

Replicability of single-cell RNA-seq cell type identification

Amy Gill

4/26/2021

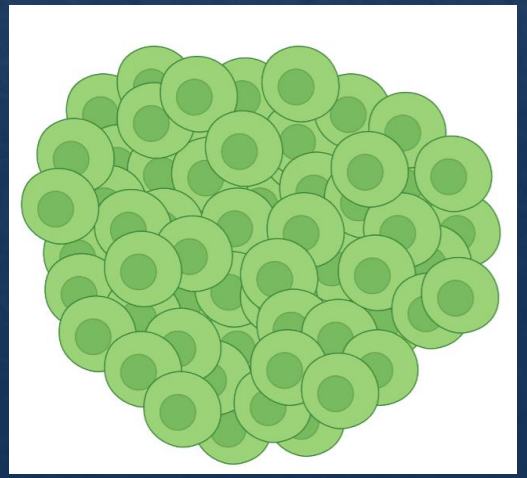
Final Project

Advanced Topics in Genomic Data Analysis

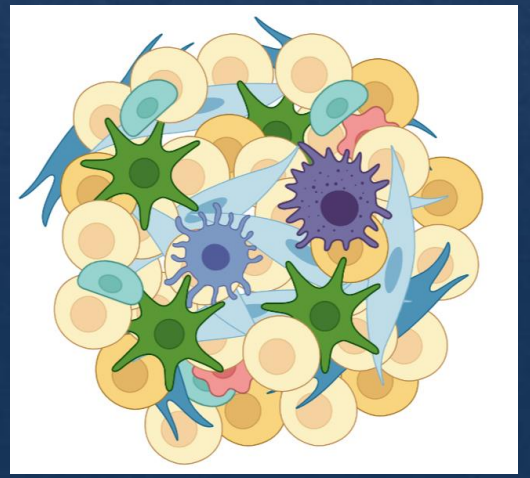
Introduction

scRNA-seq to analyze tissue heterogeneity

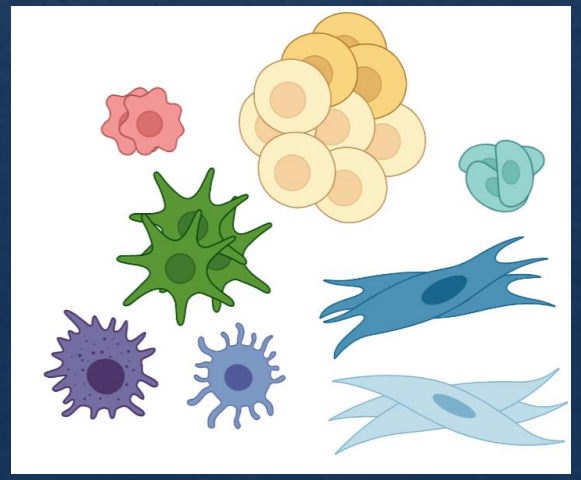
- Tissues and tumors are ecosystems of functionally distinct cell types in various states, and these differences are obscured in bulk analysis methods
- Single-cell RNA sequencing (scRNA-seq) allows transcriptome-scale investigation of tissue heterogeneity on a cellular level
- Numerous downstream applications require individual cells in a dataset to first be classified into cell types in order to analyze clusters of similar cells – how?



Bulk view of tumor



Single cell view of tumor



Goal: cluster cells by type

Overview of cell typing methods

- Devise decision rules based on expression of a small number of known marker genes
  - Pro: Uses biological knowledge for decision making
  - Cons: Markers may not be consistently expressed in all cells of a type or specific to one type; requires prior knowledge; can leave ambiguous cells unclassified
- Use supervised machine learning methods to classify unknown cells based on training data on cells of known types
  - Pro: When possible, compares unknown cells to gold standard cell identities
  - Cons: Requires reliable scRNA-seq training data with known representative cell types, which may not exist; training data cell types have to be originally determined somehow
- Use unsupervised learning methods to identify clusters of cells with similar characteristics
  - Pro: No prior knowledge or gold standard reference data required
  - Cons: Does not leverage existing biological knowledge; biological meaning of clusters unclear without further analysis (cluster-specific marker genes, etc.); “true” number of clusters unknown

Goals of this study

- Find cell clusters in triple negative breast cancer (TNBC) scRNA-seq data using marker-based cell typing and various unsupervised clustering algorithms
- Quantitatively compare composition of clusters/types produced by different approaches
  - To what degree do clusters produced by different algorithms produce the same groups of cells?
  - To what degree do clusters from by different algorithms have the same differentially expressed genes?

HYPOTHESIS

Different clustering methods will capture similar biological information and produce similar clusters, but some methods will produce more similar results than others.

- Implications of expected results
  - Some clustering methods may produce more biologically interpretable results – use these first in future research requiring cell typing
  - Ensemble approach could combine complementary clustering methods to give more robust cell typing

