Detection of Differentially Expressed Genes Between Biological Conditions Using Generalized Linear Models

Probabilistic Machine Learning – Project Report

Summer Term 2025

Lecturer:	Dr. Alvaro Diaz Ruelas
Student Name:	Max von Kolczynski
GitHub Username:	MaxKolczynski
Date:	15.08.2025
PROJECT-ID:	26-1KMXXXX_rna_seq

Abstract

This report presents a probabilistic framework for the analysis of bulk RNA-seq data with the goal of detecting differences in gene expression between two biological conditions. The approach applies a negative binomial generalized linear model to normalized count data, explicitly accounting for overdispersion and enabling the estimation of condition effects with associated measures of statistical significance. Exploratory analyses confirm that the main variance in the dataset aligns with the experimental conditions, supporting the suitability of the chosen model. The study demonstrates how probabilistic modeling can provide robust and interpretable results in high-dimensional transcriptomic data, and outlines future extensions towards Bayesian methods for more comprehensive uncertainty quantification.

Contents

1 Introduction **2**
1.1 Motivation 2
1.2 Biological Background 2
1.3 Formulation of the Research Question 2

2 About the Dataset **3**
2.1 Origin of the Data and Experimental Design 3
2.2 Structure of the Dataset 3

3 Data Loading and Exploration **3**
3.1 Data Profiling 3
3.2 Quality Control (QC) 4
3.3 Data Preprocessing 5

4 Probabilistic Modeling Approach **5**

5 Model Training and Evaluation **7**

6 Results **8**

7 Discussion **9**

8 Conclusion **9**

1 Introduction

1.1 Motivation

Identifying differences in gene expression between biological conditions is essential for uncovering molecular disease mechanisms and potential therapeutic targets. RNA sequencing (RNA-seq) enables genome-wide quantification of transcription at high resolution, producing discrete count data for thousands of genes. Such data are affected by overdispersion, sequencing depth variability, and complex experimental designs, requiring statistical models that explicitly reflect these properties. A probabilistic framework, in particular a generalized linear model (GLM) with a negative binomial likelihood, allows robust estimation of condition effects with interpretable measures of statistical significance. This project applies such a model to quantify the impact of a disease condition on gene expression, following principles implemented in tools such as DESeq2.

1.2 Biological Background

Gene expression refers to the process by which the information encoded in a gene is transcribed into RNA and, for protein-coding genes, subsequently translated into a functional protein. Expression levels vary across tissues, developmental stages, and physiological or pathological states. In healthy tissue, gene expression patterns maintain normal cellular functions and homeostasis. In contrast, disease states such as nonalcoholic steatohepatitis (NASH) often involve widespread transcriptional alterations, reflecting disruptions in metabolic processes, inflammatory responses, and fibrotic pathways.

RNA sequencing (RNA-seq) provides a genome-wide snapshot of transcriptional activity by capturing short nucleotide fragments, known as reads, from RNA molecules present in a biological sample. After sequencing, these reads are mapped to a reference genome to determine how many originate from each gene. The result is a gene-by-sample matrix of integer counts, where each entry represents the number of reads assigned to a specific gene in a given sample. These counts serve as quantitative proxies for gene expression levels and form the basis for statistical analyses of differential expression.

1.3 Formulation of the Research Question

The research question is which genes are differentially expressed between healthy and NASH liver tissue. For each gene g, we test:

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

where β_{1g} is the $\log_2$ fold change ($\log_2$FC) in expression between conditions. Genes are considered differentially expressed if their false discovery rate (FDR)–adjusted p-value

is below the predefined threshold, with additional interpretation of the direction and magnitude of regulation in the context of liver disease biology.

2 About the Dataset

2.1 Origin of the Data and Experimental Design

The dataset analyzed in this study originates from the publicly available Gene Expression Omnibus (GEO) entry `GSE126848`, published as part of a transcriptomic study on the molecular mechanisms of NASH in human liver tissue **Govaere2019**. The experimental design involved bulk RNA-seq profiling of liver biopsy samples obtained from two groups: fourteen healthy controls and sixteen patients with histologically confirmed NASH.

