Draft DE Analysis on L. plantarum on ribose and glucose

Sanne, Pim, Milain

1/9/2020

Contents

```
## Loading required package: limma
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

Title Page

Abstract

For this study RNA-seq data was used to investigate the gene expression of *L. Plantarum.* In this study two samples of WCFS1 genes were cultivated on a Ribose base and on a Glucose base. This study was carried out to test whether genes required for metabolizing ribose are upregulated in *L. Plantarum* grown on ribose-rich medium The data was normalized and filtered for low expressed genes. The fold changes were then calculated along with their adjusted p-values. The study concludes the hypothesis to be true, genes required for metabolizing ribose are upregulated when grown on a ribose rich medium. Further research could be done to determine in which capacity the genes are upregulated. Preferably this would be carried out with a larger sample size, as this study was limited by the size and reliability of the data provided.

Introduction

In 2010, a whole-transcriptome of *L. Plantarum* WFCS1 was deep-sequenced on the Illumina platform in cooperation of Wageningen University, NIZO and the Bacterial Genomics research group at the CMBI (Nijmegen, The Netherlands). NIZO and the Bacterial Genomics research group at the CMBI use *L. Plantarum* WCFS1 as a model organism to study the biology of a bacterium that is also of commercial interest, e.g. in probiotics and for metabolic engineering. This organism is a Firmicutes that is a natural inhabitant of mammalian gastrointestinal tracts (Ahrne et al. 1998). The strain Lactobacillus plantarum WCFS1 is the first Lactobacillus for which the entire genome was determined (Kleerebezem et al, PNAS, 2003). With 3.3 million nucleotides, the genome of *L. Plantarum* WCFS1 is among the largest Lactobacillus genomes (Kleerebezem, 2003) and comprises 3135 predicted genes, of which 3007 are protein-encoding (Refseq NC_004567). To further unravel the biological functions of proteins, learn about gene regulations and improve current genome annotation, whole-genome transcriptomes of *L. Plantarum* WCFS1 under various growth conditions were collected. (mogelijke begrippen uitleggen hier) For this study, a whole-genome transcriptome of *L. Plantarum* WCFS1 is used. The dataset has been generated using RNA-Seq and were derived from biological samples of *L. Plantarum* grown on two different media (glucose versus ribose). With the dawn of next generation (or deep) sequencing technologies in recent years (Ansorge, 2009; Metzker, 2010), their application to high-depth sequencing of whole transcriptomes, a technique now referred to as RNA-Seq, has been explored (Morozova et al., 2009; Wang et al., 2009; Wilhelm and Landry, 2009). RNA-Seq requires a conversion of mRNA into cDNA by reverse transcription, followed by deep sequencing of this cDNA (van Vliet, 2010). RNA-Seq was initially only used for analysing eukaryotic mRNA, as prokaryote mRNA is less stable and lacks the poly(A) tail that is used for enrichment and reverse transcription priming in eukaryotes. But these technological difficulties are being overcome, as various methods for enrichment of prokaryote mRNA and appropriate cDNA library construction protocols have been developed, some generating strand-specific libraries which provide valuable information about the orientation of transcripts. In June 2008, the first reports appeared of RNA sequencing of whole microbial transcriptomes, i.e. the yeasts Saccharomyces cerevisae (Nagalakshmi et al., 2008) and Schizosaccharomyces pombe (Wilhelm et al., 2008). Both studies demonstrated that most of the nonrepetitive sequence of the yeast genome is transcribed, and provided detailed information of novel genes, introns and their boundaries, 3 and 5 boundary mapping, 3 end heterogeneity and overlapping genes, antisense RNA and more. Starting in 2009, several examples have been reported of prokaryote whole-transcriptome analysis using tiling arrays and/or RNA-seq. The first reviews of prokaryote transcriptome sequencing have just appeared (Croucher et al., 2009; van Vliet andWren, 2009; Sorek and Cossart, 2010; van Vliet, 2010).

The dataset is used to compare the change of gene expression of *L. Plantarum* across different growth conditions and to identify genes that are significantly up- or down regulated when grown on ribose. These genes have been identified using R by first normalizing and filtering the data for low expressed genes followed by calculating fold changes and making a selecting of the top genes. These top genes were further analyzed using KEGG mapper and an annotation file for Pathway Enrichment Analysis. (waarom word dit gedaan???? Search for the catabolite responsive element upstream to the selected genes.) (bevindingen..)

Material & Methods

The dataset is used to compare the change of gene expression of *L. Plantarum* across different growth conditions and to identify genes that are significantly up- or down regulated when grown on ribose. These genes have been identified using R by first normalizing and filtering the data for low expressed genes followed by calculating fold changes and making a selection of the top genes. These top genes were further analyzed using KEGG mapper and an annotation file for Pathway Enrichment Analysis.

Data collection

The dataset has been collected and filtered via RNA-seq. Data was represented in two files, one consisting of gene identifiers and their read count for each sample, the second holding their annotation.

Data preprocessing

Normalisation

To normalize and filter the data EdgeR has been used. There has been made use of the functions DGEList for easy editing followed by calcNormFactors using the trimmed mean method(TMM) to normalize. This method normalized by removing the lowest and highest values (percentile) and calculating mean.

Check statistics

To test these results the method summary is used to show the statistics of the counts in the dataset.