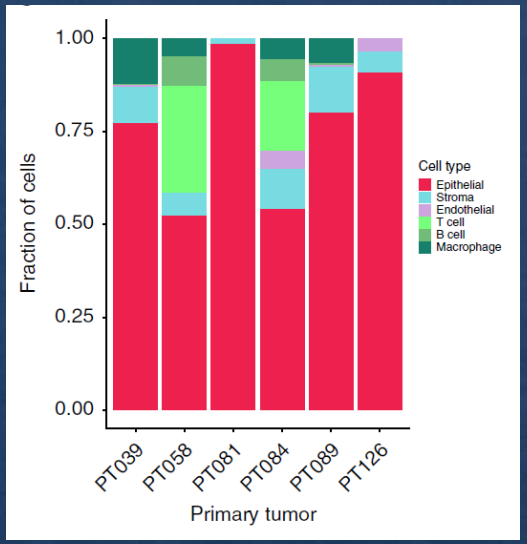
Data and Methods

Data: TNBC scRNA-seq from Karaayvaz et al., 2018

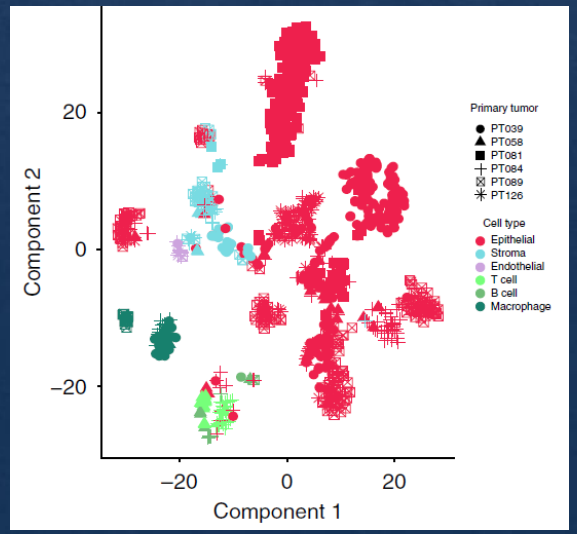
Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq

Mihriban Karaayvaz, Simona Cristea, Shawn M. Gillespie, Anoop P. Patel, Ravindra Mylvaganam, Christina C. Luo, Michelle C. Specht, Bradley E. Bernstein, Franziska Michor & Leif W. Ellisen

Nature Communications **9**, Article number: 3588 (2018) | Cite this article



- Downloaded TPM counts as well as QC-filtered, normalized data from GEO (GSE118398)
- 1189 QC-filtered cells from 6 TNBC patients
- In the paper, cells were assigned to 6 cell types using a literature-based panel of marker genes (epithelial, stromal, endothelial, T cell, B cell, macrophage)
- Markers and decision rules parsed from supplemental material



Karaayvaz et al., 2018

Workflow

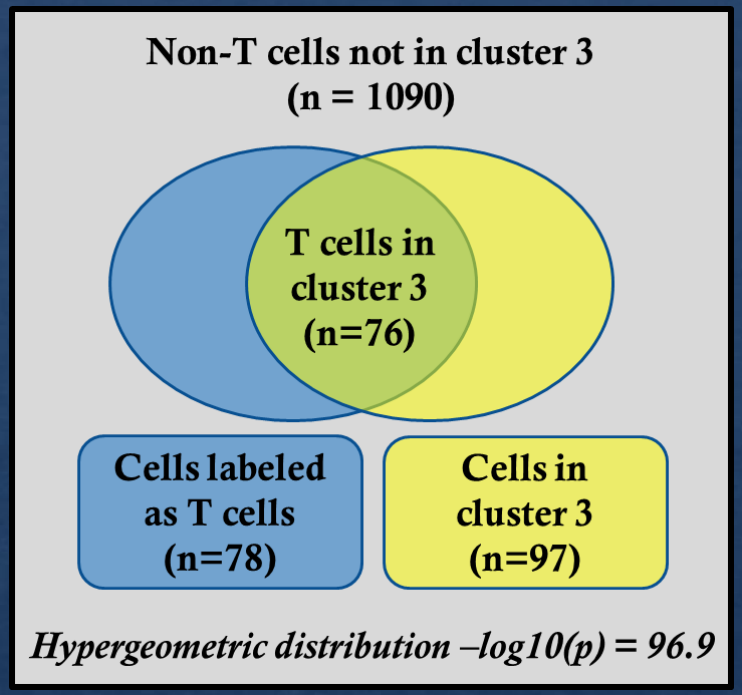
- Import data in R as SingleCellExperiment and Seurat objects
- Remove patient-specific effects by linear regression
- Find top 2000 variable features and find principal components (PCs)
- Perform tSNE and UMAP projection of cells
- Replicate marker-based cell typing using literature-based markers and decision rules
  - Visualize cell types with tSNE and UMAP
  - Find genes differentially expressed between cell types
- Perform clustering with the Louvain algorithm using the top 50 PCs (Seurat)
  - Visualize clusters with tSNE and UMAP
  - Find cells differentially expressed between clusters
- Use multimodal intersection analysis (MIA, Moncada et al., 2020) to quantify the degree of overlap between clusters/types produced by different algorithms

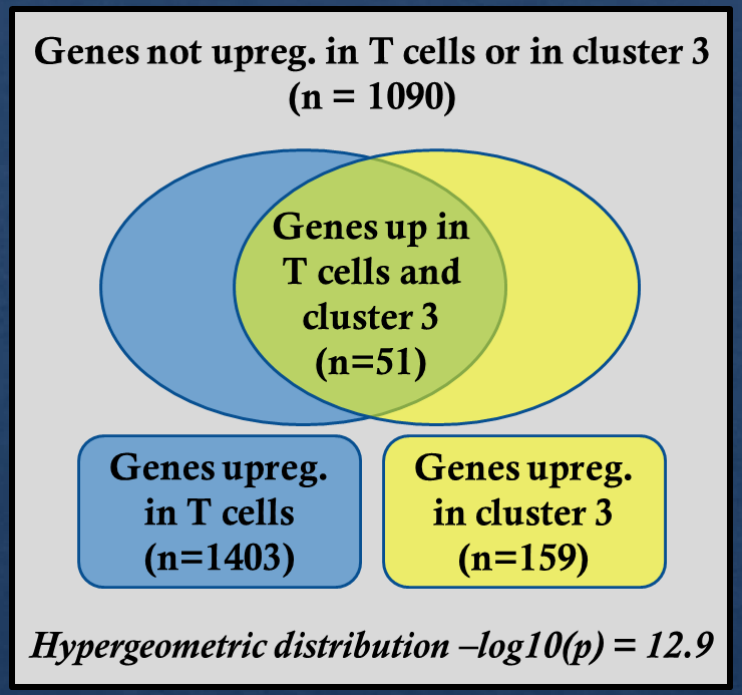
Method details: literature-based cell typing

