UNIVERSITY OF LIVERPOOL

Whole genome Sequencing and Comparative Analysis of *Streptococcus* species from Cystic Fibrosis patients infected with *Pseudomonas aeruginosa* Liverpool Epidemic Strain

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

By

Yongming.Li

June 2023

Declaration

* The wet lab results presented in this thesis were obtained from experiments carried out by Dr Fothergill Jo's lab members in University of Liverpool. I played a major role in data analysis and interpretation are entirely by my own work. Any contributions from colleagues in the collaboration, such as diagrams or tables, are explicitly referenced in the text.

I am aware of and understand the university's policy on plagiarism and I certify that this thesis is my own work, expect where indicated by referencing, and the work presented in it has not been submitted in support of another degree or qualification from this or any other university or institute of learning.

Abstract

* Cystic fibrosis (CF) is a genetically inherited disease requiring complex life-long medical treatment. people with CF have a shortened life expectancy. Although many organs in the body are affected, most morbidity and mortality is due to damage to the lungs caused by chronic, long-term bacterial infections requiring extensive treatment with antibiotics. Infections with traditional pathogens such as *Pseudomonas aeruginosa* and *Staphylococcus aureus* are acquired throughout childhood and early adulthood. Chronic lower airway infection with the bacteria *P. aeruginosa* remains the commonest cause of death for children with CF.

Next-generation sequencing techniques were employed to study CF lung bacterial pathogens and microbial communities, which gives great insights into the complex ecosystem. These studies revealed that the complex microbiological ecosystem not only includes recognized pathogens, such as *Pseudomonas aeruginosa* and *Staphylococcus aureus*, but also less recognized bacteria such as oral commensal *streptococci*. It is becoming increasingly clear that interactions between the bacterial pathogens and the microbial community in the CF lungs are crucial to understanding pathogenesis, antimicrobial resistance, and disease progression.

In our study, we isolated 60 *Streptococcus* strains from sputum samples collected from 5 CF patients during their visit to hospital. Samples were taken during 5 different time points including two stable time points with three exacerbations intervals. Interestingly these five patients were all infected with the *P. aeruginosa* Liverpool epidemic strain (LES) which is said to be the most prevalent strain of *P. aeruginosa*. We identified the *Streptococcus* species of our strains and generated 40 reference streptococci genomes for further CF studies especially with the presence of *P. aeruginosa* using next generation sequencing, followed by *de novo* assembly and species identification. A total of 14 strains were identified as novel or new *Streptococcus* species. The virulence factors in these strains which may contribute to the interaction with *P. aeruginosa* or directly with the host were also identified and compared. We also using genomic comparative analysis methods to compare the genomic diversity of these strains.

Wet lab characterization of these novel species strains was performed. Strains from the same species show similar results in the utilization of certain carbon sources and the sensitivity of different chemicals. All strains were observed as classical *Streptococcus* strains in SEM. All strains were predicted as commensal *Streptococcus* strains as these strains showed no infection ability in mice model.

Acknowledgements

Throughout my Phd study I had received a great deal of support and assistance.

I would like to thank the RDF grands from Xi’an-Jiaotong Liverpool University for supportting my Phd projects.

I would first like to thank my supervisors, Professor Xin Liu, Professor Jo Forthergill and Professor Siew Woh Choo, whose expertises were invaluable in formulating the research questions, methodology and evaluating the progress of the projects. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to thank my families. I stayed with my grandparents at my early age, they teach me the values like be friendly and be patient to other people. For my parents, they provide all the fundamentals to continue my studies until now without any complaining. For my parents in law, thanks for their trust to allow their daughter to marry me. For my brother, he is the mirror to see my weakness and remind me to be a good example in life. Most importantly, for my wife Yuan and son Yu, they are the spirit harbors to put my happiness and sadness.

Publications and Author’s Contributions

N/A

Table of Contents

List of Abbreviations

*

CF: cystic fibrosis

CFTR: Cystic fibrosis transmembrane conductance regulator

pwCF: patient with cystic fibrosis(Ratjen, Bell et al. 2015)

DDH: DNA-DNA hybridization

dDDH: digital or in-silicon DNA-DNA hybridization

OGRI: the overall genome relatedness indices

SEM: Scanning electron microscope

List of Figures

*

List of Tables

*

Chapter 1 Introduction

1.1 Cystic fibrosis: history, genetic nature

Before the 20th century, long before cystic fibrosis (CF) was recognized as a pathological entity, children with a salty taste were thought to be bewitched and recorded in popular folklores. This possible relationship with CF pertaining to a salty taste was recorded in many documents from Germany (the 15th century), Spain (1606) and other European literatures (the 17th century) (Quinton 1999). Early macroscopic and pathological description of CF causing death were contributed by autopsies, showing an enlarged, hardened and gleaming white pancreas (cirrhotic pancreas), separating the cause of disease from superstition (Rosenstein, Langbaum et al. 1984). Patients with CF were also recorded with childhood diseases including diarrhea, dystrophy, weakness, swelling in the hands and feet, a distended abdomen and a hardened pancreas (Sjogren 2006). Later, Meconium Ileus was detected in many cases using in situ techniques to examine the viscera (Sathe and Houwen 2017).

In the 20th century, before the1960s, the fundamental nature of the cause of CF remained a mystery. In terms of clinical presentation, CF was identified as a systemic illness, affecting many organs, causing pancreatic illnesses, bronchiectasis (potentially associated with vitamin A deficiency in infants), celiac disease, respiratory illness, and digestive symptoms (Ratjen, Bell et al. 2015). Sweat testing was developed in 1953 and is still a commonly used diagnostic test for CF (Servidoni, Gomez et al. 2017), moreover genetic screening is available in some countries including UK (Scotet, L'Hostis et al. 2020). Shwachman and Kulczycki stablished a ranking system of clinical severity that is still used today (Shwachman and Kulczycki 1958). They found that around 15% of patients had normal pancreatic function. During this time, advances in medicines such as the development of antibiotics and enteric-coated pancreatic enzymes contributed to the improved survival of CF patients (Nolan, Moivor et al. 1982, Taccetti, Francalanci et al. 2021).

CF is an autosomal recessive inherited disorder that is inherited in a Mendelian manner and related to the loss of function of Cystic fibrosis transmembrane conductance regulator (CFTR) gene. CFTR was first identified in 1946, and subsequently pedigree analysis of 47 families with CF was performed by Andersen and Hodges (Andersen and Hodges 1946). The inheritance pattern was further studied by Fred Allen in 1956 and Connealy in 1973 (Allen, Dooley et al. 1956, Conneally, Merritt et al. 1973). Although CF is a monogenic disease, there were rare cases of CF related to genetic heterogeneity. Due to the loss of function of CFTR protein, the Poor functioning of the epithelial tissue was commonly observed in all CF organs. Specially, the inability of the absorption of chloride ion through the tissue and the imbalance of sodium absorption. Leading to excessive retention of salts in sweat. This is the main issue in people with CF, the defective reabsorption of chloride at the cellular level which has a profound effect on epithelial cells. The channel is formed by CFTR protein and mutations in the gene encoding this leads to poor functioning of the chloride channel, however the exact defect is dependent on the type of mutation present. Thus, the loss of function or dysfunction of the CFTR gene is the initial cause of CF (Tsui, Buchwald et al. 1985, Kerem, Corey et al. 1989, Rommens, Iannuzzi et al. 1989, Berger, Anderson et al. 1991). This leads to thick mucus in the lungs, pancreas, liver, intestine, and reproductive tract. There are also other alterations to the local environment, particularly in the lungs. This altered environment can lead to bacterial colonization, recurrent infections, chronic inflammation and irreversible damage of the epithelium. Ultimately, the increased mortality in people with CF is often associated with chronic lung infections (Elborn 2016).

The incidence of CF in newborns of Caucasians (European heritage) was estimated to be 1 in 3000 from retrospective population-wide studies. Similar statistics were observed in studies from the United States, the United Kingdom, the majority of European and Australia (Conneally, Merritt et al. 1973, Levison 1980, Hammond, Abman et al. 1991, Pollitt, Dalton et al. 1997, Scotet, de Braekeleer et al. 2000, Farrell, Kosorok et al. 2001, Saiman, Chen et al. 2001, Assael, Castellani et al. 2002). However, significantly high frequencies of CF have been observed in some ethnic groups, presumably caused by genetic drift or founder effect (Super 1975, Fujiwara, Morgan et al. 1989, Rozen, Schwartz et al. 1990). For example, the Hutterites in Alberta, Afrikaners and French Canadians. The incidence of heterogenetic carriers of CF mutations is estimated to be 1 in 30 in European populations based on the common incidence of 1 in 3000 live births. There is no common agreement to explain why there is a higher frequency of CFTR mutations in Caucasian populations. However different factors have been suggested, including the coexistence of multiple CF loci (Conneally et al., 1973), high variant rate (Goodman and Reed 1952), genetic drift (Wright and Morton 1968), founder effect (Klinger 1983), sex ratio (Williams, Davies et al. 1993), segregation distortion (Williams, Davies et al. 1993), and heterozygote advantage (Bobadilla, Macek et al. 2002). So far, more than 2000 different mutations in the CFTR gene have been recorded (http://www.genet.sickkids.on.ca/cftr/Home.html).

Over the last several decades, a continuous and significant improvement in patient survival age has been observed. This is because of a multitude of scientific and standardized interventions. From the very early diagnosis before birth, the improved airway clearance of mucus, treatments controlling inflammation and bacterial infections to recently CFTR modulator therapies (Pettit and Fellner 2014). But only four CFTR modulators were in market to treat people with certain CFTR mutations (Jia and Taylor-Cousar 2023). The prevalence of cultured pathogens has also changed with available treatment methods. One example is the dramatic drop rate of the chronic infection with *P. aeruginosa* due to the widely adopted early eradication for *P. aeruginosa* in CF patients at the time of the first detection (Hansen, Pressler et al. 2008). Similarly, the prevalence of *Burkholderia cenocepacia* was decreased too (Scoffone, Chiarelli et al. 2017). While other cultured pathogens Methicillin-resistant Staphylococcus aureus (Muhlebach 2017), *Stenotrophomonas maltophilia* (Amin, Jahnke et al. 2020), *Achromobacter* spp. (Gabrielaite, Nielsen et al. 2021) and *Aspergillus* spp. (Burgel, Paugam et al. 2016) were all similarly observed with the increasing prevalence over time (Gavillet, Hatfield et al. 2022).

1.2 The respiratory microbiome in people with cystic fibrosis (pwCF)

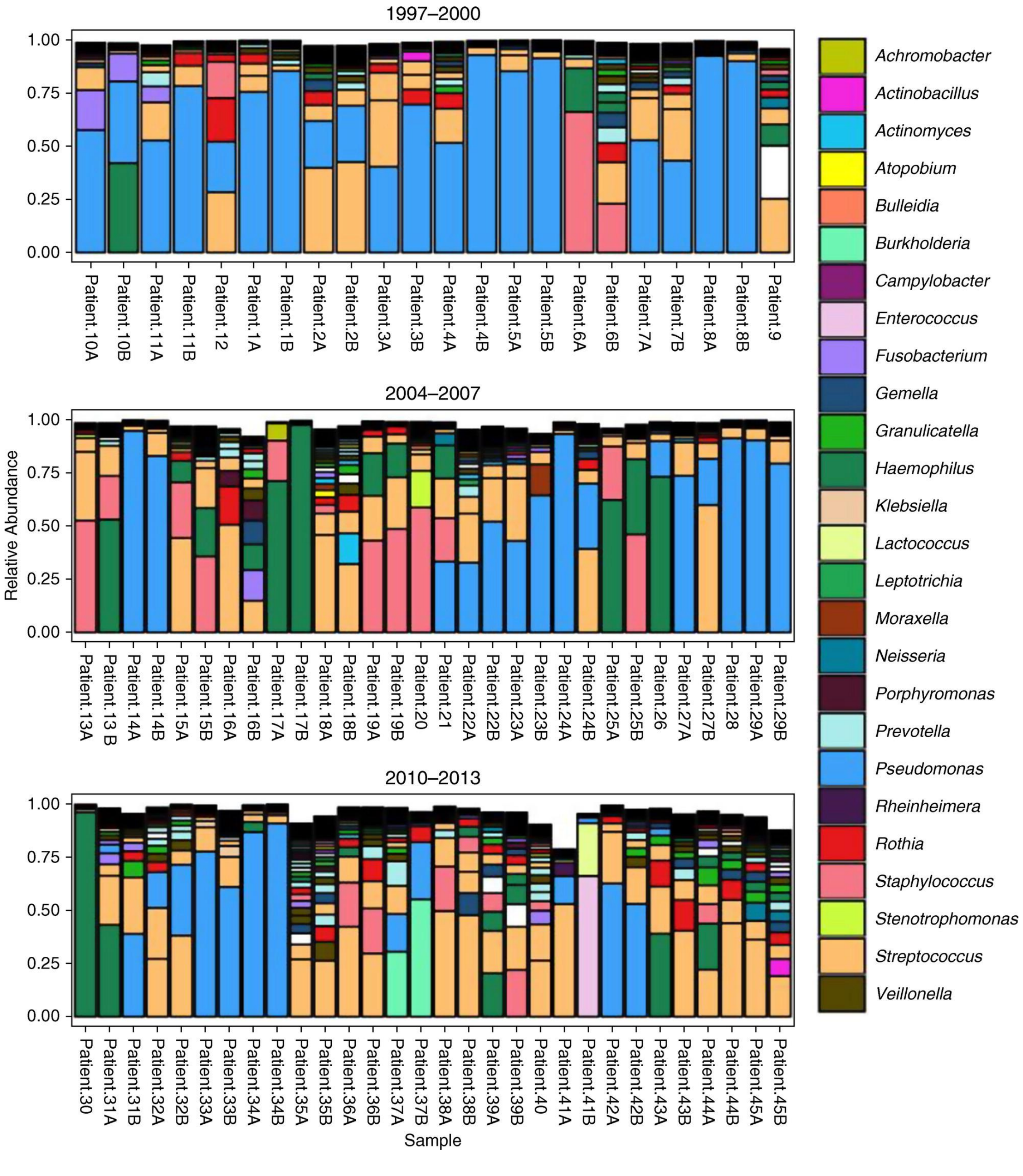
The polymicrobial nature of the CF airways has been observed in many studies (Bazett, Honeyman et al. 2015, O'Toole 2018, Coffey, Nielsen et al. 2019, Vandeplassche, Sass et al. 2019, Vandeplassche, Tavernier et al. 2019, Cuthbertson, Walker et al. 2020, Francoise and Hery-Arnaud 2020, Voronina, Ryzhova et al. 2020, van Dorst, Tam et al. 2022). The most common genera identified were *Streptococcus, Prevotella*, *Veillonella*, *Rothia*, *Actinomyces*, *Gemella*, *Granulicatella*, and *Fusobacterium* (Surette 2014). Alterations in CF-associated respiratory microbial communities have been observed and are strongly associated with age (Coburn, Wang et al. 2015). A greater diversity of organisms was associated with less disease burden. Pathogens like *P. aeruginosa* become dominant in communities in the later stage of airway disease (Fodor, Klem et al. 2012). There may be many reasons for this dominance including the production of toxic virulence factors, metabolic versatility due to a large genome and the cumulative antibiotic usage during treatment of pulmonary exacerbations which may select for resistant microorganisms (Lipuma 2010, Zhao, Schloss et al. 2012).

The study of microbiomes in CF was developed based on the availability of recent techniques for single bacteria study to a whole combination of all bacteria strains in one system. Ranging from molecular methods for cultured bacterial colonies (Pattison, Rogers et al. 2013), 16S rRNA profiling (Lucas, Yang et al. 2018), whole genome sequencing and shotgun metagenomics (Bacci, Taccetti et al. 2020) with decreased prices. Studies mainly focus on CF patients with chronic lung infection (Tuchman, Schwartz et al. 2010, Towns and Bell 2011). The main cultured pathogens in these pwCF were *Pseudomonas aeruginosa*, methicillin-resistant *Staphylococcus aureus* (MRSA), methicillin-sensitive *Staphylococcus aureus* (MSSA), *Haemophilus influenzae*, *Burkholderia cepacia* complex (BCC), *Achromobacter xylosoxidans*, *Serratia marcescens* and *Stenotrophomonas maltophilia*.

Lung disease is the most important predictor of the bacterial diversity (Turcios 2020). In sequencing studies that investigate bacterial microbiome diversity, the 16 rRNA gene is sequenced and clustering leads to the identification of operational taxonomic units (OTUs). The most abundant OTUs were *Pseudomonas*, *Streptococcus*, *Haemophilus*, *Staphylococcus*, *Prevotella*, *Rothia*, *Veillonella*, *Gemella*, and *Fusobacterium*. *Pseudomonas* strains were the most abundant taxa in a microbiome stratification and had an essential impact in the future outcomes of treatments (Rogers, Bruce et al. 2014). The number of samples with *Pseudomonas* was reduced rapidly, 16 samples collected in 1997 to 2000, 12 samples in 2004 to 2007, and 7 patients in 2010 to 2013 respectively (Figure 1.1) (Acosta, Heirali et al. 2018). But the abundance of *P. aeruginosa* in all microbiome taxa was not different in patients from three cohorts. These observations suggested that the assessment of the relative abundance of pathogens would give more information into disease progression than culture status.

It has been suggested that the reduction in diversity of the lung microbiome over time and the increasing dominance of *P. aeruginosa* is link to antibiotic treatment. Antibiotics are used in 3 scenarios for pwCF: eradication, exacerbation and chronic suppressive (maintenance) therapy. *P. aeruginosa* as a lung pathogen has been found to be closely related to the declining of lung function (Harun, Wainwright et al. 2016, Langton Hewer and Smyth 2017). Microbial interactions may also play a role in this process. Many studies have focused on interactions and co-occurrence of *P. aeruginosa* and *Staphylococcus. aureus* (Hoffman, Deziel et al. 2006, Fugere, Lalonde Seguin et al. 2014). However, it should be emphasized that *Streptococcus* and *Pseudomonas* co-occured in many samples (Figure 1.1) (Acosta, Heirali et al. 2018). In these samples, *Pseudomonas* is the dominant genus however *Streptococcus* genus often has the second-high abundance suggesting that *Streptococci* should be understood more in the context of CF.

Figure 1-1: Comparison of the cystic fibrosis related microbiome in genus level from the three successive cohorts.
The three panels (*upper*, 1997 to 2000; *middle* 2004 to 2007 and *lower* 2010 to 2013) show the relative abundance of microbiome in cystic fibrosis patients. The colored boxes represent the different bacteria genus with at least 0.05% abundance of the total operational taxonomic units (OTUs). Taxa not identified at the genus level were shown in white boxes. To simplify the analysis. Some taxa with the very low OTUs were compressed and represented in black boxes. (Cited from: (Acosta, Heirali et al. 2018))



1.3 The role of *Streptococcus* spp. in PwCF

Microbiota can be identified throughout the upper and lower airways. In pwCF, the focus of microbiota studies revolves around the lower respiratory tract, in particular the thick, dehydrated sputum. The airway-surface liquid is dehydrated and therefore hard to clear by the mucociliary clearance. Sputum samples are the most common sample used in microbiome studies. Streptococci are a core component in the CF lung microbiome (Scott and O'Toole 2019). However, historically, Streptococci were not a focus in clinical diagnostic laboratories (Sibley, Grinwis et al. 2010), or were thought to be oropharyngeal contaminants in expectoration (Sibley, Rabin et al. 2006). Therefore, they have been relatively ignored in this niche.

Nevertheless, frequent culturing and identification of Streptococci through 16S rRNA gene sequencing from thousands of sputum and lavage samples suggested their persistent presence (Sibley, Parkins et al. 2008, Filkins, Hampton et al. 2012). Moreover, Streptococci were identified from the lower airway in pwCF by collecting and analyzing multiple protected brush samples using a bronchoscope to reduce the risk of oral flora contamination (Hogan, Willger et al. 2016). One explanation for the limited presence/reporting of Streptococci in clinical diagnostics would be the specialized medium required for selecting and culturing (Vandeplassche, Coenye et al. 2017, Zachariah, Ryan et al. 2018). These media (and growth conditions) are not routinely used on sputum samples from pwCF.

The close relationships between lung microbes and oral microbiota can in part be explained by the proximity of the lower airway and the oral cavity (Whiteson, Bailey et al. 2014, Dickson and Huffnagle 2015, Dickson, Erb-Downward et al. 2016). However, in pwCF, the role of *Streptococci* in the respiratory tract is poorly understood. The *Streptococcus milleri* group (SMG) has been associated with exacerbation of lung disease in CF (Parkins, Sibley et al. 2008, Sibley, Parkins et al. 2008, Sibley, Grinwis et al. 2010). Group A *Streptococcus* has been found at a low level (4.7%) in the sputum of pwCF, and its presence may also be associated with lung exacerbation (Dennis, Coats et al. 2018). Other known pathogenic *Streptococci* such as *S. pneumoniae* have been found in 20% of oropharyngeal swabs in pwCF, and an unusual mucoid phenotype with increased biofilm and pathogenicity *in vivo* has also been identified (Esposito, Colombo et al. 2016). Together, these studies highlight a potential role for *Streptococci* in the respiratory tract of pwCF.

Streptococci may also alter the pathogenicity of known pathogens in this niche. Increased production of virulence factors by *P. aeruginosa* has been demonstrated in the presence of certain *Streptococcus* species (Whiley, Sheikh et al. 2014). Some studies in genus level have shown that *Streptococci* can also be associated with less severe lung disease (Coburn, Wang et al. 2015, Acosta, Heirali et al. 2018). Therefore, the role of *Streptococci* in this niche is likely varied. It is becoming increasingly clear that interactions between the bacterial pathogens and the microbial community in the CF lungs are crucial to understanding pathogenesis, antimicrobial resistance, and disease progression (Peters, Jabra-Rizk et al. 2012).

1.4 Genome sequencing and assembly

The double helix structure of DNA was reported by Watson and Crick in 1953 (Watson and Crick 1953). The genetic information stored in DNA was ensured by the irregular position of bases along the chain and the unrestricted bases on a single chain. This also emphasized the requirement and importance for the determination of exact sequence of bases along the chain. In a genetic study, studying the genome is the starting point. In prokaryotes, the complete genetic information in nucleotides is stored in their genomes (Loman and Pallen 2015). To determine the genome sequence, a variety of sequencing platforms are available. They are different in efficiency, accuracy, throughput, sequencing speed and cost.

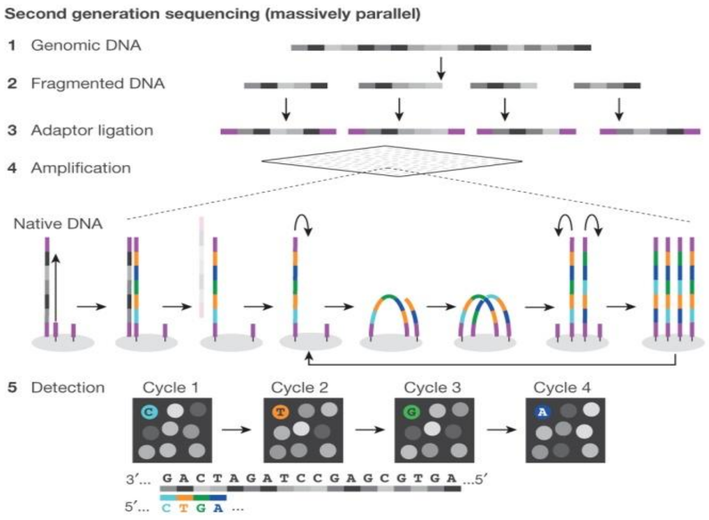
1.4.1 The next-generation sequencing (NGS)

In the 2000s, many companies participated in the racing of genome sequencing with different techniques called NGS, including 454, Solexa, Illumina, Agencourt, Complete Genomics, Applied Biosystems and Ion Torrent. Later by 2014, Illumina shared over 70% of the sequencer market and more than 90% of all DNA data was produced. In 2017, NovaSeq by

Illumina can produce 3000 Gbp reads in a single run, allowing large-scale whole genome sequencing (Pervez, Hasnain et al. 2022).

All NGS techniques require a library preparation step using native or amplified DNA fragments. The prepared library is loaded on a flow cell. Followed by massive parallel sequencing reactions (Figure 1.2) (Tucker, Marra et al. 2009). Many other applications except whole genomes sequencing are available by using NGS, including whole-exome sequencing (Rabbani, Tekin et al. 2014), high-throughput RNA sequencing (RNAseq) (Stark, Grzelak et al. 2019), chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Nakato et al., 2021), and genome-wide epigenetic landscape determination (Asp, Blum et al. 2011). Start from 2007, individual human genome re-sequencing started with Venter for the affordable cost of NGS. The first patient said to be saved by DNA sequencing was Nicholas Volker, at the age of 6 for a cord transplant to treat the XIAP gene mutation (Nielsen, Rasmussen et al. 2017). More large-scale genome projects have started to launch for the lowest of the cost compared to other sequencing generations (Perez-Sepulveda, Predeus et al. 2021).

Figure 1-2: Procedures of the next generation DNA sequencing technologies (Cited from (Tucker, Marra et al. 2009).



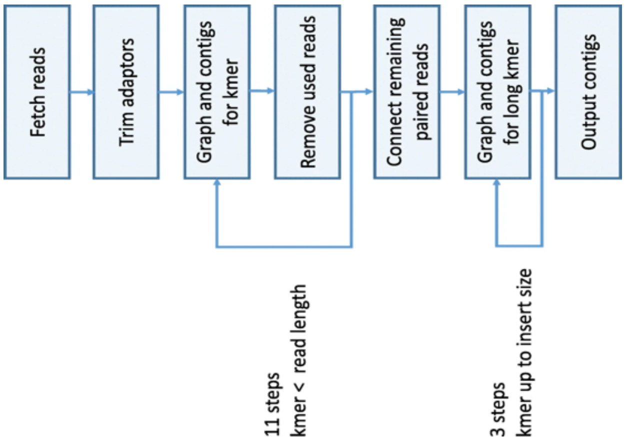
1.4.2 De novo genome assembly

Reads from random places in a genome are sequenced using a limited number of available sequencing techniques (Heather and Chain 2016). High-throughput short reads (a few hundred nucleotides) are generated by the second-generation sequencing. Longer reads with tens of thousands base pair are generated by the third-generation sequencing. Now, short reads and long reads data are used together to approach a better result.

Some WGS analysis are performed by using sequencing reads directly, for example, call the single nucleotide polymorphisms (SNPs). Other WGS analysis are based on creating an assembled genome for downstream analysis (e.g., core genome sequence typing, antimicrobial resistance gene, genomic islands) (Schurch, Arredondo-Alonso et al. 2018). Raw reads are used by an assembly software to create a representation of the actual genome. Fragmented genomes are sequenced multiple times and represented by the raw reads (Simpson and Pop 2015, Sohn and Nam 2018). Most of the genomes are represented by contigs with continuous nucleotides in assemblies.

Two steps are used by recent assemblers to deal with NGS reads. Firstly, reads are divided into suitable k-mers. Secondly, k-mers are used to produce de Bruijn graphs (Pevzner, Tang et al. 2001). SKESA based on the similar concept is used to assemble bacterial genomes in this study (Figure 1.3) (Souvorov, Agarwala et al. 2018).

Figure 1-3: Workflow of the SKESA assembler (cited from: Souvorov, Agarwala et al. 2018).



1.5 Genomic analysis

1.5.1 CheckM: assessing genome quality

Tens of thousands of draft genomes were produced as the Next-generation sequencing and computational methods continue to improve. A wide range of host-associated and environmental microorganisms, both cultivated and uncultivated, were recovered. The role of cultured microorganisms in different habitats were studied in many projects. In the Genome Encyclopedia of Bacteria and Archaea (GEBA) project, over 1250 genomes from type strains of prokaryotes associated with soil or plants were studied to understand the microbiology of soil and plants (Wu, Hugenholtz et al. 2009). Microorganisms in different locations of human were studied in The Human Microbiome Project (HMP) (Turnbaugh, Ley et al. 2007). Furthermore, single-cell genomics were used in uncultivated bacterial and archaeal lineages to complement our understanding (Kamke, Sczyrba et al. 2013). High-quality population genomes from metagenomic data were recovered in several studies (Albertsen, Hugenholtz et al. 2013). The increasing number of bacterial genomes recovered stands to improve our understanding of the microbial world. We need to automatically assess the quality of these genomes before any further downstream analysis.

Assembled genomes are subject to contaminations, and incompleteness at a level related to the experimental processes involved. A set of community-defined categories of standards for genome sequences were raised to distinguish their qualities. The standards were based on the understanding of the techniques, assemblers, and efforts to improve upon drafted genomes (Chain, Grafham et al. 2009). The assembled genomes can be divided into 6 levels, standard draft, high-quality draft, improved high-quality draft, annotation-directed improvement, noncontiguous finished and finished levels. For isolated genomes, the quality was traditionally evaluated by using assembly statistics such as N50, the number of contigs (Salzberg, Phillippy et al. 2012). For single-cell and metagenomic studies, the genome quality was estimated by the presence and absence of universal single-copy marker genes (Haroon, Skennerton et al. 2013). However, the approach is likely to be limited by the low proportions in a genome, typically accounting for only less than 10% of all genes (Sharon and Banfield 2013). Potential contamination within a genome can be reflected by the presence of multiple single-copy marker genes (Soo, Skennerton et al. 2014).

CheckM automatically and robustly estimate the level of completeness and contamination of genomes based on analyzing single-copy marker genes in genome’s inferred lineage within reference genomes (Parks, Imelfort et al. 2015, Donovan, Lynch et al. 2020). Consistently collocated marker genes in a lineage were grouped into marker sets, providing refined estimation of genome quality compared to commonly used universal or domain-level marker genes. A single copy gene in ≥97% of genomes is considered as a marker gene in CheckM. Marker genes consistently presented in nearly all genomes within a lineage (e.g., in domain level, phylum level, et al) are often organized into operons. A pair of marker genes were collocated within a lineage if the distance between each other is within 5kbp with 95% conservation in a lineage. Collocated marker genes in operon (36% on average) were grouped into a set with other marker genes to form different marker sets. A total of 5656 Trusted genomes were used to generate genome trees decorated with lineage-specific marker genes. Putative genomes for quality estimation were put into suitable positions in genomes trees. Genome quality was then estimated using specific marker gene sets.

In CheckM, the quality of draft genomes was categorized into two parts: the genome completeness level and contamination level. Draft genome completeness was classified into near, substantial, moderate and partial level with different thresholds, ≥90%, ≥70% to 90%, ≥50% to 70% and <50% separately. Similarly, the draft genome contamination levels are designated as no detectable, low, medium, high and very high, with thresholds =0%, ≤5%, 5% to ≤10%, 10% to ≤15%, and ≥15% separately. Near complete genomes with no detectable contamination or low contamination are suitable candidates for noncontiguous finished genomes after extensive additional verification.

1.5.2 Referenceseeker

Referenceseeker is an integrated, scalable, rapid and highly specific workflow for selection of proper reference genomes or closely related genomes. Newly assembled genomes are compared to their reference genomes for routine downstream in-silico analyses. It was accomplished by the integration of recent published new databases, methods and tools e.g. Refseq, average nucleotide identity (ANI) and percentage of conserved DNA (conDNA) values and Mash (Ondov, Treangen et al. 2016). The databases for five taxonomic groups bacteria, archaea, fungi, protozoa and viruses were separated built. Both genome sequences and corresponding genomic information were integrated in each database. The genome sequences were all complete, reference and representative genomes from RefSeq. The kmer profiles, related species names, NCBI Taxonomy identifiers and RefSeq assembly identifiers are also integrated in the databases.

The candidate reference genomes are identified by a two-step analysis utilizing Mash and ANI continuously. A reasonable related genomes with a Mash distance threshold of 0.1 were selected from the taxon-specific database via a kmer profile method. Then roughly selected reference genomes for input sequences were calculated and sorted based on ANI values and conDNA to generate a refined list of reference genomes. The top one in the list is the best reference genome for the input genome from the database. In this tool, the closely related genomes share an ANI value ≥90%. A candidate gnome with the best ANI and coverage would be outputted by Referenceseeker.

1.5.3 RAST: Rapid annotations

The demands for automated and reliable high through annotation are needed for the increasing number of sequenced genomes since the first complete genome released in 1995 (Fleischmann, Adams et al. 1995). The subsystem annotation approach was come up to assign expert-annotated genes into single subsystems over the complete collection of genomes. A set of predicted genes with *functional roles*, usually occur one by one as operons, participate in a specific biological process or belongs to structural complex are categorized into a *subsystem*. The SEED environment was developed for this mode of annotation, aiming to create, curate, populate and exchange of subsystems. A consistent and precise vocabulary with gene ontology terms (GO) for functional roles were produced in this environment (Overbeek, Olson et al. 2014). Similar to other projects like the KEGG (Kanehisa, Sato et al. 2016), **G**O (Ashburner, Ball et al. 2000) and MetaCyc (Karp, Riley et al. 2002), populated subsystems try to solve numerous fundamental problems in digital way.

RAST is a fully automated webserver for annotating bacterial and archaeal genomes. Predicted genes are assigned to different types, including protein-coding sequences, rRNAs sequences, tRNA sequences, CRISPR repeats, CRISPR spacers, CRISPR array. Functions of predicted genes are also assigned. The manual curated subsystems in each genome are also represented and used to build the metabolic network.

