gene-expression

Diego Cesar Villa Almeyda

2025-06-21

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

2 Methods

2.1 Experimental design and data

The study involved culturing cells from both the wild-type (WT) and FLC1Δ mutant strains under various environmental conditions. Cells were grown in four different media: YPD (Y), YPD with CFW (YC), YPD with EGTA (YE), and YPD with both CFW and EGTA (YCE), all at 37°C to induce stress. Additionally, both strains were grown under baseline conditions in standard YPD at 30°C. In total, 10 distinct experimental conditions were tested (5 environmental conditions × 2 strains), each with 3 biological replicates, resulting in 30 samples.

Table 2.1 presents the experimental design along with the sample labels for each condition. The 10 experimental conditions are labeled by combining the levels of the three experimental factors: strain, temperature, and growth media, in that order. We use the label FL to denote the FLC1Δ strain. For example, the condition involving the wild-type strain grown in YCE medium at 37°C is labeled WT-37-YCE, while the corresponding condition for the FLC1Δ strain is labeled FL-37-YCE.

The primary output of the experiment was a raw count matrix containing RNA abundance measurements for 6795 genes across 30 samples, resulting in a 6795 × 30 matrix. Each entry in the matrix represents the number of sequencing reads mapped to a specific gene in a given sample.

Table 2.1: Overview of the experimental design showing the combinations of temperature and growth media (environmental conditions) used to grow cells of the WT and FLC1Δ (FL) strains. Each cell lists the replicate sample labels (S1–S30) corresponding to each unique experimental condition.

	Enviromental conditions				
	30°C	**37°C**			
Strain	**Y**	**Y**	**YC**	**YE**	**YCE**
WT	S1, S2, S3	S4, S5, S6	S7, S8, S9	S10, S11, S12	S13, S14, S15
FL	S16, S17, S18	S19, S20, S21	S22, S23, S24	S25, S26, S27	S28, S29, S30

2.2 Pre-filtering

Some genes may exhibit low or zero counts across all samples, suggesting they are not expressed under the experimental conditions. These genes are unlikely to be identified as differentially expressed and are typically filtered out prior to analysis. The biological justification for pre-filtering low count or uninformative genes is that a gene typically needs to be expressed above a minimal threshold to be translated into a functional protein or to exert a meaningful biological effect (Chen, Lun, and Smyth 2016). Moreover, low-expression genes often reflect sampling noise rather than true biological signal (Sha, Phan, and Wang 2015).

DE analysis involves performing a hypothesis test for each gene to evaluate whether expression levels differ between experimental conditions (see Section 2.5). This results in a multiple testing problem, which is typically addressed by adjusting p-values to control the false discovery rate (FDR). However, such corrections reduce statistical power by raising the threshold for significance. As the number of tests increases, this loss of power aggravates, further limiting the ability to detect truly differentially expressed (DE) genes, especially when they represent a small proportion of the total set (Bourgon, Gentleman, and Huber 2010). To mitigate this, it is recommended to filter out low-expression genes prior to DE analysis, thereby reducing the total number of hypotheses. This makes the multiple testing correction less stringent and increases the chance of correctly identifying DE genes (Bourgon, Gentleman, and Huber 2010; Sha, Phan, and Wang 2015).

Although the software used for differential expression (DE) analysis includes an internal filtering routine (see Section 2.5), additional pre-filtering was applied using an empirical method implemented in the R package `edgeR` (Chen et al. 2025), as described in Chen, Lun, and Smyth (2016). This method retains genes whose counts-per-million (CPM) exceed a specified threshold k in at least n samples. The CPM for each gene in a sample is calculated by dividing the raw read count by the total number of reads (i.e., the library size) in that sample and scaling by one million.

To determine the CPM threshold k, the user specifies a minimum raw count that a gene must meet in n samples, and the software computes the corresponding CPM based on the smallest library size. For this study, we required a minimum count of 10 in at least 3 samples, corresponding to the number of replicates per condition. The original, unfiltered count matrix was retained to replicate the DE analysis and compare results.

2.3 Normalization and transformation of raw counts

In RNA-seq experiments, the number of reads mapped to a gene depends not only its expression level but also on between-sample and within-sample factors that make the row counts not directly comparable across genes or samples (Gierliński et al. 2015). Between-sample variation primarily arises from differences in library size, with larger library sizes producing more reads

across all genes in a sample, whereas within-sample variation occurs at a gene level and can be influenced by gene length or by the proportion of guanine (G) and cytosine (C) nucleotides in the genes (GC-content, Dillies et al. 2012).

