Enhanced Analysis of Gastric Cancer Metastasis Using Clustering Techniques and Single-Cell RNA Sequencing

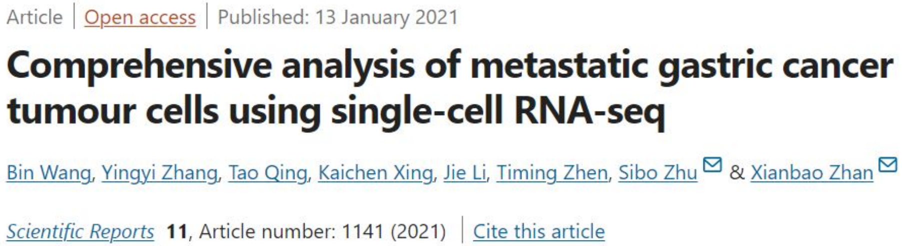
Final Project for the Machine Learning in Computational Biology Course

Konstantinos Giatras, 7115152300005
Olympia Tsiomou, 7115152200035

Professor: Elias Manolakos

DSIT 2023-2024

Overview of the Paper



Article | Open access | Published: 13 January 2021

Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq

Bin Wang, Yingyi Zhang, Tao Qing, Kaichen Xing, Jie Li, Timing Zhen, Sibo Zhu ✉ & Xianbao Zhan ✉

Scientific Reports **11**, Article number: 1141 (2021) | Cite this article

- **Purpose**: To investigate the mechanisms underlying gastric carcinoma lymph node metastasis using single-cell RNA sequencing (scRNA-seq).

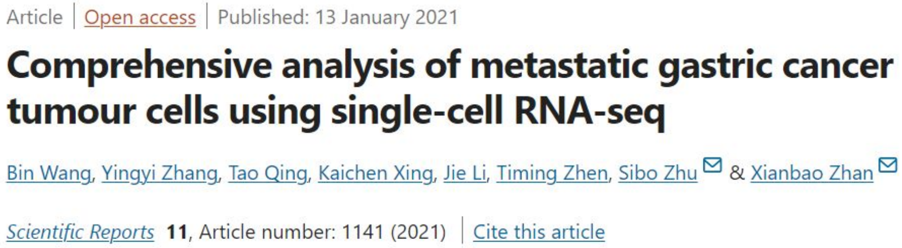
- **Importance**: Gastric cancer has a poor prognosis, particularly with metastasis, and traditional bulk-level analyses have failed to uncover the roles of cellular subpopulations in this process.

- **Experimental Design**:
  - **Sample Collection**: Primary gastric cancer and paired metastatic lymph node tissues were obtained from three patients, each at a different pathological stage and with different primary tumor sites.
  - **Single-Cell Isolation**: Tumor tissues were digested, and single cells were isolated.
  - **Sequencing**: Single-cell libraries were prepared using the Smart-seq2 protocol and sequenced on a HiSeq 2500 instrument.

Overview of the Paper

- **scRNA-seq Data Analysis**:
  - **Data Preprocessing**: Quality control & Normalization.
  - **Dimensionality Reduction**: PCA.
  - **Clustering**: Hierarchical clustering.
  - **Visualization**: t-SNE.

- **Results**:
  - Four main clusters were identified, showing significant intratumoral heterogeneity.
  - Differentially expressed marker genes were identified, both in primary and metastatic cancer tissue.



Article | Open access | Published: 13 January 2021

Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq

Bin Wang, Yingyi Zhang, Tao Qing, Kaichen Xing, Jie Li, Timing Zhen, Sibo Zhu ✉ & Xianbao Zhan ✉

Scientific Reports **11**, Article number: 1141 (2021) | Cite this article

- **Post Clustering Analysis**:
  - **Functional Annotation**: Performed using Gene Ontology (GO) analysis to understand the biological processes and pathways involved.
  - **Pseudotime Trajectory Analysis**: Conducted using TSCAN and Monocle, tracing the evolutionary states of cancer cells from primary to metastatic stages.

The Scanpy Library

- **Definition**: Scanpy (Single-Cell Analysis in Python) is an open-source Python library designed for the analysis and visualization of scRNA-seq data.

- **Functionality**: It offers a comprehensive suite of tools for preprocessing, normalization, dimensionality reduction and clustering.