- Evaluate each cell for expression of 49 marker genes validated in breast tissue
- "Expression" = normalized gene expression > 1
- Classify cell as a certain type if it meets at least 1 or more of these
- If no type, label "unknown"; if multiple, label "undecided" *(except: epithelial/stromal = epithelial)*

	Breast epithelial	Stromal	Endothelial	Immune types (T, B, macrophage)
Markers	***EPCAM, KRT8, KRT18, KRT19,*** EGFR, CDH1, KRT14, ITGA6, KRT5, TP63, KRT17, MME, FOXA1, GATA3, MUC1, CD24, KIT, GABRP	FAP, COL1A1, COL3A1, COL5A1, ACTA2, TAGLN, LUM, FBLN1, COL6A3, COL1A2, COL6A1, COL6A2	PECAM1, VWF, CDH5, SELE	**T cell:** CD2, CD3D, CD3E, CD8A, CD8B **B cell:** MS4A1, CD79A, CD79B, BLNK **Macrophage:** CD14, CD68, CD163, CSF1R **Pan-immune:** PTPRC
Decision rules	1. 2+ breast epithelial markers 2. 1 ***well-defined*** breast epithelial marker expressed above median level for that marker for that patient	1. 2+ stromal markers and no other markers 2. 3+ stromal markers and 0-1 endothelial marker	1. 2+ endothelial markers and no other markers 2. 3+ endothelial markers and 0-1 stromal marker	1. 2+ type-specific immune markers (T, B, macrophage) and no other immune markers 2. Pan-immune marker PTPRC, 1 type-specific immune marker, no other immune markers 3. 3+ type-specific immune markers and 0-1 other immune markers

Method details: multimodal intersection analysis (MIA)



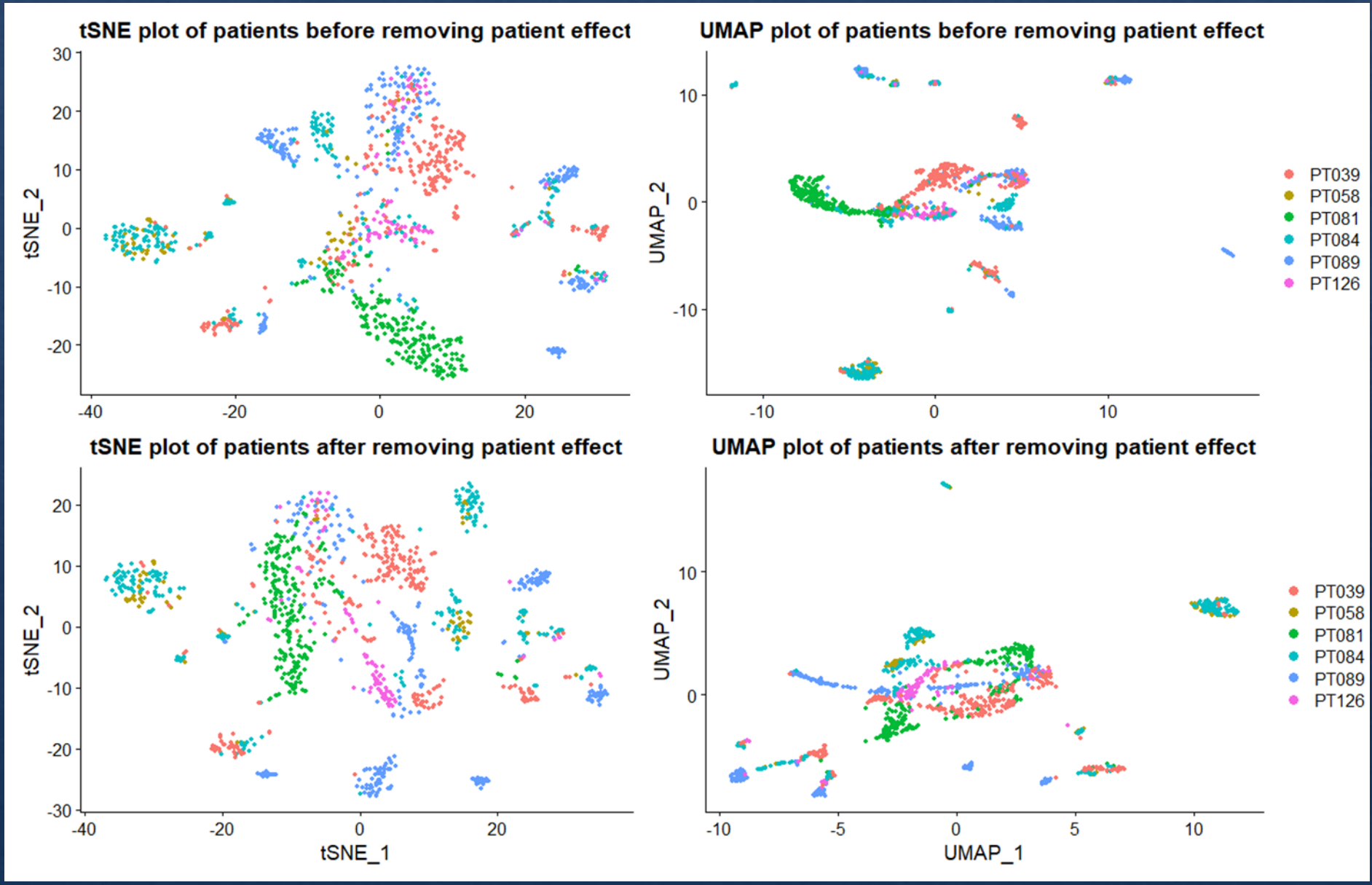


- MIA = fancy name for hypergeometric test for enrichment (Moncada et al., 2020)
- Do groups consist of overlapping cells more frequently than expected by chance?
- Do sets of differentially expressed genes overlap more than expected by chance?
- Calculate p-value for all cell-type/cluster pairs, display as heat map

