ANALYSIS OF GENES EXPRESSED IN SACCHAROMYCES CEREVISIAE IN AEROBIC AND ANAEROBIC CONDITIONS

SUBHAM DAS - 18399

Indian Institute of Science Education and Research, Bhopal

***Abstract*- RNA-seq of Baker's Yeast was analyzed using bioinformatics tools developed over the years to study differential gene expression under two (or more) different conditions. In this study, we analyze the gene expression under aerobic and anaerobic conditions. We carefully study two differentially expressed genes, THI72 and YPS1, which change in their expression levels going from one condition to another. We delve into the details of the procedure and observe how the files are being analyzed, and then we visually study the output of our two genes of interest to understand their functioning. We discuss the output and future implications of our observations.**

PURPOSE: To analyze the differentially expressed genes in baker's yeast (saccharomyces cerevisiae) and study their trend of expression by identifying two significant genes of interest of which one is upregulated and the other downregulated in the fermentation process.

I. INTRODUCTION

Here we describe a transcriptomic analysis of fermentation of Saccharomyces cerevisiae. The dataset was obtained from an experiment where fermentation was carried out in carefully controlled environmental conditions in a bioreactor to reduce transcriptomic responses that would be due to factors other than the oxygen levels. Further, we analyze two genes of interest - **THI72** and **YPS1**, the former being up-regulated and the latter being down-regulated from aerobic to anaerobic condition. Using the protocol of **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks** by Trapnell et al., we compare gene and transcript expression under the two given conditions. Finally, CummeRbund - a tool for visualizing RNA-seq analysis results, Integrative Genomics Viewer (IGV) - to observe the coverage of a gene (whether up or down-regulated) in the two conditions, and Gene Ontology Profiling - enrichment analysis to see how the differentially expressed genes contribute in different biological processes, molecular functions, cellular compartments, and pathways they are involved in.

II. METHODS USED TO ANALYZE RNA-seq AND PIPELINE OF PROCEDURE

RNA-seq experiments must be analyzed with robust, efficient and statistically principled algorithms. In view of this the following tools have been utilized to study our dataset:

1. Tophat2
2. Bowtie2
3. Cufflinks (cuffmerge, cuffdiff)
4. CummeRbund library in R
5. Integrative Genome browser (IGV)
6. Gene Ontology Profiling (g:profiler)

The RNA-seq dataset was obtained by NextSeq 500 paired end sequencing. Further details of the dataset include:

Platform - Illumina

Model - NextSeq 500

Layout - Paired

Source – Transcriptomic

Experiment Accession – SRX3871243 Strategy – RNA-seq

The dataset (FASTQ files) was obtained from ENA browser, extracted and analysis performed.

The following processes/commands were run and relevant output files obtained:

1. TopHat - Reads for each condition are mapped to the reference genome. After running TopHat, the resulting alignment files are provided to Cufflinks.

   OUTPUT: acceptedhits.bam (BAM files)

2. Cufflinks - Used to generate a transcriptome assembly for each condition.

   OUTPUT: Transcripts (GTF files)

3. Cuffmerge - Assemblies are then merged together using the Cuffmerge utility, which is included with the Cufflinks package. This merged assembly provides a uniform basis for calculating gene and transcript expression in each condition.

   OUTPUT: merged.gtf (GTF file)

4. Cuffdiff - The reads and the merged assembly are fed to Cuffdiff, which calculates expression levels and tests the statistical significance of observed changes. Cuffdiff also performs an additional layer of differential analysis. By grouping transcripts into biologically meaningful groups (such as transcripts that share the same transcription start site (TSS)), Cuffdiff identifies genes that are differentially regulated at the transcriptional or post-transcriptional level.

   OUTPUT: diff_out folder with FPKM_TRACKING, COUNT_TRACKING, READ_TRACKING, and DIFF files. These are a set of tabular files which can be indexed and visualized in Cummebund.

5. CummeRbund - Provides functions for creating commonly used expression plots such as volcano, scatter and box plots.

The entire process pipeline can be summarized by the flowchart:

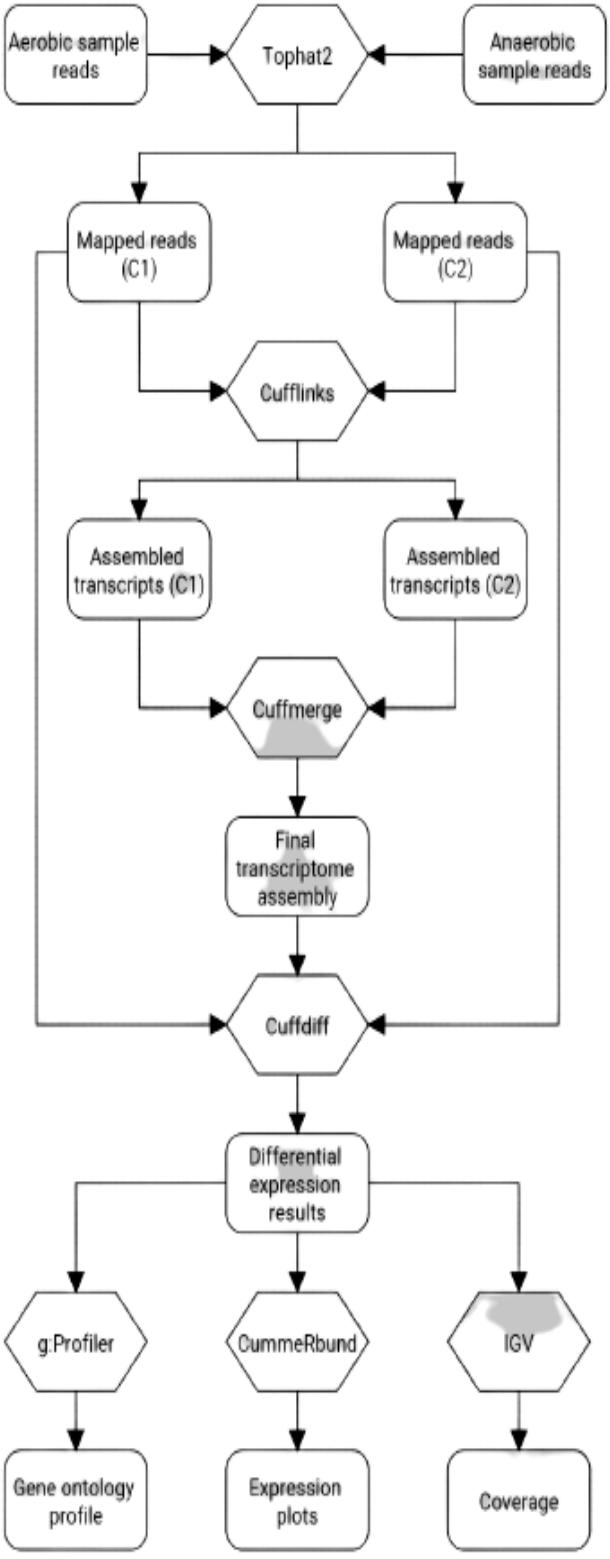


Fig.1 | An overview of the protocol.

III. RESULTS

The analysis provided a list of 6516 differentially expressed genes of which 1093 were significantly expressed ($P < 0.05$) with abs(log2_fold_change) > 1.

```
> cuff_data
CuffSet instance with:
         2 samples
         6516 genes
         8057 isoforms
         7741 TSS
         6643 CDS
         6516 promoters
         7741 splicing
         6050 relCDS
> |
```

Fig.2 | Total expressed genes

Other significant data output of coding sequence (CDS), regulated genes (relCDS), isoform, promoter, splicing, and transcription start site (TSS) is as follows:

```
> library(cummeRbund)
> cuff_data <- readCufflinks('diff_out')
> gene_diff_data <- diffData(genes(cuff_data))
> sig_gene_data <- subset(gene_diff_data, (significant == 'yes'))
> nrow(sig_gene_data)
[1] 3893
> isoform_diff_data <-diffData(isoforms(cuff_data), 'C1', 'C2')
> sig_isoform_data <-subset(isoform_diff_data,(significant == 'yes'))
> nrow(sig_isoform_data)
[1] 4333
> tss_diff_data <-diffData(TSS(cuff_data), 'C1', 'C2')
> sig_tss_data <- subset(tss_diff_data, (significant == 'yes'))
> nrow(sig_tss_data)
[1] 4304
> cds_diff_data <- diffData(CDS(cuff_data), 'C1', 'C2')
> sig_cds_data <- subset(cds_diff_data, (significant == 'yes'))
> nrow(sig_cds_data)
[1] 4007
> promoter_diff_data <- distValues(promoters(cuff_data))
> sig_promoter_data <- subset(promoter_diff_data, (significant == 'yes'))
> nrow(sig_promoter_data)
[1] 232
> splicing_diff_data <- distValues(splicing(cuff_data))
> sig_splicing_data <- subset(splicing_diff_data, (significant == 'yes'))
> nrow(sig_splicing_data)
[1] 31
> relCDS_diff_data <- distValues(relCDS(cuff_data))
> sig_relCDS_data <- subset(relCDS_diff_data, (significant == 'yes'))
> nrow(sig_relCDS_data)
[1] 132
> sig_gene_data <- subset(gene_diff_data, (significant == 'yes'), abs(log2_fold_change) > 1)
> nrow(sig_gene_data)
[1] 3893
> final_sig_gene_data <- subset(sig_gene_data, abs(log2_fold_change)>1)
> nrow(final_sig_gene_data)
[1] 1093
> |
```

