Project in Applied Bioinformatics: Reproduction of Critical Roles of the PPP and GLN3 in Isobutanol-specific Tolerance in Yeast

Emma Rydholm, Noel Waters, Leticia Castillon

March 10, 2020

Contents

1 Introduction

Concerns about climate change have motivated the search for better biofuels that can be used to replace fuels derived from petroleum. Branched-chain alcohols, such as isobuthanol, isopentanol and 2-methyl-1-butanol, are considered promising biofuels that can be used as a substitute to gasoline and have superior combustion qualities compare to ethanol (1).
To produce these branched-chain alcohols, microbes, such as *Saccharomyces cerevisiase* can be manipulated and used as a cell factories. *S. cerevisiae* is an especially a good candidate for this job because of its simple genetic manipulation, its ability to grow at low pH, immunity to phage contamination and because it is already used in the majority of large-scale bioethanol production processes, which would ease the transition to large scale production of more advanced bio-fuels.
However, the problem of using microbes as producers of branched-chain alcohols is the cellular toxicity of this desired product, which is a problem since these substances need to be produced at high titers.
With this background, a study was made by Kuroda et al (2) with the purpose to understand the mechanisms of the cellular response induced by isobutanol and eventually enhance *Saccharomyces cerevisiae* tolerance to isobutanol. What they did is to screen a gene deletion library in order to find genes that affect yeast tolerance to isobutanol. They found the gene GLN3Δ that, when deleted, resulted in the highest tolerance to isobutanol. Thereafter they sequenced the transcriptome of the GLN3Δ grown with and without isobutanol as well as the wild-type grown with and without isobutanol. From here, a differential gene expression analysis was made.
The aim with this group project is to replicate the transcriptomic analysis made in the study: Reproduction of Critical Roles of the PPP and GLN3 in Isobutanol-specific Tolerance in Yeast (2).

2 Method

The method used was to follow the instructions provided in the article and to document all commands that were used along the way.
The data for the RNA-seq analysis was obtained through a sequencing experiment performed using Illumina MiSeq. The reads obtained were 75 bp long and nucleotide paired-end sequence.
Importantly, a detailed description of our work, including processed data can be found in the repository Project_Applied_Bioinformatics on github.

2.1 Getting the raw data

The raw data obtained in the study is stored in ArrayExpress. We extracted the FASTq files following the instructions provided on Canvas.

```
#!/bin/bash
wget https://www.ebi.ac.uk/arrayexpress/files/
E-MTAB-8175/E-MTAB-8175.sdrf.txt
```

```
cat E-MTAB-8175.sdrf.txt | cut -f35 | tail -n +2 >
    links2download.txt
wget -N -i links2download.txt #download FASTq files
```

As a result we obtained 16 FASTq files corresponding to our different samples. To understand better the context of the data, the samples are divided in wild-type (WT) strains and GLN3Δ strains. For each genotype, we get two samples growing in normal conditions and two samples growing in 1.3% v/v isobutanol. The RNA-seq experiment was duplicated for each of the samples (2 replicates per strain and condition).

2.2 Alignment

The RNA-seq analysis pipeline starts by aligning the reads obtained from sequencing to a reference. The reference genome used in the original report was *S. cerevisiae* S288C version R64-1-1, from the *Saccharomyces* Genome database.

Trying to be as rigorous as possible when reproducing the analysis, we downloaded the exact same genome as they used. However, we encountered some problems regarding the use of this genome that will be described further on.

```
#!/bin/bash
wget https://downloads.yeastgenome.org/sequence/S288C_reference/
    genome_releases/S288C_reference_genome_R64-1-1_20110203.tgz
tar -xvzf S288C_reference_genome_R64-1-1_20110203.tgz
```

The tool used to perform the alignment is TopHat. TopHat uses the tool bowtie to map the reads to the reference genome, and needs an index of the reference genome. This index was created using bowtie2-build.

```
#!/bin/bash
bowtie2-build -f S288C_reference_sequence_R64-1-1_20110203.fsa \
bowtie_index/reference_genome
```

