Investigating gene expression patterns across different neuroblastoma conditions using single-cell RNA-seq (scRNA-seq) data

Sangeevan Vellappan, Jingyun Qiu
Erdős Institute
Data Science Boot Camp
Fall 2024

Motivation

scRNA-seq is essential for neuroblastoma research

- Neuroblastoma is a highly heterogeneous cancer that accounts for 15% of all pediatric cancer death.
- scRNA-seq advances neuroblastoma research by:
  - Characterizing tumor heterogeneity.
  - Profiling tumor microenvironment.
  - Revealing gene expression patterns across different cell types under different conditions

Our project

- Reanalyzed publicly available scRNA-seq data from neuroblastoma-infiltrated bone marrow.
  - Leveraging machine learning for deeper insights.
  - Integrating domain expertise for contextualized predictions.
  - Enhancing reproducibility through comprehensive pipelines.
  - Detailed rationale for preprocessing and cleanup.

Data Gathering

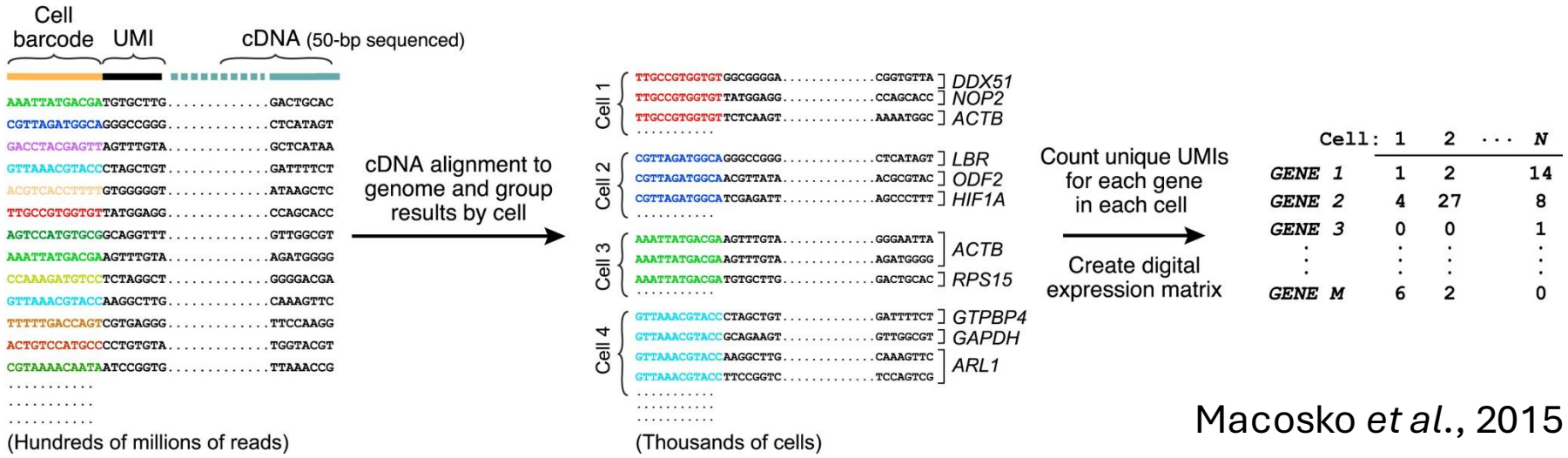
Data source

The scRNA-seq dataset for this project is obtained from a paper titled “Single-cell transcriptomics and epigenomics unravel the role of monocytes in neuroblastoma bone marrow metastasis.”