Fig.3 | Significantly expressed genes

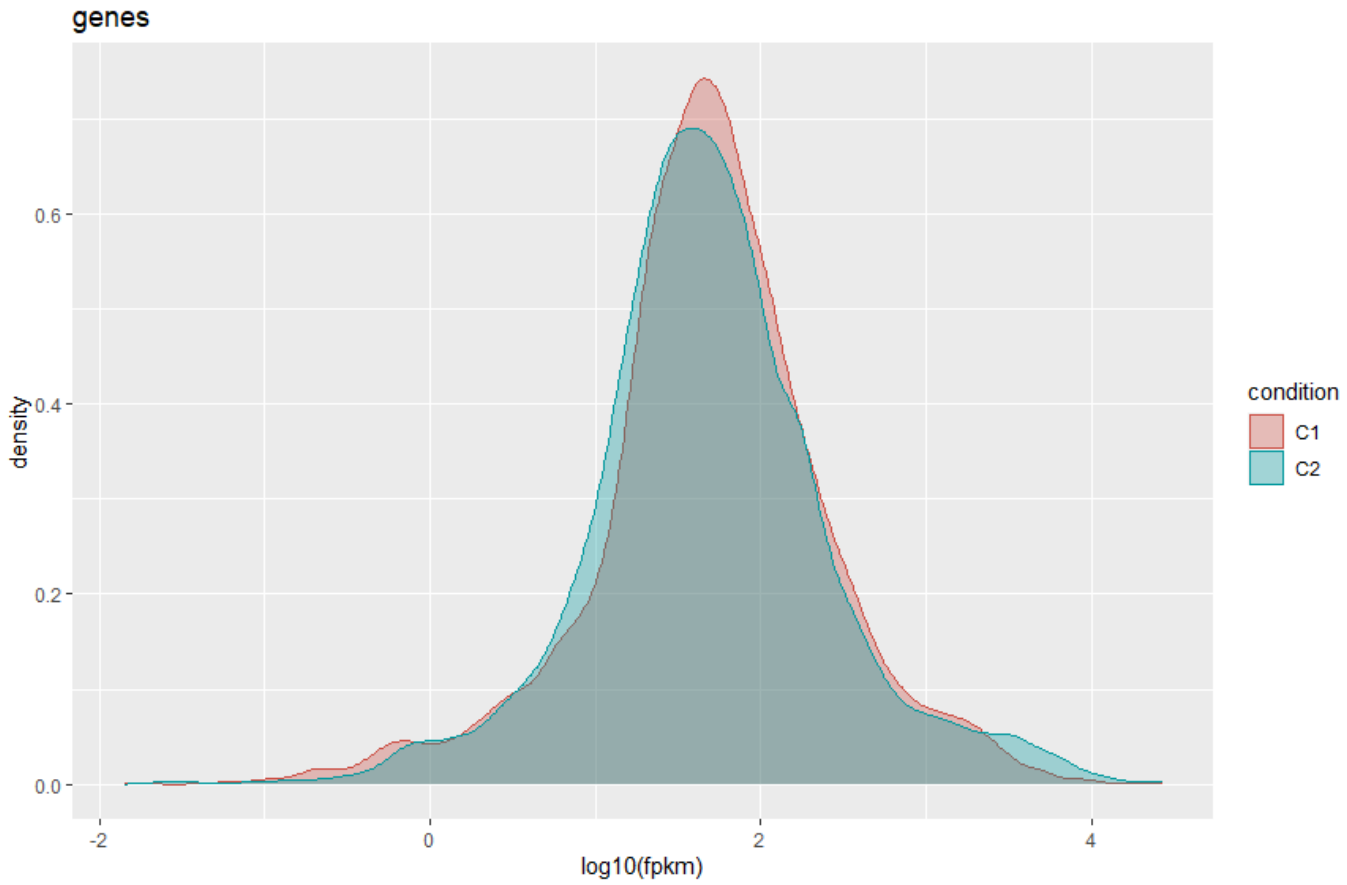


Fig.4 | CummeRbund plot of the expression level distribution for all genes in experimental conditions C1 (Aerobic condition) and C2 (Anaerobic condition). FPKM, fragments per kilobase of transcript per million fragments mapped.

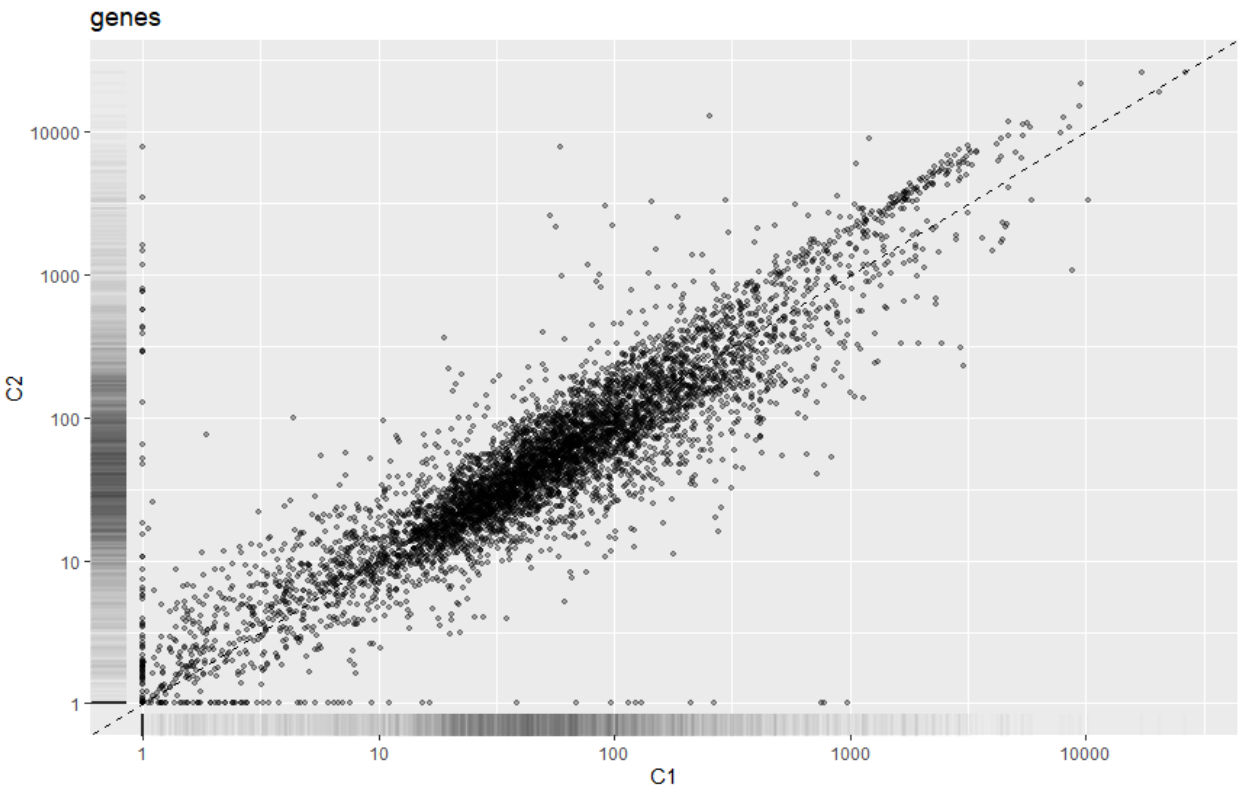


Fig.5 | CummeRbund scatter plots highlight general similarities and specific outliers of genes between aerobic and anaerobic conditions.