The version of TopHat we are using is v. 2.0.13. The version used in the study, v. 2.0.9 is no longer available via conda. We experienced some trouble when running TopHat with the flag -G, which allows the program to make use of a GTF/GFF file with known transcripts. We believe the problem was that the GFF file we were trying to use was not formatted correctly for TopHat. Originally, we decided to keep going with the pipeline and try to incorporate this information further on. However, the results we obtained following this strategy were very problematic and we decided to try with a different reference genome and associated GTF file instead. Here's the data that we downloaded to do this:

```
#!/bin/bash
wget ftp://ftp.ensembl.org:21/pub/release-67/fasta/
    saccharomyces_cerevisiae/dna/Saccharomyces_cerevisiae.EF4.67.
    dna.toplevel.fa.gz
wget ftp://ftp.ensembl.org:21/pub/release-99/gtf/
    saccharomyces_cerevisiae/Saccharomyces_cerevisiae.R64-1-1.99.
    gtf.gz
```

```
bowtie2-build -f ../new_data/Saccharomyces_cerevisiae.EF4.67.dna
    .toplevel.fa bowtie_index/reference_genome
```

Since this is the data we used to obtain our final results, we will describe the rest of the pipeline as performed using this data.

```
#!/bin/bash
tophat -G ../../new_data/Saccharomyces_cerevisiae.R64-1-1.99.gtf
    \ -o dgln3_R1/  ../bowtie_index/reference_genome \
../fasta_file_names/dgln31_1.fastq \
../fasta_file_names/dgln31_2.fastq
```

The alignment is done for 8 samples: the wild-type strains under control conditions (2 replicates) and under 1.3% isobutanol (2 replicates) and for the deletion strain $GLN3\Delta$ both under control and 1.3% of isobutanol.

The reference genome is provided in the form of the index files created by bowtie2-build in the previous step. The reads we want to map are provided as an argument as well. It is important to note here that we need to provide the forward and reverse read since we are working with paired-end reads.
The next step is to sort the alignment in order to facilitate some of our downstream procedures. For this we are going to use samtools, which has a 'sort' command that sorts alignments by leftmost coordinates.

```
#!/bin/bash
samtools sort accepted_hits.bam -o accepted_hits.sorted.bam
```

The next step is to use PicardTools to mark the PCR or optical duplicates.

```
#!/bin/bash
picard MarkDuplicates I=accepted_hits_sorted.bam \
O=accepted_hits_sorted_removed_duplicates.bam \
M=marked_dup_metrics.txt
```

Now our alignments are ready for the analysis of the differential expression.

2.3 Differential expression analysis

The differential expression analysis was done using Cufflinks, Cuffmerge and Cuffdiff in succession.
Cufflinks takes the output from TopHat and assembles the transcriptomes. It also quantifies their expression and reports it in FPKM form. Expression is reported using this expression unit because we are working with paired-end reads, where fragments and reads do not have a one to one ratio.

```
#!/bin/bash
cufflinks -g ../../new_data/Saccharomyces_cerevisiae.R64-1-1.99.
    gtf -o wt_isobutanol_R2_clout/ \
../tophat_out/wt_isobutanol_R2/
    accepted_hits_sorted_markduplicates.bam
```

The flag -g is used to incorporate the GTF annotation file. In this way Cufflinks, similarly to TopHat in the previous step, uses the annotation file to incorporate additional information to the assembly. Cufflinks is performed to all the alignments we got using TopHat. As a result we get an assembled transcriptome for each of our RNA-seq libraries.

To merge transcripts into larger, "master" transcriptomes, we use Cuffmerge. We get two merged transcriptomes, one for the wild-type strain and another one for the ΔGNL3 strain.

```
#!/bin/bash
 cuffmerge -g ../../new_data/Saccharomyces_cerevisiae.R64
    -1-1.99.gtf -s ../../new_data/Saccharomyces_cerevisiae.EF4
    .67.dna.toplevel.fa -p 8 -o cmout/ assemblies.txt
```

The merged transcripts file is an input argument needed by Cuffdiff.
Cuffdiff performs the actual differential expression analysis step by comparing the expression levels, or FPKM counts, and also reports FDR adjusted p-values. More info on the whole pipeline is found here (3).

```
#!/bin/bash
cuffdiff -M mask_gln3.gtf -o diff_out_wt -u merged.gtf \
../wt1/accepted_hits.bam,../wt2/accepted_hits.bam \
../wt_isobutanol_1/accepted_hits.bam,../wt_isobutanol_2/accepted_hits.bam
```

