Practice

BIOINFORMATICS INSTITUTE

Differential RNA expression analysis

Fedor Grigoryev[1] and Valeriya Rubinova[1]

[1]Bioinformatics Institute

Abstract

Differential expression analysis is a classical method based on the RNA-seq technique, which allows researchers to obtain dynamic results about cell behavior in the ever-changing environment Costa-Silva *et al.* (2017). It allows us to directly observe changes in expression of particular genes. As these changes proved to be adaptive Slotte *et al.* (2007), we as researchers and spectators can perform observations and further studies of the functions of genes involved in that adaptation process. In this paper, we implemented a differential expression pipeline to determine *Saccharomyces cerevisiae* genes that are up- or downregulated during the intricate fermentation process leading to dough rising.

Supplementary materials can be found via the following link: https://github.com/lear-711/Bioinformatics_practice/tree/Project_6.

Keywords: Yeast, RNA-seq, Differential expression

Introduction

Fermentation is a well-known natural process used by humanity for thousands of years with the fundamental purpose of making bread, alcoholic beverages, and many more by letting yeasts and bacteria to feed on the substrate. Upon a strictly biochemical point of view, fermentation is a process of central metabolism in which an organism converts a carbohydrate, such as starch or sugar, into an alcohol or an acid. For example, yeast performs fermentation to obtain energy by converting sugar into alcohol. Fermentation processes were spontaneously carried out before the biochemical process was fully understood. In the 1850s and 1860s, the French chemist and microbiologist Louis Pasteur became the first scientist to study fermentation, when he demonstrated that this process was performed by living cells. Fermentation processes to produce wines, beers and ciders are traditionally carried out with *Saccharomyces cerevisiae* strains, the most common and commercially available yeast.

The fermentation of the dough made by the yeasts is the most critical phase in the making of bread. The fermentative yield of yeast cells during this fermentation is crucial and determines the final quality of the bread. Despite its industrial importance, yeast physiology during solid-state fermentations is relatively unstudied.

There are powerful tools to obtain a snapshot of the cell's transcriptome and thereby gain insight into a cell's physiological state, for instance gene expression analysis approaches such as microarray analysis or transcriptome sequencing (RNA-seq). In this project we used two replicates of RNA-seq data from yeast before and during fermentation to find out the changes in gene expression and hypothesize how they might be a part of actual changes in yeast metabolism or cellular state during the switch from respiration to fermentation in bread dough.

Materials and methods

We analyzed two replicates of RNA-seq data from yeast before and during fermentation. As a reference genome we used Saccharomyces cerevisiae, in the genome database at NCBI.

Table 1 Samples used for analysis with fermentation time and rep number

	Fermentation	Replicate
SRR941816	0 minutes	1
SRR941817	0 minutes	2
SRR941818	30 minutes	1
SRR941819	30 minutes	1

At first we used the program hisat2-align-s (version 2.2.1) Zhang *et al.* (2021) to align experimental data on the reference genome. Then we had to count how many reads map to genomic features. For this purpose it is necessary to use the program featureCounts (version 2.0.3) (with parameter -g geneid) Liao *et al.* (2014). As featureCounts can not work with GFF files, we need to convert the GFF file with annotation to GTF format by using the program gffread (version 0.12.7) (with parameter -T) Pertea and Pertea (2020). After this we used two scripts Raiko (2021) for deseq2 analysis (differential gene expression analysis based on the negative binomial distribution) Love *et al.* (2014). During the result interpretation we used gene ontology terms to understand the function of individual genes GOT (2021).

Results

First we got aligned experimental data on the reference genome. Then we converted downloaded GFF file with annotation to GTF format and used the program featureCounts to assign reads

in alignment files to genes in a genome annotation file. After simplifying the file to the required number of columns, we ran the resulting file through the deseq2.r script and got 2 files as output: the first contains calculated metrics for our genes and the second contains normalised counts that we will use in visualisation in the next step. You can view these files by clicking on the link with supplementary materials in abstract.

