Practice

BIOINFORMATICS INSTITUTE

Search for the fermentation-related genes in *Saccharomyces cerevisiae*

Rustam Basyrov[1] and Anna Chechenina[2]

[1]Department 305, Moscow Aviation Institute
[2]Bioinformatics department, Escola Superior de Comerc International - Universitat Pompeu Fabra (ESCI-UPF), Barcelona, 08003, Spain

Abstract

There are many problems in biology that require comparison of genes' expression. Analysis of RNA-seq sequencing is a traditional way of comparing gene expression and identifying the ones with significantly increased or decreased activity. In this study, we decided to check which genes are being activated and actively expressed during the process of fermentation of yeast *Saccharomyces cerevisiae*. To do this we analyzed the open data of RNA-sequencing of *S. cerevisiae* before and after 30 minutes of fermentation. We found out that the expression level changed significantly for 23 genes (log2 Fold Change > 5 and p-adjusted > 0.0001). Moreover, many genes related to glycolysis and some other processes showed a significant increase in expression.
Supplementary materials can be found via the link: https://github.com/checheanya/BI_git.

Keywords: yeast; RNA-seq; fermentation; differential expression

Introduction

All biological processes require the activity of certain genes typical for this process. Along with genes, coding main proteins in the process there are many regulatory genes (for example, coding transcription factors) activated. The task of finding these genes is common in modern biology and there are many ways of identifying the level of gene activity: for example, *in situ* hybridization, Northern blot, RT-PCR, and analysis of the microarray data. However, these methods are very time and resource-consuming, while giving results for only several genes at the same time. The most used approach now is a differential expression analysis based on RNA-sequencing data or, which is becoming less and less popular, based on microarray data Costa-Silva *et al.* (2017).

There are several main problems arising during the differential expression analysis Conesa *et al.* (2016). First, even before the sequencing we have to choose control genes to compare an expression to. These genes should have a moderate constant expression not depending on the conditions we are checking in the experiment. This is an essential step because the number of copies we can see in the sample can vary depending on the sample purification, library preparation, and other manual steps. One more problem is multiple testing Soneson and Delorenzi (2013). Since we are measuring the expression of many genes, the probability to get a significant result just by chance is high, thus we have to apply corrections (for instance, Bonferonni, FDR, Holm). However, these corrections might hide some important genes.

In this study, we tried to identify genes, involved in the fermentation process of baking yeast. In order to do this we performed a differential expression analysis of the *Saccharomyces cerevisiae* RNA-sequencing before and after 30 minutes of fermentation.

Yeasts are the eukaryotic organisms widely involved in the biotechnology processes that require fermentation. Since yeasts are facultative anaerobes they can switch from anaerobic to anaerobic metabolism depending on the environmental conditions. In particular, when there is an oxygen shortage and they are not able to carry out normal respiration, yeast are switching to the fermentation process. The general fermentation process consists of converting a carbohydrate, such as starch or sugar, into an alcohol or an acid. As we know, the fermentation of yeast includes several stages and basically performs the conversion into ethanol and carbon dioxide Maicas (2020); Walker and Walker (2018). Yeasts are characterised by their capacity to ferment a high spectrum of sugars, among which glucose, fructose, sucrose, maltose, and maltotriose predominate, found both in ripe fruit and in processed cereals Maicas (2020).

This process is the keystone in biotechnological production like bakery, wine, and beer manufacturing Maicas (2020); Eldarov *et al.* (2016). During baking, carbon dioxide performs as a leaving agent, making the dough rise as the gas forms pockets or bubbles Maicas (2020).

Saccharomyces cerevisiae strains are the most common and commercially available yeast for the biotechnology industry Maicas (2020); Eldarov *et al.* (2016) and thus it is highly important to understand the gene biology behind the fermentation process in order to make production more efficient. Therefore in our study, we will try to understand which genes are being highly expressed during the fermentation and which are down-regulated. This information could be useful in further biotechnological research to improve the production strains, for example by changing promotors of the genes of interest or by doing protein engineering for better enzyme performance.