1.6 Taxonomic classification and phylogenetic analysis

1.6.1 GeneBank and RefSeq databases

Two sequence databases GeneBank and RefSeq are discussed here before the description of taxonomic analysis (O'Leary, Wright et al. 2016). All publicly available DNA sequences with annotations are continuously uploaded to GenBank, which is a NIH genetic sequence database. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. Some loci in GenBank are redundant because it is an archival database. The information in GenBank cannot be changed by third parties. Sequences records and related information for numerous organisms are provided by the NCBI Reference Sequence (RefSeq) project. Medical, functional and comparative studies are based on RefSeq projects. It is a non-redundant set of references, including chromosomes, complete genomic molecules (organelle genomes, viruses, plasmids), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins, from the GenBank database. RefSeq records can be updated by NCBI to maintain current annotation and to incorporate additional information if needed.

A bacteria strain can be grouped into Species, Genera, Families, Divisions and Kingdom levels. Interestingly, the species is the only taxonomic unit. The original definition of a species would include strains with $\geq 70\%$ DNA-DNA hybridization and $\leq 5℃\ \Delta T_m$. Phenotypes of these strains characterized should be relatively in consistency and agree with the DDH value thresholds. For the single marker gene 16S rRNA sequence in the same species, the similarity should be over 97%. But the 16S rRNA gene similarity between *S. mitis*, *S. oralis* and *S. pneumoniae* are over 99%, indicating the insufficient classification of *Streptococcus* strains into species by single marker genes (Kawamura, Hou et al. 1995, Suzuki, Seki et al. 2005).

1.6.2 Phylogenetic analysis

A phylogenetic tree is typically used to quantify the evolution relationship between different bacterial strains (Takahashi and Nei 2000). The alignment of marker gene sequences and the whole genome single nucleotide polymorphisms are two main methods for performing bacterial phylogenetic analysis. For single gene methods, sequences are firstly aligned by different tools like ClustalX (Thompson, Gibson et al. 1997), Muscle et al. Then, the neighbor-joining genetic distance method (NJ) was performed for the phylogenetic inference using tools like MEGAX (Saitou and Nei 1987, Tamura, Peterson et al. 2011). The Kimura-2-parameter is used for distance estimation (Kimura 1980). The 1000 bootstrap replications are used to check the reliability of tree topology (Felsenstein 1985). Similar to single gene sequences comparison, the whole genome SNPs are representation of differences in core genomes of a group of interested strains. For a group of bacterial strains, the pan-genome is defined as the full set of non-redundant genes in all bacterial genomes (Tettelin, Masignani et al. 2005). The pan-genome is comprised of the core and the accessory genomes. The core genome is the core complement of genes common to all members of the strains and the accessory genome is the rest of genes in this group of strains. The number of the pan-genome is based on the number of genes in the accessory genome (Medini, Donati et al. 2005).

The evolutionary relationships between different bacterial strains can be revealed by a branching diagram called a phylogenetic tree (Cerutti, Bertolotti et al. 2011). These strains are implied to evolved from a common ancestor. The tips of the tree represent groups of descendent taxa. The common ancestors for these taxa are represented by nodes on the tree. The descendants that split from the same node are called sister groups. An ancestor and all descendants of that ancestor are grouped in a clade.

1.6.3 Whole genome approach

Panseq was used in this study to determine the core regions of sequences to create files for phylogeny programs. The Core and Accessory Genome Finder (CAGF) module in Panseq is used to identify the pan-genome based on MUMmer, separate the pan-genome into core-genome and accessory-genome using the BLASTn algorithm (Altschul, Madden et al. 1997). Moreover, a SNP file containing core-genome sequence variability is generated for downstream phylogenetic applications (Altschul, Madden et al. 1997, Laing, Buchanan et al. 2010).

1.6.4 DNA-DNA hybridization (DDH)

A group of strains belonging to the same species are genetically well separated from its phylogenetic neighbours. DNA-DNA hybridization (DDH) is a pragmatic and dominant approach for species identification. A conventional DNA-DNA hybridization (DDH) similarity of $\geq 70\%$ is the gold standard for assigning two strains to the same species (Bonner, Brenner et al. 1973, Tindall, Petersen et al. 2010). Some studies even use more stringent DDH cut-off values. Ultimately, the extent of hybridization between a pair of strains is determined by their strains. With the high quality of sequenced genomes, it is not surprising that the digital DDH (dDDH, *in silico*) correlates well with and replace the conventional DDH methods. Two commonly used dDDH method, the GGDC tool showed higher correlations than the average nucleotide identity (ANI) methods (Richter and Rossello-Mora 2009). The agreement of dDDH with the conventional DDH standard is also a hallmark for the success of sequence-based methods in bacterial genome study (Stackebrandt, Frederiksen et al. 2002). Note that the pitfalls of the conventional DDH methods are overcome by the dDDH. For species delineation, the initial 95% ANI with 69% conserved DNA was corresponding to 70% DDH (Goris, Konstantinidis et al. 2007). Later, the thresholds were changed to ≥96% ANI with $\geq 90\%$ coverage for most bacterial species (Ciufo, Kannan et al. 2018). Similarly, the GGDC tool also estimated dDDH ≥70% for a species boundary using non-linear models based on a large empirical dataset (Meier-Kolthoff, Auch et al. 2013). An alternative name for dDDH, the overall genome relatedness indices (OGRI) are widely used *in silico* species classification (Chun, Oren et al. 2018).

1.7 The prediction of various Genome features

1.7.1 The restriction-modification system (RM system)

Recombination is one of the main drivers for bacterial evolution. Furthermore, the evolution of pathogens is influenced by RM systems caused recombination (Vasu and Nagaraja 2013). For example, the phylogeny of *Neisseria meningitides* is found to relate with RM systems (Budroni, Siena et al. 2011). RM systems also hinder the HGT between bacteria, limiting the spreading of mobile genetic elements or foreign DNA (Dupuis, Villion et al. 2013). Interestingly, genes in R-M systems are often carried by mobile genetic elements and disseminated through horizontal gene transfers between different bacterial species. The barriers of RM systems are not absolute as shown in many recent studies (Johnston, Cotton et al. 2019).

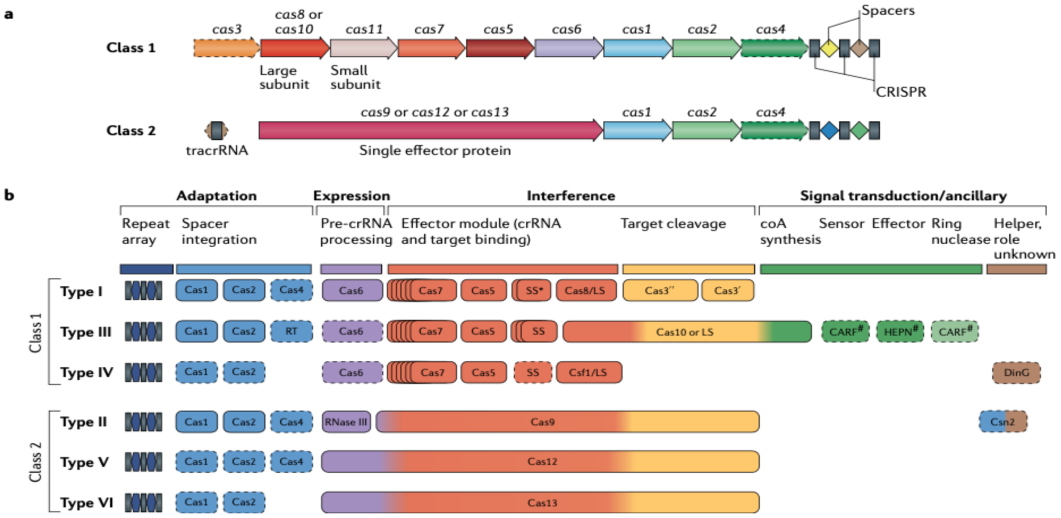
The RM systems enable bacterium to separate the methylated bacterium DNA and foreign nonmethylated DNA. In RM systems, a restriction enzyme (R) is responsible for the recognition and digestion of foreign incoming DNA and a cognate methyltransferase (M) methylate the bacterium’s own DNA to protect itself from degradation by the cognate restriction enzyme (Rodic, Blagojevic et al. 2017). The RM systems are categorized into four types, from type I to type IV, with differences in protein complexes, the subunit composition, and the functionality of the system (Vasu and Nagaraja 2013).

1.7.2 The CRISPR-Cas systems

The wide spread of adaptive immune systems in bacteria are encoded by the CRISPR-*cas* loci, resulted from the horizontal transfer of the loci and rearrangements of the locus architecture. The CRISPR-Cas modules provide sequence-specific protection against foreign DNA and RNA (Shabbir, Shabbir et al. 2019). The structure for a complete CRISPR-Cas system is a CRISPR array with flanked diverse *cas* genes. The CRISPR array is composed of short direct repeats (DR) and separated short variable DNA sequences called spacers. The CRISPR-Cas protection involves three distinct stages: adaptation, expression and interference (Rath, Amlinger et al. 2015). Fragments of foreign DNA from invading viruses and plasmids are acquired by the CRISPR array known as protospacers in adaptation stage. A target defense to against subsequent invasions by the corresponding virus or plasmid known as the sequence memory is generated by these spacers. Later, CRISPR RNAs (crRNAs), the matured form of pre-crRNAs, are transcribed from the CRISPR array during the expression stage. Finally, the nucleic acids of cognate viruses or plasmids are targeted and cleaved by the Cas proteins aided by the crRNAs during the interference stage. Besides the defense role, these systems also found to play roles in gene regulation and virulence in pathogenic bacteria. The CRISPR-*cas* loci were unambiguously separated into a relative stability of classification: 2 distinct classes, 6 types and 33 subtypes by the signature protein families and *cas* loci (Figure 1.4). (Makarova, Wolf et al. 2015, Makarova, Wolf et al. 2020).

CRISPRCasFinder is a bioinformatic tool to predict both CRISPR arrays and Cas proteins (https://crisprcas.i2bc.paris-saclay.fr/) (Couvin, Bernheim et al. 2018). It is an integrated version of CRISPRFinder and CasFinder with enhanced performance and capabilities. Moreover, the array orientations are predicted more precisely, including the presence of a leader/promoter sequence immediately before the first repeat, the existence of a diverged/truncated repeat at the 3’ end, the nature of the repeat sequence and its secondary structure, and the position of the *cas* genes clusters.

Figure 1-4: The classification of CRISPR-Cas systems and complete organization of Cas proteins in effector modules in the defense mechanism.
Both class 1 and class 2 CRISPR-Cas systems have effector modules of different Cas proteins to bind to crRNAs. The main difference of the systems in two class are the composition numbers of effector modules. For class 1 CRISPR-Cas systems, multiple Cas proteins participate in the binding. While only 1 single, multidomain Cas protein in class 2 with the similar functions to the entire *cas* proteins in class 1. The genes organization of the CRISPR-Cas systems for the two classes are illustrated in part a. Modules with different functions are organized in part b. The functional, structural and genetic organization of the 6 different represented CRISPR-Cas systems are illustrated in the scheme. Proteins are named after the current nomenclature. In the class 1 type I system, the putative small subunit highlighted with an asterisk (*) might fuse with the large subunit in several type I subtypes. The dispensable or missing components of different functions are represented in dashed outlines. The Cas6 protein is dispensable in some type I systems. But for most systems in type III, the Cas6 proteins are provided in trans by other CRISPR-cas loci. The Cas9, Cas12 and Cas13 are related to the expression and interference stages of the CRISPR-Cas immune response and represented with three colours. The CRISPR-associated Rossmann fold (CARF) and higher eukaryotes and prokaryotes nucleotide-binding (HEPN) in type III are the most common sensors and effectors in ancillary modules. Other unknown protein families highlighted in the pound (#) symbol are predicted to be involved in the same signaling pathway. LS, large subunit; SS, small subunit; tracrRNA, transactivating CRISPR RNA. (Figure modified and source from Makarova, Wolf et al. 2020)



1.7.3 Plasmids

Plasmids are autonomous replication and transformation double-stranded DNA sequences between different bacterial clones. Most of the known plasmids equipped bacteria clones with positive selection advantages or tolerance by conferring antibiotic resistance and virulence genes. Thus, the prevalence of virulent or multidrug resistant-bacteria clones are altered intensively. It is important to research both bacterial clones and the transferable plasmids in epidemiology study (Virolle, Goldlust et al. 2020).

The replication of plasmids is activated and controlled by replicons; specific regions conserved in plasmid genomes. A curated database of plasmid replicons was built and used by the PlsmidFinder Web tool for *in silico* detection of plasmids in whole-genome sequences. Possible reference plasmids are suggested by this tool for further study. The database now only used for identification of plasmids in gram-positive bacteria and *Enterobacteriaceae* species (Carattoli, Zankari et al. 2014).

1.7.4 Antibiotic resistance genes

The gold standards for measuring antimicrobial resistance (AMR) in vitro in bacteria are broth microdilution (BMD) and disc diffusion (Wheat, Connolly et al. 2001). The results generated by these typical phenotypic AST methods are hard to reproduce even following international standards (Hendriksen, Seyfarth et al. 2009). Genotypic approaches have been proposed as a valid alternative to phenotypic AST since 1991 (Courvalin 1991). With the increasing number of available bacterial genome by WGS, any known AMR gene or mutations can be detected easily (Anjum 2015). ResFinder was used in this study to not only detect genotype but also related phenotype of AMR determinants (Bortolaia, Kaas et al. 2020).

1.7.5 Genomic islands

The evolution and adaptation of bacteria in genome level via three main processes, thus mutation, recombination, and horizontal gene transfer (HGT). The dynamic nature of bacterial genome is revealed by high-throughput sequencing and related bioinformatic tools. Variations of bacterial genome sizes between different strains within Escherichia coli were easily observed two decades ago (Welch, Burland et al. 2002). Evolutionary patterns often correlate with changes in specific parts of sequenced genomes. Bacteria often acquires fragments of foreign DNA (gene clusters) with different phenotypes from outside through Horizontal gene transfer

(Schmidt and Hensel 2004). This is an important mechanism for adaptation to different environments coupled with gene loss because genome growth is limited. Thus, there is a balance between selective gene acquisition and gene loses with lower selective value (Ochman, Lawrence et al. 2000).

Genomic islands (GIs) are atypical regions of bacterial genome containing unique genes (accessory genes) acquired by horizontal gene transfer in different bacterial strains. They were classified into pathogenicity islands (PAI) with virulence factors (Hacker, Bender et al. 1990), metabolic islands (MIs) with metabolic related genes, resistance islands (RIs) with antibiotic resistance genes and symbiotic islands (SIs) with genes depend closely to the environment. Not surprisingly, the performance of each island also depends on the surrounding environments (Hacker, Blum-Oehler et al. 1997, Schmidt and Hensel 2004). All GIs have similar sizes of 10-200 kb. GIs below 10 kb are genomic islets (Hacker and Kaper 2000). GIs have GC% content and dinucleotide frequency differ from the rest of the genome, providing specific sequence compositions. These are also indicators of their presence in one genome (Juhas, van der Meer et al. 2009) and used in many prediction tools. tRNA genes located at the upstream of direct repeats of GIs are good target sites for excision. Common genes in GIs may encode conjugation related integrins, facilitator from phages, insertion elements (IS), integrases, and transposons (Gal-Mor and Finlay 2006).

HGT is an important adaptation process, receiving prepared and improved set of genes in GIs, leading to diversity and promoting the propagation of genes in bacteria (Wilson 2012). Genes in GIs perform specific and important functions are worth studying in bacteria research. For the most studied PAIs, these GIs accounts for major changes in phenotype changes in bacteria (Hacker and Carniel 2001). GIs are account for the increased distribution of virulence and antibiotic resistance factors in bacteria, leading to quicker spreading and high resistance level to various antibiotics, adding burdens to health care system (Juhas, van der Meer et al. 2009).

GIs are mosaics of genes formed by HGT, revealed by analyzing the large number of genetic sequences. Two main concepts based on characteristics are used to predict GIs for genome data. The first one is comparative analysis between genomes of close related organisms to identify variable genomic regions. The second one is analysis between sequence compositions in single genome (Lu and Leong 2016). In our study, we use IslandViewer (https://www.pathogenomics.sfu.ca/islandviewer/) for the prediction of genomic islands. IslandViewer is an integrated webserver for GIs prediction with three most accurate and complementary tools; IslandPath-DIMOB (Hsiao, Wan et al. 2003) is a tool based on nucleotide bias and presence of mobility genes, SIGI-HMM based on codon usage bias with a Hidden Markov Model approach (Waack, Keller et al. 2006) and IslandPick (Langille, Hsiao et al. 2008) based on a comparative genomics approach. Currently, there is little study on comparative analysis between genomic islands in Streptococcus species.

1.7.6 Prophages

Bacteriophages or simply phages are a group of bacterial infecting viruses. They are predicted to be 10^{31} compared to 10^{30} bacteria in the biosphere, accounting for the most abundant biological entities on Earth (Brussow, Canchaya et al. 2004). The microbial genome variation and diversity is partly contributed by phages (Fortier and Sekulovic 2013). Thus, bacteria become pathogenic, resist to antibiotic and adapt to new econiches. Bacteriophage can integrate into the host bacterial chromosome at certain insertion points through a life cycle called lysogeny. These integrated or latent phages are called prophages. In some cases, cryptic prophages are permanently integrated into the bacterial genome and becoming genetic sources for future evolution. In some cases, bacterial genomes contain 20% genomic materials from prophages and cryptic prophages (Casjens 2003). Not surprisingly, phages are closely associated with the evolution of many important bacterial pathogens (Brussow, Canchaya et al. 2004).

PHASTER (PHAge Search Tool – Enhanced Release) is used in this study for its speed and usability (Arndt, Marcu et al. 2019). This web-based tool is based on BLASTP search to match known phage sequences. The prophage detection methods include knowledge-based matric and gene function.

1.7.7 Virulence factors

Virulence factors (VFs) in bacteria are responsible for one's pathogenesis. VFs can be transferred from one strain to another via horizontal transfer. Infectious determinants can be quickly identified and comprehensively characterized from bacterial genome via WGS. In silico analysis of bacterial pathogenesis is depend on the prediction of virulence factors (VFs) responsible for the diverse clinical symptoms of pathogen infections. Up to date knowledge of VFs from various bacterial pathogens are stored in the virulence factor database (VFDB) (Chen, Yang et al. 2005). VFanalyzer was used in this study to automatically analyze potential VFs in provided draft bacterial genomes by using well-curated datasets of VFDB and a comparative pathogenomic strategy (Liu, Zheng et al. 2019).

1.8 The analysis of genetic difference between whole genomes of closely related strains

Sample genomes that are very closely related, typically over 99.9% nucleotide identity can be compared to a high-quality reference genome to identity important genetic differences between them. This often involves three steps. Firstly, each sequencing read is mapped to selected reference genomes. Secondly, genetic variations in the sample are identified by searching for differences between aligned reads and the reference genome. Lastly, genes with differences in sequences are annotated. Normally, there are three types of genetic variations to detect, the single nucleotide variants (SNVs), changes of a few nucleotides (indels) and structural variants of large chromosome sequences (SVs).

Breseq was used in this research to study diversity of strains from the same source (genome similarity over 99.9%) from mapping to annotation. It is a tool specially designed for haploid microbial-sized genomes (less than 20 Mb) to detect any key genetic change in a sample (Deatherage, Traverse et al. 2014).

1.9 Objectives

In this Phd project, we aim to

1. Sequence and assemble genomes of streptococci from pwCF infected with *P. aeruginosa* to generate useful reference genomes.

2. Perform taxonomic analysis of isolated strains.

3. Perform comparative analysis of novel *Streptococcus* species by using both dry lab and wet lab methods.

Chapter 2 Methodology

2.1 Introduction

To study *Streptococcus* strains cultured and isolated from sputum samples of pwCF, we begun by sputum sample collection, conventional bacteria strain isolation and culturing, bacterial DNA extraction, DNA library preparation, next generation DNA sequencing (Koser, Ellington et al. 2012), reads assembly, genome annotation and finally characterization of the assembled genomes. Then the phylogenetic positions of these strains were analyzed by using whole genome phylogenetic analysis. Followed by species identification using in silicon DNA-DNA hybridization methods. Furthermore, novel bacteria strains were identified and characterized by both dry and conventional experiment methods, including the prediction of various genomic features, biochemical testing, imaging of bacteria morphology, biofilm formation capacity, and finally the pathogenicity testing by two different animal models.

2.2 Bacterial isolation, culture, and genome sequencing

2.2.1 Bacterial isolation and culturing

Sputum samples were originally collected from 5 different people with cystic fibrosis (pwCF) in the UK approved by the local research ethics committee (REC reference 08/H1006/47) as described in a previous study (Mowat, Paterson et al. 2011). All pwCF were chronically infected with *Pseudomonas aeruginosa* Liverpool epidemic strain (Winstanley and Fothergill 2009). The samples used in our study were from patient 4, patient 7, patient 8, patient 9 and patient 10. A total of 12 samples were selected to represent different infection status in chronically-infected pwCF (Table 2.1) (Goss and Burns 2007). Stable samples were those taken during periods of stable infection, often during routine outpatient appointments. Exacerbations are periods of worsened infection symptoms. Although these are hard to define, the criteria for exacerbation were clinically characterized by symptoms including a 10%

drop in forced expiratory volume in 1s (FEV1), increased sputum production, discoloration, temperature measured at different parts of the body over 38℃, poor exercise tolerance et al (Rakhimova, Munder et al. 2008). Samples termed “Acute 1” were provided as the patient presented with these exacerbated symptoms but before any treatment had commenced. “Acute 2” was during intravenous antibiotic treatment and “Acute 3” was after 14 days of treatment, when exacerbation symptoms were resolved.

Following collection, sputum samples were stored at -80 degrees Celsius. To process for culture, sputum samples were thawed, homogenized with sputasol (Oxoid) at 1:1 volume ratio and incubated at room temperature for 20 minutes with shaking at 200 r.p.m. A dilution series was then prepared and plated out onto Colombia blood agar with *Streptococcus* selective supplement (Oxoid), before being incubated with microaerobic jars for 48 hours. Several colonies isolated from each sputum sample were streaked onto fresh agar. A total of 60 strains were cultured and isolated using twelve samples collected from five patients (Table 2.2) collected longitudinally.

At the stable periods, samples were collected during patients’ routine visit to the hospital, before the exacerbation of the disease. Patients at this stage are clinically well and do not need to stay in hospital. During exacerbation stage 1, patients were required to stay in hospital, samples were collected before the treatment of intravenous antibiotics. At exacerbation stage 2, samples were collected after patients received the antibiotic treatments for a few days. Then, samples were collected at exacerbation stage 3 when patients are at the end of exacerbation with the resolved symptoms, determined by a clear set of criteria motioned before (Goss and Burns 2007).

2.2.2 DNA extraction and quality check

20 mL cultures of *Streptococcus* strains were inoculated in Brain Heart Infusion (BHI) Broth and incubated overnight at 37°C. Cultures were centrifuged at 4500 rpm for 10 minutes and the supernatant removed. Pellets were resuspended in 480 µL ethylenediaminetetraacetic acid (EDTA), 120 µL lysozyme, and 20 µL mutanolysin before being incubated for one hour at 37°C, with mixing at 15 minutes intervals. Samples were then extracted using

the Wizard Genomic DNA Purification kit (Promega) according to the manufacturer’s Gram-positive bacteria protocol. DNA was rehydrated at 4°C overnight in 50 µL molecular grade water and quantified using the Qubit. Qualified DNA extraction was transferred to Beijing Genomic Institution for quality check of extracted DNA, library preparation and sequencing. Before library preparation. Quality check of extracted DNA were performed by Qubit Broad Range protocol and agarose gel electrophoresis.

2.2.3 Library preparation and whole-genome sequencing

To construct a paired end sequencing library, the extracted DNA samples were purified and sheared into smaller fragments with a desired size by sonication devices. Then the fragments were end-repaired using T4 DNA polymerase, Klenow fragment and T4 polynucleotide kinase to generate blunt ends. After adding an ‘A’ base to the 3’ end of the blunt phosphorylated DNA fragments, adapters are ligated to the ends of the DNA fragments. The fragments with the desired size were selected using electrophoresis for each sample, then selectively enriched with index tag and amplified by PCR. Before sequencing, the library quality was checked. Illumina HiSeq X10 pair-end platform was used by BGI (China) to perform whole-genome sequencing.

Table 2-1: Patients and sputum sample collection information.

Patient	CF4	CF7	CF8	CF9	CF10
Sex	Male	Female	Female	Female	Female
Age (year)	31	24	25	28	22
FEV_1 (%)	43	37	37	30	36
BMI	20.5	17	18	24	17
LES duration	> 5 yr	> 5 yr	> 5 yr	> 5 yr	2007

Definition of abbreviations: BMI = body mass index; LES = Liverpool epidemic strain; LES duration means the number of years the patient has been infected with the LES, reported as the year when infection was first confirmed or as greater than 5 years. Values for FEV_1 and body mass index from these patients in the table were collected on January 2009 as an indicator for stable period.

Table 2-2: List of 60 isolated strains from sputum samples.

Patient	Patient 4	Patient 7				Patient 8					Patient 9	Patient 10
Time point	Stable	Acute1	Acute2	Acute3	Stable	Stable	Acute1	Acute2	Acute3	Stable	Stable	Stable
Strains	4.2	7.1.8	7.2.6	7.3.3	7.1	8.1	8.1.6	8.2.1	8.3.3	8.5.11	9.1	10.1
	4.3	7.1.10	7.2.11	7.3.8	7.2	8.3	8.1.8	8.2.3	8.3.4	8.5.12	9.3	10.2
	4.4	7.1.12	7.2.13	7.3.11	7.3	8.4	8.1.9	8.2.4	8.3.6	8.5.13	9.4	10.3
	4.5	7.1.13	7.2.15	7.3.14	7.4	8.5	8.1.10	8.2.5	8.3.14	8.5.16	9.5	10.4
	-	7.1.15	7.2.16	7.3.15	7.5	8.6	8.1.11	8.2.6	8.3.15	8.5.17	-	10.5
	-	-	-	-	7.6	-	-	-	-	-	-	10.6
Sample collection time	2011/7/5	2010/3/5	2010/3/27	2010/4/6	2010/5/11	2010/5/20	2010/7/7	2010/7/12	2010/7/19	2010/9/21	2010/9/7	2009/6/2

These strains were picked randomly during the culturing stage for isolation, followed by streak plating and culture, DNA extraction, library preparation and finally next-generation sequencing of the whole genome.

Stable time point: The samples were collected during the stable condition of the patient with cystic fibrosis (PwCF) based on clinical criteria.

Acute 1: Patients are at the start of exacerbation stage of lung infection, before treating with intravenous antibiotics.

Acute 2: Patients are still in the stage of exacerbation, but receiving dual intravenous antibiotic treatment.

Acute 3: Patients are at the end of exacerbation with the resolved symptoms, determined by a clear set of criteria.

Criteria for exacerbation, clinically characterized including a 10% drop in forced expiratory volume in 1s (FEV_1), increased sputum production, discoloration,

2.3 Sequencing reads pre-processing, *de novo* genome assembly and annotation

2.3.1 Raw read pre-processing

The generated paired-end reads were around 150 bp (single read). The raw reads with at least 40% of low-quality bases (Phred ≤ 20), with at least 40% of ambiguous nucleotides, with adapter sequences and duplicates were trimmed.

2.3.2 *De novo* genome assembly

The preprocessed reads were checked and assembled using an automatic pipeline Bactopia (v. 1.4.x) (Petit and Read 2020). In the pipeline, the reads went through the quality check by FastQC (v. 0.11.9), reads assembly by Shovill (v.1.1.0) and assembled genome quality check by CheckM (v.1.1.2) (Table 3). FastQC (v. 0.11.9) controls the quality of high throughput sequencing reads. Then the assembler Shovill (v.1.1.0) was used to generate bacteria genome in contig level. Finally, CheckM (v.1.1.2) assessed the quality of assembled microbial genomes. Assemblies with low completeness (<95%) or high contamination level (>5%) were filtered out based on CheckM prediction (Parks, Imelfort et al. 2015)
Bactopia analysis was performed for each strain using the following commands:

```
# Build dataset
bactopia datasets ~/bactopia_datasets –cpus 40
# Process samples one by one
bactopia --R1 (forward_reads_work_dir)Clean.1.fq.gz --R2
(reverse_reads_work_dir)Clean.2.fq.gz --sample sample_name –datasets
bactopia_datasets/ --outdir work_dir_of_output --cpus 40
```

2.3.3 Genome annotation

The assembled genomes were annotated using the Rapid Annotation using Subsystem Technology (RAST) pipeline (Aziz, Bartels et al. 2008).

Assembled contigs were uploaded to the RAST webserver https://rast.nmpdr.org. The RAST annotation engine was specifically designed for bacterial and archaeal genomes' annotation. Genomic features (ie., protein-encoding genes and RNA) and their functions of each strain were annotated by a standard software pipeline in RAST.

2.4 Species identification

2.4.1 Phylogenetic analysis by using whole-genome approach

Appropriate reference genomes for our strains were identified by a tool called referenceseeker. The database used by referenceseeker was RefSeq release: 205 (2021-04-01).
Core-genome single-nucleotide polymorphisms (SNPs) were used to construct phylogenetic trees. To identify core-genome SNPs, we submitted all genome sequences to PanSeq v3.2.1 (Laing, Buchanan et al. 2010), which aligned the sequences and identified the core-genome sequences. SNPs in the highly conserved regions were identified for tree construction. Highly conserved regions for sequence identity cut-off were set to at least 50%, and the core-genome sequence threshold was set to be the same as the number of genomes used. Core-genome SNPs were aligned by Multiple Sequence Comparison by Log-Expectation (MUSCLE) algorithm (Edgar 2004) with MEGA X. Then, a maximum likelihood tree was built with 1,000 bootstraps replicates and the Kimura 2-parameter model using MEGA X (Kumar, Stecher et al. 2018).

2.4.2 Average nucleotide identity (ANI) analysis

To identify the species level of each strain, the average nucleotide identity (ANI) evaluated by BLAST (ANIb) was used to evaluate the relatedness between *Streptococcus* genomes, which was calculated using the webserver JSpeciesWS (Konstantinidis and Tiedje 2005). Comparison between genomes of two strains with ANIb value greater than 96% (>96%), which is equivalent to 70% relatedness by using conventional DNA-DNA Hybridization

(DDH), were considered to belong to the same species (Goris, Konstantinidis et al. 2007, Richter and Rossello-Mora 2009).

2.4.3 Genome-to-genome distance calculator (GGDC) analysis

GGDC approach was also used to identify the species level of our strains (Meier-Kolthoff, Hahnke et al. 2014). GGDC provides three formulas for calculating DDH, termed Formulas 1, 2 and 3. Formula 2 was preferred in our study as it calculates the sum of all identities found in high scoring pairs (HSP) and divides it by the length of the HSP. Thus, GGDC value from formula 2 is independent of the genome size. Therefore, it is better for incomplete genomes with contig level, which we generated in this study. Furthermore, we used a cut-off value of 70% for formula 2 to define the boundary between species. Thus, a GGDC value over 70% indicates the two strains belonged to the same species.

2.5 Characterization of novel *Streptococcus* species using bioinformatics approaches

2.5.1 Prediction of Plasmids

For each novel bacterial strain/species, the assembled genome sequences were uploaded to the web server PlasmidFinder 2 (Carattoli, Zankari et al. 2014) (https://cge.food.dtu.dk/services/PlasmidFinder/). The selected database was Gram Positive based on replicons identified in Gram positive bacteria. The selected threshold for minimum sequence identity was 50% and selected minimum sequence coverage was 50%. In this tool, the replicon sequences from the PlasmidFinder Gram Positive database were BLASTed against the complete plasmid sequences (Camacho, Coulouris et al. 2009). A comparison between plasmids in strains were performed. A plasmid presence/absence matrix was built by excel and a heatmap was graphed by heatmap illustrator in TBtools (Chen, Chen et al. 2020).

2.5.2 Prediction of the restriction-modification system

The assembled genome sequences were uploaded to Restriction-ModificationFinder webserver (https://cge.food.dtu.dk/services/Restriction-ModificationFinder/), which was based on REBASE (Roer, Hendriksen et al. 2016). Database from REBASE include type I to IV restriction-modifiction systems: restriction genes, methyltransfereases, and specificity units (Roberts, Vincze et al. 2015). The tool was built on a BLAST-based methodology for detection of genes from REBASE. The database includes putative genes as well as genes with known functions. Database was selected as All incl. putative genes. Settings were chosen as thresholds for %ID at 50% and minimum length of 60%.

Predicted genes were curated manually. Identified restriction genes, methyltransferase genes and specificity genes were inspected one by one to form a restriction-modification system. Unknown systems were assigned to types with genes identified. Incomplete systems were investigated for truncated genes by contigs with that system blasted against REBASE. Putative systems were built if all genes were clustered on the same contig, even if truncated or frame shifted. The systems were merged and named according to the type of system. A comparison of restriction modification system presence/absence was performed as described for plasmid prediction.

2.5.3 Prediction of the CRISPR-Cas system

Clustered regularly interspaced short palindromic repeats (CRISPR arrays) and their associated proteins (Cas) were predicted by CRISPRCasFinder using default settings (Couvin, Bernheim et al. 2018).

2.5.4 Prediction of antibiotic resistance genes

Assembled genomes of novel strains were uploaded to ResFinder (https://cge.food.dtu.dk/services/ResFinder/) to predict genes mediating antimicrobial resistance (Zankari, Hasman et al. 2013). Only acquired antibiotic resistance genes were predicted because *Streptococcus* strains were not included in PointFinder database. The prediction was based on sequence similarity search with select threshold for %ID at 90% and select minimum length at 60% (Camacho, Coulouris et al. 2009).Because we only did the prediction of antibiotic resistance based on whole genome, so we included all antibiotics databases in the select antibiotic configuration setting part.