- **Integration**: It is built on top of the AnnData data structure (designed to store multiple layers of annotations and metadata about cells and genes), and it seamlessly integrates with other scientific Python libraries like NumPy and Pandas.

- **Visualization**: It includes advanced plotting functions to create publication-quality visualizations, such as t-SNE, UMAP and heatmaps, enabling the exploration and interpretation of single-cell datasets.

The Dataset

- **Count Matrix**: represents the number of reads (or transcripts) that map to each gene in each cell (after data preprocessing).

- The raw counts, which are measures of gene expression, are *normalized to transcripts per million (TPM)* and *log2-transformed*.

- Dimensions: 21196 rows (genes) x 94 columns (cells).

- Number of cells from each patient (by tissue type):

	Patient 1 (PT1)	**Patient 2 (PT2)**	**Patient 3 (PT3)**
Tumor Tissue (TT)	19	27	49
Lymph Node (LN)	4	13	12

Source: Wang, B. et. al., (2021). Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq. Scientific Reports, 11, 11585. https://doi.org/10.1038/s41598-020-80881-2

Data Preprocessing

Authors' Approach

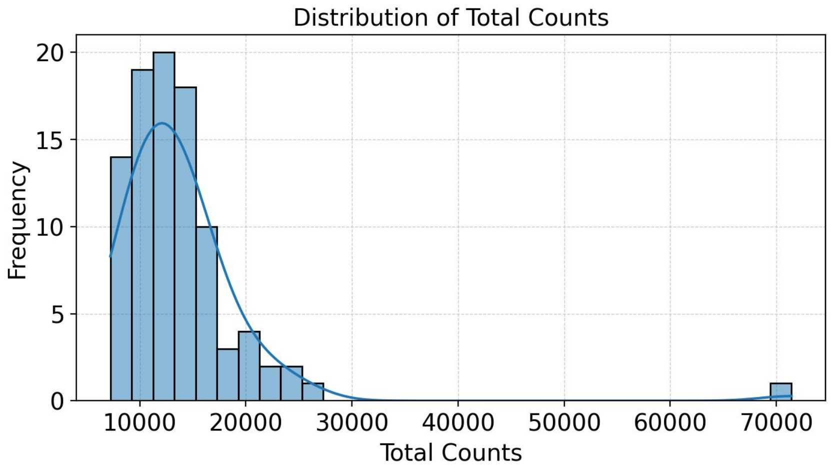
- **Trimmomatic**: Trimming sequencing adapters and removing low-quality reads (Phred < 20).
- **HiSat2**: Mapping reads to the human genome reference (UCSC hg19).
- **FeatureCounts**: Calculating expression levels of genes, producing raw count values for each gene in each cell.
- **scater (R package)**: Normalizing read counts to TPM (transcripts per million) and performing log2 transformation → count matrix.
- **Additional filtering**: Average number of reads mapped to a gene must be >1 across all cells.

Our Approach

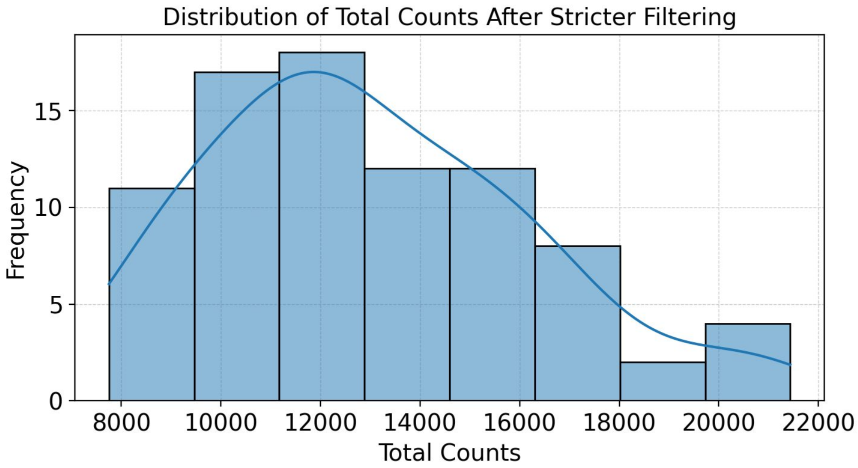
- **Plan**: apply additional QC techniques, such as doublet detection, mitochondrial content analysis and cell cycle effect adjustment
- **Implementation**:
  - Created an AnnData object from the QC-filtered and normalized count matrix.
  - Applied the same additional gene filter → 4174 out of 21196 genes passed.
  - Checked for mitochondrial gene expression (none detected).
  - Quantile Filtering: Removal of cells with abnormally high or low counts (outliers).
  - Our final preprocessed dataset: 4174 genes x 84 cells.