Table 1 Number of reads

Sample	Total number of reads	Alignment rate
0 minutes (1)	9043877	94.30%
0 minutes (2)	9929568	94.91%
30 minutes (1)	1721675	96.18%
30 minutes (2)	6172452	96.24%

Methods

Raw data

For this study, we used four datasets representing RNA-seq data from yeast before and after half an hour of fermentation. The row data reads were obtained from the EMLB EBI Short Reads Archive database through the following links: SRR941816, fermentation 0 minutes, replicate 1; SRR941817, fermentation 0 minutes replicate 2; SRR941816, fermentation 30 minutes replicate 1; SRR941816, fermentation 30 minutes replicate 2.

As a reference genome, we used *Saccharomyces cerevisiae* strain S288c genome and assembly R64, which could be found here: genome, annotation.

Main pipeline

To align our reads to the reference genome we used HISAT2 tool with default parameters (Kim *et al.* 2019). First, using *index* command we indexed the reference genome and then ran HISAT2 in single-end mode using *hisat2* command. After that alignments were sorted by samtools (Li *et al.* 2009).

To collect information about genes at alignments featureCounts Liao *et al.* (2013) tool used. But it can't work with gff-files. For this reason annotation file was converted by gffread (G and M 2020) told with '-T' parameter. For differential expression estimation DESeq2 (Michael Love 2017) library for R (R Core Team 2022) was used. Also, a heatmap was drawn by R script (Raiko 2021).

For the analysis of the deferentially expressed genes, we used the DESeq2 package for R (Michael Love 2017).

Annotation

For the annotation of the gene ontology, we used GOTermMapper (Ashburner *et al.* 2000; and Seth Carbon *et al.* 2020) tool. After the gene ontology exploration, we performed the analysis of the metabolic pathways among genes in our results. For the pathways enrichment analysis, we used GATOM tool, which is based on the KEGG pathways data (Emelianova *et al.* 2022).

Results

In order to infer genes from the given reads we aligned our reads files to the reference genome. As we can see in the Table 1, all samples showed a high percentage of mapped reads.

After performing reads mapping, we performed a differential expression analysis of the data, taking samples from the zero timepoint as a reference and a 30-minutes timepoint as an experiment. In the Figure 1 we can see that there are many genes with expression level Fold Change higher than 5. However, way more genes have increased expression rather than decreased. There are 250 genes with $\log_2$ Fold Change > 2 and p-adjusted > 0.001 and 14 genes with both $\log_2$ Fold Change > 5 and p-adjusted $>$ 10e-32, which seems to be quite significant (see Figure 1, 2).

We selected the top-50 genes by p-adjusted and performed gene ontology search on them. The result of this analysis could be found in the Table and Figure 4. We can see that rRNA processing process in highly affected and involves 13 genes out of 50 explored. Also, there are three gene ontology terms, with more then 5 genes involved, which are related to the ribosome assembly.

Since changes in metabolic pathways seem to be the key issue in the analysis related to fermentation, we also explored the pathways enrichment in the set of genes with altered expression. As we can see in the Figure 3 and in the Table 2, there were dramatic changes in the processes like glyconeogenesis, glycolysis, and secondary metabolites biosynthesis.

Discussion

High-throughput data from RNA-Seq indicated that protein synthesis and transport were the core regulatory biological processes that were up-regulated in yeast in response to the lack of oxygen and the beginning of fermentation process. We can assume that such a dramatic change in the metabolic strategy should be accompanied by an increase in the activity of a large amount of transcription factors. These transcription factors could help to switch needed genes on and off and preserve the active/inactive state of the involved genes. This hypothesis is supported by the fact that top gene ontology terms we found were related to RNA processing and ribosomal function.

Amongst down-regulated GO results there are generation of precursor metabolites and energy (GO:0006091), monocarboxylic acid metabolic process (GO:0032787) and carbohydrate metabolic process (GO:0005975). Indeed, these processes are related to aerobic metabolism, so they should be down-regulated after switching to the anaerobic type of metabolism.

If we look at the metabolic analysis we can see that many pyruvate-converting genes are down-regulated, whereas sugar-converting enzymes are on the opposite - up-regulated. The fermentation process exploits many of these enzymes (for instance, glucokinase1, GLK1) to convert them to more easily degradable biochemical units.