2.5.5 Prediction of genomic islands

Genomic islands (GIs) are genomic regions in the bacterial genome which originate from horizontal transfer. GIs were predicted using IslandViewer4 (Bertelli, Laird et al. 2017) by uploading RAST-annotated GenBank files generated in RAST annotation part mentioned above. If the predicted GI was predicted from two different contigs, it was discarded from analysis . Homologous GIs from different *Streptococcus* genomes were grouped into the same cluster using Mmseqs2 pipeline if they had at least 50% nucleotide sequence identity and at least 50% nucleotide sequence coverage. Genes in represented GIs were also analyzed. A GI presence/absence matrix was built by excel and a heatmap was graphed by heatmap illustrator in TBtools (Chen, Chen et al. 2020).

2.5.6 Prediction of prophages

For each Streptococcus genome, prophages were predicted using PHASTER (Arndt, Grant et al. 2016). We selected the Automatic Model Section and Assembled Genome/Contigs platform as parameters. The putative prophage sequences from different *Streptococcus* genomes were extracted and compared using Mmseqs2 pipeline. The highly similar prophage sequences with both >50% nucleotide sequence identity and >50% nucleotide sequence coverage was grouped into the same cluster.

2.5.7 Prediction of virulence factors

The putative virulence genes of the 14 assembled *Streptococcus* genomes were predicted by searching all RAST-predicted proteins against the Virulence Factor Database (VFDB) (Chen et al., 2012) using VFanalyzer (http://www.mgc.ac.cn/cgi-bin/VFs/v5/main.cgi) (Liu et al., 2019). The predicted virulence genes from all Streptococcus genomes were clustered using hierarchical clustering algorithm and visualized using heatmaps for comparisons.

2.5.8 Comparative analysis of closely related novel species by whole genome comparison

Based on species identification results, we compared genome variants between related strains from the same source (>99 ANI and GGDC values) based on reference-based reads alignment approaches using Breseq (Deatherage, Traverse et al. 2014).

2.6 *In vitro* and *in vivo* testing of novel species

2.6.1 Biochemical tests

The biochemical tests were performed in triplicate by determining growth of strains in the presence of 94 different compounds (Table 2.3) using Biolog Gen III GN/GP Microplate system and Inoculating Fluid C (Technopath) according to the manufacturer's protocol. Briefly, bacterial cultures were inoculated to an OD600 of 0.1 and 100 µL added to each well. Plates were incubated overnight, and absorbance readings were taken using a plate reader at 575 nm. To characterize carbon utilization, seven categories of carbon sources were tested, including 26 sugars, 5 sugar alcohols, 2 phosphorylated-hexose, 9 amino acids, 9 hexose acids, 18 carboxylic acids, and others (including D-aspartic acid and D-serine). For the chemical sensitivity, we tested the resistance activity of the strains in different conditions, including 2 pH, 4 salts, 7 antibiotics, 2 redox dyes, and others.

2.6.2 Cell morphology observation and biofilm forming capacity testing

Novel strains identified above were routinely grown on LB agar (Invitrogen Paisley, United Kingdom) at 37 °C. To prepare cultures for inoculate biofilms, liquid cultures were grown overnight at 37 °C in Todd-Hewitt broth (Oxoid, Hampshire, United Kingdom) supplemented with 0.5% BBL yeast extract

(Becton Dickinson) (THY) (Whiley, Sheikh et al. 2014). Cultured samples were transferred to wells containing 1 mL of 2.5% gutaraldehyde solution for 2h. Then, the samples were rinsed three times in phosphate-buffered saline (PBS) solution. The samples were exposed to 1% osmium tetroxide (OsO_4) for 1 h and then were dehydrated with increasing ethanol percentages (35%, 50%, 75%, 2 times 90%, and 2 times 100%) for 30 min in each solution. Samples were then immersed in hexamethyldisilazane for 1.5 h and placed in a desiccator for 12 h. Each disk was gold sputter-coated and mounted on a glass slide (Basso et al., 2011). Images were captured using the SEM (machine and detail) at a working distance of 8.5 mm and field widths of c. 100 µm, 50 µm, 20 µm, 1 µm and 500 nm (Weber, Delben et al. 2014).

Virulence testing of bacterial strains using *Galleria* model

The *Galleria mellonella* infection model was used to determine differences in virulence in a simple in vivo model. Strains grown in THY broth were washed three times and resuspended in PBS to achieve a concentration of 4 × 10^7 CFU/mL. For comparison purposes, *Streptococcus pneumoniae* D39 was used (Slager, Aprianto et al. 2018). Larvae were obtained from a standard supplier (Applied Biosystems). Larvae were kept at room temperature in darkness and used within 1 week of delivery. Before inoculation, larvae were assessed for normal movement and signs of health. For each isolate, 10 larvae were separated into a petri dish. 10 µL of each strain suspensions were inoculated using a Hamilton Syringe by injection into hemocoels of the left prolegs of *G. mellonella* larvae weighing 250-350 mg (Liverpool UK). Controls injected with PBS and controls with no injection (n=10 for each) were included in each experiment. Injected larvae were incubated at 37°C in Petri dishes lined with filter paper, and the number of viable larvae were recorded for 24h, 48h, 72h and 96h. Each experiment was performed in triplicate. Larvae were also scored for movement, melanisation and pupation. Analysis was performed in R using the Survival programme.

Virulence testing of bacterial strains using mice model

Strains with significant pathogenicity from Galleria model were performed in mice models. 10 female CD1 mice 6-8 weeks old were intranasally infected with 50µl of inoculum resulting in infection of 1x10^6 bacteria per mouse. Subjects were checked for signs of disease and weighed daily, or more often as infection progressed. Any mice reaching experimental severity limits were culled and their blood, nasopharynx and lungs were sampled for enumeration of bacterial colonization. Mice were ordered from Charles River. All animals were kept at the University of Liverpool animal facilities during the time of the experiments. All experiments were conducted in strict accordance with the recommendations of the European Convention for the Protection of Vertebrate Animals used for Experimental and Other Scientific Purposes (ETS 123) and Directive 2010/63/EU and Portuguese rules (DL 113/2013). This study was performed in strict accordance with UK Home Office guidelines, under project license PP2072053. Animal experiments were performed at the University of Liverpool with approval from local animal welfare and ethics committees. All those involved in *in vivo* experiments were UK Home Office personal license holders, having completed training modules on ethical use of animals and legal requirements for use of regulated species in research. In addition, the University of Liverpool provides bespoke training in animal handling and procedural work, including intranasal administration of infectious agents to an anesthetized animal.
All efforts were made to minimize animal suffering and to reduce the number of animals used. No animals were excluded from the analysis.
Five days post-infection, any surviving mice were culled, their tissues sampled and analyzed as previously detailed. To assess bacterial colonization, the lungs and nasopharynx were aseptically removed and homogenized in PBS. Serial dilutions were prepared in sterile saline, and plated for CFU counts.

Figure 2-1：List and positions of chemicals for biochemical analysis.

The chemicals used can be divided into different groups: control groups (2), related to pH (2), different gradients of salts (4), antibiotics (7), Hexose-PO (2), Redox Dyes (2), Sugar Alcohols (5), Amino Acids (9), Carboxylic Acids (18), Sugars (26) and others.

	1	2	3	4	5	6	7	8	9	10	11	12
A	Negative Control	Dextrin	D-Maltose	D-Trehalose	D-Cellobiose	Gentiobiose	Sucrose	D-Turanose	Stachyose	Positive Control	pH 6	pH 5
B	D-Raffinose	α-D-Lactose	D-Melibiose	β-Methyl-D-Glucoside	D-Salicin	N-Acetyl-D-Glucosamine	N-Acetyl-β-D-Mannosamine	N-Acetyl-D-Galactosamine	N-Acetyl Neuraminic Acid	1% NaCl	4% NaCl	8% NaCl
C	α-D-Glucose	D-Mannose	D-Fructose	D-Galactose	3-Methyl Glucose	D-Fucose	L-Fucose	L-Rhamnose	Inosine	1% Sodium Lactate	Fusidic Acid	D-Serine
D	D-Sorbitol	D-Mannitol	D-Arabitol	myo-Inositol	Glycerol	D-Glucose-6-PO4	D-Fructose- 6-PO4	D-Aspartic Acid	D-Serine	Troleandomycin	Rifamycin SV	Minocycline
E	Gelatin	Glycyl-L-Proline	L-Alanine	L-Arginine	L-Aspartic Acid	L-Glutamic Acid	L-Histidine	L-Pyroglutamic Acid	L-Serine	Lincomycin	Guanidine HCl	Niaproof 4
F	Pectin	D-Galacturonic Acid	L-Galactonic Acid Lactone	D-Gluconic Acid	D-Glucuronic Acid	Glucuronamide	Mucic Acid	Quinic Acid	D-Saccharic Acid	Vancomycin	Tetrazolium Violet	Tetrazolium Blue
G	p-Hydroxy-Phenylacetic Acid	Methyl Pyruvate	D-Lactic Acid Methyl Ester	L-Lactic Acid	Citric Acid	α-Keto-Glutaric Acid	D-Malic Acid	L-Malic Acid	Bromo-Succinic Acid	Nalidixic Acid	Lithium Chloride	Potassium Tellurite
H	Tween 40	γ-Amino-Butyric Acid	α-Hydroxy-Butyric Acid	β-Hydroxy-D,L-Butyric Acid	α-Keto-Butyric Acid	Acetoacetic Acid	Propionic Acid	Acetic Acid	Formic Acid	Aztreonam	Sodium Butyrate	Sodium Bromate

Chapter 3 Genome assembly and annotation

3.1 Introduction

To assemble and generate bacteria genomes for further study, quality controls of DNA sequencing reads, statistics of assembled genomes and the annotation of bacterial genomes were presented and discussed in this chapter. Specifically, the results in this chapter include: The concentration and integrity of sequenced reads before library preparation, quality control results of high throughput sequencing data, summary statistics of strains and annotations.

3.2 Quality check of extracted DNA

A total of 60 putative *S.* species strains were cultured and processed to sequencing. The quantity of DNA extracted for all isolates was over 200 ng (Table 3.1, 3.2). Gel electrophoresis was performed to visually inspect DNA degradation. Slight degradation (smearing on the gel) of DNA was observed and may have occurred during transportation. In figure 3.1 and 3.2 lanes 1, 6, 9, 13, 14, 18, 19, 20, and 24, almost all lanes in figure 3.3 and 3.4. we do observe contamination from RNA, either contaminated or slight contaminated level (Figure 3.1 to 3.4). For example, degraded RNA length around 100 to 250 bp long was appeared in lane 1, 2 and 6 in Figure 3.1. Purification needs to be performed for strains at contamination level. Strain CF4-2, CF7-2, CF7-5, CF8-3, CF8-5, CF9-4, CF9-5, CF10-1 and CF10-4 were purified before library preparation. After contamination removal, all samples were prepared as a less than 800 bp insert size sequencing library for next-generation sequencing.

Table 3-1: Summary of quality of extracted DNA from batch1 25 isolates.

No.	Sample Name	Concen-tration(ng/μL)	Volume(μL)	Total Mass(μg)	Sample Integrity	Remark
1	CF.4-2	96	82	7.87	Degraded slightly	Contaminated with RNAs.Need to be purify.
2	CF.4-3	74	80	5.92	Degraded slightly	Slight contaminated with RNAs.
3	CF.4-4	73	80	5.84	Degraded slightly	Slight contaminated with RNAs.
4	CF.4-5	84	80	6.72	Degraded slightly	N/A
5	CF.7-1	55	80	4.4	Degraded slightly	N/A
6	CF.7-2	105	80	8.4	Degraded slightly	Contaminated with RNAs.Need to be purify.
7	CF.7-3	116	80	9.28	Degraded slightly	N/A
8	CF.7-4	121	80	9.68	Degraded slightly	N/A
9	CF.7-5	99	80	7.92	Degraded slightly	Contaminated with RNAs.Need to be purify.
10	CF.7-6	78	80	6.24	Degraded slightly	N/A
11	CF.8-1	75	80	6	Degraded slightly	N/A
12	CF.8-2	162	80	12.96	Degraded slightly	Slight contaminated with RNAs.
13	CF.8-3	164	80	13.12	Degraded slightly	Contaminated with RNAs.Need to be purify.
14	CF.8-5	51	80	4.08	Degraded slightly	Contaminated with RNAs.Need to be purify.
15	CF.8-6	286	80	22.88	Degraded slightly	N/A

16	CF.9-1	84	80	6.72	Degraded slightly	N/A
17	CF.9-3	85	80	6.8	Degraded slightly	N/A
18	CF.9-4	124	80	9.92	Degraded slightly	Contaminated with RNAs.Need to be purify.
19	CF.9-5	101	80	8.08	Degraded slightly	Contaminated with RNAs.Need to be purify.
20	CF.10-1	63	80	5.04	Degraded slightly	Contaminated with RNAs.Need to be purify.
21	CF.10-2	124	80	9.92	Degraded slightly	Slight contaminated with RNAs.
22	CF.10-3	212	80	16.96	Degraded slightly	N/A
23	CF.10-4	118	80	9.44	Degraded slightly	Contaminated with RNAs.Need to be purify.
24	CF.10-5	87	80	6.96	Degraded slightly	Slight contaminated with RNAs.
25	CF.10-6	108	80	8.64	Degraded slightly	N/A

Table 3-2: DNA integrity of 8 isolates.

Lanes with more than one highlighted band showed contamination during DNA extraction or DNA degradation after extraction. The position of lanes with corresponding number in Table 3-1.

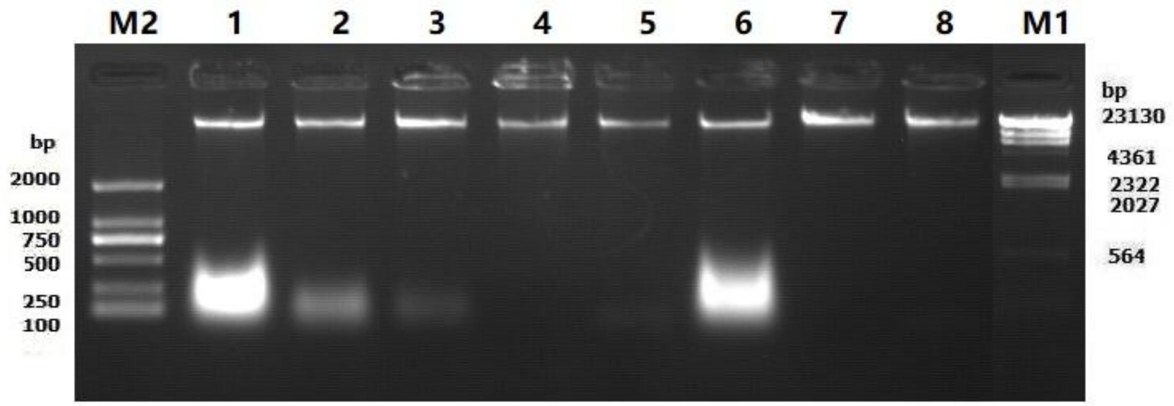


Figure 3-2: DNA integrity of 17 isolates.

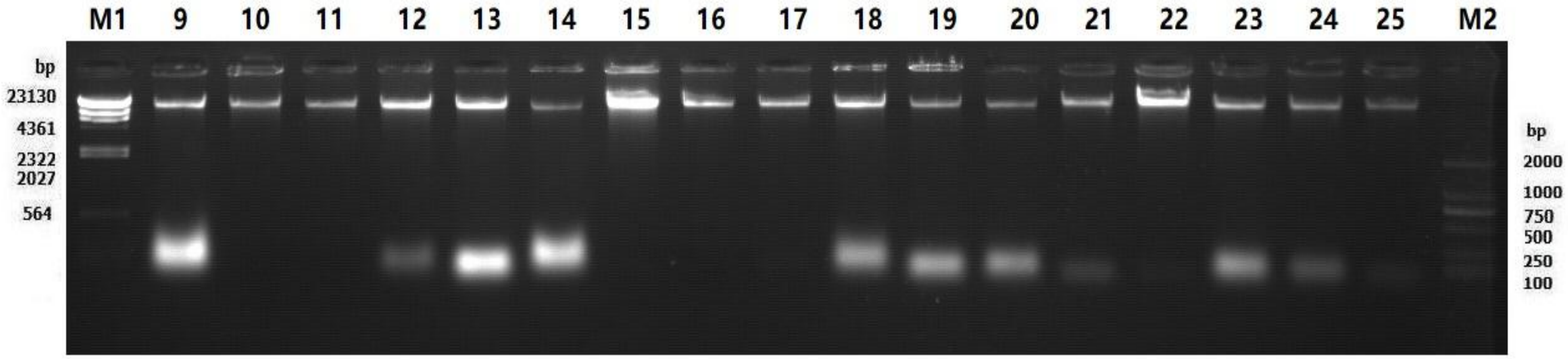


Table 3-3: Summary of quality of extracted DNA from batch2 35 isolates.

No.	Sample Name	Concen-tration(ng/μL)	Volume(μL)	Total Mass(μg)	Sample Integrity	Remark
1	CF7_Ac1-8	86.4	42	3.63	Degraded slightly	Slight contaminated with RNAs.
2	CF7_Ac1-10	159.6	42	6.7	Degraded slightly	Slight contaminated with RNAs.
3	CF7_Ac1-12	96.1	42	4.04	Degraded slightly	Slight contaminated with RNAs.
4	CF7_Ac1-13	184.3	67	12.35	Degraded slightly	
5	CF7_Ac1-15	163	67	10.92	Degraded slightly	Slight contaminated with RNAs.
6	CF7_Ac2-6	210.2	42	8.83	Degraded slightly	Slight contaminated with RNAs.
7	CF7_Ac2-11	50	42	2.1	Degraded slightly	Slight contaminated with RNAs.
8	CF7_Ac2-13	56.4	42	2.37	Degraded slightly	Slight contaminated with RNAs.
9	CF7_Ac2-15	233.1	42	9.79	Degraded slightly	Slight contaminated with RNAs.
10	CF7_Ac2-16	314.4	45	14.15	Degraded slightly	Slight contaminated with RNAs.
11	CF7_Ac3-3	30	45	1.35	Degraded slightly	
12	CF7_Ac3-8	128.9	45	5.8	Degraded slightly	Slight contaminated with RNAs.
13	CF7_Ac3-11	46.3	45	2.08	Degraded slightly	Slight contaminated with RNAs.
14	CF7_Ac3-14	35.6	41	1.46	Degraded slightly	Slight contaminated with RNAs.
15	CF7_Ac3-15	38.6	41	1.58	Degraded slightly	Slight contaminated with RNAs.
16	CF8_St5-11	70.5	43	3.03	Degraded slightly	Slight contaminated with RNAs.

17	CF8_St5-12	40.1	43	1.72	Degraded slightly	Slight contaminated with RNAs.
18	CF8_St5-13	266.2	43	11.45	Degraded slightly	Slight contaminated with RNAs.
19	CF8_St5-16	120.1	43	5.16	Degraded slightly	
20	CF8_St5-17	31.8	43	1.37	Degraded slightly	
21	CF8_Ac1-6	93.5	43	4.02	Degraded slightly	Slight contaminated with RNAs.
22	CF8_Ac1-8	223.2	43	9.6	Degraded slightly	Slight contaminated with RNAs.
23	CF8_Ac1-9	292.5	65	19.01	Degraded slightly	
24	CF8_Ac1-10	103.9	41	4.26	Degraded slightly	Slight contaminated with RNAs.
25	CF8_Ac1-11	206.7	41	8.47	Degraded slightly	Slight contaminated with RNAs.
26	CF8_Ac2-1	96.6	41	3.96	Degraded slightly	Slight contaminated with RNAs.
27	CF8_Ac2-3	64.6	41	2.65	Degraded slightly	Slight contaminated with RNAs.
28	CF8_Ac2-4	106.4	41	4.36	Degraded slightly	Slight contaminated with RNAs.
29	CF8_Ac2-5	51.4	41	2.11	Degraded slightly	Slight contaminated with RNAs.
30	CF8_Ac2-6	113	41	4.63	Degraded slightly	Slight contaminated with RNAs.
31	CF8_Ac3-3	103.6	41	4.25	Degraded slightly	
32	CF8_Ac3-4	243.7	41	9.99	Degraded slightly	Slight contaminated with RNAs.
33	CF8_Ac3-6	127.8	41	5.24	Degraded slightly	Slight contaminated with RNAs.

34	CF8_Ac3-14	72.2	41	2.96	Degraded slightly	Slight contaminated with RNAs.
35	CF8_Ac3-15	94	41	3.85	Degraded slightly	Slight contaminated with RNAs.

Table 3-4: DNA integrity of 20 isolates from batch 2.

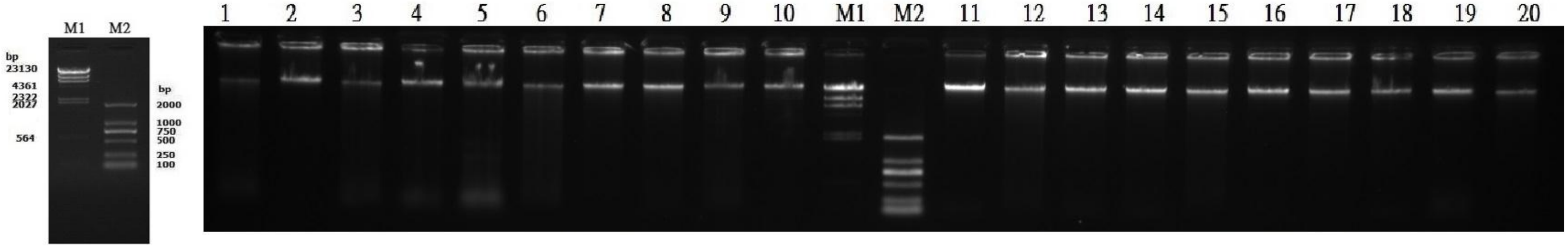
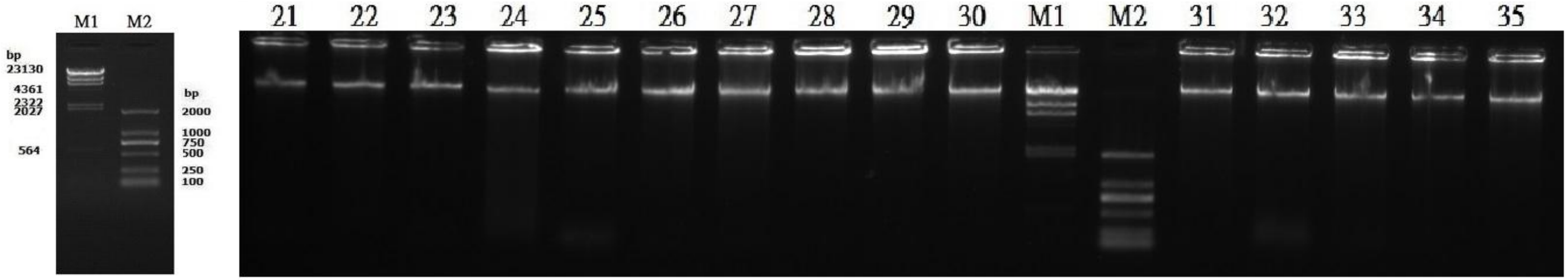


Table 3-5: DNA integrity of 15 DNA extractions from batch 2.



3.3 Quality control of sequenced reads and reads assembly

After sequencing, low quality and contaminated raw reads were filtered (Table 3.3). Using Bactopia, for each strain, reads were trimmed, quality controlled (QC), down sampled to an average of around 100× genome coverage and assembled.
For next-generation sequencing reads quality control, three measurements are often the focus in FastQC: the adapter content, the overrepresented sequences and per-base quality. The pass for these three modules indicates no systematic issue with the sequencing. For all 60 sequencing read sets, no adapter content was detected (Supplementary files for FastQC results), there was no evidence of overrepresented sequences and bases in all reads had very high-quality scores. For all samples, the read length of all strains was 35 to 150 bp with various calculated GC content. The quality score used in sequencing is Pred+33, the sanger format (Illumina 1.9). All samples contain enough read coverage to ensuring sequencing of the entire interested genomes. The mean quality scores of all bases in the reads were over 30, meaning that the sequencing error was less than 0.1%. This indicates a good quality sequencing. Similarly, for per base sequence quality figures, the y-axis of the graphs from 1 to 6 bp, meaning the start of the sequencing, the mean base quality gradually improved from 32 to 40. Then, the quality of calls will degrade as the run progresses. The background of the graph was divided in 3 parts, very good quality calls (green), reasonable quality (orange), and poor-quality calls (red). All quality scores fall into the very good quality part. These are all shown as natural illumina sequencing data.
Reads were trimmed before processing to further downstream analysis. Trimmed paired-end reads for each strain were assembled by Shovill-skesa 1.1.0 in Bactopia. Assembled genomes were quality checked by CheckM for completeness and contamination measurement using single copy marker genes. The threshold for medium contamination is over 5%. Strains with medium and higher contamination were removed before downstream analysis. The quality of 60 assembled genomes in contig level were estimated by CheckM.

Table 3-6: Statistics of DNA sequencing reads.

Sample Name	Raw Data (Mb)	Adapter (%)	Duplication (%)	Filtered Reads (%)	Low Quality Filtered Reads (%)	Clean Data (Mb)
CF4-2	5,246	0.48	27.94	31.1	2.67	3,614
CF4-3	6,262	0.36	29.7	32.61	2.54	4,219
CF4-4	6,727	0.55	27.18	30.8	3.06	4,654
CF4-5	5,210	0.74	24.91	26.96	1.29	3,805
CF7-1	5,531	0.83	27.25	30.23	2.14	3,858
CF7-2	5,719	0.45	25.83	29.2	2.9	4,049
CF7-3	5,835	0.3	27.12	29.42	1.99	4,118
CF7-4	5,529	0.44	27.69	29.89	1.75	3,876
CF7-5	5,341	0.42	26.76	29.6	2.41	3,760
CF7-6	5,309	0.98	26.24	29.25	2.02	3,756
CF8-1	5,720	1.68	24.17	27.86	2	4,126
CF8-2	5,692	0.52	25.73	28.47	2.21	4,071
CF8-3	4,198	0.84	23.55	26.81	2.41	3,072
CF8-5	5,452	1.01	26.38	30.2	2.8	3,805
CF8-6	4,973	0.39	24.95	27.47	2.11	3,607
CF9-1	6,250	0.64	27.09	29.49	1.75	4,406
CF9-3	4,941	0.84	25.6	28.52	2.06	3,531
CF9-4	5,845	0.65	27.71	30.23	1.85	4,077
CF9-5	5,055	0.72	25.42	29.03	2.88	3,587
CF10-1	5,470	1.5	25.66	30.5	3.32	3,801
CF10-2	5,743	0.36	27.17	29.85	2.31	4,028
CF10-3	6,660	0.56	25.78	29.07	2.71	4,723
CF10-4	5,076	0.8	24.68	28.1	2.61	3,649
CF10-5	6,170	2.75	26.39	31.52	2.37	4,225
CF10-6	4,751	1.31	22.11	26.19	2.75	3,507
CF7_Ac1-10	3,254	2.31	11.77	15.58	1.47	2,747
CF7_Ac1-12	3,904	2.56	11.28	15.74	1.88	3,289
CF7_Ac1-13	3,752	2.43	9.25	13.04	1.33	3,262
CF7_Ac1-15	3,911	2.59	9.09	13.17	1.46	3,396
CF7_Ac1-8	3,894	2.73	11.65	16.32	1.93	3,258
CF7_Ac2-11	3,380	2.1	10.25	14.32	1.95	2,895
CF7_Ac2-13	3,548	2.34	10.51	14.45	1.58	3,036
CF7_Ac2-15	3,604	2.59	10.37	14.8	1.83	3,071
CF7_Ac2-16	3,894	2.1	11.64	15.54	1.78	3,289
CF7_Ac2-6	3,573	2.11	10.39	14.42	1.9	3,057

CF7_Ac3-11	3,675	2.08	10.4	14.49	1.99	3,142
CF7_Ac3-14	3,796	1.93	10.44	14.2	1.8	3,257
CF7_Ac3-15	3,784	2.04	10.69	14.31	1.55	3,243
CF7_Ac3-3	3,741	2.27	10.28	14.9	2.31	3,184
CF7_Ac3-8	3,491	2.36	10.31	14.66	1.97	2,979
CF8_Ac1-10	3,666	2.57	11.01	15.44	1.84	3,099
CF8_Ac1-11	3,953	1.75	11.45	15.09	1.87	3,356
CF8_Ac1-6	3,387	3.62	8.44	13.69	1.61	2,923
CF8_Ac1-8	3,736	2	11.5	15.25	1.72	3,166
CF8_Ac1-9	3,932	1.88	10.7	14.64	2.02	3,356
CF8_Ac2-1	3,852	2.15	10.38	14.54	1.98	3,292
CF8_Ac2-3	3,687	1.7	11.08	14.57	1.76	3,150
CF8_Ac2-4	3,635	1.61	10.49	14.12	2	3,121
CF8_Ac2-5	3,483	2.61	10.79	14.61	1.19	2,974
CF8_Ac2-6	3,918	2.17	10.99	14.29	1.11	3,358
CF8_Ac3-14	3,936	3.05	10.9	15.48	1.51	3,326
CF8_Ac3-15	3,946	2.39	10.46	14.64	1.77	3,368
CF8_Ac3-3	3,691	2.1	9.05	12.81	1.64	3,218
CF8_Ac3-4	3,911	1.59	10.55	14.04	1.88	3,362
CF8_Ac3-6	3,953	2.02	11.25	15.04	1.75	3,358
CF8_St5-11	3,394	2.12	10.01	13.99	1.84	2,919
CF8_St512	3,647	2.86	9.65	14.47	1.94	3,119
CF8_St5-13	3,562	1.85	9.92	13.63	1.84	3,076
CF8_St5-16	3,860	2.23	10.24	14.35	1.86	3,305
CF8_St5-17	3,939	1.98	11.26	14.84	1.58	3,354

A total of four different maker lineages were selected by CheckM automatically for strains in this study. These marker lineages were based on tree root (UID1), k__Bacteria (UID203), o__Actinomycetales (UID1530), o__Pseudomonadales (UID4488) and o__Lactobacillales (UID544). Strains measured by k__Bacteria (UID203), o__Actinomycetales (UID1530), o__Pseudomonadales (UID4488) were all removed before downstream analysis. Genome of these strains were either with contaminations over 5% or belongs to other bacteria at least in genus level. (Table 3-4, Table 3-5). These different marker lineages include different reference genomes with different number of gene markers in different marker sets. Four strains CF4-

4, CF7_Ac2-15, CF10-2 and CF10-5 were measured by using the root marker lineage, based on 5656 genomes, with only 56 marker genes containing 24 marker sets. Interestingly, these strains contained 2 copies of these 24 marker sets, indicating 100% completeness and more than 100% contamination of genomic fragments from both closely or divergent taxa. Eleven strains CF4-5, CF7-1, CF7-2, CF8-1, CF8-2, CF8-3, CF8-5, CF7_Ac3-3, CF10-3, CF10-4 and CF10-6 were estimated by k_Bacteria with different level of completeness and contamination. These may indicate the problems happened in bacterial isolation, culturing and DNA extraction. Only CF8-2 was in substantial completeness level which lower than 90% completeness. All strains were over medium contamination level except CF10-6. CF10-6 was suitable for downstream analysis with near completeness level and low contamination. But the contamination was from bacteria belonging to another order level. Two strains CF7_Ac1-13 and CF7_Ac1-15 were qualified using o_Pseudomonadales (UID4488). This indicated that the two strains belonged to the order Pseudomonadales. These two assembled genomes were near completeness with low contamination. CF8_Ac1-6 was estimated using o_Actinomycetales (UID1530), the lineage marker based on another order Actinomycetales. The rest of the 42 strains were estimated based on o_Lactobacillales (UID544), the order Lactobacillales. Genomes for these strains were qualified as near complete with at least low contamination level. A *Streptoccous* strain would show the taxonomy levels of *Lactobacillales, Streptococcaceae, Streptococcus* from Order to Genus level. So, these 42 strains were suitable for downstream analysis.

Furthermore, these 60 assemblies were assigned with a reference genome by referenceseeker (Table 3-6 and Table 3-7). In terms of bacterial taxonomy, bacteria can be classified into Kingdom, Phylum, Class, Order, Family, Genus, and Species divisions. In the 60 strains, 53 were identified as *Streptococcus* isolates. These strains were closely related to representative genomes from *Streptococcus infantis*, *Streptococcus salivarius*, *Streptococcus oralis*, *Streptococcus mitis*, *Streptococcus sp. LPB0220*, *Streptococcus sp. oral taxon 061 F0704*. In these 53 genomes, only 40 assembled genomes with high quality were used for species identification. We will describe this in detail in the next chapter.

The remaining 7 strains were identified as other genus bacteria. Specifically, CF7_Ac2-15, CF7_Ac2-16, CF9-4 and CF9-5 were closely related to *Enterococcus faecium*. CF7_Ac1-13 and CF7_Ac1-15 were closely related to *Pseudomonas aeruginosa*. CF8_Ac1-6 was related to *Rothia mucilaginosa*.

Table 3-7: Quality estimation of 42 strains based on o__Lactobacillales (UID544) marker lineage by CheckM.

