Using RNAseq data to analyze gene expression in *Haemophilus parainfluenzae* aerobically and anaerobically

Thaís Palma

May 13, 2019

Introduction

The oral cavity harbors a very dynamic and complex microbial community. More than 700 species of microorganism colonize different niches in the mouth. Among them, one of the most abundant species is *Haemophilus parainfluenzae (H. para*). *H. para* is a Gram-negative, facultatively anaerobic coccobacillus. Even though it is a commensal bacterium in the oral cavity, in some cases it can be associated with local diseases, such as gingivitis (Park et al, 2015), or systemic diseases such as endocarditis. It is estimated that *H. para* is responsible for approximately 3% of the infective endocarditis cases (Posse et al, 2018). Despite its prevalence, *H. para* is not very well studied and more research is necessary to better understand its metabolism, and its genes expressions under different circumstances associated with its opportunistic pathogenicity characteristic. Based on that, this paper aims to investigate the difference in genes expression of *H. para* under aerobic and anaerobic conditions using an RNAseq database. Below I describe the data and report the initial results of my study. Out of 1764 observations generated by RNAseq, 258 had significantly different levels of genes expression (88 were upregulated under anaerobic conditions and 170 under aerobic conditions). Out of the 258 significant results, 48 (27 anaerobic, and 21 aerobic) presented 'fold change' equal or higher than 2.5. Future research will analyze each one of the 48 genes and, for some of them, mutant strains will be built and tested under different conditions.

Methods

Strain, growth condtion and RNA extraction

The sample *Haemophilus parainfluenzae* (ATCC 33392) was used to perform the RNAseq analyses. The sample was streaked on brain heart infusion (BHI) plates and incubated in 5% CO2 for 24 hours at 37oC. After that, one isolated colony was transfered to 3 ml of BHI liquid and incubated either aerobically or anaerobically, using a candle jar for 24 hours at 37oC. Next, 20 μl of culture was spoted on a membrane and grew for 24h under the same conditions as mentioned on the previous step. Finally, the membrane was washed with Trizol to avoid RNA degradation and frozen (-80oC). The samples were sent to have the RNA extracted and sequenced.

RNA analyses

The raw data was processed and analyzed using HTSeq to count the reads and Bowtie to align them. After that, I used an R script (available on github) to generate the table that is the base to this paper. Statistical analyzes, graphs, and tables were generated using Rstudio.

R analyses

The first step was to organize the table. To do so, I used the function "rename" to rename the columns and label the unlabeled ones. Next, I filtered all the cells that did not have value specified with the function "filter". The function "mutate" was used to create new variables that were not in the original table. For instance, I created a column named condition to specify whether the samples were incubated under aerobic or anaerobic condition.

To differentiate gene expression under the conditions aerobic and anaerobic, I considered and filtered the values of 'fold change'. This variable compares how many times (fold) the genes expression is smaller or larger in the two conditions. For instance, when the fold change was 1, that means that there is no difference in genes expression between the two conditions. Values larger than 1 indicate higher expression in anaerobic conditions, while values smaller than 1 indicated higher expression aerobically.

Given the large number of observations generated originally (1764 observation), I concentrate only on the ones that have significant fold change, i.e., where there is a significant difference in gene expression between aerobic and anaerobic conditions. To do so, I filtered for p-values smaller than 0.05 and created a new table (genes_filter). The new table also has the information specifying the condition (aerobic or anaerobic) as well as the name of the genes, proteins, and fold change

After that, for each condition (aerobic and anaerobic) I analyzed the observations that had fold change higher or equal to 2.5 and generated two tables. Table 1 and Table 2 present not only fold change values but also the protein that is codified by the genes that are upregulated. The codes below show these operations.

```
# Rename variables
genes = genes %>% rename(parent = X, name_wp = X.1, inference = X.2, protein = X.3,
                         trans_table = X.4)

# Filtering out no values
genes = genes%>% filter(fold.change != "#VALUE!") %>% mutate(fold.change = as.numeric
                                                    (as.character(fold.change)))

# Create new variables
genes = genes %>% mutate(sig = ifelse(padj >= 0.05, "Not significant","Significant"))

#Convert aerobically to same scale as anaerob.
genes = genes %>% mutate (fold_change_scaled =
                            ifelse (fold.change < 1, 1/fold.change, fold.change))
write.csv(genes, file = "gene.csv")

# Filter for conditions
genes_filter = genes%>% filter(sig == "Significant",
                               gene_description != "product=hypothetical protein")
write.csv(genes_filter, file = "genes_filter.csv")
genes_filter = genes_filter %>% mutate(condition = ifelse(fold.change > 1, "anaero", "aero"))
```

Results

Figure 1 represents the distribution of fold change in gene expression of *H. para* aerobically and anaerobically for all genes. The dashed line represents the median. This figure gives a general idea of the distribution of differences in genes expression across conditions, however, it is hard to interpret given the difference in scale. Also, it includes many tests that are not statically significant (note how the distribution is centered around 1, meaning no change). To better visualize the changes in gene expression between anaerobic and aerobic, first I separated the tests that were statistically significant (p-value < 0.05). Figure 2 shows both, significant and not significant fold change distribution in gene expression.