Observe that we are providing a mask file as indicated by the flag -M. This mask file is a GTF file with a list of transcripts. By adding it to our command we are telling Cuffdiff to ignore all transcripts in our assembly that are also present in the mask file. The reason for this is that we want to ignore rRNA, tRNA and non-coding RNA in our differential expression analysis. This is to improve the accuracy and robustness of the analysis overall, since mRNA amount is quite low in the cell when compared to all the other RNA transcripts.
To get our mask file, we identified all the transcripts marked as tRNA, rRNA, etc. in our annotation .gtf file and matched to the merged annotation:

```
#!/bin/bash
cut -f 9 Saccharomyces_cerevisiae.R64-1-1.99.gtf | cut -f 1,4 \
--delimiter=";" | grep -e "snoRNA" -e "rRNA" -e "tRNA" \
| cut -f 1 --delimiter=";" > masking_indices
cut -f 2 -d ' ' masking_indices | sed 's/"//g' > only_IDS
grep --file=only_IDS merged.gtf > mask_gln3.gtf
```

In essence, we found all rows of the merged.gtf file corresponding to undersired types of RNA and stored in another .gtf file.

2.4 GO Term association

We want to get the GO term association for the genes that are differentially expressed. In this way we can get an idea of what cellular processes are changing the most in the different conditions and/or affected by the deletion of the gene GLN3. To do this, we use the Gene Ontology Term Finder implemented in the *Saccharomyces*

Genome Database. The GO Term Finder requires that you provide a list of gene names as input. We used one the output files of Cuffdiff, named genes_exp.diff. From here we extracted all the genes that Cuffdiff marked as significantly differentially expressed:

```
#!/bin/bash
grep "yes" gene\textunderscore exp.diff \
| cut -f 3 > DE_genes_goterms.txt
```

The .txt file generated was used as input for the GO Term finder. Note here that at first the finder complained because some of our gene names were associated to more than one gene, for some of them being the official name and for a number of other genes being the alias. This was solved changing the gene name to the systematic name. An example of this is provided below:

```
#!/bin/bash
sed -i '' 's/CTR1/YPR124W/g' DE_genes_goterms.txt
```

We then performed the GO Term search by choosing the Ontology Aspect "Process" instead of "Function". The main reason for this is that we want to get an output comparable to the one they obtained in the paper, where they were mainly interested in the cellular processes affected by isobutanol or deletion of certain genes.

2.5 Building the Volcano Plot

The Volcano plot is produced in R using the package EnhancedVolcano (https://github.com/kevinblighe) from Bioconductor. The output file `gene_exp.diff` from Cuffdiff is used to create the plot and it is the -log_{10} P value that is plotted against log_2- fold change. To get the correct labels for the genes in the plot, the file containing the GO-terms result was merged with the data from `gene_exp.diff`, and the GO-terms was used to create the labels. However, the GO-term file includes hundreds of GO-annotation and we choosed to look label only a few processes, the ones that they had found to be the largest group of genes induced by isobutanol in the wild type.

3 Results

3.1 Alignment

The output of TopHat gives a summary of the alignment. This serves as an overview of the quality of the alignment. In our case, we got a concordant pair alignment rate higher than 90% for all the alignments. A concordant pair alignment means that the paired reads have been satisfactorily aligned with respect to each other based on the a priori knowledge of their relative position. This is a high percentage, which is good because anything not included in a corcondant pair will not be used in further downstream analysis. We get a high number of multiple alignments: 77.9%. Some multiple alignment is expected because a genome contains repeated regions and pseudogenes. However, this seems a bit high. Maybe it is a consequence of not having performed any preprocessing on the data prior to alignment.

3.2 Differentially Expressed genes & GO Term association

From Cuffdiff differential expression calculation we get 271 genes that are considered to be differentially expressed. This output is slightly higher than the number of differentially expressed genes found in the original paper, which is 231 in the Wildtype 0 vs 1.3% analysis.
When seeking the GO terms associated with our differentially expressed genes, we find three out of the four categories they were interested in. The processes that we get are "Cell wall organization or biogenesis", "Cellular amino acid metabolic process" and "Glycolytic process". We fail to obtain the process "Membrane lipid biosynthesis".