No.	Bin Id	Marker lineage	# genomes	Completeness	Contamination	Heterogeneity
1	CF4-2	o__Lactobacillales (UID544)	293	100	0.37	0
2	CF4-3	o__Lactobacillales (UID544)	293	100	0.37	0
3	CF7-3	o__Lactobacillales (UID544)	293	100	0.75	0
4	CF7-4	o__Lactobacillales (UID544)	293	99.88	0	0
5	CF7-5	o__Lactobacillales (UID544)	293	99.88	0.07	50
6	CF7-6	o__Lactobacillales (UID544)	293	100	0	0
7	CF8-6	o__Lactobacillales (UID544)	293	100	0	0
8	CF9-1	o__Lactobacillales (UID544)	293	100	0	0
9	CF9-3	o__Lactobacillales (UID544)	293	100	0	0
10	CF9-4	o__Lactobacillales (UID544)	293	99.63	0	0
11	CF9-5	o__Lactobacillales (UID544)	293	99.63	0	0
12	CF10-1	o__Lactobacillales (UID544)	293	100	0	0
13	CF7_Ac1-8	o__Lactobacillales (UID544)	293	100	0	0
14	CF7_Ac1-10	o__Lactobacillales (UID544)	293	100	0.75	0
15	CF7_Ac1-12	o__Lactobacillales (UID544)	293	100	0	0
16	CF7_Ac2-6	o__Lactobacillales (UID544)	293	100	0.09	0
17	CF7_Ac2-11	o__Lactobacillales (UID544)	293	100	0.09	0
18	CF7_Ac2-13	o__Lactobacillales (UID544)	293	100	0.09	0
19	CF7_Ac2-16	o__Lactobacillales (UID544)	293	99.63	0	0
20	CF7_Ac3-8	o__Lactobacillales (UID544)	293	100	0.09	0
21	CF7_Ac3-11	o__Lactobacillales (UID544)	293	100	0.09	0

22	CF7_Ac3-14	o__Lactobacillales (UID544)	293	100	0.09	0
23	CF7_Ac3-15	o__Lactobacillales (UID544)	293	100	0.09	0
24	CF8_St5-11	o__Lactobacillales (UID544)	293	100	0	0
25	CF8_St5-12	o__Lactobacillales (UID544)	293	100	0	0
26	CF8_St5-13	o__Lactobacillales (UID544)	293	100	0	0
27	CF8_St5-16	o__Lactobacillales (UID544)	293	100	0	0
28	CF8_St5-17	o__Lactobacillales (UID544)	293	100	0	0
29	CF8_Ac1-8	o__Lactobacillales (UID544)	293	100	0	0
30	CF8_Ac1-9	o__Lactobacillales (UID544)	293	100	0	0
31	CF8_Ac1-10	o__Lactobacillales (UID544)	293	100	0	0
32	CF8_Ac1-11	o__Lactobacillales (UID544)	293	100	0	0
33	CF8_Ac2-1	o__Lactobacillales (UID544)	293	100	0.75	0
34	CF8_Ac2-3	o__Lactobacillales (UID544)	293	100	0.75	0
35	CF8_Ac2-4	o__Lactobacillales (UID544)	293	100	0.75	0
36	CF8_Ac2-5	o__Lactobacillales (UID544)	293	100	0.75	0
37	CF8_Ac2-6	o__Lactobacillales (UID544)	293	100	0.75	0
38	CF8_Ac3-3	o__Lactobacillales (UID544)	293	100	0.75	0
39	CF8_Ac3-4	o__Lactobacillales (UID544)	293	100	0.75	0
40	CF8_Ac3-6	o__Lactobacillales (UID544)	293	100	0.75	0
41	CF8_Ac3-14	o__Lactobacillales (UID544)	293	100	0.75	0
42	CF8_Ac3-15	o__Lactobacillales (UID544)	293	100	0.75	0

Table 3-8: Quality estimation of 18 strains based on 4 different marker lineages.

No.	Bin Id	Marker lineage	# genomes	Completeness	Contamination	Heterogeneity
1	CF8_Ac1-6	o__Actinomycetales (UID1530)	622	98.79	6.41	4
2	CF7_Ac1-13	o__Pseudomonadales (UID4488)	185	99.68	0.45	0
3	CF7_Ac1-15	o__Pseudomonadales (UID4488)	185	99.68	0.45	0
4	CF10-6	k__Bacteria (UID203)	5449	100	3.45	84.62

5	CF4-5	k__Bacteria (UID203)	5449	100	182.12	60.62
6	CF7-1	k__Bacteria (UID203)	5449	100	5.17	71.43
7	CF7-2	k__Bacteria (UID203)	5449	99.14	98.28	83.33
8	CF8-1	k__Bacteria (UID203)	5449	98.28	145.69	36.84
9	CF8-2	k__Bacteria (UID203)	5449	84.74	113.79	49.68
10	CF8-3	k__Bacteria (UID203)	5449	100	17	93.94
11	CF8-5	k__Bacteria (UID203)	5449	100	108.97	45.9
12	CF7_Ac3-3	k__Bacteria (UID203)	5449	100	56.67	0
13	CF10-3	k__Bacteria (UID203)	5449	99.06	47.71	95.83
14	CF10-4	k__Bacteria (UID203)	5449	95.06	185.66	83.52
15	CF10-5	root (UID1)	5656	100	100	92.86
16	CF10-2	root (UID1)	5656	100	102.08	87.93
17	CF4-4	root (UID1)	5656	100	100	66.07
18	CF7_Ac2-15	root (UID1)	5656	100	100	14.29

Table 3-9: 53 out of 60 strains were identified as Streptococcus strains.

#	Strain	Reference genomes	#ID	Contamination
1	CF8_St5-17	*Streptococcus* sp. oral taxon 061 F0704	GCF_013394695.1	0
2	CF7-4	*Streptococcus salivarius* HSISS4	GCF_000448685.2	0
3	CF7-5	*Streptococcus salivarius* HSISS4	GCF_000448685.2	0.07
4	CF7_Ac2-6	*Streptococcus oralis* S.MIT/ORALIS-351	GCF_001983955.1	0.09
5	CF7_Ac3-11	*Streptococcus oralis* S.MIT/ORALIS-351	GCF_001983955.1	0.09
6	CF7_Ac3-14	*Streptococcus oralis* S.MIT/ORALIS-351	GCF_001983955.1	0.09
7	CF7_Ac2-11	*Streptococcus oralis* S.MIT/ORALIS-351	GCF_001983955.1	0.09
8	CF7_Ac3-8	*Streptococcus oralis* S.MIT/ORALIS-351	GCF_001983955.1	0.09
9	CF7_Ac3-15	*Streptococcus oralis* S.MIT/ORALIS-351	GCF_001983955.1	0.09
10	CF7_Ac2-13	*Streptococcus oralis* S.MIT/ORALIS-351	GCF_001983955.1	0.09
11	CF8-6	*Streptococcus oralis* FDAARGOS_1021	GCF_016127915.1	0
12	CF8_Ac1-11	*Streptococcus* sp. LPB0220	GCF_008727815.1	0
13	CF8_Ac1-9	*Streptococcus* sp. LPB0220	GCF_008727815.1	0
14	CF8_Ac2-6	*Streptococcus* sp. LPB0220	GCF_008727815.1	0
15	CF8_Ac3-4	*Streptococcus* sp. LPB0220	GCF_008727815.1	0

16	CF7_Ac1-10	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
17	CF8_Ac2-1	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
18	CF8_Ac3-3	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
19	CF8_Ac3-6	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
20	CF8_Ac3-14	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
21	CF8_Ac2-3	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
22	CF8_Ac2-4	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
23	CF8_Ac2-5	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
24	CF8_Ac3-15	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
25	CF7-3	*Streptococcus* sp. LPB0220	GCF_008727815.1	0.75
26	CF10-6	*Streptococcus* sp. LPB0220	GCF_008727815.1	3.45
27	CF9-1	*Streptococcus infantis* ATCC 700779	GCF_000187465.1	0
28	CF9-3	*Streptococcus infantis* ATCC 700779	GCF_000187465.1	0
29	CF8_St5-11	*Streptococcus* sp. LPB0220	GCF_008727815.1	0
30	CF8_St5-12	*Streptococcus* sp. LPB0220	GCF_008727815.1	0
31	CF8_St5-16	*Streptococcus* sp. LPB0220	GCF_008727815.1	0
32	CF8_St5-13	*Streptococcus* sp. LPB0220	GCF_008727815.1	0
33	CF7-6	*Streptococcus oralis* SF100	GCF_016549395.1	0
34	CF4-2	*Streptococcus infantis* ATCC 700779	GCF_000187465.1	0.37
35	CF4-3	*Streptococcus infantis* ATCC 700779	GCF_000187465.1	0.37
36	CF10-1	*Streptococcus oralis* FDAARGOS_1021	GCF_016127915.1	0
37	CF7_Ac1-8	*Streptococcus mitis* FDAARGOS_684	GCF_009730515.1	0
38	CF7_Ac1-12	*Streptococcus mitis* FDAARGOS_684	GCF_009730515.1	0
39	CF8_Ac1-8	*Streptococcus* sp. oral taxon 061 F0704	GCF_013394695.1	0
40	CF8_Ac1-10	*Streptococcus* sp. oral taxon 061 F0704	GCF_013394695.1	0
41	CF4-4	*Streptococcus salivarius* FDAARGOS_259	GCF_002073835.2	100, 3 species
42	CF4-5	*Streptococcus salivarius* FDAARGOS_259	GCF_002073835.2	182.12, 4 species
43	CF7-1	*Streptococcus salivarius* HSISS4	GCF_000448685.2	5.17, 4 species
44	CF7-2	*Streptococcus oralis* FDAARGOS_367	GCF_002386345.1	98.28, 5 species
45	CF8-1	*Streptococcus salivarius* FDAARGOS_259	GCF_002073835.2	145.69, 4 species
46	CF8-2	*Streptococcus oralis* subsp. *oralis* OD_332610_07	GCF_002096515.1	113.79, 5 species
47	CF8-3	*Streptococcus oralis* FDAARGOS_367	GCF_002386345.1	17, 5 species
48	CF8-5	*Streptococcus oralis* FDAARGOS_367	GCF_002386345.1	108.97, 4 species
49	CF7_Ac3-3	*Streptococcus oralis* S.MIT/ORALIS-351	GCF_001983955.1	56.67, 6 species

50	CF10-2	*Streptococcus* sp. LPB0220	GCF_008727815.1	102.08, 5 species
51	CF10-3	Streptococcus oralis subsp. oralis OD_332610_07	GCF_002096515.1	47.71, 3 species
52	CF10-4	*Streptococcus oralis* FDAARGOS_367	GCF_002386345.1	185.66, 6 species
53	CF10-5	*Streptococcus* sp. LPB0220	GCF_008727815.1	100, 6 species

Only 40 genomes were good candidate genomes for downstream analysis with low contamination level. Low quality assemblies were highlighted in red and removed before further analysis.

Table 3-10: Seven genomes were identified as Enterococcus, Rothia and Pseudomonas strains.

#	Strain	Reference genomes	#ID	Contamination
1	CF7_Ac2-16	*Enterococcus faecium* UAMSEF_20	GCF_005886735.1	0
2	CF7_Ac2-15	*Enterococcus faecium* UAMSEF_20	GCF_005886735.1	100
3	CF9-4	*Enterococcus faecalis* HA-1	GCF_006349345.1	0
4	CF9-5	*Enterococcus faecalis* HA-1	GCF_006349345.1	0
5	CF8_Ac1-6	*Rothia mucilaginosa* FDAARGOS_369	GCF_002386365.1	6.41
6	CF7_Ac1-13	*Pseudomonas aeruginosa* LESB58	GCF_000026645.1	0.45
7	CF7_Ac1-15	*Pseudomonas aeruginosa* LESB58	GCF_000026645.1	0.45

3.4 The statistics of genome annotation

The filtered 40 *Streptococcus* genomes were annotated by RAST and summary statistics are listed from Table 3-8 to Table 3.13. The strains in our study have similar genome sizes and GC contents compared to another study (Gao, Zhi et al. 2014). The genome size of *Streptococcus* group strains from previous study varied from 1.64 to 2.43 Mbps. The genome size in our study ranges from 1.68 to 2.29 Mbps. The shortest N50 length of the 40 assemblies is 116.5 kbps. The GC content of these strains range from 39.2 to 42%, is consistent with the reported range from 33.79 to 43.40%. Various genome features were predicted and annotated including number of contigs, subsystems, proteins coding genes, RNAs, CRISPRs and repeats. Interestingly, in our study, Strains with similar genome size (difference in up to only thousands of bps) were share the same genome functional prediction results of coding sequences (Figure 3-5). Only less than 31% CDSs were predicted to have a function in other assemblies due to lack of study of other unpredicted CDSs.

Table 3-12: Annotation statistics of 40 good quality Streptococcus assembles, part 1.

Strain	CF4-2	CF4-3	CF7-3	CF7-4	CF7-5	CF7-6	CF7_Ac1-8	CF7_Ac1-12	CF7_Ac1-10	CF7_Ac2-6
Size (bps)	1,735,527	1,732,098	2,072,535	2,130,083	2,137,598	2,003,414	1,962,192	1,959,363	2,105,449	1,988,664
GC Content (%)	39.2	39.2	42	39.9	39.9	41.1	40.1	40.1	42	41
N50 (bp)	270,990	233,173	196,307	195,696	195,696	1,070,883	128,406	128,416	237,648	138,487
Number of Contigs	21	19	22	26	34	13	37	42	25	31
Number of protein Coding genes	1741	1734	2047	1972	1982	1934	1894	1895	2096	1956
Number of RNAs	50	45	37	29	46	35	44	45	49	46

Table 3-11: Annotation statistics of 40 good quality Streptococcus assembles, part 2.

Strain	CF7_Ac2-11	CF7_Ac2-13	CF7_Ac3-8	CF7_Ac3-11	CF7_Ac3-14	CF7_Ac3-15	CF9-1	CF9-3	CF10-1	CF10-6
Size (bps)	1,988,856	1,989,574	1,989,301	1,988,916	1,989,106	1,989,341	1,792,476	1,792,604	2,061,503	2,293,977
GC Content (%)	41	41	41	41	41	41	39.3	39.3	40.7	41.7
N50 (bp)	185,392	185,392	261,091	261,077	185,379	185,392	1,166,477	1,166,477	265,801	219,120
Number of Contigs	26	24	22	22	23	21	13	12	22	88
Number of protein Coding genes	1955	1955	1960	1957	1958	1957	1730	1729	2117	2335
Number of RNAs	46	45	46	46	46	45	29	29	36	45

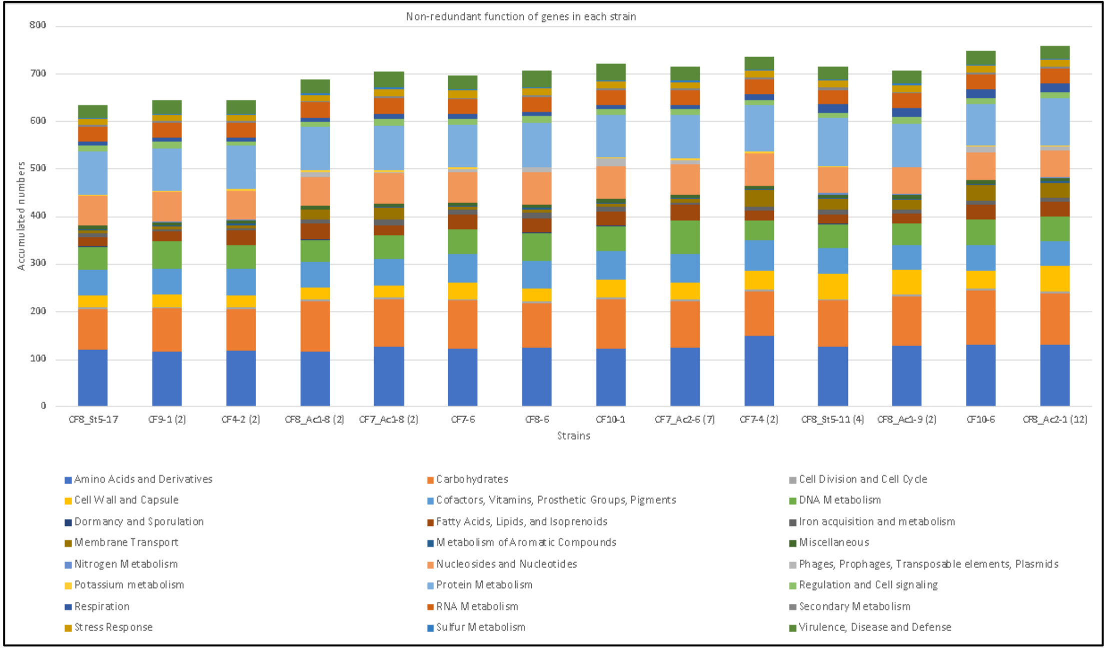
Table 3-13: Annotation statistics of 40 good quality Streptococcus assembles, part 3.

Strain	CF8_Ac1-9	CF8_Ac1-11	CF8_Ac1-8	CF8_Ac1-10	CF8_Ac2-1	CF8_Ac2-3	CF8_Ac2-4	CF8_Ac2-5	CF8_Ac2-6	CF8_Ac3-3
Size (bps)	2,112,874	2,119,345	1,683,820	1,689,616	2,106,426	2,106,446	2,106,212	2,106,652	2,107,216	2,068,924
GC Content (%)	42	41.9	39.4	39.4	42	42	42	42	42	42
N50 (bp)	278,159	214,359	878,180	878,182	123,256	116,549	233,468	233,468	239,952	239,952
Number of Contigs	21	22	12	13	30	30	26	24	25	24
Number of protein Coding genes	2020	2025	1680	1684	2096	2099	2095	2093	2094	2046
Number of RNAs	51	51	45	45	49	49	48	49	49	47

Table 3-14: Annotation statistics of 40 good quality Streptococcus assembles, part 4.

Strain	CF8_Ac3-4	CF8_Ac3-6	CF8_Ac3-14	CF8_Ac3-15	CF8_St5-17	CF8_St5-11	CF8_St5-12	CF8_St5-13	CF8_St5-16	CF8-6
Size (bps)	2,106,117	2,069,127	2,069,012	2,106,482	1,691,045	2,096,855	2,090,160	2,092,696	2,092,406	1,998,799
GC Content (%)	42	42	42	42	39.5	42	42.1	42	42	40.9
N50 (bp)	233,469	126,579	251,562	233,468	210,738	606,460	273,691	247,258	418,784	306,678
Number of Contigs	27	26	24	28	17	15	15	17	15	17
Number of protein Coding genes	2097	2045	2045	2097	1632	2014	2010	2005	2006	1984
Number of RNAs	48	49	47	49	47	48	48	47	47	28

Figure 3-1: Functional analysis of represented strains by RAST annotation.



3.5 Conclusion

For our study, restricted quality control steps were taken to ensure high coverage sequencing depth, good quality assembly, filtering assembled genomes with contaminations. A total of 40 different high quality *Streptococcus* strain genomes were generated. Only 2/3 strains were suitable for downstream analysis. This indicated the difficulties in culturing *Streptococcus* strains in practical. They were divided into 15 groups of strains based on genome distance to reference genomes in complete level. Further species identification steps need to be done to reveal the positions of these strains with reference genomes as a group.

The high-quality *Streptococcus* genomes in our study were further annotated by RAST. Only up to 31% CDSs were predicted with a function in database. This really hinders our understanding of bacterial genome by only use in silicon methods. This indicated the lack of information by studying microorganisms only use whole genome sequencing methods.

Other problems within studied strains can only be detected in downstream analysis of the assembled genome. Problems can be caused from bacterial isolation and culturing, DNA extraction and library preparation. Also, we can do little to face these problems if we encountered unless rerunning the whole process from scratch.

Chapter 4 Species identification

4.1 Introduction

We further revealed the relationships of our assembled genomes with reference genomes by using whole genome phylogenetic analysis and in-silicon species identification methods. Taxonomic analysis of isolated strains through phylogenetic tree was performed. Followed by two in silicon methods for species identification. The taxonomic position of each genome would represent positions of studied isolates in the whole Streptococcus genus as a whole group. Isolates in different species tend to have different characteristics. These two analyses were accomplished by a two-step whole genome nucleotide comparison analysis (Klenk and Goker 2010). Firstly, taxonomic position of each strain at the *Streptococcus* species level was determined by using whole genome single nucleotide polymorphism data. After this, in silico DNA-DNA hybridization analysis were performed to identify the species of each strain. We could expect strains in the same *Streptococcus* species with similar phenotypes, consume similar nutrition, suffer the same stresses and compete with other strains in a biofilm. By assign each genome into a specific *Streptococcus* species and compare genomes belong to different species, we could further reveal the differences between and within species.

4.2 Core genome single nucleotide polymorphism (SNP) phylogenetic analyses

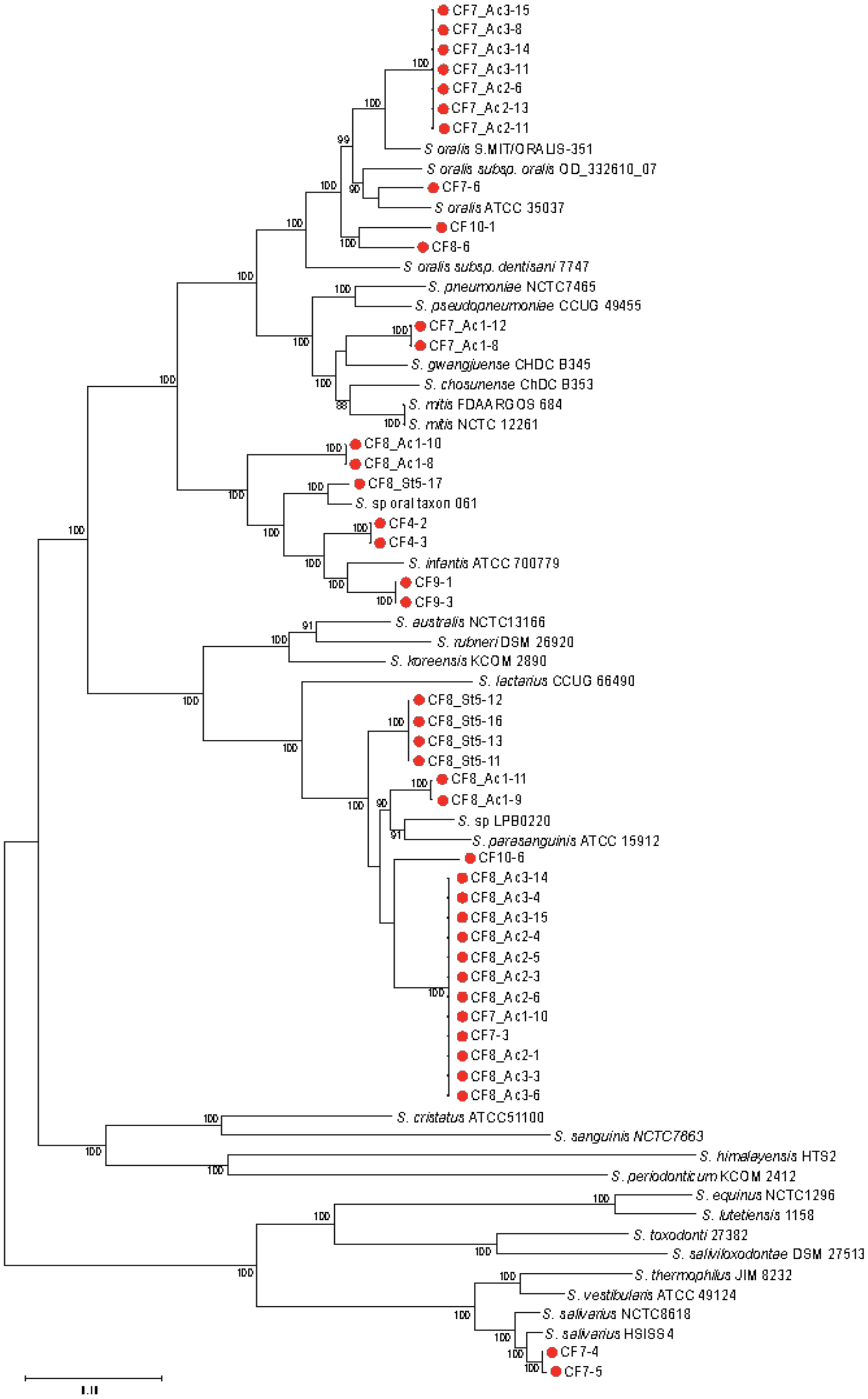
The taxonomic position of each strain was identified by constructing a phylogenetic tree using a robust whole genome approach. This was performed by comparing core genomic region single nucleotide polymorphisms between selected whole genomes. Reference whole genomes for phylogenetic analysis were choose by the similarity search of predicted 16s rRNA genes of our strains (Data not show) and the closest complete genomes from Referenceseeker results mentioned in chapter 3. Thus, a total of 71 genomes including 40 genomes from our samples and 31 reference genomes representing 23 different *Streptococcus* species were used to generate a robust phylogenetic tree.

Based on the phylogenetic tree (Figure 4.1), strains from our samples can be grouped into 2 groups. The first group was the *mitis* group (Richards, Palmer et al. 2014), which included 38 strains. These 38 strains were closely related to five different *Streptococcus* species, namely *S. oralis*, *S. mitis*, *S. infantis*, *S.* sp oral taxon 061, S. sp LPB0220 and *S. parasanguinis,* based on both Referenceseeker result and taxonomic analysis. The second group was the *salivarius* group containing only 2 of the newly sequenced strains, which were closest to *S. salivarius.*

The first group of 38 strains were divided into 4 subgroups and closely related to 4 *Streptococcus species*: *S. oralis, S. mitis, S. infantis, and S parasanguinis*. Strains with 0 distance in the same node of the whole gnome phylogenetic tree were strains from the same colony. Strains CF7_Ac2-6, CF7_Ac2-11, CF7_Ac2-13, CF7_Ac3-8, CF7_Ac3-11, CF7_Ac3-14 and CF7_Ac3-15 were from the same source and closely related to *S. oralis*. Similarly, strains CF7-6, CF10-1, and CF8-6 were isolated from different individuals but all were closely related to *S. oralis*. CF7_Ac1-12 and CF7_Ac1-8 were from the same sample and closely related to *S. mitis* with a 100% bootstrap value. A subgroup containing 7 strains were closely related to *S. infantis* and *S*. sp oral taxon 061. CF8_Ac1-10 and CF8_Ac1-8 were from the same colony, so were CF4-2 and CF4-3, CF9-1 and CF9-3. The isolate CF8_St5-17 was most closely related to *S*. sp *oral taxon 061*.

The last and largest subgroup of 19 strains were most closely related to *S. parasanguinis* and *S.* sp. LPB0220. These isolates further divided into 4 clusters. CF8_St5-12, CF8_St5-16, CF8_St5-13 and CF8_St5-11 were in the first cluster, CF8_Ac1-11 and CF8_Ac1-9 were in the second cluster, 12 strains CF7-3, CF7_Ac1-10, CF8_Ac2-1, CF8_Ac2-3, CF8_Ac2-4, CF8_Ac2-5, CF8_Ac2-6, CF8_Ac3-3, CF8_Ac3-4, CF8_Ac3-6, CF8_Ac3-14, and CF8_Ac3-15 were in the third cluster. Isolate CF10-6 clustered on its own. In the Salivarius group, two strains CF7-4 and CF7-5 were from the same sample and closely related to *S. salivarius*.

Figure 4-1: The taxonomic classification of 40 strains by whole genome phylogenetic analysis.



4.3 *In silico* DNA-DNA hybridisation

Further identification of these 40 isolates in species level was performed using golden standards in silico DNA-DNA hybridization methods. These are the average nucleotide identity (ANI) approach and Genome-to-Genome Distance Calculator (GGDC). These two methods are used to estimate the overall similarity between two genomes. Therefore, genomes of sample isolates were compared to the closest reference strains to identify how similar the closest genomes are. For ANI calculation, ANI values of over 96.00% and genome coverage of over 90% indicate that two isolates were from the same species (Goris, Konstantinidis et al. 2007). Using the genome-to-genome distance calculator (GGDC) tool, the accepted cut off for estimation of two isolates being the same species in a value of $\geq 70\%$ using formular 2. Strains with values over both thresholds would provide a confident indication of two genomes being derived from isolates of the same species.
To accurately identify all the genomes in species level. Initially, strains with zero distances at the same node of the phylogenetic tree were compared to observe their similarity. Secondly, representative genomes from the first step mentioned before were further analyzed with all genomes from closely related species.

4.3.1 The identification of rrepresentative strains from the same source by using in-silicon DDH methods

By using in-silicon DDH methods, we further prove the same recent origin of strains in several groups. The similarities by pair-wise comparison of strains from the same source were all over 99% by using both ANI and GGDC methods. Biologically, these strains contain the same genomes or only with very slight genomic changes. CF7-4 and CF7-5 were from the same source, with over 99.96% identity from both ANI and GGDC calculations. Similarly, CF7-3 and another 11 strains, CF7_Ac1-10, CF8_Ac2-1, CF8_Ac2-3, CF8_Ac2-4, CF8_Ac2-5, CF8_Ac2-6, CF8_Ac3-3, CF8_Ac3-4, CF8_Ac3-6, CF8_Ac3-14 and CF8_Ac3-15 were from the same source. CF7_Ac2-6 and

another 6 strains CF7_Ac2-11, CF7_Ac2-13, CF7_Ac3-8, CF7_Ac3-11, CF7_Ac3-14, CF7_Ac3-15 were from the source. CF8_Ac1-8 and CF8_Ac1-10 were from the same source. The other 5 groups of strains were in the same species too. The 1st groups were CF4-2 and CF-3. The 2nd was CF7_Ac1-8 and CF7_Ac1-12. The 3rd was CF8_St5-11, CF8_St5-12, CF8_St5-13 and CF8_St5-16. The 4th was CF8_Ac1-9 and CF8_Ac1-11. The last one was CF9-1 and CF9-3. The rest 5 strains CF10-6, CF7-6, CF8_St5-17, CF8-6 and CF10-1 have no other strains from the same species.
In summary, a total of 14 represented strains were selected for species identification. They were CF7-4, CF7-3, CF10-6, CF7-6, CF7_Ac2-6, CF8_Ac1-8, CF8_St5-17, CF4-2, CF7_Ac1-8, CF8_St5-11, CF8-6, CF8_Ac1-9, CF9-1 and CF10-1.

4.3.2 The species identification of representative strains from the studied 40 isolates

As previously stated, the 40 genomes were divided into 14 small groups based on phylogenetic analysis. By using in-silicon DDH methods, we further prove the same recent origin of strains in several groups.
As previously stated, 14 representative strains CF7-4, CF7-3, CF10-6, CF7-6, CF7_Ac2-6, CF8_Ac1-8, CF8_St5-17, CF4-2, CF7_Ac1-8, CF8_St5-11, CF8-6, CF8_Ac1-9, CF9-1 and CF10-1 were identified. Within these, seven isolates CF7-4, CF7-3, CF10-6, CF7-6, CF7_Ac2-6, CF8_Ac1-8, CF8_St5-17 were found to have corresponding closest strains within the same species. In another words, 7 groups of strains were found to have reported Refseq strains within the same species above pair-wise in-silicon DDHs thresholds mentioned before. The rest 7 isolates, CF4-2, CF7_Ac1-8, CF8_St5-11, CF8-6, CF8_Ac1-9, CF9-1 and CF10-1, were predicted as novel or previous unreported *Streptococcus* species (Table 4-1).
For isolated strains with a reported strain from the same species. CF7-4 and *S. salivarius* HSISS4 were from the same species. The CF7-3 group strains were belong to the same species with the representative strain of *S. parasanguinis* A1. The CF7_Ac2-6 group strains were belong to *S. oralis* 201_SPSE. The CF8_Ac1-8 group strains were belong to the *Streptococcus*

species with representative strain *S. infantis* SK1076. Another three strains CF10-6, CF7-6 and CF8_St5-17 were in different species groups with representative strains reported in Refseq as *S. parasanguinis* 392_SPAR, S. oralis 274_SPSE, and *S.* sp. oral taxon 061 F0704 separately (Table 4-2, Table 4-3).

Table 4-1: The in-silicon DDH estimation for pair-wise comparison between 14 representative strains with the closest strains in RefSeq.

# Groups	Representative	Strains	GGDC (%)	#ANIb [coverage] (%)	Remark
1	CF10-1	FDAARGOS_1021	58	94.39 [84.53]	Novel strains
2	CF9-1	X	61.8	95.12 [85.83]	
3	CF8_St5-11	349_SPAR	62.3	95.10 [86.67]	
4	CF8-6	FDAARGOS_1021	64.7	95.66 (89.60)	
5	CF4-2	SPAR10	66.2	95.79 [89.33]	
6	CF8_Ac1-9	AM25-15	66.5	95.77 [88.81]	
7	CF7_Ac1-8	SK1073	68.8	95.74 [93.06]	
8	CF8_Ac1-8	SK1076	71.3	96.60 [89.24]	With records of other strains belong to the same species
9	CF10-6	392_SPAR	72.5	96.56 [86.75]	
10	CF7-4	HSISS4	77.1	97.27 [94.25]	
11	CF7-6	274_SPSE	78	97.32 [90.97]	
12	CF7_Ac2-6	201_SPSE	82.1	97.90 [92.02]	
13	CF8_St5-17	F0704	84.1	98.08 [93.16]	
14	CF8_Ac2-1	A1	96.8	99.49 [93.63]	

Table 4-2: The in-silicon DDH estimation for pair-wise comparison between 14 representative strains with the predicted closest type species in RefSeq.

