IN SILICO MAPPING OF BIOFILM GENE CLUSTERS OF *Pseudomonas aeruginosa* IN DECIPHERING NOVEL THERAPEUTIC TARGETS FOR CLINICAL INTERVENTION

MICHAEL JOHN AMBUTSI

A thesis submitted in partial fulfillment of the requirement for the degree of Masters of Science in Bioinformatics of Masinde Muliro University of Science and Technology.

JULY, 2020

DECLARATION

This thesis is my original work prepared with no other than the indicated sources and support and has not been presented for a degree or any other award.

Signature…………………………………………… Date ………………………………

AMBUTSI MICHAEL JOHN SBF/G/01-52282/2018

CERTIFICATION

The undersigned certify that we have read and hereby recommend for acceptance of Masinde Muliro University of Science and Technology a thesis entitled “In Silico Mapping of Biofilm Gene Clusters of *Pseudomonas Aeruginosa* in Deciphering Novel Therapeutic Targets for Clinical Intervention”

Signature ………………………………………. Date…………………………………

Dr. OKOTH Patrick Ph. D

Department of Biological Sciences

Masinde Muliro University of Science and Technology

Signature…………………………………………… Date…………………………………

Prof., Dr. Oleg N. Reva

Department of Biochemistry, Centre for Bioinformatics and Computational Biology

University of Pretoria

COPYRIGHT

DEDICATION

This work is dedicated to my family for the continued love, support, and encouragement they have offered me throughout the process. May the Almighty God richly bless them.

ACKNOWLEDGEMENT

I would like to express my appreciation to the Masinde Muliro University of Technology fraternity, including the School of Natural Sciences and the Department of Biological Science for the unconditional support they offered me. Special thanks to my supervisors Dr. Patrick Okoth and Prof. Oleg Reva, from the University of Pretoria, who kept encouraging and offered timely input from the start of the research until the work was accomplished. I wish to appreciate my supervisors for their intellectual support, regular advice, encouragement and valuable scientific guidance on different aspects of this research which they offered. I am grateful to Dr Kelvin Omieno for introducing me into the field of computer programming which turned out to be worthwhile throughout my research. I wish to also thank my family for their endless financial support and encouragement throughout the course of this work. Above all, I am grateful to the almighty God for providing me with good health and strength to undertake this research work.

ABSTRACT

Advances in sequencing technology have resulted in a significant rise in the number of genome sequences deposited in the International *Pseudomonas* Consortium Database (IPCD), National Center for Biotechnology Information (NCBI) database and the DNA Databank of Japan (DDBJ). A number of special Bioinformatic algorithms have been developed to facilitate comprehensive analyses of these repositories. The profile Hidden Markov Model (pHMM) is one such tool that has successfully been applied in the characterization of protein families, gene discovery as well as the prediction of unclassified sequences. To date, approximately 176 complete genomes of *Pseudomonas aeruginosa,* an opportunistic pathogen, have been banked in the NCBI database. The large number of genomes available makes an *in silico* approach to characterize various genes of the ubiquitous organism feasible. The gram-negative bacterium is a leading cause of nosocomial infections among immunocompromised individuals which has progressively developed antibiotic resistant genes that have conferred it the ability to withstand antibiotics, further complicating the treatment. The pathogen has developed this ability through horizontal gene transfer and mutations on the variable accessory genome. Within the human host *P. aeruginosa* forms biofilms that compound its antibiotic resistance. In spite of the significance of the phenomenon in compounding antibiotic resistance, exhaustive analyses of the genes responsible for biofilm formation remain scanty and largely undocumented. This study sought to undertake an *in silico* mapping of the highly versatile biofilm formation genes to decipher novel therapeutic target regions for clinical intervention. Genes responsible for biofilm formation were identified using an Entrez search engine on the NCBI database. Complete genomes of *P. aeruginosa* were downloaded from NCBI's *P. aeruginosa* resources and the International Pseudomonas Consortium Database. A custom python script was then written to retrieve biofilm gene sequences from the annotated genomes of *P. aeruginosa* (Genbank files) and Clusters of Orthologous genes (COG) created. A phylogenetic tree representative of the evolutionary relationships of the biofilm formation genes was constructed to indicate genes that co-evolved and those that evolved differently. Specific profile Hidden Markov Models for the different classes of biofilm formation genes were constructed from the mined genes and used to analyze the different strains of *P. aeruginosa*. Overall, 13 ecological niches were deciphered. The study identified the *algD* gene as the most ubiquitous gene in the strains of this pathogen. The abscess ecological niche reported the highest density of hits. The constructed phylogenetic tree revealed *algD, algU* and *fliC* genes evolved differently indicating acquisition through horizontal gene transfer. Wilcoxon-signed rank test indicated that the density of the *htpG* pHMM hits was greater for human samples than for nonhuman samples (W=3; p-value = 0.01759) indicating a high likelihood of the gene being expressed in human hosts. This was the first ever attempt to characterize biofilm formation genes in the genomes of *P. aeruginosa* based on the profile Hidden Markov Model. The findings of this study identified four novel therapeutic targets, *algD, algU, fliC,* and *htpG* that could be explored by pharmaceutical companies in designing candidate drugs based on bioinformatics tools for management of the pathophysiology of *P. aeruginosa* infections.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEFINITION OF TERMS

Nosocomial infections These are infections that are acquired from a healthcare facility or a hospital.

Genomic Islands Discrete DNA segments between closely related strains that contribute to the adaptation and diversification of microorganisms.

GC Content The percentage of nitrogenous bases on a DNA or RNA molecule that are either guanine or cytosine.

Motif A short conserved sequence pattern associated with distinct functions of a protein or DNA. It is often associated with a distinct structural site performing a particular function.

Consensus Sequence A sequence of DNA that represents aligned, related sequences having similar structure and function in different organisms.

Orthologous genes Genes in different species that originated by vertical descent from a single gene of the last common ancestor.

LIST OF ACRONYMS AND ABBREVIATIONS

AMR	Antimicrobial Resistance
BLAST	Basic Local Alignment Search Tool
BRIG	BLAST Ring Image Generator
CARD	Comprehensive Antibiotic Resistance Database
DDBJ	DNA Databank of Japan
GEIs	Genomic Islands
HMM	Hidden Markov Model
IPCD	International *Pseudomonas* Consortium Database
MSV	Multiple Segment Viterbi
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
ORFs	Open Reading Frames
PHYLIP	Phylogeny Inference Package
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
SNPs	Single Nucleotide Polymorphisms
WGS	Whole Genome Sequencing

CHAPTER ONE
INTRODUCTION

1.1 Background Information

The National Center for Biotechnology Information (NCBI) database, International *Pseudomonas* Consortium Database (IPCD) and the DNA Databank of Japan (DDBJ) have witnessed a significant rise in the number of genomes largely due to the advances in sequencing technology over the last few years (Land *et al*., 2015). The databases have proved to be a valuable source in the search for novel and known genes from different sequences. The large amount of data that needs to be analyzed has resulted in the development of special tools like the Profile Hidden Markov Models (Francisco *et al.*, 2019). Characterization of protein families, gene discovery and sequence analyses have all been achieved using such models (Restrepo-Montoya *et al.*, 2011).

To better utilize the sequences of specific organisms already available, you could multiply align sequences drawn from a given subtype and build a probabilistic model from their consensus. The probabilistic model is the widely known profile Hidden Markov Model that can be constructed using the software package HMMER (Francisco *et al.*, 2019). Profile HMMs have previously been employed in numerous studies that involved the detection of particular sequences. Using the tool, one can determine whether or not a particular sequence belongs to a specific profile given the probabilistic and statistical intrinsic nature of the profile HMMs (Gong *et al.*, 2012). The Pfam database of protein families heavily relies on profile HMMs to facilitate the characterization and classification of different protein families (Finn *et al.*, 2016). In spite of its efficiency in sequence analysis, profile HMMs have not yet been employed in the analysis of *P. aeruginosa* genes. This study sought to construct profile HMMs of different classes of biofilm formation genes and use them to decipher different properties of the pathogen's sequences obtained from different ecological niches.

Close to 176 complete sequence genomes of *P. aeruginosa* strains isolated from different ecological niches, given the ubiquitous nature of the microorganism, have been sequenced in the National Center for Biotechnology Information (NCBI) to date. The numerous sequences are an ideal source for studies looking to investigate the diversity and complexity of strains of the pathogen (Bruggemann *et al.,* 2018). This large number of genomes available in NCBI's gene bank makes an *in silico* approach to characterize the genes of biofilm formation in *P. aeruginosa* is feasible. Comparative genomic analyses of whole genomes of the bacteria have been successfully applied in previous studies that focused on genome plasticity and the persistence of the bacterium in airway infections (Bianconi *et al.*, 2015, Bruggemann *et al.,* 2018). The findings of these studies highlighted different aspects of the bacteria ranging from its antibiotic resistance capability to its virulence properties. In each study, sequences were obtained from NCBI repositories and various bioinformatics tools were used to analyze the sequences. Exhaustive analysis of the repositories of NCBI database sequences remains scanty and largely undocumented (Freshi, 2015, Wielhmann, 2015).

The *Pseudomonas* genus comprises of both beneficial strains and opportunistic human pathogens which exhibit different lifestyles within their host (Mark *et al.,* 2011). *P.* a*eruginosa* is predominantly associated with hospital-acquired infections and accounts for 11% of the nosocomial infections (Khan *et al.,* 2015). It is an opportunistic human pathogen as it does not infect healthy individuals. Immunocompromised patients especially those with cancer, AIDS or cystic fibrosis are the most susceptible to infections caused by the bacteria (Balasubramanian and Mathee, 2009). It has been associated with high a mortality rate ranging from 18% - 60% and is one of the leading gram negative opportunistic pathogens (Kim *et al*., 2014). *P. aeruginosa* is characterized as a ubiquitous microorganism as it lives in both human and inanimate environments.

The pathogen has been implicated in a host of infections including septicemia, pneumonia, otitis media, infections of the lower respiratory tract and cystic fibrosis highlighting its medical importance. The bacterium has developed resistance to antibiotics through horizontal gene transfer and mutations in chromosomal genes (Araujo *et al.*, 2016). Its antibiotic resistance ability is compounded by plasmids that encode for beta lactamase production (Okesela and Oni, 2012). Multi-drug resistant *P. aeruginosa* complicates treatment decisions and inevitably leads to treatment failure (Zavascki AP, 2010). The organism also forms biofilms through which different groups can adhere to surfaces. These biofilms cannot be easily destroyed once they are formed compounding the microorganism's antibiotic resistance ability. The conservation and variation patterns of biofilm formation genes in *P. aeruginosa* are poorly studied in spite of the importance of the phenomenon in antibiotic resistance. This study sought to develop a specific profile to search for biofilm formation genes in the genomes of *P. aeruginosa* and identify the conservation and variation patterns of these genes among the different strains of the opportunistic pathogen to decipher novel therapeutic targets for clinical intervention. New findings have been based on chance rather than a systematic exploration (Francisco *et al.*, 2019). This was the first ever attempt to map the gene clusters of *P. aeruginosa* based on the profile Hidden Markov Model.

The pathogen's genome is composed of a conserved core and variable accessory segments that are characterized by a set of genomic islands (Mark *et al.*, 2011). The core genome has a conserved synteny of genes and a 0.5% level of nucleotide divergence. The genome has a G-C content of 65% and is about 5.2 to 7 million base pairs long (Weihlmann, 2007). *P. aeruginosa* consists of different strains that are classified under three clades which are known to occupy different niches within their host. The first clade is associated with the reference laboratory strain PAO1. Most *P. aeruginosa* strains are associated with this clade. The second clade is associated with PA14 which is more virulent and contains additional genes associated with survival in diverse environments. The third clade is associated with strain PA7 that is a non-respiratory clinical isolate from Argentina (Roy *et al.,* 2010). Previous studies have not provided reliable explanations for the population structure of the bacteria (Wielhmann, 2015). Although host associations and environmental niches have been implicated in the evolutionary differences, exhaustive analyses of the NCBI repositories remain scanty and disjointed (Freshi, 2015, Wielhmann, 2015). This study analyzed available sequence information to shed further light on the population structure and survival mechanism of *P. aeruginosa* in the human host.

The next-generation sequencing technology (NGS) is fast taking the place of traditional molecular typing techniques. It is a high throughput technique that facilitates the identification of point mutations within bacterial species and is ideal for comparative genomic studies of bacterial pathogens (Metzker, 2010). Most of the sequences deposited on NCBI repositories albeit being informative have however not been adequately explored in the efforts to better understand the survival mechanisms of *P. aeruginosa.* It is against this background that an informed decision to analyze available sequences using profile Hidden Markov Models to identify the conserved and variable biofilm formation genes was made. With NGS higher sequence resolution is guaranteed and associations can be made between the genome evolution, structure and content and the epidemiology of the pathogen (Sabat *et al.,* 2013). Whole genome sequencing data also allows us to identify biological markers that can be exploited in the development of therapies against *P.aeruginosa* infections (Bianconi *et al.,* 2015). Databases containing metadata on the strains provide relevant information that can be used in comparative genomics studies. The metadata provided insights on the specific sequences and aided in the selection of sequences to be used in the study.

The neighbor-joining algorithm was used to determine the evolutionary trends exhibited by the biofilm gene clusters of *P. aeruginosa*. Profile hidden Markov models for genes responsible for biofilm formation in gram negative bacteria were then constructed and used to characterize these genes in strains of *P. aeruginosa*. The study identified the conservation of genes in the opportunistic pathogen which were deemed as regions of interest and were identified as potential targets for novel treatment options. The profile also identified variations in the genes and reported them either as the presence/absence of particular genes or as mutations within the present genes. The findings of this study identified four novel therapeutic targets for clinical intervention, *algD, algU, fliC,* and *htpG* that could be explored by pharmaceutical companies in designing candidate drugs for management of the pathophysiology of *P. aeruginosa* infections. Further, it is hoped that the findings of this study help inform policy on management of *P. aeruginosa* pathogenesis. The findings of the study also shed more light on the survival mechanisms of *P. aeruginosa* within the human host.

1.2 Statement of the Problem

Immunocompromised individuals who visit hospitals are in danger of acquiring different nosocomial infections. Hospital-acquired *P. aeruginosa* infections cause severe illness and can lead to death in some cases. Patients with burn wounds or wounds from surgery are the most susceptible to the life-threatening infections. Antibiotics have generally been used to treat these infections. However, the bacterium has developed antibiotic resistance through horizontal gene transfer and mutations in chromosomal genes, making treatment to become more difficult. Multidrug resistant *Pseudomonas* is considered to be a serious threat according to reports from CDC (2019). Previous studies have been carried out to identify the mechanisms of antibiotic resistance of the organism. Few studies have made reference to the different niches that the bacterium occupies while in its host. One of the survival mechanisms of *P. aeruginosa* is the formation of biofilms. Biofilms often compound the pathogen’s to ability to impair the effects of therapeutic agents. The conservation and variation patterns of biofilm formation genes in *P. aeruginosa* are poorly studied in spite of the importance of the phenomenon in compounding the antibiotic resistance ability. This study sought to profile gene clusters of *P. aeruginosa* that are associated with specific biofilms through the construction of specific profile Hidden Markov Models. Sequences obtained through the next-generation sequencing technology and deposited on NCBI repositories, on the other hand, have been underutilized in the efforts to better understand the survival mechanisms of *P. aeruginosa*. Multi-drug resistant *P. aeruginosa* causes infections that result in high morbidity and mortality rates in different locations. Patients end up requiring utmost attention and the overall treatment costs are increased. The higher treatment costs are more often than not occasioned by the need to try out different treatment options before the most effective option is settled on.

1.3 Justification of the Study

Among the numerous challenges that the global healthcare sector needs to grapple with is antibiotic resistance. The misuse of antibiotics has actively contributed to the emergence of resistance genotypes due to horizontal gene transfer as well as spontaneous mutations. *P. aeruginosa* infections pose a great challenge to healthcare providers in Kenya as the bacteria has developed resistance to antibiotics used for treatment. There is need to develop therapies that are not antibiotic in nature. This is only possible if survival mechanisms of the pathogen are well understood. This study sought to lay bare the conserved and variable gene clusters that facilitate biofilm formation in a bid to provide valuable information about this survival technique. Biofilm formation has stood out as one of the mechanisms through which the pathogen facilitates its survival within the human host. The biofilms are difficult to destroy and compound the antibiotic resistance ability of *P. aeruginosa.* Development of anti-biofilm therapies could be an important step in the efforts to treat infections caused by the bacteria. This study mapped gene clusters that are associated with biofilm formation to step up the fight against *P. aeruginosa.* A better understanding of the conserved and variable biofilm formation genes within the pathogen could help to augment the efforts against the pathogen. The profile Hidden Markov Model is resourceful for determining evolutionary relationships between organisms through rapid and sophisticated analyses of gene sequences obtained from whole genome sequencing. With a single profile HMM one can accurately predict whether or not a certain sequence belongs to a particular profile (Gong *et al.,* 2012). This study sought to analyze available sequences using profile hidden Markov models to identify the conservation and variation of genes that are associated with biofilm formation and decipher novel therapeutic target regions for clinical intervention. *In silico* techniques have been developed to analyze v

1.4 General Objective

To undertake an in silico mapping of biofilm gene clusters of *P. aeruginosa* that define specific biofilm formation genes in deciphering novel therapeutic targets for clinical intervention

1.5 Specific Objectives

(i) To determine the molecular evolutionary relationship of genes responsible for biofilm formation in *P. aeruginosa* biotypes occupying different ecological niches.

(ii) To construct profile Hidden Markov Models for the whole genome sequencing genes responsible for biofilm formation in *P. aeruginosa* biotypes occupying different ecological niches.

(iii) To determine the levels of variability among biofilm formation genes in the genomes of *P. aeruginosa* strains using profile Hidden Markov Models in deciphering novel therapeutic target regions for clinical intervention.

1.6 Null Hypothesis

Ho 1. There is no significant difference between the evolutionary relationships of genes responsible for biofilm formation in *P. aeruginosa* biotypes occupying different ecological niches.

Ho 1. There is no significant difference between the relationship of the whole genome sequencing genes and biofilm formation in *P.seudomonas aeruginosa* biotypes occupying different ecological niches.

Ho 1. There is no significant difference between the variations among gene clusters of biofilm formation in *P. aeruginosa* biotypes.

1.7 Significance of the Study

The results of this study shed further light on the gene clusters responsible for the survival of *P. aeruginosa* within the human host. The results also revealed the conserved and variable biofilm formation genes in the genomes of the opportunistic pathogen. This valuable information is currently not available and has limited the therapeutic efforts of the scientific community. It was hoped that this study would contribute novel findings to help inform enhancement of the therapeutic options against infections caused by *P. aeruginosa*. Development of efficient treatment options is likely to reduce the overall cost of treatment. Mapping of gene clusters responsible for biofilm formation supplements the information already available concerning the survival mechanisms of the bacteria in the human host. The scientific community can build on this information and make clear the survival strategies of *P. aeruginosa*. The intrinsic antibiotic resistance ability of *P. aeruginosa* has brought about the need to develop non-antimicrobial treatment options against infections caused by the bacteria. These novel strategies target different survival pathways of the pathogen within the human host. Quorum sensing and formation of biofilms are paramount for the bacteria to survive and thrive in different niches. However, the gene clusters for these strategies in *P. aeruginosa* are not clearly understood. This study focused on gene clusters responsible for such mechanisms. The findings can be used to exploit novel treatment strategies that will not be affected by the antibiotic resistance ability of the pathogen. The findings can also facilitate the development of a new tool for identifying conservation and resistance patterns of biofilm formation genes.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

Pseudomonas aeruginosa is a ubiquitous microorganism that belongs to the family Pseudomonadaceae. It is a Gram-negative bacterium known to be an opportunistic human pathogen. The pathogen is responsible for a host of nosocomial infections including those of the urinary and pulmonary tracts as well as wounds and burns (Armour *et al.*, 2007, Marra *et al.*, 2006). Cystic fibrosis patients are also highly susceptible to infections that the bacterium causes.

P. aeruginosa is a persistent pathogen thanks to its antibiotic resistance ability. The resistance is both inherent and acquired. Numerous studies have been carried out in an effort to understand the mechanisms that confer these unique characteristics. The sequencing of the genome of *P. aeruginosa* was a major breakthrough in the efforts of elucidating the attributes of the pathogen (Stover *et al.*, 2000). Whole-genome analyses have replaced the conventional pathogenesis research. The latter techniques laid emphasis on individual virulence determinants while whole-genome analyses allow researchers to further interrogate the reasons behind the phenotypic characteristics. The sequences are also resourceful as large scale analyses can be efficiently carried out.

2.2 *Pseudomonas aeruginosa* Strains

Most of the strains of *P. aeruginosa* that have been sequenced to date were isolated from human infections. The strains are classified under three clades that have distinct phylogenetic origins (Freschi *et al.,* 2015) as well as a distinct evolutionary history (Bruggemann *et al.,* 2018). Group 1 contains most of the sequenced strains including PAO1 which is used as the reference strain for most studies (Stover *et al.,* 2000). The ST235 clonal linage is part of the Group 2 strains that are known to be exoU-positive (Lee *et al.,* 2006). The exoU gene contributes to the virulence ability of the pathogen. The Group 3 clade is the least populated and includes the PA7 strain that also serves as a reference strain (Roy *et al.*, 2010). A previous study has indicated that most of the newer strains, isolated between 2011 and 2016, exhibit different phylogenetic characteristics from strains that were isolated between 1994 and 1998 (Brugemann *et al.*, 2018). Very few studies have analyzed the phylogenetic differences between *P. aeruginosa* strains that were isolated over a large time period.

The PAO1 was the first strain of *P. aeruginosa* to be sequenced (Stover *et.al*., 2000) and has served as the default reference strain ever since. The genome of PAO1 has 6.26 million base pairs and is among the largest bacterial genomes to ever be sequenced. The complex genome structure allows it to thrive in highly diverse environments. Ninety percent of the genes code for different functional and structural proteins. It is no wonder that *P. aeruginosa* occupies different niches within its host. The PAO1 genome has been used to develop the *P. aeruginosa* GeneChip (Affymetrix). 117 ORFs that are not associated with the PAO1 strain are also included in the GeneChip. Other than the PAO1 strain, numerous other strains of *P. aeruginosa* have been sequenced. They include the PA14 strain, PACS2 strain, PA2192 strain as well as the C3719 strain that was isolated from the Manchester epidemic (Jones *et al.,* 2001, Mathee *et al*., 2008).

The second *P. aeruginosa* genome sequence was published for the ExoU-positive strain PA14 (Lee *et. al*, 2006), a clinical isolate displaying higher virulence than PAO1. Fifty-four PAO1 regions of at least one open reading frame (ORF) are absent in the PA14genome, and 58 PA14 regions are absent in PAO1 including the PA14 pathogenicity islands PAPI-1 and PAPI-2 (He *et al.*, 2004). LESB58, widely known as the “Liverpool epidemic strain,” was found to be highly transmissible among CF-patients and displayed the potential to cause severe infections even in non-CF human hosts (McCallum *et al*., 2002). The LESB58 genome contains previously unknown accessory genome elements (Winstanley *et al*., 2009). PA7 is a clinical isolate from Argentina with a notably unusual antimicrobial resistance pattern. Strain PA7 shares only 93.5% nucleotide identity in the core genome with the other sequenced strains confirming the previous assignment of strain PA7as a taxonomic outlier within the species *P.aeruginosa* (Roy *et al.,* 2010).

The distinct identity of the different strains is made clear by single nucleotide polymorphisms (SNPs) on the core genome. The genomic islands also indicate the individuality of strains that are isolated from the same setting. These differences demonstrate the ability of strains to adapt to their respective environments (Bruggemann *et al*., 2018). Highly related strains can be differentiated thanks to the analysis of whole genomes that is now possible. Advanced whole genome sequencing (WGS) technologies provide resolutions that can facilitate the resolution of such strains especially during routine screenings. Resistance markers in the different strains can also be identified with WGS.

The specific genomic islands on their accessory genome infer specific functions and allow different *P. aeruginosa* strains to survive in specific ecological niches. The isolated strains are acquired from different sources of the body including but not limited to the blood, surgical secretions, bronchoalveolar lavage and tracheal secretions (Snyder *et al*., 2013). The phylogeny of the various strains does not however restrict the ability of the strains to inhabit specific body sites. A previous study indicated that isolates sourced from different sample types can be found in the same clade (Yan *et al.*, 2017). The diverse strains of *P. aeruginosa* have an overall genome similarity. A good number of phenotypes however show distinct characteristics as is evident when comparing the pathogenicity of different strains (Lee *et al.,* 2006). Previous studies have indicated that the PA14 strain exhibits more virulent properties than the PAO1 strain (Choi *et al*., 2002). A functional analysis of genes of different strains reveals the genes responsible for pathogenicity as well as those that allow the bacteria to survive in different ecological niches. More often than not both of these mechanisms are multifactorial. Virulence factors are contained on the core genome and on specific virulence islands. Virulence genes in *P. aeruginosa* include those that encode secreted toxins like exotoxin A and ExoU. They also play an active role in processes that are related with the pathogenesis of the bacteria. These processes include quorum sensing and motility (Lee *et al*., 2006).

2.3 The *Pseudomonas aeruginosa* Genome

The genome of *P. aeruginosa* comprises of 5,570 open reading frames (ORFs). Simple eukaryotes have a similar number of ORFs. The genome is also characterized by a 66.6% GC content which influences the stability of the genome in different environments. The four ribosomal RNA loci (rrn) indicate the presence of long repeats in the genome (Stover *et al.*, 2000). The genome comprises of a core genome as well as accessory genomes that are specific for different strains. This was revealed when an analysis of clinical and environmental isolates was done (Lee *et al.*, 2006). The accessory genomes facilitate the survival of the bacterium in different ecological niches. The core genome comprises of 5,021 genes which perform most of the housekeeping duties (Mathee *et al*., 2008). Previous studies have associated 1,800 ORFs with the accessory genome and the ORFs perform a myriad of roles. Some are associated with niche adaptation while others are involved with bacteriophages and transposons (Mathee *et al.*, 2008). Previous studies have used DNA microarrays to study the responses of the genome to different stimuli (Goodman and Lorry, 2004). Very few such studies have exploited the use of comparative genomics tools to identify the differences and similarities of the accessory genome. This study sought to use such tools in a bid to point out gene clusters responsible for the survival mechanism of the bacteria different niches within the human host. The expression of the genes of *P. aeruginosa* is tightly controlled as is evident in the transcriptional regulators that comprise 10% of the genome (Stover *et al.,* 2000). These regulatory genes also form part of the accessory genome of different strains.

2.3.1 Genomic Islands

Genomic islands are the portion of genomes that are highly variable and are acquired through horizontal gene transfer. These sequences have been implicated in the spread of antimicrobial resistance especially in low income settings. The development of next generation sequences have improved metagenomic studies to step up the fight against antimicrobial resistance. Genomic islands are strain specific and consist of divergent DNA sequences that perform similar or related functions. Some of the genes in these islands are not present in other strains. These features were identified by whole-genome sequencing of different islands (Spencer *et al.,* 2003). The presence of genomic islands within bacterial genomes can be deciphered by a number of software tools including the Seqword GI sniffer along with the available database of genomic islands (Bezuidt *et al*., 2009). Such GI identification software tools have facilitated the analyses of important genomic islands in different studies.

2.4 Ecological Niches

P. aeruginosa strains reside in both inanimate and human environments with ease. This is possible due to the presence of numerous enzymes that allow it to use diverse substances as nutrient sources. Groups tend to form biofilms (bacterial communities that adhere to a variety of surfaces) that enhance their survival capabilities. The bacteria is characterized by an attached-for-survival mechanism and once attached they are difficult to destroy. This feature is also evident whenever the bacterium makes its way into different human hosts.

2.4.1 Biofilms

Biofilms protect *P. aeruginosa* from antibodies as well as phagocytes and this creates chronic infections in cystic fibrosis patients. During biofilm formation, the bacterium secretes polymeric substances once it has attached to its desired surface. They then develop protective communities which allow them to thrive even in the presence of antibiotics (Valentini and Filloux, 2016). Biofilms partly contribute to bacterial antibiotic resistance. Anti-biofilm therapies are potential treatment option that can be considered in the efforts to curb antimicrobial resistance in *P. aeruginosa* (Kearns, 2013). This is however only possible if the genes that encode for biofilm formation and behavior are well understood. This study sought to characterize gene clusters that are responsible for biofilms in the different strains of *P. aeruginosa* and provide insight on various aspects that may be targeted by potent therapies. Biofilm formation in bacteria is closely linked with the swarming phenomenon although the two characteristics are not similar (Kearns, 2013). The clearest difference is that biofilms are sessile while swarms are motile (Patrick and Kearns, 2012). Different sets of genes are also expressed when either of the behavior is witnessed in bacteria. The opportunistic pathogen exploits the synergy between the two, among other features, to increase its chances of survival within the human host (Kearns, 2010).

Bacteria use an intracellular messenger to decide whether to form biofilms or opt for better conditions through swarming. c-di-GMP (Bis-(3’-5’)- cyclic dimeric guanosine monophosphate), a ubiquitous secondary messenger, is greatly effective to this end (Csete and Doyle, 2004.) The levels of c-di-GMP within the cell are regulated by a host of proteins that act in response to different stimuli including contact with specific surfaces. Their regulatory activities modulate the expression of downstream genes (Hengge, 2009). High c-di-GMP levels activate biofilm matrix genes and represses flagella genes necessary for swarming motility. The situation is reversed when the intracellular c-di-GMP levels are high and is facilitated by FleQ, an enhancer-binding protein. The overall effect is the efficient co-regulation of genes responsible for swarming motility and biofilm formation (van Ditmarsch *et al.,* 2013; Matsuyama *et al.*, 2016). Antibiotic therapies against *P. aeruginosa* infections can exploit this co-regulation mechanism.

2.4.2 Quorum Sensing

P. aeruginosa secretes numerous products including lectins, rhamnolipids, elastase and pyocyanin that contribute to its pathogenic activity (Soberón-Chávez, *et al.,* 2005). Quorum-sensing regulates the production of most of these pathogenic determinants (Williams and Cámara, 2009). It is a complex regulatory network that determines most of the functions and activity of the pathogen within its host (Schuster and Greenberg, 2006). Bacteria secrete autoinducers that facilitate communication between cells during quorum sensing. The autoinducers bind to transcriptional regulators that modulate gene expression when they reach a threshold concentration depending on the bacterial-population density. Quorum sensing is therefore effective in the control of gene expression in high population densities (Popat *et al.,* 2015).

Given that *P. aeruginosa* has a high intrinsic antibiotic resistance, alternative non-antimicrobial treatment approaches are being considered. One of the approaches that has been investigated in great detail involves impairing the quorum sensing cascade in the pathogen (O'Brien *et al.*, 2015, Jakobsen *et al.,* 2013). The promise that such strategies have shown allows us to explore other pathways in the bacteria that facilitates its survival within the human host. A clear understanding of the gene clusters responsible for formation of biofilms allows the scientific body to consider developing treatment regimens that target this aspect of the pathogen.

2.5 Antibiotic Resistance

The development of antimicrobial resistance (AMR) by bacteria continuous to be a global challenge (WHO, 2014). Antibiotics are available over the counter and their use is highly unregulated thus contributing to AMR. Misuse of antibiotics has taken most of the blame for the rise of this new phenomenon in human and animal pathogens as it encourages the development of resistant genes in pathogens. These portions of the genome are acquired by mutations and horizontal gene transfer. A lot of concerted efforts have been put in place to reverse this trend. Among these efforts was an attempt to understand the dynamics of AMR. A few studies have looked at the occurrence of the phenomenon in different settings within Africa (Nyangacha *et al.*, 2017). Web based resources contain information that can be used to further demystify the whole concept of AMR. This study relied extensively on web based resources. Numerous factors have contributed to the emergence and fast rise of bacteria antimicrobial resistance. Top on that list is the variable accessory genome in different strains of *P. aeruginosa.* Horizontal gene transfer is largely responsible for the genes present in these portions of the genome (Juhas *et al.*, 2009, Bellanger *et al.*, 2014). AMR is also a result of a host of genomic islands present in the genome that can reside in the host's chromosomes and transfer between different hosts.

Resistant genes on the plasmid undergo mutations that further compound the antibiotic resistance ability of *P. aeruginosa*. The mutations can extend to chromosomal genes leading to altered functions which contribute to AMR (Lister *et al.*, 2009). Various antibiotic resistance mechanisms have also been implicated. Modifications of drug targets, alteration of membrane permeability and up regulation of efflux pumps due to mutations are some of these mechanisms. Bacteria also acquire numerous enzymes including carbapenemases, 16S rRNA methylases and extended spectrum β-lactamases that make them resistant to different antimicrobials (Lister *et al.*, 2009).

These mechanisms are further enhanced by non-synonymous Single Nucleotide Polymorphisms (nsSNPs) on the genes related with the processes.

P. aeruginosa is also known to develop resistance to particular antibiotics in the course of treatment. This poses a challenge to clinicians as they seek to prescribe the most efficient treatment regiments for their patients (Septimus and Kuper, 2009, Nathwani *et al.,* 2014). Multi-drug antibiotic resistance in *P. aeruginosa* is a combination of all these processes. Pathogenicity related genes interact in various ways and the end result is an augmented ability of the pathogen to resist the effects of different antibiotics (Alekshun and Levy, 2007, Gooderham, 2009, Jansen 2016). The complexity of antibiotic resistance in different pathogens led to a need of predicting and detecting the phenomenon *in silico.* The Comprehensive Antibiotic Resistance Database (CARD) along with novel bioinformatics tools have made this possible. The tools developed offered a more sophisticated way of analyzing the genome of antibiotic resistant pathogens. This provided new drug targets that could be exploited in the bid to identify potent treatment options. On the other hand, scientists are looking to develop therapies that won't target the bacterium directly but will rather impede its survival mechanism within the host. Such drugs are less likely to be affected by antibiotic resistance. One survival of interest is biofilm formation and this study sought to determine gene clusters that are responsible for this mechanism in a bid to decipher novel therapeutic targets for clinical intervention.

