Transcriptomic Analysis of *Pseudomonas putida* in Varied Growth Conditions

A Major Qualifying Project Report

Submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

in Bioinformatics and

Computational Biology by

Gabriella Guzman Jerry

CDR Deadline: April 25, 2024

Abstract

This project aimed at understanding the variability in gene expression of *Pseudomonas putida* between laboratory and simulated liquid soil conditions. *P. putida* was grown in LB media and solubilized extract of soil organic material (SESOM) with either log phase or stationary growth. RNA-Seq analysis was employed to understand gene expression changes under the different growth conditions. The analysis revealed that most variation occurs between changes in medium, suggesting that growth medium components strongly influence gene expression patterns. Ultimately, we aim to unravel the genetic adaptations crucial for understanding *P. putida's* survival in the soil environment, advancing scientific understanding and practical applications for engineered organisms in the soil.

Acknowledgements

I want to acknowledge and thank Professor Natalie Farny for graciously welcoming me into her laboratory for my MQP. Her swift acceptance and immediate integration of me into her team are gestures for which I am immensely grateful. Despite her demanding schedule, Professor Farny generously dedicated time each week to meet with me, providing invaluable guidance, feedback, and mentorship throughout the entirety of my project.

I also would like to acknowledge all members of the Farny Lab for their support and feedback, particularly those involved in the soil project alongside me. Their diligent efforts in collecting *Pseudomonas putida* samples across various growth conditions and meticulously preparing the datasets were instrumental in facilitating my RNA-Seq analysis. It is thanks to their contributions that I was able to leverage my bioinformatic tools effectively. The collaborative environment fostered by the Farny Lab has been instrumental in my academic and research growth, and I am deeply grateful for their support and camaraderie throughout this endeavor.

This section would be incomplete without acknowledging someone who has been instrumental in making my MQP experience enjoyable and significantly influencing the outcome of this project. I extend my heartfelt gratitude to Gabrielle Cabebe for her unwavering dedication and countless hours spent assisting me in navigating the complexities of the RNA-Seq pipeline and comprehending the bioinformatic tools essential for this project. Gabby generously devoted her time, meeting with me twice a week without hesitation, and her steadfast support throughout the entire school year has been invaluable. I cannot overstate the profound impact her guidance has had on this project; without her expertise, it would have taken a drastically different shape. I am immensely grateful for her contributions, kindness, and overwhelming support.

Table of Contents

List of Figures

1. Introduction

1.1 SynBio Organisms and GEMs

Synthetic biology is a dynamic and multidisciplinary field that encompasses the deliberate modification of genetic material within a diverse range of organisms, spanning from viruses and bacteria to plants and animals (Benner, 2005). This innovative technology, which has evolved over several decades, exerts a profound influence across various sectors of life, including medicine, agriculture, manufacturing, and environmental conservation (Cameron, 2014). The exceptional versatility of synthetic biology, combined with the ongoing advances in DNA sequencing, genome editing, and artificial intelligence, positions it as a pivotal driver of innovation, heralding sustainable solutions and playing an instrumental role in shaping the future of biotechnology (Cameron, 2014).

Genetically engineered microbes (GEMs) are central figures in the realm of synthetic biology due to their remarkable attributes, including short generation times, adaptability, and predictable behavior (Pant, 2020). GEMs have made invaluable contributions to medicine, yielding breakthroughs such as the production of therapeutic proteins and life-saving vaccines (Liu, 2022). Furthermore, GEMs have increasingly become the focus of attention in recent years, as they demonstrate significant potential in an array of soil-related applications, encompassing biosensing to detect environmental contaminants, bioremediation to mitigate the impact of pollutants, and pathogen control to safeguard public health (Rock, 2021). These applications underscore the pivotal role of GEMs as versatile tools for addressing complex challenges, exemplifying their significance in the realm of synthetic biology and beyond.

Despite the growing utilization of genetically engineered microbes, substantial challenges persist in their application. The regulatory landscape is difficult to control but pivotal in the progress of these microbes (Ezezika, 2010). For instance, a microbe engineered to degrade the toxic defoliant Agent Orange faced approval challenges, as researchers struggled to guarantee that its capacity to break down harmful chemicals would not be transferred to nearby pathogenic bacteria (Ezezika, 2010).

An additional challenge with GEMs is that they may thrive in one environment while exhibiting limitations or unforeseen behaviors in another, underscoring the complexity of biological systems (Stirling, 2020). A GEM may exhibit anticipated behavior in controlled laboratory conditions, but its performance in a natural environment, such as a soil microbiome, can deviate from expectations (Stirling, 2020). This deviation could affect the GEM's ability to execute its intended functions and potentially result in undesired consequences, including horizontal gene transfer and disruption of the natural microbiome (Rock, 2021). These variations in performance raise the need for further research and refinement to ensure the predictability and reliability of GEMs across a range of conditions.

1.2 Our Model GEM Strain

Figure 1: SEM photograph of *Pseudomonas Putida*. Adapted from LIBRARY, D. K. M. P. (n.d.) Retrieved April 5, 2024.

1.2.1 Pseudomonas Putida and SynBio

Pseudomonas putida (*P. putida)*, a member of the extensive *Pseudomonas* genus, stands out as a Gram-negative bacterium with remarkable ubiquity across diverse ecosystems (Volke et al., 2020). *P. putida* is frequently encountered among plants, waters, and within soils, particularly those contaminated with various pollutants (Volke et al., 2020). Although *P. putida* is often found in environments that would be challenging for many microbes, such as oxidative stress and exposure to toxic substances, the bacterium showcases its remarkable adaptability. *P. putida* has been found to transform industrial byproducts such as plastics, oils, food waste, and agricultural residues into valuable products. (Son et al., 2023). Notably, this bacterium also exhibits undemanding nutritional requirements, ensuring its survival and proliferation, and is easy to culture (Weimer et al., 2020). This ability has made *P. putida* of high interest for issues such as waste management or climate change (Son et al., 2023).

The journey of exploring the potential of *P. putida* has its roots in the 1960s, and over the past six decades, our understanding of this microbe has grown significantly (Nakazawa, 2002). Despite these advancements, numerous aspects of *P. putida'*s biology remain vaguely understood. While many attribute the bacterium's impressive stress tolerance to its cyclic-core metabolism and Gram-negative characteristics, the precise underlying properties enabling this resilience are yet to be fully unveiled (Weimer et al., 2020). A prominent challenge in the realm of metabolic engineering and synthetic biology with *P. putida* is unraveling the intricate metabolic and regulatory networks governing this bacterium (Weimer et al., 2020). However, the known attributes of *P. putida*, such as its adaptability and robustness, have unequivocally established it as a top-tier candidate for genetic engineering and microbiological research (Volke et al., 2020).

1.3 The *P. putida* Project

1.3.1 Project History

The presence of undiscovered landmines presents a significant and perilous risk, as they are difficult to detect due to most being buried or camouflage (Garmendia, 2008). To address this pressing issue, members of the Farny lab at WPI have embarked on a project to detect these buried landmines using a long-distance bacterial signaling system with *P. putida*. The hopes were that this engineered strain of *P. putida* would have the capability to traverse the soil, forming associations with fungal hyphae, and initiate a response upon detecting the presence of TNT buried as deep as a meter underground (Garmendia, 2008). Upon detection, the signaling cascade within *P. putida* would trigger the emission of a fluorescent signal on the surface, serving as a visible indicator of potential landmine locations (Garmendia, 2008). However, it was found that the technology has demonstrated signal propagation over only 8 cm.

While this innovative approach holds immense promise for landmine detection, it has encountered a challenge regarding the variability in *P. putida*'s gene expression patterns between laboratory conditions and its natural soil habitat for the Farny Lab. To determine the cause and levels of gene expression change in *P. putida* the team employed a liquid soil extract known as Solubilized Extract of Soil Organic Material (SESOM) to faithfully replicate the metabolic conditions of the soil (Brözel, 2009). The team also generated comprehensive transcriptomic datasets from *P. putida* cultivated in both exponential and stationary phases, comparing growth in LB medium to that in SESOM (Voltmer, 2023).

1.3.2 Next Generation Sequencing

High-throughput, next-generation sequencing (NGS) technologies have revolutionized genome research by generating vast amounts of sequence data quickly (Xu, 2014). The versatility and potency of NGS make it an essential tool across a wide range of biological sciences, transforming scientific research in numerous fields (Xu, 2014). RNA-seq analysis is an essential part of next generation sequencing.

Introduced in 2008, RNA sequencing (RNA-seq) has gained widespread adoption over the past decade due to decreasing costs and the proliferation of shared-resource sequencing cores in research institutions (Koch, 2018). RNA-seq serves various downstream analysis objectives, some being, transcript discovery, genome annotation, exploration of gene regulation mechanisms, and differential gene expression, which will be the focus of this project (Griffith, 2015). The source material for such analysis may include *in vitro* cultured cells, homogenized whole tissues, or sorted cell populations (Koch, 2018).

The growing popularity of RNA-seq has created an increasing demand for bioinformatics expertise and computational resources (Griffith, 2015). To accurately analyze and process large datasets, bench scientists must grasp the bioinformatics principles and limitations associated with the intricate RNA-seq analysis process (Koch, 2018). While RNA-seq analysis can unveil valuable insights, it presents a departure from conventional analyses, providing a massive dataset that requires extensive examination for meaningful interpretation (Koch, 2018).