A common approach to correct for between-sample variation is total count normalization, in which raw counts are divided by the total number of reads (library size) for each sample and scaled by a constant factor. CPM is a typical example of this method. However, total count normalization can be problematic in DE analysis, as a small number of highly expressed genes can disproportionately inflate the library size. This can artificially lower the normalized expression of other genes within the same sample and exaggerate differences between samples, even when no real biological variation exists (Gierliński et al. 2015). More robust normalization methods have been developed for DE analysis, such as the trimmed mean of M-values (TMM) method (Robinson and Oshlack 2010). In this study, we employed the normalization procedure implemented in the R package `DESeq2` (Love, Huber, and Anders 2014), which uses the median-of-ratios method to estimate sample-specific size factors for each gene count (Anders and Huber 2010). For comparison purposes, we also applied CPM normalization. No normalization to correct for within-sample variation (e.g., gene length or GC content) was applied, following the recommendation of the project advisor.

An additional challenge in RNA-seq analysis arises during data visualization and multivariate techniques such as clustering or principal component analysis (PCA), which are sensitive to the scale of the variables being analyzed. Even after normalization, RNA-seq data often exhibit heteroskedasticity, where genes with higher expression levels tend to show greater variance across samples than low-expressed genes. As a result, these highly variable genes can disproportionately influence the analysis, potentially obscuring meaningful biological patterns (Love, Huber, and Anders 2014). A common solution is to apply a variance-stabilizing transformation (VST) to the raw or normalized counts, which aims to place lowly and highly expressed genes on a common scale (Hafemeister and Satija 2019).

One of the most common VST is the logarithm (log) transformation. However, this transformation has the issue of exaggerating variability in low-count genes, where random noise tends to overshadow true biological signal (Love, Huber, and Anders 2014). In this study, we applied two transformations available in the `DESeq2` package: the regularized logarithm (rlog) transformation (Love, Huber, and Anders 2014) and the VST proposed by Anders and Huber (2010). The rlog transformation is closely related to the modeling framework used for differential expression analysis in `DESeq2` and will be described in more detail in Section 2.5. The latter transformation, which we will refer to simply as VST, performs a monotonic transformation of the normalized counts so that the resulting variance is approximately independent of the mean. Both transformations were applied after normalizing the counts using the `DESeq2` method. To benchmark these transformations, we also applied a log transformation with base 2 to the CPM-normalized counts having previously added a prior count of 2 to the raw counts to avoid taking logarithm of zero (undefined).

2.4 Exploratory analysis

An important first step in RNA-seq data analysis is assessing the quality and consistency of biological replicates. Ideally, replicates within the same experimental condition should exhibit similar expression profiles after normalization and transformation. To evaluate this, we computed both the correlation matrix and the pairwise distance matrix of the samples, using the Pearson correlation and the Euclidean distance, respectively. If the replicates are consistent, the correlation and distance matrices should reveal strong similarity among samples from the same condition and clear separation from those in different conditions. Prior to this analysis, the raw counts were normalized and transformed using the methods described in Section 2.3, resulting in three versions of the data: log-transformed CPM, rlog-transformed counts, and VST-transformed counts.

To complement this analysis, a more visual approach to assess replicate quality is to project the samples into a lower-dimensional space and evaluate whether replicates from the same condition cluster together. We used PCA for this purpose. PCA decomposes the correlation matrix of the genes into uncorrelated components ordered by the variance they explain, which is quantified by their corresponding eigenvalues. To determine how many dimensions to retain, we applied Horn's parallel analysis: this method compares the observed eigenvalues to those obtained from randomly generated uncorrelated data sets and retains components whose eigenvalues exceed the average from uncorrelated data sets (Dinno 2009). We visualized the samples in all pairwise combinations of the retained components. Additionally, to explore which experimental conditions or factors might be driving the separation of samples in the PCA space, we created boxplots of the sample scores (i.e., coordinates) grouped by each experimental factor, allowing us to assess potential patterns or group separations.

Furthermore, PCA also enables us to identify which genes are most strongly associated with each principal component by examining the loadings, which reflect the correlation between genes and components. For each pair of components, we generated loading plots highlighting the top 1% of genes based on the magnitude of their loadings, showing both the strength and direction of their contributions.

Prior to all PCA-related analyses, we selected the top 10% most variable genes across samples since these genes are more likely to capture biological signal, whereas low-variance genes typically contribute little or are uninformative for the purpose of visualizing the trends in the data.

2.5 Differential expression analysis

The DE analysis approach used in this study follows the methodology implemented in the `DESeq2` package. The core idea is to model the raw counts for each gene using a negative binomial (NB) distribution, where the logarithm of the normalized mean is modeled as a linear

combination of coefficients corresponding to contrasts of experimental conditions. Hypothesis testing is then performed on these coefficients to assess whether the corresponding contrasts result in statistically significant differential expression for a given gene. In the remainder of this section, we briefly outline the key features of this modeling framework, as described by Love, Huber, and Anders (2014).