2.6 Web Based Informatics

During sequence analysis, relevant sequences are sourced from public databases and refined where need be. Different computational tools are then used to predict features of interest in the retrieved sequences. The evolutionary history, structure as well as function of genes can all be revealed when sequences are analyzed. Homologues can also be identified accurately (Mehmood, Sehar and Ahmad, 2014). The *Entrez* tool of PubMed is widely used for retrieval of data as it is linked to numerous biological data domains (Geer and Sayers 2003). In a recent study that sought to characterize the chitin binding protein, the sequences were exclusively retrieved from Uniprot. The study successfully predicted the function of the gene *cbp*50 found in *Bacillus thuringiensis* using a host of sequence analysis techniques (Sehar *et al.,* 2013). Lichun and Jie, (2018) also utilized datasets from NCBI repositories as they sought to demonstrate the value of single-cell RNA-sequencing technology in type-2 diabetes studies. Their findings revealed that the technology could provide valuable information on various molecular processes of the disease. They used different machine learning classifiers to determine different influential factors in the development of the disease (Lichun and Jie, 2018)

2.7 The Hidden Markov Model (HMM)

The past few years have witnessed an astronomical rise in the number of genomes in the NCBI database (Land *et al.,* 2015). The database has become an invaluable source for the analysis of specific genes in different organisms. *In silico* mining of data is emerging as an important tool for researchers looking to shed more light on different characteristics of various organisms to address threats within the health care sector including antimicrobial resistance. The Basic Local Alignment Search Tool is the most common sequence-based approach used for mining sequence information of microorganism of interest. Along with other in silico mining tools, BLAST relies on Hidden Markov Models (HMMs) a statistical method commonly used to model biological information (Altschul *et al*., 1990, Sherlock *et al.,* 2013, Seifert *et al.,* 2014, Morton *et al.,* 2015, Weber *et al.,* 2015). The vast amount of data has necessitated the development of special tools including profile Hidden Markov Models (pHMMs) to facilitate rapid analyses of the sequences (Francisco *et al.*, 2019). The pHMMs are a specific subset of HMMs that apply a statistical model to represent the motifs and patterns of a multiple sequence alignment. With an observed frequency of a nucleotide, pHMMs are designed to estimate their true frequency at a specific position in the alignment (Yoon, 2009). From the multiple sequence alignments, profile methods build position-specific models to represent the conserved regions in the alignments. Pairwise methods like FASTA and BLAST are, however, more popular mainly due to the statistical theory that supports such methods (Altschul and Gish, 1996). The advent of hidden Markov models (HMMs), probabilistic models, has provided an elaborate theory for profile methods.

The hidden Markov model is used to describe the probability distribution of an infinite number of sequences. This model assigns constrained scores as the probability distribution sums to one. The HMM's parameters have non-trivial optima given that the probability of one sequence decreases the probability of other sequences. With HMMs, the state sequence, which in this case is a biologically meaningful alignment, cannot simply be determined from the observed symbol sequence rather it is probabilistically inferred from the observed symbol sequence.

2.7.1 Hidden Markov Model Probabilistic

HMMs work as models which generate sequences. The model starts with an initial state then moves to a new state with some transition probability. In this case, the state 1 can be maintained with a transition probability t1,1 or the move to state 2 with a transition probability t1,2. Once the next state is determined, we'll have a residue whose emission probability is specific to that state. The model iterates the transition/emission process until it arrives at the end state. Complete HMMs have both an observable symbol sequence and a hidden state sequence that cannot be seen. Standard dynamic programming algorithms are used to align and score sequences with the constructed model (Durbin *et al.*, 1998). The Viterbi algorithm is used for alignment while the Forward algorithm is used for scoring sequences.

2.7.2 Building the Hidden Markov Model

Setting parameters for an HMM can be done in two different ways. One could either train the model from unaligned (unlabeled) sequences or choose to build an HMM from pre-aligned (pre-labeled) sequences. For the pre-labeled sequences, it is assumed that the state paths are already known. With this sequences, the model simply converts state transitions and observed counts of symbol emissions into probabilities. The alignment is used as input for building the profile HMM. Training a pHMM, on the other hand, is slightly complex as it requires one to run a multiple alignment program before they can build the model (Eddy, 1998).

While it is a harder problem, training a model is an interesting alternative given that a plausible alignment of a group of sequences is not known. For training, we could either opt for the Baum-Welch expectation maximization algorithm or choose the gradient descent algorithm. Different studies have highlighted simulated annealing, genetic algorithms and Gibbs sampling as suitable training methods. Apparently, these methods can efficiently avoid spurious local optima while training HMMs (Durbin *et al.,* 1998, Eddy, 1996). Such algorithms often seek simple maximum a posteriori or maximum likelihood optimization targets. Alternatively, the algorithms can use more sophisticated optimization targets to maximize the model's ability in discriminating true positive sequences from true negative training examples (Mamistuka, 1996). The sophisticated targets also compensate for biased representation which can be attributed to non-independence of example sequences (Bruno, 1996, Sunyaev *et al.*, 1998).

Building HMMs from multiple alignments is a suitable option given that the training algorithms are simply local optimizers. Meaning that such algorithms work suitably for less complex HMMs. For complicated HMMs, the training algorithm could easily get trapped by numerous spurious local optima thanks to a complex parameter space (Eddy 1998). The existing maximum likelihood architecture construction algorithm facilitates the building process of HMMs from pre-aligned sequences (Durbin *et al.*, 1998). General HMMs can also be constructed from the architecture learning algorithms (Yada *et al.*, 1996). Alternatively, training can be done for fully connected HMMs before low-probability transitions are pruned once training is complete (Mamitsuka, 1996).

2.7.3 Profile Hidden Markov Models

Krogh *et al.*, (1994) introduced the profile HMM, a strongly linear, left-right model (Krogh *et al.*, 1994). The pHMM uses a 'match' state to model the distribution of residues in each consensus column of the multiple alignment. This state is complemented by the 'insert' and 'delete' states which account for an insertion of residues after that column or deletion of the consensus residue, respectively. With the pHMM the probability parameters are converted to additive log-odds scores. This is done before a query sequence is aligned and scored (Barrett *et al.*, 1997). The score from an aligned residue to a match state resembles how FASTA and BLAST scores are derived. For example, take the probability of match state emitting residue x as p_x and the background frequency of the same residue in the database as f_x. The score of the residue in this state is given as log p_x/f_x.

Profile HMMs have useful, non-trivial optima as their gap costs are not arbitrary numbers. While scoring pHMMs and assigning state transition probabilities, a trade-off point is made to create a balance between sequences that have an insertion against those without an insertion. The inserted residues in these models also have emission probabilities, 4 for nucleic symbols and 20 for amino acid symbols. The score of the inserted residue is given as log $f_x/f_x = 0$ if the emission probability is similar to the background amino acid frequency.

2.7.4 Profile Hidden Markov Models Optimization

To increase the accuracy of the database search using the pHMM, studies can focus solely on the analyses of conserved residues as suggested by Ahola *et al.,* 2003. In this case the efficient emission probability (EEP) estimation method is employed in the construction of the individual profile hidden Markov models. At each conserved alignment position, the *hmmbuild* algorithm divides amino acids into effective and ineffective residues. This is done to ensure that the constructed profile separates the signal from the noise in the conserved positions of the sequence alignments, hence overcoming the overfitting problem common for HMMs. With the EEP technique, the pHMMs model conserved residues rather than only focusing on the general characteristics of different amino acids. This technique combines different residue conservation scoring methods with great flexibility (Valdar, 2002). Given that only a few residues can be considered as effective for protein sequence alignments, the EEP technique helps to decrease the parameter space dimensions. With a reduced parameter space, the variance of effective residues is not compromised by the variance of ineffective residues which is likely to reduce. This is appreciated most when we calculate the confidence intervals for the different emission probabilities. The shorter confidence intervals improve the sensitivity of database search results as the model's prediction power is improved.

2.7.5 Profile Hidden Markov Model Acceleration

While pHMMs have attractive advantages, the utility of BLAST made it a suitable option over the probabilistic sequence comparison methods (Eddy, 2011). The slow analyses speed and computational expensive nature of the models restricted their use for sequence homology and similarity searches. For protein domain family analysis, however, the profile's ability to represent numerous homologous sequences compensated for the speed differential (Finn *et al.,* 2016).

HMMER3, introduced in 2011, uses the Multiple Segment Viterbi (MSV) algorithm that makes it 100-faster than previous HMMER algorithms. This implementation also uses the Forward/Backward evaluation of alignment ensembles which exploits the most of the mathematical advantages offered by probabilistic modeling techniques. With this new implementation, scientists could now run homology searches as fast as BLAST and still enjoy the multiple advantages of profile HMM methods. To achieve the MSV probabilistic model, the algorithm treats match-match state transitions as 1.0 and ignores the match, delete, and insert transitions of the original profile. While an MSV score is similar to BLAST's "sum score", the algorithm bypasses the hit extension and word hit heuristics and calculates the score by dynamic programming. This approach makes it more sensitive than BLAST's approach, although this finding is not conclusively proven (Eddy, 2011).

Besides the better sensitivity, the MSV algorithm can be used and selective sequence filter as the p-values of the algorithm can be calculated. Target sequences whose p-values are less than the chosen threshold are assumed to be non-homologous to the sequences used in model construction. This study set a threshold p-value for analyses of *P. aeruginosa* strains using the constructed models. Previous studies have indicated the MSV algorithm made HMMER3 performance comparable to other fast database search programs like NCBI BLAST, WU-BLAST, and SSEARCH (Eddy, 2011). The HMMER3 acceleration pipeline involves an MSV filter, bias filter, Viterbi filter, Forward algorithm, Backward algorithm. All these processes are put together to reduce the computational needs of the software package while maintaining its accuracy and accelerating the homology search process. The pipeline is designed to use p-values of the log-odds score to either accept or reject comparisons at different steps. The assumption made is that the residue compositions of the target sequence and query profile are close to the overall average for proteins. In cases where biased composition is expected, HMMER3 recalculates scores and p-values to compensate for the bias. The bias filter in the acceleration pipeline also reduces the problem of underestimated biased matches as it removes any additional matches that may be due to biased composition. The full Forward algorithm calculates the final reported score of the sequence. The Viterbi filter, on the other hand, is designed to reduce the computational load expected from the Forward step. Specialized memory-efficient forms are used to implement both the Forward and Backward algorithms in a bid to reduce the computational load from these processes. These two probabilities estimate local alignment "regions" with considerable posterior probability mass in the target sequence. The final step of the acceleration pipeline subjects each region to the "domain definition" pipeline which is a conceptually separate analysis pipeline. This step uses a series of algorithms to identify individual homologous regions and alignments.

2.7.6 Applications of the Profile Hidden Markov Models

HMMs have been used for numerous biological applications including, phylogenetic analysis, gene finding, protein secondary structure prediction, radiation hybrid mapping, and genetic linkage mapping (Felnestein and Churchill, 1996; Goldman *et al.,* 1996; Kruglyak *et al.,* 1996; Slonim *et al.,* 1997; Lukashin and Borodovsky, 1998). Profile HMMs are suitable for modeling protein and nucleotide sequence data as they move in one direction along the alignment and do not need cycles (Skewes-Cox *et al.,* 2014). They assess each column of a multiple sequence alignment looking out for the three types of hidden states i.e the match state, insert state or delete state. The three states respectively describe the frequencies of residues, insertions and deletions on the sequences being analyzed (Yoon, 2009). Although sequence homology approaches like BLAST are lauded for their speed, profile HMMs have been shown to demonstrate more sensitivity especially when detecting distant homologs. This superiority is achieved thanks to the models' emphasis on function-dependent conserved motifs rather than a focus on the overall sequence similarity (Park *et al.,* 1998, Madera and Gough, 2002).

Profile HMMs are highly effective in the analyses of extensive amounts of data. Characterization of protein families, comprehensive sequence analyses and gene discovery are some of the applications that the model has made possible (Restrepo-Montoya *et al.,* 2011). A single profile HMM can successfully detect sequences that belong to specific profiles as they are intrinsic in nature (Gong *et al.,* 2012). The tool has also been used in the analysis of metagenomic sequence data to identify viral protein sequences (Skewes-Cox *et al.,* 2014). In one study the model was used in the analyses of 857 genomes of *Bacillus* spp. to find Cry genes. The profile revealed that Cry proteins are not only restricted to *Bacillus thuringiensis* as it was initially presumed. Some of the protein sequences were present in *B.cereus* and other unidentified bacilli. The report also indicated the value of a systematic search for specific proteins to solve questions regarding the role of such proteins in nature (Francisco *et al.,* 2019). The Pfam database, a collection of protein families, largely depends on HMMs as well as multiple sequence alignments. The models have facilitated the characterization and classification of different protein families in the database. Different other studies have identified signal peptides, carotenoid genes as well as transmembrane proteins based on the profile Hidden Markov Model (Tonhosolo *et al.,* 2009).

This study designed, validated and implemented specific profiles to search for biofilm formation genes in the genomes of *P. aeruginosa* species. The designed profile HMMs revealed the conserved and variable patterns among gene clusters responsible for biofilm formation. The results of this study pointed out novel therapeutic targets for clinical intervention that could be explored by pharmaceutical companies in designing candidate drugs for management of the pathophysiology of *P. aeruginosa* infections. It is hoped that the findings of this study help inform policy management of *P. aeruginosa* pathogenesis. Given the large number of genomes present in the NCBI gene bank, an *in silico* approach to characterize the genes of biofilm formation in *P. aeruginosa* was a feasible attempt.

CHAPTER THREE
MATERIALS AND METHODS

3.1 Sequence Retrieval

3.1.1 Entrez Search

An Entrez gene search engine was performed on the NCBI database to identify the genes responsible for biofilm formation in *P. aeruginosa.* '**Biofilm AND *Pseudomonas aeruginosa*[ORGN]**' was used as the query term and the results were obtained from the gene database of NCBI. Both the FASTA and GenBank file formats were downloaded for analyses. Only the sequences available by December 2018 designated with the "*P. aeruginosa"* tag were selected for the analysis. The genes responsible for biofilm formation were further classified into different categories based on the functions they perform in the biofilm formation process. A literature search was performed for each gene to facilitate this categorization. The GenBank files of the individual genes also informed this categorization as they included the gene annotation information along with metadata on these sequences.

3.1.2 Multiple Sequence Alignment of Biofilm Formation Gene Sequences

Before evolutionary analyses, and other downstream analyses, of the retrieved sequences could be done, the study sought to identify whether these genes were true homologs. This was done through multiple sequence alignments of the gene sequences. Identifying homologous sequences by sequence similarity searching has been recognized as one of the first and most informative steps in any analysis of new sequences (Pearson, 2013). BLAST, FASTA, HMMER3, and PSI-BLAST, popular similarity searching programs, produce accurate statistical estimates to infer homologous sequences (Pearson and Lipman 1988, Altschul *et al.,* 1997, Johnson *et al.,* 2010). For this study, the MUSCLE algorithm available in MEGA X platform was used to run a multiple sequence alignment of the 51 biofilm formation gene sequences obtained from the initial step (Edgar 2004, Kumar *et al.*, 2018). Default parameters of MUSCLE were used for this analysis. The gap penalties were set as follows: gap opening penalty (-400.0) and the gap extend (0.00). The UPGMA algorithm was used for clustering while the number of iterations was set at 16. The Min Diag Length (Lambda) was set at 24. Sequences that shared significant similarity would be inferred to be homologous and used for downstream evolutionary analyses (Pearson WR 2013).

3.1.3 *Pseudomonas aeruginosa* Whole Genome Sequence Retrieval

The study collated whole genomes of different strains of *P. aeruginosa* to determine the distribution of biofilm formation genes in the different strains. Complete genomes were downloaded from NCBI and the International *Pseudomonas* Consortium Database (https://ipcd.ibis.ulaval.ca/) as GBK files and FASTA files. From NCBI, the '*Pseudomonas aeruginosa*' query term was used and the search was done from the 'genome' database. The study selected the 'complete genomes tag' in the list of available *P. aeruginosa* to narrow down the search further. From the IPCD database, the study downloaded the 'Genomic DNA (FASTA) complete genomes' file. A manual curation was done to identify only the strains of *P. aeruginosa* from this sequence file. The study only selected "the complete genomes" available by December 2018 from these two databases. The genome sequences were then categorized based on the ecological niches that they occupy. This classification was facilitated by the metadata (bioproject and biosample data) of the different strains available in the two databases.

3.1.4 Sequence Retrieval Using Python Script

The initial sequence retrieval of biofilm formation genes from the NCBI database successfully identified 51 non-homologous gene sequences. These sequences could not, however, be aligned against each other given that they had no homolog sequences as was indicated by the multiple sequence alignment of the initial retrieved sequences**.** For a comprehensive analysis of the phylogenetic relationship between the biofilm formation genes, the study sought to collect the same genes from representative genomes of *P. aeruginosa*. A custom python script was written to facilitate the retrieval of the biofilm formation gene sequences from the annotated genomes of *P. aeruginosa* (Appendix 1). The script was designed to create FASTA files for every gene, containing sequences of the respective genes selected from every reference genome. The GenBank files of these strains were used in this case as they are annotated – indicating the specific functions of different regions of the sequence. Custom python scripts have previously been used by Awal *et al.*, (2017)'s study to separate individual genes from the fasta files of chloroplast genomes annotated by NCBI. Python scripts were also used to select designed primers from *Trticum* genomes (Awad *et al.*, 2017). The Python scripting language has also played a pivotal role in unifying various data sources and analysis tools. With custom python scripts the bioinformatics community gets to efficiently retrieve, analyse and summarize data streams in a single workflow (Gilpin 2016).

3.1.5 Evolutionary Analyses of Biofilm Formation Genes

Besides deciphering gene and protein function, phylogenetic analysis has proven to be a useful tool for understanding organismal relationships. Using phylogenetic (gene by gene analysis) methods to compute the relatedness of organisms reveals high-quality results that can accurately demonstrate phylogenies (Mansour, 2009). This study sought to understand the evolutionary relationships between the biofilm formation genes that had been successfully retrieved. This analysis would help to identify whether these set of genes co-evolved from the same parent organism or whether they have been acquired through horizontal gene transfer in the course of the existence of *P. aeruginosa.* For this evolutionary analysis, the study used the 13 fasta files (sequence files) created from the custom python scripts. These sequences were assumed to be largely homologous as the 44 records in each file were retrieved from similar strains of *P. aeruginosa*. For every COG of biofilm formation genes found in selected reference different strains, *n*=13, the study conducted multiple sequence alignments using the MUSCLE algorithm (Edgar, 2004). This was done to test the similarity between the sets of sequences and identify any potential polymorphisms. The default values of the MUSCLE algorithm were used for the MSA. Awad *et al.,* (2017) used a similar approach as they sought to identify effective DNA barcodes for *Triticum* plants through chloroplast genome-wide analysis (Awad *et al.*, 2017). Individual COG alignments were edited by the program Gblocks and then concatenated into a super-alignment used for a phylogenetic inference by the Neighbour-Joining (NJ) algorithm implemented in the program of PHYLIP 3.69 (Phylogeny Inference Package) (Castresana, 2000, Felsenstein 2005). From the aligned sequences, the study created a maximum likelihood phylogenetic tree for each COG (specific for the individual biofilm formation genes.) To reach branch confidence values, bootstrap with 1000 iterations were set for the phylogenetic tree construction. This phylogenomic approach is used to infer phylogenetic trees for organisms based on concatenated alignments of multiple concatenated alignments of multiple orthologous genes. The initial phylogenetic analysis would be useful in identifying the evolutionary relationship between the 44 *P. aeruginosa* sequences used for downstream analyses. The resulted phylogenetic trees were compared by the program treedist of PHYLIP 3.69 using the Branch Score Distance algorithm (Felsenstein, 2005). Distances between the trees were saved into a distance matrix. The PHYLIP package was preferred as it has previously provided the foundation of population genetics (Volgyi *et al.*, 2009). The module was also used to reveal the genetic evolution of different *Klebsiella pneumoniae* strains isolated from diarrhea specimens and estimate the phylogeny of different mycobacterial strains (Guo *et al.*, 2008; Mignard and Flandrois 2008). An in-house Python script was used to reformat the treedist text output file into a matrix of distances between COG-based ML phylogenetic trees in PHYLIP format suitable for clustering of the trees by the NJ algorithm. The neighbor-joining algorithm was then used to build a tree representing the evolutionary relationships of the biofilm formation genes based on the distance matrix. The resultant dendogram was used to analyze grouping of co-evolved biofilm formation genes. Clustering of several genes together would mean their co-evolution while separating genes to different clusters would mean horizontal gene transfer exchange.

3.2 Construction of the Hidden Markov Models

3.2.1 Identification of the Protein Family of Interest

The construction of profile hidden Markov models is based on multiple sequence alignments of DNA or proteins sequences from the same functional family. The pHMM is used to represent the patterns, motifs along with other statistical properties of the alignments. Before the actual construction is performed the protein family of interest under investigation should be selected. The criteria identified by Henikoff *et al.,* was used in this study to identify the protein family of interests (Henikoff *et al.,* 1997). In this case the protein family of interest would represent a set of genes performing similar functions in different strains of *P. aeruginosa*, the pathogen under study.

3.2.2 Selection of Sequences Representative of this Family

With a protein family of interest identified, representative sequences from these sequences were selected for model construction. Sequences initially retrieved from the custom python scripts were selected in this case to provide the most informative findings in the downstream analyses. A 40% sequence similarity threshold was chosen as was suggested by Rost's study which indicated that long alignments greater than 40% are ideal for providing unambiguous results (Rost, 1999). Sequence files with similarity identities less than 40% were excluded from the downstream analyses.

3.2.3 Multiple Sequence Alignment Generation

To create a pHMM of each of the classes of biofilm formation genes, a global multiple sequence alignment was first generated using MUSCLE (v.3.8.31) (Edgar, 2004). The MUSCLE algorithm used for these analyses is available in the UGENE platform (Okonechnikov *et al.,* 2012). The study used the default values for the gap open penalty (54.00), gap extension penalty (8.00), and terminate gap penalty (4.00). These gap penalties were used to control the positions of the conserved regions within the alignment. The consensus sequence from these global alignments would inform the construction of the specific profile hidden Markov model.

3.2.4 Building of Profile HMM

The study used profile analysis to incorporate information concerning the conservation of different residues. Analyses from the constructed profiles of biofilm formation genes would be used to detect homologies and structural similarities between the sequence families of *P. aeruginosa*. To facilitate an efficient search of the database of homologous sequences, position-specific information from multiple alignments were a suitable option. From the multiple sequence alignments, the profile method built position-specific models to represent the conserved regions in the alignments. The state sequence, which in this case was a biologically meaningful alignment, was probabilistically inferred from the observed symbol sequence rather than simply being determined from the observed symbol sequence.

The parameters of the gene-specific models were set from the pre-aligned (pre-labeled) sequences. In this case the study assumed that the state paths were already known given that the multiple sequence alignments had been optimized. The model converted both the state transitions and observed counts of symbol emissions into the transition and emission probabilities, respectively. These probabilities were based on the initially set transition and emission probabilities standards.

The study used the Forward algorithms to score and optimize the gene-specific pHMMs. Alignments from the previous step were used as input for building the profile HMMs. Building HMMs from multiple alignments was preferred in this case as the training algorithms (local optimizers), are suitable for less complex HMMs**.** With a less complex parameter space, there was little chance that the spurious local optima would trap the training algorithm. Given that the study was constructing profile hidden Markov models, the probability parameters were converted to additive log-odds scores. These log-odds scores would later be used to score a query sequence once it is aligned against the constructed model. (These scores resembled the scores derived either by BLAST or FASTA)

The profile HMM was preferred to Artificial Neural Networks and PSI-BLAST algorithms given that it is a well formulated probability model for representing similarity patterns within sequence families. The other models are also known to require more computational power without necessarily providing better results. Artificial Neural Networks and PSI-BLAST are also ideal for large scale analyses that involves complete genome sequences (Altschul *et al.*, 1997). HMMs used in this case targeted only specific sections of the genome. These models also provide a precise method to search sequence databases using aligned sequences (Ahola *et al.*, 2003). From the multiple sequence alignments, the HMMER3 toolkit available on the UGENE software was used to construct profile Hidden Markov Models (pHMMs) for the twelve clusters of orthologous genes (COGs) of biofilm formation genes for analyses of genomes of *P. aeruginosa* strains (Eddy, 1998). For each of the COGs of biofilm formation genes, the study created a profile HMM. The *hmmbuild* algorithm within the HMMER3 tool (v 3.1b1) in the UGENE software was used to create a suitable profile HMM from the MSA aligned-FASTA files (http://hmmer.janelia.org).

The HMMER3 platform was accelerated by the Multiple Segment Viterbi (MSV) algorithm that is implemented in the software package (Eddy, 2011). For better accuracy of the database search using the pHMM, the study employed the efficient emission probability (EEP) estimation method to construct the gene-specific pHMMs. This estimation method ensured that the overfitting problem was overcome as signal was separated from the noise in conserved positions of the alignments, and reduced the parameter space as a result. The confidence intervals of representative emission probabilities were calculated to determine the effectiveness of the EEP estimation method. Shorter confidence intervals would indicate that the model has an improved prediction power.

3.2.5 Validation of Profile HMM

Using the *hmmsearch* algorithm on UGENE, sequences used to construct the model were searched for as positive controls. Sequences of unrelated microorganisms were also searched for as negative controls with a discrimination threshold of $E \leq 1 \times 10^{-5}$. The *hmmsearch* algorithm in the UGENE platform was used in this validation step with the constructed models serving as the query sequence while the controls served as the sequences being analyzed. The presence of signals for the positive control search and lack of signals for the negative control search demonstrated the efficiency in the prediction by the constructed models. The positive and negative controls (listed in Table 3.1) were used to evaluate the ability of the model to correctly identify biofilm formation gene sequences.

Table 3.1 Controls used in the validation of pHMM

Organism	Accession Number
Pseudomonas aeruginosa PA01	NC_002516
Bat Adenovirus 2	NC_015932
Gyrovirus 4	NC_018401
Duck circovirus	NC_007220
Domestic cat hepadnavirus	NC_040719

3.2.6 Visualization of the Model

The visualization of the constructed pHMM was performed using the HMM visual editor (HMMVE_1.2) (Dai and Cheng, 2008)

3.3 *Pseudomonas aeruginosa* Whole Genome Sequence Analyses

To determine the conservation and variation patterns of biofilm formation genes in *P. aeruginosa* strains, the study collated and analyzed whole genomes of *P. aeruginosa*. The *hmmsearch* algorithm within the HMMER3 tool in the UGENE software was used to search for the biofilm formation profile HMM against sequences of *P. aeruginosa* drawn from different ecological niches. This search revealed the number of biofilm formation genes (identified as hits) within the different *P. aeruginosa* analyzed in this study. Only 96 strains of the ubiquitous pathogen were used in this analysis as they were associated with specific ecological niches. The models were restricted to show only biofilm formation genes with more than 30% identity and an e-value lower than 1×10E-5. These parameters have been employed successfully in previous studies that used the profile HMMs to search for sequences (Gong *et al.,* 2012, Munoz-Medina *et al.,* 2015).

3.4.1 Evolutionary Analysis

The study sought to infer the sequence diversity in the sequences of different *P. aeruginosa* strains. The twelve sequence files which contained 44 records each were used for the phylogenetic analyses. Maximum parsimony algorithm on the MEGAX platform was used in the construction of the phylogenetic trees. Each bacteria strain was assigned a clade based upon its evolutionary history. This was done to determine whether the observed patterns are associated with the evolutionary history of the different pathogens. The evolutionary analyses of all the strains of *P. aeruginosa* involved 44 amino acid sequences from 12 sequence files. These analyses were conducted in MEGAX (Kumar *et al.,* 2018). For the maximum parsimony trees, the Subtree-Pruning-Regrafting (SPR) algorithm with search level 1 was used. The initial trees were obtained by the random addition of sequences**.** The bootstrap consensus tree was inferred from 1000 replicates with branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. This analyses were based on the biofilm formation genes rather than the conventional whole genome phylogenetic analyses.

3.4.2 Genome Mapping

The study also sought to determine the genotypic differences between the closely related strains of *P.aeruginosa.* A visualization of genome comparisons was performed to determine these differences. The BLAST Ring Image Generator (BRIG), a cross-platform desktop application, was used to generate circular images (maps) that indicated genome comparisons between the *P. aeruginosa* gene sequences under analysis (Alikhan *et al.,* 2011). This BRIG analysis also helped to visualize the core and flexible genomes against a reference genome. The study used default settings to generate the images of the genome data from each ecological niche initially identified. The output images from this analysis showed similarity between a central reference sequence

(PAO1 in this case) and other sequences as a set of concentric rings. BLAST matches, on the other hand, were colored on a sliding scale which indicated a defined percentage identity. Different colors on the rings indicated significant matches while non-significant matches were represented by blanks. The BLAST matches were filtered according to an E-value cut-off of 1×10^{-5}. These matches were calculated from the perspective of the reference sequence *P. aeruginosa* strain PAO1. In this case, regions of different genomes that were present in the query sequences but absent in the reference genome were not displayed in the resultant maps. For the visualization of genomes from all the ecological niches, data from different genomes were collated into a single lane. This step facilitated the visualization of all the sequences under study allowing us to compare the genomes as a group against the central reference sequence. Specific regions of the reference genome were also highlighted with custom annotations by specifying the position of features. The specific features in this case were the different biofilm formation genes under analysis. Selected annotations were uploaded from a GenBank file i.e *P.aeruginosa* PAO1 GBK file (Stover *et al.*, 2000). The annotations (biofilm formation genes) were shown in the outermost ring of the different genomic maps. A study by Bruggemann *et al.* used the visualization tool to take a closer look at the group 2 clade of *P. aeruginosa* (Bruggemann *et al.*, 2018). The reference genome used in this case was the HIAE_PA17 strain. Other than visualization of the genome of an organism BRIG is also used for genome comparison as was the case in a study by Ramanathan *et al*. In this case the sequences of clinical isolates were compared with the *P. aeruginosa* reference genome PAO1 (Ramanathan *et al.*, 2017). This study followed a similar approach as the PAO1 reference genome was used for genome comparison.

3.5 Statistical Data Analyses

Enrichment tests were performed to identify differentially abundant categories between groups of genomes based on their origin using the non-parametric Mann-Whitney Test (MWT)/Wilcoxon signed rank test. This analysis was done as the data from the profile HMMs did not indicate normal distribution. These tests were done to reveal signatures of niche specialization as was the case in the study by Bai *et al,* 2015. The analyses were performed using custom R-scripts (Appendix 2). Custom R-scripts were written and run in the R-studio version 1.3.1093.

CHAPTER FOUR
RESULTS

4.1 Sequence Retrieval

4.1.1Entrez search

The Entrez gene search identified a total of 51 biofilm formation genes associated with *P. aeruginosa*. The sequences of these genes were downloaded for downstream analyses. Table 4.1 shows the IDs, names and GenBank accession numbers of the biofilm formation genes' sequences used in this study.

Table 4.1 Biofilm formation genes retreived from the NCBI gene database

Biofilm formation genes retreived from the NCBI gene database

ID	Gene name	Accession Number
880925	*pslB*	NC_002516.2
879717	*pslA*	NC_002516.2
883079	*pslH*	NC_002516.2
882276	*pslI*	NC_002516.2
882251	*pslJ*	NC_002516.2
880828	*pslK*	NC_002516.2
879704	*pslG*	NC_002516.2
878490	*pslE*	NC_002516.2
878238	*wspC*	NC_002516.2
878103	*pslC*	NC_002516.2
878051	*pslD*	NC_002516.2
878020	*pslF*	NC_002516.2
3399421	*intl1*	NC_007100.1
882052	*fliC*	NC_002516.2
878885	*PA2824*	NC_002516.2
882125	*algU*	NC_002516.2
882792	*amrZ*	NC_002516.2

881084	*morA*	NC_002516.2
879406	*algC*	NC_002516.2
881782	*rsaL*	NC_002516.2
879474	*gshA*	NC_002516.2
882234	*PA4878*	NC_002516.2
881298	*PA5017*	NC_002516.2
877926	*Lon*	NC_002516.2
879004	*algD*	NC_002516.2
879004	*ppkA*	NC_002516.2
880125	*lipC*	NC_002516.2
882372	*ppyR*	NC_002516.2
880611	*PA0122*	NC_002516.2
879994	*mvfR*	NC_002516.2
880617	*cupC1*	NC_002516.2
879373	*PA1107*	NC_002516.2
881786	*PA1434*	NC_002516.2
878758	*cupB1*	NC_002516.2
881933	*htpG*	NC_002516.2
881493	*PA4332*	NC_002516.2
881355	*PA4398*	NC_002516.2
879143	*arnB*	NC_002516.2
878826	*estA*	NC_002516.2
877798	*amiC*	NC_002516.2
880075	*PA4781*	NC_002516.2
882208	*PA4625*	NC_002516.2
880282	*PA1823*	NC_002516.2
880470	*bfiR*	NC_002516.2
880350	*bfiS*	NC_002516.2
878223	*gshB*	NC_002516.2
878109	*PA2572*	NC_002516.2
877982	*PA4108*	NC_002516.2

882750	*PA2771*	NC_002516.2
881686	*PA1324*	NC_002516.2
881441	*PA4354*	NC_002516.2

Once the biofilm formation genes were identified by the Entrez search, the study sought to classify these sequences into different categories based on the individual functions of these genes in the biofilm formation process. This classification would inform focused analyses of these set of genes; from their evolutionary relationships to their distribution within the genomes of *P. aeruginosa*. A literature search along with the metadata and annotations of the genes informed this functional classification. Five functional classes were identified: adhesins, cell aggregation, repressors, regulatory and motility genes as indicated in table 4.2. The sixth class contained a set of genes whose functional properties have not been fully annotated. The study used the 'Unclassified' tag to identify these set of genes. Further functional analysis of these set of biofilm formation genes could be performed to enhance further studies of the ubiquitous pathogen.

