Lab vs Soil: The Transcriptome of Pseudomonas putida

By:

Stokley C. Voltmer

A Thesis
Submitted to the Faculty
of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Bioinformatics and Computational Biology

December 2023

APPROVED:

Dr. Natalie Farny, Advisor

Dr. Elizabeth Ryder, Reader

Abstract

Pseudomonas putida is a candidate for use in the soil environment, in bioremediation and as a part of the soil microbiome. However, *P. putida* often does not behave in the lab as it does in the field. This work aims to bridge the gap between the lab and field by observing the gene expression of *Pseudomonas putida* grown under lab conditions (Luria broth, LB) and a soil analog (soil extracted soluble organic matter, SESOM) created by our laboratory to determine changes in gene expression of *P. putida* between conditions, as well as the practicality of SESOM as a soil analog. Additionally, this work examines the use of Nanopore sequencing as an alternative to traditional sequencing methods for transcriptomics. The differential expression of mRNA in *P. putida* grown in LB and SESOM were determined using both Illumina and direct RNA Nanopore sequencing. We found a clear difference in expression between *P. putida* grown in the lab and soil conditions, opening the door for further research into the underlying mechanisms of transcriptional regulation under LB and SESOM conditions. Additionally, we found that Nanopore is not a suitable method for direct RNA sequencing in prokaryotes, however this may change as the technology matures. This research will ultimately improve our ability to engineer *P. putida* for future soil biosensing and bioremediation applications.

Acknowledgements

This project would not have been possible without the support of Professor Farny who I would like to thank for her excellent guidance and endless patience as this project increased in duration. In the same vein I would like to thank Professor Ryder for being flexible and accommodating with her time and committed to seeing the project through. I would like to thank the Farny lab for being an amazing group of people who made what could at times have been frustrating work a joy by virtue of their presence. I especially thank Matt and Maia for their help starting the project, and Ally and Alli for their encouragement and moral support during the challenging stages of the experiment. Lastly, I thank my family for their love and understanding throughout this journey.

Contents

Table of Figures

List of Tables

Introduction

The importance of soil health

Soil is essential for 25% of the Earth's biodiversity, 98% of global food production (Kraamwinkel et al., 2021) and 98.8% of daily calories consumed (Kopittke et al., 2019). The amount of soil used for food production totals around 37% of all ice-free soil, with croplands accounting for 12% and grazing pastures the other 25% (Kopittke et al., 2019). The amount of soil devoted to agriculture is not expected to significantly change in the coming years, while the amount of food produced must continue to increase to keep up with an ever-increasing population. For example, wheat yields are projected to increase 11% by 2026 driven by an increase in demand (OECD/FAO, 2017), while the area that produces wheat is expected to increase by only 1.8% (Kopittke et al., 2019). An even more stark example is rice, where 93% of projected increase in production is expected to come from increased yield alone (Kopittke et al., 2019). These projections indicate that soil used for agriculture is under more stress than ever before, and therefore soil health is more crucial than ever before.

Soil health goes beyond its fertility and looks at it holistically as the ability of soil to function as a living ecosystem in concert with plants, animals, and humans (Lehmann et al., 2020). Unfortunately, agricultural practices can have detrimental effects on soil health. Agriculture results in the mineralization of soil organic matter, which is used by plants to extract nutrients, breaking down the organic matter into its constituent inorganic parts. Agriculture mineralizes 30-60% of the organic carbon in soil. Mineralization of organic material results in a decrease in soil fertility over time, releasing carbon dioxide and requiring an increase in the use of fertilizer, the primary fertilizer being nitrogen. Nitrogen is applied to soil in an increasingly intense and inefficient manner. The efficiency of application has decreased from 68% in 1961 to just 47% in 2010, meaning that less than half of the nitrogen applied is used by plants, while the rest contaminates the soil. Excess nitrogen contributes to global warming with the emissions of nitrous oxide, a greenhouse gas, and acidifies the soil further reducing plant growth and yield in a self-perpetuating cycle (Kopittke et al., 2019). Approximately 30% of ice-free topsoil is acidic, as is 75% of subsoil.

Soil health is very important to agriculture as the nutritional content of crops is related to the micronutrient availability in the soil and the soil biodiversity (Lehmann et al., 2020). The soil microbiome is the diverse collection of microbial species found in soil, which includes bacteria, fungi, and other single-celled organisms (Fierer, 2017). The soil microbiome can increase plant production, and a more diverse microbiome generally leads to more crop production (Kopittke et al., 2019; Lehmann et al., 2020).

Soil health effects on agriculture also affect people, as not only is the crop yield reduced but also pollutants in the soil can be taken up through plants and accumulate in produce (Lehmann et al., 2020). In China, dietary intake of cadmium has doubled between the years of 1990 to 2015 to a rate of 15.3 micrograms per kilogram of bodyweight, additionally, in Bangladesh, arsenic contaminated ground water used to irrigate rice paddies led to elevated intake of arsenic levels (Kopittke et al., 2019).

The traditional approach to dealing with soil health, especially pollution, has been to correct it manually or through a physical process. For example, soil acidification can be corrected with lime; however, it is very financially costly. Lime has limited mobility in the soil and needs to be placed directly in the problem spots which is especially difficult in the subsoil. In Australia, it is economically viable to only lime only 4% of the total area affected by acidity; the total affected area comprising approximately 123.3 million hectares (Kopittke et al., 2019).

P. putida as a possible solution to remediating soil health

A solution to remediating soil health is through the use of genetically engineered bacteria called bioremediation. An example of a bacterium that has previously shown promise in the realm of bioremediation is the subject of this thesis, *Pseudomonas putida*, or *P. putida. Pseudomonas putida* is a gram-negative, rod shaped, aerobic soil bacteria (Timmis, 2002). *P. putida* is often found in and around water, plants, soil, and polluted areas (Weimer et al., 2020). In addition to being found around plants, *P. putida* can colonize the rhizosphere of plants and develop as a part of a plant-microbe system (Zuo et al., 2015).

P. putida has been shown to be effective in bioremediation of soil of its main pollutants, pesticides, industrial products, and oil. *P. putida* was able to degrade both organophosphate (OP) and pyrethroid pesticides simultaneously through the simultaneous expression of the OP-

degrading gene *mpd* and the pyrethroid-hydrolyzing carboxylesterase gene *pytH* (Zuo et al., 2015). The ability to degrade both contaminants at the same time is significant as the two pesticide classes are often used together, and OPs increase the toxicity of pyrethroids. p-Nitrophenol (PNP) is an intermediate in the manufacturing of azo dyes and several pesticides. PNP contaminates industrial runoff, ending up in soil and groundwater (Samuel et al., 2014). PNP will degrade in surface water but will not in soil and groundwater. *P. putida* can degrade PNP into hydroquinone rapidly (Samuel et al., 2014). Additionally, because of *P. putida's* use of carbon catabolite repression (CCR), it can consume PNP while in the presence of preferable food sources such as glucose, once again providing evidence of its adaptability and showing that a site does not need to be fully saturated with contaminant to be cleaned by *P. putida* (Samuel et al., 2014). Crude oil is another large problem affecting the soil health, especially in poorer countries where growing space is limited, and oil extraction processes are not environmentally conscious. The addition of *P. putida* along with inorganic nutrients to the contaminated soil resulted in a 98% decrease in pollution in 9 weeks and resulted in a 98.8% germination rate on reclaimed land, while contaminated land only had a 27.5% germination rate. What is especially noteworthy here is that the amount of *P. putida* added to the land was just the amount that had been there prior to the initial contamination (Nwachukwu, 2001).

(Kim & Park, 2014)Part of what allows *P. putida* to survive in such varied environments is its atypical metabolism. The metabolism is controlled on a redox demand, which attempts to maintain a balance between $NADP^+$ and NADPH while slightly favoring NADPH (Nikel et al., 2015), allowing for flexibility as well as a surplus of ATP and rapid NADPH regeneration (Kukurugya et al., 2019). Additionally, *P. putida* can use carbon catabolite repression (CCR) when grown on multiple food sources to use the most optimal carbon sources available at a given time. It is even able to make use of multiple carbon sources at the same time, partitioning them into different metabolic pathways (Sudarsan et al., 2014). *P. putida* is also able to thrive in environments that would not be conducive to the survival of other bacteria; this is in part due to *P. putida's* resistance to solvents (Ramos et al., 2015) and oxidative stress (Kim & Park, 2014).

Because of its previous success in bioremediation and its atypical metabolism, there is interest in using *P. putida* as a biological chassis organism for genetic engineering (Nikel & de Lorenzo, 2013). A chassis organism is an organism that acts as a host for genetic constructs that

serve to perform some role (de Lorenzo et al., 2021). The atypical metabolism gives *P. putida* the ability to survive in locations where the usual chassis organism *Escherichia coli* is unable to live (Zuo et al., 2015). Already there are a wide range of synthetic biology tools that can be easily used with *P. putida*, including vectors of the Standard European Vector Architecture platform (Martínez-García et al., 2020), which allows for a wide variety of modularity and customization. Commonly used stems for the manipulation of the *P. putida* genome include the Tn5 and Tn7 transposon vectors (Martínez-García et al., 2014) and the Flp-FRT recombination system (Nikel & de Lorenzo, 2013). Additionally, there have been recent advances in using modification methods that do not use selection markers such as CRISPR/Cas9 techniques and adding quorum sensing promoters using the RoxS/RoxR system (Weimer et al., 2020).

As noted by Nwachukwu (2001), the amount of *P. putida* needed to restore the land was equal to the amount that had been previously present, indicating that introducing a genetically modified version of bacterium to fulfill its previous role in the microbiome may be a path to remediation. However, often times a microbiome will not respond the same way in nature as it does in the laboratory (Jansson et al., 2023). Even more significantly, an engineered bacterium itself, let alone the microbiome, often does not behave the same way in soil as it does in the laboratory (Rebello et al., 2021). We aim to begin to bridge the gap between the laboratory and the field by introducing the use of a soil analog known as Solubilized Extract of Soil Organic Matter or SESOM to gauge the behavior of the soil chassis organism *P. putida* in laboratory and field-like conditions.

In this thesis we examined one particular strain of *P. putida*, *P. putida* KT2440. KT2440 is a TOL plasmid-free strain, meaning KT2440 contains no plasmids and does not have the gene cluster to degrade toluene, unlike the version of *P. putida* mt-2 that was originally isolated in Japan in the 1960's (Nakazawa, 2003). KT2440 has no endogenous restriction enzymes and importantly does not act as a donor of introduced DNA (Timmis, 2002), and was certified host-vector biosafety (HV1) by the FDA (Kampers et al., 2019). Additionally, it has been previously demonstrated that it is possible to add synthetic fermentation and nitrate/nitrate respiration pathways to KT2440 which allow it to perform better under anoxic conditions, making it ideal for working in bioremediation sites (Weimer et al., 2020). Using *P. putida* KT2440 as our model

organism, we have applied next generation sequencing technology to better understand gene expression of *P. putida* in soil-like growth conditions.

Transcriptomics

In order to determine the changes in behavior of an organism, we can look at its transcriptome. The transcriptome is the collection of all the RNA, all of the transcribed material, of a cell. The transcriptome contains coding, messenger, and non-coding, ribosomal, and transfer RNAs. We are most interested in the messenger RNA, mRNA. mRNA encodes for genes and is translated into proteins; thus, mRNA plays a vital role in gene expression. The amount of mRNA is related to the level of expression of a given gene. This is especially useful in comparing organisms in different conditions. A gene that has more mRNA transcripts in a given condition than the control is said to be up-regulated in that condition, while a gene with fewer mRNA transcripts is said to be down-regulated. Before we can tell whether a gene is up-regulated or down-regulated we first need to quantify the number of transcripts that are produced by each gene. This process begins with sequencing the transcripts. There are two main categories of sequencing, long-read and short-read. In short read sequencing the material to be sequenced is broken down into smaller fragments which are sequenced individually, while in long read sequencing the material is sequenced as is. The standard short-read sequencing platform is Illumina, and the long-read platform that we have chosen to use is Oxford Nanopore.

Illumina

Illumina sequencing (Fig. 1) begins with RNA extraction and fragmentation. Then the RNA fragments are bound by random hexamer primers; these primers are then extended to finish the first strand cDNA synthesis. After the first strand has been synthesized the second strand is synthesized and the ends of the strands are repaired. After the strands are repaired, the strands are dA-tailed; this involves adding a non-template deoxyadenosine monophosphate, dAMP, to the blunt 3’ end of the cDNA. Next, adapters with a single “T” base overhang are ligated to the cDNA. After the adapters have been ligated, the second strand of the cDNA is digested. To complete the library preparation the amount of cDNA is increased by PCR amplification. Illumina is a form of sequencing by synthesis, which means that the sequence of the cDNA is determined by replicating the strand that is being sequenced. In Illumina’s case, the sequencing takes place on their flow cells. The cDNA library is washed over the flow cell, where it binds to

the inside of the flow cell channel. The flow cell channel is filled with sequences that are complementary to the adapters. The sequence will bridge over to these adapters and will be amplified; this ends up creating a cluster of the same cDNA molecule. The reading for the sequence is taken from these clusters. At each sequencing step, fluorescent nucleotides are added to the strands, the strands are then fluoresced, and the fluorescent signal is recorded, the sequence of fluorescent signals is translated into the sequence of the sample.

