Project #6. Differential RNA expression analysis

Evdokimova Anastasiia, Bioinformatics Institute

Abstract

In this work we will learn to analyse data of RNA-seq for differential gene expression analysis. We will study the fermentation process and try to find out how gene expression changes in this process compared to normal growth.

Introduction

All cells of any organism, regardless of their location, function, age, have the same DNA with the same genes. The difference between them is that the expression of genes in cells of one type is clearly different from their expression in cells of another type. This idea also extends to “diseased” and “healthy” cells, or to the same cell in different states (for example, at different stages of the life cycle). Revealing such differences is the purpose of differential expression analysis.

To analyze gene expression, you need to be able to analyze mRNA. There are several ways to do it: qPCR, microarrays, RNA-seq. The last one is the most popular now due to the development of NGS technology and some significant limitations of other methods.

The differential expression analysis consists of several steps: prepare a sequencing library, sequence and data analysis. In this work we try to do the most important and difficult - the last step - on the example of the analysis of gene expression in yeast. Yeast are facultative anaerobes, which means they can switch their metabolism depending on the environmental conditions. When there is plenty of glucose and oxygen available, the yeast cells can use both to create significant amounts of ATP, the main cellular currency. They do this through aerobic respiration in the mitochondria (just like we do). However, when there is an oxygen shortage, they switch to fermentation, the process of converting sugars to acids, gases, or alcohol. We will explore how RNA expression levels change as yeast undergo fermentation to make bread rise. There are two replicates of RNA-seq data from yeast before and during fermentation, and our goal is to find out if the yeast express different genes during fermentation than they do under normal growth.

Methods

For this research we use two replicates of RNA-seq data from yeast before and during fermentation (0 min and 30 min respectively). The data was taken from [1]. As a reference genome we will use *Saccharomyces cerevisiae*, in the genome database at NCBI [2].
To check the quality of reads we use FastQC v0.11.9 [3].
To align reads to the reference sequences was used aligner HISAT2 v2.1.0 [4] and utility Samtools v1.10 for sort output SAM-files [5].
For the next feature count we need to convert a GFF file to a GTF file. For this purpose we use gffread v0.11.7 [6].
For count of features we use featureCounts v2.0.0 [7].
To find differentially expressed genes we use the R package DEseq2 v1.26.0 (R v3.6.0) [8] and ready-made scripts [9].
The results in the output file are sorted by adjusted p-values, so we take the first 50 up regulated genes and first 50 down regulated genes to interpret the results of the whole analysis. To get a sense of what these genes are doing we use the Saccharomyces Genome Database [10], maintained by Stanford, contains all of the gene ontology (GO) terms associated with yeast. GO terms are curated keywords that describe the function of individual genes and can help researchers understand what role they may be playing.

Results

The results of read quality control you can see here [11]. Briefly, we have a good quality of reads in all samples, so we don't need to trim them.
The data about the number of reads and alignment you can see in Table 1. In Table 2 you can see the summary of featureCounts run.

Sample	Total number of reads	Number of aligned reads	Percent of aligned reads
SRR941816 (0min_1)	**9043877**	**8523740**	**94.25**
SRR941817 (0min_2)	**9929568**	**9417842**	**94.85**
SRR941818 (30min_1)	**1721675**	**1655307**	**96.15**
SRR941819 (30min_2)	**6172452**	**5937923**	**96.20**

Table1. Number of reads in each sample and number of aligned reads in each sample.

	SRR941816 (0min_1)	SRR941817 (0min_2)	SRR941818 (30min_1)	SRR941819 (30min_2)
Assigned	7291724	7987002	1402168	4975467
Unassigned_Unmapped	520137	511726	66368	234529
Unassigned_MultiMapping	1299784	1654167	306538	1180651
Unassigned_NoFeatures	611845	626924	102779	380681
Unassigned_Ambiguity	26134	30665	2963	11575

Table 2. Summary of featureCounts in each sample.

The results out DEseq2 work you can find here [11]. These files contain calculated metrics for our genes (result.txt) and normalized counts that we use in visualization (norm-matrix-deseq2.txt). The visualization of the results of our differential expression analysis you can see on Fig. 1.
In total we found 1593 up regulated genes and 1594 down regulated.
Next, to understand what role these genes can play, we use GO (just for top 50 differentially expressed genes). The result of this search you can see in [11] (files GO_down_genes.html and GO_up_genes.html). Briefly, there are 40 up regulated GO terms and 51 down regulated.

Discussion

Our results demonstrate that the yeast transcriptome changes dramatically when changing from aerobic respiration to fermentation. Analysis of the function of the various differentially expressed genes yields insight into the physiological state of yeast cells as the dough fermentation process progresses. Describe some of them. Among down regulated genes there are several, responsible for ion and transmembrane transport. It can be assumed that during fermentation, yeast is in a state of osmotic stress. Generally, when faced with low water activity or high osmolarity, microbes accumulate different compatible solutes, such as ions, amino acids, or polyols, to prevent water loss. Interesting, that among both categories of genes there are genes responsible for response to osmotic stress. Therefore, this issue requires a more detailed study.
Among up regulated genes also there are very interesting results - for example, a lot of genes, connected with protein production (GO terms: rRNA/tRNA processing, ribosomal small/large subunit biogenesis, transcription by RNA polymerase I, ribosome assembly, regulation of translation, cellular amino acid

metabolic process etc). It is possible that the production of certain proteins is a mechanism for adaptation to osmotic stress. There are some works devoted to the study of the effect of osmotic stress in yeast on the processes of transcription and translation [for example 12, 13].

