7-2020

BIOINFORMATIC ANALYSIS OF THE GUT MICROBIOTA DERIVED FROM THE OIL FLY *HELAEOMYIA PETROLEI* FROM THE LA BREA TAR PITS

Brian Dillard
University of Nebraska - Lincoln, brian.dillard@huskers.unl.edu

BIOINFORMATIC ANALYSIS OF THE GUT MICROBIOTA DERIVED FROM THE OIL FLY *HELAEOMYIA PETROLEI FROM THE LA BREA TAR PITS*

By

Brian Dillard

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Biological Sciences

Under the Supervision of Professor Kenneth Nickerson

Lincoln, Nebraska

July 2020

BIOINFORMATICAL ANALYSIS OF THE GUT MICROBIOTA DERIVED FROM THE OIL FLY *HELAEOMYIA PETROLEI FROM THE LA BREA TAR PITS*

Brian Dillard, M.S.

University of Nebraska, 2020

Advisor: Kenneth Nickerson

In the early 1930s, Thorpe a prominent entomologist, called for more research into Helaeomyia petrolei larva. These larvae live in the Californian La Brea tar pits where they are exposed to large amounts of polyaromatic hydrocarbons. Molecules like anthracene, phenanthrene, and toluene which should be highly toxic to both the oil fly larvae and its enteric bacteria. This extremophilic gut microbiome has yet to be studied using current day next gen sequencing and bioinformatic techniques. In fact, since Thorpe's work in the 1930s, there have been only two publications characterizing the oil fly larvae. Both in the early 2000s by Kadavy, characterizing the abundance of enteric bacteria in the oil fly, and another describing the surprising antibiotic resistance these gut isolates possess. Almost every isolate described was resistant to over half of the 22 antibiotics tested. We hypothesize that characterizing this larval microbiota on a more intimate level could identify important extremophilic enzymes that would be useful in an industrial capacity, provide insights into the rapid natural development of antibiotic resistance, and identify an organism or collection of organisms that can metabolize the aromatic hydrocarbons that make up tar.

Using a combination of 16S and whole genomes sequencing, we have solidified taxonomic classifications for a majority of the original isolates tested, even establishing a new genus Candidatus *Petroalcaligenes*. We have also identified a large array of putative drug efflux pumps which might confer both tolerance to solvent stress and antibiotic resistance in our OF2

isolate. When looking for evidence of OF2's 16S sequence in metagenomic data sets we see that OF2 is only found in specific subsets of sample types and are found at extremely low percentages. Using Pangenomics, where you can compare genetic content amongst a large variety of genomes, we have determined that OF2 has enriched functions in the categories of transport, osmoregulation, metabolism, and antibiotic resistance.

We have also worked to show that the stress imposed by the tar pit rather than the oil gut are to blame for the increased antibiotic resistance and solvent tolerance. Using our isolates OF5, OF6, and OF10 from *Providencia*, we made a pangenome of 56 genomes comparing every known species of the *Providencia*. With particular interest on *P. rettgeri* and *P. vermicola*, gene enrichment analysis was done to multiple different groupings of genomes. These groups had an emphasis on bacterial host and host environment. While there were some enriched genes across the groups, the results were hard to parse because of the large number of species misclassifications within *Providencia rettgeri*. Through phylogenetic analysis, we highlighted some of the genomes sampled as candidates for a reclassification as *P. vermicola*.

TABLE OF CONTENTS

Page

CHAPTER 1 - Genome diversity and Pangenome Analysis of candidate genus *Petroalcaligenes* isolated from the bacterial gut flora of the oil fly *Helaeomyia petrolei* from the La Brea tar pits

Brian Dillard[1], Lisa Durso[2], Joshua R. Herr[1,3,4], and Kenneth W. Nickerson[1]*

[1] School of Biological Sciences, University of Nebraska, Lincoln, NE 68588.

[2] USDA-ARS, University of Nebraska, Lincoln, NE 68583.

[3] Department of Plant Pathology, University of Nebraska, Lincoln, NE 68583

[4] Center for Plant Science Innovation, University of Nebraska, Lincoln, NE 68588

*Corresponding author.

Kenneth W. Nickerson

Biological Sciences, University of Nebraska,

Lincoln, NE USA 68588-0666

402-472-2253 knickerson1@unl.edu

Abstract

Helaeomyia petrolei (oil fly) larvae mature in the asphaltene and polyaromatic hydrocarbon rich asphalt seeps of Rancho La Brea, Los Angeles, California. These larvae are able to pass high amounts of viscous asphalt through their digestive system with no discernible negative effects. While they do not derive nutrients from the asphalt, they can survive and grow in this harsh environment. Similar to all life, these oil fly larvae have a complex gut flora. Previous work isolated bacteria from the larval gut and characterized the antibiotic resistance of these bacteria. In the present work, we focused on an uncharacterized antibiotic resistant isolate, OF2. Using whole genome sequencing data and leveraging phylogenomic and pangenomic analysis of 162 single copy genes, the average amino acid identity to close relatives, and its prevalence in public metagenomic data sets, we suggest that OF2 should be classified in the new genus candidatus *Petroalcaligenes labreaensis*. The suite of ca. 1130 unique genes and 50 unique gene functions identified through pangenome analysis of OF2 should provide insights into how OF2 survives in a polyaromatic hydrocarbon rich extreme environment. Of particular interest are solvent tolerant enzymes of potential utility for industry and efflux pumps that confer resistance to both antibiotics and organic solvents made up of polyaromatic hydrocarbons. The array of transport, osmoprotectant, and antibiotic resistance functions positively enriched in OF2 could provide resistance to the selective pressures imposed by the La Brea tar pits.

Introduction

Antibiotics have been used for close to 80 years as a tool to promote human health by fighting infection. The initial success of the use of antibiotics has now been tempered by the development and spread of numerous mechanisms whereby microbes have become antibiotic resistant. Antibiotic resistance has become so widespread that there are serious concerns about the appearance of microbial strains that are impervious to all known antibiotics, with a projected 10 million deaths per year from untreatable antibiotic resistant infections by 2050 - equivalent to the entire 2019 population of New York City and Houston combined (1). The World Health Organization has declared antibiotic resistance a global health threat of the highest priority (2), and the United States recently developed a National Action Plan for Combating Antibiotic-Resistant Bacteria that mobilize resources across the government to address antibiotic resistance (3). Although antibiotic resistant bacteria and antibiotic resistance genes occur naturally (4,5,6,7,8) and can be found in ancient and remote samples (9,10,11), their prevalence has increased rapidly since the introduction of antibiotic drug therapy in the 20th century (12,13,14). The rapidity with which this phalanx of antibiotic resistance genes has appeared suggests that microbes have been fighting among themselves for millions of years, using antibiotics as weapons or signaling molecules (15,16,17), and/or that they emerged from obscure genomes in a stressful environmental niche where they had been serving similar protective functions (4,18,19,20).

With the goal of finding obscure microbial genomes, potentially exhibiting protective functions in extreme environments, we previously examined the microbial flora present in the larval gut of the oil fly *Helaemyia petrolei* (21). *Helaeomyia petrolei* (oil fly) larvae mature in the asphaltene and polyaromatic hydrocarbon rich asphalt seeps of Rancho La Brea, Los

Angeles, Calif. The La Brea Tar Pits are a collection of asphalt tar pits in Southern California, surrounded by urban Los Angeles. They are formed from heavy oil that seeps up from underground oil fields. The tar pits are well known for the many large Pleistocene mammal fossils unearthed from the site, and for the many uses of the tar by native Americans. The modern-day site remains dynamic, with shifting tar sands and the development of new pools of asphalt. The Oil Fly larvae are able to pass high amounts of viscous asphalt through their digestive system with no discernible negative effects. While they do not derive nutrients from the asphalt, they can survive and grow in this harsh environment. Similar to all life, these oil fly larvae have a complex gut flora that are also in contact with the asphalt and thus should possess some form of solvent tolerance. Thorpe (22,23) referred to the oil fly as "undoubtedly one of the chief biological curiosities of the world" in that the carnivorous larvae are exclusively found submerged in oil, where they ingest large quantities of oil and asphalt without suffering any ill effects (22, 24). Operating on the premise that their microbial gut flora should also be adapted to a variety of aromatic hydrocarbons contained in the oil, the microbial gut contents of oil fly larvae from the asphalt seeps of Rancho La Brea in Los Angeles, California were previously examined (25). Standard microbial counts on Luria-Bertani, MacConkey, and blood agar plates indicated ca. 2×10^5 heterotrophic bacteria per larvae (25). The culturable bacteria represented 15-20 % of the total population as determined by acridine orange staining. All the bacteria isolated were non-spore forming and gram negative. Thirteen isolates were chosen for identification using the Enterotube II and API20E systems as well as fatty acid analysis (Table 1) (25). In a subsequent paper (26), 12 of the 13 bacterial strains were tested for their resistance or sensitivity to 23 antibiotics using commercially available antimicrobial susceptibility discs. All but one of the bacteria tested were resistant to approximately 10-12 of the antibiotics. These

bacteria were antibiotic resistant without as far as we know producing any antibiotics themselves or prior exposure to such antibiotics (26). With regard to the chemical structures of the antibiotics used, the bacteria were sensitive to the penicillins, cephalosporins, streptomycins, and kanamycins, while being resistant to a diverse collection of hydrophobic antibiotics, most of which contained planar aromatic and polyaromatic ring systems (26). We previously (26, 27) suggested that the antibiotic resistance exhibited by these oil fly bacteria was due to the promiscuity of efflux pumps required to tolerate the wide range of organic solvents and polyaromatic hydrocarbons encountered in the La Brea tar pits.

In order to better characterize the microbial components of the *Helaemyia petrolei* gut and the capabilities of the oil fly microbiome to produce novel mechanisms of antibiotic resistance, we began focusing on bacterial isolates from the gut flora of oil fly larvae from the La Brea tar pits. During these previous experiments (25, 26), we identified a bacterial isolate, designated OF2, that did not match databases with both previous phenotypic (Enterotube II and API20E systems, etc.) and now genetic (16S rRNA) data. In this study, we addressed the genomic, antibiotic resistance profiles, and functional diversity of this previously undescribed bacterial strain. We hypothesize that due to the nature of living and subsisting in this extreme environment, that the oil fly microbiome would reflect novel genes with potentially novel biochemical pathways. With the intent of describing the ways in which our OF2 isolate differs from its related taxa, we have focused on the identification of OF2, characterizing its prevalence and similarity to other bacteria in publicly available data sets.

Materials and Methods

Sample Collection, Processing, and Strain Verification

Oil fly larvae and asphalt were collected from the La Brea Tar Pits in Southern California and processed as previously described (25). A total of 40 larvae were initially collected from two locations at the LaBrea tar pits on three occasions between 1994 and 1997. Briefly, larvae were shipped live to the laboratory and kept alive on egg meat medium (Difco, Detroit, MI). They were surface sterilized using a washing procedure including linoleic acid, ethanol, bleach (supplemented with Tween 20), and phosphate buffered saline (also supplemented with Tween 20). Insect gut samples were homogenized and plated, as follows, on multiple media. Colonies from the isolates were picked and stored in glycerol at -80°C. All the oil fly isolates were recovered from freezer storage by plating on Luria Bertani (LB) Agar (Difco, Franklin Lakes, NJ). Morphological features were confirmed microscopically, and cultures went through three successive rounds of isolation streaking on LB Agar (Difco, Franklin Lakes, NJ) before overnight incubation at 37°C to confirm isolate integrity.

For the initial 16S sequence determination to identify strains, cultures were grown overnight in LB broth (Difco, Franklin Lakes, NJ) at 37°C. Genomic DNA was obtained with the QIA-Amp Power Fecal DNA kit (product number 12830; Qiagen Inc., Germantown, MD) used according to the manufacturer's instructions. We then sequenced the full-length 16S ribosomal subunits of 13 strains previously studied (25,26). Extracted genomic DNA was shipped to Molecular Research LP (Shallowater, TX) overnight on dry ice. A 35-cycle PCR was performed using 27F and 1492R primers and the HotStar Taq Plus Master Mix (Qiagen, Germantown, MD). DNA quality was checked using a 2% agarose gel and purified using Ampure PB beads

(Pacific Biosciences, Menlo Park, CA). Libraries were created using the SMRTbell library kit from Pacific Biosciences (Menlo Park, CA) and sequenced on a PacBio Sequel following the manufacturer's guidelines. Data was processed using the PacBio Circular Consensus Sequencing algorithm to merge overlapping forward and reverse reads.

Genome Sequencing of the OF2 Strain Isolated from the Oil Fly Gut

On the basis of a novel 16S identification, OF2 was selected for whole genome sequencing. Pelleted cells of the OF2 oil fly gut isolate were shipped overnight for extraction and sequencing, performed by The Sequencing Center (Fort Collins, CO). Whole genome sequencing libraries were prepared using the Nextera XT Library Kit and Illumina Nextera XT Index Kit (Illumina, San Diego, CA), and sequenced on a MiSeq Sequencing System using 2 X 250 bp paired-end reads. Upon run completion, the MiSeq instrument ran an adapter trimming algorithm to remove Nextera adapter sequences from sample reads, and an algorithm that assigned reads to the previously barcoded samples subsequently demultiplexed the reads based on barcode indices assigned during the library prep step. Paired-end FASTQ files were generated for further analysis.

Due to the novel identification of the OF2 lineage inferred from the 16S rRNA sequencing, we subsequently focused specifically on characterizing the genome structure and pan-genomic relationships of OF2 to other members of the Alcaligenaceae family.

Genome Assembly and Annotation

We acquired FASTQ data in the form of paired-end 2 X 250 bp reads sequenced on the Illumina MiSeq platform. Data files were initially quality checked using the FASTQC tool, available along the HTSeq package (28). Poor quality ends of the FASTQ files were trimmed with the VSEARCH tool (29) and error-prone data was removed in the form of low abundant k-

mers using the KHMER tool (30). Trimmed FASTQ data were assembled using the SPADES tool utilizing k-mers of 21, 31, 41, 61, 91, 121, and 141 base-pair lengths (31). A consensus assembly of the different k-mer values represented our final genome assembly. Genome assembly quality was assessed using the QUAST tool (32) before assessing efflux pump, antibiotic resistance gene diversity, and pan-genome associations and phylogenetics. The final genome assembly was annotated for genes using the PROKKA tool with standard input flags for bacterial genes (33).

Pan-Genome Data Set Construction

Based on our previous 16S sequencing of the OF2 isolate, placing it in *Alcaligenaceae*, we chose for comparison 32 representative genome sequences from NCBI ref seq databases (34) representing genomes from all identified genera within the Alcaligenaceae. Genome sequences were selected based on completeness (single chromosome assembly), genome quality (lack of ambiguous assembly criteria as established using QUAST) and phylogenetic breadth within the Alcaligenaceae for pan-genome comparison (32, 34). Strains chosen for pan-genome analysis included; *Achromobacter insolitus* DSM23807, *Achromobacter spanius* DSM23806, *Achromobacter xylosoxidans* FDAARGOS150, *Alcaligenes aquatilis* QD168, *Alcaligenes faecalis* J481, *Alcaligenes faecalis* JQ135, *Alcaligenes faecalis* P156, *Alcaligenes faecalis* ZD02, *Basilea psittacipulmonis* DSM24701, *Bordetella avium* 197N, *Bordetella bronchiseptica* I328, *Bordetella hinzii* F582, *Bordetella holmesii* F627, *Bordetella holmesii* H903, *Bordetella parapertussis* FDAARGOS177, *Bordetella parapertussis* H904, *Bordetella pertussis* B1917, *Bordetella pertussis* CS, *Bordetella pertussis* B203, *Bordetella petrii* DSM12804, *Bordetella sp* H567, *Castellaniella defragrans* 65Phen, *Kerstersia gyiorum* SWMUKG01, *Oligella urethralis* FDAARGOS-329, *Pigmentiphaga aceris* Mada1488, *Pigmentiphaga sp* H8, *Pusillimonas sp*

ye3, *Taylorella asinigenitalis* MCE3, *Taylorella equigenitalis* ATCC-35865, and *Taylorella equigenitalis* MCE9. *Pseudomonas mallei* strain RKJZ01 was designated as an outgroup in the analysis.