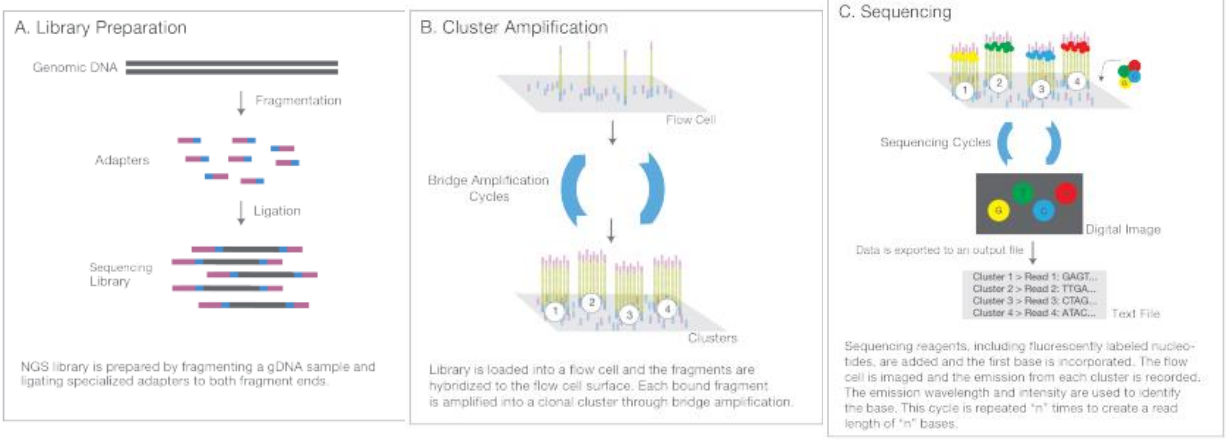

Figure 1, Illumina Sequencing: The steps to Illumina sequencing, from library preparation up through sequencing but excluding alignment and data analysis (Illumina, 2017).

One of the main drawbacks of Illumina sequencing is that it relies heavily on PCR, which allows for the introduction of errors and bias into the sequencing. Because the amount of starting material for transcriptome sequencing is so small, any errors or bias introduced prior to, or early into the PCR step will be highly amplified as numerous amplifications must be made to get the starting material to the requisite amount for sequencing. One way that bias can be introduced early on is with the fragmentation of the RNA. Analysis is performed under the assumption that the starting points of fragments relative to their original sequence are chosen approximately at random. However, positional bias can occur, so that the cDNA fragments are not uniformly distributed within the fragment that they represent, with fragments preferentially located towards the beginning or end of transcripts. This bias must be corrected for in the analysis (Roberts et al., 2011). Additionally, bias is introduced through the use of random hexamer priming to convert to

cDNA. The hexamer primers cause double stranded cDNA fragments with certain starting sequences to be over and underrepresented when compared with the frequency of their starting sequence in the transcriptome as a whole (Hansen et al., 2010).

Illumina has difficulty with extreme base compositions, namely regions that are AT or GC rich as well as poly(A) sequences. This difficulty arises from the PCR amplification stage, where the polymerase can slip on poly(A) sequences and AT repeats causing poor read quality (Aird et al., 2011) In addition, GC rich areas are strongly bound and may not dissociate. When they do dissociate, they can create a highly paired structure. These issues decrease the availability of GC-rich areas for amplification, which in turn reduces their overall expression. Further, the GC rich regions significantly impair reverse transcription itself, which also contributes to their overall under representation in the library (Sendler et al., 2011).

Oxford Nanopore

Oxford Nanopore is an alternative to traditional next generation sequencing methods. Nanopore sequencing is different primarily in that it works directly with the molecule to be sequenced and does not require PCR amplification. Nanopore instead works using a series of nanopores (Fig. 2). A motor protein attached to the sequence of interest passes the sequence through the channel of a nanopore (Oxford Nanopore, 2016b). As the sequence passes through the channel, 5 bases of the sequence block the flow of ions out from the inside of the membrane at a time (Lu et al., 2016). The blockage of the flow of ions results in a current that is measured across the membrane (Oxford Nanopore, 2016b). The current is run through a base calling algorithm which returns the human readable nucleotide form of the sequence (Oxford Nanopore, 2016b). In order to perform this process, two things need to be done to prepare the target RNA sequence (Oxford Nanopore, 2016a). First, it needs to have a motor protein attached to it. Then, it requires its complementary strand added to it. The complementary strand acts as a scaffold for the target sequence, ensuring that it does not form secondary structures and clog up the pores. To create these two features, adapters are ligated to the poly(A) tail of the sequence, and the complementary strand is extended and then the adapter containing the motor protein is ligated to the sequence (Oxford Nanopore, 2016a). A point of difficulty is readily evident here, as bacterial mRNA does not contain a poly(A) tail. This means that we will have to add a poly(A) tail to our

samples prior to sequencing (Oxford Nanopore, 2016a). Another problem with Nanopore sequencing is that the pores will clog after a certain number of sequencing steps (Kubota et al., 2019); this makes it important to minimize the amount of rRNA present, which can account for anywhere from 80-90% of the total RNA (O'Neil et al., 2013).

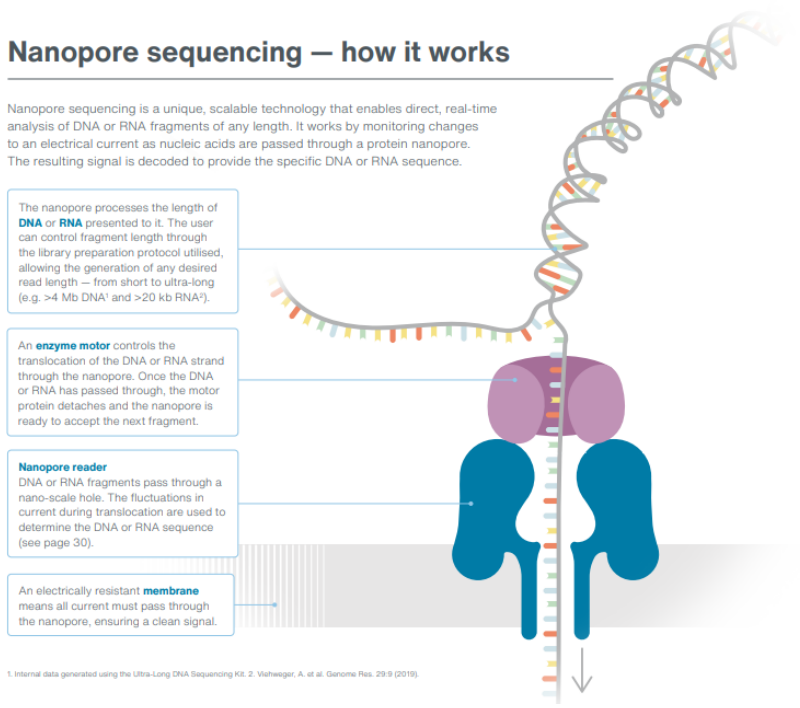

Figure 2, Nanopore Sequencing: Diagram of the active portion of Nanopore sequencing process (Oxford Nanopore, 2023).

Project Objectives

P. putida is an important soil-derived synthetic biology chassis organism, which is useful for soil bioremediation. However, in order to fully harness this organism for soil applications, we require a thorough understanding of its gene expression under soil conditions. Comparative transcriptomic analysis of *P. putida* grown under laboratory versus soil-like (SESOM) conditions has not previously been performed. Our goal is to apply both long-read and short-read next generation sequencing to compare in the transcriptomes of *P. putida* grown in SESOM and LB in both their growth and stationary phases. To that end, we prepared *P. putida* cultures grown in both SESOM and LB, then extracted and purified RNA and prepared libraries for either Illumina or Oxford Nanopore sequencing. Our results highlight the most differentially expressed genes between the two growth environments. This information will be useful in future studies to engineer *P. putida* for soil bioremediation applications.

Methods

Solubilized Extract of Soil Organic Material (SESOM)

SESOM was made in the lab using MiracleGro Performance Organics All Purpose Container Mix soil. 100 g of soil was combined with 500 ml of 10X phosphate-buffered saline (PBS) in an Erlenmeyer flask. The mixture was placed in a 37°C incubated shaker for 4 hours. The mixture was removed from the shaker and decanted into a French press. The mixture was pressed and the liquid, the solubilized extract of soil organic material (SESOM), transferred to a new container. A filtering apparatus was prepared by placing a piece of 11 µm Whatman paper under a piece of 24 µm Whatman paper in a Buchner funnel. The funnel was placed into a filtration flask with a rubber stopper. The filter paper was saturated around the edges with PBS to ensure they were sealed and SESOM was then added to the filtering apparatus. The SESOM was vacuum filtered overnight. The filtered SESOM was then sterile filtered with 0.2 µm vacuum filter.

Bacterial Samples

LB Growth

10 mL culture tubes were filled with 5 mL of LB liquid media. The tubes were inoculated with cultures of *P. putida* from glycerol stock. The cultures were incubated at 30°C overnight on a shaking incubator. After culture growth, the culture was back diluted by removing 0.5 mL of the overnight culture and placing it into a new 10 mL culture tube with 5 mL of LB. The back dilution was placed in a 30°C shaker for 90 minutes. After 90 minutes the absorbance was measured in a spectrophotometer at 260 nm. The desired optical density (OD) is around 0.5 (mid log phase). If the back dilution was not around this desired range, it was placed back into the incubator and checked every 30 minutes until it was around the desired range. Once at the desired range, 1 mL samples were taken and placed into 1.5 mL microcentrifuge tubes. The samples were centrifuged for 10 minutes at 8,000 rcf in order to pellet the cells. The supernatant was removed by pipetting or decanting, and the pellets were stored at -80°C until they were used.

LB Stationary

10 mL culture tubes were filled with 5 mL of LB liquid media. The tubes were inoculated with cultures of *P. putida* from glycerol stock. The cultures were incubated at 30°C overnight on a shaking incubator. 1 mL samples were taken and placed into 1.5 mL microcentrifuge tubes. The samples were centrifuged for 10 minutes at 8,000 rcf in order to pellet the cells. The supernatant was removed by pipetting or decanting, and the pellets were stored at -80°C until they were used.

SESOM Growth

10 mL culture tubes were filled with 5 mL of LB liquid media. The tubes were inoculated with cultures of *P. putida* from glycerol stock. The cultures were incubated at 30°C overnight on a shaking incubator. After culture growth, the culture was back diluted by removing 0.5 mL of the overnight culture and placing it into a new 10 mL culture tube with 5 mL of LB. The back dilution was placed in a 30°C shaker for 90 minutes. After 90 minutes the absorbance was measured in a spectrophotometer at 260 nm. The desired optical density (OD) is around 0.5 (mid log phase). If the back dilution was not around this desired range, it was placed back into the incubator and checked every 30 minutes until it was around the desired range. Once at the desired range, 1 mL samples were taken and placed into 1.5 mL microcentrifuge tubes. The samples were centrifuged for 10 minutes at 8,000 rcf in order to pellet the cells. The supernatant was removed by pipetting or decanting, and 1 mL of SESOM was added to the pelleted samples. The pellets were then resuspended and incubated for 2 hours with their caps partially open in a 30°C incubator. After incubating, the samples were centrifuged at 8,000 rcf until a pellet formed, around 10-30 minutes. The supernatant was removed by pipetting or decanting, and the pellets were stored at -80°C until they were used.

SESOM Stationary

10 mL culture tubes were filled with 5 mL of LB liquid media. The tubes were inoculated with cultures of *P. putida* from glycerol stock. The cultures were incubated at 30°C overnight on a shaking incubator. 1 mL samples were taken and placed into 1.5 mL microcentrifuge tubes. The samples were centrifuged for 10 minutes at 8,000 rcf in order to pellet the cells. The supernatant was removed by pipetting or decanting, and 1 mL of SESOM was added to the pelleted samples. The pellets were then resuspended and incubated for 2 hours with their caps partially open in a 30°C incubator. After incubating, the samples were centrifuged at 8,000 rcf until a pellet formed,

around 10-30 minutes. The supernatant was removed by pipetting or decanting, and the pellets were stored at -80°C until they were used.

RNA extraction with TRIzol

Adapted from *TRIzol reagent user guide* (Thermo Fisher Scientific, 2016). The sample to be extracted was removed from storage at -80°C and was allowed to come to room temperature. 1 mL of TRIzol was added to the sample and the sample was incubated for 5 minutes at room temperature. After incubation, 0.2 mL of chloroform was added to the sample. The sample was then shaken vigorously for 15 seconds and incubated at room temperature for another 2-3 minutes. The sample was then centrifuged for 15 minutes at 12,000 rcf at 4°C. After centrifuging, 400 µl of the colorless aqueous phase was transferred to a new tube, being especially careful not to touch the interphase when pipetting, the leftover sample (the interphase and organic phase) was discarded. 10 µg of RNase free glycogen was added to the sample, after which 0.5 mL of Isopropanol was added, then the sample was inverted 5 times or until the clear filament like lines disappeared. The sample was incubated for 10 minutes at 4°C and then centrifuged for 10 minutes at 12,000 rcf at 4°C. The supernatant was then discarded with a micropipette and the pellet was resuspended in 1 mL of 75% ethanol. The sample was vortexed briefly and centrifuged for 5 minutes at 7,500 rcf at 4°C. The supernatant was discarded with a micropipette and the pellet was air dried for 5-10 minutes (until the pellet changed from white to glassy). After drying the pellet was resuspended in 1 mL of 75% ethanol and vortexed briefly. The sample was then centrifuged for 5 minutes at 7,500 rcf at 4°C. The supernatant was discarded with a micropipette and the pellet was air dried for 5-10 minutes (until the pellet changed from white to glass). The pellet was then resuspended in 20 µl of RNase free water and incubated for 10 minutes at 55-60°C. The extracted RNA was stored at -80°C. The quality of the extraction was verified by gel electrophoresis. 0.5 µl of extracted RNA was combined with 3.5 µl of water and 1 µl of 5X loading dye and was run in a 1% agarose gel with Tris-acetate-EDTA (TAE) buffer at 100 v for 30 minutes.

