Erdős Data Science Bootcamp Fall 2024 Executive Summary

Team single-cell-rnaseq: Jingyun Qiu, Sangeevan Vellappan
GitHub: https://github.com/jqiu1/erdos-single-cell-rnaseq

Overview:

Single-cell RNA sequencing (scRNA-seq) enables researchers to explore tumor heterogeneity and cellular diversity at an unprecedented resolution. By combining scRNA-seq with advanced machine learning techniques, this approach provides deeper insights into disease mechanisms, identifies diagnostic markers, and opens avenues for developing targeted therapies.

In this project, we use neuroblastoma—a highly heterogeneous pediatric cancer—as a case study. Through scRNA-seq data and machine learning tools, we aim to understand how gene expression patterns vary across healthy and diseased cell types. This approach not only sheds light on the biology of neuroblastoma but also provides a framework for identifying potential therapeutic interventions.

Objectives:

1. Analyze scRNA-seq data to identify gene expression signatures and cellular heterogeneity in neuroblastoma using methods like normalization, highly variable gene selection, PCA, and UMAP for clustering and visualization.
2. Develop a reproducible analysis pipeline incorporating advanced tools like Scanpy, robust preprocessing, and machine learning methods to identify tumor-specific gene expression patterns:
   - Scrublet for doublet detection to improve dataset quality.
   - Diffusion maps for cellular trajectory analysis.
   - CellTypist for automated cell-type annotation.
3. Priorities:
   - Leverage machine learning to uncover and visualize non-linear biological relationships among individual cells, providing deeper insights into tumor heterogeneity.
   - Provide actionable insights for neuroblastoma diagnostics and therapeutics.
   - Share a transparent, reproducible pipeline for public scRNA-seq datasets.

Methodology:

1. Dataset pre-processing:
   - Quality control: Filtered cells based on the number of genes expressed, total counts, and mitochondrial gene content.
   - Normalization: Used median library size normalization to reduce sequencing depth bias.
   - Feature selection: Identified highly variable genes to focus on biologically relevant patterns.
   - Dimensionality reduction: Applied PCA and UMAP for clustering and visualization.

2. Automated annotation and quality assessment:
   - Scrublet: Detected and removed doublets to ensure dataset integrity.
   - CellTypist: Annotated cell types using immune-focused pretrained models, validated through known marker genes like MYCN (associated with neuroblastoma).

3. Evaluation metrics:
   - Visualization-based validation: Cluster separation and gene expression patterns were evaluated through UMAPs and diffusion maps to ensure biological consistency.

- Marker gene analysis: Differential expression of known markers validated cluster identity and biological relevance.

4. Pipeline evaluation:
   - Accurately separate healthy and tumor cells.
   - Highlight biologically meaningful gene expression signatures.

Results:

1. Doublet detection:
   - Doublets were effectively identified and removed using Scrublet, ensuring cleaner data and improved clustering accuracy.

2. Dimensionality reduction and clustering:
   - PCA was used to reduce the dimensionality of the dataset, capturing significant patterns while minimizing noise.
   - UMAP visualization revealed distinct clustering of healthy control and tumor cells.

3. Cellular trajectory analysis:
   - Diffusion maps 0 and 1 captured cellular transitions and trajectories, highlighting the progression from healthy to tumor states.
   - These insights are critical for understanding the mechanisms of tumor evolution and identifying potential intervention points.

4. Automated cell typing and annotation:
   - Cell typing using CellTypist successfully classified major cell types, although the limited diversity in reference datasets posed challenges in identifying rare or neuroblastoma-specific cell populations.
   - Despite these limitations, marker gene analysis validated the annotations (described below).

5. Marker gene analysis:
   - MYCN: Highly expressed in tumor clusters, consistent with its role in MYCN-amplified neuroblastoma.
   - EPCAM: Unexpectedly expressed in tumor clusters (identified through automated cell typing/prediction), suggesting partial epithelial-to-mesenchymal transition (EMT) of the tumor cell population.

Conclusion:

Conclusion: Based on our analysis, we see clear expression patterns of MYCN and EPCAM across cell clusters. MYCN, a key marker of neuroblastoma, is highly expressed in clusters corresponding to MYCN-amplified samples (AKA tumor samples), as expected. Interestingly, EPCAM, an epithelial cell marker (not typically known to be associated with tumor), is also expressed in these clusters. This suggests that neuroblastoma cells may exhibit epithelial-like traits, potentially due to interactions with the tumor microenvironment or cellular plasticity, such as partial epithelial-to-mesenchymal transition (EMT). These findings offer new insights into tumor biology and raise critical questions:

- Treatment implications: Tumors with mixed phenotypes may respond to therapies targeting both MYCN-driven proliferation and epithelial signaling pathways.

- Drug discovery: The mixed phenotype opens up potential therapeutic opportunities, such as targeting MYCN-driven tumor growth or disrupting the interplay between neuroblastoma and epithelial-like cells.

Future direction:

Further research is needed to determine whether the co-expression of MYCN and EPCAM reflects early tumor progression or advanced cellular plasticity. This investigation will deepen our understanding of neuroblastoma evolution and inform therapeutic strategies.

Limitations:

Our machine learning-based predictions were constrained by the limited diversity of cell types in the reference database, which may lead to misclassifications or insufficient resolution for rare or disease-specific cell populations, such as those in neuroblastoma.