1.3.3 Project Goals and Objectives

To gain insights into the impact of growth environment on gene expression, this project conducted comprehensive RNA-seq analyses of *P. putida* datasets cultivated in LB and SESOM media, encompassing both log phase and stationary phase growth. The primary objective was to discern patterns in gene expression dynamics, determining the genetic adaptations crucial for *P. putida's* resilience and efficacy within diverse soil environments. In addition to differential RNA-seq analysis, the most significantly upregulated and downregulated genes were compared to identify prevailing trends. Furthermore, Gene Ontology analysis was employed to investigate common pathways governing the metabolic shift necessary for survival in soil ecosystems. By examining gene expression variations, this study may not only advance fundamental scientific understanding but also may pave the way for practical applications. Specifically, the insights gained can inform the development of more robust GEMs optimized for gene expression in soil conditions. Such models hold promise for enhancing our ability to predict and engineer microbial behavior in soil environments, contributing to sustainable environmental management.

2. Methodology

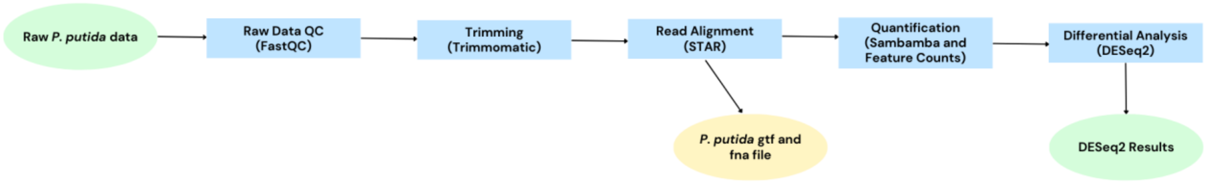

Raw *P. putida* data
Raw Data QC (FastQC)
Trimming (Trimmomatic)
Read Alignment (STAR)
Quantification (Sambamba and Feature Counts)
Differential Analysis (DESeq2)
P. putida gtf and fna file
DESeq2 Results

Figure 2: Flowchart of the RNA-Seq Analysis Pipeline used for this project.

2.1 Raw Sequence Data

2.1.1 P. putida Sample Collection

P. putida cultures were cultivated in two distinct media, LB and SESOM. Subsequently, bacterial lysis was employed to extract total RNA from the samples. Following RNA extraction, purification steps were implemented to eliminate genomic DNA (gDNA). PCR (Polymerase Chain Reaction) was then applied to amplify specific RNA regions of interest. The amplified RNA samples were subjected to gel electrophoresis, a technique used to separate and visualize nucleic acids based on size, allowing for the assessment of RNA integrity and purity. Additionally, polyadenylation, the process of adding poly(A) tails to RNA molecules, was carried out to enhance the stability and quality of the RNA samples (Voltmer, 2023).

2.1.2 P. putida Data Sets and Shell Scripts

The RNA-Seq analysis essential for this project relied on the Illumina *P. putida* datasets meticulously collected by Stokley Voltmer, a member of the Farny lab and the soil project. Stokley described that library preparations and sequencing were conducted by Azenta Life Sciences on the HiSeq platform, resulting in the generation of fastq files. FastQC was subsequently applied to assess the quality of these files. The output from FastQC informed the use of Cutadapt to clean up the fastq files, resulting in quality-controlled reads (Voltmer, 2023).

The RNA-Seq pipeline for this project employed High-Performance Computing resources at WPI, utilizing batch processing through Slurm (Griffith et al., 2015). Specifically, the Linux cluster within Slurm facilitated computational tasks (Linux Clustering | SIOS, n.d.). Utilizing Slurm as the batch manager enabled the creation of tailored shell scripts for each pipeline step. These scripts facilitated efficient batch job submissions and encapsulated essential input parameters for each tool.

2.2 RawDataQC

Before proceeding with downstream analysis, the quality of raw sequencing data was assessed using FastQC, a widely used tool for evaluating the quality of high-throughput sequencing data. FastQC provides detailed information on various quality metrics, including per-base sequence quality scores, sequence length distribution, GC content, and presence of sequencing adapters or contaminants (FastQC, n.d.).

The input files used for FastQC were the fastq.gz files created from the *P. putida* sample collection mentioned previously in section 2.1.2. The shell script ran for FastQC contained variables for the run path and results path, where run path contained the fastq files and results path was the directory where the FastQC results would go. A for loop was created to run through each file in the run path and run FastQC on each.

Upon completion of the FastQC analysis, comprehensive reports were generated for each sample some being the per base sequence quality, per sequence quality scores, quality score distribution, and sequence length distribution. These reports served as valuable insights for subsequent data preprocessing steps.

2.3 Trimming

Raw sequencing reads often contain low-quality bases and adapter sequences that can affect downstream analysis. Therefore, trimming of raw reads is important in ensuring that read mapping and assembly is accurate. Trimming was performed using Trimmomatic, a tool designed to remove sequencing adapters and trim low-quality bases from reads (Bolger et al., 2014).

Trimming was conducted on each pair of input FASTQ files within the specified directory. An adapter file called TruSeq3-PE.fa was used to notify Trimmomatic on what adapter sequences should be removed from the files. The Trimmomatic parameters were set as follows: a sliding window size of 4 bases with an average quality threshold of 15, ensuring the removal of low-quality bases from the ends of reads. Additionally, a minimum quality score of 3 was applied for both the beginning and end of the reads. Any bases with a quality score below this threshold were trimmed. The minimum length of the trimmed reads was set to 36 bases to ensure data quality.

The script facilitated automated processing of paired-end RNA-Seq data, providing an efficient and reproducible method for quality control and preprocessing. This approach ensured the removal of low-quality bases and adapter sequences, thereby enhancing the accuracy and reliability of downstream analyses, such as alignment and differential expression analysis.

2.4 Read Alignment

2.4.1 STAR Indexing

Effective alignment is crucial for accurately mapping sequencing reads to a reference genome. To achieve this, a specific indexing process called STAR Indexing was employed (Piper, 2017). This index was generated utilizing both the genomic FASTA (fna) file and the gene annotation file (gtf) of the KT2440 strain of *P. putida*, ensuring alignment compatibility.

The indexing process was initiated with the STAR command, specifying the **--**runMode genomeGenerate option to indicate the genome indexing mode. The script required the genome FASTA file and the gene annotation file, along with additional parameters such as **--**sjdbOverhang to specify the length of spliced alignments and **--**runThreadN to allocate computational threads. Specifically, the **--s**jdbOverhang parameter was set to 49, corresponding to the maximum spliced alignment overhang. Finally, the --quantMode GeneCounts option was used to enable the generation of gene-level expression counts during the indexing process.

This comprehensive indexing procedure facilitated the creation of a STAR index tailored to the reference genome and gene annotation, ensuring accurate alignment and quantification of RNA-Seq data in subsequent analyses. The resulting index files were stored in the designated directory to use in downstream transcriptome analysis pipelines.

2.4.2 STAR Alignment

Alignment of trimmed reads to the reference genome of *P. putida* was conducted using the STAR aligner (v2.7.9a). The script locates the trimmed FASTQ files in a specified directory and initiates a loop to process each file individually. For each iteration, the STAR aligner was invoked with specific parameters for alignment. The --genomeDir option specified the directory containing the pre-built STAR index for the reference genome. The --readFilesIn option provided the path to the current trimmed FASTQ file being processed.

Additionally, various other options were set to configure the alignment process, including specifying the output file prefix, setting the maximum number of multiple alignments per read to one (--outFilterMultimapNmax), defining the output format (--outSAMtype), and setting the threading options to two (--runThreadN) (Piper, 2017). During the execution of the loop, a counter variable was incremented for each iteration, ensuring unique identifiers for the output files. This facilitated the organization and identification of alignment results for each sample.

The alignment process generated aligned read files in the Sequence Alignment/Map (SAM) format, which were subsequently converted to the Binary Alignment/Map (BAM) format for further analysis (Freeman, 2016). Overall, this script enabled the efficient alignment of trimmed RNA-Seq reads to the reference genome using the STAR aligner, producing aligned BAM files ready for downstream analysis and interpretation.

2.5 Quantification

2.5.1 Sambamba

Quantification of gene expression levels was performed using Sambamba. Sambamba is utilized to filter aligned reads based on their mapping quality scores, enabling the removal of low-quality reads and improving the accuracy of downstream analyses in the RNA-seq pipeline (Tarasov et al., 2015). To execute the MAPQreads filtering step as part of the RNA-seq pipeline, the directory containing the aligned BAM files generated from the previous STAR alignment was identified. Subsequently, and output directory was created to store the filtered BAM files, each preserving reads with MAPQ scores surpassing a predefined threshold of 20. The script iterated over each BAM file, applying the built-in filtering criteria of the sambamba tool, effectively removing reads with suboptimal mapping quality. This process optimizes the quality of the RNA-seq data by retaining only the most reliable alignments, thereby facilitating more accurate gene expression quantification and enhancing the overall robustness of downstream analyses.

2.5.2 FeatureCounts

FeatureCounts was employed to count the number of reads mapping to each gene in the reference genome. This quantification step provided a measure of gene expression abundance, which served as input for the subsequent differential gene expression analysis. The program featureCounts version 2.0.1 was utilized for read summarization, a crucial process in RNA-seq data analysis for quantifying gene expression levels (Liao et al., 2014). The specific command used was "featureCounts", where the "-a" flag specified the input annotation file containing genomic feature information, in this case, the KT2440_110_modified.gtf file. The "-o" flag indicated the output file name, designated as "counts.txt". By executing this command, featureCounts processed the aligned reads and assigned them to genomic features, ultimately generating a summary file containing the counts of reads mapped to each feature.