# Groups	Representative	Species	GGDC (%)	# ANIb [coverage] (%)
1	CF8_Ac1-8	*S. infantis*	38.1	89.19 [69.99]
2	CF8_St5-17	*S. infantis*	44.7	91.46 [73.40]
3	CF4-2	*S. infantis*	56.5	94.04 [82.03]
4	CF9-1	*S. infantis*	60.3	94.61 [76.41]
5	CF7_Ac1-8	*S. mitis*	54.9	93.44 [76.37]
6	CF7-6	*S. oralis*	59.4	94.62 [85.44]
7	CF10-1	*S. oralis*	56.1	93.87 [79.62]
8	CF7_Ac2-6	*S. oralis*	57	94.20 [80.66]
9	CF8_Ac2-1	*S. parasanguinis*	55.2	93.43 [73.89]
10	CF10-6	*S. parasanguinis*	54.8	93.04 [72.44]
11	CF8_Ac1-9	*S. parasanguinis*	55.5	93.46 [77.80]
12	CF8_St5-11	*S. parasanguinis*	54.5	93.42 [76.97]
13	CF8-6	*S. parasanguinis*	57.7	94.21 [83.68]
14	CF7-4	*S. salivarius*	64.8	95.42 [87.56]

Table 4-3: The in-silicon DDH estimation for pair-wise comparison between 13 reference genomes and the predicted closest type species in RefSeq.

#	Strain	Best-match type	GGDC	# ANIb [Coverage]
1	FDAARGOS_1021	Streptococcus oralis	57.7	94.29 [83.49]
2	201_SPSE	Streptococcus oralis	58.5	94.47 [79.56]
3	274_SPSE	Streptococcus oralis	60.5	94.79 [83.61]
4	SK1073	Streptococcus mitis	57	93.84 (73.65)
5	HSISS4	Streptococcus salivarius	65.9	95.64 [86.25]
6	SPAR10	S. infantis ATCC 700779	57	94.22 [74.19]
7	X	S. infantis ATCC 700779	57.6	94.13 [78.96]
8	F0704	S. infantis ATCC 700779	44.7	91.36 [73.69]
9	SK1076	S. infantis ATCC 700779	38.6	89.18 [71.08]
10	349_SPAR	S. parasanguinis ATCC15912	55.2	93.56 [78.65]
11	AM25-15	S. parasanguinis ATCC15912	56.2	93.46 [77.80]
12	A1	S. parasanguinis ATCC15912	55.3	93.46 [74.14]
13	392_SPAR	S. parasanguinis ATCC15912	55.2	93.57 [73.97]

4.4 Conclusion

A total of 14 represented strains were selected from these 40 strains based on both phylogenetic analysis and dDDHs methods. Seven out of 14 strains were predicted to present in a known species with reported strains in Refseq database. The other 7 strains were from novel species without any previously reported strains in database.

From previous published studies, to identify the species of one strain, pair-wise comparison of genomes between studied genome with predicted type strains were performed. Interestingly, by doing this, all the pair-wise dDDH values were below the thresholds (Table 4-2). This means all strains in our study were formed different novel species. We refine the process to include all genomes in reported species. Thus, 7 of 14 of our selected strains were from a known species with a reported reference strain. The other 7 strains CF4-2, CF7_Ac1-8, CF8_St5-11, CF8-6, CF8_Ac1-9, CF9-1 and CF10-1 formed 7 group of novel species.

For 7 known species strains, we identified 7 closest reference genomes within the same species group in database. But the species identification of these 7 reference genomes were not convinced by using our refinement methods (Table 4-3). Thus, we can only conclude that these 7 group strains CF7-4, CF7-3, CF10-6, CF7-6, CF7_Ac2-6, CF8_Ac1-8 and CF8_St5-17 had reference genomes within the same species. In contrast, the other 7 group strains CF4-2, CF7_Ac1-8, CF8_St5-11, CF8-6, CF8_Ac1-9, CF9-1 and CF10-1 were identified as novel species because no reported reference genomes within the same species. The identification of the novel species strains will contribute to the research of the new *Streptococcus* strains from CF samples.

Chapter 5 In-silico characterization and comparative analysis of novel species strains

5.1 Introduction

Important bacterial genomic features including the CRISPR-Cas system, the restriction-modification system, plasmids, antibiotic resistance genes, prophages, genomic islands, and virulence factors were predicted to in novel species strains to identify coding sequences (CDSs) encoding important phenotypes. Variation between different strains from the same novel species groups were analysed to predict CDSs with minor differences in genomes of the same novel species group.

Bacteria developed an adaptive immune system called CRISPR-cas system to protect foreign genomic sequences invasion (Barrangou, Fremaux et al. 2007). These foreign genomic sequences called genomic islands were from phages, and different mobile genetic elements. The system is thought to be a result of parallel evolution of the bacterial immune system between bacteria and GIs (Shabbir, Hao et al. 2016). The restriction-modification system is another immunity system for bacteria (Rodic, Blagojevic et al. 2017). Plasmids were short extrachromosomal nucleotide sequences in bacteria cells. They play an important role in bacterial evolution and manipulation of bacterial phenotypes (Billane, Harrison et al. 2022). Genes in plasmids often show genetic advantages such as antibiotic resistance. They were transferred by plasmid conjugation to other bacterial lineages. Prophages may carry new genes that play important roles in the acquisition of new traits and the generation of genetic diversity (Pallen and Wren 2007). GIs in bacteria harbour genes encoding important traits such as antibiotic resistance, symbiosis and fitness (Dobrindt, Hochhut et al. 2004). The pathogenicity of a bacterial was determined by the virulence factors in the genome.

The Different genomic features in a total of 14 strains in 7 groups of *Streptococcus* species were analysed. By comparative study of these different genome features above, we can reveal the difference and similarity between these novel species strains.

Comparative Genome Island (GI) analysis. Streptococci encounter significant fluctuations in environmental conditions such as surrounding pH, oxygen tension or osmolarity when growing on the surface of organs of CF patients. The transition to the bloodstream environment involves an even greater shift in the conditions of the external environment. We postulated that the adaptation and evolution of streptococci to cope with different environments within the human body may have been mediated through the acquisition of gene clusters or GIs by horizontal gene transfer. Typically, Therefore, horizontally transferred GIs in these strain genomes were predicted using the IslandViewer software tool (Langille and Brinkman 2009).

5.2 Comparative analysis of different genomic features

5.2.1 Analysis of the CRISPR-Cas system

Only CF8-6, CF9-1 and CF9-3 were predicted with a complete set of CRISPR-Cas system (Figure 5-1). Other two strains CF7-6 and CF10-1 were only predicted with two truncated CRIPSR arrays (Table 5-1). According to the Cas proteins, the CRISPR-Cas systems predicted in these 3 strains were class II. We may observe less genomic islands in CF9-1, CF9-3 and CF8-6 strains considering the stability of genomes provided by CRISPR-Cas system.

5.2.2 Analysis of the restriction-modification system (RM)

The total presence of predicted single genes or operons related to separate RM systems in these 7 novel species groups is 22. For strains in the same novel species group, the same RM system components were observed. Thus, the 7 novel species groups were represented by 7 strains: CF8_Ac1-9,

Figure 5-1: Illustration of the complete CRISPR-Cas systems in 3 of 14 novel strains. The number and orientation of each Cas gene is also drawn. The lengths of cas genes and CRISPR arrays are also labelled.

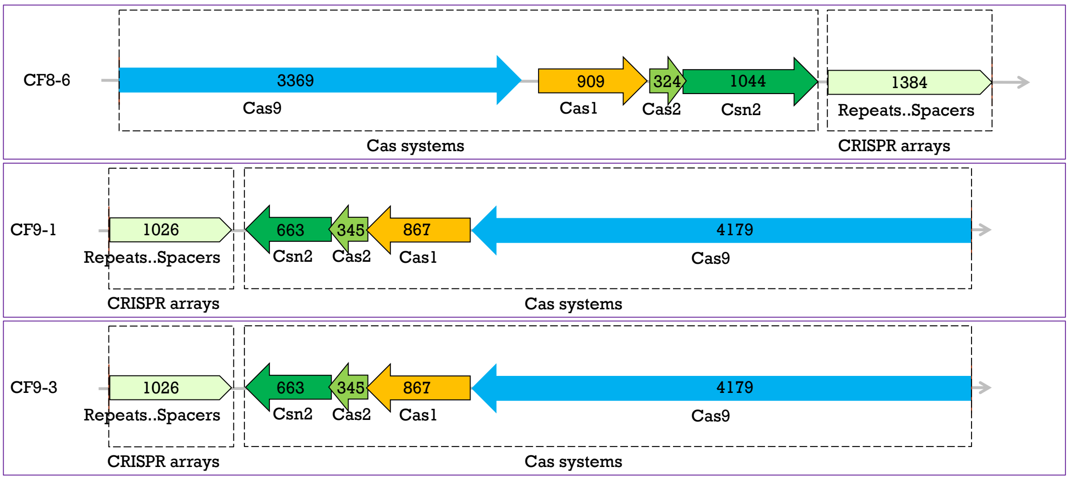
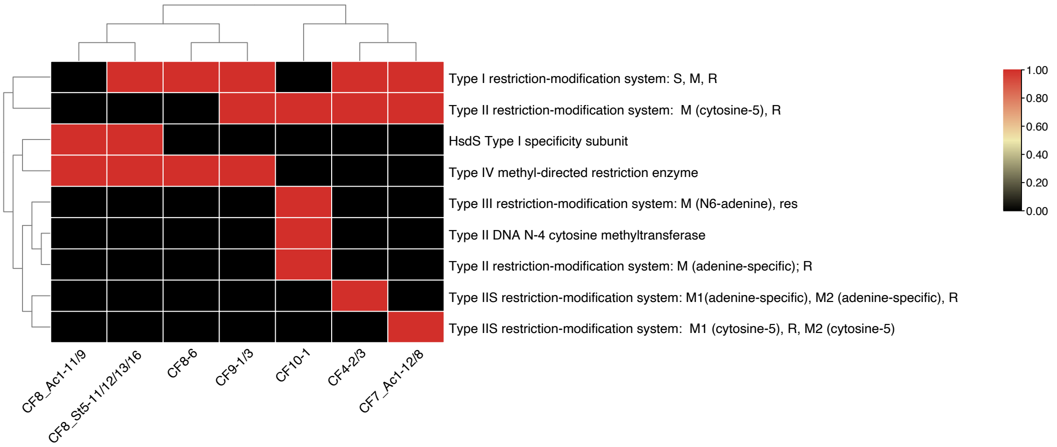


Table 5-1: The predicted CRISPR-Cas systems in 14 novel strains.

Strain	CRISPR_Length	Repeat_Length	Spacers_Nb	Evidence_Level	Number of Cas genes
CF8_6	1420	36	21	4	4
CF9-3	1025	36	15	4	4
CF10-1	816	35	12	4	N/A
CF7-6	295	35	3	2	N/A

CF8_St5-11, CF8-6, CF10-1, CF9-1, CF4-2 and CF7_Ac1-8. Although 4 types of RM systems were observed in these novel species. The number of a particular RM type was different in different novel species (Table 5-2). The type 3 RM system was only observed in CF10-1. Other RM systems were distributed in at least 4 novel species groups. Strains in CF8_Ac1-9 novel species group only contain 1 type IV RM system. The other novel species strains contain 2 types of RM systems except CF9-1 with the existence of three RM systems. Different distribution of the same RM system was observed in the closely related novel specie groups. CF10-1 and CF8-6 were closely related compared to other novel species strains with distinct types of RM systems. CF10-1 contained a total of 5 RM systems, 3 type II and 2 type III RM systems. While CF8-6 only contained 1 type I and 1 type IV. CF8_St5-11 and CF8-6 were closely related to two different *Streptococcus* sp. with the same number and type of RM systems. The existence of a particular type of RM system seemed random in species level.

Figure 5-2: Comparison of predicted the RM systems components.



5.3.3 Analysis of the plasmids

Plasmids were predicted to be existed in five strains CF4-2, CF4-3, CF9-1, CF9-3, and CF10-1 based on the identification of the replicons. CF4-2, CF4-3 and CF10-1 were predicted to have the plasmid repUS43. The reference for these three strains was *Enterococcus faecium* DO plasmid 1 (DOp1, CP003584) (Table 5-2). CF9-1 and CF9-3 were predicted to have the plasmid repUS38. The reference for these two strains was pFW213 (EU685104) (Chen, Shieh et al. 2011).

By tracing back, the replicons into sequences in each strain, comparative analysis of conserved genes in sequences with plasmid origin between novel strains and their corresponding references were performed and highlighted in numbers with red (Figure 5-3). The reference DOp1 is 36262 bp with 35 predicted coding sequences containing replicon Rep_trans (Balson and Shaw 1990). Three contigs containing the replicon Rep_trans in CF4-2, CF4-3 and CF10-1 are the pCF4-2 (CF4-2 contig 11, 38967bp, 44 protein coding sequences (CDSs), JANCPO010000011), the pCF4-3 (CF4-3 contig 11, 20849 bp, 25 CDSs, JANCPN010000011), and pCF10-1 (CF10-1 contig 1, 464941 bp, 463 CDSs, JANCPQ010000001). A total of 13 coding sequences were conserved in these 4 nucleotide sequences. The predicted phenotype conferred by these 4 sequences is antibiotic resistance by *tet(M)*. An important conserved gene antirestriction *ardA*, which encodes ArdA, facilitating the envasion of plasmids in bacteria genome (Nekrasov, Agafonova et al. 2007). At least 7 conjugal transfer proteins were predicted. One conjugative transposon is predicted as YtxH domain-containing protein. This gene contains a SLC5-6-like_sbd region. SLC5 proteins co-transport Na+ with sugars, amino acids, inorganic ions or vitamins (Kristensen, Andersen et al. 2011). A total of 25 coding sequences were conserved in *Streptocococus* strains CF4-2, CF4-3 and CF10-1. Interestingly, another antibiotic resistance gene *erm(B)* is only shared by CF4-2, CF4-3 and CF10-1 strains (Figure 5-3).

Plasmids were conserved in strains from the same species. The presence and absence of certain regions in predicted conserved plasmids from different species strains may result from the activities of lost or gain CDSs after conjugative transfer. Conserved circular plasmids were observed in CF9-1 and CF9-3 and *S. parasanguinis* FW213 strain (Figure 5-4).

Table 5-2: The existence of plasmids predicted in novel strains by the identification of replicons in each strain.

Assembly	Database	Plasmid	Identity (%)
4.3 (4.2)	Rep_trans	repUS43	100
9.1 (9.3)	Rep3	repUS38	100
10.1	Rep_trans	repUS43	99.83

Figure 5-3: Synteny analysis of conserved genes by comparing the predicted genes in four nucleotide sequences with plasmid origin from strain CF4-2, CF4-2, CF10-1 and Enterococcus faecium DO.

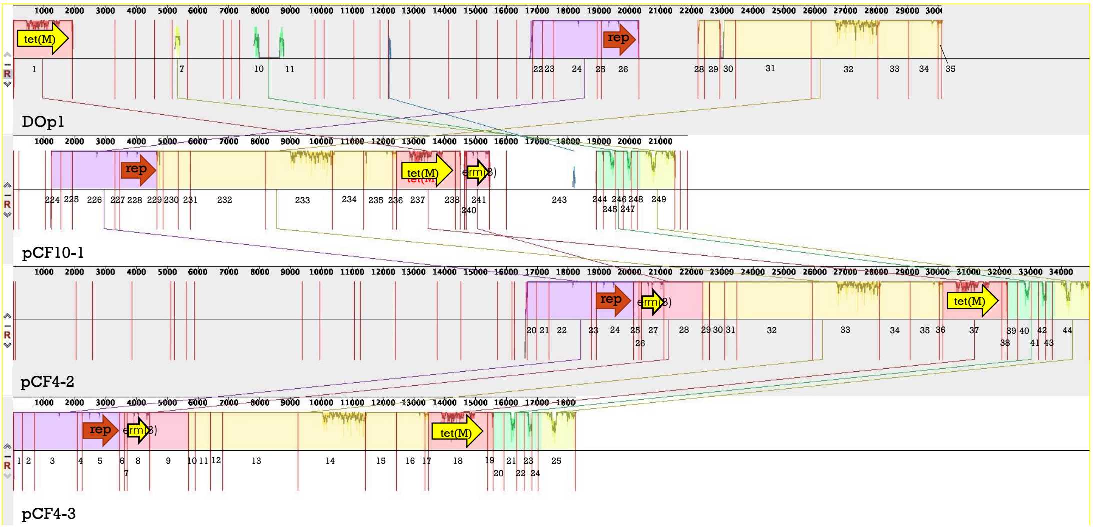
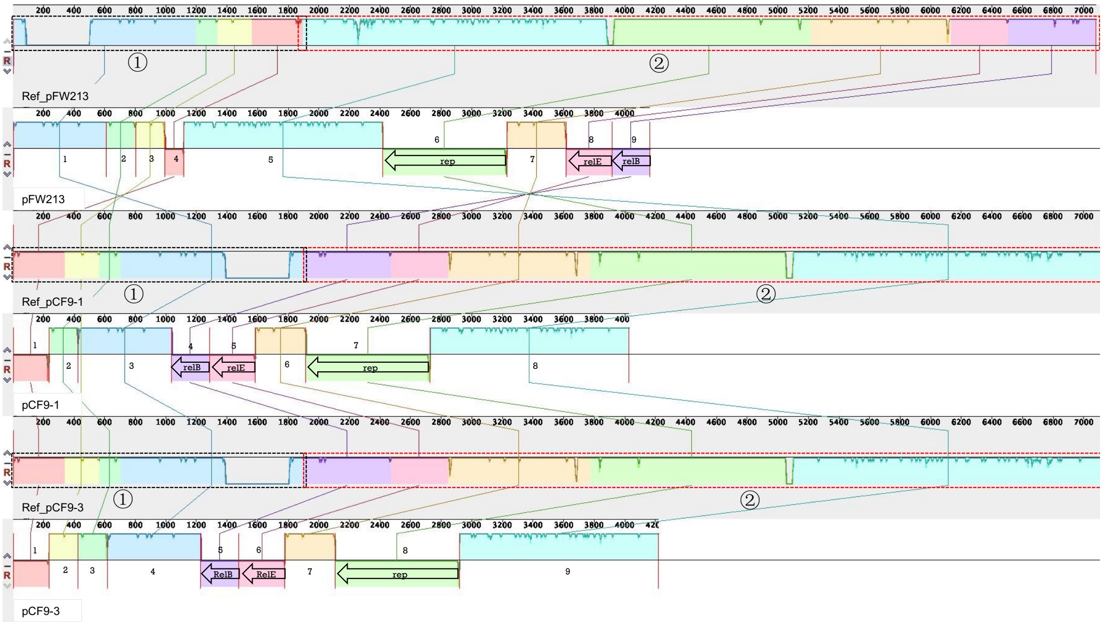


Figure 5-4: Synteny analysis between 3 predicted plasmid nucleotides and the 3 corresponding predicted gene in nucleotide form.
Two large LCBs were predicted in three plasmid nucleotides. For the 3 corresponding gene sequences, each gene was separated by two vertical red lines. Genes below the middle line indicated the reverse direction of the genes in the whole nucleotides. Because the reference plasmid is a circular one. The other two putative plasmids were also circular ones based on the composition of the conserved genes and the orientation of the genes.

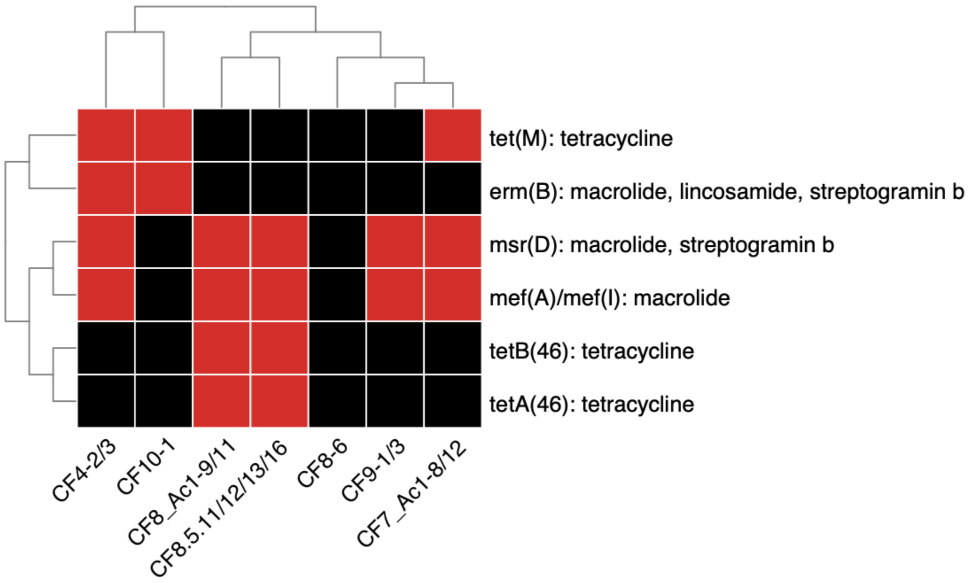


5.3.4 Comparative analysis of the antibiotic resistance genes and the corresponding phenotypes

Strains in the same novel species group were predicted to contain the same antibiotic resistance genes. A total of 6 antibiotic resistance genes were predicted (Figure 5-5). CF8-6 was predicted to contain no antibiotic resistance genes. Three genes *tet(M)*, *tetB(46)* and *tetA(46)* confer tetracycline resistance. It seems that strains in our study confer tetracycline by either the expression of a single *tet(M)* or two genes (*tetB(46)* and *tetA(46)*). This may indicate that the cooperation of both TetB(46) and TetA(46) in tetracycline resistance. Three other predicted genes *erm(B)*, *msr(D)*, and *mef(A)* confer macrolide, lincosamide, and streptogramin b resistance. CF10-1 with only two genes *tet(M)* and *erm(B)* was predicted to confer resistant to the four kinds of antimicrobial drugs mentioned before. CF4-2 novel species group also confer the resistance to the same four kinds of antimicrobial drugs mentioned before but with 4 genes *tet(M)*, *erm(B)*, *msr(D),* and *mef(A).* Closely related novel species represented by CF8_Ac1-

9 and CF8_St5-11 were predicted to contain the same antibiotic resistance genes conferring the same resistance to macrolide, streptogramin b and tetracycline.

Figure 5-5: Comparison of predicted antibiotic resistance genes in 14 novel strains.



Resistance testing of 7 different antimicrobial chemicals were also performed on these 14 strains in our study (Table 5-3). These 7 chemicals belong to 7 different family of commonly used antibiotics for gram-positive and gram-negative bacterial strains (Table 5-4). Only chemicals belonged to tetracycline, lincosamide, glycopeptide, naphthyridone and Monobactams were resisted in some strains. No macrolide resistance was detected. Tetracycline was resisted only in CF4-3. Both CF4-3 and CF10-1 showed resistance to linosamide. Glycopepetide resistance was only detected in CF9-3. Naphthyridone was resisted in all strains except. Monobactams resistance was conferred by all strains.

Differences were observed by comparing both in-silicon antibiotic resistance prediction and chemical testing of commonly used antimicrobial chemicals. Strains containing antibiotic resistance genes cannot confer resistance for expected antimicrobial chemicals. CF4-3 with the three tetracycline resistance genes conferring resistance to Minocycline may indicate the cooperation of multiple genes in bacterial antibiotic resistance.

Table 5-4: Biochemical tests of antibiotics.

Well Contents	CF4-2	CF4-3	CF9-1	CF9-3	CF8_Ac1-11/9	CF8_St5-11/12/13	CF8_St5-16	CF8-6	CF10-1	CF7_Ac1-8	CF
Troleandomycin	N/A	N/A	N/A	N/A	-	-	N/A	N/A	N/A	-	
Rifamycin SV	-	-	N/A	N/A	-	-	N/A	N/A	N/A	-	
Minocycline	N/A	+	N/A	N/A	-	-	N/A	N/A	N/A	-	
Lincomycin	N/A	+	N/A	N/A	-	-	-	N/A	+	-	
Vancomycin	-	N/A	N/A	+	-	-	N/A	N/A	N/A	-	
Nalidixic Acid	+	+	+	+	+	+	+	+	-	+	
Aztreonam	+	+	+	+	+	+	+	+	+	+	

Positive (+), negative (-) and N/A represent resistant to, surrender to and hard to decide whether affect cell growth with the existence of certain antibiotics.

Table 5-3: The classification of chemicals into different antibiotic families.

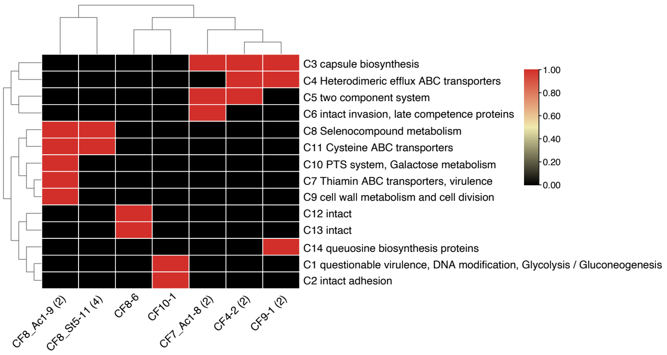
Well Contents	Family	Antibiotic resistance genes	Strains (14 in total)	Range	Reference
Troleandomycin	macrolide	erm (B), mef (A), msr(D)		G+	Gürel et al., 2009
Rifamycin SV	ansamycins	/	/	G+ (G-)	Lin et al., 2017
Minocycline	tetracycline	tet (M), tetB(46), tetA (46)	9	G+, G-	Alano et al., 2006
Lincomycin	lincosamide	erm (B)	CF10-1, CF4-2 (2)	G+	Macleod et al., 1964
Vancomycin	glycopeptide	/	/	G+	Liu et al., 2011
Nalidixic Acid	naphthyridone	/	/	G-	Emmerson et al., 2003
Aztreonam	Monobactams	/	/	G- (G+)	Quon et al., 2014

5.3.5 The prediction and comparative analysis of prophages

Prophages were predicted in the 14 novel species strains by PHASTER. Strains in the same novel species groups contained the same prophages contents after manual curation although some showed different prediction results. For novel species groups with multiple strains, we only use predicted prophages results in one strain to represent the whole novel species group. A total of 20 prophage regions were clustered into 14 non-redundant prophages and compared (Figure 5-6). By comparing the number of certain phage organism and the total number of CDS in the same region, PHASTER marked predicted prophages with incomplete, questionable and intact levels to indicate the prophage completeness. Only 3 intact and 2 questionable prophages were predicted. They were either CF8-6 strain specific or CF10-1 specific prophages. The existence of the other 4 species-specific prophages were also observed in rest novel species groups. Only cluster 3 prophages were shared by three novel species groups. The rest 4 clusters of prophages were shared by only two novel species groups. The total number of predicted prophages varied differently in novel species groups with CF8_Ac1-9 contained 5 incomplete prophages. While CF8-6 and CF10-1 contained only 2 prophages separately.

Although CF8-6 contained two intact prophages, no CDS were annotated to confer essential roles. In the contrast, the two intact prophages in CF10-1 can provide the strain with virulence factors and participate in many metabolisms. Other prophages with different length were predicted to confer useful phenotypes in studied strains.

Figure 5-6: The distribution of predicted prophages in 7 novel species groups.



5.3.6 The prediction and comparative analysis of genomic islands

A total of 97 GIs were predicted in single contigs of strains from the 7 novel species groups. Not surprisingly, strains in the same group shared the same contents of predicted genomic islands.

For the novel species group containing CF4-2 and CF4-3, a total of 13 different genomic islands were predicted and shared (Figure 5-7). A total of 4 GIs were predicted to have a phage origin by the existence of phage genes. These GIs may also contain CDs involve in different metabolisms and antibiotic resistance including benzoate degradation, short chain fatty acid metabolism, pyruvate metabolism and nisin-resistance. Cds in other GIs were annotated as proteins in bacterial IV secretory system, peptidases, type II R-M system, ABC transporters and quorum sensing in lantibiotic biosynthesis response. It looks like part of the 1st and all part of 2nd GIs from CF4-2 were missing in CF4-3. Further curation confirmed the existence of the missing GIs in CF4-3.

For novel species group containing CF7_Ac1-8 and CF7_Ac1-12, these two strains shared 14 GIs (Figure 5-8). Five GIs were regions of prophages. Most of the cds in the 4 prophage regions were phage related proteins indicating the newly acquired of these prophages. Only 1 prophage region with phage associated CI-like repressor contained cds in maltose metabolism. One GI was from the transposon 916 with antibiotic resistance genes *tet(M)*, *msr(D)* and *mef(A)* and conjugative transposon protein. Other GIs were predicted with the virulence factors including various membrane proteins, IgA-specific metalloendopeptidase (EC 3.4.24.13), C5a peptidase precursor (EC 3.4.21.-) and cds in various carbohydrate metabolisms including fructose, mannose induced PTS system components and galactose metabolism.

A total of 10 GIs were shared by CF9-1 and CF9-3 (Figure 5-9). Two GIs were regarded as prophage regions because of the prediction of phage related proteins. One of these GIs contained cell divisome proteins FtsA, FtsZ, DivIVA, Isoleucyl-tRNA synthetase (EC 6.1.1.5) involved in aminoacyl-tRNA biosynthesis, and Phosphoglycerate mutase (EC 5.4.2.11) in various metabolisms. The rest 8 GIs were predicted to be involved in chromosome partitioning, Phenylalanine, tyrosine and tryptophan biosynthesis, antibiotic resistance, purine metabolism and the bacterial immune system.

For CF8_St5-11, CF8_St5-12, CF8_St5-13 and CF8_St5-16, these strains shared 16 GIs (Figure 5-10). A total of three slight differences occurred in three GIs. For a shared GI with 10848 bp long, the only difference was the absence of a 152 bp long hypothetical protein in CF8_St5-13. For a 32899 bp GIs, two neighbour hypothetical proteins were predicted in only CF8_St5-11, CF8_St5-13 and CF8_St5-16 but missing in CF8_St5-12. Lastly, a shared 4585 bp GIs containing a 146 bp hypothetical protein were predicted in only CF5_St5-11 and CF8_St5-16. Thus, GIs in CF8_St5-11 were chosen to represent the CF8_St5-11 species group strains. These 16 GIs contained cds encoding proteins annotated as NG,NG-dimethylarginine dimethylaminohydrolase 1 (EC 3.5.3.18), Ornithine racemase (EC 5.1.1.12), ABC transporters, Toxin HigB, Antitoxin HigA, Type I restriction-modifcation system enzyme RSM, Rhoptry protein, Type II restriction and modification enzymes, Phosphoribosylanthranilate isomerase like (EC 5.3.1.24), cyanophycinase [EC:3.4.15.6], Lantibiotic ABC transporters, antibiotic resistance Mef(A), Msr(D), NisR, NisK, and Glutathione biosynthesis bifunctional protein gshF (EC 6.3.2.2)(EC 6.3.2.3). These proteins were predicted to be involved in D-amino aicd metabolism, bacterial pathogenicity, bacterial immunity, tryptophan biosynthesis, Cysteine and methionine metabolism and Glutathione metabolism.

Interestingly, in novel species group containing CF8_Ac1-9 and CF8_Ac1-11, one part of the GI, named CF8_Ac1-11_G13, had an extra 5238 bp long nucleotides in CF8_Ac1-11 containing a CAAX amino terminal protease family protein and the Xre family transcriptional regulators (Figure 5-11). Cds in other shared regions of this GI were antibiotic resistance related genes encoding Mef(A) and Msr(D), which responsible for antibiotic resistance. Other 12 GIs were shared by these two strains. In these GIs, 7 GIs were prophage related by the identification of phage related proteins, specifically phage integrases. The cds in these 7 GIs were annotated as the Xre family transcriptional regulator, Efflux ABC transporters, DNA primase-like protein, the MutR family positive transcriptional regulator, UDP-N-acetylglucosamine kinase (EC 2.7.1.176), Epsilon antitoxin to Zeta toxin, D-alanine--D-alanine ligase (EC 6.3.2.4), Peptidoglycan hydrolase, Autolysin2 (EC 3.5.1.28), GTP-binding protein EngB, ATP-dependent Clp protease ATP-binding subunit ClpX, Dihydrofolate reductase (EC 1.5.1.3), Thymidylate synthase (EC 2.1.1.45), Glucokinase (EC 2.7.1.2), Bis-ABC ATPase Uup, CCA tRNA nucleotidyltransferase (EC 2.7.7.72), 4-hydroxy-tetrahydrodipicolinate reductase (EC 1.17.1.8), the XRE family pleiotropic regulator, virulence-associated protein E, the phage associate CI-like repressor, , the MerR family transcriptional regulator, the short-chain dehydrogenase/reductase family oxidoreductase, the Fic/DOC family protein. These proteins participated in quorum sensing, thiamine biosynthesis, pathogenicity, D-Amino acid metabolism, methionine metabolism, pyrimidine metabolism, Glycolysis / Gluconeogenesis, Galactose metabolism, Starch and sucrose metabolism, lysine biosynthesis. Cds in the rest 5 GIs were annotated as the MutT/Nudix family protein, the GNAT family acetyltransferase, the PEP-utilizing enzymes family protein, and transcriptional regulatory protein NisR, the permuted papain-like amidase enzyme YaeF/YiiX. These proteins played a role in carbohydrate phosphotransferase system (PTS), lantibiotic nisin biosynthesis and pathogenicity.

For CF10-1, a total 17 genomic islands were predicted. Among these, 10 were phage related because of the prediction of phage proteins, especially the prediction of phage integrases. For these 10 phage related GIs, two GIs encoding late competence proteins and various putative virulence factors involved in cell motility, intracellular trafficking, secretion, and vesicular transport, coenzyme metabolism and carbohydrate metabolism, one phage contained a plasmid recombination enzyme indicating the phenomenon invasion of phage to the bacterial plasmid, two GIs contained gene operons encoding epsilon antitoxin to zeta toxin, one GI revealed the ABC transporters involved in inorganic ion transport metabolism. For the rest 7 GIs, one contained mostly mobile genetic elements, one contained only putative glycosyltransferases involved in cell wall, membrane and envelope biogenesis, two contained only hypothetical proteins, one contained antibiotic resistance genes involved in translation, ribosomal structure biogenesis and proteins involved in replication, recombination, repair and Cell wall, membrane and envelope biogenesis, one contained a valyl-tRNA synthetase (EC 6.1.1.9) and a superfamily I DNA/RNA helicase protein involved in translation, ribosomal structure biogenesis, replication, recombination and repair, one contained mostly bacteriocin immunity proteins.