We explored further four genes that showed the highest levels of p-adjusted and log2 Fold Change: YER062C, YDR536W, YHR094C, and YNL065W.

YER062C gene codes DL-glycerol-3-phosphate phosphatase. This enzyme is involved in the glycerol biosynthesis. It was shown that the expression of this gene is induced in response to hyperosmotic or oxidative stress Påhlman *et al.* (2001). In addition, this protein is responsible for the glycerol transport and accumulation, which was shown to serve as a nontoxic redox sink during fermentation Norbeck *et al.* (1996).

YDR536W (STL) gene codes glycerol proton symporter of the plasma membrane. It was shown to be glucose-induced inactivated and induced during the osmotic shock. As we mentioned above, glycerol as well as other polyols can be used as an osmoprotectant in yeasts and other organisms. Therefore, regulation of its transport and synthesis is essential, especially, in anaerobic conditions Ferreira *et al.* (2005).

The protein coded by YHR094C is the low-affinity glucose transporter. Expression of the HXT1 gene, which encodes a low affinity glucose transporter in Saccharomyces cerevisiae, is regulated positively in response to glucose by the general glucose induction pathway, involving the Snf3/Rgt2 membrane glucose sensors, the SCF-Grr1 ubiquitination complex and the Rgt1 transcription factor (s Cobos *et al.* 2004).

YNL065W codes a probable transporter that confers resistance to short-chain monocarboxylic acids and quinidine (Tenreiro *et al.* 2002). YNL065W is important for a moderate increase of yeast tolerance to ketoconazole and to the cationic dye crystal violet; it protects the cell against short-chain monocarboxylic acids (C(2)-C(6)).

Overall, we can say that the switch to fermentation is a fundamental change that involves many genes, related to the core metabolic pathways. We showed that among more than 200 genes that dramatically changed their expression, many are related to the processes of glucose, fructose, and other sugars' conversion, which appears to participate in the first stages of the fermentation pathway.

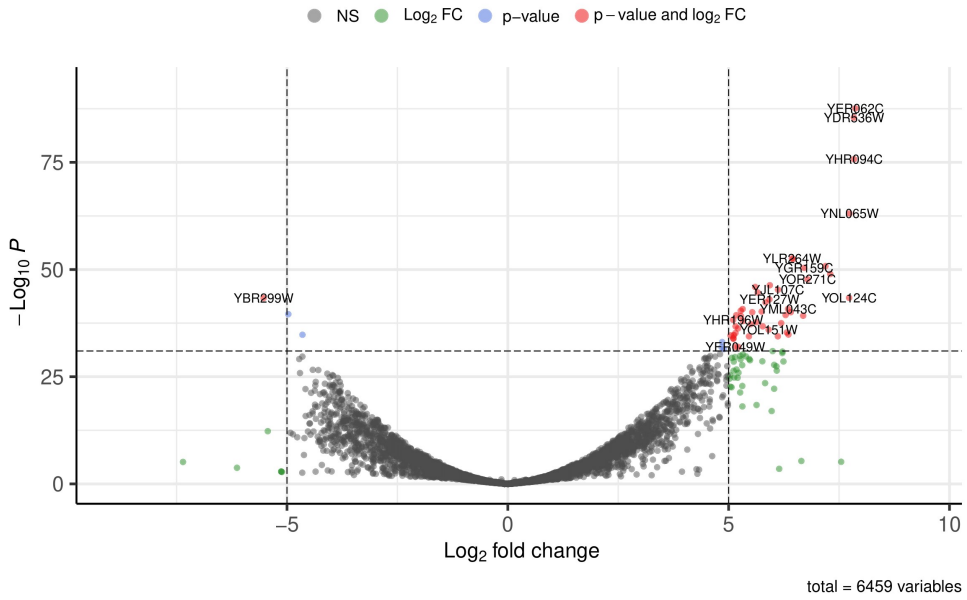

Figure 1 Distribution of the p-adjusted and log Fold Change values for genes expression. Thresholds: $\log_2$ FoldChange $= 5$, p-adjusted $= 10^{-32}$.