Pan-Genomic Analysis of the OF2 isolate

In order to identify the core genome of our OF2 isolate and related taxa, everything was annotated in the standard GFF3 format using the PROKKA tool with standard input flags for bacterial functional genes (33). We then used the ANVI'O tool (35) to construct a pan-genome of the Alcaligenaceae. This workflow consisted of first generating a genome database of both DNA and amino acid sequences, as well as the functional annotation, from the selected genomes in the previous section. This database was then used to identify gene clusters across the genome database. We used the criteria for gene clusters – predicted open readings frames exhibiting homology at the level of DNA translation – defined as the standard default in the ANVI'O tool (35). Sequences were aligned using MUSCLE (36) and Similarities in amino acid sequence were identified via the blastp algorithm (37). Using tools provided through ANVI'O (35), we ran an analysis to pull out both the enriched functions and unique gene clusters found in OF2 when compared to the rest of the pan-genome. Enriched functions were classified based on the PROKKA (33) functional annotations and were classed as either negatively or positively enriched based on the proportion of genomes that carried each specific function compared to OF2. In the case of this analysis we used the q-value cutoff of .5% to determine which functions could accurately be described as enriched. The unique gene clusters of OF2 were found by binning the annotated genes that were only present in OF2. This bin was summarized, and clusters were categorized into subsets using BlastKOALA (38) based on KEGG categories.

Single copy genes used to construct the pan-genome phylogeny were selected from published lists of single copy genes for both bacteria and archaea. FAST-TREE (39) was used to run a Bayesian phylogenetic inference with posterior probabilities calculated for each node of the phylogenetic tree. The subsequent consensus tree was rooted with strain *Pseudomonas mallei* RKJZ01 using the FIGTREE tool (40).

Pathway mapping of OF2 and the Greater Pan-genome

In addition to OF2, 11 genomes including; *Achromobacter insolitus* DSM23807, *Alcaligenes faecalis* ZD02, *Basilea psittacipulmonis* DSM24701, *Bordetella pertussis* B1917, *Castellaniella defragrans* 65Phen, *Kerstersia gyiorum* SWMUKG01, *Oligella urethralis* FDAARGOS-329, *Pigmentiphaga aceris* Mada1488, *Pseudomonas mallei* RKJZ01, *Pusillimonas sp* ye3, *Taylorella equigenitalis* ATCC-35865 were chosen to represent every genus the greater pan genome. The PROKKA annotations were submitted to the BlastKOALA (38) tool in order to reannotate the sequences with the KEGG database making them compatible with the KEGG mapper tool (41). Duplicate gene annotations were removed, and the KEGG mapper tool was used to map the cumulative annotations of the 12 chosen genomes and OF2 against all of the available pathways.

Environmental Sequence Search for OF2-like Sequences

There were no annotated OF2 16S genes in curated reference databases, so we wanted to identify environmental niches and quantify the abundance of the OF2 taxonomic unit in various public data sets. To do this, we queried publicly available 16S marker based sequencing data sets using the V4-V6 region primers from the Earth Microbiome Project (42). The number of public 16S data sets we were able to query (queried on September 15th, 2019) totaled 422877. We used the VSEARCH tool to cluster reads showing a 95, 97, and 99 percent similarity for every 100

bases to our OF2 isolate 16S sequence (29). This provided us with 29246, 4400, and 510 hits respectively from a total of 72 different ecosystems with which we were able to measure overall abundance and average proportion in microbiome samples. We aligned the 510 sequences with 99 percent similarity and constructed a population level tree using a GTR+G+I model for 16S with the SPLITSTREE tool and rooted at the midpoint (Fig. 1) (43).

Antibiotic Resistance Testing

Following the protocol and according to the interpretive standards outlined by the manufacturer (standardized in April 1999), antibiotic resistance testing of OF2 was done with BBL Sensi-Disc Antimicrobial Susceptibility Test Discs (Becton-Dickinson, Sparks, Maryland). A total of 18 antibiotics were tested against OF2, including: ampicillin, aztreonam, bacitracin, cefotaxime, cefoxitin, chloramphenicol, ciprofloxacin, colistin, erythromycin, kanamycin, nalidixic acid, neomycin, nitrofurantoin, piperacillin, polymyxin B, streptomycin, tetracycline, and vancomycin. Mueller-Hinton agar (Difco, Detroit, Michigan) was used as the growth medium and the control organisms used in the experiment were *Escherichia coli* ATCC 25922 and *Staphylococcus aureus* ATCC 25923.

Results

Genome and 16S rRNA sequencing of the OF2 isolate

Twenty years after they were originally isolated, we re-examined the oil fly gut bacteria using modern sequencing techniques (Table 1). Thirteen of the original oil fly isolates, designated OF1 to OF14, were still viable and single colony isolates of each were obtained for identification by 16S rRNA sequencing (Table 1). With the exception of OF2, the identifications by 16S sequencing confirmed or closely approximated those obtained previously by the BBL

Enterotube II and API 20E systems and fatty acid analysis (25). With more complete sequence databases amassed over the last 20 years, the nine isolates previously identified as *Providencia rettgeri* were now recognized as *Providencia vermicola* and the isolate originally identified as *Pseudomonas maltophila* is now more accurately identified as *Pseudomonas aeruginosa* (Table 1). The nine *Providencia* isolates were not clonally related in that they had previously given different identities by their fatty acid profiles and BBL Enterotube II analyses. OF2 was chosen for whole genome sequencing, based on its unique 16S rRNA sequence and resistance to antibiotics, to characterize its bacterial lineage and identify presumptive antibiotic resistance genes and efflux pumps. Thus, we are merging our antibiotic resistance and phenotypic data with corresponding genomic data highlighting the novelty of OF2 and its genomic diversity with a view towards bolstering the evidence that antibiotic resistance may have arisen from niche systems such as the La Brea tar pits that force bacteria to combat chemically similar stressors. (Table 1).

OF2 as a new genus, *Candidatus Petroalcaligenes labreaensis*, and pangenome relationships

The bacterium designated OF2 was hard to identify during previous studies as it did not have close matches based on any of the bacterial identification systems implemented (25). During our initial 16S identification this view was strengthened, the closest hit was to an uncharacterized *Alcaligenes* species (Table 1). When constructed, the phylogenetic tree of all 16S sequences exhibiting 95% base pair similarity to our OF2 isolate (Fig. 1) showed that there were clearly defined lineages for Alcaligenes, Paenalcaligenes, and a clade of sequences exhibiting homology to our OF2 isolate.

We continued our phylogenetic characterization with a larger concatenated gene phylogeny and pangenome analysis (Fig. 2). The phylogeny shown in figure 2 was built using

concatenated single copy genes from 32 different taxa. Compared to the 16S phylogeny (Fig. 1), which is built using about 1500 nucleotide positions, the concatenated tree is much more informative. According to our pangenome analysis, OF2 contains a unique set of approximately 1130 gene clusters that are not found in the other taxa sampled. Of these unique gene clusters, we were able to classify and sort 421 into functional categories using BlastKOALA (38) (Fig. 3). Though multiple gene clusters may code for the same function, using a homogeneity index from ANVI'O (35), each cluster meets the threshold to be described as a different homologous gene. In addition to unique gene clusters, unique functions were also analyzed using ANVI'O. There are 50 functions that are positively enriched in OF2 and 7 functions that are negatively enriched when compared to the rest of the pangenome. These positively enriched genes include functions that have to do with transport, osmoregulation, metabolism, and antibiotic resistance. The 7 functions that were absent from OF2 include gene relating to amino acid biosynthesis, queuosine biosynthesis, and queuosine tRNA modification (Suppl Table. 1.1). Considering both the more robust phylogeny and the pangenome analysis, it is clear that OF2 clusters by itself and should be considered a new genus.

The above conclusion is further backed up when we look at *Paenalcaligenes hominis.* According to 16S phylogenies, *P. hominis* is the most closely relate genus to OF2, and like OF2 is closely related to Alcaligenes (Fig 1.) (44). The average amino acid identity of OF2 versus *P. hominis* is 64% across 52.1% of the proteome (1854 of 3558 proteins) while OF2 versus *Alcaligenes faecalis* is 69.9% across 50.8% of the proteome. For context, the average amino acid identity of *P. hominis* versus *A. faecalis* is 69.8% across 56.6% of the proteome (45, 46). Thus, OF2 is as different from *P. hominis* and *A. faecalis* as they are from each other, which in conjunction with Fig 2. leads us to suggest that OF2 warrants its own genus, possibly

Petroalcaligenes. *P. hominis* is not included in either analysis contained in figure 2 because to our knowledge, an assembled genome from the genus Paenalcaligenes that meets our stringent sampling criteria does not exist at this time.

Pathway Mapping: Giving Depth to the Unique Clusters and Enriched Functions

In order to give context to the pangenome analysis done with OF2, its genome was mapped against all of the available pathways using the KEGG Mapper tool (38). Of the pathways available there were zero instances of OF2 having a complete pathway that the combined 12 genomes did not have. On the other hand, there were an array of pathways consisting mostly of amino acid synthesis that OF2 was lacking compared to the other 12 genomes. The lack of these specific pathways was consistent with the ANVI'O functional enrichment analysis.

Environmental Sequence Search for OF2-like Sequences

Homologous hits in curated public 16S databases for our OF2 isolate have not been present, so we decided to search raw data sets of 16S sequencing. We downloaded and searched through 422877 public 16S datasets for homologous sequences to our OF2 isolate. We found that OF2 was most commonly found associated with fly and insect data sets (including house fly guts and bodies) and at lower abundances in sediment and plant samples (Fig. 4). OF2 was rarely found on other host organisms, but while it was found on plants in a few data sets there were rare occasions when the OF2 isolates were found in abundances between .01 and .08 percent (Fig. 5).

Antibiotic resistance pattern for OF2

OF2 was not included among the bacterial antibiotic resistance patterns identified previously (26) because at the time OF2 and OF3 were thought to be close relatives and OF3 was tested in its place. This assumption turned out to be an incorrect, and accordingly we have now tested OF2 against 18 antibiotics (Table 1). OF2 was resistant to 8 antibiotics: bacitracin,

chloramphenicol, ciprofloxacin, erythromycin, neomycin, nitrofurantoin, streptomycin, and vancomycin. OF2 was sensitive to 9 antibiotics: ampicillin, aztreonam, cefotaxime, cefoxitin, colistin, nalidixic acid, piperacillin, polymyxin B, and tetracycline, and showed intermediate susceptibility to kanamycin. This resistance/sensitivity pattern differs from those exhibited by any of the other oil fly bacterial isolates (26).

Discussion

We report here the results of the sequencing and analysis of OF2 in an attempt to use this information to explore more deeply its resistance to antibiotics and organic solvents. This information expands our fundamental understanding of the mechanisms, origin, and evolution of antibiotic resistance. Larval guts of oil flies are adapted to the highly viscous tar of the La Brea tar pits which constitute a unique habitat requiring unique genetic capabilities. Our current work reinforces earlier evidence (25, 26) that identifying natural reservoirs where those antibiotic resistance genes can be found provides information on the possible evolutionary origins of such resistances. The results of our phylogenetic and pangenomic analysis of the oil fly bacterium OF2 not only identifies an array of unique cellular processes when compared to bacteria from the family Alcaligenaceae, but also shows that a new bacterial genus is warranted. Not only does OF2 warrant placement in a novel genus, according to a search of ecological datasets using 16S RNA, it is exceedingly rare in the environment. Being mostly associated with insects and plants at very low levels.

Many efflux pumps are noted for the broad range of substrates they can extrude (47). The composition of the La Brea tar is highly weathered with very few linear alkanes remaining. Thin

layer chromatographic analysis has showed that the La Brea tar is composed of 10% branched alkanes and alkylated cyclic alkanes, 47% aromatic, 30% resins, and 13% polars, while GC/MS analysis showed significant hopanes, phenanthrene, and C1, C2, and C3 phenanthrenes (27). Many of these compounds are highly toxic and/or mutagenic to bacteria. What would be more natural than finding polyaromatic hydrocarbon-extruding bacteria in the guts of oil fly larvae which are continuously ingesting tar/asphalt estimated to be 47% aromatic? If weathered petrochemicals such as the La Brea tar pits were an original selective pressure for antibiotic resistant bacteria, then they are also a continuing source for that selective pressure. This line of reasoning supports the conclusion (48) that the complete eradication of antibiotic resistance in populations of microbes following reduced selective pressure from antimicrobials would not be straightforward. Finally, our previous research output in this area (25,26) and our current work characterizing the OF2 oil fly gut isolate illustrate the potential for finding novel efflux pumps. A minor limitation of this study is that these findings were based on the analysis of only 40 oil fly larvae taken from two locations at the La Brea tar pits on three occasions. Future studies could address the diversity of bacterial isolates in other tar pits, or petrochemical seeps inhabited with the oil fly on a larger scale. There could be a great many other resistance mechanisms out there waiting to be discovered.

Acknowledgments

Data on the chemical composition of the La Brea tar from reference 27 was provided by Roger C. Prince, ExxonMobil, Annandale, New Jersey. We thank Alyssa Damke for her technical assistance during the antibiotic resistance testing portion of this study.

References

1. O'Neill J. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. Rev Antimicrob Resist. 2014; Available from: http://amr-review.org/Publications.
2. United Nations General Assembly. Political Declaration of the High-Level Meeting of the General Assembly on Antimicrobial Resistance: draft resolution / submitted by the President of the General Assembly. Decision. 2016.Available from: https://digitallibrary.un.org/record/842813
3. Interagency Task Force for Combating Antibiotic-Resistant Bacteria. National Action Plan for Combating Antibiotic-Resistant Bacteria. United States Government Publishing Office (GPO), Washington, DC, United States of America, 2013. 63 pp. Available from: https://obamawhitehouse.archives.gov/sites/default/files/docs/national_action_plan_for_combating_antibotic-resistant_bacteria.pdf (accessed on 18 January 2019).
4. Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J, Handelsman J. Call of the wild: antibiotic resistance genes in natural environments. Nature Rev Microbiol. 2010; 8: 251-259.
5. D'Costa VM, McGrann KM, Hughes DW, Wright GD. Sampling the antibiotic resistome. Science. 2006; 311: 374–377. doi:10.1126/ science.1120800
6. D'Costa VM, Griffiths E, Wright G.D. Expanding the soil antibiotic resistome: Exploring environmental diversity. Curr. Opin. Microbiol. 2007; 10: 481–489. doi:10.1016/j.mib.2007.08.009
7. Durso LM, Wedin, D, Gilley JE, Miller DN, Marx DB. Assessment of selected antibiotic resistances in ungrazed native Nebraska prairie soils. J Environ Qual. 2016; 45: 454-462. doi:10.2134/jeq2015.06.0280
8. Cytryn, E. The soil resistome: The anthropogenic, the native, and the unknown. Soil Biol. Biochem. 2013; 63: 18–23. doi:10.1016/j.soilbio.2013.03.017
9. D'Costa VM, King CE, Kalan L, Morar M, Sung WWL, Schwarz C, et al. Antibiotic resistance in ancient. Nature. 2011; 477: 457–461.
10. Bhullar K, Waglechner N, Pawlowski A, Koteva K, Banks ED, Johnston MD et al. Antibiotic resistance is prevalent in an isolated cave microbiome. PLoS ONE. 2012; 7, 2012;e34953.