2.6 Differential Gene Expression

2.6.1 Calculating Differential Gene Expression

Differential gene expression analysis was conducted using the DeSeq2 package in the R programming environment. DeSeq2 is used for detecting differentially expressed genes between different experimental conditions, considering factors such as biological variability and sequencing depth (Piper, 2017). Input data for DeSeq2 analysis included the raw read counts obtained from FeatureCounts and experimental metadata specifying the experimental conditions for each sample (e.g., LB vs. SESOM medium, stationary vs. log growth phase). Each DGE tool has their own methods for modeling read counts and estimating fold change values. DESeq2 uses sample-wise size factor for nomarlization, Cox-Reid approximate conditional inference with a focus of maximum individual dispersion estimate, negative binomial for the assumed distribution, low false positives, support for multifaceted experiments, no detection of differential isoforms, and a runtime of 3-5 replicates (Dundar, 2015).

First, a design formula was defined to model the experimental conditions, where each condition is represented as a separate group. Next, pairwise comparisons of interest were specified using contrast vectors, enabling comparison between different sample groups. Subsequently, a DESeqDataSet object was created from the count matrix (readcounts) and sample information, adhering to the design formula specified earlier. The DESeq analysis was then performed using the DESeq function, which estimated size factors and dispersions and fitted a negative binomial model to the count data. The logfold threshold for the DESeq2 analysis was 1 where upregulated genes were larger than 1 and downregulated genes were less than -1. The DESeq2 analysis would consider any data that had a p-value of less than 0.05.

Following the DESeq analysis, differential expression contrasts were conducted for each pairwise comparison of interest using the results function. This step allowed for the identification of genes that were differentially expressed between the specified sample groups. Finally, summary statistics were generated for each contrast, providing insights into the magnitude and significance of differential expression for the respective comparisons. Overall, this workflow in DESeq2 enables comprehensive differential expression analysis of RNA-seq data, facilitating the identification of genes associated with specific experimental conditions or biological phenomena.

2.6.2 Visualizations

2.6.2.1 Venn Diagrams

To visualize the transcriptional responses of Pseudomonas putida to different growth conditions, significant genes were identified from the RNA-seq comparisons based on adjusted p-values and log fold changes. Upregulated genes, defined as those with a padj < 0.05 and log2FoldChange > 1.0, were extracted from each comparison. Similarly, downregulated genes, meeting the criteria of padj < 0.05 and log2FoldChange < -1.0, were identified for the same comparisons. Venn diagrams were then generated using the R package VennDiagram to visualize the overlap of significant genes across the different growth conditions. Separate diagrams were created for upregulated and downregulated genes, each illustrating the shared and unique gene sets among the pairwise comparisons. The category names in the Venn diagrams corresponded to the specific growth conditions being compared.

2.6.2.2 Volcano Plots

For each pairwise comparison of conditions defined in the contrasts, contrast analysis was performed using the DESeq2 package in R. This analysis generated contrast-specific results, capturing differential gene expression between conditions. Subsequently, thresholds for significance were established, with a false discovery rate (padj) threshold set at 0.05 and a log2 fold change (log2FC) threshold set at 1. Genes meeting these criteria were categorized as exhibiting high expression, while others were categorized as showing low expression. Volcano plots, visual representations of differential expression, were then generated using the ggplot2 package. These plots depict log2 fold changes on the x-axis and the negative logarithm of adjusted p-values on the y-axis, with significant genes highlighted in color. Dashed lines indicate the significance thresholds for padj and log2FC. Each volcano plot was saved as a PNG file, with the filename corresponding to the specific contrast being analyzed. This approach allowed for the comprehensive visualization of differential gene expression patterns across the various conditions under investigation.

2.6.2.3 Heatmaps

The first heatmap focuses on visualizing the expression levels of the top 20 highly expressed genes across different experimental conditions. These genes were identified by ordering them based on their mean expression levels across all samples, with the highest expressing genes selected for visualization. The heatmap was generated using the pheatmap package in R, which allowed for the creation of a graphical representation where rows represent genes and columns represent individual samples. To maintain the specific order of genes and samples, clustering of both rows and columns was disabled. Additionally, sample annotations such as experimental conditions and size factors were included as column annotations to provide contextual information about each sample.

The second heatmap visualizes the pairwise distances between samples based on their gene expression profiles. To prepare the data for visualization, the raw count data from the DESeq2 object (dds1) underwent variance-stabilizing transformation (VST) and regularized log transformation (rlog) to normalize the gene expression values and mitigate technical noise. The transformed expression values were then used to compute the pairwise distances between samples using the Euclidean distance metric. The resulting sample distance matrix, which quantifies the similarity or dissimilarity between each pair of samples, was visualized as a heatmap using the pheatmap package. Clustering of both rows and columns in the heatmap was performed based on the sample distance matrix to reveal any inherent patterns or groupings within the data.

2.6.2.4 PCA Plot

In this study, PCA was performed on the variance-stabilized transformed (VST) data obtained from the DESeq2 object (dds1). The plotPCA function from the DESeq2 package was utilized to generate a PCA plot, where each sample is represented as a point in a two-dimensional space defined by the first two principal components (PC1 and PC2). To facilitate the interpretation of the PCA plot, additional information about the experimental conditions and size factors associated with each sample was incorporated. The intgroup parameter was specified to indicate which experimental factors should be included as grouping variables, enabling the visualization of sample clustering based on these factors. Additionally, the returnData parameter was set to TRUE to return the PCA data, including the coordinates of each sample in the principal component space.

Subsequently, a scatter plot was created using the ggplot2 package to visualize the PCA results. Each sample was plotted based on its coordinates in the PC1 and PC2 dimensions, with points colored according to their respective experimental conditions. The axes of the PCA plot were labeled with the percentage of variance explained by each principal component, providing insight into the contribution of each component to the overall variance in the data. By examining the distribution of samples in the PCA plot, patterns of similarity or dissimilarity between samples can be discerned, aiding in the identification of potential clusters or trends in the data.

2.6.2.5 GO Term Analysis

The custom *P. putida* Org package was created to facilitate gene ontology (GO) term analysis specifically tailored for Pseudomonas putida KT2440. The necessary R packages including clusterProfiler, AnnotationDbi, tidyr, dplyr, and AnnotationForge were loaded to support the analysis. The first step involved identifying housekeeping genes and sigma factor genes relevant to *P. putida* KT2440. The presence of eight housekeeping genes, including "argS", "gyrB", "ileS", "nuoC", "ppsA", "recA", "rpoB", and "rpoD", was confirmed in the gene annotation file obtained from a previous study (Ogura K. et al, 2019). Additionally, sigma factor genes, encompassing "rpoD", "rpoS", "fliA", "rpoH", "rpoE", and "rpoN", along with ECF family RNA polymerase sigma factors, were extracted from the annotation file. Subsequently, a list of genes of interest, required for GO term analysis, was compiled.

The next step involved preparing a data frame containing gene ID information using the Pseudomonas putida KT2440 gene annotation CSV file. This file was processed to extract relevant columns including gene ID, locus tag, and product name. Duplicate entries and rows with missing information were removed to ensure data integrity. Moreover, chromosome information was extracted to provide additional context for gene analysis. Another CSV file containing GO annotations was imported and filtered to obtain GO terms associated with the genes of interest. The resulting data frame was refined to include columns for gene ID, GO accession, and GO evidence code.

Finally, the custom *P. putida* Org package was generated using the makeOrgPackage function from the AnnotationForge package. This function utilized the processed gene information and GO annotations to create an Org package tailored for *P. putida* KT2440. The package was assigned a version number, taxonomic identifiers, and taxonomic classification parameters. Additionally, the package was configured with contact information for maintenance and authorship. Following package creation, the install.packages function was used to install the generated Org package locally for subsequent use in GO term analysis.

For the contrasts, upregulated genes and downregulated genes, determined by a p-value of less than 0.05 and a log fold threshold of 1, were subjected to GO analysis to elucidate their functional roles. The *P. putida* KT2440 gene annotation CSV file was imported and processed to create an annotation data frame. Subsequently, upregulated/downregulated genes specific to the contrasts were filtered from the annotation data frame based on their gene IDs. The resulting subset of genes was further refined to extract locus tags, which served as identifiers for GO analysis.

The enrichGO function from the clusterProfiler package was employed to perform GO analysis on the genes. This function utilized the gene locus tags along with the *P. putida* KT2440 organism database (org.Pputida.eg.db) to retrieve GO annotations. Gene symbols were specified as the key type for gene mapping, and the analysis focused on biological process ontology. To control multiple testing, the Benjamini-Hochberg method was applied with a significance threshold set at 1 for both raw and adjusted p-values.

The results of the GO analysis for the upregulated/downregulated genes in the contrasts were obtained and inspected. This included significant enriched GO terms, along with associated statistical measures such as p-values, adjusted p-values, and gene counts. The top enriched GO terms were extracted and visualized as a data frame for further interpretation and the creation of pie charts.

To generate pie charts illustrating the GO term analysis results for the upregulated and downregulated genes of each contrast, a GO_data data frame was constructed. This data frame includes columns for the GO term ID, description, and gene ratio. The gene ratio represents the ratio of genes associated with each GO term that were differentially expressed over the total number of genes analyzed. The gene ratio was split into numerator and denominator components using the strsplit function, and then converted into numeric values. Subsequently, the proportions of differentially expressed genes for each GO term were calculated by dividing the numerator by the denominator. Labels were created to display both the GO term description and the gene ratio on the pie chart slices. Finally, a pie chart was generated using the pie function, with the proportions as input data and the labels displaying the GO term descriptions and gene ratios. This approach facilitated the visualization of the distribution of differentially expressed genes across various biological processes associated with the GO terms analyzed in the contrasts.