Data Preprocessing

Before Quantile Filtering



After Quantile Filtering



- Extreme outliers with very high or very low total counts are removed from the dataset. This helps in reducing the noise and potential artifacts in the data.
- We also created 3 additional columns in the AnnData object: ‘patient’, ‘tissue’ and ‘cell_type’, to use for further analysis.

Dimensionality Reduction

Authors' Approach

- **Principal Component Analysis (PCA)** was used for dimensionality reduction with the 'prcomp' R function.

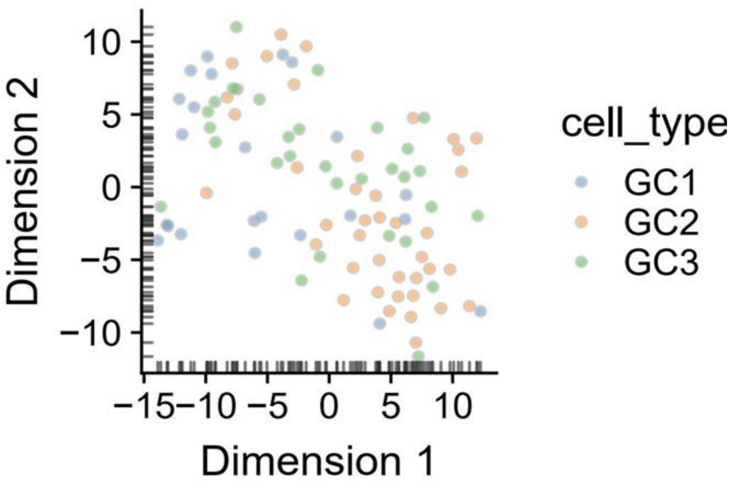
- **12 principal components** were extracted and used for further analysis (no mention of total explained variance).

Our Approach

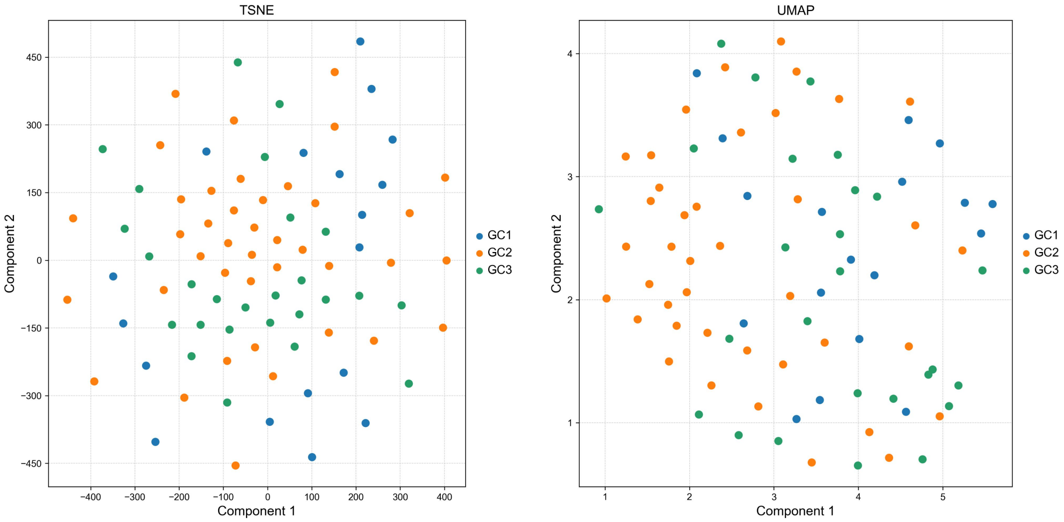
- **Plan**: Use PCA, t-SNE and UMAP.
- **Implementation**:
  - **PCA**: Determined optimal number of principal components for further use (range 39-42, optimal: 39)
  - **t-SNE and UMAP**: optimized the parameters of both methods to best represent out dataset, based on PCA's reduction
  - t-SNE parameters: number of PCs (42), perplexity (50)
  - UMAP parameters: number of neighbors (20), number of PCs (40), spread (1.5) and min distance (0.1)
  - All 3 methods for plotting before clustering

