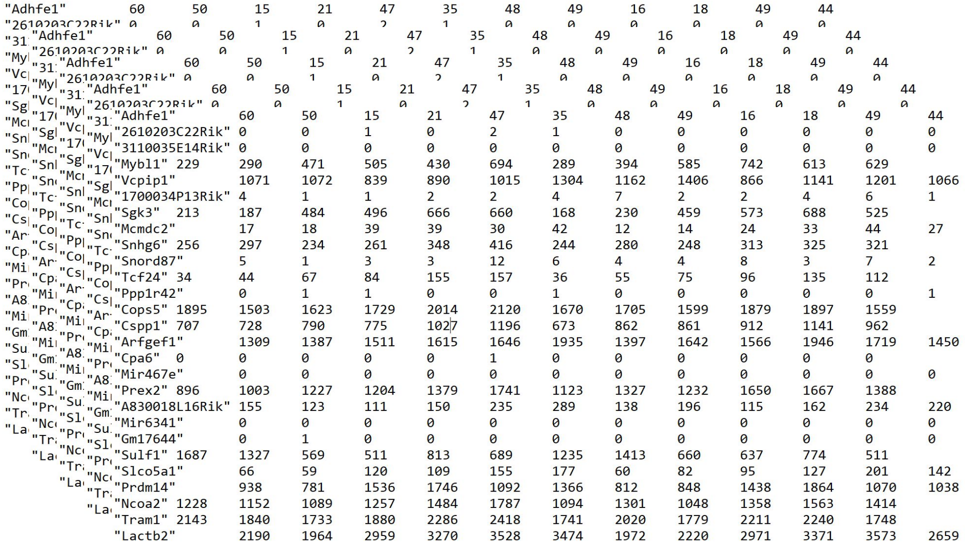


```
"Adhfe1"        60      50      15      21      47      35      48      49      16      18      49      44
"2610203C22Rik" 0       0       1       0       2       1       0       0       0       0       0       0
"3110035E14Rik" 0       0       0       0       0       0       0       0       0       0       0       0
"Mybl1" 229     290     471     505     430     694     289     394     585     742     613     629
"Vcpip1"        1071    1072    839     890     1015    1304    1162    1406    866     1141    1201    1066
"1700034P13Rik" 4       1       1       2       2       4       7       2       2       4       6       1
"Sgk3"  213     187     484     496     666     660     168     230     459     573     688     525
"Mcmdc2"        17      18      39      39      30      42      12      14      24      33      44      27
"Snhg6" 256     297     234     261     348     416     244     280     248     313     325     321
"Snord87"       5       1       3       3       12      6       4       4       8       3       7       2
"Tcf24" 34      44      67      84      155     157     36      55      75      96      135     112
"Ppp1r42"       0       1       1       0       0       1       0       0       0       0       0       1
"Cops5" 1895    1503    1623    1729    2014    2120    1670    1705    1599    1879    1897    1559
"Cspp1" 707     728     790     775     1027    1196    673     862     861     912     1141    962
"Arfgef1"       1309    1387    1511    1615    1646    1935    1397    1642    1566    1946    1719    1450
"Cpa6"  0       0       0       0       0       1       0       0       0       0       0       0
"Mir467e"       0       0       0       0       0       0       0       0       0       0       0       0
"Prex2" 896     1003    1227    1204    1379    1741    1123    1327    1232    1650    1667    1388
"A830018L16Rik" 155     123     111     150     235     289     138     196     115     162     234     220
"Mir6341"       0       0       0       0       0       0       0       0       0       0       0       0
"Gm17644"       0       1       0       0       0       0       0       0       0       0       0       0
"Sulf1" 1687    1327    569     511     813     689     1235    1413    660     637     774     511
"Slco5a1"       66      59      120     109     155     177     60      82      95      127     201     142
"Prdm14"        938     781     1536    1746    1092    1366    812     848     1438    1864    1070    1038
"Ncoa2" 1228    1152    1089    1257    1484    1787    1094    1301    1048    1358    1563    1414
"Tram1" 2143    1840    1733    1880    2286    2418    1741    2020    1779    2211    2240    1748
"Lactb2"        2190    1964    2959    3270    3528    3474    1972    2220    2971    3371    3573    2659
```