3. Results

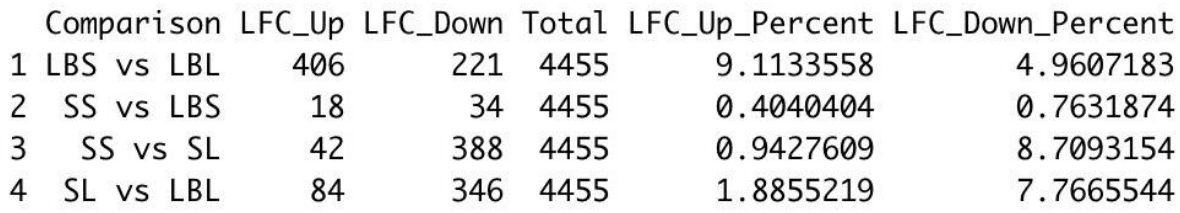
In the following sections of the report, the samples will be referenced as follows: *P. putida* grown in LB medium during the log phase of growth will be denoted as LBL, while samples from LB medium during the stationary growth phase will be labeled as LBS. Similarly, samples cultivated in SESOM medium during the log growth phase will be referred to as SL, and those from SESOM medium during the stationary growth phase will be represented by SS.

3.1 Differential Gene Expression Analysis

To identify genes whose expression levels significantly change between the different conditions or sample groups, four DESeq2 reports were generated and formatted into one table (Table 1). After running DESeq2 analysis, significance was determined by a p-value less than 0.05, and a log fold change threshold of 1 was used to measure upregulation or downregulation. The differential expression analysis across all four contrasts yielded insightful findings regarding the gene expression dynamics of *P. putida*￼

In the comparison between LBS and LBL, 406 genes exhibited upregulation while 221 genes showed downregulation, out of the total 4455 genes analyzed (Table 1). Similarly, the contrast between SS and LBS conditions revealed 18 upregulated genes, 34 downregulated genes, (Table 1). Likewise, the comparison between SS and SL demonstrated 42 upregulated genes, and 388 downregulated genes (Table 1). Lastly, the contrast between SL and LBL unveiled 84 upregulated genes and 346 downregulated genes, providing further insights into the molecular responses of *P. putida* to variations in growth medium and phase (Table 1).

The most upregulated genes can be found in the LBS vs LBL contrast, where the most downregulated genes can be found in the SS vs SL contrast. The contrast with the least differentially expressed genes is the SS vs LBS, most likely due to the stationary growth phases causing lower activity than log growth phase. A notable variation is seen in the SL vs LBL contrast where 84 genes upregulated when going from SL to LBL and 346 are downregulated. This means that those 346 genes are upregulated in SL demonstrating the impact that growth media has on variation.



```
  Comparison LFC_Up LFC_Down Total LFC_Up_Percent LFC_Down_Percent
1 LBS vs LBL    406      221  4455      9.1133558        4.9607183
2  SS vs LBS     18       34  4455      0.4040404        0.7631874
3   SS vs SL     42      388  4455      0.9427609        8.7093154
4  SL vs LBL     84      346  4455      1.8855219        7.7665544
```

Table 1: DESeq2 results for all four contrasts demonstrating the amount of upregulated or downregulated genes, the total amount of genes in the sample, and the percent of upregulated genes downregulated genes in the sample.

3.2 Visualizing Differential Gene Expression for Contrasts

To represent and further understand the differential gene expression for each contrast four volcano plots were created. The significance threshold for the data was a p-adjusted value of below 0.05 and a log change threshold of 1. Along the x-axis, the log2-transformed fold change in gene expression is depicted. Genes positioned to the right of the plot demonstrate upregulated genes, while those on the left exhibit downregulated genes. On the y-axis, the negative log10-transformed adjusted p-value is represented, indicating the level of statistical significance associated with each gene. Points positioned higher on the y-axis indicate greater statistical significance. Each point on the plot corresponds to a specific gene, with its position determined by both its fold change and adjusted p-value. Red points denote genes that exhibit significant differential expression, characterized by substantial fold changes. Conversely, blue points represent genes that fail to meet significant criteria or demonstrate low fold changes.

In Figures 3 and 4 volcano plots are depicted for contrasts of the same medium but different growth rate. In Figure 3, the volcano plot illustrates the comparison between LBS and LBL conditions. Red points scatter across both sides of the plot, indicating genes with varied expression levels. Notably, these points are generally evenly distributed, with slightly more red points on the right side of the plot than the left. Meanwhile, blue points, signifying genes with expression or low fold changes, are positioned lower on the plot compared to the red points. Since *P. putida's* gene expression levels remained fairly constant amongst both LB samples, this suggests that growth rate does not have much of an impact on variation, and growth medium may have more of an impact. Figure 4 illustrates the comparison between SS and SL conditions. A notable observation is the abundance of red points clustered predominantly on the left side of the plot, indicating a higher prevalence of downregulated genes in SS compared to SL. Meanwhile, the blue points, appear to be noticeably lower on the plot compared to the red points.

Figures 5 and 6 portray the volcano plots for contrasts of differing mediums. Figure 6 depicts the comparison between SL and LBL conditions. A greater abundance of red points is observed on the left side of the plot, indicating a higher number of downregulated genes compared to upregulated ones. Similarly to Figures 3 and 4, the blue points in Figure 6 also appear lower than the red points, denoting their lower statistical significance or minimal fold changes. This suggests *P. putida* may prefer SL conditions over LBL conditions, showing the impact of growth media. In Figure 5 the SS vs LBS contrast is shown but differs slightly in appearance from the others. In this plot there are significantly fewer red points than in the other contrasts, the number of blue points may surpass the amount of red. The red points on the plot exhibit a relatively even distribution, with a slightly greater concentration observed on the left side. This is understandable that gene expression is low for both samples in the context of the data since growth tends to slow down or stop completely in stationary growth.