Table 4.2 Classes of biofilm formation genes, number of sequences and the individual genes

Classes	**Sequences**	**Percentage**	**Genes**
Adhesins	3	6%	*PA1107, PA1434, ppyR*
Cell aggregation	2	4%	*cupB1, cupC1*
Repressors	2	4%	*gshB, PA0122*
Regulatory	19	37%	*algC, algD, algU, amiC, arnB, bfiR, bfiS, lon, mvfR, PA1324, PA2572, PA2824, PA4332, PA4354, PA4625, PA4871, PA4878, ppkA, rsaL*
Motility	11	22%	*amrZ, estA, fliC, gshA, htpG, lipC, morA, PA2771, PA4108, PA4398, PA5017*
Unclassified	14	27%	*pslC, pslJ, intl1, PA1823, pslA, pslB, pslD, pslE, pslF, pslG, pslH, pslI, pslK, wspC*
Total	**51**		

Besides the overall statistics, the study sought to further identify the representative percentages of each class of the biofilm formation genes. This was done to clearly represent the class with the highest number of genes. It is important to note that the 51 biofilm formation genes retrieved from the GenBank database were all associated with the *P. aeruginosa PAO1* strain (the reference strain).

4.1.2 Multiple Sequence Alignment of Biofilm Formation Gene Sequences

The multiple sequence alignments revealed high levels of dissimilarity between the retrieved biofilm formation gene sequences. From the extensive gaps and minimal regions of similarity, the study inferred that the sequences retrieved from the *Entrez* gene search were not homologous and could not be used for downstream analyses of these set of genes. These results informed the decision by the study to retrieve sequences associated with these genes from the whole genome sequences of different strains of *P. aeruginosa.*

4.1.3 *Pseudomonas aeruginosa* Sequence Retrieval

The study further sought to retrieve complete genome sequences of *P. aeruginosa* strains isolated from various ecological niches. These sequences would be analyzed by different comparative genomics tools in a bid to characterize the biofilm formation genes with respect to different strains of the pathogen. A total of 194 complete genome sequences of *P. aeruginosa* strains from the NCBI and IPCD databases were retrieved for analysis. The study then classified the retrieved strains based on the ecological niches they occupy. This classification was done to facilitate an informed analysis of the *P. aeruginosa* strains. The study used metadata and GenBank annotations on each strain to complete this classification. 13 ecological niches were identified, 11 in the human host and two catering for environmental isolates. The human ecological niches include abscess, blood, bronchial, clinical, dental, eye, lungs, sputum, trachea, wound and urine. The environmental samples were classified as environmental and cell culture isolates. The remaining isolates lacked comprehensive annotations and were therefore categorized into an unclassified group. Only the classified *P. aeruginosa* strains (n=96) were used for downstream analyses given that they were associated with different ecological niches. Table 4.3 shows niche-specific categories of *P. aeruginosa* isolates which were used for downstream analyses.

Table 4.3 Statistics of Pseudomonas aeruginosa sequences and their ecological niches

Ecological niche	**Analyzed sequences**	**Percentage**
Abscess	2	1%
Blood	15	17%
Bronchial	6	6%
Cell culture	4	4%
Clinical	10	11%
Dental	1	1%
Environment	8	8%
Eye	2	2%
Lung	1	1%
Sputum	26	27%
Trachea aspirates	5	5%
Urine	7	7%
Wound	9	9%
Total	**96**	

This selection does not represent any real prevalence of *P. aeruginosa* in nature. It is, however, biased by how the different strains of the ubiquitous pathogen are selected for various sequencing projects.

Strains isolated from the sputum niche were the most abundant while the isolates from the lungs and dental niches were the least abundant. This statistics is consistent with the fact that P.aeruginosa is mainly associated with cystic fibrosis and most of the sequencing efforts have been biased towards strains isolated from the airways.

4.1.4 Python Scripts Sequence Retrieval

The custom python script targeted all the initial biofilm formation genes that were retrieved from the *Entrez* gene search. Out of the possible 51 biofilm formation genes associated with *P.aeruginosa* PAO1, the reference strain, the study successfully retrieved 13 biofilm formation genes which were common in most of the strains of the ubiquitous microorganism. These genes were identified and grouped into corresponding clusters of orthologous genes (COGs) represented by individual FASTA files as indicated in table 4.4. The custom python scripts successfully created 13 fasta files for the biofilm formation genes, each file containing 44 sequences of the respective genes selected from every *P. aeruginosa* reference genome. This represented 25.49% of the total number of biofilm formation gene sequences. In this case the amino sequences were retrieved from the GenBank files of different *P. aeruginosa* strains. Table 4.5 indicates the names, GenBank accession number, size and number of annotated genes of the 44 *P. aeruginosa* genomes.

Table 4.4 Classes of biofilm formation genes retrieved using the custom python scripts

Classes	No of Genes	Genes
Adhesins	1	*ppyR (psl)*
Repressors	1	*gshB*
Regulatory	5	*algC, algD, algU, arnB, rsaL*
Motility	3	*fliC, gshA, htpG*
Unclassified	3	*pslJ, pslE, pslG,*
Total	**13**	

Table 4.5 Information about the 44 Pseudomonas aeruginosa genomes

Scientific Names	GenBank Accession Number	Size (kbps)	Number of annotated genes	GC (%)	Ecological niche
P. a PAO1	NC_002516	6,264.404	5700	66.56	Unclassified
P. a strain 24Pae112	NZ_CP029605	7097.241	6596	65.99	
P. a strain 268	NZ_CP032761	7030.474	6604	65.91	
P. a strain B17932	NZ_CP034436	6744.658	5943	65.94	

P. a strain BA15561	NZ_CP033432	6793.961	5813	65.84	Blood
P.a strain NCTC 12903	NZ_LR134309	6839.985	6431	66.09	
P. a strain PA1207	NZ_CP022001	7411.863	6813	65.70	
P. a strain PA1242	NZ_CP022002	7050.510	6303	65.80	
P. a strain PABL012	NZ_CP031659	6546.467	6089	66.29	
P. a strain PABL017	NZ_CP031660	6503.460	6019	66.31	
P. a strain Pa58	NZ_CP021775	7241.575	6673	65.80	Bronchial
P. a strain Pa84	NZ_CP021999	6566.724	6058	66.23	
P. a strain Pa124	NZ_CP021774	7008.516	6479	65.84	
P. a strain Pa127	NZ_CP022000	7148.302	6565	65.74	
P. a strain GIMC5015:PAKB6	NZ_CP034429	6258.491	5772	66.53	
P. a strain H26023	NZ_CP033685	6729.216	6260	66.21	
P. a strain NCTC11445	NZ_LR134308	6766.292	6378	66.06	Clinical
P. a paerg002	NZ_LR130527	6451.470	5935	66.40	
P. a paerg003	NZ_LR130530	6433.962	5945	66.40	
P. a paerg004	NZ_LR130531	6452.809	5936	66.40	
P. a paerg005	NZ_LR130534	6931.425	6427	66.00	
P. a paerg009	NZ_LR130533	6941.287	6352	65.98	
P. a paerg010	NZ_LR130536	6433.960	5950	66.40	
P. a paerg011	NZ_LR130535	6434.133	5946	66.40	
P. a paerg012	NZ_LR130537	6434.020	5948	66.40	
P. a strain L10	NZ_CP019338	6661.962	6119	66.13	Environment
P. a strain PA34	NZ_CP032552	6810.079	6314	66.07	Eye
P. a C-NN2 isolate	NZ_LT883143	6902.967	6412	66.12	Lung
P. a strain H25883	NZ_CP033686	6706.800	6236	66.15	Wound
P. a strain H26027	NZ_CP033684	7079.598	6650	66.07	
P. a strain MRSN12280	NZ_CP028162	7070.928	6597	66.02	
P. a PAO1161	NZ_CP032126	6383.803	5918	66.42	
P. a strain NCTC13715	NZ_LR134330	6765.311	6288	66.12	Urine
P. a strain FDAARGOS_505	NZ_CP033832	7029.824	6520	65.87	Trachea
P. a strain AES1M	NZ_CP037925	6373.139	5848	66.48	Sputum
P. a strain AES1R	NZ_CP037926	6373.893	5833	66.48	
P. a strain CCUG 70744	NZ_CP023255	6859.232	6422	66.04	
P. strain LW	NZ_CP022478	6824.837	6271	65.97	
P. a strain PASGNDM345	NZ_CP020703	6893.164	6432	66.07	

P. a strain PASGNDM699	NZ_CP020704	6985.102	6545	66.00
P. a strain SP2230	NZ_CP034434	6976.603	6067	65.74
P. a strain SP4527	NZ_CP034409	7005.215	6123	65.79
P. a strain SP4528	NZ_CP033439	6877.287	6082	65.85
P. a strain Y31	NZ_CP030910	6831.076	6322	66.15

4.1.5 Evolutionary Analyses of Biofilm Formation Genes

From the comparison of the 13 COG-based ML phylogenetic trees, the study created a tree of relationships between different biofilm related genes using the treedist distance matrix (Figure 4.1). The phylogenetic analyses revealed four clusters. From the 13 biofilm formation genes analyzed, 10 genes fell into a single cluster. The *algD* and *algU* genes diverged the most from the other biofilm formation genes, while *fliC* was not completely divergent from the other genes which seem to have co-evolved together. While the study assumed that all the biofilm formation genes co-evolved together given that they belong to a group of functionally related genes that generally was confirmed by the obvious co-evolution of these genes – the divergence of the three genes may result from a horizontal gene transfer.

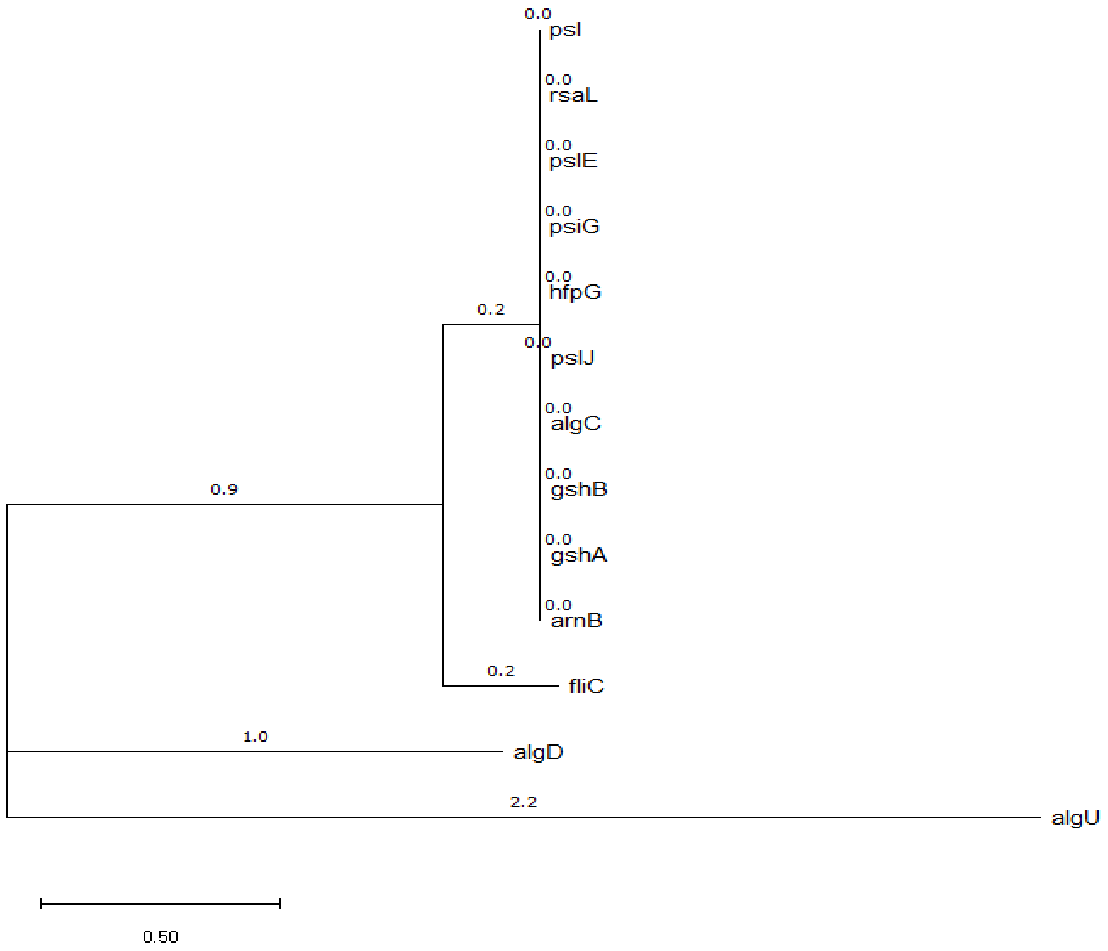


Figure 4.1 Co-evolution analyses of biofilm formation genes in strains of P. aeruginosa.

The NJ dendogram shows that the majority of gene COGs produced identical phylogenetic trees of the selected reference genomes that indicates a strong co-evolution of these genes. Exceptions were the genes fliC, algD and algU which may be exchanged by horizontal gene transfer or evolved faster than other genes of this functional group. The tree is drawn to scale with branch lengths in the same units as those of the evolutionary distances used to infer the COG phylogenetic tree.

4.2 Construction of the Profile Hidden Markov Models

4.2.1 Identification of Protein Family of Interest

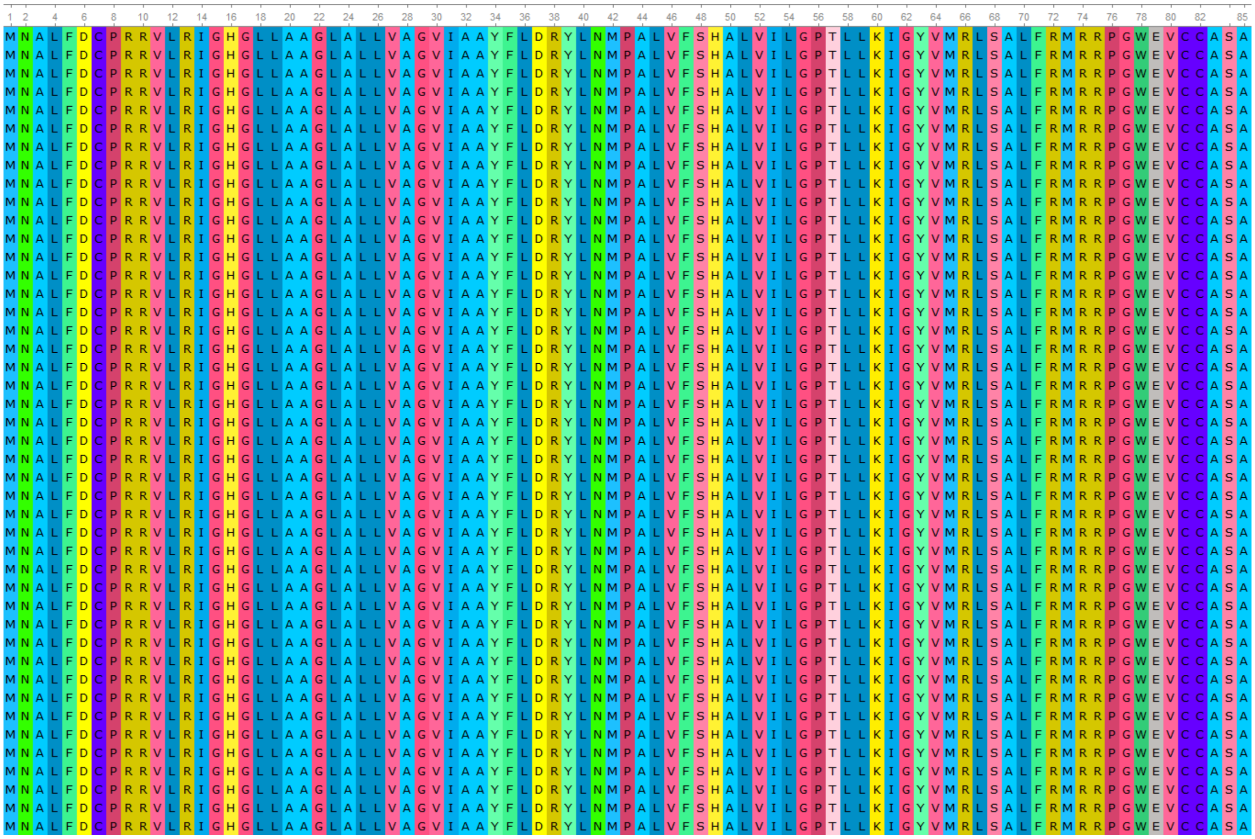
Homologous genes with highly similar functions are often classified as gene families. For this study, genes responsible for biofilm formation in different strains of *P. aeruginosa* were identified and selected as the protein family of interest. Using the criteria identified by Henikoff *et al.*, the study classified these set of genes into a family of related sequences (Henikoff *et al.*, 1997). These genes were then used to inform the construction and validation of the profile hidden Markov models.

4.2.2 Select Sequences Representative of this Family

Homologous sequences files obtained in the preceding sequence analyses were selected as the representative sequences for the biofilm formation genes protein family (clusters of orthologous genes). 12 sequence files created by the custom python script were selected for the downstream analyses as they contained sequences from different strains of the ubiquitous pathogen. Each of these sequences represented a single biofilm formation gene containing 44 records of *P. aeruginosa* sequences. Gong *et al.*, has previously reported that the choice of representative sequences of a protein family of interest inevitably affects the outcomes of downstream analyses performed on these set of sequences (Gong *et al.*, 2012). An extensive biological knowledge of the protein family under study is necessary for one to make an informed decision. The study relied on the previous analyses to make a judgment of the sequence homology. In his study, Rost mentioned that pair-wise sequence identity of long alignments that are less than 40% could result in ambiguous results in downstream analyses (Rost, 1999). Homologous sequence files were preferred in this case as they would contain patterns and motifs which could be identified by the pHMM and used to analyze different strains of *P.aeruginosa*. The amino acid sequences were also preferred given that they provide adequate information that can be modelled in a pHMM.

4.2.3 Building Multiple Sequence Alignment

The study sought to create multiple sequence alignments that would later be used to construct pHMMs. For each family of sequence, a multiple sequence alignment was created using the MUSCLE algorithm in UGENE. Figure 4.2 indicates a portion of the multiple sequence alignment file. Gaps are indicated by dashes without amino acid sequences while the matching regions of the alignment are indicated by continuous columns of similar amino acid bases. A total of 12 multiple sequence alignments were created by the study. The alignment length (indicating the consensus sequence length) and number of sequences of each the 12 sequence alignments are indicated in table 4.6.



```
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
MNALFDCPRRVLRIGHGLLAAGLALLVAGVIAAYFLDRYLNMPALVFSHALVILGPTLLKIGYVMRLSALFRMRRPGWEVCCASA
```

Figure 4.2 A portion of the multiple sequence alignment of the psl cluster of orthologous genes.

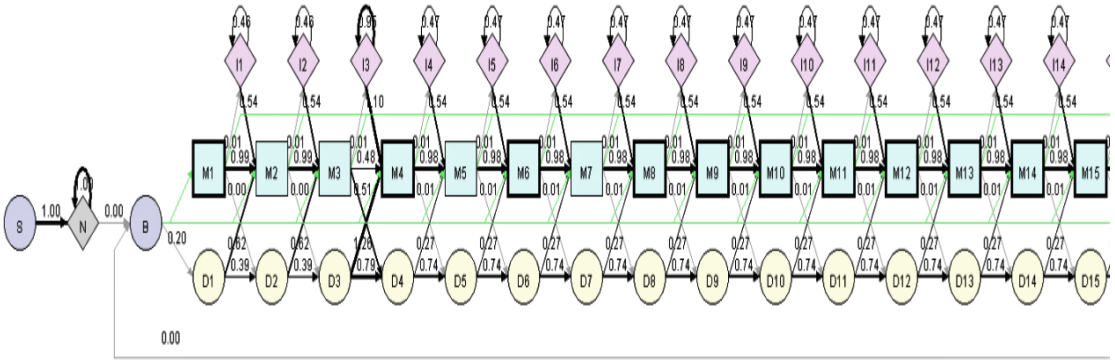
The lack of dashes in this alignment indicates the high levels of homology (similarity) in this set of genes.

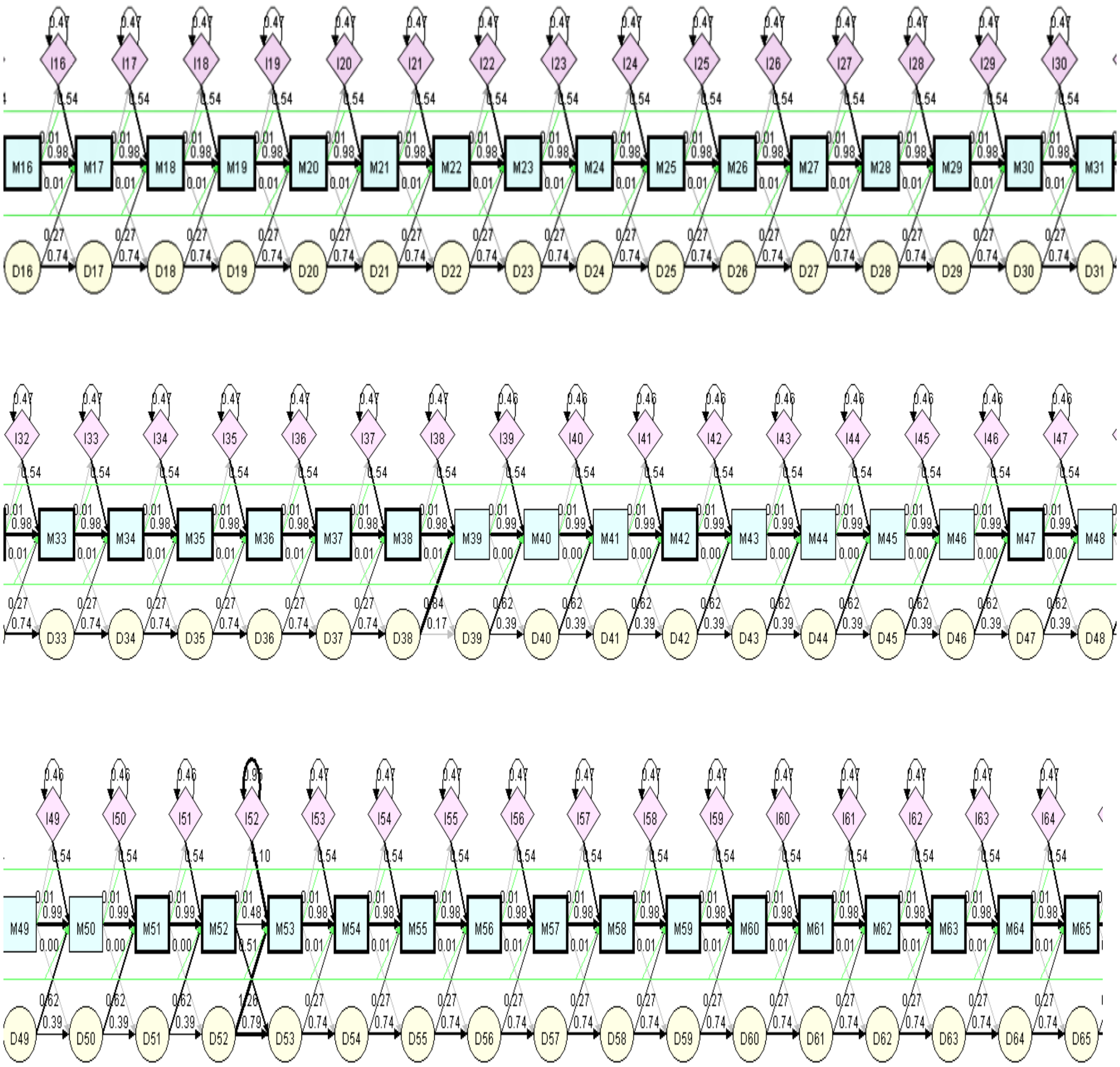
Table 4.6 Names of alignments, consensus sequences length and the total sequences in each alignment

Alignment name	Consensus sequence length	Number of sequences
algU	870	44
algD	872	44
arnB	382	44
fliC	489	44
gshA	427	44
gshB	317	44
htpG	649	44
Psl	85	44
pslE	662	44
pslG	442	44
pslJ	478	44
rsaL	80	44

4.2.4 Build Profile HMM

The study successfully constructed 12 pHMMs from the clusters of orthologous genes created using the python scripts. A representative section of one pHMM is shown in figure 4.3.




I16 I17 I18 I19 I20 I21 I22 I23 I24 I25 I26 I27 I28 I29 I30
M16 M17 M18 M19 M20 M21 M22 M23 M24 M25 M26 M27 M28 M29 M30 M31
D16 D17 D18 D19 D20 D21 D22 D23 D24 D25 D26 D27 D28 D29 D30 D31
0.47
0.54
0.01
0.98
0.01
0.27
0.74
I32 I33 I34 I35 I36 I37 I38 I39 I40 I41 I42 I43 I44 I45 I46 I47
M33 M34 M35 M36 M37 M38 M39 M40 M41 M42 M43 M44 M45 M46 M47 M48
D33 D34 D35 D36 D37 D38 D39 D40 D41 D42 D43 D44 D45 D46 D47 D48
0.46
0.99
0.00
0.84
0.17
0.62
0.39
I49 I50 I51 I52 I53 I54 I55 I56 I57 I58 I59 I60 I61 I62 I63 I64
M49 M50 M51 M52 M53 M54 M55 M56 M57 M58 M59 M60 M61 M62 M63 M64 M65
D49 D50 D51 D52 D53 D54 D55 D56 D57 D58 D59 D60 D61 D62 D63 D64 D65
0.95
0.10
0.48
0.51
1.26
0.79

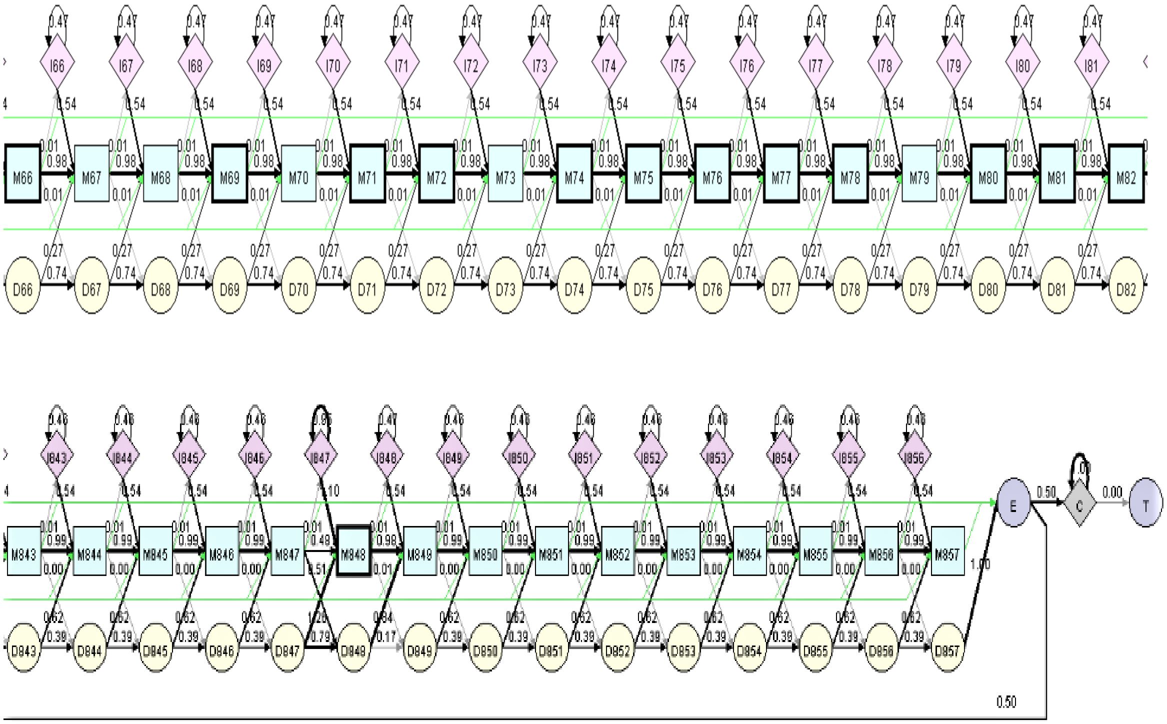


Figure 4.3 A representative profile HMM indicating the architecture of the constructed models.

The different shapes indicate states while the arrows indicate state transitions. The ‘S’ represents the start position of the model. The ‘N’ represents the null model that the HMMER algorithm constructed first before creating the rest of the representative model. The ‘M’ (squares) represents the match states which indicate the frequencies of the most probable amino acid in those different locations. The ‘D’ (circles) represents the delete states while the ‘I’ (diamonds) represents the insert states.

The profile HMM had three important states, the match state, delete state, and insert state as indicated on figure 4.3, page 52. The match state three transition probabilities i.e. 0.01 for the insert state, 0.00 for the delete state, and 0.99 for the next match state. The insert state had two transition probabilities 0.54 for the match state and 0.47 to remain on the insert state. The delete state, on the other hand, also had two transition probabilities 0.27 for the match state and 0.74 to remain on the delete state. These patterns were used to identify the biofilm formation genes in the sequences of *P. aeruginosa*.

Besides the visualization of the model architecture the HMM, HMMVE_1.2 was used to visualize the HMM logo which indicated the most likely amino acid for specific positions (Dai and Cheng, 2008). A representative HMM logo is described in figure 4.4. From figure 4.4, it is important to note that the most conspicuous amino acid in the first position is methionine which is expected to be the first amino acid in a gene sequence. Such findings highlight the accuracy of the constructed models. The models were built to optimally represent the common motifs and patterns from the multiple sequence alignment of the biofilm formation genes. A clear representation of these common patterns would help the study to clearly point out both the conserve and variable regions within the sequences of *P. aeruginosa*. The pHMM was chosen for this study as it is useful in creating specific architectures suitable for modeling sequence profiles.

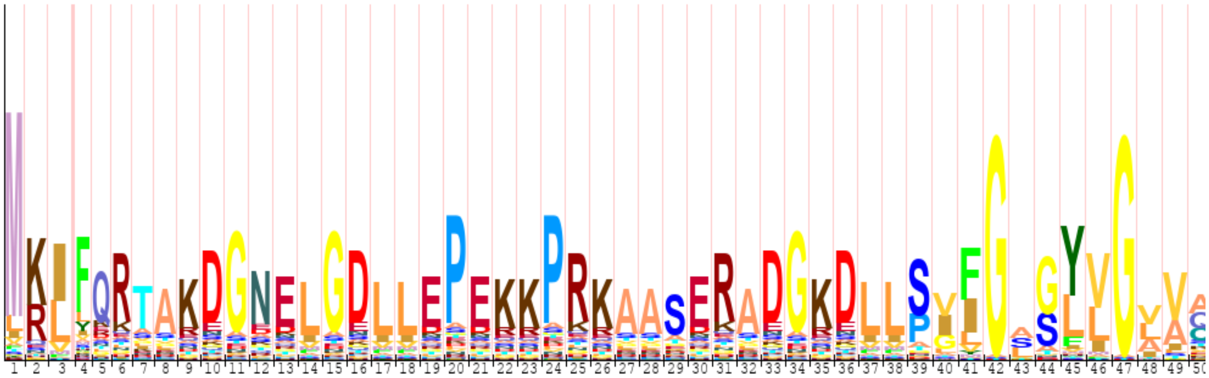


Figure 4.4 Frequencies of the different amino acid sequences in their relative positions.

The most conspicuous amino acid represents the most frequent amino acid in that specific position. Such amino acids can easily be used to identify the consensus sequences for our alignments.

4.2.5 Validation of the Profile HMM

The ability of the developed profile HMM to detect biofilm formation genes was analyzed using the positive and negative controls listed in Table 3.1. The reference *P. aeruginosa* PA01 was chosen as the positive control. All the constructed pHMMs correctly identified different biofilm formation genes in the positive control. The negative controls were chosen because they are of different species and do not exhibit biofilm formation as one of their survival mechanisms. When searched against the four negative controls, the 12 pHMMs showed no identification of the biofilm formation genes as was expected.

4.3 *Pseudomonas aeruginosa* Sequence Analyses

The search performed by the developed profile HMM against the *P. aeruginosa* sequences identified a total of 197 hits for the 13 different ecological niches as indicated in table 4.7, page 56. The hits identified represent the total number of biofilm formation genes identified within the genome of various strains of the ubiquitous pathogen. Of the 197 hits, 144 hits (73%) belonged to the human samples while 53 hits (27%) belonged to the nonhuman samples. 38% of the human sample hits were recorded from ecological niches that were respiratory in nature. 62% of the hits were associated with non-respiratory niches within the human host. 22.34% of the biofilm formation gene sequences identified by the profile HMMs were identified in the blood ecological niche. The lung and dental ecological niches, on the other hand, indicated the least number of biofilm formation genes, 1 hit each, representing 0.51% of the identified genes (Table 4.8, page 56). The *algD* gene was most commonly found, 44 hits (22.34%) in the different strains of *P. aeruginosa* sequences, followed by the *rsaL* gene, 31 hits (15.74%). The *gshB* gene sequences were the least abundant sequences, 2 hits (1.02%) (Table 4.9, page 57).

