Bioinformatics Lab Report

Daniel Gehlbach A10661606
Orysya Stus A10743411
Gavin Tse A10567068

Beng 160: Chemical & Molecular Bioengineering Techniques

Section 1, Group 3

June 10, 2015

Background

In **lab 6a** we extracted RNA from different yeast strains and then isolated the RNA via column purification. We then moved on to **lab 6b** where we quantified the concentration of RNA using the Nanodrop machine and proceeded to perform cDNA synthesis in preparation for **lab 6c** PCR amplification. **Lab 6c** included protocols to amplify the RNA sequencing libraries through illumina sequencing and a combination of primers, H2O, OneTaq PCR master mix, and cDNA synthesis reaction. To conclude **lab 6c** PCR is conducted. **Lab 6d** has 3 main purposes: to perform PCR Column Purification, Gel Electrophoresis, and Size Selection. **Lab 9a** transitioned to data analysis and bioinformatics, utilizing bowtie, samtools, and various others commands in order to perform alignment. Then in **lab 9b** the protocol gave instructions for how to create an expression matrix, where the columns are the samples and the rows are the genes, in order to determine differential gene expression between different knock-out strains. In **lab 10a** we then proceeded to use DESeq in order to determine those samples that are significantly up/down regulated. In **lab 10b** we utilize the Database for Annotation, Visualization, and Integrated Discovery (DAVID). David takes a gene list input and interprets which biological pathways are upregulated/downregulated.

DESeq contains various normalization methods at several steps of the process. DESeq uses the median of scaled counts. Data is normalized in one way by the normalized counts which is the counts/ size factors. Another way is that there is a variance stabilizing transformation provided by DESeq that is very useful. It has been recommended to be used for clustering, heatmaps, or other visualization.

In the process of DESeq you first take the geometric mean of each condition for a gene and use this as a reference value. Next you can divide the gene expression value for each condition by the reference value you just obtained in order to create a list of values for each condition. Then you obtain the size factor, which is the median value of this list. Due to the fact that there exists a list per condition it will then generate a size factor value for each condition. The counts are then normalized by dividing each column by its size factor. This procedure is one of the ways DESeq produces normalized data.

Our purpose in analyzing the reads of the genes are to determine if the results are statistically significant or not. If they are, this would mean that the difference between read count is greater than would be expected just by natural random variation. Since we know that not all gene libraries are of the same size, we take this into account by taking the median ratio correction factor for each strain as a means of normalization. We apply this normalization value to all the read counts through the functions *estimatesizefactors* and *sizefactors* in DESeq. After calculating this factor for each strain we use the read counts divided by the factor for each strain, which gives us a value that we desire to use.

The null hypothesis for DESeq is that there is no difference between the strains that we annotated, or in other words no gene has been knocked out. Significantly low p-values ($p<0.05$) signifies that we reject the null hypothesis, thus resulting in the conclusion that there is a difference between the strains due to more than just the random inherent differences. For DAVID, the null hypothesis is that the enrichment of each annotation is due to chance. By selecting ($p<0.05$) we are able to test if our annotations are significant.