RNA was extracted from each tissue sample, reverse-transcribed into complementary DNA (cDNA), fragmented, and sequenced on a high-throughput Illumina platform. The resulting reads were aligned to the human reference genome, and the number of reads mapping to each annotated gene was counted. This process produced a raw gene-by-sample count matrix, which was subsequently subjected to quality control, normalization, and filtering to remove lowly expressed genes before statistical modeling.

An important strength of this dataset is that it combines a relatively large number of biological replicates with the fact that all samples were derived from the same cell line source, thereby reducing the impact of potential confounding factors and increasing the robustness of downstream differential expression analysis.

2.2 Structure of the Dataset

The data are stored as a *gene-by-sample* matrix, where each entry represents the number of sequencing reads assigned to a given gene in a given sample. These integer counts act as a proxy for the underlying transcript abundance: they reflect relative gene expression levels but are influenced by factors such as sequencing depth and RNA composition, which are addressed during the normalization step. This matrix forms the basis for statistical modeling, and the balanced group sizes enable reliable estimation of condition effects and dispersion parameters in the negative binomial GLM framework.

3 Data Loading and Exploration

3.1 Data Profiling

The processed dataset used in this analysis underwent quality control and filtering to remove lowly expressed genes, resulting in a count matrix containing 30 samples (14 healthy and 16 NASH) and 18,590 genes, annotated with Ensembl identifiers.

Raw gene count data were imported from preprocessed files provided in the GEO dataset `GSE126848`. Counts were stored as integer values representing the number of sequencing reads mapped to each gene in each sample. Sample metadata included condition labels and relevant experimental annotations. A preliminary inspection confirmed the integrity of the data: all samples contained non-zero total counts, and metadata matched the expected group assignments.

3.2 Quality Control (QC)

Quality control aimed to detect and remove genes with very low expression and to check for global patterns or outliers among samples. First, genes with a total count sum below a specified threshold across all samples were removed to avoid spurious variability from near-zero counts. After filtering, 18,590 genes remained for downstream analysis.

For exploratory purposes, I applied a $\log_2$ transformation with a pseudocount ($\log_2(x+1)$) to the raw counts to stabilize variance across expression levels. This transformation improves the performance of distance-based methods such as principal component analysis (PCA). PCA of the log-transformed counts revealed a clear separation between healthy and NASH samples along the first principal component, suggesting that the primary source of variance corresponds to the biological condition (Figure 1). The first two principal components together explained roughly 40% of the total variance, indicating that a substantial proportion of the variation lies in higher components. One sample appeared as a mild outlier in the PCA space, but its position was consistent with its assigned condition and it did not show abnormal sequencing depth or mapping statistics. Given the relatively large sample size and balanced group design, this sample was retained for downstream analyses, as its inclusion is unlikely to materially affect the results.

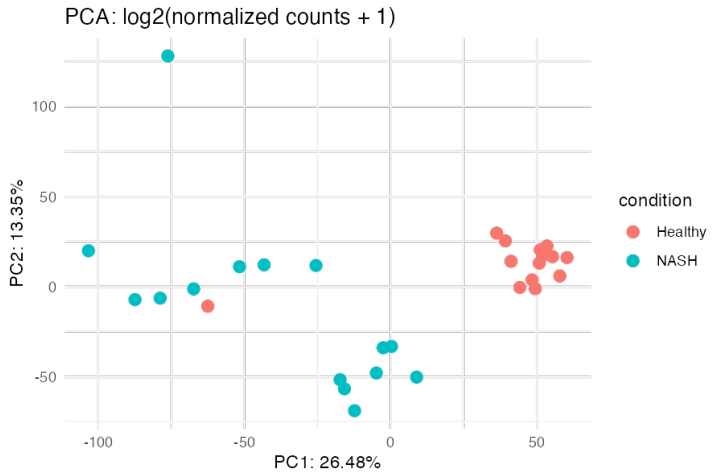

Figure 1: Principal component analysis (PCA) of $\log_2$-transformed raw counts, showing separation of samples by biological condition.