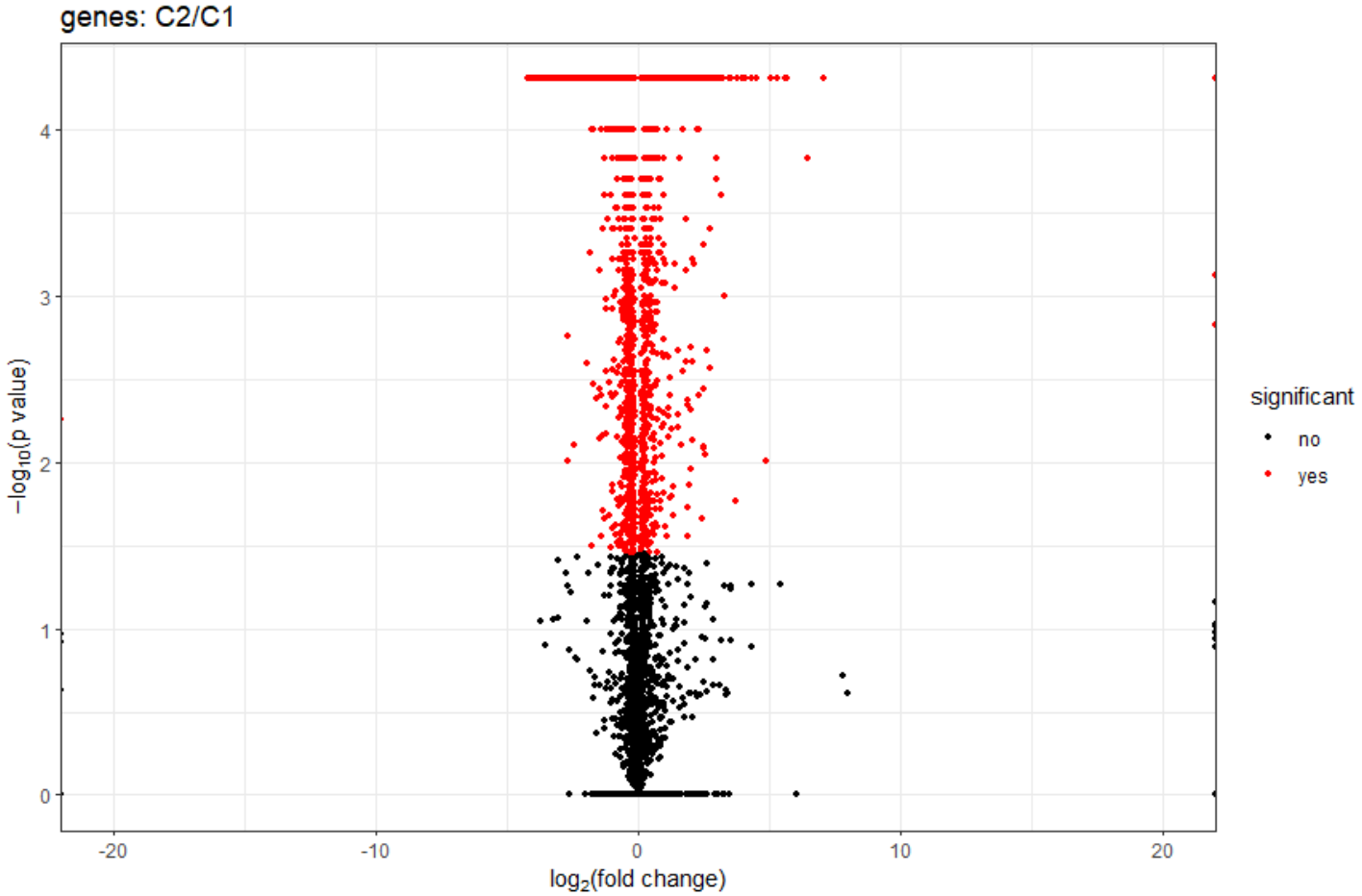


Fig.6 | CummeRbund volcano plot reveal genes that differ significantly between the pairs of conditions.

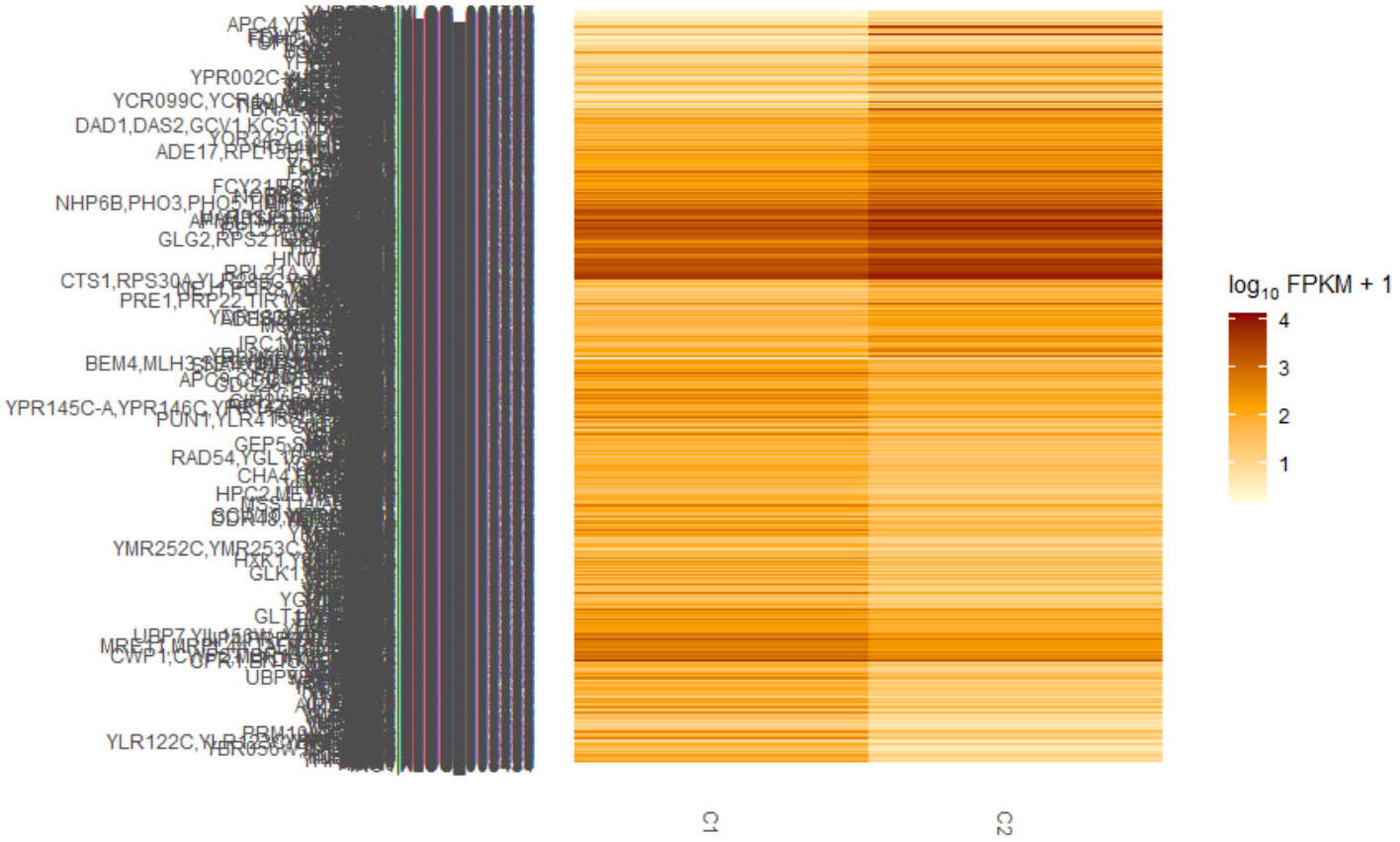


Fig.7 | Heat map of differentially expressed genes

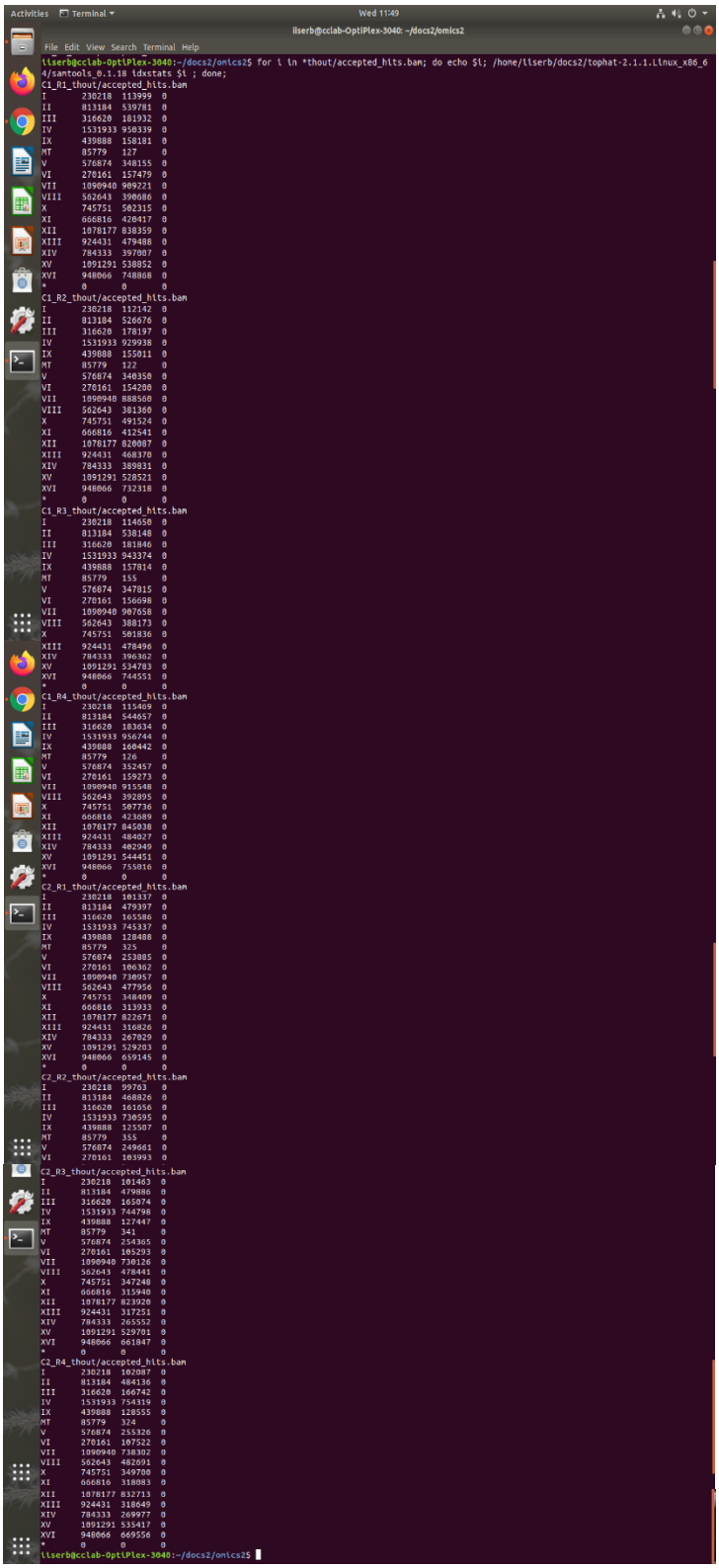


Fig.8 | Report of the number of fragments that map to each chromosome.