Table 4.7 Comparison of pHMM hits of the genes across the 13 ecological niches

Niche	algD	algU	pslJ	arnB	gshB	htpG	psl	pslE	PslG	rsaL	gshA	fliC
Abscess	2	0	2	0	0	0	0	2	2	2	1	1
Blood	10	5	2	1	0	9	4	3	4	3	3	0
Bronchial	4	0	0	0	0	0	1	0	0	2	3	0
cell culture	2	1	1	1	0	1	2	1	1	3	0	1
Clinical	7	1	0	1	0	1	3	0	0	4	2	0
Dental	1	0	0	0	0	0	0	0	0	0	0	0
Environment	7	2	1	0	1	2	1	1	1	1	2	1
Eye	1	1	0	0	1	0	1	0	0	0	0	0
Lung	0	0	0	0	0	0	0	0	0	0	1	0
Sputum	5	3	1	1	0	0	6	1	1	9	8	1
Trachea	1	0	0	0	0	0	2	0	0	1	4	0
Urine	1	0	0	0	0	0	6	0	0	4	1	0
Wound	3	0	1	0	0	0	3	1	1	2	5	0

Table 4.8 Frequency (%) of the pHMM hits across the different ecological niches

Niche	No. of sequences	Frequency (%)	Type 1	Type 2
Blood	44	22.34	Human	non respiratory
Sputum	36	18.27	Human	Respiratory
Environment	20	10.15	non-human	
Clinical	19	9.64	non-human	
Wound	16	8.12	Human	non respiratory
Cell culture	14	7.11	non-human	
Abscess	12	6.09	Human	non respiratory
Urine	12	6.09	Human	non respiratory
Bronchial	10	5.08	Human	Respiratory
Trachea	8	4.06	Human	Respiratory
Eye	4	2.03	Human	non respiratory
Dental	1	0.51	Human	non respiratory
Lung	1	0.51	Human	Respiratory

Table 4.9 Frequency of the identified biofilm formation genes

Biofilm formation genes	No. of Sequences	Frequency (%)
algD	44	22.34
rsaL	31	15.74
gshA	30	15.23
pslJ	29	14.72
algU	13	6.60
htpG	13	6.60
pslG	10	5.08
pslE	9	4.57
pslJ	8	4.06
arnB	4	2.03
fliC	4	2.03
gshB	2	1.02

The study sought to put these results into context and identified the density of the hits which represented the hit per Megabases as is indicated in table 4.10 (page 58)**.** Figure 4.5 (page 59) indicates the distribution of the density of hits per ecological niche. In this case, the abscess ecological niche had the highest density of hits while the lung niche had the lowest density of hits. Figure 4.6 (page 59) compares the density of hits between the human and non-human samples.

Table 4.10 Distribution of the density of pHMM hits (hits/MB) across the ecological niches

Niche	algDM	algUM	pslJM	arnBM	gshBM	htpGM	pslM	pslEM	pslGM	rsaLM	gshAM	fliCM
Abscess	0.000152	0	1.52E-04	0	0	0	0	1.52E-04	1.52E-04	1.52E-04	7.60E-05	7.60E-05
Blood	9.69E-05	4.85E-05	1.94E-05	9.69E-06	0	8.72E-05	3.88E-05	2.91E-05	3.88E-05	2.91E-05	2.91E-05	0
Bronchial	9.77E-05	0	0	0	0	0	2.44E-05	0	0	4.88E-05	7.33E-05	0
cell culture	7.95E-05	3.98E-05	3.98E-05	3.98E-05	0	3.98E-05	7.95E-05	3.98E-05	3.98E-05	1.19E-04	0.00E+00	3.98E-05
Clinical	1.06E-04	1.52E-05	0.00E+00	1.52E-05	0	1.52E-05	4.56E-05	0.00E+00	0.00E+00	6.09E-05	3.04E-05	0
Dental	1.44E-04	0.00E+00	0.00E+00	0.00E+00	0	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0
Environment	1.32E-04	3.79E-05	1.89E-05	0.00E+00	1.89E-05	3.79E-05	1.89E-05	1.89E-05	1.89E-05	1.89E-05	3.79E-05	1.89E-05
Eye	7.29E-05	7.29E-05	0.00E+00	0.00E+00	7.29E-05	0.00E+00	7.29E-05	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0
Lung	0	0.00E+00	0.00E+00	0.00E+00	0	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.45E-05	0
Sputum	2.85E-05	1.71E-05	5.69E-06	5.69E-06	0	0.00E+00	3.42E-05	5.69E-06	5.69E-06	5.12E-05	4.56E-05	5.69E-06
Trachea	2.88E-05	0.00E+00	0.00E+00	0.00E+00	0	0.00E+00	5.75E-05	0.00E+00	0.00E+00	2.88E-05	1.15E-05	0
Urine	2.13E-05	0.00E+00	0.00E+00	0.00E+00	0	0.00E+00	0.00E+00	0.00E+00	0.00E+00	8.51E-05	2.13E-05	0
Wound	4.85E-05	0.00E+00	1.62E-05	0.00E+00	0	0.00E+00	4.85E-05	1.62E-05	1.62E-05	3.24E-05	8.09E-05	0

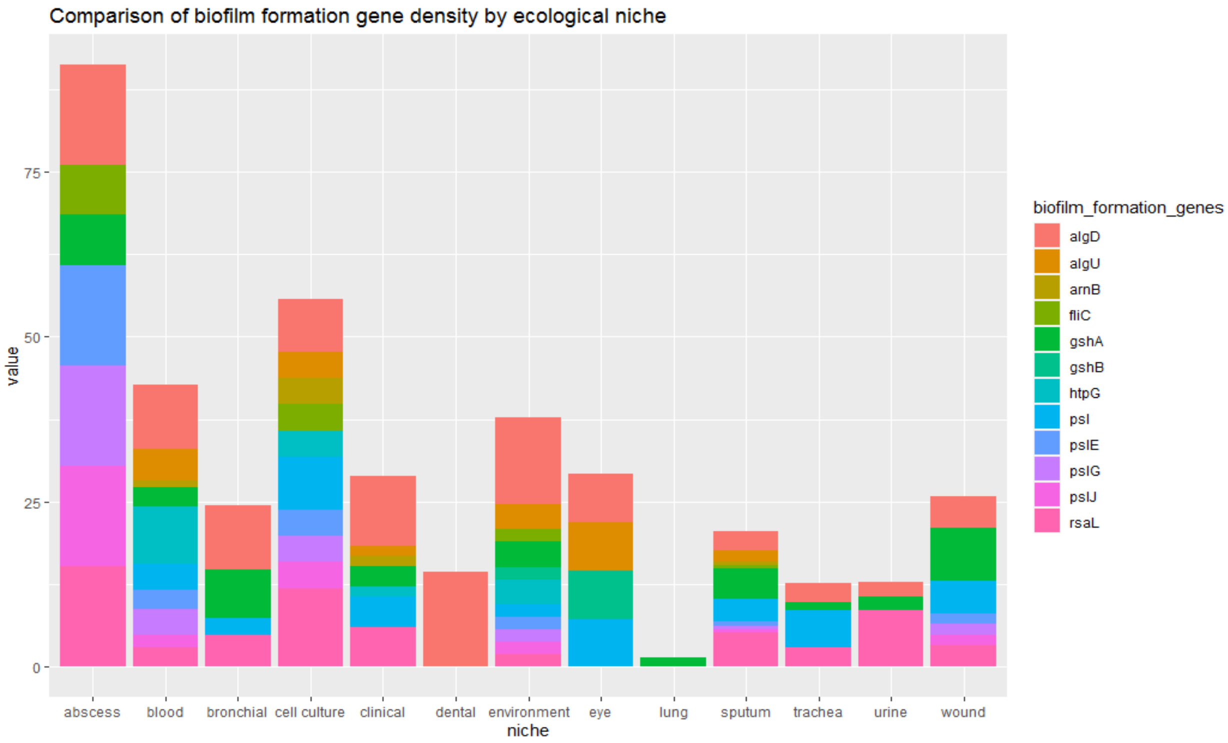


Figure 4.5 Comparison of biofilm formation genes by ecological niche reported by the pHMMs

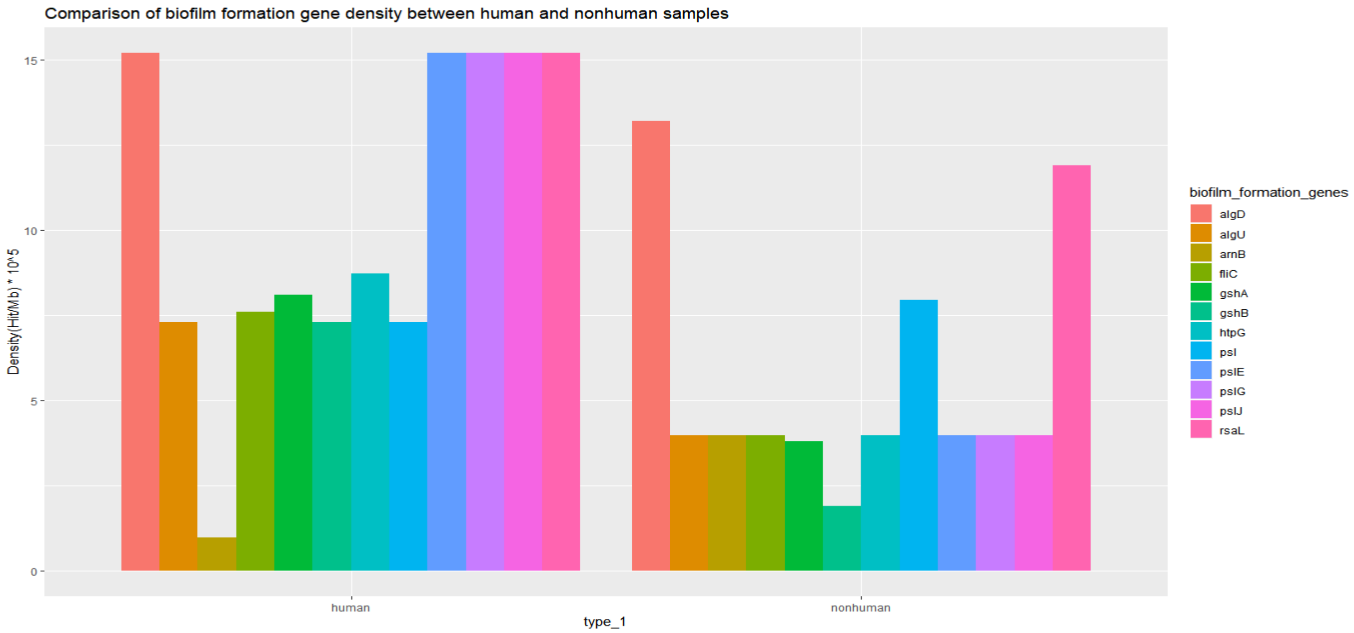


Figure 4.6 Comparison of biofilm formation gene hits between human and non-human strains

Strains of the ubiquitous pathogen from human samples had a higher density of hits compared to the strains from non-human samples. The only different observation was indicated in the density of hits for the *arnB* gene where the nonhuman samples had a significantly higher density of hits compared to their human sample counterparts. The overall result also indicated a significant variation in density of hits between the different sites within the human metagenomes. This pattern was also reflected in four of the respiratory subsites, namely bronchial, lung, sputum and trachea. The lung metagenomes had the lowest biofilm formation gene density and exhibited significantly lower densities than all the other ecological niches.

With regards to the biofilm formation genes, the *algD* gene had the highest number of hits and highest density of hits compared to the models of the other biofilm formation genes as indicated in both figures 4.5 and 4.6 (page 59). The Wilcoxon rank test indicated that the density of the *htpG* pHMM hits was greater for human samples than for nonhuman samples, W=3, p = 0.01759.

Table 4.11 Wilcoxon rank test results comparing the hits in human and non-human strains

Analyzed strains n = 97				
Gene	Human strains n =75	Non-human strains n= 22		
	Mean rank	Mean rank	Wilcoxon-test value	p-value
algD	0.05	0.11	8	0.287
algU	0.02	0.03	7	0.168
pslJ	0.01	0.01	10	0.408
arnB	0.01	0.01	9	0.079
gshB	0.002	0.01	12	0.501
htpG	0.02	0.03	3	**0.018***
Psl	0.05	0.04	9	0.360
pslE	0.01	0.01	10	0.408
pslG	0.02	0.01	10	0.408
rsaL	0.05	0.06	10	0.444
gshA	0.05	0.03	17	0.799
fliC	0.002	0.02	8	0.180

4.4 Genomic Analyses

4.4.1 Evolutionary Analyses

The study sought to determine the evolutionary relationship of the different strains of *P. aeruginosa*. 13 phylogenetic trees were constructed in the MEGA X platform to elucidate this relationship. Out of the 13 set of sequences, 3 sequences files did not indicate parsimonious sites and therefore couldn’t be inferred by the Maximum parsimony method. The maximum likelihood method, along with the JTT matrix-based model, was used instead to infer the evolutionary history of these set of sequences (Felsenstein, 1985). The initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The trees were drawn to scale, with branch lengths measured in the number of substitutions per site. Similar to the maximum parsimony method, the bootstrap consensus tree inferred from 100 replicates was taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. All these analyses were conducted in MEGA X (Kumar *et al.*, 2018).

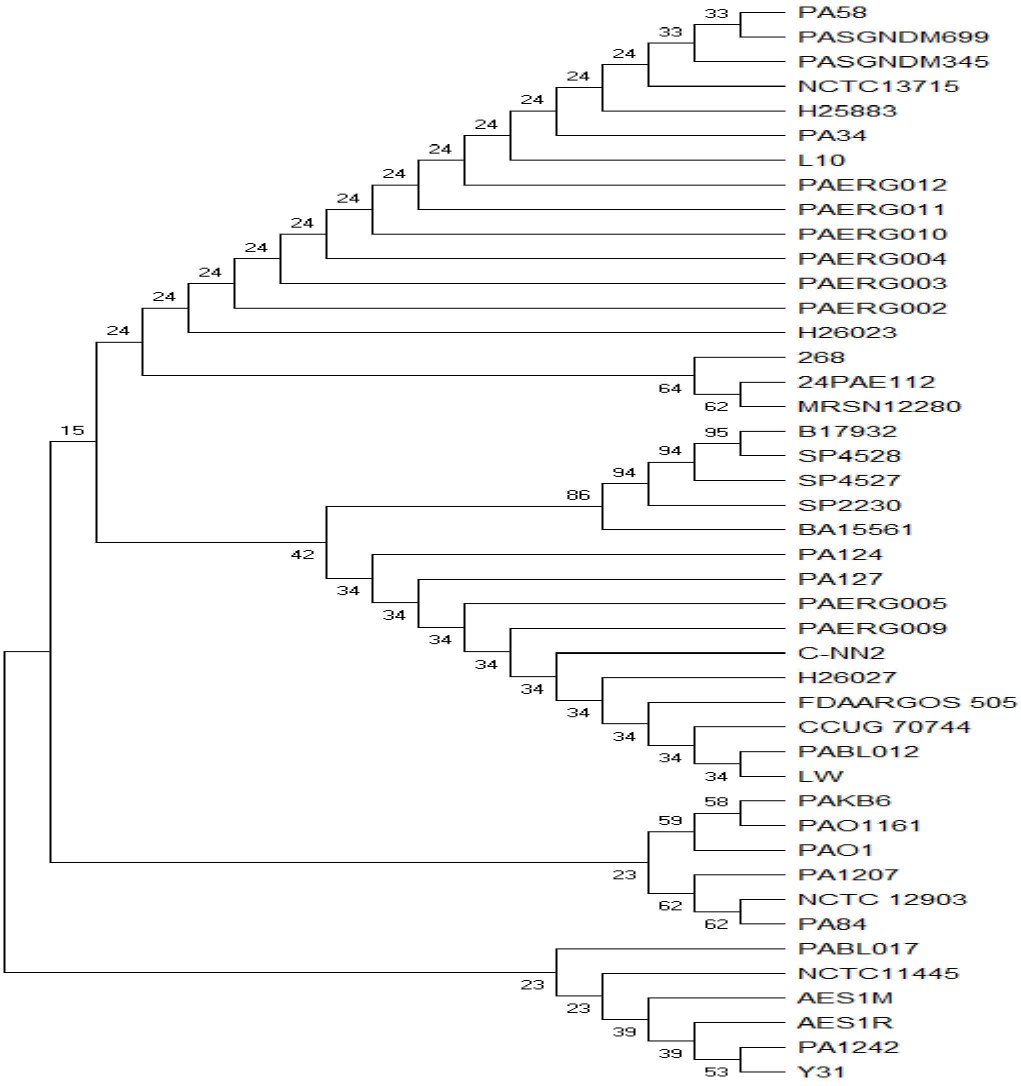


Figure 4.7 Evolutionary history of arnB gene sequences

The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. There were a total of 382 positions in the final dataset.

Figure 4.7 represented the arnB phylogenetic tree identified two clades. In this case five genome sequences clustered separately from the other gene sequence. PABL017, PA1242 (blood), NCTC 1145 (clinical isolates), AES1M, AES1R and Y31 (sputum isolates), were clustered in group 2.

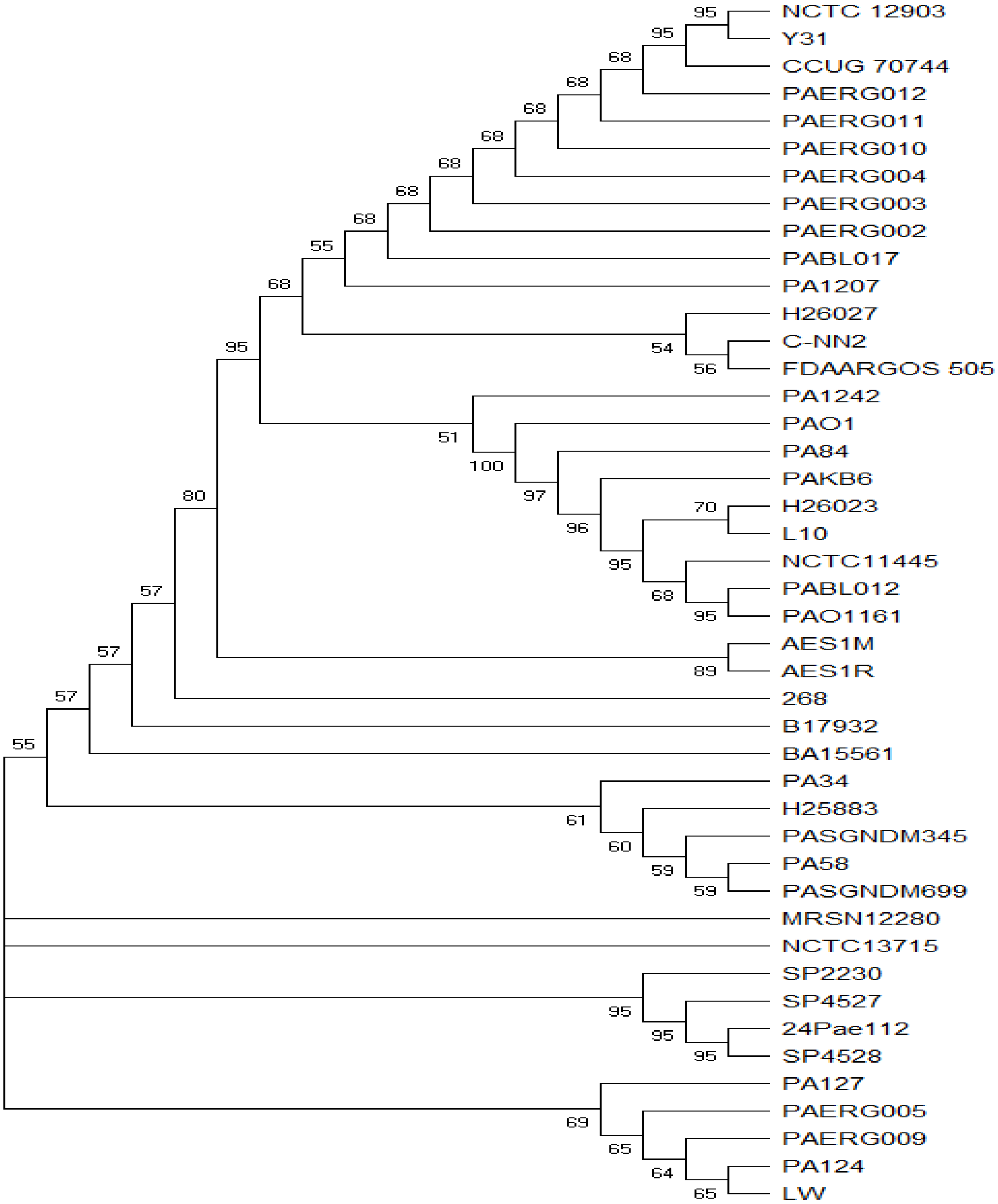


Figure 4.8 Evolutionary history of fliC gene sequences.

The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. There were a total of 489 positions in the final dataset.

P. aeruginosa strain SP4527 clustered differently from the rest of the other strains in figure 4.8 phylogenetic tree that indicated evolution of *fliC*.

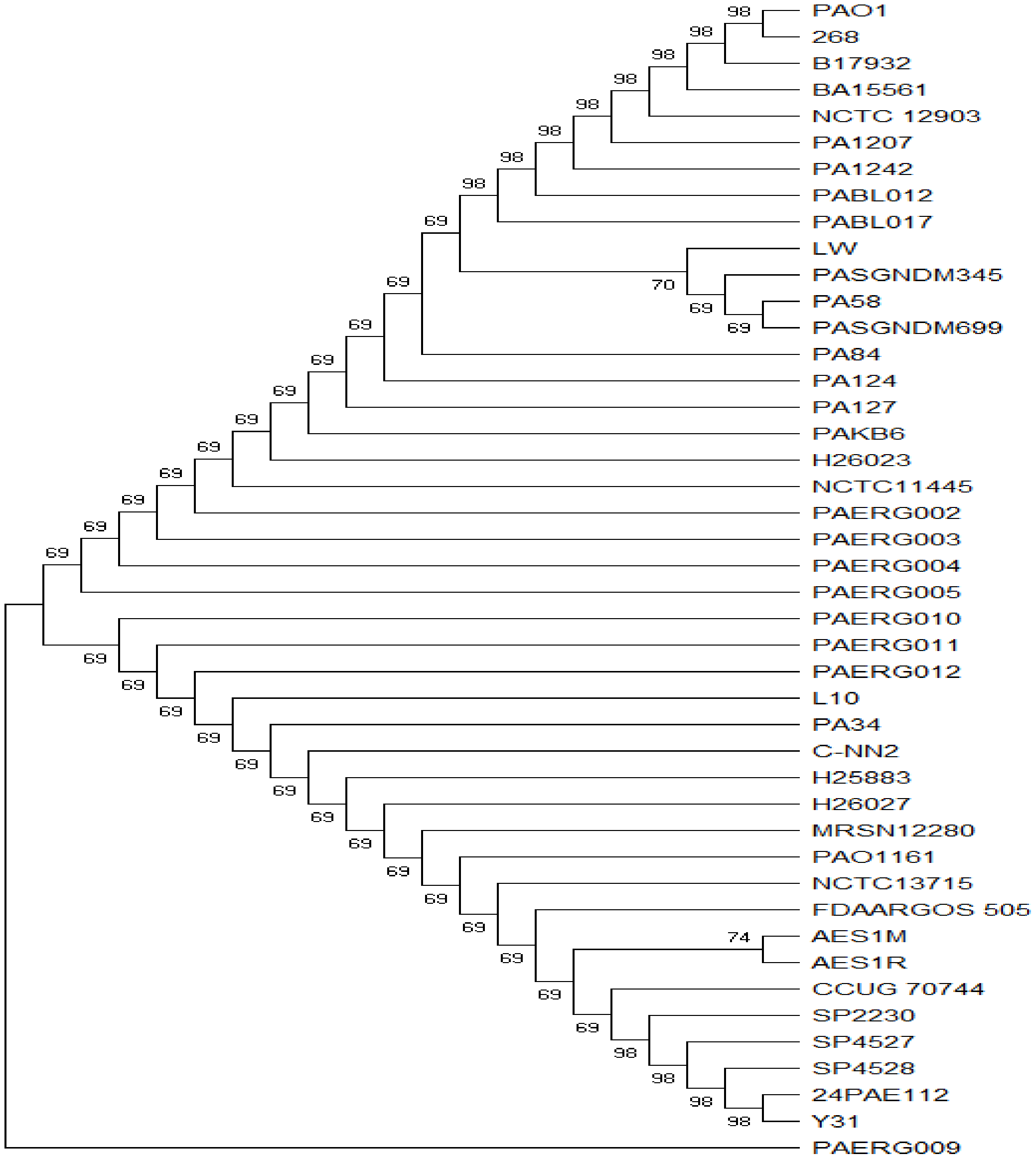


Figure 4.9 Evolutionary history of gshB gene sequences.

The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. There were a total of 317 positions in the final dataset.

PAERG009 (clinical isolate) clustered separately from the other sequences in figure 4.9 which identified two separate clades in the evolution of *gshB.*

The pslJ based tree identified two clades with the PABL012 strain clustering separately from the other sequence (figure 4.20). This strain had been classified under sequences obtained from blood samples. The other sequences sourced from the blood samples clustered close to each other.

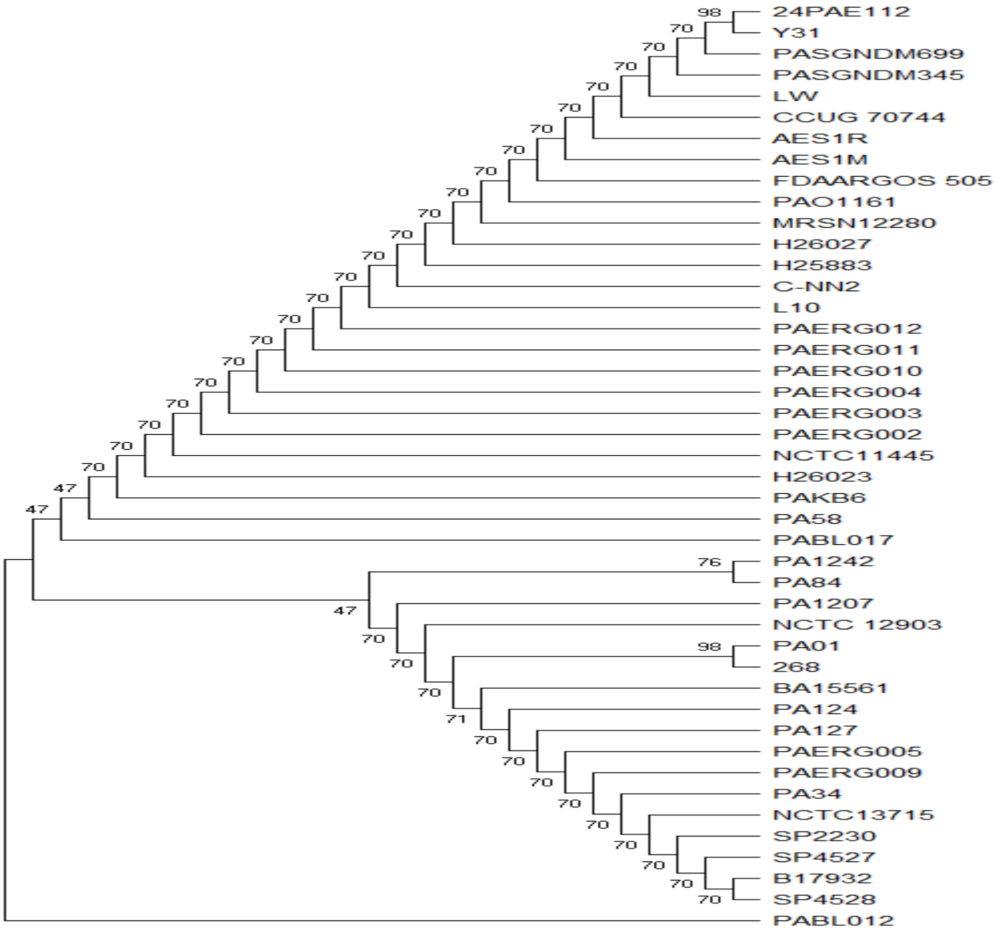


Figure 4.10 Evolutionary history of pslJ gene sequences

The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. This analysis involved 44 amino acid sequences. There were a total of 478 positions in the final dataset.

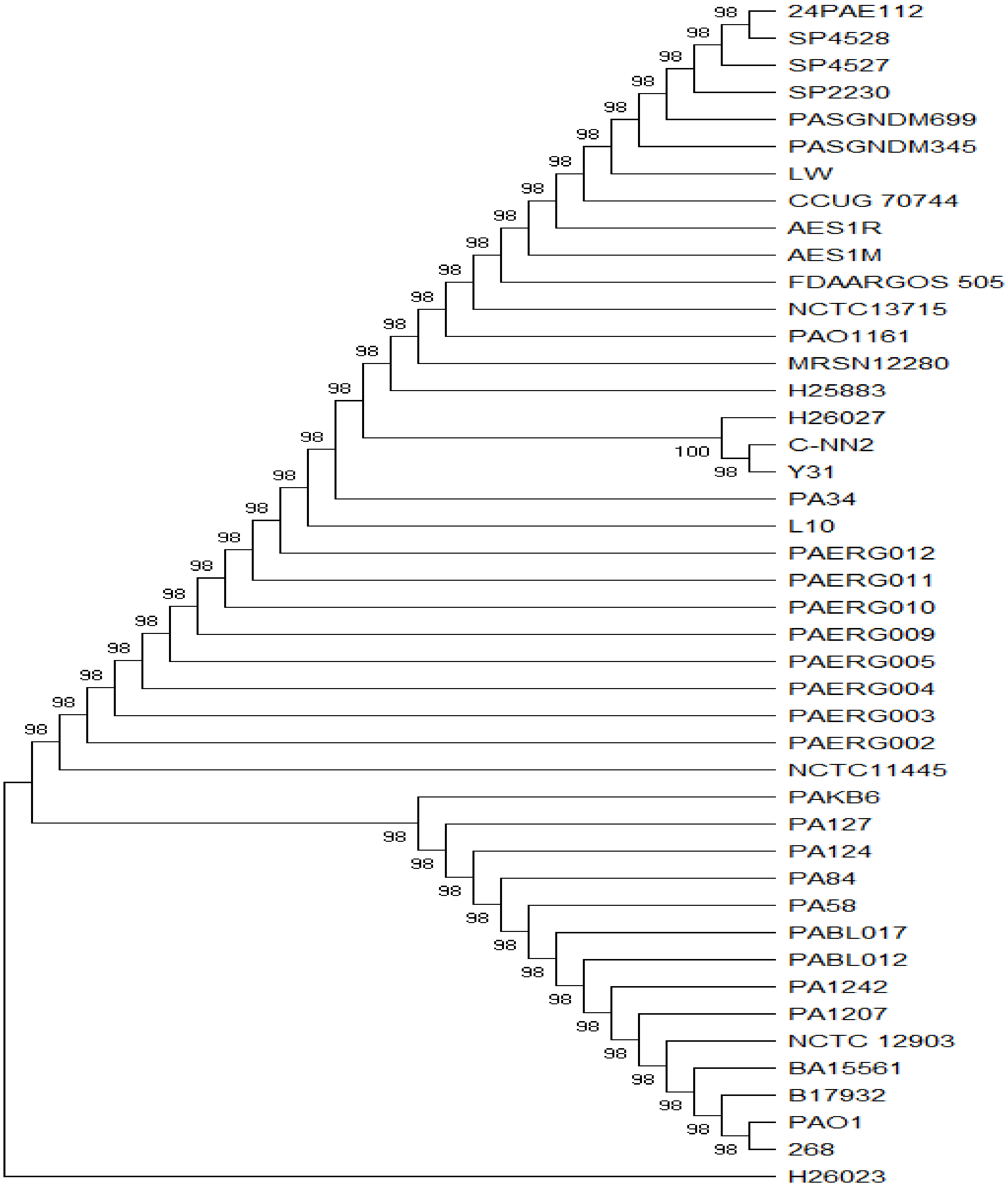


Figure 4.11 Evolutionary history of htpG gene sequences

The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. This analysis involved 44 amino acid sequences. There were a total of 649 positions in the final dataset.

The htpG tree identified only two clades with the H26023 clustering separately from the other *P.aeruginosa* sequences. This gene sequence was obtained from the bronchial isolates (figure 4.11).

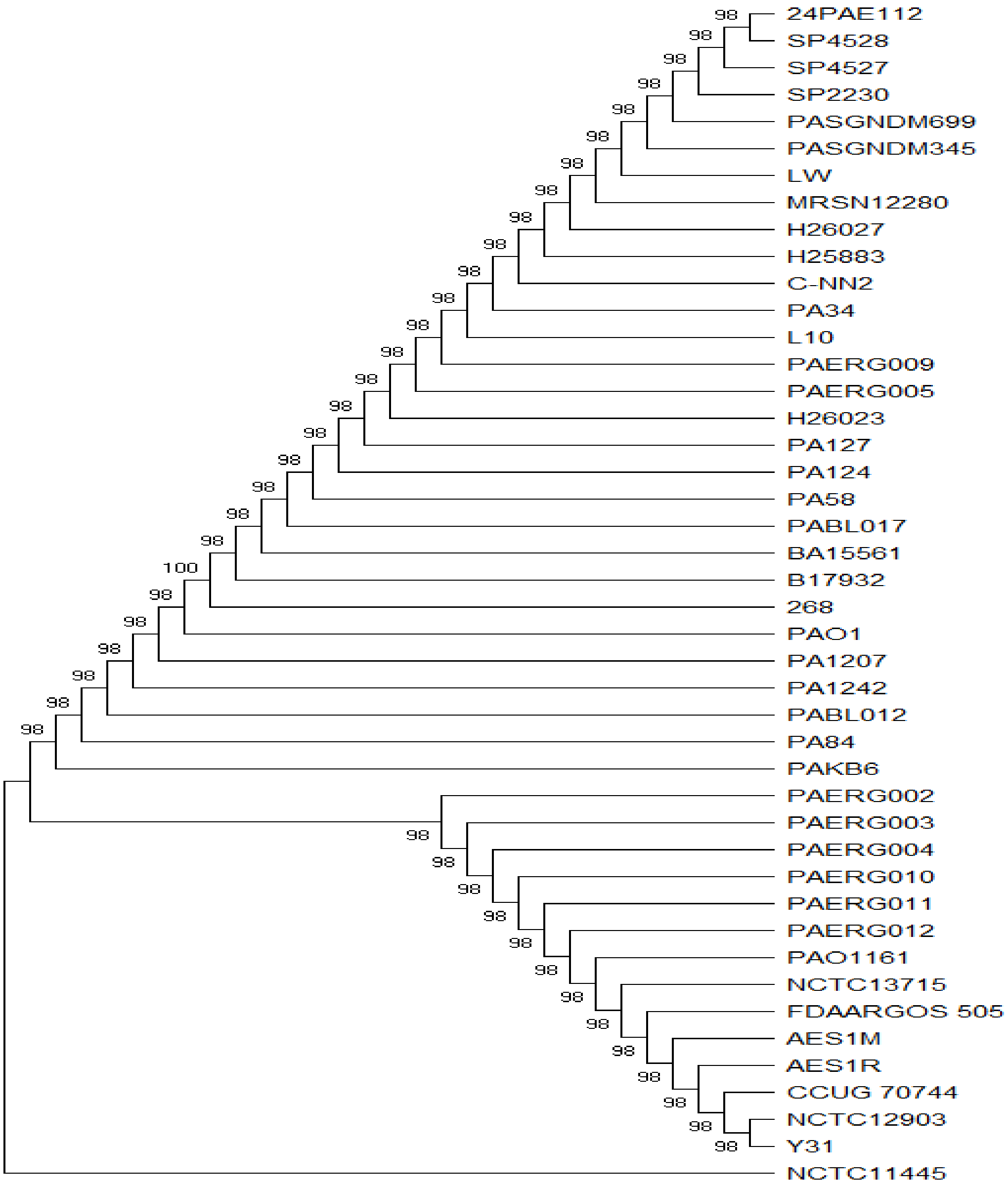


Figure 4.12 Evolutionary history of pslE gene sequences

The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. This analysis involved 44 amino acid sequences. There were a total of 662 positions in the final dataset.