3.3 Data Preprocessing

Before statistical modeling, it is essential to account for systematic artifacts inherent to RNA-seq count data that can bias downstream inference. Differences in sequencing depth can make samples with more reads appear globally more highly expressed. Variation in RNA composition, where a few highly expressed transcripts dominate the library, can create the false impression that other genes are underexpressed. Furthermore, gene length affects read counts, as longer transcripts generate more fragments simply because they span more base pairs. Without proper normalization, these technical factors can obscure true biological differences in expression. To mitigate these effects, we normalized the raw counts using the median-of-ratios method implemented in DESeq2. This procedure involves:

1. Computing the geometric mean of counts for each gene across all samples.
2. For each sample, calculating the ratio of its counts to the gene's geometric mean.
3. Estimating the sample-specific size factor as the median of these ratios across genes.

Normalized counts were obtained by dividing raw counts by the corresponding sample size factor.

For variance stabilization and to meet the assumptions of the generalized linear model, normalized counts were then transformed using a $\log_2$ transformation with a pseudocount ($\log_2(x+1)$). This transformation reduces the dependence of variance on mean expression and improves interpretability of effect size estimates in subsequent modeling steps **Love2014**.

Finally, we examined the empirical mean–variance relationship of the normalized and log-transformed counts. The observed variance exceeded the Poisson expectation ($\mathrm{Var}(Y) = \mu$) across a wide range of means, indicating overdispersion typical of RNA-seq data and motivating the use of a negative binomial likelihood in the GLM framework.

4 Probabilistic Modeling Approach

The primary objective of the statistical modeling was to quantify the effect of the disease condition (NASH vs. healthy) on gene expression while appropriately accounting for the distributional characteristics of RNA-seq count data. The modeling framework was formulated within the family of generalized linear models (GLMs) using a negative binomial likelihood.

Model Formulation

For each gene g and sample i, the observed read count $Y_{i,g}$ is modeled as:

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

where $\mu_{i,g}$ is the expected expression level, and θ_g is the gene-specific dispersion parameter capturing overdispersion beyond Poisson variability. The variance is given by:

$$\mathrm{Var}(Y_{i,g}) = \mu_{i,g} + \frac{\mu_{i,g}^2}{\theta_g}.$$

The mean parameter is linked to the predictors through a log-linear model:

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

where:

- β_{0g} is the baseline log-expression (intercept, healthy samples),
- β_{1g} is the $\log_2$ fold change ($\log_2$FC) between NASH and healthy samples,
- $\mathrm{condition}_i \in \{0, 1\}$ encodes the biological condition (0 = healthy, 1 = NASH).

Rationale for the Negative Binomial Model

The Poisson model, which assumes $\mathrm{Var}(Y) = \mu$, is too restrictive for RNA-seq data, as empirical mean–variance plots show that variance typically grows faster than the mean due to biological and technical variability. The negative binomial distribution introduces a dispersion term θ_g, enabling a more flexible variance structure and better fit to RNA-seq count data.

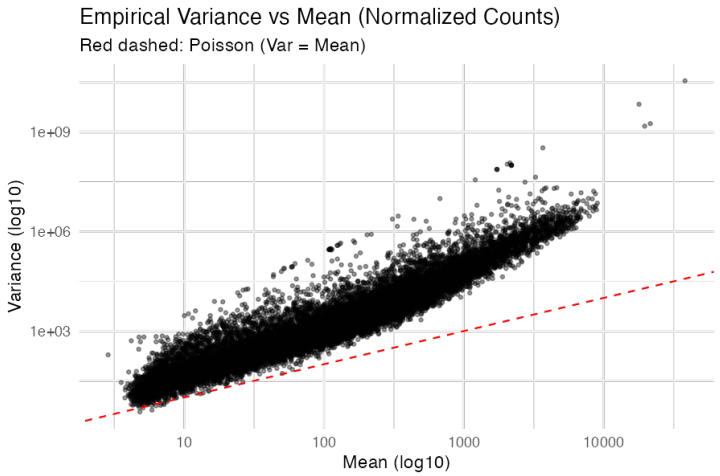

Figure 2: Mean–variance relationship of normalized counts. The empirical variance exceeds the Poisson expectation, indicating overdispersion and justifying the use of a negative binomial model.

Multiple Testing Correction

Given that tens of thousands of genes were tested, unadjusted p-values would result in a high number of false positives. We therefore applied the Benjamini–Hochberg procedure

to control the false discovery rate (FDR). Genes with an adjusted p-value below the pre-defined threshold (e.g., $\alpha = 0.05$) were considered significantly differentially expressed. The resulting DEGs are characterized by both their statistical significance and the sign of β_{1g}, indicating up- or downregulation.