Results

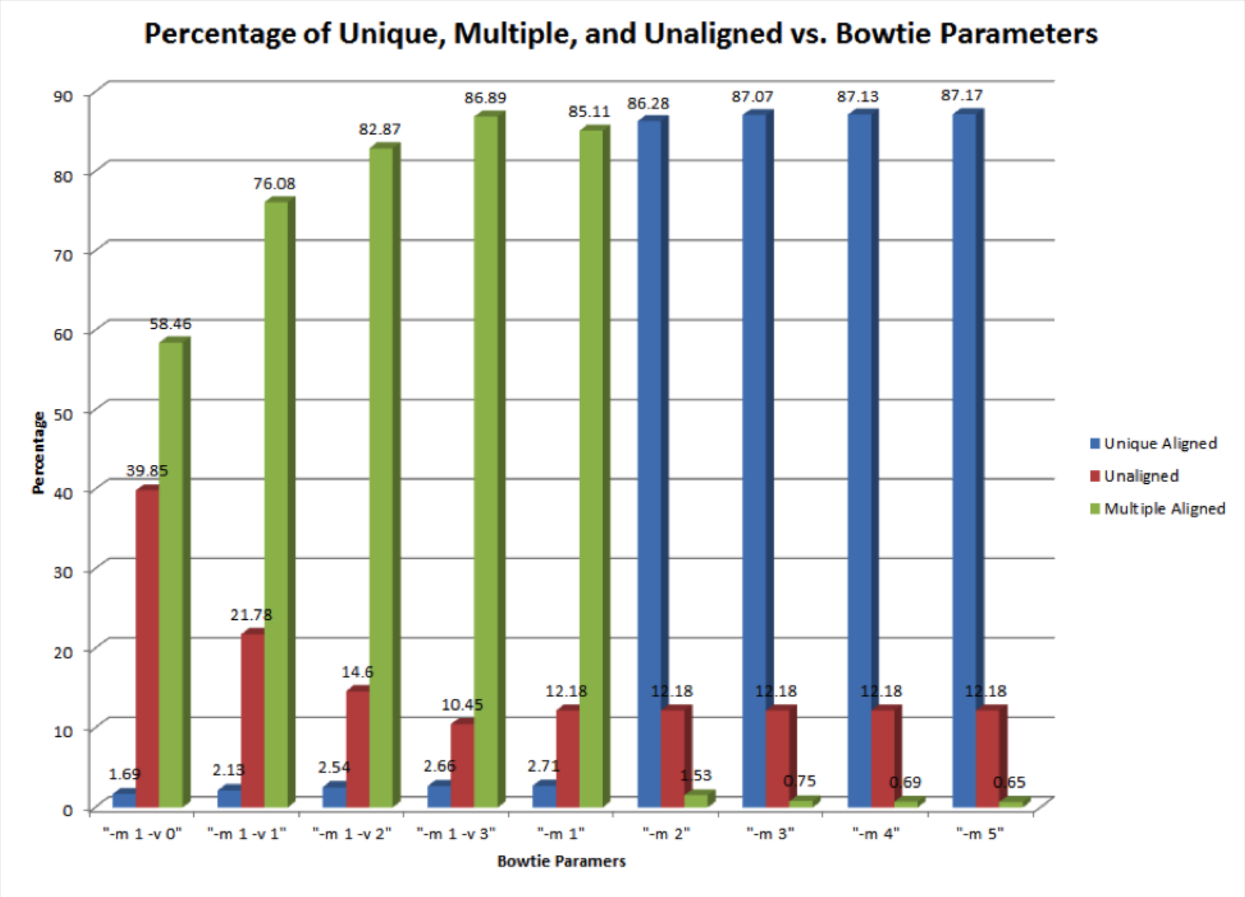


Figure 1 The effect of changing parameters on Bowtie alignment statistics. This graph shows the percentages of unique aligned, unaligned, and multiple aligned with varying Bowtie parameters. The first four columns show the effect of increasing the number of mismatches of one mapping location. The next five columns show the effect of increasing the number of maximum mapping locations. As shown in the bar graphs above, increasing values of "-v" results in increasing values of both unique aligned and multiple aligned, and decreasing values of unaligned. Increasing values of "-m" results in increasing values of unique aligned, decreasing values of multiple aligned, and unchanged unaligned values.

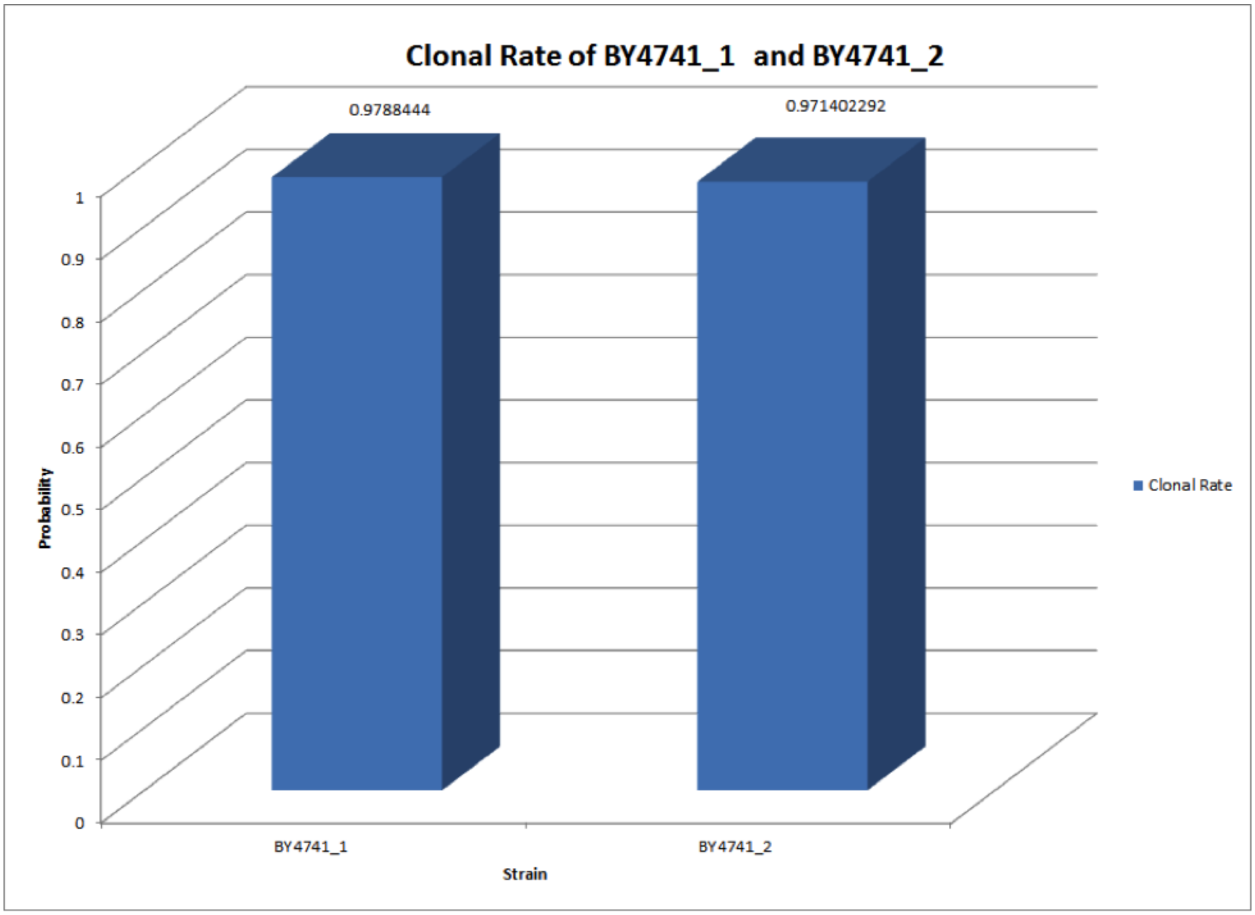


Figure 2 Clonal rate of BY4741_1 and BY4741_2. These two bar graphs compare the clonal rates between BY4741_1 and BY4741_2 which are 97.88% and 97.14% respectively. Since the given equation for clonal rate is:

clonal rate = 1 - (remaining hits after clonal removal/successful alignments reported by Bowtie) we can conclude that the lower the clonal rate, the higher number of remaining hits after clonal removal, which would be better for sequencing data because there would be more data to work with for analysis. Thus, the clonal rate of BY4741_2 is slightly better than that of BY4741_1.