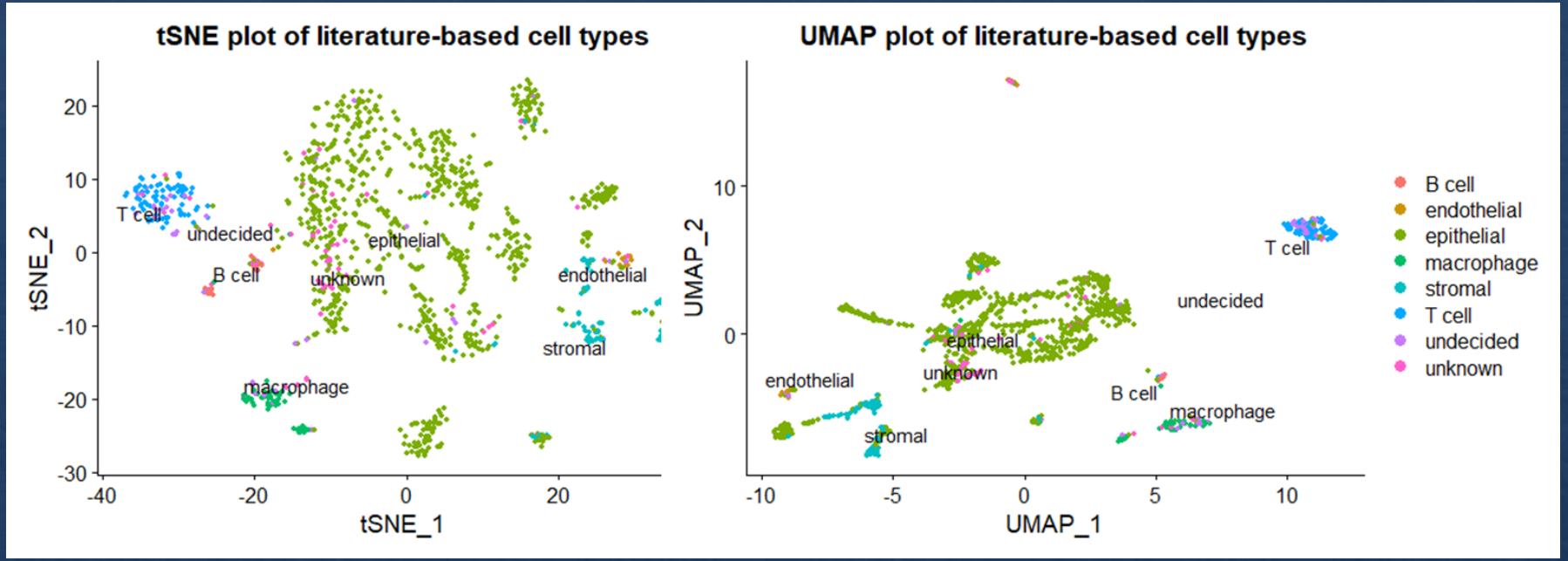
Figures adapted from Moncada et al., 2020