Dimensionality Reduction

Authors' Approach



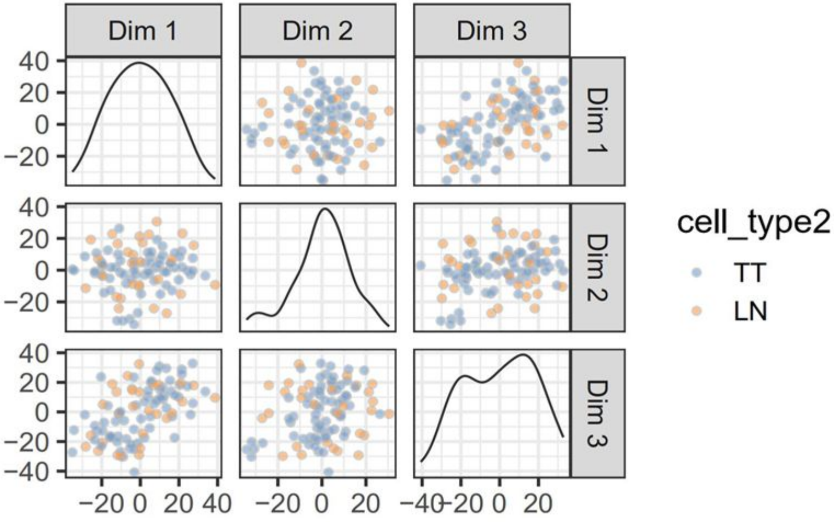
Our Approach



Source: Wang, B. et. al., (2021). Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq. Scientific Reports, 11, 11585. https://doi.org/10.1038/s41598-020-80881-2

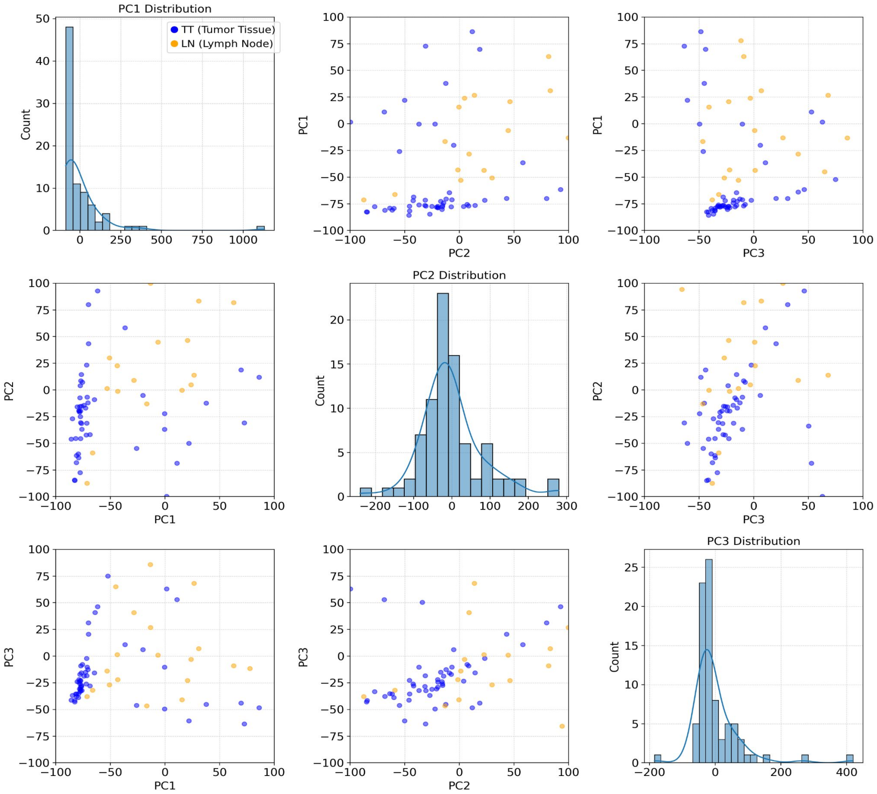
Dimensionality Reduction

Authors’ Approach



Source: Wang, B. et. al., (2021). Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq. Scientific Reports, 11, 11585. https://doi.org/10.1038/s41598-020-80881-2