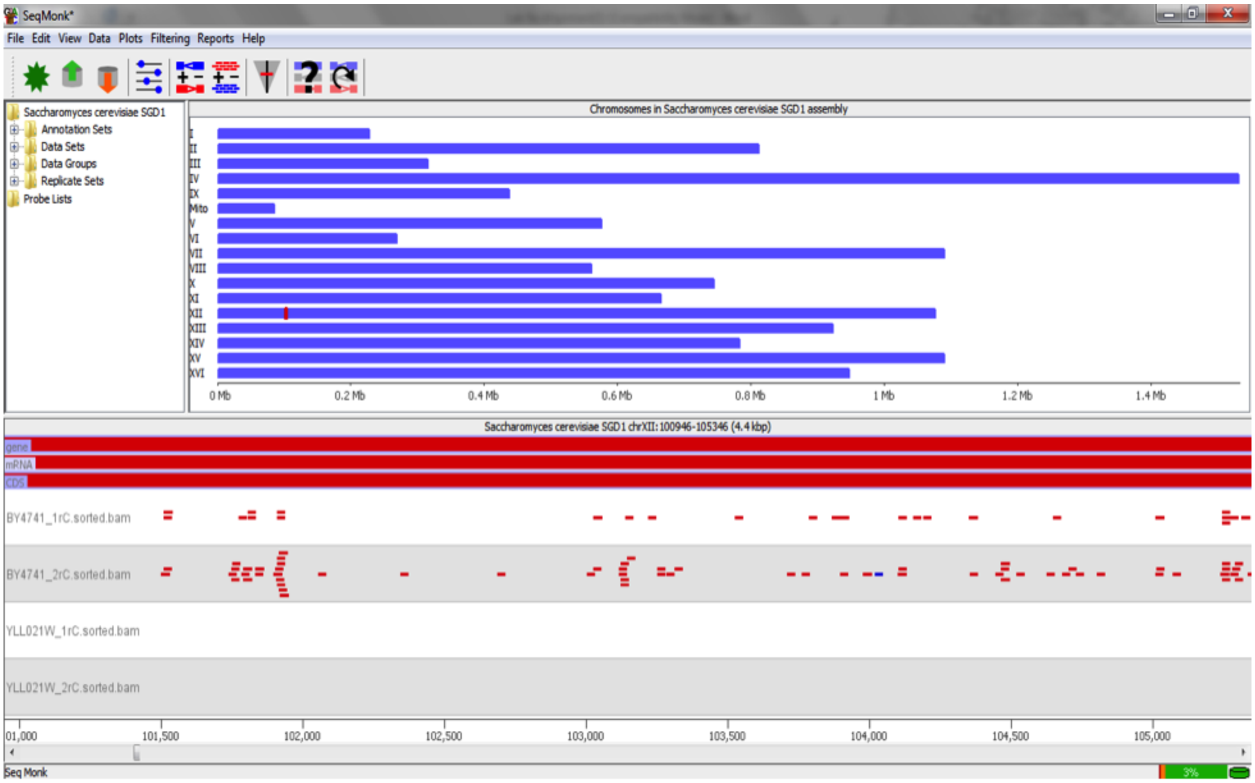


Figure 3 SeqMonk visualization of gene expression in BY4741 and YLL021W. This figure depicts the gene expression of the SPA2 gene in the wild type strain BY4741 and the mutated knockout strain YLL021W. As shown above, there are no reads in the two knockout strains compared to the noticeable number of reads in the two wild type strains. This is consistent with what is expected; the knockout strain has no reads because the SPA2 gene was knocked out to remove expression, but the wild type strain has normal expression of the gene.

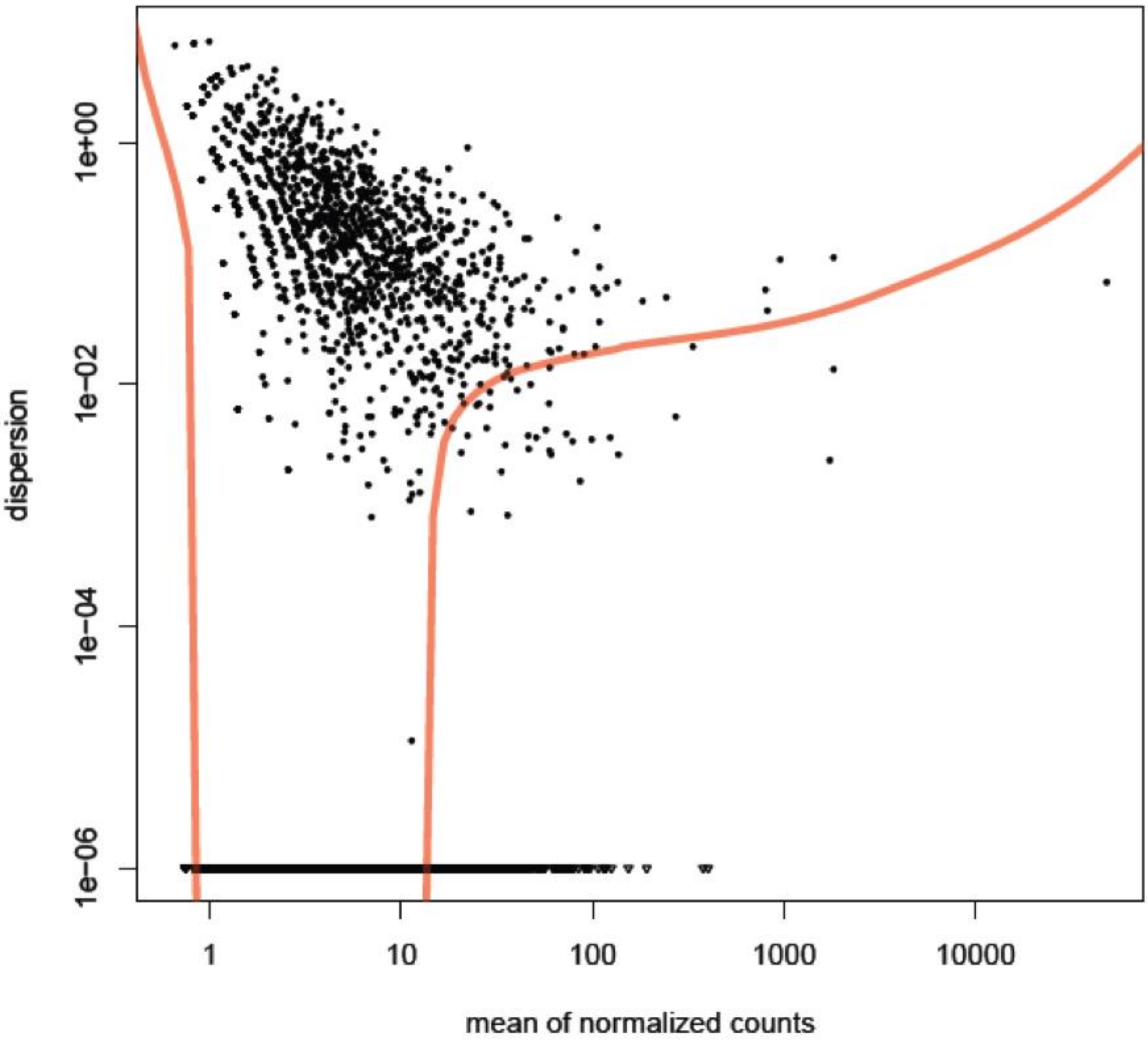


Figure 4a Dispersion vs mean of normalized counts. This graph displays estimated dispersion for each mean of normalized counts. When analyzing differential gene expression, estimated dispersion values are essential in determining whether results are true, or whether they are false positives or negatives. Overestimation of dispersion can lead to over filtering of true differentially expressed genes. Underestimation can lead to false positives.

