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.

Contents

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.

1.2 Biological Background

Gene expression is the process by which a gene's information is transcribed into RNA and, for protein-coding genes, translated into a functional protein. Expression varies across tissues, developmental stages, and physiological or pathological states. In healthy tissue, patterns maintain normal function, whereas in nonalcoholic steatohepatitis (NASH) they often show widespread changes, reflecting disrupted metabolism, inflammation, and fibrosis. RNA sequencing (RNA-seq) captures a genome-wide snapshot of transcription by sequencing short fragments (reads) from RNA molecules in a sample. Reads are mapped to a reference genome to count how many originate from each gene, producing a gene-by-sample matrix of integer counts that proxy expression levels and serve as the basis for differential expression analysis.

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 comes from the publicly available GEO entry `GSE126848`, part of a transcriptomic study on NASH in human liver tissue **Govaere2019**. Bulk RNA-seq was performed on liver biopsies from 14 healthy controls and 15 patients with histologically confirmed NASH. RNA was extracted, reverse-transcribed into cDNA, fragmented, and sequenced on an Illumina platform. Reads were aligned to the human reference genome, and counts per annotated gene were obtained. The resulting gene-by-sample matrix underwent quality control, normalization, and low-expression filtering before modeling.

A key strength of this dataset is the number of biological replicates and the consistent tissue source and sequencing platform, reducing confounding and improving robustness of the 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 29 samples (14 healthy and 15 NASH) and 16,849 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. Genes with a total count sum below a specified threshold across all samples were removed to avoid spurious variability from near-zero counts, leaving 16,849 genes for downstream analysis. For exploratory purposes, a $\log_2$ transformation with a pseudocount ($\log_2(x+1)$) was applied 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. One sample appeared as a mild outlier in the PCA space but showed no 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.

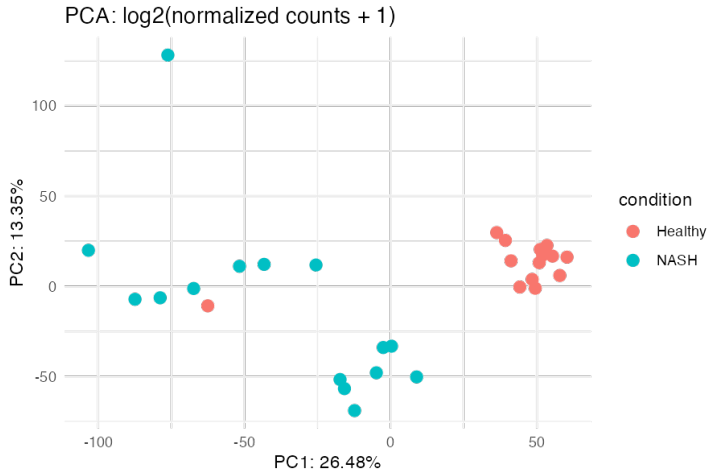

Figure 1: Principal component analysis (PCA) of $\log_2$-transformed raw counts, showing separation of samples by biological condition.

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 factors, raw counts were normalized using the DESeq2 median-of-ratios method: geometric means were computed per gene, each sample's counts were divided by these means to obtain ratios, and the sample-specific size factor was defined as the median ratio across genes. Normalized counts were then obtained by dividing raw counts by the respective size factor. For variance stabilization and to better meet GLM assumptions, normalized counts were $\log_2$-transformed with a pseudocount ($\log_2(x+1)$), reducing the dependence of variance on mean expression and improving interpretability of effect sizes.

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 accounting for the distributional characteristics of RNA-seq count data. From a modeling perspective, it is intuitive to relate the condition factor **linearly** to the expected expression, allowing the effect of disease status to be captured by a single coefficient β_{1g} per gene, which quantifies the proportional change in expression between conditions. Coding the condition as a binary variable (0 = healthy, 1 = NASH) makes β_{0g} the baseline expression in healthy samples and β_{1g} the change associated with NASH. The modeling framework was implemented within the family of generalized linear models (GLMs) assuming a negative binomial distribution.

4.1 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,

- $\text{condition}_i \in \{0, 1\}$ encodes the biological condition (0 = healthy, 1 = NASH).

4.2 Rationale for the Negative Binomial Model

The Poisson model, which assumes $\text{Var}(Y) = \mu$, 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.

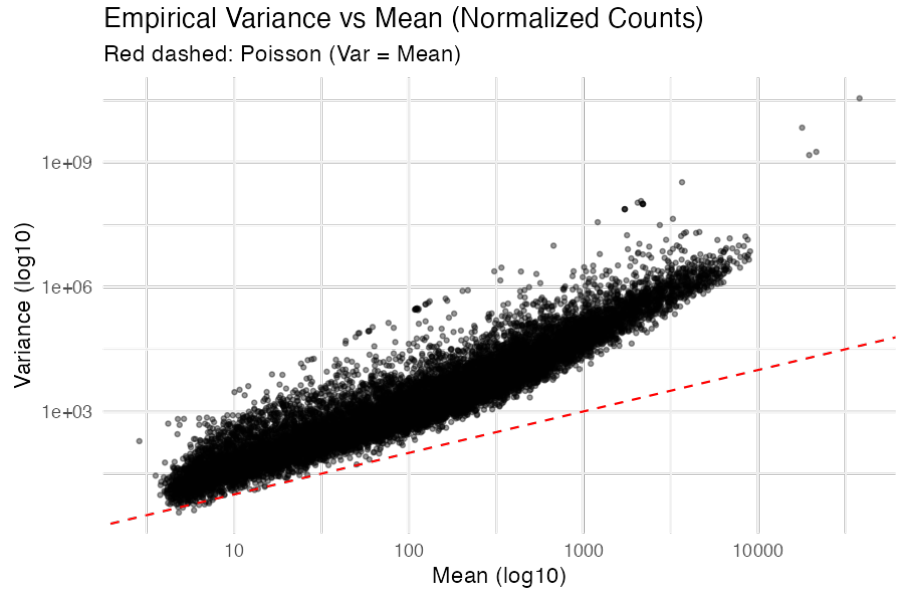

Figure 2: Mean–variance relationship of normalized counts. The empirical variance exceeds the Poisson expectation, indicating overdispersion and justifying the use of a negative binomial model.