3.3 Volcano plot

The resulting Volcano plots from our analysis are shown in figure 2 and 1. Comparing our result in figure 2 and 1 with the original result in figure 3 it seems like our results agree with theirs.

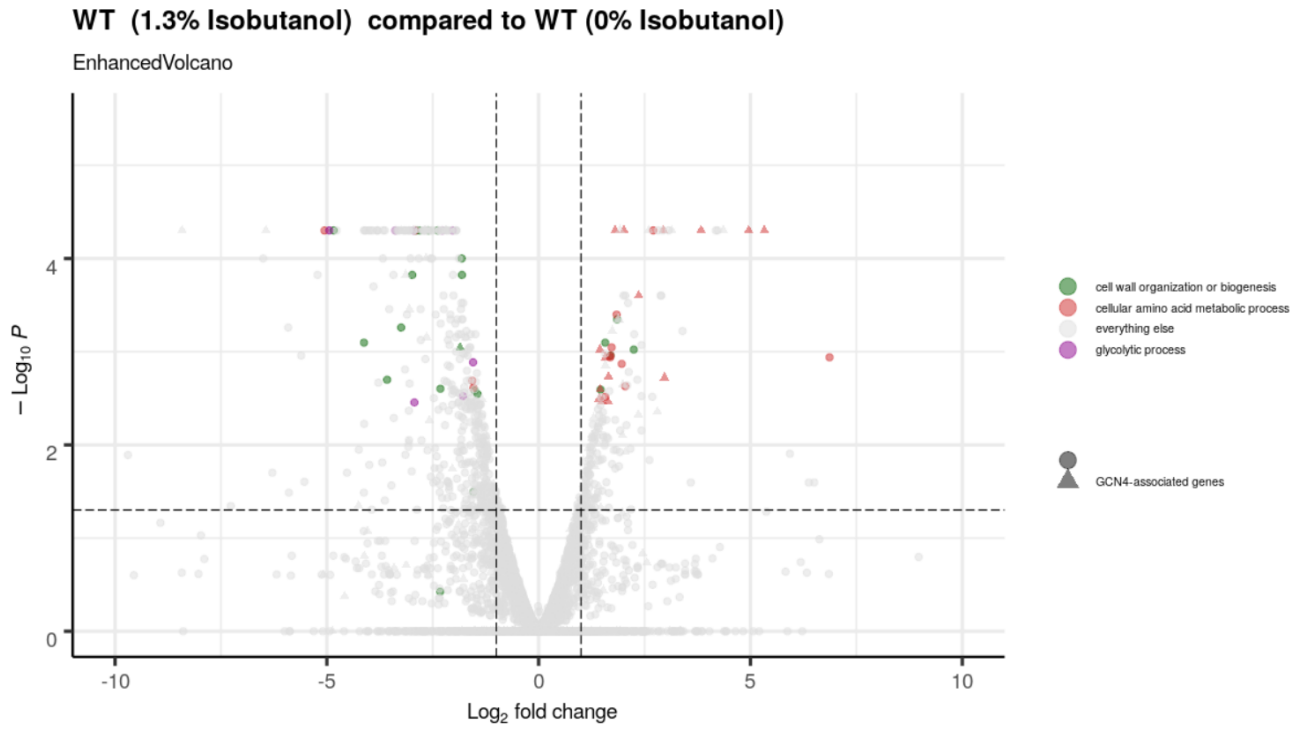

Figure 1: Resulting volcano plot from our analysis. -log_{10} P- value is plotted against log_2 fold change for WT in 1.3 % Isobutanol in comparision to WT in 0% isobutanol.