Poly(A) Tailing

Adapted from *Poly(A) Tailing of RNA using E. coli Poly(A) Polymerase* (New England BioLabs, n.d.-a). The RNA samples were removed from the -80°C storage and allowed to come to room temperature. The following components were combined in the order specified: 12 µl of sample

RNA, 2.5 µl of RNase free water, 0.5 µl of RNase inhibitor. 2 µl of 10X *E. coli* Poly(A) Polymerase Reaction Buffer, 2 µl of 10 mM ATP, and 1 µl of *E. coli* Poly(A) Polymerase. The combined solution was incubated at 37°C for 30 minutes, then proceeded directly to clean up.

Poly(A) Tailing Cleanup

Adapted from *Instructions For Use / RNAClean XP / In Vitro Produced / RNA and cDNA Purification* (Beckman Coulter, 2020). RNAClean XP beads were vortexed, then 36 µl of the beads were added to the Poly(A) tailed sample. The sample was mixed by pipetting up and down 15 times, then the sample was incubated for 15 minutes on ice. The tube containing the sample was placed on a magnetic rack in order to separate the beads from the supernatant (approximately 5 minutes). After the beads separated from the supernatant the supernatant was carefully removed and discarded using a micropipette. Freshly prepared 80% alcohol was then added to the tube (enough to cover the beads) and was incubated for 30 seconds at room temperature prior to removing the supernatant using a micropipette. The ethanol wash was repeated once for a total of two washes. After completely removing the ethanol the second time, the tube was allowed to air dry for 5 minutes with the lid open. The tube was removed from the rack and the RNA was eluted by adding 16 µl of nuclease free water. The solution was mixed by pipetting up and down 10 times and was briefly spun down. The solution was incubated for 2 minutes at room temperature and then placed on the magnetic rack until the beads separated from the supernatant (approximately 2 minutes). 15 µl of the supernatant were then removed and stored at -80°C until proceeding to the next step.

rRNA depletion

Adapted from *Protocol for rRNA depletion using NEBNext rRNA Depletion Kit* (New England BioLabs, n.d.-b) and *Instructions For Use / RNAClean XP / In Vitro Produced / RNA and cDNA Purification* (Beckman Coulter, 2020). Poly(A) tailed samples were removed from storage and allowed to come to room temperature. 5 µl of sample was combined with 6 µl of nuclease free water and kept on ice. The following were combined on ice to make the hybridization reaction: 11 µl of sample and water, 2 µl of NEBNext Bacterial rRNA depletion solution, 2 µl of NEBNext probe hybridization buffer. The hybridization reaction was mixed by pipetting up and down 10 times and was briefly centrifuged. The tube was then placed in a thermocycler and the following program was run with the heated lit set to 105°C: 95°C for 2 minutes, ramp down to

22°C at a rate of 0.1°C/sec, 22°C for 5 minutes. The tube was briefly centrifuged and placed on ice. The following were combined on ice to make the RNase digest reaction: 15 µl of hybridized RNA, 2 µl of RNase H reaction buffer, µl of NEBNext thermostable RNase H, 1 µl of nuclease free water. The RNase digest reaction was mixed by pipetting up and down 10 times and was briefly centrifuged. The tube was incubated in a thermocycler for 30 minutes at 50°C with the lid set to 55°C. The tube was then briefly centrifuged and placed on ice. The following were combined on ice to make the DNase digest reaction: 20 µl of RNase digested RNA, 5 µl of DNase 1 reaction buffer, 2.5 µl of NEBNext DNase 1, 22.5 µl of nuclease free water. The DNase digest reaction was mixed by pipetting up and down 10 times and was then briefly centrifuged. The tube was incubated in a thermocycler for 30 minutes at 37°C with the lid set to 40°C. The tube was then briefly centrifuged and placed on ice.

RNAClean XP beads were vortexed, then 90 µl of the beads were added to the DNase digested RNA. The sample was mixed by pipetting up and down 15 times, then the sample was incubated for 15 minutes on ice. The tube containing the sample was placed on a magnetic rack in order to separate the beads from the supernatant (approximately 5 minutes). After the beads separated from the supernatant the supernatant was carefully removed and discarded using a micropipette. 200 µl of freshly prepared 80% alcohol was then added to the tube and was incubated for 30 seconds at room temperature prior to removing the supernatant using a micropipette. The ethanol wash was repeated once for a total of two washes. After completely removing the ethanol the second time, the tube was allowed to air dry for 5 minutes with the lid open. The tube was removed from the rack and the RNA was eluted by adding 7 µl of nuclease free water. The solution was mixed by pipetting up and down 10 times and was briefly spun down. The solution was incubated for 2 minutes at room temperature and then placed on the magnetic rack until the beads separated from the supernatant (approximately 2 minutes). 5 µl of the supernatant were then removed and stored at -80°C until proceeding to the next step.

Nanopore Library Preparation

The rRNA depleted samples were thawed at room temperature. 5 µl of the sample was combined with 4 µl of nuclease free water. The sample was mixed by flicking and then briefly spun down. The following were combined in a 0.2 ml thin-walled PCR tube in the following order: 3 µl of NEBNext quick ligation reaction buffer, 9 µl of RNA, 0.5 µl of RNA Calibrant Strand, 1 µl of

RT adapter, 1.5 µl mix the reaction by pipetting and spin down then incubate for 10 minutes at room temperature. The reverse transcription master mix was made by mixing 9 µl of nuclease free water, 2 µl of 10 mM dNTPs, 8 µl of 5x first strand buffer and 4 µl of 0.1 M DTT. The master mix was added to the 0.2 ml thin-walled PCR tube containing the RT adapter-ligated RNA and was mixed by pipetting. 2 µl of SuperScript III reverse transcriptase was added to the reaction, the reaction was mixed by pipetting. The reaction was placed in a thermal cycler and incubated at 50°C for 50 minutes, then for 10 minutes at 70°C before being brought to 4°C before proceeding. The sample was transferred to a 1.5 ml DNA LoBind tube. RNAClean XP beads were vortexed, then 72 µl of the beads were added to the reverse transcription reaction and were mixed by pipetting. The reaction was incubated on a rotary mixer for 5 minutes at room temperature then 200 µl of freshly prepared 70% ethanol were added to it. After adding the ethanol, the sample was spun down and placed on a magnetic rack in order to separate the beads from the supernatant. After the beads separated from the supernatant, the supernatant was pipetted off and discarded. Keeping the tube on the magnetic rack, 150 µl of freshly prepared 70% ethanol was added to the tube. To wash the beads the tube was spun 180° on the rack and was left until the beads migrated and pelleted on the other side, the tube was then rotated back to the starting position and the beads were allowed to pellet again. The supernatant was then removed with a pipette, the tube was spun down and the paced back on the rack and the supernatant was removed again. The tube was removed from the magnetic rack and the pellet was resuspended in 20 µl of water. The sample was incubated at room temperature for 5 minutes and then placed back on the magnetic rack. 20 µl of eluate was removed and placed in a 1.5 ml DNA LoBind tube. In the DNA LoBind tube, the following were added in order: 8 µl of NEBNext quick ligation reaction buffer, 6 µl of RNA Adapter, 3 µl of nuclease-free water, 3 µl of T4 DNA ligase. The reaction was mixed by pipetting and incubated at room temperature for 10 minutes. 16 µl of RNAClean XP beads were added to the adapter ligation reaction and were mixed by pipetting. The reaction was incubated on a rotary mixer for 5 minutes at room temperature and then was spun down and placed on a magnetic rack. The supernatant was removed and 150 µl of wash buffer was added to the beads. The tube was removed from the magnetic rack and the beads were resuspended by gently flicking. The tube was then returned to the magnetic rack and the beads were allowed to pellet before pipetting off the supernatant. The adding and removing of the wash buffer was repeated, then the tube was removed from the

magnetic rack and the beads were resuspended in 21 µl of elution buffer by gently flicking the tube. The reaction was incubated for 10 minutes at room temperature and was then placed back on the magnetic rack to pellet the beads. After the beads were pelleted 21 µl of eluate was removed into a 1.5 DNA LoBind tube.

Bioinformatics

Illumina

Figure 3 details the bioinformatics pipeline for processing the Illumina data. Library preparations and sequencing were performed by Azenta Life Sciences (South Plainfield, NJ) on the HiSeq platform, and fastq files were generated. Then FastQC (Andrews, 2010) was run on the files to get information about their quality. The information from FastQC was used to run Cutadapt (Martin, 2011) to clean up the fastq files, giving us quality controlled reads. The *P. putida KT2440* genome was downloaded from the National Center for Biotechnology Information as a fasta file and used Bowtie2 (Langmead & Salzberg, 2012) to create an alignment index. The alignment index was used with Bowtie 2 to turn the quality-controlled reads into aligned reads. The software used to assemble the transcripts requires that reads are sorted by position along the chromosome, so SAMtools (Danecek et al., 2021) was used to sort the aligned reads. The sorted aligned reads were then analyzed by StringTie (Pertea et al., 2015) to merge the transcripts and create read coverage tables for use with DESeq2 (Love et al., 2014), which was used for differential expression analysis.

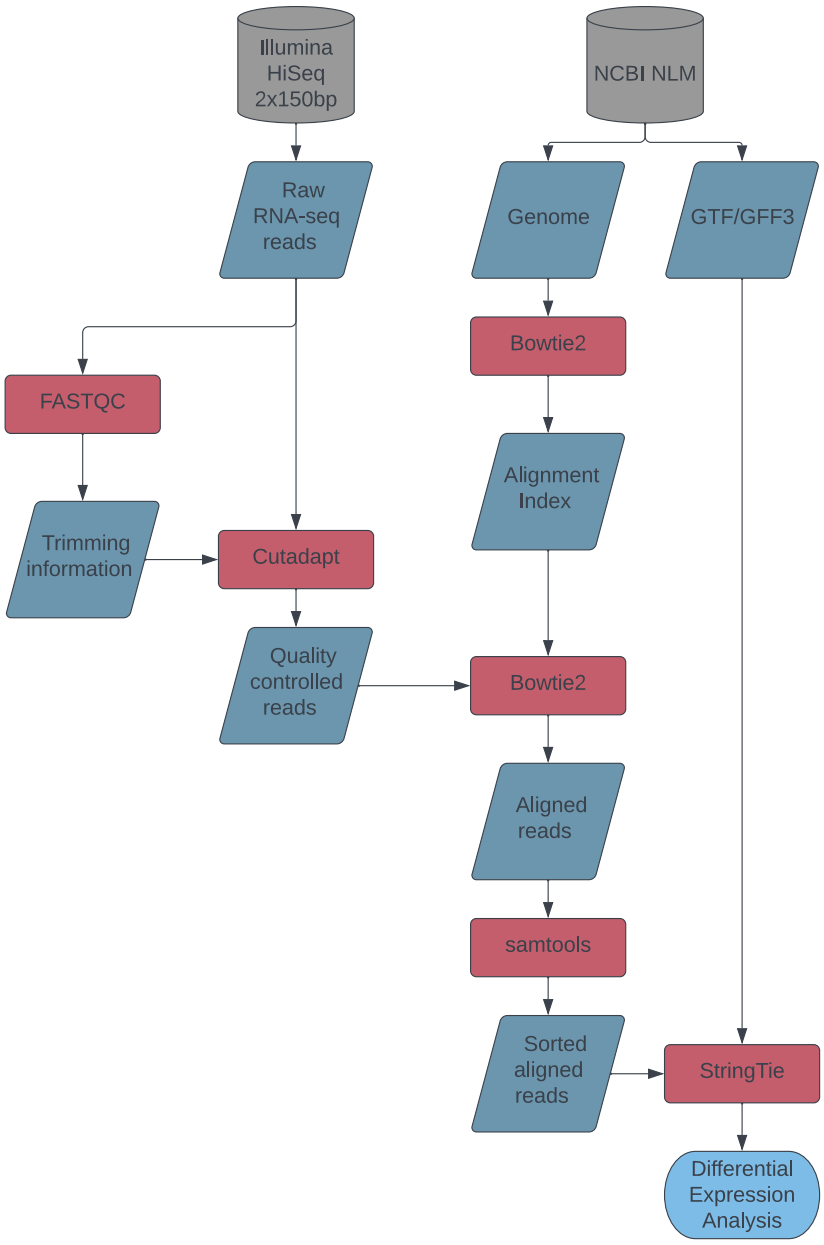

Figure 3, Illumina Pipeline: Pipeline showing the methods used to analyze the Illumina data. Data origins are shown in grey; data are shown in blue, and programs are shown in red.

Nanopore

A similar method to the Illumina pipeline was used to analyze the Nanopore data, which is shown in Figure 4. The pipeline begins with Fast5 files, which are electrical signals output by the Nanopore. The Fast5 files were converted into reads by a the Nanopore basecaller, Guppy, and was run with the high-accuracy split settings (Oxford Nanopore, 2016b). We then aligned the basecalled reads using minimap2 (Li, 2018), an aligner that is specific for long-read data. Once again, SAMtools was used to sort the aligned reads and then StringTie was used in its long-read mode to merge the transcripts and create read coverage tables for use with DESeq2.

StringTie

The pipeline that we used specifically for StringTie is seen in Figure 5. Using a GFF3 (gene feature format) file that was retrieved from the NCBI website, the GFF3 file and the sorted aligned reads were used to create assembled transcripts. When assembling the transcripts from the Nanopore data, StringTie was run in its special long-read mode. We then used StringTie to merge the transcripts and re-ran StringTie using the merged transcripts as the basis rather than the GFF3 file. This allowed us to account for differences in sample sizes in the aligned transcripts.