Example count data of 12 averaged cell populations (columns) for a few given genes (rows)

- Often researchers are only interested in a single gene
- However when performing exploratory experiments or trying to characterize a new system, a more full view is more valuable
- Based on how the technologies have developed over the past two decades, RNA sequencing is the leading method for profiling gene expression
- Classically, cells are taken as a population, and RNA is isolated and sequenced, producing a count for each gene
- However this approach misses heterogeneity in the averaged sample, erasing the evidence for rare cell types or states

Single-cell RNA sequencing (scRNAseq)

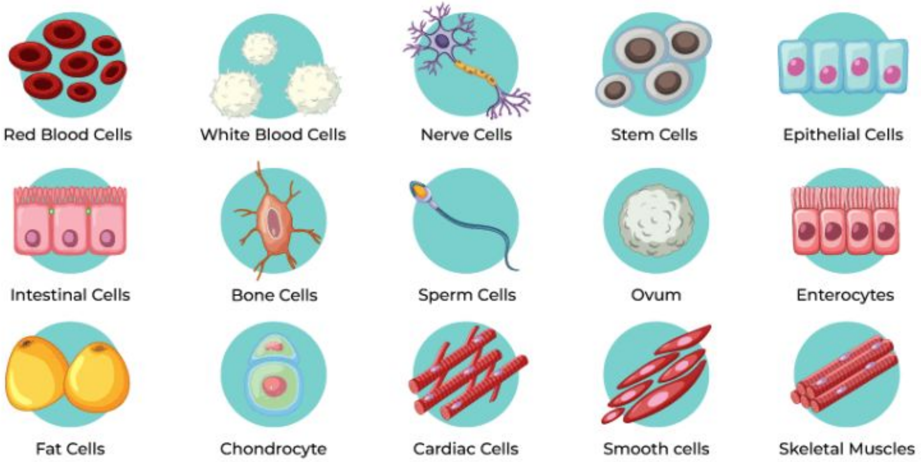


```
"Adhfe1"        60      50      15      21      47      35      48      49      16      18      49      44
"2610203C22Rik" 0       0       1       0       2       1       0       0       0       0       0       0
"3110035E14Rik" 0       0       0       0       0       0       0       0       0       0       0       0
"Mybl1" 229     290     471     505     430     694     289     394     585     742     613     629
"Vcpip1"        1071    1072    839     890     1015    1304    1162    1406    866     1141    1201    1066
"1700034P13Rik" 4       1       1       2       2       4       7       2       2       4       6       1
"Sgk3"  213     187     484     496     666     660     168     230     459     573     688     525
"Mcmdc2"        17      18      39      39      30      42      12      14      24      33      44      27
"Snhg6" 256     297     234     261     348     416     244     280     248     313     325     321
"Snord87"       5       1       3       3       12      6       4       4       8       3       7       2
"Tcf24" 34      44      67      84      155     157     36      55      75      96      135     112
"Ppp1r42"       0       1       1       0       0       1       0       0       0       0       0       1
"Cops5" 1895    1503    1623    1729    2014    2120    1670    1705    1599    1879    1897    1559
"Cspp1" 707     728     790     775     1027    1196    673     862     861     912     1141    962
"Arfgef1"       1309    1387    1511    1615    1646    1935    1397    1642    1566    1946    1719    1450
"Cpa6"  0       0       0       0       0       1       0       0       0       0       0       0
"Mir467e"       0       0       0       0       0       0       0       0       0       0       0       0
"Prex2" 896     1003    1227    1204    1379    1741    1123    1327    1232    1650    1667    1388
"A830018L16Rik" 155     123     111     150     235     289     138     196     115     162     234     220
"Mir6341"       0       0       0       0       0       0       0       0       0       0       0       0
"Gm17644"       0       1       0       0       0       0       0       0       0       0       0       0
"Sulf1" 1687    1327    569     511     813     689     1235    1413    660     637     774     511
"Slco5a1"       66      59      120     109     155     177     60      82      95      127     201     142
"Prdm14"        938     781     1536    1746    1092    1366    812     848     1438    1864    1070    1038
"Ncoa2" 1228    1152    1089    1257    1484    1787    1094    1301    1048    1358    1563    1414
"Tram1" 2143    1840    1733    1880    2286    2418    1741    2020    1779    2211    2240    1748
"Lactb2"        2190    1964    2959    3270    3528    3474    1972    2220    2971    3371    3573    2659
```