After the global gene differentiation visualization was performed, we used the output file: **gene_exp.diff** to sort the log2(fold_change) column, which is the effect size estimate. This value indicates how much the gene or transcript's expression seems to have changed between the two conditions. A high negatives value indicates a highly down-regulated gene, while a positive value indicates an up-regulated gene. Using this data two genes of interest were selected **THI72** and **YPS1,** and further analyzed.

The expression bar plot, isoform bar, line plot, and read coverage from IGV are as follows:

(1) THI72

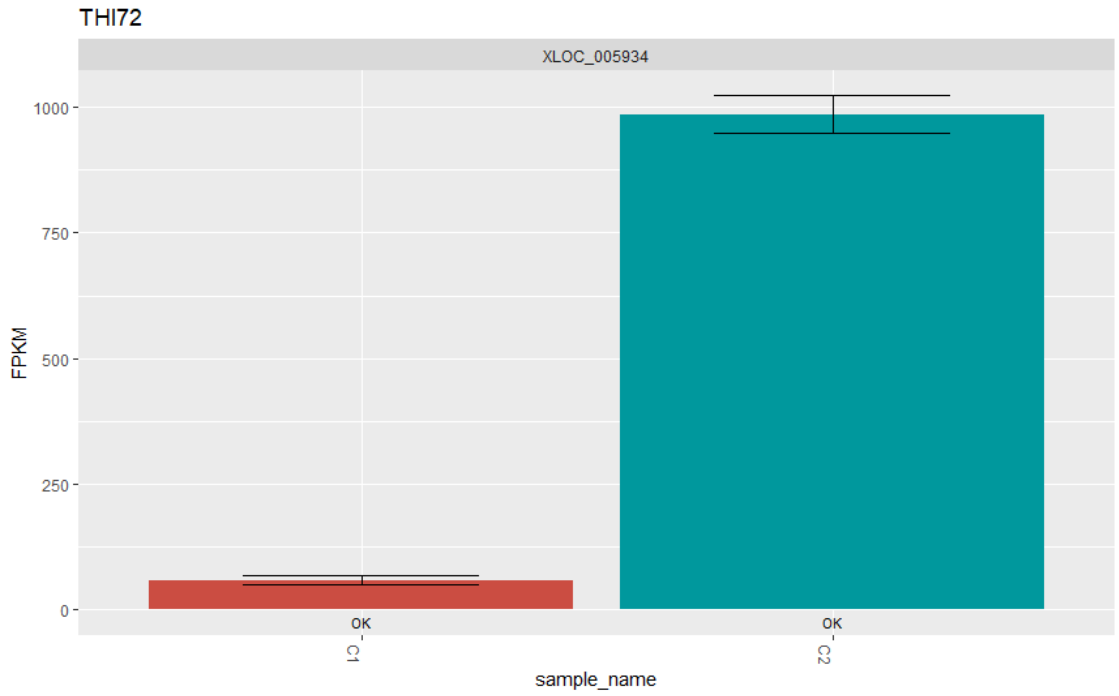


Fig.9 | Differential analysis results. Expression plot shows clear differences in the expression of THI72 across conditions C1 (aerobic) and C2 (anaerobic), measured in FPKM

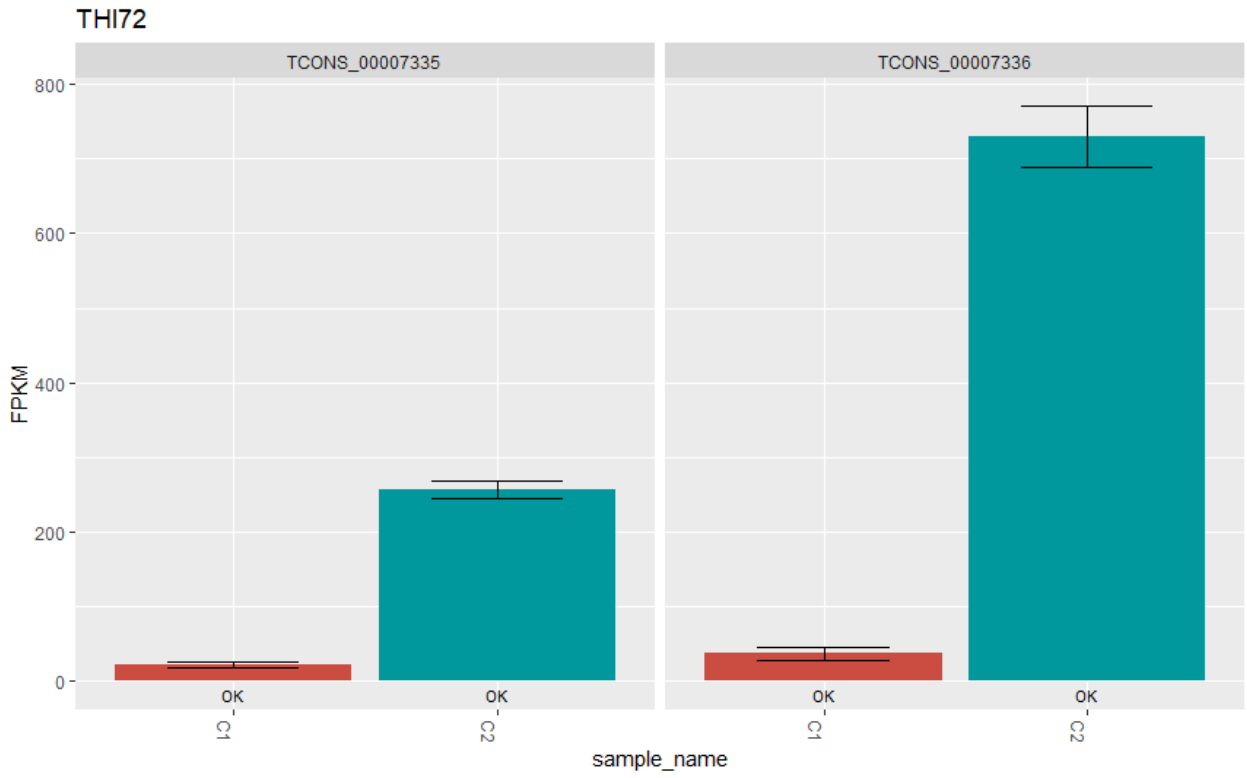


Fig.10 | Changes in THI72 expression are attributable to a large increase in the expression of one of two alternative isoforms.

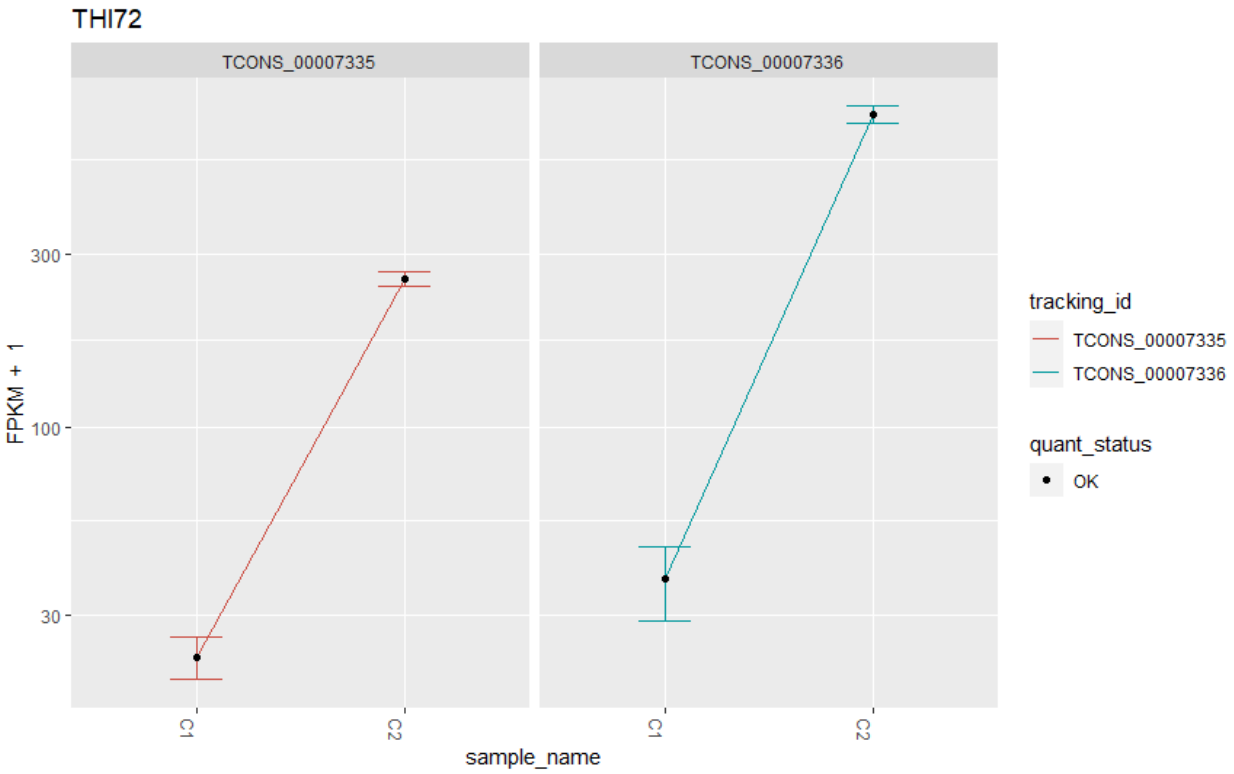


Fig.11 | Figure shows linear plot of expression of THI72 isoforms. The gene is up-regulated.