- **GEO accession (GSE216176)**
  - Healthy control
  - MYCN-amplified tumor samples

Data structure

scRNA-seq data matrix



Macosko *et al.*, 2015

Data Analysis Pipeline

- **Quality control**
- **Data distribution**
- **Data cleaning**
- **Data normalization**

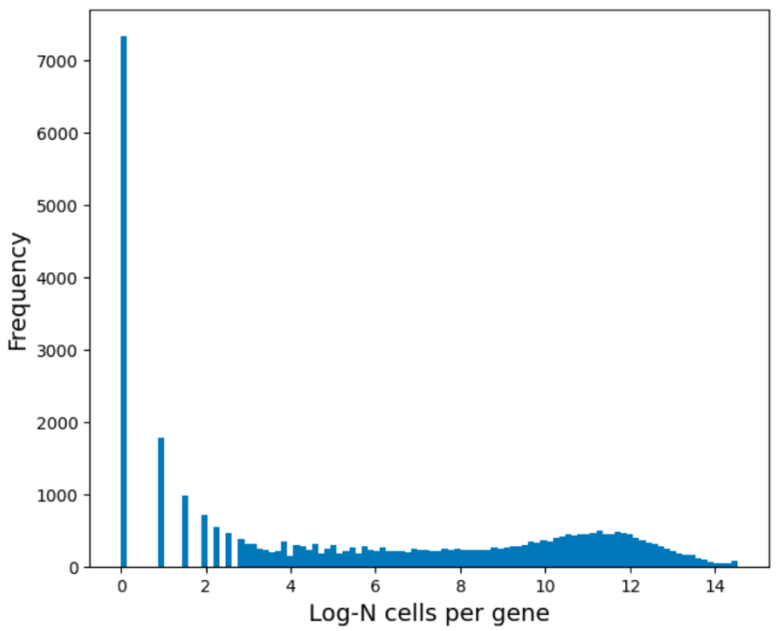
- **PCA (Principal Component Analysis)**
- **UMAP (Uniform Manifold Approximation and Project)**