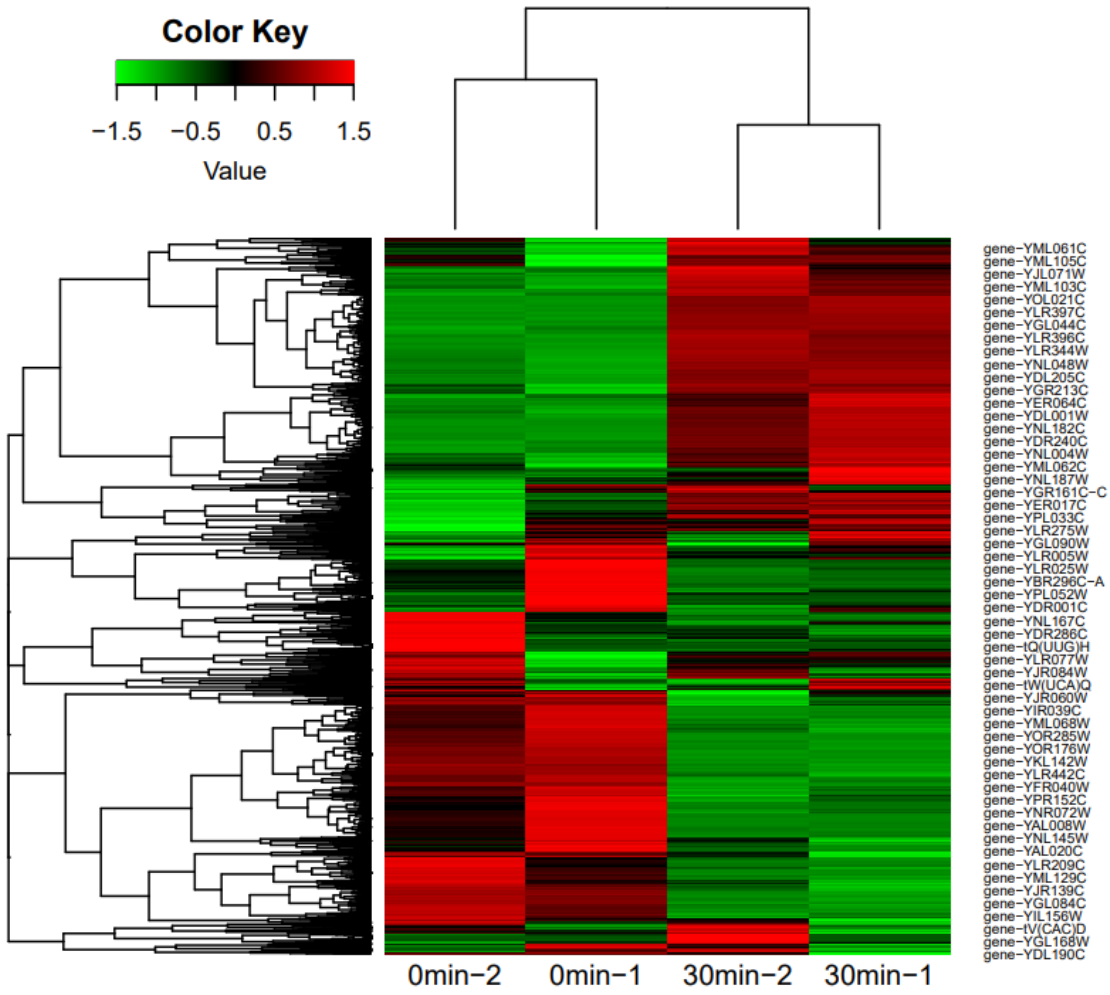

Fig. 1. The heatmap illustrated the results of our differential expression analysis; red - upregulated genes, green - downregulated genes.

Citation

1. RNA-seq data: https://www.ebi.ac.uk/ena/browser/view/PRJNA212389

2. Reference genome and annotation: https://www.ncbi.nlm.nih.gov/assembly/GCA_000146045.2
3. FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
4. HISAT2: http://daehwankimlab.github.io/hisat2/
5. Samtools: http://www.htslib.org/
6. Gffread: https://github.com/gpertea/gffread
7. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30.
8. DEseq2: https://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
9. Scripts for DEseq2: https://figshare.com/articles/software/Scripts_for_RNA-seq_project/14239304
10. Saccharomyces Genome Database: http://www.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl
11. Results: https://drive.google.com/drive/folders/1P2cK52-w4dVacRrhTJpT5ZdxqfvRS6f6?usp=sharing
12. Uesono Y, Toh-e A. 2002. Transient inhibition of translation initiation by osmotic stress. *J. Biol. Chem.* 277:13848–13855.
13. Rep M, Krantz M, Thevelein JM, Hohmann S. 2000. The transcriptional response of Saccharomyces cerevisiae to osmotic shock: Hot1p and Msn2p/Msn4p are required for the induction of subsets of high osmolarity glycerol pathway-dependent genes. *J. Biol. Chem.* 275:8290–8300