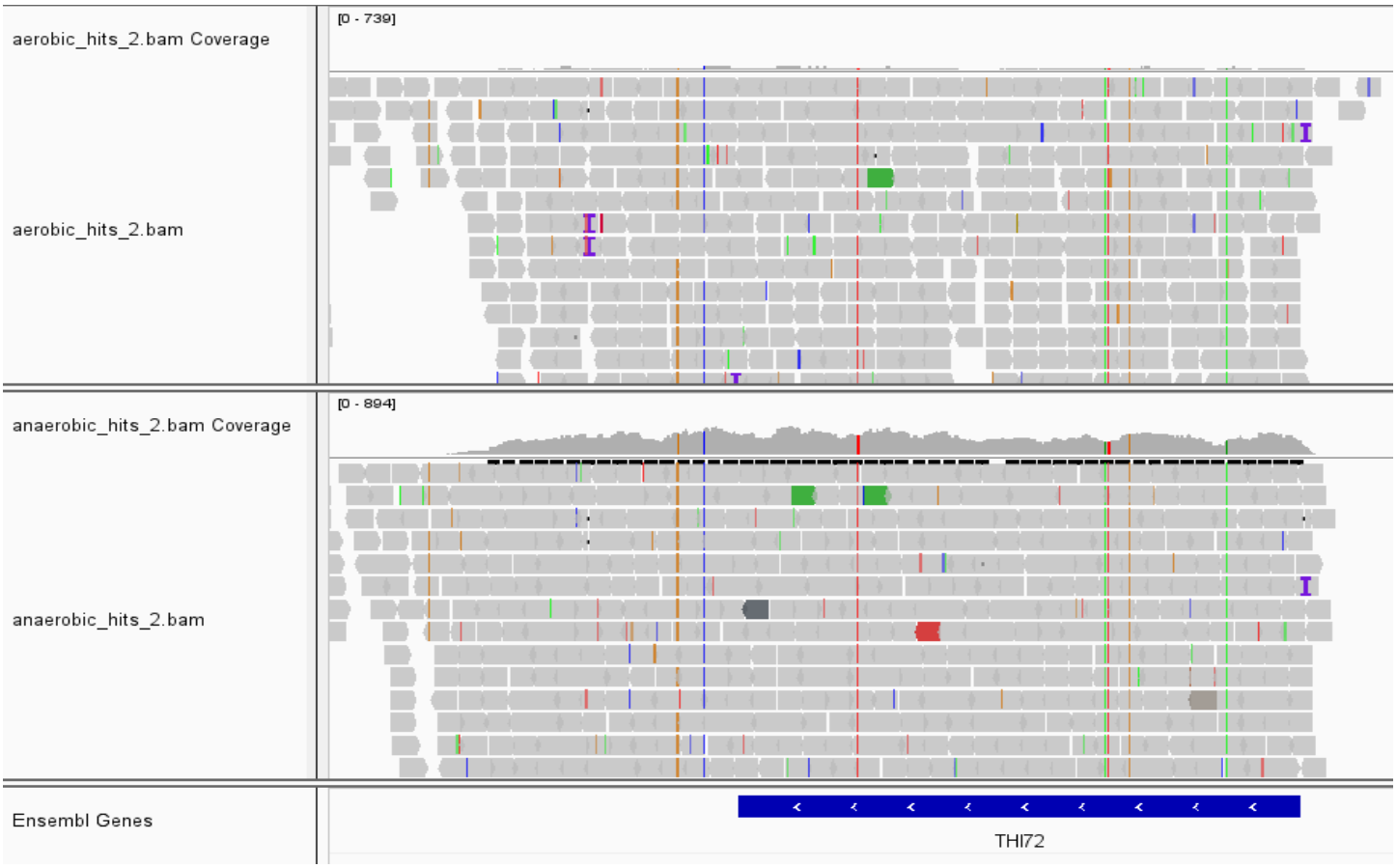


Fig.12 |The read coverage, viewed through the genome browsing application IGV42, shows an increase in sequencing reads originating from the gene in condition C2.

(2) YPS1

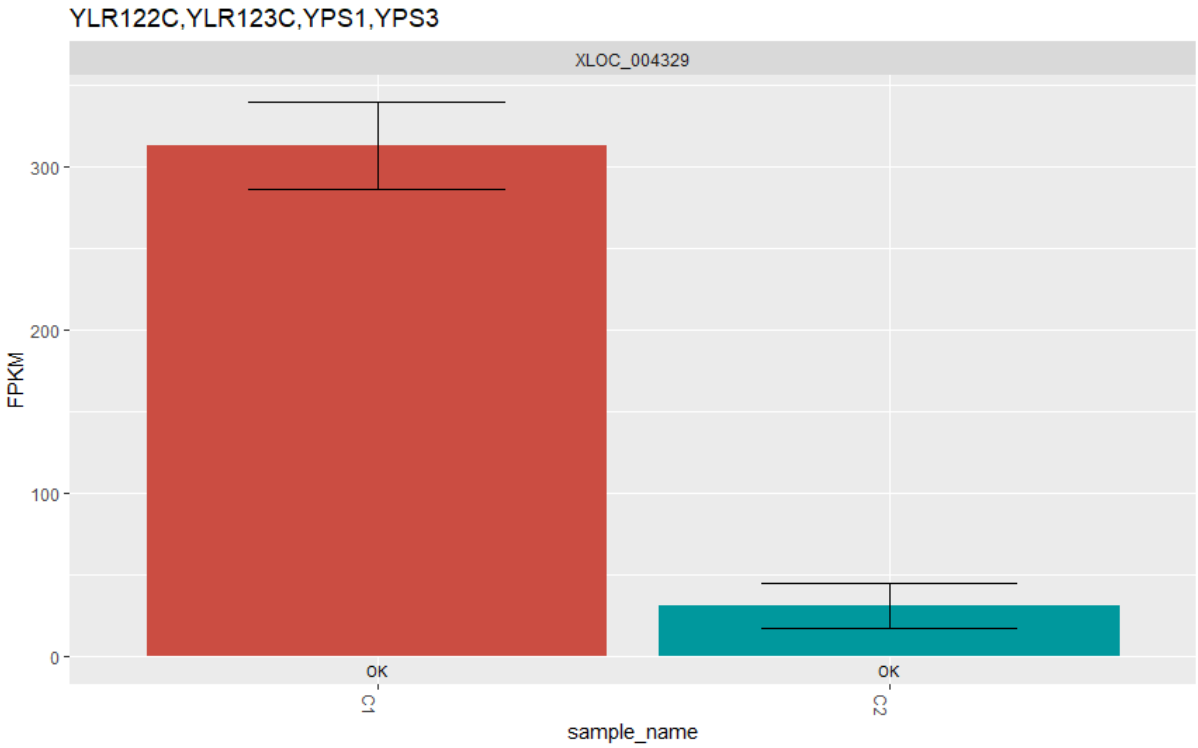


Fig.13 | Differential analysis results. Expression plot shows clear differences in the expression of YPS1 across conditions C1 (aerobic) and C2 (anaerobic), measured in FPKM

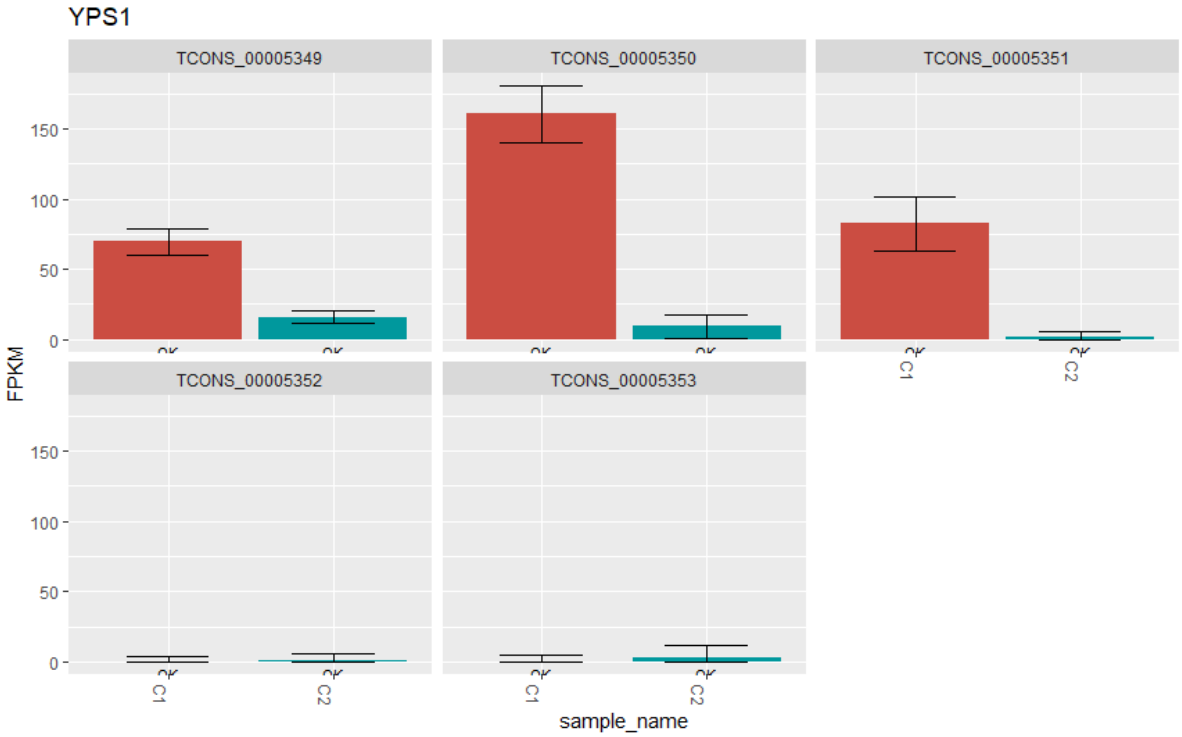


Fig.14 | Changes in YPS1 expression are attributable to a large increase in the expression of one of five alternative isoforms, of which two are insignificant.

