a). Aligning with HISAT2

build genome index:

```
hisat2-build GCF_000146045.2_R64_genomic.fna GCF_000146045.2_R64_genomic.gff
```

run hisat2 in single-end mode:

```
hisat2 -x hisat_indeces/index -U SRR941816.fastq | samtools sort > out.bam
```

```
#!/bin/bash
path_out=/home/rybina/BI2020prak/Project6

#build genome index
hisat2-build ${path_out}/GCF_000146045.2_R64_genomic.fna ${path_out}/ref_yeast

#run hisat2 in single-end mode:
for f in ${path_out}/*.fastq ; do FILENAME=${f##*/}; hisat2 -p 4 -x ${path_out}/ref_yeast -U ${f} | samtools sort > ${path_out}/${FILENAME%.*}.bam; done;
```

b) Quantifying with featureCounts

Convert from GFF to GTF:

```
../../progs/gffread/gffread GCF_000146045.2_R64_genomic.gff -T -o GCF_000146045.2_R64_genomic.gtf
```

Run the feature counts program:

```
../../progs/subread-2.0.2-source/bin/featureCounts -g gene_id -a GCF_000146045.2_R64_genomic.gtf -o  featureCounts_output/SRR941817_fc SRR941817.bam

../../progs/subread-2.0.2-source/bin/featureCounts -g gene_id -a GCF_000146045.2_R64_genomic.gtf -o  featureCounts_output/all_fc.txt SRR941816.bam SRR941817.bam SRR941818.bam SRR941819.bam
```

Simplify the counts:

```
cat SRR941816_fc | cut -f 1,7-10 > SRR941816_simple_counts.txt

cat all_fc.txt | cut -f 1,7-10 > all_simple_counts.txt

cat featureCounts_output/all_simple_counts.txt | R -f deseq2.r
cat norm-matrix-deseq2.txt | R -f draw-heatmap.r

head result.txt -n 50 | cut -f 1 > genes.txt
```

```
cat genes.txt | grep -o Y.* > genes_id.txt

head result.txt -n 50 | cut -f 1 > genes.txt

cat genes.txt | grep -o Y.* > genes_id.txt
```

c) KEGG over-representation test and visualization

```
library('pathview')
library('enrichplot')
library(clusterProfiler)



# select those with padj < 0.05
geneList <- as.numeric(sorted.df[sorted.df$padj<0.05,]$log2FoldChange)
names(geneList)                                                    <-
as.character(row.names(sorted.df[sorted.df$padj<0.05,]))
geneList <- sort(geneList, decreasing = TRUE)


kk <- enrichKEGG(gene         =names(geneList)[abs(geneList) > 2],
                 organism     = 'sce',
                 pvalueCutoff = 0.05)

barplot(kk, showCategory=20)


# sce03008 ribosome biogenesis - select genes which absolute value of
LogFold > 2
pathview(gene.data=geneList[abs(geneList)     >     2],     species='sce',
pathway.id="sce03008", gene.idtype = "KEGG",
         limit= list(gene=max(abs(geneList)), cpd=1))


# sce00020 Citrate cycle (TCA cycle) - select genes which absolute value
of LogFold > 2
pathview(gene.data=geneList[abs(geneList)     >     2],     species='sce',
pathway.id="sce00020", gene.idtype = "KEGG",
         limit= list(gene=max(abs(geneList)), cpd=1))

# sce00010 Glycolysis / Gluconeogenesis
pathview(gene.data=geneList[abs(geneList)     >     2],     species='sce',
pathway.id="sce00010", gene.idtype = "KEGG",
         limit= list(gene=max(abs(geneList)), cpd=1))
```