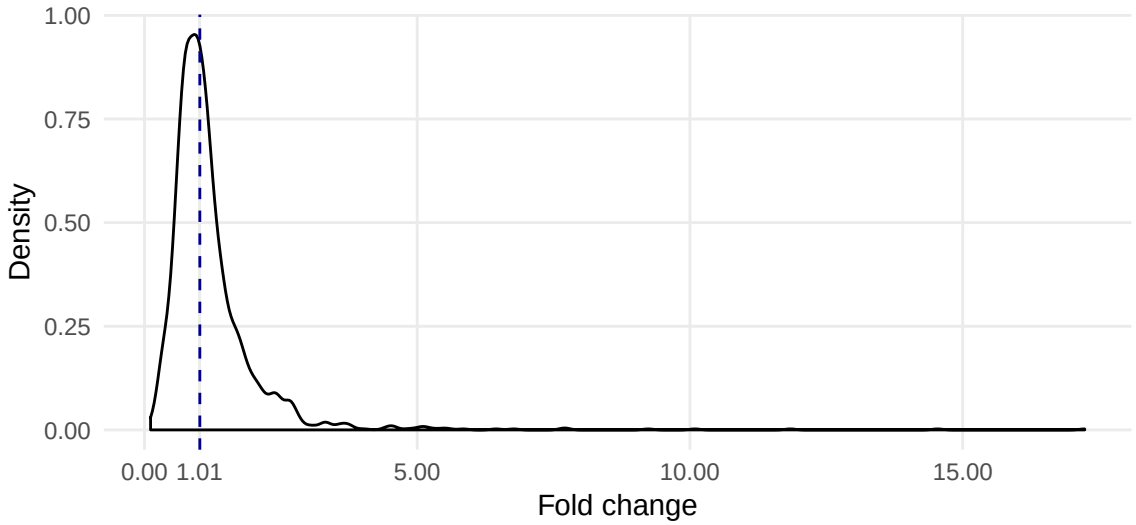

Fig.1: Distribution of fold change in gene expression.

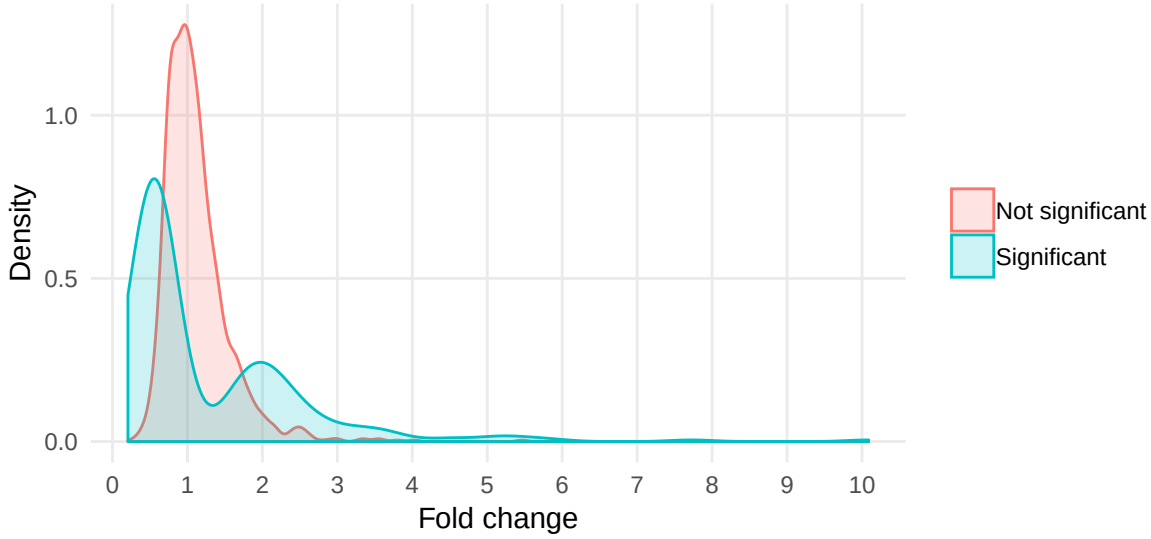

Fig.2: Fold change in gene expresion in *Haemophilus parainfluenzae*

Analyzing Figure 2 it is possible to see that when the "not significant" fold changes are excluded, there are many cases of actual change, in which fold changes are either clearly smaller or larger than one.