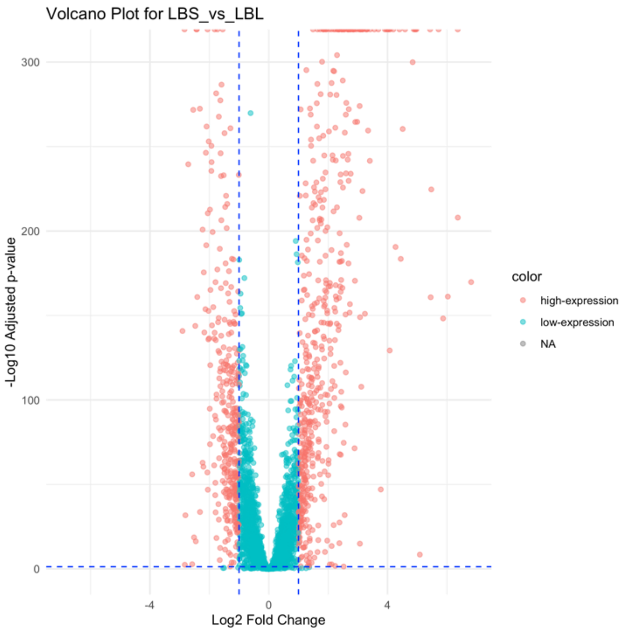


Figure 3: LBS vs LBL volcano plot

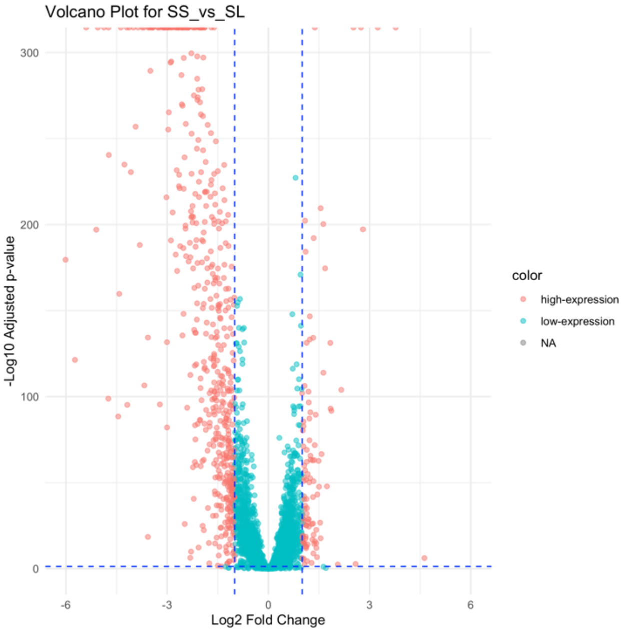


Figure 4: SS vs SL volcano plot

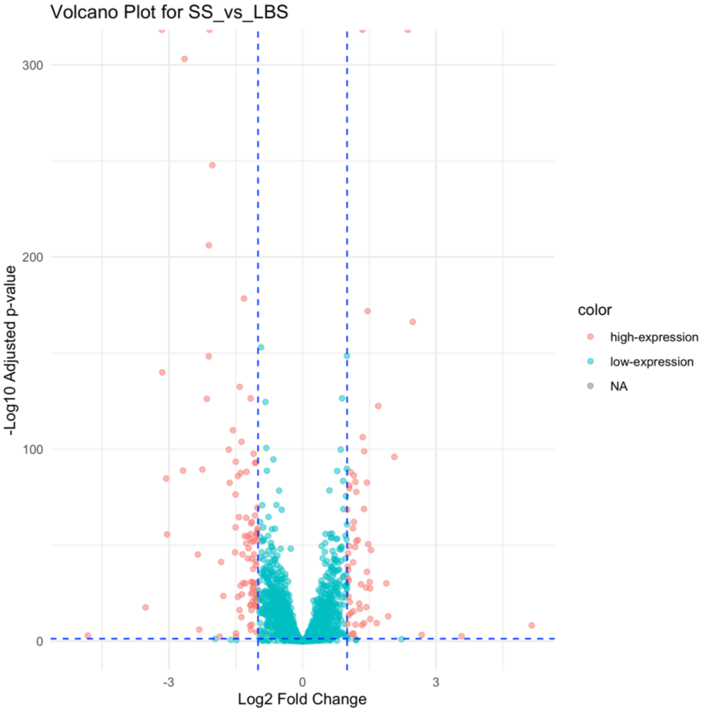


Figure 5: SS vs LBS volcano plot

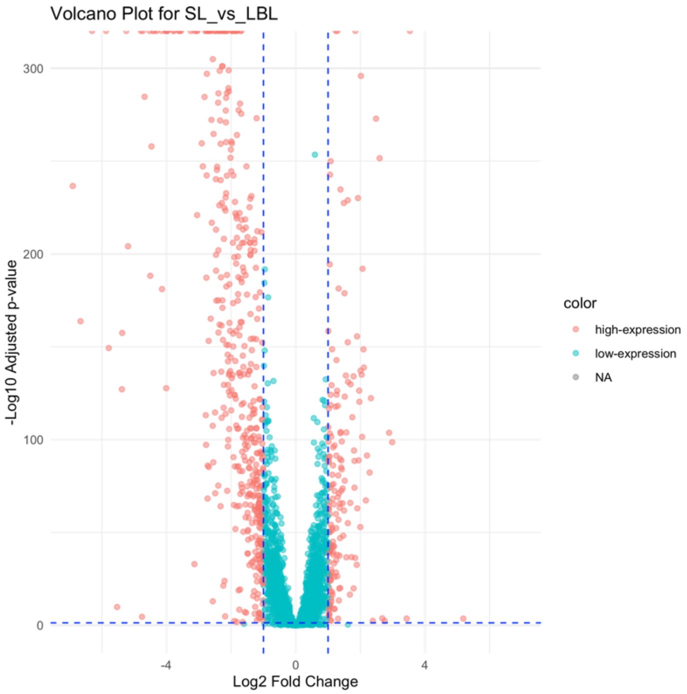


Figure 6: SL vs LBL. volcano plot

3.3 Assessing Data Quality and Sample Similarity

3.3.1 Data Quality and Sample Similarity with Heatmaps

To observe overall data quality and sample similarity two heatmaps were created and utilized expression levels or sample-sample relationships. The heatmap shown in Figure 7 provides a graphical representation of the expression levels of the top 20 genes, in terms of their mean expression levels, across the different conditions. Each row in the heatmap represents a gene, while each column corresponds to a sample. Each condition has three sample replicates; however, six samples are shown for each condition due to the paired end analysis described in the Methods section of this report. The color intensity in the heatmap indicates the relative expression level of each gene across the samples, with higher expression levels depicted in darker shades and lower expression levels in lighter shades (Figure 7). The heatmap reveals similar patterns of gene expression across the sample replicates, which demonstrates that the data is good quality and is being interpreted in the way that would be expected. The highest gene expression levels are found in LBL, SL, and SS replicates. This suggests a preference for these conditions for *P. putida* growth and should be considered when cultivating the bacteria for future manipulation. (Figure 7).

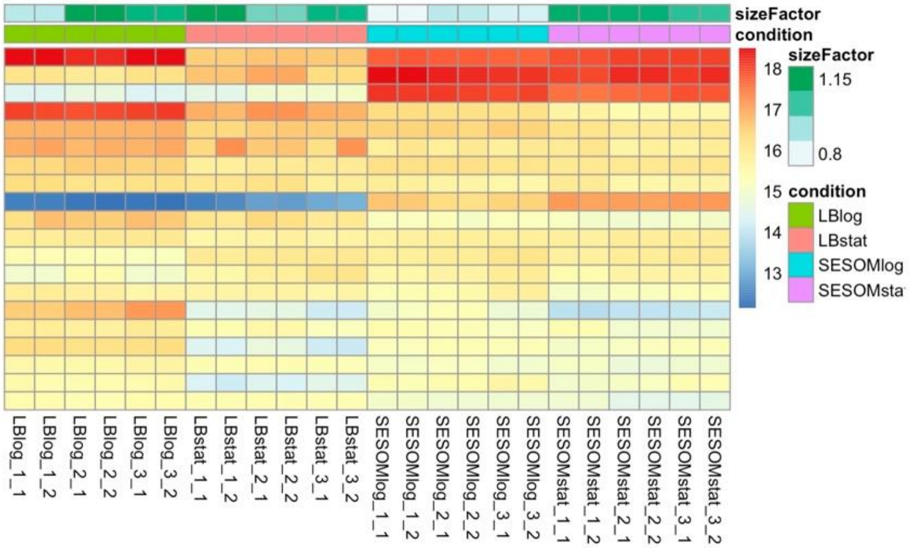


Figure 7: Heatmap of top 20 genes for the samples. Three sample replicates were collected for each condition, but six points appear for each due to the paired end analysis ran on the data.

In the heatmap shown in Figure 8, the Euclidean distance of samples is represented where the blue color depicts high similarity or closeness between samples, while lighter colors such as white or light blue indicate greater dissimilarity or distance. Therefore, if a cell in the heatmap is blue, it suggests that the corresponding samples are similar or closely related based. In this heatmap, sample replicates are not labeled with their pair end numbering as they are in Figure 7. However, each paired end replicate should be next to its pair within the condition for this heatmap they are just not labeled as such. With this considered, as expected sample replicates of the conditions have the closest clustering as observed in this heatmap by the dark blue shade shared by each replicate within the samples. This highlights that the data is good quality due to the sample replicates showing expected similarity.

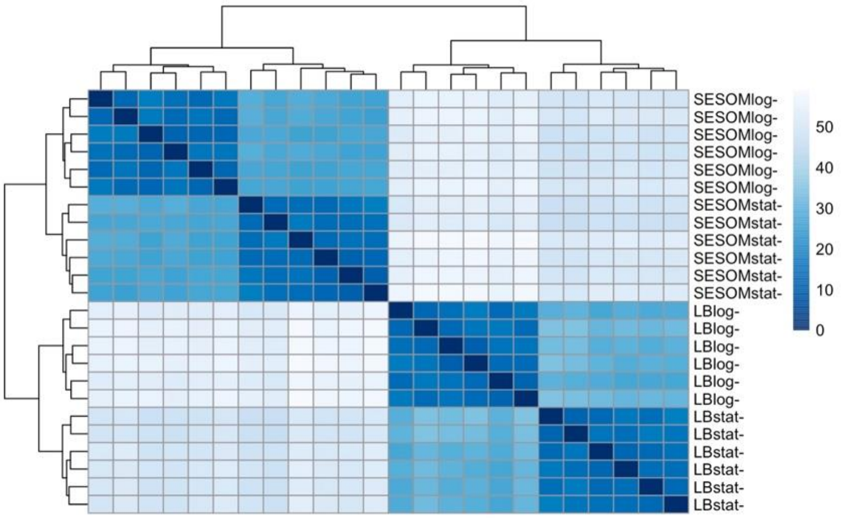


Figure 8: Euclidean distance heatmap for sample-sample relationships. Three sample replicates were collected for each condition, but six points appear for each due to the paired end analysis ran on the data.

3.3.2 Data Quality and Sample Similarity with PCA Plot

To observe data variability and overall quality one PCA plot was created for the samples. (Figure 9). PC1, which accounts for 76% of the total variance, represents the primary axis of variation in the data. This suggests that the largest source of variation among the samples is captured along PC1. Furthermore, PC2, capturing an additional 7% of the variance, represents a secondary axis of variation. While PC2 explains a smaller proportion of the total variance compared to PC1, it still contributes significantly to the overall structure of the data. SL and SS appear to the far right of PC1 axis clustering close to one another between -5 and 5 of the PC2 axis. Whereas LBS and LBL are to the left of the PC1 axis and are far apart where LBL clustering is close to -10 on the PC2 axis and LBS clustering is close to 10 on the PC2 axis. These plot results suggest that the data is good quality due to the high variability within the data set but low variability within sample replicates demonstrated by the close clustering of the replicates. This separation between LB samples and SESOM samples also highlights the impact of growth medium on variation over growth phase.

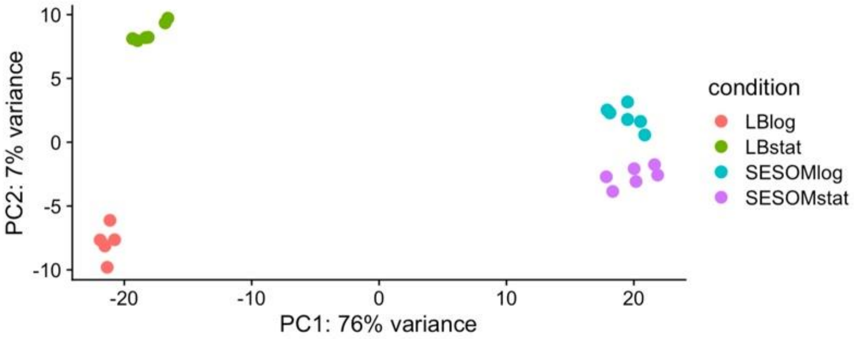


Figure 9: PCA Plot of variance after DESeq2 analysis. Three sample replicates were collected for each condition, but six points appear for each due to the paired end analysis ran on the data.