Our Approach



Clustering

Authors' Approach

- **Hierarchical clustering** analysis was employed using the 'cluster' R function.
- No specific clustering algorithm was mentioned any further.

Our Approach

- **Plan**: Try different clustering algorithms, including GMMs, spectral clustering or DBSCAN
- **Implementation**:
  - We employed **Gaussian Mixture Models (GMMs)**, **average linkage** and **Ward's algorithm** (agglomerative clustering methods), **spectral clustering** and the **Louvain** and **Leiden algorithms** (graph-based clustering)
  - We run these algorithms for each dimensionality reduction method (PCA, t-SNE, UMAP)

Clustering

Our Implementation (continued):

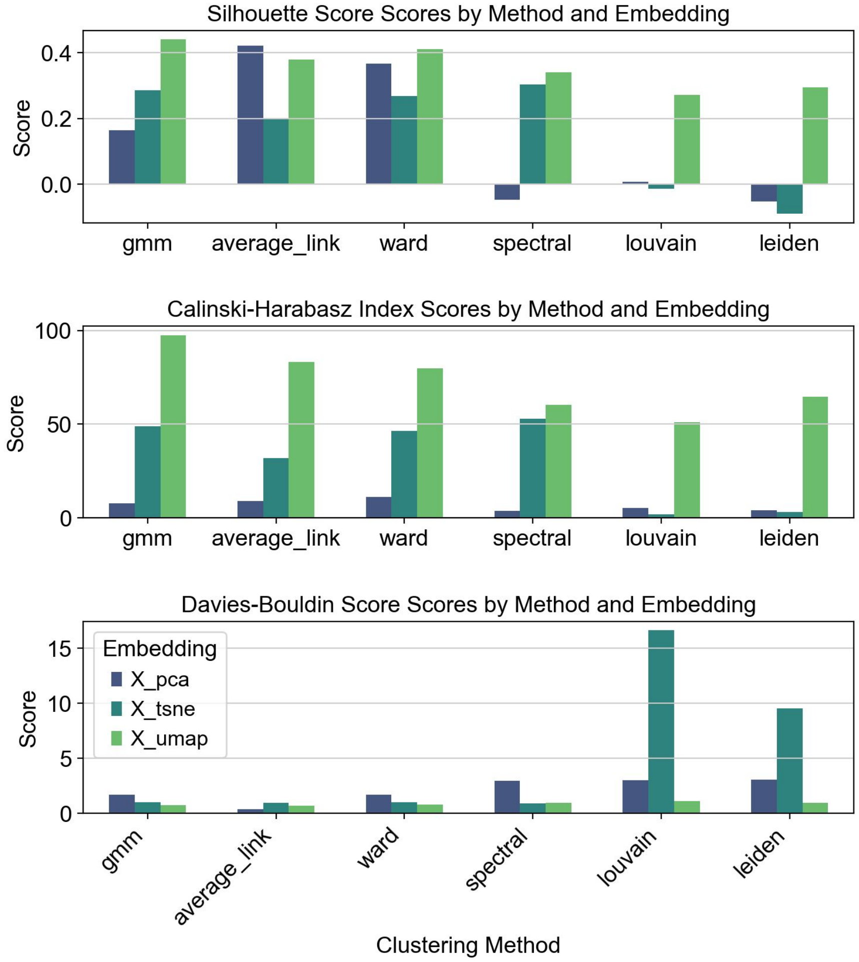
- For each combination of dimensionality reduction method - clustering algorithm, certain evaluation metrics were calculated to determine the best clustering result. These metrics included:
    - **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters, with higher values indicating better-defined clusters.
    - **Calinski-Harabasz Index**: Evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion, with higher values indicating better clustering performance.
    - **Davies-Bouldin Index**: Assesses the average similarity ratio of each cluster with its most similar cluster, where lower values indicate better-defined clusters.