4.3 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

5.1 Model Fitting

The statistical analysis was conducted in R using the `MASS::glm.nb()` function to fit negative binomial generalized linear models (NB-GLMs) via maximum likelihood estimation. For each of the 16,849 genes, the biological condition (0 = healthy, 1 = NASH) was the sole predictor, allowing the gene-specific dispersion parameter θ_g to adapt to individual variability.The modeling workflow comprised: (1) extracting each gene's count vector across all 29 samples; (2) normalizing counts using sample-specific size factors from the median-of-ratios method; (3) constructing a binary-coded design matrix; (4) estimating β_{0g} (baseline expression), β_{1g} ($\log_2$ fold change), and θ_g; and (5) computing the standard error and Wald statistic for β_{1g}.

5.2 Significance Thresholds

Genes were classified as differentially expressed if their Benjamini–Hochberg–adjusted p-value was below 0.05. The sign of $\hat{\beta}_{1g}$ indicated the direction of regulation, with positive values representing upregulation in NASH and negative values indicating downregulation. The resulting DEGs formed the basis for downstream interpretation and visualization, including the volcano plot (see Results section).

5.3 Evaluation Metrics

Inference quality for each gene was summarized by:

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

6 Results

Applying the NB-GLM framework and Benjamini–Hochberg correction (FDR 0.05) identified **7,887** differentially expressed genes, with 4,070 upregulated and 3,817 downregulated in NASH compared with healthy controls. The analysis pipeline exports a CSV file containing all identified DEGs to the `outputs/tables` directory.

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

Table 1: Five DEGs with largest $\log_2$ FC values.

Gene ID	baseMean	$\log_2$FC	SE	padj
26835	43.54	7.50	1.56	1.32×10^{-5}
7503	124.71	6.93	1.38	5.35×10^{-6}
26870	109.00	6.81	1.03	1.28×10^{-9}
6060	111.40	6.78	1.08	9.76×10^{-9}
26863	109.79	6.70	1.03	2.50×10^{-9}

($\log_2$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.

6.1 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.

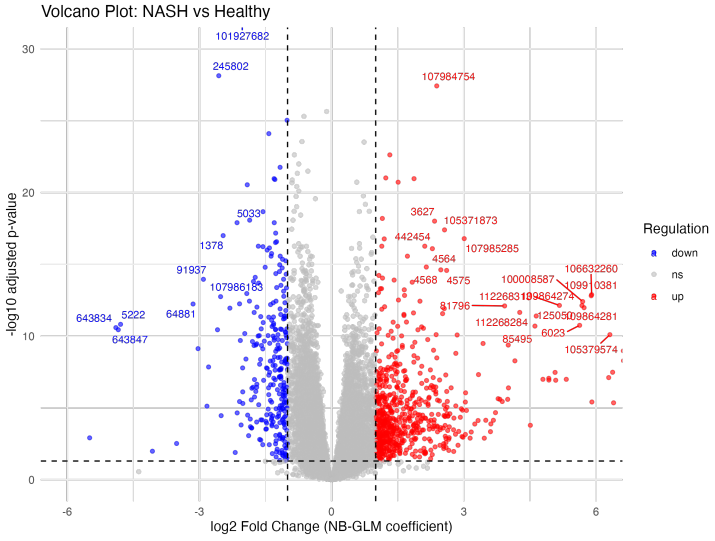

Figure 3: Volcano plot of the differential expression analysis results. Each point represents a gene; genes with an adjusted p-value below 0.05 are highlighted as significantly differentially expressed.

6.2 Effect Size Interpretation

The coefficient β_1 represents the estimated $\log_2$ fold change in expression between NASH and healthy samples. An absolute $\log_2$FC of 1 corresponds to a twofold difference in expression; for example, $\beta_1 = 2$ indicates a fourfold higher expression in NASH, while $\beta_1 = -1.5$ corresponds to approximately 2.8-fold lower expression. The sign of β_1 specifies the direction of regulation: positive values indicate higher expression in NASH, negative values lower expression. Effect sizes should be interpreted together with their standard errors and adjusted p-values, as large coefficients with high uncertainty may not represent robust differential expression.

7 Discussion

The analysis identified many genes with differential expression between NASH and healthy liver tissue. Statistically, the negative binomial GLM modeled overdispersion effectively, offering more reliable inference than a Poisson model. The balanced design with 29 samples provided sufficient power for moderate effect sizes, and FDR control limited false positives. Limitations include reliance on point estimates and standard errors, which only partly capture uncertainty, and residual variability from factors beyond condition, such as genetics or lifestyle. Results may also depend on preprocessing choices like filtering thresholds and normalization, warranting robustness checks. Future work could explore a *Bayesian* negative binomial GLM, to obtain posterior distributions for effect sizes. Such an approach would enable richer uncertainty quantification through credible intervals and posterior probabilities (e.g., $P(\beta_{1g} > 0)$), potentially enhancing both interpretability and biological relevance.

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 **7,887** DEGs at a 5% false discovery rate from 16,849 filtered genes across 29 samples (14 healthy, 15 NASH). These results indicate 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.

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/`.