3.4 Intersection Analysis of Upregulated and Downregulated Genes

To visualize and understand the commonality between the upregulated genes and downregulated genes of the contrasts, two Venn diagrams were generated containing the upregulated genes for all four contrasts, and the downregulated genes for all contrasts. The first Venn diagram illustrates the intersection of upregulated genes (Figure 10), while the second diagram depicts the overlap of downregulated genes (Figure 11).

In the Venn diagram representing upregulated genes, it was observed that 497 genes were upregulated in the LBS vs LBL contrast, 39 in SS vs SL, 112 in SL vs LBL, and 19 in SS vs LBS (Figure 10). Notably, regions where contrasts overlap without shared genes are represented as zero, indicating exclusive upregulation in specific comparisons. Of interest are the gene overlaps revealed in the diagram. Notable findings include 56 genes shared between SS vs SL and SL vs LBL, as well as 28 genes shared between LBS vs LBL and SS vs LBS (Figure 10). The 56 shared genes amongst SS vs SL and SL vs LBL reveal genes that are upregulated in the log phase compared to stationary phase regardless of the growth medium. The overlap between LBS vs LBL and SS vs LBS indicates shared genes affected by LB medium compared to soil stationary phase.

In the Venn diagram depicting downregulated genes, a total of 312 genes exhibited downregulation in the LBS vs LBL contrast, while 94 genes showed decreased expression in SS vs SL, 78 in SL vs LBL, and 28 in SS vs LBS (Figure 11). An intriguing observation arises from the overlap between SS vs SL and SL vs LBL, where 377 downregulated genes were shared. This substantial overlap may suggest common regulatory pathways or biological processes affected by the transition from soil stationary phase to LB stationary phase. Additionally, noteworthy overlaps include 32 genes shared between SS vs SL and SS vs LBS which represents connections between soil stationary phase and LB stationary phase, distinct from the LB soil medium, as well as 28 downregulated genes shared between LBS vs LBL and SS vs LBS demonstrating genes that are affected by LB medium compared to SS (Figure 11).

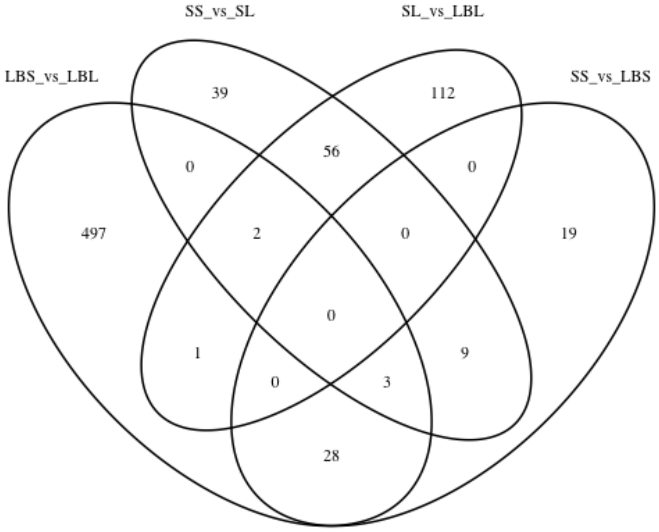


Figure 10: Venn diagram of upregulated genes amongst contrasts

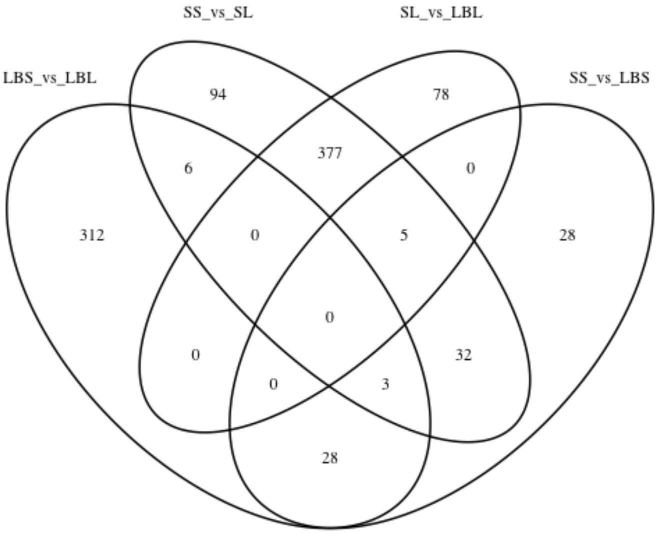


Figure 11: Venn diagram of downregulated genes amongst contrasts

3.5 Biological Processes of Samples

GO enrichment analysis was used to understand the biological processes of the differentially expressed genes amongst the contrasts, revealing common pathways or functionally related genes. The results of GO enrichment analysis provided the ID for the GO term, description of the biological process, gene ratio of my input list, Bg ratio of my dataset, p-value, p-adjusted value, qvalue, and the geneID. Utilizing the enriched Gene Ontology terms obtained from the analysis, visually informative pie charts were generated. Each pie chart represents the GO terms associated with either upregulation or downregulation in a specific contrast, providing a depiction of the enriched biological processes. In total, eight pie charts were generated, illustrating the GO terms for both upregulated and downregulated genes across each contrast.

For the LBS vs LBL contrast, Figures 12 and 13 illustrate the GO terms for upregulated and downregulated genes, respectively. Similarly, Figures 14 and 15 depict the GO terms for SL vs LBL contrast, while Figures 16 and 17 represent the GO terms for upregulated and downregulated genes in SS vs LBS. Finally, Figures 18 and 19 visualize the GO terms for upregulated and downregulated genes in SS vs SL. Each pie chart encapsulates six GO terms, each segment labeled with the specific GO term description and its corresponding gene ratio. It's important to note that while all pie charts contain six GO terms, the specific terms vary between contrasts.

GO terms in Figure 12 reveal biological processes enhanced in the log phase independent of the LB media. GO terms found in Figure 13 reveal biological processes that are suppressed in *P. putida* in the log phase independent of LB media. GO terms found in Figure 14 reveal genes that are enhanced in going from SESOM to LB independent of the growth phase. Whereas GO terms found in Figure 15 reveal genes that are suppressed going from SESOM to LB independent of the growth phase. GO terms found in Figure 16 reveal genes that are enhanced going from SESOM to LB regardless of the growth phase. Whereas Figure 17 shows genes that are suppressed going from SESOM to LB regardless of the growth phase. Lastly GO terms found in Figure 18 reveal genes that are activated going from stationary to log phase growth independent of the media. Whereas Figure 19 reveals genes that are suppressed going from stationary to log phase growth independent of the media.

The GO term analyses revealed distinct biological processes influenced by growth phase and media type in *P. putida*. Upregulated genes in the LBS vs LBL comparison (Figure 12) highlighted processes enhanced during log phase growth, irrespective of the LB medium, while downregulated genes (Figure 13) indicated suppressed processes during this phase. In contrast, genes upregulated in the transition from SESOM to LB media (Figure 14) suggested adaptations specific to LB, while downregulated genes (Figures 15) indicated suppressed functions under LB conditions. Figure 16 reveals genes that are enhanced going from SESOM to LB regardless of the growth phase. Whereas Figure 17 shows genes that are suppressed going from SESOM to LB regardless of the growth phase. Furthermore, comparison between stationary and log phase growth (Figures 18 and 19) revealed activated and suppressed genes independent of media type. These findings can be used to further understand the biological processes of *P. putida* in the varied growth conditions.

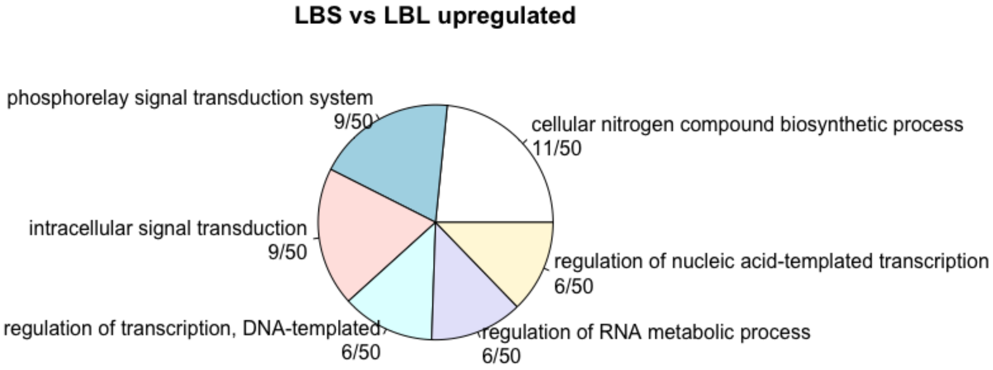


Figure 12: GO terms pie chart for upregulated genes for LBS vs LBL

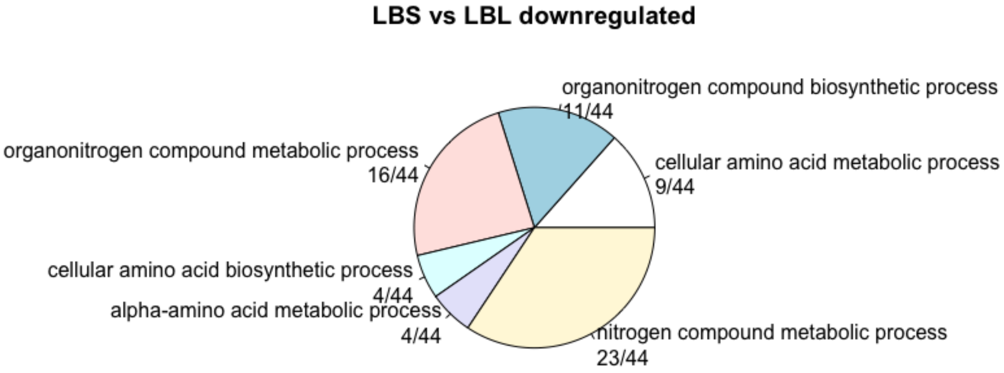


Figure 13: GO terms pie chart for downregulated genes for LBS vs LBL.

