SEQUENCING AND COMPARATIVE GENOME ANALYSIS OF *STREPTOCOCCUS SANGUINIS* AND *STREPTOCOCCUS GORDONII*

ZHENG WENNING

DEPARTMENT OF ORAL AND CRANIOFACIAL SCIENCES
FACULTY OF DENTISTRY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

SEQUENCING AND COMPARATIVE GENOME ANALYSIS OF *STREPTOCOCCUS SANGUINIS* AND *STREPTOCOCCUS GORDONII*

ZHENG WENNING

THESIS SUBMITTED IN FULFILMENTOF THE REQUIREMENTSFOR THE DEGREE OF DOCTOR OF PHILOPSOPHY

DEPARTMENT OF ORAL AND CRANIOFACIAL SCIENCES
FACULTY OF DENTISTRY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: ZHENG WENNING (I.C/Passport No: 911207015734)

Registration/Matric No: DHA130013

Name of Degree: Doctor of Philosophy

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

SEQUENCING AND COMPARATIVE GENOME ANALYSIS OF *STREPTOCOCCUS SANGUINIS* AND *STREPTOCOCCUS GORDONII*

Field of Study:

BIOINFORMATICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

Mitis group oral streptococci are opportunistic human pathogens that live primarily in the oral cavity, and potentially cause infective endocarditis (IE) in neutropenic patients with haematological disease. Among the members of Mitis group, *Streptococcus sanguinis* and *Streptococcus gordonii* are two pioneer colonizing species of dental plaque that are often associated with streptococcal IE infection. In this study, comparative genome analyses of these two closely-related species were performed in order to provide a better understanding of the co-existence of *S. gordonii* and *S. sanguinis* and their biology, evolution, genomics and virulence in invasive infections. Here I have successfully sequenced, assembled, identified and annotated 27 strains of 6 species of Mitis group oral streptococci that were isolated from patients in different geographical areas.

A comparative whole-genome study was performed on 14 *S. gordonii* strains and 5 *S. sanguinis* strains along with the reference strains of *S. gordonii* Challis and *S. sanguinis* SK36 using different bioinformatics approaches such as phylogenetic analysis, functional enrichment analysis, orthologous genes and pan-genome analysis, comparative pathogenomics analysis, comparative prophage analysis and genomic island (GI) analysis. The data showed that both species have generally high sequence homology with evidence of their considerable number of core genes and virulence genes. Significantly, *S. sanguinis* carries genes involved in nickel, cobalt and cobalamin utilization in their core genomes. Interestingly, both *S. sanguinis* and *S. gordonii* harbour open pan-genomes, indicating their potential in exhibiting greater virulence by acquiring antibiotic resistance and new virulence genes in the future. While *S. gordonii* has been found to recruit additional copies of *ComCDE* quorum-sensing system as competence mechanism to support its virulence, *S. sanguinis* has acquired a broad array

of potential antibiotic resistance genes including the drug/metabolite transporter (DMT) superfamily, *rsmE,* TetR/AcrR family transcriptional regulator (TFR) and GNAT acetyltransferase through horizontally-transferred GIs to further support its bacterial adaptation in the host cells during infections.

To facilitate the expanding *Streptococcus* genus research worldwide, I developed a Mitis group oral streptococci genomic resource and analysis platform, StreptoBase (http://streptococcus.um.edu.my), allowing researchers to access and browse the comprehensive *Streptococcus* genomes and annotations. It currently hosts 104 Mitis group genomes including 27 strains which were sequenced using the high-throughput Illumina HiSeq technology platform, enabling comparative analyses and visualization of both cross-species and cross-strain characteristics of Mitis group bacteria. StreptoBase incorporates sophisticated in-house designed bioinformatics web tools such as Pairwise Genome Comparison (PGC) tool and Pathogenomic Profiling Tool (PathoProT), which facilitate comparative pathogenomics analysis of *Streptococcus* strains.

In conclusion, this comparative genome study has successfully characterised the core genomes of *S. sanguinis* and *S. gordonii*, and identified key differences between the species. These new insights into the genomic differences, biology and virulence of the two closely-related species will provide a foundation for further investigations into how these bacteria make the transition from oral commensal species into important pathogens in IE. Ultimately, this may lead to improved measures for dental plaque control and/or better management of diseases caused by these opportunistic pathogens. With addition of new genome sequences of *S. sanguinis* and *S. gordonii* in StreptoBase, it will be an invaluable platform to accelerate Mitis group streptococci research in their impact on human health and disease.

ABSTRAK

Kumpulan Mitis streptokoki mulut merupakan patogen oportunis manusia yang berhabitat di rongga mulut dan berpotensi menyebabkan Endokarditis Infektif (IE) kepada pesakit neutropenic yang menjangkiti penyakit hematologi. Antara ahli-ahli kumpulan Mitis, *Streptococcus sanguinis* dan *Streptococcus gordonii* adalah dua spesies perintis menjajah plak gigi yang sering diasosiasikan dengan penyakit streptococcal IE. Untuk kajian penyelidikan ini, genom perbandingan analisis bagi kedua-dua spesies yang berkait rapat telah dilakukan untuk menimbul pengetahuan mengenai *S. gordonii* dan *S. sanguinis* dalam aspek biologi, evolusi, genomik dan virulen dalam jangkitan pergigian invasif. Di sini saya telah berjaya menyusunkan, menentukan dan menganotasikan 27 strain dari 6 spesies Mitis kumpulan streptococci oral yang diperolehi melalui klinikal prosedur dari pesakit antarabangsa.

Saya juga telah melancarkan penyelidikan genom perbandingan bagi 14 strain *S. gordonii* dan 5 *S. sanguinis* strain bersampingan dengan strain rujukan *S. gordonii* Challis dan *S. sanguinis* SK36 dengan mengaplikasikan pendekatan bioinformatik yang berbeza seperti analisis filogenetik, analisis gen berfungsi, kesamaan gen dan analisis pan-genom, pathogenomics perbandingan analisis, analisis prophage perbandingan dan analisis pulau genomik (GI). Data saya menunjukkan bahawa kedua-dua spesies mempunyai homologi yang tinggi dibuktikan dengan sebilangan besar kesamaan gen dan gen virulen. Secara khususnya, *S. sanguinis* membawa gen yang melibati dalam proses nikel, kobalt dan kobalamin di dalam kesamaan genom *S. sanguinis*. Tambahan pula, kedua-dua *S. sanguinis* dan *S. gordonii* mempunyai pan-genom terbuka. Ini menunjukkan potensi *S. sanguinis* and *S. gordonii* dalam penguguhan virulen dengan memperolehi gen rintangan antibiotik dan gen baru pada masa depan. *S. gordonii* telah memiliki salinan tambahan ComCDE sistem penderiaan kuorum sebagai mekanisme

untuk menabahkan virulen. Jika dibandingkan dengan *S. gordonii, S. sanguinis* telah memperolehi pelbagai jenis gen rintangan antibiotik termasuk kumpulan dadah / metabolit pengangkut (DMT), kumpulan RSME, TetR / AcrR kumpulan pengawal selia transkripsi (TFR) dan GNAT acetyltransferase melalui GI untuk terus menyokong penyesuaian bakteria dalam sel pesakit semasa jangkitan.

Untuk melajukan penyelidikan genus *Streptococcus* yang berkembang di seluruh dunia, saya menubuhkan StreptoBase (http://streptococcus.um.edu.my) sebagai satu sumber sumber genomik dan analisis plakfom bagi streptokoki mulut kumpulan Mitis. StreptoBase berfungsi untuk memudahkan penyelidik untuk mengakses dan melayari keseluruhan genom *Streptococcus* dengan lebih teliti dan menyeluruh. Kini, StreptoBase mengandungi 104 genom kumpulan Mitis termasuk 27 jenis yang telah disusun menggunakan platform teknologi Illumina HiSeq yang berprosesan tinggi. Selain itu, StreptoBase dapat menandaskan analisis perbandingan dan visualisasi bagi kedua-dua spesies sekali gus dengan ciri-ciri bakteria. StreptoBase melindungi peralatan bioinformatik yang canggih seperti alat perbandingan selaras gen (PGC) dan alat patogenomik (PathoProT) untuk melampiaskan perbandingan analisis patogenomik bagi bakteria *Streptococcus*.

Kesimpulannya, ulasan sistematik ini telah berjaya memperbandingkan ciri-ciri genom bagi *S. sanguinis* dan *S. gordonii* disamping itu juga menilaikan perbezaan utama antara spesies *Streptococcus*. Bukti-bukti novel ini yang melampaikan perbezaan genomik, biologi dan kebisaan *S. gordonii* dan *S. sanguinis* akan menyediakan asas untuk pengajian lanjut dalam cara evolusi peralihan bakteria ini dari bakteria mulut sehingga menjadi patogen penting IE. Seterusnya, penyelidikan ini dapat memanfaatkan dalam pencegahan dan pengurusan penyakit *Streptococcus* melalui langkah-langkah

pengawalan plak gigi. Dengan penambahan genom baru *S. sanguinis* dan *S. gordonii* dalam StreptoBase, ia akan memuncul sebagai satu plakfom yang berharga dan penting untuk memacukan penyelidikan kumpulan Mitis streptokoki mulut bagi menjamin kesihatan pergigian masyarakat.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisors, Dr Lawrence Choo Siew Woh, Dr. Nick Jakubovics and Prof. Dr. Ian Charles Paterson for their continuous support of my Ph.D study, for his enormous patience, motivation, enthusiasm, and immense knowledge. In particular, I am grateful to Dr. Lawrence Choo Siew Woh for his guidance and invaluable advice throughout my entire research period in producing my published journals and writing of this thesis.

Moreover, I would also like to thank my fellow colleagues in Genome Informatics Research Group (GIRG) especially Wei Yee Wee, Shi Yang Tan and Mui Fern Tan for their stimulating discussions, insightful opinions and useful suggestions. I would also like to take this opportunity to thank the staffs from the High Impact Research (HIR) secretariat and the Faculty of Dentistry for their efiicient administration and management works.

Additionally, I extend my thanks to University of Malaya and Ministry of Education (MOHE), Malaysia which funded my study under the High Impact Research (HIR) Grant UM.C/HIR/MOHE/08 and UM Research Grant (UMRG) [Account No. UMRG: RG541-13HTM]. Furthermore, my sincere appreciation goes to the High Impact Research Committee for the financial aid of four semesters via the Graduate Research Assistantship Scheme (GRAS) for my PhD study. Besides, I am truly grateful for the two semesters Skim Biasiswa Universiti Malaya (SBUM) scholarship offered by University of Malaya.

Last but not the least, I would like to express my heartily thanks to my beloved family: my parents Clemence Chong and Nyok Chin Choo, for providing moral support and spiritual encouragement during the hardship of my Ph.D study.

TABLE OF CONTENTS

Abstract iii

Abstrak v

Acknowledgements viii

Table of Contents x

List of Figures xiv

List of Tables xvi

List of symbols and abbreviations xvii

CHAPTER 1: INTRODUCTION 1

1.1 Overview 1

1.2 Project Objectives 4

CHAPTER 2: LITERATURE REVIEW 5

2.1 Oral streptococci: Mitis group 5

2.2 Streptococcal Infective endocarditis (IE) 6

2.3 Next-Generation Sequencing (NGS) technologies 9

2.4 Phylogenetic analyses 12

2.5 Pan-genome analyses 18

2.6 Horizontal Gene Transfer (HGT) analysis 22

2.7 Genomic Island (GI) analysis 23

2.8 Prophage analysis 24

2.9 Summary 25

CHAPTER 3: MATERIALS AND METHODS 27

3.1 Bacterial strains and DNA extraction 27

3.2 Library preparation and next-generation sequencing 27

3.3 Raw read quality checking and preprocessing 28

3.4 Genome assembly and contamination checking 28

3.5 Genome annotation 29

3.6 Multiple sequence alignment (MSA) and phylogenetic inference 30

3.7 Orthologous gene family comparisons and pan-genome analysis 30

3.8 Functional enrichment analysis 31

3.9 Virulence gene prediction 32

3.10 Comparative prophage analysis 33

3.11 Comparative Genomic Island (GI) analysis 34

3.12 Development of Mitis group oral *Streptococcus* database – StreptoBase 34

CHAPTER 4: RESULTS – COMPARATIVE GENOMIC ANALYSIS OF *STREPTOCOCCUS SANGUINIS* AND *STREPTOCOCCUS GORDONII* 37

4.1 Sample source and genome assemblies 37

4.2 Genome Overview 37

4.3 Phylogenetic inference 39

4.4 Open pan-genomes of *S. sanguinis* and *S. gordonii* 40

4.5 Orthologous gene family comparisons 43

4.5.1 Porphyrin-containing compound biosynthetic process 45

4.5.2 Cobalamin biosynthetic process 46

4.6 Comparative prophage analysis 47

4.7 Comparative Pathogenomics Analysis 51

4.8 Comparative Genomic Island (GI) analysis 60

CHAPTER 5: RESULTS – DEVELOPMENT OF STREPTOBASE 69

5.1 Datasets of StreptoBase 69

5.2 Database Structure and Composition 75

5.2.1 *Streptococcus* Genome Browser (SGB) 79

5.2.2 Real-time keyword search engine 81

5.3 Database Features and Incorporated Bioinformatics Tools 81

5.3.1 Pairwise Genome Comparison (PGC) tool 82

5.3.2 Pathogenomics Profiling (PathoProT) tool 86

5.3.3 Sequence search tools 93

5.4 Availability and System Requirements 93

CHAPTER 6: DISCUSSION 94

6.1 Overview 94

6.2 Comparative genomic analyses of two closely related *S. sanguinis* and *S. gordonii* 94

6.3 Development of StreptoBase and bioinformatics tools 100

6.4 Limitations 102

6.5 Future Work 103

CHAPTER 7: CONCLUSION 105

REFERENCES 107

LIST OF PUBLICATIONS AND PAPERS PRESENTED 123

LIST OF PUBLICATIONS AND PAPERS PRESENTED 124

LIST OF PUBLICATIONS AND PAPERS PRESENTED 125

APPENDICES……126

LIST OF FIGURES

Figure 2.1: The mechanism of streptococcal IE..... 7

Figure 2.2: A series of species-specific of Streptococcal cell surface adhesions targeted by *Streptococcus* bacteria during dental plaque formation (Source: (Tilley and Kerrigan 2013) 9

Figure 2.3: The phylogenetic tree constructed based on 34 *Streptococcus* species using 16S rRNA sequences of the type strain of oral streptococci (Source: (Stackebrandt and Goebel 1994) 14

Figure 2.4: The *sodA*-based phylogenetic tree of *Streptococcus* Mitis group species with the 100 bootstrappings (Kawamura, Whiley et al. 1999). 15

Figure 2.5: The 50 ribosomal protein genes-based phylogenetic tree of *S. sinensis* HKU4T and the other 28 *Streptococcus* species.. 17

Figure 2.6: *S. agalactiae* core genome plot..... 19

Figure 2.7: Pan-genome plot of *S. agalactiae.*..... 20

Figure 2.8: Pan-genome of *S. pneumoniae*. 22

Figure 4.1: Phylogenetic inference of *S. gordonii* and *S. sanguinis*.. 40

Figure 4.2: Pan-genome analyses..... 43

Figure 4.3: Venn diagram showing comparative analysis of orthologous genes in *S. gordonii* and *S. sanguinis*. …..... 44

Figure 4.4: Functional enrichment analysis of *S. sanguinis*-specific core genes..... 45

Figure 4.5: Predicted intact prophages in *S. gordonii* and *S. sanguinis*..... 49

Figure 4.6: Illustration of rps/polysaccharide gene clusters of *S. gordonii* 38 and *S. gordonii* Challis in *Streptococcus* genomes..... 54

Figure 4.7: The visualization of *S. gordonii* Challis-type polysaccharide gene cluster structure in *S. gordonii* and *S. sanguinis* using Mauve software. 56

Figure 4.8: A heat map shows the main differences of virulence genes harbored by *S. sanguinis* and *S. gordonii.*..... 58

Figure 5.1: StreptoBase structure and composition.. 76

Figure 5.2: A flowchart shows the sequential processed web interfaces while browsing on StreptoBase. 79

Figure 5.3: A screenshot for visualising a genomic region of *S. sanguinis* NCTC 7863 in the SGB browser. 80

Figure 5.4: Pairwise genome comparison between *S. mitis* B6 *and S. mitis* 17/34 using PGC tool incorporated in StreptoBase. 84

Figure 5.5: A putative intact prophage detected in the genome of *S. mitis* B6. 85

Figure 5.6: A PathoProT flowchart.. 88

Figure 5.7: An informative heat map generated by PathoProT tool.. 89

LIST OF TABLES

Table 2.1: Summary of benchmarking analysis on different type of NGS technologies.. 10

Table 4.1: Summary of the genome features of 19 newly sequenced *S. gordonii* and *S. sanguinis* strains 38

Table 4.2: Summary of the comparative prophage analyses of *S. sanguinis* and *S. gordonii*... 48

Table 4.3: Overview of putative prophages including the size of the prophage, the number of CDS, ATT-site (special attachment site in the bacterial and phage genomes) status and GC content. 50

Table 4.4: Summary of predicted GIs in the genomes of *S. gordonii* and *S. sanguinis*.... 63

Table 4.5: The details of the putative GI including the size of the GI, the number of CDS, GC contents and key genes incorporated in each GI. 64

Table 5.1: The isolation details of 27 *Streptococcus* strains including the isolation source, geographical area and strain author..... 71

Table 5.2: The genome sequencing statistics of 27 oral streptococci strains using Next Generation Sequencing Illumina Hiseq 2000 platform..... 73

Table 5.3: The genome identity of the 27 isolated *Streptococcus* strains with the summary assembly statistics. 74

Table 5.4: The species table summarizes the total number of draft and complete genomes of each *Streptococcus* Mitis group species accordingly..... 75

Table 5.5: StreptoBase summary statistics..... 77

Table 5.6: The ATP synthases within the atp operon of *S. mitis* B6..... 86

LIST OF SYMBOLS AND ABBREVIATIONS

IE	Infective Endocarditis
NJ	Neighbour-Joining
MSA	Multiple Sequence Alignment
GI	Genomic Island
SNP	Single Nucleotide Polymorphism
HTML	HyperText Markup Language
PHP	HyperText Preprocessor
CSS	Cascading Style Sheets
LAMP	Linux, Apache, MySQL and PHP
MVC	Model-view-controller
RGP	Rhamnose glucose polymers
LCB	Locally Collinear Blocks
ORF	Open reading frame
TFR	TetR/AcrR family transcriptional regulator
DMT	Drug/metabolite transporter
PGC	Pairwise Genome Comparison
SGB	Streptococcus Genome Browser
HAD	Haloacid dehalogenase
NAD	Nicotinate-nucleotide adenylyltransferase
PEP	Phosphoenol pyruvate
GNAT	GCN5-related N-acetyltransferase

MGC	Microbial Genome Comparison
ACT	Artemis Comparison Tool
TB	Tuberculosis
MCL	Markov Cluster Algorithm

CHAPTER 1: INTRODUCTION

1.1 Overview

The human oral streptococci are commensals which often inhabit the gastrointestinal and the oral mucosa and tooth surfaces. In healthy individuals, streptococci can constitute more than 50% of the oral microbiota (Human Microbiome Project 2012) and these bacteria generally possess low pathogenic potential. However, oral streptococci can invade the bloodstream, and have the potential to cause infective endocarditis (IE) or septicaemia following antineoplastic therapy in neutropenic patients with haematological disease (Westling 2005). Other oral *Streptococcus*-associated conditions including odontofacial infections, brain abscesses and abdominal infections have also been reported (Westling 2005). The largest and most abundant group of oral streptococci is the Mitis group, which comprised of 13 species including some of the most common human oral colonizers such as *Streptococcus mitis*, *Streptococcus oralis*, *Streptococcus sanguinis* and *Streptococcus gordonii* as well as species such as *Streptococcus tigurinus*, *Streptococcus oligofermentans* and *Streptococcus australis* that have only recently been classified and are poorly understood at present.

Within the Mitis group, there are two distinct groups: Sinensis group (*S. sinensis, S. oligofermentans,* and *Streptococcus cristatus*) and Sanguinis group (*S. sanguinis* and *S. gordonii*) (Teng, Huang et al. 2014). The Sanguinis group members of *S. sanguinis* and *S. gordonii* were considered as a single species until the late 1980's, as their 16S rRNA sequences are highly homologous (Kilian, MIKKELSEN et al. 1989). *S. sanguinis* and *S. gordonii* are closely related not only in terms of their phylogenetic relationship but also in their biological traits (Teng, Hsueh et al. 2002). These two opportunistic pathogens are

often isolated from the same intraoral sites (Nobbs, Zhang et al. 2007) and also sometimes from patients with endocarditis or with neutropenic bloodstream infections (Presterl, Grisold et al. 2005). It is expected that these two *Streptococcus* species are likely to compete for the same array of host receptors, particularly as they express a similar set of surface proteins (Nobbs, Zhang et al. 2007). It has also been reported that *S. sanguinis* and *S. gordonii* can antagonize *S. mutans* through the production of H_2O_2, whereas *S. mutans* also competes with *S. sanguinis* and *S. gordonii* through bacteriocin secretion (Kreth, Zhang et al. 2008). Despite its low-level presence in the oral cavity, *S. gordonii* tends to persist and co-exist with *S. sanguinis* over time even though *S. sanguinis* is almost always more abundant (Nobbs, Zhang et al. 2007). Investigating the genomic differences between *S. gordonii* and *S. sanguinis* is critical for understanding the different strategies of these species for colonizing and co-existing within dental plaque. *S. sanguinis* and *S. gordonii* are important early colonizers and potentially have a strong influence on the subsequent accumulation of dental plaque Insights gained from these intermicrobial interactions might contribute to the development of new interventions and potential medical treatment for maintaining oral health in the future (Nobbs, Zhang et al. 2007).

To obtain a better understanding of these two closely related human colonizers, the present study sequenced 19 whole-genomes of *S. sanguinis* and *S. gordonii*; 13 were isolated from the United Kingdom, 4 from United States, 1 from Denmark and 1 from Australia. Six strains were isolated from dental plaque or the oral cavity; 10 strains were sub-acute bacterial endocarditis isolates and three were of unknown origin. Comparative genome analyses were then performed using different bioinformatics approaches to investigate the phylogeny, virulence, biology and genomics of *S. sanguinis* and *S. gordonii.* To classify the 19 *Steptococcus* strains, phylogenetic analyses, using both whole genome data and single

gene markers, were performed to verify the taxonomic position of the 19 strains. After the genome identification and phylogeny inference studies, I performed orthologous gene comparison and pan-genome analyses in order to identify the core genomes shared between the two closely related species. Functional enrichment study was conducted to examine the enrichment of the specific categories of genes of interest in different biological processes. Since regulation of virulence factor expression plays a major role for pathogenic species in adapting to their dynamic host environments, a comparative virulence gene analysis study was performed to identify potential virulence genes in *S. gordonii* and *S. sanguinis*. As the pan-genome analysis study suggested the plasticity of the *Streptococcus* genomes, I aimed to investigate evidence that lateral gene transfer has occurred within genomes of *S. sanguinis* and *S. gordonii* and to assess the impact of this on the acquisition of species-specific genes for possible virulence enhancement and potential adaptive evolution.

To support the expanding research into the Mitis group of oral streptococci, I have also developed StreptoBase, which provides a resource and analysis platform for the research community. Through this platform and the provided in-house designed analysis tools, I hope to provide insights into the biology, phylogeny, genetic variation and virulence of particular strains or species of interest for research worldwide. In addition to the 77 public available genomes downloaded from the National Center for Biotechnology Information (NCBI) resources, I have included my 27 newly sequenced, assembled and annotated genomes of novel strains from six different species of Mitis group into the StreptoBase. These new genomes include 5 genome sequences of the recently classified species *S. oligofermentans* and *S. tigurinus*. The ultimate objective of StreptoBase is to provide a user-friendly database resource and analysis platform particularly for comparative analyses. Users can search, browse, visualize, download and analyze the Mitis group genomes, and

conduct comparative whole-genome analysis on the fly using the in-house developed bioinformatics tools. StreptoBase is designed to support the expanding *Streptococcus* genus research community.

1.2 Project Objectives

The overarching aim of this project was to employ genome sequencing and genomic analyses for the identification of genotypic differences between *S. gordonii* and *S. sanguinis* that will help to understand their subtly different ecological niches and their pathogenic potentials. In addition, I aimed to developed tools to facilitate research into Mitis group streptococci more broadly. The objectives of this study were:

1. To sequence, assemble and annotate the functional elements in the genomes of *S. gordonii* and *S. sanguinis*
2. To study the phylogenetic relationships between all sequenced clinical isolates
3. To perform comparative analyses between the closely related *S. gordonii* and *S. sanguinis*
4. To design and develop a new genomic resource and comparative analysis platform for oral streptococci

CHAPTER 2: LITERATURE REVIEW

2.1 Oral streptococci: Mitis group

Streptococcus is a major genus of spherical Gram-positive bacteria which belong to the phylum *Firmicutes*. Streptococci are classified as alpha-hemolytic, beta-hemolytic or gamma-hemolytic according to their appearance on blood agar. Alpha-hemolysis involves the bleaching of heme iron by streptococcal hydrogen peroxide (H_2O_2), resulting in a greenish tinge on blood agar (Barnard and Stinson 1996). Alpha-hemolytic streptococci used to be known as the 'Viridans group' for the greenish color produced by hemolysis. However, alpha-hemolysis is not entirely consistent between different strains of individual streptococcal species, and therefore the term 'Viridans' is somewhat misleading and is no longer used. These organisms are now more commonly known as the oral streptococci. Overall, the streptococci are divided into six groups, namely the Mitis, Anginosus, Salivarius, Mutans, Bovis and Pyogenic groups, using sequence analysis of the 16S rRNA gene or of a group of housekeeping genes (Bentley, Leigh et al. 1991, Kawamura, Hou et al. 1995, Jakubovics, Yassin et al. 2014). In 2002, Facklam proposed a phenotypic identification scheme which included an additional new cluster called the Sanguinis group which includes *S. gordonii* and *S. sanguinis* (Facklam 2002). More recently, Teng and his colleagues proposed another new cluster named Sinesis which encompasses three species, namely *S. sinensis*, *S. oligofermentans* and *S. cristatus* (Teng, Huang et al. 2014). In the present study, both the Sanguinis group and Sinesis groups are categorized under the Mitis group of oral streptococci.

The Mitis group is comprised of 13 known species including *S. australis, S. cristatus* (formerly *S. crista*), *S. gordonii, S. infantis, S. mitis, S. oligofermentans, S. oralis, S. parasanguinis* (formerly *S. parasanguis*), *S. peroris, S. pneumoniae, S. pseudopneumoniae,*

S. sanguinis (formerly *S. sanguis*), and the latest grouped species, *S. tigurinus.* Currently, the complete genome sequences of 7 species of this Mitis group (*S. pneumoniae, S. pseudopneumoniae, S. mitis, S. oralis, S. gordonii, S. sanguinis* and *S. parasanguinis*) are stored on the National Center for Biotechnology Information (NCBI)'s FTP site. Recent work has shown that Mitis group streptococci play a major role in exacerbating influenza infection particularly among immunocompromised individuals; *S. oralis* and *S. mitis* were found to produce neuraminidase (NA), a vital target of anti-influenza drugs. The NA activity exhibited by these oral bacteria stimulates the release of influenza virus, boosts viral M1 protein expression levels and activates the ERK cell signaling pathway, potentially enhancing viral infections (Kamio, Imai et al. 2015).

2.2 Streptococcal Infective endocarditis (IE)

The oral streptococci are common commensals which are usually found at sites within the oral cavity, such as buccal epithelium, palate, tongue, teeth, epithelial linings of gingival crevices and periodontal pockets. Excessive buildup of plaque biofilms can trigger inflammatory conditions such as gingivitis or chronic periodontitis, which in turn leads to proliferation of the periodontal vasculature, permitting entry of these oral streptococci into the bloodstream. Occasionally, this may result in infective endocarditis (IE) which is a potentially lethal endovascular disease caused by a bacterial infection on heart valves (Wilson, Taubert et al. 2007, Parahitiyawa, Jin et al. 2009). Platelets and fibrin are deposited on exposed extracellular matrix proteins induced by trauma of the damaged endocardium (Ruggeri 2009). Subsequently, a sterile platelet-fibrin vegetation is formed on the endocardium where bacteria from the bloodstream enter, adhere and colonize during bacteremia (Durack and Beeson 1972). These bacteria then recruit platelets from the blood circulation, stimulating platelet activation and platelet aggregation. Large septic thrombi

occur on the heart valves, and eventually disrupt cardiac hemodynamic patterns. The extreme pressure acts on the compromised valves and typically causes congestive heart failure (Yvorchuk and Chan 1994).. The detailed mechanisms of IE are described in Figure 2.1.

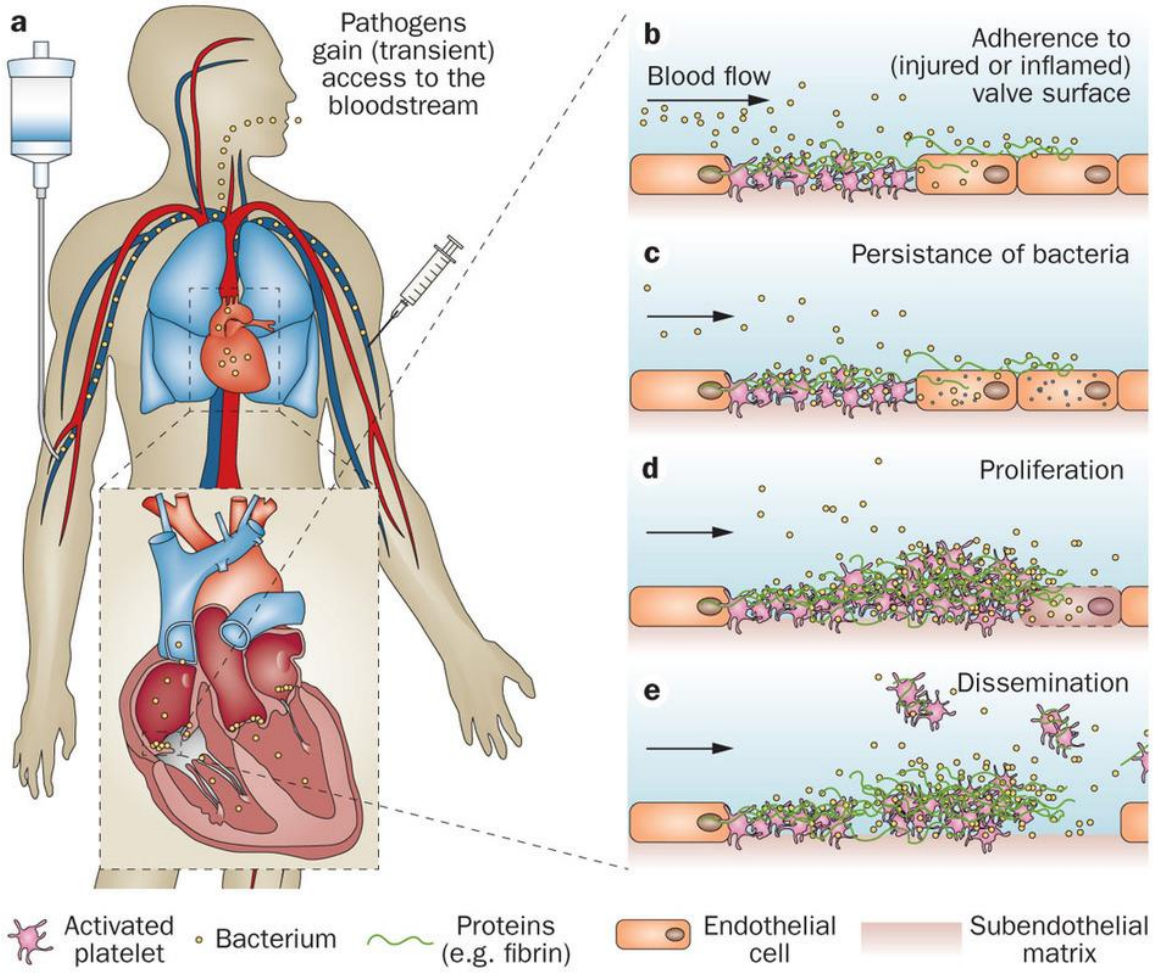
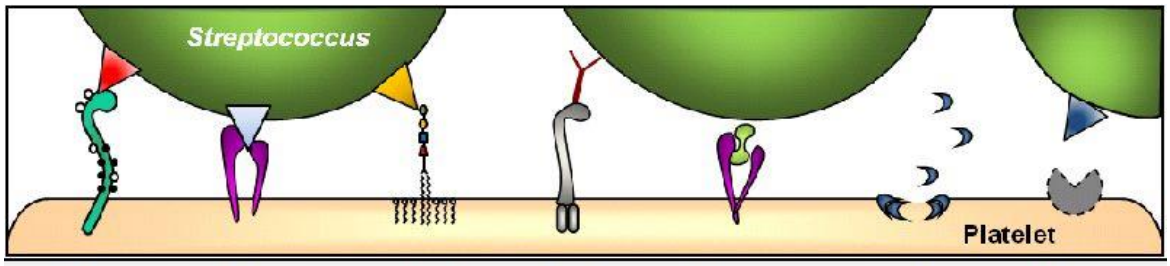


Figure 2.1: The mechanism of streptococcal IE. (a) Pathogens gain access to the bloodstream via health-care procedures or intravenous drug use. (b) Pathogens adhere via platelet fibrin deposition to an inflammed or damaged valves surfaces. (c) Pathogens access to the valve endothelium, causing inflammation and aggressive tissue destruction. (d) Proliferation of the pathogens on the valve endothelium leads to formation of vegetations. (e) Embolization of vegetation particles and dissemination of pathogens causes downstream clinical implications (Source: (Werdan, Dietz et al. 2014).

Among the multiple species of bacteria which have been isolated from patients with IE, streptococci are the second most common (Durack and Beeson 1972, Schierholz, Beuth et

al. 1999). In fact, previous studies have identified streptococci as a major cause of IE in the normal population, although the high incidence of staphylococcal IE on prosthetic valves and in intravenous drug users means that the occurrence of streptococcal IE is often overshadowed (Fowler, Miro et al. 2005, Tleyjeh, Steckelberg et al. 2005, Vogkou, Vlachogiannis et al. 2016). Within the oral Mitis group streptococci, *S. gordonii* and *S. sanguinis* are prominent agents of biofilms on tooth surfaces called dental plaque (Kreth, Merritt et al. 2009). These two pioneer colonizing *Streptococcus* species contribute to dental plaque formation by attaching to tooth surfaces via specific cell surface adhesions such as PAAP, Hsa, Srp and *S. sanguinis*p (Figure 2.2). In human streptococcal infections, these oral pathogens enter the bloodstream from the dental plaque biofilm by surgery or routine oral hygiene. This results in bacteraemia which can lead to infection of the endocardial surfaces of heart valves and streptococcal infective endocarditis (IE) (Hall‐Stoodley, Stoodley et al. 2012). Recent studies have reported that formation of the sterile platelet fibrin thrombus is the key approach of *S. gordonii* and *S. sanguinis* in the development of IE with the predisposition of the heart valve endothelium defects (Turner 2008, Keane 2010). .



GPIbα	αIIbβ3 (direct)	Ganglioside GD3	IgG - FcγRIIA	Fibrinogen - αIIbβ3 (indirect)	?	Unknown receptors
S. gordonii (Hsa/GspB) *S. sanguinis* (SrpA)	*S. gordonii* (PadA)	*S. mitis* (PblA, PblB)	*S. sanguinis* *S. mitis* *S. mutans* *S. pyogenes* *S. agalactiae* *S. oralis*	*S. agalactiae* (FbsA) *S. mitis* (Lysin$_{SMI}$) *S. pyogenes* (M protein)	*S. mitis* (sm-hPAF/ lectinolysin)	*S. sanguinis* (PAAP) *S. gordonii* (SspA/SspB) *S. mutans* (PAc,RGPs)

Figure 2.2: A series of species-specific of Streptococcal cell surface adhesions targeted by *Streptococcus* bacteria during dental plaque formation (Source: (Tilley and Kerrigan 2013).

2.3 Next-Generation Sequencing (NGS) technologies

In the early 1990s, Sanger sequencing was the widely employed method of sequencing which was performed either through shotgun de novo sequencing (cloning of fragmented DNA into a high-copy-number plasmid) or targeted resequencing (PCR amplification involving primers-flanked target) (Shendure and Ji 2008). At present, there are several next-generation sequencing technologies which utilize real time, cyclic-array sequencing such as 454 sequencing (454 Genome Sequencers and Roche Applied Science), Solexa technology (Illumina platforms), the SOLiD platform (Applied Biosystems (ABI)), the Polonator (Dover) and the HeliScope Single Molecule Sequencer technology (Helicos BioSciences) (Shendure and Ji 2008). Next-generation DNA sequencing allows comprehensive analysis of genomes and transcriptomes which ultimately accelerates the progress of biological and biomedical research (Fullwood, Wei et al. 2009). The recent NGS benchmark analysis studies demonstrated the differences of the 5 most

commercialized NGS technologies as summarized in Table 2.1 (Dunne Jr, Westblade et al. 2012, van Dijk, Auger et al. 2014):

Table 2.1: Summary of benchmarking analysis on different type of NGS technologies.

Type of NGS technologies	Chemistry	Pros	Cons
454	• Pyrosequencing	• Suitable for de novo genome assemblies or metagenomics applications facilitated by the efficient mapping of long reads to a reference genome • Faster run times (approximately 23 hours)	• Low throughput • High reagent cost. • High error rates in homopolymer repeats (Metzker 2010)
Illumina	• Dye termination or synthesis	• High compatibility with most of the library preparation protocols • Highest throughput • Lowest per-base cost (read lengths up to 300 base pairs) (Liu, Li et al. 2012)	• Risk of overloading (resulting in overlapping clusters and poor sequence quality) • Requirement for sequence complexity
SOLiD	• Ligation	• High throughput • Low error rates	• Shortest reads (not more than 75 nucleotide read length)

		• High accuracy (two times reading of each base) (Liu, Li et al. 2012)	• Long run times • Less-well-suited for de novo genome assembly • Less-well-developed panel of sample preparation kits and services
Ion Torrent	• Semi-conductor	• Fast run times • Wide range of applications	• High error rates in homopolymer repeats
PacBio	• Direct detection	• Enable optimal performance in genome assembly with its extreme long reads • Fast run times	• Expensive (US$2–17 per Mb) • High overall error rates (approximately 14%) • Lowest throughput • Limited range of applications

Bacterial NGS have been proved to provide better understanding in the bacterial evolution and diagnostic of bacterial infections (Hasman, Saputra et al. 2013). In a recent Tuberculosis (TB) study conducted in the UK, NGS has been demonstrated in successfully investigating the microevolution within *Mycobacterium tuberculosis* genomes by determining the genetic diversity of the bacterial strains (Walker, Ip et al. 2013). Besides, previous whole-genome sequencing study using NGS technology has shown the dynamic

populations of *Staphylococcus aureus* nasal carriage which carry genetic variation that contributes to its bacterial evolution over time (Young, Golubchik et al. 2012). More broadly, the use of NGS enabled the early inspection of the enterohemorrhagic *Escherichia coli* O104:H4 outbreak in Germany which enabled the characterization of the whole bacterial genome (Mellmann, Harmsen et al. 2011).

2.4 Phylogenetic analyses

Bacterial phylogenetic analyses involve the alignment of marker gene sequences or the identification of whole genome core-SNPs to reconstruct the evolutionary relationships between bacteria and to identify their taxonomic position. Stackebrandt and Goebel proposed that 16S rRNA gene-based phylogenetic analysis is an appropriate tool for classification when the sequence homologies of prokaryotic species are below 97% (Stackebrandt and Goebel 1994). When the homology values are more than 97%, DNA-DNA hybridization can be conducted, in order to verify the taxonomic position of the prokaryotic strains. Kawamura and colleagues performed phylogenetic analysis on 34 *Streptococcus* species using 16S rRNA sequences of the type strain of oral streptococci (Kawamura, Hou et al. 1995). They extracted all the 16S rRNA sequences of each oral *Streptococcus* member from the GenBank and the European Molecular Biology Laboratory (EMBL) databases except for two type strains, *S. mitis* NCTC 12261 and *S. gordonii* NCTC 7865. The 16S rRNA of these two type strains were determined from position 8 to position 1392 (*E. coli* numbering) of their sequenced genomes. Next, they utilized the ODEN program set of the DNA Data Bank of Japan to align the 16S rRNA sequences. Using neighbour-joining (NJ) method to construct the phylogenetic tree, Kawamura and

colleagues identified six major clusters of the *Streptococcus* genus: pyogenic group, anginosus group, Mitis group, salivarius group, bovis group and mutans group. Additionally, they successfully classified the species of oral streptococci Mitis group into the same cluster. This cluster of Mitis group species included *S. mitis, S. gordonii, S. pneumoniae, S. sanguinis, S. oralis* and *S. parasanguinis* (Figure 2.3). Kawamura et al. indicated that *S. mitis, S. oralis* and *S. pneumoniae* were closely related, with 16S rRNA gene sequence homology of more than 99%. Since the members of the Mitis group achieved less than 60% of DNA similarity, the authors concluded the distinct taxa of each of these *Streptococcus* Mitis group species. Nevertheless, they found that *S. suis* and *S. acidominimus* were unrelated to any of these groups. Lastly, they excluded *S. pleomorphus* from the genus of *Streptococcus* based on both sequence homology data (less than 85% sequence homology with the other *Streptococcus* species) and neighbour-joining data. (Teng, Huang et al. 2014)

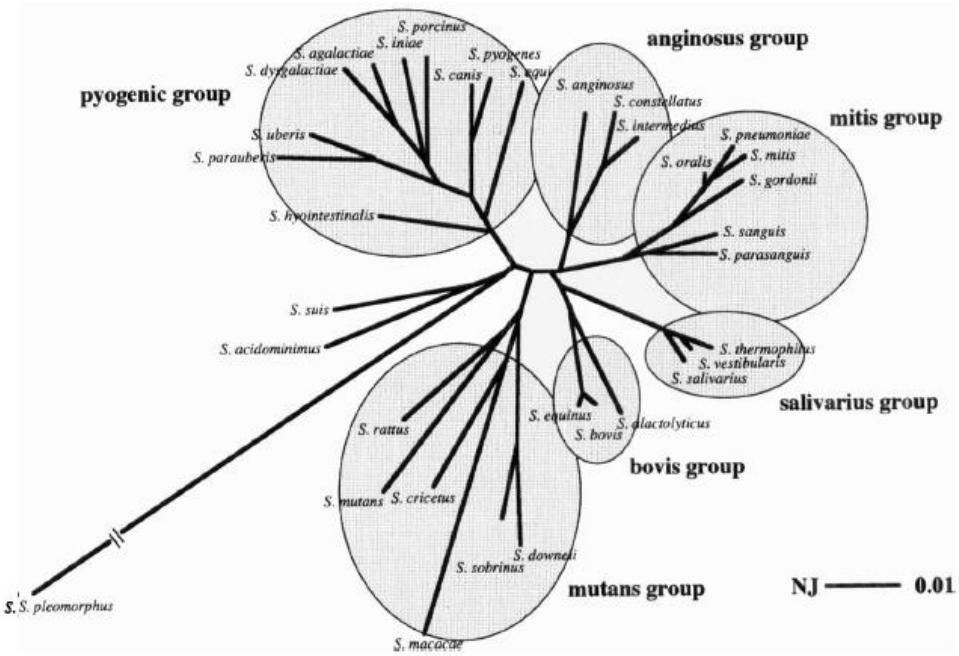


Figure 2.3: The phylogenetic tree constructed based on 34 *Streptococcus* species using 16S rRNA sequences of the type strain of oral streptococci. (Source: (Kawamura, Hou et al. 1995)

In 1999, Kawamura and colleagues generated a *Streptococcus* Mitis group species neighbour-joining (NJ) phylogenetic tree using the partial sequences of *sodA* gene from 96 strains (Figure 2.4). The resulting eight clusters formed were corresponded to the DNA–DNA hybridization results. Noticeably, *S. pneumoniae* strains formed a distinct sub-cluster within the *S. mitis* cluster on the *sodA*-based phylogenetic tree. Judging from the species-specific base differences, *S. pneumoniae* can be clearly distingushed from other species including *S. mitis* (Kawamura, Whiley et al. 1999).

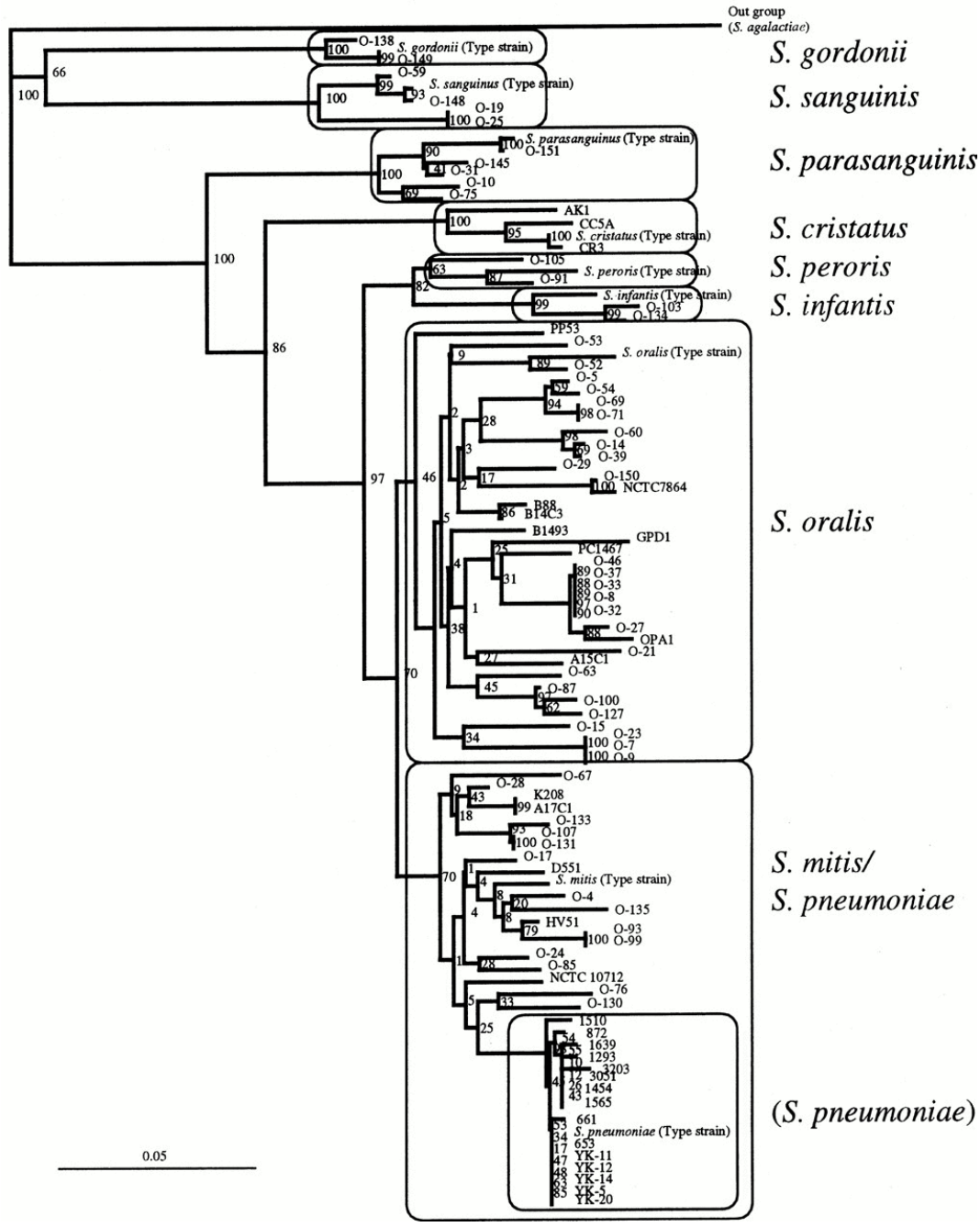


Figure 2.4: The *sodA*-based phylogenetic tree of *Streptococcus* Mitis group species with the 100 bootstrappings (Kawamura, Whiley et al. 1999).

In the most recent Mitis group oral streptococci phylogenetic study, Teng and co-workers investigated the phylogenetic relationship among 87 *Streptococcus* genomes using two

single gene loci, namely 16S rRNA and *groEL*. In the 16S rRNA gene phylogenetic tree, the group observed that *S. sinensis* clearly distinguished from the Anginosus, Mitis, and Sanguinis group members but *S. sinensis* closely clustered with *S. oligofermentans* and *S. cristatus.* However, *S. sinensis* was clustered with the Anginosus and Sanguinis group members in the *groEL* gene-based phylogenetic tree. Further phylogenetic analysis using 50 ribosomal protein genes and hierarchical cluster analysis of MALDI-TOF MS spectra using *S. sinensis* HKU4T and 28 nonduplicated *Streptococcus* species (Figure 2.5) indicated that *S. sinensis*, *S. oligofermentans* and *S. cristatus* formed a distinct clade known as the Sinesis group (Teng, Huang et al. 2014).

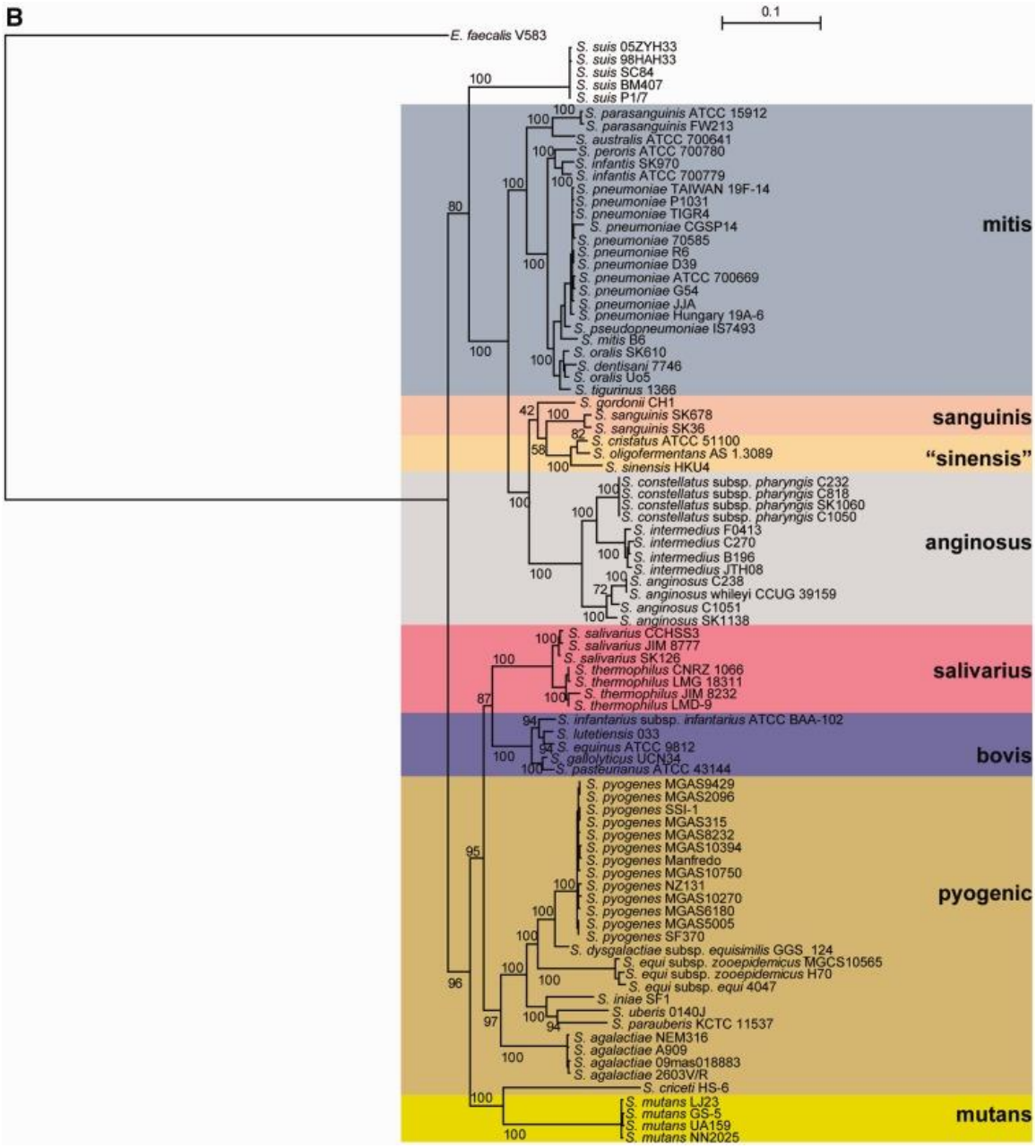


Figure 2.5: The 50 ribosomal protein genes-based phylogenetic tree of *S. sinensis* HKU4T and the other 28 *Streptococcus* species. The phylogenetic tree was generated via maximum-likelihood method with bootstrap values of 1,000 replicates. The scale bar indicated the mean number of nucleotide substitutions per site on each branch (Teng, Huang et al. 2014).

2.5 Pan-genome analyses

Pan-genome analyses describe a complete gene set of all strains of a species, namely the core genome (genes present in all strains), the accessory genome which is comprised of both the dispensable genome (genes present in two or more strains) and the unique genome (genes specific to single strains) (Tettelin, Masignani et al. 2005). A bacterial pan-genome can be either closed or open. An open pan-genome is indicated by an infinite size with new genes continually being taken up by the bacteria. By contrast, a closed pan-genome harbors a definite size, with a constant pan-genome size that represents the whole bacterial species (Tettelin, Masignani et al. 2005).