Results

tSNE/UMAP before and after removing patient effect

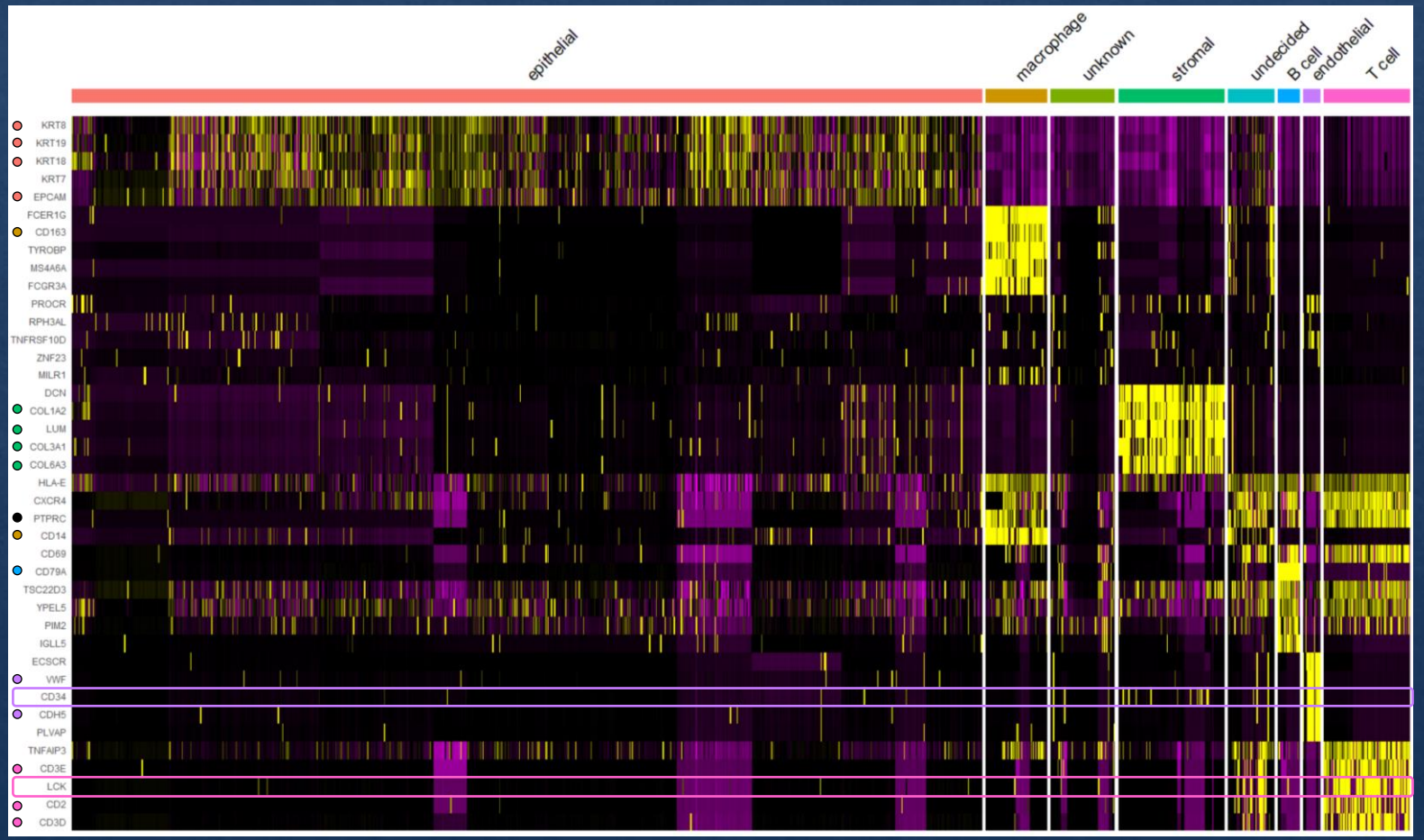


Literature-based cell typing: tSNE/UMAP



- 3 major divisions: epithelial, stromal/endothelial, immune
- Most cells are epithelial, with multiple apparent subtypes
- Other cell types tend to form fairly consistent and compact clusters
- Some epithelial cells cluster near stroma – either reflects EMT or decision rule failure
- Most undecided cells are immune, most unknown cells are epithelial

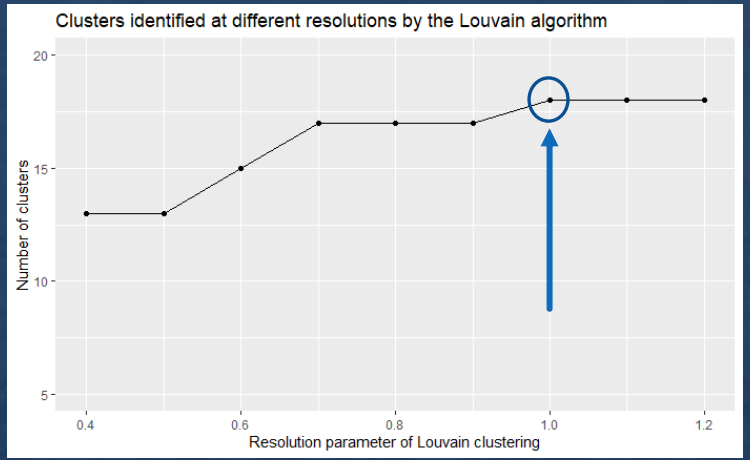
Literature-based cell typing: differentially expressed genes

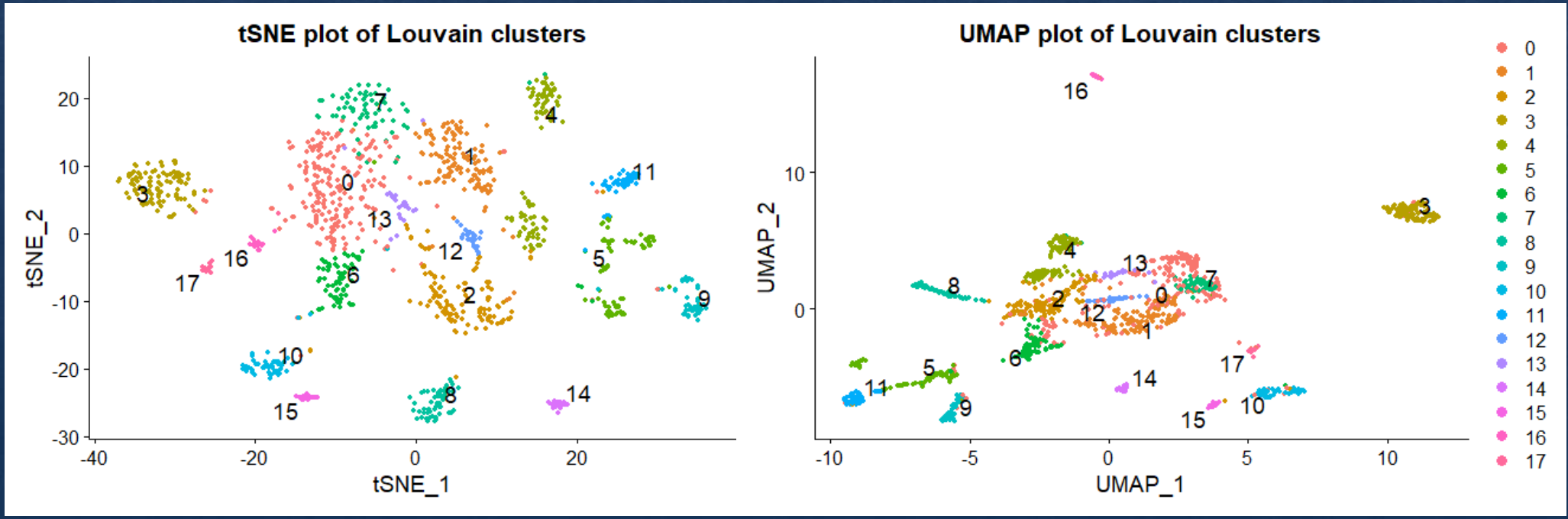


- Includes several markers used to assign cell types
- Also includes some known markers not used by decision rules (CD34, LCK, …)
- Consistent gene expression profiles suggest marker-based cell typing reflects true biological differences
- But several “undecided” cells clearly match certain cell-type-specific expression profiles