```
## [1] "Count statistics"
```

```
##   WCFS1.glc.1      WCFS1.glc.2      WCFS1.rib.1      WCFS1.rib.2   
##  Min.   :     0   Min.   :     0   Min.   :     0   Min.   :     0  
##  1st Qu.:   115   1st Qu.:   120   1st Qu.:   129   1st Qu.:   144  
##  Median :   610   Median :   639   Median :   634   Median :   684  
##  Mean   :  3666   Mean   :  3818   Mean   :  3611   Mean   :  3778  
##  3rd Qu.:  2362   3rd Qu.:  2422   3rd Qu.:  2331   3rd Qu.:  2538  
##  Max.   :304949   Max.   :348490   Max.   :299861   Max.   :351280  
```

```
##                 group lib.size norm.factors
## WCFS1.glc.1 WCFS1.glc 11206738    1.0348799
## WCFS1.glc.2 WCFS1.glc 11671184    1.0278238
## WCFS1.rib.1 WCFS1.rib 11039613    0.9468222
## WCFS1.rib.2 WCFS1.rib 11549518    0.9929399
```

Create matrix

Next the dispersion was estimated this was done by first creating a design matrix using the model.matrix function. This design is used to let R know the 2 groups.

```
##             WCFS1.glc WCFS1.rib
## WCFS1.glc.1         1         0
## WCFS1.glc.2         1         0
## WCFS1.rib.1         0         1
## WCFS1.rib.2         0         1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

Estimate Dispersion

To calculate the dispersion three methods have been used `estimateGLMCommonDisp`, `estimateGLMTrendedDisp(method="bin` and `estimateGLMTagwiseDisp`. The recommended parameter for datasets with more than 200 genes for the `estimateGLMTrendedDisp` function is `method="bin.spline"`.

Checking Dispersion

The outcomes of these functions were added to the DGE list and inspected by making an MDS plot and a BCV plot.

Determine fold changes

Construct the contrast matrix corresponding to specified contrasts of a set of parameters.

Data processing

To process the collected data first the foldchange was determined. This was done by fitting a negative binomial generalized log-linear model to the read counts for each gene using the glmFit method followed by making a contrast using the makeContrast method to determine "glc vs rib" or "rib vs glc"

and making a glmLRT with the data gathered by the glmFit.