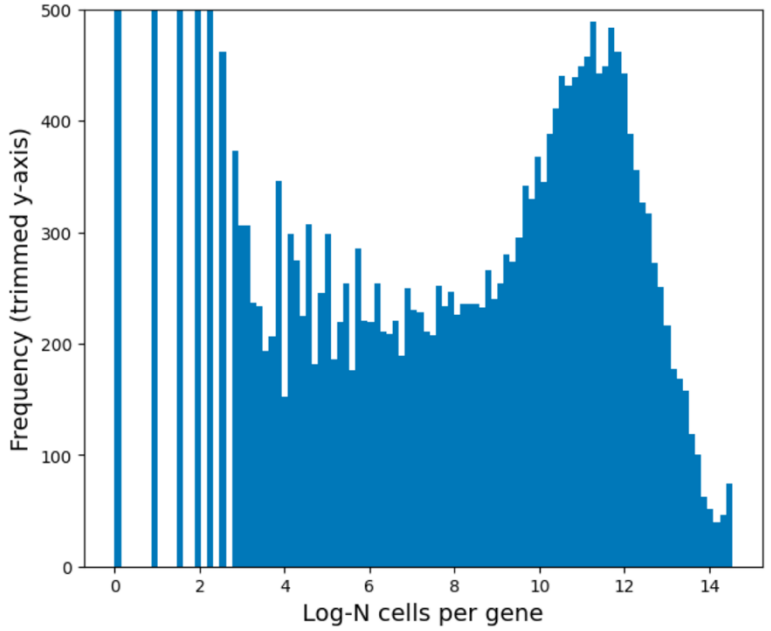
- **UMAP**

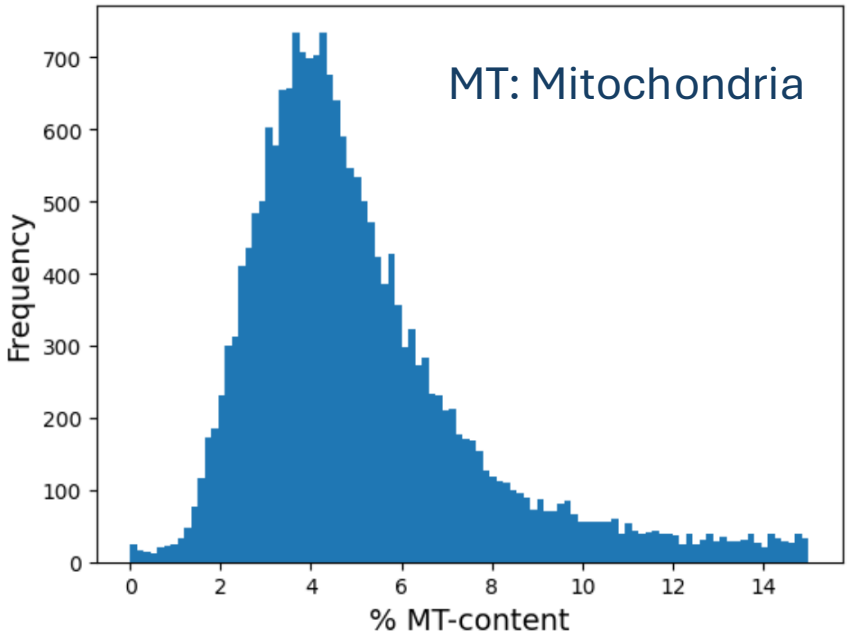
- **Phenograph clustering**
- **CellTypist**
- **Diffusion maps**

For further details on data analysis pipeline, please check Github_Repository

Preprocessing







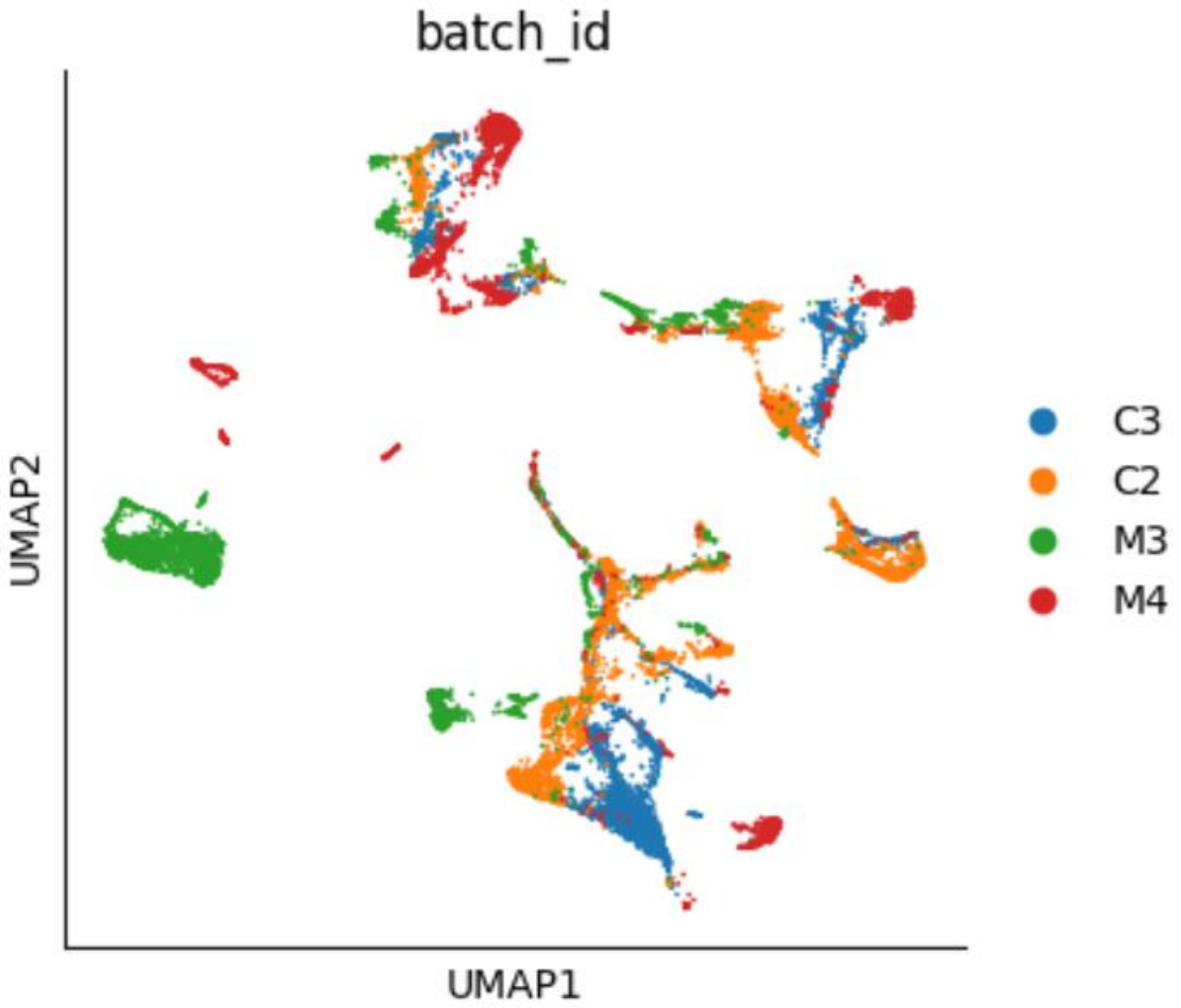
The distribution plots help assess the spread of gene expression across cells and help identify an appropriate threshold for filtering out low-expressed genes.