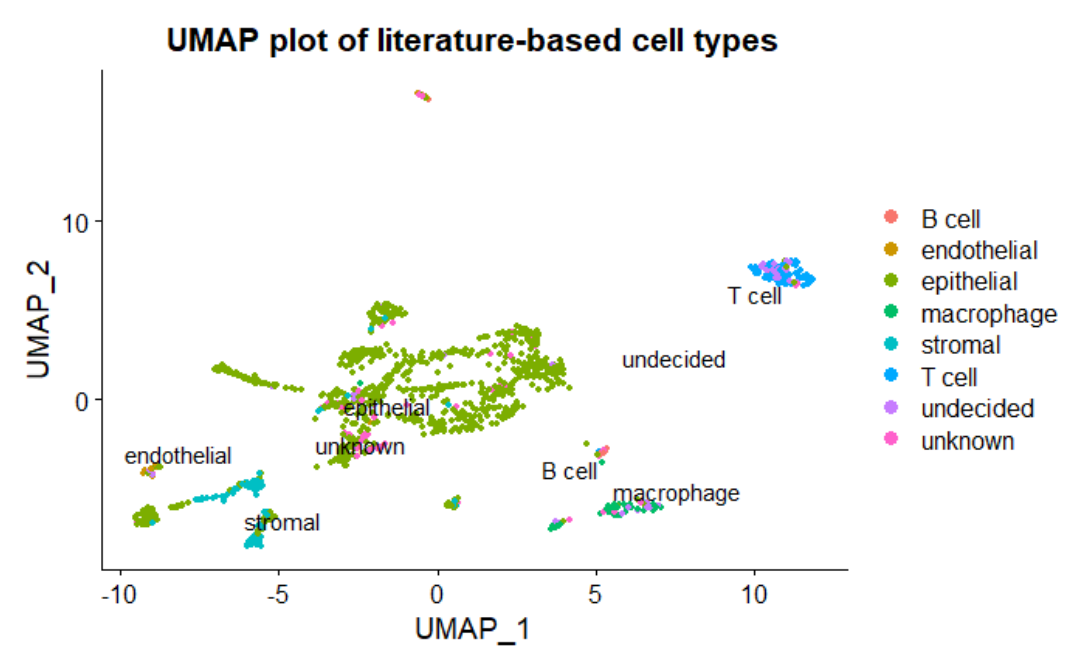
Clustering with the Louvain algorithm

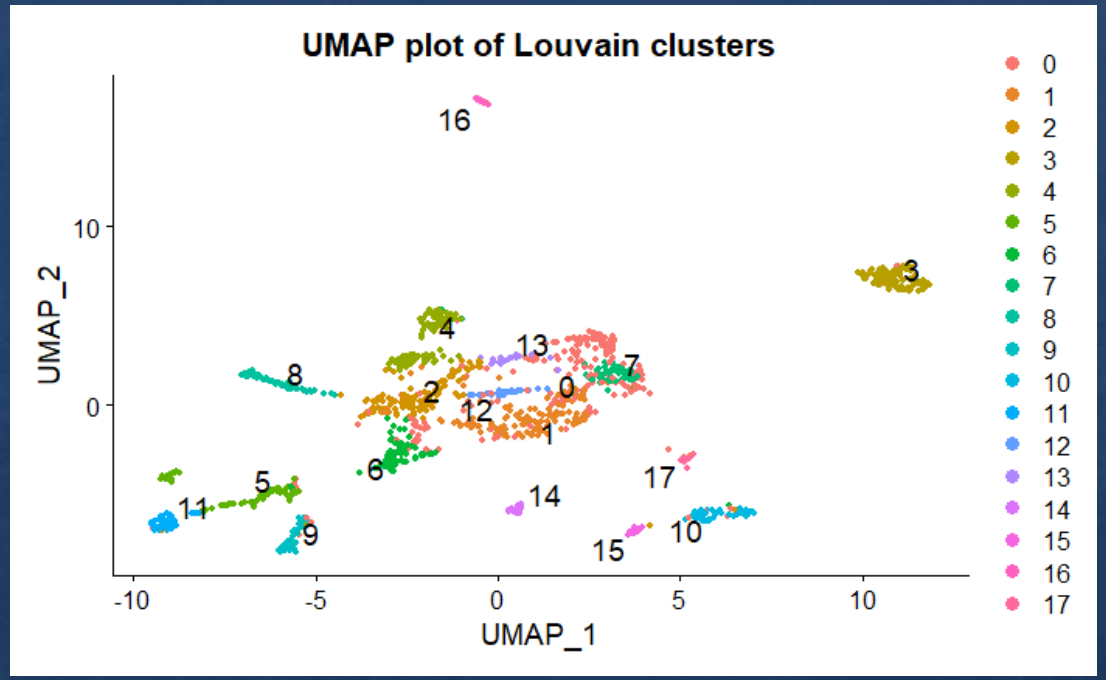
- Tested a variety of resolutions (0.4-1.2): 18 clusters at resolution 1.0
- Clusters reflect consistent visual groupings in projected data





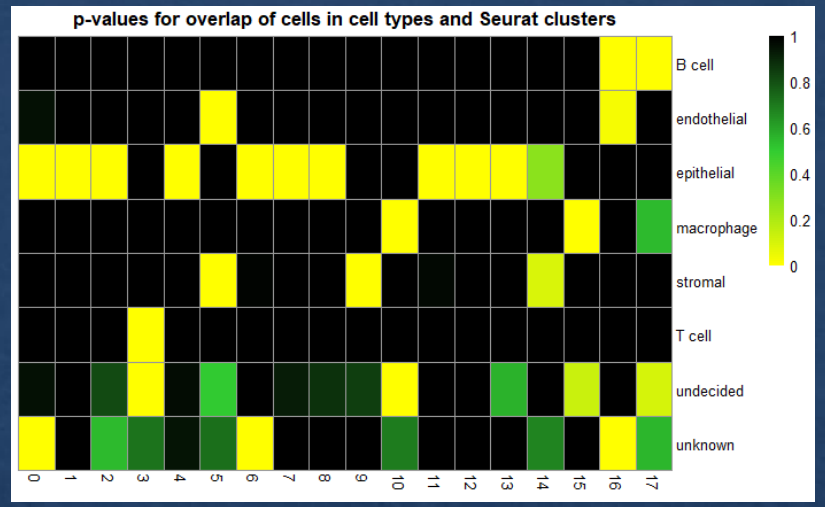
Visual comparison: cell types and Louvain clusters

