Search for changes in yeast RNA expression during the fermentation process using differential expression analysis methods.

Vinogradova Sofiya, Zherko Nikita
Moscow Institute of Physics and Technology

March 2, 2024

Abstract

RNA sequencing (RNA-seq) offers a powerful approach to study differential gene expression across various conditions, facilitating a deeper understanding of cellular responses and regulatory mechanisms. In this study, we utilized RNA-seq to investigate the alterations in gene expression of Saccharomycetes, a group of yeast, under aerobic and anaerobic conditions. Our analysis revealed significant changes in the expression levels of numerous genes, including those involved in ribosomal RNA (rRNA) biosynthesis and overall translation processes, as well as those associated with carbohydrate metabolism. These findings highlight the complex regulatory adjustments these organisms make in response to oxygen availability, impacting fundamental biological processes such as protein synthesis and energy production. Our results contribute to the understanding of the molecular basis of yeast adaptation to different oxygen levels, offering insights into the metabolic flexibility and resilience of Saccharomycetes.

1 Introduction

Differential expression analysis is a fundamental technique in molecular biology and bioinformatics used to identify genes or proteins that demonstrate significant changes in expression levels between different biological conditions. These conditions could include various experimental treatments, disease states, developmental stages, or other factors.

Differential expression analysis typically involves comparing gene expression profiles obtained from high-throughput techniques such as RNA sequencing [1] or microarray experiments [2]. Statistical methods are then applied to detect genes whose expression levels are significantly altered between the compared conditions. Subsequent analysis of differentially expressed genes can provide valuable insights into the biological pathways and networks involved in the represented phenotypic differences.

2 Materials and Methods

Work was performed with two copies of *yeast* RNA-seq data before and during fermentation, which are represented by 4 files of reads. Reads in SRR941816 and SRR941817 in the absence of fermentation, and in SRR94181818 and SRR941819 30 min after the start of fermentation. We used Saccharomyces cerevisiae strain S288c and assembly R64 as a reference genome and also used annotation file of this strain.

The pipeline for the analysis was built on the basis of two programs: HISAT2 [3] and deseq2 [4]. We used HISAT2 to align reads files (for single-end mode used -U flag) to a reference genome using pre-prepared index files of a given genome (genome index construction was also done through HISAT2) and after used samtools to sort [5]. After alignments we used featureCounts [6]. This tool allows us to estimate read counts of genes from 4 files obtained after alignment. FeatureCounts can not work with GFF files therefore we convert the GFF file to GTF format using gffread [7].

To find genes that are differentially expressed, we will use two scripts: *deseq2.r* *draw-heatmap.r*. With the first one, we get the metrics for our genes and matrix, which contain normalised counts that we will use in visualisation. With the help of the second script we build a heatmap.

We also used a service to determine the GO for the top 50 differentially expressed genes [8].

3 Results

Alignments and subsequent sorting with samtools were performed for each reads file, after which we got 4 files with single-end sequencing results in bam format. Before alignment, we made genome indices and take 8 files in ht2 format.

Using featureCounts, we obtained the SRR and SRR.summary files. After we left only the columns we needed and made a file containing gene names and number of reads for each of the 4 different bam files obtained earlier.

When we ran the *deseq2.r*, we got two files: result.txt, containing the calculated metrics for our genes (baseMean, log2FoldChange, lfcSE, sta, pvalue, padj) and norm-matrix-deseq2.txt, containing the normalized counts that we will use in the visualization.

The result of running the second script is the heat map shown in picture 1.

The list of top 50 genes is as follows: YER062C YDR536W YHR094C YNL065W YKL120W YJL122W, YLR264W, YGR159C, YNL112W, YOR271C, YJL108C, YJL107C, YOR360C, YOL124C, YOL136C, YER127W, YLR180W, YCR072C, YNL141W, YML043C, YMR300C, YEL026W, YOR359W, YHR066W, YGR079W, YIR012W, YJL148W, YOL041C, YLR327C, YOL151W, YBL028C, YOL080C,

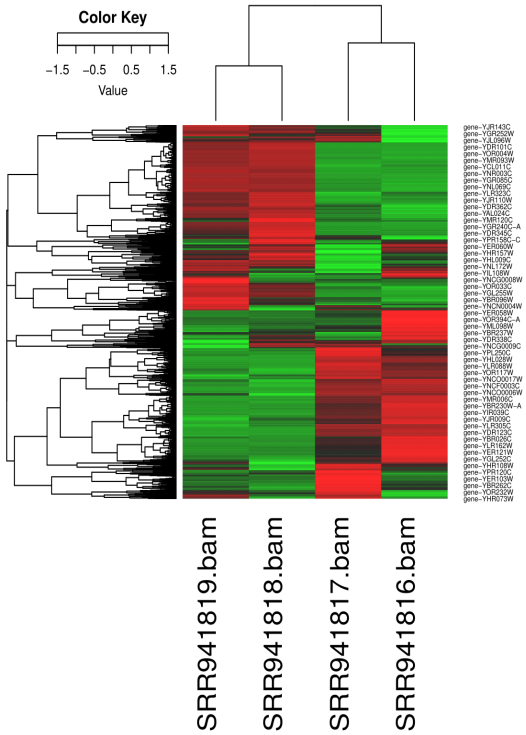

Figure 1: Genes heatmap

YDL063C, YML018C, YHR196W, YNR060W, YPL212C, YBL039C, YLR264C, YJR063W, YDR037W, YNL182C, YDR449C, YMR093W, YLR224W, YJL069C, YKR097W, YBR105C, YNL248C.

4 Discussion

Guided by the obtained heatmap we can establish that most of the genes changed their activity after 30 min of fermentation. There are two situations: initially genes were turned off, but during fermentation they were turned on, and the situation is the opposite.

Consider transmembrane transport, in which 270 of 6486 annotated genes are involved. One type of transport is the transmembrane transport of sugars (GO:0008643), which are needed directly during fermentation. Carboxyl ions are also transported (GO:0015695). These types of transport are directly dependent on fermentation processes within the yeast.

There is also an activation of processes related to transcription and translation in the cell, indicating that the cell is rearranging its metabolism and beginning to express other genes (GO:0006401)

5 References

[1] Aishwarya Gondane and Harri M. Itkonen. "Revealing the History and Mystery of RNA-Seq". In: *Current Issues in Molecular Biology* 45 (3 Feb. 2023), pp. 1860–1874. ISSN: 1467-3045. DOI: 10.3390/cimb45030120.

[2] Muhammad Farooq Rai et al. "Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears". In: *Journal of Orthopaedic Research* 36 (1 Jan. 2018), pp. 484–497. ISSN: 0736-0266. DOI: 10.1002/jor.23661.

[3] Daehwan Kim et al. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype". In: *Nature Biotechnology* 37 (8 Aug. 2019), pp. 907–915. ISSN: 1087-0156. DOI: 10.1038/s41587-019-0201-4.

[4] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15 (12 Dec. 2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

[5] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". In: *GigaScience* 10 (2 Jan. 2021). ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008.

[6] Yang Liao, Gordon K. Smyth, and Wei Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30 (7 Apr. 2014), pp. 923–930. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt656.

[7] Geo Pertea and Mihaela Pertea. "GFF Utilities: GffRead and GffCompare". In: *F1000Research* 9 (Apr. 2020), p. 304. ISSN: 2046-1402. DOI: 10.12688/f1000research.23297.1.

[8] Yeast genome. URL: https://www.yeastgenome.org/goSlimMapper.