- One recent method that addresses heterogeneity is single cell RNA sequencing
- Rather than averaging the RNA across all cells, each cell is isolated first, then RNA is collected
- This method generates orders of magnitude more data
- Rather than each sample or population average having a count per gene, each cell has a count per gene
- Experiments can include >20 000 genes and >100 000 cells with multiple “batches” of cells

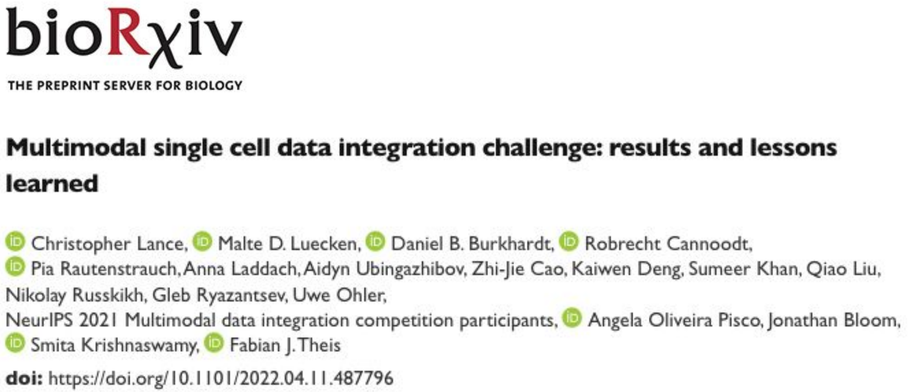
Cell type labelling task

- Cells are dynamic entities that are constantly changing in response to internal and external factors.
- Assigning a discrete label to cells (often based on gene expression) is common practice to ease the process of interpreting sequencing data
- Large efforts focused on linking changes in gene expression to the behaviour of certain cell types
  - Typically, disease states only appear in specific cell types
  - Connecting cell type to disease to potential intervention is a common goal



https://www.geeksforgeeks.org/structure-and-types-of-animal-tissues/

Dataset



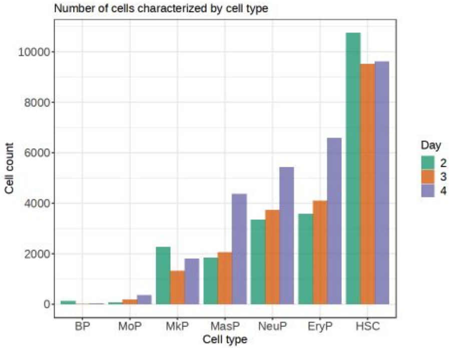
bioRχiv

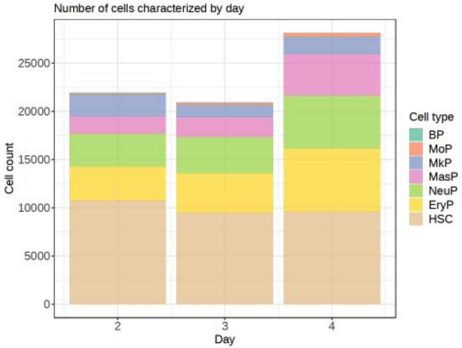
THE PREPRINT SERVER FOR BIOLOGY

Multimodal single cell data integration challenge: results and lessons learned

Christopher Lance, Malte D. Luecken, Daniel B. Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, Qiao Liu, Nikolay Russkikh, Gleb Ryazantsev, Uwe Ohler, NeurIPS 2021 Multimodal data integration competition participants, Angela Oliveira Pisco, Jonathan Bloom, Smita Krishnaswamy, Fabian J. Theis

doi: https://doi.org/10.1101/2022.04.11.487796





kaggle

Open Problems - Multimodal Single-Cell Integration

Predict how DNA, RNA & protein measurements co-vary in single cells

- RNA expression data is provided as a flat table with 70,988 items (cells) and 22,050 attributes (genes)
- The value for each item-attribute pair is a quantitative value representing the normalized and transformed RNA counts (quantitative information)
- An additional 5 attributes encode:
  - cell_id (*categorical*): Unique alphanumeric string that is assigned to each cell
  - day (*sequentially ordered quantitative*): the time point at which sequencing was performed
  - donor (*categorical*): a unique identifying number that is assigned to the 4 healthy adult donors
  - cell_type (categorical): inferred cell type label