The plot confirms the removal of unhealthy cells with MT content greater than 15%.

Dimensionality Reduction and Visualization

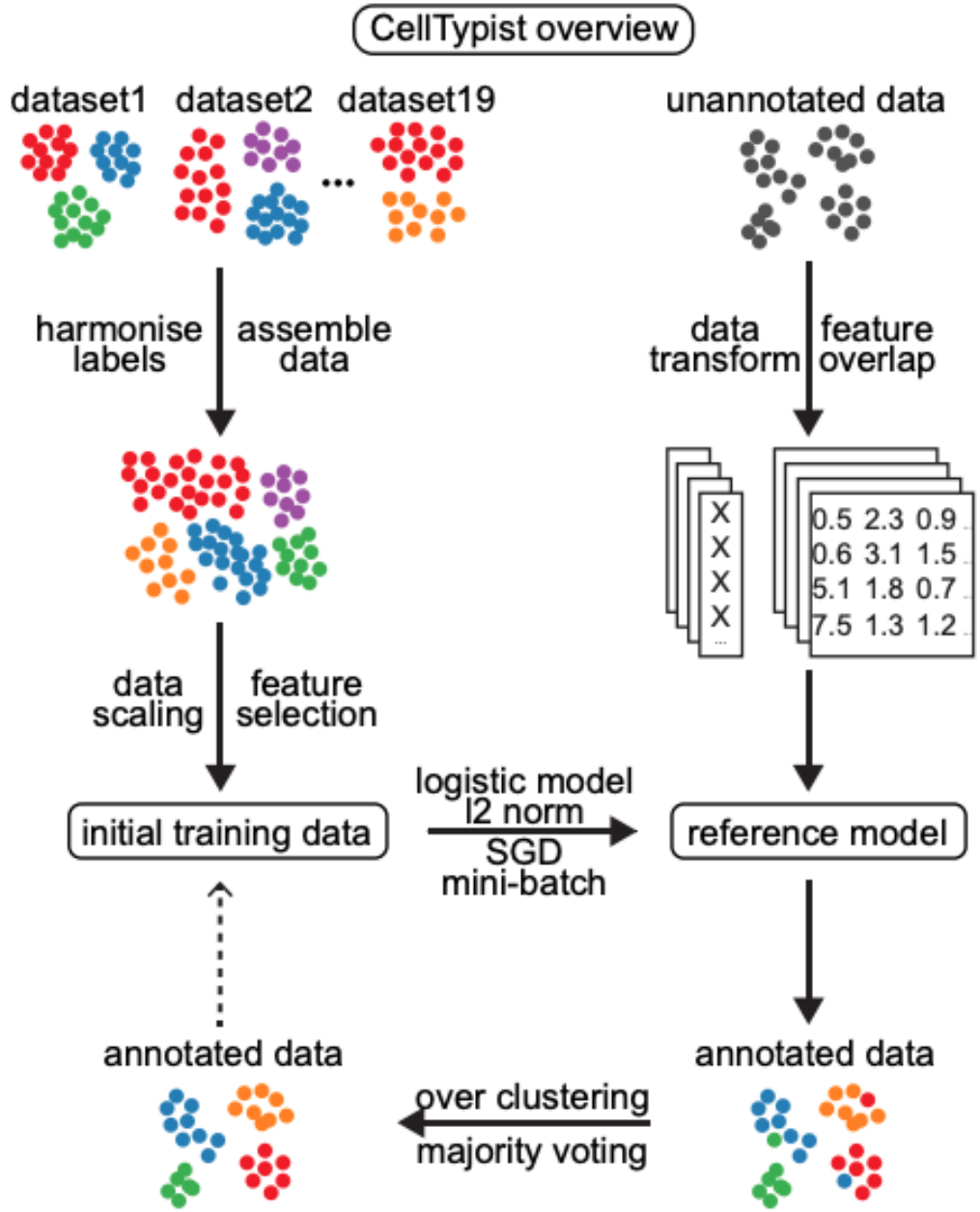
UMAP helps distinguish tumor cells from healthy cells.