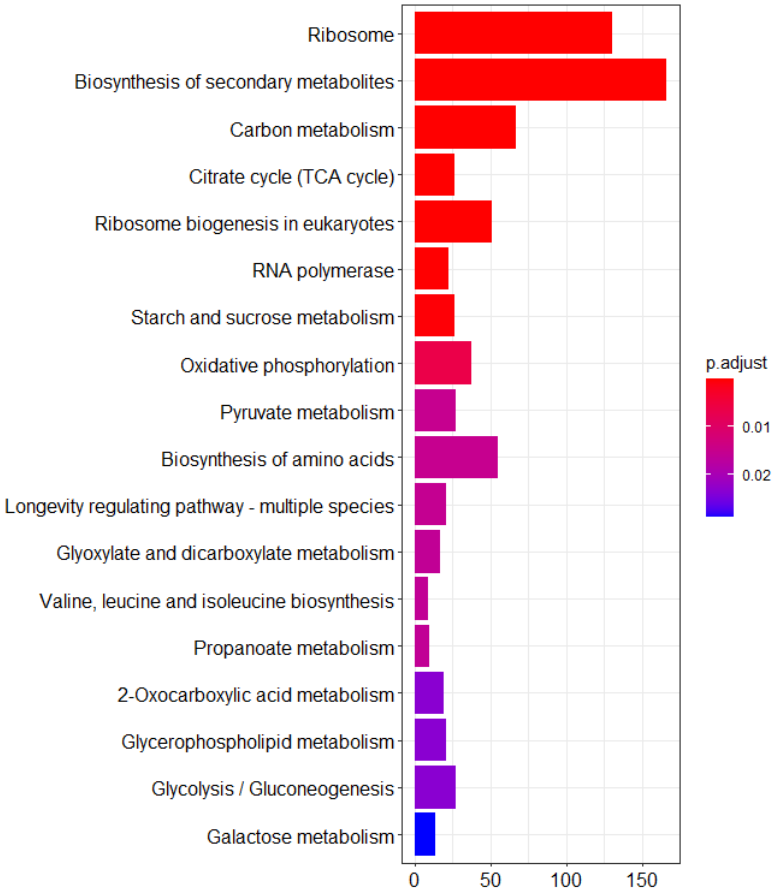

Figure S1. Barplot visualization of 20 top enriched terms found by enrichKEGG for significantly differentially expressed genes (padj<0.05) of absolute log2FoldChange value greater than 2.