11. Durso LM, Miller DN, Wienhold BJ. Distribution and quantification of antibiotic resistant genes and bacteria across agricultural and nonagricultural metagenomes. PLoS One. 2012; 7:e48325. doi:10.1371/journal.pone.0048325
12. Knapp CW, Dolfing J, Ehlert PAI, Graham DW. Evidence of increasing antibiotic resistance gene abundances in archived soils since 1940. Environ Sci Technol, 2010; 44: 580-587.
13. Gillings MR Integrons: Past, Present, and Future. Microbiol and Molecular Biol Rev. 2014; 78: 257-277.
14. Perry J, Waglechner N, Wright G. The pre-history of antibiotic resistance. Cold Spring Harbor Perspectives in Medicine. 2016; 6(6):a025197.
15. Waksman SA. Antagonistic relations of microorganisms. Bact Rev. 1941; 5: 231-291.
16. Davies J. Are antibiotics naturally antibiotics? J Ind Microbiol Biotechnol. 2006:33:496-9
17. Ghoul M, Mitri S. The ecology and evolution of microbial competition. Trends in Microbiol. 2016; 2016; 24:833-845.
18. Martinez, JL. Antibiotics and antibiotic resistance genes in natural environments. Science. 2008; 321: 365-367.
19. Aminov RI. The role of antibiotics and antibiotic resistance in nature. Environ Microbiol. 2009; 11: 2970-2988.
20. Davies J, Davies, D. Origins and evolution of antibiotic resistance. Microbiol and Molec Biol Rev. 2010; 74: 417-433.
21. Chopard L. La mouche du petrole et les questions qu'elle pose. La Nat (Paris). 1963; 3338: 255–256
22. Thorpe WH. The biology of the petroleum fly (Psilopa petrolei). Trans Entomol Soc London. 1930; 78:331–344.
23. Thorpe WH. The biology of the petroleum fly. Science. 1931; 73: 101-103.
24. Hogue CL. Insects of the Los Angeles basin. Third Edition. Natural History Museum Foundation, Natural History Museum of Los Angeles County, Los Angeles, Calif; 1974.
25. Kadavy DR, Plantz B, Shaw CA, Myatt J, Kokjohn TA, Nickerson KW. Microbiology of the Oil Fly, Helaeomyia petrolei. Appl Environ Microbiol. 1999; 65: 1477-1482.

26. Kadavy DR, Hornby JM, Haverkost T, Nickerson KW. Natural antibiotic resistance of bacteria isolated from larvae of the oil fly, Helaeomyia petrolei. Appl Environ Microbiol. 2000; 66: 4615-4619.
27. Nickerson K.W., Plantz B. Microbiology of Oil Fly Larvae. In: Krell T. (eds) Cellular Ecophysiology of Microbe. Handbook of Hydrocarbon and Lipid Microbiology. Springer, Heidelberg; 2017. Doi: 10.1007/978-3-319-20796-4 37-1.
28. Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166-169.
29. Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584.
30. Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., ... & Fenton, J. (2015). The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research*, *4*.
31. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pyshkin, A. V. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), 455-477.
32. Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.
33. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069.
34. NCBI Resource Coordinators (2017). Database resources of the national center for biotechnology information. *Nucleic acids research*, *45* (Database issue), D12.
35. Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319
36. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research. 2004; 32(5), 1792-97.
37. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.
38. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. Journal of molecular biology. 2016 Feb 22;428(4):726-31.

39. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one. 2010 Mar 10;5(3):e9490.
40. Rambaut A. FigTree v1. 4.
41. Kanehisa M, Sato Y. KEGG Mapper for inferring cellular functions from protein sequences. Protein Science. 2020 Jan;29(1):28-35.
42. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017 Nov;551(7681):457.
43. Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics (Oxford, England)*, *14*(1), 68-73.
44. Kämpfer P, Falsen E, Langer S, Lodders N, Busse H-J. Paenalcaligenes hominis gen. nov., sp. nov., a new member of the family Alcaligenaceae. Intl J Syst Evol Microbiol. 2010; 60: 1537-1542.
45. Rodriguez LM., Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Prepr. 2016; 4, e1900v1
46. Medlar AJ, Toronen P, Holm L. AAI-profiler: fast proteome-wide exploratory analysis reveals taxonomic identity, misclassification and contamination. Nucleic Acids Res. 2018; 46:W479-W485.
47. Du D, Wang-Kan X, Neuberger A, vanVeen HW, Pos KM, Piddock LJV, Luisi BF. Multidrug efflux pumps: structure, function, and regulation. Nature Rev Microbiol. 2018; 16: 523-539.
48. Holmes AH, Moore LS, Sundsfjord A, Steinbakk M, Regmi S, Karkey A, Guerin PJ, Piddock LJ. Understanding the mechanisms and drivers of antimicrobial resistance. Lancet. 2016; 387: 176-187.

Table 1 - Overview of the classification and antibiotic resistance annotations of the 13 Oil fly bacterial isolates[a] tested in this study.

Strain	Original Classification	16S Classification	Number of Antibiotic Resistances	NRRL Culture Collection #	NCBI Accession #
OF001	Providencia rettgeri	Providencia vermicola	12	B-65562	NA
OF002	Acinetobacter [b] lwoffi	Uncultured Alcaligenes sp.	8[c]	B-65563	MN527032
OF004	Pseudomonas [b] maltophila	Pseudomonas aeruginosa	NA	NA	MN547155
OF005	Providencia rettgeri	Providencia vermicola	11	B-65564	MN547314
OF006	Providencia rettgeri	Providencia vermicola	11	B-65565	MN547625
OF007	Providencia rettigeri	Providencia vermicola	12	B-65567	NA
OF008	Morganella morganii	Morganella morganii	10	B-65567	MN547625
OF009	Providencia rettgeri	Providencia vermicola	9	B-65568	NA
OF010	Providencia rettgeri	Providencia vermicola	11	B-65569	MN547993
OF011	Providencia rettgeri	Providencia vermicola	11	B-65570	NA
OF012	Providencia rettgeri	Providencia vermicola	12	B-65571	NA
OF013	Undetermined	Uncultured Providencia sp.	8	B-65572	NA
OF014	Providencia rettgeri	Providencia vermicola	11	B-65573	NA

A/ Strain numbers and original classifications from Kadavy 1999 (25) and antibiotic resistances are from Kadavy 2000 (26). 16S classification is from this paper. Strain OF003 was no longer viable after 20 years in storage. B/ Both OF002 and OF004 were classified as nonenteric by API 20E and as either *Acinetobacter lwoffi* or *Pseudomonas maltophila* by Enterotube. C/ OF002 is resistant to 8 of 17 antibiotics (this paper) while the other strains are resistant to 8-12 of 23 antibiotics from Kadavy 2000 (26).

Figure 5

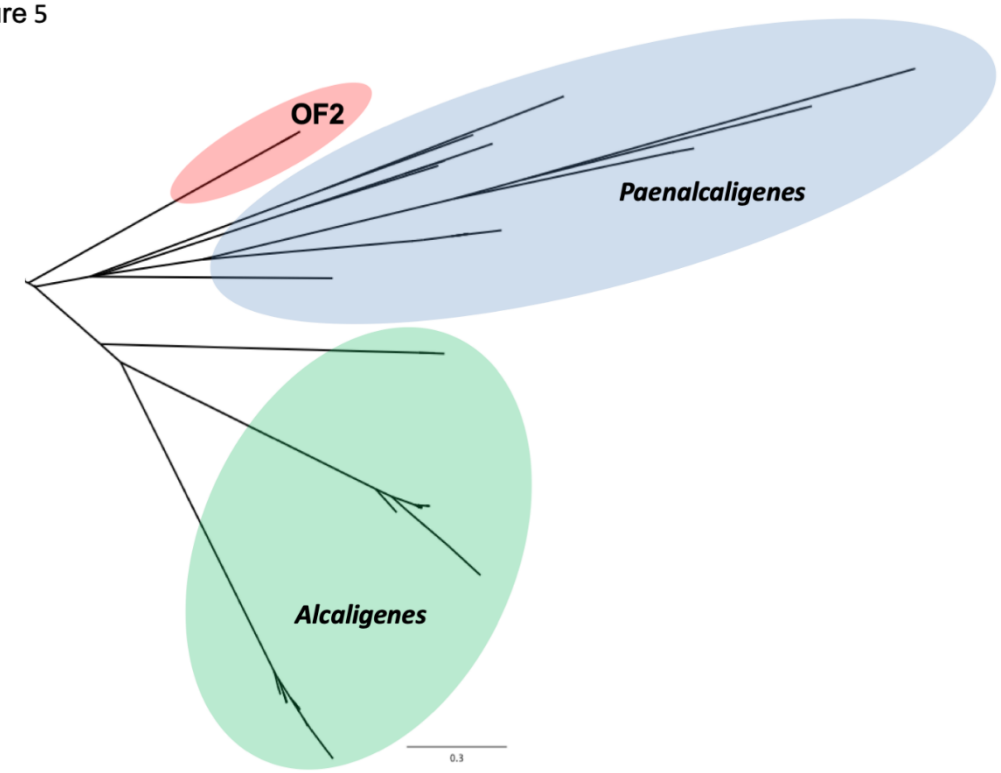


Fig. 1. Maximum likelihood analysis of the Alcaligenes, Paenalicaligenes, and Candidatus Petroalcaligenes genera

Phylogenetic tree of 16S sequences taken from NCBI denoting the placement of OF2 relative to the Alcaligenes and the Paenalcaligenes taxa in the Alcaligenaceae.

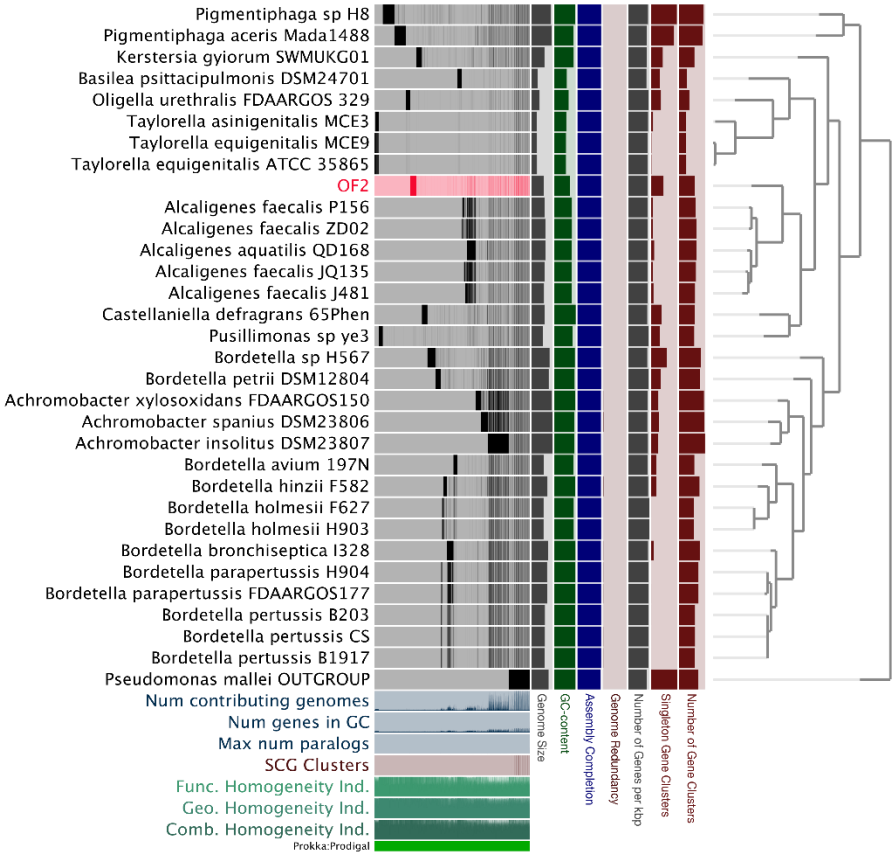


Fig. 2. Pangenome analysis of the Alcaligenaceae

On the right, a phylogenetic tree consisting of concatenated single copy from 31 taxa related to OF2. On the left, a Pangenome description of the Alcaligenaceae. The grey bars closest to the taxa names highlight the unique gene clusters for each strain.

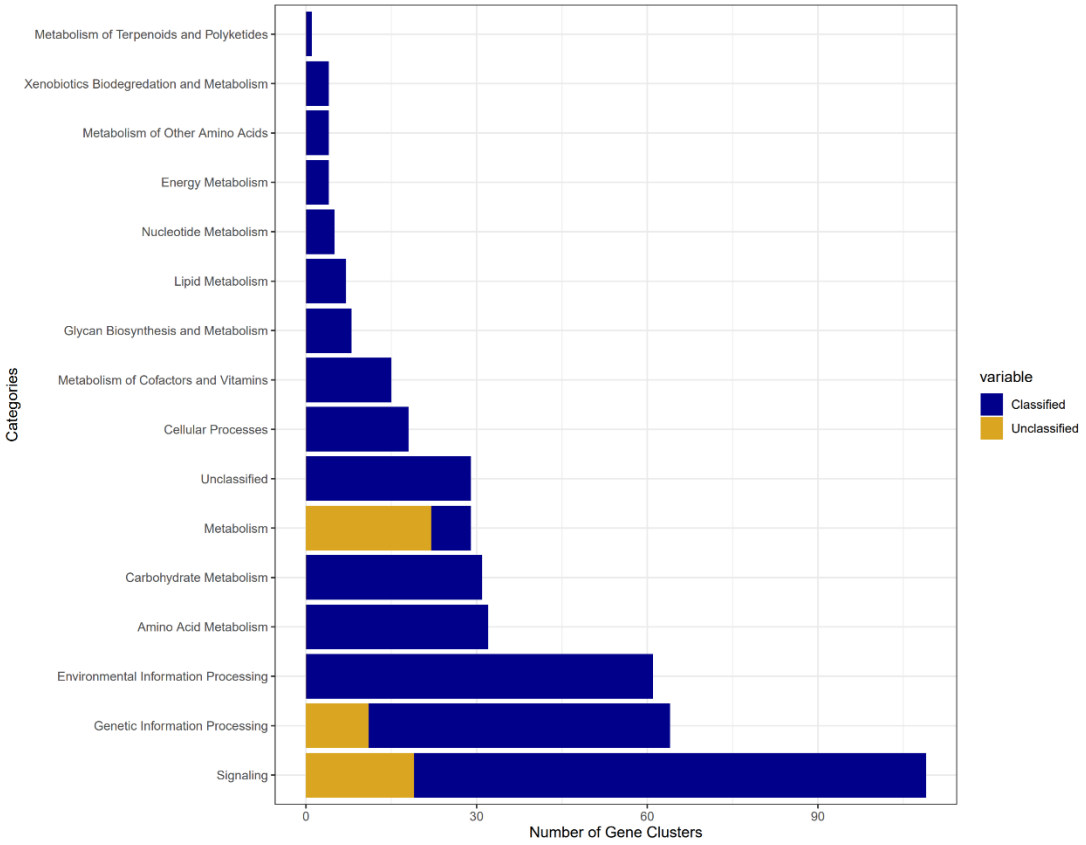


Fig. 3 Functional Categories of OF2's Unique Gene Clusters

Gene clusters unique to OF2 were annotated by BlastKOALA and then assigned a functional category based on KEGG orthology. The unclassified genes are able to be placed in a category, but do not have a putative or known function.