- Healthy cells: C2, C3
- Tumor cells: M3, M4



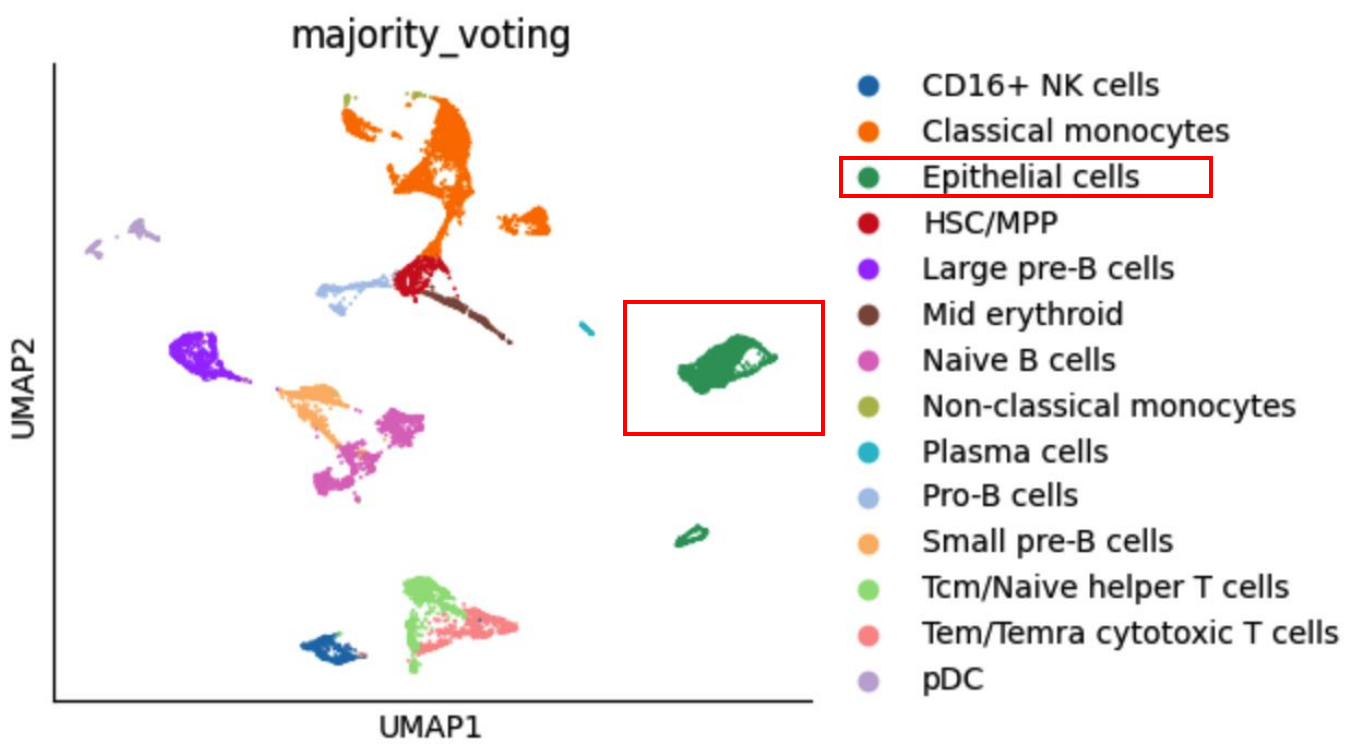
Cluster annotation to know what are the cell types (supervised model-Celltypist)