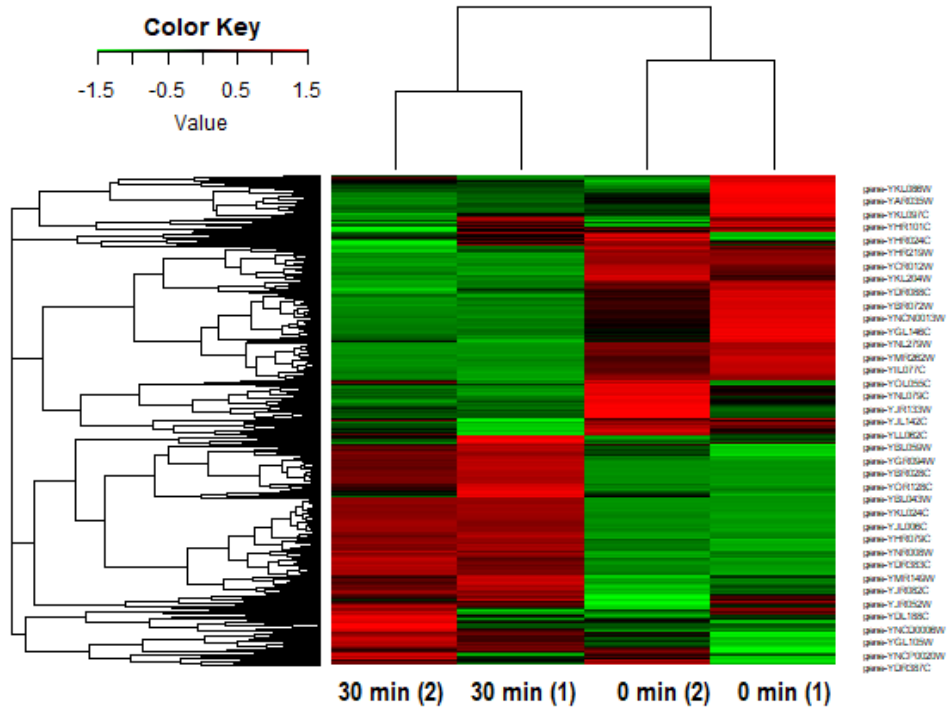

Figure 2 Heatmap for the three groups of samples comparing levels of expression. Red strips - decreased expression, green - increased.