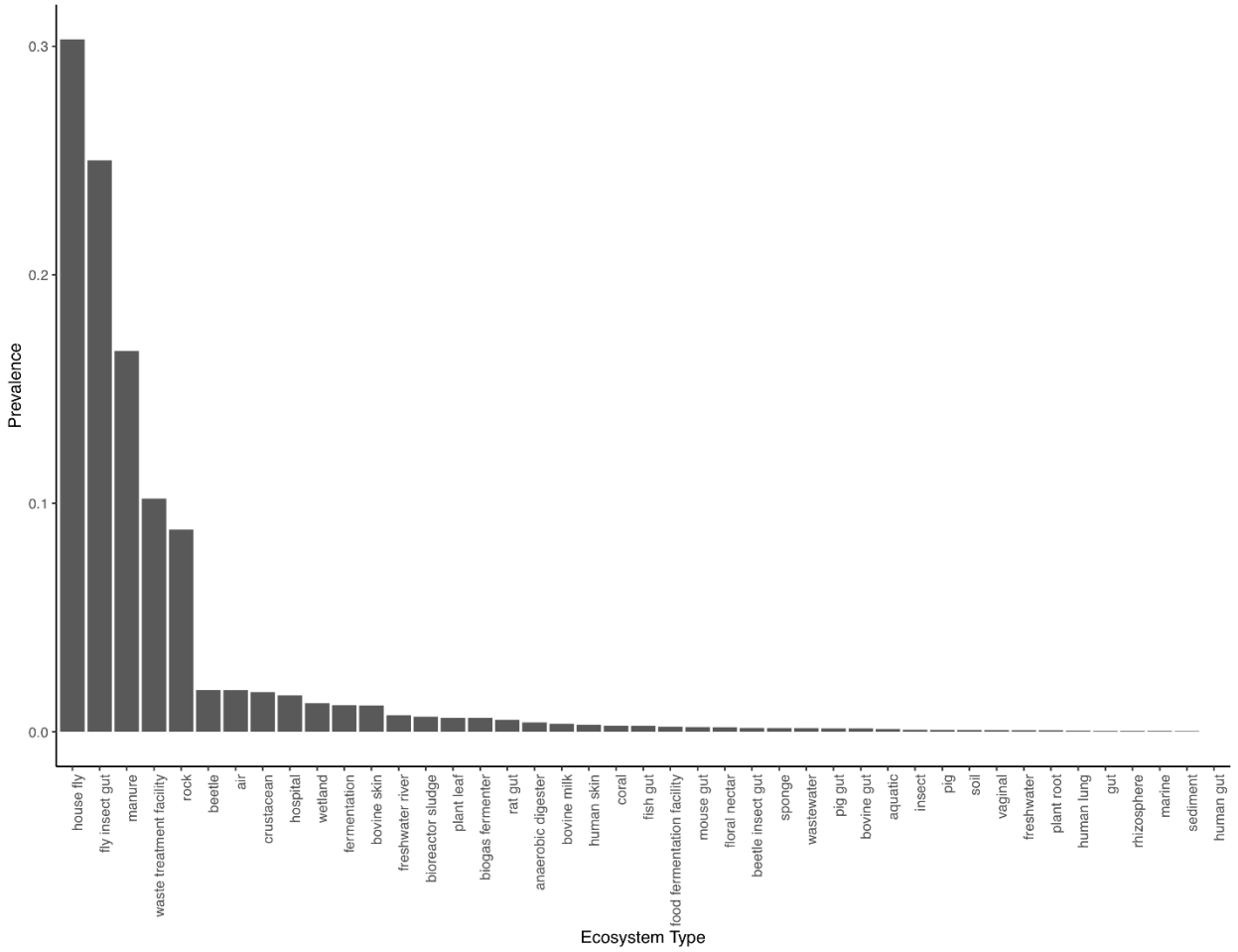


Fig. 4 Prevalence of OF2's 16S rRNA in publicly available ecological datasets

OF2's 16S rRNA sequence was queried against 422877 public datasets with environment data. Hits with 99% similarity were used to construct this graph showing that OF2 is most closely associated with flies, manure, waste treatment facilities, and rock albeit at a low level.

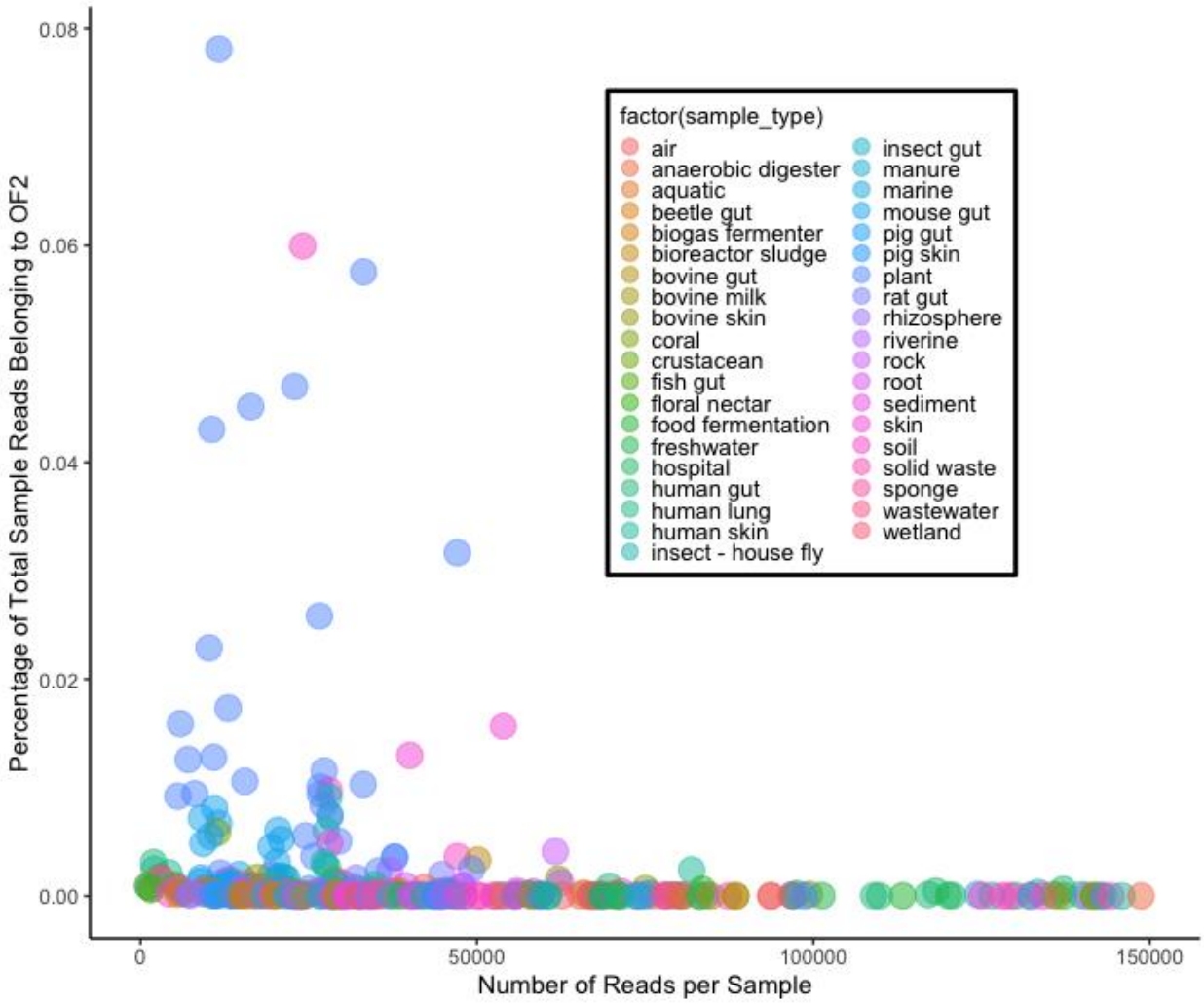


Fig. 5 Depth of the OF2 OTU at 99% similarity versus size of sample

OF2's abundance in environmental databases when mapping against 16S rRNA. Showing that when OF2 is found in natures, it is found at low abundances.

Table 1 - Overview of the classification and antibiotic resistance annotations of the 13 Oil fly bacterial isolates[a] tested in this study.

Strain	Original Classification	16S Classification	Number of Antibiotic Resistances	NRRL Culture Collection #	NCBI Accession #
OF001	Providencia rettgeri	Providencia vermicola	12	B-65562	NA
OF002	Acinetobacter [b] lwoffi	Uncultured Alcaligenes sp.	8[c]	B-65563	MN527032
OF004	Pseudomonas [b] maltophila	Pseudomonas aeruginosa	NA	NA	MN547155
OF005	Providencia rettgeri	Providencia vermicola	11	B-65564	MN547314
OF006	Providencia rettgeri	Providencia vermicola	11	B-65565	MN547625
OF007	Providencia rettigeri	Providencia vermicola	12	B-65567	NA
OF008	Morganella morganii	Morganella morganii	10	B-65567	MN547625
OF009	Providencia rettgeri	Providencia vermicola	9	B-65568	NA
OF010	Providencia rettgeri	Providencia vermicola	11	B-65569	MN547993
OF011	Providencia rettgeri	Providencia vermicola	11	B-65570	NA
OF012	Providencia rettgeri	Providencia vermicola	12	B-65571	NA
OF013	Undetermined	Uncultured Providencia sp.	8	B-65572	NA
OF014	Providencia rettgeri	Providencia vermicola	11	B-65573	NA

A/ Strain numbers and original classifications from Kadavy 1999 (25) and antibiotic resistances are from Kadavy 2000 (26). 16S classification is from this paper. Strain OF003 was no longer viable after 20 years in storage. B/ Both OF002 and OF004 were classified as nonenteric by API 20E and as either *Acinetobacter lwoffi* or *Pseudomonas maltophila* by Enterotube. C/ OF002 is resistant to 8 of 17 antibiotics (this paper) while the other strains are resistant to 8-12 of 23 antibiotics from Kadavy 2000 (26).

Figure 5

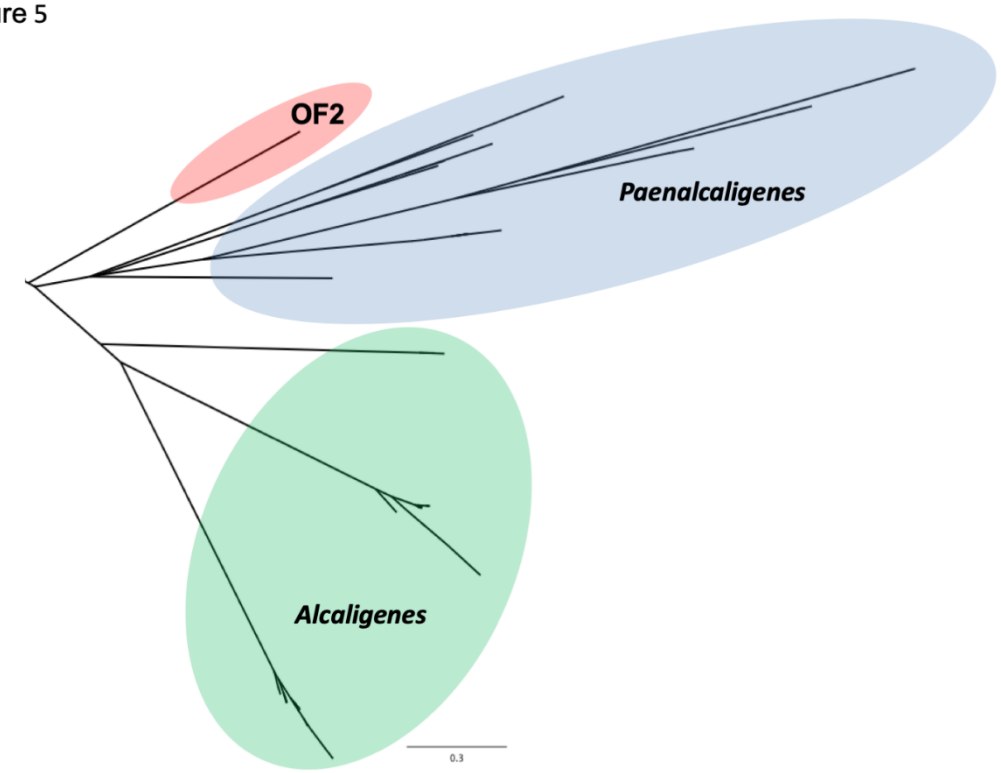


Fig. 1. Maximum likelihood analysis of the Alcaligenes, Paenalicaligenes, and Candidatus Petroalcaligenes genera

Phylogenetic tree of 16S sequences taken from NCBI denoting the placement of OF2 relative to the Alcaligenes and the Paenalcaligenes taxa in the Alcaligenaceae.

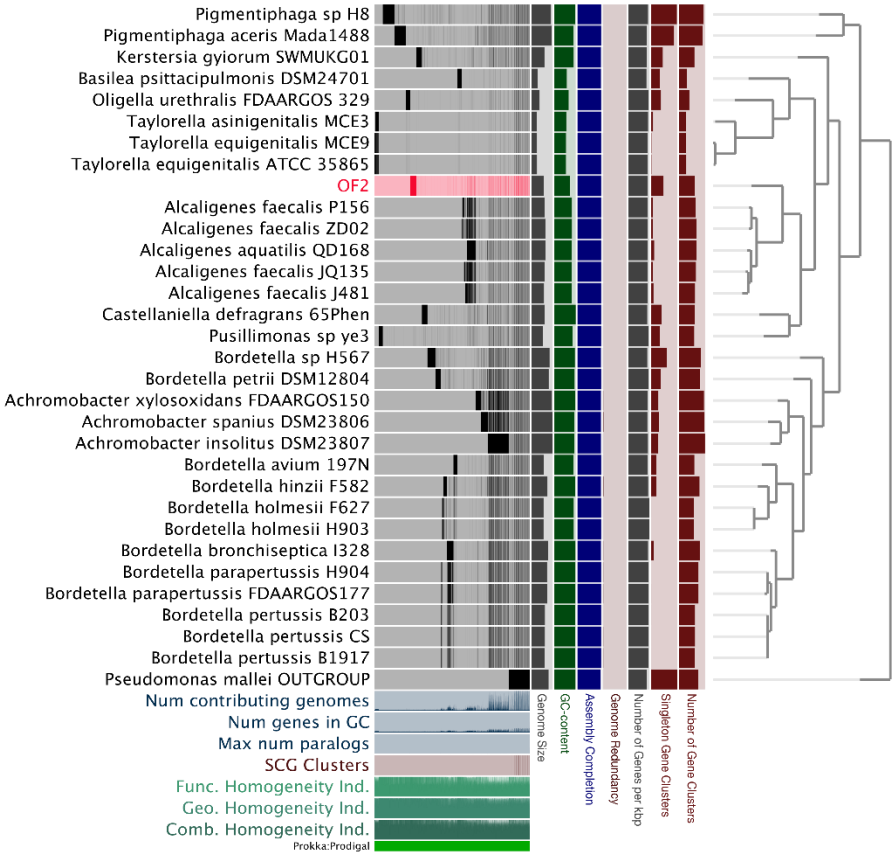


Fig. 2. Pangenome analysis of the Alcaligenaceae

On the right, a phylogenetic tree consisting of concatenated single copy from 31 taxa related to OF2. On the left, a Pangenome description of the Alcaligenaceae. The grey bars closest to the taxa names highlight the unique gene clusters for each strain.

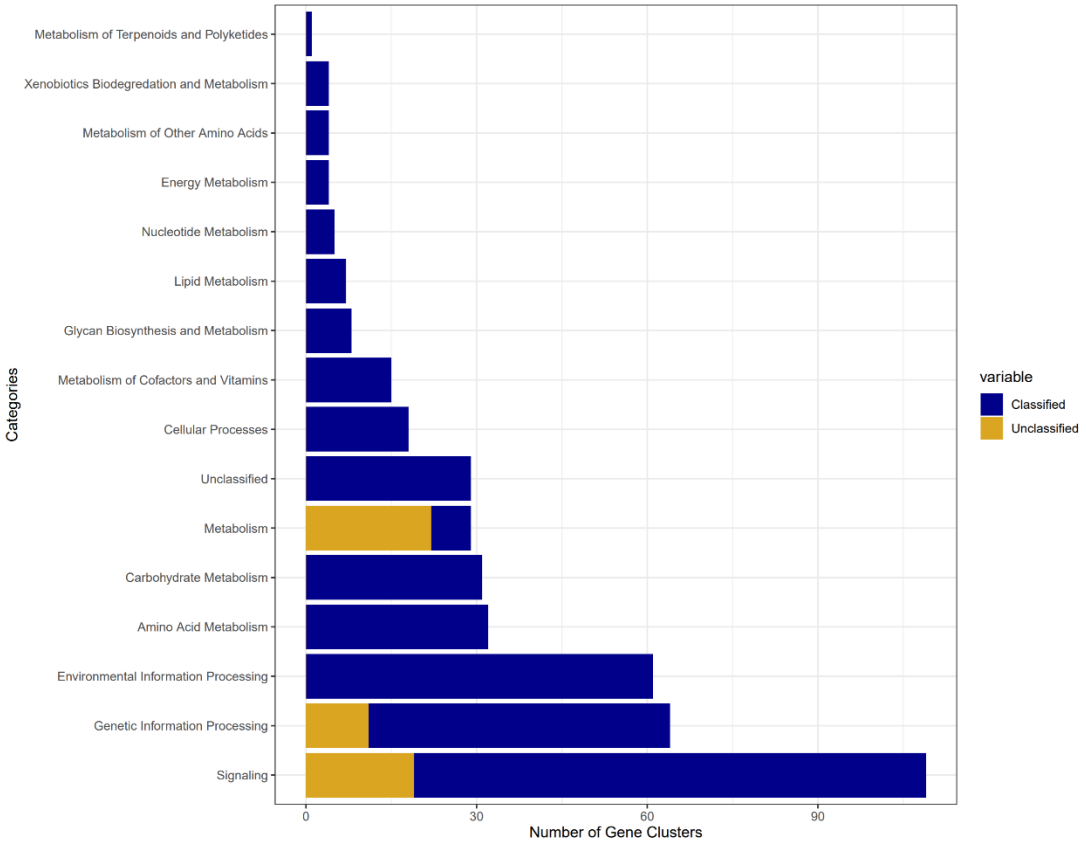


Fig. 3 Functional Categories of OF2's Unique Gene Clusters

Gene clusters unique to OF2 were annotated by BlastKOALA and then assigned a functional category based on KEGG orthology. The unclassified genes are able to be placed in a category, but do not have a putative or known function.