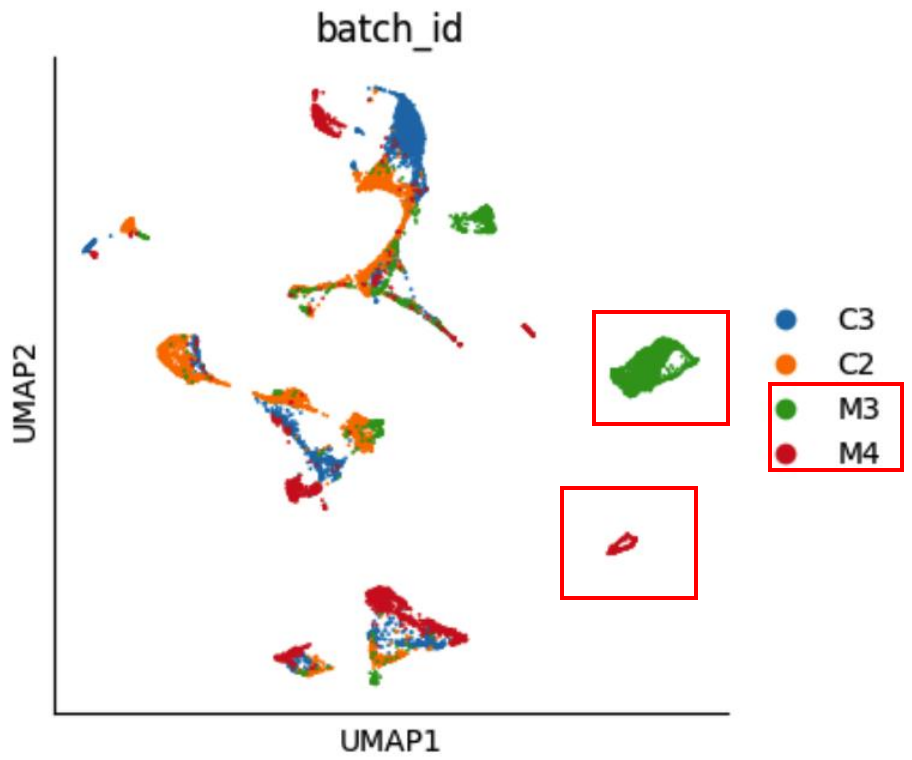
- Logistic Regression with Stochastic gradient descent (SGD)
- L2 regularization/ Ridge regularization
- Train model, select features, train model again
- Using CellTypist models to annotate gives you two different kinds of labels and a confidence score:
  - **Option 1: Most confident cell type**
  - Option 2: A list of possible cell types with confidence level
- Use of Immune cell markers for our purpose



