Course: Probabilistic Machine Learning (SoSe 2025) **Lecturer:** Dr. Alvaro Diaz Ruelas
Student(s) Name(s): Max von Kolczynski
GitHub Username(s): MaxKolczynski
Date: 02.06.2025
PROJECT-ID: 26-1KMXXXX_rna_seq

---

1. Introduction

1.1 Motivation for my project

Advances in single-cell RNA sequencing (scRNA-seq) now allow transcriptomic profiling at single-cell resolution, enabling the investigation of cellular heterogeneity and gene expression differences across experimental conditions. This opens up new possibilities for uncovering cell-type-specific regulatory patterns, understanding developmental processes, and exploring the effects of external stimuli at high resolution.

My project focuses on detecting differentially expressed genes between treated and untreated biological samples using probabilistic Bayesian modeling. Additionally I hope to not only identifies expression differences but also provide quantitative estimates of uncertainty and effect sizes, offering a more nuanced and robust interpretation of the results.

1.2 Brief description of the dataset and problem

1.2.1 Biological model and experimental design

> original study and biological model
> https://doi.org/10.1038/s41467-025-58295-3

The dataset analyzed in this project stems from the study *“Astrocyte-secreted cues promote neural maturation and augment activity in human forebrain organoids.”* It uses 90-day-old human **dorsal forebrain organoids derived from embryonic stem cells**, serving as an in vitro model to study early brain development. The study investigates how astrocyte-secreted signals influence neural maturation and functionality. For this purpose, **two experimental conditions** were established: organoids **treated** with astrocyte-

conditioned medium (ACM) and **untreated controls** cultured in standard medium. This setup allows the identification of transcriptional changes driven by astrocyte-derived cues.

1.2.2 How the data was generated

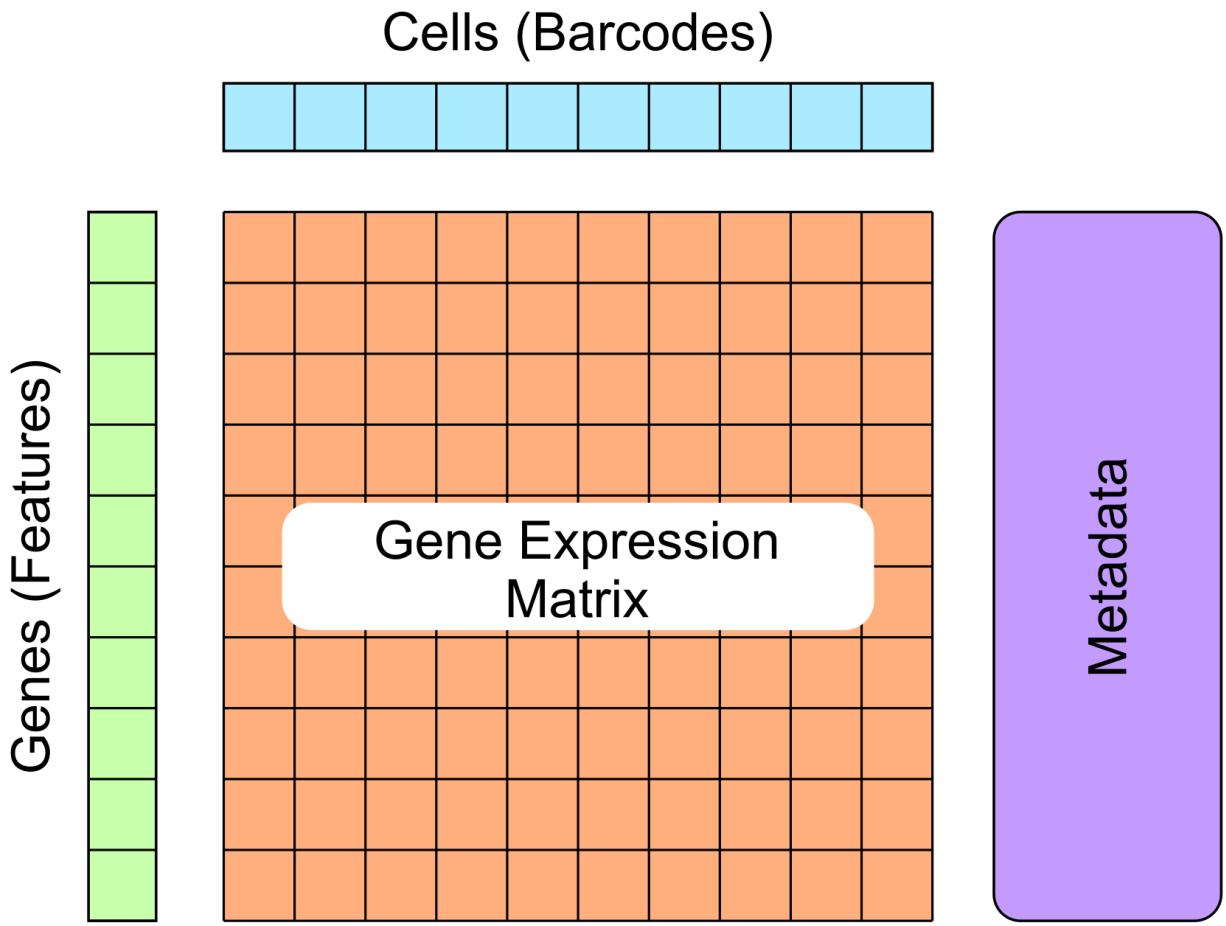
The data were generated using **single-cell RNA sequencing (scRNA-seq)** from a total of ten human dorsal forebrain organoids (five per condition, cultured for 90 days). After enzymatic dissociation of the organoids into single cells using a papain-based method, **approximately 12,000 cells per sample** were individually encapsulated into gel emulsions (GEMs) using the **10x Genomics Chromium** system. Library preparation was performed using the Chromium Single Cell 3′ Library & Gel Bead Kit v2. The resulting cDNA libraries were sequenced on an **Illumina NovaSeq 6000 platform** using a paired-end 150 bp sequencing protocol. The raw sequencing output was processed with Cell Ranger (version 7.0.0), **aligning reads to the human reference genome GRCh38** to quantify cells and their transcript counts.