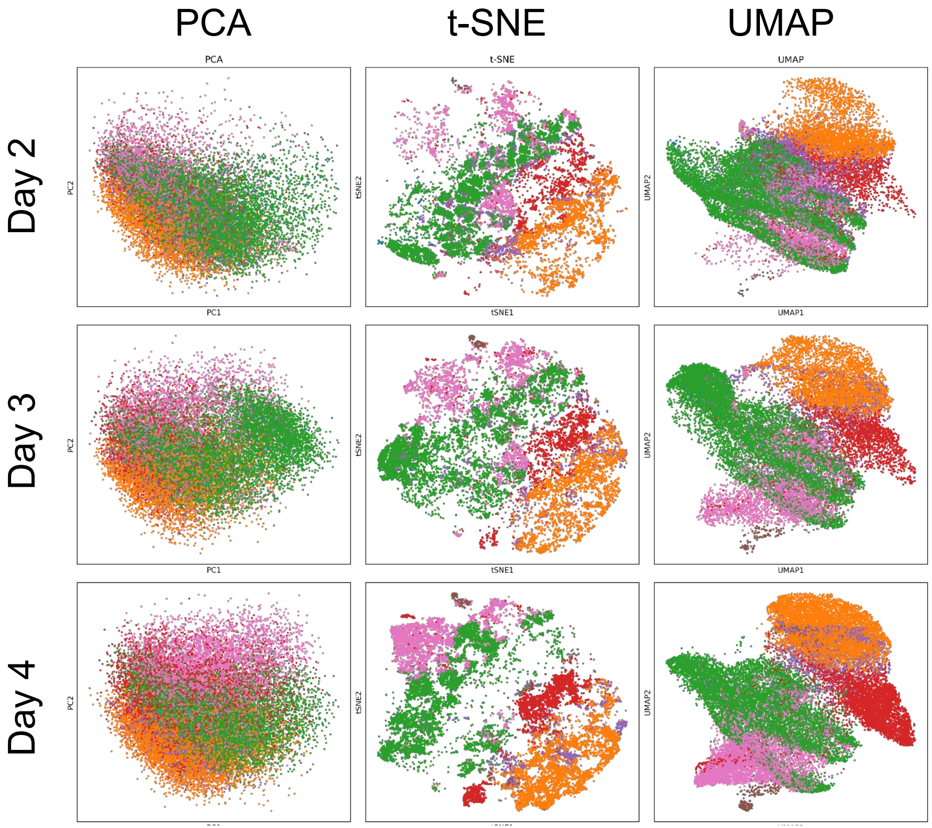
Goals of our analysis project

- Explore a scRNAseq dataset using common and accessible analysis packages and analyze how the different visual encodings can affect interpretations and guide downstream analyses
    - Seurat (https://github.com/satijalab/seurat)
    - Scanpy (https://github.com/scverse/scanpy)
    - Additional softwares: trajectory inference (Monocle3), graphs (PAGA)