The first ever pan-genome analysis was performed on *Streptococcus agalactiae*, a pathogenic species, which occasionally infects the elderly and and is a major cause of lethal infections in newborn infants (Remm, Storm et al. 2001). In the study conducted by Tettelin and colleagues, eight genomes of *S. agalactiae* were selected and their predicted genes clustered using the Jaccard algorithm (Tettelin, Masignani et al. 2005). The thresholds were set as 80% identity and Jaccard coefficient above 0.6 in order to classify the core and accessory genes. For the prediction of pan-genome size of *S. agalactiae*, the identified core genes were extrapolated by fitting the decaying function $Fc = Kc\exp[-n/Tc] + \Omega$ and $Fc = Ks \exp[-n/Ts] + tg(\Theta)$, where n is the number of strains and Kc, Ks, Tc, Ts, Ω and $tg(\Theta)$ are free parameters and the number of core genes were plotted (Figure 2.6).

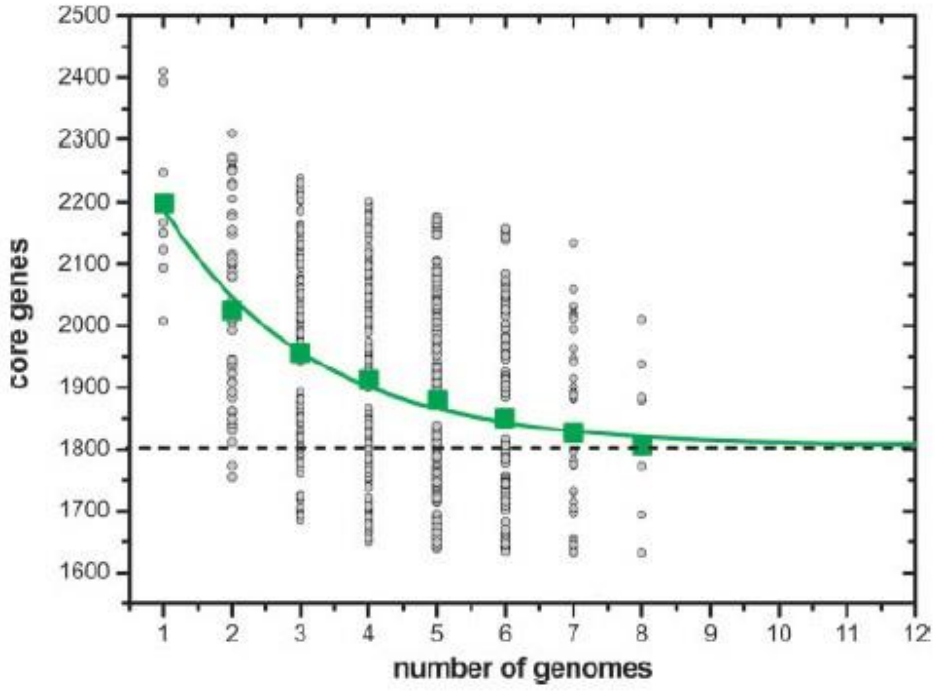


Figure 2.6: *S. agalactiae* core genome plot. The number of core genes was plotted as a function of the number of genome (n) sequentially added. Using the formula 8!/[(n-1)!.(8-n)!], each circle indicates the number of core genes by different strains combination. The extrapolated core genes are shown as a dashed line (Source: (Tettelin, Masignani et al. 2005)).

Tettelin and co-workers observed that the number of core genes was reducing with the addition of newly sequenced genomes from the pan-genome analysis. Based on the curve (Figure 2.7), the decline of core genome reached a plateau around 1,806 genes (95% confidence interval = 1,750-1,841) and remained constant as the number of sequenced genomes grows. Meanwhile, the authors estimated the number of new genes added by a new genome, applying the decaying exponential function as mentioned previously. An average number of 161 new genes were recorded with the addition of the second genome of *S. agalactiae* while this number dropped to 54 genes when the fifth genome of *S. agalactiae* was included. Using a threshold of 95% confidence interval, Tettelin deduced that average of 33 new genes will be recruited to the pan-genome of *S. agalactiae* for every

newly sequenced *S. agalactiae* genome, further indicating the open pan-genome characteristic of *S. agalactiae.*

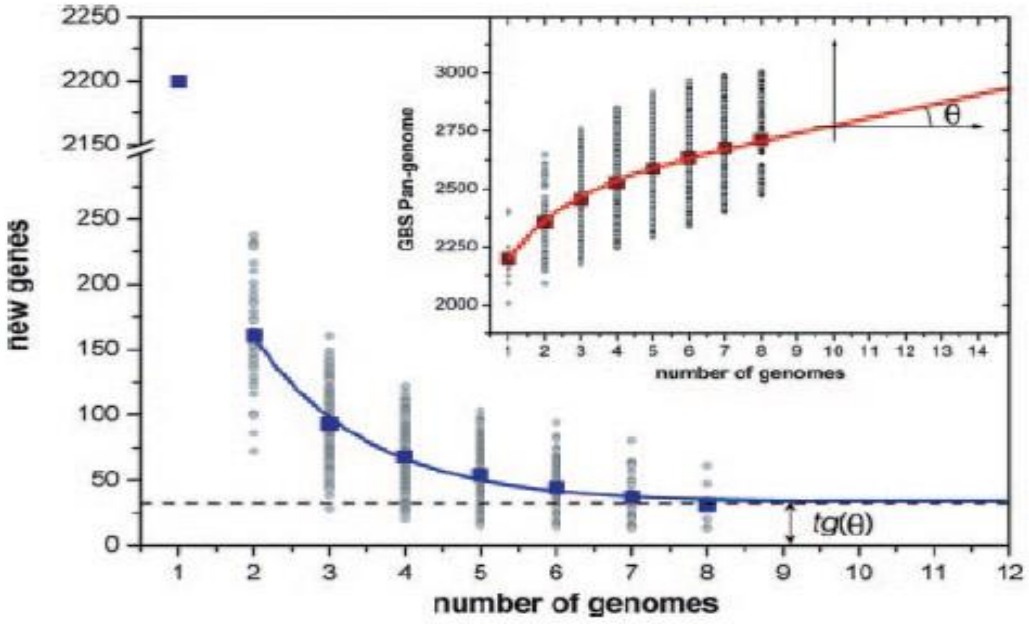


Figure 2.7: Pan-genome plot of *S. agalactiae.* The blue curve showed the least square fit of the decaying function $Fc = Kc\exp[-n/Tc] + \Omega$ and $Fc = Ks \exp[-n/Ts] + tg(\Theta)$ with its extrapolated average number of new genes shown in dashed line. The red curve indicated the pan-genome size of *S. agalactiae* with new added different number of genomes (Source:(Tettelin, Masignani et al. 2005).

In contrast, Donati's group performed a pan-genome analysis of 44 strains of *S. pneumoniae* using two different approaches: the finite supragenome model and the power law regression model (Donati, Hiller et al. 2010) The finite supragenome model enables prediction of the number of genes within a particular fraction of the *S. pneumoniae* genomes, varying from rare genes which are present in less than 3% of the genomes to core genes. On the contrary, the power law regression model which is similar to the method used by Tettelin et al. allows the extrapolation to an infinite number of genomes, predicting genes found in *S. pneumoniae* is finite (closed pan-genome) or unlimited (open pan-

genome). Both approaches showed consistent pan-genome results when the number of genomes is less than 40 strains. Based on the finite supragenome model, Donati et al. observed a significant drop in number of new genes identified for each genome at 100 strains and stabilization in the number of core genes at 1,647 (Figure 2.8b). Besides, the supragenome model identified 48% *of* core genes and approximately 27% are rare genes. They estimated *S. pneumoniae* posseses an open pan-genome with 44 strains including 92.7% of the pneumococcal pan-genome. On the contrary, as the number of genomes grows above 40 strains, the power law regression model portrayed an average number of new genes as a function of the number of genomes, well-fitted with exponent $\xi = -1.0 \pm 0.15$ (Figure 2.8). Hence, Donati et al. suggested open pan-genome of *S. pneumoniae* as the pneumococcal pan–genome size increases logarithmically ($\xi > -1$). In short, this study conducted by Donati et al. proved that *S. pneumoniae* harbors an open pan-genome using both finite supragenome and power law regression models.

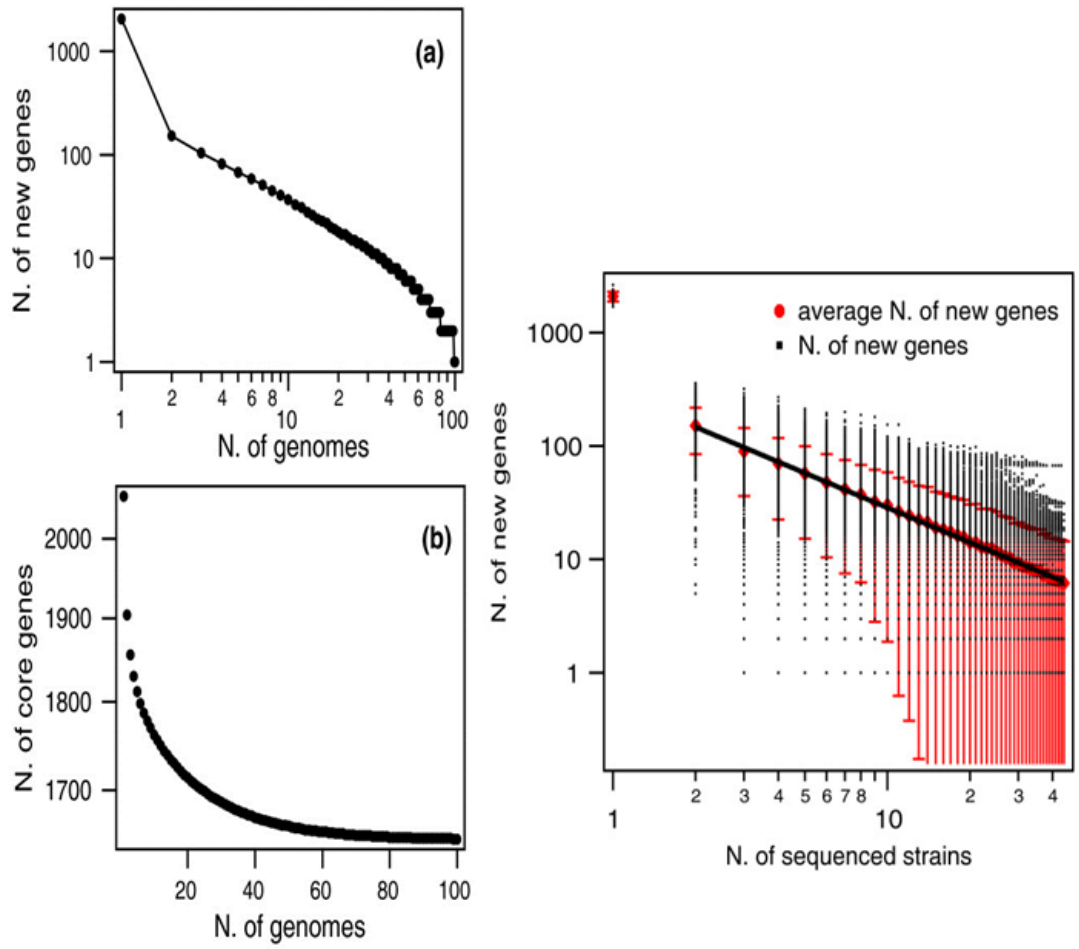


Figure 2.8: Pan-genome of *S. pneumoniae*. The number of specific genes is plotted as a function of sequentially added number of strains (n), fitted with a decaying power law y = A/nB. Each n refers to the values obtained for the different strain combinations while red symbols indicate the average of these values, and error bars represent standard deviations.

2.6 Horizontal Gene Transfer (HGT) analysis

Horizontal gene transfer (HGT) is a lateral transmission of genetic material between organisms, possibly of different species. The concept of Genomic Island (GI) originated from Pathogenicity Islands (PAIs) and was first described by Hacker and his colleagues when they discovered a genomic region of virulence gene clusters in *E. coli* (Hacker and Kaper 2000). Later, they revealed antibiotic resistance islands which carry cluster of genes encoding adaptive resistance and metabolic islands which harbor gene groups that encode adaptive metabolic properties. GIs often carry large sets of genes (mostly strain-specific

genes) (Hacker and Kaper 2000). Furthermore, GIs can confer significant genetic differences particularly between closely related species and their analysis is able to reveal ecologically relevant features of genomes (Coleman, Sullivan et al. 2006, Cuadros-Orellana, Martin-Cuadrado et al. 2007). On the contrary, prophages are viral sequences integrated into bacterial genomes by bacteriophage that infect and reproduce within their bacterial hosts. Likewise, prophages have significant impacts on bacterial evolution and ecology by influencing the pathogenicity and virulence of their bacterial hosts (Mitchell 2014). Furthermore, prophages have had some success in coping with bacterial infections and antibiotic resistance (Keen 2012, Rea, Alemayehu et al. 2013).

2.7 Genomic Island (GI) analysis

A Genomic Island (GI) is a cluster of genes acquired by horizontal transfer. In general, there are two different types of GIs: 1) 'ecological islands' which usually found in environmental bacteria and 2) 'saprophytic islands', 'symbiosis islands' or 'pathogenicity islands' (PAIs) which are frequently detected in bacteria associated with a host (Hacker and Carniel 2001). GIs potentially change bacterial traits in antibiotic resistance, virulence, symbiosis and fitness as well as microbial adaptation, resulting in a great impact on bacterial evolution (Langille, Hsiao et al. 2010). In a comprehensive analysis of GIs done by Gómez and his colleagues, 70 marine bacterial genomes were studied to explore the distribution, patterns and functional gene content of GIs found in their genomes (Fernández-Gómez, Fernàndez-Guerra et al. 2012). The GIs of 53 genomes were extracted from IslandViewer database (Langille and Brinkman 2009). IslandViewer is a web-based interface that integrates three methods for GI identification and visualization: IslandPick, IslandPath-DIMOB and SIGI-HMM. IslandPick (Langille, Hsiao et al. 2008) compares phylogenetically related genomes; SIGI-HMM (Langille and Brinkman 2009) estimates

codon usage and IslandPath (Hsiao, Wan et al. 2003) measures abnormal sequence composition or genes related to mobile elements withinpredicted GIs. The remaining 17 genomes which consisted of 8 Alphaproteobacteria and 9 marine Bacteroidetes, were downloaded from the NCBI website and the J. Craig Venter Institute (http://www.jcvi.org/). These 17 genomes were then uploaded to IslandViewer database in order to predict GIs in their genomes. IslandPick GIs prediction involved selection of minimum of three closely related genomes along with a distant genome. IslandPath-DIMOB and SIGI-HMM were then used to detect overlapped GIs predicted by at least two of the predictors. Additionally, Gómez et al. performed manual refining of these putative GIs using Artemis and Integrated Microbial Genomes (IMG) genome browser (Markowitz, Chen et al. 2012) to insert tRNA found within or flanking the GIs. They successfully detected 438 GIs which carry a total of 8152 genes. They indicated the overall GI number per genome was strongly and positively correlated with the total GI size. An average GI genome length of 3% to as high as 12% was detected in half of the marine bacterial genomes analysed.

2.8 Prophage analysis

A prophage is an integrated bacteriophage genome found in bacteria. It can either be incorporated into the circular bacterial DNA chromosome or exist as an extrachromosomal plasmid. A recent study has reported the significance of prophages in altering the lifestyle, fitness, virulence, and evolution of their bacterial host (Fortier and Sekulovic 2013). In *S. pneumoniae*, temperate prophages have been linked to affect the dynamics of oral biofilm establishment with a localized release of eDNA during spontaneous phage induction (Carrolo, Frias et al. 2010).In 2014, Scott Mitchell carried out a benchmark study by comparing different tools and methods for detecting prophages such as the PHAST,

Prophage Finder and Basic Local Alignment Search Tool (BLAST) search method (Mitchell 2014). Based on the prophage prediction results of 41 sequenced *S. aureus* genomes, he concluded that the PHAST program had significant advantages (Zhou, Liang et al. 2011) over BLAST and Prophage Finder (Bose and Barber 2006). PHAST which analyzes a variety of genome elements such as unusual genes, attachment site recognition and tRNA, was found to include more prophage sequences than traditional BLAST search. In addition, the latest released PHAST program has been shown to have greater sensitivity and higher accuracy in attachment site prediction while more defective prophages (false-positive) were detected by Prophage Finder (Mitchell 2014). PHAST applies Gene Locater and Interpolated Markov ModelER (GLIMMER) gene prediction to detect prophages along with the position, length, number and boundaries of genes. Remarkably, PHAST concatenates a variety of different prediction information including open reading frame prediction through GLIMMER, proteins and phage sequences identification via BLAST, tRNA analysis using tRNAscan-SE, attachment site recognition via ARAGORN as well as gene clustering density readings through Density-Based Spatial Clustering of Applications with Noise (Zhou, Liang et al. 2011). Identified prophages are classified into intact, questionable or incomplete prophage based on their 'completeness score'. An intact prophage has a score above 90; questionable prophage achieved scores between 60 to 90 and incomplete prophage has less than 60 score. In this study, Scott has revealed six virulent prophages in *S. aureus* subspecies TCH60: one intact prophage, two questionable prophages and three incomplete prophages.

2.9 Summary

Based on the literatures reviewed, I selected the methods and bioinformatics softwares/platforms which are able to achieve the optimal results of the Mitis group oral

streptococci comparative analysis. Illumina Hiseq2000 platform was selected for NGS analysis due to its relatively low price but efficient sequencing performance by producing high-throughput parallel sequencing (Minoche, Dohm et al. 2011). Further, Illumina Hiseq is very well-established and it is one of the most widely used sequencing devices in many genome sequencing projects (Minoche, Dohm et al. 2011, van Dijk, Auger et al. 2014). In reference to the recent *Streptococcus* Mitis group species phylogenetic study performed by Teng and colleagues (Teng, Huang et al. 2014), the widely used 16S rRNA was selected as the single gene marker to construct the phylogenetic tree using the Neighbour-Joining method. The pan-genome study conducted by Tettelin and co-workers suggested that both *S. gordonii* and *S. sanguinis* might have a tendency to harbor an open pan-genome correspond to their *Streptococcus* Mitis group species of *S. pneumoniae* (Tettelin, Masignani et al. 2005). For horizontal gene transfer study, the efficiency of the IslandViewer tool and PHAST pipeline have been well-demonstrated in Gómez and Mitchell studies (Fernández-Gómez, Fernàndez-Guerra et al. 2012, Mitchell 2014). In the present study, I aimed to apply these bioinformatics tools for the analysis of two important dental plaque colonisers and occasional invasive pathogens, *S. sanguinis* and *S. gordonii*.

CHAPTER 3: MATERIALS AND METHODS

3.1 Bacterial strains and DNA extraction

77 genome sequences of Mitis group streptococci were downloaded from the public NCBI database. Additionally, 27 novel strains/genomes of Mitis group streptococci generated from the laboratory were included in the sequencing project. All 27 Mitis group streptococci strains were originally isolated from oral sites or infective endocarditis cases at a variety of different geographical locations and were used in this study due to their strain availability. These *Streptococcus* strains were cultured in THYE medium (30 g/L Todd Hewitt broth, 5 g/L yeast extract) for 16 hours at 37°C prior to DNA extraction using standard protocols (Old, Lowes et al. 2006). The bacterial culture and DNA extraction were performed by Lesley A. Old, a laboratory technician of School of Dental Sciences in Newcastle University, United Kingdom. .

3.2 Library preparation and next-generation sequencing

The *Streptococcus* bacterial DNA library preparation involved fragmentation of DNA samples using Covaris S2 for 120 seconds at temperature of 5.5 – 6.0 degree Celsius. The quantity and quality of the fragmented DNA were evaluated by Agilent BioAnalyzer 2100. The sample size was selected using Invitrogen 2% agarose E-gels. For DNA library construction, only the fragments tagged with adapter molecules at both ends underwent 10 cycles of PCR. The constructed genomic library was validated using Agilent BioAnalyzer 2100. The 19 *Streptococcus* genomes were sequenced using Next Generation Sequencing Illumina Hiseq2000 platform. The paired-end sequencing of *Streptococcus* genomes uses a standard read length of 100 base pairs. The *Streptococcus* genomes run on a single lane, employing the TruSeq LT assay. The paired-end sequencing generates two FASTQ output

data files: one containing the forward primer ("AGATCGGAAGAGCACACGTCTGAACTCCAGTCA") derived reads "_R1" and one containing the reverse primer ("AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT") derived reads "_R2". The library preparation and sequencing services were outsourced to Source BioScience Ltd in the United Kingdom.

3.3 Raw read quality checking and preprocessing

The raw read quality was verified through FastQC software (Andrews 2011). The overall sequencing reads showed satisfactory results of per base N content and optimal per base sequence quality with no over-represented sequences. The quality score is directly proportional to the level of base call. Data pre-processing was completed by a trimming approach using CLC Genomic Workbench V6.5 (CLC BIO Inc., Aarhus, Denmark). A series of trimming operations offered by CLC Genomic Workbench V6.5: quality trimming based on quality scores, ambiguity trimming of gaps in scaffold genomes, adapter trimming, base trimming by removing a specified number of bases at either 3' or 5' end of the reads and length trimming within a specified threshold. Quality trimming which applies the Modified-Mott trimming algorithm was selected. All sequencing reads were trimmed based on a Phred score of Q20. The default parameter for quality trimming allows a maximum number of two ambiguities.

3.4 Genome assembly and contamination checking

The preprocessed reads were *de novo* assembled using CLC Workbench 6.5 (CLC BIO Inc., Aarhus, Denmark). In general, the assembly involved two steps: 1) Generation of simple contig sequences using the information within the read sequences, 2) Mapping of reads

based on the previously generated simple contig sequence. The later step serves to show coverage levels along the contigs and facilitate downstream analysis activity such as Single Nucleotide Polymorphism detection (CLC BIO Inc., Aarhus, Denmark). The N50 contig is estimated by summarizing the lengths of the largest contig number until half of the total contig length. Significantly, high N50 values and low contig numbers of genomes indicate a greater assembly performance. After the genome assembly, potential contaminated sequences were filled out by searching against common contaminant databases including VecScreen database (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) and mitochondrial disease database (http://www.mitodb.com/) through sequence continuity and contamination screening.

3.5 Genome annotation

To facilitate comparative analysis across different Mitis group *Streptococcu*s genomes, consistency in genome annotation is important. Therefore, all 104 genome sequences were annotated using the Rapid Annotation using Subsystem Technology (RAST) pipeline, which is a well-established and fully open web-based engine, supporting annotation of both complete and draft genomes (Aziz, Bartels et al. 2008). The RAST pipeline enables genome identification of an array set of distinct genome components including protein-coding genes, ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and gene function prediction. The RAST genome annotation works by mapping a set of genes to their correspond subsystems as well as their metabolic reconstructions. Moreover, it predicts functional proteins assignment according to their relatedness in the subsystems of FIGfams database.

3.6 Multiple sequence alignment (MSA) and phylogenetic inference

To reconstruct a single gene marker 16S rRNA-based phylogeny tree, 16S rRNA gene sequences were predicted from every single *Streptococcus* genome using RNAmmer 1.2 Server (Lagesen, Hallin et al. 2007). All extracted 16S rRNA sequences were aligned using MAFFT web-based program (Katoh, Kuma et al. 2005). The core-genome SNP sequences of each *Streptococcus* genome were determined via Panseq web server (Laing, Buchanan et al. 2010) by searching the SNPs within core genome and identifying the distribution of their accessory genomic regions. The Multiple Sequence Alignment (MSA) of these core genome SNPs was done using ClustalW from the European Bioinformatics Institute (Thomopson, Higgins et al. 1994). Ultimately, the generated MSA results from both the MAFFT and Panseq servers were imported into MEGA6 (Molecular Evolutionary Genetics Analysis 6) software (Tamura, Stecher et al. 2013) in order to construct the phylogenetic trees. Both 16S rRNA-based and core-genome SNP-based phylogenetic trees were constructed using 1000 times bootstrapping replication using the Neighbour-Joining (NJ) algorithm method (Saitou and Nei 1987).

3.7 Orthologous gene family comparisons and pan-genome analysis

The pan-genomes of the *Streptococcus* isolates were analyzed using the pan-genome analysis pipeline (PGAP) which implements functional ortholog clustering (Zhao et al., 2012). Each RAST-predicted protein sequence was labelled with unique strain identifiers that were later concatenated into a single input sequence file. Using the BLASTALL algorithm, the min. score value was set to 50 and E-value<1e-8 (Altschul, Gish et al. 1990). Based on the Markov Cluster Algorithm (MCL) algorithm, the amino acid cutoff was adjusted to 50% sequence identity and 50% sequence coverage in order to group two genes into the same cluster (Enright, Van Dongen et al. 2002). Finally, in-house perl scripts were

used to retrieve the protein sequences of accessory genes and search against all oral *Streptococcus* genome sequences used in this analysis using TBLASTN program. This method is used to identify any genes in the genomes that might be overlooked by the RAST pipeline and any genes identified by this approach were retrieved and included in the genes list of the genome/strain.

3.8 Functional enrichment analysis

To examine whether the unique core genes of both *S. sanguinis* and *S. gordonii* implicated in the biological aspects are functionally relevant and to discover unexpected shared functions between these unique core genes, functional enrichment analysis was performed using Blast2GO software (Conesa and Götz 2008). Firstly, the Blast2GO predicted the function of each gene which involved three steps: BLAST to find homologous protein sequences, MAPPING to retrieve Gene Ontology (GO) terms and ANNOTATION to select reliable functions. BLAST mappings were implemented using the protein sequences of *S. sanguinis* and *S. gordonii* unique core genes against their reference strains of *S. sanguinis* SK36 and *S. gordonii* Challis, respectively. After MAPPING and ANNOTATION, I ran the InterPro Scan prior to functional enrichment process. InterPro scan helps to improve the outcome of the annotations by adding the GO terms retrieved through motifs to the current annotations in a sequence-wise manner. This step is crucial prior to the functional enrichment test which involved comparison of two of GO function between two sets of functional annotations. The functional enrichment was performed by a Fisher's Exact Test along with the False Discovery Rate (FDR) which used for correction in multiple test comparisons. A node fillter value of 0.05 is set as the adjusted FDR p-value to define statistically significant enriched GO terms. Target lists of both *S. sanguinis* and *S. gordonii* unique core genes were selected respectively for specialized functional enrichment

analysis, generating GO graphs from tables of under- and over-enriched *Streptococcus* unique core genes.

3.9 Virulence gene prediction

All virulence genes of Mitis group *Streptococcus* strains were predicted by BLASTing all RAST-predicted protein sequences against the Virulence Factors Database (VFDB) that stores experimentally verified known virulence genes (L. Chen, Xiong, Sun, Yang, & Jin, 2011). In-house Perl scripts were then used to process BLAST outputs (generated by searching these query sequences against VFDB) for each RAST-predicted protein (query sequence) in the Mitis group *Streptococcus* genomes. Based on the sequence homology, a gene is defined as a putative virulence gene if it has at least 50% sequence identity and at least 50% sequence completeness with the BLAST hit from the VFDB database. The filtered BLAST results were consolidated the virulence genes with minimum mapped sequence identity and sequence coverage of 50% in both query and subject were organized in a matrix table. Lastly, in-house R scripts were used for hierarchical clustering and a heat map was generated for visualization of the virulence gene profiles across all studied bacterial strains/genomes. The presence of a virulence gene in a genome was shown in red, whereas the absence of the virulence gene was shown in black in the heat map.

The predicted *rps* gene locus in the sequenced genomes of 19 *S. gordonii* and *S. sanguinis* strains was analyzed manually using in-house scripts. The protein sequences of the first four regulatory genes: *wzg* (gi|157075510|gb|ABV10193.1|:1-486), *wzh* (gi|157075683|gb|ABV10366.1|:1-243), *wzd* (gi|157076133|gb|ABV10816.1|:1-231) and *wze* (gi|157076456|gb|ABV11139.1|:1-231) were extracted from the *S. gordonii* Challis genome stored in the NCBI database, while the predicted protein sequences of the 10

genes: *wchA* (Q83YQ3), *wchF* (Q83YS0), *wefA* (Q83YR9), *wefB* (Q83YQ5), *wefC* (Q83YR8), *wefD* (Q83YR4), *wzy* (Q83YR3), *wzx* (Q83YR2), *glf* (A0A0F2CL65) and *wefE* (Q83YR0) were retrieved from the same species in the Universal Protein Resource (UniProt) resource (www.uniprot.org/). Next, protein BLAST was performed using these *rps* gene sequences against *Streptococcus* protein sequences. The protein BLAST results were then filtered based on the threshold of 50% sequence identity and 50% sequence coverage. To determine whether similar genome arrangements were present in other *S. gordonii* and *S. sanguinis* strains, their genomes were further analyzed for the presence of 'locally collinear blocks' (LCBs) via the Mauve genome analysis tool.

3.10 Comparative prophage analysis

Putative prophages of *S. gordonii* and *S. sanguinis* were identified using the PHAST (Phage Search Tool) web server (Zhou, Liang et al. 2011). The genome contig sequences of the *Streptococcus* species were concatenated to serve as input files to locate, annotate and visualize prophage sequences and prophage features. The identification and completeness of these putative prophages were evaluated through a series of operations including the genome-scale ORF prediction and translation via GLIMMER, protein, phage sequence and tRNA identification, attachment site recognition and gene clustering density measurements as well as sequence annotation text mining. The predicted putative prophages were verified by eliminating the prophages mapped located within two contigs. All putative prophages were then BLAST searched across strains of *S. sanguinis* and *S. gordonii* for genome completeness checking to verify their presence in oral streptococci genomes with cutoff values (at least 70% sequence identity and 70% sequence coverage). An intact prophage was defined by achieving a score of at least 90 by the PHAST software.

To predict the lifestyle of the prophage, I utilized Phage Classification Tool Set (PHACTS) which involved a novel similarity algorithm and Random Forest Classifier (Pal 2005). The amino acids sequences of each phage were uploaded for phage lifestyle annotation using similarity algorithm. Datasets consisting of various sizes of partial proteomes were created. Each proteome was created by randomly selecting a replacement phage with a known lifestyle followed by randomly choosing a set of contiguous proteins in that phage. Lastly, the classification of the lifestyle of a phage (either 'virulent' or 'temperate') was performed by Random Forest classifier (Pal 2005).

3.11 Comparative Genomic Island (GI) analysis

All putative GIs in *S. sanguinis* and *S. gordonii* were predicted using the IslandViewer software tool (Langille and Brinkman 2009), involving three different GI identification approaches: sequence composition-based approaches SIGI-HMM and IslandPath-DIMOB, and the comparative genomics approach IslandPick. The predicted GIs were then further filtered by removing GIs with genomic length less than 10 Kbp. BLASTClust was utilized to perform clustering of all predicted GIs, eliminating duplicated GIs based on threshold of at least 50% sequence identity and 50% sequence coverage. Likewise, the predicted putative GIs were further filtered by omitting any GI mapped across two different contigs. To cluster homologous GIs, all putative GIs from *S. sanguinis* and *S. gordonii* strains were clustered if they have >50% sequence identity and >50% sequence coverage.

3.12 Development of Mitis group oral *Streptococcus* database – StreptoBase

A total number of 104 Mitis group oral *Streptococcus* genomes including 77 genome sequences from the public NCBI database and the genome sequences of 27 isolated bacterial strains from 11 species: *S. australis, S. cristatus, S. gordonii, S. infantis, S. mitis,*

S. oligofermentans, S. oralis, S. parasanguinis, S. peroris, S. sanguinis, and *S. tigurinus* have been incorporated in StreptoBase. To systematically predict subcellular localization of each RAST-predicted gene, I utilized the latest PSORTb subcellular localization tool (version 3.0) program (Nancy, Wagner et al. 2010). PSORTb is an efficient, open-source tool which supports high precision of proteome-scale prediction coverage and refined sub-categories localization. The predicted subcellular localization sites were computationally calculated based on the values of feature variables which infer the sequences characteristics. Each of the generated values was then sorted to their respective candidate site through their estimated relativity. Besides the subcellular localization information, I ran an in-house Perl script to estimate the GC content, hydrophobicity and molecular weight of each protein or gene.

The web interface of StreptoBase was developed using HyperText Markup Language (HTML), HyperText Preprocessor (PHP), JavaScript, jQuery, Cascading Style Sheets (CSS) and AJAX. StreptoBase is supported by Linux, Apache, MySQL and PHP (LAMP) architecture. The Apache web server is equipped with Linux Operating System to manage the comprehensive *Streptococcus* genomic data housed in StreptoBase. The front end PHP framework of CodeIgniter version 2.1.3 was implemented to offer model-view-controller, dividing application data, presentation and background logic and process into three distinct modules. With this advanced feature, all *Streptococcus* related sources codes and biological data are arranged in a clear and organized fashion which facilitate future updating of new *Streptococcus* genomes into the existing database system. For *Streptococcus* biological data storage and management, I utilized MySQL version 14.12 in order to store the extensive *Streptococcus* genome information into a well-designed database schema and tables. The

backend process of StreptoBase is monitored by Perl script, Python script and R script which support the efficiency and functionality of the integrated bioinformatics tools.

CHAPTER 4: RESULTS – COMPARATIVE GENOMIC ANALYSIS OF *STREPTOCOCCUS SANGUINIS* AND *STREPTOCOCCUS GORDONII*

4.1 Sample source and genome assemblies

Total cellular DNA was extracted from fourteen *S. gordonii* strains and five *S. sanguinis* isolates present in the culture collection. Of these, thirteen were originally isolated from the United Kingdom, four from the United States and one each from Denmark and Australia. Six strains were originated from dental plaque or the oral cavity; ten strains were from subacute bacterial endocarditis and the origin of the other three was not known. The genomes of these *Streptococcus* isolates were sequenced using the Illumina HiSeq2000 sequencing technology platform. The generated raw sequencing reads were pre-processed using a trimming approach at a Phred quality score of 20 and *de novo* assembled using CLC Genomic Workbench V6.5 (CLC BIO Inc., Aarhus, Denmark). The assembled genomes had an average genomic size of 2,290,927bp with an average G+C content of 41.2%.

4.2 Genome Overview

To gain better insights into the assembled genomes and to evaluate the coverage of how complete are these *Streptococcus* genomes, all assemblies were mapped onto the complete reference genomes of *S. sanguinis* SK36 and *S. gordonii* Challis using the NUCmer program (Delcher, Phillippy, Carlton, & Salzberg, 2002). The genome coverages and genome identities of 14 *S. gordonii* strains ranged from 88%-95% and 95-98%, respectively. The other 5 *S. sanguinis* strains achieved genome coverages between 84% to 97% and genome identities between 95 to 96%. The RAST annotation pipeline predicted approximately 2,117 to 2,429 functional genes and 2-12 ribosomal RNA genes in both *Streptococcus* species. In general, *S. gordonii* genomes harbored between 38-59 transfer

RNAs with an average GC content of 40.5%, whereas *S. sanguinis* genomes had 40-61 transfer RNAs with a relatively higher average GC content of 43.2% compared to *S. gordonii* (Table 4.1).

Table 4.1: Summary of the genome features of 19 newly sequenced *S. gordonii* and *S. sanguinis* strains. The details include genome size, GC content (%), number of coding sequences (CDS), tRNAs, rRNA, genome identity and genome coverage. There are a total of 14 *S. gordonii* strains (grey) and five *S. sanguinis* strains (yellow). The reference genomes of *S. gordonii* Challis and *S. sanguinis* SK36 are highlighted in red.

Strain	PV40	Blackburn	Channon	FSS2	FSS3
Status of genome	Contigs	Contigs	Contigs	Contigs	Contigs
Genome Size (Mbp)	2.19	2.16	2.23	2.19	2.31
GC Content (%)	40.5	40.5	40.6	40.5	40.2
Number of CDS	2170	2132	2236	2165	2212
Number of tRNAs	46	42	42	46	42
Number of rRNAs	3	3	3	3	5
Genome Identity (%)	98	96	96	98	96
Genome Coverage (%)	95	90	89	92	92
Strain	FSS8	M5	M99	MB666	MW10
Status of genome	Contigs	Contigs	Contigs	Contigs	Contigs
Genome Size (Mbp)	2.15	2.16	2.17	2.31	2.19
GC Content (%)	40.6	40.6	40.5	40.3	40.5
Number of CDS	2132	2117	2128	2314	2158
Number of tRNAs	38	41	43	46	38
Number of rRNAs	2	3	3	3	3
Genome Identity (%)	95	95	95	96	98
Genome Coverage (%)	90	88	89	90	92
Strain	NCTC 7863	FSS4	FSS9	MB451	PJM8
Status of genome	Contigs	Contigs	Contigs	Contigs	Contigs
Genome Size (Mbp)	2.3	2.31	2.43	2.45	2.37
GC Content (%)	43.3	43.2	43.1	42.9	43.2
Number of CDS	2284	2294	2418	2429	2326
Number of tRNAs	40	49	47	47	42
Number of rRNAs	6	3	3	3	5
Genome Identity (%)	95	95	95	96	95
Genome Coverage (%)	84	85	97	94	92

Strain	PK488	SK12	SK120	SK184	***S. gordonii* Challis**	***S. sanguinis* SK36**
Status of genome	Contigs	Contigs	Contigs	Contigs	Complete	Complete
Genome Size (Mbp)	2.2	2.15	2.16	2.26	2.2	2.39
GC Content (%)	40.4	40.6	40.4	40.5	40.5	43.4
Number of CDS	2176	2143	2119	2273	2173	2385
Number of tRNAs	37	47	47	42	59	61
Number of rRNAs	3	3	3	3	12	12
Genome Identity (%)	96	95	96	97	100	100
Genome Coverage (%)	91	89	90	92	100	100

4.3 Phylogenetic inference

To identify the taxonomic position of each sequenced isolate, I reconstructed phylogenetic trees using both single gene and core-genome SNP-based approaches. The single gene approach utilized the 16S rRNA housekeeping gene to construct a phylogenetic tree using *S. parasanguinis* as an outgroup species (Figure 4.1A). 16S rRNA gene sequences have been widely used as gene markers to differentiate species of *Streptococcus* genus, particularly for α-hemolytic streptococci including *S. sanguinis* and *S. gordonii* (Thompson, Emmel et al. 2013). The 16S rRNA-based phylogenetic tree clearly classified the 19 *Streptococcus* strains into two clades: 14 strains of *S. gordonii* (PV40, Blackburn, Channon, FSS2, FSS3, FSS8, M5, M99, MB666, MW10, PK488, SK12, SK120 and SK184) and five strains of *S. sanguinis* (NCTC 7863, FSS4, FSS9, MB451 and PJM8). *S. gordonii* and *S. sanguinis* are closely related and are approximately 97% identical across the 16S rRNA genes.

To further support the classification results, a more robust and reliable phylogenetic tree was constructed using the core-genome SNP data. Encouragingly, the data showed that the classification results from the core-genome SNP-based tree (Figure 4.1B) were consistent

with the classification from the 16S rRNA-based tree. Interestingly, the *S. gordonii* FSS2, MW10 and PV40 are almost identical at the level of 16S rRNA gene sequence and core-genome SNP, even though these strains were isolated from different sources at different times. *S. gordonii* FSS2 and PV40 were from Newcastle upon Tyne, UK, whereas the *S. gordonii* MW10 strain was isolated in Sydney Australia; *S. gordonii* PV40 and MW10 were from the oral cavity, whereas FSS2 originated from a case of bacterial infective endocarditis.

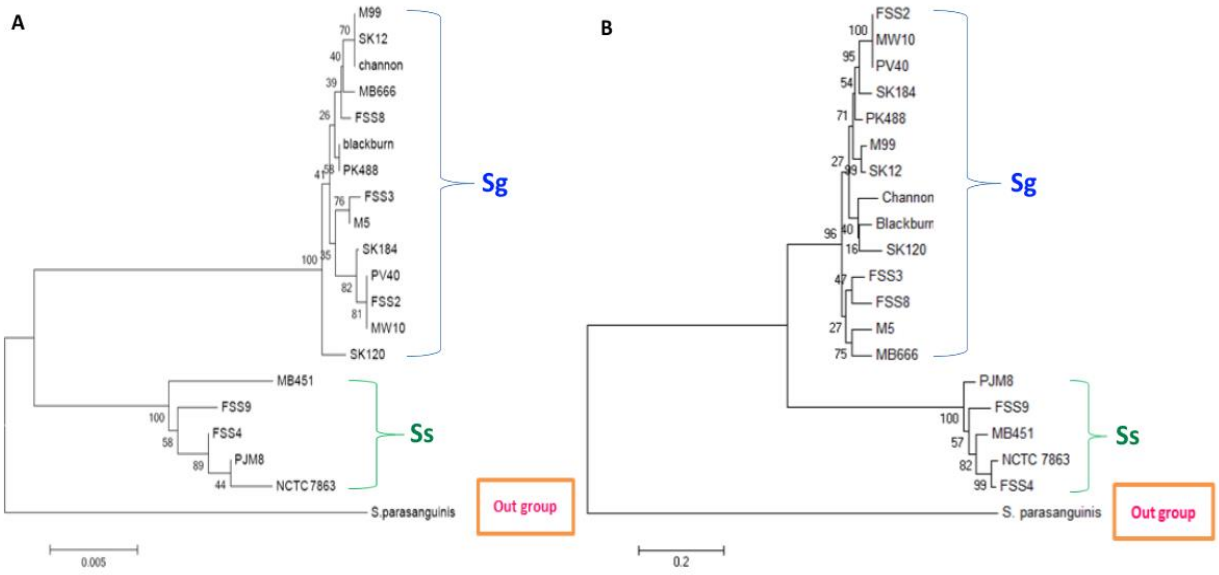


Figure 4.1: Phylogenetic inference of *S. gordonii* and *S. sanguinis*. (A) single gene marker 16S rRNA-based phylogenetic tree (B) a core genome-SNP based phylogenetic tree. Both approaches consistently identified 14 strains as *S. gordonii* and five strains as *S. sanguinis*. S. *parasanguinis* was used as outgroup in the phylogenetic analyses.

4.4 Open pan-genomes of *S. sanguinis* and *S. gordonii*

Gathering all the functional genes of 14 strains of *S. gordonii*, a total number of 4,401 pangenomic gene families of *S. gordonii* were determined. The accessory gene families contributed a larger part of the pan-genome composition (2,774 genes) than the core gene

families (1,627 genes). The accessory gene families were further classified into 1,968 dispensable genes (shared by 2 to 13 strains) and 806 strain-specific genes (shared by only one strain). The core gene families of *S. gordonii* accounted for approximately 37.0% of the total gene families. Due to the low number of *S. sanguinis* isolated strains (5 strains), I included 22 other *S. sanguinis* genomes from the public NCBI database in this analysis in order to have a better representation of this species as a whole. These were all the *S. sanguinis* genomes available at the time of conducting the analysis. Based on the 27 *S. sanguinis* strains, a total of 5,100 pangenomic gene families were identified. The core gene families comprise 1,739 genes (34.1%) and the remainders are accessory gene families. Of the 3,361 accessory gene families, 7% are strain-specific. The pan-genome and core-genome sizes of *S. sanguinis* and *S. gordonii* were estimated by extrapolation of the above genome data. Briefly, the gene clusters and core gene families of *Streptococcus* genomes were calculated, represented by N (N = 1,2,3…..25,26,27). All permutations of genome comparisons for every pan-genome size and core genome of N genomes were analyzed to avoid random bias. Simultaneously, their mean values were predicted and depicted along the core genome family curve and pan-genome family curve. The generated pan-genome curves of both *S. gordonii* and *S. sanguinis* are well-represented by the Heaps law mathematical functions: $Y = 573.705131118841\ X^{0.603} + 1559.42450454357$ and $Y = 816.330402837524\ X^{0.455} + 1410.909236541$, respectively, where Y refers to the pan-genome size while X refers to the number of sequenced *Streptococcus* genomes. According to these equations, the pan-genome size (Y) of both *S. gordonii* and *S. sanguinis* appeared to reach infinity when the number of genomes (X) increased to infinity (Figure 4.2A and Figure 4.2C). Therefore, my data suggest that both *S. gordonii* and *S. sanguinis* have open pan-genomes, which indicates that both species have infinite genomes. The infinite pan-

genome of *S. sanguinis* and *S. gordonii* suggests these bacterial species may keep acquiring new genes as they evolve independently over evolutionary time.

For *S. gordonii,* the rate of new discovery stabilizes at approximately 110 new genes per additional new genome (Figure 4.2B). For example, 295 new genes were detected when a second genome was added to the first *S. gordonii* genome. The mathematical equation predicted 119 new genes gained by the *S. gordonii* species with every new *S. gordonii* genome added. For *S. sanguinis*, I estimated about 61 new genes detected when each additional genome is added (Figure 4.2D). Here, I inferred that *S. gordonii* and *S. sanguinis* have approximately 34 - 37% of core genes of their total gene clusters, probably inclining to an open pan-genome. The intake of new genes may alter the bacterial genome structure and facilitate adaptation of *Streptococcus* species to a dynamic or changing niche (Kurland, Canback et al. 2003).

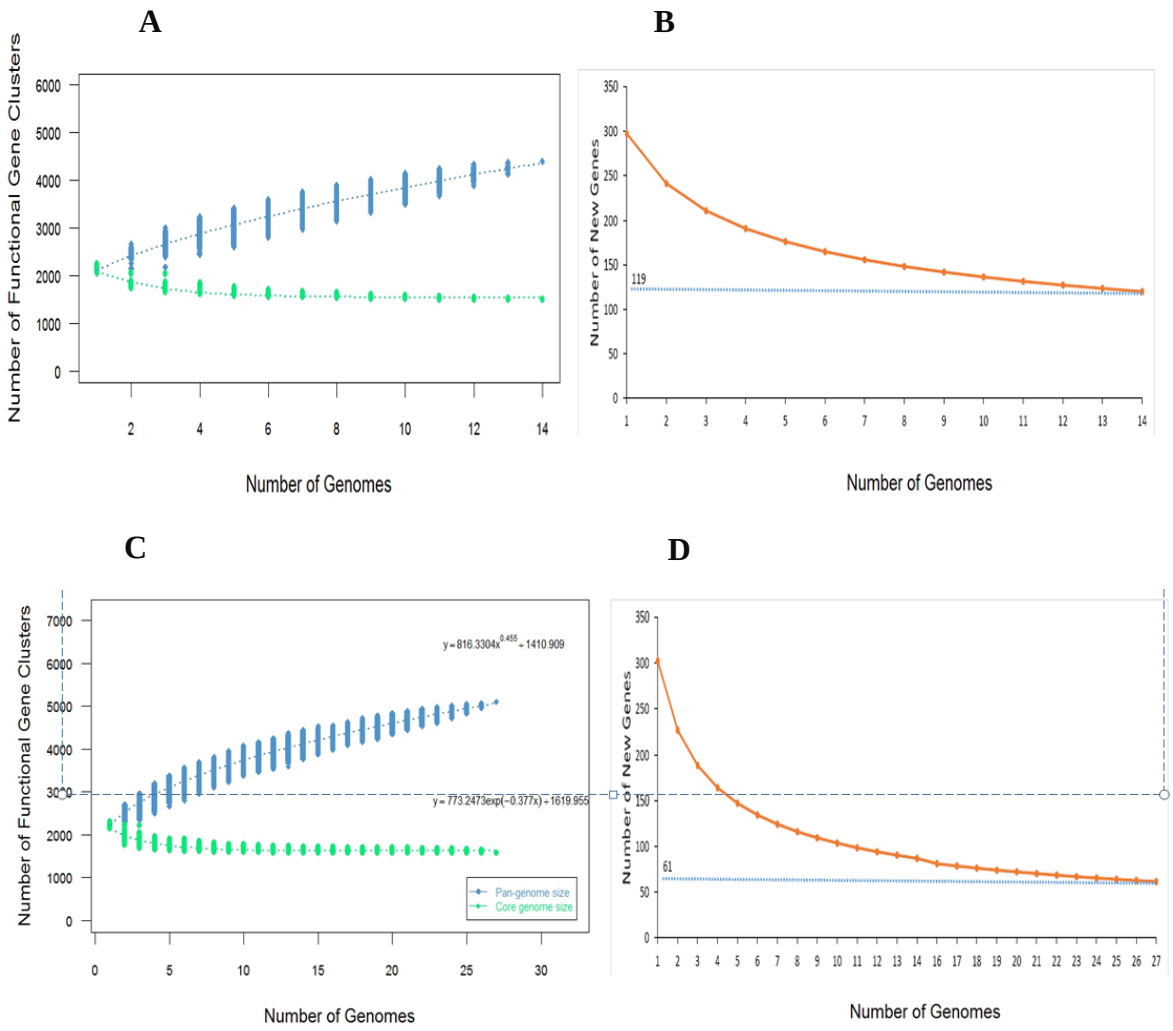


Figure 4.2: Pan-genome analyses. Curves for *S. gordonii* (A) and *S. sanguinis* (C) pan-genomes and core genomes. The blue dots denote the *Streptococcus* pan-genome size for each genome comparison whereas the green dots indicate the *Streptococcus* core genome size for each genome comparison. The median values were connected to represent the relationship between number of genomes and gene families. Curves for *S. gordonii* (B) and *S. sanguinis* (D) illustrate the number of expected new genes detected with every increase in the number of *Streptococcus* genomes.

4.5 Orthologous gene family comparisons

To identify the overlap between the predicted gene functions within the *S. gordonii* and *S. sanguinis* genomes, I clustered all predicted genes from both species that were generated during the pan-genome analysis. I compared the core genes of *S. gordonii* and *S. sanguinis*

and found they shared a large set of gene families (1,372), reflecting a high similarity between the two species (Figure 4.3). Notably, *S. sanguinis* has a relatively higher number of unique core gene families (367) compared to the unique core genes of *S. gordonii* (255).

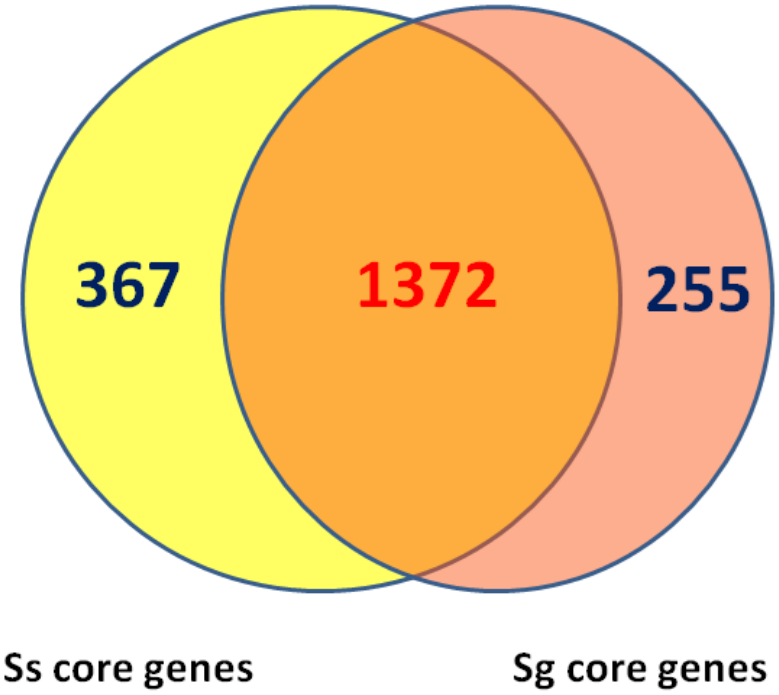


Figure 4.3: Venn diagram showing comparative analysis of orthologous genes in *S. gordonii* and *S. sanguinis*. Both species shared a high number of core genes. *S. sanguinis* has relatively higher species-specific genes compared to *S. gordonii*.

To examine the biological functions of unique core genes, I performed a functional enrichment analysis using Blast2GO software (Conesa, Götz et al. 2005). No statistically enriched functions of the unique core genes of *S. gordonii* were observed. In contrast, I found the unique core genes of *S. sanguinis* are significantly over-represented in the porphyrin-containing compound biosynthetic and cobalamin biosynthetic processes (Figure 4.4).

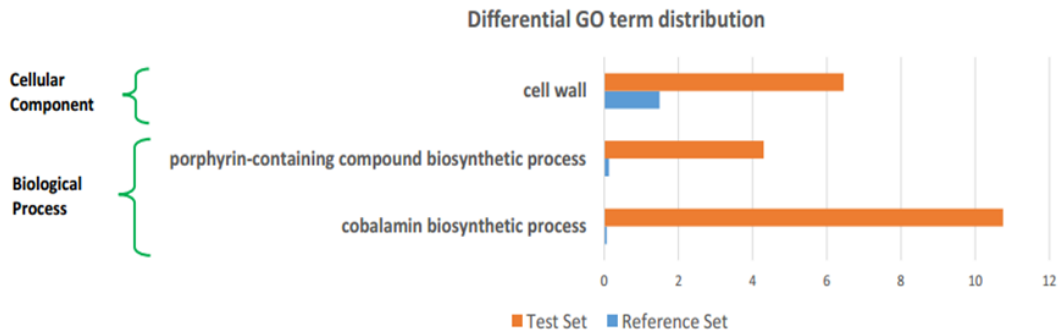


Figure 4.4: Functional enrichment analysis of *S. sanguinis*-specific core genes. The functional enrichment analysis indicates *S. sanguinis* unique core genes (orange bars) are statistically enriched in two conserved biological processes: cobalamin biosynthesis and biosynthesis of porphyrin-containing compounds. *S. sanguinis* SK36 genes were used as background dataset for comparison.