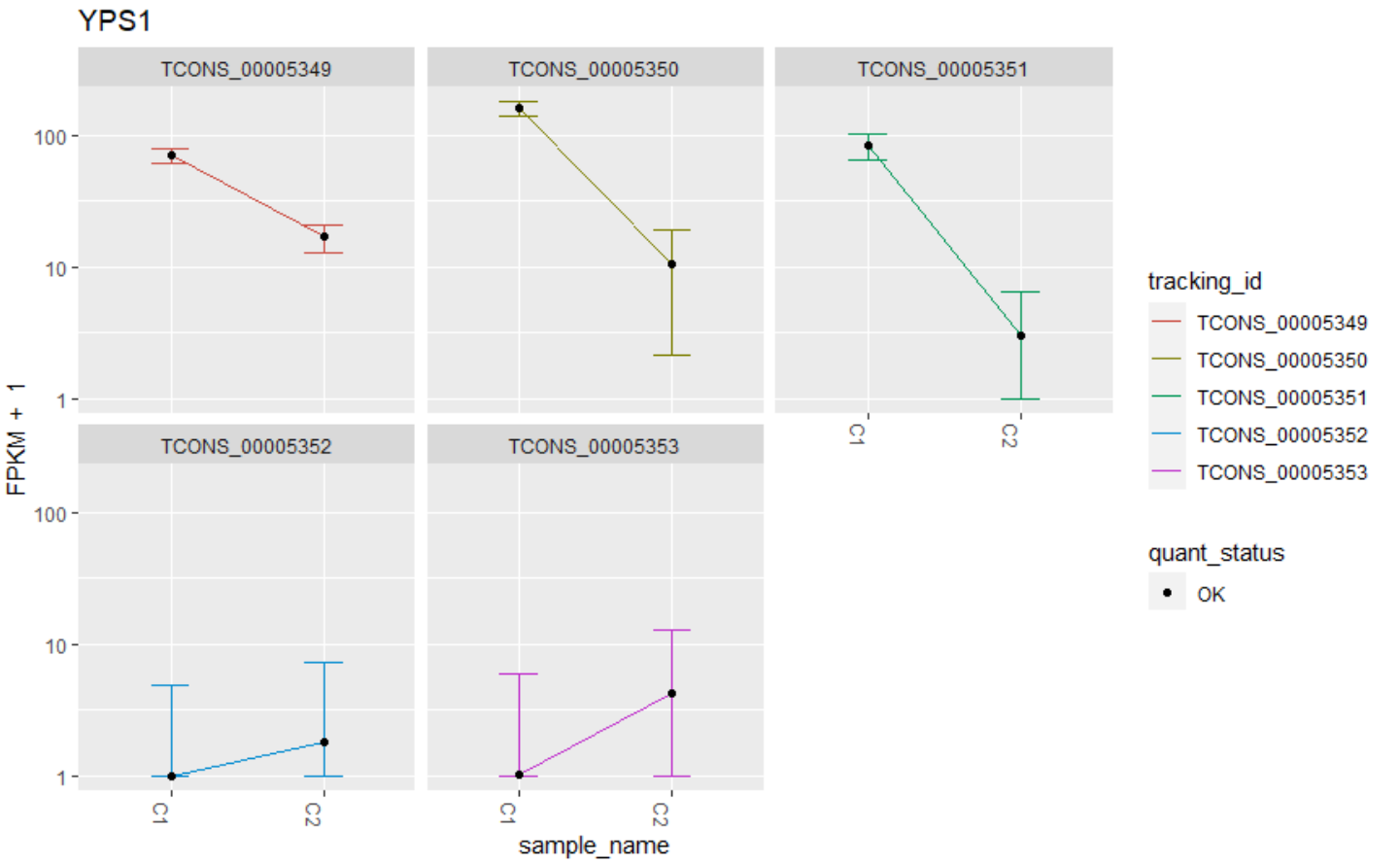


Fig.15 | Figure shows linear plot of expression of YPS1 isoform. The gene is down-regulated. Bottom two plots can be ignored as they are insignificant.

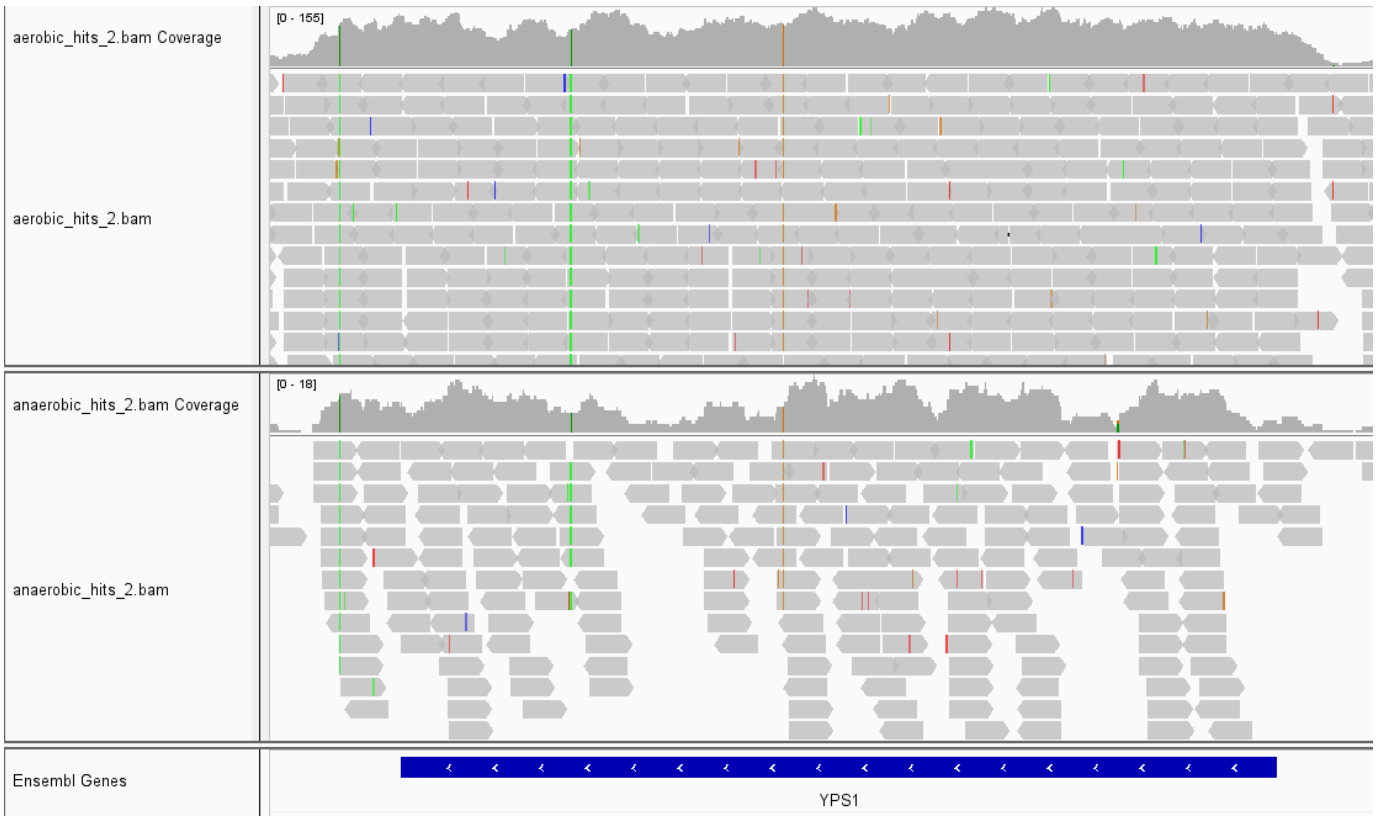


Fig.16 | The read coverage, viewed through the genome browsing application IGV42, shows a decrease in sequencing reads originating from the gene in condition C2.

Finally, a list was created of only the names of the significantly expressed genes and uploaded to **g:Profiler** to check the gene ontology expression (large scale effects on cellular pathways and functions). GO profiler performs functional enrichment analysis, also known as over-representation analysis (ORA) or gene set enrichment analysis, on the input gene list. It maps genes to known functional information sources and detects statistically significantly enriched terms.

