**Paulo Tanicala**

**RNA-seq Project**

**Understanding the Data**

*Trichomonas tenax (T.tenax)* is a flagellated protozoan parasite that belongs to the Trichomonadida class. It typically inhabits the oral cavity and respiratory tract of humans with previous studies suggesting that it can cause cell damage and phagocytose host epithelial cells. *T.tenax* on its own typically does not cause symptomatic diseases. It can, however, potentially contribute to the development of oral and respiratory infections (Hong et al., n.d.). In order to better understand the pathogenic mechanism of its interaction with respiratory infections, the investigators did a comparative gene expression profiling analysis of RNA-seq from *T.tenax* treated and untreated lung epithelial cells.

The investigators utilized NCI-H292 lung epithelial cells and compared the expression profiles of samples treated with *T.tenax* and untreated samples (NCBI GEO Accession Number: GSE228216).The RNA-seq data was obtained using standard RNA-seq protocols (Berge et al. 2019). RNA extraction and purification was done using Direct-zolTM RNA MiniPrep (AYMO RESEARCH) which yielded 1 ug of total RNA that was used for sequencing library construction. Enrichment of Target RNAs was done using KAPA mRNA HyperPrep Kit, which utilizes poly(A) capture method to isolate mRNA from ribosomal and transfer RNAs. The kit also includes necessary reagents for complementary DNA (cDNA) synthesis, library construction, and amplification. Sequencing of the cDNA libraries were done using Illumina’s NovaSeq 6000 instrument. The resulting sequencing data can be found in the NCBI website using the following SRA study ID: [SRP429347](https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP429347&o=acc_s%3Aa). The data used for this project was obtained from the NCBI FTP site, using the SRR ID runs at a total of 6 runs (NCBI GEO Accession Number: GSE228216)

This analysis project seeks to determine which genes are over expressed in cells treated with *T.tenax* and what are the potential consequences of the expression profile.

Although any insights from this analysis could prove to be valuable, its application in real world scenarios could vary due to the differences in immune responses of individuals. There is also the possibility that there is no significant difference in the gene expression profiles of the treated cells vs. the untreated cells which would suggest that the interaction between *T.tenax* and oral/respiratory infections might not have a genetics related component.

Understanding which genes are the most affected can be useful for the development of targeted therapeutics or could warrant a more focused study. A potential study for an identified upregulated gene would be inhibitions studies and its effects on infection development. On the other hand, any downregulated genes can be upregulated and observed for any interactions with infection development. In either case, insights from such studies have the potential to be targets for gene based or small molecule therapeutics.

A diagram of a diagram of a company

Description automatically generated with medium confidence

**Figure 1**. This shows the general abstract of the analysis project starting from data processing and quality control and ending in data analysis and visualization. An experiment, library preparation, and NGS abstract was included to provide context and background to the analysis project.

**Data Processing and Analysis**

FastQC was performed to assess the overall quality of the sequencing data obtained. It was determined that the sequencing adapters needed to be trimmed before proper alignment could be done. Initially, a base count of 15 was trimmed, however alignment yielded less than ideal results. Average successfully assigned elements after alignment were shown to be at less than 1%. To account for this, the base count of 200 was trimmed to achieve better alignment and quicker processing. Figure 2 shows the per base sequences before and after trimming.

A graph showing a graph of a graph

Description automatically generated with medium confidenceA graph with lines in different colors

Description automatically generated with medium confidence

**Figure 2.** Per base sequence content graphs from FastQC illustrating the distribution of nucleotides at specific read position. The graph on the left shows the distribution before trimming with Cutadapt. The graph on the right shows the distribution after trimming the first 200 bases.

**Table 1**. Summary of counting results from featureCounts.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Treatment | Successfully assigned alignments | Proportion of assigned alignments |
| SRR23972383 | Cocultured | 3,355,176 | 58.3% |
| SRR23972384 | Cocultured | 3,337,609 | 57.2% |
| SRR23972385 | Cocultured | 3,353,226 | 55.7% |
| SRR23972386 | Untreated | 3,333,710 | 54.4% |
| SRR23972387 | Untreated | 3,344,595 | 53.9% |
| SRR23972388 | Untreated | 3,331,932 | 55.5% |

The trimmed sequences were then aligned to the GRCh38 reference sequence using Hisat2. After alignment, a counts table was generated using featureCounts function from the Subread package. A summary of the counting results is shown in Table 1. The counts table was then analyzed in RStudio with a metadata corresponding to the experiment.

Gene expression profile of samples cocultured with *T.tenax* was obtained using DESeq2. Figure 3 shows the mean of normalized counts against the log2 fold changes of the cocultured samples.