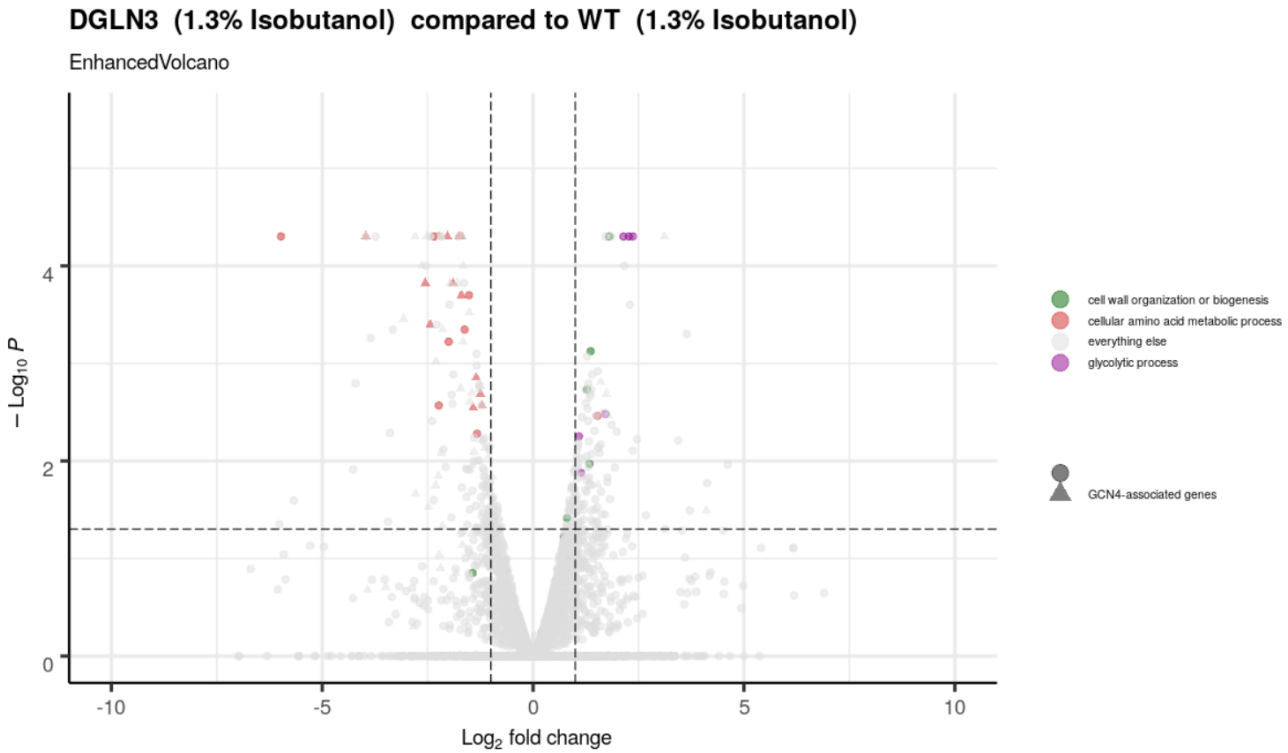

Figure 2: Resulting volcano plot from our analysis. -log_{10} P- value is plotted against log_2 fold change for GLN3Δ in comparision to WT in 1.3 % Isobutanol.