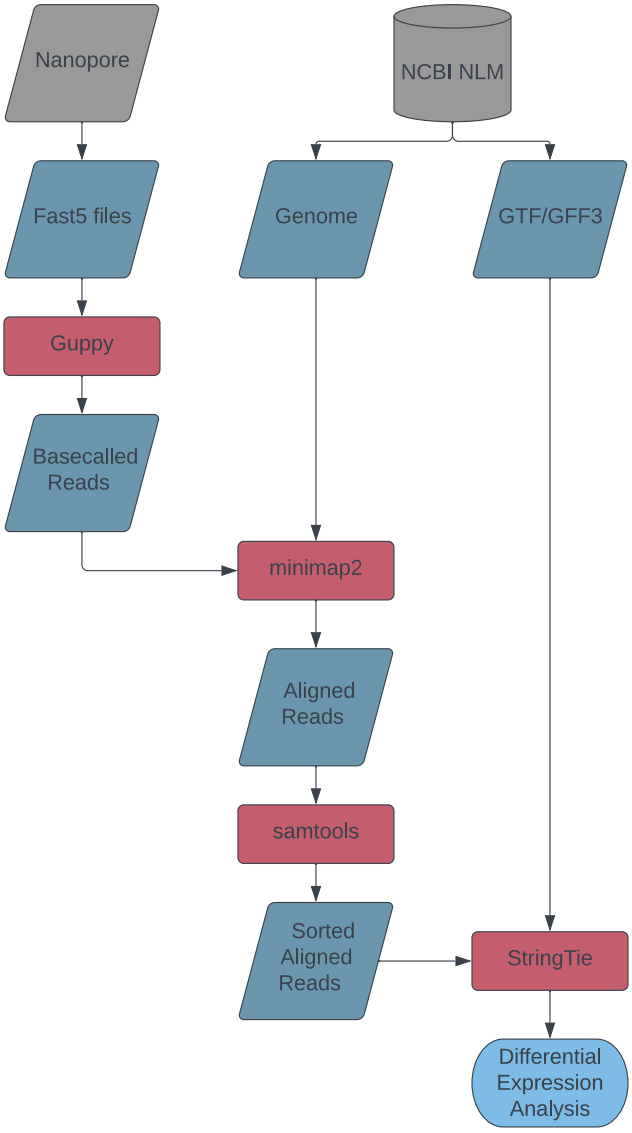

Figure 4, Nanopore Pipeline: Pipeline showing the methods used to analyze the Nanopore data. Data origins are shown in grey; data are shown in blue, and programs are shown in red.

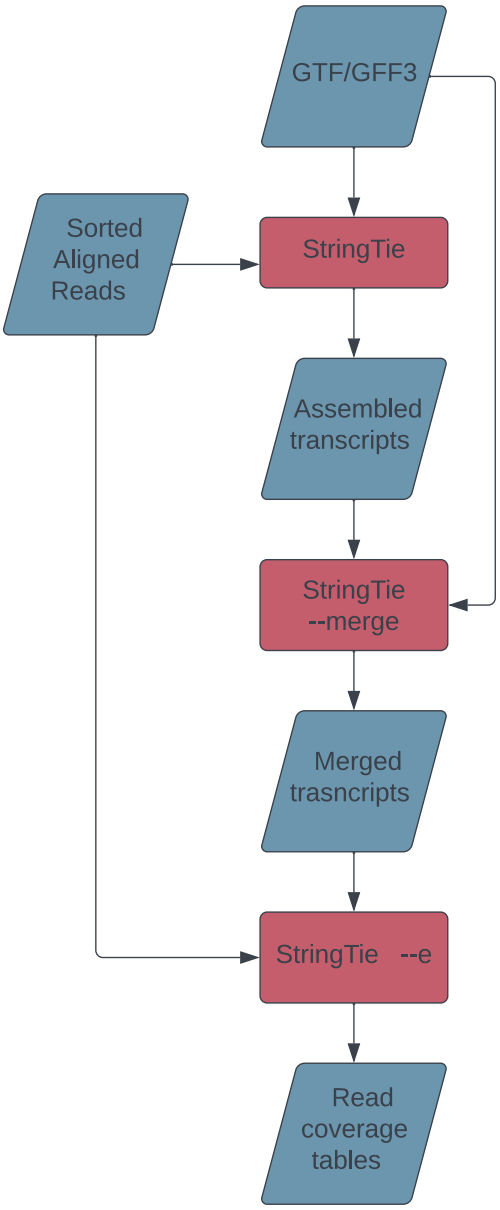

Figure 5, StringTie Pipeline: Pipeline showing the methods used in the StringTie step of the pipelines. Data are shown in blue, and programs are shown in red.

Results

Sample Generation Results and Analysis

We isolated RNA from *P. putida* in four conditions. Grown in LB in stationary and growth phase; and grown in SESOM in stationary and growth phase. Ultimately, twelve samples (three replicates of each of the four conditions) were sent for Illumina RNA-seq, and ten samples (two or three replicates of each condition), were processed for Nanopore sequencing (Table 1). The number of samples processed for Nanopore sequencing was limited by the number of flow cells available (two) and the minimum target read depth.

Table 1, Transcriptomics Samples

Transcriptomics Samples					
Illumina			Nanopore		
Sample #	Media	Phase	Sample #	Media	Phase
1	LB	Stationary			
2	LB	Stationary	2	SESOM	Growth
3	LB	Stationary	3	SESOM	Growth
4	LB	Growth	4	SESOM	Stationary
5	LB	Growth	5	SESOM	Stationary
6	LB	Growth	6	SESOM	Stationary
7	SESOM	Stationary	7	LB	Growth
8	SESOM	Stationary	8	LB	Growth
9	SESOM	Stationary			
10	SESOM	Growth	10	LB	Stationary
11	SESOM	Growth	11	LB	Stationary
12	SESOM	Growth	12	LB	Stationary

Table of the transcriptomic samples that were isolated from P. putida under the four conditions given as well as the method that was used to sequence each sample.

Choice of extraction protocol

The first step in the process of sample generation is RNA extraction. We tested both spin column purification (Qiagen RNeasy) and TRIzol reagent to extract total RNA from *P. putida* KT2440 cultures grown in LB or SESOM, to log or stationary phases, as described in the methods. Genomic DNA (gDNA) contamination of RNA can lead to amplification artifacts in the Illumina library preps, and therefore must be avoided. As shown in Figure 6, the TRIzol extraction that we performed was successful, with no gDNA present in the extracted sample

(Fig. 6, right image). This is in direct contrast to the RNA extraction column where there was often visible gDNA when running the unamplified sample (Fig. 6, left image). When we ran the sample in a PCR reaction there was still no visible gDNA contamination, which validated our choice of the TRIzol method. A lack of visible gDNA after PCR meant that we did not have to perform a DNase digest and an additional cleanup step which would have caused us to lose additional RNA.

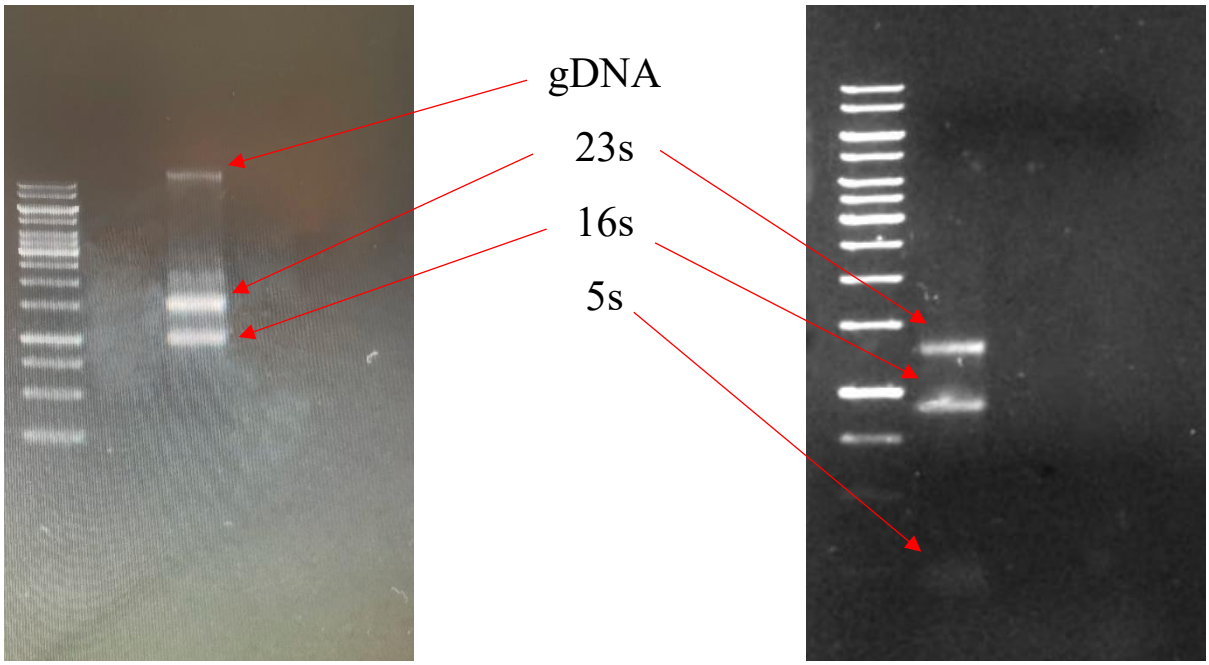

Figure 6, RNA extraction from P. putida cultures: Gel electrophoresis of RNA isolated using spin column purification (left, Qiagen RNeasy) and after TRIzol extraction (right). Cultures were grown in LB for both cases shown (see Methods). rRNA 23s, 16s, and 5s bands and contaminating DNA bands (gDNA) are shown. Marker bands are, in decreasing bp length, 10000, 8000, 6000, 5000, 4000, 3500, 3000, 2500, 2000, 1500, 1000, 750, 500, 250.

Another advantage of the TRIzol that we discovered after successfully performing the extraction was that it allowed us to capture more breadth of the RNA spectrum. We were able to visibly identify the band of the 5s rRNA subunit after the TRIzol extraction, while the lowest molecular weight band visible after performing an RNA extraction column was the 16s rRNA subunit, implying that we were able to capture smaller RNAs with the TRIzol extraction that with the RNA extraction column. TRIzol extractions were sent without further processing for

Illumina sequencing or continued through polyadenylation and rRNA depletion for Oxford Nanopore sequencing.

Preparation for Nanopore sequencing – RNA Polyadenylation

The poly(A) tail of mRNAs is used by the nanopores to capture the mRNA for nanopore sequencing. As bacterial mRNAs are not polyadenylated, we used poly(A) polymerase to add a poly(A) tail to TRIzol-extracted RNA. Through the use of gel electrophoresis to visualize the rRNA (Fig. 7), we were able to verify that we added around 100 bases of adenosine to the rRNAs during the polyadenylation process. We assume that mRNAs, which are not directly visible by this method, were similarly polyadenylated. Per the nanopore protocol, we only needed to add around 10 base pairs. However, we were limited by the resolution of the gel electrophoresis that we used to verify the polyadenylation, so the exact length of tails added is not known.

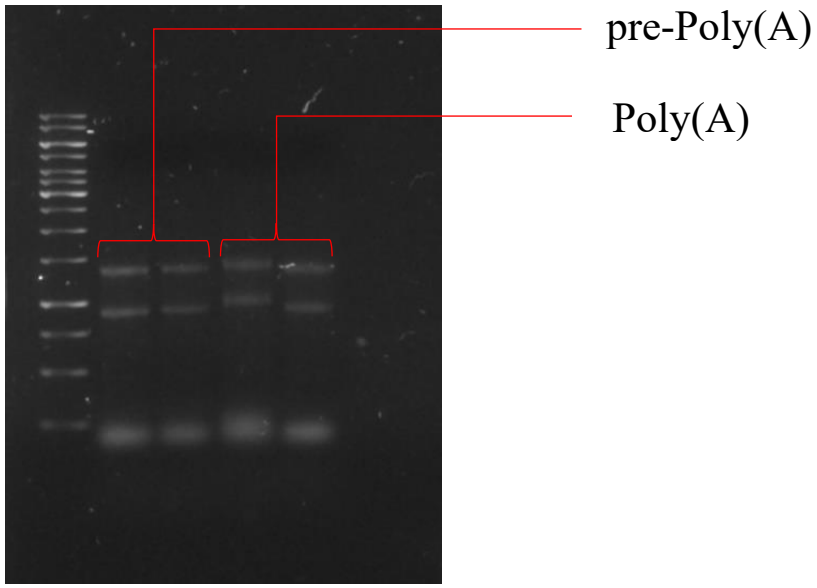

Figure 7, polyadenylation of extracted P. putida RNA: Gel electrophoresis of RNA polyadenylated using E. coli Poly(A) Polymerase. The pair of bands on the left is pre-polyadenylation rRNA, and the pair of bands on the right are polyadenylated rRNA. The shift in the bands in the third lane indicates a successful polyadenylation, the polyadenylation of the fourth lane is indeterminate due to the resolution of instrumentation at hand. Marker bands are, in decreasing bp length, 10000, 8000, 6000, 5000, 4000, 3500, 3000, 2500, 2000, 1500, 1000, 750, 500, 250.

Preparation for Nanopore sequencing -- rRNA depletion

Ribosomal RNAs (rRNAs) comprise the vast majority of mRNA in most cells. In eukaryotes, mRNAs are polyadenylated and rRNAs are not, permitting these populations to be separated by the poly(A) tails. However, in prokaryotes, neither RNA is polyadenylated. Thus, in order to enrich for mRNAs and improve the read depth, rRNAs must be depleted using an RNAse-H-based targeting method. The rRNA depletion visually appeared to be successful, as indicated by the lack of rRNA bands in the post depletion gel (Fig. 8). However, after using the expected concentration of the depleted samples to estimate depletion efficiency, it was apparent that the depletion was not completely successful.