A graph of a normalized count

Description automatically generated

**Figure 3.** Mean of normalized counts after Deseq2 analysis showing the presence of differentially expressed genes.

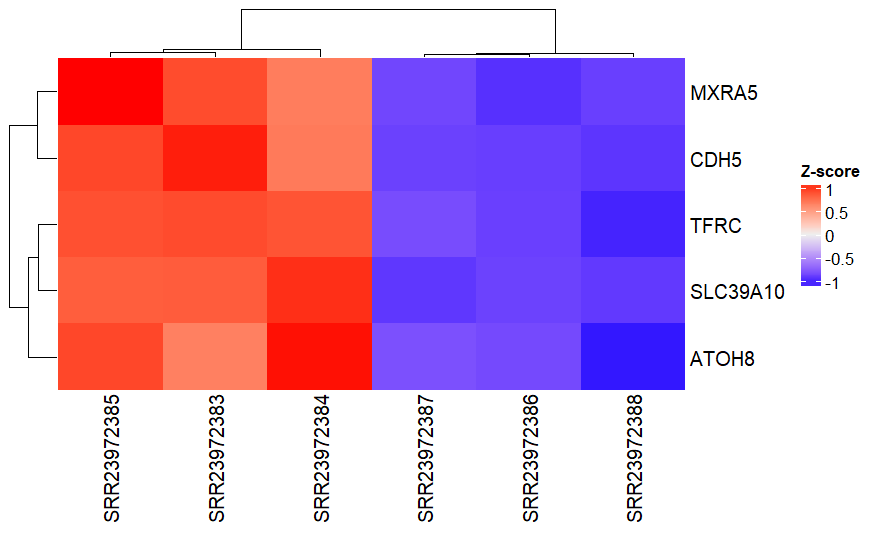
The resulting DESeq data was filtered using a significance threshold of 0.05. Annotations were done using the Genome wide annotation for Human package. Overexpressed and under expressed genes were visualized through a volcano plot shown in Figure 4.

A graph with numbers and letters

Description automatically generated with medium confidence

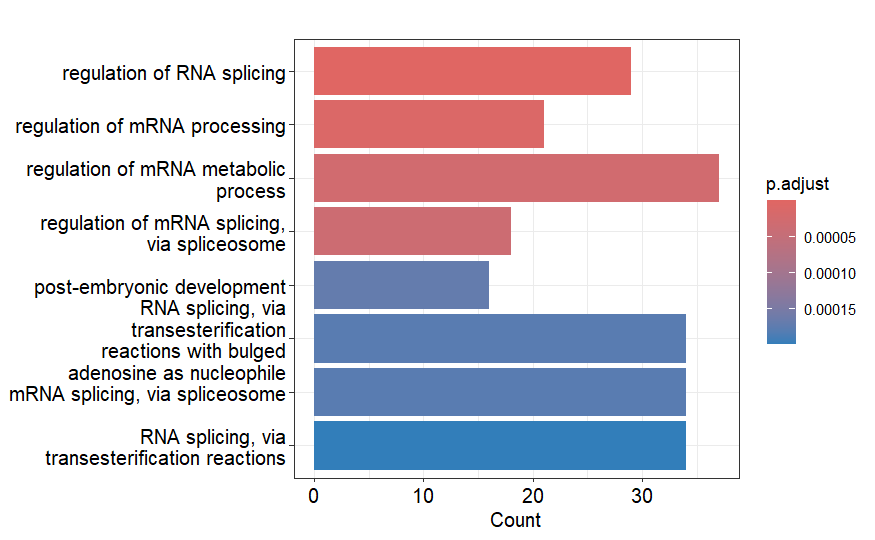
**Figure 4.** Volcano plot of genes in T.tenax cocultured samples. Data points labeled are the significant genes. Genes on the right side of the figure are overexpressed while the genes on the left are under expressed.

Comparisons of gene expression profiles between the treated and untreated groups were done using the ComplexHeatmap package. The resulting expression profile comparison is shown in Figure 5.



**Figure 5.** Top 5 genes overexpressed in the T.tenax treated samples ("SRR23972383", "SRR23972384", "SRR23972385"). Consequently, these genes are under expressed in the untreated groups.

Genome ontology enrichment was also performed to illustrate which biological processes correspond to the overrepresented genes. The results are illustrated in Figure 6 as a bar plot and a Manhattan plot in Figure 7.



**Figure 6**. Overrepresented biological functions from the gene expression profile obtained through DESeq2. RNA processing is overrepresented in the expression profile of samples cocultured with T.tenax.