glmLRT conducts likelihood ratio tests for one or more coefficients in the linear model and so decides which genes are upregulated, downregulated or not significant. To get a list of the most differentially expressed genes ,ranked on either p-value or absolute log fold change, the toptags method has been used on the glmLRT data.

```
## Coefficient:  1*WCFS1.glc -1*WCFS1.rib
##              logFC    logCPM       LR        PValue           FDR
## lp_3539 -12.277544 11.815752 1918.311  0.000000e+00  0.000000e+00
## lp_3538 -11.879920 13.813096 3385.032  0.000000e+00  0.000000e+00
## lp_3540 -11.683235 12.411008 1931.621  0.000000e+00  0.000000e+00
## lp_3542 -11.427840 11.875177 1662.353  0.000000e+00  0.000000e+00
## lp_3570  -9.041299 11.561244 1389.846 3.379310e-304 2.066110e-301
## lp_3658  -8.107730 11.744789 1320.546 3.874534e-289 1.974075e-286
## lp_3569  -8.632696 11.871533 1265.186 4.157698e-277 1.815726e-274
## lp_3541 -10.920173 10.415195 1201.651 2.669622e-263 1.020129e-260
## lp_3537 -12.282873  9.935179 1187.072 3.934253e-260 1.336335e-257
## lp_3660  -8.372607 11.052275 1130.010 9.917972e-248 3.031924e-245
```

To inspect this data it was visualized using the plotMD method

Now that the foldchange was determined, a selection of differentially expressed genes had been made using only the genes with a p-value lower than 0.05 and a fold change lower than -2 or higher than 2.

This selection was then merged with an annotation file containing the ORF, start, stop , orientation, name, function, class, subclass, EC number and subcellular location.

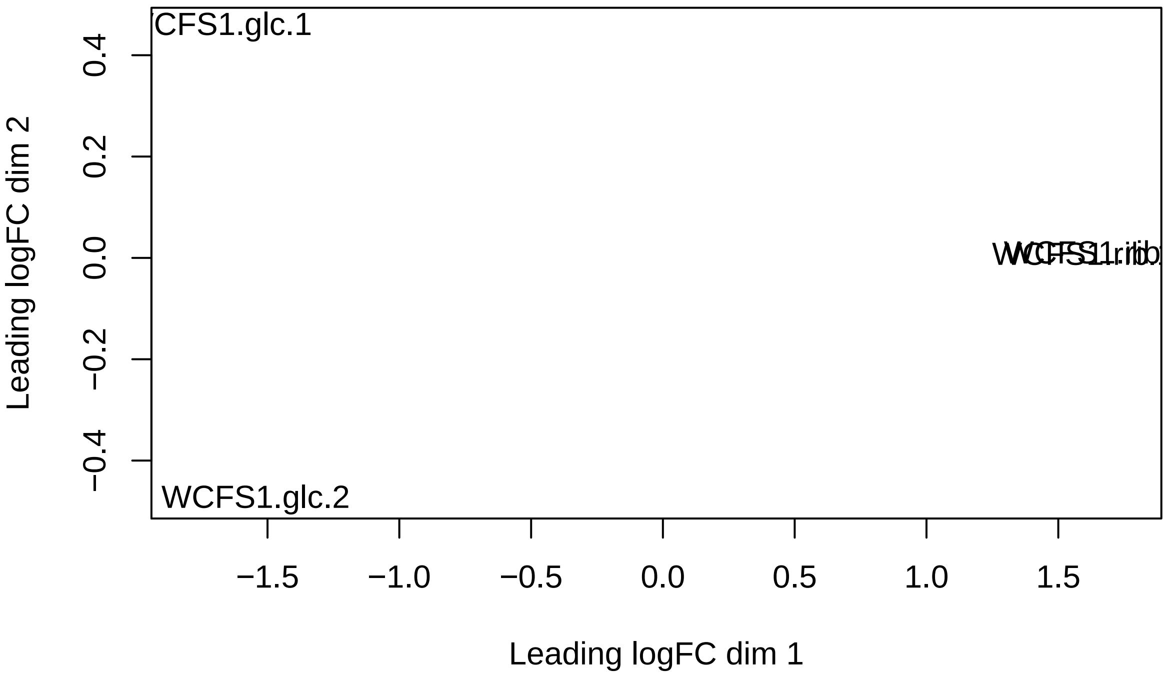
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Analyzing To gain insight in what genes were connected to metabolizing of ribose, keggmaper was used together with the annotation file using the function, class and subclass to determine if ribose was involved.