4.5.1 Porphyrin-containing compound biosynthetic process

The porphyrin-containing compound biosynthetic pathway leads to biosynthesis of porphyrin-containing compounds such as heme or siroheme (Ryter and Tyrrell 2000). In *S. sanguinis* (NCTC 7863), the superpathway of heme biosynthesis includes a number of branch points that lead to the biosynthesis of a variety of important compounds such as vitamin B_{12} (cobalamin), siroheme and heme D (Caspi, Altman et al. 2014). Eight genes involved in the porphyrin-containing compound biosynthetic pathway were identified in the unique core genome of *S. sanguinis* (S1 Table). Four of these genes encode enzymes predicted to be involved in the biosynthesis of uroporphyrinogen III from glutamyl-tRNA: glutamyl-tRNA reductase (EC 1.2.1.70), glutamate-1-semialdehyde aminotransferase (EC 5.4.3.8), porphobilinogen deaminase (EC 2.5.1.61) and uroporphyrinogen III synthase (EC 4.2.1.75). Therefore, the ability to synthesise uroporphyrinogen III appears to be conserved among *S. sanguinis* strains.

4.5.2 Cobalamin biosynthetic process

Besides the porphyrin-containing compound biosynthetic process, the unique core genes of *S. sanguinis* in the cobalamin biosynthetic process were also over-represented. Uroporphyrinogen III is the first macrocyclic intermediate in the biosynthesis of tetrapyrroles. In *S. sanguinis* it is likely that uroporphyrinogen III is particularly important for cobalamin biosynthesis since genes encoding all components of the cobalamin biosynthetic pathway were present in the unique core genes of *S. sanguinis*. Interestingly, two types of gene clusters, *cobCMTU* and *cbiACDGHKMNP* are primary cobalamin (vitamin B12) biosynthesis genes have been well-characterized in *Salmonella* Typhimurium (Raux, Lanois et al. 1996). The *cbi* genes located at the 5' end of the operon are devoted to synthesis of the corrin ring, while the *cob* genes located at the 3' end of the operon are required for the assembly of the nucleotide loop of cobalamin (Banerjee 1999). Cobalamin is required as a cofactor in the enzymatic pathways for degradation of ethanolamine into ammonia and acetaldehye and breakdown of propanediol. Previous studies have reported that cobalamin can enable different bacterial species to obtain carbon and nitrogen in anaerobic conditions within the host when ethanolamine and propanediol are abundant (Khatri, Khatri et al. 2012).

Cobalamin is a cobalt-containing vitamin and genes associated with cobalt/nickel uptake *cbi/nikMNQO* were also present in the unique core genome of *S. sanguinis*. These were functionally annotated under the membrane transport group. This gene cluster was first identified in the genome sequence of *S. sanguinis* SK36 (Xu, Alves et al. 2007). These genes are encoded within the upstream region of the cobalamin biosynthesis genes in bacterial genomes including *S. sanguinis* (Chen and Burne 2003). Previous research reported that the periplasmic binding protein NikA and ATPase NikE transporters from the

NikABCDE system of *E. coli* belong to the nickel/peptide/opine ABC transporter family (Navarro, Wu et al. 1993). The *cbiMNQO* operon encodes an Energy Coupling Factor (ECF) transporter. These systems are a subgroup of ABC transporters and CbiMNQO is essential for cobalt and nickel uptake in bacteria (Rodionov, Hebbeln et al. 2006). Moreover, the transport of nickel and cobalt along with cobalamin synthesis is particularly important in bacteria to support survival in host environments (Zhang, Rodionov et al. 2009). Hence, cobalamin synthesis and high-affinity cobalt/nickel uptake might contribute to the survival and growth of *S. sanguinis* in dental plaque and/or to its ability to cause infective endocarditis (Xu, Alves et al. 2007).

4.6 Comparative prophage analysis

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). Prophages in the genomes of *S. gordonii* and *S. sanguinis* were predicted using the Phage Search Tool (PHAST) software (Zhou, Liang et al. 2011). In total, twelve putative prophages were identified: eight in *S. gordonii* and four in *S. sanguinis* (Table 4.2). These included five intact prophages, four of which were *S. gordonii*-specific prophage and one was *S. sanguinis*-specific prophages (Figure 4.5).

Table 4.2: Summary of the comparative prophage analyses of *S. sanguinis* and *S. gordonii*. Four intact prophages (pink), six incomplete prophages (green) and a questionable prophage (blue) were identified by PHAST software. The presence of the predicted prophages in the bacterial genomes are colored in yellow for *S. gordonii* and orange for *S. sanguinis*. Two conserved prophages were identified in the genomes of all *S. sanguinis* strains. The cutoff was set as 70% prophage sequence coverage and 70% prophage sequence identity.

		Sg														Ss					
Prophage	**Length**	PV40	Blackburn	Channon	FSS3	FSS8	M5	M99	MB666	MW10	PK488	SK12	SK120	SK184	**Challis**	NCTC 7863	MB451	PJM8	FSS4	FSS9	**SK36**
7863_1	5.8kb																				
FSS8_1	43.2kb																				
SK12_1	36.5kb																				
SK184_1	59.2kb																				
SK184_3	48.7kb																				
Channon_2	39.4Kb																				
FSS4_1	30.9Kb																				
FSS4_2	16.4Kb																				
MB451_1	23.3Kb																				
SK184_2	36kb																				
SK184_4	6.9kb																				
MB666_1	47Kb																				

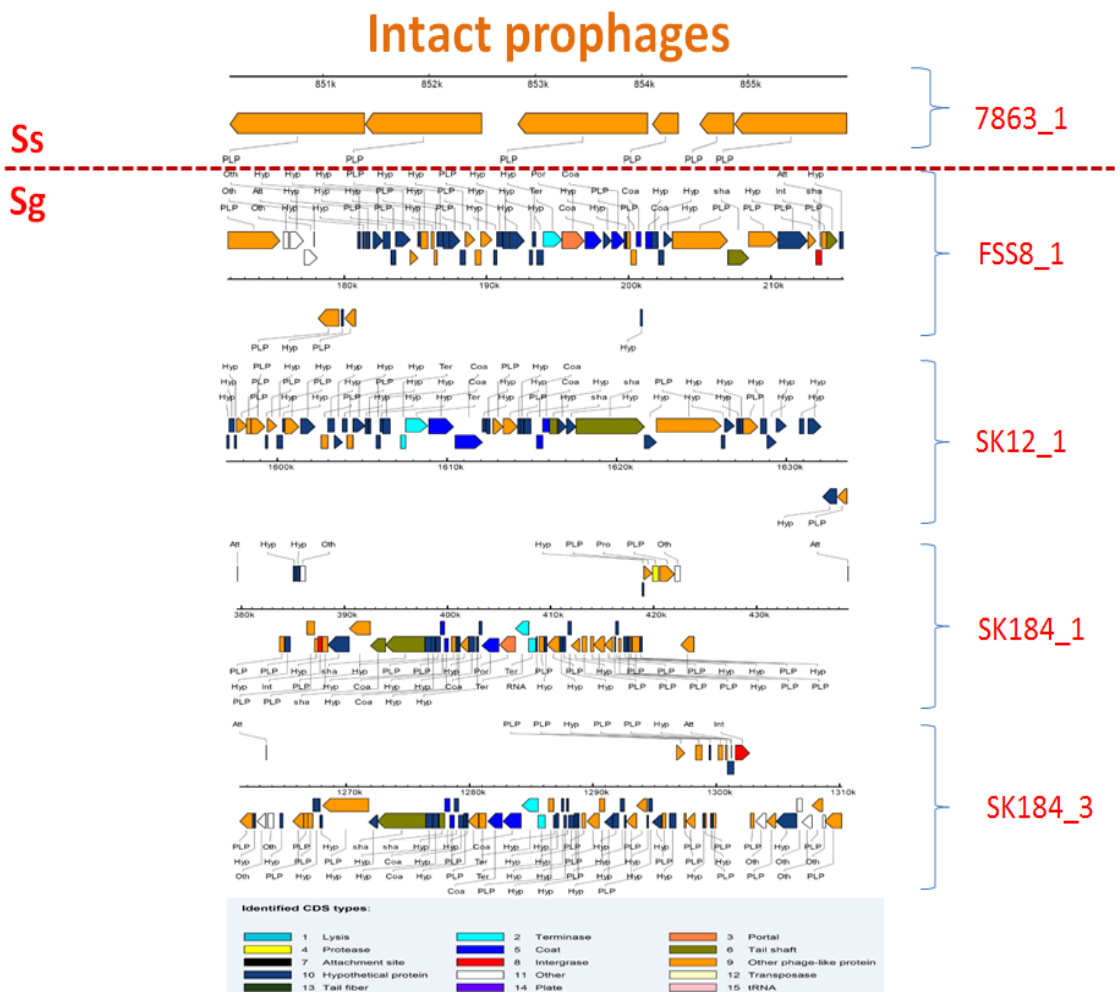


Figure 4.5: Predicted intact prophages in *S. gordonii* and *S. sanguinis*. Five putative prophages were detected, of which four were present in *S. gordonii* (FSS8_1, SK12_1, SK184_1 and SK184_3) and only one was found in *S. sanguinis* (7863_1).

Only two prophages (FSS4_1 and MB451_1) are conserved across all *S. sanguinis* strains. In addition to the phage protein orthologs, two putative attachment sites: *attL* and *attR* and ancillary enzymes such as integrase were detected in most of these prophages, providing further evidence to support the likely possibility that they were acquired by horizontal gene transfer (Table 4.3).

Table 4.3: Overview of putative prophages including the size of the prophage, the number of CDS, ATT-site (special attachment site in the bacterial and phage genomes) status and GC content.

Prophages	Genomic size (kb)	CDS	ATT-site identified	GC content
7863_1	5.8	6	No	40.99%
FSS8_1	43.2	58	Yes	38.80%
SK12_1	36.5	54	No	41.25%
SK184_1	59.2	57	Yes	41.08%
SK184_3	48.7	75	Yes	40.74%
Channon_2	39.4	62	Yes	38.76%
FSS4_1	30.9	25	Yes	43.52%
FSS4_2	16.4	29	Yes	40.31%
MB451_1	23.3	26	Yes	43.51%
SK184_2	36	21	Yes	37.94%
SK184_4	6.9	11	No	43.49%
MB666_1	47	32	Yes	40.29%

Interestingly, an operon composed of the *efeUOB* system along with genes of the twin-arginine translocation (Tat) pathway, *tatA* (Sec-independent protein secretion pathway component) and *tatC* (Sec-independent protein translocase) was found within the conserved prophage FSS4_1 in all six *S. sanguinis* genomes. The EfeUOB system can import ferrous iron under acid conditions whereas the Tat system exports folded proteins across bacterial cytoplasmic membranes (Lee, Tullman-Ercek et al. 2006, Cornelis and Andrews 2010). *Streptococcus thermophilus* was the first *Streptococcus* species reported to possess genes of the Tat system (Bolotin, Quinquis et al. 2004). Subsequently, *tatA* and *tatC* genes were detected in *S. sanguinis* SK36, encoded by SSA_1133 and SSA_1132, respectively (Lee, Tullman-Ercek et al. 2006, Xu, Alves et al. 2007).

Another conserved prophage MB451_1 found in *S. sanguinis* contains a gene encoding N-acetylmuramoyl-L-alanine amidase, a streptococcal phage lysin found in Streptococcal C1 bacteriophage (Oliveira, Melo et al. 2013). This enzyme hydrolyzes the N-acetylmuramoyl-l-alanine amide bond between the glycan strand and the cross-linking peptide of peptidoglycan (Llull, López et al. 2006). The Phage Classification Tool Set (PHACTS), which is an online computational tool, was used to classify the lifestyle of the MB451_1 prophage (McNair, Bailey et al. 2012). PHACTS predicted the prophage MB451_1 to have a temperate lifestyle (including both lytic and lysogenic phases) with an averaged probability of 0.55 and standard deviation of 0.045. Hence, I deduced that the lysogenic phase might enable the prophage MB451_1 which carries N-acetylmuramoyl-L-alanine amidase to survive without killing the host.

4.7 Comparative Pathogenomics Analysis

The genetic basis that underlies the transition of oral streptococci from commensals in the mouth to pathogens in infective endocarditis is currently unclear. To identify potential virulence factors of *S. gordonii* and *S. sanguinis,* I performed a comparative virulence gene profiling analysis using 27 *S. sanguinis* genomes and 15 *S. gordonii* genomes.

I screened for putative virulence genes in all genomes by BLASTing all protein-coding genes against the VFDB with stringent criteria (section 3.9). In total, 150 non-redundant virulence genes were identified across all 42 *Streptococcus* genomes. Of the 150 genes, *S. gordonii* strains possessed 97 to 126 of the virulence genes, whereas *S. sanguinis* strains had 101-139 putative virulence genes (Appendix DD). In total, 79 of these genes were shared between both *S. gordonii* and *S. sanguinis*. These common virulence genes include a variety of loci involved in polysaccharide biosynthesis, including homologues of *cps*, *rml*

and *rgp* gene clusters. Interestingly, the core loci for polysaccharide production appear to fall into two distinct groups that are fairly evenly distributed across *S. gordonii* and *S. sanguinis*. This provides further evidence that these species are continually evolving and exchanging genetic material in order to adapt and thrive within the host.

In *S. pneumoniae*, synthesis of capsular polysaccharides is dependent upon a large gene cluster that consists of four regulatory genes followed by serotype-specific *cps* genes (Mavroidi, Aanensen et al. 2007). This locus encodes the machinery to synthesize and export capsular polysaccharides from the cell (Bentley, Aanensen et al. 2006). Oral streptococci generally do not produce clear capsules *in vitro*, but most strains examined to date include homologous loci with four regulatory genes upstream of genes for polysaccharide biosynthesis and export. In many oral streptococci, including strains of *S. gordonii* and *S. sanguinis*, these genetic loci mediate production of receptor polysaccharides (RP*S. sanguinis*) that participate in cell-cell adhesion (coaggregation) with other oral bacteria (Yang, Yoshida et al. 2014). The structure and function of these RP*S. sanguinis* are determined by the precise combinations of transferases and polymerases present in a particular strain. For example, *S. gordonii* 38 and *S. sanguinis* SK45 contain similar *rps* gene clusters located downstream of the *nrdG* gene, but produce antigenically distinct RP*S. sanguinis*, probably due to the presence of glycosyl transferases encoded by *wefB* and *wefC* in *S. gordonii*38, compared with *wefH* in *S. sanguinis* SK45 (Yoshida, Ganguly et al. 2006, Yang, Yoshida et al. 2014). Polysaccharides produced by some strains of *S. gordonii* and *S. sanguinis*, including *S. gordonii* Challis and *S. sanguinis* SK36, are not involved in coaggregation (Yang, Yoshida et al. 2014). Disruption of the polysaccharide gene locus in *S. gordonii* Challis abrogated adhesion to collagen type I or II

indicates that the *S. gordonii* Challis polysaccharide may be more important for the recognition of host tissue rather than other bacteria (Giomarelli, Visai et al. 2006).

Closer examination of genome sequences in the strains presented here identified *rps* gene clusters similar to those of *S. gordonii* 38 and *S. sanguinis* SK45 in several *S. gordonii* strains, but not in the *S. sanguinis* strains (Figure 4.6). Only *S. gordonii* MB666 contained *wefB*, whereas *S. gordonii* M99, SK12 and SK120 contained similar gene clusters without *wefB*. All other streptococci sequenced here contained the first four genes downstream of *nrdG* (*wzg*, *wzh*, *wzd* and *wze*) but lacked clear homologues of the *S. gordonii* 38 genes such as *wchA*, *wchF*, *wefA*, *wefB*, *wefC*, *wefD*, *wzy*, *wzx*, *glf* and *wefE*. Homologues of *wchF* were identified, but these were always at a separate locus from *nrdG*-*wze*. Analysis of the *S. gordonii* Challis genome region downstream of *wze* identified a number of other putative glycosyltransferases and polysaccharide production enzymes that have not yet been well characterized (Figure 4.6).

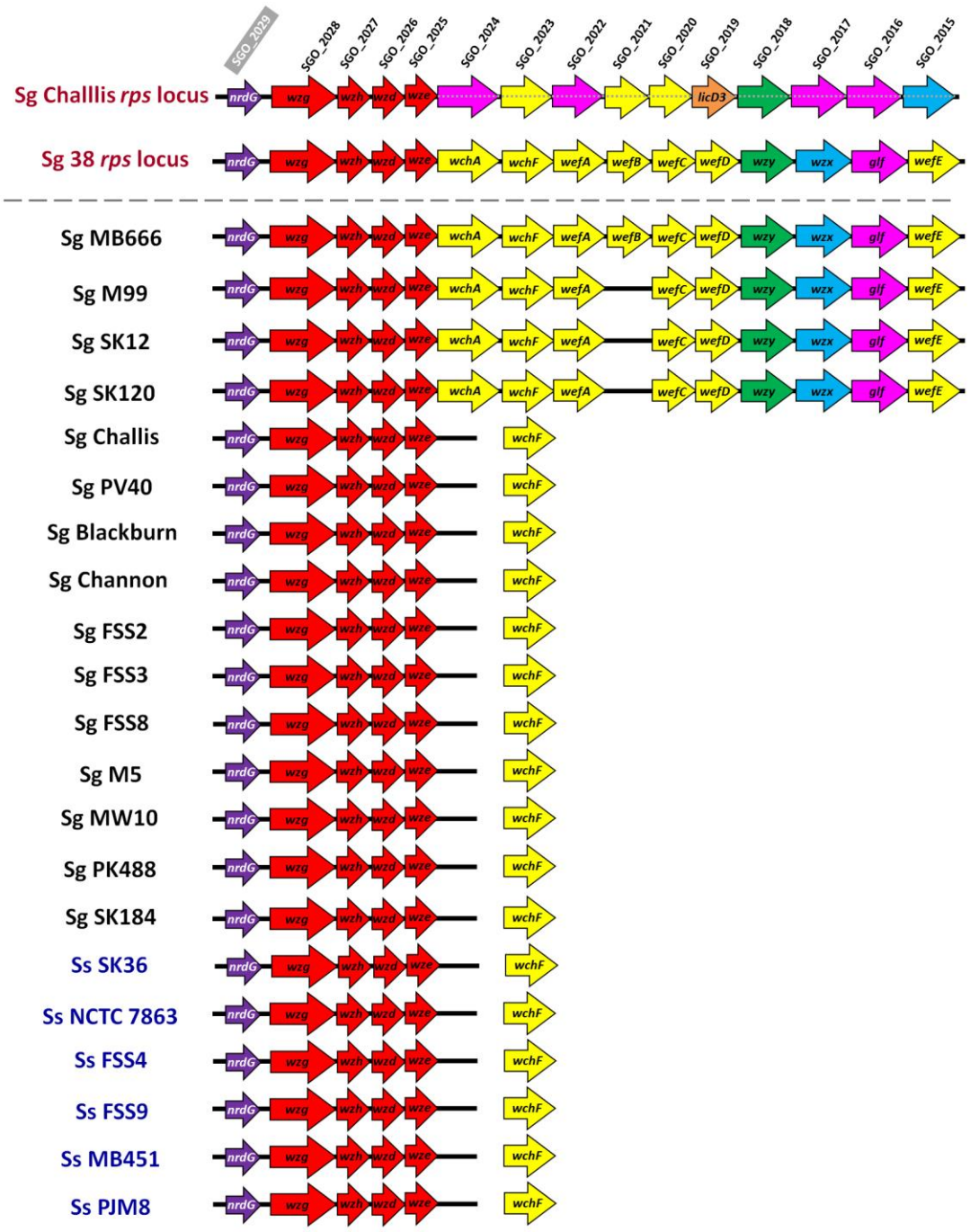


Figure 4.6: Illustration of rps/polysaccharide gene clusters of *S. gordonii* 38 and *S. gordonii* Challis in *Streptococcus* genomes. Color coding is as follows: *nrdG* gene upstream of the polysaccharide gene cluster (purple), regulatory genes (red), transferases (yellow), putative phosphorylcholine transferase licD3 (orange), polysaccharide polymerases (green), flippases (blue), nucleotide-linked sugar synthesis (magenta).

The Mauve program was used to facilitate the visualization of the *S. gordonii* Challis-type polysaccharide gene cluster structure in both *S. gordonii* and *S. sanguinis* genomes. The

Mauve genome analysis tool separated the *S. gordonii* Challis polysaccharide biosynthesis locus intonine locally contiguous blocks (LCB's) (S2 Figure). All of these were present in the same order in the *S. gordonii* strains PV40, FSS3, Blackburn, MW10, SK184, PK488 and FSS2. *S. gordonii* FSS8 lacked a large central region containing 5 LCB's. *S. gordonii* Channon displayed an absence of a smaller region of 2 LCB's and *S. gordonii* M5 was missing a region of 2 LCB's at the 3' end of the locus. All *S. sanguinis* strains shared the core polysaccharide locus structure with *S. gordonii* Challis, with the exception that they lacked the 3' LCB. Moreover, the *S. gordonii* 38-type *rps* loci in *S. gordonii* MB666, SK12, SK120 and M99 were clearly distinguishable from *S. gordonii* Challis (Figure 4.7).

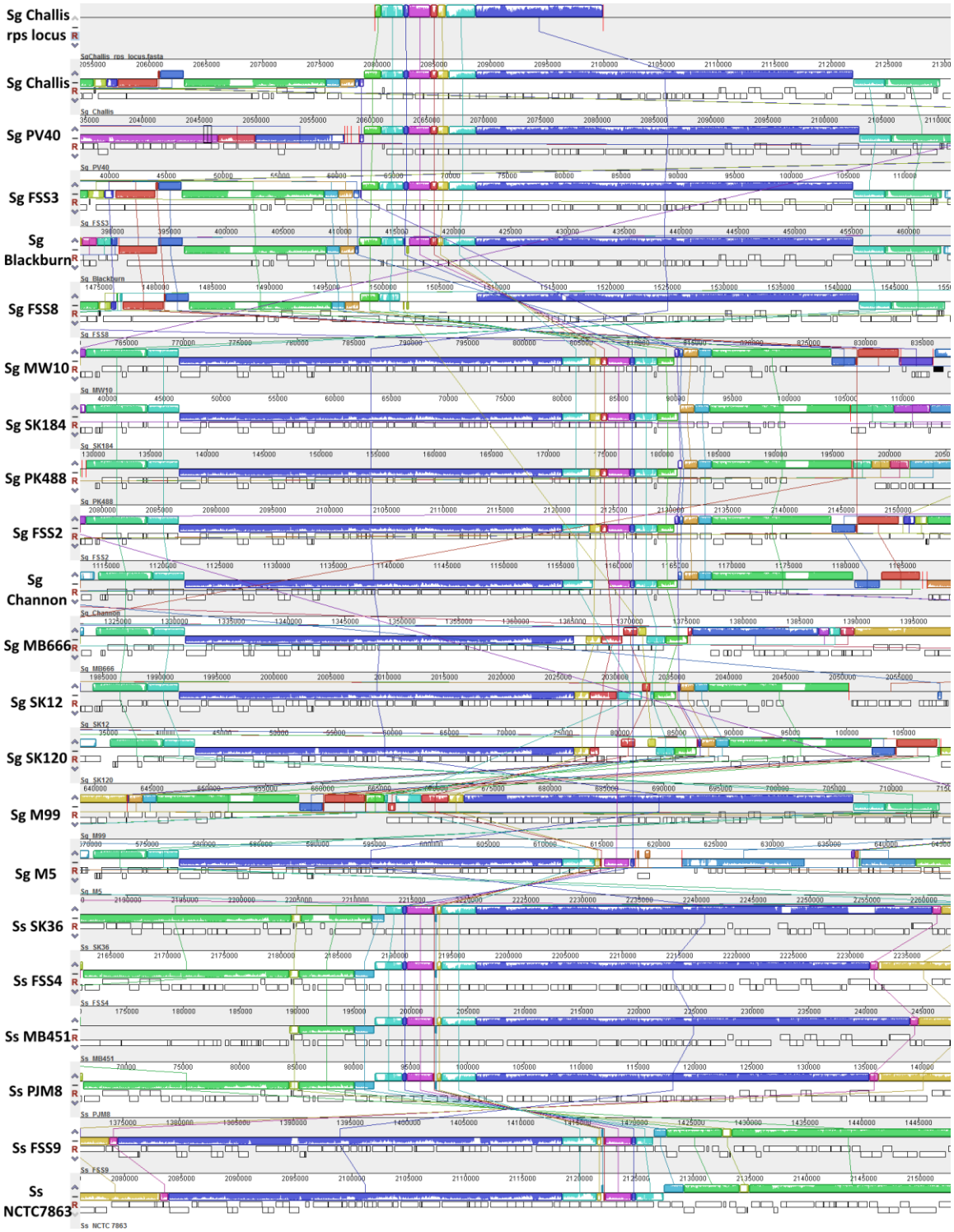


Figure 4.7: The visualization of *S. gordonii* Challis-type polysaccharide gene cluster structure in *S. gordonii* and *S. sanguinis* using Mauve software. Seven strains of *S. gordonii* (PV40, FSS3, Blackburn, MW10, SK184, PK488 and FSS2) shared the same *rps* locus structure with *S. gordonii* Challis while all *S. sanguinis* strains have clearly different *rps* locus structure with *S. gordonii* Challis.

Genes encoding enzymes involved in the production of key substrates for polysaccharide biosynthesis are located at a number of loci that are distinct from the polysaccharide biosynthesis/export operons. For instance, dTDP-L-rhamnose, is synthesized by the products of the *rml* genes. Of these, *rmlACB* are located downstream of *gufA*, whilst *rmlD* is on a separate operon downstream of *orf15*(Yang, Yoshida et al. 2014). These *rml* genes appear to be conserved in *S. gordonii* and *S. sanguinis* strains, indicating that they play key functions in the metabolism of these species. The *rml* genes, together with *rgp* genes, may also be involved in the synthesis of other rhamnose glucose polymers (RGPs) that have been identified in a range of known streptococci (Yamashita, Tsukioka et al. 1998). In *S. suis*, RGPs have been reported to be linked to several pathology-induced functions such as triggering sepsis, stimulating release of inflammatory cytokines and provoking nitric oxide production (Holden, Hauser et al. 2009). Further, the RGPs of oral streptococci have been shown to stimulate platelet aggregation, a process that is thought to be important in the pathogenesis of streptococcal infective endocarditis (Kerrigan and Cox 2012). and play significant roles in assisting bacteria to escape killing by human polymorphonuclear leukocytes (Tsuda, Yamashita et al. 2000). Overall, it is suggested that the synthesis of RGPs by *S. sanguinis* and *S. gordonii* may contribute to their pathogenesis in infective endocarditis, as well as modulating initial adhesion during the colonization of tooth surfaces and the formation of dental plaque.

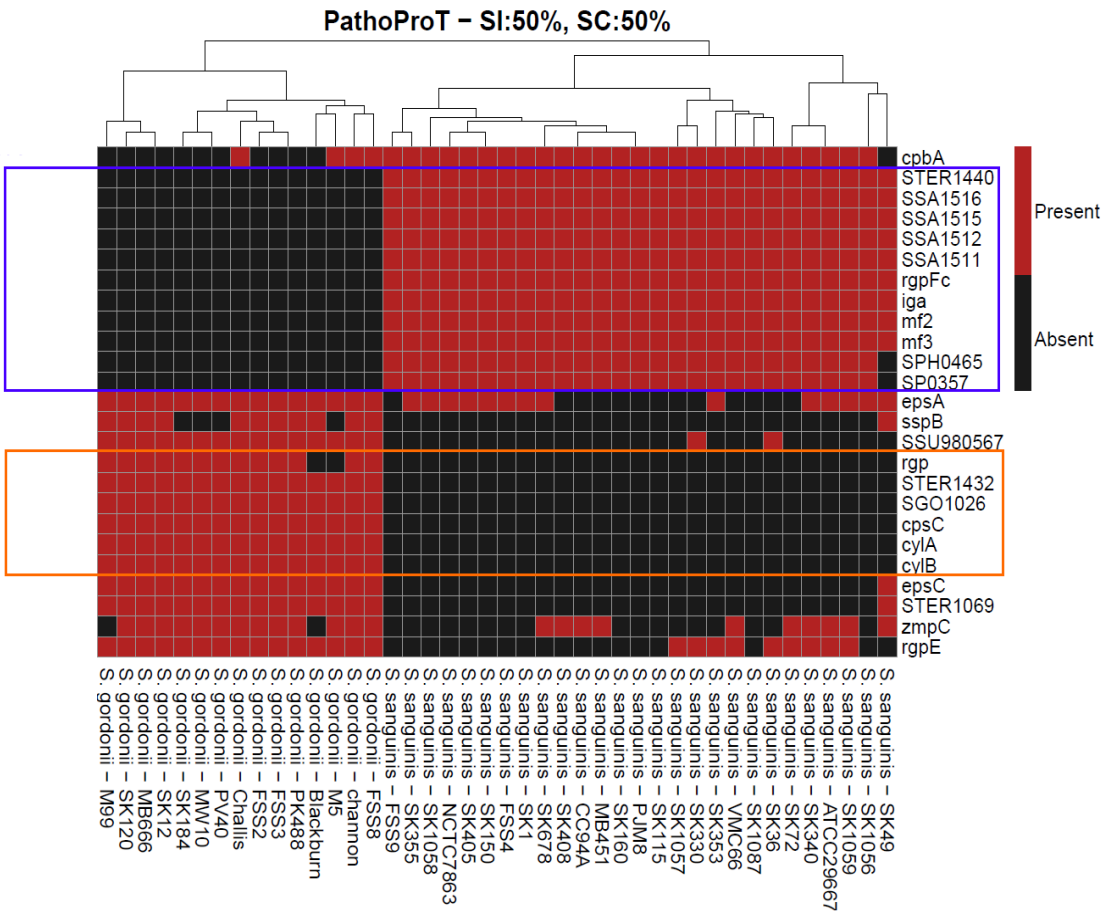


Figure 4.8: A heat map shows the main differences of virulence genes harbored by *S. sanguinis* and *S. gordonii*. (A) shows the *S. sanguinis*-specific virulence genes(highlighted in purple box). (B) depicts the *S. gordonii*-specific virulence genes (highligted in orange box).

Figure 4.8 shows the comparison ofthe virulence gene profiles between *S. gordonii* and *S. sanguinis*. Virulence-associated genes present uniquely in *S. sanguinis* include *SSA1511*, *SSA1512*, *SSA1515* and *SSA1516*, which encode hypothetical membrane proteins and glycosyltransferases. Additionally, *mf2* and *mf3* (mitogenic factor 2 and 3), which were only detected in *S. sanguinis*, encode DNases which have been reported in other streptococci to reduce the viscosity of pus via their enzymatic activity, facilitating the colonization of bacteria across tissue surfaces during invasive streptococcal infections (Relman, Falkow et al. 2000).

My virulence gene analysis also identified the *iga* gene among the list of the unique genes of *S. sanguinis*. The *iga* gene encodes IgA protease, and previous studies have shown IgA protease activity in *S. sanguinis* but not in *S. gordonii* (Kilian, MIKKELSEN et al. 1989). The IgA protease has been shown to enhance adhesion of oral bacteria to saliva-coated hydroxyapatite (Reinholdt, Tomana et al. 1990). The proteolytic activity of IgA proteases decreases the efficiency of secretory antibodies (Reinholdt and Kilian 1987). However, Fab alpha fragments are generated to sustain the antigen-binding function on the bacterial cell surface, promoting *S. sanguinis* adherence to tissues in the oral cavity (Reinholdt and Kilian 1987). The IgA proteases have exquisite specificity for human IgA, and therefore the presence of IgA proteases in *S. sanguinis* suggests an independent evolution of the enzymes in proteolysis during colonization or infection of humans (Gilbert, Plaut et al. 1991).

Strikingly, the *S. gordonii*-specific *cyl* gene cluster appears to be unique to *S. gordonii* and the β-haemolytic Group B streptococci (Rosa-Fraile, Dramsi et al. 2014). Together, *cylA* and *cylB* encode an ATP-binding cassette (ABC) transporter (Spellerberg, Pohl et al. 1999) that plays important roles in antibiotic resistance as multidrug resistance (MDR) transporters in addition to its core function as an exporter of the Cyl cytolysin (Gottschalk, Bröker et al. 2006). I investigated the homologs of *cylA* and *cylB* genes and found three homologs for each gene in *S. agalactiae*, which are currently annotated as hypothetical proteins, *cylA/cylB* proteins and *cylA/cylB* permeases separately. I assessed the completeness of the *S. agalactiae cyl* genes against the *S. gordonii cyl* genes. There was remarkably high sequence coverage and sequence identity for *cylA* and *cylB* genes which were (100/78.64)% and (100/80.82)%, respectively. To further verify this finding, I also tested the sequence coverage of *cyl* genes in the complete whole-genome of *S. sanguinis* SK36 and the results showed that *cyl* genes are likely to be absent in *S. sanguinis* genomes.

Given the presence of these genes in all *S. gordonii* strains, this may provide the first evidence of CylA and CylB production by the α-haemolytic *S. gordonii*. The role of CylA/B in multidrug resistance in *S. gordonii* remains to be determined.

4.8 Comparative Genomic Island (GI) analysis

Oral streptococci encounter significant fluctuations in environmental conditions such as surrounding pH, oxygen tension or osmolarity when growing in dental plaque. The transition to the bloodstream environment involves an even greater shift in the conditions of the external environment. It is 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, GIs in bacteria harbor genes encoding important traits such as antibiotic resistance, symbiosis and fitness (Dobrindt, Hochhut et al. 2004). Therefore, horizontally transferred GIs in the genomes of *S. gordonii* and *S. sanguinis* were predicted using the IslandViewer software tool (Langille and Brinkman 2009).

In total, 13 putative GIs were identified: two conserved GIs shared by all *S. gordonii* and *S. sanguinis* strains, 6 *S. gordonii*-specific GIs and five *S. sanguinis*-specific GIs (Table 4.3 and Table 4.4). For example, GI_55 was found to be conserved in *S. gordonii* and *S. sanguinis* and is composed of a series of putative V-type ATP synthase subunits (C, E, F,G, I and K) and a GCN5-related N-acetyltransferase (GNAT) family acetyltransferase. V-type ATP synthases are exclusively found in low GC, gram-positive bacteria and utilize the free energy released from phosphoenol pyruvate (PEP) or ATP hydrolysis to pump solutes across the membrane against concentration gradients (Samuels 2010). A recent report has suggested V-type ATPases in *S. pyogenes* are regulated by a group of small RNAs. Most

V-type ATPases pump hydrogen ions from the cytosol, ensuring the survival of *Streptococcus* species by overcoming acid stress during growth or infection. It is possible that these systems help *S. sanguinis* and *S. gordonii* to survive cycles of acidification within dental plaque. Alternatively, these systems may pump Na^+ ions rather than H^+ since it has been shown that the *Enterococcus hirae* V-type ATPase pumps Na+ ions, and promotes survival in high pH (Soontharapirakkul and Incharoensakdi 2010). However, the actual function of this system is still unclear and further work is required to determine the substrate specificity and physiological roles of streptococcal V type ATPases. On the other hand, GNAT acetyltransferase is believed to convey aminoglycoside antibiotic resistance for *S. sanguinis* and *S. gordonii* (Vetting, de Carvalho et al. 2005) which has also been reported in other oral streptococci (Richards, Palmer et al. 2014). Overall, it is likely that the acquisition of the 5,516 bp GI_55 by *S. gordonii* and *S. sanguinis* through lateral gene transfer may have enhanced their ability to survive in low-pH environments such as cariogenic dental plaque.

Another conserved GI, GI_16, consists of: iojap (Iowa-japonica) protein, a methyltransferase, a hydrolase from the Haloacid Dehalogenase (HAD) superfamily, *yqeK* gene and nicotinate-nucleotide adenylyltransferase (NAD). In bacteria, the *ybeB* gene is the ortholog of *iojap* protein which usually forms a conserved operon with the *ybeA* gene encoding a predicted methyltransferase. This *ybe* operon gene is often found adjacent to the *nadD* gene encoding nicotinate-nucleotide adenylyltransferase in nicotinamide-adenine dinucleotide (NAD) biosynthesis (Bernhardt and De Boer 2004). Additionally, this *ybe* operon has been reported to have an overlapping coding region with the *yqeK* gene, encoding a metal-dependent phosphatase (Branda, González-Pastor et al. 2004). Together, *nadD* and *ybeB* appear to form a two-domain fusion protein (Bernhardt and De Boer 2004).

Hence, I deduced the methyltransferase found in GI_16 is a likely a homologue of the *ybeA* gene which shares an operon with *ybeB* gene. However, the significance of the association between *yqeK* and *nadD* as well as the structural terminology of *nadD-YbeB* complex remains unknown.

Table 4.4: Summary of predicted GIs in the genomes of *S. gordonii* and *S. sanguinis*. Two conserved GIs were shared by *S. sanguinis* and *S. gordonii* (coloured in blue), six *S. gordonii*-specific GIs (coloured in green) and five *S. sanguinis*-specific GIs (coloured in orange).

		S. gordonii													*S. sanguinis*				
Genomic Island	Size (bp)	PV40	Blackburn	Channon	FSS2	FSS3	FSS8	M5	M99	MB666	MW10	PK488	SK12	SK120	SK184	NCTC 7863	MB451	PJM8	FSS4	FSS9
GI_5	5253																			
GI_14	10312																			
GI_16	5085																			
GI_31	5557																			
GI_43	7035																			
GI_45	5556																			
GI_47	7627																			
GI_51	7355																			
GI_53	4194																			
GI_55	5516																			
GI_58	7364																			
GI_67	4094																			
GI_75	4183																			

Table 4.5: The details of the putative GI including the size of the GI, the number of CDS, GC contents and key genes incorporated in each GI.

GI	Size (bp)	Number of CDS. *sanguinis*	GC content	Key Genes
GI_5	5253	5	33.70%	DNA recombination and repair protein RecF; FIG001621: Zinc protease; FIG009210: peptidase, M16 family and Transcriptional regulator in cluster with unspecified monosaccharide ABC transport system
GI_14	10312	6	20.50%	hypothetical proteins
GI_16	5085	6	38.90%	FIG007079: UPF0348 protein family; FIG145533: Methyltransferase (EC 2.1.1.-); Iojap protein;Hydrolase (HAD superfamily), YqeK and Nicotinate-nucleotide adenylyltransferase (EC 2.7.7.18)
GI_31	5557	6	33.60%	Permease of the drug/metabolite transporter (DMT) superfamily; TetR/AcrR family transcriptional regulator
GI_43	7035	10	42.20%	Integrase; ž Chromosome segregation helicase and MutT/nudix family protein; 7,8-dihydro-8-oxoguanine-triphosphatase
GI_45	5556	8	33.60%	Chromosome (plasmid) partitioning protein ParB / Stage 0 sporulation protein J; Serine protease, DegP/HtrA, do-like (EC 3.4.21.-); LSU m3Psi1915 methyltransferase RlmH and Competence pheromone precursor
GI_47	7627	12	43.20%	Integrase; Chromosome segregation helicase; MutT/nudix family protein; 7,8-dihydro-8-oxoguanine-triphosphatase; acetyltransferase,GNAT family;Ribosomal protein L11 methyltransferase (EC 2.1.1.-); Ribosomal RNA small subunit methyltransferase E (EC 2.1.1.-), and Mobile element protein (2 units)
GI_51	7355	9	31.90%	Chromosome (plasmid) partitioning protein ParB / Stage 0 sporulation protein J; Serine protease, DegP/HtrA, do-like (EC 3.4.21.-);LSU m3Psi1915 methyltransferase RlmH; Competence pheromone precursor; Histidine kinase of the competence regulon ComD; Response regulator of the competence regulon ComE; GTP-binding and nucleic acid-binding protein YchF; Peptidyl-tRNA hydrolase (EC 3.1.1.29) and Transcription-repair coupling factor
GI_53	4194	5	44.40%	CAAX amino terminal protease family family and Transcriptional regulator, TetR family
GI_55	5516	7	48.40%	V-type ATP synthase subunit C, E, F,G, I and K (EC 3.6.3.14) and Acetyltransferase, GNAT family
GI_58	7364	8	32.10%	Chromosome (plasmid) partitioning protein ParB / Stage 0 sporulation protein J; Serine protease, DegP/HtrA, do-like (EC 3.4.21.-); LSU m3Psi1915 methyltransferase RlmH; Competence pheromone precursor; Histidine kinase of the competence regulon ComD; Response regulator of the competence regulon ComE; GTP-binding and nucleic acid-binding protein YchF and Peptidyl-tRNA hydrolase (EC 3.1.1.29)
GI_67	4094	5	41.90%	Topoisomerase IV subunit B (EC 5.99.1.-) and lipoprotein, putative
GI_75	4183	5	44.60%	CAAX amino protease and Transcriptional regulator, TetR family

Out of the six *S. gordonii*-specific GIs detected, GI_67 is comprised of genes *camG*, encoding a putative lipoprotein, and *parE*, encoding topoisomerase IV subunit B. In *S. pneumoniae*, fluoroquinolone resistance is often associated with mutations in genes encoding subunits of topoisomerase IV, including *parE* (Varon and Gutmann 2000). The *camG* gene encodes a lipoprotein, with a leader sequence that includes a 7-amino acid peptide pheromone known as gordonii-cAM373 heptapeptide SVFILAA (Vickerman, Flannagan et al. 2010). This pheromone is required for transfer of plasmid DNA from *Enterococcus faecalis* into *S. gordonii* and has been associated with multiple antibiotic resistance (Vickerman, Flannagan et al. 2010). I hypothesize that genes on GI_67 influences antibiotic resistance in *S. gordonii* and facilitate the exchange of resistance genes between oral bacteria within dental plaque.

Interestingly, the putative *S. gordonii*-specific GI_45, GI_51 and GI_58 which vary in size from 5,556 to 7,364 bp share a large group of paralogous genes. The *com* gene cluster, *comCDE*, is located in all three putative GIs. These genes encode a peptide pheromone (*comC*) and a sensing system (*comDE*) that are involved in quorum sensing, transformation and biofilm formation (Cheng, Campbell et al. 1997, Jack, Daniels et al. 2015). Inactivation of *comD* and *comE* leads to abnormal biofilm formation which eventually decreased plaque biomass (Li, Tang et al. 2002, Jack, Daniels et al. 2015). Hence, the competence regulation operon found in GI_45, GI_51 and GI_58 of *S. gordonii* activates streptococcal cell-cell peptide signaling systems of *S. gordonii* via exogenous DNA incorporation, enabling acid tolerance of *S. gordonii* in oral biofilm formation (Matsui and Cvitkovitch 2010) Apart from its role in oral biofilm formation, *comCDE* has also been implicated in increasing

genome plasticity via uptake of new genes (Claverys, Prudhomme et al. 2000), DNA repair (Prudhomme, Attaiech et al. 2006), as well as providing nutrition of carbon, nitrogen, phosphorus, and energy source for *S. gordonii* (Finkel and Kolter 2001). It is likely that the presence of multiple *comCDE* systems may enhance the capacity of *S. gordonii* to uptake genetic material, and increase its rate of evolution. Within GI_45, GI_51 and GI_58, I identified another streptococcal plasmid acquired gene, *parB*, which is associated with important biological processes of DNA replication, cell division and cell growth (Varon and Gutmann 2000). In other bacteria such as *Vibrio cholerae* and *E. coli*, *parB* is part of an operon along with the *parA* gene that together have been implicated in drug resistance, stress response, and pathogenesis (Baek, Rajagopala et al. 2014). It is unclear whether *parB* is important in *S. gordonii* since *parA* is absent.

Another important gene, present within the GI_45, GI_51 and GI_58, is the *degP/htrA* gene, which encodes a protein responsible for folding, maturation and degradation of secreted proteins (Kim and Kim 2005). Recently, the *htrA* gene has been shown to play a key role in the repair of reactive oxygen species (ROS)-damaged DNA and protein (Henningham, Döhrmann et al. 2015). The accumulation of misfolded proteins causes the susceptibility of bacteria to high temperatures and reactive oxygen intermediates stresses. In *S. pyogenes, degP* gene knockout is impaired in virulence in a mouse model of streptococcal infection (Jones, Tove'C et al. 2001). Therefore, the presence of *degP/htrA* may enable *S. gordonii* to overcome thermal, oxidative and osmotic stresses, thus indirectly enhancing its virulence in infections.

I identified five putative *S. sanguinis*-specific GIs known as GI_31, GI_43, GI_47, GI_53, and GI_75. Of these GIs, the GI_31 carries a permease of the drug/metabolite transporter

(DMT) superfamily and a TetR/AcrR family transcriptional regulator (TFR), and thus is potentially an antibiotic resistance island. The DMT Superfamily which consists of 35 distinctive subfamilies is associated with multi-drug and various antibiotic resistances (Västermark, Almén et al. 2011). In addition, the TFRs have been reported to be overarching regulators involved in numerous processes including biosynthesis or degradation of fatty acids (Feng and Cronan 2011), antibiotic biosynthesis or activation (Uguru, Stephens et al. 2005), biofilm formation (Croxatto, Chalker et al. 2002), toxin production (MacEachran, Stanton et al. 2008), and cell-cell signaling (Pompeani, Irgon et al. 2008). Therefore, GI-31 may enhance antibiotic resistance in *S. sanguinis* and potentially may be a source of antibiotic resistance genes that can be transferred to other oral bacteria.

An intrinsic putative GI_47, which houses different functional gene components, within six *S. sanguinis* genomes was also identified. This GI includes a GNAT acetyltransferase that may convey aminoglycoside resistance. In addition, a ribosomal RNA small subunit methyltransferase E (*rsmE*) is also found in GI_47. This gene encodes an enzyme that methylates DNA, RNA, proteins or small molecules such as catechol and is also associated with antibiotic resistance (Vester and Long 2000, Morić, Savić et al. 2010). In addition, GI_47 includes the "housecleaning" gene *mutt* encoding a nudix family protein that catalyzes pyrophosphohydrolase activity directed at the removal of mutagens arising from inappropriate methylation by *rsmE* as well as reactive oxygen species generated by endogenous metabolites (Bessman, Frick et al. 1996). Two mobile elements and an integrase found within this putative GI_47 provide evidence that this region might have been horizontally transferred to *S. sanguinis*.

In addition, two putative *S. sanguinis*-specific GIs, GI_53 and GI_75, were found to include genes encoding CAAX amino protease family members and TetR family transcriptional regulators (TFR). Two genes, *bfrH1* and *bfrH2* encode CAAX family proteins. In *S. sanguinis*, these two genes are regulated by the BfrABss two-component system which controls the expression of two *bfrCD*-homologous operons (*bfrCDss* and *bfrXYss*), a *bfrH*-homologous gene (*bfrH1ss*) and another CAAX amino-terminal protease family protein gene (*bfrH2ss*). Homologues of this BfrABss system are required for biofilm formation by oral streptococci (Zhang, Whiteley et al. 2009). According to a recent report from Jimin and colleagues (Pei, Mitchell et al. 2011), *S. sanguinis* has the highest known level of CAAX amino protease compared to other member species in CPBP (CAAX proteases and bacteriocin-processing enzymes) family such as *S. pneumoniae* and *S. pyogenes*. It is likely that these CAAX effector proteases are important for the biological function of *S. sanguinis*, perhaps by contributing to establishment and survival within dental plaque.

CHAPTER 5: RESULTS – DEVELOPMENT OF STREPTOBASE

Several studies have reported that the oral streptococci are among the most common causative agents of bacterial IE and are also important agents in septicaemia in neutropenic patients. The Mitis group of oral streptococci is comprised of 13 species including some of the most common human oral colonizers such as *S. mitis*, *S. oralis*, *S. sanguinis* and *S. gordonii* as well as species such as *S. tigurinus*, *S. oligofermentans* and *S. australis* that have only recently been classified and are poorly understood at present. The availability of more Mitis group oral streptocci genomes sequences would enable researchers to gain a better understanding of these *Streptococcus* bacteria at the genomic level. Therefore, I have developed a specialized online biological database called StreptoBase in order to facilitate the ongoing research of Mitis group oral streptococci. All the comprehensive set of Mitis group oral *Streptococcus* genome sequences, annotations and results generated were collected and stored in StreptoBase. Users are able to browse, search and download the Mitis group oral *Streptococcus* genome annotations, gene sequences information and genome data as well as to perform comparative genomics analysis across different species of Mitis group oral *Streptococcus* strains. In short, StreptoBase offers access to a range of streptococci genomic resources and in-house designed analysis tools particularly for comparative genome analysis and will be an invaluable platform to accelerate research on Mitis group oral streptococci.

5.1 Datasets of StreptoBase

Seventy-seven genome sequences of Mitis group streptococci were downloaded from the public NCBI database. In addition, I have included 27 novel strains/genomes of Mitis group oral streptococci in the database. All 27 strains were clinical isolates from individuals with

dental plaque or infective endocarditis from different geographical locations (Table 5.1). Of these strains, 14 strains were isolated in the United Kingdom, 10 in United States, 2 in Australia and 1 in Denmark (Table 5.1). *S. sanguinis* NCTC 7863 is also known as ATCC 10556 while *S. gordonii* Blackburn and Channon are designated NCTC 10231 and NCTC 7869, respectively. Additionally, a number of these Mitis group strains including JPIIBBV4, JPIIBV3, JPIBVI, LRIIBV4, DGIIBVI and DOBICBV2 have been previously described (McAnally and Levine 1993). The isolation of strain M99 was described in a study of mechanisms of platelet aggregation by oral streptococci (Sullam, Valone et al. 1987). The other two oral isolates, SK120 and SK184 have also been described by Mogens Kilian and his fellow researchers in their taxonomic study of 'Viridans' Streptococci conducted in 1989 (Kilian, MIKKELSEN et al. 1989).

Briefly, the 27 Mitis group *Streptococcus* genomes were sequenced using Next-Generation Sequencing Illumina HiSeq2000 platform (Table 5.2). Data pre-processing was performed by a trimming approach (Phred score Q20) and assembled using CLC Genomic Workbench V6.5 (CLC BIO Inc., Aarhus, Denmark). In general, these assemblies showed high N50 values and low contig numbers, indicating high quality genome assemblies. The assembled Mitis group genomes harbor an average GC content of 35% to 45% and with an average genome size of approximately 2MB (Table 5.3). Using the RAST pipeline, I predicted 213,268 Coding Sequences (CD*S. sanguinis*), 5,140 RNAs and 4,542 tRNAs in all 104 genomes in the Mitis group genomes.

Table 5.1: The isolation details of 27 *Streptococcus* strains including the isolation source, geographical area and strain author.

Strain Name	Identified Species	Isolation source	Country	Strain Author	References
PV40	*S. gordonii*	Infective endocarditis	UK	P.M. Vesey, S.D. Hogg and R.R.B. Russell, Newcastle University	
NCTC 7863	*S. sanguinis*	Infective endocarditis	USA	White and Niven 1946	*Streptococcus sanguinis* (ATCC® 10556™)
Blackburn	*S. gordonii*	Human isolate	UK	R. Hare, P.H.L.S. Colindale, London	Describe in Nobbs, A. H., et al (2007). Journal of bacteriology, 189(8), 3106-3114.
BVME8	*S. parasanguinis*	Human oral cavity	UK	J. Manning, S.D. Hogg, Newcastle University	
Channon	*S. gordonii*	Not recorded	UK	R. Hare, Queen Charlotte's Hospital, London	Described in Millsap, K. W. et al (1999). FEMS Immunology & Medical Microbiology, 26(1), 69-74.
DGIIBVI	*S. tigurinus*	Dental plaque	USA	M. Levine, Oklahoma University	Described in McAnally & Levine (1993) Oral Microbiol Immunol 8: 69-74
DOBICBV2	*S. oligofermentans*	Dental plaque	USA	M. Levine, Oklahoma University	Described in McAnally & Levine (1993) Oral Microbiol Immunol 8: 69-74
FSS2	*S. gordonii*	Infective endocarditis	UK	S.D. Hogg, Newcastle University	
FSS3	*S. gordonii*	Infective endocarditis	UK	S.D. Hogg, Newcastle University	
FSS4	*S. sanguinis*	Infective endocarditis	UK	S.D. Hogg, Newcastle University	
FSS8	*S. gordonii*	Infective endocarditis	UK	S.D. Hogg, Newcastle University	
FSS9	*S. sanguinis*	Infective endocarditis	UK	S.D. Hogg, Newcastle University	
JPIIBBV4	*S. oligofermentans*	Dental plaque	USA	M. Levine, Oklahoma University	Described in McAnally & Levine (1993) Oral Microbiol Immunol 8: 69-74
JPIIBV3	*S. oralis*	Dental plaque	USA	M. Levine, Oklahoma University	Described in McAnally & Levine (1993) Oral Microbiol Immunol 8: 69-75
JPIBVI	*S. tigurinus*	Dental plaque	USA	M. Levine, Oklahoma University	Described in McAnally & Levine (1993) Oral Microbiol Immunol 8: 69-76
LRIIBV4	*S. oligofermentans*	Dental plaque	USA	M. Levine, Oklahoma University	Described in McAnally & Levine (1993) Oral Microbiol Immunol 8: 69-77

M5	*S. gordonii*	Dental plaque	USA	Rosan, B., University of Pennsylvania	Described in Rosan B (1973) Infect Immun 7 (2):205
M99	*S. gordonii*	Infective endocarditis	USA	P.M. Sullam, UCSF	Isolation described in Sullam, P.M., Valone, F.H., and Mills, J. (1987) Infect Immun 55: 1743–1750.
MB451	*S. sanguinis*	Infective endocarditis	UK	S.D. Hogg, Newcastle University	
MB666	*S. gordonii*	Infective endocarditis	UK	S.D. Hogg, Newcastle University	
MW10	*S. gordonii*	Not recorded	Australia	J. Manning, Sydney Dental School	
PJM8	*S. sanguinis*	Human oral cavity	UK	J. Manning, S.D. Hogg, Newcastle University	
PK488	*S. gordonii*	Subgingival dental plaque	USA	P. E. Kolenbrander, National Institutes of Health, MD	
POW10	*S. parasanguinis*	Not recorded	Australia	J. Manning, Sydney Dental School	
SK12	*S. gordonii*	Human oral cavity	Denmark	M. Kilian, Aarhus, Denmark	
SK120	*S. gordonii*	Human oral cavity	UK	P. H. A. Sneath (provided by M. Kilian)	Described in Kilian et al (1989) INTERNATIONAL JOURNAL OF SYSTEMATIC BACTERIOLOGY, 39: 471-484.
SK184	*S. gordonii*	Dental plaque	UK	P. Handley (provided by M. Kilian)	Described in Kilian et al (1989) INTERNATIONAL JOURNAL OF SYSTEMATIC BACTERIOLOGY, 39: 471-484.