The pslE tree also identified two clades with the NCTC11445 (from blood) clustering separately from the other gene sequences. This was similar to what was observed in the pslG tree where two clades were identified and NCTC11445 clustered separately (figure 4.12).

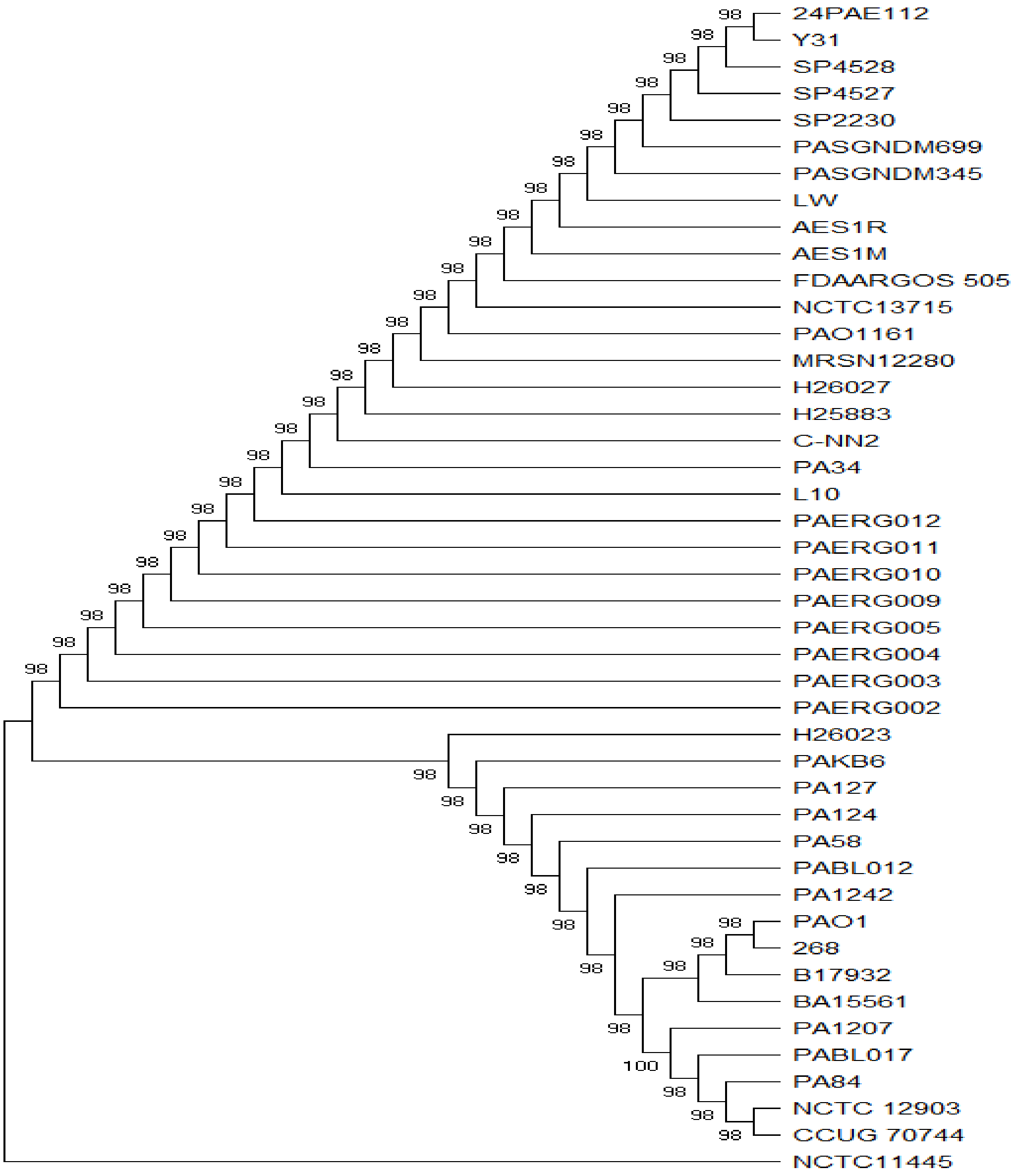


Figure 4.13 Evolutionary history of pslG gene sequences

The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. This analysis involved 44 amino acid sequences. There were a total of 442 positions in the final dataset. Evolutionary analyses were conducted in MEGA X.

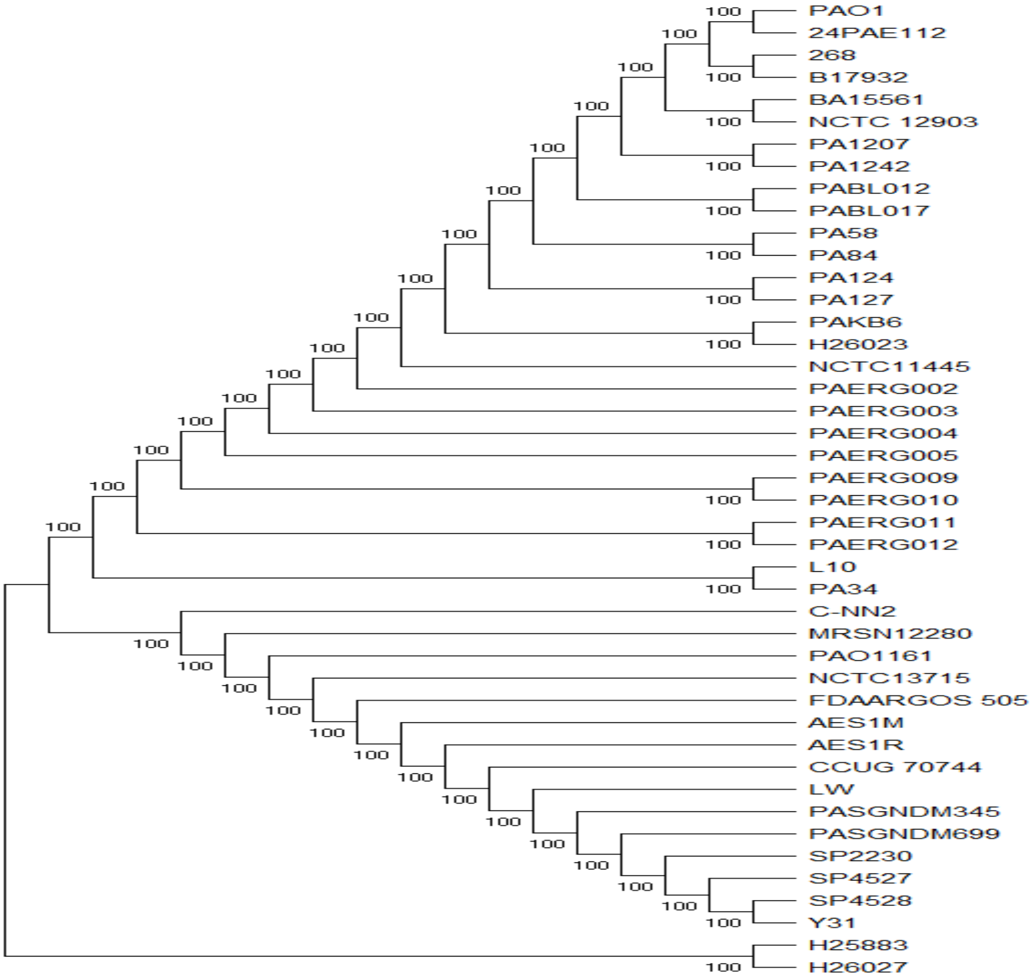


Figure 4.14 Evolutionary history of psl gene sequences

The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. This analysis involved 44 amino acid sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 85 positions in the final dataset.

The rsaL gene tree which identified 2 clades and clustered H25883 and H26027 separately from the other gene sequences. These two sequences had been retrieved from wound isolates (figure 4.13). The psl tree identified two clades with the H25883 and H26027 clustering separately from the other gene sequences (figure 4.14). It is important to note that the final three phylogenetic trees were constructed by the neighbor joining algorithm. Attempts to construct the tree with the maximum parsimony algorithm were futile as no parsimonious sites were present in these set of sequences.

4.4.2 Biofilm Formation Gene Distribution

The study aimed to find the distribution of biofilm formation genes in the genome sequences of various strains of *P. aeruginosa*. The study used the PAO1 genome as the reference genome for this analyses. Different colors on the rings indicated significant matches while non-significant matches were represented by blanks. A BRIG analysis was conducted for each ecological niche and 14 BRIG images were produced during this analysis (Figure 4.15 to figure 4.28). Figure 4.28 indicated the distribution of the biofilm formation genes among the sequences of the strains isolated from the 13 ecological niches that were identified by this study. Figure 4.16 to figure 4.28 indicated the distribution of the genes among sequences of strains isolated from each individual ecological niche. Besides the distribution of biofilm formation genes, the BRIG analysis was also used to identify the conserved and variable regions of the genome of the pathogen. This was done to further elucidate the differences in the distribution of genes in strains from different ecological niches. The key on each of the gene maps indicates the individual genomes in each of the concentric rings.

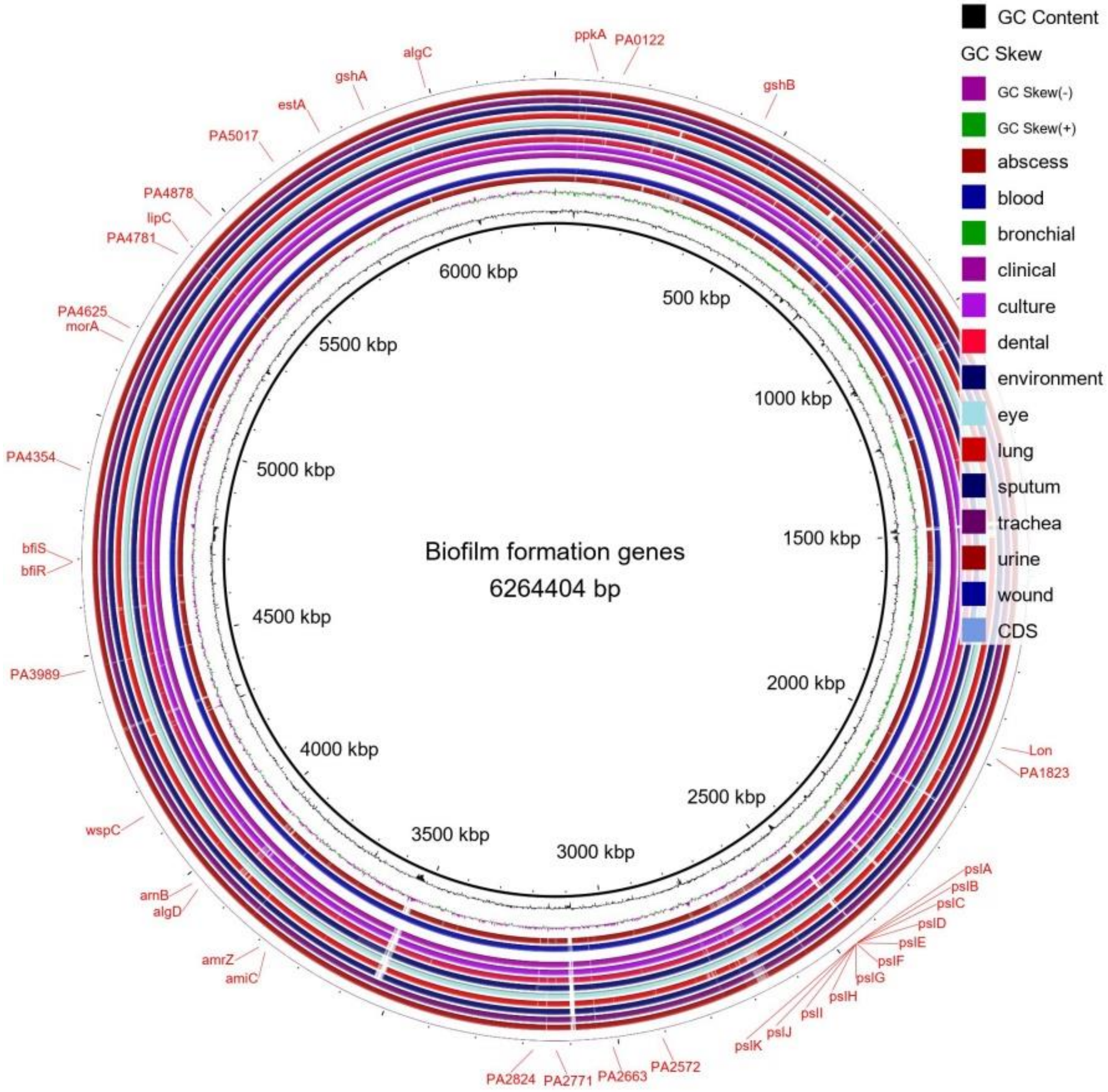


Figure 4.15 Visualization of genome comparison of the different strains of P. aeruginosa.

The BRIG analysis also included the distribution of biofilm formation genes within the genomes of different strains. Differences regarding the flexible genome can be seen. Different colors on the *rings indicated significant matches while non-significant matches were represented by blanks.*

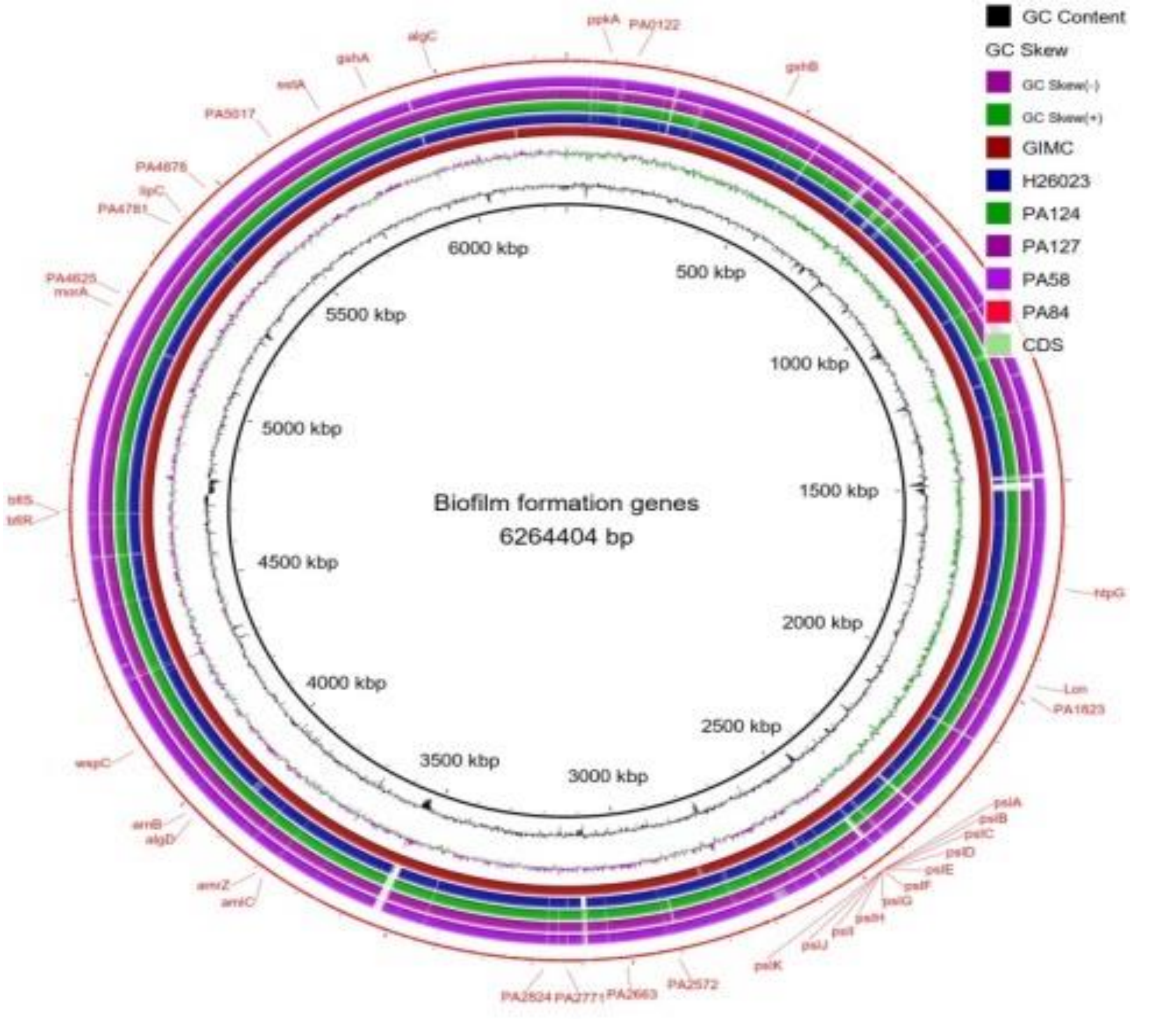


Figure 4.16 BRIG analysis for bronchial sequences.

Genes in the variable regions include PA4878 and lipC at locus 5400kbps, bfiS and bfiR at locus 4600kbps, amiC at locus 3800kbps and the pclass at locus 2400kbps

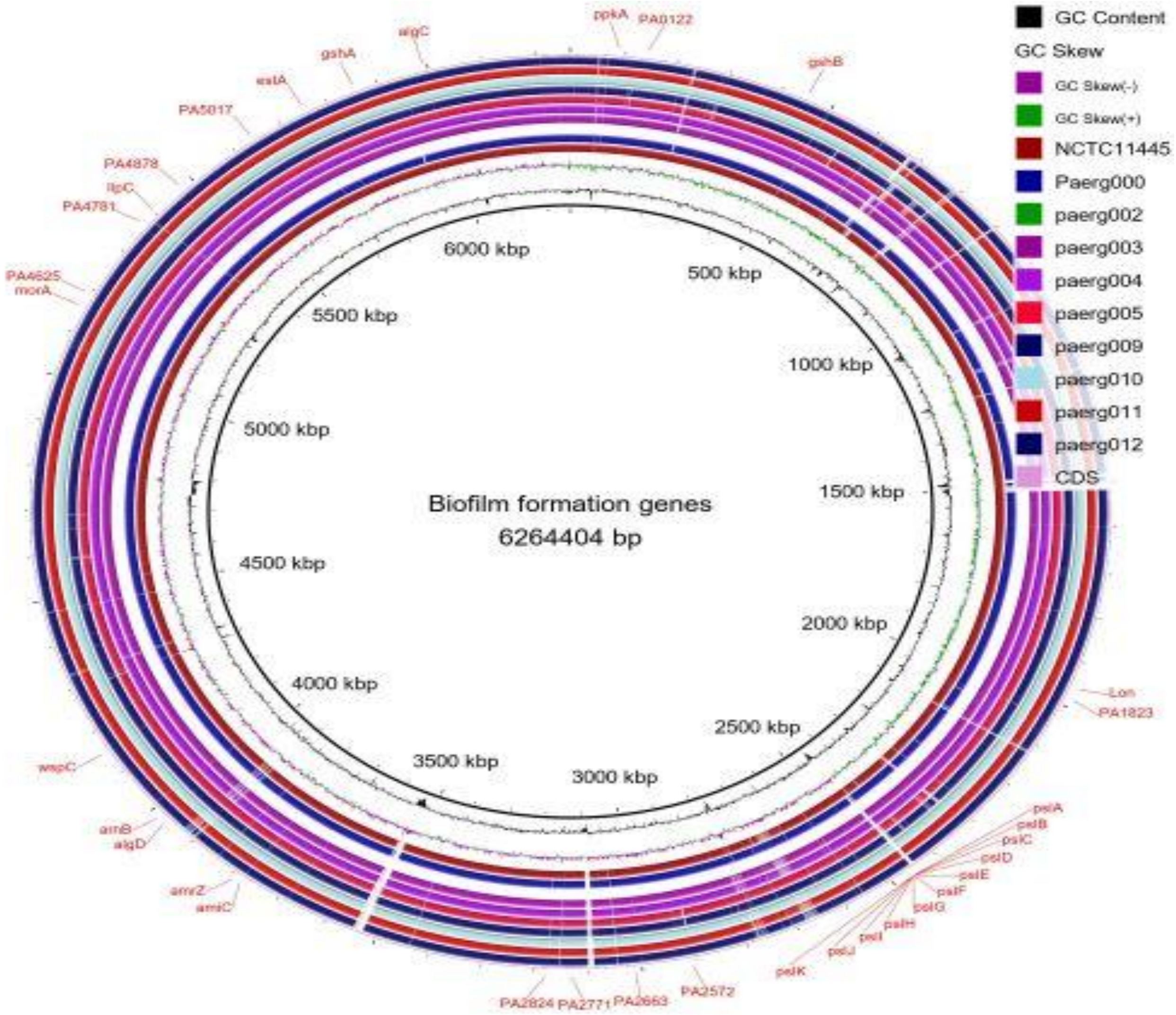


Figure 4.17 BRIG analysis for clinical sequences.

Genes in the variable regions included ppkA at locus 100kbp, p class at locus 2450kbps, amiC at locus 3750kbps, arnB and arnD at locus 3950kbps, wspC at locus 4100kbps, pA4625 at locus 5150kbos and algC at locus 6000kbps.

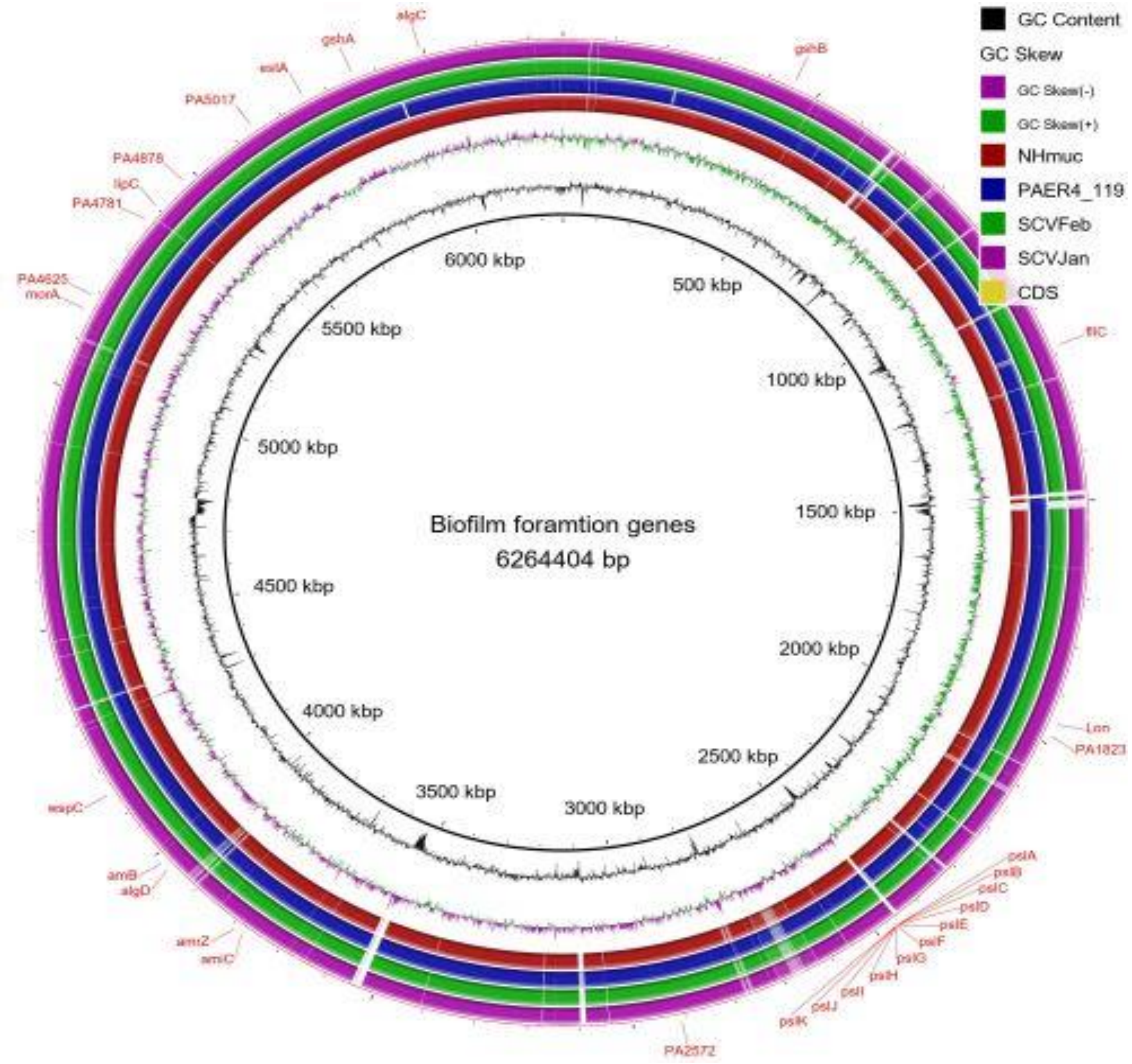


Figure 4.18 BRIG analysis for cell culture sequences.

Genes in the variable regions included the p class at locus 2450kbps, and the lipC and PA4781 at locus 5400kbps

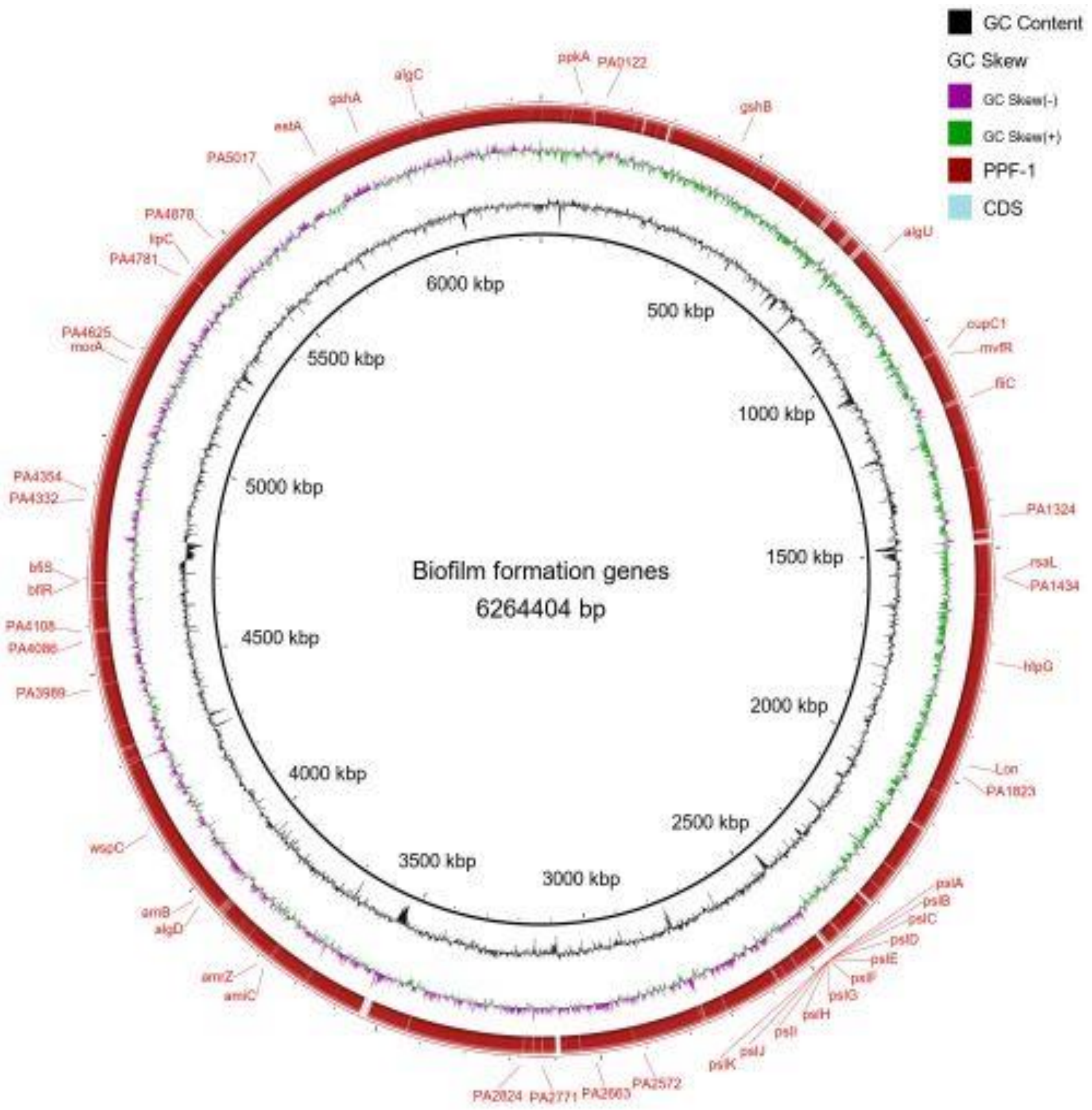


Figure 4.19 BRIG analysis for dental sequences.
Genes in the variable regions included cupC1at locus 1050kbps, fliC at locus 1200kbps, p class at locus 2450kbps, PA2572 at locus 3000kbps, PA2771 at locus 3150kbps, amiC at locus 3700kbps, PA3989 at locus 4500kbps, PA4108 at locus 4600kbps and bfiS and bfiR at locus 4700kbps.

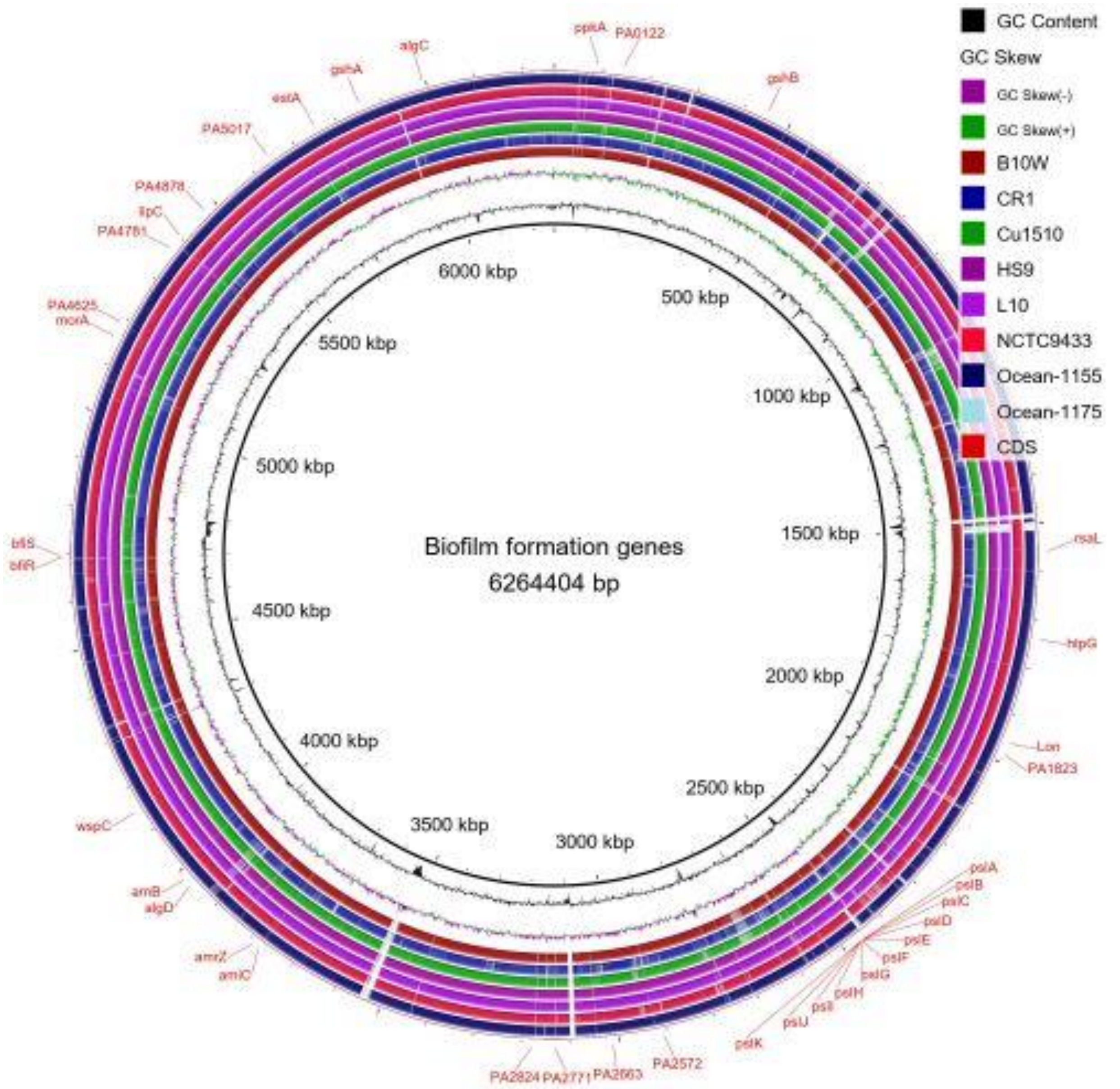


Figure 4.20 BRIG analysis for environmental sequences.

Genes in the variable regions included the p class at locus 2450kbps, PA2824 at locus 3150kbps, amiC at locus 3750kbps, morA at locus 5150kbps and PA4625 at locus 5200kbps.