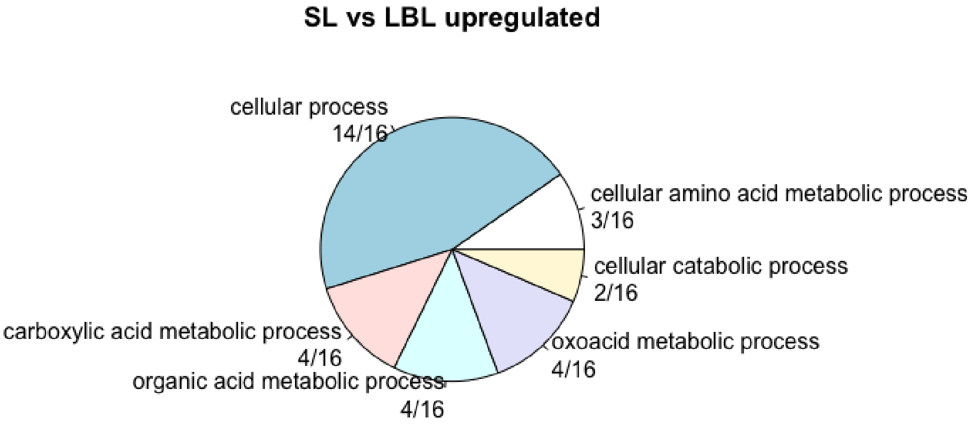


Figure 14: GO terms pie chart for upregulated genes for SL vs LBL.

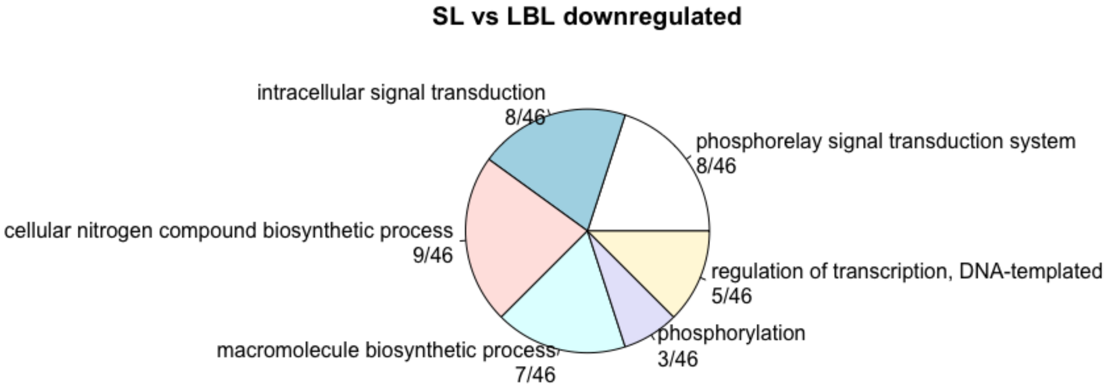


Figure 15: GO terms pie chart for downregulated genes for SL vs LBL.

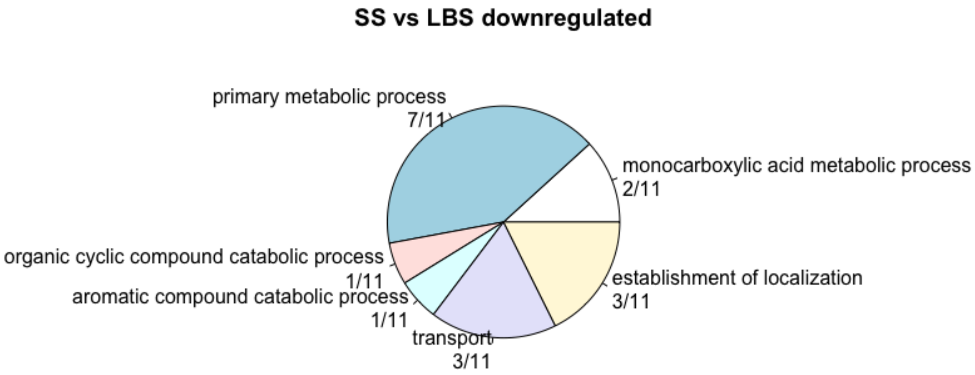


Figure 16: GO terms pie chart for downregulated genes for SS vs LBS.

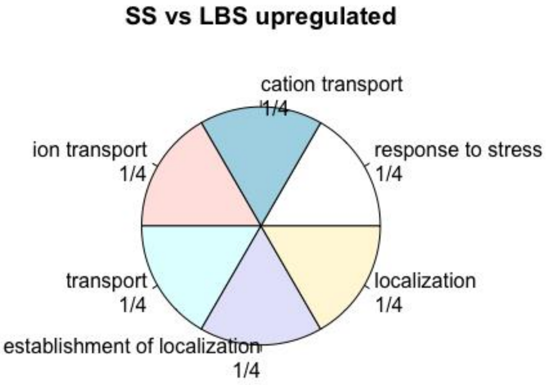


Figure 17: GO terms pie chart for upregulated genes for SS vs LBS..

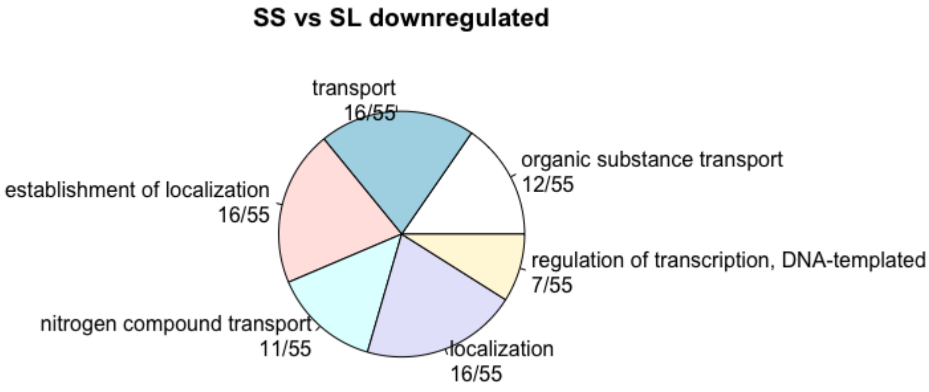


Figure 18: GO terms pie chart for downregulated genes for SS vs SL.

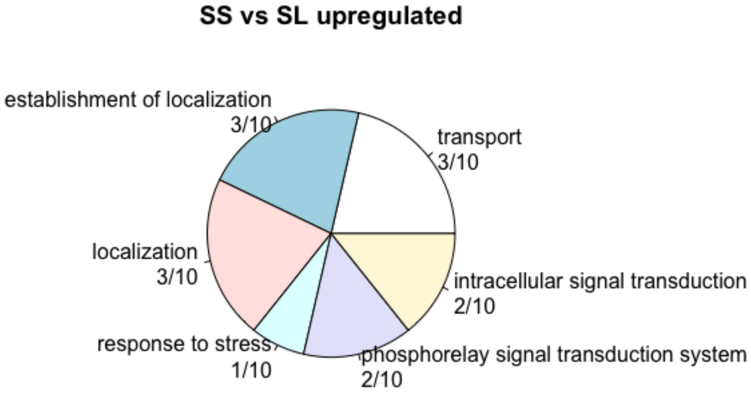


Figure 19: GO terms pie chart for upregulated genes for SS vs SL.

4. Discussion

4.1 Impact of Growth Media and Growth Phase on Gene Expression Variation

The analysis of differential gene expression across varying growth conditions provided valuable insight into the impact of growth media and growth phase on *P. putida*. Both growth media and growth phase altered gene expression amongst conditions. When examining the volcano plots and venn diagrams we see instances of gene expression changing due to change in media. For instance, in Figure 6 we see the impact of changing from SL to LBL and the large number of downregulated genes in comparison to the number of upregulated genes. Similarly in Figure 10 and Figure 11 when looking at contrasts SL vs LBL and SS vs LBS there is evident gene variation amongst contrasts due to the change in media. Growth phase also creates variation in gene expression as seen in Figure 4, 10, and 11. In Figure 4 the volcano plot of SS vs SL shows significantly more downregulated genes showing the impact of growth phase on gene expression. In Figure 10 and 11 the venn diagrams show the impact of growth phase between the contrasts LBS vs LBL and SS vs SL. However, while both environmental conditions are important to *P. putida* persistence and gene expression behavior, the condition which has the most impact on variability is the type of growth media.