Table 5.2: The genome sequencing statistics of 27 oral streptococci strains using Next Generation Sequencing Illumina Hiseq 2000 platform.

STRAIN NAME	YIELD (MBASES)	NUMBER OF PAIRED-END READS	MEAN QUALITY SCORE (PF)
PV40	826	8264126	36.24
NCTC 7863	822	8222024	35.69
BLACKBURN	1180	11797640	36.55
BVME8	721	7213268	36.64
CHANNON	695	6949680	36.49
DGIIBVI	1159	11591112	36.76
DOBICBV2	1037	10370504	36.56
FSS2	882	8823944	36.67
FSS3	944	9442900	37.11
FSS4	1294	12943882	36.6
FSS8	1010	10102148	36.69
FSS9	988	9877224	36.61
JPIIBBV4	1328	13283576	36.55
JPIIBV3	746	7462264	36.75
JPIBVI	1017	10165948	36.7
LRIIBV4	1273	12727786	36.65
M5	678	6784012	36.43
M99	666	6657462	36.19
MB451	1127	11271462	36.59
MB666	1095	10949508	36.81
MW10	1069	10693318	36.94
PJM8	1054	10543052	36.63
PK488	624	6240768	36.14
POW10	1074	10738992	36.58
SK12	878	8782388	36.81
SK120	732	7324194	36.96
SK184	680	6795252	36.36

Table 5.3: The genome identity of the 27 isolated *Streptococcus* strains with the summary assembly statistics.

Strain Name	K-mer	Contig no.	N50 (bp)	Genome Size (bp)	Identified Species	Genome coverage (%)	Genome Identity (%)	NCBI Accession numbers
PV 40	32	43	233745	2191051	*S. gordonii*	95	98	SAMN03480623
NCTC 7863	24	110	45631	3078022	*S. sanguinis*	84	95	SAMN03480625
Blackburn	24	50	158790	2164532	*S. gordonii*	90	96	SAMN03480626
BVME8	17	109	53977	2122687	*S. parasanguinis*	86	97	SAMN03480630
Channon	28	33	174000	2233600	*S. gordonii*	89	96	SAMN03480628
DGIIBVI	26	44	229281	1885841	*S. tigurinus*	79	94	SAMN03480631
DOBICBV2	21	99	45179	1979216	*S. oligofermentans*	77	94	SAMN03480632
FSS2	28	19	575926	2185874	*S. gordonii*	92	98	SAMN03481559
FSS3	21	398	172943	2312061	*S. gordonii*	92	96	SAMN03481560
FSS4	28	63	389092	2312671	*S. sanguinis*	85	95	SAMN03480635
FSS8	25	41	286373	2151860	*S. gordonii*	90	95	SAMN03480641
FSS9	25	20	356680	2429261	*S. sanguinis*	97	95	SAMN03480643
JPIIBBV4	30	95	48467	1991853	*S. oligofermentans*	78	94	SAMN03480680
JPIIBV3	31	75	209178	1990145	*S. oralis*	79	94	SAMN03480681
JPIBVI	28	37	940267	1792994	*S. tigurinus*	87	96	SAMN03480682
LRIIBV4	24	373	44211	2097683	*S. oligofermentans*	76	94	SAMN03481561
M5	28	67	145888	2157832	*S. gordonii*	88	95	SAMN03480683
M99	29	45	134448	2167061	*S. gordonii*	89	95	SAMN03480687
MB451	26	27	382788	2452806	*S. sanguinis*	94	96	SAMN03480686
MB666	25	20	313888	2308142	*S. gordonii*	90	96	SAMN03480688
MW10	28	27	247835	2186113	*S. gordonii*	92	98	SAMN03480689
PJM8	25	163	396031	2368281	*S. sanguinis*	92	95	SAMN03480699
PK488	38	46	183297	2262708	*S. gordonii*	91	96	SAMN03480700
POW10	14	117	30074	2042518	*S. parasanguinis*	77	96	SAMN03480701
SK12	25	28	235294	2164760	*S. gordonii*	89	95	SAMN03480703
SK120	36	27	200167	2145851	*S. gordonii*	90	96	SAMN03480740
SK184	26	53	210865	2255121	*S. gordonii*	92	97	SAMN03480741

StreptoBase currently comprises a total of 104 Mitis group oral *Streptococcus* genomes from 11 known species: *S. australis, S. cristatus, S. gordonii, S. infantis, S. mitis, S. oligofermentans, S. oralis, S. parasanguinis, S. peroris, S. sanguinis*, and *S. tigurinus* (Table 5.4).

Table 5.4: The species table summarizes the total number of draft and complete genomes of each *Streptococcus* Mitis group species accordingly.

Species	Draft Genomes	Complete Genome
S. australis	1	0
S. cristatus	1	0
S. gordonii	14	1
S. infantis	6	0
S. mitis	21	1
S. oligofermentans	3	1
S. oralis	10	1
S. parasanguinis	8	2
S. peroris	1	0
S. sanguinis	26	1
S. tigurinus	6	0

5.2 Database Structure and Composition

StreptoBase was designed to provide a wide range of useful information and functionalities (Figure 5.1).

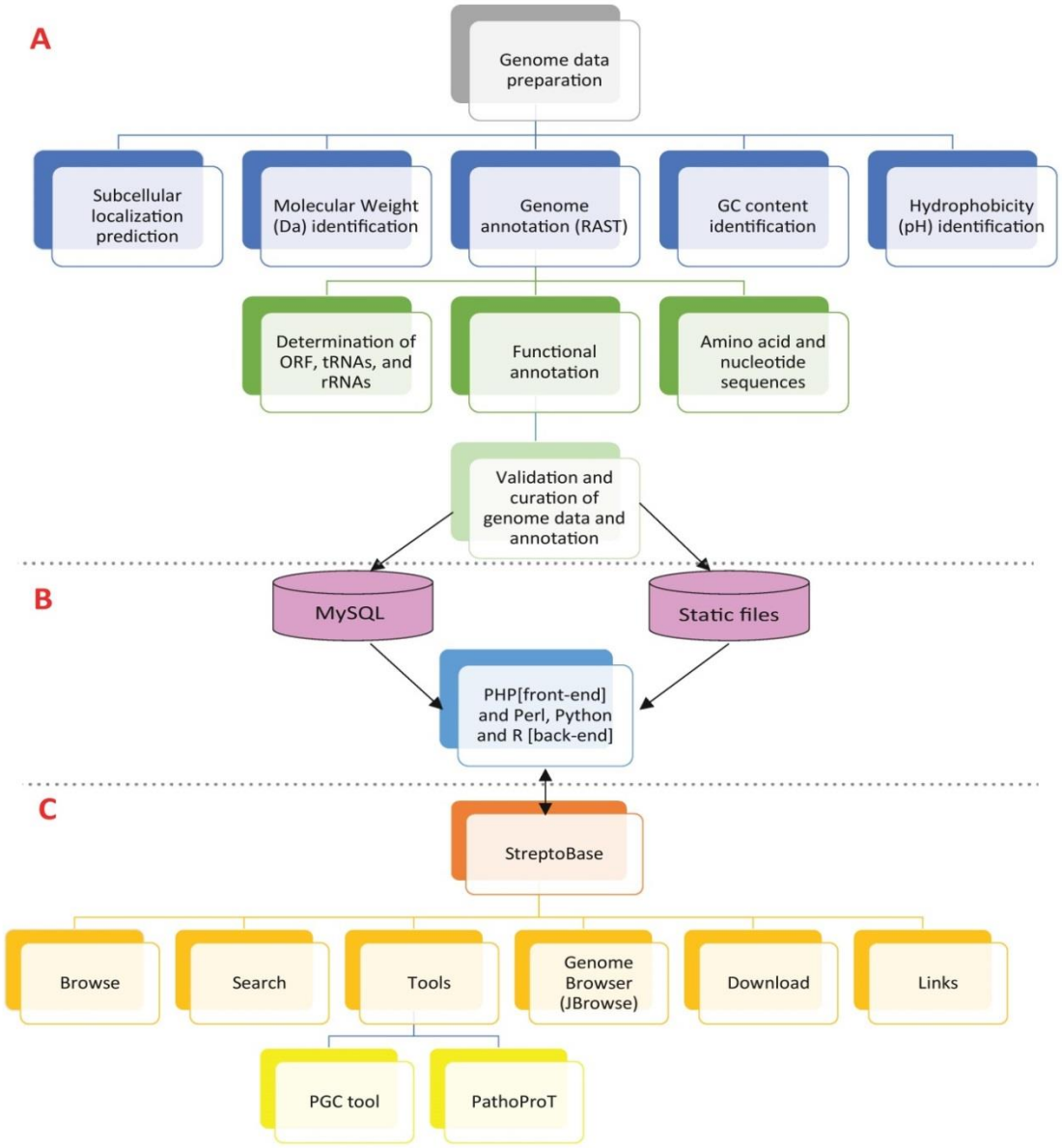


Figure 5.1: StreptoBase structure and composition. (A) Flowchart of functional annotation of *Streptococcus* genomes. (B) Diagram of the StreptoBase web server. (C) Web interface of the StreptoBase sitemap.

For instance, StreptoBase provides users with some background information about Mitis group *Streptococcus* species. Within the homepage of StreptoBase, there is a summary box which comprises the genome information stored in the database, such as number of species,

strains, number of CDS, number of RNAs and number of tRNAs (Table 5.5), which are useful before users proceed to further genome details and downstream analyses.

Table 5.5: StreptoBase summary statistics.

Database Summary	Counts
Number of Species:	11
Number of Strains/Genomes:	104
Number of CDS:	213,268
Number of RNAs:	5,140
Number of tRNAs:	4,542

Furthermore, I have compiled and gathered information from various sources on *Streptococcus* Mitis group species, for example, news and conferences, blogs and information and recently published papers, which are available in the StreptoBase homepage. By clicking on "Browse" menu, users will see the list of 11 *Streptococcus* Mitis group species along with their respective number of draft and complete genomes, while each "View Strains" button, enabling users to visualize all available *Streptococcus* genomes of any particular species, respectively. Under the "Browse Strains" page, a summarized genome description which encompasses genome size (Mbp), GC content (%) and a list of contigs, genes and rRNAs of that particular species strain are tabulated and displayed. The "Details" button allows users to access further detailed and additional data of that particular strain such as a complete list of ORFs in the genome, their corresponding functions, start and stop chromosomal positions of each ORF/gene in the "Browse ORF" page. To display all information about an ORF or gene, users can click on the "Details" button associated

with the ORF. This will open the “ORF Detail” page, displaying information such as their gene type, start and stop positions, nucleotide length, amino acid sequences, functional classification, SEED subsystem (if available), direction of transcription (strand), subcellular localization, hydrophobicity (pH) as well as molecular weight (Da) will be displayed. The demonstration of the workflow while browsing on StreptoBase is shown as following (Figure 5.2):

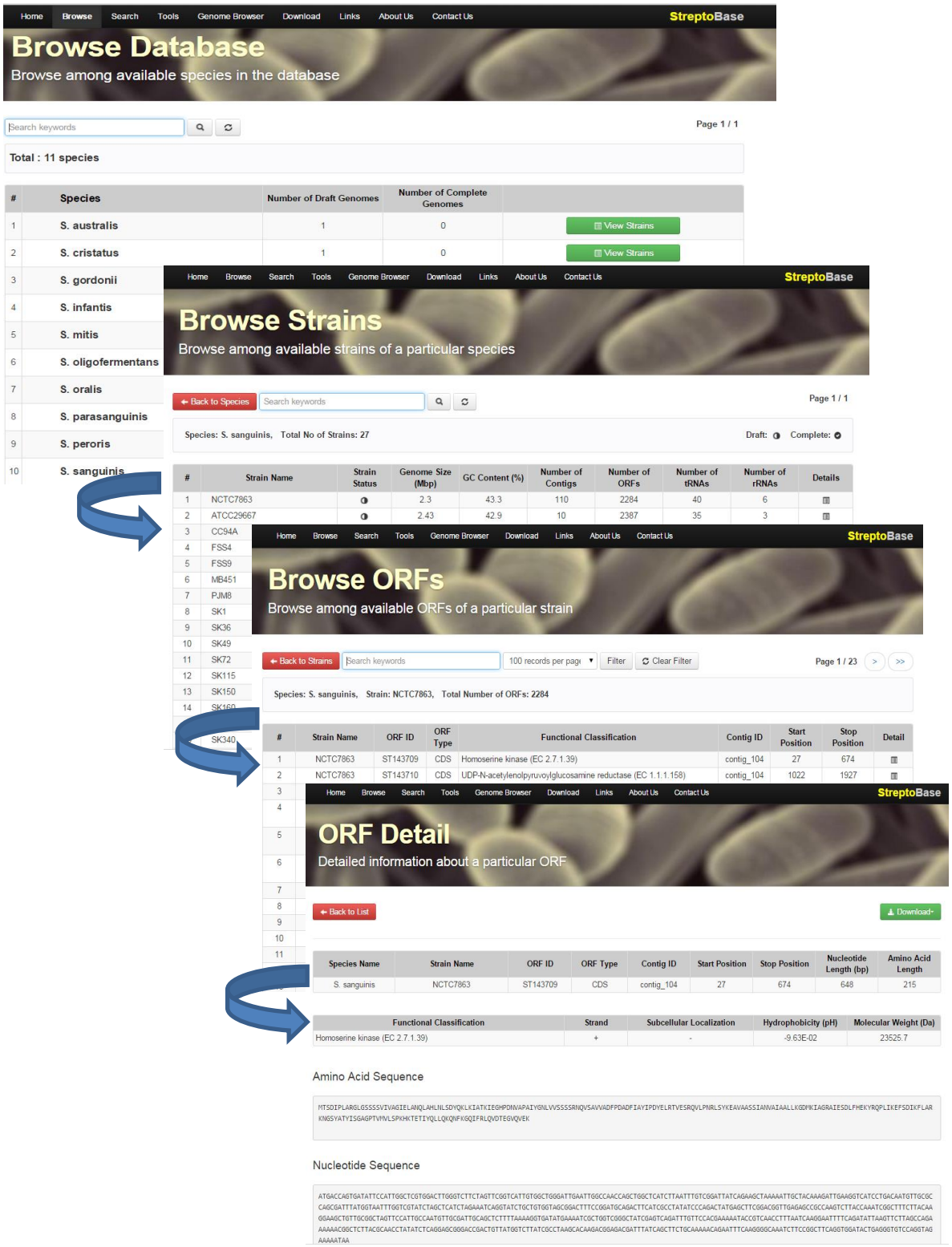


Figure 5.2: A flowchart shows the sequential processed web interfaces while browsing on StreptoBase.

(SGB) (Figure 5.3), which was customised from a well-established genome browser,

JBrowse (Skinner, Uzilov et al. 2009), a fast and modern JavaScript-based genome browser which performs navigation on genome annotations and visualization of the location of genes and flanking genomic regions/genes of a selected *Streptococcus* strain. This interactive SGB enables users to browse genes or genomic regions with graphic-wise motion smoothly and rapidly. SGB overcomes the discontinuous transitions and provides efficient panning and zooming of a specific genomic region in each *Streptococcus* genome. Furthermore, users can remotely turn on or off the DNA, RNA, and CDS tracks during the navigation process, providing flexibility in customizing what to view in the SGB viewer. I have also implemented a "Search" feature in the genome browser page, allowing users to quickly search a gene by keyword or ORF ID which is not provided by JBrowse.

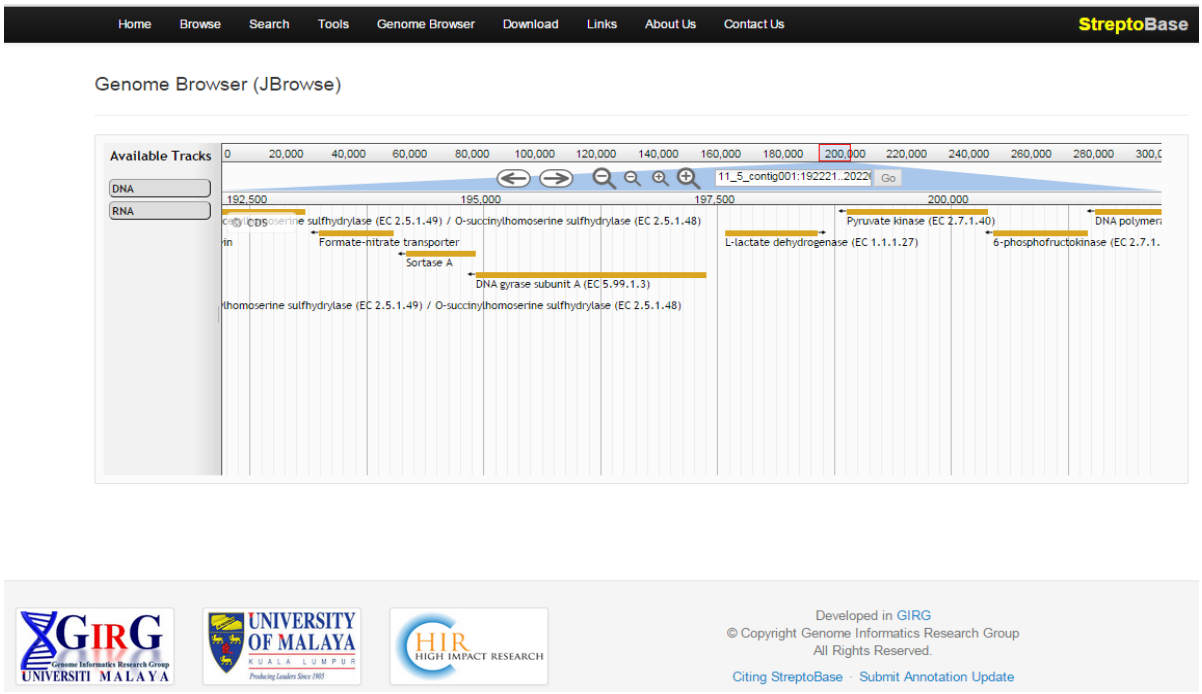


Figure 5.3: A screenshot for visualising a genomic region of *S. sanguinis* NCTC 7863 in the SGB browser.

5.2.2 Real-time keyword search engine

Considering the fact that StreptoBase would host an extensive number of genes and their annotation and this information will increase periodically, the ability to rapidly search a gene in the database is crucial. To address this issue, I implemented a real-time search engine in StreptoBase using Asynchronous JavaScript and XML (AJAX) technology. This real-time search engine was designed to support asynchronous communications between web interface and MySQL database, avoiding the need to refresh the web page and allowing the visualization of search results seamlessly. The real-time search engine retrieves a list of suggested functional classifications of genes that are related to the entered keyword once a keyword is typed.

5.3 Database Features and Incorporated Bioinformatics Tools

The Mitis group streptococci are important colonizers of the oral cavity, and are occasionally associated with serious infections (Bancescu, Dumitriu et al. 2004). In addition, these organisms have recently been suggested to play important roles in the pathogenesis of influenza (Kamio, Imai et al. 2015). Therefore, the genomic study of diverse Mitis group streptococci is essential in order to understand how these microorganisms transit from a commensal lifestyle in the mouth to subsequent pathogenesis. However, there is no existing specialized genome database available for the wide array of Mitis group streptococcal genomes for comparative genomics. While most biological genome databases only focus on the genome content and genetic variation, I have identified a need to create functional bioinformatics tools to investigate virulence determinants within genomes through comparative pathogenomics, as well as to compare the genome content and genetic variation within the Mitis group streptococci.

5.3.1 Pairwise Genome Comparison (PGC) tool

I designed and customised a web-based PGC tool for Mitis group streptococci, enabling users to select and perform pairwise comparisons between two user-selected *Streptococcus* genomes. A list of *Streptococcus* genomes is available on PGC tool of StreptoBase, allowing users to choose two *Streptococcus* genomes for cross strain or cross species comparison. Alternatively, users can upload their own genome sequences, either nucleotides or protein, and compare with the *Streptococcus* genomes in StreptoBase.

Briefly, the PGC pipeline is supported by NUCmer (Delcher, Phillippy et al. 2002) that is designed to align whole-genome sequences, and Circos (Krzywinski, Schein et al. 2009) that is a well-established tool for genome visualisation. Once users submit their jobs to the server, PGC will call NUCmer program to align user-selected genomes and in-house scripts will be used to process the genome alignment output and generate input files parsed to Circos in order to generate a circular ideogram layout of alignments. Unlike the conventional linear display of alignments, the circular layout shows the relationship between pairs of positions with karyotypes and links encoding the position, size and orientation of the related genomic elements.

Three user-defined parameters are provided in the PGC web interface including minimum percent identity (%), merge threshold (bp) and link threshold (bp). The minimum percent identity cut-off defines a homologous region (represented by links/ribbons in the Circos plot) between the two compared genomes. The merge threshold allows merging of two links/ribbons which have distance within the user-defined threshold, and the link threshold allows users to eliminate any mapped/homologous regions that have genomic size less than the user-defined cut-off. A histogram track is added in the outer ring of the circular plot to

indicate the percentage of mapped regions, allowing users to quickly identify potential indels (indicated by white gaps) and mapping regions (indicated by green charts) between the two aligned genomes. The implementation of the PGC pipeline is governed using Perl scripts. This pipeline produces two types of outputs: NUCmer alignment results and the high quality Circos plot (SVG format). Users can freely download these results for publication or further analyses in the PGC result page.

To demonstrate the utility of PGC, I compared *Streptococcus mitis* B6 (complete genome) and 17/34 (draft genome) as a case study in Figure 5.4. The parameters were set as 80% of minimum percent, default value of 1000bp link threshold and 2000bp merge threshold. *S. mitis* B6 was isolated in Germany, whereas *S. mitis* 17/34 was isolated from the urethra of a Russian patient with urethritis. Based on the generated PGC plot, both *S. mitis* genomes generally shared high similarity as most of their genomic regions could be aligned (Figure 5.4). One of the features of PGC plot is its ability to quickly identify putative indels via visualization of the gaps in the plot chart which is supported by information displayed in the histogram track. For instance, two of the gap occurrences (Figure 5.4) indicate the absence of genomic regions in the *S. mitis* 17/34 genome. The external circular bar of the plot shows the genome size measurements which are approximately 2MB for both *S. mitis* genomes. Based on the gap observed in Figure 5.4 (indel 'A'), the gene loss is likely to occur close to position 400,000bp.

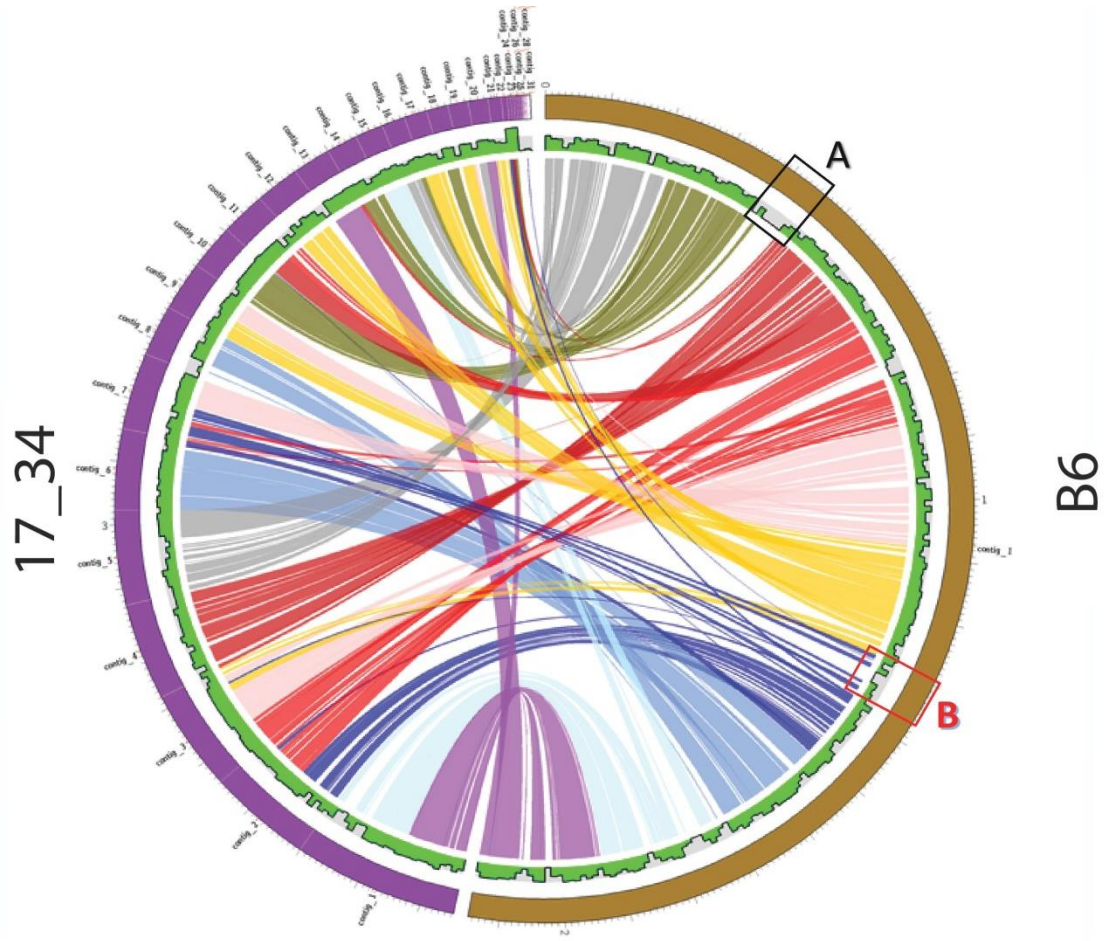


Figure 5.4: Pairwise genome comparison between *S. mitis* B6 *and S. mitis* 17/34 using PGC tool incorporated in StreptoBase. 50% sequence identity and 50% sequence coverage were used to compare the genomes of the two bacterial strains. A and B highlight the putative indels found in the genome comparison between *S. mitis* B6 *and S. mitis* 17/34.

Next, I examined the genes located at indel 'A' in *S. mitis* B6 (Figure 5.4) by visualising this region using SGB. I identified many phage-related genes associated with this region. To further examine this region, I utilized PHAST (PHAge Search Tool) to annotate and identify prophages sequences found within the genome of *S. mitis* B6 (You Zhou et al., 2011). A 56Kb intact prophage with 82 CDSs and GC content of 39.9% was detected from 390,924bp to 446,969bp. Since *S. mitis* B6 is a complete genome, I can therefore imply the base pair position directly into the B6 annotation file. According to PHAST results, this

intact prophage of *S. mitis* B6 comprised phage-associated genes including phage integrase protein, phage CI-like repressor, phage binding protein, phage portal protein, SPP1 family phage head morphogenesis protein and phage capsid proteins. Therefore, I suggest that *S. mitis* B6 might have recently acquired this intact prophage. The graphical display of the intact prophage with different types of phage-related genes is shown in Figure 5.5.

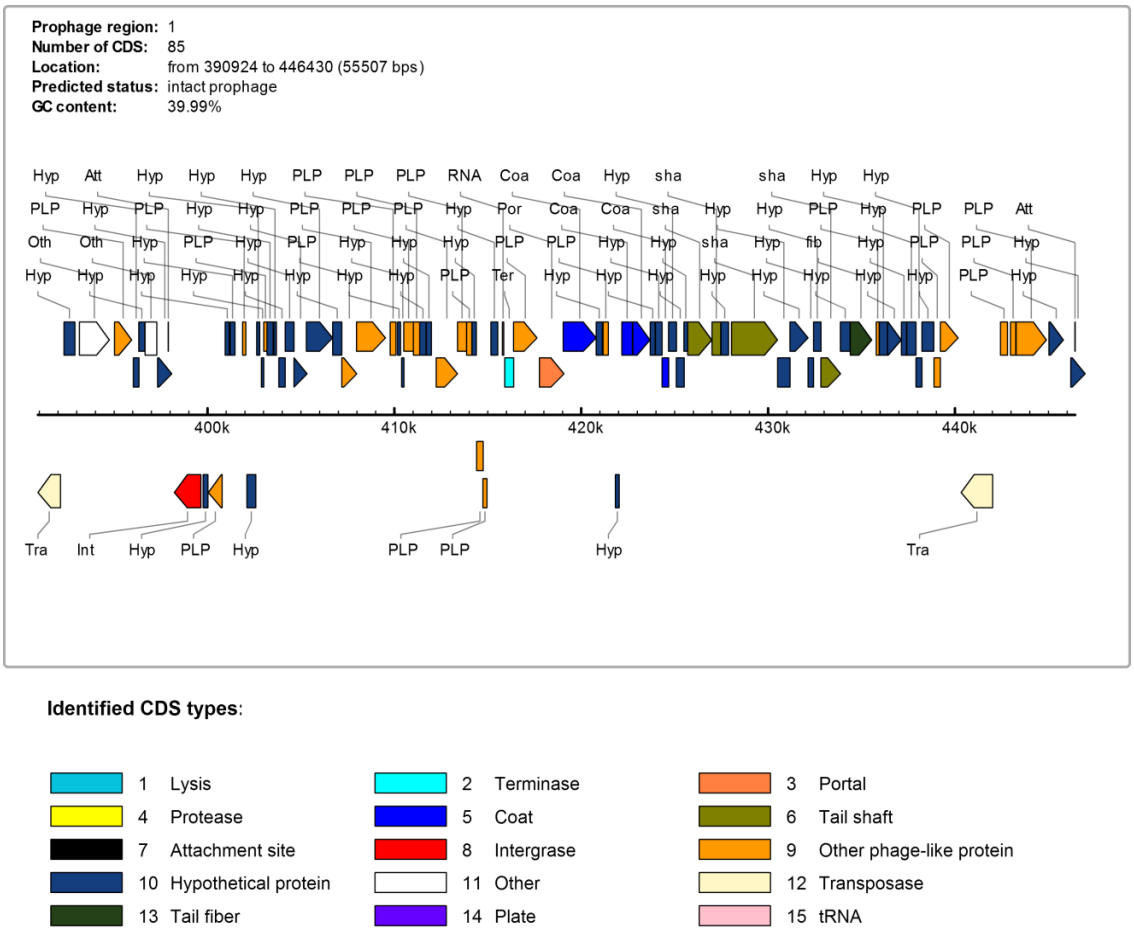


Figure 5.5: A putative intact prophage detected in the genome of *S. mitis* B6. This prophage has 85 putative genes.

Based on the indel 'B' detected on the PGC plot in Figure 5.4, I have revealed a 24Kb incomplete prophage with GC content of 39.17% located at position 1356040bp to 1380128bp. Interestingly, this region contains a complete *atp* operon regulated by the CcpA protein within this incomplete prophage of *S. mitis* B6 genome. The genes of the *atp*

operon are shown in Table 5.6. These genes encoding ATP synthases are commonly possessed by oral streptococci for adaptation to the acidic host environment by creating a more alkaline internal system.

Table 5.6: The ATP synthases within the *atp* operon of *S. mitis* B6.

Locus Tag	Gene Name	Functional annotation
smi_1315	*atpE*	ATP synthase C chain (EC 3.6.3.14)
smi_1314	*atpB*	ATP synthase A chain (EC 3.6.3.14)
smi_1313	*atpF*	ATP synthase B chain (EC 3.6.3.14)
smi_1312	*atpH*	ATP synthase delta chain (EC 3.6.3.14)
smi_1311	*atpA*	ATP synthase alpha chain (EC 3.6.3.14)
smi_1310	*atpG*	ATP synthase gamma chain (EC 3.6.3.14)
smi_1309	*atpD*	ATP synthase beta chain (EC 3.6.3.14)
smi_1308	*atpC*	ATP synthase epsilon chain (EC 3.6.3.14)

It has been reported that the protective mechanism is critical especially for streptococcal acid-sensitive glycolytic enzymes (Lemos, Abranches et al. 2005). Hence, there is a possibility that the acquisition of this *atp* operon carried by the incomplete prophage of *S. mitis* B6 via horizontal gene transfer might have assisted its commensal status in maintaining the optimal pH level for bioenergetics processes of *S. mitis* B6 cells.

5.3.2 Pathogenomics Profiling (PathoProT) tool

PathoProT was designed to predict virulence genes by comparing *Streptococcus* protein sequences against the Virulence Factors Database (VFDB) (Chen, Yang et al. 2005). PathoProT utilizes the stand-alone BLAST tools downloaded from the NCBI website.

VFDB (Version 2012) currently hosts a set of 19,775 experimentally verified virulence genes originating from a wide range of different bacterial species, providing a useful resource for sequence homology searches. Users can select a list of *Streptococcus* strains for comparative analysis and set the cut-off, for example, genome identity and completeness for the BLAST search through the provided online web form. The default parameters of PathoProT pipeline are set at 50% sequence identity and 50% sequence completeness for searching and identifying orthologous virulence genes across the selected *Streptococcus* genomes. However, users can apply their desired cut-offs for the homology search in order to achieve the optimal stringency levels in their analyses.

Briefly, PathoProT pipeline was mainly implemented using Perl. In-house Perl scripts will process BLAST outputs (generated by searching these query sequences against VFDB) for each RAST-predicted protein (query sequence) in the user-selected genomes and identify putative virulence based on user-defined parameters. The filtered BLAST results are consolidated and organised in a matrix table containing information of presence or absence of virulence genes (rows) and *Streptococcus* strain names (columns). Finally, PathoProT will pass and process this output with the in-house R scripts for hierarchical clustering (complete-linkage algorithm) and generating a heat map for visualisation. The *Streptococcus* strains will be sorted based on their virulence gene profiles (Figure 5.6) and a phylogenetic tree will be drawn, users are able to gauge the relationships among the closely-related *Streptococcus* Mitis group species/strains as well as their corresponding putative virulence genes form noticeable clusters through the dendrograms. Therefore, this comparative pathogenomics analysis pipeline is able to provide excellent insight into the virulence gene profiles across different species of *Streptococcus*.

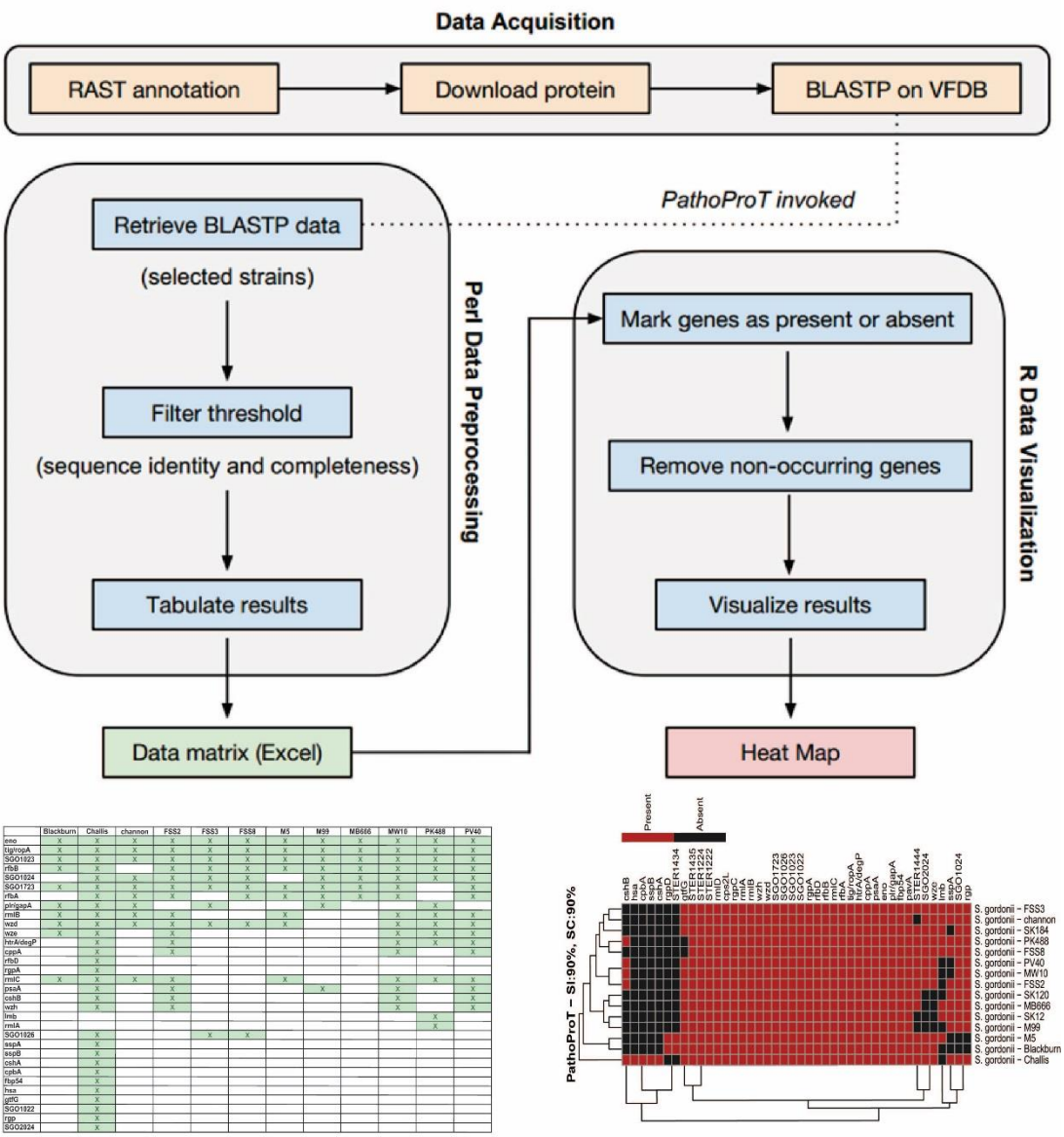


Figure 5.6: A PathoProT flowchart. PathoProT was mainly implemented using Perl and R scripts. The input of PathoProT would be lists of genes for the selected strains/genomes and the pipeline will generate a heat map at the end of the process.

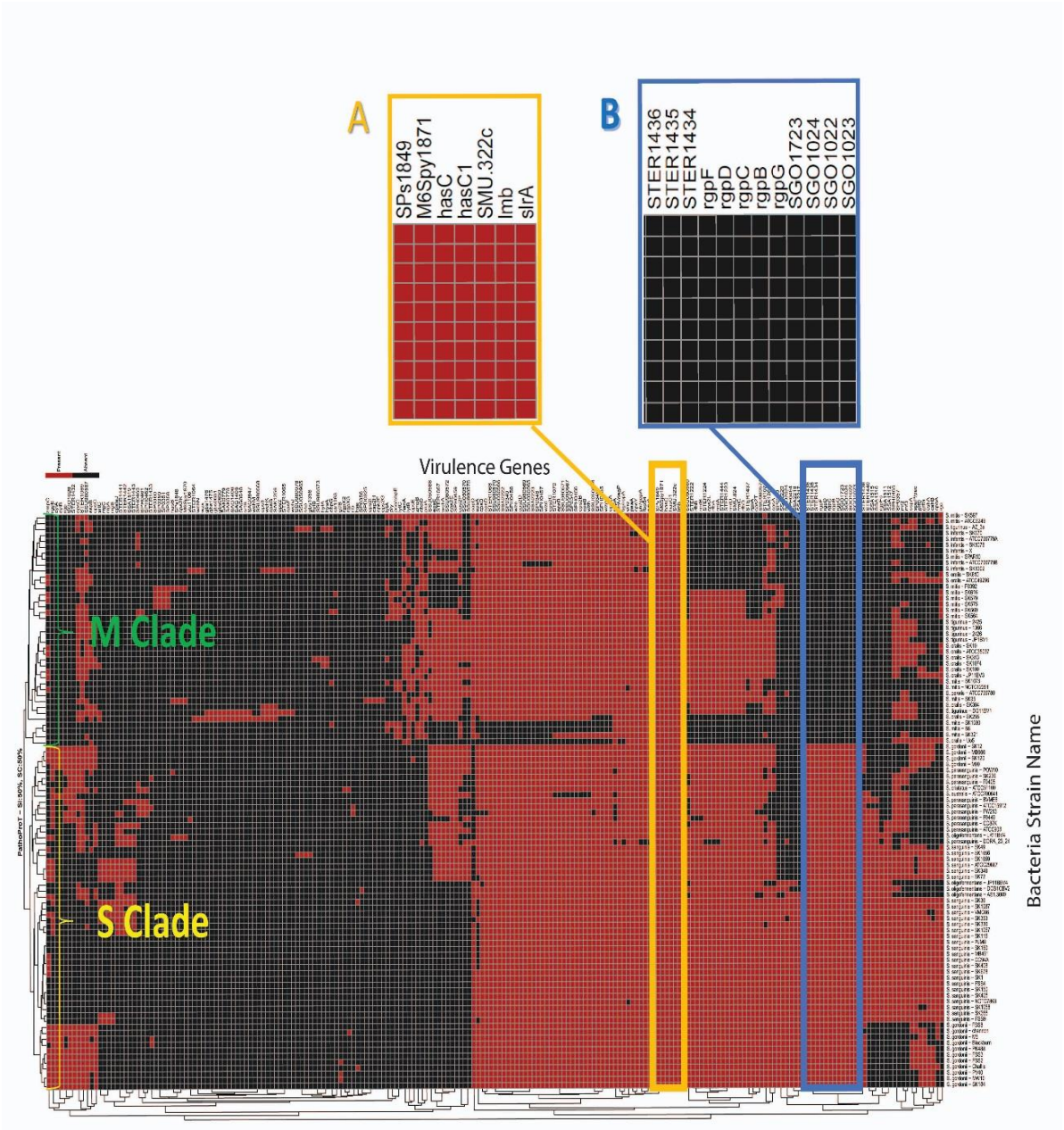


Figure 5.7: An informative heat map generated by PathoProT tool. (A) A list of conserved putative virulence genes carried by all Mitis group species and (B) The RGP synthesis related genes which can differentiate M Clade and S Clade. Presence of the virulence gene was indicated in red and absence of the virulence genes was indicated in black.

To demonstrate the features or functionalities of PathoProT, I present a comparative virulence gene study among the Mitis group streptococci using a threshold of 50% for both

sequence identity and coverage to give an insight into their virulence gene profiles. Based on the generated PathoProT heat map, a number of putative virulence genes appear to be conserved among all the Mitis group species (Figure 5.7). The conserved genes *hasC (hasC1*or*SMU.322c)* which encodes UTP-glucose-1-phosphate uridylyltransferase (or UDP–glucose pyrophosphorylase)(M6Spy1871) is involved in synthesis of the hyaluronic acid (HA) capsule along with two neighboring genes: *hasA* and *hasB* within the *has* operon.(Crater and Van de Rijn 1995). In *Streptococcus,* HA is found as streptococcal capsule material in *S. pyogenes* and related species and is an important virulence factor, camouflaging the bacteria efficiently against the recognition of host immune system (Wessels, Moses et al. 1991, Schmidt, Günther et al. 1996) as well as protecting them against reactive oxides released by leukocytes (Cleary and Larkin 1979). Additionally, it is possible that HA plays a significant role in Mitis group streptococcal adherence and colonization of epithelial cells, leading to bacterial resistance against phagocytosis by macrophages (Wibawan, Pasaribu et al. 1999, Kim, Park et al. 2006, Chen, Marcellin et al. 2009).

Another conserved virulence gene, *slrA* encodes streptococcal lipoprotein rotamase A, which is one of the major surface proteins expressed by *S. pneumoniae.* This gene is an important cyclophilin that modulates biological function of virulence proteins during the first stage of pneumococcal infection (Hermans, Adrian et al. 2006). It is likely that the *slrA* gene promotes invasion of host cells and facilitates pneumococcal colonization and adherence in Mitis group streptococci (Moscoso, García et al. 2006, Sanchez, Kumar et al. 2011). Furthermore, it has been reported that deficiency in *slrA* can reduce bacterial virulence due to its impact on the adherence and internalization by epithelial and endothelial cells (Hermans, Adrian et al. 2006). Likewise, the conserved *lmb* gene encodes

a laminin-binding protein which was first identified in *S. agalactiae* (Dmitriev, Shen et al. 2004). Virtually identical adhesins were later discovered in both *S. suis* (Zhang, Shao et al. 2014) and *S. pyogenes* (Elsner, Kreikemeyer et al. 2002, Terao, Kawabata et al. 2002). The *lmb* adhesins have been proposed to help in bacterial pathogenesis via invasion of the damaged epithelium (Spellerberg, Rozdzinski et al. 1999). Overall my data showed that many surface lipoproteins and adhesins that are important in virulence and pathogenic infections are highly conserved across the Mitis group streptococci.

According to the dendrogram generated on the left side of the PathoProT heat map (Figure 5.7), the Mitis group can be clearly categorized into two major clades based on their virulence gene profiles: S Clade (*S. sanguinis, S. gordonii, S. parasanguinis, S. australis, S. cristatus* and *S. oligofermentans*) and M Clade (*S. mitis, S. infantis, S. tigurinus, S. oralis* and *S. peroris*). This phylogeny relationship of the oral streptococci Mitis group species indicates the close relatedness of cross-species within M Clade and species-to-species of S Clade. Interestingly, I found the *rgp* genes may be able to use as markers to differentiate the two different clades in the heat map. For instance, these marker genes are present in all S Clade species but absent in all the M Clade species.

The *rgp* genes cluster (B, C, D, F and G) is responsible for the synthesis of rhamnose-glucose polysaccharide (RGP) in *Streptococcus mutans.* Notably, similar genes have been found to be involved in rhamnan synthesis in *E. coli* (Shibata, Yamashita et al. 2002). In fact, it has been suggested that *E. coli* and *S. mutans* share a common pathways for rhamnan synthesis based on their similarities in RGP synthesis (Shibata, Yamashita et al. 2002). The function of *rgpB* is to transfer the second rhamnose residue to a rhamnose residue on *N*-acetylglucosamine linked to the lipid carrier, followed by *rgpF* which later

catalyzes the transfer of the third rhamnose residue to the second rhamnose residue of the resultant glycolipid carrier. Both *rgpB* and *rgpF* have presumably to work alternately in the elongation of the rhamnan chain. Homologous rhamnosyl transferases of *rgpB* and *rgpF* have been detected in *S. thermophilus* (STER1436) and *S. gordonii* (SGO1022). On the other hand, *rgpC* and *rgpD* genes encode the putative ABC transporters specific for RGP (homologous STER1434 in *S. thermophilus* and homologous SGO1024 in *S. gordonii*) which play role in polysaccharide export (Shibata, Yamashita et al. 2002). The *rgpG* gene (*S. gordonii* SGO1723 homolog) initiates the RGP synthesis by transferring ***N***-acetylglucosamine-1-phosphate to a lipid carrier (Yamashita, Shibata et al. 1999).

The *rgp* genes are also implicated in pathogenesis in several *Streptococcus* species. For instance, *rgp* plays an essential role in bacterial virulence as well as eliciting an inflammatory response in *S. suis* (Holden, Hauser et al. 2009). Induction of infective endocarditis by *S. mutans* has been reported to be triggered by *rgp* genes via nitric oxide release (Martin, Kleschyov et al. 1997), platelet aggregation (Chia, Lin et al. 2004) and conferring resistance to phagocytosis by human polymorphonuclear leukocytes (Tsuda, Yamashita et al. 2000). Therefore, S Clade Mitis group streptococci which produce these rhamnose rich polymers might exhibit a different pattern of pathogenesis from M Clade *Streptococcus* species in order to establish greater virulence and increased survival in host cells. A previous study has identified the *Sanguinis* group of streptococci as a common causative agent of transient bacteremia which potentially can lead to infective endocarditis (Widmer, Que et al. 2006). This group has also been reported to be present in a few cases of virulent septicemic infection in neutropenic patients (Shelburne, Sahasrabhojane et al. 2014).

5.3.3 Sequence search tools

I have incorporated two types of BLAST engines, standard BLAST and VFDB BLAST, into StreptoBase to search for the closest *Streptococcus* strains to the query strain. These exclusive BLAST searches are functionally based on the stand-alone BLAST tool (Johnson, Zaretskaya et al. 2008) downloaded from NCBI. Both BLAST engines support three types of BLAST functions, namely, BLASTN, BLASTP and BLASTX. Users are allowed to define the genome completeness (%) and genome identity (%) on the BLAST tools submission forms. These specialized BLAST tools are aimed to facilitate users to perform similarity searches of their query sequences against *Streptococcus* genome sequences, gene sequences (standard BLAST) as well as against the virulence genes of VFDB (VFDB BLAST), which allows users to examine whether their genes of interest are potential virulence genes using a sequence homology approach.

5.4 Availability and System Requirements

StreptoBase is available online at http://streptococcus.um.edu.my. Users can download and visualize all sequences and annotations described in this paper on the StreptoBase website. This analysis platform is generally compatible with multiple type of browsers including Internet Explorer 8.x or higher, Mozilla Firefox® 10.x or higher, Safari 5.1 or higher, Chrome 18 or higher and any other equivalent browser software. This web site is best viewed at a screen resolution of 1024 × 768 pixels or higher.

CHAPTER 6: DISCUSSION

6.1 Overview

In the present study, I have successfully isolated, sequenced, assembled and annotated the whole-genome of 27 Mitis group streptococci: fourteen *S. gordonii* strains, five *S. sanguinis* strains, three *S. oligofermentans* strains, two *S. parasanguinis* strains, two *S. tigurinus* strainsand one *S. oralis* strain. Among the 27 Mitis group streptococci genomes, 14 strains are oral isolates, 10 strains are IE isolates and the origin of three strains were not recorded. Of these strains, 14 strains were isolated in the United Kingdom, ten in United States, two in Australia and one in Denmark. In general, this study supports the expanding *Streptococcus* research with the genome announcements of these 27 strains of Mitis group streptococci which contributed to the existing genome database of the species of Mitis group streptococci. Significantly, I present a comparative genome study of 19 *S. gordonii* and *S. sanguinis* clinically-derived isolates using different bioinformatics genomic approaches encompassing phylogenetic analysis, pan-genome analysis, function annotation and enrichment analysis, prophages and GI analysis and pathogenomic analysis. This comparative study revealed the genomic similarities and differences between these two closely related species which greatly impact on their virulence in oral biofilm formation in dental plaque as well as their ultimate pathogenesis of streptococcal infections in host body. Furthermore, I have also successfully developed StreptoBase, a database housing all 27 *Streptococcus* genomes along with the other 77 existing genomes of the 11 species of Mitis group streptococci in order to facilitate the ongoing research studies.

6.2 Comparative genomic analyses of two closely related *S. sanguinis* and *S. gordonii*

Fourteen strains of *S. gordonii* and five strains of *S. sanguinis* have been sequenced and

compared along with their reference complete genome strains of *S. gordonii* Challis and *S. sanguinis* SK36. The taxonomic position of each isolate was identified, supported by evidence from molecular phylogenetic analyses using the 16S rRNA single gene marker and the core-genome SNP approaches. A previous study has reported high level of 16S rRNA sequence homology between *S. gordonii* and *S. sanguinis* (Kilian, MIKKELSEN et al. 1989). The results of the present study revealed high orthologous gene similarity of *S. gordonii* and *S. sanguinis* with the later species harboring a higher number of *S. sanguinis*-specific core genes. In the functional enrichment analysis, I found the *S. sanguinis*-specific core genes (e.g. *cob*, *cbi* and *nik* gene clusters) were enriched in nickel and cobalt utilization and cobalamin biosynthesis. These gene clusters support the energy utilization of *S. sanguinis* bacteria for greater adaptation to growth and survival within the human host (Khatri, Khatri et al. 2012). Moreover, the *efeOUB-tat* system discovered in the conserved prophage FSS4_1 shared across the *S. sanguinis* genomes is predicted to support bacterial iron uptake and protein transport, further conferring its virulence. Previous research has reported truncated genes of the Tat system found in the genome of *S. pneumoniae* as evidence of loss of potential virulence genes during the divergent evolution of oral streptococci species (Denapaite, Brückner et al. 2010). Since the Tat system was also detected in *S. sanguinis* SK 36, it is possible that the acquisition of the FSS4_1 prophage containing the *efeUOB-tat* operon by *S. sanguinis* might occur in early stage after the separation of *S. sanguinis* from *S. gordonii*.

In addition, the *S. sanguinis*-specific core genes were statistically enriched in the process of cell wall components is supported by the virulence gene analysis of *S. sanguinis*, where a series of RGPs synthesis associated genes were identified. In fact, both *S. gordonii* and *S. sanguinis* harbor sets of virulence-associated genes including *rps*, *rml* and *rgp* gene loci

which are responsible for streptococcal polysaccharide biosynthesis. These homologous genes which participate in housekeeping functions such as polysacccharide synthesis, amino acid and nucleic acid synthesis as well as the bacterial survival in anaerobic conditions have been proven to be essential virulence factors for *S. sanguinis*-associated infective endocarditis (Paik, Senty et al. 2005).

The results of this study clearly showed that both *S. gordonii* and *S. sanguinis* have open pan-genomes, reflecting the genome variation of these pathogens as part of the evolutionary process over the time. I have also observed in my analysis that these two *Streptococcus* species have acquired new putative genes that potentially enhance pathogenesis via lateral gene transfer elements of prophages and GIs. These results might reflect the ability of both *Streptococcus* species to adapt to divergent and harsh host environmental conditions, For instance, *S. gordonii* and *S. sanguinis* might have acquired capability to generate energy and create an acidic environment for the bacteria in the host through the acquisition of the V-type ATPase in GI_55 (as was shown in the GI analysis) (Tesorero, Yu et al. 2013). On the other hand, I discovered another putative conserved GI_16 which brings together in physical proximity of the *yqeK*, *nadD* and ybeB genes. A relevant functional association between *yqeK*, *nadD* and *ybeB* genes is suggested, which would be interesting for future research. Noticeably, the importance of repairing mutated proteins due to reactive oxygen species (ROS) is indicated by the insertion of horizontally transferred *DegP/HtrA* gene and mutT/nudix family proteins by *S. gordonii* and *S. sanguinis,* respectively.