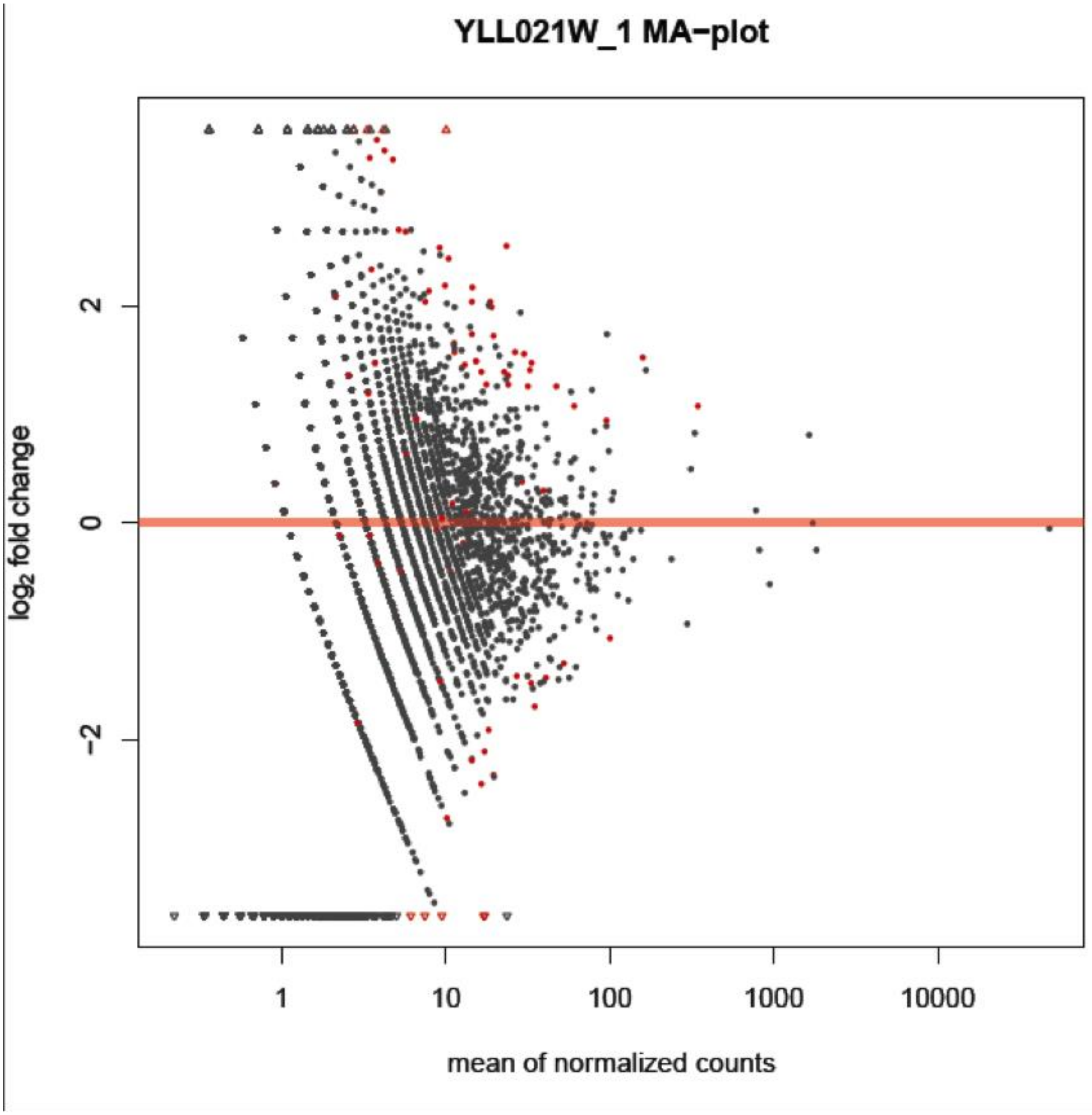


Figure 4b MA plot for YLL021W_1. This graph gives a representation of difference, differential expression, versus intensity, or number of counts. As shown above, increasing the mean of normalized counts results in a decrease in log2 fold change. In other words, as the number of counts increases for BY4741_1 and YLL021W_1, the difference between the expression counts decreases.