- Based on these metrics, the best performing combination was **GMMs with UMAP**.

Clustering

- **Highest Silhouette (0.440) and Calinski-Harabasz (97.512) Scores:** Demonstrates exceptional internal cohesion and separation among clusters.

- **Low Davies-Bouldin Score (0.748):** Indicates minimal variation within clusters, although not the lowest, it ensures highly distinct clusters.



Visualization of the Results

Authors’ Approach

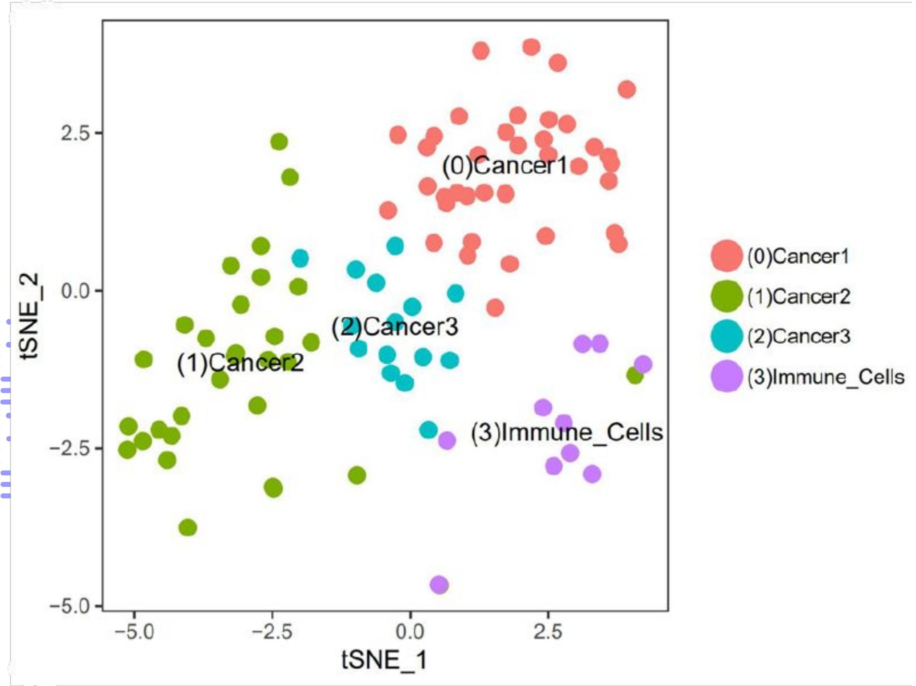
- t-SNE is used to visualize the clusters in two-dimensional space

Our Approach

- **Plan**: Use appropriate visualization method based on best performing clustering result
- **Implementation**:
  - Our best performing clustering based on the evaluation metrics was the combination of GMM with UMAP
  - Thus, UMAP was used to visualize our results in two dimensional space