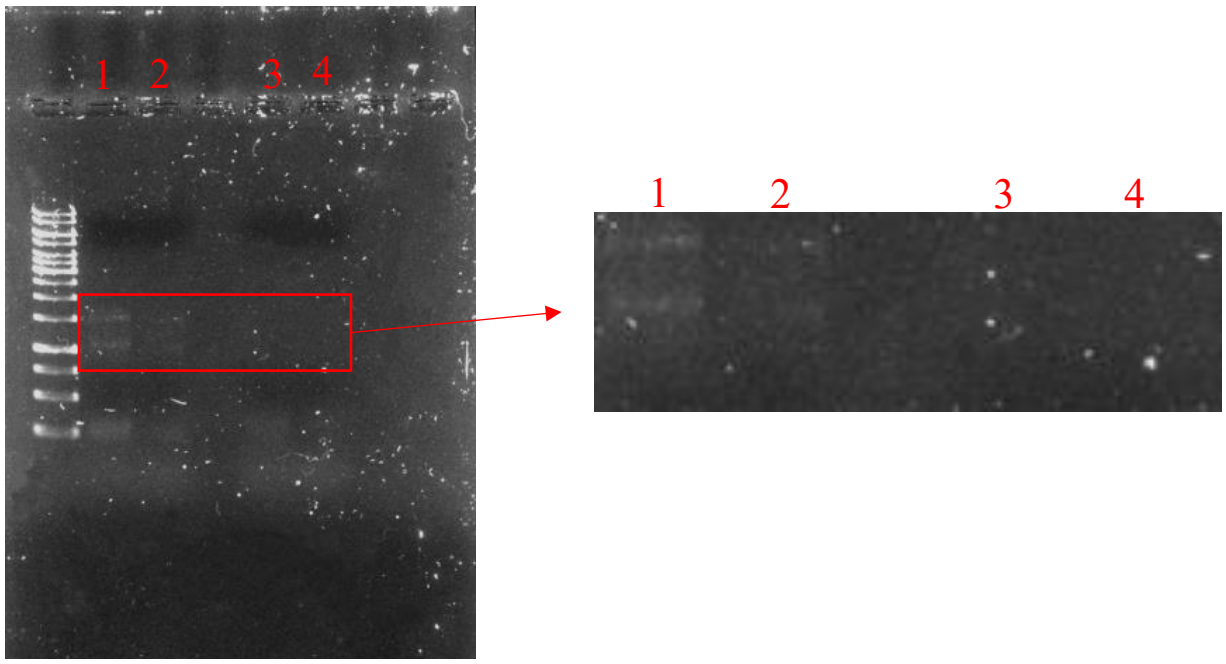

Figure 8, depletion of P. putida polyadenylated rRNA: Lanes 1 and 2 contain samples prior to rRNA depletion, and lanes 3 and 4 contain the samples from 1 and 2 after depletion using NEBNext rRNA Depletion Kit. The lack of bands in lanes 3 and 4 indicate that depletion was successful. Marker bands are, in decreasing bp length, 10000, 8000, 6000, 5000, 4000, 3500, 3000, 2500, 2000, 1500, 1000, 750, 500, 250.

By taking concentration of the initial polyadenylated RNA we were able to determine the amount of poly(A) RNA and estimate the amount of RNA that should remain after the depletion process, as seen in Table 2. For this estimation we assumed that ribosomal RNA makes up

around 80 to 90 percent of the total RNA; specifically, we chose to assume it made up 85 percent of total RNA. We used the concentration of the depleted poly(A) RNA to determine the amount of poly(A) RNA remaining post depletion and compared it to our estimate. We determined that at best we were able to deplete around only 55 percent of the Poly(A) RNA and at worst only 24 percent.

Table 2, rRNA Depletion was Incomplete

rRNA Depletion						
Condition	Phase	Sample	Initial Amount (ng)	Expected Final Amount (ng)	Final Amount (ng)	Percent Depleted
SESOM	Growth	2	567.5	85.125	207	41%
		5	782	117.3	213	55%
		8	718	107.7	220	49%
	Stationary	9	417	62.55	161	39%
		10	573	85.95	168	51%
		11	871	130.65	220	59%
LB	Growth	2	662	99.3	222	45%
		4	1317.5	197.625	741	27%
		8	1342.5	201.375	655	31%
	Stationary	2	656.5	98.475	408	24%
		6	882.5	132.375	466	28%
		10	1427	214.05	827	26%

The starting amount of polyadenylated RNA, as well as the actual vs. expected amount of poly(A) RNA after performing the rRNA depletion on all samples.

Bioinformatics Results

Sample clustering across all conditions shows strong reproducibility from Illumina but not Nanopore data

It is helpful to be able to determine how similar samples are to each other, especially since so much of our analysis depends on comparing various expressions of genes in different sample conditions. One way to determine how similar samples are to each other is to calculate the Euclidean distance between each of the samples. Once the distances have been calculated they can be plotted as a heat map, with the darker colors indicating that any two given samples are closer to each other. This can help give us some indication of what we might expect from our differential expression results.

Figure 9 shows the clustering of samples by Euclidian distance for the Illumina samples. As expected, experimental replicates of all conditions have the closest clustering. We further observe from this analysis that the grown medium (LB versus SESOM) defines these relationships more strongly than the growth phase (log versus exponential).

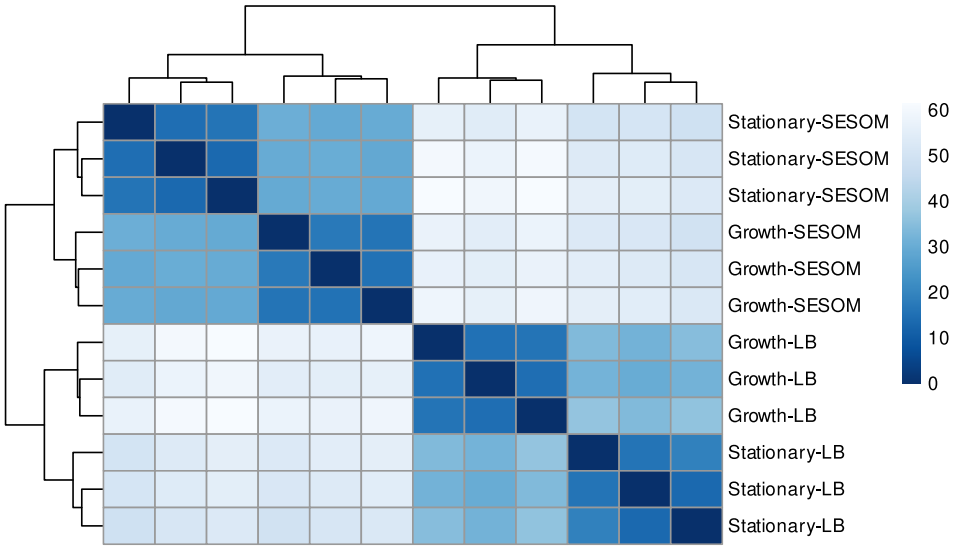

Figure 9, Illumina all conditions sample clustering: Heatmap of the Euclidian distance between each sample. The data comes from the Illumina sequencing results.

Figure 10 shows the clustering of samples by Euclidian distance for the Nanopore samples. Here we observe no strong relationship between replicates of all conditions and see no indication that any specific condition or set of conditions strongly defines the relationship between the samples.

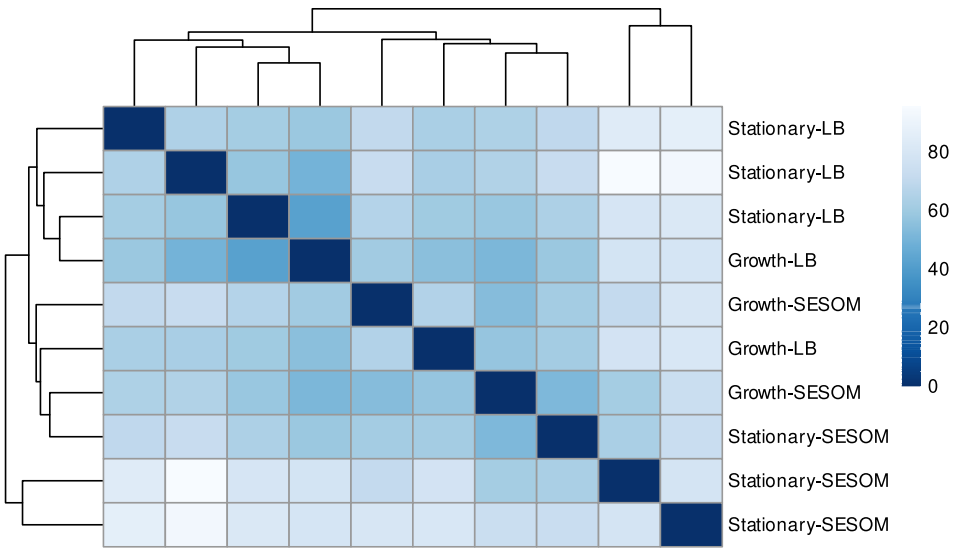

Figure 10, Nanopore all conditions sample clustering: Heatmap of the Euclidian distance between each sample. The data comes from the Nanopore sequencing results.

Gene clustering across all conditions shows strong reproducibility from Illumina but not Nanopore data

Another way to determine whether samples are showing good reproducibility by condition is to look at the variance in gene expression. When we look at gene variance by clustering, we do not take into account the conditions, the phase or the media, of the different samples; we are not looking at specific changes in expression or degree of expression between conditions, rather we look at the variance across conditions. We look at the amount by which a gene deviates in a specific sample from the gene's average across all samples. In Figure 11 we look at the 20 most variable genes, ordered top to bottom most to least variable, where the blue colors indicates that the expression of the gene in that sample is lower than the average expression of the gene and the red indicates the opposite.

Figure 11 shows gene variance between conditions in the Illumina samples and further points to the observation that the growth medium (LB vs SESOM) defines the relationship in the samples as there is a clear delineation between the variance in each condition.

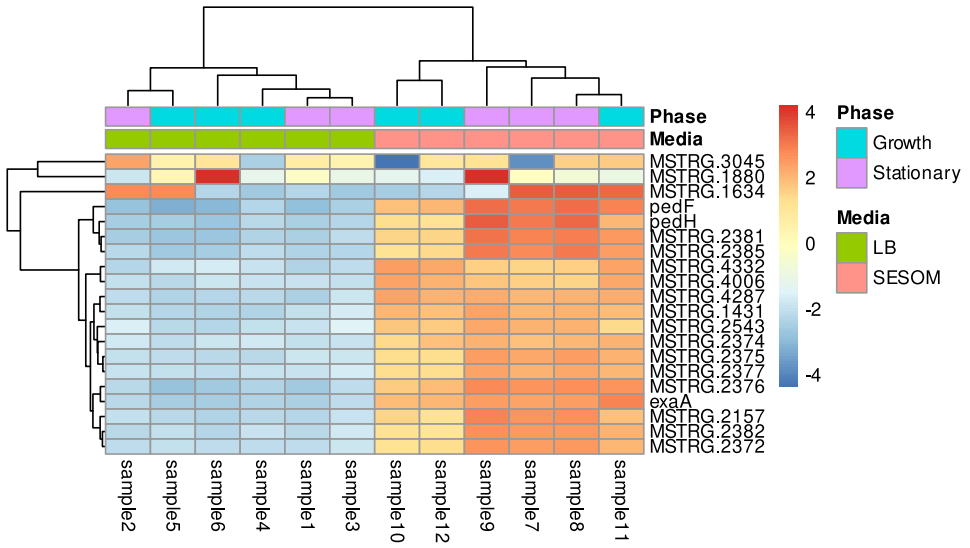

Figure 11, Illumina all conditions gene variance: Heatmap of the difference in expression of genes as measured by the number of transcripts. The data comes from the Illumina sequencing results. The samples correspond to the samples in Table 1. Blue colors indicates that the expression of the gene in that sample is lower than the average expression of the gene and the red indicates the opposite.

Figure 12 shows gene variance in the Nanopore samples and here the apparent lack of relationship between variance and condition is apparent by the lack of discernable pattern in the heatmap, especially when compared to Figure 11 and the Illumina samples.

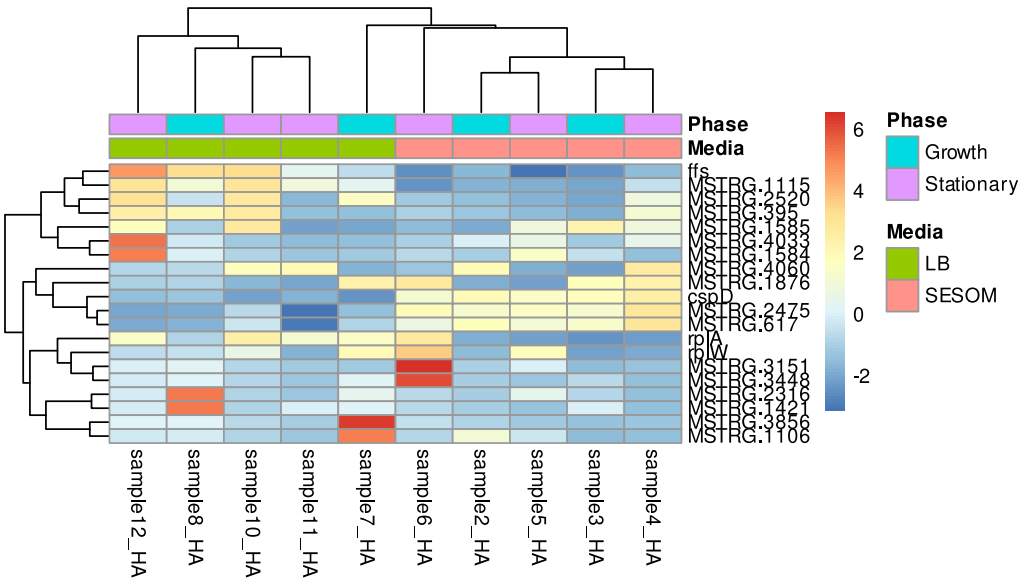

Figure 12, Nanopore all conditions gene variance: Heatmap of the difference in expression of genes as measured by the number of transcripts. The data comes from the nanopore sequencing results. The samples correspond to the samples in Table 1. Blue colors indicates that the expression of the gene in that sample is lower than the average expression of the gene and the red indicates the opposite.