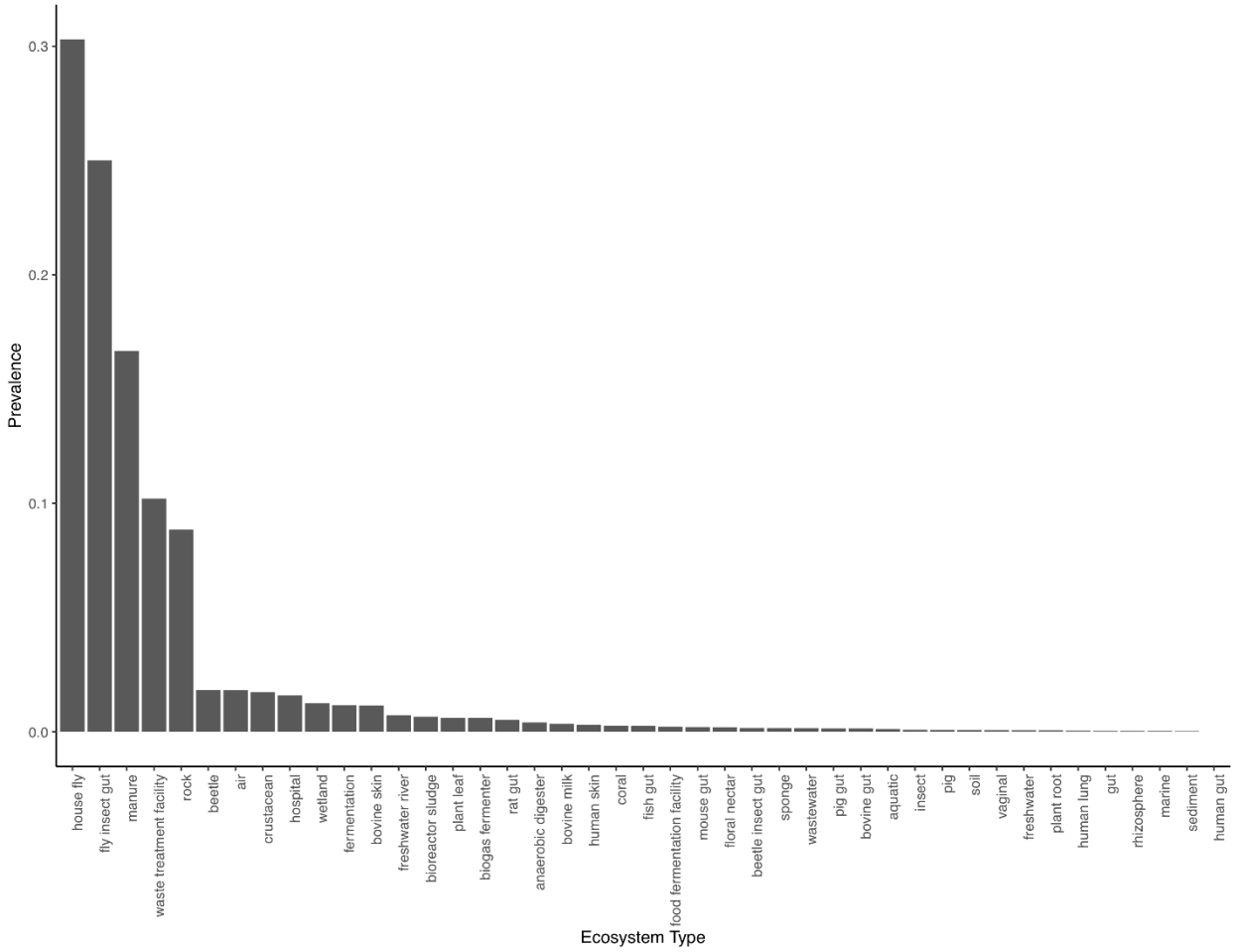


Fig. 4 Prevalence of OF2's 16S rRNA in publicly available ecological datasets

OF2's 16S rRNA sequence was queried against 422877 public datasets with environment data. Hits with 99% similarity were used to construct this graph showing that OF2 is most closely associated with flies, manure, waste treatment facilities, and rock albeit at a low level.

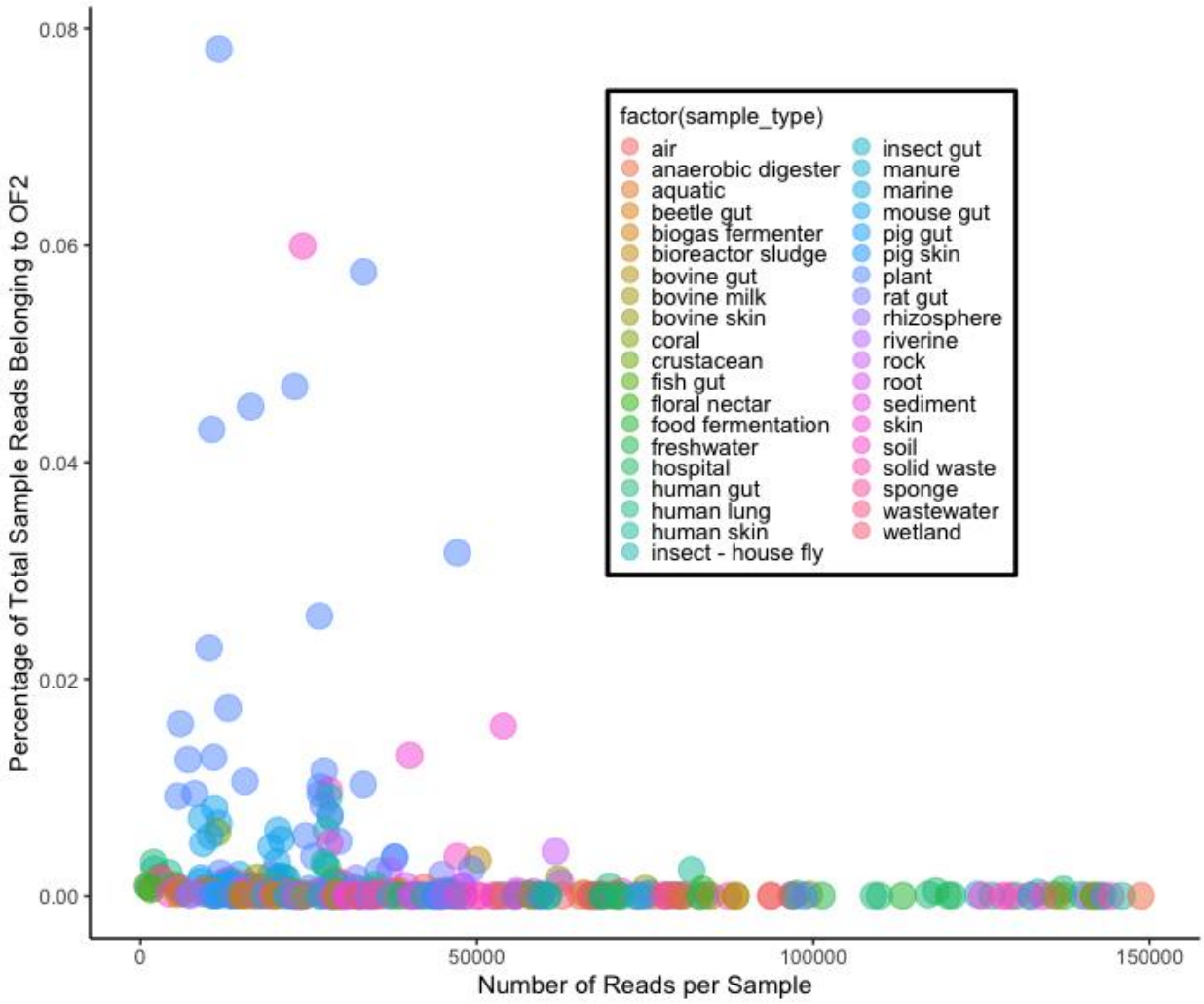


Fig. 5 Depth of the OF2 OTU at 99% similarity versus size of sample

OF2's abundance in environmental databases when mapping against 16S rRNA. Showing that when OF2 is found in natures, it is found at low abundances.

CHAPTER 2- Pangenome analysis of *Providencia* isolates derived from the gut flora of the oil fly *Helaeomyia petrolei* from the La Brea tar pits

Brian Dillard[1], Lisa Durso[2], and Kenneth W. Nickerson[1]*

[1] School of Biological Sciences, University of Nebraska, Lincoln, NE 68588.

[2] USDA-ARS, University of Nebraska, Lincoln, NE 68583.

*Corresponding author.

Kenneth W. Nickerson

Biological Sciences, University of Nebraska,

Lincoln, NE USA 68588-0666

402-472-2253 knickerson1@unl.edu

Abstract

Helaeomyia petrolei (oil fly) larvae live in the asphaltene rich oil seeps of Rancho La Brea, Los Angeles, California where they are constantly in the presence of aromatic hydrocarbons like toluene. These larvae can pass this oil through their digestive system without any harmful effects. Previous research has characterized the presence of bacteria in the oil fly gut, and isolates taken from the oil fly larvae were shown to be resistant to a large array of antibiotics. This study aims to determine whether an extremophilic environment like the La Brea oil seeps provides a positive selective pressure on these gut bacteria. Such a selective pressure could lead to new functional traits in an attempt to combat the solvent stress imposed by the tar. This study will combine phylogenomic and pangenomic analysis to compare a large array of *Providencia* against three strains isolated from the oil fly larvae gut - *P. rettgeri* OF5, *P. vermicola* OF6, and *P. vermicola* OF10. These tools will help us to better refine the current classification structure of *Providencia* while identifying functions within our isolates that are enriched or suppressed relative to other groups in the pangenome. The information provided will help us identify whether host derived bacteria change when confronted with a hostile environment.

Introduction

Extremophilic environments can spur the evolution and adaptation of bacterial communities (1), which may in turn can help to combat abiotic stressors for the host organism (2). This accelerated evolution to combat stress has been highlighted as an area of interest for industrial enzymes that are suitable for unique production processes (3) and as a way for environmental antibiotic resistance genes to develop (4).

Helaeomyia petrolei (oil fly) larvae live in the hydrocarbon and asphaltene rich asphalt seeps of the La Brea tar pits in Los Angeles California (5). Thorpe (6) previously identified that the oil fly larvae were living in what should be an extremely toxic environment. Specifically, he observed that the oil fly larvae spend much of this stage submerged in asphaltene rich oil and pass the oil through their digestive system with zero negative physiological effects. This asphalt is known to be composed of a wide variety of aromatic hydrocarbons including substances like toluene and anthracene. Many of these compounds readily pass through cellular membranes and thus can be highly toxic (7). These larvae have a complex microbial gut flora that is likely resistant to the aromatic hydrocarbons passing through the digestive system. For this reason, we have focused our attention on bacterial isolates from the larvae's gut, in order to understand the effects that this extreme environment will have on the evolution of the gut microbiota.

In previous studies we quantified the amount of bacteria in the oil fly larvae gut, classified select isolates phenotypically with Enterotube II and API20E systems, and characterized the antibiotic resistance profiles using zone of inhibition assays (5,8). Through this work, we have found that there are about 2×10^5 heterotrophic bacteria per larvae (5) and when 12 of the 13 bacterial strains were submitted to zone of inhibition testing of 23 different antibiotics they were

all resistant to about half (8). Because of *Providencia*'s status as a human pathogen and the large number of available genomes with which to compare the genomics of our isolates, we picked *Providencia rettgeri* OF5, *Providencia vermicola* OF6, and *Providencia vermicola* OF10 to study further. More recently as a part of this work, we first did full length 16S sequencing to better taxonomically classify our oil fly isolates, and then completed whole genome sequencing for a subset of our isolates. The strains OF5, OF6, and OF10 identified above, were loosely classified via their 16S sequences, and then placed with a more robust concatenated phylogeny using the whole genome sequencing results.

Using a combination of pangenomics and phylogenomics, we plan to compare our *Providencia* oil fly isolates to a host of other *Providencia* strains derived from different eukaryotic hosts. Using functional enrichment analysis, we will compare different subsets of isolates against the greater pangenome and determine if either group has specific functions enriched or suppressed. We hypothesize that the environment in which the host lives will influence the genomic capabilities of each group of organisms to a greater extent than the host within which these bacteria live. In addition, using this large number of sampled isolates, we can use phylogenomics to confirm the placement of certain strains into the correct species clades.

Materials and Methods

Sample Collection, Sequencing, and Assembly

As described previously by Kadavy et al., 40 Oil fly larvae were collected from the La Brea tar pits in California between the years 1994 and 1997 (5). These larvae were shipped alive back to the laboratory in the presence of oil also collected from the La Brea tar pits. The larvae were sustained in large petri dishes with a thin layer of oil and fed 40 mg of egg meat medium

(Difco, Detroit, MI). In preparation for sampling the bacterial gut contents, larvae were surface sterilized through a series of washes including linoleic acid, 70% ethanol, 15% bleach supplemented with Tween 20, and phosphate buffered saline supplemented with Tween 20. Sterile larvae were homogenized and streaked onto either yeast extract and peptone (YEP) or Luria-Bertani (LB) agar plates. Plates were incubated for 24 hours at 37°C, colonies were picked, and then stored as glycerol stocks at -80°C. Coming out of the -80°C freezer, isolates were streaked and incubated at 37°C overnight three times consecutively to confirm strain purity.

In order to confirm the strain identifications previously described by Kadavy et al., 16S sequencing was performed on select isolates (5). Strains OF5, OF6, and OF10 were grown on LB agar overnight at 37°C, and then genomic DNA was extracted according to the manufacturer's instructions with the QIA-Amp Power Fecal DNA kit (product number 12830; Qiagen Inc., Germantown, MD). The extracted DNA was shipped overnight on dry ice to Molecular Research LP (Shallowater, TX) for full length 16S ribosomal DNA sequencing. DNA amplification was carried out with a 35-cycle PCR using 27F and 1492R primers, and Hotstar Taq Plus Master Mix (Qiagen, Germantown, MD). DNA quality was checked on a 2% agarose gel and then purified using Ampure PB beads (Pacific Biosciences, Menlo Park, CA). Libraries were made with the DMRTbell library kit (Pacific Biosciences, Menlo Park, CA) and then sequenced according to the manufacturer's instructions on a PacBio sequel. Overlapping forward and reverse reads were merged using the PacBio Circular Consensus Sequencing algorithm.

Of the isolates chosen for 16S sequencing, three *Providencia* isolates were chosen for whole genome sequencing. Cell pellets of OF5, OF6, and OF10 were sent to The Sequencing Center (Fort Collins, CO) for DNA extractions and whole genome sequencing. Sequencing libraries were prepared with the Nextera XT Library Kit and Illumina Nextera XT Index Kit

(Illumina, San Diego, CA) and then sequenced on a MiSeq Sequencing System resulting in 2 X 250 bp paired-end reads. The MiSeq system subsequently removed the Nextera adapter sequences and subsequently demultiplexed the reads, resulting in paired-end FASTQ files.