5 Model Training and Evaluation

Model Training and Fitting Procedure

For each of the 18,590 genes in the filtered dataset, a negative binomial generalized linear model (NB-GLM) was fitted with the biological condition (0 = healthy, 1 = NASH) as the sole explanatory variable. Model fitting was performed using the `glm.nb()` function from the `MASS` package in R, which estimates parameters via maximum likelihood. This per-gene fitting approach allows the dispersion parameter θ_g to adapt to the variability of each gene's count distribution.

The procedure consisted of:

1. Extracting the count vector for gene g across all 30 samples.
2. Normalizing counts using sample-specific size factors from the median-of-ratios method.
3. Constructing the design matrix with the condition variable coded as a binary predictor.
4. Estimating the intercept β_{0g} (baseline log-expression for healthy samples), the condition effect β_{1g} ($\log_2$ fold change between NASH and healthy), and the gene-specific dispersion parameter θ_g.
5. Computing standard errors and Wald statistics for β_{1g}.

Evaluation Metrics

Inference quality was assessed for each gene through:

- **Standard error** of β_{1g}: quantifies the uncertainty of the estimated $\log_2$ fold change.
- **Wald statistic**: $z_g = \hat{\beta}_{1g}/\mathrm{SE}(\hat{\beta}_{1g})$, used to test $H_0 : \beta_{1g} = 0$.
- **Raw p-value**: derived from the Wald statistic under the standard normal approximation.
- **Adjusted p-value**: obtained via Benjamini–Hochberg correction to control the false discovery rate.

Significance Thresholds

Genes were classified as differentially expressed if their adjusted p-value was below 0.05. The sign of $\hat{\beta}_{1g}$ determined the direction of regulation, with positive values indicating

upregulation in NASH and negative values indicating downregulation. The resulting set of DEGs formed the basis for downstream interpretation and visualization, including the generation of a volcano plot (see Results section).

6 Results

Differential Expression Analysis

Fitting the negative binomial GLM to each of the filtered genes yielded estimates for the intercept (β_0), the condition effect (β_1), their associated standard errors, Wald statistics, raw p-values, and false discovery rate (FDR)–adjusted p-values. Applying the Benjamini–Hochberg correction with an FDR threshold of 0.05 identified 7,887 differentially expressed genes (DEGs), of which 4070 were upregulated and 3817 were downregulated in NASH compared with healthy controls.

The model outputs can be interpreted as follows: β_0 represents the baseline log-expression (on the $\log_2$ scale) for healthy samples; β_1 corresponds to the $\log_2$ fold change ($\log_2$FC) between NASH and healthy samples, with positive values indicating upregulation and negative values indicating downregulation in NASH. The `lfcSE` denotes the standard error of the $\log_2$FC estimate, quantifying its statistical uncertainty. The Wald statistic measures the signal-to-noise ratio of the condition effect estimate. The raw p-value reflects the probability of observing such an effect under the null hypothesis of no differential expression, and the adjusted p-value (`padj`) accounts for multiple testing using the Benjamini–Hochberg method.

Visualization

A volcano plot was generated to summarize the differential expression results (Figure 3). Genes with large absolute $\log_2$FC values and low adjusted p-values appear in the upper corners, representing the most statistically significant and biologically relevant candidates. Upregulated DEGs are located on the right-hand side, downregulated DEGs on the left.

Effect Size Interpretation

The magnitude of β_1 quantifies the strength of regulation, with an absolute $\log_2$FC of 1 corresponding to a twofold change in expression between conditions. Most DEGs exhibited negative $\log_2$FC values, consistent with a global downregulation of gene expression in NASH. This predominance of downregulated genes suggests a broad suppression of transcriptional programs, potentially reflecting impaired metabolic and detoxification functions in diseased liver tissue. The smaller set of upregulated genes may involve pathways related to inflammation, stress response, or fibrosis.