For CF8-6, a total of 14 GIs were predicted. Similar to CF10-1, 8 GIs were predicted with a phage origin. Cds in these 8 GIs were annotated as the AcrR family transcriptional regulator, the competence regulon ComE and ComD, competence-stimulating peptide (CSP), the FtsK/SpoIIIE family protein, the Cro/CI family transcriptional regulator, DNA-cytosine methyltransferase (EC 2.1.1.37), N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28), the ArpU family transcriptional regulator. These proteins were involved in bacterial competence, DNA segregation, the persistence of phage, type II R-M system, phage lysin, enhance pathogenicity. Cds in the rest 6 GIs were annotated as Type III restriction-modification system methylation subunit (EC 2.1.1.72), the arsR family regulatory protein, Chaperone protein ClpB, GNAT family acetyltransferase BA2701, Undecaprenyl-phosphate galactosephosphotransferase (EC 2.7.8.6) RfbP, CRISPR-associated protein Csn2, Cas1, Cas2, and Type IV secretory system Conjugative DNA transfer protein. These CDs may involve in type III restriction-modification system, cadmium efflux system, antibiotic resistance, antiviral system and plasmid transfer.

By combining these 97 GI sequences into one large sequence, we performed pair-wise sequence alignment. A total of 1457 pairs of regions in these GIs were found by sequence-to-sequence Blastn search. The matched regions in these GIs were from either the intergenic regions, the coding sequence regions (CDs) or the combinations of both intergenic and CDs regions. By manually curation, only 84 matches representing 28 similar sequences with annotations from 34 GIs were identified (Figure 5-12). These shared regions were related to different phenotypes in bacteria, including antimicrobial resistance, rRNA modification, arginine and proline metabolism, biofilm formation, quorum sensing, two component system, Phenylalanine, tyrosine and tryptophan biosynthesis, toxin-antitoxin system, Aminoacyl-tRNA biosynthesis and enhance pathogenicity.

Figure 5-7: The comparison of predicted genomic islands in CF4-2 and CF4-3.
Most of the genomic islands predicted in this group share the same coding sequences and compositions. #: Some of coding sequences were the same in this group. But CF4-2 contained extra coding sequences. *: This region only predicted in CF4-2.

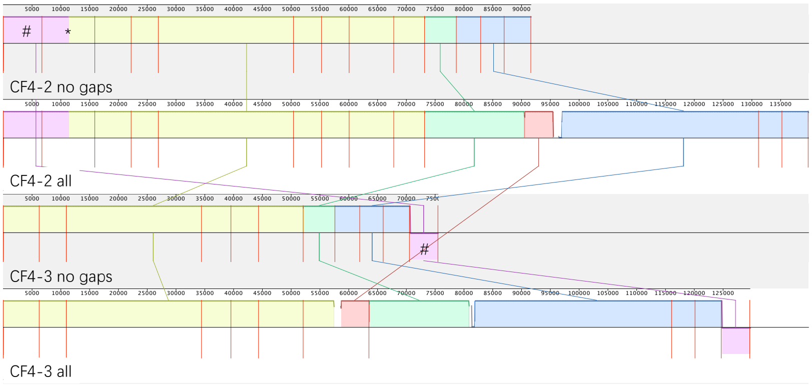


Figure 5-8: CF7_Ac1-8 and CF7_Ac1-12 shared the same genomic islands.

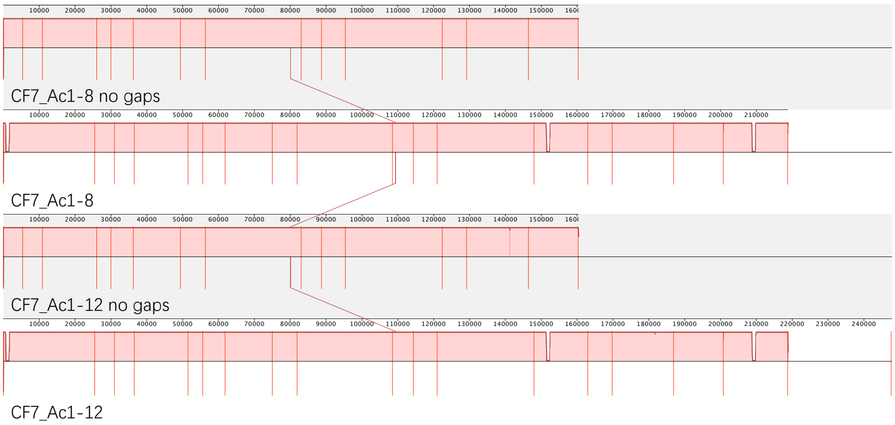


Figure 5-9: The comparison of predicted genomic islands in CF9-1 and CF9-3.

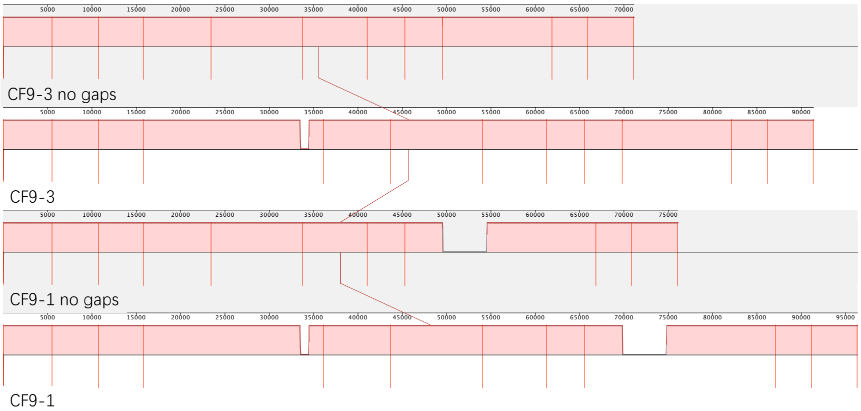


Figure 5-10: The comparison of predicted genomic islands in strains CF8_St5-11, CF8_St5-12, CF8_St5-13 and CF8_St5-16.
Only CF8_St5-11 contained a genomic island from two contigs. The four strains shared the same genomic islands. The only difference laid in the missing of some coding sequences in CF8_St5-12 (#).

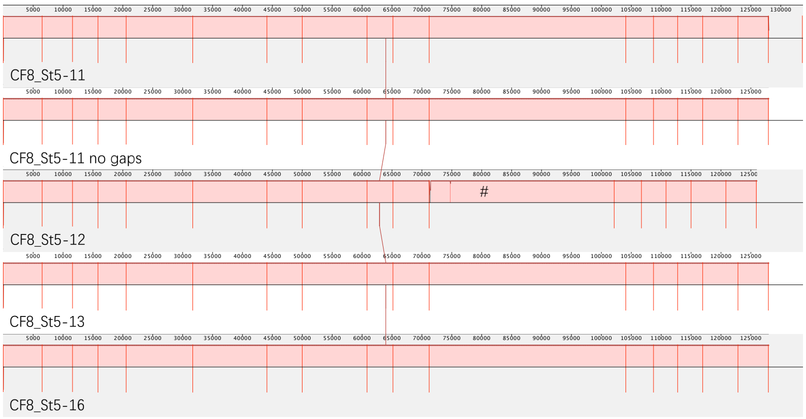


Figure 5-11: The comparison of predicted genomic islands in CF8_Ac1-9 and CF8_Ac1-11. The coding sequences in GIs from CF8_Ac1-11 can be fully represented by the gis in CF8_Ac1-9 because gis in CF8_Ac1-9 covered most of the cds in CF8_Ac1-11. Also, GIs in CF8_Ac1-9 could be use as representatives for this group due to the total longer length of gis in these two strains. #: regions with this mark represent the difference of coding sequences in a shared GI. *: regions with this mark represent gis from two contigs and connected by Ns.

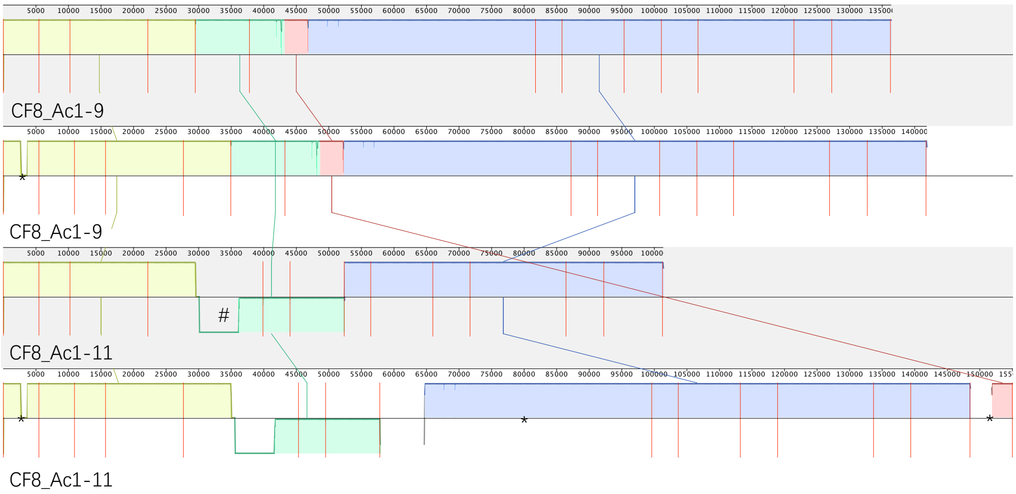
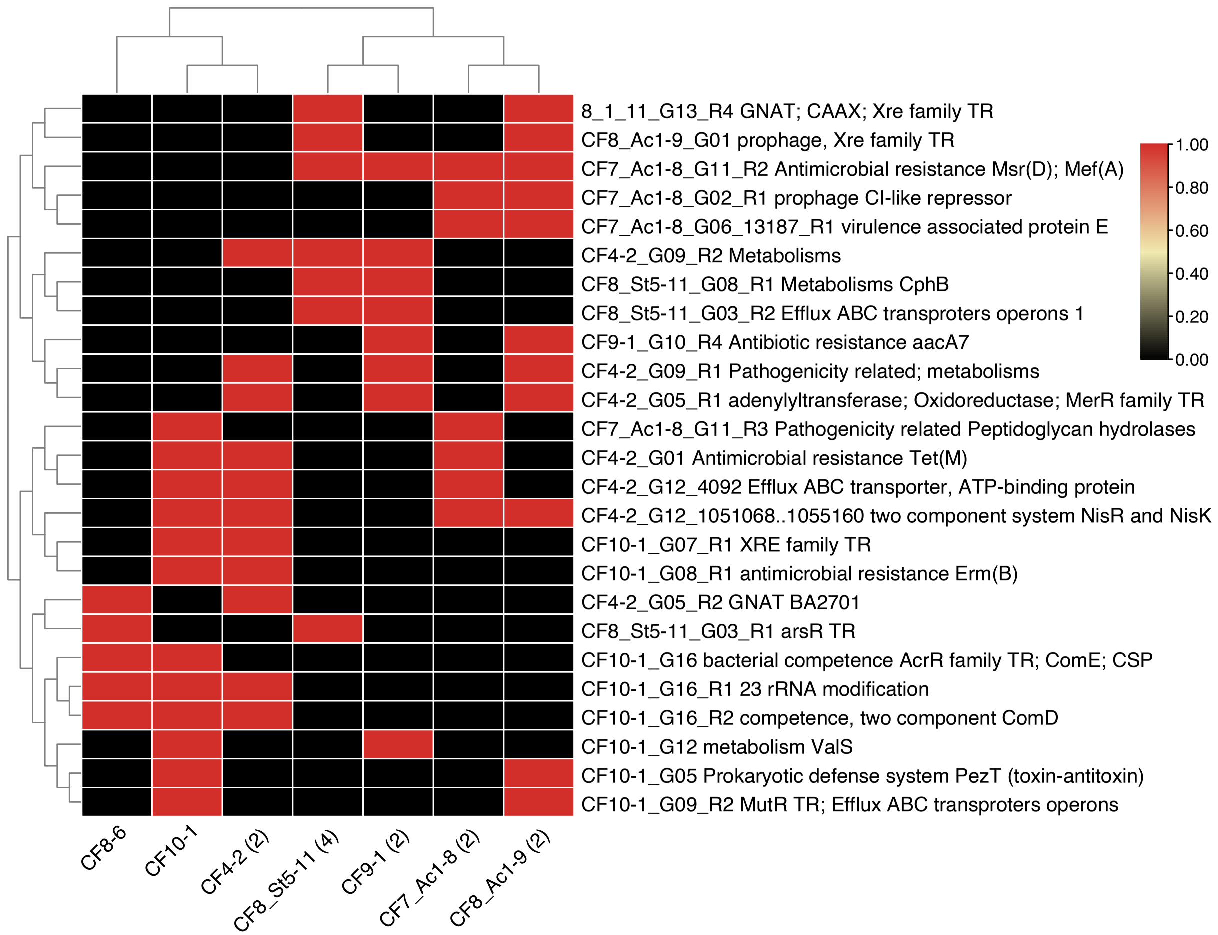


Table 5-5: Comparative analysis of genomic islands in novel species isolates.

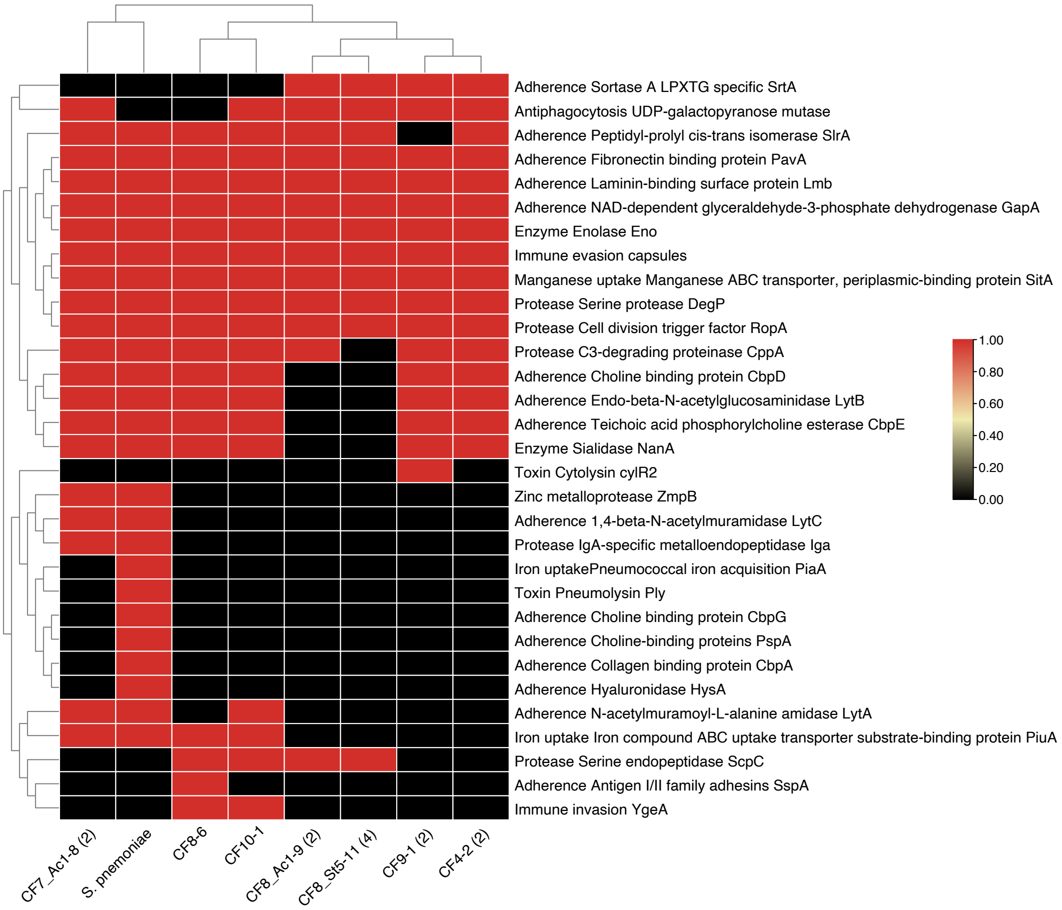


5.3.7 The prediction and comparative analysis of virulence factors

The possibility of bacterium to cause disease can be measured by the virulence factors predicted in the corresponding strains. Virulence factors in a total of 15 strains including novel strains and a reference *S. pneumoiae* D39 were systematically screened by using VFanalyzer (Liu, Zheng et al. 2019) and compared (Figure 5-13). For strains in the same novel species group, the same virulence factors were predicted. A total of 31 predicted virulence factors were predicted. These virulence factors fall into 8 virulence factor classes based on their function including adherence, enzyme, immune evasion, iron uptake, manganese uptake, protease, toxin and antiphagocytosis. Among these 31 virulence factors, 8 virulence genes, pavA, lmb, gapA, eno, cps operons, sitA, degP, ropA were shared by all these strains. They function as adherences, enzyme, immune evasion, manganese uptake, proteases. Strains in these 7 novel species group all lack 6 virulence factors compared to the reference pathogen *S. pneumoniae*, thus pnumococal ion acquisition piaA, pneumolysin ply, 2 choline-binding protein cbpG and pspA, collagen binding protein cbpA and hyaluronidase hysA. Within all these novel species groups, strains in CF7_Ac1-8 novel species group were predicted to share the most abundant virulence factors with *S. pnuemoniae*, with zmpB, lytC, iga were only predicted in these 3 strains. Interestingly, most of the analysed strains except CF8-6 and *S. pnuemoniae* were predicted to contain a UDP-galactopyranose mutase act as an antiphagocytosis protein in bacterial pathogenicity. Only strains in CF9-1 novel species group were predicted to contain a toxin called cytolysin.

The distribution of virulence factors in these strains were corresponded well with the distance of these strains predicted before. Based on virulence factor comparison, these 14 strains can be further classified into 5 groups. For example, CF4-2 and CF9-1 novel species groups were more closely related compared to the other novel specie group strains. Similarly, CF8_Ac1-9 group and CF8_St5-11 group, CF8-6 and CF10-1 group and lastly CF7_Ac1-8 alone.

Figure 5-12: Comparison of virulence genes identified in represented strain in novel species group and selected reference genome.



5.4 Variations of closely related strains from the same novel species group using reads mapping method

The analysis of the variations between genomes of different strains in the same species groups were performed by Breseq. Variations caused by base substitution, short insertions and deletions, large deletions, mobile elements insertions and gene duplications can be identified.

The genome variations in 5 novel species groups were identified (Table 5-5). The variations fall into both intergenic regions and CDSs regions. There were very few variations in novel species groups CF4-2, CF9-1, CF7_Ac1-8 and CF8_St5-11. A total of 10 variations in 10 CDSs were identified (Table 5-6). CF4-2 and CF4-3 only had 1 variation in a PTS fructose and mannose component IIC, which may affect the sucrose metabolism in novel species group. The three variations in CF7_Ac1-8 species groups may result in the change of GMP synthase (glutamine-hydrolysing) [EC:6.3.5.2] and type I restriction subunit R, which may affect the purine metabolism (Hirst, Haliday et al. 1994), bacterial defence system (Murray 2000)and bacterial survival during starvation of this species group strains. Surprisingly, there were a total of 1179 variations between strains in CF8_Ac1-9 species groups. These variations were predicted to cause changes to 32 CDSs (Table 5-7). In this novel species group, the same virulence factors varied between two examined strains. The virulence factor zinc metalloprotease with 121 variations were identified.

Table 5-6: Statistics of the genomic variation of strains in the same novel species group.

Reference Genome	Query Sequence	% of Mapped Reads	# of Variations	# of Genes with annotations	# Intergenic Regions
CF4-2	CF4-3	97.6	5	1	0
CF7_Ac1-8	CF7_Ac1-12	97.5	3	3	0
CF9-3	CF9-1	97.4	2	1	1
CF8_Ac1-9	CF8_Ac1-11	96.8	1179	32	10
CF8_St5-11	CF8_St5-12	98.6	5	4	0
	CF8_St5-13/16	98.7	6	5	0
	CF8_Ac1-9/11	82.4%/81.3%	62015	Not calculated	Not calculated

Table 5-7: Overview of genomic variations in strains from the same novel species groups, part 1

Reference	Isolates	#	Annotations	Detail	Remark
CF4-2	CF4-3	1	PTS system, fructose- and mannose-inducible IIC component	T→C	
CF7_Ac1-8	CF7_Ac1-12	1	Type I restriction-modification system, restriction subunit R (EC 3.1.21.3)	G→C	
		2	GMP synthase [glutamine-hydrolyzing], amidotransferase subunit (EC 6.3.5.2) / GMP synthase [glutamine-hydrolyzing], ATP pyrophosphatase subunit (EC 6.3.5.2)	A→C	
		3	DNA protection during starvation protein	A→T	
CF9-3	CF9-1	1	N-acyl-L-amino acid amidohydrolase (EC 3.5.1.14)	A→G	
CF8_St5-11	CF8_St5-13/14/16	1	FIG001553: Hydrolase, HAD subfamily IIIA	C→T	not in CF8_St5-12
		2	Limit dextrin alpha-1,6-maltotetraose-hydrolase (EC 3.2.1.196) / Pullulanase (EC 3.2.1.41)	C→G	
		3	3-isopropylmalate dehydratase large subunit (EC 4.2.1.33)	C→T	
		4	Xanthine phosphoribosyltransferase (EC 2.4.2.22)	A→T	
		5	Manganese ABC transporter, periplasmic-binding protein SitA	A→G	

Table 5-8: Overview of genomic variations in strains from the same novel species groups, part 2

#	Annotation	Number of positions with mutations
1	Beta-glucoside bgl operon antiterminator, BglG family	1
2	2-hydroxymuconate tautomerase-like protein	1
3	C5a peptidase (EC 3.4.21.-)	1
4	FtsK/SpoIIIE family protein, putative EssC/YukB component of Type VII secretion system	1
5	Ribonucleotide reductase of class III (anaerobic), large subunit (EC 1.17.4.2)	1
6	UPF0291 protein YnzC	1
7	Peptide-methionine (S)-S-oxide reductase MsrA (EC 1.8.4.11) / Peptide-methionine (R)-S-oxide reductase MsrB (EC 1.8.4.12)	1
8	Spermine/spermidine acetyltransferase	6
9	Neopullulanase (EC 3.2.1.135)	11
10	Flavodoxin	13
11	GMP reductase (EC 1.7.1.7)	16
12	Thymidine kinase (EC 2.7.1.21)	17
13	Pneumococcal vaccine antigen A homolog	18
14	Xanthine phosphoribosyltransferase (EC 2.4.2.22)	18
15	ABC transporter, ATP-binding protein	21
16	Transcriptional regulator, MerR family	27
17	Transcriptional regulator, GntR family	31
18	FIG036672: Nucleoside-diphosphate-sugar epimerase	32
19	MBL-fold metallo-hydrolase superfamily	34
20	Pseudouridylate synthases, 23S RNA-specific	35
21	Serine hydroxymethyltransferase (EC 2.1.2.1)	37
22	Threonylcarbamoyl-AMP synthase (EC 2.7.7.87)	39
23	Na+-driven multidrug efflux pump	45
24	Extracellular protein	46
25	Xanthine permease	50
26	GMP synthase [glutamine-hydrolyzing], amidotransferase subunit (EC 6.3.5.2)	53
27	Peptide chain release factor N(5)-glutamine methyltransferase (EC 2.1.1.297)	58
28	EriC-type fluoride/proton exchange protein	58
29	Peptide chain release factor 1	61

30	Streptococcal extracellular nuclease 2; Mitogenic factor 2	68
31	Dihydroxy-acid dehydratase (EC 4.2.1.9)	71
32	Zinc metalloprotease	121

5.5 Conclusion

To gain insights of the difference of novel species in our study. In this chapter, we first analysed various genomic features predicted by using bioinformatic tools. Interestingly, genomic features of all kinds were conserved in all strains of the same novel species group. So, the genomic features in each novel species group were represented by either single strain in that group. For relatively small genomic features including the restriction modification system, plasmids, the CRISPR-Cas system, we observed the random distribution of these features in species. But we did observe the antibiotic resistance genes in plasmids from our strains.

Although plasmids were thought to confer important phenotypes for bacteria. We observed only small number of plasmids in our novel species strains. Then main evolution method for our strains were driven by HGT. We observed really a high number of genomic islands and prophages. Some CDSs regions of GIs or prophages were conserved. These regions were predicted to play roles in various metabolisms, toxin-antitoxin components, pathogenicity related, the RM system components, antibiotic resistance. We even identified extra antibiotic resistance genes directly from GIs predicted.

We also tried to reveal variations between strains from the same novel species group. For most novel species groups, the number of genomic variations between strains were low, from 2 to 1000 variations. Part of these variations were in CDS regions. So different phenotypes would be expected in some novel species strains.

Chapter 6 Characterization of novel strains by wet lab methods

6.1 Introduction

For *Streptococcus* species to colonize and persist in CF patients, they need to have the capacity to form biofilms in different organs and adapt to the surrounding environments.

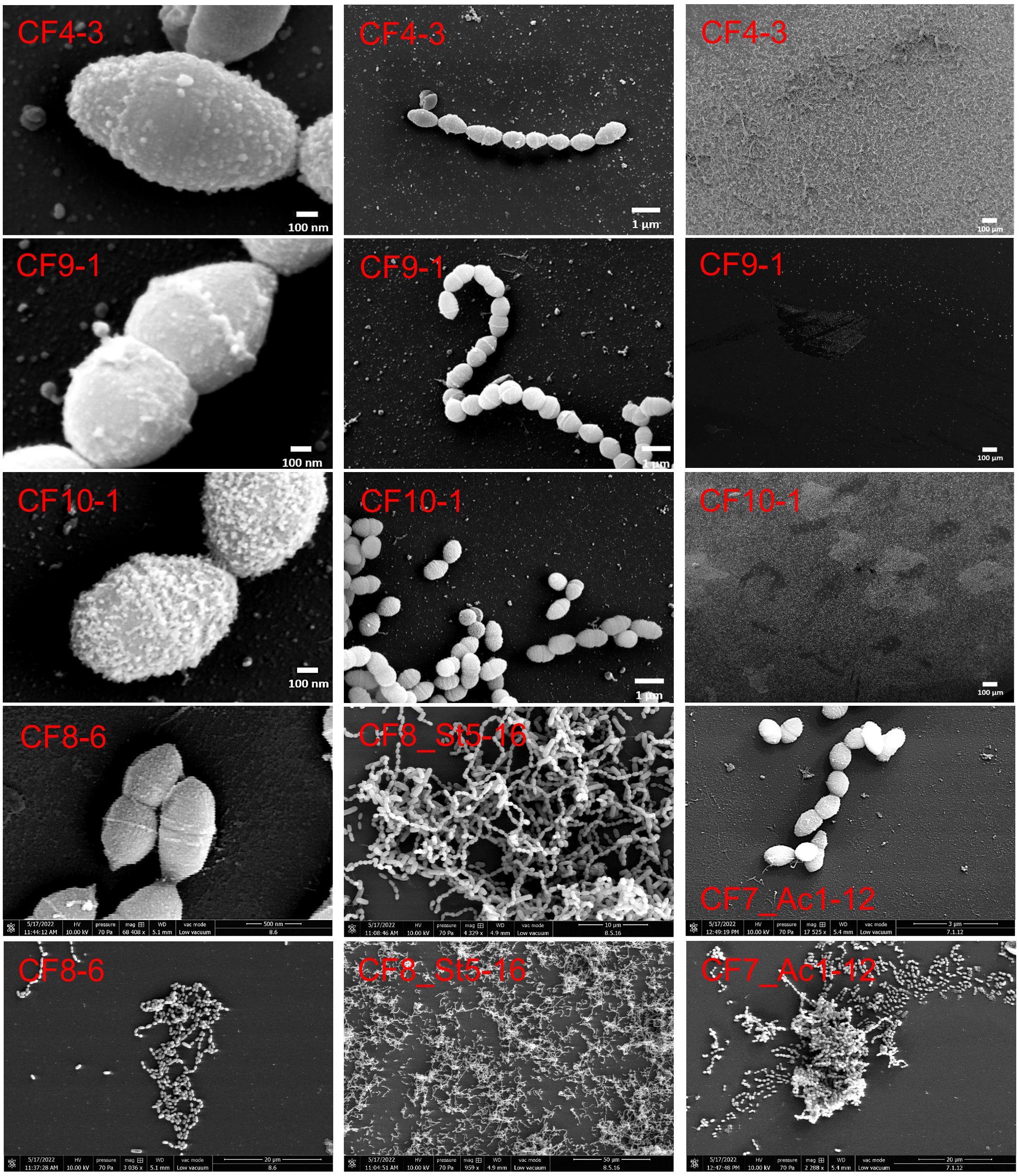
In this chapter, we firstly examined the morphology and biofilm forming capacity of different novel strains identified in this study. Then, we performed biochemical tests to identify single source carbon utilisation and chemical sensitivity of these strains. Lastly, we examined the pathogenicity of these strains in two animal models.

The experiments in this chapter were performed by the colleagues and the main supervisor Jo Fothergil in the University of Liverpool. I performed the analysis of the results. Jo also plotted the survival probability plots and CFU counting in mice.

6.2 Cell Morphology and biofilm forming capacity

SEM images confirmed the Streptococcus morphologies of six represented isolates in terms of diameters and growth (Figure 6-1). All strains were relatively 0.5 to 1 μm in diameter. They grew either singly, in pairs, short chains and clusters. These six strains showed distinct biofilm forming capacity, CF4-3 was the strongest, followed by CF10-1, CF8_St5-16, CF7_Ac1-12 and CF8-6. CF9-1 showed no capacity of biofilm formation in single culture condition. The colonization of Streptococcus strains in the host was determined by the biofilm formation. This reflected the different biofilm formation and colonization ability between strains.

Figure 6-1: The SEM images of cell morphology and biofilm formation of novel strains.



6.3 Biochemical characterisation

Fourteen novel strains were cultured on customized Libby Biolog GEN III Gram-positive bacterial chemicals 96-well plates to compare their difference in carbon utilization and chemical sensitivity. Three levels from positive, borderline to negative were used to represent the difference of tested strains under certain chemical environments. Because it is hard for us to distinguish from positive to borderline reaction and from borderline to negative reaction, we only compare positive and negative reactions between strains.

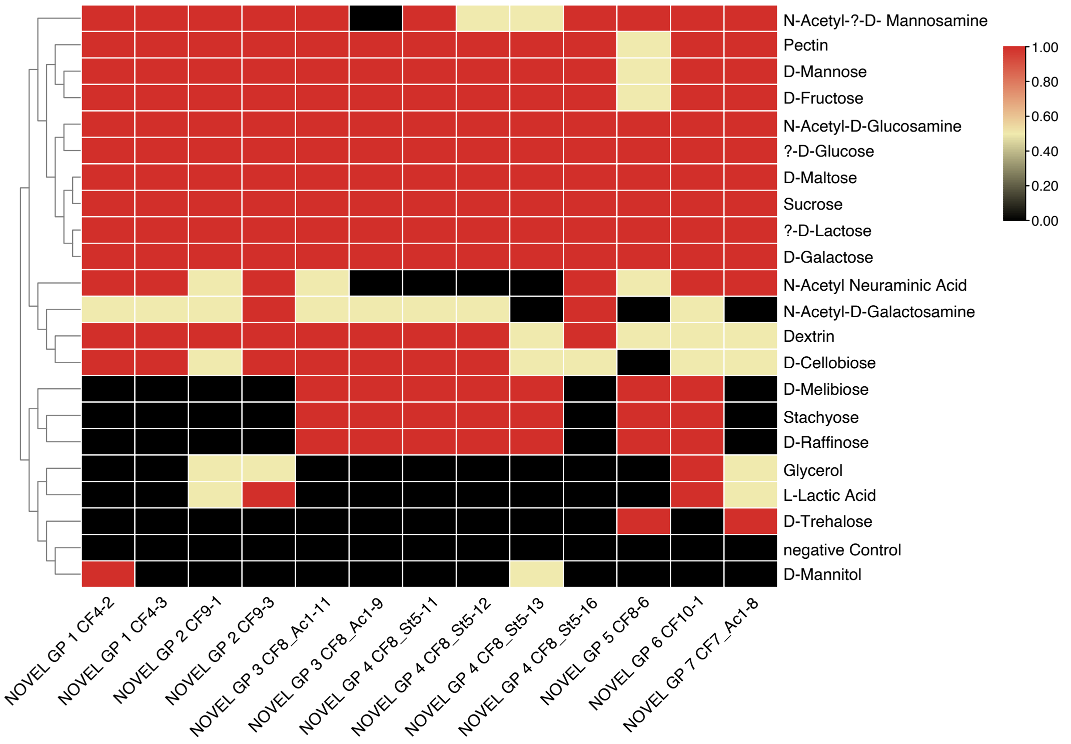
Based on the previous species identification results, these 14 strains were divided into 7 groups of novel species. The group 1 novel *Streptococcus* species strains contain CF4-2 and CF4-3, the group 2 strains contain CF9-1 and CF9-3, the group 3 strains contain CF8_Ac1-9 and CF8_Ac1-11, the group 4 strains contain CF8_St5-11, CF8_St5-12, CF8_St5-13 and CF8_St5-16, the group4 5 contain only CF8-6, the group 6 contain only CF10-1, the group 7 contain only CF7_Ac1-12 and CF7_Ac1-8. Because there was no data for some of the testing wells in CF7_Ac1-12, the comparison of carbon utilization and chemical sensitivity were performed between only 13 strains and within 4 groups.