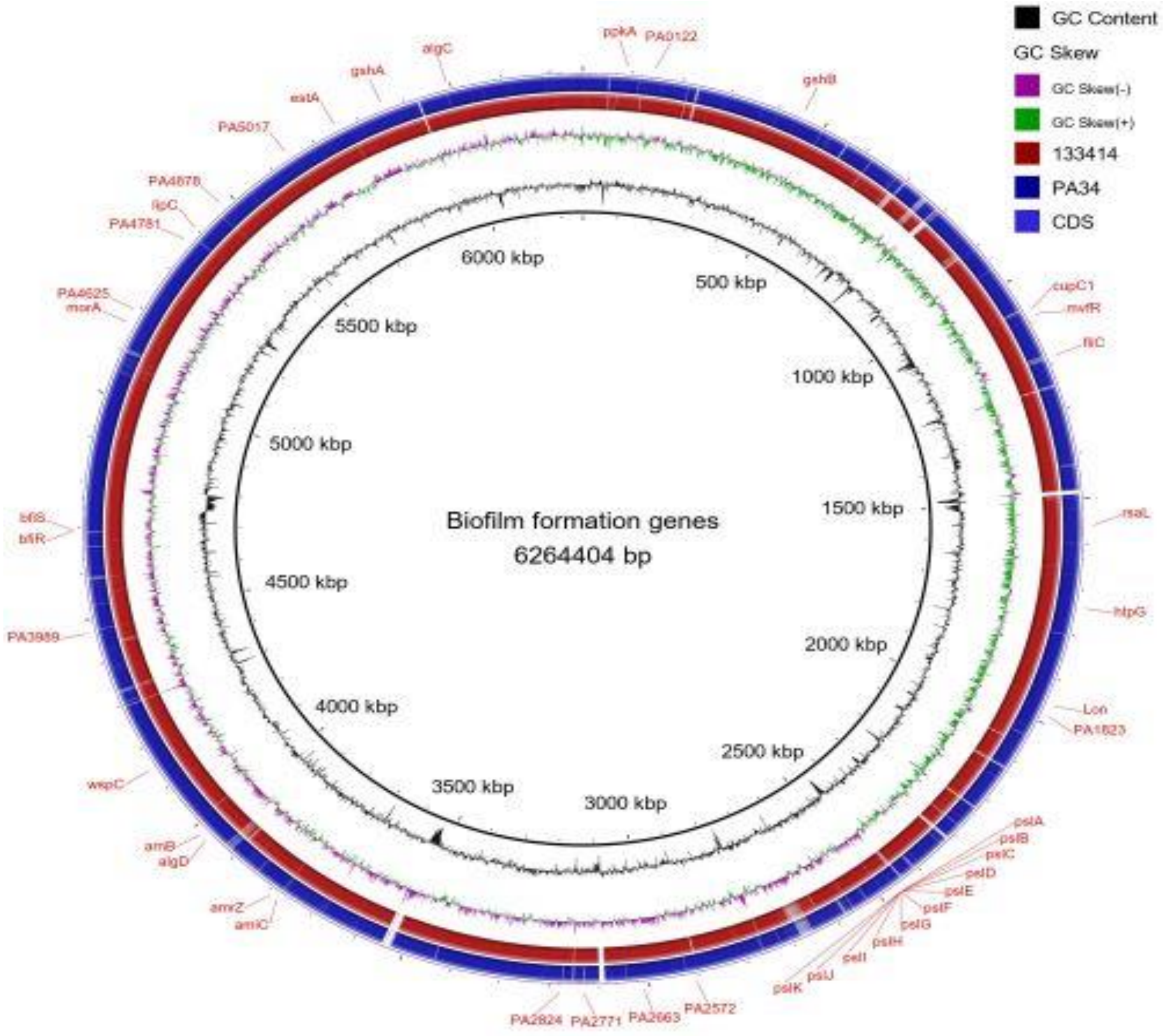


Figure 4.21 BRIG analysis for sequences from the eye.
Genes in the variable regions included fliC at the 1050kbps locus, amiC at the locus 3750kbps, PA3989 at the locus 4450kbps and bfiS and bfiR at the locus 4700kbps.

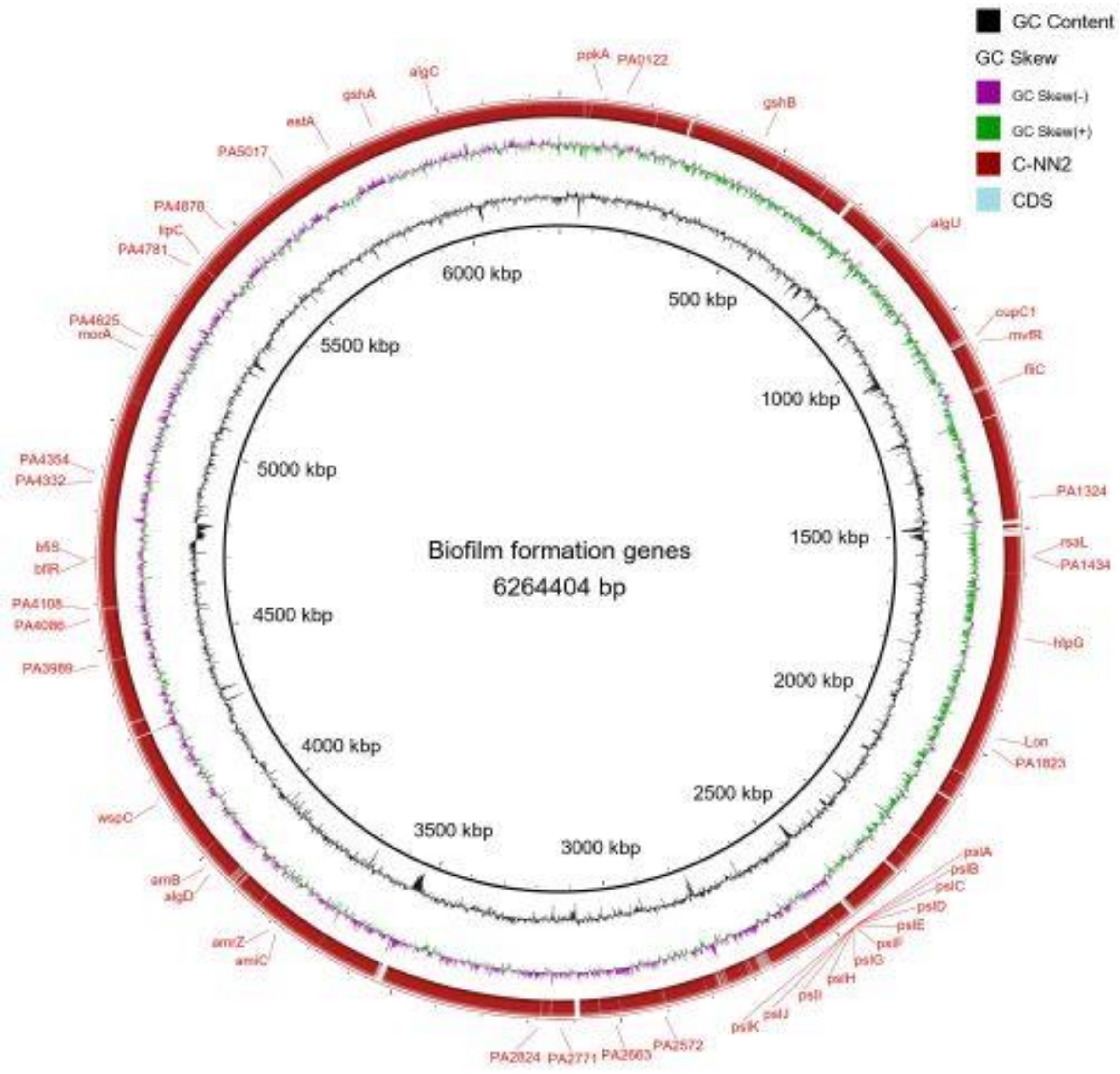


Figure 4.22 BRIG analysis of sequences from the lung.

Genes in the variable regions included cupC1 at locus 1050kbps, fliC at locus 1150kbps, pclass at locus 2400kbps, amiC at locus 3750 kbps, PA3989 at locus 4450kbps and PA4108 at locus 4600kbps.

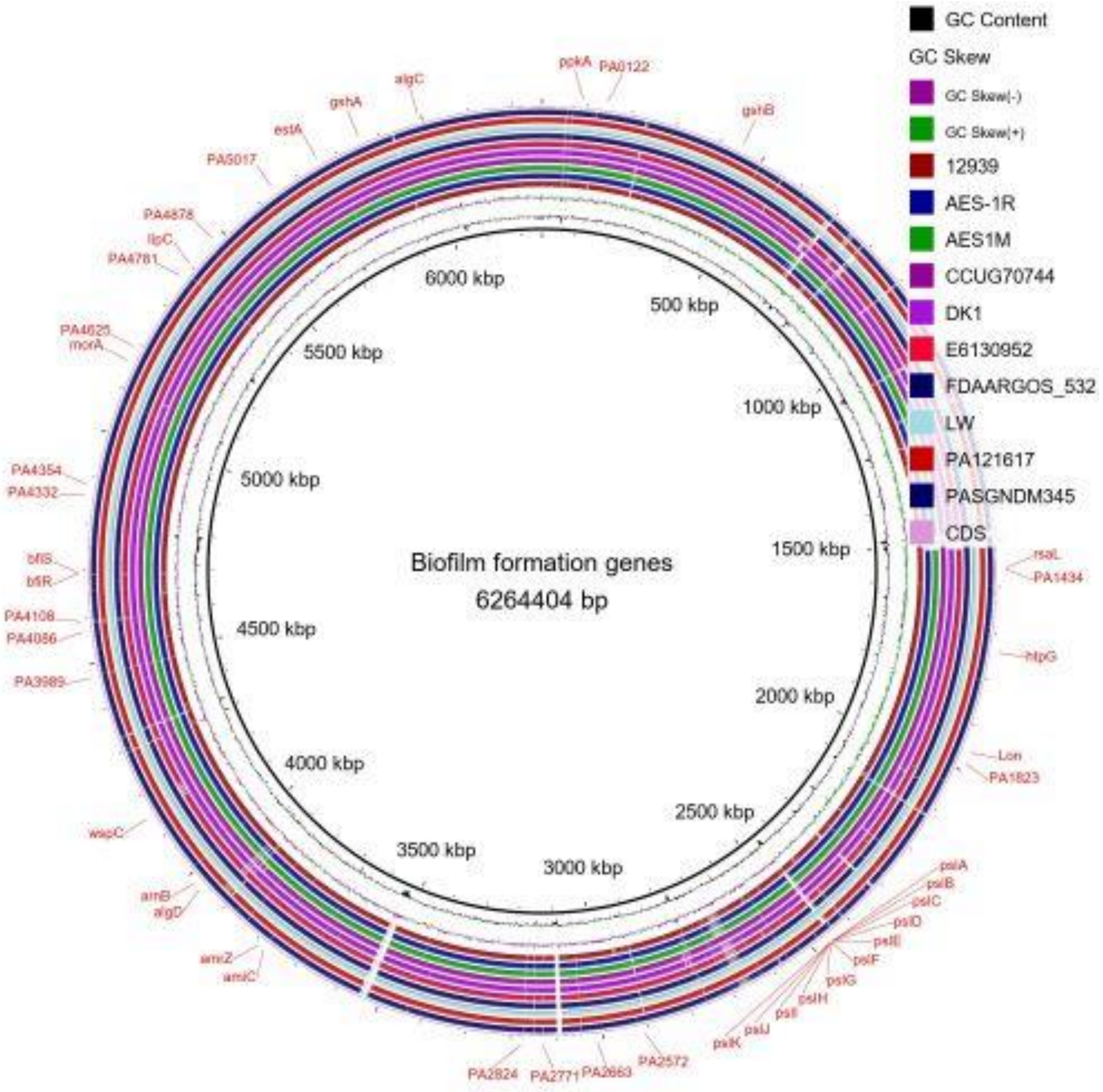


Figure 4.23 BRIG analysis of sequences from the sputum.

Genes located in variable regions included p class at locus 2450kbps, PA2771 at locus 3150kbps, amiC at locus 3800kbps, PA4108 at 4600kbps, bfiS and bfiR at 4700kbps and PA4354 at 4900kbps.

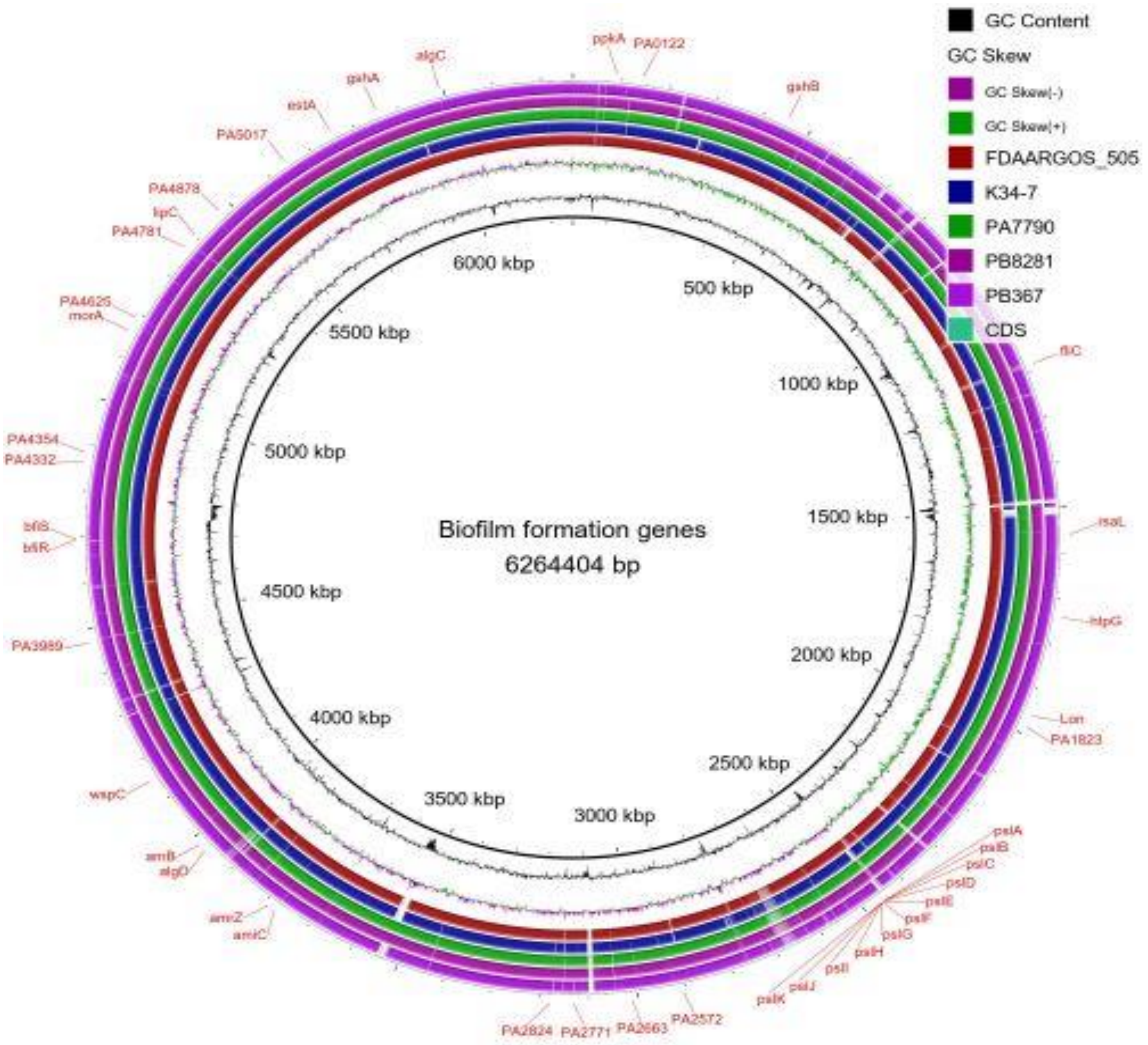


Figure 4.24BRIG analysis for sequence from the trachea.

Genes in the variable regions included fliC at locus 1150kbps, p class at locus 2400kbs, PA2771 at locus 3150kbps, amiC at locus 3750kbps, PA3989 at locus 4450kbps, bfiS and bfiR at locus 4700kbps, PA4625 at locus 5200kbps and algC at locus6000kbps

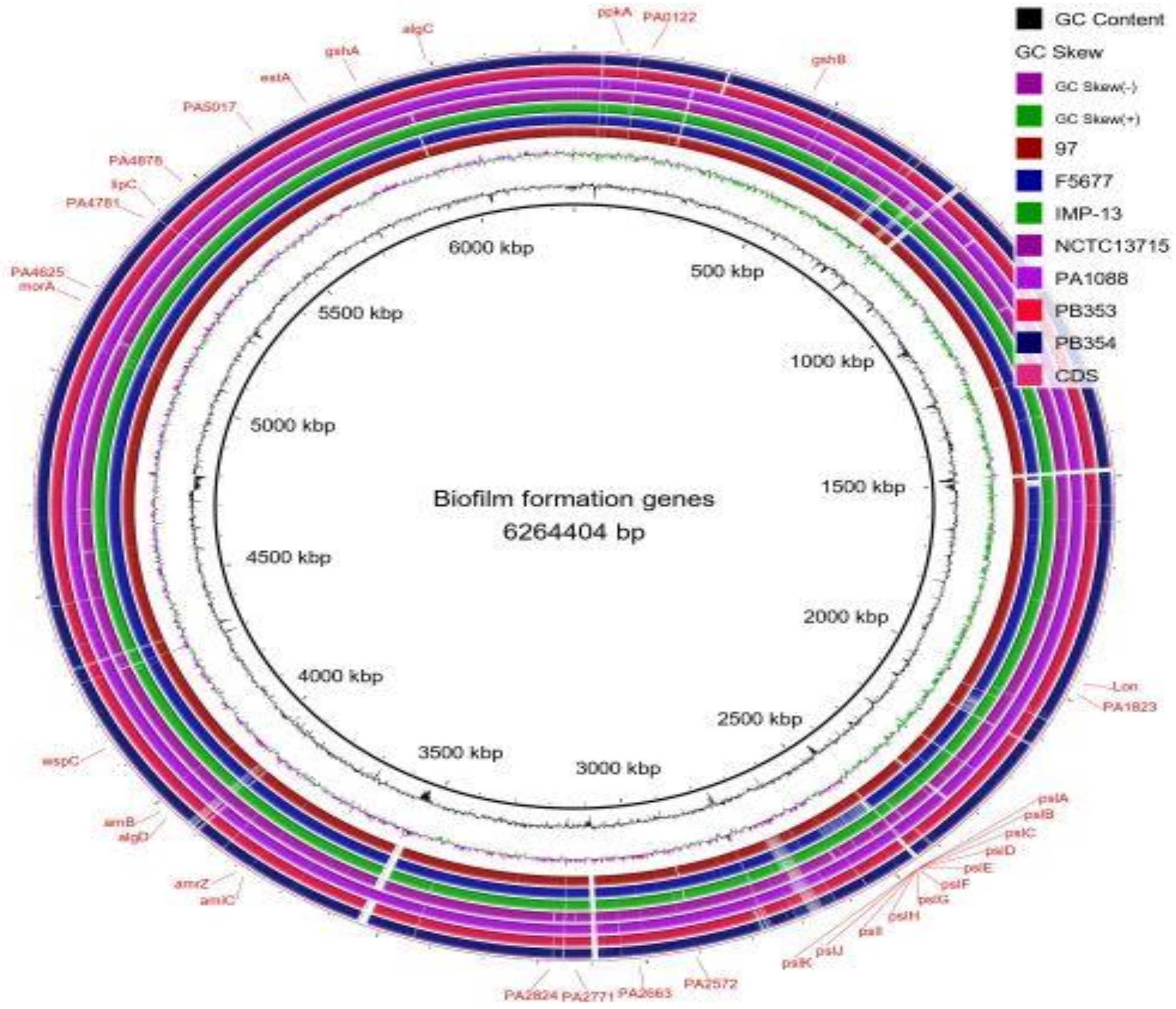


Figure 4.25 BRIG analysis for sequence from the urine samples.
Genes in variable regions included p class at 2400kbp and amiC at locus 3750kbps.

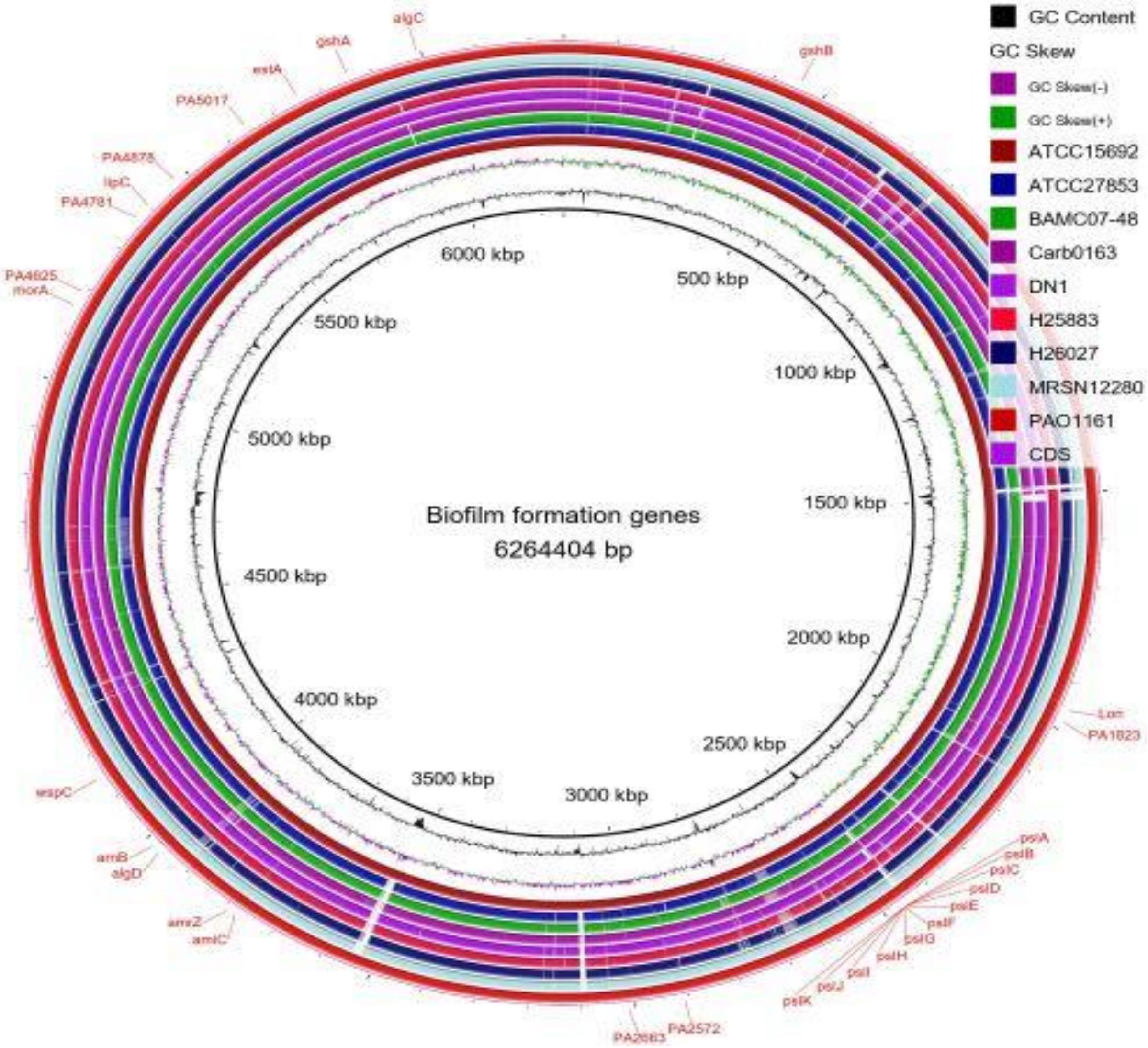


Figure 4.26 BRIG analysis for sequences from wound infections.

Genes in the variable regions included p class at locus 2400kbps, PA2752 at locus 2850kbps and amiC at locus 3750kbps.

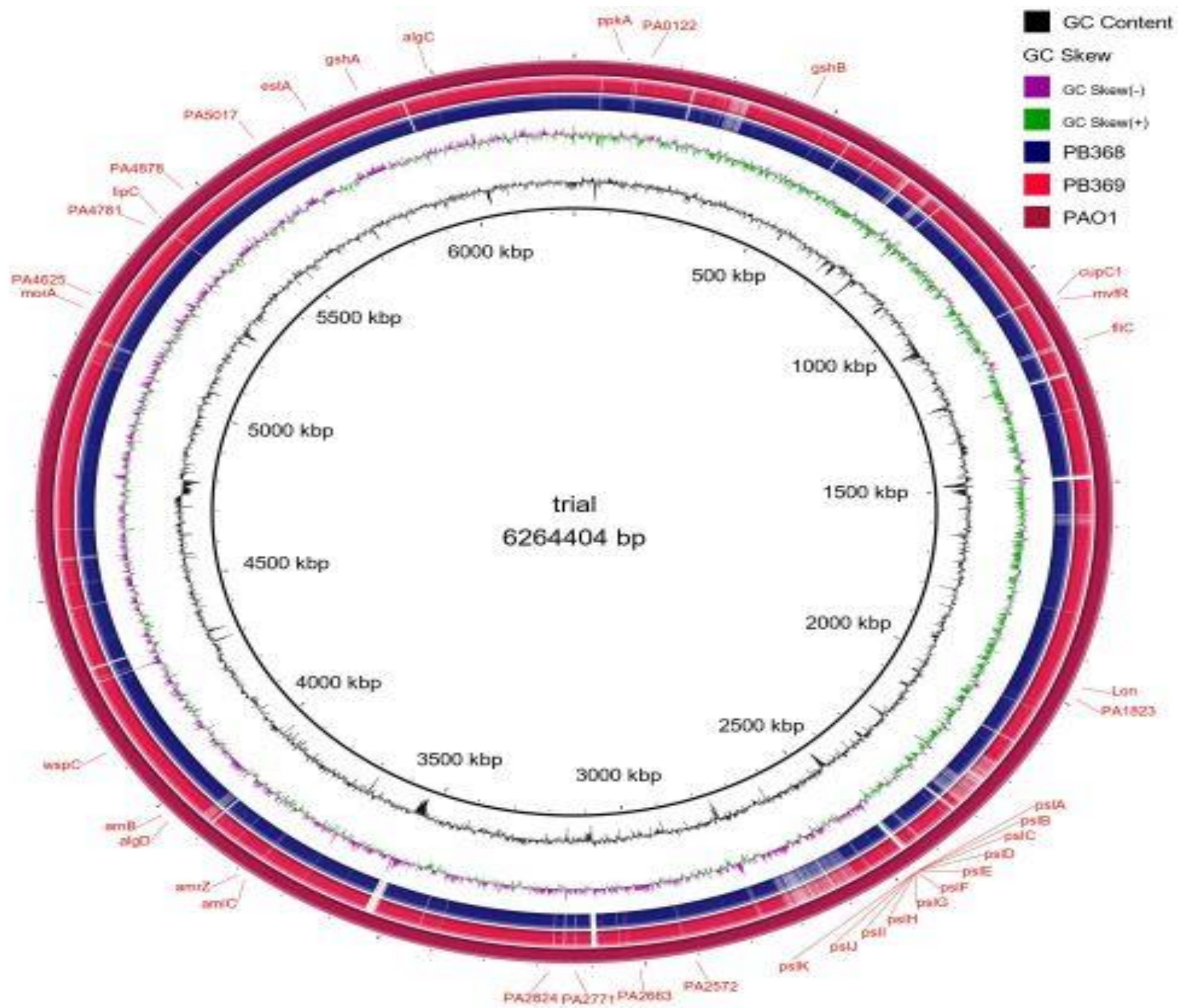


Figure 4.27 BRIG analysis for sequences from the abscess.
Genes in the variable regions included cupC1 at locus 1050kbps, fliC at locus 1150kbps, p class at locus 2400kbps, amiC at locus 3750kbps and algC at locus 6000kbps.

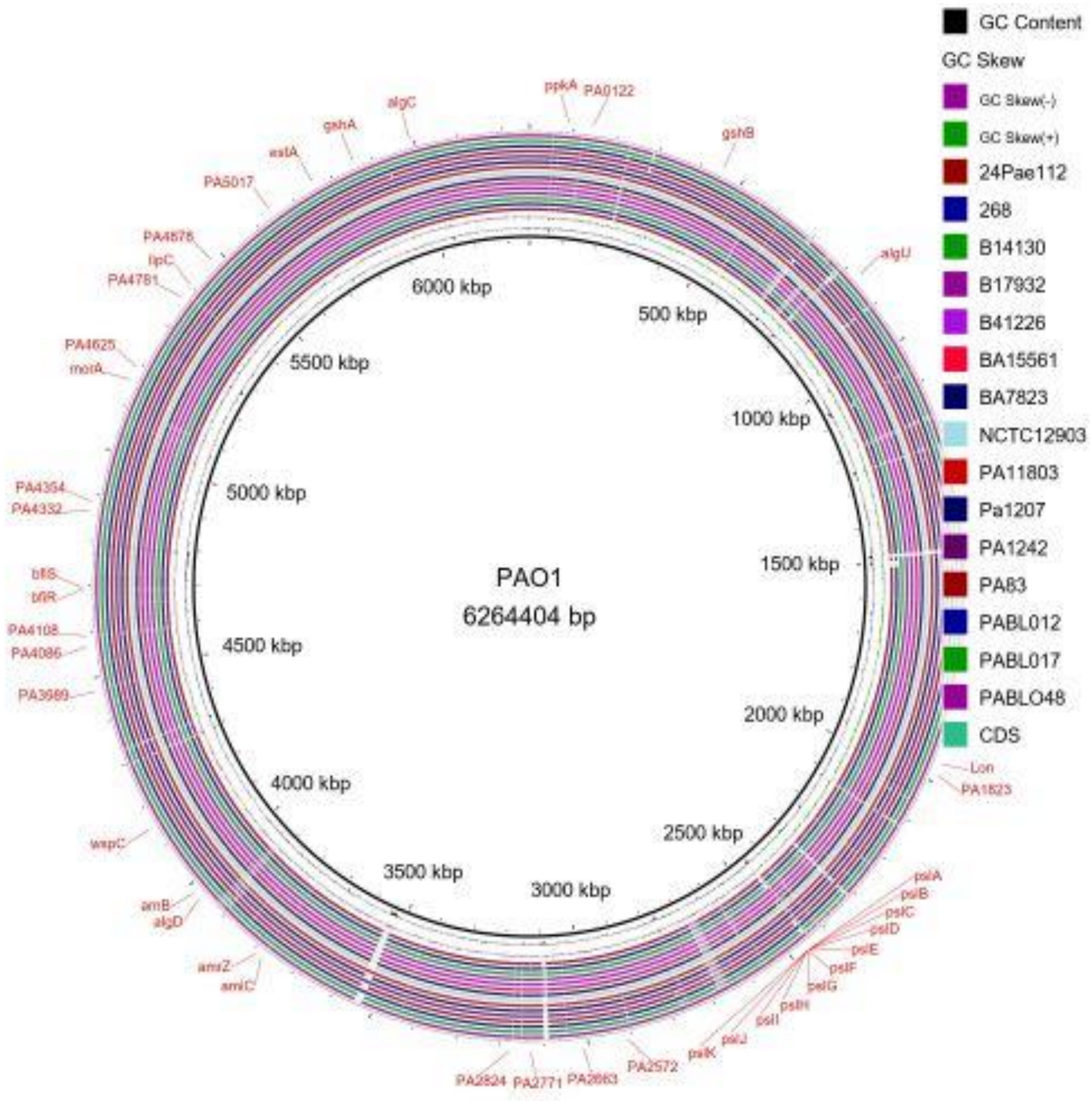


Figure 4.28 BRIG analysis for sequences from blood.

Genes present in the variable regions included p class at locus 2400kbps, PA2824 at locus3150 kbps, amiC at locus 3750kbps, PA3989 at locus 4450kbps, PA4108 at locus 4600kbps, bfiS and bfiR at locus 4700kbps

The overall result of the BRIG analysis indicated high sequence similarity between the different strains of *P. aeruginosa* that were analyzed. The presence of a number of gaps in the analysis can be attributed to the mutations necessary for different strains to adapt and survive in their specific ecological niches.

CHAPTER FIVE
DISCUSSION

5.1 Introduction

Pseudomonas aeruginosa, a Gram-negative bacterium, is a leading cause of nosocomial (hospital-acquired) infections among immunocompromised individuals which has progressively developed resistant genes that have conferred it the ability to withstand the effects of antibiotics, further complicating the treatment of the infections it causes. Strains of the ubiquitous pathogen form biofilms within the human host, further compounding its antibiotic resistance ability and resulting in serious infections that could overwhelm the public health sector (Francisco *et al.*, 2019). In spite of the significance of this phenomenon, exhaustive analyses of the genes responsible for biofilm formation remain scanty and largely undocumented. The use of bioinformatics tools in deciphering and designing candidate drug targets for clinical intervention could be an important step in mitigating the *P. aeruginosa* disease burden. With approximately 176 complete genomes of the ubiquitous pathogen available in the NCBI database, an *in silico* mapping of the highly versatile biofilm formation genes to decipher novel therapeutic target regions was a viable approach (Bruggermann *et al.*, 2018). This study aimed to identify biofilm formation genes, classify them based on the role they play in the biofilm formation process and analyze the distribution of these genes in the genomes of various *P. aeruginosa* strains. All these objectives were aimed at identifying genes and processes that could serve as potential targets for novel antibiofilm therapies. The neighbor-joining algorithm was used to determine the evolutionary trends exhibited by the biofilm gene clusters of *P. aeruginosa* while profile hidden Markov models for genes responsible for biofilm formation in gram negative bacteria were constructed and used to characterize these genes in strains of *P. aeruginosa*. The study identified the conservation of genes in the opportunistic pathogen which were deemed as regions of interest and were identified as potential biomarkers for novel treatment options. The profile also identified variations in the genes.