Celltypist – annotate at a cluster level

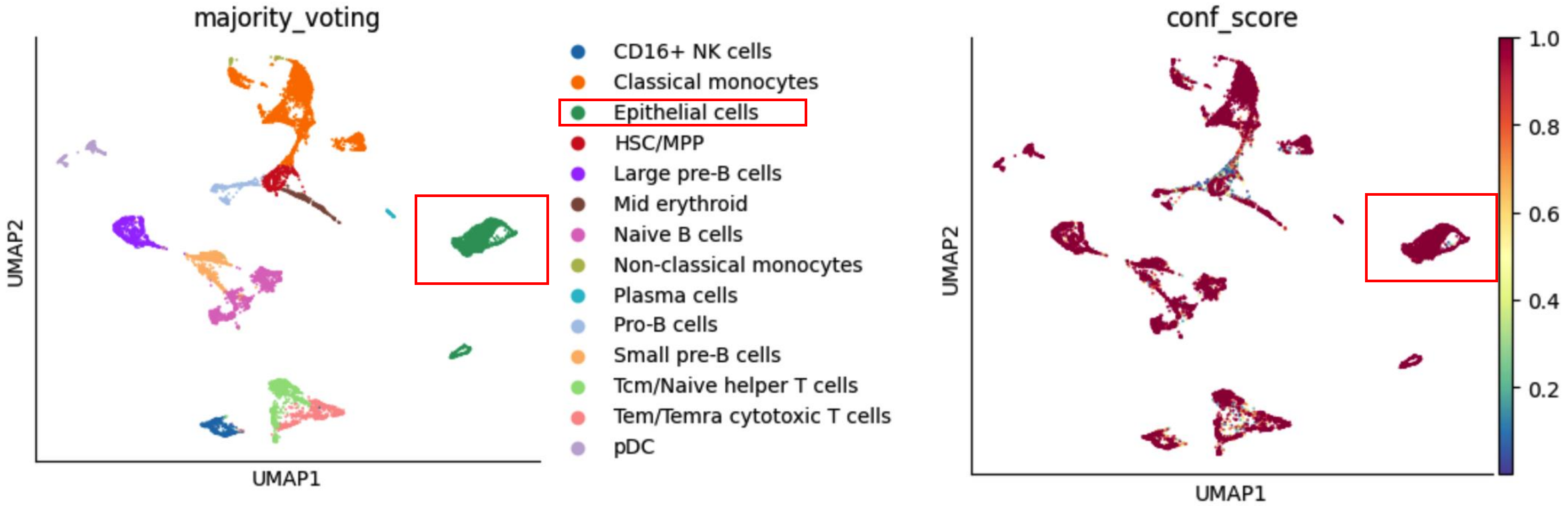
Epithelial cells are non-tumor/cancer cells