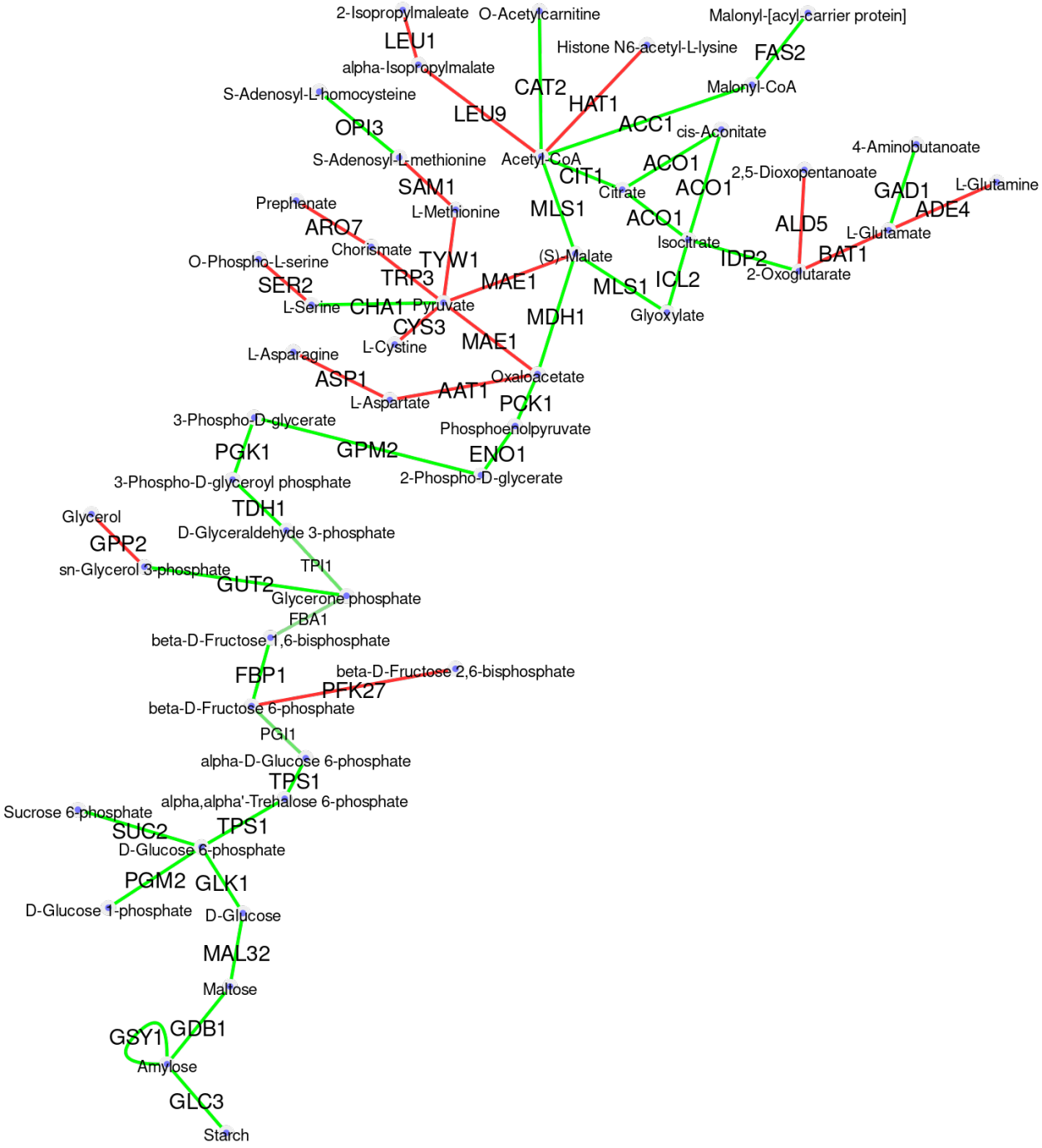

Figure 3 Map of the metabolic pathways for the genes with altered expression level.