Evidence of horizontal gene transfer in *S. sanguinis* and *S. gordonii* is strongly indicated by two distinct groups of core loci for streptococcal polysaccharide production which are shared across both *S. gordonii* and *S. sanguinis* strains. This indicates that both species are

constantly exchanging genetic material to support their adaptation, survival and evolution. Since the *S. gordonii* Challis-type polysaccharide gene cluster structure is so widely conserved in both *S. gordonii* and *S. sanguinis* strains, I propose that this is the ancestral gene cluster in *S. gordonii* and *S. sanguinis* strains. Presumably, the *S. gordonii* 38-type gene cluster arrangement has arisen at least twice by horizontal gene transfer since it is present in at least one strain of both *S. gordonii* and *S. sanguinis*, although it was not observed in any *S. sanguinis* strains analyzed here. It is notable that the strains harboring *S. gordonii* 38-type *rps* gene loci did not cluster together by either 16S rRNA or whole genome SNP analysis (Figure 4.1). Nevertheless, this does not exclude the possibility that these strains might have diverged from a common ancestor after acquiring the *S. gordonii* 38-type *rps* locus.

Competence is an essential virulence determinant in a majority of the pathogenic streptococci (Li, Tian et al. 2008) and additional copies of *comCDE* quorum-sensing system components have been determined in three GIs of *S. gordonii*. Based on the comparative GI analysis, I found that *S. gordonii* has been well-equipped with the *ComCDE* quorum-sensing system as competence mechanism through the acquisition of putative genomic islands (GI_45, GI_51 and GI_58) likely through horizontal gene transfer over the evolutionary time. Apart from its role in oral biofilm, the *ComCDE* is potentially important for several functions such as increasing genome plasticity via uptake of new genes (Claverys, Prudhomme et al. 2000), DNA repair (Prudhomme, Attaiech et al. 2006), as well as providing nutrition of carbon, nitrogen, phosphorus, and energy source for *S. gordonii* (Finkel and Kolter 2001). Based on the comparative genome study results, this *comCDE* could possibly be a significant virulence factor for *S. gordonii* to enhance its pathogenesis in biofilm development and eventually streptococcal infective endocarditis.

However, I observed the potential of *S. gordonii* to facilitate intergeneric DNA transfer from *E. faecalis* with the insertion of GI_67 that may confer fluoroquinolone resistance which has recently emerged among enterococci (Vickerman, Flannagan et al. 2010). This genetic transfer is further affirmed by the discovery of *cylA* and *cylB* genes in *S. gordonii*, which presumably have originated from the α-hemolytic *Enterococcus faecalis* (previously known as Group D Streptococci). In *β*-hemolytic *S. agalactiae*, the *cyl* genes are responsible for hemolysin synthesis (Spellerberg, Martin et al. 2000). To the best of my knowledge, this study provides the first reported identification of *cyl* genes in α-hemolytic *S. gordonii* which may have contribute to antibiotic resistance. I suggest that these *cyl* genes were part of an ancestral form of the Gram positive cocci which was eventually lost in majority of the oral streptococci. In short, the ability of intergeneric gene acquisition by *S. gordonii* in oral biofilms might give rise to enhanced antibiotic resistance and virulence of *S. gordonii* in the foreseeable future.

It is noteworthy that *S. gordonii* is known to be the most effective competitor of *S. sanguinis* compared to other species of oral streptococci for adherence to saliva-coated hydroxyapatite (Kreth, Merritt et al. 2009). Apart from sharing the same colonization sites in the human oral cavity, these two genetically similar pioneer colonizing species often exhibit interspecies antagonism for similar host-derived salivary receptors (Nobbs, Zhang et al. 2007). Since the etiology of oral diseases such as dental caries, involves a primary step of bacterial colonization and adherence in the developing biofilm (Busscher and Van der Mei 1997), the ability of *S. gordonii* to compete with *S. sanguinis* in the initial adhesion to saliva pellicle may play an important role in determining the development of dental plaque-related diseases (Kreth, Merritt et al. 2009).

The *S. sanguinis*-specific virulence gene, *iga* which is important for adhesion and

colonization of tooth surfaces is a key potential virulence factor that distinguishes *S. sanguinis* from *S. gordonii*. This is because oral streptococci tend to bind selectively to the acquired enamel pellicle and then become coated with salivary proteins such as α-amylase and secretory IgA in dental plaque (Stevens and Kaplan 2000). Therefore, *S. gordonii* may be disadvantaged from this IgA binding specificity which favors *S. sanguinis* if α-amylases are denatured. Furthermore, IgA agglutinates *S. sanguinis* which may reduce its ability to adhere to oral surfaces. In addition, IgA has been proved to be a strong competitor over other immunoglobulin isotypes as the binding of IgA-Fab fragments blocked access of other immunoglobulin isotypes to mediate host-effector functions against *S. sanguinis* (Russell, Reinholdt et al. 1989). Overall, the production of IgA protease may be a key factor that enables *S. sanguinis* to colonise and grow to higher cell densities than *S. gordonii* in dental plaque and saliva (Kreth, Merritt et al. 2009).

Through genes on a horizontally transferred GI, *S. sanguinis* has also recruited a putative BfrABss system encoding CAAX proteases which typically contribute to biofilm formation by oral streptococci. In addition, I ascertained that drug or antimicrobial resistance is potentially conferred by GI's in *S. sanguinis* as this oral-biofilm pathogen acquired a series of antibiotic resistance genes including drug/metabolite transporter (DMT) superfamily, *rsmE,* TetR/AcrR family transcriptional regulator (TFR) and GNAT acetyltransferase through lateral gene transfer GIs. The GNAT acetyltransferase is responsible for aminoglycoside antibiotic resistance; the TFRs account for the tetracycline antibiotic resistance while *rsmE* confers resistance via methylation of DNA, RNA, proteins or small molecules. The DMT superfamily supports a variety of antibiotic resistance phenotypes through its different phylogenetic families such as the drug/metabolite exporter (DME) family, the 4 TMS Small Multidrug Resistance (SMR) family, the Glucose/Ribose Porter

(GRP) family and others. Since *S. sanguinis* exploits multiple antibiotic resistance over a wide array of drugs, it has been reported that combined treatment may be required to battle this opportunistic pathogen (Martinez, Martin-Luengo et al. 1995). Overall, this comparative genome analysis may provide better insights into how *S. sanguinis* generally achieves greater cell densities in oral biofilms than *S. gordonii* in respect to their co-existence within dental plaque. I have also identified a series of potential virulence genes, essential metabolism-related genes as well as horizontally acquired genes from GIs and prophages in both species. In fact, most epidemiological studies indicate that *S. gordonii* and *S. sanguinis* are very similar in their pathogenicity where both species exhibit equivalent virulence (Kumar, van der Linden et al. 2014). .

6.3 Development of StreptoBase and bioinformatics tools

With advances in NGS technology, further *Streptococcus* Mitis group species or strains will be sequenced and this creates an urgent need to store, browse, retrieve and analyze vast amounts of genome data and the development of specialized tools for comparative analyses of these genomes. I have successfully described the functionalities of StreptoBase particularly the in-house designed bioinformatics pipelines for the analyses of the genomic data of 11 species of *Streptococcus* Mitis group including *S. australis, S. cristatus, S. gordonii, S. infantis, S. mitis, S. oligofermentans, S. oralis, S. parasanguinis, S. peroris, S. sanguinis*, and *S. tigurinus*. This specialized biological database will be constantly updated in order to provide the latest genome updates and research developments associated with the *Streptococcus* Mitis group, and to ensure the accuracy and usefulness of the *Streptococcus* Mitis group species genome data and annotation.

The Perl scripts goverened-PGC tool allows users to perform and visualize cross species or same species pairwise genome comparison of two strains of *Streptococcus* Mitis group

species. The existing Microbial Genome Comparison (MGC) tool utilizes an *in silico* genome subtraction method to identify genetic elements specific to a group of strains (Argimón, Konganti et al. 2014). While PGC tool uses genome files and NUCmer to perform pairwise genome alignment, the MGC tool uses *in silico* fragmented genome sequences and performs BLASTN on groups of queries. On the contrary, the VISTA Browser which is well-known for its biological application is able to perform pre-computed pairwise and multiple whole-genome alignments using both global and local alignments (Frazer, Pachter et al. 2004). In contrast to circular plots and histograms that are generated by the PGC tool, the alignment results generated by VISTA Browser are displayed using VISTA track in graph plot format to show conserved regions. Additionally, the open source Java-based Artemis Comparison Tool (ACT) requires users to generate a comparison file which identifies homology regions between assembly and reference genome using programs such as BLASTN, TBLASTX or Mummer to be loaded on ACT (Carver, Rutherford et al. 2005). The comparative ACT visualization is performed using Artemis components. By contrast, the PGC tool on StreptoBase enables a single-flow process of pairwise genome alignment and instant display of the comparative alignment Circos plot.

Moroever, PathoProT enables rapid and effective prediction of putative virulence genes across different species of *Streptococcus* Mitis group based on the protein sequences of the oral streptococci. In conjunction with the existing *Streptococcus pneumoniae* Genome Database (SPGDB) (Swetha, Sekar et al. 2014), I anticipate that StreptoBase will serve as a useful resource and analysis platform particularly for comparative analyses of the *Streptococcus* Mitis group genomes for research communities.

6.4 Limitations

In this study, a limited number of genomes were used due to the availability of bacterial strains. The analysis of additional strains would be required to provide a more comprehensive comparative analysis of different species of Mitis group oral streptococci. Moreover, this study was not designed or powered to make associations with geographical origin or disease. The 27 Mitis group oral streptococci were collected mainly in Europe and Australia apart from the other continents around the world. Previous report has linked the association of substantial variation in different geographical areas in the United States to the rate of increase in antibiotic resistance of *S. pneumoniae* **(**McCormick, Whitney et al. 2003). Therefore, it will be crucial to include a greater geographical regional coverage of Mitis group oral streptococci isolates in future comparative genomic studies of *Streptococcus* genus.

Furthermore, the relevance of the assignation of genes as virulence genes and whether strains of these species have evolved to cause endocarditis or if it was a pure chance occurrence could possibly have been misguided via the method of assigning a biological function (virulence) on the basis of gene homology in VFDB. Such complications could be further exacerbated by circumstances in which homologous genes from different species carry different molecular functions, as well as misannotations of genes in many large public databases (Schnoes, Brown et al. 2009). Hence, the identified potential virulence factors from the comparative pathogenomics analysis in the present study might not have the capacity to cause streptococcal infections, but they represent possible candidates for further Mitis group oral streptococci studies.

6.5 Future Work

Since the importance of RGPs and Ni/Co/cobalamin system which are enriched in the core genes of *S. sanguinis* has been proven in several studies, an important future milestone would be to understand the biological roles of these genes via mutational analysis. Notably, the *cylA* and *cylB* discovered in *S. gordonii* do not harbor the complete functional cytolysin machinery as those in *S. agalactiae*, supported by the fact that *S. gordonii* is non-cytolytic. I propose a hypothesis that these *cyl* genes might have evolved a slightly different function in the import or export of some biological product. Validation of this hypothesis will require experimental verification, for example via gene knockout techniques and insertion of the knockout strains into an animal model of IE. A similar biological approach is required to examine the relevance of the assignation of genes as virulence genes by comparing the mutant with the wild-type in some relevant animal models of disease. In future studies, it would be beneficial to expand the pan-genome, GI and prophage analyses by involving other *Streptococus* Mitis group species. Significantly, it will be worth looking into the *rps* locus structure in the genomes of the rest of *Streptococus* Mitis group species apart from *S. gordonii* and *S. sanguinis*.

Furthermore, additional bioinformatics analysis such as toxin-antitoxin system analysis and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas system analysis on the genomes of *Streptococcus* Mitis group species can be included in the future work. The type II chromosomal toxin-antitoxin systems (TAS) has been reported to be involved in stress response which mediates growth by decreasing the biosynthesis of macromolecular (Gerdes, Christensen et al. 2005). It has also been linked to bacterial programmed cell death (Kolodkin-Gal and Engelberg-Kulka 2006). Therefore, prediction of the TAS in *Streptococcus* genomes might contributes to development and design of a

novel class of antibiotics by targeting these predicted TAS. On the other hand, CRISPR-Cas system serves as microbial defence system via its adaptive sequence-specific immunity in bacteria (van der Oost, Westra et al. 2014). Recent studies have revealed the existence of CRISPR-Cas system in *S. thermophilus* and *S. pyogenes* (Ferretti, Stevens et al. 2016). Determination of the CRISPR-Cas system in *S. gordonii* and *S. sanguinis* may lead to better understanding in the bacterial specific immunity against undesirable foreign genes and possibly the viral resistance of the oral streptococci (Horvath and Barrangou 2010).

In terms of improving the StreptoBase, in-house designed bioinformatics tools such as, a Multiple Sequence Alignment (MSA) visualization tool, toxin-antitoxin system prediction tool, phylogenetic inference analysis tool and prophage prediction tool to enhance the efficiency and functionality of this *Streptococcus* Mitis group biological database would enable future studies in this field. For example, a phylogenetic tree generation tool can be incorporated on StreptoBase which allows users to select several Mitis group oral streptococci genomes in StreptoBase by using a series of widely used *Streptococcus* genus marker genes including *gdh*, *GspB, Hsa*, *ddl*, *rpoB* and *sodA* to draw a phylogenetic tree. It would be interesting to include some novel *Streptococcus* bacterial strains isolated in Malaysia future studies along with the need for substantial sequencing effort for the Mitis group oral streptococci. These novelgenome sequences and more new publicly available genomescould then be added into this database in order to support the expanding Mitis group oral streptococci research worldwide. Since many of the Mitis group oral streptococci harbor open pan-genomes, new insights on a greater genetic diversity of the *Streptococcus* species can be revealed.

CHAPTER 7: CONCLUSION

In conclusion, I have successfully sequenced and assembled the genomes of two closely related oral streptococcal species, *S. gordonii* and *S. sanguinis*. I have also presented a comparative genome analysis of clinically-derived mitis group oral streptococci species, particularly on *S. sanguinis* and *S. gordonii*. Significantly, this study provides better insights into the differing ecological strategies of *S. gordonii* and *S. sanguinis*. Both species are common within dental plaque and both have the potential to cause infective endocarditis. However, *S. sanguinis* is usually present in higher numbers than *S. gordonii*, and differing associations between these species and oral disease have been shown. Functions such as cobalamin biosynthesis, IgA protease activity and CAAX proteases may contribute to the expansion of *S. sanguinis* within dental plaque. On the other hand, the presence of *cylA* and *cylB* within the core genome of *S. gordonii* is interesting and warrants further studies. There are no genes that are clearly enriched in endocarditis isolates, and this is in keeping with the observation that oral and endocarditis isolates of *S. sanguinis* do not form distinct subclones (Do, Gilbert et al. 2011). My data clearly showed that both *S. gordonii* and *S. sanguinis* have open pan-genomes, proposing that they may continue to evolve and acquire new genes in future. Potentially, the exchange of genetic information between bacteria in biofilms may accelerate the spread of antibiotic resistance between bacteria in the oral cavity. Overall, the comparative analyses of *S. gordonii* and *S. sanguinis* will provide a basis for understanding how these species establish within dental plaque and how they transition from commensal species within the mouth to important pathogens in infective endocarditis.

Furthermore, I have successfully developed the first mitis group oral streptococci genomic resource and analysis platform database, StreptoBase which provides free and direct access

of comprehensive *Streptococcus* genomes and information, as well as the new in-house designed analysis tools particularly for comparative analyses, is believed to be an invaluable resource to accelerate and support the expanding *Streptococcus* genus research worldwide.

REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." Journal of molecular biology **215**(3): 403-410.

Andrews, S. (2011). FastQC—a quality control tool for high throughput sequence data. Babraham Bioinformatics.

Argimón, S., K. Konganti, H. Chen, A. V. Alekseyenko, S. Brown and P. W. Caufield (2014). "Comparative genomics of oral isolates of *Streptococcus mutans* by *in silico* genome subtraction does not reveal accessory DNA associated with severe early childhood caries." Infection, Genetics and Evolution **21**: 269-278.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass and M. Kubal (2008). "The RAST Server: rapid annotations using subsystems technology." BMC genomics **9**(1): 75.

Baek, J. H., S. V. Rajagopala and D. K. Chattoraj (2014). "Chromosome segregation proteins of *Vibrio cholerae* as transcription regulators." MBio **5**(3): e01061-01014.

Bancescu, G., S. Dumitriu, A. Bancescu, C. Defta, M. Pana, D. Ionescu, S. Alecu and M. Zamfirescu (2004). "Susceptibility testing of *Streptococcus* mitis group isolates." Indian journal of medical research **119**: 257-261.

Banerjee, R. (1999). Chemistry and Biochemistry of B12, John Wiley & Sons.

Barnard, J. P. and M. W. Stinson (1996). "The alpha-hemolysin of *Streptococcus gordonii* is hydrogen peroxide." Infect Immun **64**(9): 3853-3857.

Bentley, R. W., J. A. Leigh and M. D. Collins (1991). "Intrageneric structure of *Streptococcus* based on comparative analysis of small-subunit rRNA sequences." International journal of systematic bacteriology **41**(4): 487-494.

Bentley, S. D., D. M. Aanensen, A. Mavroidi, D. Saunders, E. Rabbinowitsch, M. Collins, K. Donohoe, D. Harris, L. Murphy and M. A. Quail (2006). "Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes." PLoS Genet **2**(3): e31.

Bernhardt, T. G. and P. A. De Boer (2004). "Screening for synthetic lethal mutants in *Escherichia coli* and identification of EnvC (YibP) as a periplasmic septal ring factor with murein hydrolase activity." Molecular microbiology **52**(5): 1255-1269.

Bessman, M. J., D. N. Frick and S. F. O'Handley (1996). "The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed,"housecleaning" enzymes." Journal of Biological Chemistry **271**(41): 25059-25062.

Bolotin, A., B. Quinquis, P. Renault, A. Sorokin, S. D. Ehrlich, S. Kulakauskas, A. Lapidus, E. Goltsman, M. Mazur and G. D. Pusch (2004). "Complete sequence and

comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*." Nature biotechnology **22**(12): 1554-1558.

Bose, M. and R. D. Barber (2006). "Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences." *In silico* biology **6**(3): 223-227.

Branda, S. S., J. E. González-Pastor, E. Dervyn, S. D. Ehrlich, R. Losick and R. Kolter (2004). "Genes involved in formation of structured multicellular communities by *Bacillus subtilis*." Journal of bacteriology **186**(12): 3970-3979.

Busscher, H. and H. Van der Mei (1997). "Physico-chemical interactions in initial microbial adhesion and relevance for biofilm formation." Advances in dental research **11**(1): 24-32.

Carrolo, M., M. J. Frias, F. R. Pinto, J. Melo-Cristino and M. Ramirez (2010). "Prophage spontaneous activation promotes DNA release enhancing biofilm formation in *Streptococcus pneumoniae*." PLoS One **5**(12): e15678.

Carver, T. J., K. M. Rutherford, M. Berriman, M.-A. Rajandream, B. G. Barrell and J. Parkhill (2005). "ACT: the Artemis comparison tool." Bioinformatics **21**(16): 3422-3423.

Caspi, R., T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari and A. Kubo (2014). "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases." Nucleic acids research **42**(D1): D459-D471.

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 research **33**(suppl 1): D325-D328.

Chen, W. Y., E. Marcellin, J. Hung and L. K. Nielsen (2009). "Hyaluronan molecular weight is controlled by UDP-N-acetylglucosamine concentration in *Streptococcus zooepidemicus*." Journal of Biological Chemistry **284**(27): 18007-18014.

Chen, Y.-Y. M. and R. A. Burne (2003). "Identification and characterization of the nickel uptake system for urease biogenesis in *Streptococcus salivarius* 57. I." Journal of bacteriology **185**(23): 6773-6779.

Cheng, Q., E. Campbell, A. Naughton, S. Johnson and H. Masure (1997). "The com locus controls genetic transformation in *Streptococcus pneumoniae*." Molecular microbiology **23**(4): 683-692.

Chia, J.-S., Y.-L. Lin, H.-T. Lien and J.-Y. Chen (2004). "Platelet aggregation induced by serotype polysaccharides from *Streptococcus mutans*." Infection and immunity **72**(5): 2605-2617.

Claverys, J. P., M. Prudhomme, I. Mortier ‐ Barrière and B. Martin (2000). "Adaptation to the environment: *Streptococcus pneumoniae*, a paradigm for recombination ‐ mediated genetic plasticity?" Molecular microbiology **35**(2): 251-259.

Cleary, P. P. and A. Larkin (1979). "Hyaluronic acid capsule: strategy for oxygen resistance in group A streptococci." Journal of Bacteriology **140**(3): 1090-1097.

Coleman, M. L., M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. DeLong and S. W. Chisholm (2006). "Genomic islands and the ecology and evolution of *Prochlorococcus*." Science **311**(5768): 1768-1770.

Conesa, A. and S. Götz (2008). "Blast2GO: A comprehensive suite for functional analysis in plant genomics." International journal of plant genomics **2008**.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón and M. Robles (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Cornelis, P. and S. C. Andrews (2010). Iron uptake and homeostasis in microorganisms, Horizon Scientific Press.

Crater, D. L. and I. Van de Rijn (1995). "Hyaluronic acid synthesis operon (has) expression in group A streptococci." Journal of Biological Chemistry **270**(31): 18452-18458.

Croxatto, A., V. J. Chalker, J. Lauritz, J. Jass, A. Hardman, P. Williams, M. Cámara and D. L. Milton (2002). "*VanT*, a homologue of *Vibrio harveyi LuxR*, regulates serine, metalloprotease, pigment, and biofilm production in *Vibrio anguillarum*." Journal of Bacteriology **184**(6): 1617-1629.

Cuadros-Orellana, S., A.-B. Martin-Cuadrado, B. Legault, G. D'Auria, O. Zhaxybayeva, R. T. Papke and F. Rodriguez-Valera (2007). "Genomic plasticity in prokaryotes: the case of the square haloarchaeon." The ISME journal **1**(3): 235-245.

Delcher, A. L., A. Phillippy, J. Carlton and S. L. Salzberg (2002). "Fast algorithms for large-scale genome alignment and comparison." Nucleic acids research **30**(11): 2478-2483.

Denapaite, D., R. Brückner, M. Nuhn, P. Reichmann, B. Henrich, P. Maurer, Y. Schähle, P. Selbmann, W. Zimmermann and R. Wambutt (2010). "The genome of *Streptococcus mitis* B6-what is a commensal?" PLoS One **5**(2): e9426.

Dmitriev, A., A. Shen, Ľ. Tkáčiková, I. Mikula and Y. Yang (2004). "Structure of scpB-lmb intergenic region as criterion for additional classification of human and bovine group B Streptococci." Acta Veterinaria Brno **73**(2): 215-220.

Do, T., S. Gilbert, J. Klein, S. Warren, W. Wade and D. Beighton (2011). "Clonal structure of *Streptococcus sanguinis* strains isolated from endocarditis cases and the oral cavity." Molecular oral microbiology **26**(5): 291-302.

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

Donati, C., N. L. Hiller, H. Tettelin, A. Muzzi, N. J. Croucher, S. V. Angiuoli, M. Oggioni, J. C. D. Hotopp, F. Z. Hu and D. R. Riley (2010). "Structure and dynamics of the pan-genome of S*treptococcus pneumoniae* and closely related species." Genome biology **11**(10): R107.

Dunne Jr, W., L. Westblade and B. Ford (2012). "Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory." European journal of clinical microbiology & infectious diseases **31**(8): 1719-1726.

Durack, D. and P. Beeson (1972). "Experimental bacterial endocarditis: I. Colonization of a sterile vegetation." British journal of experimental pathology **53**(1): 44.

Durack, D. T. and P. B. Beeson (1972). "Experimental bacterial endocarditis: II. Survival of bacteria in endocardial vegetations." British journal of experimental pathology **53**(1): 50.

Elsner, A., B. Kreikemeyer, A. Braun-Kiewnick, B. Spellerberg, B. A. Buttaro and A. Podbielski (2002). "Involvement of Lsp, a member of the LraI-lipoprotein family in *Streptococcus pyogenes*, in eukaryotic cell adhesion and internalization." Infection and immunity **70**(9): 4859-4869.

Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic acids research **30**(7): 1575-1584.

Facklam, R. (2002). "What happened to the streptococci: overview of taxonomic and nomenclature changes." Clinical microbiology reviews **15**(4): 613-630.

Feng, Y. and J. E. Cronan (2011). "Complex binding of the FabR repressor of bacterial unsaturated fatty acid biosynthesis to its cognate promoters." Molecular microbiology **80**(1): 195-218.

Fernández-Gómez, B., A. Fernàndez-Guerra, E. O. Casamayor, J. M. González, C. Pedrós-Alió and S. G. Acinas (2012). "Patterns and architecture of genomic islands in marine bacteria." BMC genomics **13**(1): 347.

Ferretti, J., D. Stevens and V. Fischetti (2016). "The CRISPR-Cas system of *Streptococcus pyogenes*: function and applications--*Streptococcus pyogenes*: Basic Biology to Clinical Manifestations."

Finkel, S. E. and R. Kolter (2001). "DNA as a nutrient: novel role for bacterial competence gene homologs." Journal of Bacteriology **183**(21): 6288-6293.

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

Fowler, V. G., J. M. Miro, B. Hoen, C. H. Cabell, E. Abrutyn, E. Rubinstein, G. R. Corey, D. Spelman, S. F. Bradley and B. Barsic (2005). "*Staphylococcus aureus* endocarditis: a consequence of medical progress." Jama **293**(24): 3012-3021.

Frazer, K. A., L. Pachter, A. Poliakov, E. M. Rubin and I. Dubchak (2004). "VISTA: computational tools for comparative genomics." Nucleic acids research **32**(suppl 2): W273-W279.

Fullwood, M. J., C.-L. Wei, E. T. Liu and Y. Ruan (2009). "Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses." Genome research **19**(4): 521-532.

Gerdes, K., S. K. Christensen and A. Løbner-Olesen (2005). "Prokaryotic toxin–antitoxin stress response loci." Nature Reviews Microbiology **3**(5): 371-382.

Gilbert, J. V., A. G. Plaut and A. Wright (1991). "Analysis of the immunoglobulin A protease gene of *Streptococcus sanguis*." Infection and immunity **59**(1): 7-17.

Giomarelli, B., L. Visai, K. Hijazi, S. Rindi, M. Ponzio, F. Iannelli, P. Speziale and G. Pozzi (2006). "Binding of *Streptococcus gordonii* to extracellular matrix proteins." FEMS microbiology letters **265**(2): 172-177.

Gottschalk, B., G. Bröker, M. Kuhn, S. Aymanns, U. Gleich-Theurer and B. Spellerberg (2006). "Transport of multidrug resistance substrates by the *Streptococcus agalactiae* hemolysin transporter." Journal of bacteriology **188**(16): 5984-5992.

Hacker, J. and E. Carniel (2001). "Ecological fitness, genomic islands and bacterial pathogenicity." EMBO reports **2**(5): 376-381.

Hacker, J. and J. B. Kaper (2000). "Pathogenicity islands and the evolution of microbes." Annual Reviews in Microbiology **54**(1): 641-679.

Hall‐Stoodley, L., P. Stoodley, S. Kathju, N. Høiby, C. Moser, J. William Costerton, A. Moter and T. Bjarnsholt (2012). "Towards diagnostic guidelines for biofilm‐associated infections." FEMS Immunology & Medical Microbiology **65**(2): 127-145.

Hasman, H., D. Saputra, T. Sicheritz-Ponten, O. Lund, C. A. Svendsen, N. Frimodt-Møller and F. M. Aarestrup (2013). "Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples." Journal of clinical microbiology: JCM. 02452-02413.

Henningham, A., S. Döhrmann, V. Nizet and J. N. Cole (2015). "Mechanisms of group A *Streptococcus* resistance to reactive oxygen species." FEMS microbiology reviews: fuu009.

Hermans, P. W., P. V. Adrian, C. Albert, S. Estevão, T. Hoogenboezem, I. H. Luijendijk, T. Kamphausen and S. Hammerschmidt (2006). "The streptococcal lipoprotein rotamase A (SlrA) is a functional peptidyl-prolyl isomerase involved in pneumococcal colonization." Journal of Biological Chemistry **281**(2): 968-976.

Holden, M., H. Hauser, M. Sanders, T. H. Ngo, I. Cherevach, A. Cronin, I. Goodhead, K. Mungall, M. A. Quail and C. Price (2009). "Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*." PLoS One **4**(7): e6072.

Horvath, P. and R. Barrangou (2010). "CRISPR/Cas, the immune system of bacteria and archaea." Science **327**(5962): 167-170.

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.

Human Microbiome Project, C. (2012). "Structure, function and diversity of the healthy human microbiome." Nature **486**(7402): 207-214.

Jack, A. A., D. E. Daniels, M. A. Jepson, M. M. Vickerman, R. J. Lamont, H. F. Jenkinson and A. H. Nobbs (2015). "*Streptococcus gordonii* comCDE (competence) operon modulates biofilm formation with Candida albicans." Microbiology **161**(2): 411-421.

Jakubovics, N. S., S. A. Yassin and A. H. Rickard (2014). "Community interactions of oral streptococci." Adv Appl Microbiol **87**: 43-110.

Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis and T. L. Madden (2008). "NCBI BLAST: a better web interface." Nucleic acids research **36**(suppl 2): W5-W9.

Jones, C. H., B. Tove'C, K. F. Jones, G. O. Zeller and D. E. Hruby (2001). "Conserved DegP Protease in Gram-Positive Bacteria Is Essential for Thermal and Oxidative Tolerance and Full Virulence in *Streptococcus pyogenes*." Infection and immunity **69**(9): 5538-5545.

Kamio, N., K. Imai, K. Shimizu, M. E. Cueno, M. Tamura, Y. Saito and K. Ochiai (2015). "Neuraminidase-producing oral mitis group streptococci potentially contribute to influenza viral infection and reduction in antiviral efficacy of zanamivir." Cellular and Molecular Life Sciences **72**(2): 357-366.

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*." International journal of systematic bacteriology **45**(2): 406-408.

Kawamura, Y., R. A. Whiley, S.-E. Shu, T. Ezaki and J. M. Hardie (1999). "Genetic approaches to the identification of the mitis group within the genus *Streptococcus*." Microbiology **145**(9): 2605-2613.

Keane, C. (2010). "The molecular mechanisms of *streptococcus gordonii*-platelet interactions involved in the pathogenesis of cardiovascular infection."

Keen, E. C. (2012). "Paradigms of pathogenesis: targeting the mobile genetic elements of disease." Frontiers in cellular and infection microbiology **2**.

Kerrigan, S. W. and D. Cox (2012). "Platelet-Bacterial Interactions as Therapeutic Targets in Infective Endocarditis."

Khatri, N., I. Khatri, S. Subramanian and S. Raychaudhuri (2012). "Ethanolamine utilization in *Vibrio alginolyticus*." Biology direct **7**(1): 1.

Kilian, M., L. Mikkelsen and J. Henrichsen (1989). "Taxonomic study of viridans streptococci: description of *Streptococcus gordonii* sp. nov. and emended descriptions of *Streptococcus sanguis* (White and Niven 1946), *Streptococcus oralis* (Bridge and Sneath 1982), and *Streptococcus mitis* (Andrewes and Horder 1906)." International Journal of Systematic Bacteriology **39**(4): 471-484.

Kim, D.-Y. and K.-K. Kim (2005). "Structure and function of *HtrA* family proteins, the key players in protein quality control." BMB Reports **38**(3): 266-274.

Kim, S.-J., S.-Y. Park and C.-W. Kim (2006). "A novel approach to the production of hyaluronic acid by *Streptococcus zooepidemicus*." Journal of microbiology and biotechnology **16**(12): 1849-1855.

Kolodkin-Gal, I. and H. Engelberg-Kulka (2006). "Induction of *Escherichia coli* chromosomal *mazEF* by stressful conditions causes an irreversible loss of viability." Journal of bacteriology **188**(9): 3420-3423.

Kreth, J., J. Merritt and F. Qi (2009). "Bacterial and host interactions of oral streptococci." DNA and cell biology **28**(8): 397-403.

Kreth, J., Y. Zhang and M. C. Herzberg (2008). "Streptococcal antagonism in oral biofilms: *Streptococcus sanguinis* and *Streptococcus gordonii* interference with *Streptococcus mutans*." Journal of bacteriology **190**(13): 4632-4640.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones and M. A. Marra (2009). "Circos: an information aesthetic for comparative genomics." Genome research **19**(9): 1639-1645.

Kumar, V. N., M. van der Linden, T. Menon and D. P. Nitsche-Schmitz (2014). "Viridans and bovis group streptococci that cause infective endocarditis in two regions with contrasting epidemiology." International Journal of Medical Microbiology **304**(3): 262-268.

Kurland, C., B. Canback and O. G. Berg (2003). "Horizontal gene transfer: a critical view." Proceedings of the National Academy of Sciences **100**(17): 9658-9662.

Langille, M. G. and F. S. Brinkman (2009). "IslandViewer: an integrated interface for computational identification and visualization of genomic islands." Bioinformatics **25**(5): 664-665.

Langille, M. G., W. W. Hsiao and F. S. Brinkman (2008). "Evaluation of genomic island predictors using a comparative genomics approach." BMC bioinformatics **9**(1): 329.

Langille, M. G., W. W. Hsiao and F. S. Brinkman (2010). "Detecting genomic islands using bioinformatics approaches." Nature Reviews Microbiology **8**(5): 373-382.

Lee, P. A., D. Tullman-Ercek and G. Georgiou (2006). "The bacterial twin-arginine translocation pathway." Annual review of microbiology **60**: 373.

Lemos, J. A., J. Abranches and R. A. Burne (2005). "Responses of cariogenic streptococci to environmental stresses." Current issues in molecular biology **7**(1): 95-108.

Li, Y.-H., N. Tang, M. B. Aspiras, P. C. Lau, J. H. Lee, R. P. Ellen and D. G. Cvitkovitch (2002). "A quorum-sensing signaling system essential for genetic competence in *Streptococcus mutans* is involved in biofilm formation." Journal of bacteriology **184**(10): 2699-2708.

Li, Y.-H., X.-L. Tian, G. Layton, C. Norgaard and G. Sisson (2008). "Additive attenuation of virulence and cariogenic potential of *Streptococcus mutans* by simultaneous inactivation of the *ComCDE* quorum-sensing system and HK/RR11 two-component regulatory system." Microbiology **154**(11): 3256-3265.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law (2012). "Comparison of next-generation sequencing systems." BioMed Research International **2012**.

Llull, D., R. López and E. García (2006). "Skl, a novel choline-binding N-acetylmuramoyl-L-alanine amidase of *Streptococcus mitis* SK137 containing a CHAP domain." FEBS letters **580**(8): 1959-1964.

MacEachran, D. P., B. A. Stanton and G. A. O'Toole (2008). "Cif is negatively regulated by the TetR family repressor CifR." Infection and immunity **76**(7): 3197-3206.

Markowitz, V. M., I.-M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang and P. Williams (2012). "IMG: the integrated microbial genomes database and comparative analysis system." Nucleic acids research **40**(D1): D115-D122.

Martin, V., A. L. Kleschyov, J.-P. Klein and A. Beretz (1997). "Induction of nitric oxide production by polyosides from the cell walls of *Streptococcus mutans* OMZ 175, a gram-positive bacterium, in the rat aorta." Infection and immunity **65**(6): 2074-2079.

Martinez, F., F. Martin-Luengo, A. Garcia and M. Valdes (1995). "Treatment with various antibiotics of experimental endocarditis caused by penicillin-resistant *Streptococcus sanguis*." European heart journal **16**(5): 687-691.

Matsui, R. and D. Cvitkovitch (2010). "Acid tolerance mechanisms utilized by *Streptococcus mutans*." Future microbiology **5**(3): 403-417.

Mavroidi, A., D. M. Aanensen, D. Godoy, I. C. Skovsted, M. S. Kaltoft, P. R. Reeves, S. D. Bentley and B. G. Spratt (2007). "Genetic relatedness of the *Streptococcus pneumoniae* capsular biosynthetic loci." Journal of bacteriology **189**(21): 7841-7855.

McAnally, J. and M. Levine (1993). "Bacteria reactive to plaque - toxin - neutralizing monoclonal antibodies are related to the severity of gingivitis at the sampled site." Oral microbiology and immunology **8**(2): 69-74.

McCormick, A. W., C. G. Whitney, M. M. Farley, R. Lynfield, L. H. Harrison, N. M. Bennett and M. H. Samore (2003). "Geographic diversity and temporal trends of antimicrobial resistance in *Streptococcus pneumoniae* in the United States." Nature medicine **9**(4), 424-430.

McNair, K., B. A. Bailey and R. A. Edwards (2012). "PHACTS, a computational approach to classifying the lifestyle of phages." Bioinformatics **28**(5): 614-618.

Mellmann, A., D. Harmsen, C. A. Cummings, E. B. Zentz, S. R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji and W. Zhang (2011). "Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104: H4 outbreak by rapid next generation sequencing technology." PloS one **6**(7): e22751.

Metzker, M. L. (2010). "Sequencing technologies—the next generation." Nature reviews genetics **11**(1): 31-46.

Minoche, A. E., J. C. Dohm and H. Himmelbauer (2011). "Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems." Genome biology **12**(11): 1.

Mitchell, S. (2014). "Zombies in bacterial genomes: identification and analysis of previously virulent phage."

Morić, I., M. Savić, T. Ilić-Tomić, S. Vojnović, S. Bajkić and B. Vasiljević (2010). "rRNA Methyltransferases and their role in resistance to antibiotics." Journal of Medical Biochemistry **29**(3): 165-174.

Moscoso, M., E. García and R. López (2006). "Biofilm formation by *Streptococcus pneumoniae*: role of choline, extracellular DNA, and capsular polysaccharide in microbial accretion." Journal of bacteriology **188**(22): 7785-7795.

Nancy, Y. Y., J. R. Wagner, M. R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S. C. Sahinalp, M. Ester and L. J. Foster (2010). "PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes." Bioinformatics **26**(13): 1608-1615.

Navarro, C., L. F. Wu and M. A. Mandrand - Berthelot (1993). "The nik operon of *Escherichia coli* encodes a periplasmic binding - protein - dependent transport system for nickel." Molecular microbiology **9**(6): 1181-1191.

Nobbs, A. H., Y. Zhang, A. Khammanivong and M. C. Herzberg (2007). "*Streptococcus gordonii Hsa* environmentally constrains competitive binding by *Streptococcus sanguinis* to saliva-coated hydroxyapatite." Journal of bacteriology **189**(8): 3106-3114.

Old, L., S. Lowes and R. Russell (2006). "Genomic variation in *Streptococcus mutans*: deletions affecting the multiple pathways of β - glucoside metabolism." Oral microbiology and immunology **21**(1): 21-27.

Oliveira, H., L. D. Melo, S. B. Santos, F. L. Nóbrega, E. C. Ferreira, N. Cerca, J. Azeredo and L. D. Kluskens (2013). "Molecular aspects and comparative genomics of bacteriophage endolysins." Journal of virology **87**(8): 4558-4570.

Paik, S., L. Senty, S. Das, J. C. Noe, C. L. Munro and T. Kitten (2005). "Identification of virulence determinants for endocarditis in *Streptococcus sanguinis* by signature-tagged mutagenesis." Infection and immunity **73**(9): 6064-6074.

Pal, M. (2005). "Random forest classifier for remote sensing classification." International Journal of Remote Sensing **26**(1): 217-222.

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

Parahitiyawa, N., L. Jin, W. Leung, W. Yam and L. Samaranayake (2009). "Microbiology of odontogenic bacteremia: beyond endocarditis." Clinical microbiology reviews **22**(1): 46-64.

Pei, J., D. A. Mitchell, J. E. Dixon and N. V. Grishin (2011). "Expansion of type II CAAX proteases reveals evolutionary origin of γ-secretase subunit APH-1." Journal of molecular biology **410**(1): 18-26.

Pompeani, A. J., J. J. Irgon, M. F. Berger, M. L. Bulyk, N. S. Wingreen and B. L. Bassler (2008). "The Vibrio harveyi master quorum - sensing regulator, LuxR, a TetR - type protein is both an activator and a repressor: DNA recognition and binding specificity at target promoters." Molecular microbiology **70**(1): 76-88.

Presterl, E., A. Grisold, S. Reichmann, A. Hirschl, A. Georgopoulos and W. Graninger (2005). "Viridans streptococci in endocarditis and neutropenic sepsis: biofilm formation and effects of antibiotics." Journal of Antimicrobial Chemotherapy **55**(1): 45-50.

Prudhomme, M., L. Attaiech, G. Sanchez, B. Martin and J.-P. Claverys (2006). "Antibiotic stress induces genetic transformability in the human pathogen *Streptococcus pneumoniae*." Science **313**(5783): 89-92.

Raux, E., A. Lanois, F. Levillayer, M. J. Warren, E. Brody, A. Rambach and C. Thermes (1996). "*Salmonella typhimurium* cobalamin (vitamin B12) biosynthetic genes: functional studies in *S. typhimurium* and *Escherichia coli*." Journal of bacteriology **178**(3): 753-767.

Rea, M. C., D. Alemayehu, R. P. Ross and C. Hill (2013). "Gut solutions to a gut problem: bacteriocins, probiotics and bacteriophage for control of *Clostridium difficile* infection." Journal of medical microbiology **62**(Pt 9): 1369-1378.

Reinholdt, J. and M. Kilian (1987). "Interference of IgA protease with the effect of secretory IgA on adherence of oral streptococci to saliva-coated hydroxyapatite." Journal of dental research **66**(2): 492-497.

Reinholdt, J., M. Tomana, S. Mortensen and M. Kilian (1990). "Molecular aspects of immunoglobulin A1 degradation by oral streptococci." Infection and immunity **58**(5): 1186-1194.

Relman, D. A., S. Falkow, W. A. Petri Jr, B. J. Mann, C. D. Huston, E. L. Hewlett, M. A. Hughes, C. W. Dieffenbach, E. C. Tramont and S. F. Plaeger (2000). "Mandell, Douglas and Bennett's principles and practice of infectious diseases." Complement **77**: 22.

Remm, M., C. E. Storm and E. L. Sonnhammer (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-1052.

Richards, V. P., S. R. Palmer, P. D. P. 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 biology and evolution **6**(4): 741-753.

Rodionov, D. A., P. Hebbeln, M. S. Gelfand and T. Eitinger (2006). "Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters." Journal of bacteriology **188**(1): 317-327.

Rosa-Fraile, M., S. Dramsi and B. Spellerberg (2014). "Group B streptococcal haemolysin and pigment, a tale of twins." FEMS microbiology reviews **38**(5): 932-946.

Ruggeri, Z. M. (2009). "Platelet adhesion under flow." Microcirculation **16**(1): 58-83.

Russell, M. W., J. Reinholdt and M. Kilian (1989). "Anti‐inflammatory activity of human *IgA* antibodies and their Fab α fragments: inhibition of *IgG*‐mediated complement activation." European journal of immunology **19**(12): 2243-2249.

Ryter, S. W. and R. M. Tyrrell (2000). "The heme synthesis and degradation pathways: role in oxidant sensitivity: heme oxygenase has both pro-and antioxidant properties." Free Radical Biology and Medicine **28**(2): 289-309.

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

Samuels, D. S. (2010). Borrelia: molecular biology, host interaction and pathogenesis, Horizon Scientific Press.

Sanchez, C. J., N. Kumar, A. Lizcano, P. Shivshankar, J. C. Dunning Hotopp, J. H. Jorgensen, H. Tettelin and C. J. Orihuela (2011). "*Streptococcus pneumoniae* in biofilms are unable to cause invasive disease due to altered virulence determinant production." PLoS One **6**(12): e28738.

Schierholz, J., J. Beuth and G. Pulverer (1999). "" Difficult to treat infections" pharmacokinetic and pharmacodynamic factors--a review." Acta microbiologica et immunologica Hungarica **47**(1): 1-8.

Schmidt, K.-H., E. Günther and H. S. Courtney (1996). "Expression of both M protein and hyaluronic acid capsule by group A streptococcal strains results in a high virulence for chicken embryos." Medical microbiology and immunology **184**(4): 169-173.

Schnoes, A. M., S. D. Brown, I. Dodevski and P. C. Babbitt (2009). "Annotation error in public databases: misannotation of molecular function in enzyme superfamilies." PLoS Comput Biol **5**(12): e1000605.

Shelburne, S. A., P. Sahasrabhojane, M. Saldana, H. Yao, X. Su, N. Horstmann, E. Thompson and A. R. Flores (2014). "*Streptococcus mitis* strains causing severe clinical disease in cancer patients." Emerging infectious diseases **20**(5): 762.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nature biotechnology **26**(10): 1135-1145.

Shibata, Y., Y. Yamashita, K. Ozaki, Y. Nakano and T. Koga (2002). "Expression and characterization of streptococcal rgp genes required for rhamnan synthesis in *Escherichia coli*." Infection and immunity **70**(6): 2891-2898.

Skinner, M. E., A. V. Uzilov, L. D. Stein, C. J. Mungall and I. H. Holmes (2009). "JBrowse: a next-generation genome browser." Genome research **19**(9): 1630-1638.

Soontharapirakkul, K. and A. Incharoensakdi (2010). "Na+-stimulated ATPase of alkaliphilic halotolerant cyanobacterium Aphanothece halophytica translocates Na+ into proteoliposomes via Na+ uniport mechanism." BMC biochemistry **11**(1): 30.

Spellerberg, B., S. Martin, C. Brandt and R. Lütticken (2000). "The *cyl* genes of *Streptococcus agalactiae* are involved in the production of pigment." FEMS microbiology letters **188**(2): 125-128.

Spellerberg, B., B. Pohl, G. Haase, S. Martin, J. Weber-Heynemann and R. Lütticken (1999). "Identification of genetic determinants for the hemolytic activity of *Streptococcus agalactiae* by ISS1Transposition." Journal of bacteriology **181**(10): 3212-3219.

Spellerberg, B., E. Rozdzinski, S. Martin, J. Weber-Heynemann, N. Schnitzler, R. Lütticken and A. Podbielski (1999). "*Lmb*, a protein with similarities to the LraI adhesin family, mediates attachment of *Streptococcus agalactiae* to human laminin." Infection and immunity **67**(2): 871-878.

Stackebrandt, E. and B. Goebel (1994). "Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology." International Journal of Systematic Bacteriology **44**(4): 846-849.

Stevens, D. L. and E. L. Kaplan (2000). Streptococcal infections: clinical aspects, microbiology, and molecular pathogenesis, Oxford University Press, USA.

Sullam, P., F. Valone and J. Mills (1987). "Mechanisms of platelet aggregation by viridans group streptococci." Infection and immunity **55**(8): 1743-1750.

Swetha, R. G., D. K. K. Sekar, E. D. Devi, Z. Z. Ahmed, S. Ramaiah, A. Anbarasu and K. Sekar (2014). "Streptococcus pneumoniae Genome Database (SPGDB): A database for strain specific comparative analysis of Streptococcus pneumoniae genes and proteins." Genomics **104**(6): 582-586.

Teng, J. L., Y. Huang, H. Tse, J. H. Chen, Y. Tang, S. K. Lau and P. C. Woo (2014). "Phylogenomic and MALDI-TOF MS analysis of *Streptococcus sinensis* HKU4T reveals a distinct phylogenetic clade in the genus *Streptococcus*." Genome biology and evolution **6**(10): 2930-2943.

Teng, L.-J., P.-R. Hsueh, J.-C. Tsai, P.-W. Chen, J.-C. Hsu, H.-C. Lai, C.-N. Lee and S.-W. Ho (2002). "groESL sequence determination, phylogenetic analysis, and species differentiation for viridans group streptococci." Journal of clinical microbiology **40**(9): 3172-3178.

Terao, Y., S. Kawabata, E. Kunitomo, I. Nakagawa and S. Hamada (2002). "Novel laminin-binding protein of *Streptococcus pyogenes*, *Lbp*, is involved in adhesion to epithelial cells." Infection and immunity **70**(2): 993-997.

Tesorero, R. A., N. Yu, J. O. Wright, J. P. Svencionis, Q. Cheng, J.-H. Kim and K. H. Cho (2013). "Novel regulatory small RNAs in *Streptococcus pyogenes*." PloS one **8**(6): e64021.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones and A. S. Durkin (2005). "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"." Proceedings of the National Academy of Sciences of the United States of America **102**(39): 13950-13955.

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.

Thomopson, J., D. G. Higgins and T. J. Gibson (1994). "ClustalW." Nucleic Acids Res **22**: 4673-4680.

Thompson, C. C., V. E. Emmel, E. L. Fonseca, M. A. Marin and A. C. P. Vicente (2013). "Streptococcal taxonomy based on genome sequence analyses." F1000Research **2**.

Tilley, D. and S. W. Kerrigan (2013). "Platelet-bacterial interactions in the pathogenesis of infective endocarditis. Part I: The *Streptococcus.*"

Tleyjeh, I. M., J. M. Steckelberg, H. S. Murad, N. S. Anavekar, H. M. Ghomrawi, Z. Mirzoyev, S. E. Moustafa, T. L. Hoskin, J. N. Mandrekar and W. R. Wilson (2005). "Temporal trends in infective endocarditis: a population-based study in Olmsted County, Minnesota." Jama **293**(24): 3022-3028.

Tsuda, H., Y. Yamashita, K. Toyoshima, N. Yamaguchi, T. Oho, Y. Nakano, K. Nagata and T. Koga (2000). "Role of serotype-specific polysaccharide in the resistance of *Streptococcus mutans* to phagocytosis by human polymorphonuclear leukocytes." Infection and immunity **68**(2): 644-650.

Turner, L. (2008). "Identification of Virulence Determinants for *Streptococcus sanguinis* Infective Endocarditis."

Uguru, G. C., K. E. Stephens, J. A. Stead, J. E. Towle, S. Baumberg and K. J. McDowall (2005). "Transcriptional activation of the pathway‐specific regulator of the actinorhodin biosynthetic genes in *Streptomyces coelicolor*." Molecular microbiology **58**(1): 131-150.

van der Oost, J., E. R. Westra, R. N. Jackson and B. Wiedenheft (2014). "Unravelling the structural and mechanistic basis of CRISPR-Cas systems." Nature Reviews Microbiology **12**(7): 479-492.

van Dijk, E. L., H. Auger, Y. Jaszczyszyn and C. Thermes (2014). "Ten years of next-generation sequencing technology." Trends in genetics **30**(9): 418-426.

Varon, E. and L. Gutmann (2000). "Mechanisms and spread of fluoroquinolone resistancein *Streptococcus pneumoniae*." Research in microbiology **151**(6): 471-473.

Västermark, Å., M. S. Almén, M. W. Simmen, R. Fredriksson and H. B. Schiöth (2011). "Functional specialization in nucleotide sugar transporters occurred through differentiation of the gene cluster EamA (DUF6) before the radiation of Viridiplantae." BMC evolutionary biology **11**(1): 123.

Vester, B. and K. S. Long (2000). "Antibiotic resistance in bacteria caused by modified nucleosides in 23S ribosomal RNA."

Vetting, M. W., L. P. S. de Carvalho, M. Yu, S. S. Hegde, S. Magnet, S. L. Roderick and J. S. Blanchard (2005). "Structure and functions of the GNAT superfamily of acetyltransferases." Archives of biochemistry and biophysics **433**(1): 212-226.

Vickerman, M., S. Flannagan, A. Jesionowski, K. Brossard, D. Clewell and C. Sedgley (2010). "A genetic determinant in *Streptococcus gordonii* Challis encodes a peptide with activity similar to that of enterococcal sex pheromone cAM373, which facilitates intergeneric DNA transfer." Journal of bacteriology **192**(10): 2535-2545.

Vogkou, C. T., N. I. Vlachogiannis, L. Palaiodimos and A. A. Kousoulis (2016). "The causative agents in infective endocarditis: a systematic review comprising 33,214 cases." European Journal of Clinical Microbiology & Infectious Diseases: 1-19.

Walker, T. M., C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey and D. W. Crook (2013). "Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study." The Lancet infectious diseases **13**(2): 137-146.

Werdan, K., S. Dietz, B. Löffler, S. Niemann, H. Bushnaq, R.-E. Silber, G. Peters and U. Müller-Werdan (2014). "Mechanisms of infective endocarditis: pathogen-host interaction and risk states." Nature Reviews Cardiology **11**(1): 35-50.

Wessels, M. R., A. E. Moses, J. B. Goldberg and T. J. DiCesare (1991). "Hyaluronic acid capsule is a virulence factor for mucoid group A streptococci." Proceedings of the National Academy of Sciences **88**(19): 8317-8321.

Westling, K. (2005). Viridans group streptococci septicaemia and endocarditis: Molecular diagnostics, antibiotic susceptibility and cinical aspects, Institutionen för medicin/Department of Medicine.

Wibawan, I. W. T., F. Pasaribu, I. Utama, A. Abdulmawjood and C. Laemmler (1999). "The role of hyaluronic acid capsular material of *Streptococcus equi subsp. zooepidemicus* in mediating adherence to HeLa cells and in resisting phagocytosis." Research in veterinary science **67**(2): 131-135.

Widmer, E., Y.-A. Que, J. M. Entenza and P. Moreillon (2006). "New concepts in the pathophysiology of infective endocarditis." Current infectious disease reports **8**(4): 271-279.