6.3.1 Carbon utilization

A total of 71 different carbon sources were used to test the utilization by novel species strains. These 13 strains showed the positive utilization of 21 different carbon sources ranging from 17 different sugars, 2 sugar alcohols, 1 hexose acid and 1 carboxylic acid. In KEGG database, the 21 positively utilized carbon sources were found mainly involved in starch and sucrose metabolism, galactose metabolism, glycerolipid metabolism, amino sugar and nucleotide sugar metabolism, O-antigen nucleotide sugar biosynthesis, biosynthesis of nucleotide sugars, glycolysis/gluconeogenesis, fructose and mannose metabolism, pentose and glucuronate interconversions, pyruvate metabolism, propanoate metabolism, glucagon signalling pathway.

These 13 strains from the 7 novel species groups were all found to ferment 6 sugars, thus D-Maltose, Sucrose, α-D-Lactose, D-Galactose, N-Acetyl-D-Glucosamine, α-D-Glucose. N-Acetyl-β-D-Mannosamine was positively utilized in all 7 novel species groups but varies between group 3 strains and group 4 strains. In group 3 novel species strains, it is positively used in CF8_Ac1-11 but negatively used in CF8_Ac1-9. In group 4 novel species strains, the utilization was uncertain in CF8_St5-12 and CF8_St5-13. D-Mannose, D-Fructose and Pectin were utilized by 6 groups except the novel species group 5. Stachyose, D-Raffinose and D-Melibiose were positively utilized by strains in group 3, 4, 5 and 6. Furthermore, these three sugars were not used by CF8_St5-16 but utilized by the other 3 strains in the same novel group. Dextrin and D-Cellobiose were positively used by strains from novel species group 1, 2, 3 and 4. D-Trehalose was positively utilized by group 5 and group 7. N-Acetyl-D-Galactosamine was positively utilized by 2 out of 13 strains from group 2 and group 4. N-Acetyl Neuraminic Acid was positively utilized by strains in group 1, 2, 4, 6 and 7. D-Mannitol was only utilized by CF4-2 from group 1. L-Lactic Acid were utilized by strains in group 2 and group 6 (Figure 6-2).

Figure 6-2: The carbon source utilization by strains in different novel species groups.

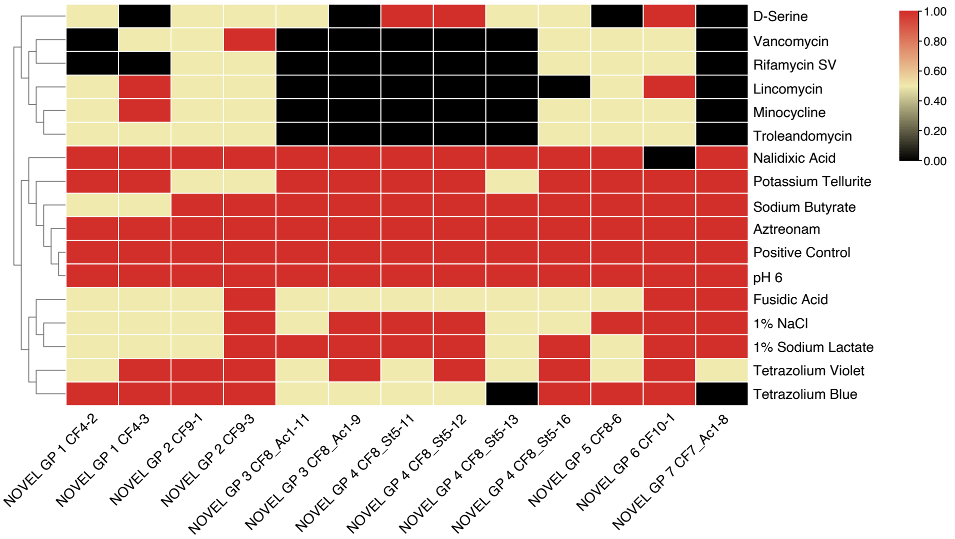


6.3.2 Chemical sensitivity

As mentioned in 6.3.1, only 13 strains except CF7_Ac1-12 were involved in chemical sensitivity analysis.

A total of 23 different chemical ranging from pH, antibiotics, salts, redox dyes were used for testing. Only 14 chemicals were resisted by at least 1 out 13 strains. These strains showed resistance to pH 6 and aztreonam. All strains showed resistant to Nalidixic Acid except CF10-1. Although the resistance to other chemicals varies between strains from different novel species groups, it is still hard to determine the chemical sensitivity between strains within the same novel species group due to the lack of measurement for borderline (Figure 6-3).

Figure 6-3: the comparison of chemical sensitivity in 13 strains.



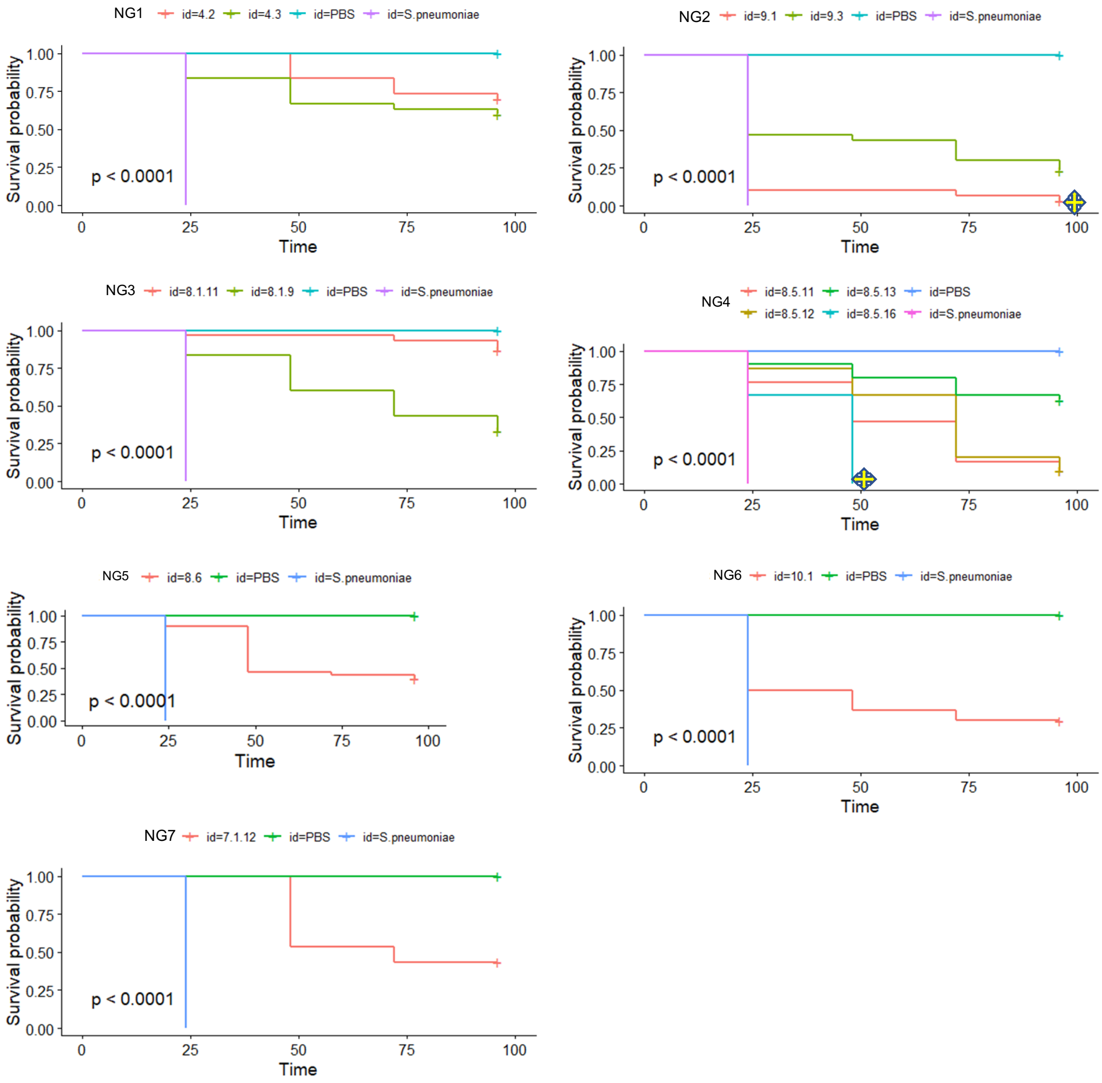
6.4 Pathogenicity detection using different animal models

A total of 15 different strains including a *Streptococcus* pathogen S. pneumoniae D39 and 14 novel species strains in this study were tested for pathogenicity difference. Furthermore, strains with relatively higher pathogenicity from these 14 strains were further examined by using mice models.

6.4.1 Validation of strain virulence using *in vivo* galleria model

A total of 30 galleria were treated as 3 replicates for each strain and observed during 4 days. The survival probability of all tested strains was illustrated in Figure 6-4. S. pneumoniae killed all galleria after only 24 hours. Similar to *S. pneumoniae*, strain CF9-1 was the quickest to decreased the survival probability to around 10% in only 24 hours. Interestingly, after 48 hours, only strain CF8_St5-16 among all 4 strains in the same novel species group 4 was found to kill all galleria. All tested strains except CF9-1, CF9-3 and CF10-1, influenced the survival of galleria greatly until 48 hours. The survival probability of treated galleria was positively affected by the pathogenicity of tested strain. For the tested strains, CF9-1 and CF8_St5-16 were the two with highest pathogenicity.

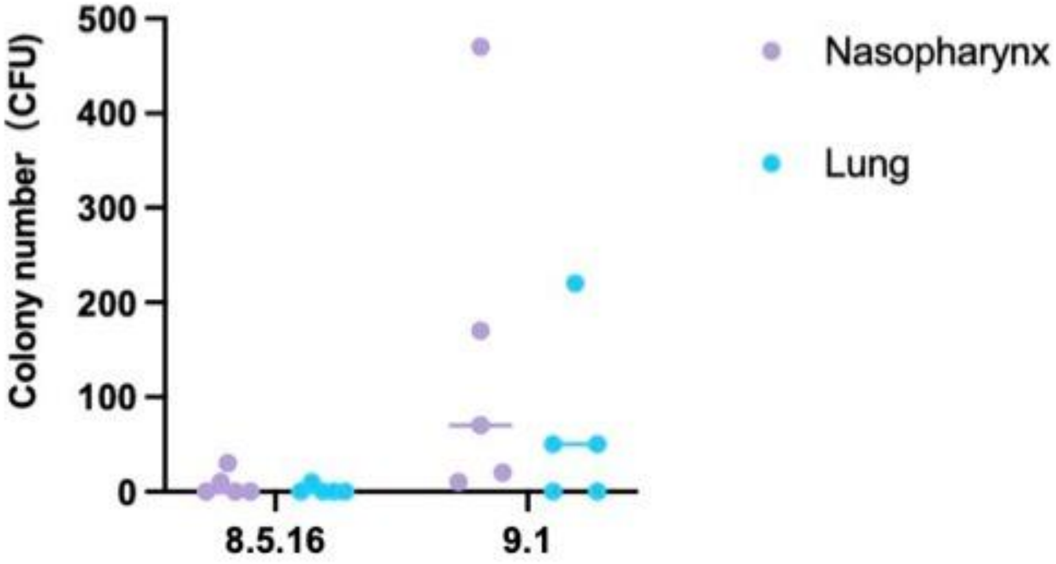
Figure 6-4: The pathogenicity of 14 novel strains on galleria model. Strains CF9-1 and CF8_St5-16 showed the highest pathogenicity among all tested strains.



6.4.2 Validation of strain virulence using In vivo Mice model

Based on the observations of pathogenicity testing in the *Galleria* model, an in vivo mouse model was used to further test the pathogenicity potential of CF8_St5-16 and CF9-1. These two strains were injected to 10 mice in total, 5 mice for each strain. Later, for each mouse, nasopharynx and lung were plated for counting bacterial colony number to infer the survival of each strain. There seemed no observation of CF8_St5-16 in mice samples. CF9-1 showed some level of pathogenicity in mice models. But the average number of CF9-1 in two organs were similar and less than 100 CFU (Figure 6-5). The survival number of tested bacteria strains in mice were determined by the pathogenicity and the immune systems in mice models.

Figure 6-5: The number of CF8_St5-16 and CF9-1 at nasopharynx and lungs of tested mice.



6.5 Discussion and conclusion

Biofilm formation is known to play an important role in Streptococcal infection (Stevens 2003). The different biofilm forming capacity reflect different pathogenicity and the colonization ability of different strains in the same species (Fothergill, Neill et al. 2014). It is interesting to see, strain CF9-1 and CF8_St5-11 out of the other strains showed the strongest pathogenicity in the *Gelleria* model survival testing. Later CF9-1 showed some level of colonization ability in mice. It is not surprising to see the phage variation of pathogenicity within the same bacterial species strains because within the same bacterial species, some strains developed adaptations in the host by sacrificed the level of pathogenicity. This phenomenon was observed and studies in different strains of *Pseudomonas aeruginosa*.

Streptococcus strains relay on carbohydrate metabolisms to generate energy for other metabolisms. *S. thermophilus* fermented lactose, fructose, sucrose, and glucose (Harnett, Davey et al. 2011). Similarly, *S. pneumoniae* was reported to use fructose, galactose, sucrose, glucose, raffinose, inulin, trehalose, and maltose as energy sources (Bergey and Holt 1994). So, it is not surprising for these streptococci to use the same carbon sources as *S. thermophilus* and *S. pneumoniae*. Moreover, these strains were capable to use other carbon sources reflecting the complex carbohydrate metabolisms in streptococci. The different utilization of carbohydrates in these strains were determined by the different enzymes within each strain.

Strains in each group of species were identified as from the same source and assumed to behave the same phenotypes due to the very close similarity (over 99% in both similarity and identity) in genomic level. From our results, we observed the distinct utilization of the same carbon source by strains belongs to the same novel species groups. Thus, the utilization of D-Mannitol by only CF4-2 in group1, the different utilization of D-Melibiose, Stachyose, D-Raffinose, N-Acetyl Neuraminic Acid and N-Acetyl-D-Galactosamine by 4 strains in group 4, the varies of N-Acetyl-β-D- Mannosamine utilization by group 3 strains. The shared positive utilization of certain carbon sources reflected the conservation of key enzymes involved in related carbohydrate metabolisms in the *Streptococcus* strains. The occurrence of SNPs in strains from the same novel groups may have effects on bacterial carbon metabolisms. For the novel species group 1, CF4-2 and CF4-3 were predicted to have 1 nucleotide difference between a coding sequence annotated as PTS system, fructose- and mannose-inducible IIC component. This enzyme was precited to involve in the phosphotransferase system and used by bacteria for uptake of D-Mannose. Interestingly, the different utilization of Mannose was observed between CF4-2 and CF4-3.

Chapter 7 Conclusion

In this Phd journey, we focused on three aims. The aim one was to generate reference *Streptococcus* genomes for downstream analysis and future study. The second aim was to perform taxonomic analysis of these strains. We further assign each strain into a *Streptococcus* species level. This will benefit the choosing of suitable references genomes for in-silicon genomic analysis. The aim 3 was to analyze the selected novel strain genomes. Represented species genomes between different species and genomes in the same species were compared via comparative genomic features and whole genomes analysis.

Although a total of 60 isolates were sequences, 40 high quality incomplete *Streptococcus* genomes were assembled and processed for downstream analysis. Annotation by RAST revealed only 30% of CDSs were predicted with a function in these genomes. This is really normal for bacterial genomes were always predicted to contain a large number of hypothetical proteins even for model species. The CDSs with annotations still gave us some important functions preserved by the bacteria like prophage, virulence related components, capsules and components participated in essential metabolisms.

The 40 genomes were further processed to phylogenetic analysis and in-silicon species identification. These genomes can be assigned in *S. oralis*, *S. mitis*, *S. infantis*, and *S. parasanguinis*, *S.* sp. *CF4*-2, *S.* sp. CF10-1, *S.* sp.CF7_Ac1-8, *S.* sp. CF8_Ac1-9, *S.* sp.8-6, *S.* sp. CF8_St5-11 and *S.* sp. 9-1. This high number of *Streptococcs* species level identification in one study maybe the first case. Think about the co-occur of *Streptococcus* and *Pseudomonase*. It would be really interesting to performing future study of strains from the two genera. A total of 14 genomes were predicted to from 7 novel *Streptococcus* species groups. Five of these novel species groups contain multiple isolates. Isolates CF4-2 and CF4-3 were from *S.* sp. CF4-2 group, Isolates CF7_Ac1-8 and CF7_Ac1-8 formed *S.* sp. CF7_Ac1-8 group, Isolates CF9-1 and CF9-3 were from *S.* sp. CF9-1 group, four isolates CF8_St5-11, CF8_St5-12, CF8_St5-13 and CF8_St5-16 were from *S.* sp.

CF8_St5-11 group, finally, CF8_Ac1-9 and CF8_Ac1-11 formed a *S*. sp. CF8_Ac1-9 group.

Interestingly, although we observed different number of genomic features like prophage and genomic islands in genomes from the same novel species group. After manual curation, the contents genomic features predicted in genomes from the same species in our study were actually the same. For genomic features predicted from different novel species groups, we observed matches in some regions of these features. The commonly shared regions were CDSs for various antibiotic resistance, pathogenicity related, specifically for biofilm formation, invasion and finally limited kinds of metabolisms It is hardly to see genomes from the same novel species belongs to different patients. Different patients provide environments for the evolution of the *Streptococcus* in our study. We also expect more kinds of novel species groups in more large-scale study. Or we may need to think of other good ways to separate strains from the same sample. For examples, strains in the same genera with different genomes, but cooccurred in biofilm may group together.

We also used reads-mapping methods to compare variations in genomes from the same novel species. The variations number were similar to genomes from the same group but different greatly between different species. This may reflect the dynamic evolution of *Streptococcus* in CF patients. But the different variations between species may reflect the difference of *Streptococcus* species react to survival pressure or sensitivity to dynamic environments in different host. This also reflect the genome stability difference in *Streptococcus* species.

We also performed some wet lab experiments to compare the difference and similarity between novel species genomes. Similar to in-silicon prediction, strains in the same novel species have similar cell morphology and biofilm formation capacity (Data not shown). Large difference was observed between isolates from different novel species groups. In biochemical tests part, we observed a few difference in carbon utilization by strains in the same species. We also observed slight differences in survival plots of Galleria but with the same trend of affection.

For future study, we can focus our interests in Streptococcus-to-*Pseudomonase aeruginosa* interactions.

Bibliography

Acosta, N., A. Heirali, R. Somayaji, M. G. Surette, M. L. Workentine, C. D. Sibley, H. R. Rabin and M. D. Parkins (2018). "Sputum microbiota is predictive of long-term clinical outcomes in young adults with cystic fibrosis." Thorax **73**(11): 1016-1025.

Albertsen, M., P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson and P. H. Nielsen (2013). "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes." Nat Biotechnol **31**(6): 533-538.

Allen, F. H., Jr., R. R. Dooley, H. Shwachman and A. G. Steinberg (1956). "Linkage studies with cystic fibrosis of the pancreas." Am J Hum Genet **8**(3): 162-176.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Amin, R., N. Jahnke and V. Waters (2020). "Antibiotic treatment for Stenotrophomonas maltophilia in people with cystic fibrosis." Cochrane Database Syst Rev **3**(3): CD009249.

Andersen, D. H. and R. G. Hodges (1946). "Celiac syndrome; genetics of cystic fibrosis of the pancreas, with a consideration of etiology." Am J Dis Child (1911) **72**: 62-80.

Anjum, M. F. (2015). "Screening methods for the detection of antimicrobial resistance genes present in bacterial isolates and the microbiota." Future Microbiol **10**(3): 317-320.

Arndt, D., J. R. Grant, A. Marcu, T. Sajed, A. Pon, Y. Liang and D. S. Wishart (2016). "PHASTER: a better, faster version of the PHAST phage search tool." Nucleic Acids Res **44**(W1): W16-21.

Arndt, D., A. Marcu, Y. Liang and D. S. Wishart (2019). "PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes." Brief Bioinform **20**(4): 1560-1567.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Asp, P., R. Blum, V. Vethantham, F. Parisi, M. Micsinai, J. Cheng, C. Bowman, Y. Kluger and B. D. Dynlacht (2011). "Genome-wide remodeling of the epigenetic landscape during myogenic differentiation." Proc Natl Acad Sci U S A **108**(22): E149-158.

Assael, B. M., C. Castellani, M. B. Ocampo, P. Iansa, A. Callegaro and M. G. Valsecchi (2002). "Epidemiology and survival analysis of cystic fibrosis in an area of intense neonatal screening over 30 years." Am J Epidemiol **156**(5): 397-401.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke and O. Zagnitko (2008). "The RAST Server: rapid annotations using subsystems technology." BMC Genomics **9**: 75.

Bacci, G., G. Taccetti, D. Dolce, F. Armanini, N. Segata, F. Di Cesare, V. Lucidi, E. Fiscarelli, P. Morelli, R. Casciaro, A. Negroni, A. Mengoni and A. Bevivino (2020). "Untargeted Metagenomic Investigation of the Airway Microbiome of Cystic Fibrosis Patients with Moderate-Severe Lung Disease." Microorganisms **8**(7).

Balson, D. F. and W. V. Shaw (1990). "Nucleotide sequence of the rep gene of staphylococcal plasmid pCW7." Plasmid **24**(1): 74-80.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero and P. Horvath (2007). "CRISPR provides acquired resistance against viruses in prokaryotes." Science **315**(5819): 1709-1712.

Bazett, M., L. Honeyman, A. N. Stefanov, C. E. Pope, L. R. Hoffman and C. K. Haston (2015). "Cystic fibrosis mouse model-dependent intestinal structure and gut microbiome." Mamm Genome **26**(5-6): 222-234.

Berger, H. A., M. P. Anderson, R. J. Gregory, S. Thompson, P. W. Howard, R. A. Maurer, R. Mulligan, A. E. Smith and M. J. Welsh (1991). "Identification and regulation of the cystic fibrosis transmembrane conductance regulator-generated chloride channel." J Clin Invest **88**(4): 1422-1431.

Bergey, D. H. and J. G. Holt (1994). Bergey's Manual of Determinative Bacteriology, Williams & Wilkins.

Bertelli, C., M. R. Laird, K. P. Williams, G. Simon Fraser University Research Computing, B. Y. Lau, G. Hoad, G. L. Winsor and F. S. L. Brinkman (2017). "IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets." Nucleic Acids Res **45**(W1): W30-W35.

Billane, K., E. Harrison, D. Cameron and M. A. Brockhurst (2022). "Why do plasmids manipulate the expression of bacterial phenotypes?" Philos Trans R Soc Lond B Biol Sci **377**(1842): 20200461.

Bobadilla, J. L., M. Macek, Jr., J. P. Fine and P. M. Farrell (2002). "Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening." Hum Mutat **19**(6): 575-606.

Bonner, T. I., D. J. Brenner, B. R. Neufeld and R. J. Britten (1973). "Reduction in the rate of DNA reassociation by sequence divergence." J Mol Biol **81**(2): 123-135.

Bortolaia, V., R. S. Kaas, E. Ruppe, M. C. Roberts, S. Schwarz, V. Cattoir, A. Philippon, R. L. Allesoe, A. R. Rebelo, A. F. Florensa, L. Fagelhauer, T. Chakraborty, B. Neumann, G. Werner, J. K. Bender, K. Stingl, M. Nguyen, J. Coppens, B. B. Xavier, S. Malhotra-Kumar, H. Westh, M. Pinholt, M. F. Anjum, N. A. Duggett, I. Kempf, S. Nykasenoja, S. Olkkola, K. Wieczorek, A. Amaro, L. Clemente, J. Mossong, S. Losch, C. Ragimbeau, O. Lund and F. M. Aarestrup (2020). "ResFinder 4.0 for predictions of phenotypes from genotypes." J Antimicrob Chemother **75**(12): 3491-3500.

Brussow, H., C. Canchaya and W. D. Hardt (2004). "Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion." Microbiol Mol Biol Rev **68**(3): 560-602, table of contents.

Budroni, S., E. Siena, J. C. Dunning Hotopp, K. L. Seib, D. Serruto, C. Nofroni, M. Comanducci, D. R. Riley, S. C. Daugherty, S. V. Angiuoli, A. Covacci, M. Pizza, R. Rappuoli, E. R. Moxon, H. Tettelin and D. Medini (2011). "Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination." Proc Natl Acad Sci U S A **108**(11): 4494-4499.

Burgel, P. R., A. Paugam, D. Hubert and C. Martin (2016). "Aspergillus fumigatus in the cystic fibrosis lung: pros and cons of azole therapy." Infect Drug Resist **9**: 229-238.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. L. Madden (2009). "BLAST+: architecture and applications." BMC Bioinformatics **10**: 421.

Carattoli, A., E. Zankari, A. Garcia-Fernandez, M. Voldby Larsen, O. Lund, L. Villa, F. Moller Aarestrup and H. Hasman (2014). "In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing." Antimicrob Agents Chemother **58**(7): 3895-3903.

Casjens, S. (2003). "Prophages and bacterial genomics: what have we learned so far?" Mol Microbiol **49**(2): 277-300.

Cerutti, F., L. Bertolotti, T. L. Goldberg and M. Giacobini (2011). "Taxon ordering in phylogenetic trees by means of evolutionary algorithms." BioData Min **4**: 20.

Chain, P. S., D. V. Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H. M. Khouri, C. D. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D. Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam, C. Genomic Standards Consortium Human Microbiome Project Jumpstart and J. C. Detter (2009). "Genomics. Genome project standards in a new era of sequencing." Science **326**(5950): 236-237.

Chen, C., H. Chen, Y. Zhang, H. R. Thomas, M. H. Frank, Y. He and R. Xia (2020). "TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data." Mol Plant **13**(8): 1194-1202.

Chen, L., J. Yang, J. Yu, Z. Yao, L. Sun, Y. Shen and Q. Jin (2005). "VFDB: a reference database for bacterial virulence factors." Nucleic Acids Res **33**(Database issue): D325-328.

Chen, Y. Y., H. R. Shieh, C. T. Lin and S. Y. Liang (2011). "Properties and construction of plasmid pFW213, a shuttle vector with the oral Streptococcus origin of replication." Appl Environ Microbiol **77**(12): 3967-3974.

Chun, J., A. Oren, A. Ventosa, H. Christensen, D. R. Arahal, M. S. da Costa, A. P. Rooney, H. Yi, X. W. Xu, S. De Meyer and M. E. Trujillo (2018). "Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes." Int J Syst Evol Microbiol **68**(1): 461-466.

Ciufo, S., S. Kannan, S. Sharma, A. Badretdin, K. Clark, S. Turner, S. Brover, C. L. Schoch, A. Kimchi and M. DiCuccio (2018). "Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI." Int J Syst Evol Microbiol **68**(7): 2386-2392.

Coburn, B., P. W. Wang, J. Diaz Caballero, S. T. Clark, V. Brahma, S. Donaldson, Y. Zhang, A. Surendra, Y. Gong, D. Elizabeth Tullis, Y. C. Yau, V. J. Waters, D. M. Hwang and D. S. Guttman (2015). "Lung microbiota across age and disease stage in cystic fibrosis." Sci Rep **5**: 10241.

Coffey, M. J., S. Nielsen, B. Wemheuer, N. O. Kaakoush, M. Garg, B. Needham, R. Pickford, A. Jaffe, T. Thomas and C. Y. Ooi (2019). "Gut Microbiota in Children With Cystic Fibrosis: A Taxonomic and Functional Dysbiosis." Sci Rep **9**(1): 18593.

Conneally, P. M., A. D. Merritt and P. L. Yu (1973). "Cystic fibrosis: population genetics." Tex Rep Biol Med **31**(4): 639-650.

Courvalin, P. (1991). "Genotypic approach to the study of bacterial resistance to antibiotics." Antimicrob Agents Chemother **35**(6): 1019-1023.

Couvin, D., A. Bernheim, C. Toffano-Nioche, M. Touchon, J. Michalik, B. Neron, E. P. C. Rocha, G. Vergnaud, D. Gautheret and C. Pourcel (2018). "CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins." Nucleic Acids Res **46**(W1): W246-W251.

Cuthbertson, L., A. W. Walker, A. E. Oliver, G. B. Rogers, D. W. Rivett, T. H. Hampton, A. Ashare, J. S. Elborn, A. De Soyza, M. P. Carroll, L. R. Hoffman, C. Lanyon, S. M. Moskowitz, G. A. O'Toole, J. Parkhill, P. J. Planet, C. C. Teneback, M. M. Tunney, J. B. Zuckerman, K. D. Bruce and C. J. van der Gast (2020). "Lung function and microbiota diversity in cystic fibrosis." Microbiome **8**(1): 45.

Deatherage, D. E., C. C. Traverse, L. N. Wolf and J. E. Barrick (2014). "Detecting rare structural variation in evolving microbial populations from new sequence junctions using breseq." Front Genet **5**: 468.

Dennis, E. A., M. T. Coats, S. Griffin, B. Pang, D. E. Briles, M. J. Crain and W. E. Swords (2018). "Hyperencapsulated mucoid pneumococcal isolates from patients with cystic fibrosis have increased biofilm density and persistence in vivo." Pathog Dis **76**(7).

Dickson, R. P., J. R. Erb-Downward, F. J. Martinez and G. B. Huffnagle (2016). "The Microbiome and the Respiratory Tract." Annu Rev Physiol **78**: 481-504.

Dickson, R. P. and G. B. Huffnagle (2015). "The Lung Microbiome: New Principles for Respiratory Bacteriology in Health and Disease." PLoS Pathog **11**(7): e1004923.

Dobrindt, U., B. Hochhut, U. Hentschel and J. Hacker (2004). "Genomic islands in pathogenic and environmental microorganisms." Nat Rev Microbiol **2**(5): 414-424.

Donovan, M., M. D. J. Lynch, C. S. Mackey, G. N. Platt, B. K. Washburn, D. L. Vera, D. J. Trickey, T. C. Charles, Z. Wang and K. M. Jones (2020). "Metagenome-Assembled Genome Sequences of Five Strains from the Microtus ochrogaster (Prairie Vole) Fecal Microbiome." Microbiol Resour Announc **9**(2).

Dupuis, M. E., M. Villion, A. H. Magadan and S. Moineau (2013). "CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance." Nat Commun **4**: 2087.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Elborn, J. S. (2016). "Cystic fibrosis." Lancet **388**(10059): 2519-2531.

Esposito, S., C. Colombo, A. Tosco, E. Montemitro, S. Volpi, L. Ruggiero, M. Lelii, A. Bisogno, C. Pelucchi, N. Principi and F. Italian Pneumococcal Study Group on Cystic (2016). "Streptococcus pneumoniae oropharyngeal colonization in children and adolescents with cystic fibrosis." J Cyst Fibros **15**(3): 366-371.

Farrell, P. M., M. R. Kosorok, M. J. Rock, A. Laxova, L. Zeng, H. C. Lai, G. Hoffman, R. H. Laessig and M. L. Splaingard (2001). "Early diagnosis of cystic fibrosis through neonatal screening prevents severe malnutrition and improves long-term growth. Wisconsin Cystic Fibrosis Neonatal Screening Study Group." Pediatrics **107**(1): 1-13.

Felsenstein, J. (1985). "Confidence Limits on Phylogenies: An Approach Using the Bootstrap." Evolution **39**(4): 783-791.

Filkins, L. M., T. H. Hampton, A. H. Gifford, M. J. Gross, D. A. Hogan, M. L. Sogin, H. G. Morrison, B. J. Paster and G. A. O'Toole (2012). "Prevalence of streptococci and increased polymicrobial diversity associated with cystic fibrosis patient stability." J Bacteriol **194**(17): 4709-4717.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick and et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Fodor, A. A., E. R. Klem, D. F. Gilpin, J. S. Elborn, R. C. Boucher, M. M. Tunney and M. C. Wolfgang (2012). "The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations." PLoS One **7**(9): e45001.

Fortier, L. C. and O. Sekulovic (2013). "Importance of prophages to evolution and virulence of bacterial pathogens." Virulence **4**(5): 354-365.

Fothergill, J. L., D. R. Neill, N. Loman, C. Winstanley and A. Kadioglu (2014). "Pseudomonas aeruginosa adaptation in the nasopharyngeal reservoir leads to migration and persistence in the lungs." Nat Commun **5**: 4780.

Francoise, A. and G. Hery-Arnaud (2020). "The Microbiome in Cystic Fibrosis Pulmonary Disease." Genes (Basel) **11**(5).

Fugere, A., D. Lalonde Seguin, G. Mitchell, E. Deziel, V. Dekimpe, A. M. Cantin, E. Frost and F. Malouin (2014). "Interspecific small molecule interactions between clinical isolates of Pseudomonas aeruginosa and Staphylococcus aureus from adult cystic fibrosis patients." PLoS One **9**(1): e86705.

Fujiwara, T. M., K. Morgan, R. H. Schwartz, R. A. Doherty, S. R. Miller, K. Klinger, P. Stanislovitis, N. Stuart and P. C. Watkins (1989). "Genealogical analysis of cystic fibrosis families and chromosome 7q RFLP haplotypes in the Hutterite Brethren." Am J Hum Genet **44**(3):

327-337.