For each of the *Providencia* isolates chosen, the 2 X 250 bp paired-end reads from the Illumina sequencing were quality checked with the FASTQC tool (9) and then assembled with SPADES (10), using the built in error correcting and k-mer base-pair lengths of 21, 33, 55, and 77. After assembly, the QUAST tool (11) was used to assess assembly quality.

Dataset construction and Annotation

In addition to the three oil fly isolates, we chose 49 genome assemblies from *Providencia* and 4 genome assemblies from *Morganella* from NCBI based on genome quality and the presence of biosample information providing the source of sample collection (Table 1). We sampled all the species present within *Providencia* that were available through the NCBI database, and we used *Morganella* as an outgroup. All 56 genomes were annotated with the PROKKA tool (12) using the default settings for bacterial genomes resulting in the standard GFF3 format.

Pangenome Construction

Using the annotations from the previously generated GFF3 files and the original assembly files downloaded from NCBI, we used ANVI'O tool (13) to generate a genome database of all 56 genomes. Using the default flags for gene homology, ANVI'O used the blastp algorithm (14) for identifying similarities in amino acid sequences, MCL (15) for gene clustering, and MUSCLE (16) for multiple sequence alignment to create a pangenome of the 56 genomes sampled.

Phylogenomic Analysis

Using ANVI'O (13), the 1011 single copy genes that were present in all of the 56 genomes were binned and the amino acid sequences were extracted. We concatenated the sequences for each genome and then aligned them using the multiple sequence alignment tool MUSCLE (16), resulting in an alignment sequence length of 297,642 amino acid residues. We used FastTree (17) to produce a maximum likelihood tree with a log likelihood of -1,118,668. The tree was imported into the ANVI'O pangenome as a layer, rearranged according to the topology, and then rooted with the *Morganella* clade.

Functional enrichment analysis

Again using ANVI'O (13) and based on the PROKKA (12) annotations, we identified functions within certain groups of genomes that were either positively or negatively enriched with a p-value cutoff of 5%. Genomes for this analysis included the *Providencia rettgeri* that are derived from eukaryotic hosts and the Oil Fly larvae isolates. We did not include *Providencia vermicola* P8538 as it did not group phylogenetically with the other rettgeri or vermicola strains. We tested eight groups of isolates that are outlined in table 2. The full output of the gene enrichment analysis for each of the groups is available in the supplemental materials (Suppl Material. 1).

Results

Pangenome and Phylogenomic Relationships

The *P. vermicola* species was first described in 2006 (18), and since then there have been no published whole genome sequencing efforts of the species. Currently the only way to bioinformatically classify a new isolate as *P. vermicola* is to compare the 16S rRNA to the type strain *P. vermicola* OP1. In Somvanshi's 2006 paper (18) describing the species, the type strain *P. vermicola* OP1 had a 99.5% 16S rRNA sequence similarity to *P. rettgeri*. While this sequence similarity doesn't necessarily meet the suggested threshold of 98.65% (19), this cutoff has been shown to inaccurate (20) and better tools such as pangenomics or more robust phylogenies have been recommended. Many of the currently classified *P. vermicola* strains have 16S rRNA sequence similarities of less than 99.5% with OP1. One such strain is P8538, the first whole genome sequence on NCBI claiming to be *P. vermicola*. When placed in a phylogenetic tree comparing the 16S rRNA sequences of other *Providencia* species, OP1 grouped most closely with *P. rettgeri*. As shown in figure 1, strain P8538 does not group closely with *P. rettgeri*. Instead strain P8538 groups by itself, and just based on the placement in figure 1, could likely be described as a new species. The clade from figure 1 that is diverging from the other *rettgeri* taxa and composed of strains OF6, OF10, RB151, FDAARGOS_330, and 594M10B all have at least 99.73% 16S rRNA sequence similarities to the *vermicola* type strain OP1 and

Functional Enrichment Analysis

The functional enrichment analysis looks at the presence and absence of gene functions (Fig 2.) within different groupings of isolates from the *P. rettgeri* and *P. vermicola* clades (table 2.). When looking at just insect derivatives, there is a shift in functional content when compared

to the Nonhuman Host grouping. Within the Insect Control grouping the larger number of suppressed functions is due to the absence of OF6 and OF10 without the inclusion of any *P. vermicola* strains on the side of the insects. If instead we look at the Oil Fly grouping, we can see that there is a greater number of enriched genes and a smaller number of suppressed genes. If we control for the different functions in each species clade by just comparing isolates from *P. rettgeri*, we can see that while Dmel1 has more enriched functions, OF5 has many fewer suppressed functions and should lead to an overall more complete set of *P. rettgeri* functional characteristics for OF5.

The Presumptive *vermicola* grouping has a large number of enriched and suppressed genes. This supports the idea that this clade of bacteria is sufficiently different from *P. rettgeri* to be classified as *P. vermicola*. The grouping representing OF6 and OF10 while different from the Presumptive *vermicola* group gains many of its enriched and suppressed functions because of its *vermicola* classification. However, unlike any other groupings tested OF6 and OF10 have 32 enriched and 28 suppressed transposases of varying types. Transposases are not present to this degree in any other enrichment analysis.

Discussion

To characterize the result of living in an extremophilic environment, we observed the phylogenetic groupings of *Providencia* (Fig 1.) and measured the functional differences across various groups of isolates (Fig 2.). This analysis should highlight any functional shifts that arise because of the host organism or environmental conditions. Overall, there were some functional differences for insect gut isolates as a whole and for the oil fly isolates. However, even with the

slight positive shift in functional enrichments for these two groupings (Fig 2.), there is no clear indication that the extreme environment or the host organism is the cause. More likely many of the differences seen are due to the functional divide between the *vermicola* and *rettgeri* clades. In fact, the large differentiation of enriched and suppressed genes between the two clades represented in the Presumptive vermicola grouping is further evidence that these classifications need to be reassessed. The large number of functions both enriched and suppressed between the *vermicola* and *rettgeri* clades shows that even though the 16S rRNA sequence similarity is relatively high compared to standard thresholds (19), *vermicola* is functionally distinct. As such, both of our *P. vermicola* oil fly isolates OF6 and OF10 should be the first available whole genome sequences for the species. Based on the phylogenetic placement and the low 16S rRNA sequence similarity to other *vermicola* isolates, strain P8538 described in table 1 should be classified as a new species.

While OF6 and OF10 had similar numbers of enriched and suppressed genes to the proposed *vermicola* group, there were a significant number of transposase genes in both categories. The transposase genes listed in the supplemental materials do not have a representative class or any identifiable unifying characteristic. This might suggest that there are high levels of genetic motility within the genome, or there are different plasmids present in the oil fly isolates than in other related taxa. Work done on the *Drosophila* isolates described in table 1 (21) outline the presence of varying plasmids across three different species of *Providencia*.

Overall, I would agree with the notion that there is little variation within *Providencia* (21). The short branch lengths, the greater than normal degree of 16S rRNA sequence similarity, and the low functional variation within species clades suggests that the *Providencia* studied might not vary based on their host or environmental niches as much as hypothesized. Pangenome

analysis is well suited to differentiate genomic content even without functional annotations with a tool like PROKKA (12), but this study differentiating based on functional differences relies on these annotations. The genes that are at this time unannotated due to an insufficient knowledge base could skew the results further in one direction or another. This study also only looked at the volume of enriched for suppressed functions. This data could be looked at more in depth or mapped against established pathways for trends conferring utility like antibiotic resistance or solvent tolerance.

References

1. Li SJ, Hua ZS, Huang LN, Li J, Shi SH, Chen LX, et al. Microbial communities evolve faster in extreme environments. Sci Rep [Internet]. 2014 Aug 27 [cited 2020 Jul 23];4(1):1–9. Available from: www.nature.com/scientificreports

2. Bang C, Dagan T, Deines P, Dubilier N, Duschl WJ, Fraune S, et al. Metaorganisms in extreme environments: do microbes play a role in organismal adaptation? Vol. 127, Zoology. Elsevier GmbH; 2018. p. 1–19.

3. Van den Burg B. Extremophiles as a source for novel enzymes. Vol. 6, Current Opinion in Microbiology. Elsevier Ltd; 2003. p. 213–8.

4. Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J, Handelsman J. Call of the wild: Antibiotic resistance genes in natural environments [Internet]. Vol. 8, Nature Reviews Microbiology. Nature Publishing Group; 2010 [cited 2020 Jul 24]. p. 251–9. Available from: www.nature.com/reviews/micro

5. Kadavy DR, Plantz B, Shaw CA, Myatt J, Kokjohn TA, Nickerson KW. Microbiology of the oil fly, Helaeomyia petrolei. Appl Environ Microbiol [Internet]. 1999 Apr 1 [cited 2020 Jul 19];65(4):1477–82. Available from: http://aem.asm.org/

6. Thorpe WH. The biology of the petroleum fly. Vol. 73, Science. 1931. p. 101–3.

7. Sikkema J, De Bont JAM, Poolman B. Mechanisms of membrane toxicity of hydrocarbons. Vol. 59, Microbiological Reviews. American Society for Microbiology; 1995. p. 201–22.

8. Kadavy DR, Hornby JM, Haverkost T, Nickerson KW. Natural antibiotic resistance of bacteria isolated from larvae of the oil fly, Helaeomyia petrolei. Appl Environ Microbiol [Internet]. 2000 Nov 1 [cited 2020 Jul 24];66(11):4615–9. Available from: http://aem.asm.org/

9. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2020 Jul 21]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

10. Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. A and PAP. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19(5):455–77.

11. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. 2013 [cited 2020 Jul 21];29(8):1072–5. Available from: http://bioinf.spbau.ru/quast

12. Seemann T. Prokka: rapid prokaryotic genome annotation. 2014 [cited 2020 Jul 21];30(14):2068–9. Available from: https://academic.oup.com/bioinformatics/article-abstract/30/14/2068/2390517

13. Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: An advanced analysis and visualization platformfor 'omics data. PeerJ [Internet]. 2015 Oct 8 [cited 2020 Jul 22];2015(10):e1319. Available from: http://merenlab.org/projects/anvio.

14. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics [Internet]. 2009 Dec 15 [cited 2020 Jul 22];10(1):1–9. Available from: https://link.springer.com/articles/10.1186/1471-2105-10-421

15. Van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. Methods Mol Biol [Internet]. 2012 [cited 2020 Jul 22];804:281–95. Available from: https://link.springer.com/protocol/10.1007/978-1-61779-361-5_15

16. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. [cited 2020 Jul 22]; Available from: http://www.drive5.com/muscle.

17. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One [Internet]. 2010 Mar 10 [cited 2020 Jul 22];5(3):e9490. Available from: www.plosone.org

18. Somvanshi VS, Lang E, Stäubler B, Spröer C, Schumann P, Ganguly S, et al. Providencia vermicola sp. nov., isolated from infective juveniles of the entomopathogenic nematode Steinernema thermophilum. Int J Syst Evol Microbiol [Internet]. 2006 Mar 1 [cited 2020 Jul 19];56(3):629–33. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.63973-0

19. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol [Internet]. 2014 Feb 1 [cited 2020 Jul 26];64(PART 2):346–51. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.059774-0

20. Rossi-Tamisier M, Benamar S, Raoult D, Fournier PE. Cautionary tale of using 16s rRNA gene sequence similarity values in identification of human-associated bacterial species. Int J Syst Evol Microbiol [Internet]. 2015 Jun 1 [cited 2020 Jul 26];65(6):1929–34. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.000161

21. Galac MR, Lazzaro BP. Comparative genomics of bacteria in the genus Providencia isolated from wild Drosophila melanogaster. BMC Genomics [Internet]. 2012 Nov 13 [cited 2020 Jul 27];13(1):1–18. Available from: https://link.springer.com/articles/10.1186/1471-2164-13-612

Table 1 – Information from NCBI for bacterial isolates making up the pangenome (Fig 1.).

Strain	Ascension	Source	Strain	Ascension	Source
Providencia alcalifaciens					
205/92 (JALD)	GCA_000527335.1	*Homo sapiens*	NCTC10286	GCA_900478095.1	Type Strain
Dmel2	GCA_000314875.2	*Drosophila melanogaster*	PAL-1	GCA_000527275.1	*Homo sapiens*
DSM 30120	GCA_000173415.1	*Homo sapiens*	PAL-2	GCA_000527255.1	*Homo sapiens*
FDAARGOS_408	GCA_002393505.1	*Homo sapiens*	PRM-2	GCA_003057415.1	*Homo sapiens*
MGYG-HGUT-01465	GCA_902375285.1	*Homo sapiens*	R90-1475	GCA_000527315.1	*Homo sapiens*
MGYG-HGUT-01466	GCA_902375275.1	*Homo sapiens*	RIMD 1656011	GCA_000527295.1	*Homo sapiens*
Providencia burhodogranariea					
DSM19968	GCA_000314855.2	*Drosophila melanogaster*			
Providencia heimbachae					
99101	GCA_005157325.1	*Sus scrofa domesticus*	NCTC12003	GCA_900475855.1	*Spheniscidae*
ATCC35613	GCA_001655055.1	*Spheniscidae*	P12672	GCA_900061445.1	*Homo sapiens*
Providencia rettgeri					
297	GCA_007644115.1	*Homo sapiens*	NVIT03	GCA_003426175.1	*Nasonia vitripennis*
594m/10B	GCA_011683805.1	*Corvus brachyrhynchos*	PR1	GCA_002265395.1	*Homo sapiens*
BML2531	GCA_010320145.1	*Homo sapiens*	PR_162	GCA_003936755.1	Hospital Sink
Dmel1	GCA_000314835.2	*Drosophila melanogaster*	PR-15-2-50	GCA_005155965.1	*Homo sapiens*
DSM 1131	GCA_000158055.1	*Homo sapiens*	RB151	GCA_001874625.1	*Homo sapiens*
FDAARGOS_330	GCA_002984195.1	*Homo sapiens*	YPR31	GCA_013255915.1	*Anatidae*
MGYG-HGUT-01323	GCA_902373935.1	*Homo sapiens*			
Providencia rustigianii					
DSM 4541	GCA_000156395.1	*Homo sapiens*	NCTC6933	GCA_900635875.1	Type Strain
MGYG-HGUT-01708	GCA_902377615.1	*Homo sapiens*	NCTC8113	GCA_900637755.1	Type Strain
NCTC11667	GCA_900455235.1	*Spheniscus humboldti*			
Providencia sneebia					
DSM 19967	GCA_000314895.2	*Drosophila melanogaster*			
Providencia stuartii					
ASO12334	GCA_010597545.1	*Homo sapiens*	FDAARGOS_291	GCA_002983665.1	*Homo sapiens*
ATCC 25827	GCA_000154865.1	*Homo sapiens*	FDAARGOS_294	GCA_002206175.2	*Homo sapiens*
BE2467	GCA_001888205.1	*Homo sapiens*	FDAARGOS_645	GCA_008693805.1	*Homo sapiens*
Crippen	GCA_001853385.1	*Lucilia sericata*	MGYG-HGUT-01307	GCA_902373775.1	*Homo sapiens*
FDAARGOS_87	GCA_000783455.2	*Homo sapiens*	MRSN 2154	GCA_000259175.1	*Homo sapiens*
FDAARGOS_145	GCA_001558855.2	*Homo sapiens*	PS901	GCA_012956045.1	*Homo sapiens*
Providencia vermicola					
P8538	GCA_010748935	*Homo sapiens*			
Morganella morganii					
AR_0057	GCA_002968775.1	Clinical	MGYG-HGUT-02512	GCA_902387845.1	*Homo sapiens*
FDAARGOS_63	GCA_000783955.2	*Homo sapiens*	NCTC12028	GCA_900478755.1	*Homo sapiens*

Table 2 – Groups of isolates used for functional enrichment analysis (Fig 2.)