2.5.1 Modelling gene expression

Let c_{ij} denote the observed raw count for gene i in sample j. We assume that c_ij has a NB distribution with mean μ_{ij} and gene-specific dispersion parameter α_i; that is, $c_{ij} \sim NB(\mu_{ij}, \alpha_i)$. The mean μ_{ij} is modelled as the product of a sample-specific size factor s_j and a normalized expression level q_{ij}, such that $\mu_{ij} = s_j q_{ij}$. The size factor s_j is estimated using the median-of-ratios method, as mentioned in Section 2.3. The normalized expression level, q_{ij}, is modelled in this study as

$$\begin{aligned}\log(q_{ij}) = \beta_{i0} + \beta_{i1}x_j^{(\mathrm{FL})} + \beta_{i2}x_j^{(\mathrm{YC})} + \beta_{i3}x_j^{(\mathrm{YE})} + \beta_{i4}x_j^{(\mathrm{YCE})} + \beta_{i5}x_j^{(30)} + \\ \beta_{i6}x_j^{(\mathrm{FL})}x_j^{(\mathrm{YC})} + \beta_{i7}x_j^{(\mathrm{FL})}x_j^{(\mathrm{YE})} + \beta_{i8}x_j^{(\mathrm{FL})}x_j^{(\mathrm{YCE})} + \beta_{i9}x_j^{(\mathrm{FL})}x_j^{(30)},\end{aligned}$$

where the intercept term β_{i0} denotes the baseline expression level of gene i under the reference condition. The coefficients β_{i1} to β_{i5} represent the main effects of the experimental factors: $x_j^{(\mathrm{FL})}$ is an indicator variable for the FLC1Δ strain, with the WT strain serving as the reference level; $x_j^{(\mathrm{YC})}$, $x_j^{(\mathrm{YE})}$, and $x_j^{(\mathrm{YCE})}$ are indicators for the different media conditions, with Y as the reference level; and $x_j^{(30)}$ indicates the 30°C temperature condition, with 37°C as reference. Together, this implies that the reference experimental condition is WT-37-Y. The interaction terms β_{i6} to β_{i9} model how the effects of media and temperature differ under the presence or absence of the FLC1 gene. No interaction between media and temperature was included, as these factors are not fully crossed in the experimental design.

2.6 Motif enrichment analysis

3 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

`r dim(counts_raw)[1]`

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Nature Precedings*, 1–1.

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber. 2010. "Independent Filtering Increases Detection Power for High-Throughput Experiments." *Proceedings of the National Academy of Sciences* 107 (21): 9546–51.

Chen, Yunshun, Lizhong Chen, Aaron T L Lun, Pedro Baldoni, and Gordon K Smyth. 2025. "edgeR V4: Powerful Differential Analysis of Sequencing Data with Expanded Functionality and Improved Support for Small Counts and Larger Datasets." *Nucleic Acids Research* 53 (2): gkaf018. https://doi.org/10.1093/nar/gkaf018.

Chen, Yunshun, T. L. Lun Lun, and Gordon K. Smyth. 2016. "From Reads to Genes to Pathways: Differential Expression Analysis of RNA-Seq Experiments Using Rsubread and the edgeR Quasi-Likelihood Pipeline." *F1000Research* 5 (1438). https://doi.org/10.12688/f1000research.8987.2.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2012. "A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis." *Briefings in Bioinformatics* 14 (6): 671–83. https://doi.org/10.1093/bib/bbs046.

Dinno, Alexis. 2009. "Exploring the Sensitivity of Horn's Parallel Analysis to the Distributional Form of Random Data." *Multivariate Behavioral Research* 44 (3): 362–88.

Gierliński, Marek, Christian Cole, Pietà Schofield, Nicholas J Schurch, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, et al. 2015. "Statistical Models for RNA-Seq Data Derived from a Two-Condition 48-Replicate Experiment." *Bioinformatics* 31 (22): 3625–30.

Hafemeister, Christoph, and Rahul Satija. 2019. "Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression." *Genome Biology* 20 (1): 296.

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15: 1–21.

Robinson, Mark D, and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology* 11: 1–9.

Sha, Ying, John H Phan, and May D Wang. 2015. "Effect of Low-Expression Gene Filtering on Detection of Differentially Expressed Genes in RNA-Seq Data." In *2015 37th Annual In-*

ternational Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 6461–64. IEEE.