When looking at the PCA plot which measures variance (Figure 9) there is a distinct separation between LB samples and SESOM samples. This highlights the impact of growth medium on variation as the media used has more of an impact on variation than growth phase. An additional example of this is when looking at Figure 7 which displays the Euclidian distance between samples. When analyzing this heatmap the samples that show the most similarity are the sample replicates and the samples which have differing growth phases but the same media. This means that what begins to make the samples the most different is the change in media. When looking at the volcano plot of SL vs LBL we once again see that the impact of change in media is a lot more impactful than change in growth phase where we have many more downregulated genes than upregulated genes (Figure 6). However this volcano plot also reveals an interesting takeaway, that when *P. putida* is grown at the optimal growth phase for microbial growth, there are more enhanced genes in SESOM vs LB. This preference may reflect the physiological adaptation of *P. putida* to the soil environment, where nutrient availability and growth conditions differ from laboratory settings. Similarly, the contrast of SS vs LBS conditions seen in Table 1, although fewer differentially expressed genes than the other contrasts most likely due to the stationary growth phase creating reduced metabolic activity, a similar pattern was observed where there were more downregulated genes than upregulated. This means that there were more upregulated genes in SS than LBS. Overall this makes sense since the SESOM mimics the natural habitat of *P. putida* so there being better gene expression would almost be expected. This is a great point to consider when cultivating *P. putida* for future genetic engineering and cultivation.

Further exploration of the overlapping gene sets in the Venn diagrams (Figure 10 and 11) provided additional insights into common regulatory pathways and biological processes affected by the transition between growth conditions and phases. Notably, the substantial overlap between SS vs SL and SL vs LBL contrasts suggests shared regulatory mechanisms or biological responses associated with the transition from soil stationary phase to LB stationary phase. Similarly, the overlap between LBS vs LBL and SS vs LBS contrasts highlights genes affected by LB medium compared to soil stationary phase, emphasizing the influence of growth media on gene expression dynamics.

The GO term analysis results also shed light on the biological processes of *P. putida* in its preferred growth media and growth phase. This can be seen in Figure 13 where the six GO terms that were downregulated in LBL are actually upregulated in SL. Examining these GO terms can relay potential processes necessary for *P. putida* survival. For the GO terms, for example the phosphorelay signal transduction and intracellular signal transduction, these processes may be important for the bacteria to adapt to its soil environment which may be filled with other microbes, harsh pollutants, and other living organisms.

Overall, these findings underscore the intricate interplay between growth media, growth phase, and gene expression regulation in Pseudomonas putida, shedding light on the adaptive strategies employed by this bacterium in response to environmental changes. The comprehensive analysis of gene expression dynamics provides a foundation for further investigations into the molecular mechanisms governing *P. putida* physiology and its potential applications in bioremediation and environmental monitoring.

4.2 Recommendations and Conclusions

Based on the findings of this study, several recommendations and conclusions can be drawn regarding gene expression dynamics and future research directions. Firstly, the most dramatic changes in gene expression were observed in the transition from SL to LBL conditions, where a substantial difference in the number of upregulated and downregulated genes was evident. This highlights the significant impact of growth media transitions on Pseudomonas putida gene regulation and suggests the need for further investigation into the underlying molecular mechanisms driving these changes. Additionally, the use of pairwise comparisons for Venn diagrams may provide a more focused and informative analysis compared to the inclusion of all four contrasts in a single diagram. This approach would allow for a clearer interpretation of overlapping gene sets and facilitate the identification of common regulatory pathways or biological processes shared between specific contrasts.

While this study primarily focused on genes exhibiting significant changes in expression levels, future research should also consider genes that remain relatively constant across different growth conditions and phases. Analyzing genes with stable expression patterns can provide valuable insights into core metabolic processes and regulatory networks essential for *P. putida* survival and adaptation.

Furthermore, the evaluation of Gene Ontology (GO) terms for each contrast should be given more attention in future studies. GO enrichment analysis provides valuable information about the biological processes, molecular functions, and cellular components associated with differentially expressed genes. A more comprehensive analysis of GO terms can enhance our understanding of the functional implications of gene expression changes and help identify key biological pathways influenced by growth conditions and phase transitions.

In conclusion, this study sheds light on the dynamic nature of Pseudomonas putida gene expression in response to varying growth conditions and phases. The findings underscore the importance of growth media and growth phase transitions in shaping gene regulation dynamics and highlight avenues for future research aimed at elucidating the molecular mechanisms underlying these changes. By considering the recommendations outlined above, future studies can further advance our understanding of *P. putida* physiology and its potential applications in bioremediation and environmental monitoring.

References

Benner, S. A., & Sismour, A. M. (2005). Synthetic biology. *Nature Reviews Genetics*, *6*(7), Article 7. https://doi.org/10.1038/nrg1637

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Cameron, D. E., Bashor, C. J., & Collins, J. J. (2014). A brief history of synthetic biology. *Nature Reviews Microbiology*, *12*(5), Article 5. https://doi.org/10.1038/nrmicro3239

Dundar, F., et al. (2015). Introduction to Differential Gene Expression Analysis Using RNA-seq. Applied Bioinformatics. *Weill Cornell Medical College*, 63.

Ezezika, O. C., & Singer, P. A. (2010). Genetically engineered oil-eating microbes for bioremediation: Prospects and regulatory challenges. *Technology in Society*, *32*(4), 331–335. https://doi.org/10.1016/j.techsoc.2010.10.010

FastQC. (n.d.). Retrieved March 19, 2024, from https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/fastqc.html

Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J., & Griffith, O. L. (2015). Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLOS Computational Biology*, *11*(8), e1004393. https://doi.org/10.1371/journal.pcbi.1004393

Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T., & Winter, D. R. (2018). A Beginner's Guide to Analysis of RNA Sequencing Data. *American Journal of Respiratory Cell and Molecular Biology*, *59*(2), 145–157. https://doi.org/10.1165/rcmb.2017-0430TR

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

LIBRARY, D. K. M. P. (n.d.). Pseudomonas putida, Gram-negative, SEM - Stock Image—C032/1985. Science Photo Library. Retrieved April 22, 2024, from https://www.sciencephoto.com/media/798804/view/pseudomonas-putida-gram-negative-sem

Liebeke, M., Brözel, V. S., Hecker, M., & Lalk, M. (2009). Chemical characterization of soil extract as growth media for the ecophysiological study of bacteria. *Applied microbiology and biotechnology*, *83*(1), 161–173. https://doi.org/10.1007/s00253-009-1965-0

Linux Clustering | SIOS. (n.d.). SIOS Technology Corp. Retrieved March 19, 2024, from https://us.sios.com/linux-clustering/

Liu, Y., Feng, J., Pan, H., Zhang, X., & Zhang, Y. (2022). Genetically engineered bacterium: Principles, practices, and prospects. *Frontiers in Microbiology*, *13*. https://www.frontiersin.org/articles/10.3389/fmicb.2022.997587

Nakazawa, T. (2002). Travels of a Pseudomonas, from Japan around the world. *Environmental Microbiology*, *4*(12), 782–786. https://doi.org/10.1046/j.1462-2920.2002.00310.x

Pant, G., Garlapati, D., Agrawal, U., Prasuna, R. G., Mathimani, T., & Pugazhendhi, A. (2021). Biological approaches practised using genetically engineered microbes for a sustainable environment: A review. *Journal of Hazardous Materials*, *405*, 124631. https://doi.org/10.1016/j.jhazmat.2020.124631

Piper, M. M., Bob Freeman, Mary. (2017, June 7). Alignment with STAR. Introduction to RNA-Seq Using High-Performance Computing - ARCHIVED. https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

Piper, M. M., Radhika Khetani, Mary. (2017, May 12). Gene-level differential expression analysis with DESeq2. Introduction to DGE - ARCHIVED. https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html

RNAseq-tutorial—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved December 15, 2023, from https://research.csc.fi/rnaseq-tutorial

Rock, A. (2021, December 6). *Microbial Community Models for Measuring Survival and Persistence of SynBio Microbes in Soil*. Worcester Polytechnic Institute. https://digital.wpi.edu/show/v692t9438

Services | VANGARD. (n.d.). Retrieved December 15, 2023, from https://bioinfo.vanderbilt.edu/vangard/services-rnaseq.html

Son, J., Lim, S. H., Kim, Y. J., Lim, H. J., Lee, J. Y., Jeong, S., Park, C., & Park, S. J. (2023). Customized valorization of waste streams by Pseudomonas putida: State-of-the-art, challenges, and future trends. Bioresource Technology, 371, 128607. https://doi.org/10.1016/j.biortech.2023.128607

Stirling, F & Silver, P A. Controlling the implementation of Transgenic Microbes: Are We Ready for What Synthetic Biology Has to Offer? *Mol. Cell* **78**, 614-623 (2020).

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. Bioinformatics, 31(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Volke, D. C., Calero, P., & Nikel, P. I. (2020). Pseudomonas putida. *Trends in Microbiology*, *28*(6), 512–513. https://doi.org/10.1016/j.tim.2020.02.015

Voltmer, S. (2023). *Lab vs Soil: The Transcriptome of Pseudomonas putida*.: Worcester Polytechnic Institute. https://digital.wpi.edu/show/h415pf44s

Weimer, A., Kohlstedt, M., Volke, D. C., Nikel, P. I., & Wittmann, C. (2020). Industrial biotechnology of Pseudomonas putida: Advances and prospects. *Applied Microbiology and Biotechnology*, *104*(18), 7745–7766. https://doi.org/10.1007/s00253-020-10811-9

Xu, J. (2014). *Next-Generation Sequencing: Current Technologies and Applications*. Caister Academic Press. http://ebookcentral.proquest.com/lib/wpi/detail.action?docID=5897784