5.2 Evolutionary Relationships of Biofilm Formation Genes

The study successfully identified 51 biofilm formation genes from the *Entrez* gene search. The regulatory genes were the most frequent (19 out of the 51), highlighting the importance of this step in the biofilm formation process (Table 4.2). This finding also revealed the need for strains of *P.aeruginosa* to constantly regulate the process based on prevailing conditions or the specific ecological niche. The motility genes at 22% (11 out of the 51) indicated the need for the pathogen to regulate mobility during quorum sensing to determine whether or not they would form biofilms once they colonize a viable host. Adhesins, repressors and cell aggregation genes recorded the least percentages. From these findings, it is possible to assume that these processes of biofilm formation are most likely cast on stone and depend on the pathogen's decision to form or not to form biofilms. It is also possible that these processes are not premeditated and may be straight-forward once a group of bacteria opt to form biofilms. These set of genes could be an important target for new drug therapies given that they are highly likely to be expressed once the bacteria achieve a quorum that allows for the formation of biofilms (Freschi *et al.*, 2015). The unclassified set of genes indicates a group of genes that are expressed at different stages of the biofilm formation process. Their expression may be dependent on the ecological niche of the bacteria as well as the specific host that the bacteria colonize.

Besides the biofilm formation genes, the study also collated different strains of *P. aeruginosa* which were classified based on the ecological niches that the different strains occupied. The identification of 13 different ecological niches occupied by the pathogen is consistent with reports that describe *P. aeruginosa* as a ubiquitous pathogen: one that thrives in a wide range of animate and inanimate hosts (Klockgether and Tummler, 2017). The pathogen has also been closely associated with numerous infections among Cystic fibrosis patients (Talwalkar and Murray 2016). This study's results agreed with this association given that 38 of the 84 retrieved human-associated strains (45.23%) were isolated from the respiratory system. 26 of these strains were isolated from sputum samples, one of the hallmarks of cystic fibrosis samples, representing the most preferred ecological niche of *P. aeruginosa*. Blood, clinical and wound samples were the other statistically significant ecological niches.

A custom python script search of genbank files successfully retrieved 13 biofilm formation genes out of the possible 51 genes. This represented 25.49% of the number of genes that were originally obtained. The 25.49% of the genes retrieved were probably a result of incomplete sequence annotation of the retrieved *P. aeruginosa* strain sequences. This finding also indicated that while the Entrez search tool is reliable for sequence search and gene identification, it might not be the best database to use when trying to retrieve and analyze sequences of protein families (Pearson, 2013**)**. For one, the Entrez search only identified sequences associated with the reference strain. This finding may be due to the extensive studies that have already been done on the PAO1 strain. Unfortunately, such information does not give conclusive knowledge on the presence of the biofilm formation genes in the other strains of the bacteria which are of clinical importance. This finding also gives more credence to the use of profile hidden Markov models (pHMMs) to analyze sequences of genes responsible for biofilm formation in *P. aeruginosa.* With these models, the study created a multiple sequence alignment of DNA or protein sequences belonging to the same functional family and built a HMM that effectively represents the common motifs, patterns and statistical properties of the alignment (Eddy, 1998). With a specific architecture, the pHMMs allowed us to suitably model sequence profiles enabling better analyses of sequence families – in our case biofilm formation genes. A study by Madera and Gough, (2002) indicated that profile methods show better performance compared to pairwise methods like NCBI BLAST when finding sequence homologs. Although with a slower speed, the profile-based methods can detect approximately 10% more true homologues than pairwise methods (Madera and Gough, 2002). On the other hand, it is possible that the biofilm formation genes are not present in most strains of the *P. aeruginosa,* meaning that the not all the stains of the pathogen form biofilms to facilitate their survival (Kamali *et al.,* 2020).

The results also inform the need to create a database designated for genes responsible for the phenomenon in the bacteria. Such a database will give the scientific body an easier time as they seek to understand the biofilm formation process further and use this information to find potent mitigation measures against the pathogen. The IPCD database has done a commendable job in facilitating advanced studies on *P. aeruginosa* (Freschi *et al,* 2015). This repository contains thousands of *P. aeruginosa* isolates from plants, human infections, the environment, and animals. As of October 2019, the database contained 1763 isolates and 1165 draft genomes of these isolates. The database facilitates metadata analyses that links bacterial phenotype, genotype and clinical data. While the benefits of the IPCD cannot be overstated, it has laid a lot of emphasis on Cystic Fibrosis (development of prognostic approaches to treating these infections). A database dedicated to the biofilm formation process will only help to augment these efforts.

While the study assumed that all the biofilm formation genes co-evolved together given that they belong to a group of functionally related genes, that generally was confirmed by the obvious co-evolution of these genes – the divergence of three genes (*fliC*, *algD* and *algU*) may have been as a result of a horizontal gene transfer. Alternatively, it may be assumed that these genes evolved faster than other genes as they are more important in terms of a proper response of the biofilm formation to specific environmental stimuli in different habitats. It makes these genes potential targets for antibiofilm therapies. Both the *algD* and *algU* genes were classified as regulatory genes responsible for the regulatory stage of the biofilm formation process.

Protein AlgU has previously been identified as a founding member of the ExtraCytoplasmic Function (ECF) family – a group of products responsible for transcriptional regulation and response to environmental stress (Sineva *et al.,* 2017). This sigma factor has been associated with regulation of genes having the AlgD's promoter (Yin *et al.,* 2013). Previous studies indicate that mutations in *algU* can affect mucoidy in *P. aeruginosa* or lead to a partially active AlgU (Damkiaer *et al.,* 2013). A study by Scalan *et al.* (2015), also indicated that this gene is important for *P. aeruginosa* to withstand various treatments hence the need for an adjustment in its activity by specific mutations. It may explain a rather specific evolution pattern displayed by this protein. Divergence of this gene as seen in figure 4.1 may result from selective accumulation of such mutations. The mucoidy phenotype improves the pathogenesis of *P. aeruginosa* infections as the pathogen acquires increased resistance to antibiotics and phagocytic killing while allowing it to evade the host's immune response (Leid *et al.,* 2005). This prominent role of the gene makes it a suitable target for antibiofilm therapy. Further studies should be conducted to understand the exact mutations of this sigma factor and improve efforts to tackle antibiotic resistance witnessed in infections caused by the ubiquitous pathogens.

Flagellin, on the other hand, has been identified as an important virulence factor in the pathogenesis of *P. aeruginosa* infections. Non-flagellated mutant strains often exhibit less virulence and can hardly invade deeper tissues (Ahmadi *et al.,* 2017). Different strains of *P. aeruginosa* rely especially on flagella during lifestyle switches from planktonic state to biofilms and back to planktonic state. The flagella plays a key role in the attachment and detachment of biofilms from different surfaces (Suriyanarayanan *et al.,* 2016). The study's evolutionary analysis indicated that the *fliC* gene, which encodes flagellin, had a different evolutionary trend compared to the other gene sequence. This was an important finding as mutations in this gene have not been reported by previous studies to the best of our knowledge. Given the importance of the flagellin in the lifestyle of different strains of *P. aeruginosa* it is highly unlikely that this gene was acquired through horizontal gene transfer. This study assumed that this gene has a faster evolutionary rate compared to the other biofilm formation genes under study.

5.2 Creating Profile Hidden Markov Models

There has been a geometrical increase in the number of genomes available in the NCBI database, making it an ideal source to search for specific genes (Land *et al.,* 2015). This database along with other sequence storage databases, like IPCD and DDBJ, however, have exceptionally large amounts of data to be analyzed. This has led to the development of special tools like the profile hidden Markov models to facilitate the analysis of the available data (Eddy, 1998, Yoon, 2009). Profile HMMS apply statistical models that estimate the true frequency of an amino acid or a nucleotide at a specific position of a multiple sequence alignment from the observed frequency. This property allows the subset of HMMs to represent motifs and patterns of multiple sequence alignments (Yoon, 2009). The development of specific profiles to search for biofilm formation genes in 96 genomes of *P. aeruginosa* clearly indicated that this family of proteins have significant levels of variability depending on the ecological niches that different strains of the pathogen occupies.

This study opted to use profile hidden Markov models to represents patterns and motifs of specific biofilm formation genes. Previous studies have proven that the hidden Markov models are highly effective in the analyses of massive amounts of data. Applications of these models ranges from sequence analysis, protein characterization, and gene discovery (Francisco *et al.,* 2019, Restrpo-Montoya *et al.*, 2011). These models are also effective in evaluating whether individual sequences belong to specific profiles (Gong *et al.,* 2012). This study used this property of the models to determine the presence and distribution of different biofilm formation genes in various genomes of *P. aeruginosa* occupying different ecological niches. Analyses of the pathogens' genomes identified genes that were deemed of importance in the biofilm formation process and that could serve as potential targets for novel therapeutic agents against infections caused by *P. aeruginosa*.

The study created profile hidden Markov models that would represent sequence patterns of the biofilm formation gene sequences retrieved by the custom python custom scripts. The models were based on the multiple sequence alignments of the clusters of orthologous genes. Unlike general HMMs, profile HMMs do not contain any circles as they move from left to right, making them suitable for modelling protein and nucleotide sequence data. In the pHMM, the match state 'M' represented the case when a signal in the new sequence matches the symbol in the same position of the original alignment. The match states primarily modeled conserved positions of the alignments and the residue frequencies, consistent with what is described in previous studies (Yoon, 2009). The insert 'I' and delete 'D' states accounted for insertions and deletions in new observation sequences. The insert states represented additional symbols not present in the consensus sequences while the delete states handled amino acids present in the consensus sequence but absent in the original sequence. The emission probabilities of the resulting pHMMs at a specific position represented the observed symbol frequencies in that specific column in the consensus column.

From the representative pHMM present on figure 4.2, it is important to note that each of the match states (M) had four transition probabilities: for the next match state, for the insert state, for the delete state and the end state. Each insert state had two transition probabilities, one for the match state and the other for the same insert state. The delete states also had two transition probabilities, one for the next match state and the other for the next delete states. The transition probabilities from/to the insert and delete states catered for the gap penalties for insertions and deletions in the alignments. These findings were consistent with what is expected for profile HMMs (Eddy, 2011). On the emission probabilities, both the match and insert states had 20 probabilities, each representing one of the twenty amino acids given that the sequences used for these analyses were protein sequences. The delete states, on the other hand, had no emission probabilities as is expected. Given that the delete state represents missing symbols, it is described as a silent state which simply serves as a place-holder to connect neighboring states i.e. match and delete states. It is for this reason that the delete states have no emission probabilities.

The HMM profiles that were developed here proved to be valid to detect biofilm formation genes as it verified the detection of different genes from the reference genome, but excluded the genes in the negative controls during the validation test of the model. The controls were selected based on reported biofilm formation; the positive control was known for forming biofilms while the negative controls don't form biofilms to facilitate their survival (Gong *et al.*, 2012). This validation also indicated that the developed HMM profiles were able to detect specific sequences of biofilm formation, that is, the *algD* profile HMM only detected *algD* gene sequences, the *algM* profile HMM detected *algM* gene sequences and so on. However, the shortcomings of the profiles were evident as no single profile could detect all the different biofilm formation genes. This was probably due to the structural similarities seen in the sequences of the genes as was indicated by the multiple sequence alignments. The study had to create gene-specific profile hidden Markov models to circumvent this challenge. Nevertheless, the HMM profiles developed were totally efficient as they did not detect any false positives in the negative controls.

5.3 Conservation and Variation Patterns of Biofilm Formation Genes

The tendency of different strains of *P. aeruginosa* to form biofilms that facilitate its survival in different ecological niches demonstrates the presence of biofilm formation genes which influenced this process. Different studies have demonstrated the presence of these family of genes in various sequences of the ubiquitous pathogen. Previous reports have also indicated that the *P. aeruginosa* tends to occupy different ecological niches, both in the environment and within the human host. Nevertheless, the findings of these studies were more as a result of chance, rather than a systematic search based on in silico procedures or using large collections of the different strains of *P. aeruginosa.* The present study hypothesized that there are no variations among gene clusters of biofilm formation in *P. aeruginosa* biotypes. This was made based on the fact that this set of genes perform a similar function in the survival mechanism of this organism. The study sought to find out if there are any variations in these genes, especially with regards to the different ecological niches that the pathogen is known to occupy. The approach followed here helped to identify the distribution of biofilm formation genes among strains occupying different ecological niches while identifying ecological niches with the highest abundance of hits of biofilm formation genes.

On screening the retrieved sequences of *P. aeruginosa* from different ecological niches, the study identified a significant number of hits for most of the ecological niches apart from the lung and dental ecological niches. These results give more credence to the importance of biofilm formation to the survival of the ubiquitous in different environments. Human niches also indicated a significantly higher number and density of hits compared to the non-human niches. The results indicated that *P. aeruginosa* is more likely to form biofilms that increases its chances of survival once it colonizes the human host. This finding is consistent with previous studies which have correctly indicated that biofilm formation significantly contributes to the antibiotic resistance ability of this pathogen resulting in chronic illnesses for susceptible patients (Olsen, 2015).

The *algD* gene, previously described as a component of the alignate operon, demonstrated the highest number of hits compared to the other biofilm formation genes. Alignate biosynthesis, modification and export is important to chronic *P. aeruginosa* as these processes contribute significantly to antibiotic resistance and opsonization, resulting in highly potent pathogens (Okkotsu *et al.*, 2014). The significantly higher number of hits indicate an insistent need by the pathogen to express the *algD* gene. A clear understanding of the expression and mutation habits of this gene could prove worthwhile in the bid of developing novel treatment options against pathogenic strains of *P. aeruginosa* pHMM screening of the biofilm formation genes also indicated that the *htpG* gene had a significantly higher number of hits in human isolates compared to the non-human isolates. The *htpG* gene has previously been described as a heat protein gene which helps *P. aeruginosa* bacteria survive in environmental stress. A recent study indicated that Δ*htpG* mutant strains have significantly diminished adhesion, swimming, swarming and twitching motility as compared to the wildtype strains. Mutation of this gene also resulted in reduced biofilm formation as it affected the processes of adhesion, bacterial motility, and cellular aggregation (Grudniak *et al.,* 2018). The significantly higher occurrence of this gene in human isolates understates its importance in the survival of the pathogen when it colonizes the human environment. *P. aeruginosa* relies on functional flagella to swim and colonize different surfaces. Pili-mediated twitching motility also facilitates the movement and colonization ability of this pathogen. Both these processes are *htpG* dependent as indicated by previous studies (Kazmierczak *et al.,* 2015). Novel therapeutic agents targeting this gene can impair virulence determinants of *P. aeruginosa* increasing the likelihood of completely wiping out infections that the pathogen causes.

It is important to note that the results from this work showed that the biofilm formation genes are distributed among *P. aeruginosa* strains occupying different ecological niches. This observation is consistent with previous reports that have indicated that the pathogen forms biofilms to facilitate its survival within different hosts. These results validate the question about the variability of the biofilm formation genes based on different ecological niches occupied by different strains of *P. aeruginosa*. As shown here, the biofilm formation genes are highly dispersed and frequently found, regardless of the niche occupied by different strains of the pathogen. Much work is still to be done to reveal how the genes specifically affect the survival of the pathogen once it occupies different ecological niches.

The study further performed genome mapping analyses to determine the distribution of the biofilm formations genes among the different strains of *P. aeruginosa*. These analyses were niche specific to help determine if these genes were present in the variable and conserved regions of the pathogen's genome. The analyses also helped us determine the locus of the genes with the pathogen's sequences. Previous studies have revealed that *cupB* and *cupC* genes which were classified as cell aggregation genes play an important role in the biofilm formation process, especially through micro-colony formation and bacterial clustering (Segolene *et al.,* 2007). In the absence of appendages like the flagella and type IVa, the CupC system has more significant contributions making it a more viable target for anti-biofilm therapies. This study determined that the *cupC1* gene is located around the 1050kbps in the genomes of *P. aeruginosa* PAO1 according to the BRIG analysis (Figure 4.16**).** These analyses also revealed that this region is variable in the sequences of strains of *P. aeruginosa* isolated from the lungs, abscess and the dental area (Figure

4.22, 4.23 and 4.19). This finding could indicate that the gene continues to experience constant mutations depending on the ecological niche further highlighting its influence in the biofilm formation process. The gene was, however, located in constant regions of the genomes of *P. aeruginosa* strains isolated from the other ecological niches (Figures 4.17, 4.18, 4.19, 4.20, 4.21, 4.22, 4.23, 4.24 and 4.25). This could probably downplay the need of mutations of this gene in these ecological niches. It is also possible that the cell aggregation property of the bacterium has not been impaired when it colonizes these niches (Segolene *et al.,* 2007). The two sets of genes were not retrieved by the python scripts and were not analyzed by the PHYLIP phylogenetic tree (Figure 4.1) and the pHMMs.

Interestingly, none of the BRIG analyses was able to pick out the *cupB1* gene. This could indicate that its role in the process could easily be downplayed with the *cupC1* gene playing a more prominent role in the cell aggregation and microcolony formation which are important steps of biofilm formation. These findings further heighten the possibility of new therapies that could target pathways and processes that are mediated by the *cupC1* gene. The study by Segolene *et al.,* 2007 indicated a near complete failure by bacteria to form aggregates when the *cupC3* gene, part of the *cupC* system, was deleted. This was contrasted by a strong aggregative phenotype when there was an overexpression of this gene (Segolene *et al.,* 2007). The study however, suggested a synergy between the *cupB* and *cupC* systems that facilitate the assembly of fimbrial structures for better cell-to-cell interactions leading to a proper architecture of the biofilm.

At 2%, genes responsible for repressing the biofilm formation process were among the least abundant class of genes. The present study classified *gshA* and *gshB* differently with the two sets of genes occupying the repressors and motility classes respectively (Table 4.2). The two genes, however, demonstrated a close ancestral relationship. A finding that was highlighted by the

PYHLIP phylogenetic tree which clustered the two genes in the same clade (Figure 4.1). *gshB* has previously been shown to facilitate glutathione (GSH) biosynthesis. In a study that deleted the *gshA,* or its mutant variant *gshB,* the role of GSH in the biofilm formation process was highlighted by reporting an increase in biofilm formation (Wongsaroj *et al.,* 2018). Those results suggested that GSH had a role to play in repression of biofilm formation. A previous observation has also revealed that GSH can disrupt immature and mature biofilms formed by *P. aeruginosa* (Klare *et al.,* 2016). The bacterium tends to suppress the expression of these genes especially during cell aggregation. The BRIG analysis revealed that the *gshA* gene is located at 5680kbps on the genome of *P. aeruginosa.* This region is highly conserved in all the *P. aeruginosa* sequences analyzed during this study (Figure 4.16 – 4.28). This finding highlights the lack of variability in the evolution of this gene. Novel antibiofilm therapies could target the *gshA* expression by targeting its possible repressors. Such therapies could facilitate the suppression of biofilm formation by the pathogen to improve the antibiotic therapy. The *gshB* gene, on the other hand, is located at the 450kbps conserved region of the genome of *P. aeruginosa* (Figure 4.16). The conservation of this position indicates less probability of notable mutations within this gene providing a probable target for novel antibiofilm therapies. From the pHMM screening results, the *gshB* indicated hits in only two ecological niches, the environment and eye niches. The *gshA* gene, on the other hand, indicated hits in 10 of the 13 ecological niches. It only lacked hits in the dental, eye and lung ecological niches. None of these genes exhibited significant results when the Wilcoxon signed rank test was performed to determine the difference in the number of hits between the human and non-human samples. This result indicated that the gene may not necessarily have a preference in strains that occupy human or environmental samples. A study by Rao *et al.,* (2011) indicated that *P. aeruginosa* PA0122 negative mutant strains had a significant increase in biofilm formation compared to the strains that contained this gene. This highlights the role of the PA0122 gene as a repressor of the biofilm formation process. The BRIG analysis revealed that the PA0122 gene is located at the 1050kbps position in the reference genome (Figure 4.16). This is a conserved region on the genome sequences of *P. aeruginosa*. The PA0122 gene was, however, not retrieved using the python script and was therefore not used in downstream analyses.

The study by Wilhelm *et al.,* (2007) demonstrated that mutant strains of *P. aeruginosa* lacking the gene *estA* showed no swarming motility. Swimming and twitching motility, the other forms of surface motility, were also absent in the mutant strains. It is known that the formation of the three-dimensional biofilm architecture relies on the swarming motility of bacterial cells. The lack of this gene impaired biofilm formation in the mutant strains. The present study identified this gene at 5750 kbps locus that is highly conserved among all sequences of *P. aeruginosa* (Figure 4.16). The *estA* gene was, however, not retrieved by the python script and was thus not part of the PHYLIP phylogenetic analyses. Further analyses needs to be done to demystify the evolutionary relatedness between classes of such biofilm formation genes. The functional dissimilarities between the two sets of genes are rather clear given that the PA4878 gene has been found to be a transcriptional regulator (Chambers *et al.,* 2014). The gene effects its action through the regulation of the c-di-GMP molecule that has a hand in the correlation between the formation of biofilms and the antibiotic resistance ability of the refractory pathogen (Gupta *et al.,* 2014). The BRIG analysis revealed that it is located at 5450 kbps highly conserved locus in the reference genome (Figure 4.16). A study by Jones *et al.,* (2013) indicated that *amrZ* repressed the transcription of the psl operon which in turn modulated the architecture of biofilm layers in *P.aeruginosa.* The protein encoded by this gene binds to a site overlapping the *pslA* promoter (Jones *et al.,* 2013). This observation is consistent with reports describing *amrZ* as a multifunctional regulator activating the *algD* operon besides the repression of the *psl* operon. Given that this process annuls biofilm tower formation in *P. aeruginosa,* various treatment strategies are compromised. The BRIG analysis identified the location of *amrZ* at 3700kbps locus in the reference genome. This region is highly conserved in the genome of different strains of *P. aeruginosa* (Figure 4.16). The *amrZ* gene was, however, not retrieved by the python script hence it was part of the PHYLIP phylogenetic analysis. Proper understanding of interactions mediated by this gene could potentially result in development of therapeutic agents that impair biofilm formation. Infections that are refractory to antibiotic treatments could be kept in check.

CHAPTER SIX
CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This study aimed to identify biofilm formation genes, classify them based on the role they play in the biofilm formation process and analyze the distribution of these genes in the genomes of various strains of *P.aeruginosa.* All these objectives were aimed at identifying genes and processes that could serve as potential targets for novel antibiofilm therapies.

The study successfully identified 51 biofilm formation genes using the *Entrez* search protocol available on NCBI. This number was, however, reduced to 13 after a python script search was performed. While the Entrez search is important for identification of different proteins of interest, it might not be ideal to identify homologous sequences. A BLAST search of individual genes or writing custom scripts to select the gene sequences from annotated genomes is an ideal option for such undertakings. Biofilm formation genes generally seemed to have evolved together overtime. The *algD, algU* and *fliC* which were the exceptions in this study may have been obtained via horizontal gene transfer from other bacteria in the respective niches of *P. aeruginosa*. The study also assumed that these set of genes may have indicated faster evolution than the other set of genes. The study's hypothesis that gene clusters with similar functions evolved similarly was negated with our findings that the evolutionary trends of the biofilm formation genes in *P. aeruginosa* did not necessarily depend on the functional relationships. Given that the biofilm formation genes of the same class are clustered differently in the cladogram, further analyses are needed to decipher the relationships between their functions and evolution. From the phylogenetic analyses of the pathogen, it was clear that *P. aeruginosa* isolates mostly cluster into two major clades. These phylogenetic analyses also indicated that different genomic regions of the *P. aeruginosa* have different evolutionary mechanisms. It is also clear that some of these genes may have been obtained through horizontal gene transfer to enable the pathogen to survive in different ecological niches. These findings further pointed to the high diversification of different strains of *P. aeruginosa* present in different ecological niches.

The study successfully used profile hidden Markov models to determine the levels of variability in the genes responsible for biofilm formation in different strains of the pathogen under study. These models correctly represented the sequence patterns in the different biofilm formation genes given that they were constructed from the multiple sequence alignments of these sequences. Protein sequences were ideal for construction of the models as they factored in mutations likely to have occurred within the DNA sequences.

From the analyses of different *P. aeruginosa* strains using the gene-specific pHMMs, the study indicated that *P. aeruginosa* is more likely to form biofilms once it colonizes the human host than when it colonizes a non-human host as indicated by the different density of hits in the two hosts. *P. aeruginosa* has a high tendency to form biofilms once they colonize the abscess ecological niche, most likely to be found in wounds. The *algD* gene was identified as a possible target for novel antibiofilm therapies given that it was reported as the most abundant biofilm formation gene between the various strains of the pathogen. This result along with *algD's* different evolutionary characteristic lays further claim to its potential as a novel drug target site. The diverse distribution of genes involved in biofilm formation in *P.aeruginosa* highlights the diversity of pathways the bacterium can explore in different ecological niches. The conservation and variability of some of these genes in particular niches offers the scientific body a lot to think about especially on the possibility of new antibiofilm therapies.

5.2 Recommendations for Future Studies

Further research should be carried on:

1. The role of horizontal gene transfer in the evolution differences witnessed among biofilm formation genes in different strains of *P. aeruginosa*. The expression patterns of biofilm formation genes in *P. aeruginosa* strains occupying different ecological niches.

2. Comparison between pHMM and other well-known search engines.

3. *algU, algD, fliC* and *htpG* genes as potential target sites for novel antibiofilm therapies.

REFERENCES

Ahmadi H, Behrouz B, Irajian G, Amirmozafari N & Naghazi S (2017). Bivalent flagellin immunotherapy protects mice against *Pseudomonas aeruginosa* infections in both acute pneumonia and burn wounds models. *Biologicals: Journal of the International Association of Biological Standardization, 46*, 26-37.

Ahola V, Aittokallio T, Uusipaikka E & Vihinen M (2003). Efficient estimation of emission probabilities in profile hidden Markov models. *Bioinformatics*, 2003, 19, 2359-2368.

Alikhan NF, Petty NK, Ben Zakour NL & Beatson SA (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics, 12,* 402.

Alekshun MN & Levy SB (2007). Molecular mechanisms of antibacterial multidrug resistance. *Cell*. 128 (6):1037±50.

Araujo BF, Ferreira ML, Campos PA Royer S, Batistão DW, Dantas RC, Gonçalves IR, Faria AL, Brito CS, Yokosawa J, Gontijo-Filho PP, & Ribas RM (2016). Clinical and molecular epidemiology of multidrug-resistant P. aeruginosa carrying aac (60)-Ib-cr, qnrS1 and blaSPM genes in Brazil. *PLoS One*; 11: e0155914.

Armour AD, Shankowsky HA, Swanson T, Lee J & Tredget EE (2007). The impact of nosocomially-acquired resistant *Pseudomonas aeruginosa* infection in a burn unit. *Journal of Trauma 63* 164–171.

Altschul SF & Gish W (1996). Local alignment statistics. *Methods in Enzymology. 266,* 460-80.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ. (1990). Basic local alignment search tool. *Journal of Molecular Biology 215,* 403-410.

Altschul SF, Madden TL, Schaffer AA, Shang J, Zhang Z, Miller W & Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research 25*, 3389-3402.

Awad M, Fahmy RM, Mosa KA, Helmy M & El-Feky FA (2017). Identification of effective DNA barcodes for Triticum plants through chloroplast genome-wide analysis. *Computational Biology and Chemistry. 71,* 20-31.

Bai Y, Muller DB, Srinivas G, GArrido-Oder R, Potthoff E, Rott M, Nina D, Philipp CM, Stijn S, Mitja RE, Bruno H, Alice CM, Julia A V & Paul S (2015). Functional overlap of the Arabidopsis leaf and root microbiota. *Nature 528*, 364-369.

Balasubramanian D & Mathee K (2009). Comparative transcriptome analyses of *Pseudomonas aeruginosa*. Hum Genomics 3:349-361.

Barrett C, Hughey R & Karplus K (1997). Scoring hidden Markov models. *Comput. Applic. Biosci.*, **13**, 191-199.

Bellanger X, Payot S, Leblond-Bourget N, and Guédon G (2014). Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol Rev*. Jul; 38(4):720-60.

Bezuidt O, Lima-Mendez G, Oleg R (2009). SeqWord Gene Island Sniffer: A program to study the lateral genetic exchange among bacteria. World Academy of Science, Engineering and Technology. 58. 1169-1174.

Bianconi I, Jeukens J, Freschi L Beatriz AF, Marcella F, Brian B,Antonio M, Irena K, Burkhard T, Roger CL, Alessandra B (2015). Comparative genomics and biological characterization of sequential *Pseudomonas aeruginosa* isolates from persistent airways infection. *BMC Genomics* ; 16:1105.

Bruggemann H, Leticia B, Romario O, Paula C, Souza AV, Jensen A, Poehlein A, Brzuszkiewicz E, Doi AM, Pasternak J, Martino MDV, Severino P (2018). Comparative Genomics of Non-outbreak Pseudomonas aeruginosa Strains Underlines Genome Plasticity and Geographic Relatedness of the Global Clone ST235. Genome Biol. Evol. 10(7):1852–1857.

Bruno WJ (1996). Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.*, **13**, 1368-1374.

Castresana J (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol. 17*(4):540-52. doi: 10.1093/oxfordjournals.molbev.a026334

Chambers JR, Liao J, Schurr MJ and Sauer K (2014). BrlR from Pseudomonas aeruginosa is a c-di-GMP-responsive transcription factor. Mol. Microbiol. 92, 71–487.

Centers for Disease Control and Prevention (2019). Antibiotic Resistance Threats in the United States. *U.S Department of Health and Human Services*

Choi JY, Sifri CD, Goumnerov BC, Rahme LG, Ausubel FM, Calderwood SB (2002) Identification of virulence genes in a pathogenic strain of *Pseudomonas aeruginosa* by representational difference analysis. *Journal Of Bacteriology*, 184**:**952-961.

Csete M, Doyle J (2004). Bow ties, metabolism and disease. TRENDS in Biotechnology 22: 446-450.

Dai, J., & Cheng, J. (2008). HMMEditor: a visual editing tool for profile hidden Markov model. *BMC genomics*, *9 Suppl 1*(Suppl 1), S8.

Damkiaer S, Yang L, Molin S & Jelsbak L (2013). Evolutionary remodeling of global regulatory networks during long-term bacterial adaptation to human hosts. *Proc Natl Acad Sci USA. 110*(19):7766-7771. doi: 10.1073/pnas.1221466110

Durbin R, Eddy SR, Krogh A and Mitchison GJ (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK.

Eddy S (1998). Profile hidden Markov models. Bioinformatics 14:755–763.

Eddy S (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*., 7:e1002195

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797

Felsenstein J (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791.

Felsenstein J (2005). PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author*. *Department of Genome Sciences, University of Washington, Seattle.*

Finn RD, Coggill P, Eberhardt RY, Eddy SZ, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*; 44:D279–D285

Francisco CE, Ismael HG, Jorge EI (2019). Search for Cry proteins expressed by *Bacillus* spp. genomes, using hidden Markov model profiles. 3 Biotech 9:13.

Freschi L, Jeukens J, Kukavica-Ibrulj I, Boyle B, Dupont MJ, Laroche J, Larose S, Maaroufi H, Fothergill JL, Moore M, Winsor GL, Aaron SD, Barbeau J, Bell SC, Burns JL, Camara

M (2015). Clinical utilization of genomics data produced by the international *Pseudomonas aeruginosa* consortium. *Front Microbiol* 6:1036

Geer RC, Sayers EW (2003) Entrez: making use of its power. Brief Bioinform 4: 179-184.

Gilpin W (2016). PyPDB: a Python API for the Protein Data Bank. *Bioinformatics* 32(1): 159-160

Goldman N, Thorne JL and Jones DT (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.,* **263**, 196-208.

Gong YN, Chen GW, Shih SR (2012). Characterization of subtypes of the influenza A hemagglutinin (HA) gene using profile hidden Markov models. J Microbiol Immunol Infect 45:404–410.

Gooderham WJ, Hancock RE (2009). Regulation of virulence and antibiotic resistance by two-component regulatory systems in Pseudomonas aeruginosa. FEMS microbiology reviews. 33(2):279-94.

Goodman AL and Lory S (2004). Analysis of regulatory networks in Pseudomonas aeruginosa by genomewide transcriptional profiling. Curr. Opin. Microbiol. Vol. 7, pp. 39–44.

Guo XL, Wang DC, Zhang YM, Wang XM, Zhang Y, Zuo Y, Zhang DM, Kan B, Wei L and Gao Y (2008). Isolation, identification and 16S rDNA phylogenetic analysis of *Klebsiella pneumonia* from diarrhea specimens. *Chinese Epidemiology Journal* **29**, 1225-1229.

Grudniak AM, Klecha B and Wolska KI (2018). Effects of null mutation of the heat-shock gene *htpG* on the production of virulence factors by *Pseudomonas aeruginosa. Future microbiology, 13,* 69-80.

Gupta K, Liao J, Petrova OE, Cherny KE and Sauer K (2014). Elevated levels of the second messenger c-di-GMP contribute to antimicrobial resistance of Pseudomonas aeruginosa. Mol. Microbiol. 92, 488–506.

He J, Baldini RL, Déziel E, Saucier M, Zhang Q, Liberati NT, Lee D, Urbach J, Goodman HM, and Rahme LG (2004). The broad host range pathogen *Pseudomonas aeruginosa* strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2530–253510.1073/pnas.0304622101

Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK and Hood L (1997). Gene families: The taxonomy of protein paralogs and chimeras. *Science, 278,* 609-614.

Hengge R (2009). Principles of c-di-GMP signalling in bacteria. *Nature reviews Microbiology* 7: 263-273.

HMMER. http://hmmer.janelia.org

Jakobsen TH, Bjarnsholt T, Jensen PØ, Givskov M, Høiby N (2013). Targeting quorum sensing in Pseudomonas aeruginosa biofilms: current and emerging inhibitors. *Future Microbiol* **8**:901–21.

Jansen G, Mahrt N, Tueffers L, Barbosa C, Harjes M, Adolph G Anette F, Annegret KW, Philip R, Hinrich S (2016). Association between clinical antibiotic resistance and susceptibility of Pseudomonas in the cystic fibrosis lung. Evolution, Medicine, and Public Health.