We obtained data on:

(i) GO molecular function

(ii) GO cellular component

(iii) GO biological process

(iv) KEGG pathway

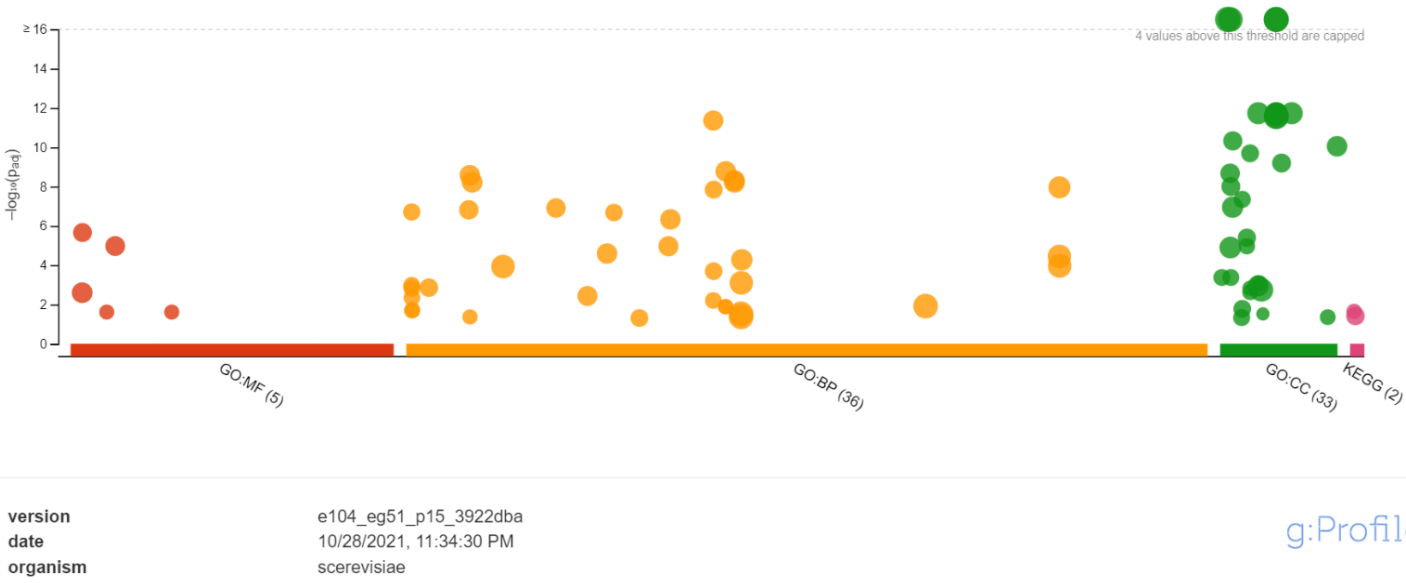
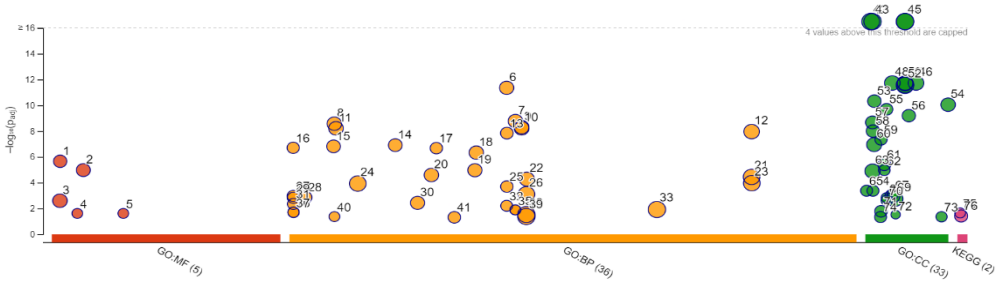


Fig.17 | Gene ontology profile plot



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0003735	structural constituent of ribosome	2.195×10^{-6}
2	GO:MF	GO:0005198	structural molecule activity	1.060×10^{-5}
3	GO:MF	GO:0003723	RNA binding	2.519×10^{-3}
4	GO:MF	GO:0004812	aminoacyl-tRNA ligase activity	2.434×10^{-2}
5	GO:MF	GO:0016875	ligase activity, forming carbon-oxygen bonds	2.434×10^{-2}
6	GO:BP	GO:0042254	ribosome biogenesis	4.390×10^{-12}
7	GO:BP	GO:0043043	peptide biosynthetic process	1.682×10^{-9}
8	GO:BP	GO:0006412	translation	2.650×10^{-9}
9	GO:BP	GO:0043604	amide biosynthetic process	4.919×10^{-9}
10	GO:BP	GO:0043603	cellular amide metabolic process	5.989×10^{-9}
11	GO:BP	GO:0006518	peptide metabolic process	6.100×10^{-9}
12	GO:BP	GO:1901566	organonitrogen compound biosynthetic process	1.094×10^{-8}
13	GO:BP	GO:0042274	ribosomal small subunit biogenesis	1.437×10^{-8}
14	GO:BP	GO:0016072	rRNA metabolic process	1.234×10^{-7}
15	GO:BP	GO:0006364	rRNA processing	1.531×10^{-7}
16	GO:BP	GO:0000462	maturation of SSU-rRNA from tricistronic rRNA tra...	1.966×10^{-7}
17	GO:BP	GO:0030490	maturation of SSU-rRNA	2.102×10^{-7}
18	GO:BP	GO:0034660	ncRNA metabolic process	4.674×10^{-7}
19	GO:BP	GO:0034470	ncRNA processing	1.086×10^{-5}
20	GO:BP	GO:0022613	ribonucleoprotein complex biogenesis	2.557×10^{-5}
21	GO:BP	GO:1901564	organonitrogen compound metabolic process	3.589×10^{-5}
22	GO:BP	GO:0044281	small molecule metabolic process	5.309×10^{-5}
23	GO:BP	GO:1901576	organic substance biosynthetic process	1.096×10^{-4}
24	GO:BP	GO:0009058	biosynthetic process	1.176×10^{-4}
25	GO:BP	GO:0042273	ribosomal large subunit biogenesis	2.014×10^{-4}
26	GO:BP	GO:0044249	cellular biosynthetic process	7.955×10^{-4}
27	GO:BP	GO:0000460	maturation of 5.8S rRNA	1.070×10^{-3}
28	GO:BP	GO:0002181	cytoplasmic translation	1.394×10^{-3}
29	GO:BP	GO:0000466	maturation of 5.8S rRNA from tricistronic rRNA tra...	1.492×10^{-3}
30	GO:BP	GO:0019637	organophosphate metabolic process	3.683×10^{-3}
31	GO:BP	GO:0000469	cleavage involved in rRNA processing	4.497×10^{-3}
32	GO:BP	GO:0042255	ribosome assembly	6.244×10^{-3}
33	GO:BP	GO:0071704	organic substance metabolic process	1.209×10^{-2}
34	GO:BP	GO:0043038	amino acid activation	1.297×10^{-2}
35	GO:BP	GO:0043039	tRNA aminoacylation	1.297×10^{-2}
36	GO:BP	GO:0000479	endonucleolytic cleavage of tricistronic rRNA trans...	1.980×10^{-2}
37	GO:BP	GO:0000478	endonucleolytic cleavage involved in rRNA proces...	1.980×10^{-2}
38	GO:BP	GO:0044238	primary metabolic process	2.974×10^{-2}
39	GO:BP	GO:0044237	cellular metabolic process	4.242×10^{-2}
40	GO:BP	GO:0006418	tRNA aminoacylation for protein translation	4.302×10^{-2}
41	GO:BP	GO:0032543	mitochondrial translation	4.907×10^{-2}
42	GO:CC	GO:0005622	intracellular anatomical structure	6.553×10^{-38}
43	GO:CC	GO:0005737	cytoplasm	5.312×10^{-29}
44	GO:CC	GO:0043229	intracellular organelle	2.176×10^{-22}
45	GO:CC	GO:0043226	organelle	4.793×10^{-22}
46	GO:CC	GO:0070013	intracellular organelle lumen	1.853×10^{-12}
47	GO:CC	GO:0043233	organelle lumen	1.853×10^{-12}
48	GO:CC	GO:0031974	membrane-enclosed lumen	1.853×10^{-12}
49	GO:CC	GO:0043232	intracellular non-membrane-bounded organelle	1.985×10^{-12}
50	GO:CC	GO:0043228	non-membrane-bounded organelle	1.985×10^{-12}
51	GO:CC	GO:0043231	intracellular membrane-bounded organelle	2.548×10^{-12}
52	GO:CC	GO:0043227	membrane-bounded organelle	2.836×10^{-12}
53	GO:CC	GO:0005840	ribosome	4.748×10^{-11}
54	GO:CC	GO:1990904	ribonucleoprotein complex	8.905×10^{-11}
55	GO:CC	GO:0030684	preribosome	2.038×10^{-10}
56	GO:CC	GO:0044391	ribosomal subunit	6.342×10^{-10}
57	GO:CC	GO:0005730	nucleolus	2.122×10^{-9}
58	GO:CC	GO:0005759	mitochondrial matrix	9.933×10^{-9}
59	GO:CC	GO:0015935	small ribosomal subunit	4.499×10^{-8}
60	GO:CC	GO:0005829	cytosol	1.125×10^{-7}
61	GO:CC	GO:0022626	cytosolic ribosome	3.995×10^{-6}
62	GO:CC	GO:0022627	cytosolic small ribosomal subunit	1.088×10^{-5}
63	GO:CC	GO:0005739	mitochondrion	1.268×10^{-5}
64	GO:CC	GO:0005761	mitochondrial ribosome	4.233×10^{-4}
65	GO:CC	GO:0000313	organellar ribosome	4.233×10^{-4}
66	GO:CC	GO:0032040	small-subunit processome	8.139×10^{-4}
67	GO:CC	GO:0031981	nuclear lumen	1.120×10^{-3}
68	GO:CC	GO:0030686	90S preribosome	1.493×10^{-3}
69	GO:CC	GO:0032991	protein-containing complex	1.809×10^{-3}
70	GO:CC	GO:0030687	preribosome, large subunit precursor	2.372×10^{-3}
71	GO:CC	GO:0015934	large ribosomal subunit	1.668×10^{-2}
72	GO:CC	GO:0033176	proton-transporting V-type ATPase complex	3.015×10^{-2}
73	GO:CC	GO:1905369	endopeptidase complex	4.417×10^{-2}
74	GO:CC	GO:0010494	cytoplasmic stress granule	4.651×10^{-2}
75	KEGG	KEGG:00970	Aminoacyl-tRNA biosynthesis	2.234×10^{-2}
76	KEGG	KEGG:03010	Ribosome	3.846×10^{-2}

version	e104_eg51_p15_3922dba
date	10/28/2021, 10:43:06 PM
organism	scerevisiae

g:Profiler

Fig.18 | Differentially expressed genes involved in enriched processes in S. cerevisiae in response to anaerobic conditions compared to aerobic conditions