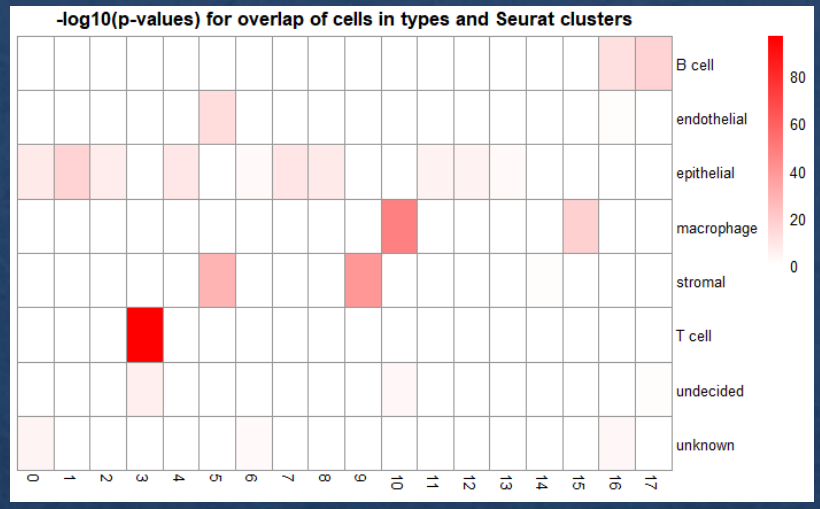


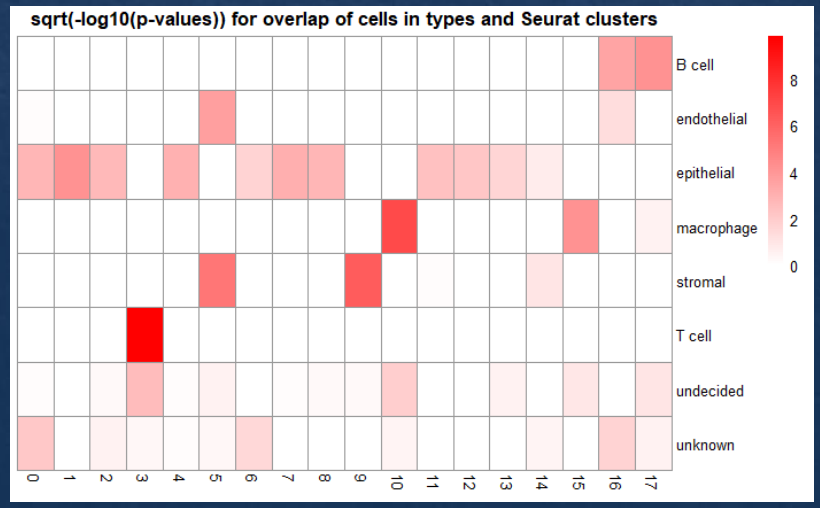


- Cell types well-represented by 1-2 clusters:
  T cells (3), B cells (16, 17), macrophages (10, 15), stroma (5, 9)
- Cell type without an independent cluster: endothelial (5 with stroma, but visually distinct)
- Numerous clusters with epithelial identity
- Undecided/unknown cells are spread throughout existing clusters, not independent