Wilson, W., K. A. Taubert, M. Gewitz, P. B. Lockhart, L. M. Baddour, M. Levison, A. Bolger, C. H. Cabell, M. Takahashi and R. S. Baltimore (2007). "Prevention of Infective Endocarditis Guidelines From the American Heart Association: A Guideline From the American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee, Council on Cardiovascular Disease in the Young, and the Council on Clinical Cardiology, Council on Cardiovascular Surgery and Anesthesia, and the Quality of Care and Outcomes Research Interdisciplinary Working Group." Circulation **116**(15): 1736-1754.

Xu, P., J. M. Alves, T. Kitten, A. Brown, Z. Chen, L. S. Ozaki, P. Manque, X. Ge, M. G. Serrano and D. Puiu (2007). "Genome of the opportunistic pathogen *Streptococcus sanguinis*." Journal of bacteriology **189**(8): 3166-3175.

Yamashita, Y., Y. Shibata, Y. Nakano, H. Tsuda, N. Kido, M. Ohta and T. Koga (1999). "A novel gene required for rhamnose-glucose polysaccharide synthesis in *Streptococcus mutans*." Journal of bacteriology **181**(20): 6556-6559.

Yamashita, Y., Y. Tsukioka, K. Tomihisa, Y. Nakano and T. Koga (1998). "Genes Involved in Cell Wall Localization and Side Chain Formation of Rhamnose-Glucose Polysaccharide in *Streptococcus mutans*." Journal of bacteriology **180**(21): 5803-5807.

Yang, J., Y. Yoshida and J. O. Cisar (2014). "Genetic basis of coaggregation receptor polysaccharide biosynthesis in *Streptococcus sanguinis* and related species." Molecular oral microbiology **29**(1): 24-31.

Yoshida, Y., S. Ganguly, C. A. Bush and J. O. Cisar (2006). "Molecular basis of L-rhamnose branch formation in streptococcal coaggregation receptor polysaccharides." Journal of bacteriology **188**(11): 4125-4130.

Young, B. C., T. Golubchik, E. M. Batty, R. Fung, H. Larner-Svensson, A. A. Votintseva, R. R. Miller, H. Godwin, K. Knox and R. G. Everitt (2012). "Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease." Proceedings of the National Academy of Sciences **109**(12): 4550-4555.

Yvorchuk, K. J. and K.-L. Chan (1994). "Application of transthoracic and transesophageal echocardiography in the diagnosis and management of infective endocarditis." Journal of the American Society of Echocardiography **7**(3): 294-308.

Zhang, Y.-M., Z.-Q. Shao, J. Wang, L. Wang, X. Li, C. Wang, J. Tang and X. Pan (2014). "Prevalent distribution and conservation of Streptococcus suis Lmb protein and its protective capacity against the Chinese highly virulent strain infection." Microbiological research **169**(5): 395-401.

Zhang, Y., D. A. Rodionov, M. S. Gelfand and V. N. Gladyshev (2009). "Comparative genomic analyses of nickel, cobalt and vitamin B12 utilization." BMC genomics **10**(1): 78.

Zhang, Y., M. Whiteley, J. Kreth, Y. Lei, A. Khammanivong, J. N. Evavold, J. Fan and M. C. Herzberg (2009). "The two-component system BfrAB regulates expression of ABC transporters in *Streptococcus gordonii* and *Streptococcus sanguinis*." Microbiology **155**(1): 165-173.

Zhou, Y., Y. Liang, K. H. Lynch, J. J. Dennis and D. S. Wishart (2011). "PHAST: a fast phage search tool." Nucleic acids research: gkr485.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Zheng, W., Tan, T. K., Paterson, I. C., Mutha, N. V., Siow, C. C., Tan, S. Y., Old, L.A., Jakubovics, N.S. and Choo, S. W. (2016). StreptoBase: An Oral *Streptococcus* Mitis Group Genomic Resource and Analysis Platform. PloS one, 11(5), e0151908.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

UNIVERSITI TUNKU ABDUL RAHMAN

Certificate of Attendance

This is to certify that

Zheng Wenning

has presented in the

UTAR National Postgraduate Fundamental and Applied Science Seminar 2014 (UTAR NPFASS 2014)

Organised by
Centre for Biodiversity Research
Faculty of Science
Institute of Postgraduate Studies & Research

held on 14 – 15 th June 2014,
at UTAR Perak Campus, Kampar.

Mr. Ng Wen Jie
Chairperson
Organising Committee
UTAR NPFASS 2014

Assoc. Prof. Dr. Lim Tuck Meng
Dean, Faculty of Science
UTAR Perak Campus

LIST OF PUBLICATIONS AND PAPERS PRESENTED

UNIVERSITI MALAYA
The Leader in Research & Innovation

CERTIFICATE OF PARTICIPATION

this is to certify that

ZHENG WENNING

has attended the

DENTAL CONGREGATION
organized by Faculty of Dentistry, University of Malaya

THE ROYALE CHULAN DAMANSARA
PETALING JAYA, SELANGOR

2016
AUGUST 13-14

"111TH YEAR OF EXCELLENCE: INTEGRATING RESEARCH AND PRACTICE"

Prof. Dato' Dr.
Zainal Ariff Abdul Rahman
Dean
Faculty of Dentistry
University of Malaya

Dr. Zubaidah Zanul Abidin
Chairperson
Dental Congregation 2016

[A2 • 6 CPD hours rewarded]

APPENDICES

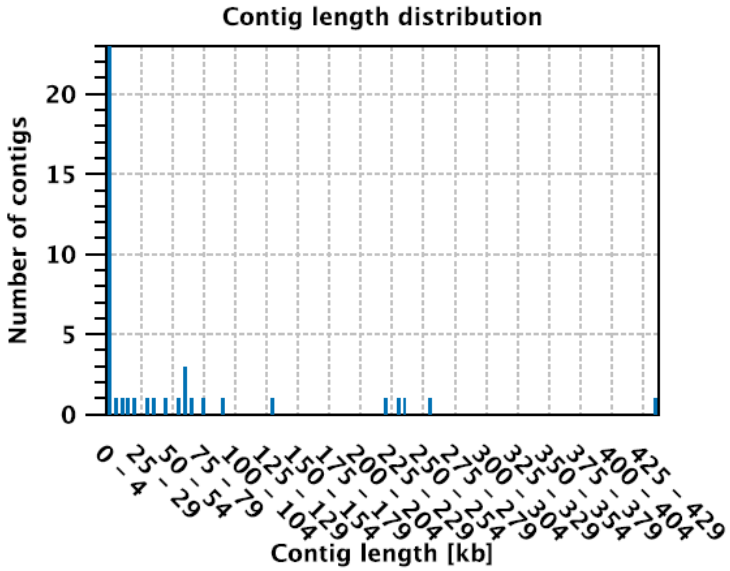
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	652,839	29.8%
Cytosine (C)	440,784	20.1%
Guanine (G)	446,116	20.4%
Thymine (T)	651,312	29.7%

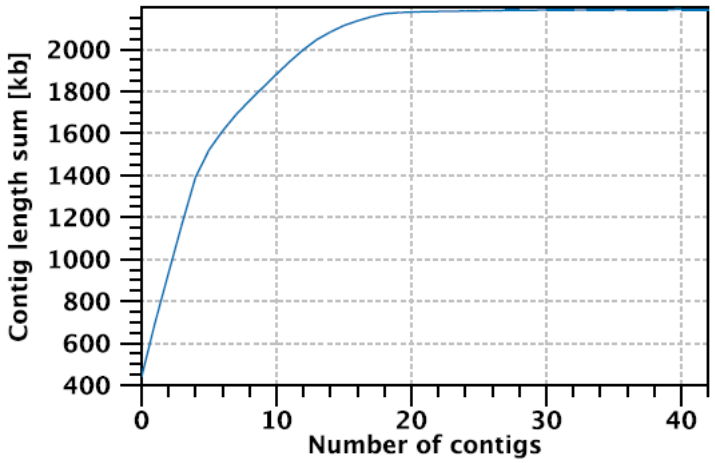
1.2 Contig measurements

N75	77,940
N50	233,745
N25	258,448
Minimum	207
Maximum	439,437
Average	50,955
Count	43

Total	2,191,051



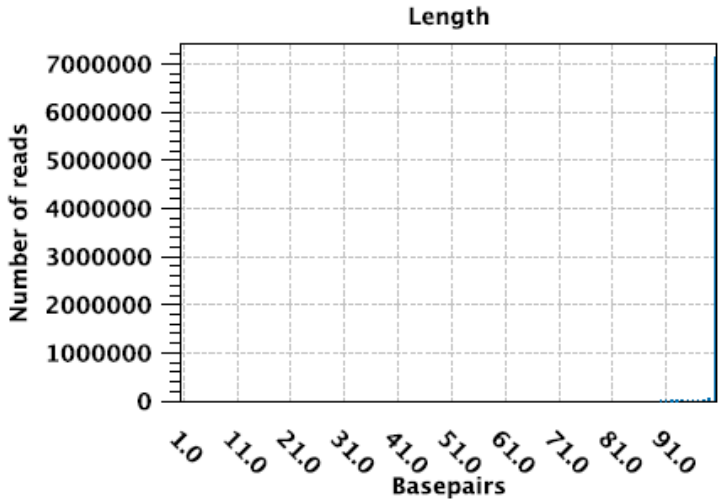
1.3 Accumulated contig lengths



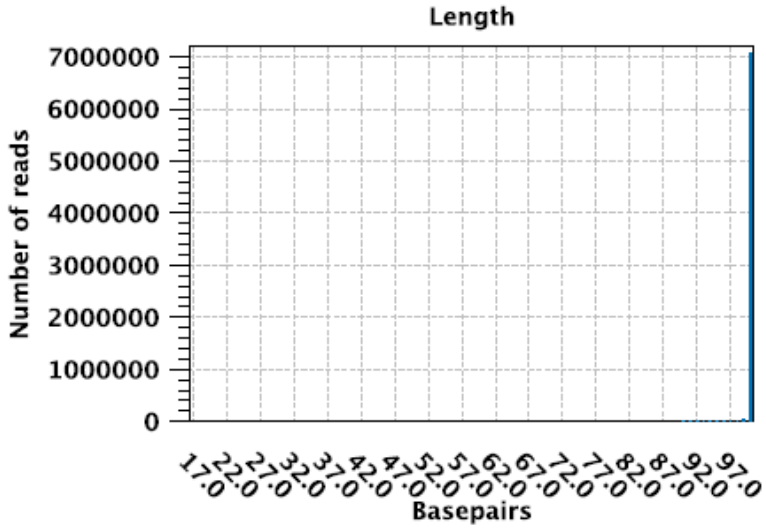
1.4 Summary statistics

	Count	Average length	Total bases
Reads	8,146,818	95.7	779,676,884
Matched	7,993,309	96.55	771,777,013
Not matched	153,509	51.46	7,899,871
Contigs	43	50,954	2,191,051
Reads in pairs	6,915,550	352.86	
Broken paired reads	1,077,759	91.98	

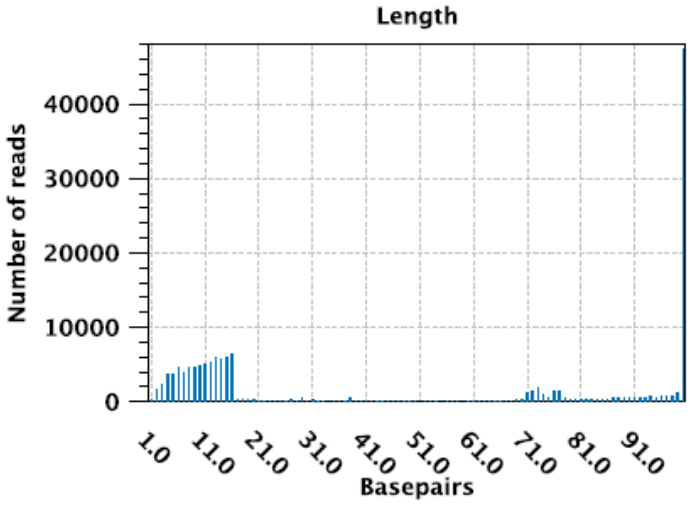
1.5 Distribution of read length



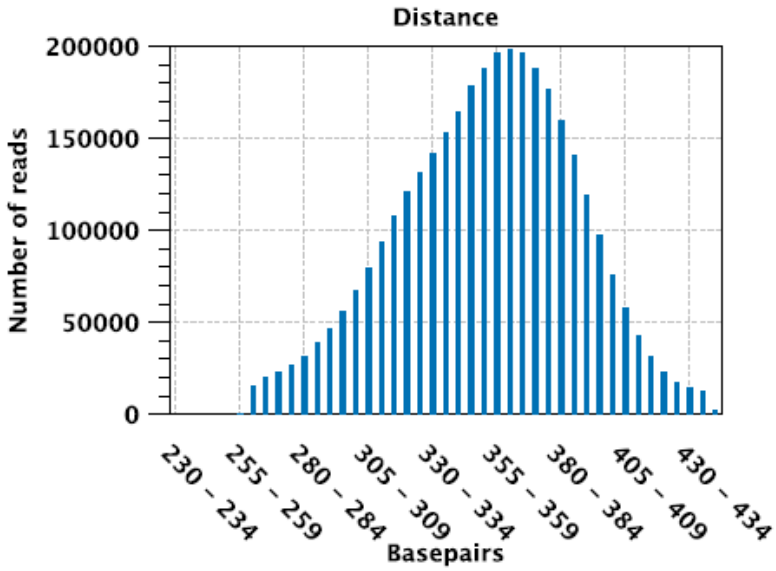
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix A: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* PV40 with Phred score 20.

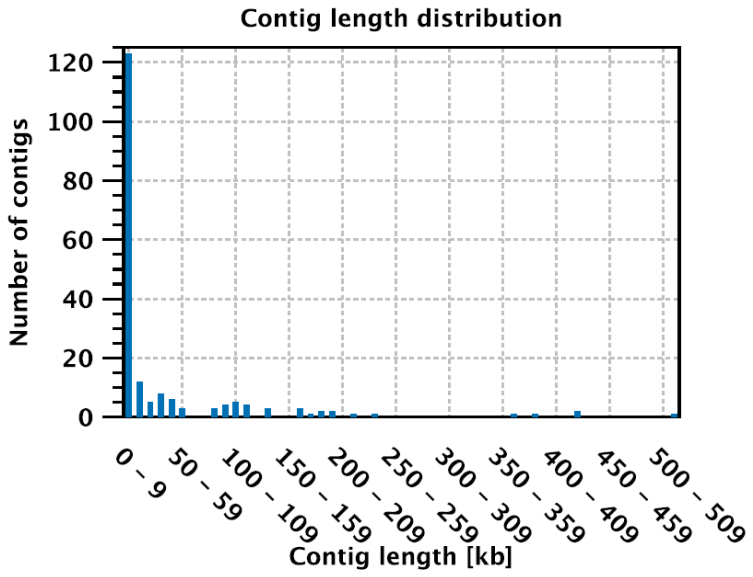
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	1,877,939	25.9%
Cytosine (C)	1,730,511	23.9%
Guanine (G)	1,759,595	24.3%
Thymine (T)	1,877,683	25.9%

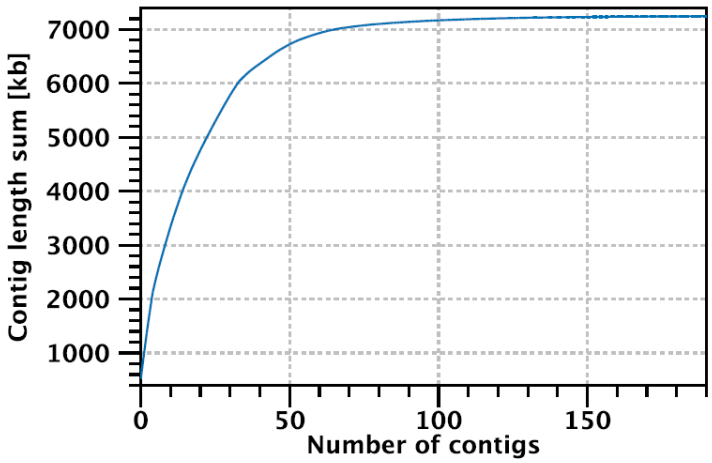
1.2 Contig measurements

N75	98,642
N50	164,052
N25	364,438
Minimum	215
Maximum	518,021
Average	37,936
Count	191

Total	7,245,728



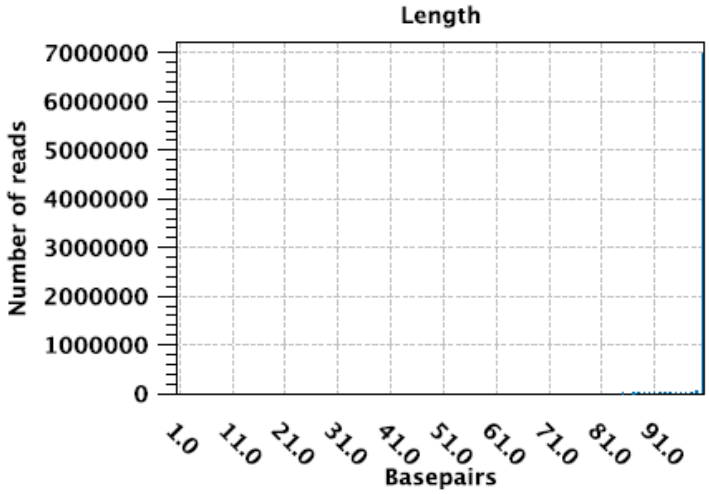
1.3 Accumulated contig lengths



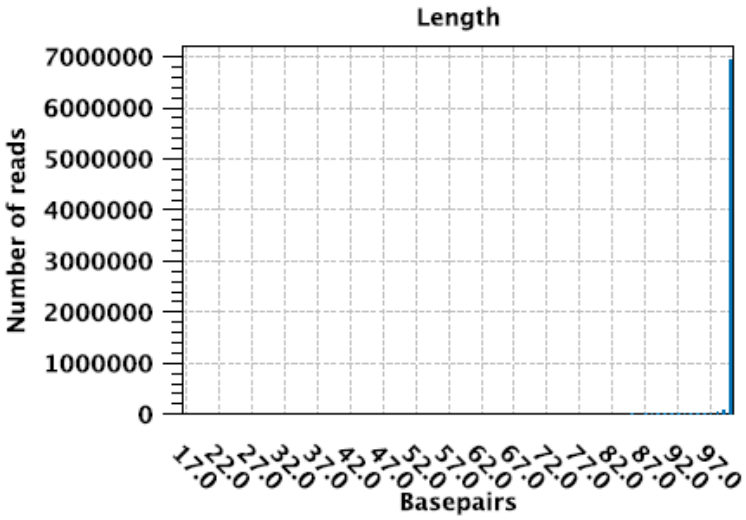
1.4 Summary statistics

	Count	Average length	Total bases
Reads	8,181,512	95.62	782,339,389
Matched	8,090,008	96.14	777,769,196
Not matched	91,504	49.95	4,570,193
Contigs	191	37,935	7,245,728
Reads in pairs	7,247,810	286.35	
Broken paired reads	842,198	94.54	

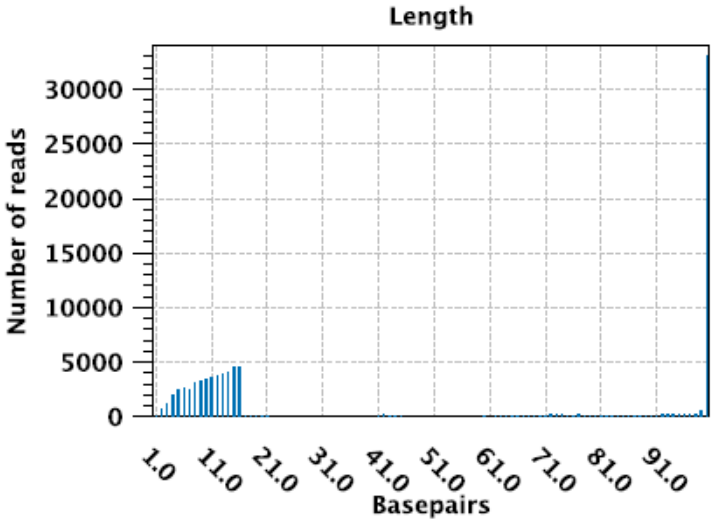
1.5 Distribution of read length



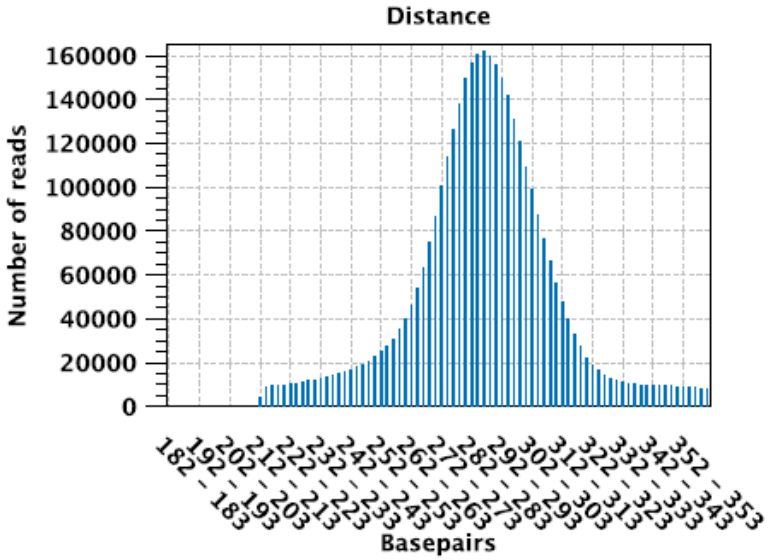
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix B: The CLC Genomics Workbench de novo assembly summary report of *S. sanguinis* NCTC 7863 with Phred score 20.

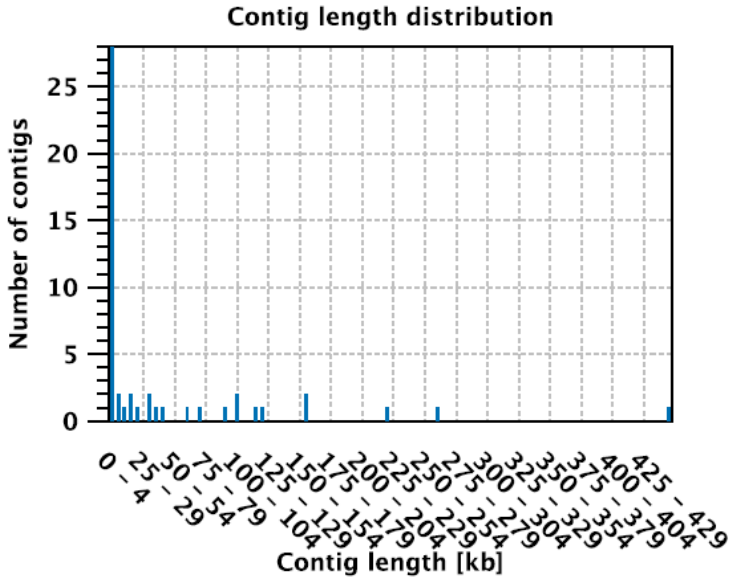
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	645,576	29.8%
Cytosine (C)	432,611	20.0%
Guanine (G)	443,399	20.5%
Thymine (T)	642,946	29.7%

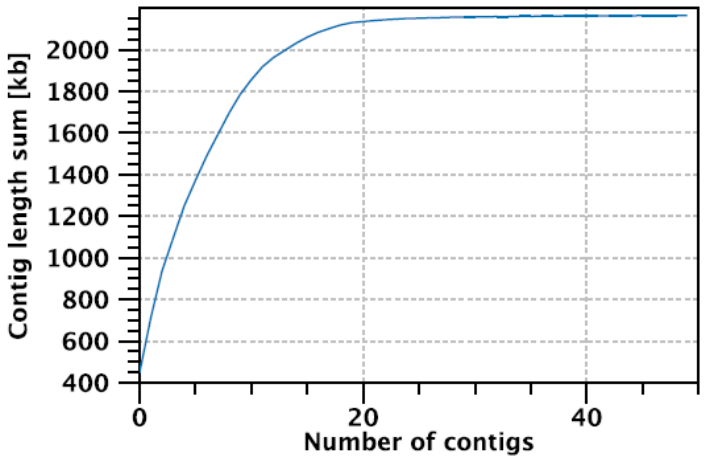
1.2 Contig measurements

N75	102,876
N50	158,790
N25	262,985
Minimum	203
Maximum	448,113
Average	43,291
Count	50

Total	2,164,532



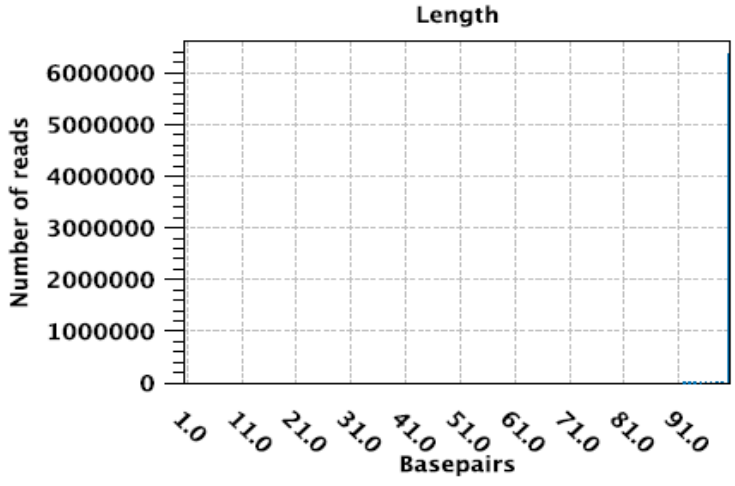
1.3 Accumulated contig lengths



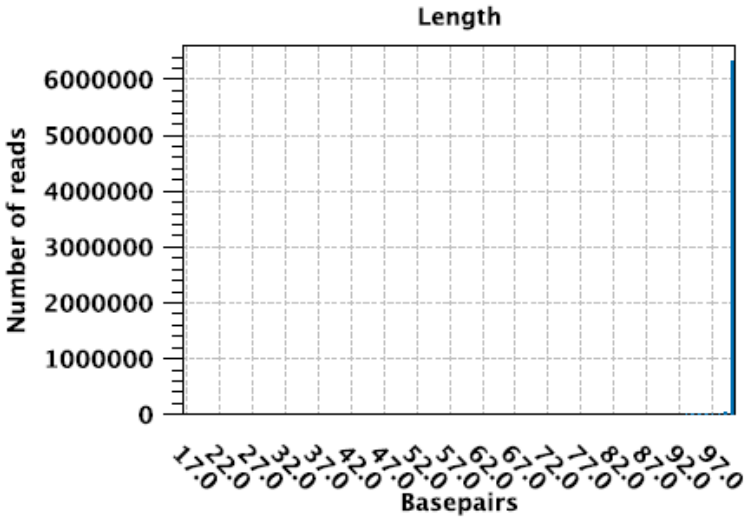
1.4 Summary statistics

	Count	Average length	Total bases
Reads	7,161,562	96.37	690,179,613
Matched	7,051,125	96.96	683,689,807
Not matched	110,437	58.76	6,489,806
Contigs	50	43,290	2,164,532
Reads in pairs	6,237,434	323.44	
Broken paired reads	813,691	93.07	

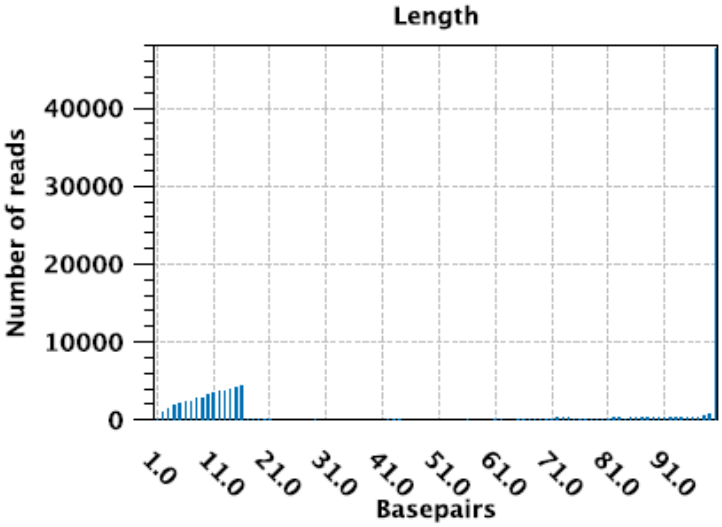
1.5 Distribution of read length



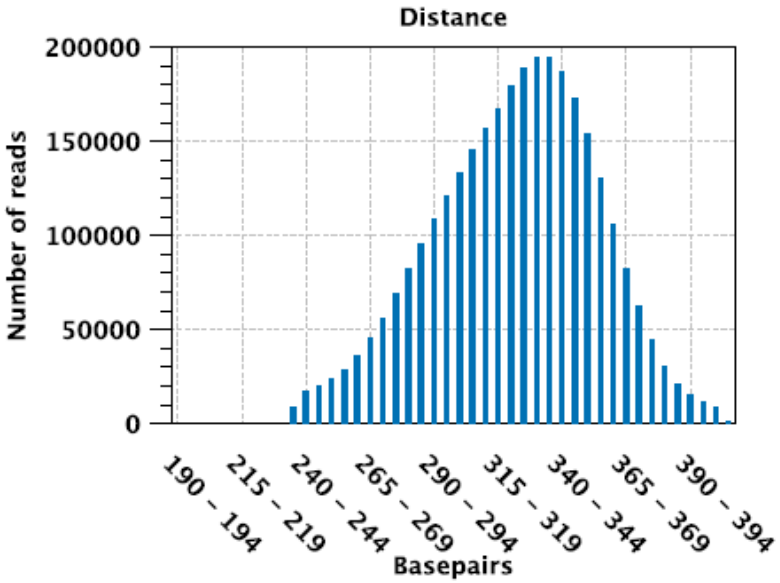
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix C: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* Blackburn with Phred score 20.

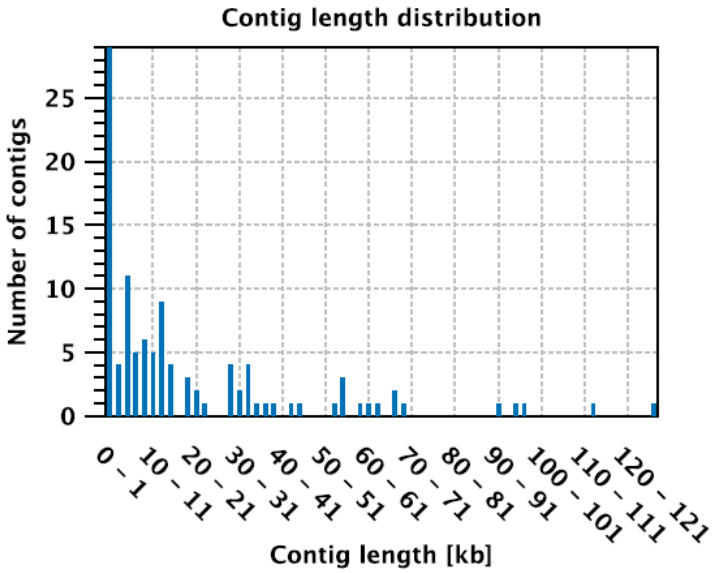
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	620,080	29.2%
Cytosine (C)	441,863	20.8%
Guanine (G)	445,452	21.0%
Thymine (T)	615,292	29.0%

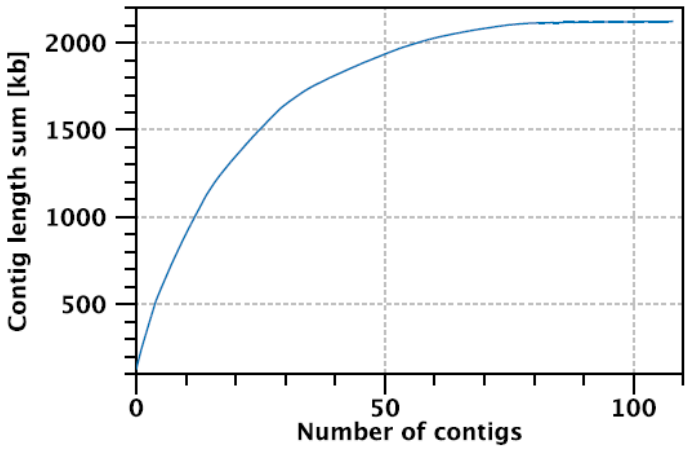
1.2 Contig measurements

Length	
N75	29,288
N50	53,977
N25	69,079
Minimum	203
Maximum	126,572
Average	19,474
Count	109

Length	
Total	2,122,687



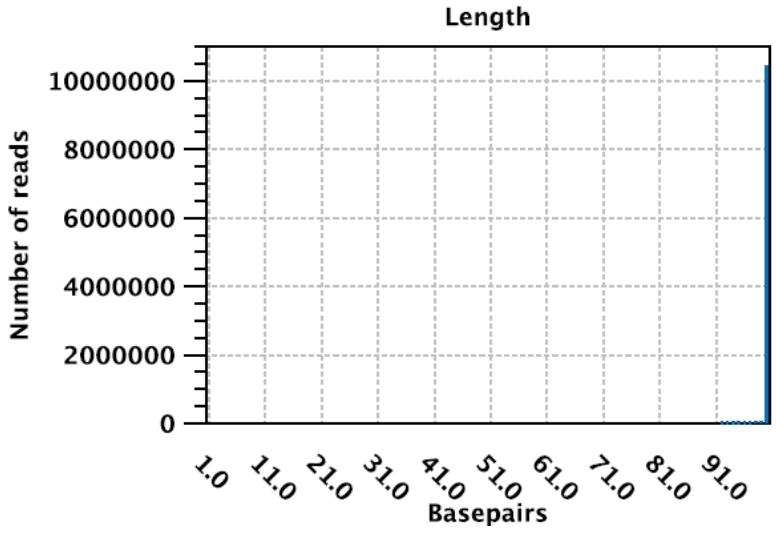
1.3 Accumulated contig lengths



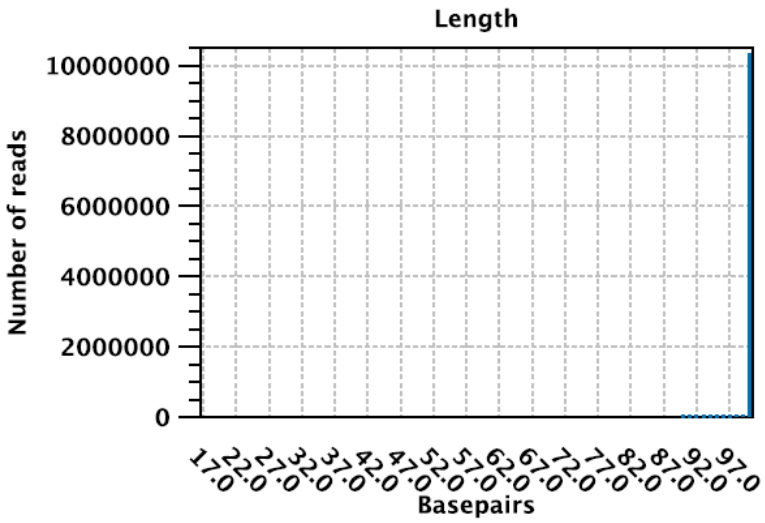
1.4 Summary statistics

	Count	Average length	Total bases
Reads	11,731,358	96.41	1,130,980,007
Matched	11,556,016	96.99	1,120,850,979
Not matched	175,342	57.77	10,129,028
Contigs	109	19,474	2,122,687
Reads in pairs	9,965,744	306.72	
Broken paired reads	1,590,272	94.99	

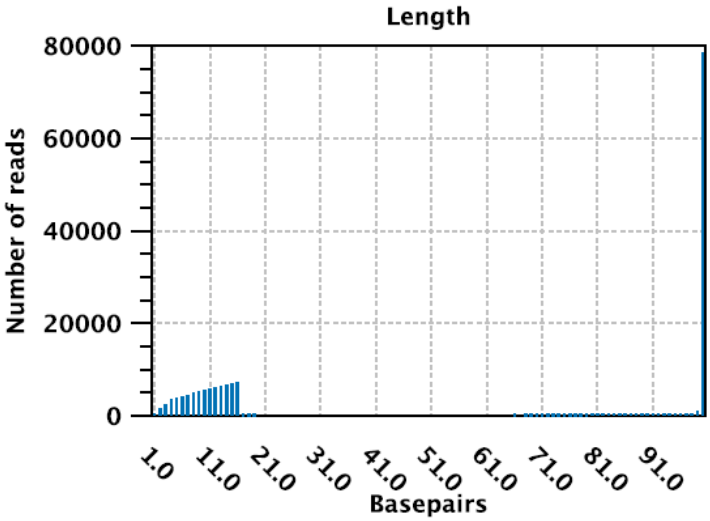
1.5 Distribution of read length



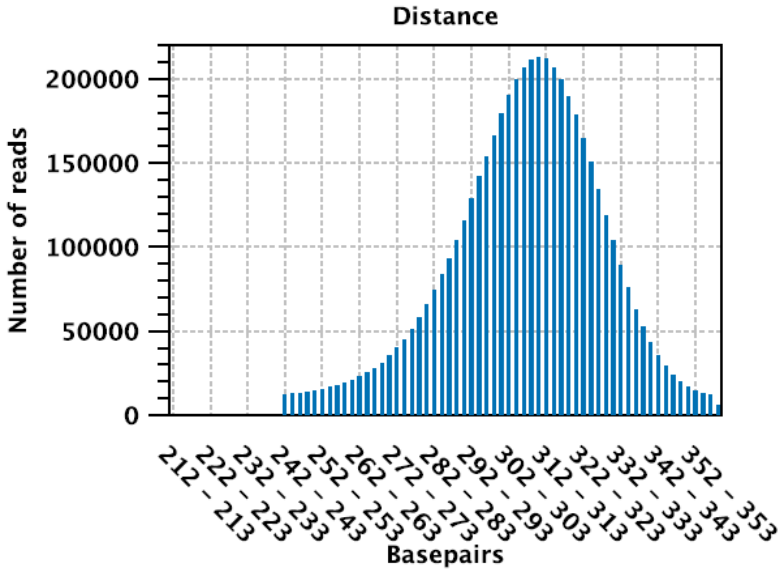
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix D: The CLC Genomics Workbench de novo assembly summary report of *S. parasanguinis* BVME8 with Phred score 20.

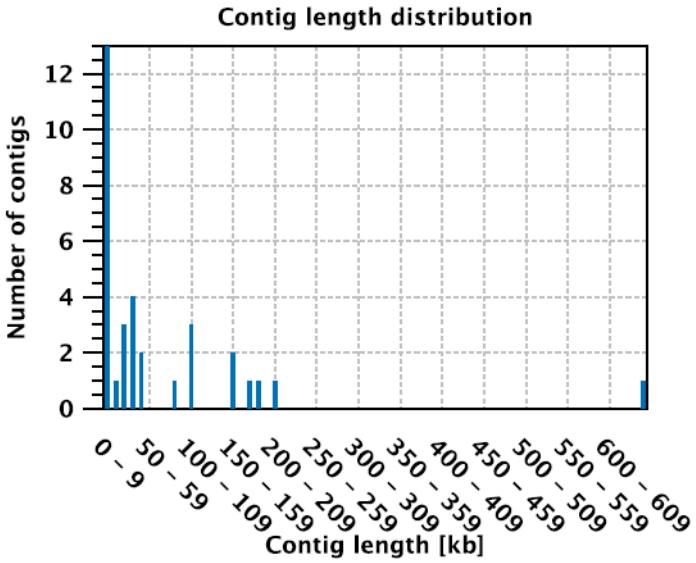
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	663,841	29.7%
Cytosine (C)	461,610	20.7%
Guanine (G)	446,203	20.0%
Thymine (T)	661,946	29.6%

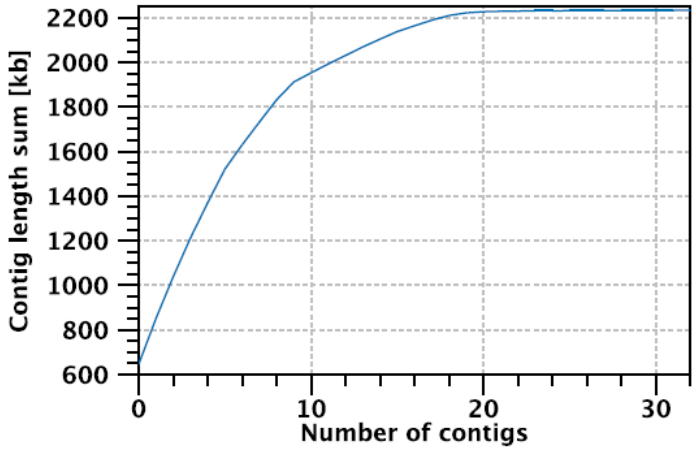
1.2 Contig measurements

Length	
N75	101,656
N50	174,000
N25	646,319
Minimum	233
Maximum	646,319
Average	67,685
Count	33

Length	
Total	2,233,600



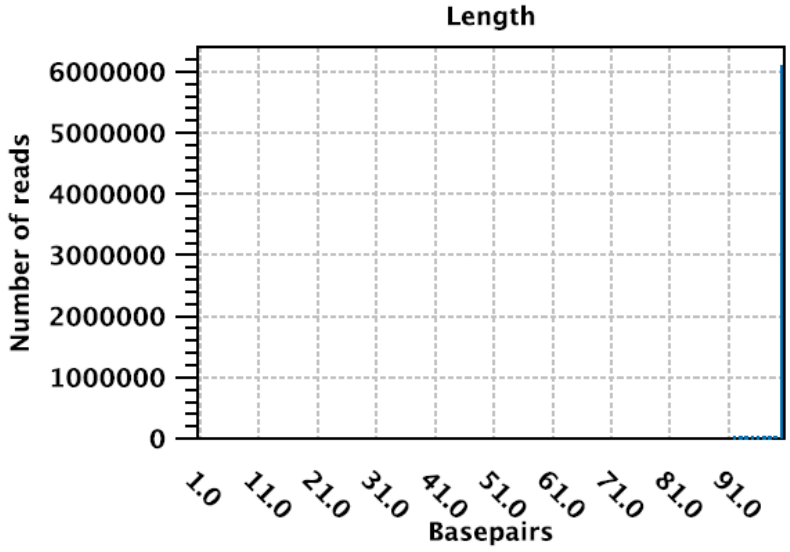
1.3 Accumulated contig lengths



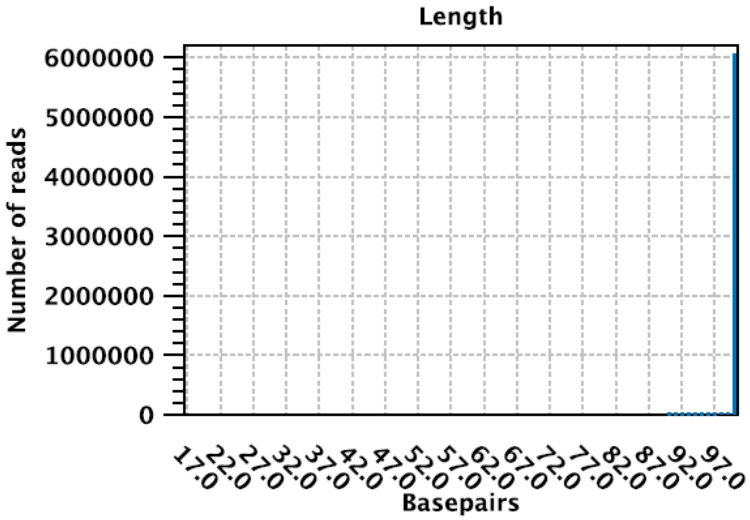
1.4 Summary statistics

	Count	Average length	Total bases
Reads	6,884,592	96.11	661,660,680
Matched	6,785,884	96.78	656,753,343
Not matched	98,708	49.72	4,907,337
Contigs	33	67,684	2,233,600
Reads in pairs	5,987,218	342.86	
Broken paired reads	798,666	92.64	

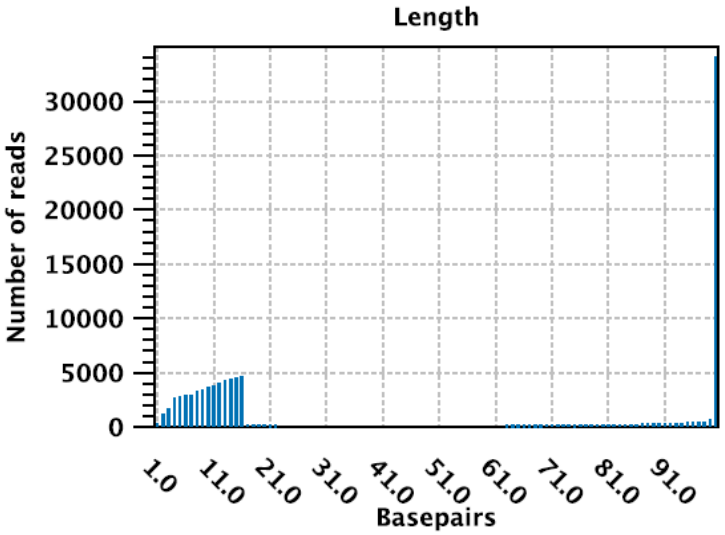
1.5 Distribution of read length



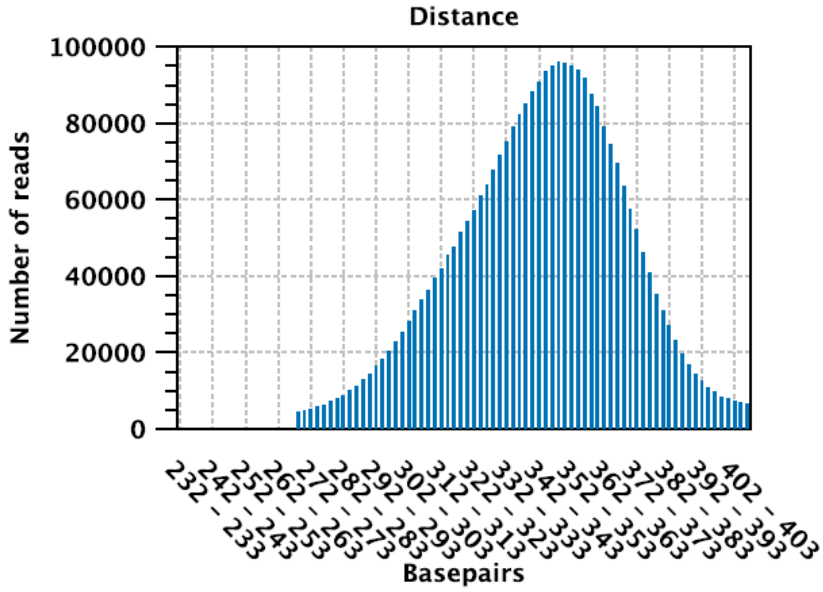
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix E: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* Channon with Phred score 20.

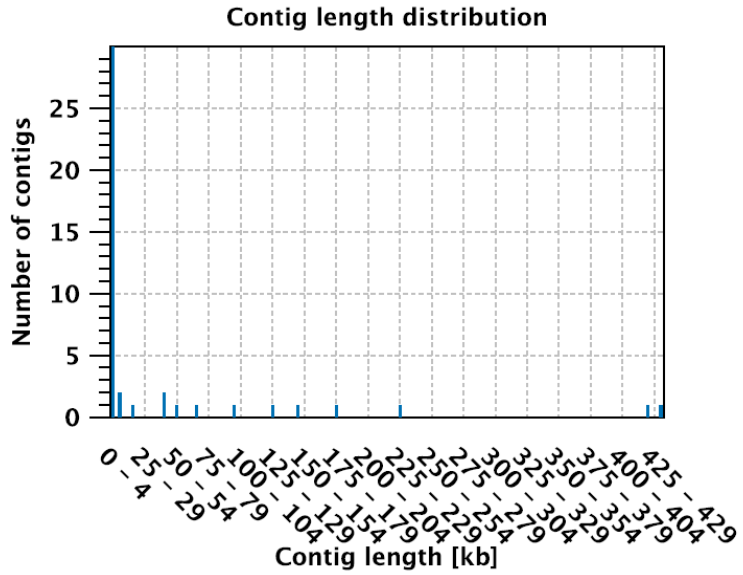
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	550,645	29.2%
Cytosine (C)	398,501	21.1%
Guanine (G)	375,910	19.9%
Thymine (T)	560,785	29.7%

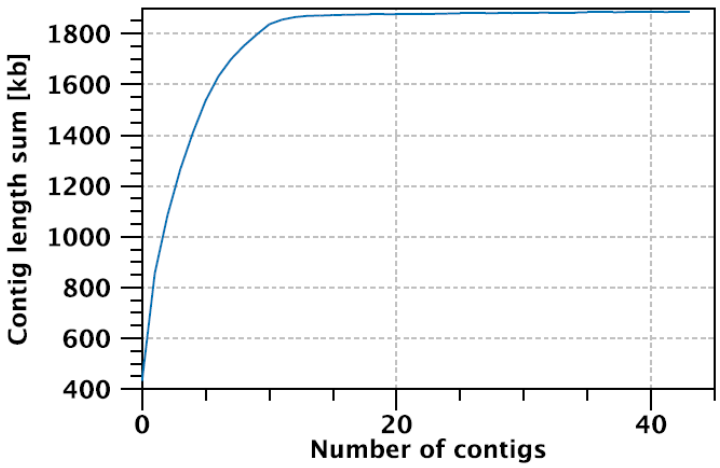
1.2 Contig measurements

N75	125,768
N50	229,281
N25	423,561
Minimum	213
Maximum	432,612
Average	42,860
Count	44

Total	1,885,841



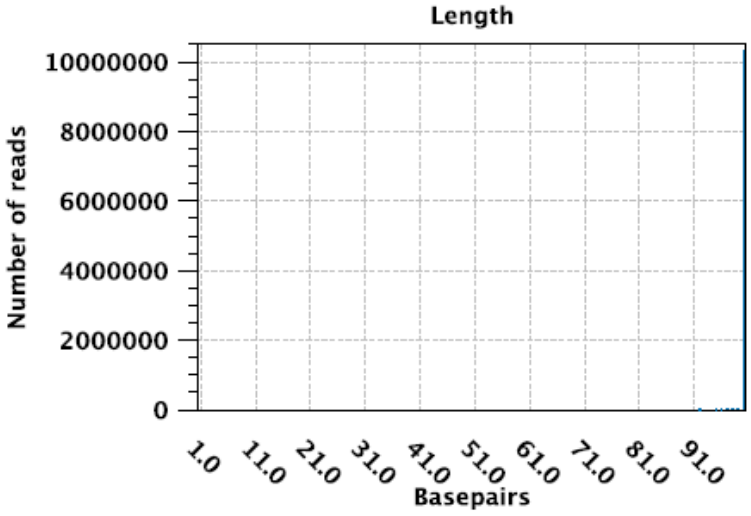
1.3 Accumulated contig lengths



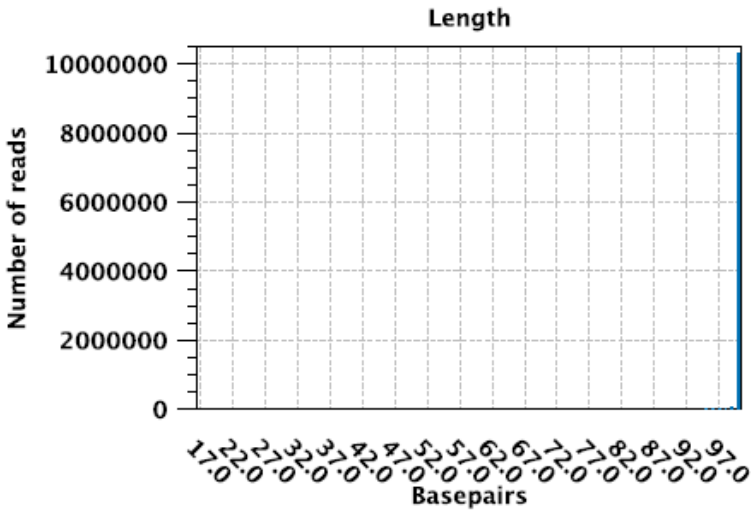
1.4 Summary statistics

	Count	Average length	Total bases
Reads	11,523,042	96.77	1,115,067,262
Matched	11,407,708	97.28	1,109,798,015
Not matched	115,334	45.69	5,269,247
Contigs	44	42,860	1,885,841
Reads in pairs	9,948,918	303.91	
Broken paired reads	1,458,790	95.85	

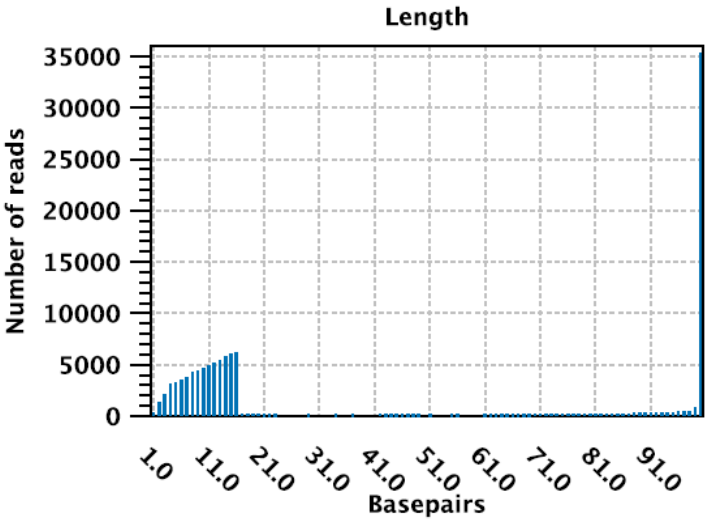
1.5 Distribution of read length



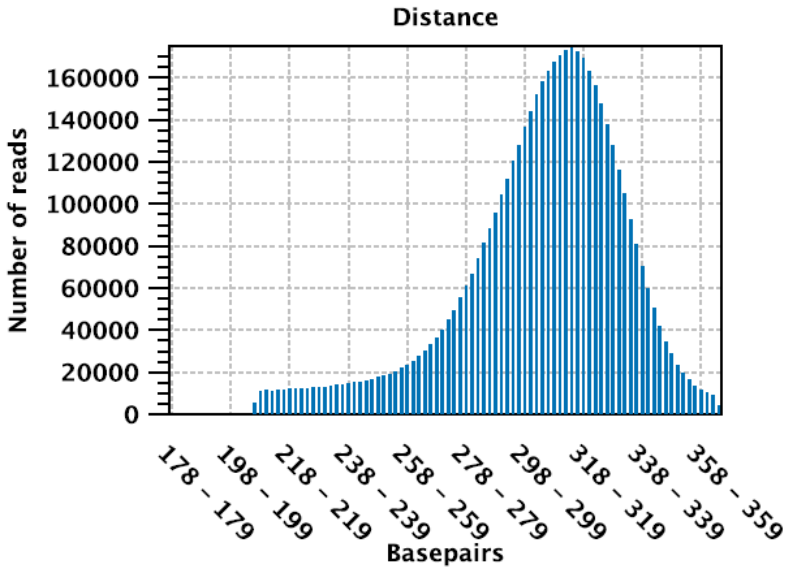
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix F: The CLC Genomics Workbench de novo assembly summary report of *S. tigurinus* DGIIBVI with Phred score 20.