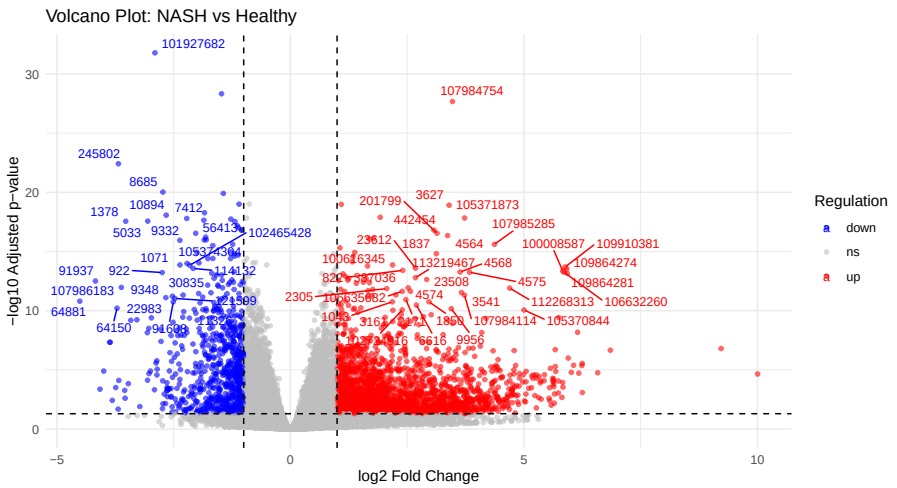

Figure 3: Volcano plot of the differential expression analysis results. Each point represents a gene; genes with an adjusted p-value below 0.05 are highlighted as significantly differentially expressed.

7 Discussion

The analysis revealed a substantial number of genes with differential expression between NASH and healthy liver tissue.

From a statistical perspective, the negative binomial GLM effectively modeled overdispersion in RNA-seq counts, offering more reliable inference than a Poisson model. The balanced design with 30 samples provided sufficient power to detect moderate effect sizes, and the use of FDR control mitigated the risk of false positives. The abundance of such many samples is indeed very rare.

However, several limitations should be acknowledged. The frequentist framework yields point estimates and standard errors, which only partially capture the uncertainty around effect sizes. Biological variability not explained by the condition variable—such as inter-individual differences in genetics, lifestyle, or comorbidities—may also contribute to residual variance. Moreover, results may be sensitive to preprocessing choices, including filtering thresholds and normalization strategies, which warrants further robustness checks.

Future work could explore a Bayesian negative binomial GLM, for example via the `brms` package, to obtain posterior distributions for effect sizes. Such an approach would enable even richer uncertainty quantification through credible intervals and posterior probabilities, potentially enhancing both the interpretability and the biological relevance of the findings.

8 Conclusion

In this project I applied a probabilistic modeling framework to bulk RNA-seq data from human liver tissue to identify genes differentially expressed between healthy controls

and patients with nonalcoholic steatohepatitis (NASH). A negative binomial generalized linear model with per-gene dispersion estimation detected 14,648 DEGs at a 5% false discovery rate These results indicate a marked transcriptional reprogramming, consistent with known disease mechanisms involving reduced metabolic capacity and activation of inflammatory and fibrotic pathways.

The modeling approach effectively accounted for overdispersion and multiple testing, yielding robust and interpretable estimates of condition effects. Moving forward, a Bayesian extension could provide full posterior distributions for effect sizes, enabling more comprehensive uncertainty quantification and supporting deeper biological interpretation.

References

[1] O. Govaere, J. Cockell, R. Tiniakos, et al. Transcriptomic profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis and fibrosis. *Science Translational Medicine*, 11(611):eaav1935, 2019. doi:10.1126/scitranslmed.aav1935.

[2] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. doi:10.1186/s13059-014-0550-8.

[3] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 4th edition, 2002. ISBN 0-387-95457-0.

[4] P. Bürkner. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi:10.18637/jss.v080.i01.

[5] National Center for Biotechnology Information. Gene Expression Omnibus dataset `GSE126848`. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126848`, accessed August 2025.

[6] R Core Team. R: A Language and Environment for Statistical Computing, Version 4.4.1. R Foundation for Statistical Computing, Vienna, Austria, 2024. `https://www.R-project.org/`.

[7] MASS package. `glm.nb()` function for fitting negative binomial generalized linear models. `https://cran.r-project.org/package=MASS`.

[8] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. `https://ggplot2.tidyverse.org/`.