MIA for cells: cell type and Seurat cluster

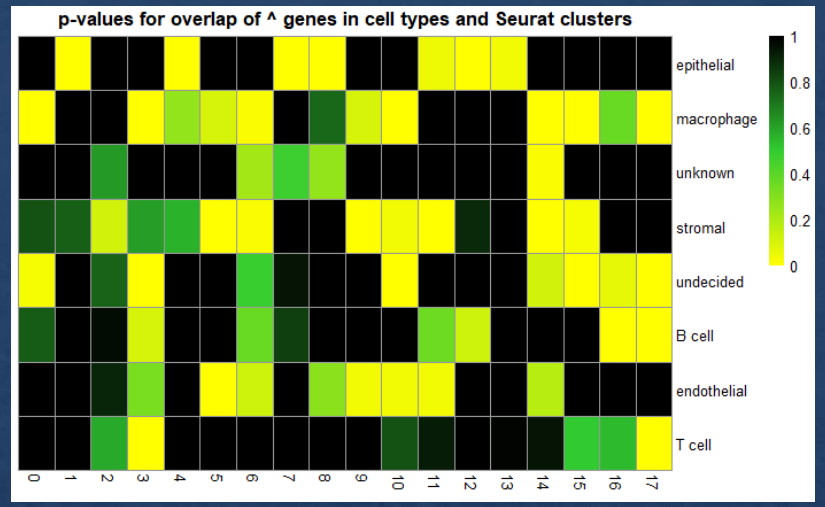


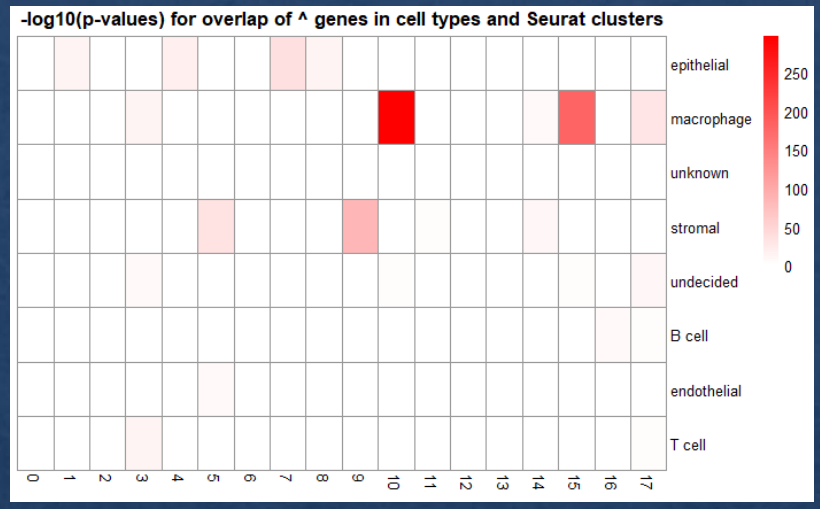




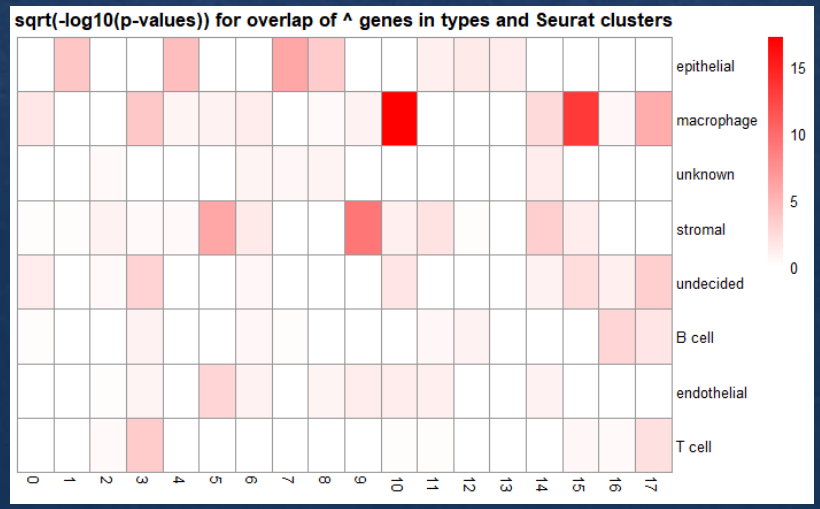
- NOTE: only looking at enrichment of cells
- Certain cell types are statistically enriched in certain clusters, matching visual inspection
- Overlap is most significant for T cells, macrophages, stroma, B cells
- Less enrichment for epithelial clusters – likely reflects multi-cluster split of epithelial cells

MIA for upregulated genes: cell type and Seurat cluster





- NOTE: row order differs from last slide (will fix in longer term)
- NOTE: only looking at upregulated genes and only looking at enrichment
- Similar gene sets are enriched in some cell types and some clusters, reflecting a shared biological identity
- Most coherent gene sets: macrophage and 10, 15



Discussion

Cell typing with literature-based markers

Pros

- Leverages existing knowledge of cell biology
- Can separate cells into biologically meaningful cell types with similar gene expression patterns (if markers are well-selected)
- Groups have clear biological interpretability
- Can define arbitrarily complex rules to separate out cell subtypes

Cons

- Can overlook cells with highly similar expression profiles that happen not to express the chosen markers (heterogeneity, technical dropout)
- Unclassified or multiply-classified cells are not easily interpreted
- Can obscure differences between subtypes of larger cell groups
- Subjectivity and bias in defining robust markers and rational decision rules
- Markers may be inconsistent in different biological conditions

Cell typing with Louvain clustering

Pros

- Does not depend on existing biological knowledge
- Can separate cells into biologically meaningful cell types with similar gene expression patterns
- No unclassified cells
- Can identify cell subtypes within larger clusters
- More independent of researcher bias

Cons

- Does not leverage existing knowledge
- Not immediately interpretable, requires biological knowledge to annotate the clusters (subject to selective inference)
- No clear rules for deciding how many clusters to generate from the data – final clusters may not be parsimonious

Can we find a better way?

- Probably, especially if annotated reference datasets exist to train classifiers
- Numerous supervised cell typing methods (scmap, scVI, SingleCellNet, …) train various algorithms to assign cells to types, but performance varies by dataset
  - Extensive benchmarking has been performed (e.g. Abdelaal et al., 2019)
  - To benchmark methods, ground truth cell type must be known
- Approach is unclear when annotated reference datasets do not exist, which is currently true in several tissues and contexts
- Literature-based methods and unsupervised clustering have complementary strengths and weaknesses – can they be combined?
  - Probably would increase performance
  - But difficult or impossible to assess performance with accuracy/precision/recall when true annotations are unknown

Future Directions

Short term future directions: extend to additional unsupervised clustering methods

- Options
  - Graph-based methods: Leiden, infomap
  - Distance-based methods: Hierarchical clustering with different linkages
  - Centroid-based methods: k-means
  - Density-based methods: DBSCAN, mean shift
- Workflow
  - Apply clustering, label cells, find differentially expressed genes
  - Use MIA to compare to literature-based cell types and to previous clustering results

Longer term future directions

- Find supervised cell typing methods that can be performed with existing databases
  - Caveat: currently human breast tissue is not in the Human Cell Atlas or other databases
  - Mouse breast tissue has been analyzed
  - Some methods (e.g. SingleCellNet) can perform cross-species classification
- Experiment with ensemble methods that combine different unsupervised (and possibly supervised) clustering approaches
  - Would need to switch/expand to a dataset with high confidence cell type annotations to assess performance
- Decide on "best" cell type classifications and perform downstream single cell analysis
  - Infer cell-cell interactions important to these tumors
  - Use to build mechanistic models of tumor ecosystem signaling

References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., & Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, *20*(1), 194. https://doi.org/10.1186/s13059-019-1795-z

*** Karaayvaz, M., Cristea, S., Gillespie, S. M., Patel, A. P., Mylvaganam, R., Luo, C. C., Specht, M. C., Bernstein, B. E., Michor, F., & Ellisen, L. W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nature Communications*, *9*(1), 3588. https://doi.org/10.1038/s41467-018-06052-0

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., & Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, *38*(3), 333–342. https://doi.org/10.1038/s41587-019-0392-8

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Acknowledgments

- Alexis Battle
- Ashton Omdahl
- Feilim Mac Gabhann
- And you!

Thank you!

Questions?