The separation between statistically significant and not significant tests around the 0.05 threshold could be misleading if there were many tests in which the p-value was close but above 0.05. Figure 3 shows the distribution of p-values for all tests, providing reassurance that this is not the case. Figure 3 indicates that out of 1764 expressed genes, 275 (15.6%) are statistically significant.

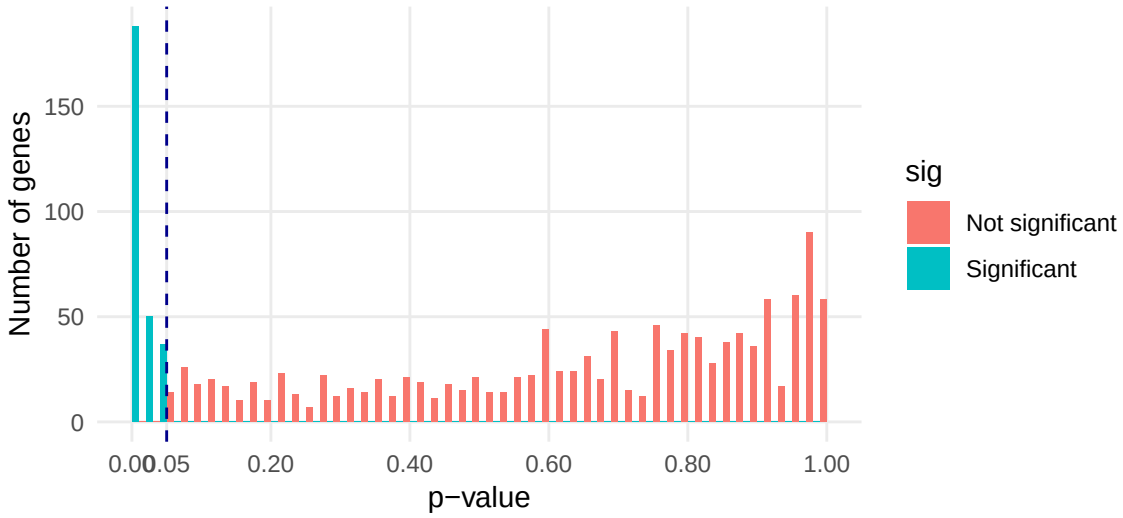

Fig.3: Distribution of p−values over the genes.

The difference between aerobic and anaerobic genes expression is represented in Figure 4. In this plot only significant fold changes ($p < 0.05$) were considered. According to this analysis, it is possible to affirm that under aerobic (aero) conditions, the fold change varies from 1.35 to 5. On the other hand, under anaerobic (anaero) conditions, the fold change is more spread and varies from 1.35 to 10. However, it is important to note that large part of the genes is concentrated around 2-fold change.

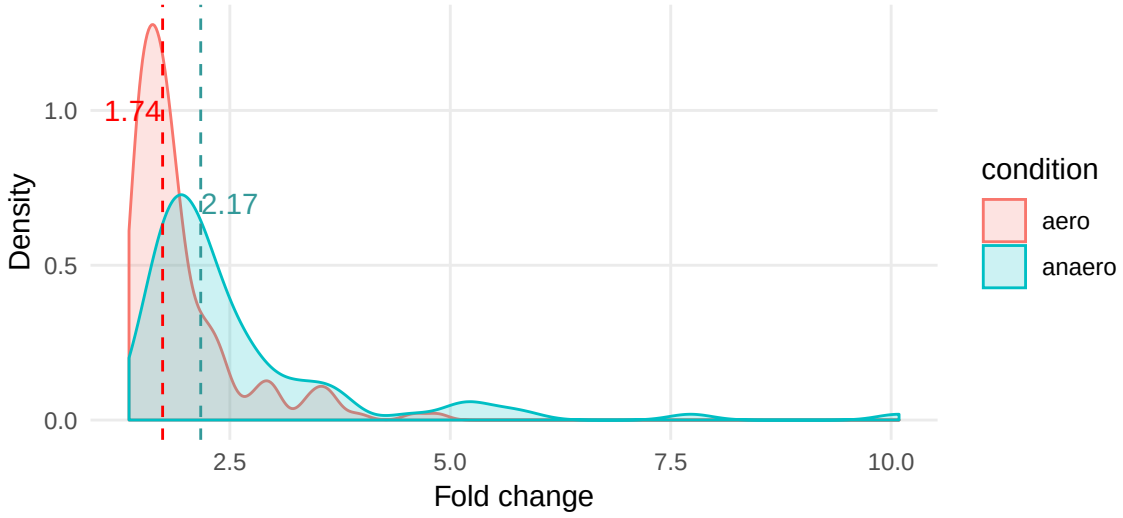

Fig.4: Difference in the expression of genes.

Due to the large number of observations created by the RNAseq analyses, I decided to analyze the cases in which the upregulated genes were the most different across the two conditions. To do so, I filtered the results based on the p-value (< 0.05), and extreme fold change (larger than 2.5).