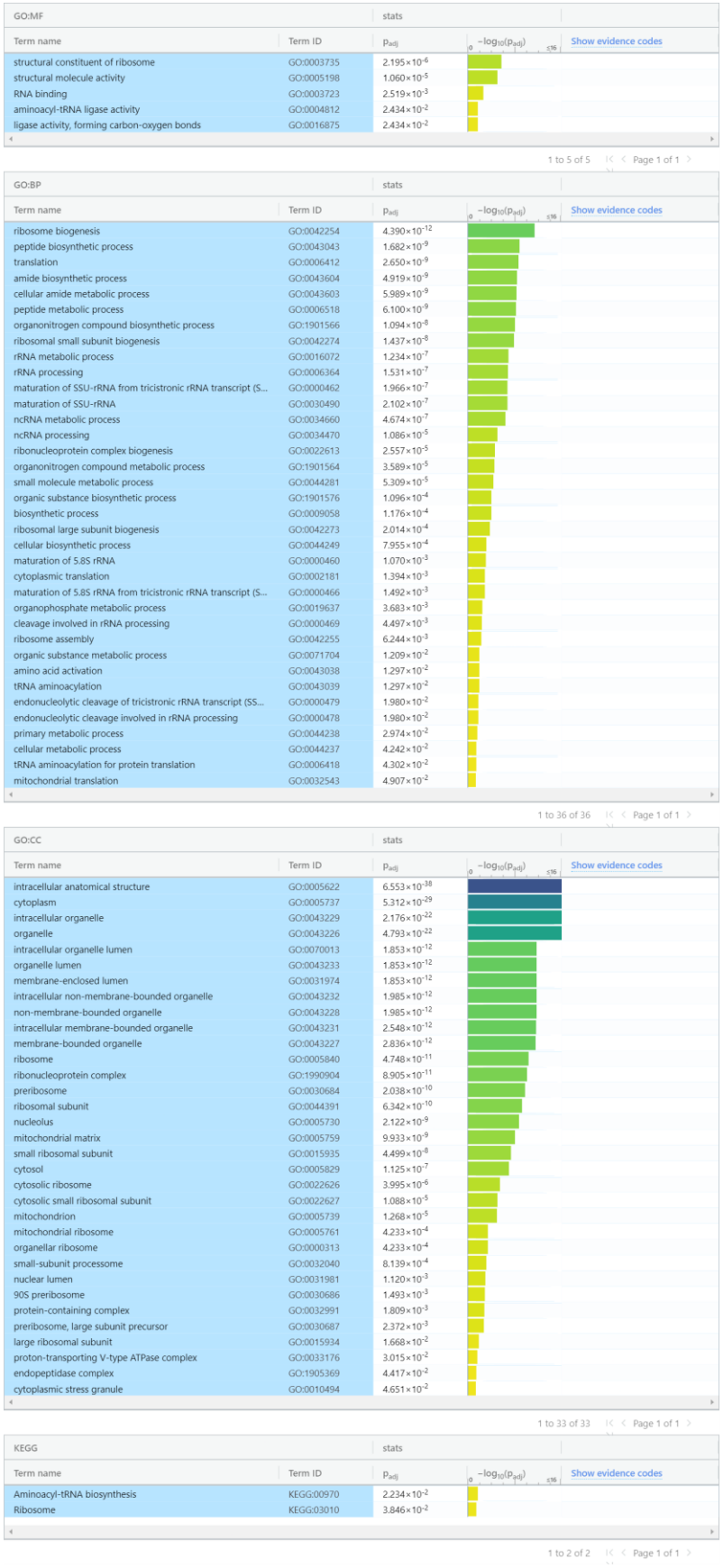


Fig.19 | Enriched GO terms (Biological Processes (BP), Molecular Function (MF), and Cellular component (CC)) in Saccharomyces cerevisiae under both the conditions.

IV. DISCUSSION

Low-affinity thiamine transporter, **THI72**, is responsible for the intake of thiamine. Its primary function is the uptake of closely related compounds, and that thiamine transport is a secondary activity of these proteins. The analysis reported a 17 fold (log2_fold_change = 4.06446) increase in the expression of THI72.

Aspartic protease, a hyperglycosylated member of the yapsin family of proteases, **YPS1**, stays attached to the plasma membrane via a glycosylphosphatidylinositol (GPI) anchor. It is involved in nutrient limitation-induced cleavage of the extracellular inhibitory domain of signaling mucin Msb2p, resulting in activation of the filamentous growth MAPK pathway. It is also involved with other yapsins in the cell wall integrity response; role in KEX2-independent processing of the alpha-factor precursor. Studies have shown that yeast cells grow equally well in nutrient-rich environment under aerobic and anaerobic conditions, Taherzadeh et al. Filamentous growth restricts cell multiplication of budding yeast cells under nutrient-poor conditions. During fermentation, yeast cells have a surplus of nutrients that can be utilized by fermenting to ethanol by partial oxidation, and therefore cells prefer the budding form of reproduction over filamentous growth. The analysis reported a 10 fold (log2_fold_change = -3.31345) decrease in the expression of YPS1.

Of the two genes selected, THI72 is up-regulated when conditions change from aerobic to anaerobic, while YPS1 is down-regulated. From the gene ontology profiling data, we can predict how the processes that use these genes would be affected in the duration of the experiment. All processes utilizing the THI72 gene would be expressed highly, and the changes in the cell would occur accordingly, while those processes in which YPS1 is involved would be reduced. The extent to which the process would be affected depends on how significantly the gene is involved in that particular process.

Output from Cuffdiff tells us whether a gene's observed fold change is significant. In the case of YPS1, it has significantly differentially expressed isoforms. However, as a gene's expression level is the sum of the expression levels of its isoforms, and some YPS1 isoforms are increased while others are decreased, the fold change in overall gene expression is modest. While for THI72, which has only two isoforms, that to both having increased expression levels, the fold change in overall gene expression is relatively high.

The experiment reveals markedly differentially expressed genes and transcripts between the two conditions. The protocol does not result in more spurious genes and transcripts than expected (the default false discovery rate for Cuffdiff is 5%). However, poorly replicated conditions, inadequate depth or quality of sequencing, and errors in the underlying annotation used to quantify genes and transcripts can all lead to artifacts during differential analysis.

S. cerevisiae is better able to deal with the fermentation environment possibly due to its efficient competitive uptake of sterols, copper and iron, accompanied by cell wall remodeling to accommodate additional mannoproteins and PAU proteins (information inferred from gene ontology expression). These strategies allow the yeast to regulate membrane fluidity and cell wall porosity, and withstand an anaerobic, high ethanol environment.

V. FUTURE DIRECTIONS AND IMPLICATIONS

The protocol followed to obtain differential gene expression for baker's yeast under the given conditions (aerobic and anaerobic) can be extrapolated to another species and any required condition. For example, we can follow this protocol to study human RNA-seq in a diseased and healthy state and compare the expression of different genes. This data would be helpful in developing a cure or modifying pathways to alter the function.

The strong cell wall-related responses in S. cerevisiae suggest the importance of this organelle in the cellular response to other species. In particular, the data support that the regulation of adhesion properties may play a central role in modulating the physical and ecological interactions between species

This analysis revealed genes possibly responsible for adapting cells better to anaerobic conditions. Thus, important cellular pathways and key players in fermentation can be studied and industrially utilized for optimal wine production.

REFERENCES

[1] Trapnell, C., Roberts, A., Goff, L. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2012). https://doi.org/10.1038/nprot.2012.016

[2] Shekhawat, K., Patterton, H., Bauer, F.F. et al. RNA-seq based transcriptional analysis of Saccharomyces cerevisiae and Lachancea thermotolerans in mixed-culture fermentations under anaerobic conditions. BMC Genomics 20, 145 (2019). https://doi.org/10.1186/s12864-019-5511-x

[3] Taherzadeh MJ, Karimi K. Pretreatment of lignocellulosic wastes to improve ethanol and biogas production: a review. Int J Mol Sci. 2008 Sep;9(9):1621-51. doi: 10.3390/ijms9091621. Epub 2008 Sep 1. PMID: 19325822; PMCID: PMC2635757.