Then we drew a heatmap with different gene expression in four experimental samples (Figure 1 in supplementary materials). Here we can see that repetitions of the same sample are quite similar to each other, but different samples differ from each other.

Next, we interpreted the results: from the file that contains calculated metrics for our genes, we took the names of 50 genes that differ most between samples before and during fermentation. The file with the names of the genes was loaded into the generic gene ontology (GO) term mapper for analysis. In supplementary materials you can see a partial result of this analysis: table 2 shows gene GO terms that contain 3 or more genes, the expression of which changes after the start of fermentation. In the table, you can see that a large number of functions are associated with ribosome biogenesis and rRNA processing.

Discussion

According to the Saccharomyces Genome Database, the 48 chosen differentially expressed genes can be classified into several gene ontology terms. Interestingly, the majority of upregulated genes are in some ways involved in synthetic pathways, and could be narrowed into rRNA processing and ribosome biogenesis, ribosome assembly, transcription of ribosomal genes by RNA polymerase I. This can be indicative of a highly active protein synthesis and cell growth, logically suggesting that yeast during the fermentation process have highly active metabolic rate and have to produce a lot of enzymes to maintain it.

Also we notice a number of genes associated with carbohydrate metabolic processes, for example GPP2. In *Saccharomyces cerevisiae* GPP2 is involved in glycerol biosynthesis. In this pathway a glycolytic intermediate, dihydroxyacetonephosphate is converted into glycerol-3 phosphate by Glycerol-3-phosphate dehydrogenase encoded by GDP1 and GDP2. GPP1 and GPP2 encode isoforms of glycerol-3-phosphate phosphatase; these proteins conduct redundant catalytic functions at the last step in this biosynthetic route, the conversion of glycerol-3-phosphate into glycerol. GPP1 and GPP2 are induced under hyperosmotic shock, partially by the High Osmolarity Glycerol (HOG) pathway Martho *et al.* (2019).

Furtermore, there was a group of ion transport genes such as AQR1. This gene codes a cell membrane transporter that mediates excretion of amino acids under conditions in which, despite an abundant availability of carbon, cell growth is limited by a second factor such as the lack of an essential compound. In bread dough fermentation, osmotic stress is likely to be the growth-restrictive factor that leads to a striking upregulation of AQR1 Velasco *et al.* (2004).

Literature cited

2021. Generic gene ontology term mapper.

Costa-Silva J, Domingues D, Lopes FM. 2017. Rna-seq differential expression analysis: An extended review and a software tool. PloS one. 12:e0190152.

Liao Y, Smyth GK, Shi W. 2014. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30:923–930.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology. 15:1–21.

Martho KF, Brustolini OJ, Vasconcelos AT, Vallim MA, Pascon RC. 2019. The glycerol phosphatase gpp2: a link to osmotic stress, sulfur assimilation and virulence in cryptococcus neoformans. Frontiers in Microbiology. 10:2728.

Pertea G, Pertea M. 2020. Gff utilities: Gffread and gffcompare. F1000Research. 9:304.

Raiko M. 2021. Scripts for RNA-seq project. .

Slotte T, Holm K, McIntyre LM, Lagercrantz U, Lascoux M. 2007. Differential expression of genes important for adaptation in capsella bursa-pastoris (brassicaceae). Plant physiology. 145:160–173.

Velasco I, Tenreiro S, Calderon IL, André B. 2004. Saccharomyces cerevisiae aqr1 is an internal-membrane transporter involved in excretion of amino acids. Eukaryotic Cell. 3:1492–1503.

Zhang Y, Park C, Bennett C, Thornton M, Kim D. 2021. Rapid and accurate alignment of nucleotide conversion sequencing reads with hisat-3n. Genome Research. 31:1290–1295.