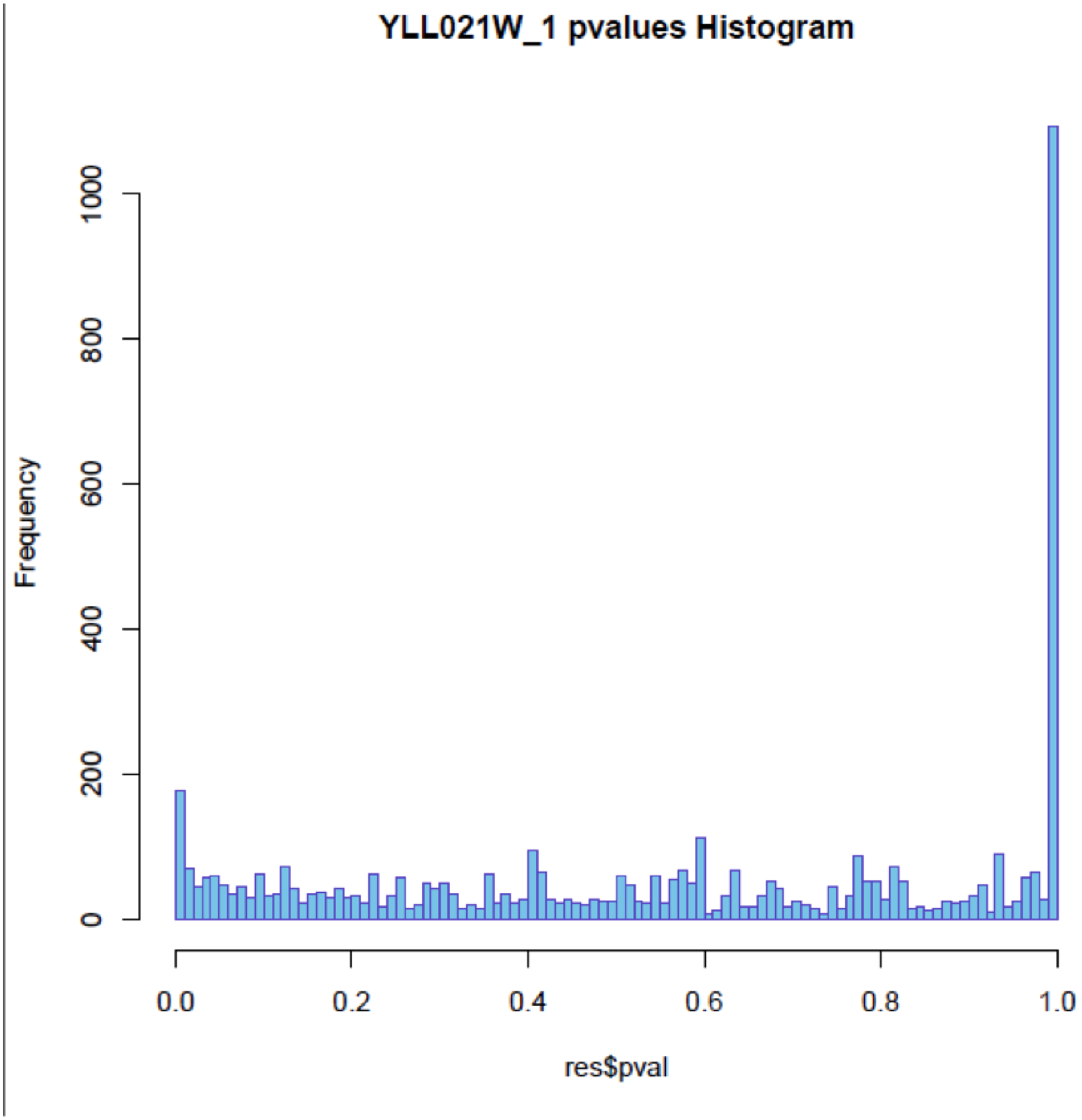


Figure 4c YLL021W_1 p-value distribution. This figure shows a histogram of p-value distribution in gene expression for the knockout YLL021W_1 strain compared to the wild type BY4141_1 strain. In this case, the null hypothesis is that there is no difference between the counts of the two strains for a given gene. Assuming a 99% confidence interval, values of less than 0.01 would reject the null hypothesis, meaning that the difference between gene expression is significant, implying differential expression. However, the most frequent p-value is 1.0, which means that most genes are not are not affected by the knockout.