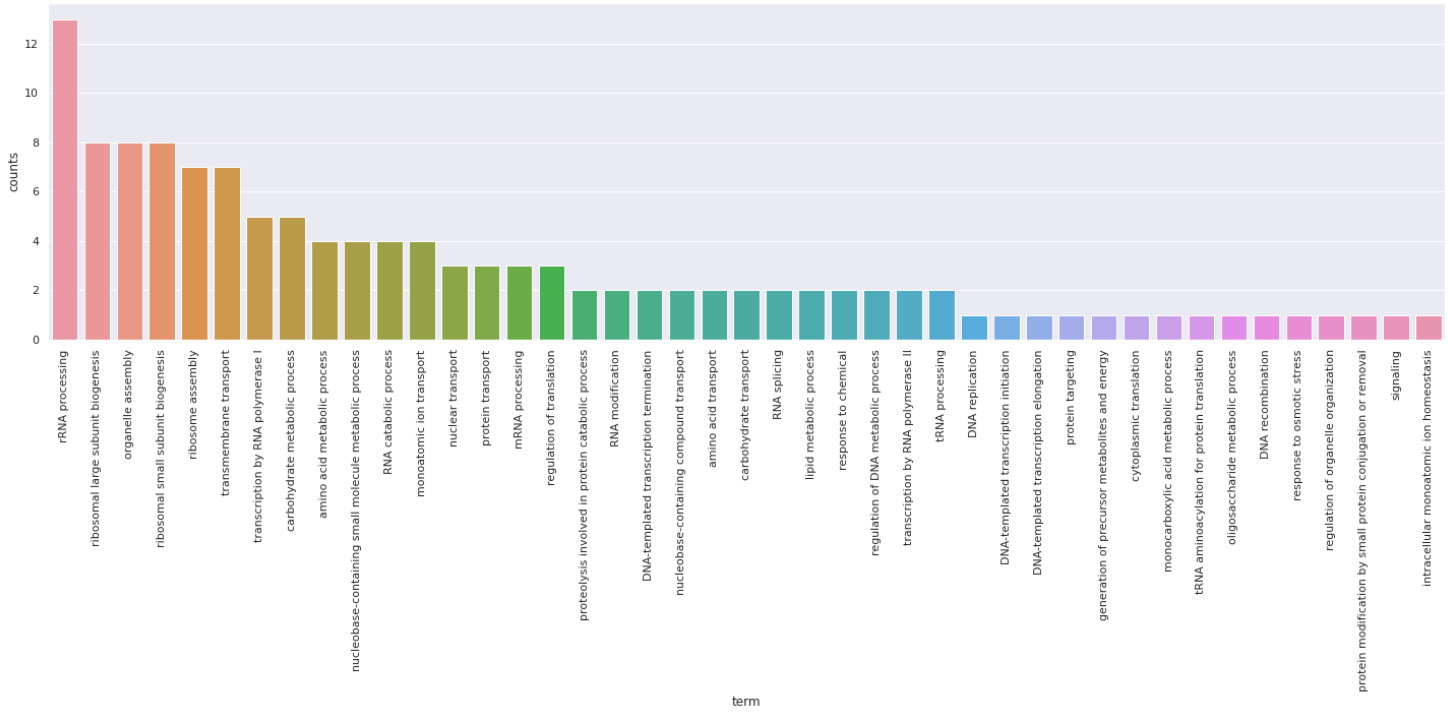

Figure 4 GO terms distribution among top-50 genes with changed expression.

Table 2 Enrichment analysis - the most enriched pathways.

pathway	pval	padj	overlap	size	overlap Genes
sce_M00003: Gluconeogenesis, oxaloacetate => fructose-6P	3.57e-08	2.89e-06	8	8	YKR097W, YGR254W, YDL021W, YCR012W, YJL052W, YDR050C, YKL060C, YLR377C
sce_M00001: Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	2.91e-07	1.18e-05	8	9	YCL040W, YBR196C, YKL060C, YDR050C, YJL052W, YCR012W, YDL021W, YGR254W
sce01110: Biosynthesis of secondary metabolites	9.59e-07	2.04e-05	39	189	YAL012W, YBR126C, YBR196C, YCL040W, YCL064C, YCR012W, YDL021W, YDR050C, YDR321W, YEL011W, YER073W, YFR015C, YGL009C, YGR208W, YGR254W, YHR208W, YIL155C, YIL162W, YJL052W, YJR073C, YKL060C, YKL085W, YKL106W, YKL211C, YKR097W, YLR174W, YLR180W, YLR304C, YLR377C, YMR105C, YMR250W, YMR300C, YNL117W, YNR001C, YNR016C, YOR108W, YPR006C, YPR060C, YPR184W
sce00010: Glycolysis / Gluconeogenesis	1.01e-06	2.04e-05	12	22	YBR196C, YCL040W, YCR012W, YDL021W, YDR050C, YER073W, YGR254W, YJL052W, YKL060C, YKR097W, YLR377C, YMR105C
sce00500: Starch and sucrose metabolism	4.65e-06	7.54e-05	9	14	YBR126C, YBR196C, YBR299W, YCL040W, YEL011W, YFR015C, YIL162W, YMR105C, YPR184W
sce_M00012: Glyoxylate cycle	2.58e-05	3.49e-04	5	5	YNR001C, YLR304C, YPR006C, YNL117W, YKL085W
sce_M00002: Glycolysis, core module involving three-carbon compounds	1.40e-04	1.62e-03	5	6	YDR050C, YJL052W, YCR012W, YDL021W, YGR254W
sce_M00165: Reductive pentose phosphate cycle (Calvin cycle)	2.74e-03	2.78e-02	4	6	YCR012W, YJL052W, YKL060C, YLR377C
sce00051: Fructose and mannose metabolism	3.96e-03	3.56e-02	5	10	YCL040W, YDR050C, YKL060C, YLR377C, YOL136C