All genes that had a known EC number were selected and put into keggmapper, the found pathways were examined. To visualize the data in the annotation file a barplot was made using ggplot2, showing the foldchange of every gene colored by their class and another plot with the same but colored by their subclass. Catabolite responsive elements were analyzed???

results

Figure 1 - MDS plot



This is a two dimensional scatter plot showing the distances approximating the log2 fold changes between the two samples. Relatively the difference between the glucose and ribose samples are larger than the difference between the two types of samples. To be exact 0.9 between the glucose samples in the y dimension and almost no difference between the ribose samples on both dimension, while the difference between the different types of samples is at minimum 3.0 and 0.4 in the x and y dimensions respectively.

Figure 2 - BCV plot

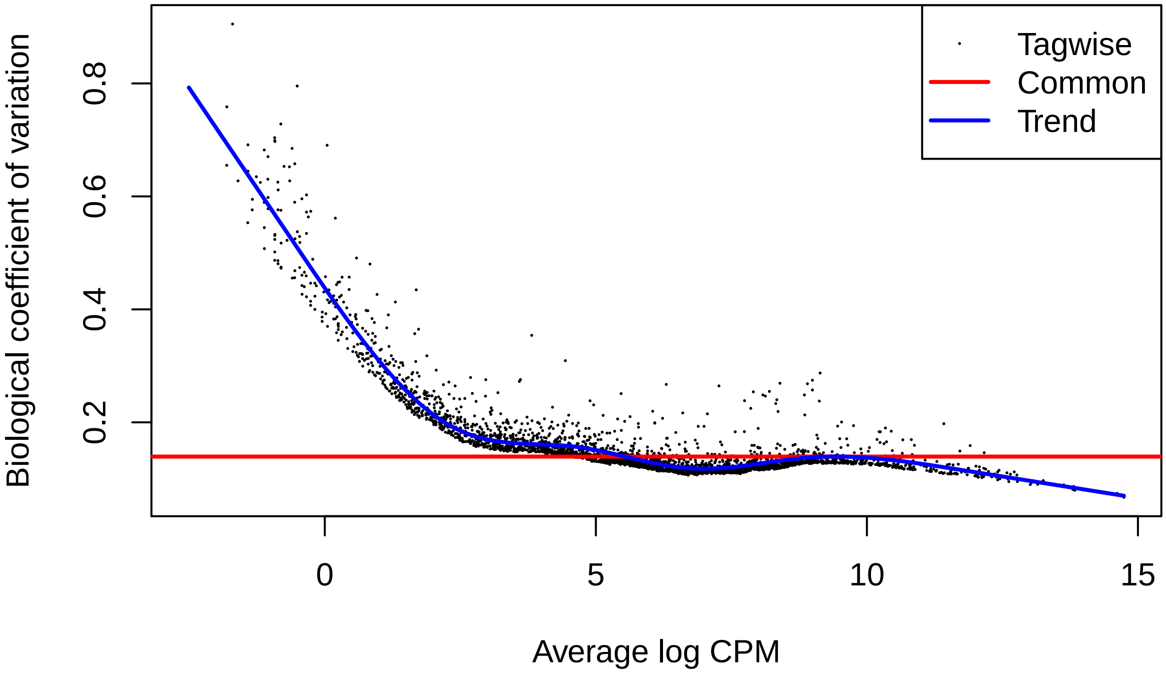
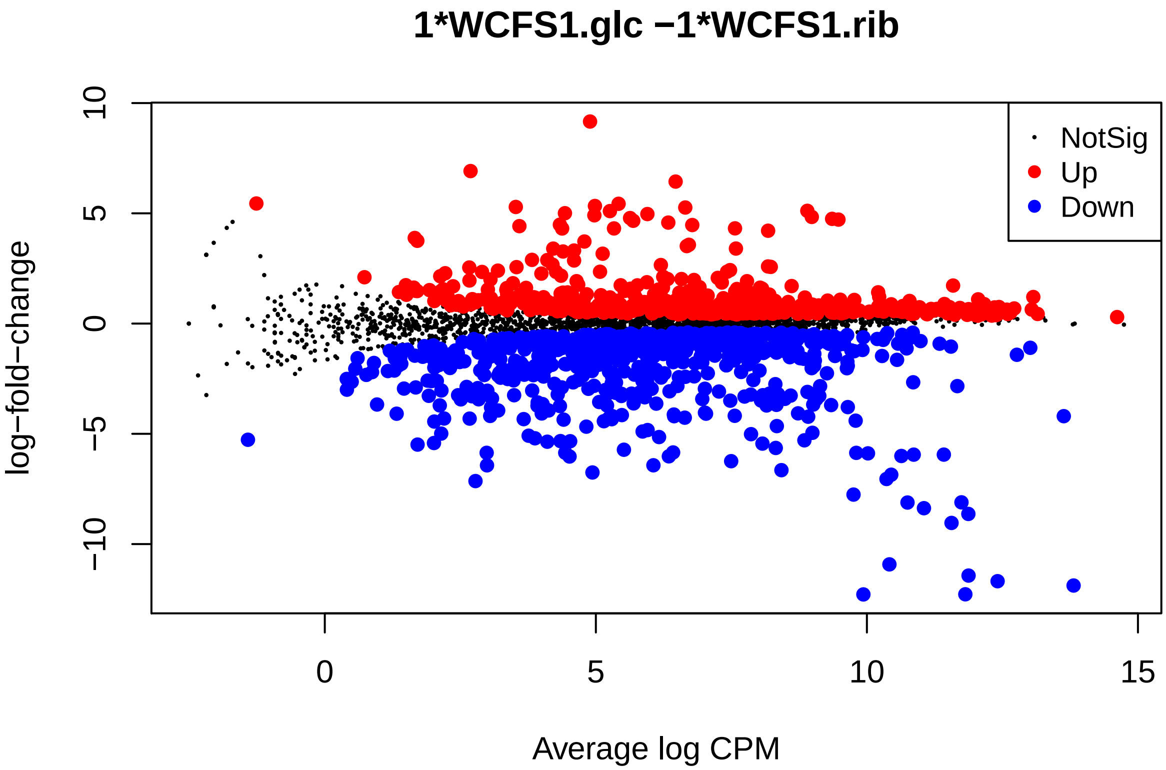