Grouping	Oil Fly	Insects	OF5	OF6 & OF10	Nonhuman Hosts	Insects Control	Presumptive *vermicola*	OF5 *rettgeri*	Dmel1 *rettgeri*
PR_297_Human	N	N	N	N	N	N	N	N	N
PR_594M10B_AmericanCrow	N	N	N	N	Y	N	Y	NA	NA
PR_BML2531_Human	N	N	N	N	N	N	N	N	N
PR_DSM1131_Human	N	N	N	N	N	N	N	N	N
PR_Dmel1_FruitFly	N	Y	N	N	Y	Y	N	NA	Y
PR_FDAARGOS330_Human	N	N	N	N	N	N	Y	NA	NA
PR_HGUT01323_Human	N	N	N	N	N	N	N	N	N
PR_NVIT03_ParasitoidWasp	N	Y	N	N	Y	Y	N	N	N
PR_PR15250_Human	N	N	N	N	N	N	N	N	N
PR_PR162_HospitalSink	NA	NA	NA	NA	NA	NA	N	NA	NA
PR_PR1_Human	N	N	N	N	N	N	N	N	N
PR_RB151_Human	N	N	N	N	N	N	Y	NA	NA
PR_YPR31_Duck	N	N	N	N	Y	N	N	N	N
PV_OF0010_OilFly	Y	Y	NA	Y	Y	NA	Y	NA	NA
PR_OF005_OilFly	Y	Y	Y	NA	Y	NA	N	Y	NA
PV_OF006_OilFly	Y	Y	NA	Y	Y	NA	Y	NA	NA
PV_P8538_Human	NA	NA	NA	NA	NA	NA	NA	NA	NA

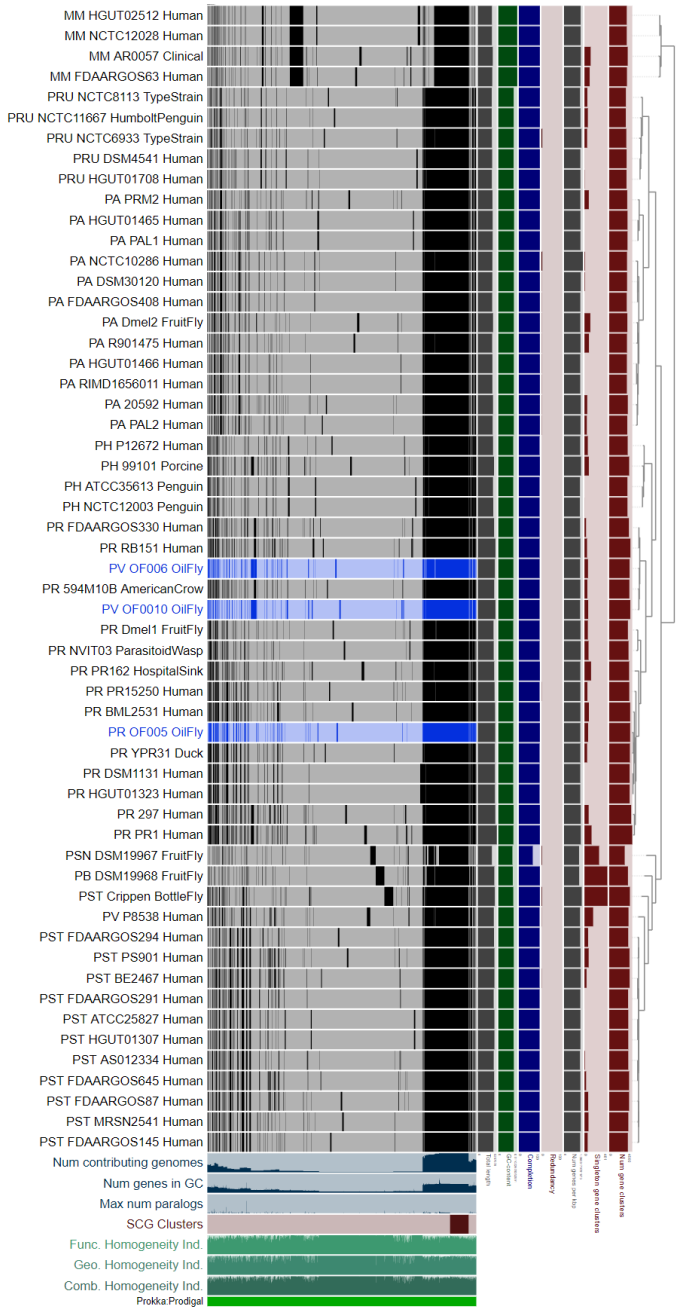


Fig. 1 Pangenome and Phylogenetic of analysis of *Providencia*

A combined pangenome and concatenated phylogeny of *Providencia*, rooted on *Morganella*. The grey bars on the left denote the gene cluster arrangements. Table 1 lists relevant information for the strains making up the pangenome.

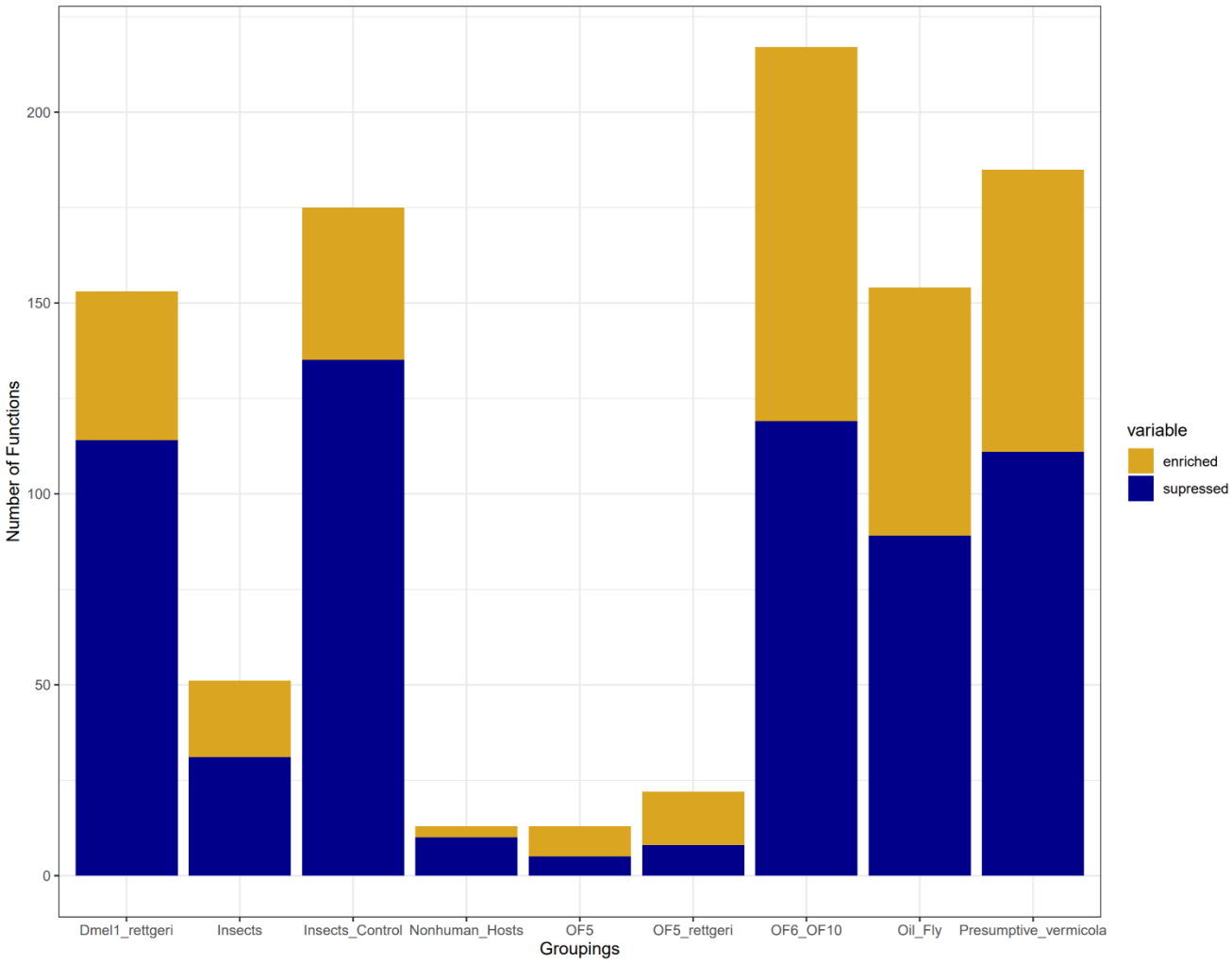


Fig. 2 Functional Enrichment Analysis of *P. rettgeri* and *P. vermicola*

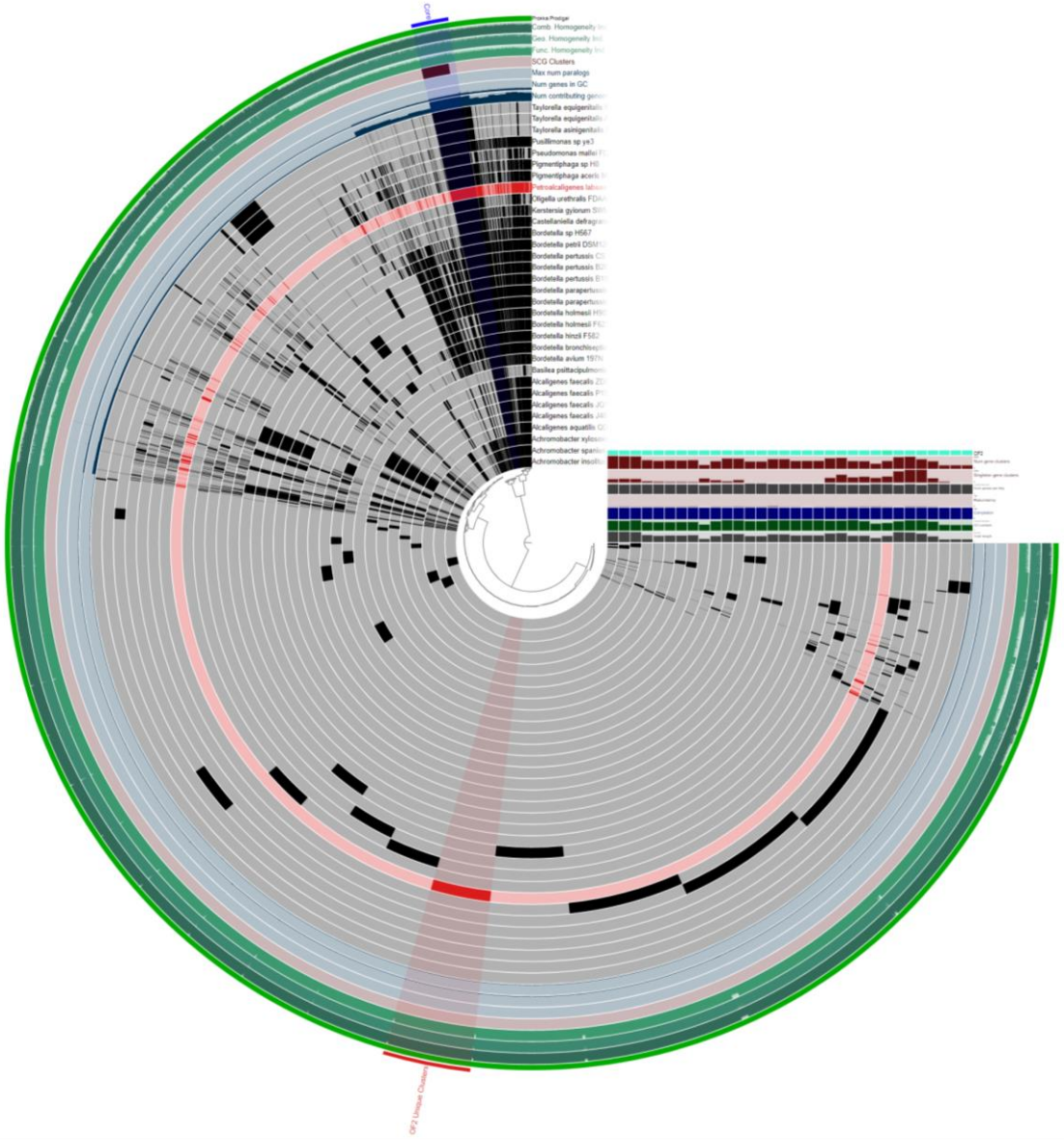
Number of functions for each grouping that are enriched and suppressed. Strains included in each grouping is outlined in table 2.

Conclusions

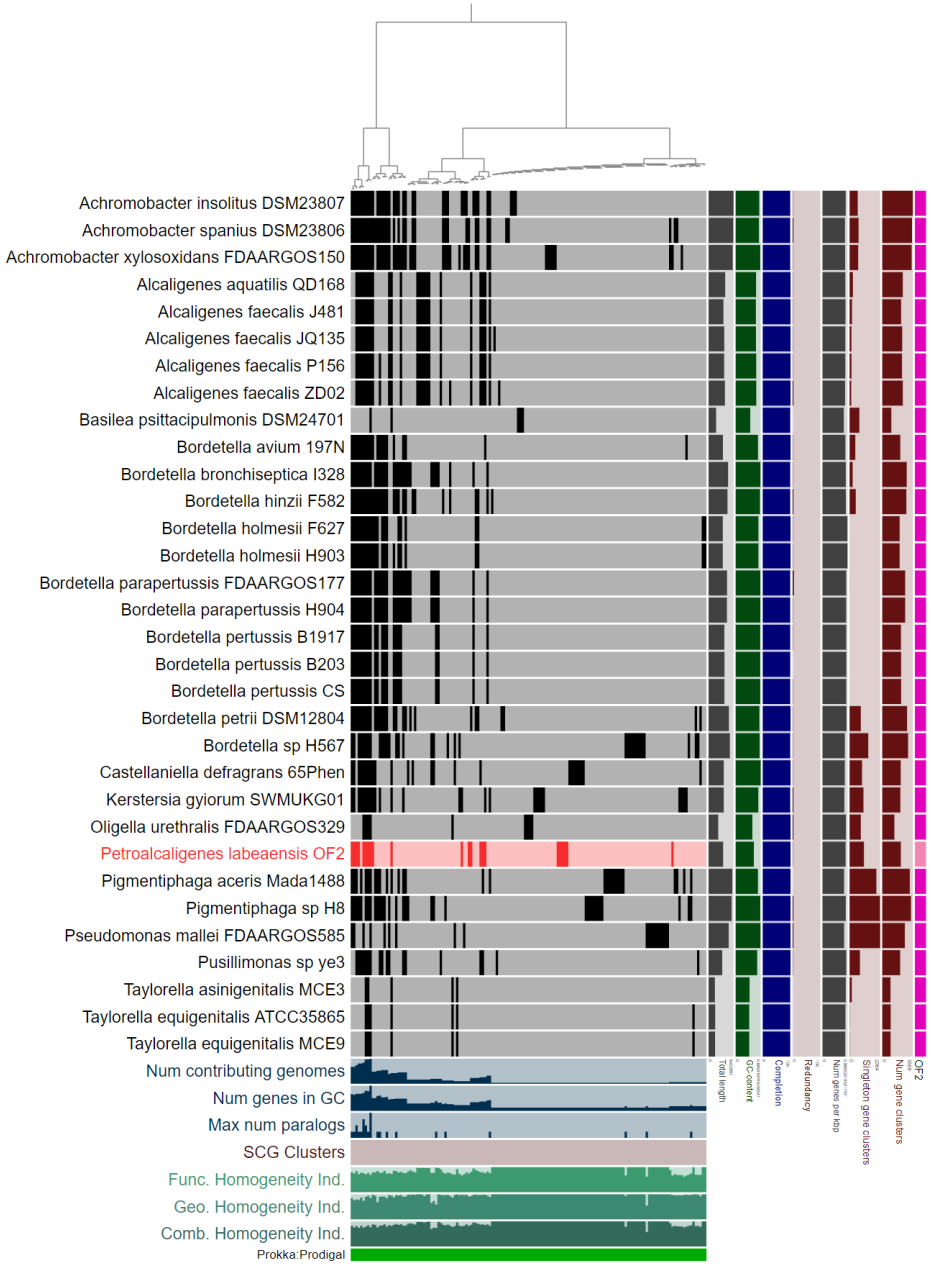
The goal with this research was the functional characterization of the gut microflora of a unique eukaryotic host. A host that lives in the presence of tremendous solvent stress with no known negative effects. The oil fly was first highlighted for study in the 1930s, and unfortunately to this day that research has been minimal. Already with this research we have established the novel Candidatus genus *Petroalcaligenes*, used pangenomics to highlight functional diversity, and used phylogenomics to identify standing problems in the genus *Providencia*. There is tremendous potential in the continuation and expansion of this study system. There is little know about the possible unique physiology and life cycle of the oil fly itself, and there currently no opportunities for testing of larval solvent tolerance in a germ-free environment.