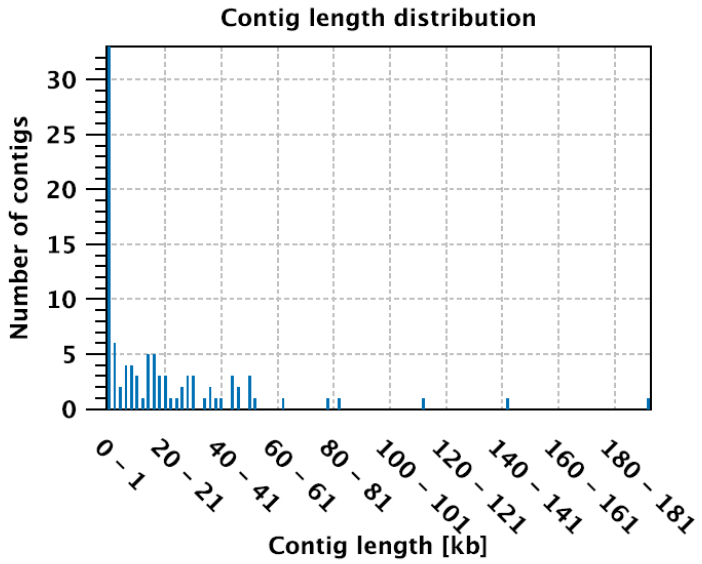
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	570,776	28.8%
Cytosine (C)	424,665	21.5%
Guanine (G)	418,948	21.2%
Thymine (T)	564,827	28.5%

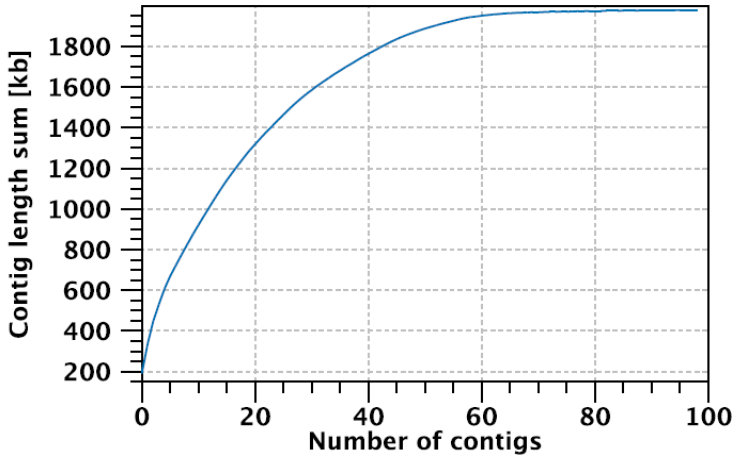
1.2 Contig measurements

N75	27,472
N50	45,179
N25	82,902
Minimum	204
Maximum	192,392
Average	19,992
Count	99

Total	1,979,216



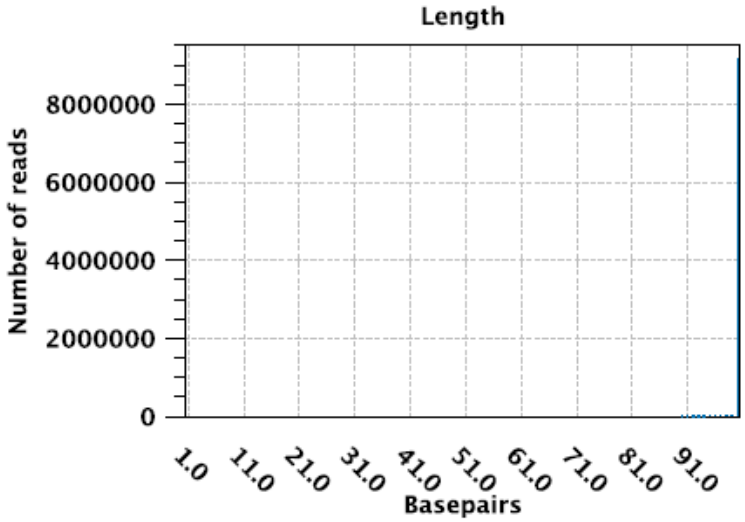
1.3 Accumulated contig lengths



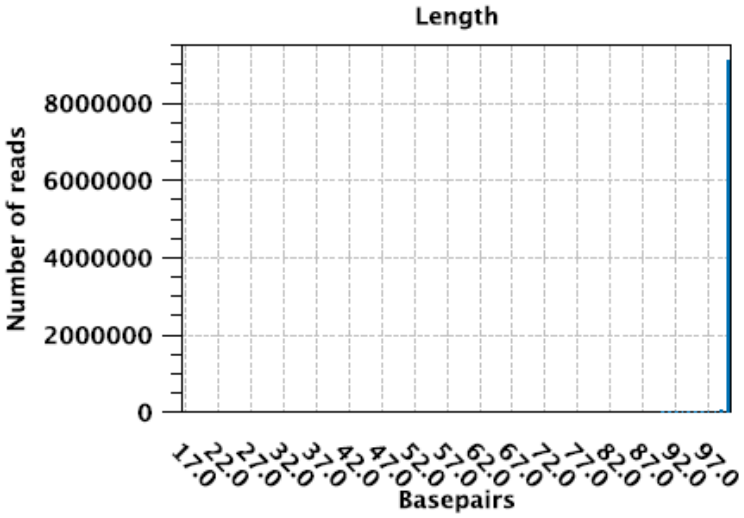
1.4 Summary statistics

	Count	Average length	Total bases
Reads	10,303,700	96.55	994,826,036
Matched	10,168,849	97.07	987,091,084
Not matched	134,851	57.36	7,734,952
Contigs	99	19,992	1,979,216
Reads in pairs	8,909,234	279.54	
Broken paired reads	1,259,615	95.48	

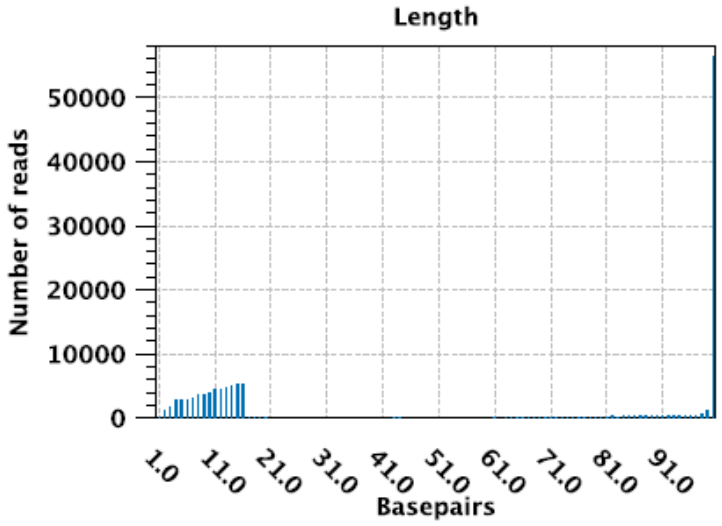
1.5 Distribution of read length



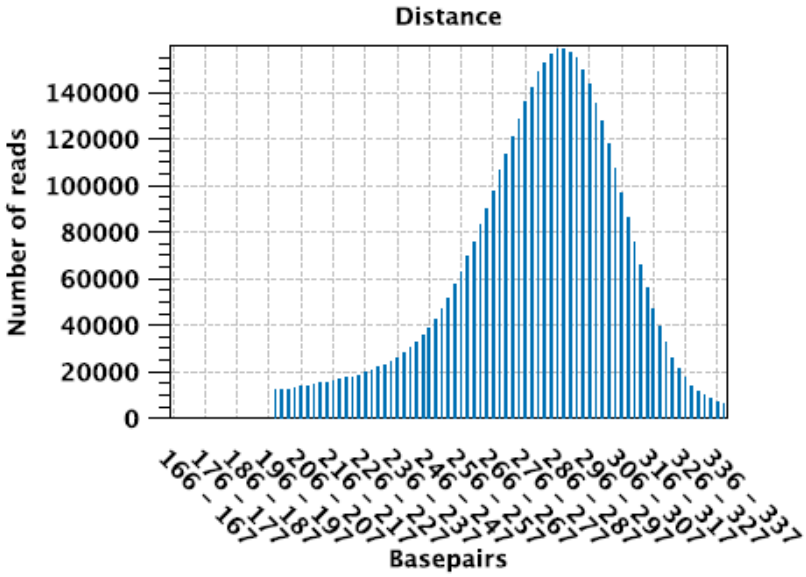
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix G: The CLC Genomics Workbench de novo assembly summary report of *S. oligofementans* DOBICBV2 with Phred score 20.

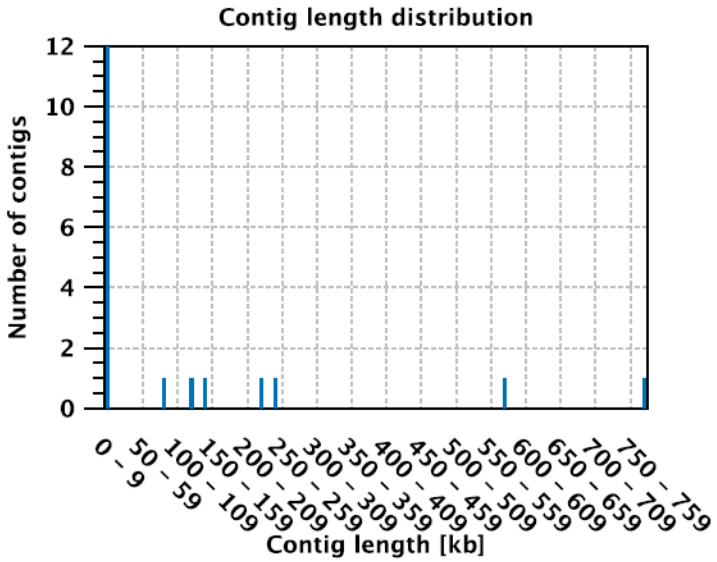
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	648,873	29.7%
Cytosine (C)	458,947	21.0%
Guanine (G)	425,917	19.5%
Thymine (T)	652,137	29.8%

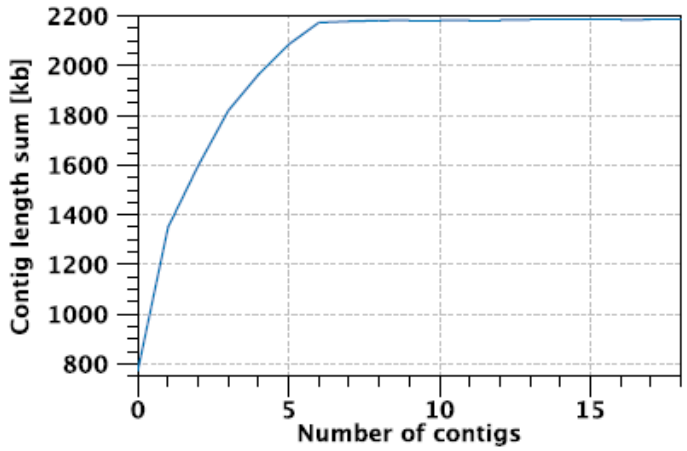
1.2 Contig measurements

Length	
N75	221,892
N50	575,926
N25	774,004
Minimum	183
Maximum	774,004
Average	115,046
Count	19

Length	
Total	2,185,874



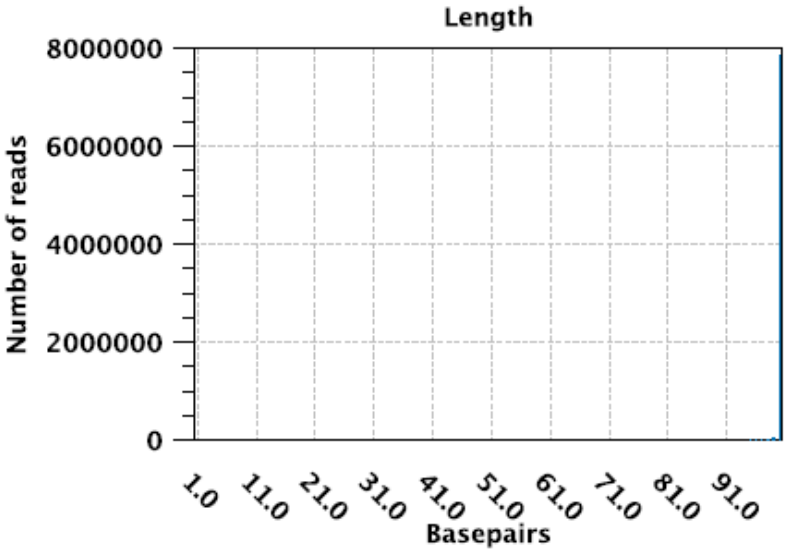
1.3 Accumulated contig lengths



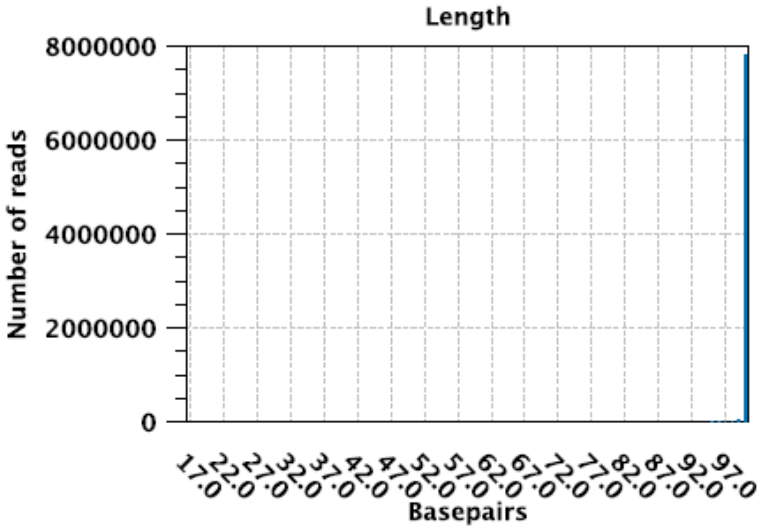
1.4 Summary statistics

	Count	Average length	Total bases
Reads	8,698,012	96.98	843,508,189
Matched	8,584,918	97.57	837,646,696
Not matched	113,094	51.83	5,861,493
Contigs	19	115,046	2,185,874
Reads in pairs	7,422,344	311.72	
Broken paired reads	1,162,574	96.1	

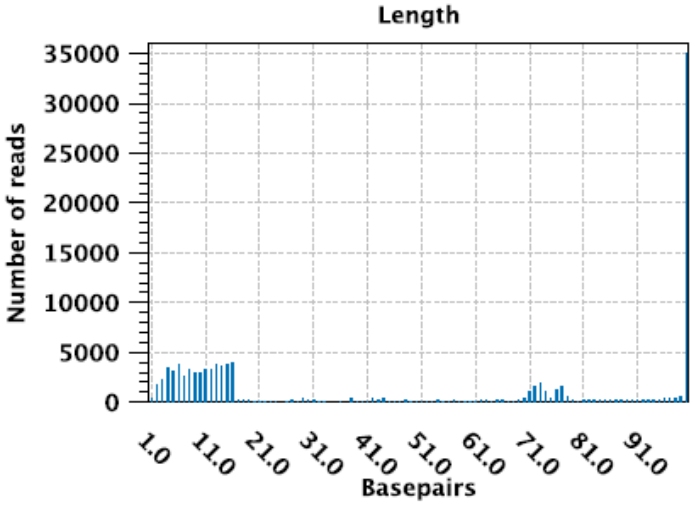
1.5 Distribution of read length



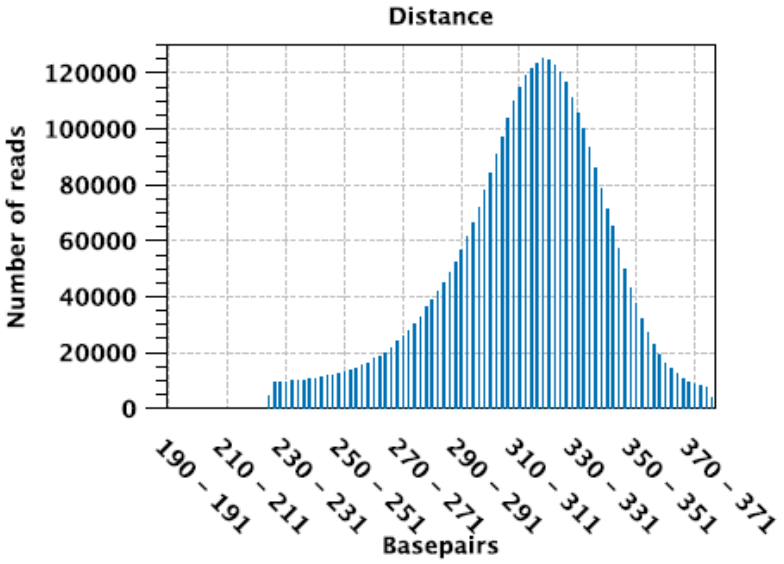
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix H: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* FSS2 with Phred score 20.

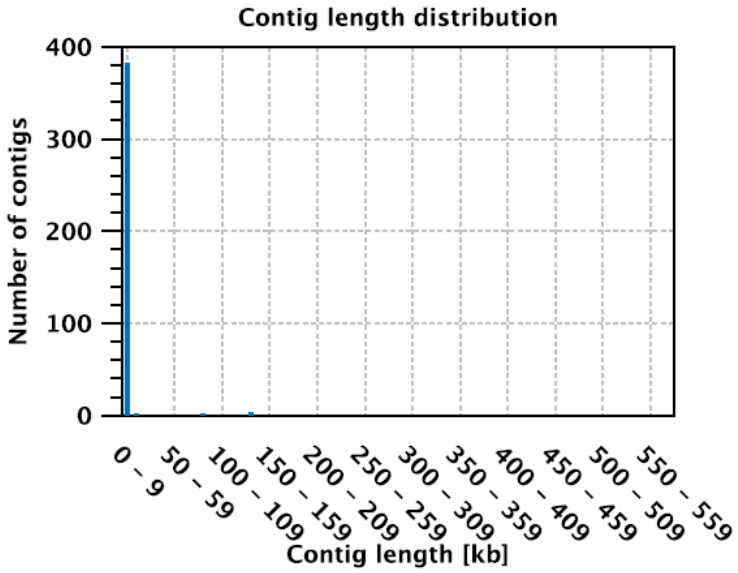
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	686,195	29.7%
Cytosine (C)	479,985	20.8%
Guanine (G)	450,365	19.5%
Thymine (T)	695,516	30.1%

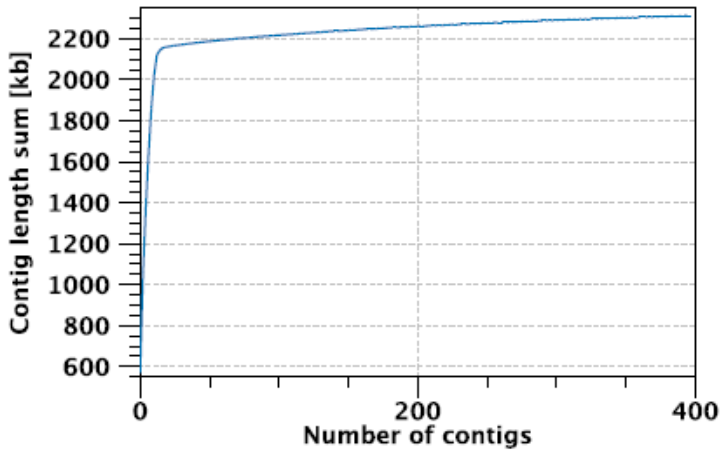
1.2 Contig measurements

Length	
N75	130,893
N50	172,943
N25	268,087
Minimum	186
Maximum	572,997
Average	5,809
Count	398

Length	
Total	2,312,061



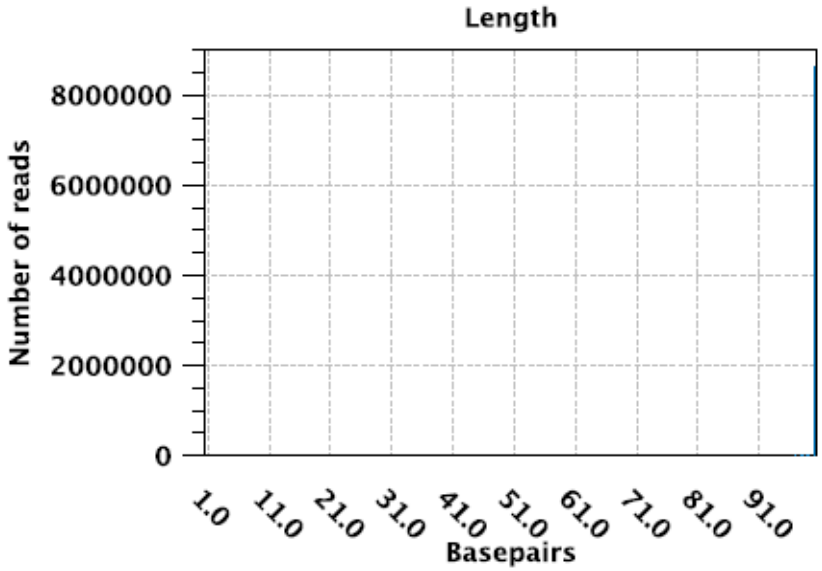
1.3 Accumulated contig lengths



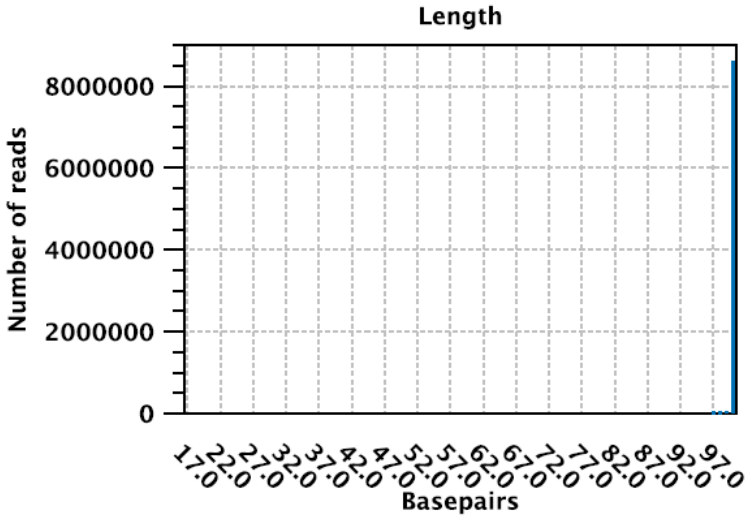
1.4 Summary statistics

	Count	Average length	Total bases
Reads	9,414,890	97.59	918,761,128
Matched	9,348,188	97.87	914,949,881
Not matched	66,702	57.14	3,811,247
Contigs	398	5,809	2,312,061
Reads in pairs	8,178,300	277.29	
Broken paired reads	1,169,888	96.5	

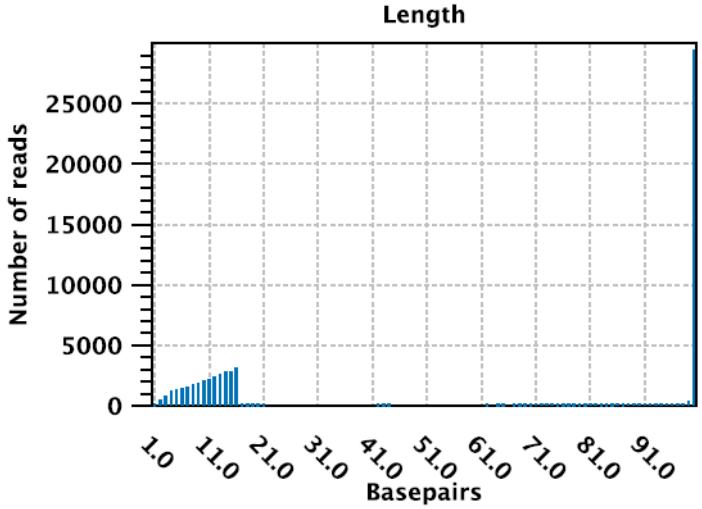
1.5 Distribution of read length



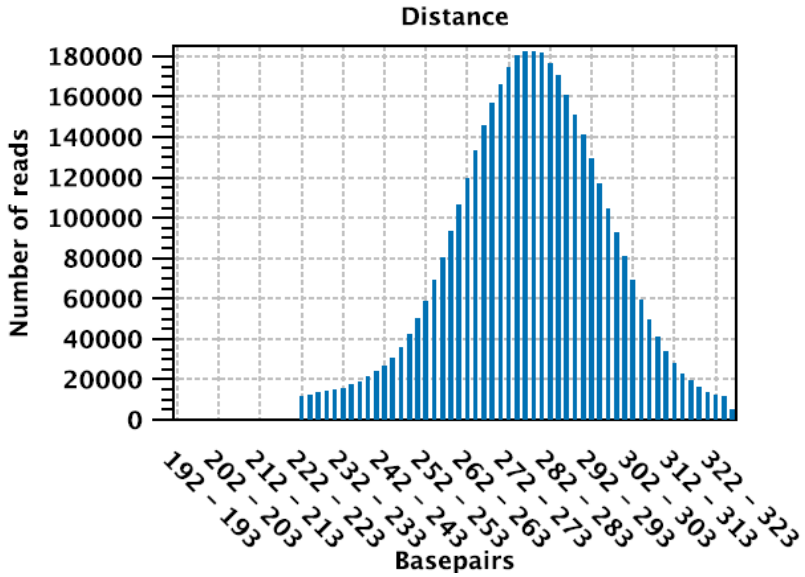
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix I: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* FSS3 with Phred score 20.

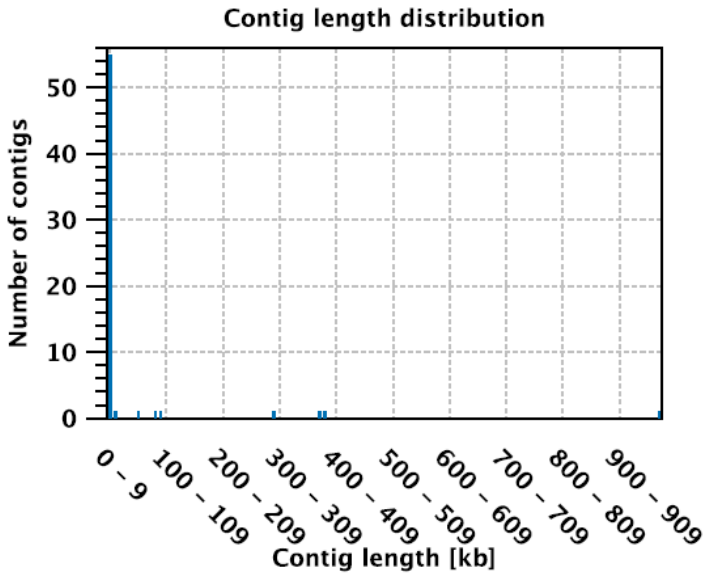
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	657,659	28.4%
Cytosine (C)	513,758	22.2%
Guanine (G)	484,314	20.9%
Thymine (T)	656,940	28.4%

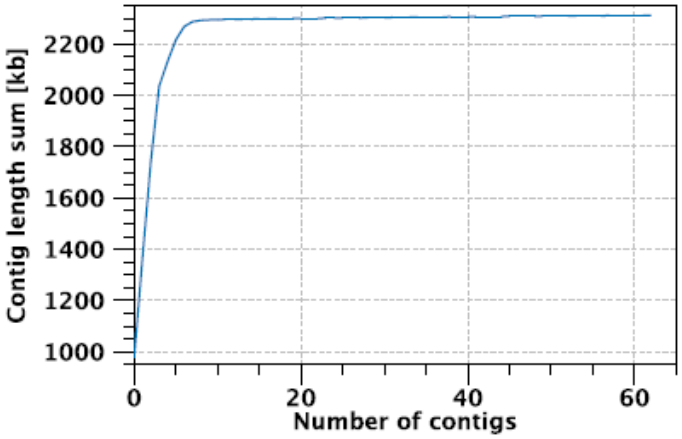
1.2 Contig measurements

Length	
N75	376,417
N50	389,092
N25	975,846
Minimum	205
Maximum	975,846
Average	36,709
Count	63

Length	
Total	2,312,671



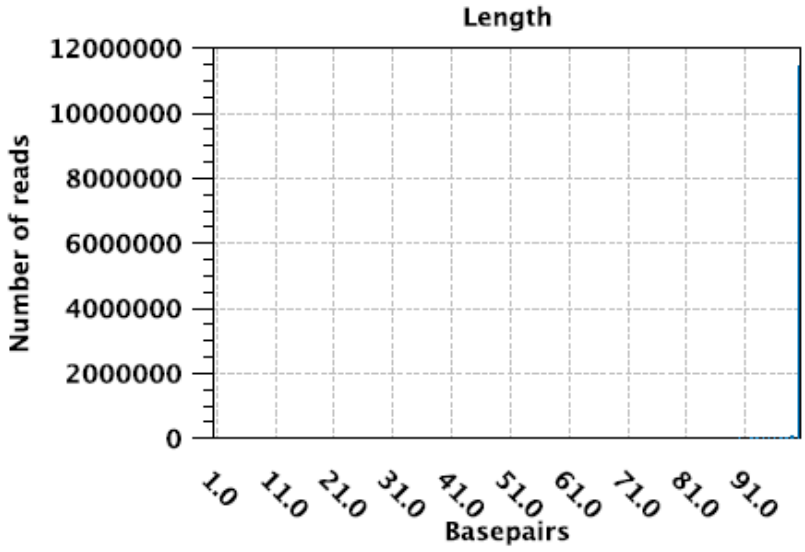
1.3 Accumulated contig lengths



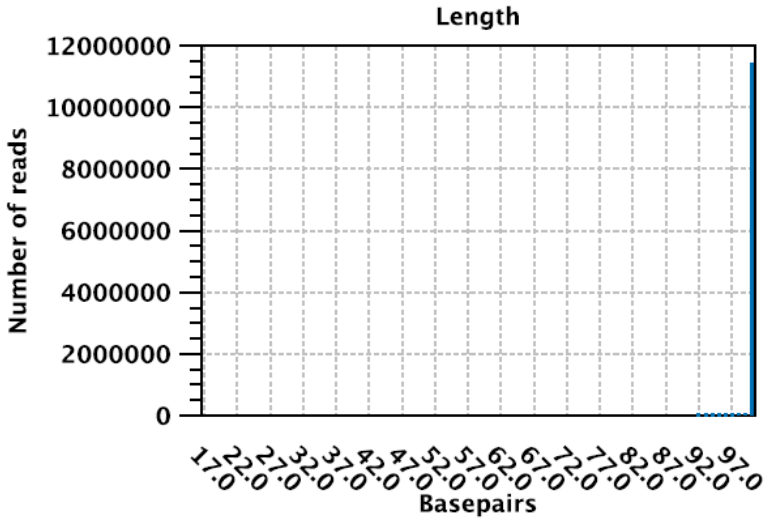
1.4 Summary statistics

	Count	Average length	Total bases
Reads	12,852,976	96.78	1,243,947,165
Matched	12,697,923	97.28	1,235,205,242
Not matched	155,053	56.38	8,741,923
Contigs	63	36,709	2,312,671
Reads in pairs	11,329,656	306.19	
Broken paired reads	1,368,267	95.82	

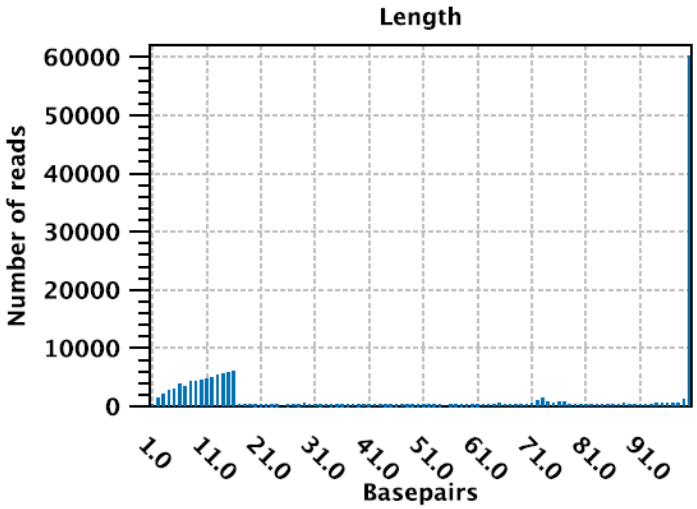
1.5 Distribution of read length



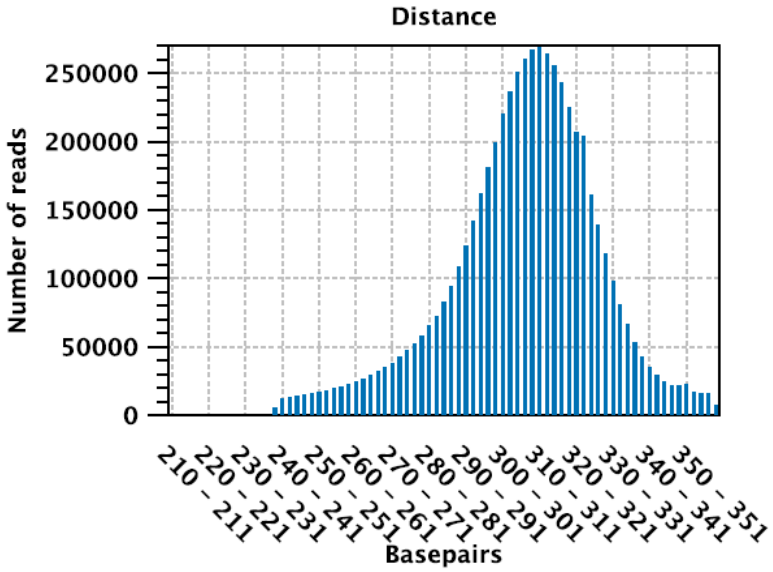
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix J: The CLC Genomics Workbench de novo assembly summary report of *S. sanguinis* FSS4 with Phred score 20.

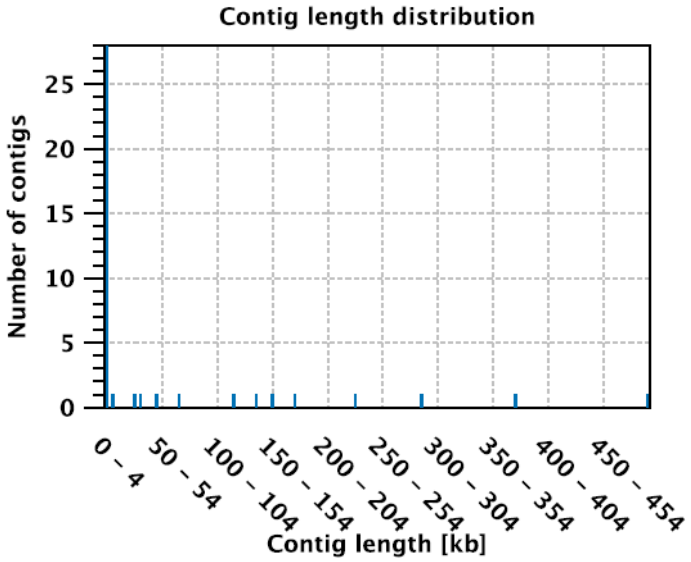
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	641,233	29.8%
Cytosine (C)	440,239	20.5%
Guanine (G)	433,615	20.2%
Thymine (T)	636,773	29.6%

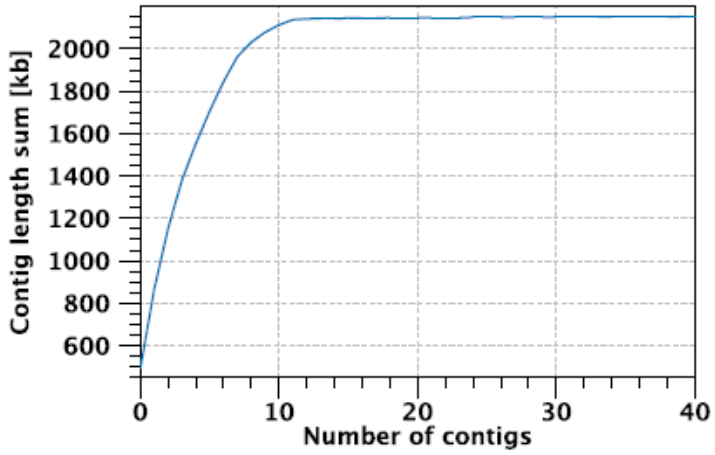
1.2 Contig measurements

Length	
N75	153,822
N50	286,373
N25	374,027
Minimum	204
Maximum	491,752
Average	52,484
Count	41

Length	
Total	2,151,860



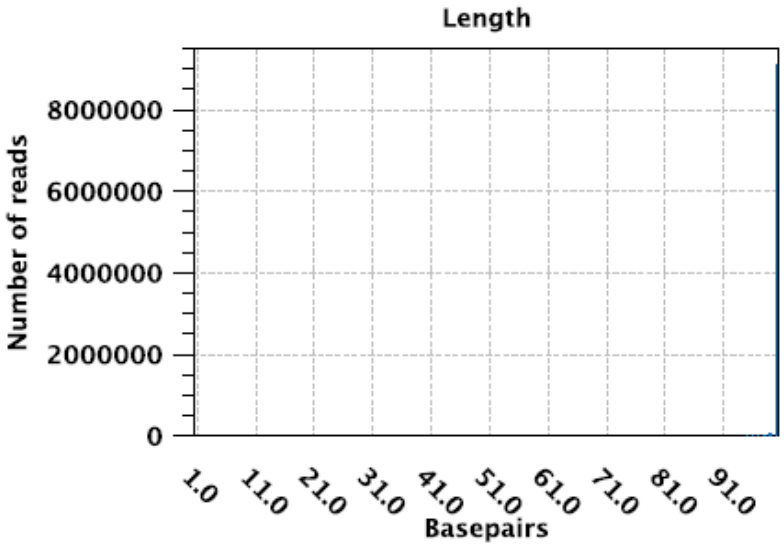
1.3 Accumulated contig lengths



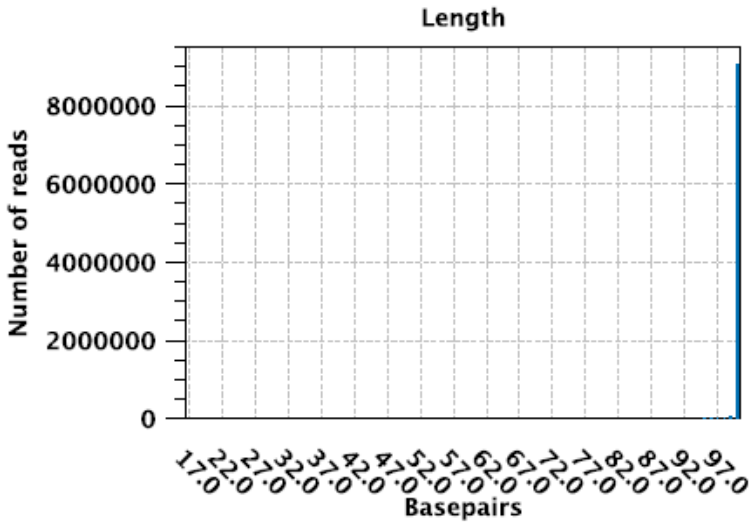
1.4 Summary statistics

	Count	Average length	Total bases
Reads	10,058,360	97.21	977,760,943
Matched	9,961,076	97.6	972,212,551
Not matched	97,284	57.03	5,548,392
Contigs	41	52,484	2,151,860
Reads in pairs	8,512,384	340.45	
Broken paired reads	1,448,692	96.94	

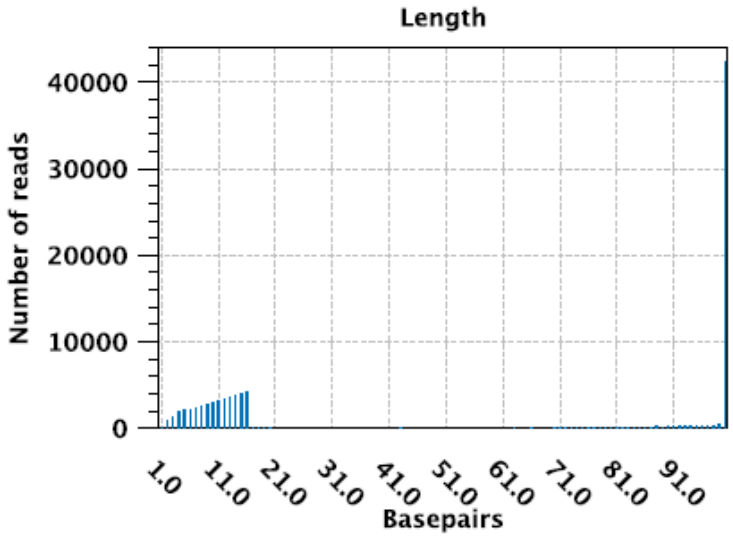
1.5 Distribution of read length



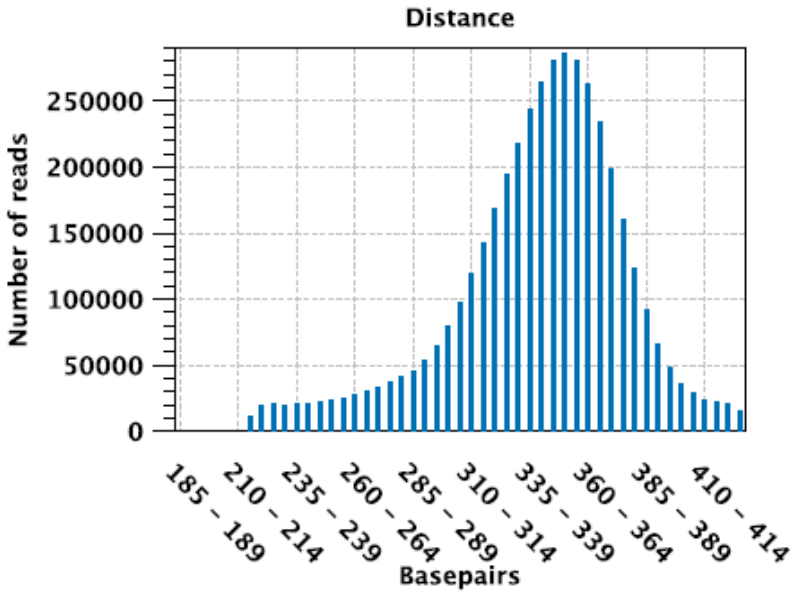
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix K: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* FSS8 with Phred score 20.

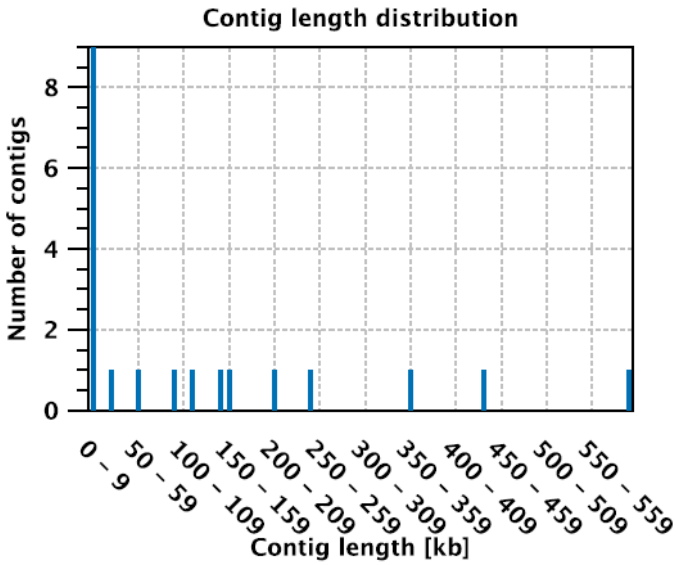
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	693,957	28.6%
Cytosine (C)	510,508	21.0%
Guanine (G)	536,520	22.1%
Thymine (T)	688,276	28.3%

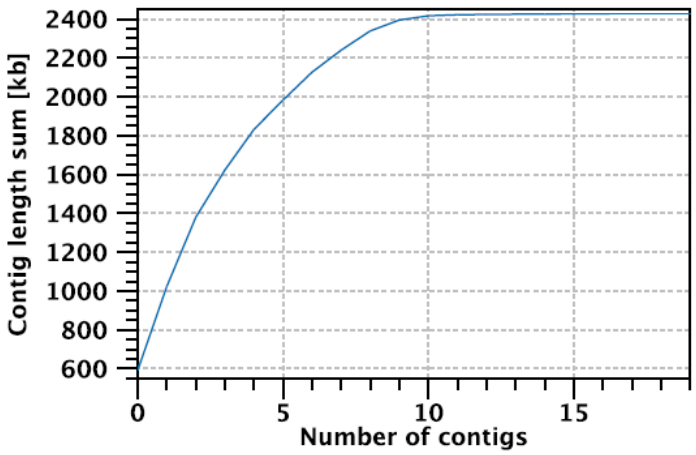
1.2 Contig measurements

N75	209,248
N50	356,680
N25	432,321
Minimum	205
Maximum	590,673
Average	121,463
Count	20

Total	2,429,261



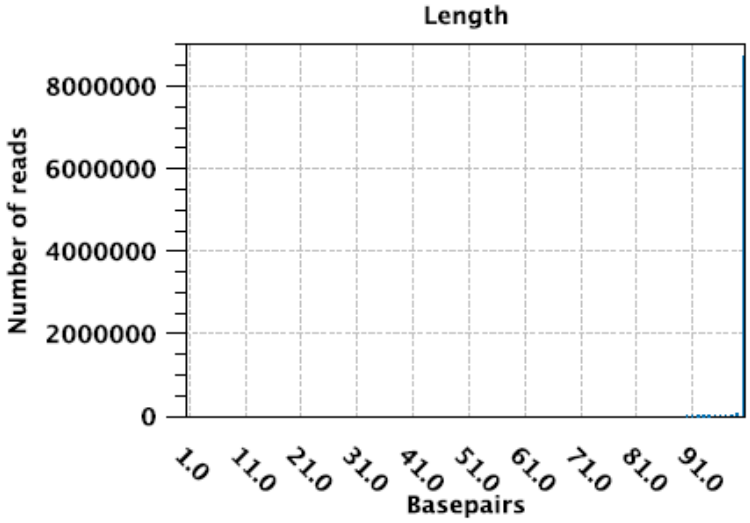
1.3 Accumulated contig lengths



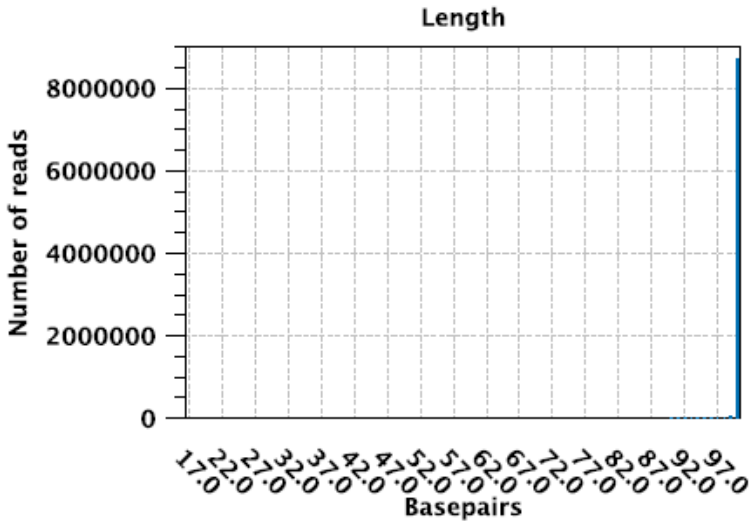
1.4 Summary statistics

	Count	Average length	Total bases
Reads	9,834,566	96.64	950,434,170
Matched	9,751,845	97.12	947,123,396
Not matched	82,721	40.02	3,310,774
Contigs	20	121,463	2,429,261
Reads in pairs	8,782,296	328.51	
Broken paired reads	969,549	95.51	

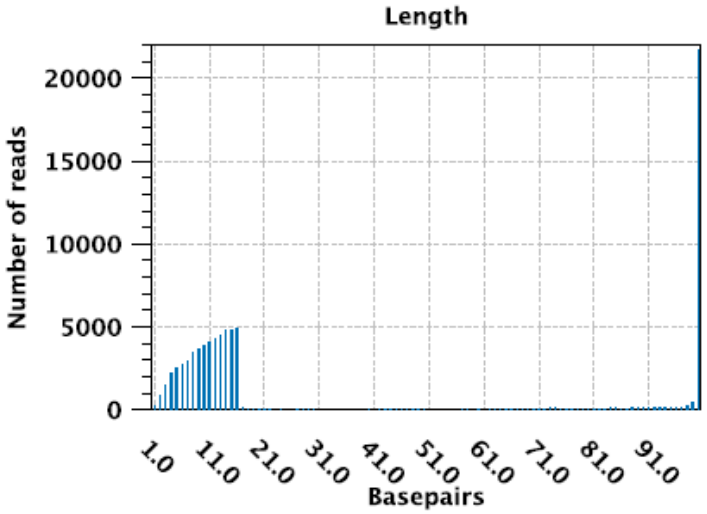
1.5 Distribution of read length



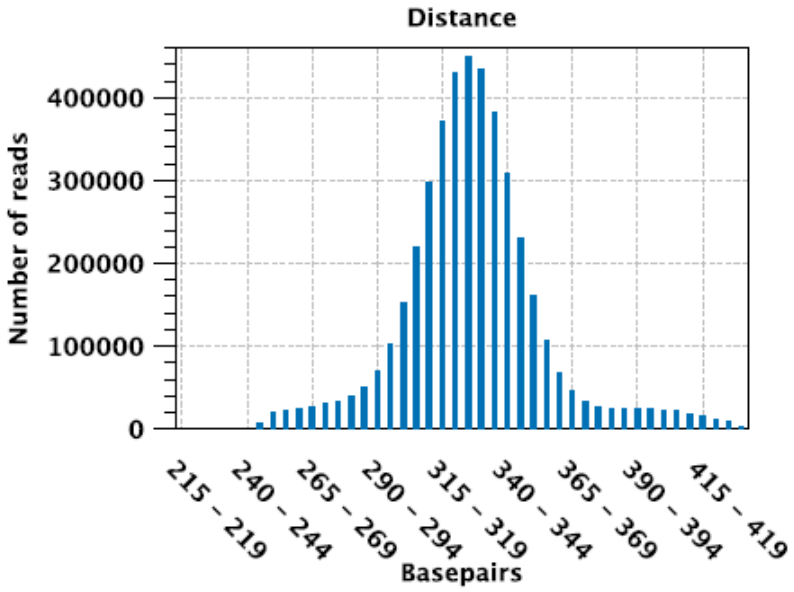
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix L: The CLC Genomics Workbench de novo assembly summary report of *S. sanguinis* FSS9 with Phred score 20.