1. Evaluate how well are overall cell type differences visualized between and within timepoints.

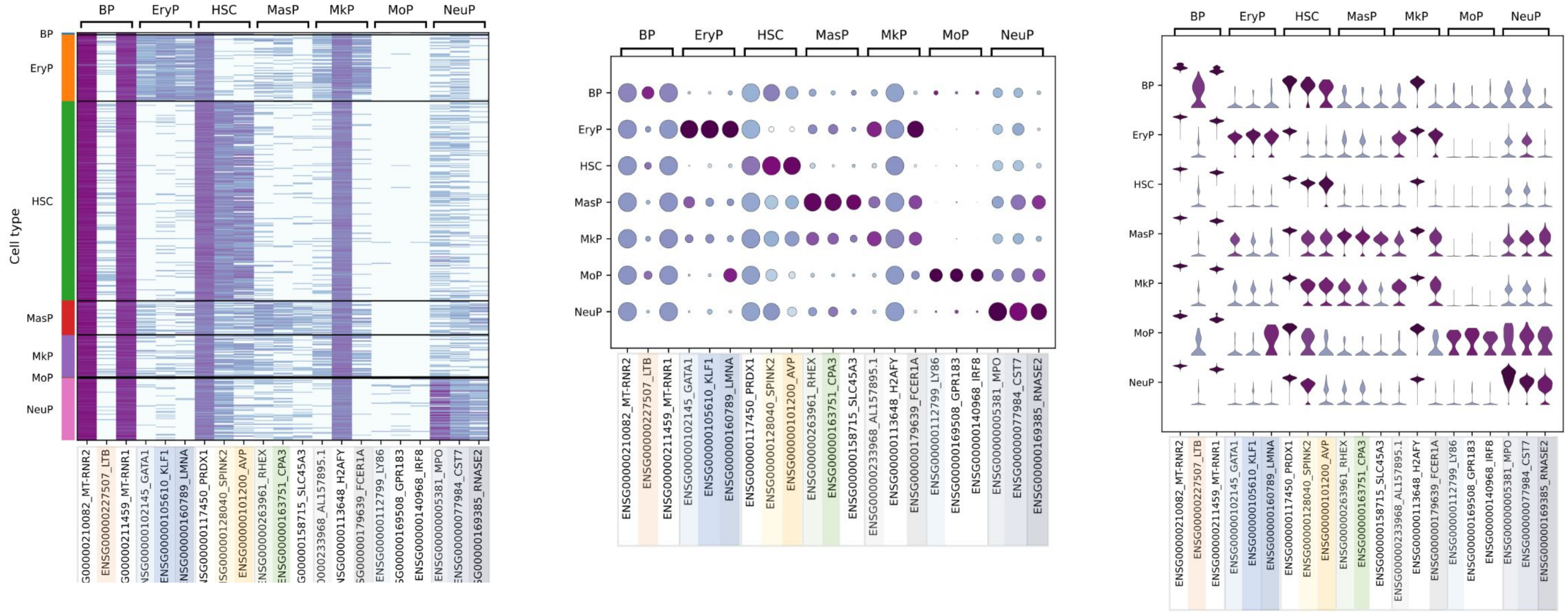
2. Evaluate how well are changes in specific genes visualized between and within timepoints.

Cell types between and within time points with DR



- Dimensionality reduction is a common tool for global analysis of scRNAseq data
- Each point represents a cel
- Three common DR tools
  - PCA
  - t-SNE
  - UMAP
- Clear differences between linear (PCA) and nonlinear DR viz
- Some noticeable differences between UMAP and t-SNE for finer details or clustering
- Qualitative interpretations

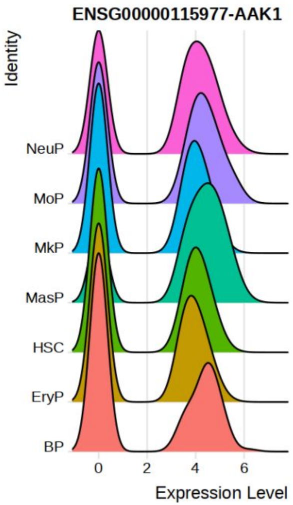
Cell types between and within time points: subsets of genes

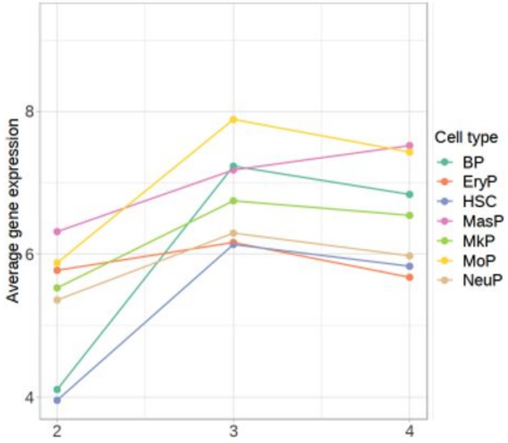


- Three different representations of the same heatmap
- Left: each row is a cell
- Middle: each row is the aggregation of all the cells of that cell type
- Right, each row shows the distribution of expression values for that cell type