Principal Component Analysis of All Conditions confirmed that samples were separable by medium type

In a principal component analysis (PCA), the samples are projected onto a 2D plane and spread out in two directions that capture as much of the variance of the samples as possible. The first principle component, the horizontal axis, captures the most variance of the samples, while the second principle component, the vertical axis, captures the second most variance of the samples. The sum of the variances do not add to 100% as there are other causes of variances (other directions that are not plotted), but these other causes are less important as they do not capture as much of the variance as the main two.

The PCA plots of the Illumina data (Fig. 13, left graph) reveal that the majority of the variance in the samples is attributable to the growth medium (LB vs SESOM) and that this accounts for 77% of the variance. This observation is consistent with the Euclidian clustering in Figure 9. An additional 9% of the variance is attributable to the growth phase (log versus stationary).

In contrast, the PCA plots of the Nanopore data (Fig. 13, right graph) reveal that the majority of the variance in the samples is again still attributable to the growth medium, however this only accounts for 37% of the variance and the samples are not as consistently separated. This observation is also consistent with the Euclidian clustering in Figure 10, namely the lack of discernable clustering. An additional 15% of the variance is attributable to some other factor, possibly the growth phase, but the separation is not clear enough to make a distinction.

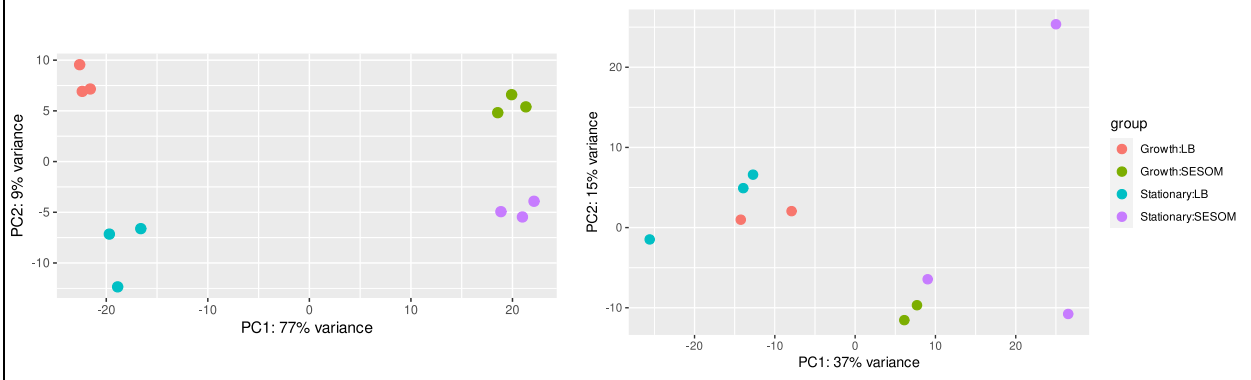

Figure 13, All conditions PCA: Principal Component Analysis of the Illumina and Nanopore data. The plot on the left is the Illumina data, and the plot on the right is the Nanopore Data.

Focusing on analysis of Stationary Phase samples

After looking at the all condition data we decided to focus on just the stationary phase. We had done more replicates for the stationary phase because we reasoned that *P. putida* in the soil was not likely to exist in the log phase, but that it would be more or less stationary and in equilibrium with the existing microbial community. Thus, we felt the stationary condition was a better model for the transcriptional changes that we believed were relevant in the soil. An advantage of looking at only one condition is that DESeq2 no longer had to account for variance from the other condition, which we hoped would clean up the data.

Sample clustering across stationary samples shows strong reproducibility from Illumina and marginal reproducibility from Nanopore data

Figure 14 shows the clustering of the stationary phase samples by Euclidian distance. The Illumina samples, the left graph, reveals the same that we saw in Figure 9, the experimental replicates have strong clustering. What is interesting is the Nanopore samples, the right graph, where the experimental replicates appear to have some form of clustering whereas in Figure 10 they did not.

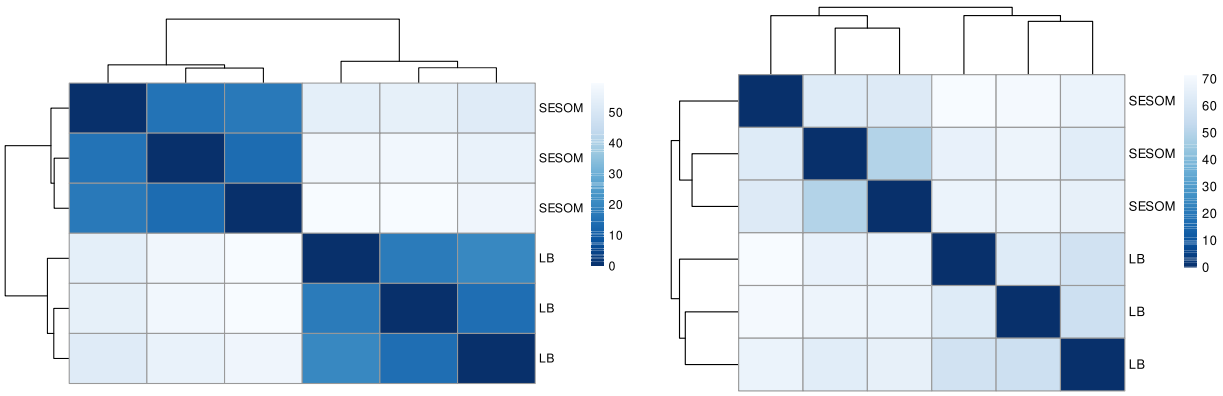

Figure 14, Sample Distances: Sample distances for the Illumina and Nanopore data only including samples that were in the stationary phase. The Illumina data is on the left and the Nanopore data is on the right.

Gene clustering in the stationary phase shows strong reproducibility from Illumina and marginal reproducibility from Nanopore data

Figure 15 shows gene variation between conditions in the Illumina stationary phase samples and looks very similar to the Illumina samples in Figure 11, further cementing the observation that the growth medium (LB vs SESOM) defines the relationship in the samples. If anything, the degree of variance in the genes is higher than that in Figure 11.

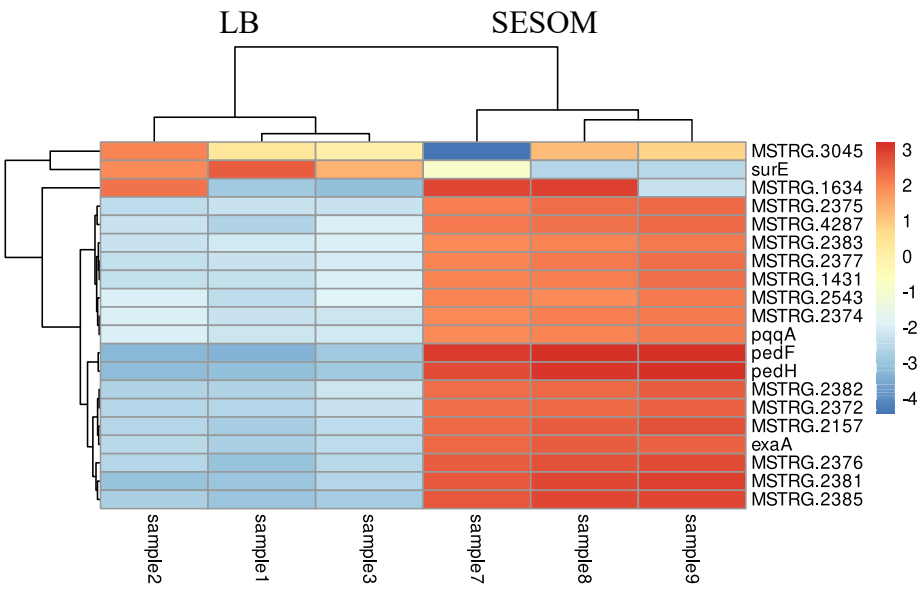

Figure 15, Illumina Stationary Phase: Heatmap of the difference in expression of genes as measured by the number of transcripts. The data comes from the Illumina sequencing results and only includes samples that were in the Stationary phase. The samples correspond to the samples in Table 1. Blue colors indicates that the expression of the gene in that sample is lower than the average expression of the gene and the red indicates the opposite.

Figure 16 shows gene variance in the Nanopore stationary phase samples. To an extent Figure 16 mirrors the change we saw in Figure 14, where the Nanopore samples are more clustered related to each other by their growth medium than the Nanopore all conditions samples.

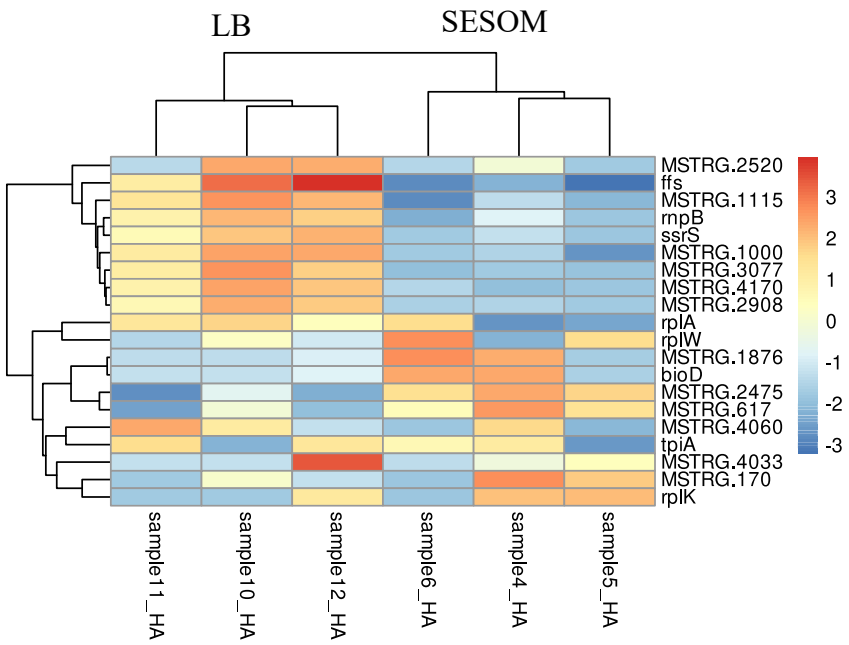

Figure 16, Nanopore Stationary Phase: Heatmap of the difference in expression of genes as measured by the number of transcripts. The data comes from the Nanopore sequencing results and only includes samples that were in the Stationary phase. The samples correspond to the samples in Table 1. Blue colors indicates that the expression of the gene in that sample is lower than the average expression of the gene and the red indicates the opposite.

Principal Component Analysis of Stationary Phase confirmed that samples were separable by medium type

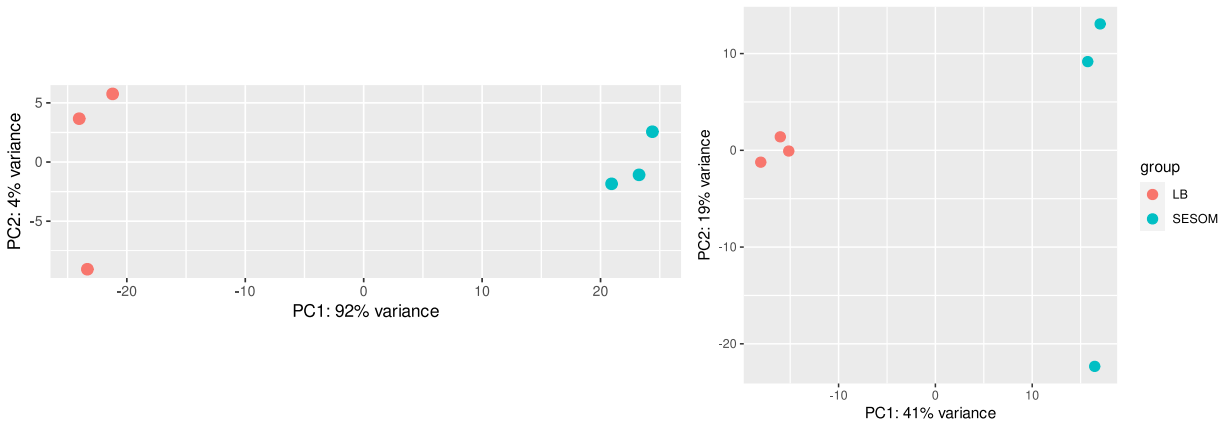

Figure 17, SESOM vs LB PCA: Principal Component Analysis of the Illumina and Nanopore data where the phase is stationary. The plot on the left is the Illumina data, and the plot on the right is the Nanopore Data.

The PCA plots of the Illumina data (Fig. 17, left graph) reveal that the majority of the variance in the samples is attributable to the growth medium and that this accounts for 92% of the variance which is a great deal higher than the 77% of variance that we saw in Figure 13 for the all conditions data. This pattern is repeated in the Nanopore data (Fig 15, right graph) where the PCA plots reveal that the growth medium accounts for 41% of the variance, which is higher than the 37% of variance in the all conditions data.

MA plot of Stationary Phase confirmed differential expression of Illumina data and was inconclusive for Nanopore data

An MA plot is a way to visualize differential expression. The x-axis is the average of the normalized expression of the gene across all samples, and the y-axis is log2 fold change between the two conditions that are being looked at. If a gene is upregulated in the condition, it will appear above the x-axis and if it is downregulated it will appear below the x-axis. The distance from the x-axis also indicates the degree to which the expression level is different in the conditions.