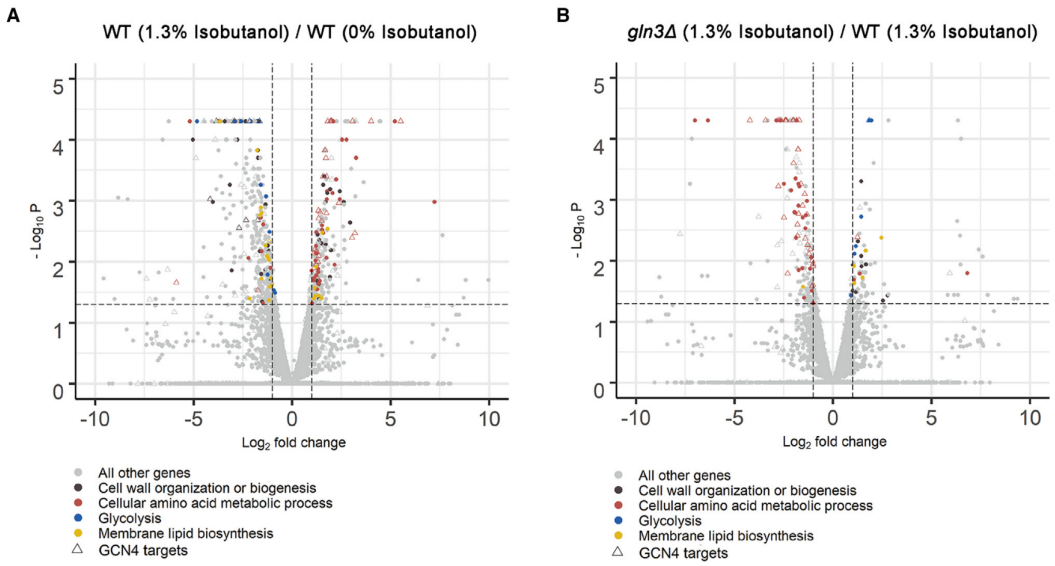

Figure 3: Volcano plots obtained from the original article (2)

In the result for the wildtype in 1.3% isobutanol compared to Wild type in 0% isobutanol, we can see that our result in figure 1 shows an increase in the gene expression for the genes involved in *cellular amino acid metabolic process* and a decrease in the expression of genes involved in *cell wall organization or biogenesis*

and *glycolytic processes* which is also shown in figure 3.
When looking at GLN3Δ in in 1.3% isobutanol compared to the wildtype in 1.3% isobutanol the gene expression for *cellular amino acid metabolic process* is lower for GLN3Δ and the gene expression for *cell wall organization or biogenesis* and *glycolytic processes* is higher in GLN3Δ compared to wildtype, which also agrees with what is shown in figure 3.

4 Discussion

The purpose of our project was to try to reproduce the results obtained by Kuroda et al in (2), which was roughly successful judging by the number of differentially expressed genes and the volcano plots. However, it could also be interesting to look closer to the fold changes for specific genes and see if we obtained similar results. Next, we discuss some differences and possible reasons for them.
We compared their differentially expressed genes against ours and we obtained 227 genes in common. This means that we just missed 4 genes from their results, but got 44 genes that were not considered differentially expressed by their analysis.
Also, in our results, none of our differentially expressed genes seemed to be involved in any process involving lipid biosynthesis or metabolism, a process they clearly found since it is labeled in yellow in their volcano plots. However, this could be an artifact of how the GO Term finder algorithm works and how this changes over time.
They didn't mention any kind of pre-processing on the RNA-seq raw data, something we assume that they did since it's standard procedure. However, as they omitted it in the report, we ignored this step. Lower quality data increases the risk of misleading conclusions, and may well have contributed e.g to us finding more differentially expressed genes. Furthermore, as mentioned in section 2.2, the reference genome that was used for the alignment in our final version is not the same that they used because we encountered some problems using those files. On top of this, the version of TopHat that we used is also different, as indicated in section 2.2 as well.
In the cufflinks pipeline, they did not specify how they removed the duplicates using the PicardTool package MarkDuplicates, which is another issue since that carried multiple options.
Other things they failed to mention which could have had an effect in the output of the analysis are the sources for the Mask file necessary for Cuffdiff, and the list of genes regulated by GCN4. The maskfile facilitates a more accurate calculation of the differential expression of the genes. Therefore, if we failed to ignore some transcripts that they did ignore it will have affected our results. However, most likely this is less significant than the other issues mentioned above. The GCN4 issue affects the labelling of the Volcano plot and is therefore not central to the computation steps.
To summarise, there are some parts of the pipeline where they could have been more elaborate, but overall the pipeline was possible to follow and the final results are not too different. Furthermore, the origin of some of the problems is beyond their control, like the version of the programs available or the algorithm used by external tools like GO Term Finder. Thus, reproducibility will always be affected by how much time has passed since the study was performed.

References

[1] Wess J, Brinek M, Boles E. Improving isobutanol production with the yeast Saccharomyces cerevisiae by successively blocking competing metabolic pathways as well as ethanol and glycerol formation. Biotechnology for biofuels. 2019;12(1):173.

[2] Kuroda K, Hammer SK, Watanabe Y, López JM, Fink GR, Stephanopoulos G, et al. Critical Roles of the Pentose Phosphate Pathway and GLN3 in Isobutanol-Specific Tolerance in Yeast. Cell systems. 2019;9(6):534–547.

[3] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley D, et al. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. Nature protocols. 2012 03;7:562–78.