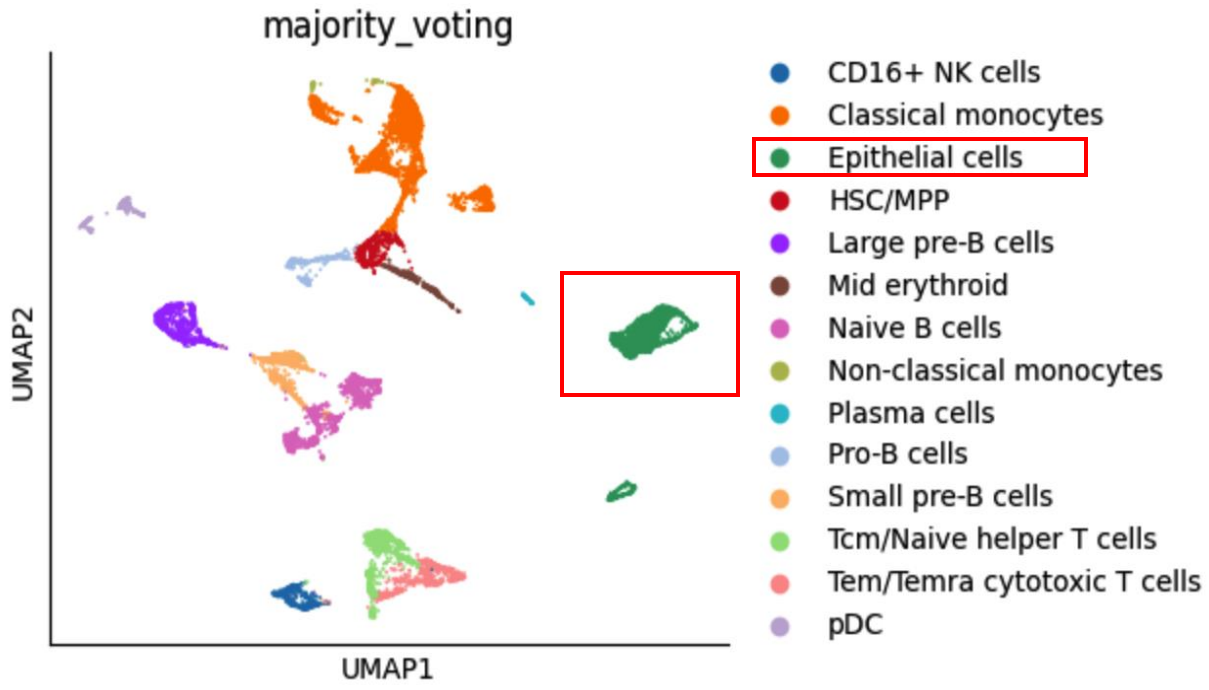


M3 & M4 are cancer/tumor samples

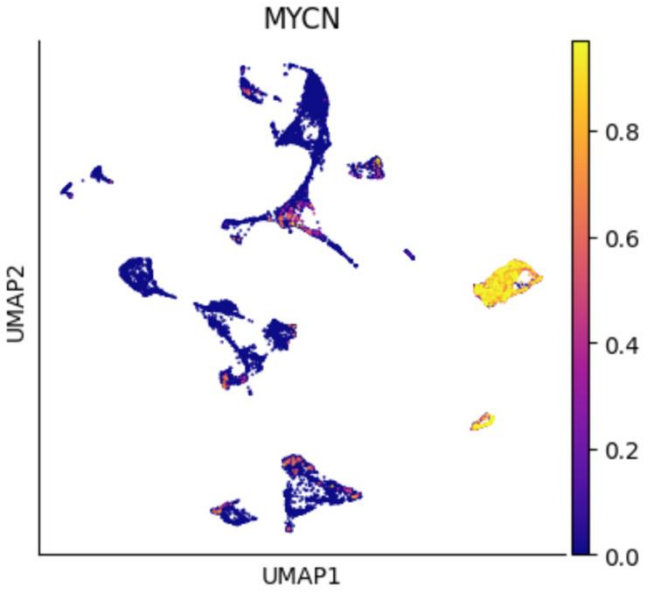
Quantitative validation of the prediction-confidence score



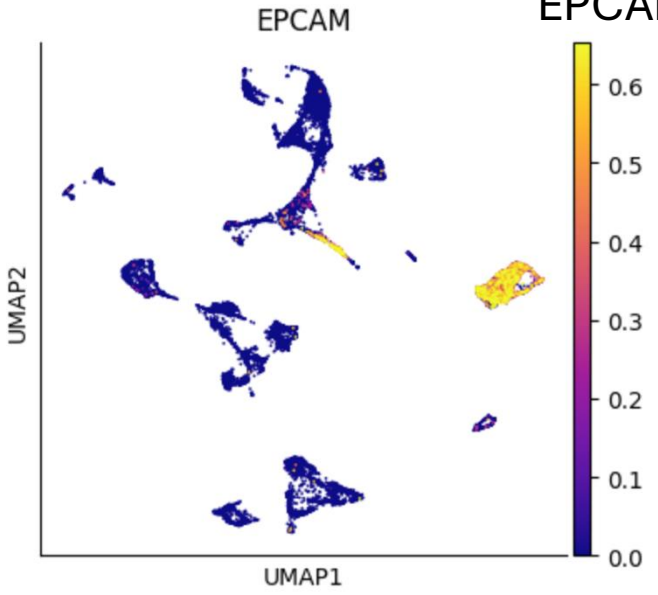
Biological validation of the prediction



MYCN- tumor gene



EPCAM- Epithelial gene



Strengths/weakness/implication of the approach

- Strengths:
  - high confidence scores for most clusters
  - biological validation confirms the prediction

- Weakness:
  - Reference datasets are limited (the field is expanding)

- Implication:
  - This approach uncovers tumor-epithelial interactions within tumor clusters, providing critical insights for therapies targeting tumor plasticity and microenvironment dynamics.

- Next step:
  - Targeted therapy that could disrupt tumor-epithelial interactions