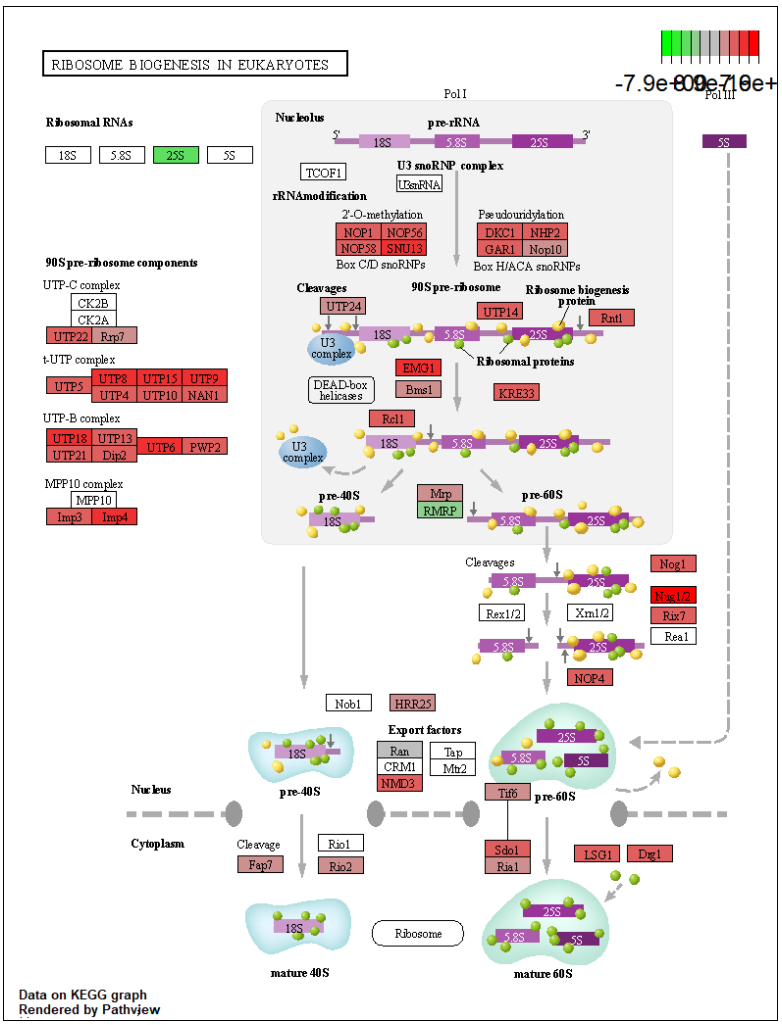

Figure S2. Ribosome biogenesis KEGG pathway (sce03008). Up-regulated genes marked red, down-regulated marked green. Differentially expressed genes (padj < 0.05) with absolute value of log2FoldChange > 2 are highlighted.