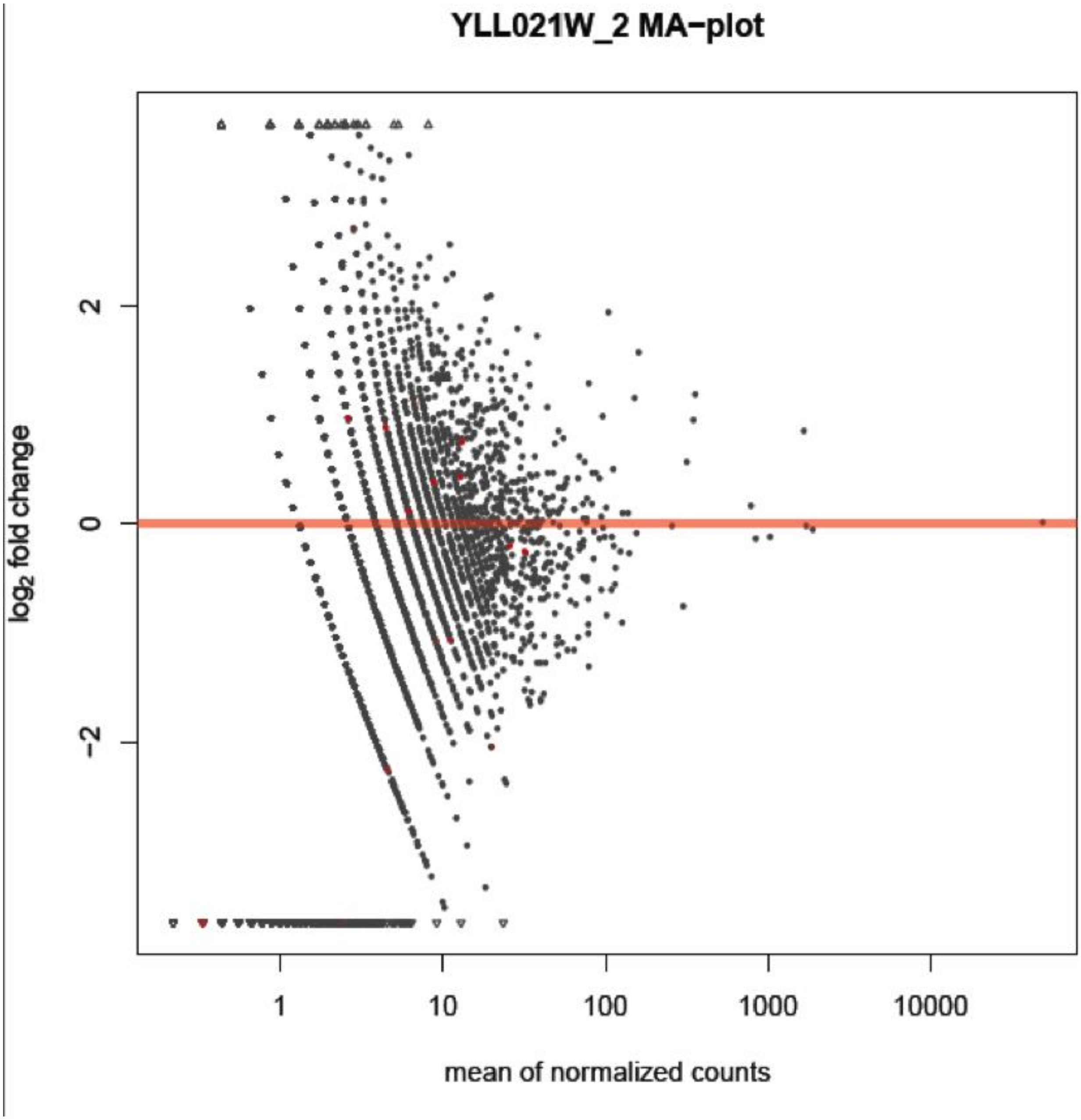


Figure 4d MA plot for YLL021W_2. As in the previous MA plot for YLL021W_1, this graph for YLL021W_2 shows that an increase in the mean of normalized counts results in a decrease in log2 fold change. In other words, as the number of counts increases for BY4741_1 and YLL021W_1, the difference between the expression counts decreases.

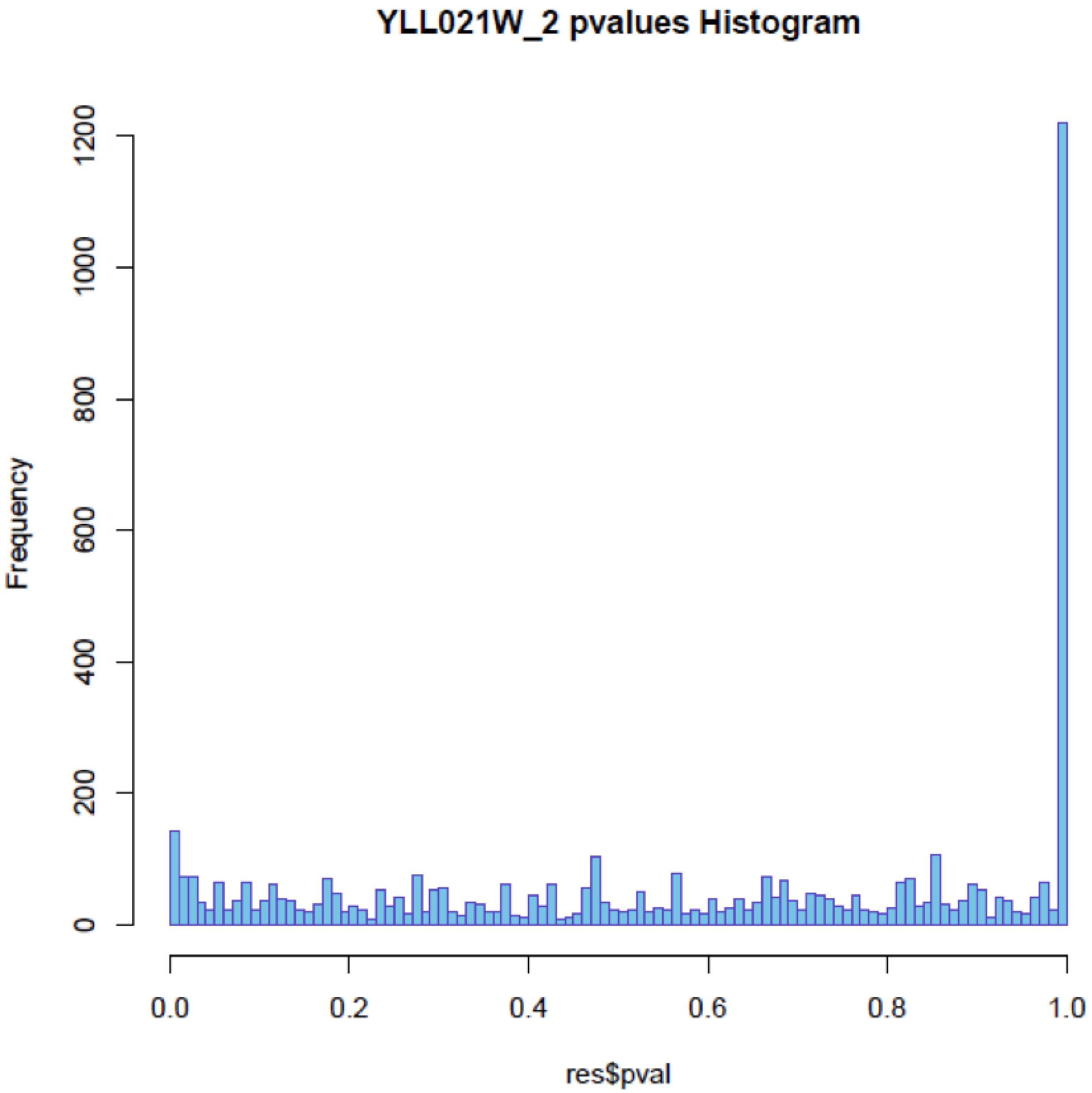


Figure 4e YLL021W_2 p-value distribution. This figure shows a histogram of p-value distribution in gene expression for the knockout YLL021W_2 strain compared to the wild type BY4141_2 strain. In this case, the null hypothesis is that there is no difference between the counts of the two strains for a given gene. Assuming a 99% confidence interval, values of less than 0.01 would reject the null hypothesis, meaning that the difference between gene expression is significant, implying differential expression. However, the most frequent p-value is 1.0, which means that most genes are not are not affected by the knockout.