Supplementary materials

Table 2 GO Terms from the biological process Ontology

GO Term (GO ID)	Genes Annotated to the GO Term	GO Term Usage in Gene List	Genome Frequency of Use
rRNA processing (GO:0006364)	"YDR449C, YEL026W, YER127W, YGR159C, YHR066W, YHR196W, YJL069C, YLR264W, YMR093W, YNL112W, YNL182C, YOL041C, YOL080C"	"13 of 48 genes, 27.08%"	"373 of 6985 annotated genes, 5.34%"
ribosomal large subunit biogenesis (GO:0042273)	"YCR072C, YDL063C, YHR066W, YIR012W, YJL122W, YNL182C, YOL041C, YOL080C"	"8 of 48 genes, 16.67%"	"128 of 6985 annotated genes, 1.83%"
organelle assembly (GO:0070925)	"YCR072C, YGR159C, YHR066W, YIR012W, YLR180W, YLR264W, YNL182C, YOL080C"	"8 of 48 genes, 16.67%"	"214 of 6985 annotated genes, 3.06%"
ribosomal small subunit biogenesis (GO:0042274)	"YDR449C, YEL026W, YER127W, YGR159C, YHR196W, YJL069C, YLR264W, YMR093W"	"8 of 48 genes, 16.67%"	"148 of 6985 annotated genes, 2.12%"
ribosome assembly (GO:0042255)	"YCR072C, YGR159C, YHR066W, YIR012W, YLR264W, YNL182C, YOL080C"	"7 of 48 genes, 14.58%"	"83 of 6985 annotated genes, 1.19%"
transmembrane transport (GO:0055085)	"YDR536W, YHR094C, YJL107C, YJL108C, YKL120W, YNL065W, YOR271C"	"7 of 48 genes, 14.58%"	"491 of 6985 annotated genes, 7.03%"
transcription by RNA polymerase I (GO:0006360)	"YHR196W, YJL148W, YJR063W, YML043C, YMR093W, YNL248C"	"6 of 48 genes, 12.50%"	"78 of 6985 annotated genes, 1.12%"
amino acid metabolic process (GO:0006520)	"YBL039C, YDR037W, YLR180W, YMR300C"	"4 of 48 genes, 8.33%"	"256 of 6985 annotated genes, 3.66%"
nucleobase-containing small molecule metabolic process (GO:0055086)	"YBL039C, YMR300C, YNL141W, YOL136C"	"4 of 48 genes, 8.33%"	"246 of 6985 annotated genes, 3.52%"
carbohydrate metabolic process (GO:0005975)	"YBR105C, YER062C, YKR097W, YOL136C"	"4 of 48 genes, 8.33%"	"291 of 6985 annotated genes, 4.17%"
RNA catabolic process (GO:0006401)	"YGR159C, YLR264W, YNL112W, YOR359W"	"4 of 48 genes, 8.33%"	"180 of 6985 annotated genes, 2.58%"
monoatomic ion transport (GO:0006811)	"YDR536W, YHR094C, YNR060W, YOR271C"	"4 of 48 genes, 8.33%"	"235 of 6985 annotated genes, 3.36%"
nuclear transport (GO:0051169)	"YDL063C, YHR196W, YLR264W"	"3 of 48 genes, 6.25%"	"191 of 6985 annotated genes, 2.73%"
protein transport (GO:0015031)	"YBR105C, YDL063C, YOR359W"	"3 of 48 genes, 6.25%"	"646 of 6985 annotated genes, 9.25%"
mRNA processing (GO:0006397)	"YEL026W, YGR159C, YPL212C"	"3 of 48 genes, 6.25%"	"241 of 6985 annotated genes, 3.45%"
regulation of translation (GO:0006417)	"YLR264W, YNL112W, YOR359W"	"3 of 48 genes, 6.25%"	"240 of 6985 annotated genes, 3.44%"