Table 1: Proteins coded by uperregulated genes and fold change (>= 2.5) in gene expression under aerobic condition

Protein	Fold change
NAD-dependent deacetylase	4.9
ribosomal subunit interface protein	4.6
iron ABC transporter substrate-binding protein	3.9

Protein	Fold change
autonomous glycyl radical cofactor GrcA	3.7
galactose/glucose ABC transporter substrate-binding protein MglB	3.6
rhodanese-like domain-containing protein	3.6
peptidyl-prolyl cis-trans isomerase	3.6
solute:sodium symporter family transporter	3.5
galM	3.4
ATP-dependent chaperone ClpB	3.4
TonB-dependent receptor	3.3
substrate-binding domain-containing protein	3.0
peptide ABC transporter substrate-binding protein	3.0
molecular chaperone DnaK	3.0
carboxymuconolactone decarboxylase family protein	3.0
cytochrome bd oxidase subunit I	2.9
galactose/methyl galactoside ABC transporter permease MglC	2.9
cytochrome d ubiquinol oxidase subunit II	2.8
Bax inhibitor-1/YccA family protein	2.8
cyd operon protein YbgE	2.7
galactokinase	2.5

Table 1 shows proteins coded by upregulated genes under aerobic condition. Here, it is possible to see that 21 genes have fold change higher or equal to 2.5. The gene that presents the highest value codifies for a NAD-dependent deacetylase which is an important regulator in cellular stress response and energy metabolism (Gou et at, 2012). The table also presents genes that are associated with ABC transporter proteins, chaperones, and others. On table 2, 27 genes have expression higher or equal to 2.5. The gene that has the highest value anaerobically codifies a protein associated with glutathione biosynthesis pathway. Other genes codify proteins important for ABC transporter, and cell wall synthesis. The two tables generated in this paper will be used as base for future research.

Table 2: Proteins coded by uperregulated genes and fold change (>= 2.5) in gene expression under anaerobic condition

Protein	Fold change
glutathione synthetase	10.1
MFS transporter	7.7
DNA polymerase IV	5.8
dihydroneopterin aldolase	5.6
iron ABC transporter permease	5.3
monofunctional biosynthetic peptidoglycan transglycosylase	5.1
DNA polymerase I	5.1
molybdopterin-guanine dinucleotide biosynthesis protein B	4.5
septation protein A	3.8
bifunctional tRNA (5-methylaminomethyl-2-thiouridine)	3.7
DNA-formamidopyrimidine glycosylase	3.7
beta-phosphoglucomutase family hydrolase	3.6
phosphoribosylformylglycinamidine synthase	3.5
30S ribosome-binding factor RbfA	3.3
tRNA pseudouridine(55) synthase TruB	3.3
sulfurtransferase TusD	3.3
tRNA (uridine(54)-C5)-methyltransferase TrmA	3.1
YfcC family protein	3.1
AmpG family muropeptide MFS transporter	2.8

Protein	Fold change
bifunctional chorismate mutase/prephenate dehydrogenase	2.8
YbaK/prolyl-tRNA synthetase associated domain-containing protein	2.8
DedA family protein	2.8
3-deoxy-D-manno-octulosonic acid transferase	2.7
HTH domain-containing protein	2.7
MBL fold metallo-hydrolase	2.6
ribosome maturation factor RimP	2.6
FAD-binding oxidoreductase	2.5

Conclusion

This paper presents results of initial analyzes of RNAseq data of *H. para* aerobically and anaerobically. It is possible to observe that 258 genes are upregulated in those conditions. Large part of the genes (170) are upregulated aerobically while 88 are upregulated anaerobically. Also, 48 genes are 2.5 times more expressed either aerobically or anaerobically. Future directions include more detailed analyzes of each of those 48 genes and, for some of them, construction of mutants that will be tested in specific conditions such as aerobically, anaerobically, and in co-cultures with other species that colonize the oral cavity to investigate the interaction between them.

References

Guo, X. Kesimer, M. Tolun, G. Zheng, X, Xu,Q. Lu,J. Sheehan, J.K. Griffith, J.D. Li,X. (2012). The NAD+-dependent protein deacetylase activity of SIRT1 is regulated by its oligomeric status. Scientific Reports volume 2, Article number: 640.

Park, O.J., Yi, H., Jeon, J. H., Kang, S.S., Koo, K.T., Kum, K.Y., . Han, S. H. (2015). Pyrosequencing Analysis of Subgingival Microbiota in Distinct Periodontal Conditions. Journal of Dental Research, 94(7), 921-927. https://doi.org/10.1177/0022034515583531.

Posse, J.L. Dios, P.D. Scully,C. *Saliva Protection and Transmissible Diseases.* Academic Press, 2018. Print.