Comparing between within timepoints with density plots

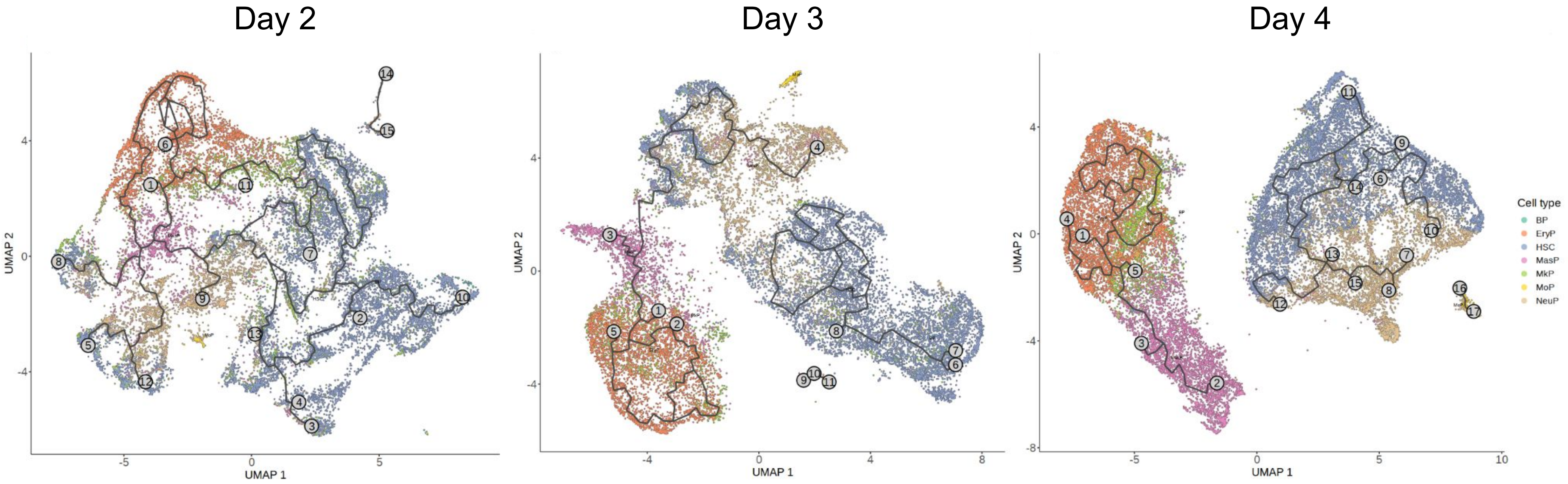
Looking at individual genes





- Each cell type is represented by the density of expression values
- Each column of the graph is a selected ene
- lote that while this gives information on a single gene, it also gives us an idea of the distribution of that gene
- Note the large number of cells that have no expression level of a given gene

Trajectory inference



Day 2

Day 3

Day 4

- UMAP dimensionality reduction
- Learn a trajectory (pseudotemporal ordering) that fits the cells’ lower-dimensional coordinates (principal graph embedding algorithms)

Take away from standard pipeline

- Many options for examining expression within a single time point
- No inbuilt tools for looking at genes that are dynamic across timepoints in a given cell type
- Comparisons with DR tools also tend to be very qualitative and many key comparisons cannot be made between timepoints
- All of the tools tested were very computationally intensive - issues for accessibility
  - DR plots took ~30 minutes to run per plot
  - Unable to run many existing trajectory inference softwares given our hardware (laptops)
- Dichotomy between available visual encodings provided by these widely used toolkits.
  - Attempting to capture global variation in the dataset where metrics are purely qualitative (dimensionality reduction)
  - Selection of a few specific genes for more quantitative analyses
- Future direction of visual encodings: interactivity, combine dimensionality reduction methods with visual encodings that can lead to quantitative interpretations