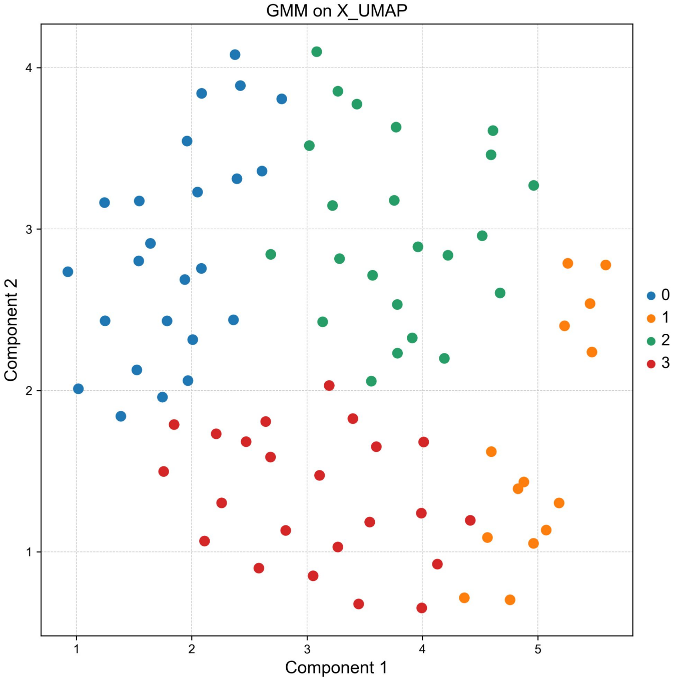
Visualization of the Results

Authors' Approach



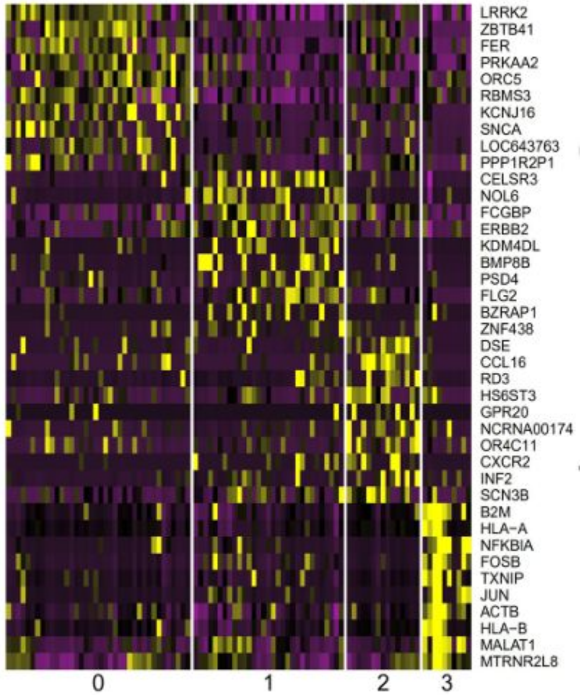
Source: Wang, B. et. al., (2021). Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq. Scientific Reports, 11, 11585. https://doi.org/10.1038/s41598-020-80881-2

Our Approach



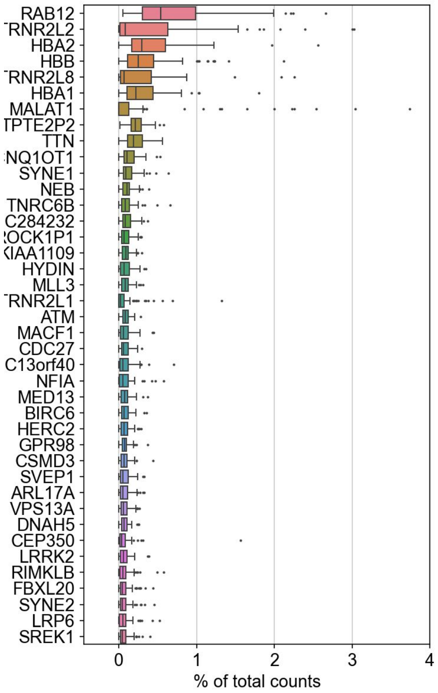
Visualization of the Results

Authors’ Approach



Source: Wang, B. et. al., (2021). Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq. Scientific Reports, 11, 11585. https://doi.org/10.1038/s41598-020-80881-2

Our Approach



- We found the marker genes with scanpy based on our clustering results
- We plan to attempt plotting the heatmap of markers highly expressed in each cluster
- We will evaluate our results and compare with the authors’ approach

Conclusions

- We provided a more comprehensive machine learning scRNA-seq workflow, which included a variety of dimensionality reduction methods, clustering algorithms and evaluation metrics to get the best possible results.
- Our preliminary results about differential marker gene expression between cell clusters somewhat agree with the authors, emphasizing the usefulness of this methodology in identifying cell heterogeneity.

Future Work

- Determine optimal number of clusters (instead of 4) for input in the GMMs, average linkage, Ward's and spectral clustering algorithms.
- Plot heatmap of marker genes based on clustering result and compare results with authors' work.
- Run the entire pipeline without our own preprocessing and compare.
- Experiment with different clustering algorithms (e.g. DBSCAN) and/or evaluation metrics (e.g. BIC criterion).
- Functional annotation of clusters using Gene Ontology (GO).
- Pseudotime trajectory analysis of clustering results.

Thank you for your attention.

Do you have any questions?

Credits: This presentation template was created by Slidesgo, and includes icons by Flaticon, and infographics & images by Freepik