This physiological testing along with a modern metagenomic approach to the gut microflora could provide insights into the evolution of antibiotic resistance in natural environments, identify solvent tolerant genes of interest to industry, or unveil additional novel species or genera. Specifically, with a broader push to study the microflora with metagenomics, the holobiont could provide insights into functional pathways not available to a single species of bacteria.

Chapter 1 Supplemental Figures



Supplementary Figure 1.1 – duplicate pangenome of figure 2 from chapter 1 in a circular format to better display gene clusters. The core genome is binned and displayed in blue, and the OF2 unique gene clusters are binned and displayed in red. OF2 is highlighted in red as well.



Supplementary Figure 1.2 - A pangenome split describing the gene clusters representing multidrug resistance or multidrug efflux functions.

Supplementary Table 1.1 – Functional enrichment data of OF2 against the greater pangenome.

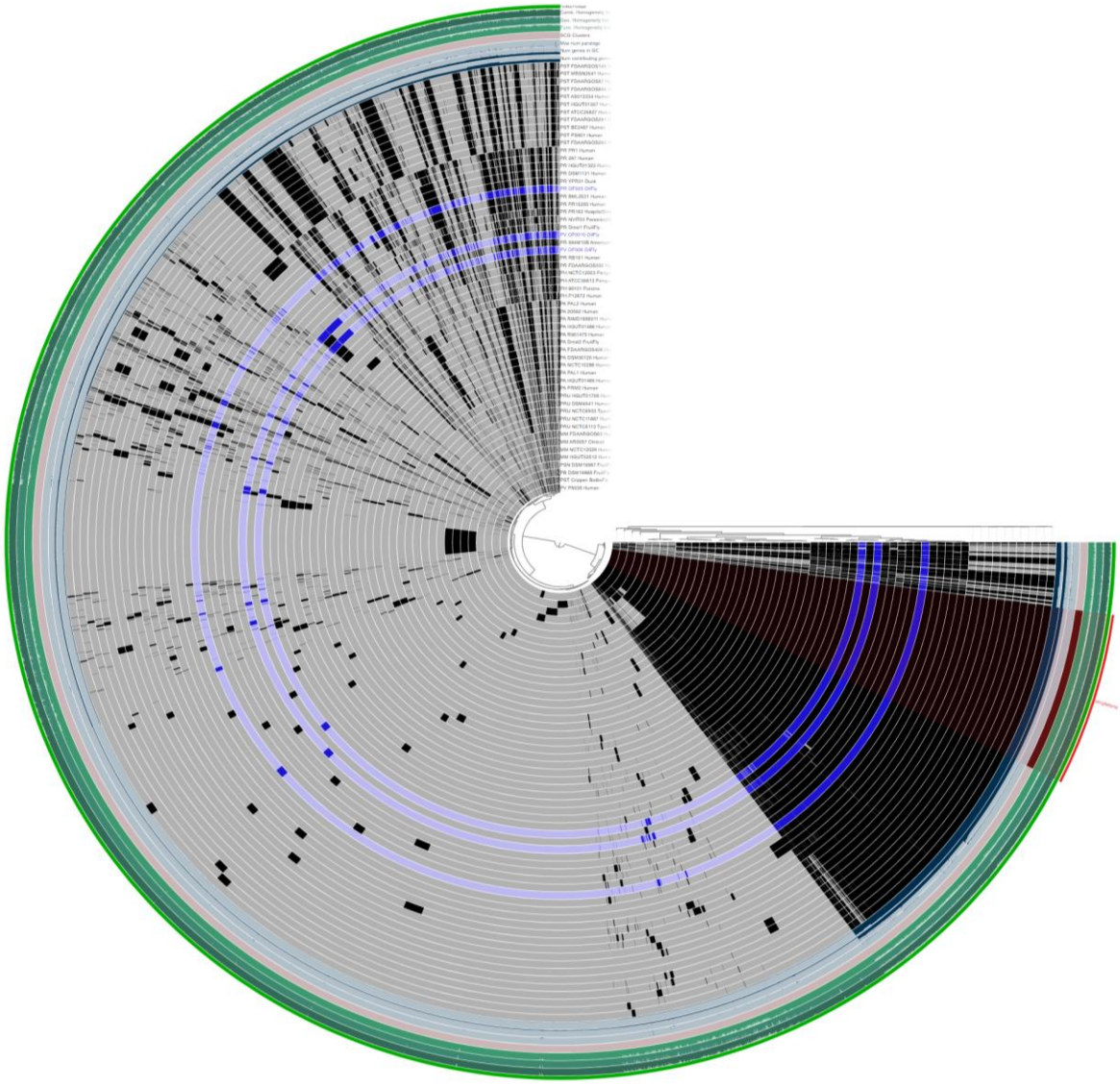
Gene Identifier	Function	Enrichment Score	Q-value	Enriched Group
abaQ	Quinolone resistance transporter	32.00007	1.93E-06	OF2
acnR	HTH-type transcriptional repressor AcnR	32.00007	1.93E-06	OF2
aldA_1	Lactaldehyde dehydrogenase	32.00007	1.93E-06	OF2
ansA_1	L-asparaginase 1	15.47315	0.005318	OF2
araH_1	L-arabinose transport system permease protein AraH	15.47315	0.005318	OF2
aruI	putative 2-ketoarginine decarboxylase AruI	32.00007	1.93E-06	OF2
bioH	Pimeloyl-[acyl-carrier protein] methyl ester esterase	32.00007	1.93E-06	OF2
camD	5-exo-hydroxycamphor dehydrogenase	32.00007	1.93E-06	OF2
cfiA	2-oxoglutarate carboxylase large subunit	32.00007	1.93E-06	OF2
csqR	HTH-type transcriptional repressor CsqR	32.00007	1.93E-06	OF2
dapX	putative N-acetyl-LL-diaminopimelate aminotransferase	32.00007	1.93E-06	OF2
dctA	Aerobic C4-dicarboxylate transport protein	32.00007	1.93E-06	OF2
dddP	Dimethlysulfonioproprionate lyase DddP	32.00007	1.93E-06	OF2
elfD	putative fimbrial chaperone protein ElfD	32.00007	1.93E-06	OF2
fieF	Ferrous-iron efflux pump FieF	15.47315	0.005318	OF2
garP	putative galactarate transporter	15.47315	0.005318	OF2
gbuA	Guanidinobutyrase	15.47315	0.005318	OF2
gbuA_2	Glycine betaine/carnitine transport ATP-binding protein GbuA	32.00007	1.93E-06	OF2
gdnC	putative guanidinium efflux system subunit GdnC	32.00007	1.93E-06	OF2
glpC	Anaerobic glycerol-3-phosphate dehydrogenase subunit C	32.00007	1.93E-06	OF2
hsdS	Type-1 restriction enzyme EcoKI specificity protein	15.47315	0.005318	OF2
lpfB	putative fimbrial chaperone LpfB	15.47315	0.005318	OF2
mamZ	Magnetosome protein MamZ	32.00007	1.93E-06	OF2
mtnC	Enolase-phosphatase E1	32.00007	1.93E-06	OF2
nlpD_2	Murein hydrolase activator NlpD	32.00007	1.93E-06	OF2
nucH	Thermonuclease	32.00007	1.93E-06	OF2
oleD	2-alkyl-3-oxoalkanoate reductase	15.47315	0.005318	OF2
opgE	Phosphoethanolamine transferase OpgE	15.47315	0.005318	OF2
opuAB	Glycine betaine transport system permease protein OpuAB	15.47315	0.005318	OF2
ousX	Glycine betaine-binding periplasmic protein OusX	32.00007	1.93E-06	OF2
peb1A	Major cell-binding factor	32.00007	1.93E-06	OF2
pmfR	Transcriptional activator PmfR	32.00007	1.93E-06	OF2
prr	Gamma-aminobutyraldehyde dehydrogenase	15.47315	0.005318	OF2
putP	Sodium/proline symporter	11.15105	0.052447	OF2
recE	Exodeoxyribonuclease 8	15.47315	0.005318	OF2
rhmT	Inner membrane transport protein RhmT	32.00007	1.93E-06	OF2
rocA	1-pyrroline-5-carboxylate dehydrogenase	32.00007	1.93E-06	OF2
sbcD	Nuclease SbcCD subunit D	15.47315	0.005318	OF2
sfaA	S-fimbrial protein subunit SfaA	32.00007	1.93E-06	OF2
tcyN	L-cystine import ATP-binding protein TcyN	15.47315	0.005318	OF2
tdiR	Transcriptional regulatory protein TdiR	15.47315	0.005318	OF2
torT	Periplasmic protein TorT	15.47315	0.005318	OF2
tyrP	Tyrosine-specific transport protein	32.00007	1.93E-06	OF2

uehC	Ectoine/5-hydroxyectoine TRAP transporter large permease protein UehC	32.00007	1.93E-06	OF2
viuB_2	Vibriobactin utilization protein ViuB	15.47315	0.005318	OF2
yceD	Large ribosomal RNA subunit accumulation protein YceD	15.47315	0.005318	OF2
ychN	Protein YchN	15.47315	0.005318	OF2
yfnB	Putative HAD-hydrolase YfnB	32.00007	1.93E-06	OF2
ygcS	Inner membrane metabolite transport protein YgcS	15.47315	0.005318	OF2
yraJ	Outer membrane usher protein YraJ	15.47315	0.005318	OF2
dapD	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase	15.47315	0.005318	Pan
dapE	Succinyl-diaminopimelate desuccinylase	15.47315	0.005318	Pan
dapF	Diaminopimelate epimerase	15.47315	0.005318	Pan
dat	D-alanine aminotransferase	15.47315	0.005318	Pan
gluQ	Glutamyl-Q tRNA(Asp) synthetase	15.47315	0.005318	Pan
queC	7-cyano-7-deazaguanine synthase	15.47315	0.005318	Pan
queF	NADPH-dependent 7-cyano-7-deazaguanine reductase	32.00007	1.93E-06	Pan

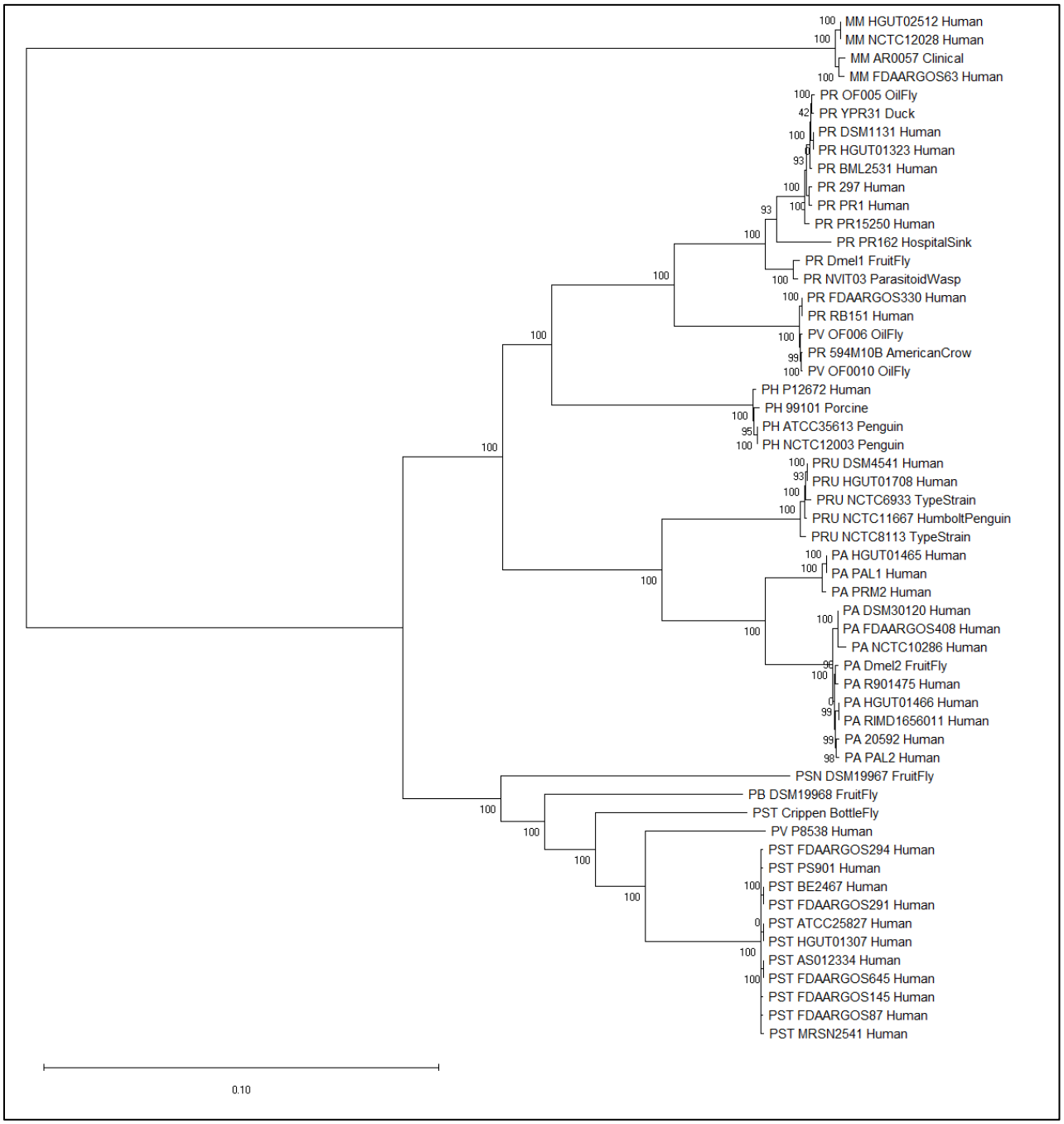
Supplementary Table 1.2 – Overview of multidrug resistance genes from *P. labreaensis*

Gene Cluster ID	Gene Identifier	Function
GC_00000031	ttgB	putative efflux pump membrane transporter TtgB
GC_00000086	mdtC	Multidrug resistance protein MdtC
GC_00000111	mdtB_1	Multidrug resistance protein MdtB
GC_00000680	emrE	Multidrug transporter EmrE
GC_00000977	mexB	Multidrug resistance protein MexB
GC_00000989	acrA	Multidrug efflux pump subunit AcrA
GC_00001067	emrB	Multidrug export protein EmrB
GC_00001244	emrA	Multidrug export protein EmrA
GC_00003496	mdtN	Multidrug resistance protein MdtN
GC_00003696	norM	Multidrug resistance protein NorM
GC_00003782	None	putative multidrug-efflux transporter
GC_00004092	None	putative multidrug-efflux transporter
GC_00008450	stp_1	Multidrug resistance protein Stp
GC_00008797	mdtL	Multidrug resistance protein MdtL
GC_00011991	stp_2	Multidrug resistance protein Stp
GC_00013140	mdtA_2	Multidrug resistance protein MdtA
GC_00027092	mdtA_4	Multidrug resistance protein MdtA
GC_00028341	mdtA_3	Multidrug resistance protein MdtA
GC_00028751	mdtA_1	Multidrug resistance protein MdtA
GC_00028819	mdtB_2	Multidrug resistance protein MdtB

Chapter 2 Supplemental Figures



Supplementary Figure 2.1 – Duplicate pangenome of figure 1 from chapter 2 in its circular form. The core singleton genes are binned and highlighted in red. OF5, OF6, and OF10 are highlighted in blue.



Supplementary Figure 2.2 – Phylogeny from chapter 1 that was attached to the pangenome. The larger format and node support values aid in clarity. Maximum likelihood phylogeny made with 56 genomes, 1011 single copy genes, and 297,642 amino acid positions.