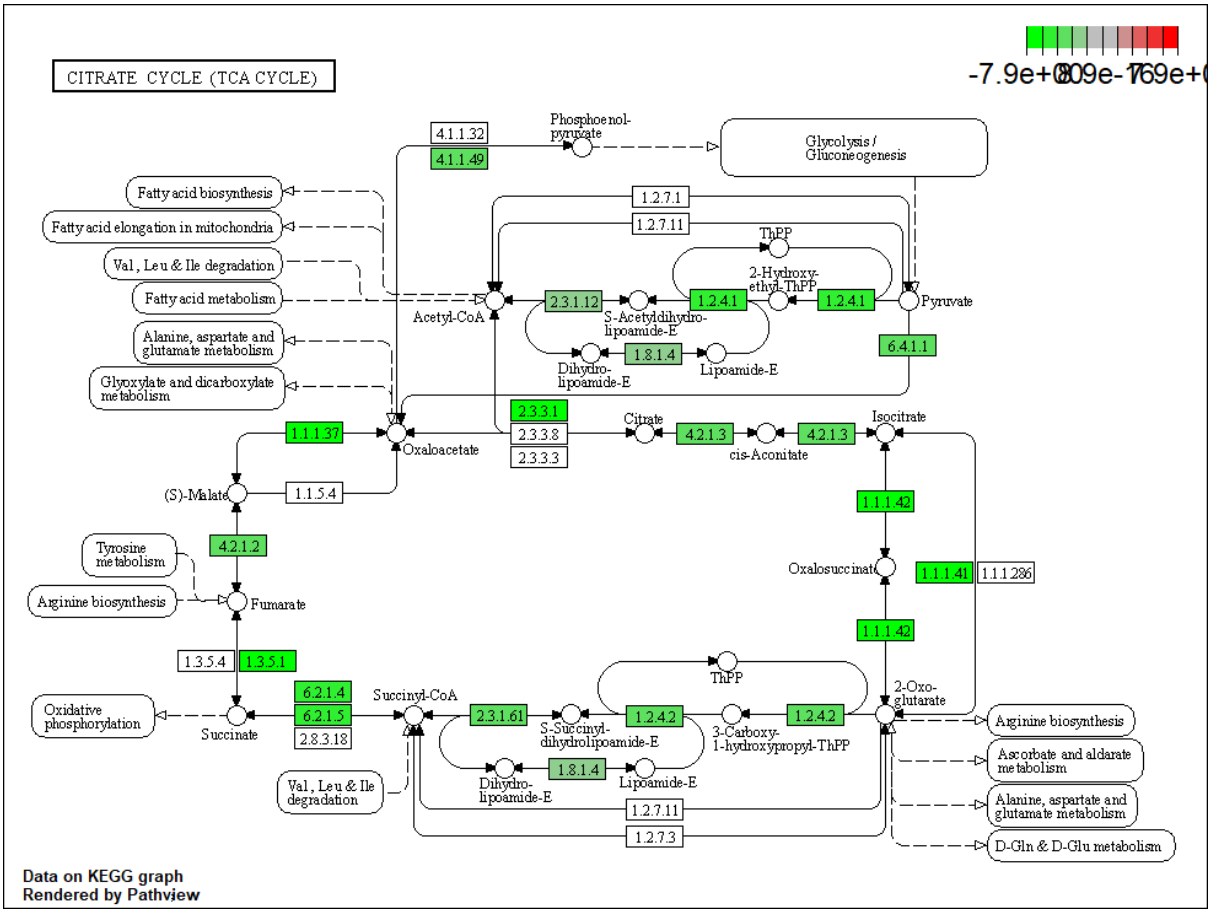

Figure S3. Citrate cycle (TCA cycle) KEGG pathway (sce00020). Up-regulated genes marked red, down-regulated marked green. Differentially expressed genes (padj < 0.05) with absolute value of log2FoldChange > 2 are highlighted.

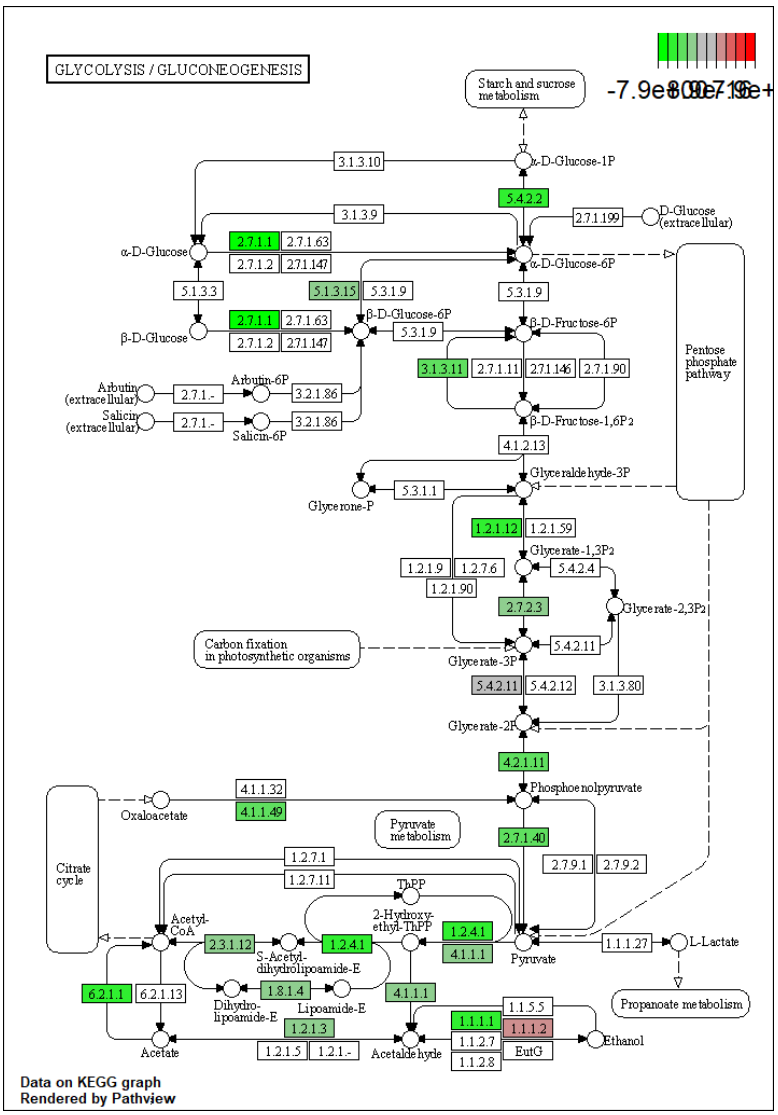

Figure S4. Glycolysis / Gluconeogenesis KEGG pathway (sce00010). Up-regulated genes marked red, down-regulated marked green. Differentially expressed genes (padj < 0.05) with absolute value of log2FoldChange > 2 are highlighted.