Gabrielaite, M., F. C. Nielsen, H. K. Johansen and R. L. Marvig (2021). "Achromobacter spp. genetic adaptation in cystic fibrosis." Microb Genom **7**(7).

Gal-Mor, O. and B. B. Finlay (2006). "Pathogenicity islands: a molecular toolbox for bacterial virulence." Cell Microbiol **8**(11): 1707-1719.

Gao, X. Y., X. Y. Zhi, H. W. Li, H. P. Klenk and W. J. Li (2014). "Comparative genomics of the bacterial genus Streptococcus illuminates evolutionary implications of species groups." PLoS One **9**(6): e101229.

Gavillet, H., L. Hatfield, D. Rivett, A. Jones, A. Maitra, A. Horsley and C. van der Gast (2022). "Bacterial Culture Underestimates Lung Pathogen Detection and Infection Status in Cystic Fibrosis." Microbiol Spectr **10**(5): e0041922.

Goodman, H. O. and S. C. Reed (1952). "Heredity of fibrosis of the pancreas; possible mutation rate of the gene." Am J Hum Genet **4**(2): 59-71.

Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme and J. M. Tiedje (2007). "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities." Int J Syst Evol Microbiol **57**(Pt 1): 81-91.

Goss, C. H. and J. L. Burns (2007). "Exacerbations in cystic fibrosis. 1: Epidemiology and pathogenesis." Thorax **62**(4): 360-367.

Hacker, J., L. Bender, M. Ott, J. Wingender, B. Lund, R. Marre and W. Goebel (1990). "Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates." Microb Pathog **8**(3): 213-225.

Hacker, J., G. Blum-Oehler, I. Muhldorfer and H. Tschape (1997). "Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution." Mol Microbiol **23**(6): 1089-1097.

Hacker, J. and E. Carniel (2001). "Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes." EMBO Rep **2**(5): 376-381.

Hacker, J. and J. B. Kaper (2000). "Pathogenicity islands and the evolution of microbes." Annu Rev Microbiol **54**: 641-679.

Hammond, K. B., S. H. Abman, R. J. Sokol and F. J. Accurso (1991). "Efficacy of statewide neonatal screening for cystic fibrosis by assay of trypsinogen concentrations." N Engl J Med **325**(11): 769-774.

Hansen, C. R., T. Pressler and N. Hoiby (2008). "Early aggressive eradication therapy for intermittent Pseudomonas aeruginosa airway colonization in cystic fibrosis patients: 15 years experience." J Cyst Fibros **7**(6): 523-530.

Harnett, J., G. Davey, A. Patrick, C. Caddick and L. Pearce (2011). Lactic Acid Bacteria | Streptococcus thermophilus. Encyclopedia of Dairy Sciences (Second Edition). J. W. Fuquay. San Diego, Academic Press: 143-148.

Haroon, M. F., C. T. Skennerton, J. A. Steen, N. Lachner, P. Hugenholtz and G. W. Tyson (2013).

"In-solution fluorescence in situ hybridization and fluorescence-activated cell sorting for single cell and population genome recovery." Methods Enzymol **531**: 3-19.

Harun, S. N., C. Wainwright, K. Klein and S. Hennig (2016). "A systematic review of studies examining the rate of lung function decline in patients with cystic fibrosis." Paediatr Respir Rev **20**: 55-66.

Heather, J. M. and B. Chain (2016). "The sequence of sequencers: The history of sequencing DNA." Genomics **107**(1): 1-8.

Hendriksen, R. S., A. M. Seyfarth, A. B. Jensen, J. Whichard, S. Karlsmose, K. Joyce, M. Mikoleit, S. M. Delong, F. X. Weill, A. Aidara-Kane, D. M. Lo Fo Wong, F. J. Angulo, H. C. Wegener and F. M. Aarestrup (2009). "Results of use of WHO Global Salm-Surv external quality assurance system for antimicrobial susceptibility testing of Salmonella isolates from 2000 to 2007." J Clin Microbiol **47**(1): 79-85.

Hirst, M., E. Haliday, J. Nakamura and L. Lou (1994). "Human GMP synthetase. Protein purification, cloning, and functional expression of cDNA." J Biol Chem **269**(38): 23830-23837.

Hoffman, L. R., E. Deziel, D. A. D'Argenio, F. Lepine, J. Emerson, S. McNamara, R. L. Gibson, B. W. Ramsey and S. I. Miller (2006). "Selection for Staphylococcus aureus small-colony variants due to growth in the presence of Pseudomonas aeruginosa." Proc Natl Acad Sci U S A **103**(52): 19890-19895.

Hogan, D. A., S. D. Willger, E. L. Dolben, T. H. Hampton, B. A. Stanton, H. G. Morrison, M. L. Sogin, J. Czum and A. Ashare (2016). "Analysis of Lung Microbiota in Bronchoalveolar Lavage, Protected Brush and Sputum Samples from Subjects with Mild-To-Moderate Cystic Fibrosis Lung Disease." PLoS One **11**(3): e0149998.

Hsiao, W., I. Wan, S. J. Jones and F. S. Brinkman (2003). "IslandPath: aiding detection of genomic islands in prokaryotes." Bioinformatics **19**(3): 418-420.

Jia, S. and J. L. Taylor-Cousar (2023). "Cystic Fibrosis Modulator Therapies." Annu Rev Med **74**: 413-426.

Johnston, C. D., S. L. Cotton, S. R. Rittling, J. R. Starr, G. G. Borisy, F. E. Dewhirst and K. P. Lemon (2019). "Systematic evasion of the restriction-modification barrier in bacteria." Proc Natl Acad Sci U S A **116**(23): 11454-11459.

Juhas, M., J. R. van der Meer, M. Gaillard, R. M. Harding, D. W. Hood and D. W. Crook (2009). "Genomic islands: tools of bacterial horizontal gene transfer and evolution." FEMS Microbiol Rev **33**(2): 376-393.

Kamke, J., A. Sczyrba, N. Ivanova, P. Schwientek, C. Rinke, K. Mavromatis, T. Woyke and U. Hentschel (2013). "Single-cell genomics reveals complex carbohydrate degradation patterns in poribacterial symbionts of marine sponges." ISME J **7**(12): 2287-2300.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe (2016). "KEGG as a reference resource for gene and protein annotation." Nucleic Acids Res **44**(D1): D457-462.

Karp, P. D., M. Riley, S. M. Paley and A. Pellegrini-Toole (2002). "The MetaCyc Database."

Nucleic Acids Res **30**(1): 59-61.

Kawamura, Y., X. G. Hou, F. Sultana, H. Miura and T. Ezaki (1995). "Determination of 16S rRNA sequences of Streptococcus mitis and Streptococcus gordonii and phylogenetic relationships among members of the genus Streptococcus." Int J Syst Bacteriol **45**(2): 406-408.

Kerem, E., M. Corey, B. Kerem, P. Durie, L. C. Tsui and H. Levison (1989). "Clinical and genetic comparisons of patients with cystic fibrosis, with or without meconium ileus." J Pediatr **114**(5): 767-773.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." J Mol Evol **16**(2): 111-120.

Klenk, H. P. and M. Goker (2010). "En route to a genome-based classification of Archaea and Bacteria?" Syst Appl Microbiol **33**(4): 175-182.

Klinger, K. W. (1983). "Cystic fibrosis in the Ohio Amish: gene frequency and founder effect." Hum Genet **65**(2): 94-98.

Konstantinidis, K. T. and J. M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes." Proc Natl Acad Sci U S A **102**(7): 2567-2572.

Koser, C. U., M. J. Ellington, E. J. Cartwright, S. H. Gillespie, N. M. Brown, M. Farrington, M. T. Holden, G. Dougan, S. D. Bentley, J. Parkhill and S. J. Peacock (2012). "Routine use of microbial whole genome sequencing in diagnostic and public health microbiology." PLoS Pathog **8**(8): e1002824.

Kristensen, A. S., J. Andersen, T. N. Jorgensen, L. Sorensen, J. Eriksen, C. J. Loland, K. Stromgaard and U. Gether (2011). "SLC6 neurotransmitter transporters: structure, function, and regulation." Pharmacol Rev **63**(3): 585-640.

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

Laing, C., C. Buchanan, E. N. Taboada, Y. Zhang, A. Kropinski, A. Villegas, J. E. Thomas and V. P. Gannon (2010). "Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions." BMC Bioinformatics **11**: 461.

Langille, M. G. and F. S. Brinkman (2009). "Bioinformatic detection of horizontally transferred DNA in bacterial genomes." F1000 Biol Rep **1**: 25.

Langille, M. G., W. W. Hsiao and F. S. Brinkman (2008). "Evaluation of genomic island predictors using a comparative genomics approach." BMC Bioinformatics **9**: 329.

Langton Hewer, S. C. and A. R. Smyth (2017). "Antibiotic strategies for eradicating Pseudomonas aeruginosa in people with cystic fibrosis." Cochrane Database Syst Rev **4**(4): CD004197.

Levison, H. (1980). "Perspectives in cystic fibrosis." Proc Inst Med Chic **33**(1): 2-6.

Lipuma, J. J. (2010). "The changing microbial epidemiology in cystic fibrosis." Clin Microbiol Rev **23**(2): 299-323.

Liu, B., D. Zheng, Q. Jin, L. Chen and J. Yang (2019). "VFDB 2019: a comparative pathogenomic platform with an interactive web interface." Nucleic Acids Res **47**(D1): D687-D692.

Loman, N. J. and M. J. Pallen (2015). "Twenty years of bacterial genome sequencing." Nat Rev Microbiol **13**(12): 787-794.

Lu, B. and H. W. Leong (2016). "Computational methods for predicting genomic islands in microbial genomes." Comput Struct Biotechnol J **14**: 200-206.

Lucas, S. K., R. Yang, J. M. Dunitz, H. C. Boyer and R. C. Hunter (2018). "16S rRNA gene sequencing reveals site-specific signatures of the upper and lower airways of cystic fibrosis patients." J Cyst Fibros **17**(2): 204-212.

Makarova, K. S., Y. I. Wolf, O. S. Alkhnbashi, F. Costa, S. A. Shah, S. J. Saunders, R. Barrangou, S. J. Brouns, E. Charpentier, D. H. Haft, P. Horvath, S. Moineau, F. J. Mojica, R. M. Terns, M. P. Terns, M. F. White, A. F. Yakunin, R. A. Garrett, J. van der Oost, R. Backofen and E. V. Koonin (2015). "An updated evolutionary classification of CRISPR-Cas systems." Nat Rev Microbiol **13**(11): 722-736.

Makarova, K. S., Y. I. Wolf, J. Iranzo, S. A. Shmakov, O. S. Alkhnbashi, S. J. J. Brouns, E. Charpentier, D. Cheng, D. H. Haft, P. Horvath, S. Moineau, F. J. M. Mojica, D. Scott, S. A. Shah, V. Siksnys, M. P. Terns, C. Venclovas, M. F. White, A. F. Yakunin, W. Yan, F. Zhang, R. A. Garrett, R. Backofen, J. van der Oost, R. Barrangou and E. V. Koonin (2020). "Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants." Nat Rev Microbiol **18**(2): 67-83.

Medini, D., C. Donati, H. Tettelin, V. Masignani and R. Rappuoli (2005). "The microbial pan-genome." Curr Opin Genet Dev **15**(6): 589-594.

Meier-Kolthoff, J. P., A. F. Auch, H. P. Klenk and M. Goker (2013). "Genome sequence-based species delimitation with confidence intervals and improved distance functions." BMC Bioinformatics **14**: 60.

Meier-Kolthoff, J. P., R. L. Hahnke, J. Petersen, C. Scheuner, V. Michael, A. Fiebig, C. Rohde, M. Rohde, B. Fartmann, L. A. Goodwin, O. Chertkov, T. Reddy, A. Pati, N. N. Ivanova, V. Markowitz, N. C. Kyrpides, T. Woyke, M. Goker and H. P. Klenk (2014). "Complete genome sequence of DSM 30083(T), the type strain (U5/41(T)) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy." Stand Genomic Sci **9**: 2.

Mowat, E., S. Paterson, J. L. Fothergill, E. A. Wright, M. J. Ledson, M. J. Walshaw, M. A. Brockhurst and C. Winstanley (2011). "Pseudomonas aeruginosa population diversity and turnover in cystic fibrosis chronic infections." Am J Respir Crit Care Med **183**(12): 1674-1679.

Muhlebach, M. S. (2017). "Methicillin-resistant Staphylococcus aureus in cystic fibrosis: how should it be managed?" Curr Opin Pulm Med **23**(6): 544-550.

Murray, N. E. (2000). "Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle)." Microbiol Mol Biol Rev **64**(2): 412-434.

Nekrasov, S. V., O. V. Agafonova, N. G. Belogurova, E. P. Delver and A. A. Belogurov (2007). "Plasmid-encoded antirestriction protein ArdA can discriminate between type I methyltransferase and complete restriction-modification system." J Mol Biol **365**(2): 284-297.

Nielsen, T. K., M. Rasmussen, S. Demaneche, S. Cecillon, T. M. Vogel and L. H. Hansen (2017). "Evolution of Sphingomonad Gene Clusters Related to Pesticide Catabolism Revealed by Genome Sequence and Mobilomics of Sphingobium herbicidovorans MH." Genome Biol Evol **9**(9): 2477-2490.

Nolan, G., P. Moivor, H. Levison, P. C. Fleming, M. Corey and R. Gold (1982). "Antibiotic prophylaxis in cystic fibrosis: inhaled cephaloridine as an adjunct to oral cloxacillin." J Pediatr **101**(4): 626-630.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy and K. D. Pruitt (2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." Nucleic Acids Res **44**(D1): D733-745.

O'Toole, G. A. (2018). "Cystic Fibrosis Airway Microbiome: Overturning the Old, Opening the Way for the New." J Bacteriol **200**(4).

Ochman, H., J. G. Lawrence and E. A. Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.

Ondov, B. D., T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren and A. M. Phillippy (2016). "Mash: fast genome and metagenome distance estimation using MinHash." Genome Biol **17**(1): 132.

Overbeek, R., R. Olson, G. D. Pusch, G. J. Olsen, J. J. Davis, T. Disz, R. A. Edwards, S. Gerdes, B. Parrello, M. Shukla, V. Vonstein, A. R. Wattam, F. Xia and R. Stevens (2014). "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)." Nucleic Acids Res **42**(Database issue): D206-214.

Pallen, M. J. and B. W. Wren (2007). "Bacterial pathogenomics." Nature **449**(7164): 835-842.

Parkins, M. D., C. D. Sibley, M. G. Surette and H. R. Rabin (2008). "The Streptococcus milleri group--an unrecognized cause of disease in cystic fibrosis: a case series and literature review." Pediatr Pulmonol **43**(5): 490-497.

Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz and G. W. Tyson (2015). "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." Genome Res **25**(7): 1043-1055.

Pattison, S. H., G. B. Rogers, M. Crockard, J. S. Elborn and M. M. Tunney (2013). "Molecular detection of CF lung pathogens: current status and future potential." J Cyst Fibros **12**(3): 194-205.

Perez-Sepulveda, B. M., A. V. Predeus, W. Y. Fong, C. M. Parry, J. Cheesbrough, P. Wigley, N. A. Feasey and J. C. D. Hinton (2021). "Complete Genome Sequences of African Salmonella enterica Serovar Enteritidis Clinical Isolates Associated with Bloodstream Infection." Microbiol Resour Announc **10**(12).

Pervez, M. T., M. J. U. Hasnain, S. H. Abbas, M. F. Moustafa, N. Aslam and S. S. M. Shah (2022). "A Comprehensive Review of Performance of Next-Generation Sequencing Platforms." Biomed Res Int **2022**: 3457806.

Peters, B. M., M. A. Jabra-Rizk, G. A. O'May, J. W. Costerton and M. E. Shirtliff (2012). "Polymicrobial interactions: impact on pathogenesis and human disease." Clin Microbiol Rev **25**(1): 193-213.

Petit, R. A., 3rd and T. D. Read (2020). "Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes." mSystems **5**(4).

Pettit, R. S. and C. Fellner (2014). "CFTR Modulators for the Treatment of Cystic Fibrosis." P T **39**(7): 500-511.

Pevzner, P. A., H. Tang and M. S. Waterman (2001). "An Eulerian path approach to DNA fragment assembly." Proc Natl Acad Sci U S A **98**(17): 9748-9753.

Pollitt, R. J., A. Dalton, S. Evans, H. N. Hughes and D. Curtis (1997). "Neonatal screening for cystic fibrosis in the Trent region (UK): two-stage immunoreactive trypsin screening compared with a three-stage protocol with DNA analysis as an intermediate step." J Med Screen **4**(1): 23-28.

Quinton, P. M. (1999). "Physiological basis of cystic fibrosis: a historical perspective." Physiol Rev **79**(1 Suppl): S3-S22.

Rabbani, B., M. Tekin and N. Mahdieh (2014). "The promise of whole-exome sequencing in medical genetics." J Hum Genet **59**(1): 5-15.

Rakhimova, E., A. Munder, L. Wiehlmann, F. Bredenbruch and B. Tummler (2008). "Fitness of isogenic colony morphology variants of Pseudomonas aeruginosa in murine airway infection." PLoS One **3**(2): e1685.

Rath, D., L. Amlinger, A. Rath and M. Lundgren (2015). "The CRISPR-Cas immune system: biology, mechanisms and applications." Biochimie **117**: 119-128.

Ratjen, F., S. C. Bell, S. M. Rowe, C. H. Goss, A. L. Quittner and A. Bush (2015). "Cystic fibrosis." Nat Rev Dis Primers **1**: 15010.

Richards, V. P., S. R. Palmer, P. D. Pavinski Bitar, X. Qin, G. M. Weinstock, S. K. Highlander, C. D. Town, R. A. Burne and M. J. Stanhope (2014). "Phylogenomics and the dynamic genome evolution of the genus Streptococcus." Genome Biol Evol **6**(4): 741-753.

Richter, M. and R. Rossello-Mora (2009). "Shifting the genomic gold standard for the prokaryotic species definition." Proc Natl Acad Sci U S A **106**(45): 19126-19131.

Roberts, R. J., T. Vincze, J. Posfai and D. Macelis (2015). "REBASE--a database for DNA restriction and modification: enzymes, genes and genomes." Nucleic Acids Res **43**(Database issue): D298-299.

Rodic, A., B. Blagojevic, E. Zdobnov, M. Djordjevic and M. Djordjevic (2017). "Understanding key features of bacterial restriction-modification systems through quantitative modeling." BMC Syst Biol **11**(Suppl 1): 377.

Roer, L., R. S. Hendriksen, P. Leekitcharoenphon, O. Lukjancenko, R. S. Kaas, H. Hasman and F. M. Aarestrup (2016). "Is the Evolution of Salmonella enterica subsp. enterica Linked to Restriction-Modification Systems?" mSystems **1**(3).

Rogers, G. B., K. D. Bruce, M. L. Martin, L. D. Burr and D. J. Serisier (2014). "The effect of long-term macrolide treatment on respiratory microbiota composition in non-cystic fibrosis bronchiectasis: an analysis from the randomised, double-blind, placebo-controlled BLESS trial." Lancet Respir Med **2**(12): 988-996.

Rommens, J. M., M. C. Iannuzzi, B. Kerem, M. L. Drumm, G. Melmer, M. Dean, R. Rozmahel, J. L. Cole, D. Kennedy, N. Hidaka and et al. (1989). "Identification of the cystic fibrosis gene: chromosome walking and jumping." Science **245**(4922): 1059-1065.

Rosenstein, B. J., T. S. Langbaum and K. Winn (1984). "Unexpected diagnosis of cystic fibrosis at autopsy." South Med J **77**(11): 1383-1385.

Rozen, R., R. H. Schwartz, B. C. Hilman, P. Stanislovitis, G. T. Horn, K. Klinger, J. Daigneault, M. De Braekeleer, B. Kerem, L. Tsui and et al. (1990). "Cystic fibrosis mutations in North American populations of French ancestry: analysis of Quebec French-Canadian and Louisiana Acadian families." Am J Hum Genet **47**(4): 606-610.

Saiman, L., Y. Chen, S. Tabibi, P. San Gabriel, J. Zhou, Z. Liu, L. Lai and S. Whittier (2001). "Identification and antimicrobial susceptibility of Alcaligenes xylosoxidans isolated from patients with cystic fibrosis." J Clin Microbiol **39**(11): 3942-3945.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**(4): 406-425.

Salzberg, S. L., A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop and J. A. Yorke (2012). "GAGE: A critical evaluation of genome assemblies and assembly algorithms." Genome Res **22**(3): 557-567.

Sathe, M. and R. Houwen (2017). "Meconium ileus in Cystic Fibrosis." J Cyst Fibros **16 Suppl 2**: S32-S39.

Schmidt, H. and M. Hensel (2004). "Pathogenicity islands in bacterial pathogenesis." Clin Microbiol Rev **17**(1): 14-56.

Schurch, A. C., S. Arredondo-Alonso, R. J. L. Willems and R. V. Goering (2018). "Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches." Clin Microbiol Infect **24**(4): 350-354.

Scoffone, V. C., L. R. Chiarelli, G. Trespidi, M. Mentasti, G. Riccardi and S. Buroni (2017). "Burkholderia cenocepacia Infections in Cystic Fibrosis Patients: Drug Resistance and

Therapeutic Approaches." Front Microbiol **8**: 1592.

Scotet, V., M. de Braekeleer, M. Roussey, G. Rault, P. Parent, M. Dagorne, H. Journel, A. Lemoigne, J. P. Codet, M. Catheline, V. David, A. Chaventre, I. Dugueperoux, C. Verlingue, I. Quere, B. Mercier, M. P. Audrezet and C. Ferec (2000). "Neonatal screening for cystic fibrosis in Brittany, France: assessment of 10 years' experience and impact on prenatal diagnosis." Lancet **356**(9232): 789-794.

Scotet, V., C. L'Hostis and C. Ferec (2020). "The Changing Epidemiology of Cystic Fibrosis: Incidence, Survival and Impact of the CFTR Gene Discovery." Genes (Basel) **11**(6).

Scott, J. E. and G. A. O'Toole (2019). "The Yin and Yang of Streptococcus Lung Infections in Cystic Fibrosis: a Model for Studying Polymicrobial Interactions." J Bacteriol **201**(11).

Servidoni, M. F., C. C. S. Gomez, F. A. L. Marson, A. Toro, M. Ribeiro, J. D. Ribeiro, A. F. Ribeiro and C. Grupo Colaborativo de Estudos em Fibrose (2017). "Sweat test and cystic fibrosis: overview of test performance at public and private centers in the state of Sao Paulo, Brazil." J Bras Pneumol **43**(2): 121-128.

Shabbir, M. A., H. Hao, M. Z. Shabbir, Q. Wu, A. Sattar and Z. Yuan (2016). "Bacteria vs. Bacteriophages: Parallel Evolution of Immune Arsenals." Front Microbiol **7**: 1292.

Shabbir, M. A. B., M. Z. Shabbir, Q. Wu, S. Mahmood, A. Sajid, M. K. Maan, S. Ahmed, U. Naveed, H. Hao and Z. Yuan (2019). "CRISPR-cas system: biological function in microbes and its use to treat antimicrobial resistant pathogens." Ann Clin Microbiol Antimicrob **18**(1): 21.

Sharon, I. and J. F. Banfield (2013). "Microbiology. Genomes from metagenomics." Science **342**(6162): 1057-1058.

Shwachman, H. and L. L. Kulczycki (1958). "Long-term study of one hundred five patients with cystic fibrosis; studies made over a five- to fourteen-year period." AMA J Dis Child **96**(1): 6-15.

Sibley, C. D., M. E. Grinwis, T. R. Field, M. D. Parkins, J. C. Norgaard, D. B. Gregson, H. R. Rabin and M. G. Surette (2010). "McKay agar enables routine quantification of the 'Streptococcus milleri' group in cystic fibrosis patients." J Med Microbiol **59**(Pt 5): 534-540.

Sibley, C. D., M. D. Parkins, H. R. Rabin, K. Duan, J. C. Norgaard and M. G. Surette (2008). "A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in cystic fibrosis patients." Proc Natl Acad Sci U S A **105**(39): 15070-15075.

Sibley, C. D., H. Rabin and M. G. Surette (2006). "Cystic fibrosis: a polymicrobial infectious disease." Future Microbiol **1**(1): 53-61.

Simpson, J. T. and M. Pop (2015). "The Theory and Practice of Genome Sequence Assembly." Annu Rev Genomics Hum Genet **16**: 153-172.

Sjogren, I. (2006). "Nils Rosen von Rosenstein--the father of paediatrics." Ups J Med Sci **111**(1): 3-16.

Slager, J., R. Aprianto and J. W. Veening (2018). "Deep genome annotation of the opportunistic human pathogen Streptococcus pneumoniae D39." Nucleic Acids Res **46**(19): 9971-9989.

Sohn, J. I. and J. W. Nam (2018). "The present and future of de novo whole-genome assembly." Brief Bioinform **19**(1): 23-40.

Soo, R. M., C. T. Skennerton, Y. Sekiguchi, M. Imelfort, S. J. Paech, P. G. Dennis, J. A. Steen, D. H. Parks, G. W. Tyson and P. Hugenholtz (2014). "An expanded genomic representation of the phylum cyanobacteria." Genome Biol Evol **6**(5): 1031-1045.

Souvorov, A., R. Agarwala and D. J. Lipman (2018). "SKESA: strategic k-mer extension for scrupulous assemblies." Genome Biol **19**(1): 153.

Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. D. Grimont, P. Kampfer, M. C. J. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward and W. B. Whitman (2002). "Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology." Int J Syst Evol Microbiol **52**(Pt 3): 1043-1047.

Stark, R., M. Grzelak and J. Hadfield (2019). "RNA sequencing: the teenage years." Nat Rev Genet **20**(11): 631-656.

Stevens, D. L. (2003). "Dilemmas in the treatment of invasive Streptococcus pyogenes infections." Clin Infect Dis **37**(3): 341-343.

Super, M. (1975). "Cystic fibrosis in the South West African Afrikaner. An example of population drift, possibly with heterozygote advantage." S Afr Med J **49**(20): 818-820.

Surette, M. G. (2014). "The cystic fibrosis lung microbiome." Ann Am Thorac Soc **11 Suppl 1**: S61-65.

Suzuki, N., M. Seki, Y. Nakano, Y. Kiyoura, M. Maeno and Y. Yamashita (2005). "Discrimination of Streptococcus pneumoniae from viridans group streptococci by genomic subtractive hybridization." J Clin Microbiol **43**(9): 4528-4534.

Taccetti, G., M. Francalanci, G. Pizzamiglio, B. Messore, V. Carnovale, G. Cimino and M. Cipolli (2021). "Cystic Fibrosis: Recent Insights into Inhaled Antibiotic Treatment and Future Perspectives." Antibiotics (Basel) **10**(3).

Takahashi, K. and M. Nei (2000). "Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used." Mol Biol Evol **17**(8): 1251-1258.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar (2011). "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods." Mol Biol Evol **28**(10): 2731-2739.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli and C. M. Fraser (2005). "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"." Proc Natl Acad Sci U S A **102**(39): 13950-13955.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res **25**(24): 4876-4882.

Tindall, E. A., D. C. Petersen, S. Nikolaysen, W. Miller, S. C. Schuster and V. M. Hayes (2010). "Interpretation of custom designed Illumina genotype cluster plots for targeted association studies and next-generation sequence validation." BMC Res Notes **3**: 39.

Towns, S. J. and S. C. Bell (2011). "Transition of adolescents with cystic fibrosis from paediatric to adult care." Clin Respir J **5**(2): 64-75.

Tsui, L. C., M. Buchwald, D. Barker, J. C. Braman, R. Knowlton, J. W. Schumm, H. Eiberg, J. Mohr, D. Kennedy, N. Plavsic and et al. (1985). "Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker." Science **230**(4729): 1054-1057.

Tuchman, L. K., L. A. Schwartz, G. S. Sawicki and M. T. Britto (2010). "Cystic fibrosis and transition to adult medical care." Pediatrics **125**(3): 566-573.

Tucker, T., M. Marra and J. M. Friedman (2009). "Massively parallel sequencing: the next big thing in genetic medicine." Am J Hum Genet **85**(2): 142-154.

Turcios, N. L. (2020). "Cystic Fibrosis Lung Disease: An Overview." Respir Care **65**(2): 233-251.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight and J. I. Gordon (2007). "The human microbiome project." Nature **449**(7164): 804-810.

van Dorst, J. M., R. Y. Tam and C. Y. Ooi (2022). "What Do We Know about the Microbiome in Cystic Fibrosis? Is There a Role for Probiotics and Prebiotics?" Nutrients **14**(3).

Vandeplassche, E., T. Coenye and A. Crabbe (2017). "Developing selective media for quantification of multispecies biofilms following antibiotic treatment." PLoS One **12**(11): e0187540.

Vandeplassche, E., A. Sass, A. Lemarcq, A. A. Dandekar, T. Coenye and A. Crabbe (2019). "In vitro evolution of Pseudomonas aeruginosa AA2 biofilms in the presence of cystic fibrosis lung microbiome members." Sci Rep **9**(1): 12859.

Vandeplassche, E., S. Tavernier, T. Coenye and A. Crabbe (2019). "Influence of the lung microbiome on antibiotic susceptibility of cystic fibrosis pathogens." Eur Respir Rev **28**(152).

Vasu, K. and V. Nagaraja (2013). "Diverse functions of restriction-modification systems in addition to cellular defense." Microbiol Mol Biol Rev **77**(1): 53-72.

Virolle, C., K. Goldlust, S. Djermoun, S. Bigot and C. Lesterlin (2020). "Plasmid Transfer by Conjugation in Gram-Negative Bacteria: From the Cellular to the Community Level." Genes (Basel) **11**(11).

Voronina, O. L., N. N. Ryzhova, M. S. Kunda, E. V. Loseva, E. I. Aksenova, E. L. Amelina, G. L. Shumkova, O. I. Simonova and A. L. Gintsburg (2020). "Characteristics of the Airway Microbiome of Cystic Fibrosis Patients." Biochemistry (Mosc) **85**(1): 1-10.

Waack, S., O. Keller, R. Asper, T. Brodag, C. Damm, W. F. Fricke, K. Surovcik, P. Meinicke and R. Merkl (2006). "Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models." BMC Bioinformatics **7**: 142.

Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." Nature **171**(4356): 737-738.

Weber, K., J. Delben, T. G. Bromage and S. Duarte (2014). "Comparison of SEM and VPSEM imaging techniques with respect to Streptococcus mutans biofilm topography." FEMS Microbiol Lett **350**(2): 175-179.

Welch, R. A., V. Burland, G. Plunkett, 3rd, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. Mobley, M. S. Donnenberg and F. R. Blattner (2002). "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli." Proc Natl Acad Sci U S A **99**(26): 17020-17024.

Wheat, L. J., P. Connolly, M. Smedema, E. Brizendine, R. Hafner, A. C. T. Group, A. the Mycoses Study Group of the National Institute of and D. Infectious (2001). "Emergence of resistance to fluconazole as a cause of failure during treatment of histoplasmosis in patients with acquired immunodeficiency disease syndrome." Clin Infect Dis **33**(11): 1910-1913.

Whiley, R. A., N. P. Sheikh, N. Mushtaq, E. Hagi-Pavli, Y. Personne, D. Javaid and R. D. Waite (2014). "Differential potentiation of the virulence of the Pseudomonas aeruginosa cystic fibrosis liverpool epidemic strain by oral commensal Streptococci." J Infect Dis **209**(5): 769-780.

Whiteson, K. L., B. Bailey, M. Bergkessel, D. Conrad, L. Delhaes, B. Felts, J. K. Harris, R. Hunter, Y. W. Lim, H. Maughan, R. Quinn, P. Salamon, J. Sullivan, B. D. Wagner and P. B. Rainey (2014). "The upper respiratory tract as a microbial source for pulmonary infections in cystic fibrosis. Parallels from island biogeography." Am J Respir Crit Care Med **189**(11): 1309-1315.

Williams, C., D. Davies and R. Williamson (1993). "Segregation of delta F508 and normal CFTR alleles in human sperm." Hum Mol Genet **2**(4): 445-448.

Wilson, D. J. (2012). "Insights from genomics into bacterial pathogen populations." PLoS Pathog **8**(9): e1002874.

Winstanley, C. and J. L. Fothergill (2009). "The role of quorum sensing in chronic cystic fibrosis Pseudomonas aeruginosa infections." FEMS Microbiol Lett **290**(1): 1-9.

Wright, S. W. and N. E. Morton (1968). "Genetic studies on cystic fibrosis in Hawaii." Am J Hum Genet **20**(2): 157-169.

Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'Haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J. F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H. P. Klenk and J. A. Eisen (2009). "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea." Nature **462**(7276): 1056-1060.

Zachariah, P., C. Ryan, S. Nadimpalli, G. Coscia, M. Kolb, H. Smith, M. Foca, L. Saiman and P. J. Planet (2018). "Culture-Independent Analysis of Pediatric Bronchoalveolar Lavage Specimens." Ann Am Thorac Soc **15**(9): 1047-1056.

Zankari, E., H. Hasman, R. S. Kaas, A. M. Seyfarth, Y. Agerso, O. Lund, M. V. Larsen and F. M. Aarestrup (2013). "Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing." J Antimicrob Chemother **68**(4): 771-777.

Zhao, J., P. D. Schloss, L. M. Kalikin, L. A. Carmody, B. K. Foster, J. F. Petrosino, J. D. Cavalcoli, D. R. VanDevanter, S. Murray, J. Z. Li, V. B. Young and J. J. LiPuma (2012). "Decade-long bacterial community dynamics in cystic fibrosis airways." Proc Natl Acad Sci U S A **109**(15): 5809-5814.

Appendix 1

*
N/A