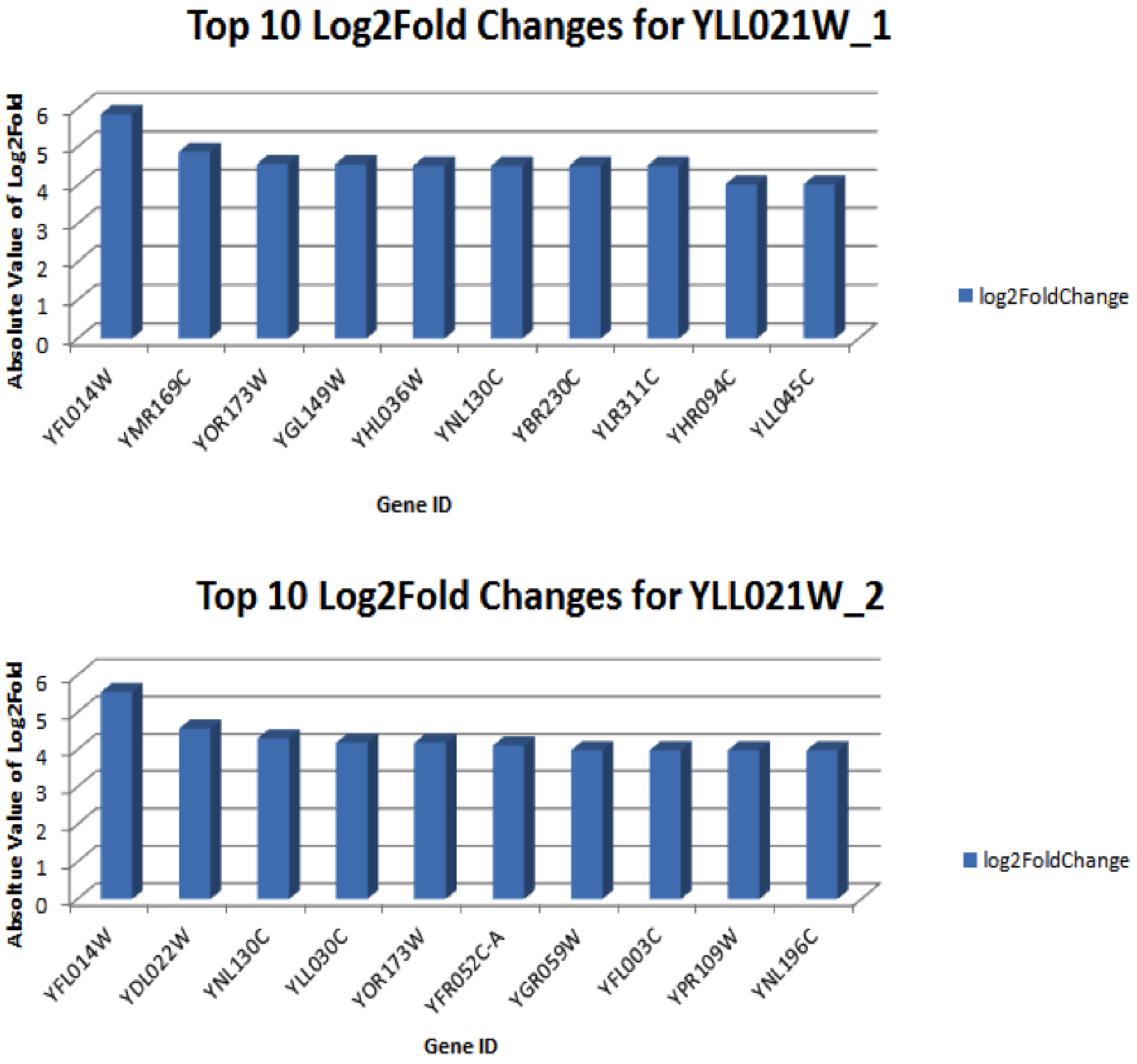
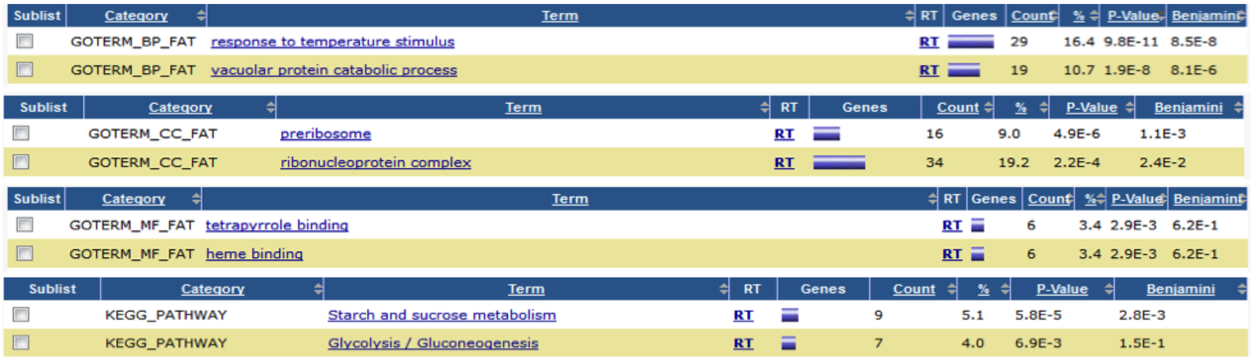


Figure 5 Significant differentially expressed genes for YLL021W. This figure shows the top ten differentially expressed genes for the two strains YLL021W_1 and YLL021W_2 in comparison to BY4741_1 and BY4741_2. Differentiation is quantified by the absolute value of log2 fold change. As shown above, genes such as YFL014W, YOR173W, and YNL130C are common as differentially expressed genes for both the knockout strains.



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_FAT	response to temperature stimulus	RT	▬	29	16.4	9.8E-11	8.5E-8
□	GOTERM_BP_FAT	vacuolar protein catabolic process	RT	▬	19	10.7	1.9E-8	8.1E-6

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_CC_FAT	preribosome	RT	▬	16	9.0	4.9E-6	1.1E-3
□	GOTERM_CC_FAT	ribonucleoprotein complex	RT	▬	34	19.2	2.2E-4	2.4E-2

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_MF_FAT	tetrapyrrole binding	RT	▬	6	3.4	2.9E-3	6.2E-1
□	GOTERM_MF_FAT	heme binding	RT	▬	6	3.4	2.9E-3	6.2E-1

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	KEGG_PATHWAY	Starch and sucrose metabolism	RT	▬	9	5.1	5.8E-5	2.8E-3
□	KEGG_PATHWAY	Glycolysis / Gluconeogenesis	RT	▬	7	4.0	6.9E-3	1.5E-1

Figure 6a Top pathways from DAVID gene annotation for various genes. The top two pathways for GOTERM_BP_FAT, GOTERM_CC_FAT, GOTERM_MF_FAT, and KEGG_PATHWAY. The top two pathways for GOTERM_BP_FAT are “response to temperature stimulus,” which refers to the change in state or activity of the cell towards of external temperature, and “vacuolar protein catabolic process,” which refers to the breakdown of protein in the vacuole. The top two pathways for GOTERM_CC_FAT are “preribosome,” which refers to pre-rRNAs, ribosomal proteins, and associated proteins formed during ribosome biogenesis, and “ribonucleoprotein complex,” which refers to the macromolecular complex containing both RNA and proteins. The top two pathways for GOTERM_MF_FAT are “tetrapyrrole binding,” which refers to selective and non-covalent interactions with tetrapyrrole, and “heme binding,” which refers to the selective and non-covalent interactions with heme. The top two pathways for KEGG_PATHWAY are “starch and sucrose metabolism” and “glycolysis/gluconeogenesis.”