Figure 18 shows the MA plot of the Illumina stationary phase data. The difference in expression is apparent by the large number of genes that have a high fold change, additionally a large number of genes are statistically significant.

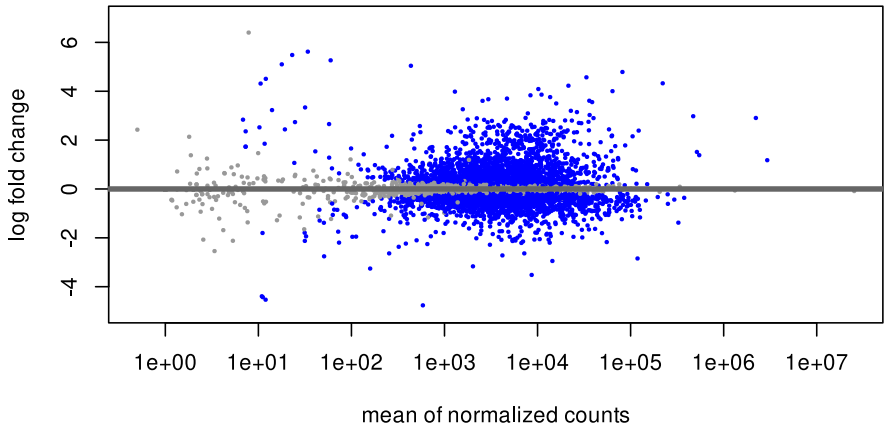

Figure 18, Illumina Stationary Phase MA Plot: MA plot of the Illumina stationary phase samples. Genes with an adjusted p value below 0.1 are shown in blue.

Figure 19 shows the MA plot of the Nanopore stationary phase data. The distribution of the genes is not similar to the genes in Figure 18, additionally a lot fewer genes are statistically significant.

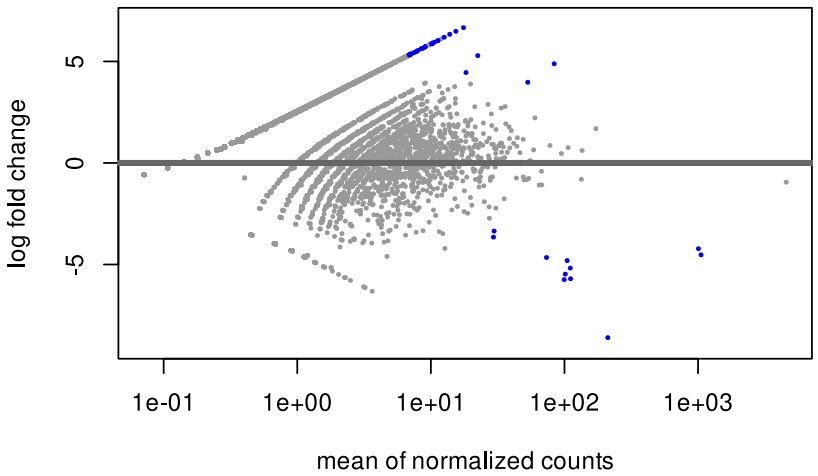

Figure 19, Nanopore Stationary Phase MA Plot: MA plot of the Nanopore stationary phase samples. Genes with an adjusted p value below 0.1 are shown in blue.

Volcano plot of Stationary Phase confirmed differential expression for Illumina and Nanopore data

A volcano plot is another way to visualize differential expression of genes and is named so because it is visually similar to an exploding volcano. In a volcano plot the negative log base-10 of the p-value is plotted against the log base-2 of the fold change. The higher on the y-axis a point is, the lower p-value that gene has, meaning it is more statistically significant. The further a point is from the center on the x-axis, the greater the difference in expression levels between conditions. A wide dispersion indicates a high level of difference in gene expression between conditions.

Figure 20 shows the volcano plot of the Illumina stationary phase samples. Figure 20 shows a wide dispersion of genes indicating differential expression, much like Figure 18, and a large number of statistically significant genes.

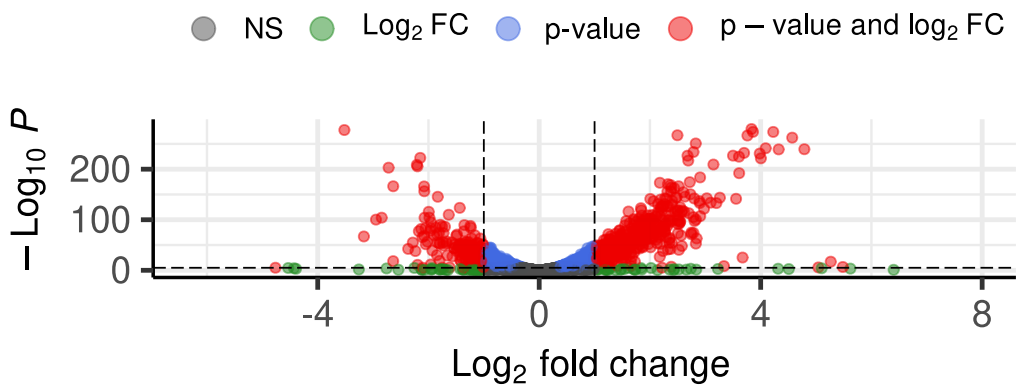

Figure 20, Illumina Stationary Phase Volcano Plot: Volcano plot created with a log fold-change threshold of -1 and 1 and a p-value threshold of 0.00001.

Unlike Figure 20, Figure 21, the volcano plot of the Nanopore stationary phase samples, has a wide dispersion, but does not have a large number of statistically significant genes. The structure of Figure 19 is not readily apparent in the plot.

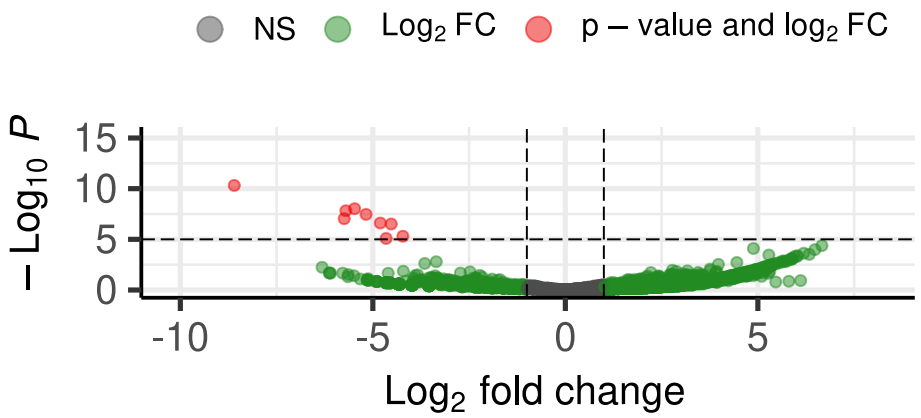

Figure 21, Nanopore Stationary Phase Volcano Plot: Volcano plot created with a log fold-change threshold of -1 and 1 and a p-value threshold of 0.00001.

We looked at all of the genes that were above the p-value threshold in the Nanopore data, and checked to see whether they were also significant in the Illumina data and to what extent they were represented. Our hypothesis was that if genes were present and significant in both forms of sequencing then they most likely were of real importance. Table 3 shows the conversion of names given to the genes in the Nanopore data by StringTie into their equivalent names in the Illumina data. For those which did not have an equivalent name listed we ran a protein blast on the sequence and used the resulting protein.

Table 3, Significant Names

Significant Names			
Nanopore ID	Gene ID	Protein Blast	Illumina ID
ffs	ffs	N/A	ffs
MSTRG.3077	None	tpx	tpx
MSTRG.1000	None	PP_RS05895	MSTRG.1001
MSTRG.4170	None	Unknown Protein	N/A
MSTRG.1115	None		N/A
MSTRG.2908	None		N/A
ssrS	ssRs	N/A	ssrS
rnpB	rnpB	N/A	rnpB
MSTRG.474	PP_RS02785	N/A	MSTRG.453

The names of the genes that were the most statistically significant in the Nanopore data, and their counterpart in the Illumina data.

After we found the names of the genes in the Illumina data, we checked to see what their log-fold change and p value were to see how they compare to their nanopore counterpart, this can be seen in Table 4. Two genes are not included in the table because their p value was greater than 0.05.

Table 4, Significant Genes

Significant Genes				
	Nanopore		Illumina	
Gene	Log2 Fold Change	-log10(p)	Log2 Fold Change	-log10(p)
ffs	-8.600419344	10.31873271	-0.391240396	2.054987393
tpx	-5.473781767	8.024390204	-0.282194788	2.268236372
PP_RS05895	-5.701191305	7.822246689	-0.386721184	4.677954384
ssrS	-4.524193746	6.512166407	N/A	N/A
rnpB	-4.220283732	5.306480853	N/A	N/A
PP_RS02785	-4.655519646	5.08710644	-0.263481393	1.828390188

Table of the genes that were most statistically significant in the Nanopore data, their log2 Fold change and -log10(p) value as well as the log2 Fold change and -log10(p) value of their equivalent in the Illumina data.

Discussion

All Conditions discussion

Our results showed that the expression of the *P. putida* samples differed based on the media they were grown in, LB or SESOM, and the phase of growth they were in, log or stationary (Fig. 9). What is interesting is the degree to which media (LB or SESOM) seems to play a much larger role in determining sample difference than does the growth phase (log or stationary). The sample clustering heatmap for the Illumina data is divided into two distinct sections based on the media that the samples were grown in. The samples that are the most similar are the biological replicates, i.e., the same media and growth Phase. The samples that are next closest in similarity are those samples that share the same media but are in a different phase of growth. The samples appear to share little similarity with those that are grown in a different media, regardless of whether or not the samples are in the same phase. The idea that media is what matters for sample difference is further enforced in the Illumina gene clustering in Figure 11. We can see that the while the top 3 genes are variable in different conditions with no apparent pattern, the remainder of the genes have a strict line that divides them between media, The tree at the top of the diagram in Figure 11 is another way to visualize this division, where the two media types are the roots of the tree and the different phases are intermingled seemingly at random. Another point of interest is that aside from the first 3 of the most variably expressed genes, the other 17 most variably expressed genes displayed a clear pattern where they were below average in their expression in the LB media and above average in their expression in the SESOM media.

Interestingly, the Nanopore data showed very different results from the Illumina data. Unlike the gene clustering heatmap of the Illumina data in Figure 9, there is no clear division in the Nanopore data in Figure 10. It is clear that the samples are different from each other, and are further away from certain samples than others, but we do not see the same division that is apparent in the Illumina data. By looking at the tree on the top and sides of the heatmap we can see that generally speaking the samples with the same media are more similar to each other, however this is not as clear cut as it is in the Illumina data. Furthering this narrative, The Nanopore data in Figure 12 appears in contrast to the Illumina data in Figure 11, the first two genes are especially interesting, as instead of being variable in many conditions like the most variable Illumina genes, they are instead more expressed than average in the LB and less

expressed in the SESOM. Some of the later genes follow the familiar pattern from Figure 11, less expression in the LB and more expression in the SESOM, but the majority of the genes have no discernable pattern of expression, certainly not to the same extent as Figure 11, where one could determine the change in media by the color of the genes alone. The tree at the top of the table indicates that the samples are broadly more related to each other by their media than their phase, however even in the tree there is some confusion, as one of the LB samples gets grouped with the SESOM. Certainly, when compared to the Illumina data in Figure 11, the Nanopore data does not show strong consistency across the behavior of samples from the same condition.

The inconsistency of the Nanopore data is evident in the principal component analysis as well (Fig. 12). The first principal component for both the Illumina and Nanopore data is the media. For the Illumina data the media accounts for 71% of the variance, while for the Nanopore data the media accounts for only 31% of the variance. The second principal component for the Illumina data is the growth phase which accounts for 9% of the variance for the samples in the Illumina. The second principal component for the Nanopore data accounts for 15% of the variance, but we are not sure what that component is. The order that was seen in the Illumina data in Figure 11 is very apparent in the PCA, as is the disordered nature of the Nanopore data. However, while the Nanopore data did appear very disordered in Figure 12, it is clear from the PCA analysis of Illumina data that both media and phase play an important role in gene expression and media the larger role of the two.

We expect that the greater variability and poorer correlations within the Nanopore data arises from a combination of the sample preparation, sequencing method and available replicates. We know that the rRNA depletion was not complete for any of the samples, and varied greatly between the samples, which affected the number of useful reads in each sample. Additionally, because we did not have access to the proper equipment, we were not able to verify the number of bases added in the polyadenylation, and it is possible that they were much more than were required, which may negatively affect the sample sequencing. Due to time constraints, we were not able to run the samples for their required 48-72 hours; instead, the samples were run for around 10-12 hours each. This most likely decreased the number of reads that we got for each sample. Due to budgetary constraints, we were forced to re-use the flow cell 5 times. We observed a significant decrease in read number and quality after each run. The least correlated

samples (LB log and SESOM log, two replicates each) were also those run last on each flow cell. Finally, there are only two biological replicates of some samples because we again were limited with the available flow cells and calculated that five was the greatest number of samples that we could run on each flow cell and still return a reasonable read count. We prioritized the stationary phase samples as we predicted these to be the most relevant to the native soil condition of *P. putida*. These caveats and limitations are all contributing to the lack of consistency in the Nanopore samples.