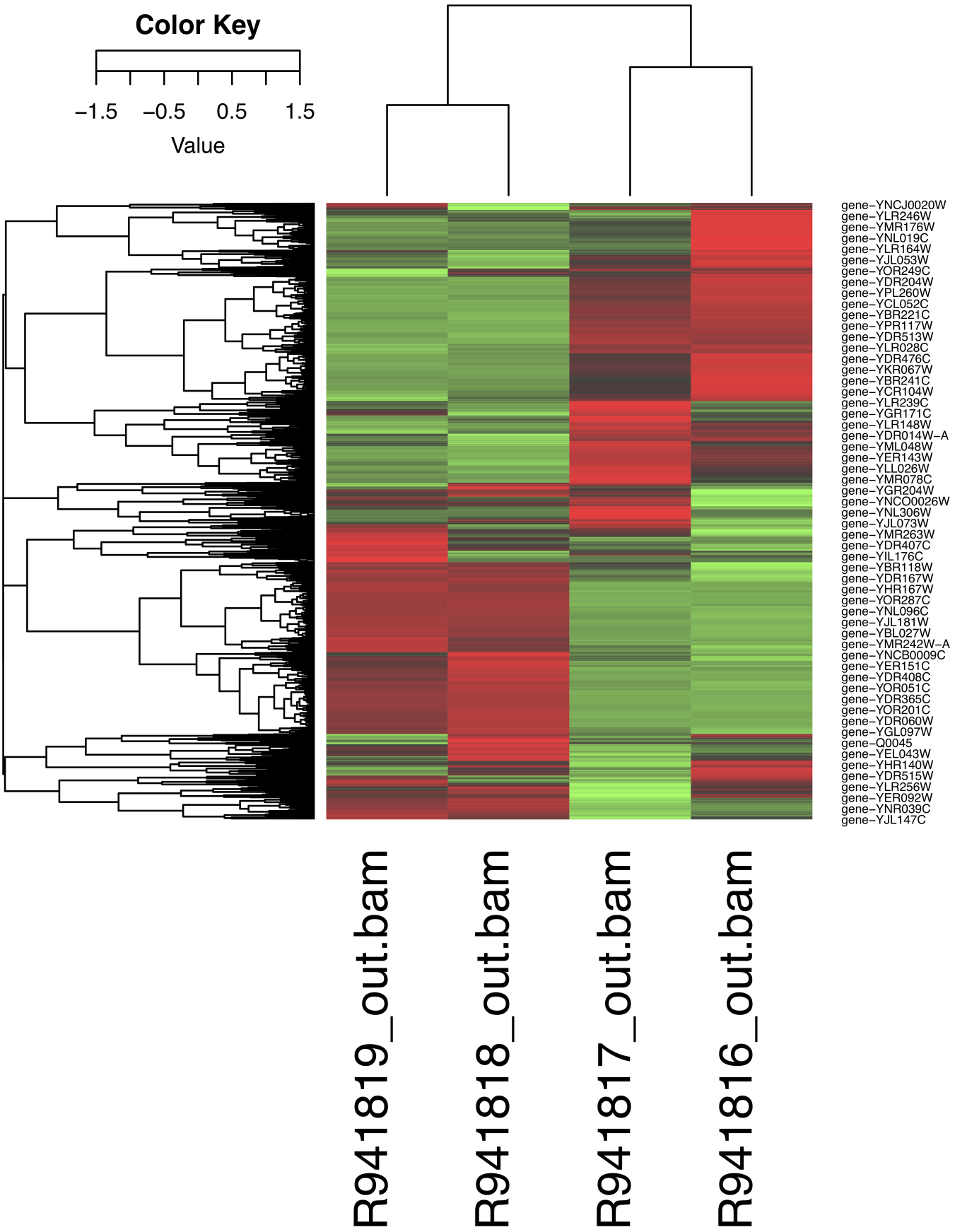

Figure 1 Expression clustering tree in experimental samples