Figure2. In this Biological Coefficient of Variation (BCV) plot you can see the genewise biological coefficient of variation against gene abundance, in log2 Counts Per Million (CPM). The BCV appears to be higher when the average log2 CPM is lower .

Figure 3 - MD plot



In this Mean Difference (MD) plot the log fold changes are plotted against the average log CPM. Here you can see that there is a difference in the amount of up- or downregulated genes and the variation in the log fold change. There are more downregulated genes and between these genes the variation in log fold change appears bigger.

```
##        1*WCFS1.glc -1*WCFS1.rib
## Down                        676
## NotSig                     1886
## Up                          495
```

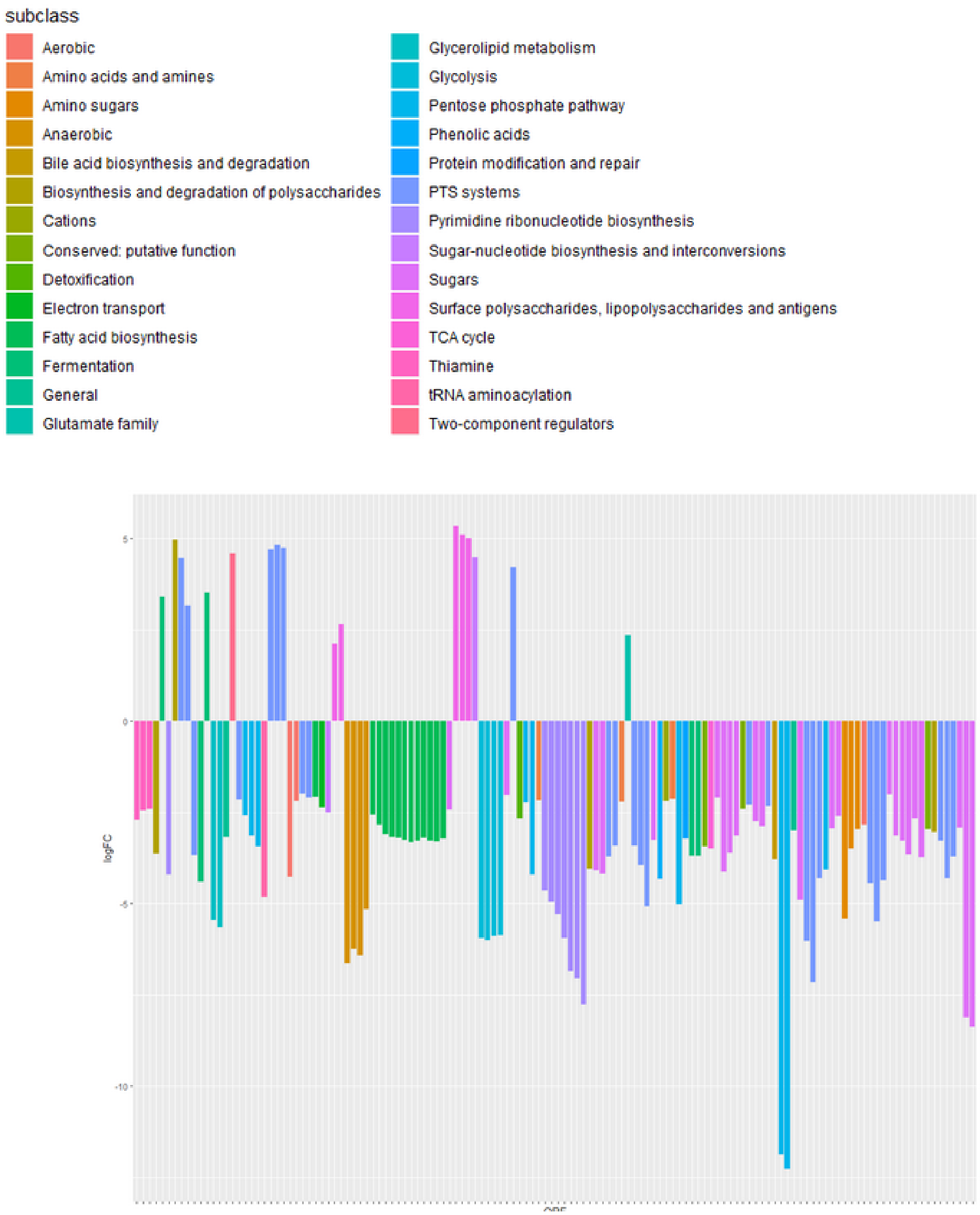


Figure 1: Here you can see the log fold changes per gene coloured by subclass on glucose. The Pentose phosphate pathway appears to be heavily downregulated on glucose. Everything below the 0 line is downregulated on glucose and everything above the 0 line is upregulated on glucose. If the plot was from the ribose samples the opposite would be visible.