Table 3 GO terms for top-50 genes

GO term	GO ID	Number of genes	Genes
rRNA processing	GO:0006364	13 of 48 genes (27.08%)	"YDR449C, YEL026W, YER127W, YGR159C, YHR066W, YHR196W, YJL069C, YLR264W, YMR093W, YNL112W, YNL182C, YOL041C, YOL080C"
ribosomal large subunit biogenesis	GO:0042273	8 of 48 genes (16.67%)	"YCR072C, YDL063C, YHR066W, YIR012W, YJL122W, YNL182C, YOL041C, YOL080C"
organelle assembly	GO:0070925	8 of 48 genes (16.67%)	"YCR072C, YGR159C, YHR066W, YIR012W, YLR180W, YLR264W, YNL182C, YOL080C"
ribosomal small subunit biogenesis	GO:0042274	8 of 48 genes (16.67%)	"YDR449C, YEL026W, YER127W, YGR159C, YHR196W, YJL069C, YLR264W, YMR093W"
ribosome assembly	GO:0042255	7 of 48 genes (14.58%)	"YCR072C, YGR159C, YHR066W, YIR012W, YLR264W, YNL182C, YOL080C"
transmembrane transport	GO:0055085	7 of 48 genes (14.58%)	"YDR536W, YHR094C, YJL107C, YJL108C, YKL120W, YNL065W, YOR271C"
transcription by RNA polymerase I	GO:0006360	5 of 48 genes (10.42%)	"YHR196W, YJL148W, YJR063W, YML043C, YMR093W"
carbohydrate metabolic process	GO:0005975	5 of 48 genes (10.42%)	"YBR105C, YBR299W, YER062C, YKR097W, YOL136C"
amino acid metabolic process	GO:0006520	4 of 48 genes (8.33%)	"YBL039C, YDR037W, YLR180W, YMR300C"
nucleobase-containing small molecule metabolic process	GO:0055086	4 of 48 genes (8.33%)	"YBL039C, YMR300C, YNL141W, YOL136C"
RNA catabolic process	GO:0006401	4 of 48 genes (8.33%)	"YGR159C, YLR264W, YNL112W, YOR359W"
monoatomic ion transport	GO:0006811	4 of 48 genes (8.33%)	"YDR536W, YHR094C, YNR060W, YOR271C"
nuclear transport	GO:0051169	3 of 48 genes (6.25%)	"YDL063C, YHR196W, YLR264W"
protein transport	GO:0015031	3 of 48 genes (6.25%)	"YBR105C, YDL063C, YOR359W"
mRNA processing	GO:0006397	3 of 48 genes (6.25%)	"YEL026W, YGR159C, YPL212C"
regulation of translation	GO:0006417	3 of 48 genes (6.25%)	"YLR264W, YNL112W, YOR359W"
proteolysis involved in protein catabolic process	GO:0051603	2 of 48 genes (4.17%)	"YBR105C, YLR224W"
RNA modification	GO:0009451	2 of 48 genes (4.17%)	"YOL124C, YPL212C"
DNA-templated transcription termination	GO:0006353	2 of 48 genes (4.17%)	"YJR063W, YNL112W"
nucleobase-containing compound transport	GO:0015931	2 of 48 genes (4.17%)	"YHR196W, YLR264W"
amino acid transport	GO:0006865	2 of 48 genes (4.17%)	"YNL065W, YOR271C"
carbohydrate transport	GO:0008643	2 of 48 genes (4.17%)	"YDR536W, YHR094C"
RNA splicing	GO:0008380	2 of 48 genes (4.17%)	"YEL026W, YGR159C"
lipid metabolic process	GO:0006629	2 of 48 genes (4.17%)	"YBL039C, YOL151W"
response to chemical	GO:0042221	2 of 48 genes (4.17%)	"YLR224W, YOR271C"