1.2.3 Description of the dataset

> conceptual architecture of the dataset
> how to interpret the data

The core output of single-cell RNA-seq is the **expression matrix**, a numerical representation of gene expression levels across individual cells. This matrix is typically structured with **rows representing genes** and **columns representing single cells**. Each entry in the matrix corresponds to the **count of a transcript** mapped to a specific gene in

a particular cell.



The count matrix, representing gene expression in thousands of single cells from treated and untreated human forebrain organoids, serves as the basis for all downstream analyses. It enables identification of cell types, states, and treatment effects through clustering, differential expression, and trajectory inference.

1.2.2 My Aim / Hypothesis / Goal

My project specifically aims to apply **probabilistic differential expression modeling** to determine which genes g show significant expression differences between the two conditions, quantifying these differences with posterior probabilities.

Which genes show significant differences between conditions?

$$H_0 : \beta_{1g} = 0 \quad \text{against} \quad H_1 : \beta_{1g} \neq 0$$

2. Data Loading and Exploration

For both conditions (treated / untreated) I downloaded the Count Matrix, Cell Barcodes and Feature List (Genes) form Original Data Source

Link to Google Drive Folder - contains the dataset

Notebook - data loading - #work_in_progress

3. Data Preprocessing

#work_in_progress

> Steps taken to clean or transform the data

Preprocessing is a crucial step in single-cell RNA-seq analysis to ensure data quality and interpretability. Raw count matrices typically include technical noise, low-quality cells, and other artifacts that can obscure biological signals. Standard preprocessing workflows aim to filter unreliable cells, normalize expression levels, and identify informative genes for downstream analyses such as clustering or differential expression.

I follow a preprocessing and QC strategy closely aligned with the original study:

Given the expression count matrix

1. **Cell Filtering**

- **Gene Count Thresholds:** Retain cells with a number of detected genes between 200 and 6000.
  → Filters out empty droplets and potential doublets or multiplets.
- **UMI Count Threshold:** Retain cells with at least 500 total UMIs.
  → Ensures sufficient transcriptional information per cell.
- **Complexity Filter:** Keep cells with $\log_{10}$(genes per UMI) > 0.8.
  → Removes low-complexity libraries, often indicative of poor-quality capture.
- **Mitochondrial and Ribosomal Content:** Remove cells with >20% mitochondrial or >30% ribosomal gene content.
  → High values indicate stressed or dying cells.

2. **Normalization**

- **Library Size Normalization:** Scale gene counts per cell to 10,000 total counts.
  → Corrects for differences in sequencing depth.

- **Log Transformation:** Apply natural log to normalized counts.
  → Stabilizes variance and brings expression levels onto a comparable scale.

3. **Feature Selection**

- **Highly Variable Genes (HVGs):** Select the top 2000 HVGs using Variance Stabilizing Transformation (VST).
  → Focuses downstream analysis on genes that carry the most biological signal.

4. Probabilistic Modeling Approach

> Description of the models chosen
> Why they are suitable for your problem
> Mathematical formulations (if applicable)

Goal: Model the gene expression per gene as a function of the group variable (treated vs. untreated)

Model idead: Bayesian Negative Binomial Model (GLM)

RNA-seq data is **count** data and typically **overdispersed**
→ the **negative binomial distribution** is the standard choice.

For each gene g and cell i:

$$Y_{i,g} \sim \mathrm{NB}(\mu_{i,g}, \theta_g)$$

- $Y_{i,g}$: **measured** count-expression (scRNA-seq)
- $\mu_{i,g}$: **expected** Expression
- θ_g: overdispersion parameter for gene g

Linear Modell for $\mu_{i,g}$:

$$\log(\mu_{i,g}) = \beta_{0g} + \beta_{1g} \cdot \mathrm{condition}_i$$

Priors:

Goal: **Posterior Estimation** via *fold-change* per Gen

---

5. Model Training and Evaluation

6. Results

> Present key findings
> Comparison of models if multiple approaches were used

Visualisierung:
Volcano Plots, Posterior Intervalls, Heatmaps der DE-Gene

7. Discussion

> Interpretation of results
> Limitations of the approach
> Possible improvements or extensions

8. Conclusion

> Summary of main outcomes

9. References

> Cite any papers, datasets, or tools used