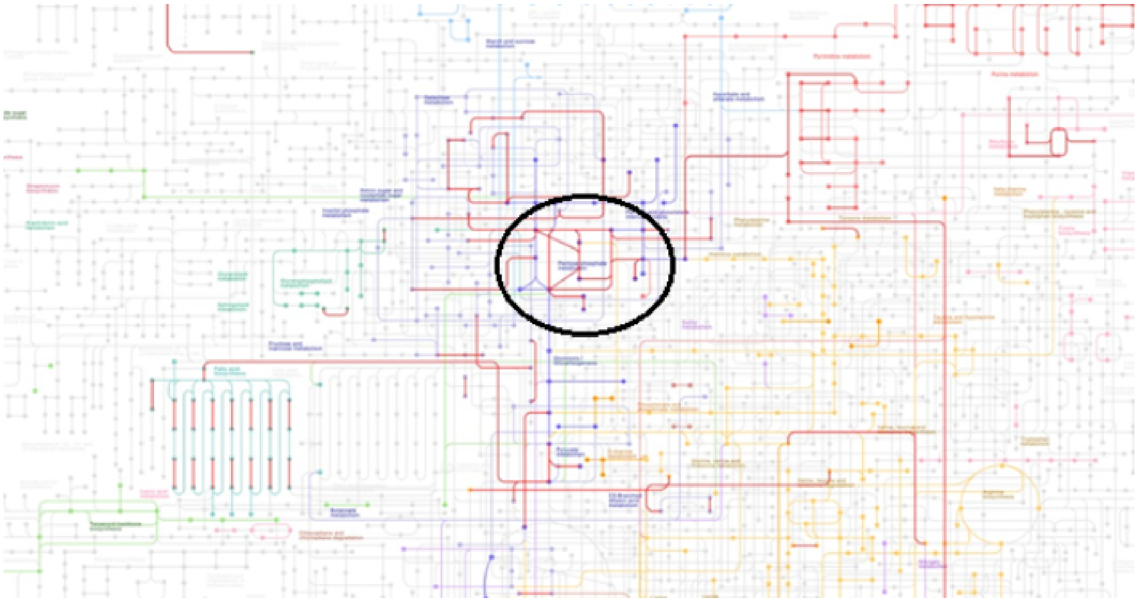


Figure 2: In this figure the pathways of the upregulated genes on ribose are shown in red on the KEGG (KEGG reference) metabolic reference pathway map. The encircled pathways are pathways that contain the most heavily upregulated genes, these lay within the pentose phosphate pathway. This would indicate that this pathway would almost certainly be more active in L. plantarum grown on ribose.

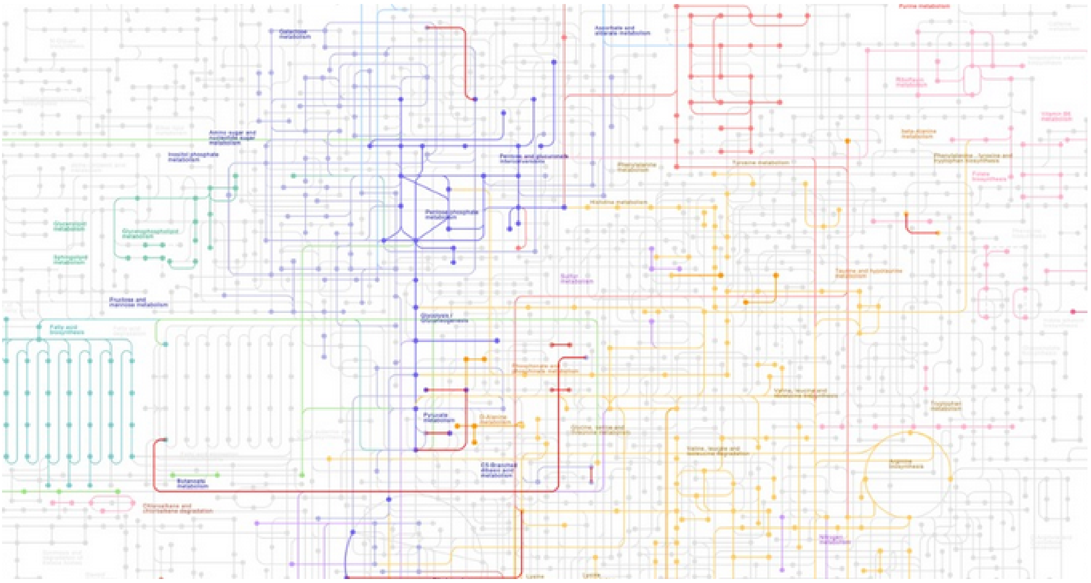


Figure 3: This figure below shows the upregulated pathways on glucose. The major difference is the lack of upregulated pathways