Johnson LS, Eddy SR, Portugaly E (2010). Hidden markov model speed heuristic and iterative hmm search procedure. *BMC Bioinformatics. 11*:431

Jones AM, Govan JR, Doherty CJ, Dodd ME, Isalska BJ, Stanbridge TN, Webb AK (2001). Spread of a multiresistant strain of Pseudomonas aeruginosa in an adult cystic fibrosis clinic, Lancet Vol. 358, pp. 557–558.

Jones CJ, Ryder CR, Mann EE and Wozniak DJ (2013). AmrZ modulates *Pseudomonas aeruginosa* biofilm architecture by directly repressing transcription of the *psl* operon. J Bacteriol 195:1637-1644.

Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, and Crook DW (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS microbiology reviews*, *33*(2), 376–393.

Kamali E, Jamali A, Ardebili A, Ezadi F and Mohebbi A (2020). Evaluation of antimicrobial resistance, biofilm forming potential, and the presence of biofilm-related genes among clinical isolates of *Pseudomonas aeruginosa*. *BMC Res Notes* 13**(**27).

Kazmierczak BI, Schniederberend M and Jain R (2015). Cross-regulation of Pseudomonas motility systems: the intimate relationship between flagella, pili and virulence. *Current opinion in microbiology*, *28,* 78-82.

Kearns DB (2010). A field guide to bacterial swarming motility. Nature reviews Microbiology 8: 634-644.

Kearns DB (2013). You get what you select for: better swarming through more flagella. Trends Microbiol 21: 508-509.

Khan HA, Ahma A and Mehboob R (2015). Nosocomial infections and their control strategies. Asian Pacific Journal of Tropical Biomedicine, 5, 509-514.

Klockgether J and Tummler B (2017). Recent advances in understanding *Pseudomonas aeruginosa* as a pathogen. *F1000Research, 6*, 1261.

Kim YJ, Jun YH, Kim YR, Park KG, Park YJ, Kang JY and Kim SI (2014). Risk factors for mortality in patients with *Pseudomonas aeruginosa* bacteremia; retrospective study of impact of combination antimicrobial therapy. *BMC infectious diseases, 14,* 161.

Klare W, Das T, Ibugo A, Buckle E, Manefield M and Manos J (2016). Glutathione-disrupted biofilms of clinical *Pseudomonas aeruginosa* strains exhibit an enhanced antibiotic effect and a novel biofilm transcriptome. *Antimicrob Agents Chemother*. 60(8):4539–51

Krogh A, Brown M, Mian IS, Sjolander K and Haussler D (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.,* **235**, 1501-1531.

Kruglyak L, Daly MJ, Reeve-Daly MP and Eeckman FH (1996). A generalized hidden Markov model for the recognition of human genes in DNA. In *Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology,* **4**, 134-141. AAAI Press, Menlo Park.

Kumar S, Stecher G, Li M, Knyaz C and Tamura K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol*. Jun 1;35(6):1547-1549.

Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW (2015). Insights from 20 years of bacterial genome sequencing. Funct Integr Genom 15:141–161.

Lee DG, Urbach JM, Wu G, Liberati NT, Feinbaum RL, Miyata S, Diggins LT, He J, Saucier M, Déziel E, Friedman L, Li L, Grills G, Montgomery K, Kucherlapati R, Rahme LG, Ausubel FM (2006). Genomic analysis reveals that Pseudomonas aeruginosa virulence is combinatorial, Genome Biol. Vol. 7, p. R90.

Leid JG, Willson CJ, Shirtliff ME, Hassett DJ, Parsek MR & Jeffers AK (2005). The exopolysaccharide alignate protects *Pseudomonas aeruginosa* biofilm bacteria from IFN-gamma-mediated macrophage killing. *J.Immunol 175*(11): 7512-7518. doi: 10.4049/jimmunol.175.11.7512

Lichun Ma and Jia Zheng (2018). Single-cell gene expression analysis reveals β-cell dysfunction and deficit mechanisms in type 2 diabetes. BMC Bioinformatics 19: 515.

Lister PD, Wolter DJ, Hanson ND (2009). Antibacterial-resistant Pseudomonas aeruginosa: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. Clinical microbiology reviews. 22(4):582-610.

Lukashin AV and Borodovsky M (1998). GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res.,* **26**, 1107-1115.

Madera M, Gough J. (2002). A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Research 30:4321-4328.

Mamistuka H (1996). A learning method of hidden Markov models for sequence discrimination. *J. Comput. Biol.,* **3**, 361-373.

Mansour A (2009). Phylip and Phylogentics.

Mark WS, Craig W, Scott AC, Stuart BL, Robert WJ (2011). *Pseudomonas* genomes: diverse and adaptable, *FEMS Microbiology Reviews*, Volume 35, Issue 4, Pages 652-680.

Marra AR, Bar K, Bearman GM, Wenzel RP, Edmond MB (2006). Systemic inflammatory response syndrome in adult patients with nosocomial bloodstream infection due to Pseudomonas aeruginosa, J. Infect. Vol. 53, pp. 30–35.

Mathee K, Narasimhan G, Valdes C, Qiu X, Matewish JM, Koehrsen M, Rokas A, Yandava CN, Engels R, Zeng E, Olavarietta R, Doud M, Smith RS, Montgomery P, White JR, Godfrey PA, Kodira C, Birren B, Galagan JE, Lory S (2008). Dynamics of Pseudomonas aeruginosa genome evolution. Proc. Natl. Acad. Sci. USA Vol. 105, pp. 3100–3105.

Matsuyama BY, Krasteva PV, Baraquet C, Harwood CS, Sondermann H, Navarro MV (2016). Mechanistic insights into c-di-GMP±dependent control of the biofilm regulator FleQ from Pseudomonas aeruginosa. Proceedings of the National Academy of Sciences 113: 209-218.

McCallum SJ, Gallagher MJ, Corkill JE, Hart CA, Ledson MJ, and Walshaw MJ (2002). Spread of an epidemic *Pseudomonas aeruginosa* strain from a patient with cystic fibrosis (CF) to non-CF relatives. *Thorax* 57**:**559-560.

Mehmood MA, Sehar U, Ahmad N (2014). Use of Bioinformatics Tools in Different Spheres of Life Sciences. *J Data Mining Genomics Proteomics* 5: 158.

Metzker ML (2010). Sequencing technologies the next generation. Nature reviews Genetics. 11(1):31-46.

Mignard S, Flandrois JP (2008). A seven-gene, multilocus, genus-wide approach to the phylogeny of mycobacteria using supertrees. *International Journal of Systematic and Evolutionary Microbiology* **58,** 1432-1441.

Morton JT, Freed SD, Lee SW, Friedberg I (2015). A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins. BMC Bioinformatics 16:1_9

Muñoz-Medina JE, Sánchez-Vallejo CJ, Méndez-Tenorio A Irma E, Javier A, Andrea S, Clara ES, Joaquín G, Yu-Mei AH, César RG, Eva RG, José AD (2015). In silico identification of highly conserved epitopes of influenza A H1N1, H2N2, H3N2, and H5N1 with diagnostic and vaccination potential. Biomed Res Int 813047.

Nathwani D, Raman G, Sulham K, Gavaghan M, Menon V (2014). Clinical and economic consequences of hospital-acquired resistant and multidrug-resistant Pseudomonas aeruginosa infections: a systematic review and meta-analysis. Antimicrob Resist Infect Control. 3(1):1-16.

NCBI. https://www.ncbi.nlm.nih.gov/.

Nyangacha RM, Odongo D, Oyieke F, Ochwoto M, Korir R, Ngetich RK, Nginya G, Makwaga O, Bii C, Mwitari P and Tolo F (2017). Secondary bacterial infections and antibiotic resistance among tungiasis patients in Western, Kenya. *PLoS neglected tropical diseases*, *11*(9).

O'Brien KT, Noto JG, Nichols-O'Neill L, Lark JP (2015). Potent irreversible inhibitors of LasR quorum sensing in Pseudomonas aeruginosa. *ACS Med Chem Lett* **6**:162–7.

Okesola AO, Oni AA (2012). Occurrence of Extended-Spectrum Beta-Lactamase-Producing Pseudomonas aeruginosa Strains in South-West Nigeria. *Research Journal of Medical Sciences*. 6 (3):93–96.

Okkotsu Y, Little AS and Schurr MJ (2014). The Pseudomonas aeruginosa AlgZR two-component system coordinates multiple phenotypes. *Frontiers in cellular and infection microbiology*, *4*, 82.

Okonechnikov K, Golosova O, Fursov M and the UGENE team (2016). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28(8): 1166-1167

Olsen I (2015). Biofilm-specific antibiotic tolerance and resistance. *Eur J Clin Microbiol Infect Dis.* 34(5):877-86.

Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. Journal of Molecular Biology 284:1201-1210

Patrick, J.E. and Kearns, D.B. (2012) Swarming motility and the control of master regulators of flagellar biosynthesis. Mol. Microbiol. 83, 14–23

Pearson WR (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics, Chapter 3,* Unit3.1. doi:10.1002/0471250953.bi0301s42

Pearson WR and Lipman DJ (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA. 85*:2444-2448

Popat R, Cornforth DM, McNally L, Brown SP (2015). Collective sensing and collective responses in quorum-sensing bacteria. *J R Soc Interface* **12**:20140882-.

Ramanathan B, Jindal HM, Le CF, Gudimella R, Anwar A, Razali R, Poole-Johnson J, Manikam R, Sekaran SD (2017). Next generation sequencing reveals the antibiotic resistant variants in the genome of *Pseudomonas aeruginosa*. PLoS ONE 12(8): e0182524.

Rao, J., DiGiandomenico, A., Artamonov, M., Leitinger, N., Amin, A. R., & Goldberg, J. B. (2011). Host derived inflammatory phospholipids regulate rahU (PA0122) gene, protein, and biofilm formation in Pseudomonas aeruginosa. *Cellular immunology*, *270*(2), 95–102.

Restrepo-Montoya D, Becerra D, Carvajal-Patiño J, Mongui A, Niño L, Patarroyo M, Patarroyo M (2011). Identification of plasmodium vivax proteins with potential role in invasion using sequence redundancy reduction and profile hidden Markov models. PLoS One 6:e25189.

Rost B (1999). Twilight zone of protein sequence alignments. *Protein Eng* **12:**85-94

Roy PH, Sasha G, André L, Liam E, Simon T, Qinghu R, Robert D, Derek H, Ryan S, Kisha W, Yasmin M, Ian TP (2010). Complete genome sequence of the multiresistant taxonomic outlier Pseudomonas aeruginosa PA7. PLoS One 5(1):e8842.

Sabat AJ, Budimir A, Nashev D, Sa-Leao R, van Dijl J, Laurent F, Grundmann H, Friedrich AW (2013). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro surveillance: European communicable disease bulletin. 18(4):20380.

Scalan PD, Hall AR, Blackshields G, Friman VP, Davis MR, Goldberg JB & Buckling A (2015). Coevolution with bacteriophages drives genome-wide host evolution and constrains the acquisition of abiotic-beneficial mutations. *Molecular biology and evolution, 32*(6), 1425-1435. doi:10.1093/molbev/msv032

Schuster M, Greenberg EP (2006). A network of networks: Quorum-sensing gene regulation in Pseudomonas aeruginosa. *Int J Med Microbiol* **296**:73–81.

Segolene R, Silke S, Alain F and Sophie B (2007). Assembly of Fimbrial Structures in *Pseudomnonas aeruginosa*: Functionality and Specificity of Chaperone-Usher Machineries. *Journal of Bacteriology.* April 189 (9) 3547-3555. doi:10.1128/JB.00093-07

Sehar U, Mehmood MA, Hussain K, Nawaz S, Nadeem S, Siddique MH, Nadeem H, Gull M, Ahmad N, Sohail I, Gill SS, Majeed S (2013). Domain wise docking analyses of the modular chitin binding protein CBP50 from Bacillus thuringiensis serovar konkukian S4. Bioinformation 9: 901-907.

Seifert M, Abou-El-Ardat K, Friedrich B, Klink B, Deutsch A. (2014). Autoregressive higher-order hidden Markov models: exploiting local chromosomal dependencies in the analysis of tumor expression profiles. PLOS ONE 9:e100295

Septimus EJ, Kuper KM. (2009). Clinical Challenges in Addressing Resistance to Antimicrobial Drugs in the Twenty-First Century. Clin Pharmacol Ther. 86:336-9.

Sherlock C, Xifara T, Telfer S, Begon M. 2013. A coupled hidden Markov model for disease interactions. Journal of the Royal Statistical Society Series C, Applied Statistics 62:609_627

Sineva E, Savkina M, & Ades SE (2017). Themes and variations in gene regulation by extracytoplasmic function (ECF) sigma factors. *Current opinion in microbiology, 36,* 128-137. doi:10.1016/j.mi.2017.05.004.

Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL. 2014. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLOS ONE 9:e105067

Slonim D, Kruglyak L, Stein L and Lander E (1997). Building human genome maps with radiation hybrids. *J. Comput. Biol.*, **4**, 487-504.

Snyder LA, Loman NJ, Faraj LA, Levi K, Weinstock G, Boswell TC, Pallen MJ, Ala'Aldeen DA (2013). Epidemiological investigation of Pseudomonas aeruginosa isolates from a six-year-long hospital outbreak using high-throughput whole genome sequencing. Euro surveillance: European communicable disease bulletin. 18(42).

Soberón-Chávez G, Lépine F, Déziel E (2005). Production of rhamnolipids by Pseudomonas aeruginosa. *Appl Microbiol Biotechnol* 718–25.

Stover CK, Pham XQ, Erwin AL, Mitoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E (2000). Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen, Nature Vol. 406, pp. 959–964.

Spencer DH, Kas A, Smith EE, Raymond CK, Sims EH, Hastings M, Burns JL, Kaul R, Olson MV (2003). Whole-genome sequence variation among multiple isolates of *Pseudomonas aeruginosa*. *J Bacteriol*, 185:1316-1325.

Suriyanarayanan T, Periasamy S, Lin MH, Ishihama Y and Swarup S (2016). Flagellin FliC Phosphorylation Affects Type 2 Secretion and Biofilm Dispersal in Pseudomonas aeruginosa PAO1. *PloS one, 11*(10), e0164155.

Sunyaev SR, Rodchenkov IV, Eisenhaber F and Kuznetsov EN (1998). Analysis of the position dependent amino acid probabilities and its application to the search for remote homologues. In *RECOMB '98,* pp.258-265.

Talwalkar JS and Murray TS (2016). The Approach to *Pseudomonas aeruginosa* in Cystic Fibrosis. *Clin Chest Med.* 37(1):69-81.

Tonhosolo R, D'Alexandri FL, de Rosso VV, Gazarini ML, Matsumura MY, Peres VJ, Merino EF, Carlton JM, Wunderlich G, Mercadante AZ, Kimura EA, Katzin AM (2009). Carotenoid biosynthesis in intraerythrocytic stages of *Plasmodium falciparum*. J Biol Chem 284:9974–9985.

Valdar WSJ (2002). Scoring residue conservation. *Proteins,* 48:227-241.

Valentini M, Filloux A (2016). Biofilms and Cyclic di-GMP (c-di-GMP) Signaling: Lessons from Pseudomonas aeruginosa and Other Bacteria. Journal of Biological Chemistry 291: 12547-12555

Volgyi A, Zalan A, Svetnik E and Pamjav H (2009). Hungarian population data for 11 Y-STR and 49 Y-SNP markers. *Forensic Science International: Genetics* **3(2)**: 27-28

Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH (2015). antiSMASH 3.0_a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research 43:W237_W243

Wiehlmann L, Cramer N, and Tümmler B (2015). Habitat-associated skew of clone abundance in the *Pseudomonas aeruginosa* population. *Environmental Microbiology Reports* 7(6):955-60.

Wilhelm S, Gdynia A, Tielen P, Rosenau F and Jaeger KE (2007). The autotransporter esterase EstA of *Pseudomonas aeruginosa* is required for rhamnolipid peoduction, cell motility, and biofilm formation. *Journal of Bacteriology, 189*(18), 6695-6703. doi:10.1128/JB.00023-07

Williams P, Cámara M, Ca M. (2009). Quorum sensing and environmental adaptation in Pseudomonas aeruginosa: a tale of regulatory networks and multifunctional signal molecules. *Curr Opin Microbiol* **12**:182–91.

Winstanley C, Langille MG, Fothergill JL, Kukavica-Ibrulj I, Paradis-Bleau C, Sanschagrin F, Thomson NR, Winsor GL, Quail MA, Lennard N, Bignell A, Clarke L, Seeger K, Saunders D, Harris D, Parkhill J, Hancock RE, Brinkman FS, and Levesque RC(2009). Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome research*, *19*(1), 12–23.

Wongsaroj L, Saninjuk K, Romsang A, Duang-nkern J, Trinachartyanit W, Vattavaviboon P and Mongkolsuk S (2018). *Pseudomonas aeruginosa* glutathione biosynthesis genes play multiple roles in stress protection, bacterial virulence and biofilm formation. *PLoS ONE.* 13 (10) e0205815.

World Health Organization (2014). *Antimicrobial resistance: global report on surveillance 2014*. Geneva, Switzerland: WHO.

Yada T, Ishikawa M, Tanaka H and Asai K (1996). Extraction of hidden Markov model representations of signal patterns in DNA sequences. *Pac. Symp. Biocomput.*, World Scientific, Singapore, pp. 686-696.

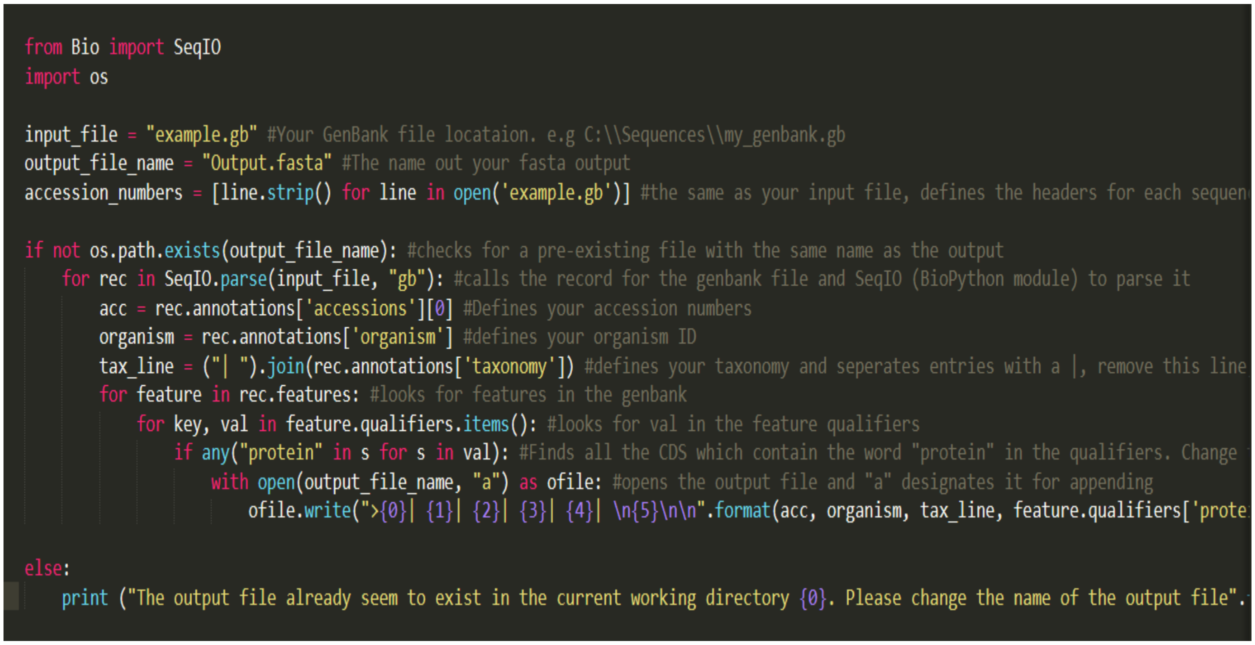
Yan J, Deforet M, Boyle KE, Rahman R, Liang R, Okegbe C, Lars EP, Weigang Qiu,Joao B (2017). Bow-tie signaling in c-di-GMP: Machine learning in a simple biochemical network. PLoS Comput Biol 13(8):e1005677.

Yin Y, Damron FH, Withers TR, Pritchett CL, Wang X, Schurr MR & Yu HD (2013). Expression of mucoid induction factor MucE is dependent upon the alternate sigma factor AlgU in *Pseudomonas aeruginosa*. *BMC Microbiol* **13**, 232. doi:10.1186/1471-2180-13-232

Yoon BJ (2009). Hidden Markov models and their applications in biological sequence analysis. Curr Genom 10:402–415.

van Ditmarsch D, Boyle KE, Sakhtah H, Oyler JE, Nadell CD, Déziel É, Dietrich LE, Xavier JB (2013). Convergent Evolution of Hyperswarming Leads to Impaired Biofilm Formation in Pathogenic Bacteria. Cell Reports 4: 697-708

APPENDICES



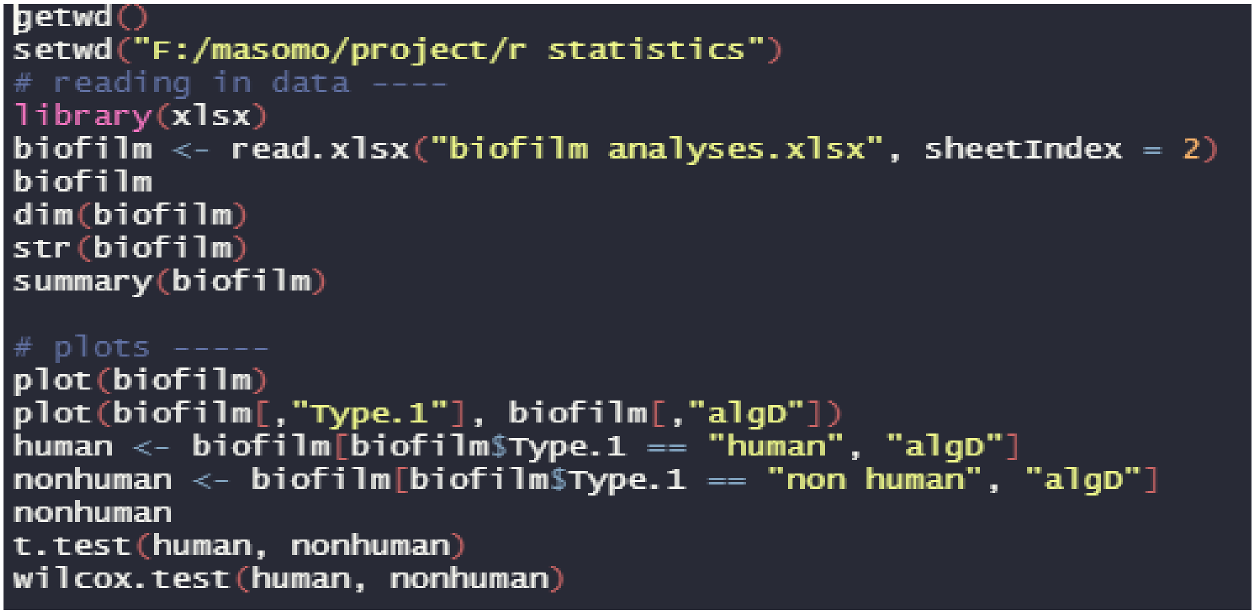
```python
from Bio import SeqIO
import os

input_file = "example.gb" #Your GenBank file locataion. e.g C:\\Sequences\\my_genbank.gb
output_file_name = "Output.fasta" #The name out your fasta output
accession_numbers = [line.strip() for line in open('example.gb')] #the same as your input file, defines the headers for each sequen

if not os.path.exists(output_file_name): #checks for a pre-existing file with the same name as the output
    for rec in SeqIO.parse(input_file, "gb"): #calls the record for the genbank file and SeqIO (BioPython module) to parse it
        acc = rec.annotations['accessions'][0] #Defines your accession numbers
        organism = rec.annotations['organism'] #defines your organism ID
        tax_line = ("| ").join(rec.annotations['taxonomy']) #defines your taxonomy and seperates entries with a |, remove this line
        for feature in rec.features: #looks for features in the genbank
            for key, val in feature.qualifiers.items(): #looks for val in the feature qualifiers
                if any("protein" in s for s in val): #Finds all the CDS which contain the word "protein" in the qualifiers. Change
                    with open(output_file_name, "a") as ofile: #opens the output file and "a" designates it for appending
                        ofile.write(">{0}| {1}| {2}| {3}| {4}| \n{5}\n\n".format(acc, organism, tax_line, feature.qualifiers['prote

else:
    print ("The output file already seem to exist in the current working directory {0}. Please change the name of the output file".
```

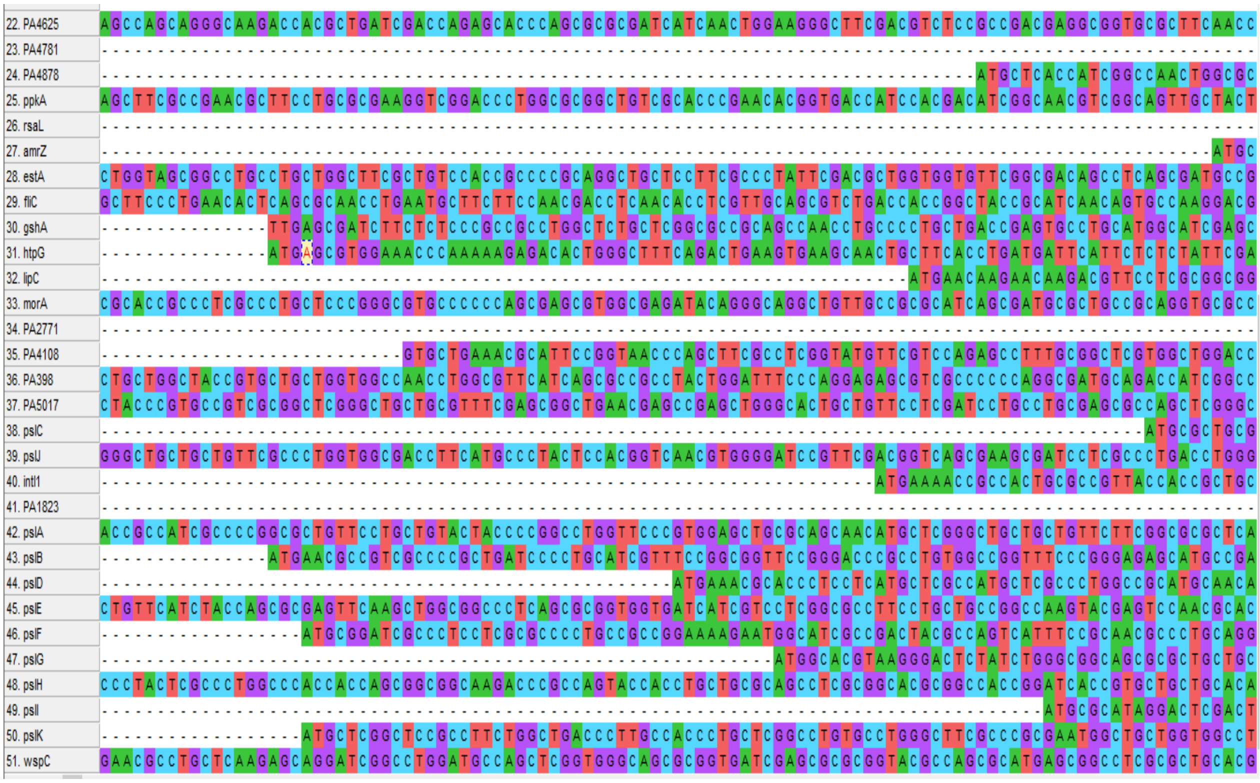
Appendix I. Custom python script used to retrieve biofilm formation genes from the genomes of *P. aeruginosa*.



```r
getwd()
setwd("F:/masomo/project/r statistics")
# reading in data ----
library(xlsx)
biofilm <- read.xlsx("biofilm analyses.xlsx", sheetIndex = 2)
biofilm
dim(biofilm)
str(biofilm)
summary(biofilm)

# plots -----
plot(biofilm)
plot(biofilm[,"Type.1"], biofilm[,"algD"])
human <- biofilm[biofilm$Type.1 == "human", "algD"]
nonhuman <- biofilm[biofilm$Type.1 == "non human", "algD"]
nonhuman
t.test(human, nonhuman)
wilcox.test(human, nonhuman)
```

Appendix II. Custom R-script used to perform the Wilcoxon-Signed rank test



```
22. PA4625  AGCCAGCAGGGCAAGACCACGCTGATCGACCAGAGCACCCAGCGCGCGATCATCAACTGGAAGGGCTTCGACGTCTCCGCCGACGAGGCGGTGCGCTTCAACC
23. PA4781  -------------------------------------------------------------------------------------------------------
24. PA4878  ------------------------------------------------------------------------------ATGCTCACCATCGGCCAACTGGCGC
25. ppkA    AGCTTCGCCGAACGCTTCCTGCGCGAAGGTCGGACCCTGGCGCGGCTGTCGCACCCGAACACGGTGACCATCCACGACATCGGCAACGTCGGCAGTTGCTACT
26. rsaL    -------------------------------------------------------------------------------------------------------
27. amrZ    ---------------------------------------------------------------------------------------------------ATGC
28. estA    CTGGTAGCGGCCTGCCTGCTGGCTTCGCTGTCCACCGCCCCGCAGGCTGCTCCTTCGCCCTATTCGACGCTGGTGGTGTTCGGCGACAGCCTCAGCGATGCCG
29. fliC    GCTTCCCTGAACACTCAGCGCAACCTGAATGCTTCTTCCAACGACCTCAACACCTCGTTGCAGCGTCTGACCACCGGCTACCGCATCAACAGTGCCAAGGACG
30. gshA    ---------------TTGAGCGATCTTCTCTCCCGCCGCCTGGCTCTGCTCGGCGCCGCAGCCAACCTGCCCCTGCTGACCGAGTGCCTGCATGGCATCGAGC
31. htpG    ---------------ATGAGCGTGGAAACCCAAAAAGAGACACTGGGCTTTCAGACTGAAGTGAAGCAACTGCTTCACCTGATGATTCATTCTCTCTATTCGA
32. lipC    ------------------------------------------------------------------------ATGAACAAGAACAAGACGTTCCTCGCGGCGG
33. morA    CGCACCGCCCTCGCCCTGCTCCCGGGCGTGCCCCCCAGCGAGCGTGGCGAGATACAGGGCAGGCTGTTGCCGCGCATCAGCGATGCGCTGCCGCAGGTGCGCC
34. PA2771  -------------------------------------------------------------------------------------------------------
35. PA4108  ---------------------------GTGCTGAAACGCATTCCGGTAACCCAGCTTCGCCTCGGTATGTTCGTCCAGAGCCTTTGCGGCTCGTGGCTGGACC
36. PA398   CTGCTGGCTACCGTGCTGCTGGTGGCCAACCTGGCGTTCATCAGCGCCGCCTACTGGATTTCCCAGGAGAGCGTCGCCCCCCAGGCGATGCAGACCATCGGCC
37. PA5017  CTACCCGTGCCGTCGCGGCTCGGGCTGCTGCGTTTCGAGCGGCTGAACGAGCCGAGCTGGGCACTGCTGTTCCTCGATCCTGCCTGCGAGCGCCAGCTCGGGC
38. pslC    ---------------------------------------------------------------------------------------------ATGCGCTGCG
39. pslJ    GGGCTGCTGCTGTTCGCCCTGGTGGCGACCTTCATGCCCTACTCCACGGTCAACGTGGGGATCCGTTCGACGGTCAGCGAAGCGATCCTCGCCCTGACCTGGG
40. intI1   ---------------------------------------------------------------------ATGAAAACCGCCACTGCGCCGTTACCACCGCTGC
41. PA1823  -------------------------------------------------------------------------------------------------------
42. pslA    ACCGCCATCGCCCCGGCGCTGTTCCTGCTGTACTACCCCGGCCTGGTTCCCGTGGAGCTGCGCAGCAACATGCTCGGGCTGCTGCTGTTCTTCGGCGCGCTCA
43. pslB    ---------------ATGAACGCCGTCGCCCCGCTGATCCCCTGCATCGTTTCCGGCGGTTCCGGGACCCGCCTGTGGCCGGTTTCCCGGGAGAGCATGCCGA
44. pslD    ---------------------------------------------------ATGAAACGCACCCTCCTCATGCTCGCCATGCTCGCCCTGGCCGCATGCAACA
45. pslE    CTGTTCATCTACCAGCGCGAGTTCAAGCTGGCGGCCCTCAGCGCGGTGGTGATCATCGTCCTCGGCGCCTTCCTGCTGCCGGCCAAGTACGAGTCCAACGCAC
46. pslF    ------------------ATGCGGATCGCCCTCCTCGCGCCCCTGCCGCCGGAAAAGAATGGCATCGCCGACTACGCCAGTCATTTCCGCAACGCCCTGCAGG
47. pslG    ------------------------------------------------------------ATGGCACGTAAGGGACTCTATCTGGGCGGCAGCGCGCTGCTGC
48. pslH    CCCTACTCGCCCTGGCCCACCACCAGCGGCGGCAAGACCCGCCAGTACCACCTGCTGCGCAGCCTCGCGGCACGCGGCCACCGGATCACCGTGCTGCTGCACA
49. pslI    ------------------------------------------------------------------------------------ATGCGCATAGGACTCGACT
50. pslK    ------------------ATGCTCGGCTCCGCCTTCTGGCTGACCCTTGCCACCCTGCTCGGCCTGTGCCTGGGCTTCGCCCGCGAATGGCTGCTGGTGGCCT
51. wspC    GAACGCCTGCTCAAGAGCAGGATCGGCCTGGATGCCAGCTCGGTGGGCAGCGCGGTGATCGAGCGCGCGGTACGCCAGCGCATGAGCGGCCTCGCGCTGCACG
```

Appendix III Representative multiple sequence alignment of the 51 biofilm formation genes indicating extensive gaps and few regions of similarity