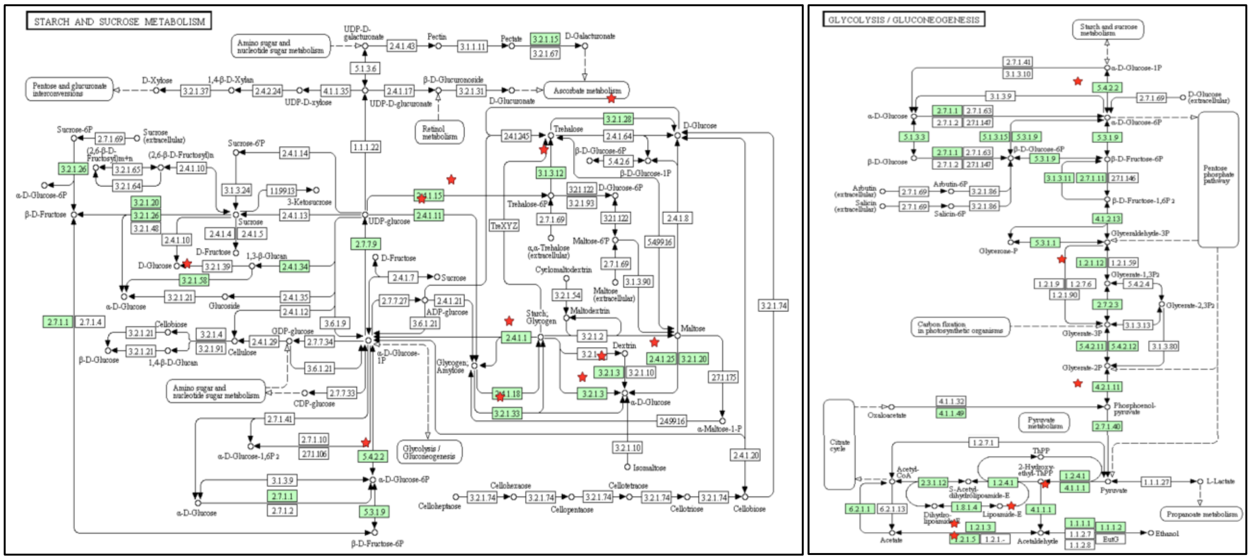


Figure 6b Metabolic pathways of starch and sucrose, and glycolysis/gluconeogenesis. For starch and sucrose metabolism, enzymes such as glycogen debranching enzyme and glycogen phosphorylase are upregulated due to increased gene expression, which causes increased catabolic activity for knockout strains. This also implies downregulation of carbohydrate synthesis. For glycolysis/gluconeogenesis, enzymes such as enolase 1 and phosphoglucomutase-2 are upregulated due to increased gene expression, which causes increased glycolysis activity or downregulation of gluconeogenesis.

Discussion

1. Gene annotation comparisons between S. *cerevisiae* WT and YLL02W strains using DAVID, revealed that the YLL02W, a SPA2 knockout strain, are upregulated for enzymatic starch breakdown genes but downregulated for complex carbohydrate synthesis from sucrose and upregulated for enzymatic glycolysis genes like GAPDH but downregulated for gluconeogenesis. Analysis of the metabolic abnormalities suggest that the SPA2 gene is involved in regulating the direction and control of cell division and polarizing cell growth. [1] SPA2 is involved in pheromone-induced morphogenesis and efficient mating, as positive regulator of actin cytoskeleton reorganization. The Snyder group showed that SPA2 is a necessary gene for efficient mating; upon stimulus via mating pheromone WT S. cerevisiae undergo cellular differentiation to form a morphologically different “schmoo” cell during mating. Double staining showed the SPA2 and actin localize at the “schmoo” tip, gene comparisons between SPA2 sequence and proteins containing coiled structures revealed homology, and although SPA2 mutant cells have increased or similar growth rate the cells are unable to mate efficiently with other SPA2 mutant cells. [2] Thus YLL02W can be used as a model for studying mating impairments in yeast as well as morphologic changes induced by pheromones. Our data collected from DAVID supports morphologic changes that are phenotypically expressed by the SPA2 knockout strain, YLL02W; but, to verify the necessity of SPA2 in mating discussed in literature further experiments would need to be performed such as immunohistochemistry and determining mating efficiency over time.

2. Since clonal rate is defined as:

$$1 - \left(\frac{\text{remaining hits after clonal removal}}{\text{successful alignment reported by Bowtie}}\right) = \text{clonal rate}$$

, in order to reduce the clonal rate we must alter the RNA-seq protocol so that the we must increase the number of remaining hits after clonal removal and/or reduce the successful alignment reported by Bowtie. The RNA-seq protocol can be altered in vitro by changing the concentration of the sample RNA used, the time that PCR was conducted for, and also the temperatures at which PCR was executed at. Concentration, time, and temperature all play important roles in the amount of sample produced. The RNA-seq protocol can be altered in silico by changing the -v command in the bowtie code to control the number of alignments found. The -v command requires a proceeding variable input ranging from 0 to 3; for the analysis performed above -v 3 was set. In order to decrease the number of alignments, meaning that there will be no more than the inputted number of mismatches found (ie. -v 2 means that alignments will have no more than 2 mismatches found while keeping –m value constant), the clonal rate can be increased.

References

[1] Michael Snyder. (1989) The SPA2 protein of yeast localizes to sites of cell growth. *J. Cell Biology*; 108(4):1419-29.

[2] Sonja Gehrung and Michael Snyder. (1990) The SPA2 gene of Saccharomyces cerevisiae is important for pheromone-induced morphogenesis and efficient mating. *J. Cell Biology*; 111(4):1451-64.