A colorful dots on a white background

Description automatically generated

**Figure 7**. Manhattan plot generated using the gene ontology enrichment analysis. Of note are the yellow bubbles indicating that the expression profile heavily influences biological processes.

Kegg pathways were then generated using GO datasets, Kegg pathway datasets, and the Generally Applicable Gene-set Enrichment for Pathway Analysis package. The pathways use the expression profiles obtained from DESeq2 analysis and determines their consequences on biological processes using the GO datasets and the Kegg pathway datasets. One of the top upregulated pathways is illustrated in Figure 8.

A diagram of a cell

Description automatically generated

**Figure 8.** One of the most upregulated pathways that is supported by gene expression profile and distribution of overrepresented genes. Gene products in red are upregulated while gene products in green are downregulated. Gene products in gray have no significant difference in expression levels compared to untreated samples.

**Interpretation and Conclusion**

Differential gene expression analysis through DESeq2 identified *MXRA5, CDH5, TFRC, SLC39A10,* and *ATOH8* as upregulated genes in samples treated with *T.tenax.* Four out of the 5 genes are shown in Figure 4 and all the genes are shown in Figure 6. The analysis determined that biological processes are overrepresented as shown by the large amount of “yellow bubbles” in Figure 7. Consequently, the specific biological processes affected are the RNA splicing and mRNA processing functions. The spliceosome pathway illustrated in Figure 8 shows the specific gene products that are upregulated and downregulated indicating that this specific pathway is highly affected by the presence of *T.tenax*. The implication is that this parasite worsens pre-existing periodontal disease and infections by affecting the mRNA processing and slicing mechanisms in lung epithelial cells.

**References**

Berge, Koen Van den, Katharina M. Hembach, Charlotte Soneson, Simone Tiberi, Lieven Clement, Michael I. Love, Rob Patro, and Mark D. Robinson. 2019. “RNA Sequencing Data: Hitchhiker’s Guide to Expression Analysis.” *Annual Review of Biomedical Data Science* 2 (Volume 2, 2019): 139–73. <https://doi.org/10.1146/annurev-biodatasci-072018-021255>.

“GEO Accession Viewer.” n.d.-a. Accessed April 20, 2024. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE228216>.

Hong, Zih-Bin, Yu-Ting Lai, Chun-Hsien Chen, Yi-Jen Chen, Chien-Chin Chen, and Wei-Chen Lin. n.d. “Trichomonas Tenax Induces Barrier Defects and Modulates the Inflammatory Cytotoxicity of Gingival and Pulmonary Epithelial Cells.” *Parasite* 30: 7. <https://doi.org/10.1051/parasite/2023010>.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. “fastp: an ultra-fast all-in-one FASTQ preprocessor.” Bioinformatics 34 (17): i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>.

Soneson C (2024). Rhisat2: R Wrapper for HISAT2 Aligner. R package version 1.20.0, <https://github.com/fmicompbio/Rhisat2>.

Liao Y, Smyth GK, Shi W (2019). “The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads.” Nucleic Acids Research, **47**, e47. [doi:10.1093/nar/gkz114](https://doi.org/10.1093/nar/gkz114).

Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, **15**, 550. [doi:10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).

Blighe K, Rana S, Lewis M (2024). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version 1.22.0, <https://github.com/kevinblighe/EnhancedVolcano>.

Gu Z, Eils R, Schlesner M (2016). “Complex heatmaps reveal patterns and correlations in multidimensional genomic data.” Bioinformatics. [doi:10.1093/bioinformatics/btw313](https://doi.org/10.1093/bioinformatics/btw313).

Gu Z (2022). “Complex Heatmap Visualization.” iMeta. [doi:10.1002/imt2.43](https://doi.org/10.1002/imt2.43).

Carlson M (2019). org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2.

Luo, Weijun, Brouwer, Cory (2013). “Pathview: an R/Bioconductor package for pathway-based data integration and visualization.” Bioinformatics, **29**(14), 1830-1831. [doi:10.1093/bioinformatics/btt285](https://doi.org/10.1093/bioinformatics/btt285).

Luo, Weijun, Friedman, Michael, Shedden, Kerby, Hankenson, Kurt, Woolf, Peter (2009). “GAGE: generally applicable gene set enrichment for pathway analysis.” BMC Bioinformatics, **10**, 161.

Luo, Weijun, Brouwer, Cory (2013). “Pathview: an R/Bioconductor package for pathway-based data integration and visualization.” Bioinformatics, **29**(14), 1830-1831. [doi:10.1093/bioinformatics/btt285](https://doi.org/10.1093/bioinformatics/btt285).

Luo W (2024). gageData: Auxillary data for gage package. R package version 2.42.0.