regulation of DNA metabolic process	GO:0051052	2 of 48 genes (4.17%)	"YNL182C, YOR359W"
transcription by RNA polymerase II	GO:0006366	2 of 48 genes (4.17%)	"YJR063W, YNL112W"
tRNA processing	GO:0008033	2 of 48 genes (4.17%)	"YOL124C, YPL212C"
DNA replication	GO:0006260	1 of 48 genes (2.08%)	YNL182C
DNA-templated transcription initiation	GO:0006352	1 of 48 genes (2.08%)	YML043C
DNA-templated transcription elongation	GO:0006354	1 of 48 genes (2.08%)	YJL148W
protein targeting	GO:0006605	1 of 48 genes (2.08%)	YBR105C
generation of precursor metabolites and energy	GO:0006091	1 of 48 genes (2.08%)	YOL136C
cytoplasmic translation	GO:0002181	1 of 48 genes (2.08%)	YLR264W
monocarboxylic acid metabolic process	GO:0032787	1 of 48 genes (2.08%)	YOL136C
tRNA aminoacylation for protein translation	GO:0006418	1 of 48 genes (2.08%)	YDR037W
oligosaccharide metabolic process	GO:0009311	1 of 48 genes (2.08%)	YBR299W
DNA recombination	GO:0006310	1 of 48 genes (2.08%)	YGR159C
response to osmotic stress	GO:0006970	1 of 48 genes (2.08%)	YER062C
regulation of organelle organization	GO:0033043	1 of 48 genes (2.08%)	YLR180W
protein modification by small protein conjugation or removal	GO:0070647	1 of 48 genes (2.08%)	YLR224W
signaling	GO:0023052	1 of 48 genes (2.08%)	YOR360C
intracellular monoatomic ion homeostasis	GO:0006873	1 of 48 genes (2.08%)	YNR060W

Literature cited

and Seth Carbon, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, Basu S, Chisholm RL, Dodson RJ, Hartline E *et al*. 2020. The gene ontology resource: enriching a GOld mine. Nucleic Acids Research. 49:D325–D334.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*. 2000. Gene ontology: tool for the unification of biology. Nature Genetics. 25:25–29.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X *et al*. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology. 17.

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

Eldarov MA, Kishkovskaia SA, Tanaschuk TN, Mardanov AV. 2016. Genomics and biochemistry of saccharomyces cerevisiae wine yeast strains. Biochemistry (Moscow). 81:1650–1668.

Emelianova M, Gainullina A, Poperechnyi N, Loboda A, Artyomov M, Sergushichev A. 2022. Shiny GATOM: omics-based identification of regulated metabolic modules in atom transition networks. Nucleic Acids Research. 50:W690–W696.

Ferreira C, van Voorst F, Martins A, Neves L, Oliveira R, Kielland-Brandt MC, Lucas C, Brandt A. 2005. A member of the sugar transporter family, stl1p is the glycerol/hsup/supsymporter inisaccharomyces cerevisiae/i. Molecular Biology of the Cell. 16:2068–2076.

G P, M P. 2020. Gff utilities: Gffread and gffcompare [version 1; peer review: 3 approved]. .

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 37:907–915.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Liao Y, Smyth GK, Shi W. 2013. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30:923–930.

Maicas S. 2020. The role of yeasts in fermentation processes. Microorganisms. 8:1142.

Michael Love SA. 2017. Deseq2.

Norbeck J, Påhlman AK, Akhtar N, Blomberg A, Adler L. 1996. Purification and characterization of two isoenzymes of DL-glycerol-3-phosphatase from saccharomyces cerevisiae. Journal of Biological Chemistry. 271:13875–13881.

Påhlman AK, Granath K, Ansell R, Hohmann S, Adler L. 2001. The yeast glycerol 3-phosphatases gpp1p and gpp2p are required for glycerol biosynthesis and differentially involved in the cellular responses to osmotic, anaerobic, and oxidative stress. Journal of Biological Chemistry. 276:3555–3563.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

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

s Cobos L, é L, Mas G, Sanz P, Posas F. 2004. Expression of the HXT1 low affinity glucose transporter requires the coordinated activities of the HOG and glucose signalling pathways. J Biol Chem. 279:22010–22019.

Soneson C, Delorenzi M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 14.

Tenreiro S, Nunes PA, Viegas CA, Neves MS, Teixeira MC, Cabral MG, Correia I. 2002. AQR1 gene (ORF YNL065w) encodes a plasma membrane transporter of the major facilitator superfamily that confers resistance to short-chain monocarboxylic acids and quinidine in Saccharomyces cerevisiae. Biochem Biophys Res Commun. 292:741–748.

Walker GM, Walker RS. 2018. Enhancing yeast alcoholic fermentations, In: , Elsevier. pp. 87–129.