Stationary Phase discussion

When we look at the sample clustering in the stationary phase in Figure 14 the Illumina data looks the same as in Figure 9, however the difference is apparent when we look at the Nanopore data. The samples appear much more similar, and they vaguely resemble the shape of the Illumina data. It appears that when we look at only the stationary phase, we remove a lot of noise from the samples and are able to see that the replicate samples that come from the same media are more closely related to each other. This is confirmed by the trees on the side and top of the heatmap that show that like the Illumina data, the Nanopore data is evenly divided into blocks of related samples based on their media. This is reflected in the gene clustering data in Figures 15 and 16 for the Illumina and Nanopore data respectively. The Illumina data looks very similar to the gene clustering of the all conditions Illumina data in Figure 11. We see the same pattern where the most variable genes are variant in all conditions with no discernable pattern, and then the remaining genes are less present than their average in LB and more present than their average in the SESOM. A difference here from Figure 11 is that the degree of variation in the genes appears to be higher between conditions than in the all condition data. The Nanopore stationary phase gene clustering data in Figure 16 is very different from the all conditions data in Figure 12. In Figure 16 we can clearly see the division between the two media, just like we can with the Illumina data. Interestingly, unlike the Illumina data, the larger section of genes is expressed more on average in the LB than it is in the SESOM.

The trends from the gene clustering are reflected in the PCA in Figure 17. By looking only at the stationary phase, the degree to which the media plays an effect is even more apparent. For the Illumina data we see that 92 percent of the variance is accounted for by the difference in media, and for the Nanopore data the difference in media accounts for 41 percent of the variance,

in both cases this is up from the principal component analysis of the all conditions data. What is especially interesting is the second principal component, which in this case is unknown. The unknown component contributes to 4 percent of the variance in the Illumina data, which is close to the same amount of contribution that the difference in phase had in the all conditions data, and in the Nanopore data the unknown second principal component contributes to 19 percent of the variance, which is greater than the contribution of the phase in the all conditions data.

Looking more closely at the differential expression of genes in the stationary phase data, we can see in Figure 18 that for the Illumina data, there are many genes that are statistically significant, however, there do not appear to be that many genes that have a very large log2 fold change and the amount of up and down regulated genes visually appear to be similar. The Nanopore data seen in Figure 19, however, is vastly different from its Illumina counterpart. There are not many statistically significant genes in the nanopore data, and the genes fall into very interesting and unexplained patterns, where there seems to be a strong correlation between the expression and the log fold change.

By looking at volcano plots of the data we see more clearly the degree to which differential expression occurs. For the Illumina data in Figure 20 we have a solid spread of genes that are significant and have a log fold-change greater than the threshold. This is what we would expect for two differently expressed conditions, further confirming our belief that *P. putida* behaves differently in the lab, LB conditions than it does in soil conditions, as the SESOM is directly derived from soil. The Nanopore data in Figure 21 only somewhat appears how we expected it to, the dispersion that we expect to see in differently expressed conditions is there, however, almost none of the genes are statistically significant, there are only 9 genes that are above the p-value threshold. When we looked at those 9 genes in the Illumina data we saw that of the 4 that were statistically significant in the Illumina data all of them had the same direction of fold change as the Nanopore data. While their degree of fold change was a lot less, this indicates that to a very limited extent the Nanopore data may have detected the most robustly differentially expressed genes in the conditions. The analysis further suggests that with alterations to the rRNA depletion protocol and additional flow cells to improve read depth, the Nanopore approach could, in future experiments, yield accurate data reflecting the transcriptome of *P. putida* under varying growth conditions.

While we were able to show the potential of using SESOM as a method to validate the behavior of *P. putida,* we were not able to demonstrate the usefulness of Nanopore in direct RNA sequencing at least at its current stage. While the drawbacks to Illumina that we mentioned previously are important to keep in mind, we discovered that while Nanopore does not have those drawbacks it does have issues of its own. Specifically, Nanopore is not optimized for working with prokaryotic RNA. The necessity to add a poly(A) tail to the sample and then eliminate rRNA adds a lot of complexity to the process and results in loss of the sample.

The Illumina data set will be analyzed in future studies to reveal more information about specific gene expression pathways and programs that are active in the soil environment. One specific form of analysis that we are interested in is gene ontology (GO) term analysis. GO term analysis makes use of the Gene Ontology (GO) system of classification, which is a formal way of representing genes by describing the biological process, molecular function, and cellular components of gene products (Yon Rhee et al., 2008). We are interested in using GO term analysis to perform enrichment analysis where we look at a set up genes that are up regulated in a condition, and see what biological process and functions those genes correspond to, this will allow us to get a better understanding of how P. putida behaves when it is in the LB vs SESOM environment.

Bibliography

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. https://doi.org/10.1186/gb-2011-12-2-r18

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data* (0.12.0). https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Beckman Coulter. (2020). *Instructions For Use RNAClean XP In Vitro Produced RNA and cDNA Purification*. www.beckman.com/techdocs

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

de Lorenzo, V., Krasnogor, N., & Schmidt, M. (2021). For the sake of the Bioeconomy: define what a Synthetic Biology Chassis is! *New Biotechnology*, *60*, 44–51. https://doi.org/10.1016/j.nbt.2020.08.004

Fierer, N. (2017). Embracing the unknown: Disentangling the complexities of the soil microbiome. In *Nature Reviews Microbiology* (Vol. 15, Issue 10, pp. 579–590). Nature Publishing Group. https://doi.org/10.1038/nrmicro.2017.87

Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, *38*(12).

Illumina. (2017). *An introduction to Next-Generation Sequencing Technology*.

Jansson, J. K., McClure, R., & Egbert, R. G. (2023). Soil microbiome engineering for sustainability in a changing environment. In *Nature Biotechnology*. Nature Research. https://doi.org/10.1038/s41587-023-01932-3

Kampers, L. F. C., Volkers, R. J. M., & Martins dos Santos, V. A. P. (2019). Pseudomonas putida KT2440 is HV1 certified, not GRAS. In *Microbial Biotechnology* (Vol. 12, Issue 5, pp. 845–848). John Wiley and Sons Ltd. https://doi.org/10.1111/1751-7915.13443

Kim, J., & Park, W. (2014). Oxidative stress response in Pseudomonas putida. In *Applied Microbiology and Biotechnology* (Vol. 98, Issue 16, pp. 6933–6946). Springer Verlag. https://doi.org/10.1007/s00253-014-5883-4

Kopittke, P. M., Menzies, N. W., Wang, P., McKenna, B. A., & Lombi, E. (2019). Soil and the intensification of agriculture for global food security. In *Environment International* (Vol. 132). Elsevier Ltd. https://doi.org/10.1016/j.envint.2019.105078

Kraamwinkel, C. T., Beaulieu, A., Dias, T., & Howison, R. A. (2021). Planetary limits to soil degradation. In *Communications Earth and Environment* (Vol. 2, Issue 1). Nature Publishing Group. https://doi.org/10.1038/s43247-021-00323-3

Kubota, T., Lloyd, K., Sakashita, N., Minato, S., Ishida, K., & Mitsui, T. (2019). Clog and release, and reverse motions of DNA in a nanopore. *Polymers*, *11*(1). https://doi.org/10.3390/polym11010084

Kukurugya, M. A., Mendonca, C. M., Solhtalab, M., Wilkes, R. A., Thannhauser, T. W., & Aristilde, L. (2019). Multi-omics analysis unravels a segregated metabolic flux network that tunes co-utilization of sugar and aromatic carbons in Pseudomonas putida. *Journal of Biological Chemistry*, *294*(21), 8464–8479. https://doi.org/10.1074/jbc.RA119.007885

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lehmann, J., Bossio, D. A., Kögel-Knabner, I., & Rillig, M. C. (2020). The concept and future prospects of soil health. In *Nature Reviews Earth and Environment* (Vol. 1, Issue 10, pp. 544–553). Springer Nature. https://doi.org/10.1038/s43017-020-0080-8

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. In *Genomics, Proteomics and Bioinformatics* (Vol. 14, Issue 5, pp. 265–279). Beijing Genomics Institute. https://doi.org/10.1016/j.gpb.2016.05.004

Martin, M. (2011). Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet.Journal*, *17*(1), 10–12. http://www-huber.embl.de/users/an-

Martínez-García, E., Aparicio, T., de Lorenzo, V., & Nikel, P. I. (2014). New transposon tools tailored for metabolic engineering of Gram-negative microbial cell factories. *Frontiers in Bioengineering and Biotechnology*, *2*(OCT). https://doi.org/10.3389/fbioe.2014.00046

Martínez-García, E., Goñi-Moreno, A., Bartley, B., McLaughlin, J., Sánchez-Sampedro, L., Pascual Del Pozo, H., Prieto Hernández, C., Marletta, A. S., De Lucrezia, D., Sánchez-Fernández, G., Fraile, S., & De Lorenzo, V. (2020). SEVA 3.0: An update of the Standard European Vector Architecture for enabling portability of genetic constructs among diverse bacterial hosts. *Nucleic Acids Research*, *48*(D1), D1164–D1170. https://doi.org/10.1093/nar/gkz1024

Nakazawa, T. (2003). Travels of a Pseudomonas, from Japan around the world (Environmental Microbiology 4:12 (782-786)). In *Environmental Microbiology* (Vol. 5, Issue 1, p. 78). https://doi.org/10.1046/j.1462-2920.2002.00310.x

New England BioLabs. (n.d.-a). *Poly(A) Tailing of RNA using E. coli Poly(A) Polymerase*. Retrieved June 11, 2023, from https://www.neb.com/protocols/2014/08/13/poly-a-tailing-of-rna-using-e-coli-poly-a-polymerase-neb-m0276

New England BioLabs. (n.d.-b). *Protocol for rRNA depletion using NEBNext rRNA Depletion Kit*. Retrieved June 11, 2023, from https://www.neb.com/protocols/2019/09/18/protocol-for-rrna-depletion-using-nebnext-rrna-depletion-kit-bacteria-neb-e7850-neb-e7860

Nikel, P. I., Chavarría, M., Fuhrer, T., Sauer, U., & De Lorenzo, V. (2015). Pseudomonas putida KT2440 strain metabolizes glucose through a cycle formed by enzymes of the Entner-Doudoroff, embden-meyerhof-parnas, and pentose phosphate pathways. *Journal of Biological Chemistry*, *290*(43), 25920–25932. https://doi.org/10.1074/jbc.M115.687749

Nikel, P. I., & de Lorenzo, V. (2013). Implantation of unmarked regulatory and metabolic modules in Gram-negative bacteria with specialised mini-transposon delivery vectors. *Journal of Biotechnology*, *163*(2), 143–154. https://doi.org/10.1016/j.jbiotec.2012.05.002

Nwachukwu, S. U. (2001). Bioremediation of sterile agricultural soils polluted with crude petroleum by application of the soil bacterium, Pseudomonas putida, with inorganic nutrient supplementations. *Current Microbiology*, *42*(4), 231–236. https://doi.org/10.1007/s002840110209

OECD/FAO. (2017). *OECD-FAO Agricultural Outlook 2017-2026*. OECD. https://doi.org/10.1787/agr_outlook-2017-en

O'Neil, D., Glowatz, H., & Schlumpberger, M. (2013). Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity. *Current Protocols in Molecular Biology*, *103*(1). https://doi.org/10.1002/0471142727.mb0419s103

Oxford Nanopore. (2016a). *Nanopore Technical Document: Chemistry Technical Document*.

Oxford Nanopore. (2016b). *Nanopore Technical Document: Data Analysis*.

Oxford Nanopore. (2023). *Nanopore product brochure*.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Ramos, J. L., Cuenca, M. S., Molina-Santiago, C., Segura, A., Duque, E., Ǵomez-Garciá, M. R., Udaondo, Z., & Roca, A. (2015). Mechanisms of solvent resistance mediated by interplay of cellular factors in Pseudomonas putida. In *FEMS Microbiology Reviews* (Vol. 39, Issue 4, pp. 555–566). Oxford University Press. https://doi.org/10.1093/femsre/fuv006

Rebello, S., Nathan, V. K., Sindhu, R., Binod, P., Awasthi, M. K., & Pandey, A. (2021). Bioengineered microbes for soil health restoration: present status and future. In *Bioengineered* (Vol. 12, Issue 2, pp. 12839–12853). Taylor and Francis Ltd. https://doi.org/10.1080/21655979.2021.2004645

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, *12*(3), R22. https://doi.org/10.1186/gb-2011-12-3-r22

Samuel, M. S., Sivaramakrishna, A., & Mehta, A. (2014). Bioremediation of p-Nitrophenol by Pseudomonas putida 1274 strain. *Journal of Environmental Health Science and Engineering*, *12*(1). https://doi.org/10.1186/2052-336X-12-53

Sendler, E., Johnson, G. D., & Krawetz, S. A. (2011). Local and global factors affecting RNA sequencing analysis. *Analytical Biochemistry*, *419*(2), 317–322. https://doi.org/10.1016/j.ab.2011.08.013

Sudarsan, S., Dethlefsen, S., Blank, L. M., Siemann-Herzberg, M., & Schmid, A. (2014). The Functional Structure of Central Carbon Metabolism in Pseudomonas putida KT2440. *Applied and Environmental Microbiology*, *80*(17), 5292–5303. https://doi.org/10.1128/AEM.01643-14

Thermo Fisher Scientific. (2016). *TRIzol Reagent User Guide*.

Timmis, K. N. (2002). Pseudomonas putida: a cosmopolitan opportunist par excellence. *Environmental Microbiology*, *4*(12), 779–781. https://doi.org/10.1046/j.1462-2920.2002.00365.x

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

Yon Rhee, S., Wood, V., Dolinski, K., & Draghici, S. (2008). Use and misuse of the gene ontology annotations. In *Nature Reviews Genetics* (Vol. 9, Issue 7, pp. 509–515). https://doi.org/10.1038/nrg2363

Zuo, Z., Gong, T., Che, Y., Liu, R., Xu, P., Jiang, H., Qiao, C., Song, C., & Yang, C. (2015). Engineering Pseudomonas putida KT2440 for simultaneous degradation of organophosphates and pyrethroids and its application in bioremediation of soil. *Biodegradation*, *26*(3), 223–233. https://doi.org/10.1007/s10532-015-9729-2