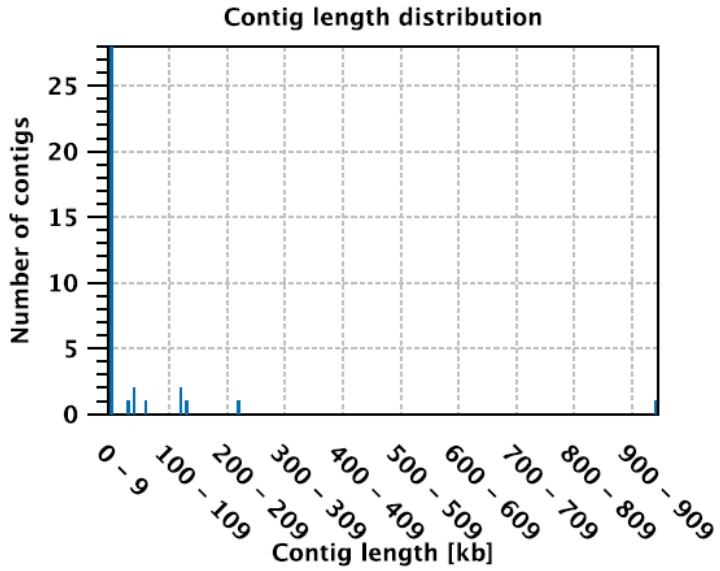
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	527,553	29.4%
Cytosine (C)	369,037	20.6%
Guanine (G)	370,496	20.7%
Thymine (T)	525,908	29.3%

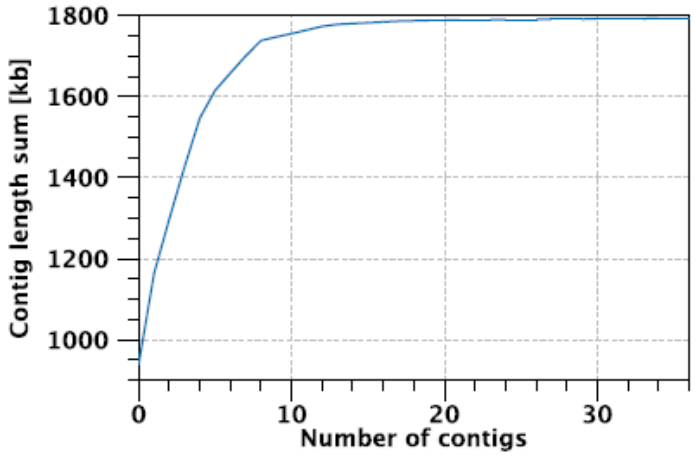
1.2 Contig measurements

Length	
N75	128,495
N50	940,267
N25	940,267
Minimum	201
Maximum	940,267
Average	48,459
Count	37

Length	
Total	1,792,994



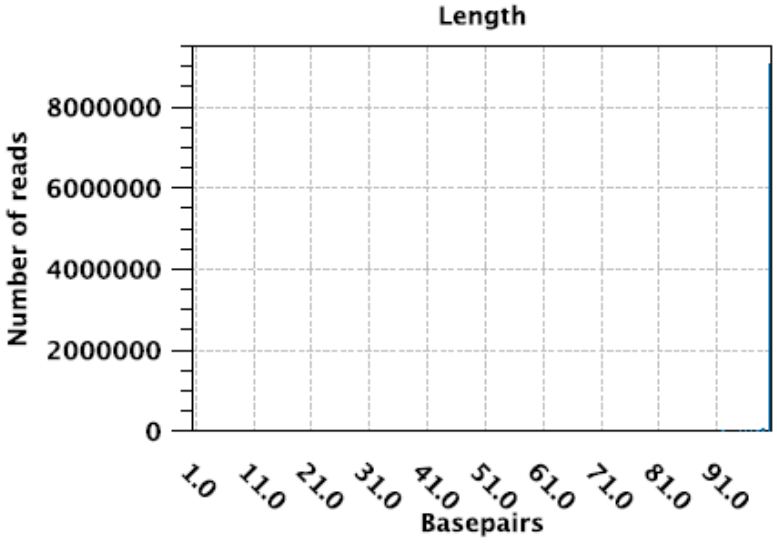
1.3 Accumulated contig lengths



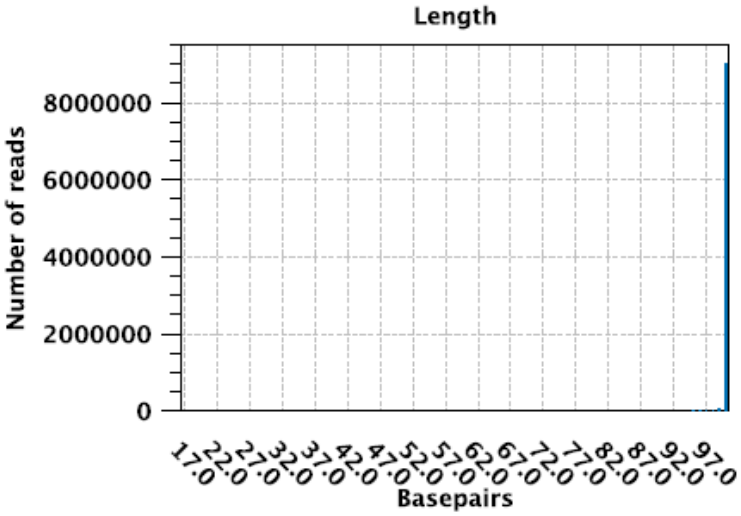
1.4 Summary statistics

	Count	Average length	Total bases
Reads	10,099,826	96.69	976,586,381
Matched	9,995,959	97.23	971,873,347
Not matched	103,867	45.38	4,713,034
Contigs	37	48,459	1,792,994
Reads in pairs	8,675,744	304.71	
Broken paired reads	1,320,215	95.74	

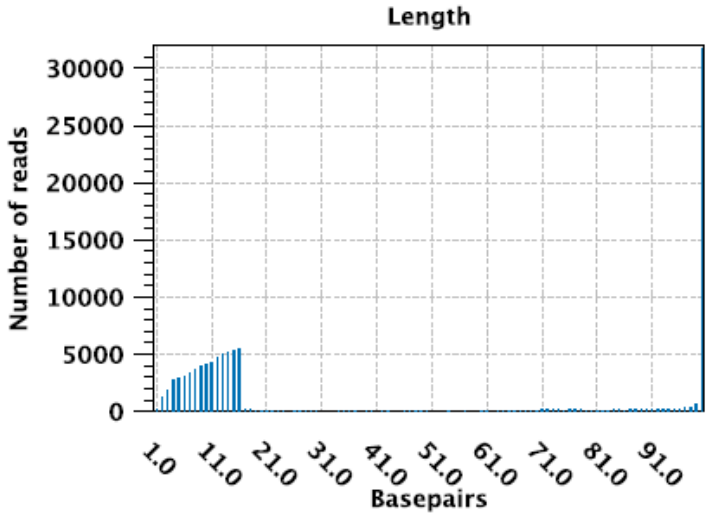
1.5 Distribution of read length



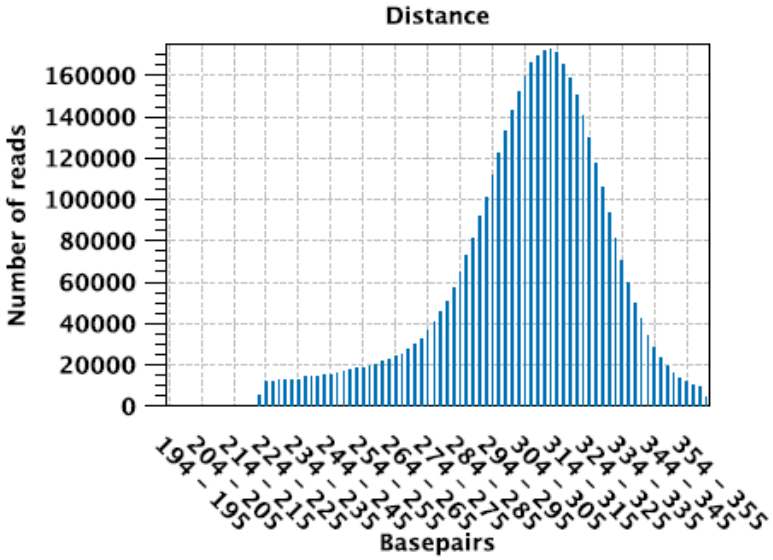
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix M: The CLC Genomics Workbench de novo assembly summary report of *S. tigurinus* JPIBVI with Phred score 20.

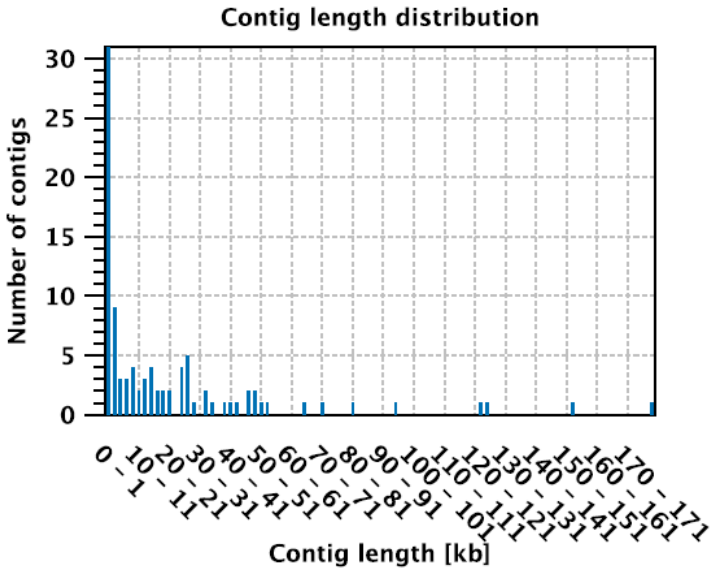
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	574,103	28.8%
Cytosine (C)	414,363	20.8%
Guanine (G)	430,023	21.6%
Thymine (T)	573,364	28.8%

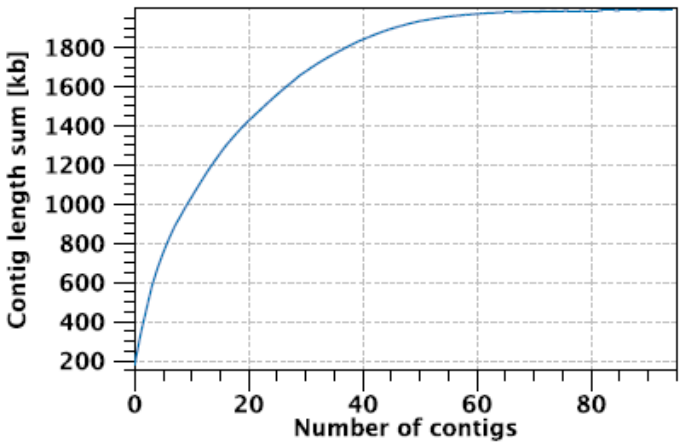
1.2 Contig measurements

Length	
N75	27,015
N50	48,467
N25	122,398
Minimum	203
Maximum	178,005
Average	20,967
Count	95

Length	
Total	1,991,853



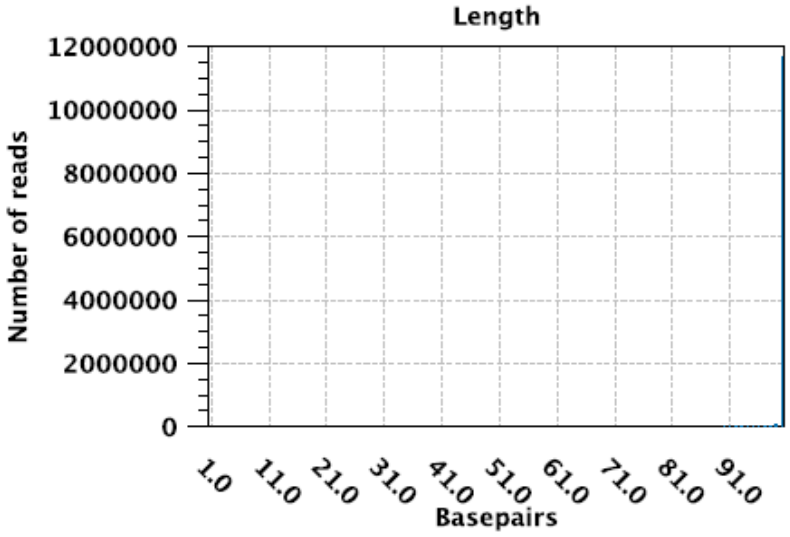
1.3 Accumulated contig lengths



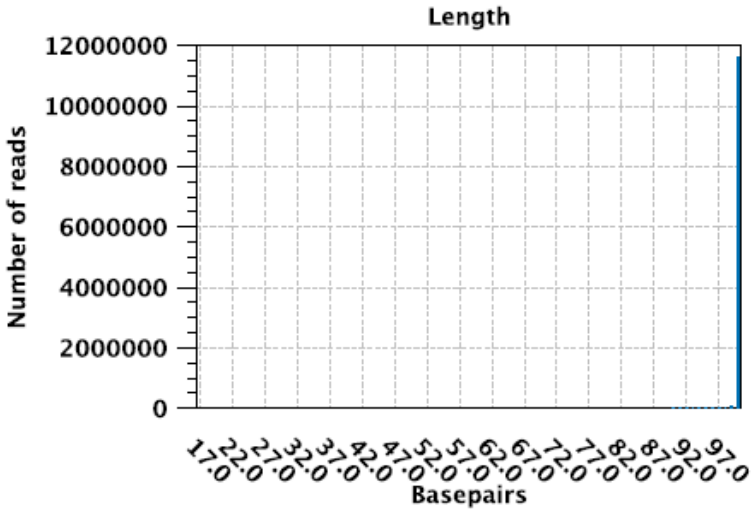
1.4 Summary statistics

	Count	Average length	Total bases
Reads	13,207,950	96.37	1,272,822,259
Matched	13,024,866	96.93	1,262,531,409
Not matched	183,084	56.21	10,290,850
Contigs	95	20,966	1,991,853
Reads in pairs	11,399,280	333.08	
Broken paired reads	1,625,586	95.37	

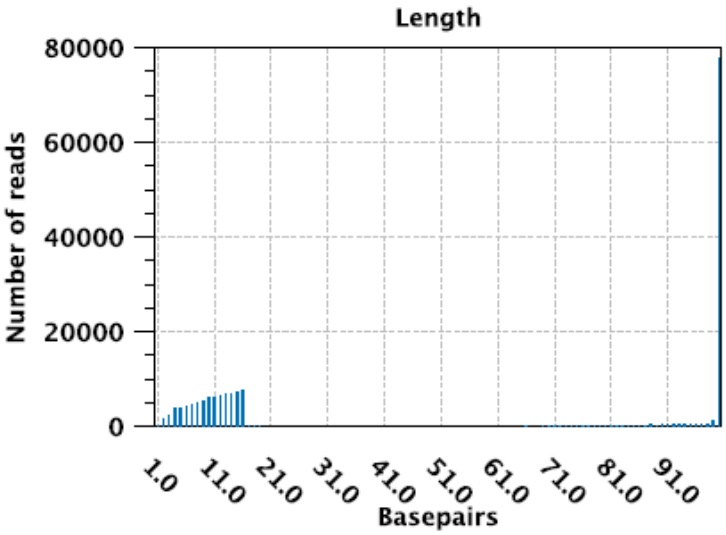
1.5 Distribution of read length



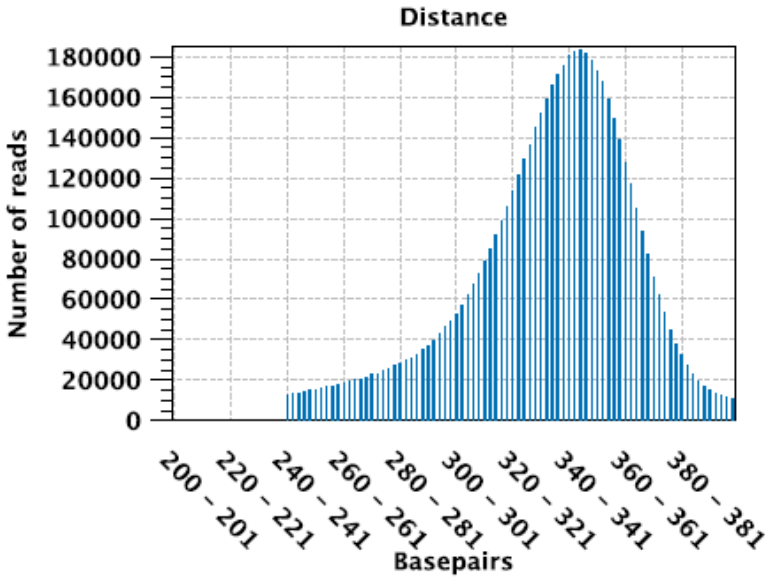
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix N: The CLC Genomics Workbench de novo assembly summary report of *S. oligofermentans* JPIIBBV4 with Phred score 20.

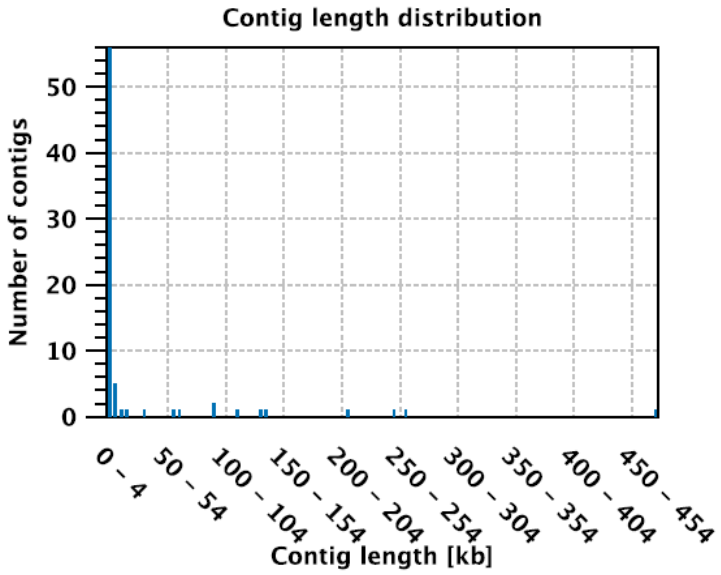
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	584,254	29.4%
Cytosine (C)	401,961	20.2%
Guanine (G)	413,964	20.8%
Thymine (T)	589,966	29.6%

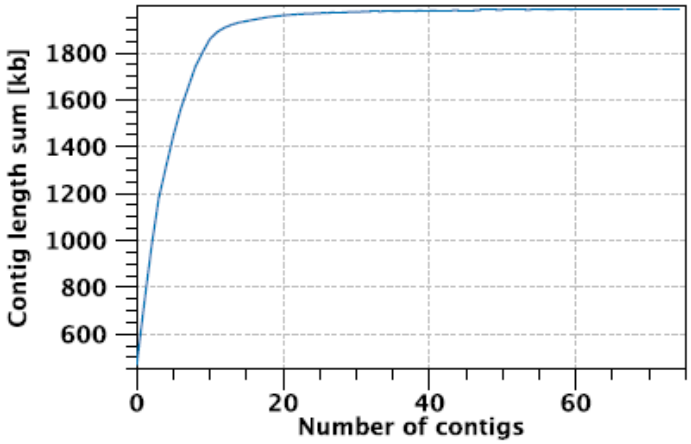
1.2 Contig measurements

Length	
N75	113,998
N50	209,178
N25	255,496
Minimum	210
Maximum	470,303
Average	26,535
Count	75

Length	
Total	1,990,145



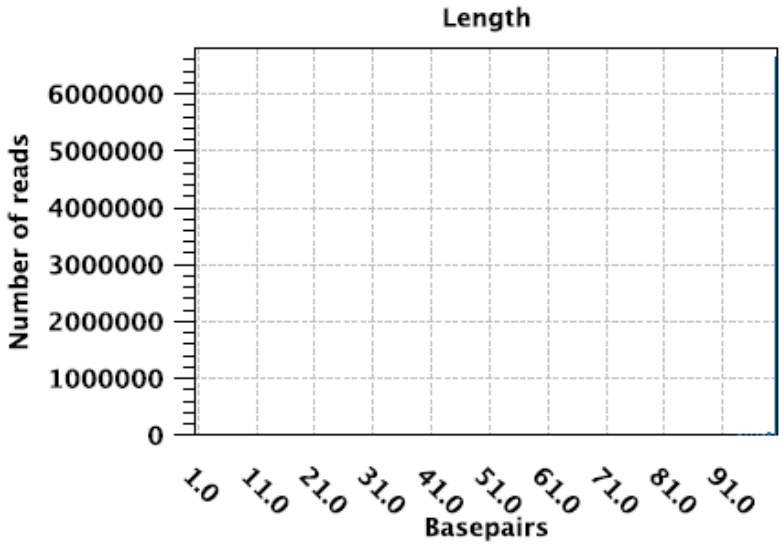
1.3 Accumulated contig lengths



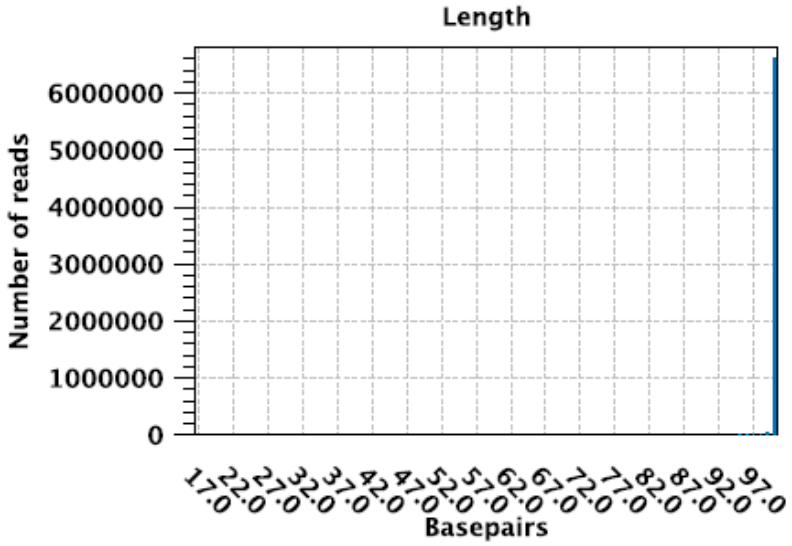
1.4 Summary statistics

	Count	Average length	Total bases
Reads	7,409,458	96.8	717,270,752
Matched	7,326,622	97.32	713,062,641
Not matched	82,836	50.8	4,208,111
Contigs	75	26,535	1,990,145
Reads in pairs	6,623,600	291.8	
Broken paired reads	703,022	95.03	

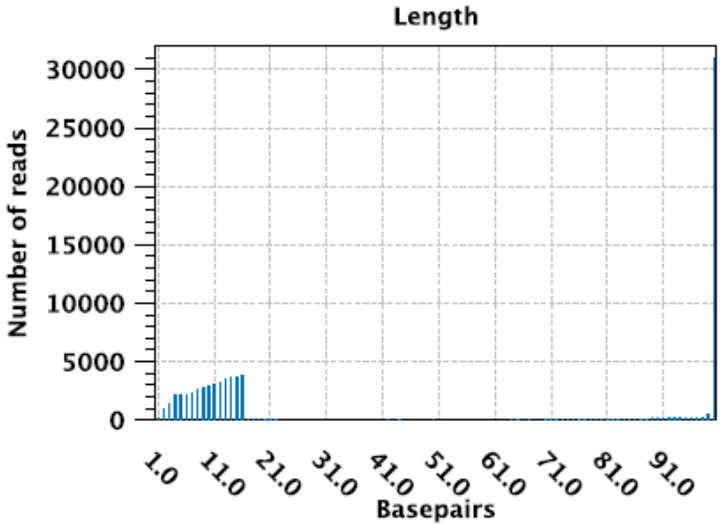
1.5 Distribution of read length



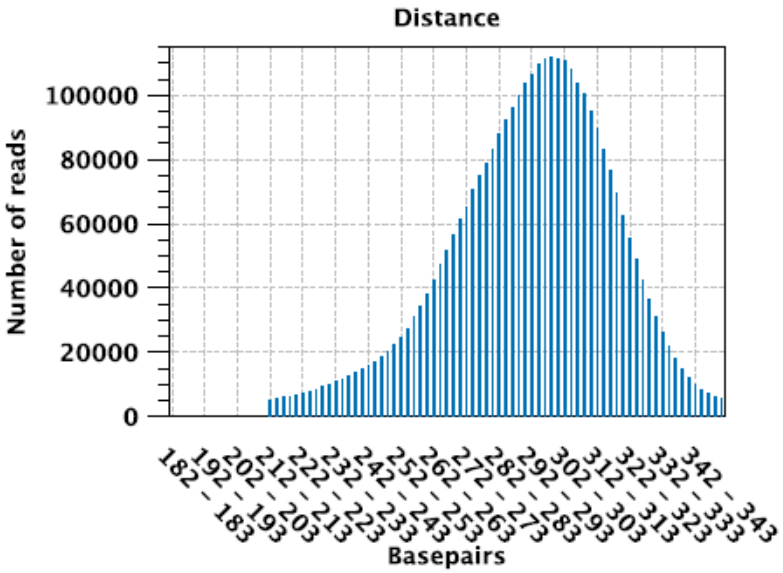
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix O: The CLC Genomics Workbench de novo assembly summary report of *S. oralis* JPIIV3 with Phred score 20.

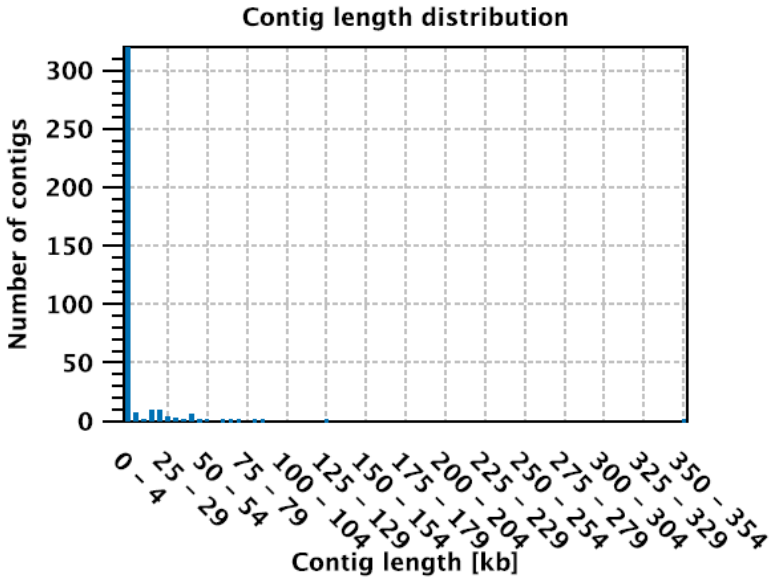
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	602,584	28.7%
Cytosine (C)	449,754	21.4%
Guanine (G)	440,941	21.0%
Thymine (T)	604,404	28.8%

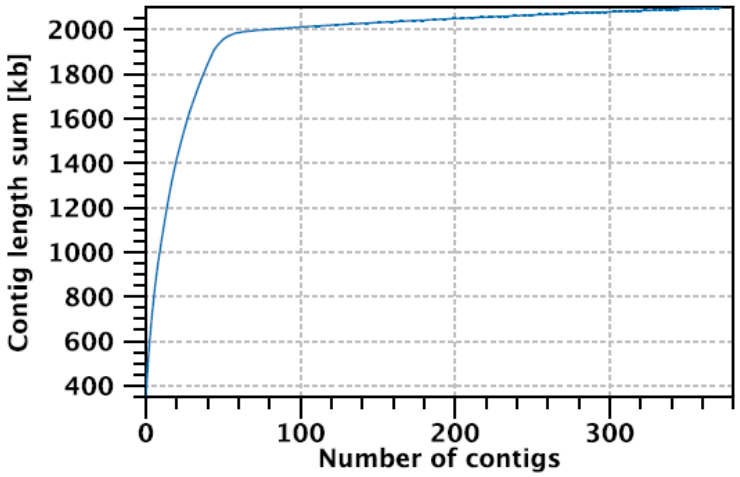
1.2 Contig measurements

N75	24,083
N50	44,211
N25	87,556
Minimum	203
Maximum	353,499
Average	5,624
Count	373

Total	2,097,683



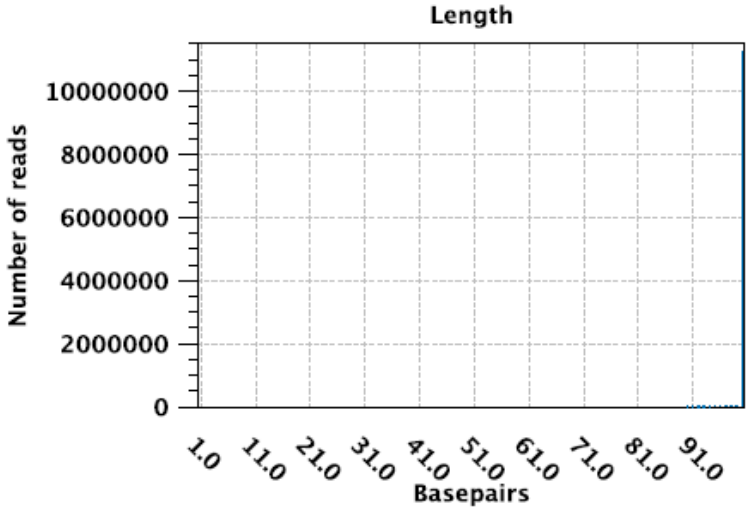
1.3 Accumulated contig lengths



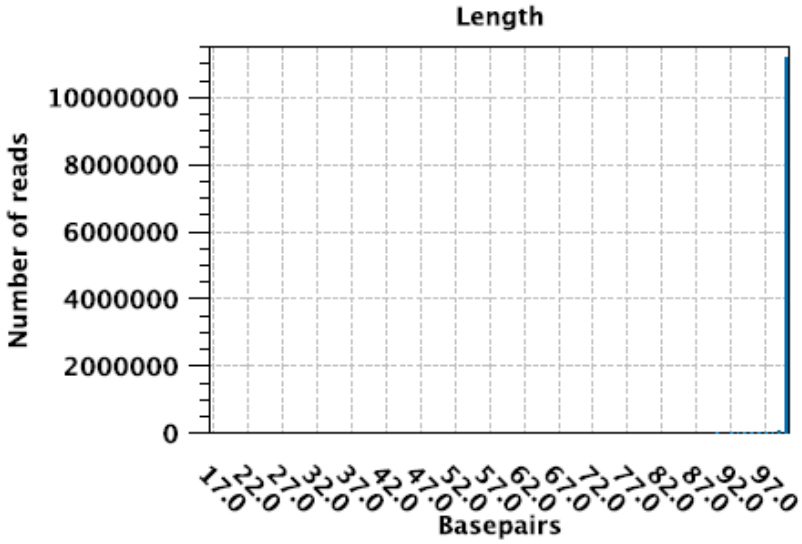
1.4 Summary statistics

	Count	Average length	Total bases
Reads	12,653,658	96.69	1,223,453,056
Matched	12,472,538	97.16	1,211,891,231
Not matched	181,120	63.84	11,561,825
Contigs	373	5,623	2,097,683
Reads in pairs	10,987,682	309.73	
Broken paired reads	1,484,856	95.58	

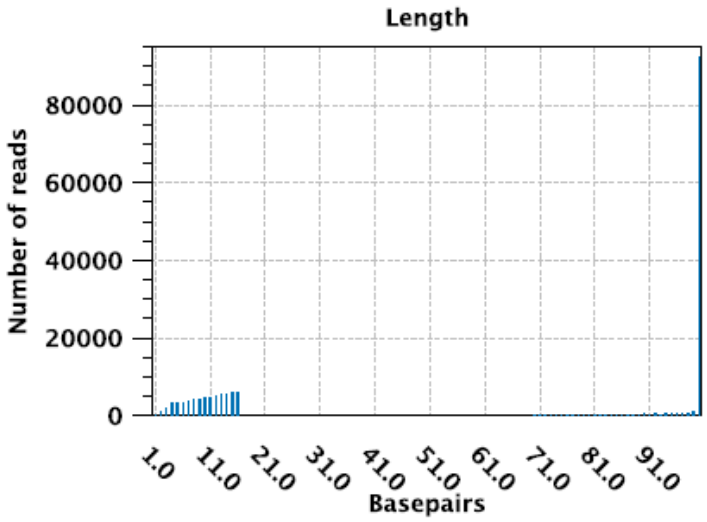
1.5 Distribution of read length



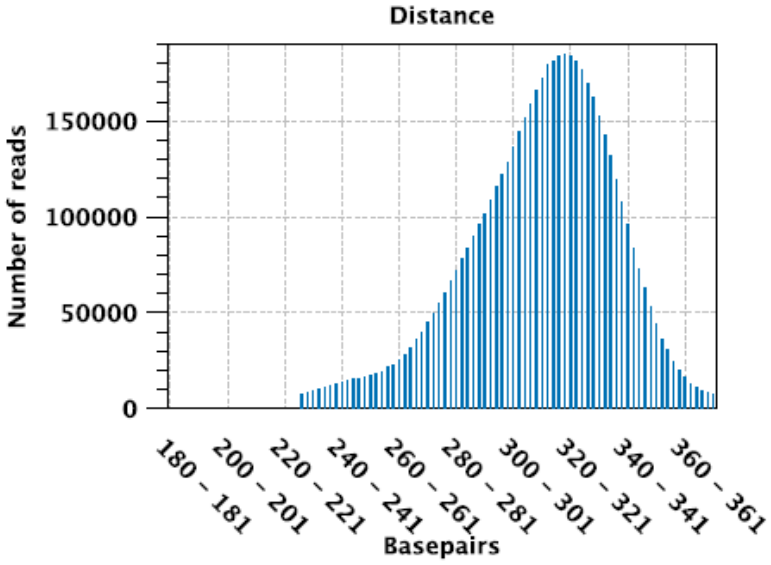
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix P: The CLC Genomics Workbench de novo assembly summary report of *S. oligofermentans* LRIIBV4 with Phred score 20.

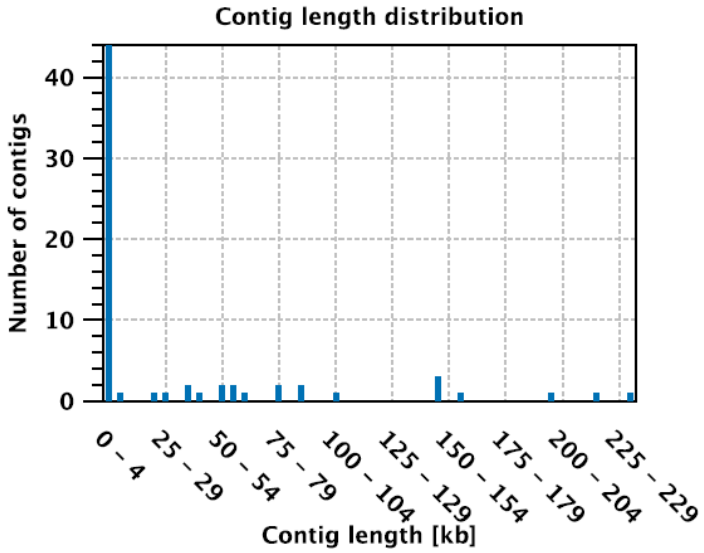
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	635,159	29.4%
Cytosine (C)	441,093	20.4%
Guanine (G)	434,083	20.1%
Thymine (T)	647,497	30.0%

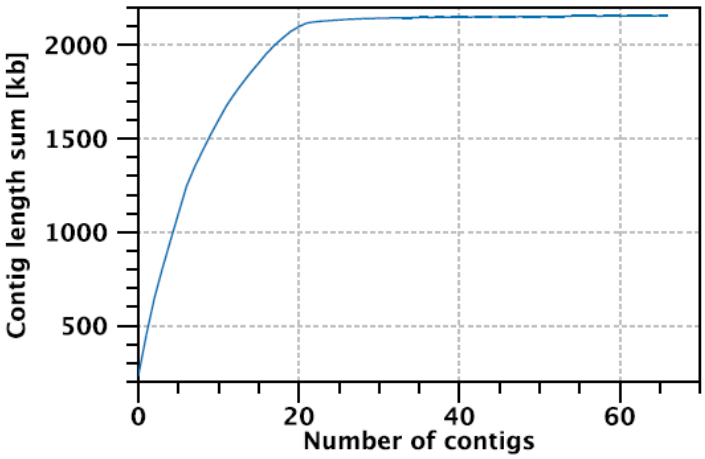
1.2 Contig measurements

N75	77,048
N50	145,888
N25	196,117
Minimum	223
Maximum	231,833
Average	32,206
Count	67

Total	2,157,832



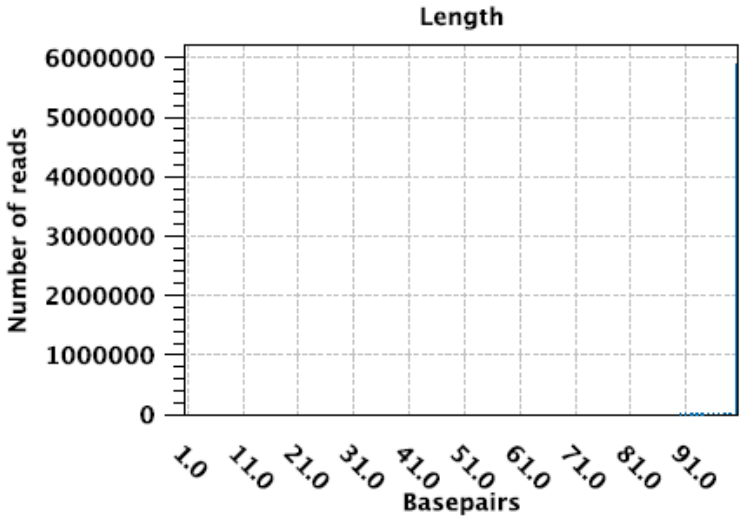
1.3 Accumulated contig lengths



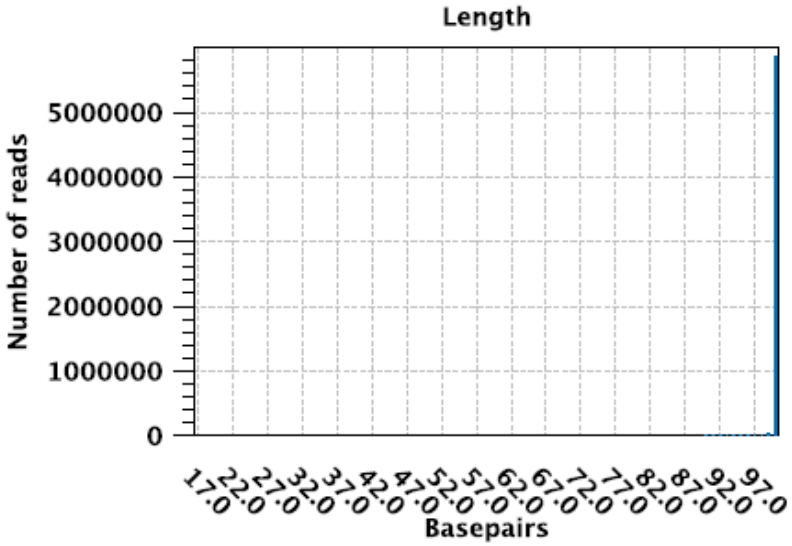
1.4 Summary statistics

	Count	Average length	Total bases
Reads	6,723,804	95.86	644,546,911
Matched	6,610,163	96.58	638,414,786
Not matched	113,641	53.96	6,132,125
Contigs	67	32,206	2,157,832
Reads in pairs	5,642,544	343.95	
Broken paired reads	967,619	92.62	

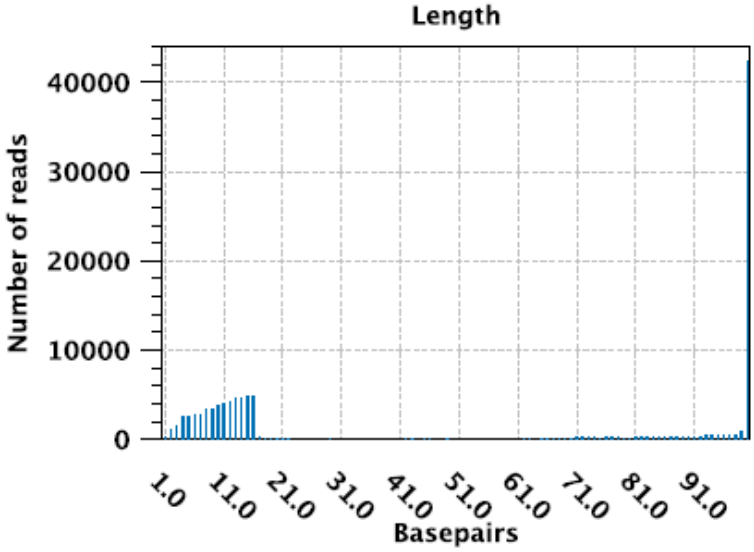
1.5 Distribution of read length



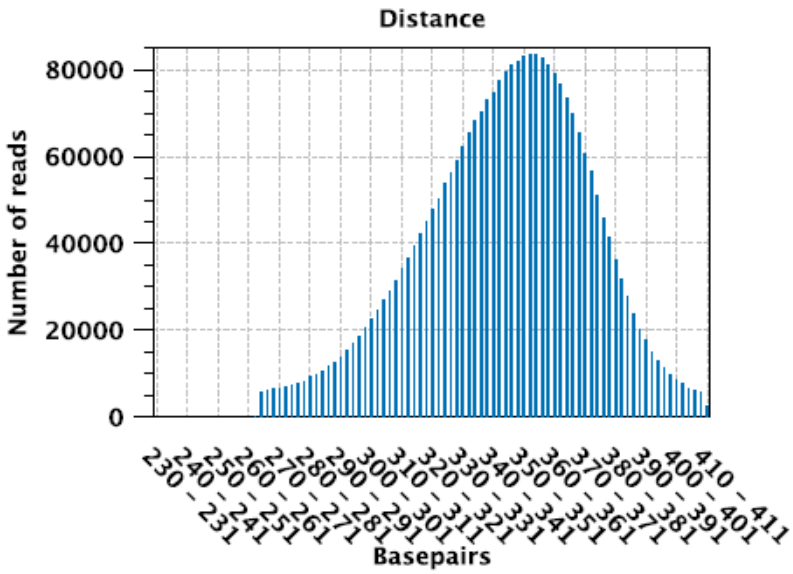
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix Q: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* M5 with Phred score 20.

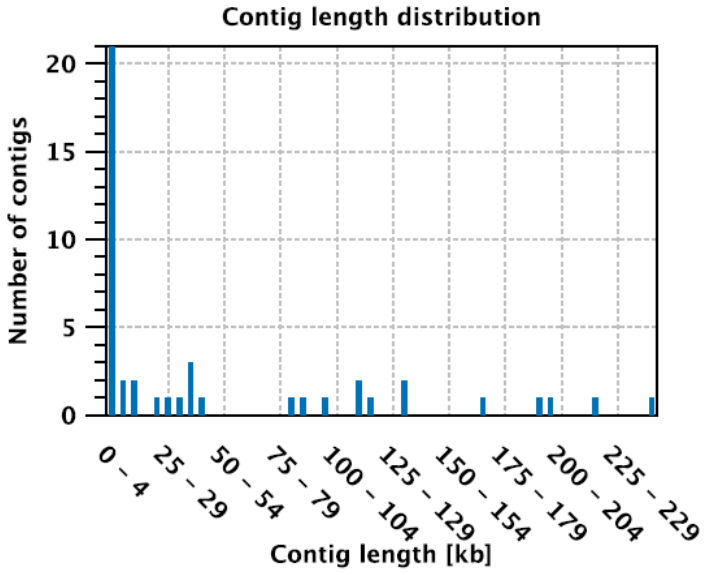
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	642,844	29.7%
Cytosine (C)	457,038	21.1%
Guanine (G)	420,154	19.4%
Thymine (T)	647,025	29.9%

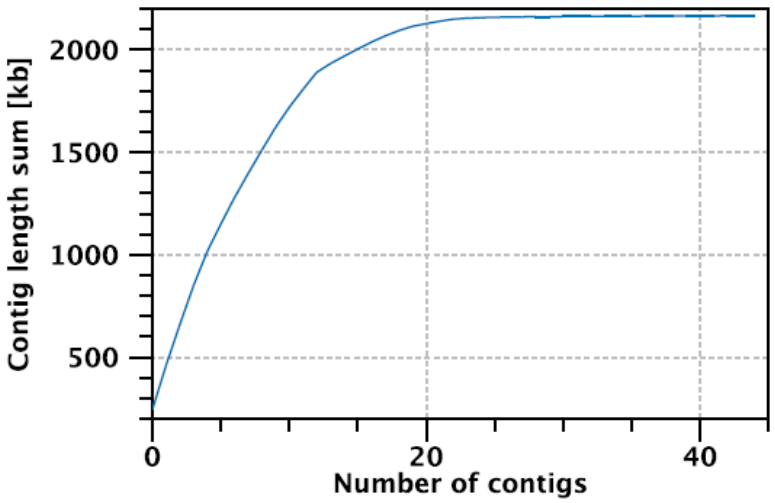
1.2 Contig measurements

N75	98,444
N50	134,448
N25	198,363
Minimum	244
Maximum	243,635
Average	48,157
Count	45

Total	2,167,061



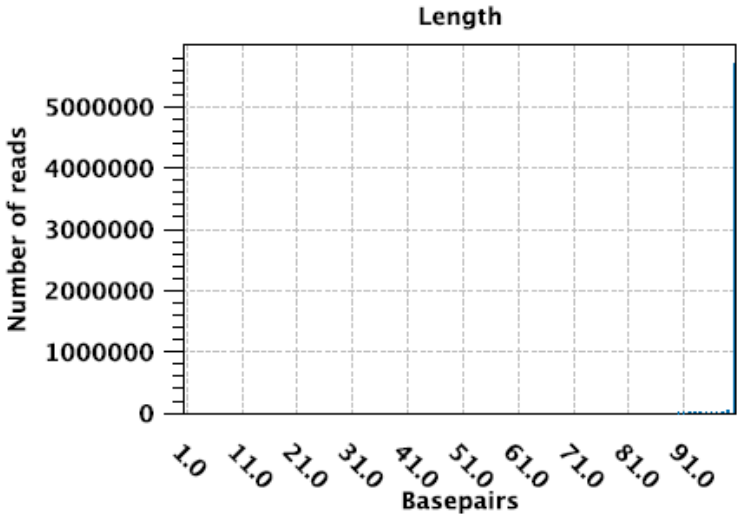
1.3 Accumulated contig lengths



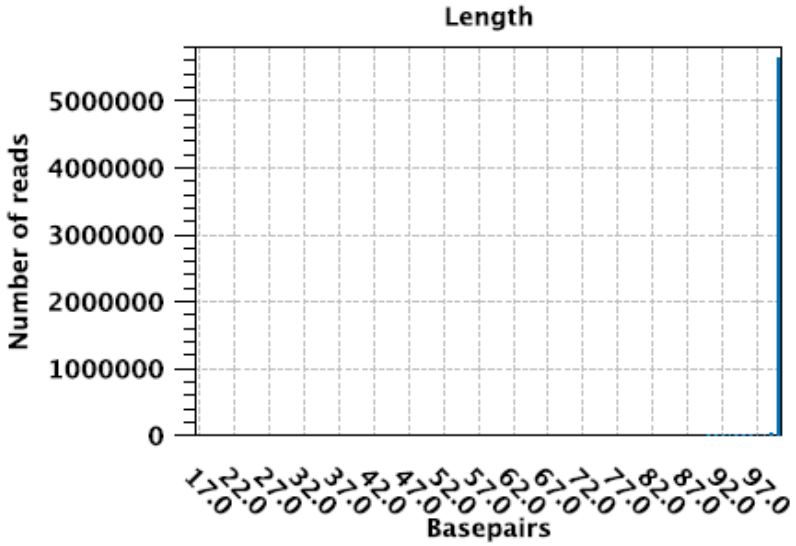
1.4 Summary statistics

	Count	Average length	Total bases
Reads	6,594,634	95.26	628,178,840
Matched	6,440,322	96.16	619,296,493
Not matched	154,312	57.56	8,882,347
Contigs	45	48,156	2,167,061
Reads in pairs	5,051,012	383.2	
Broken paired reads	1,389,310	94.04	

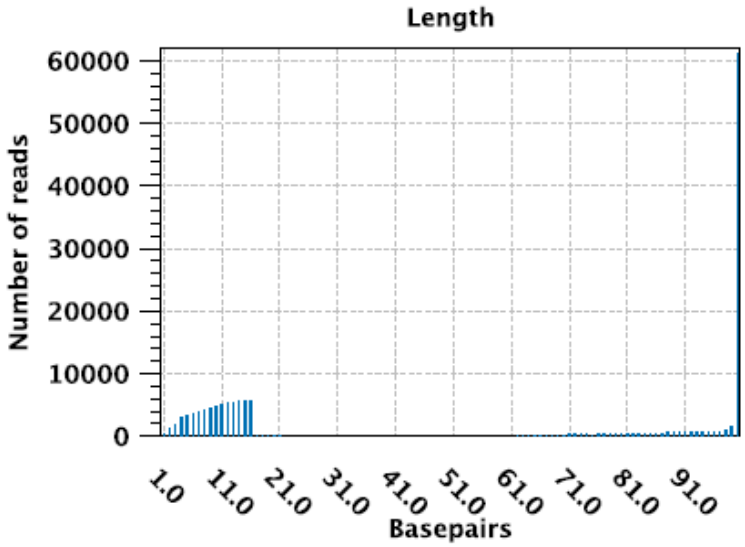
1.5 Distribution of read length



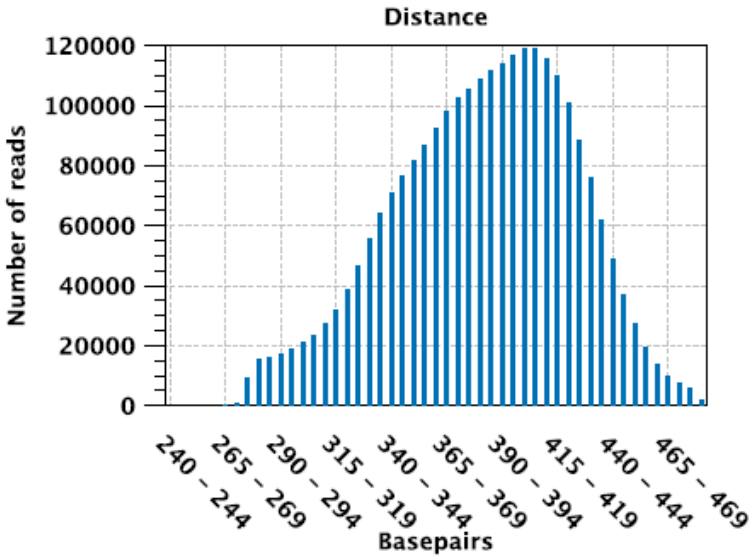
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix R: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* M99 with Phred score 20.

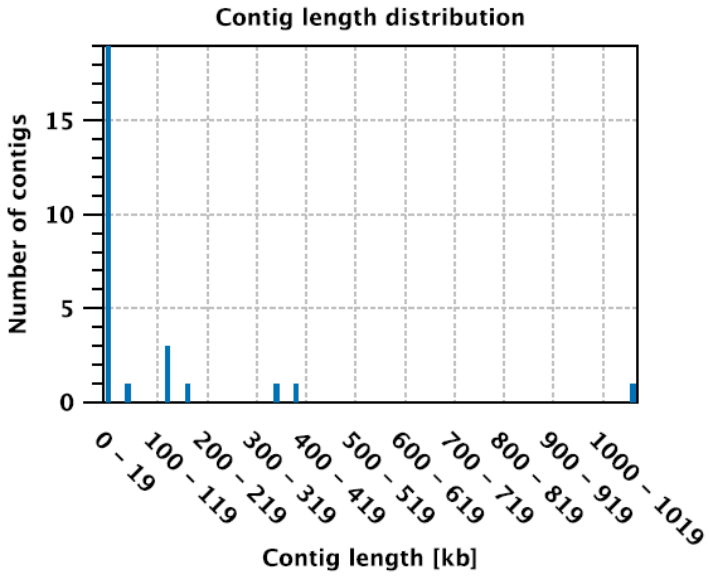
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	701,764	28.6%
Cytosine (C)	536,789	21.9%
Guanine (G)	514,566	21.0%
Thymine (T)	699,687	28.5%

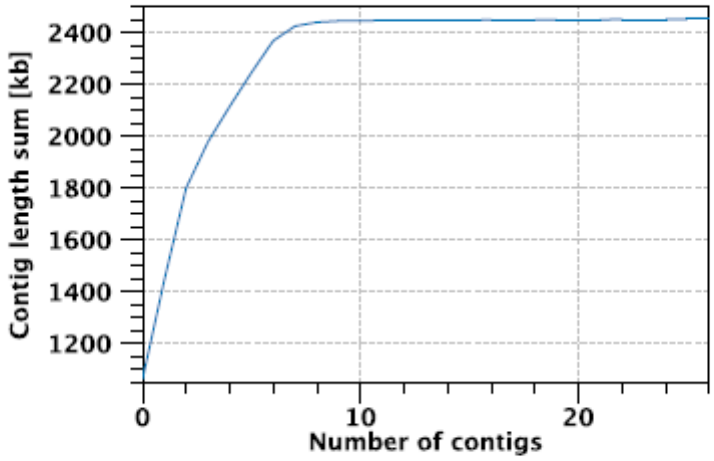
1.2 Contig measurements

Length	
N75	178,173
N50	382,788
N25	1,063,370
Minimum	257
Maximum	1,063,370
Average	90,845
Count	27

Length	
Total	2,452,806



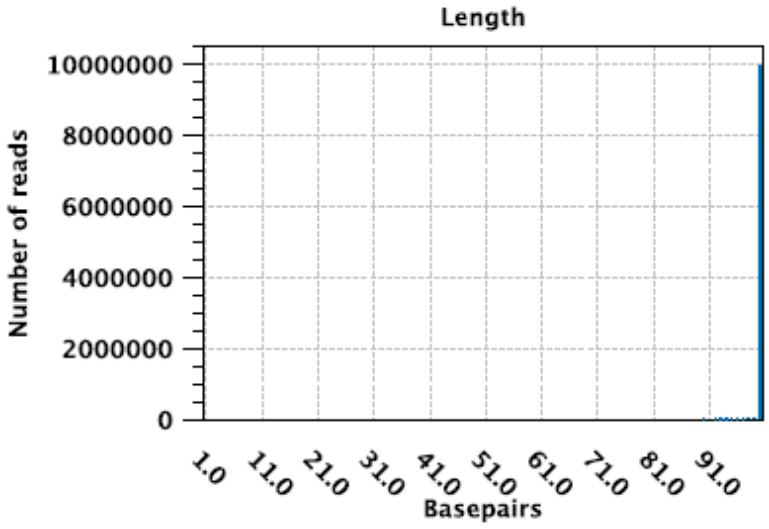
1.3 Accumulated contig lengths



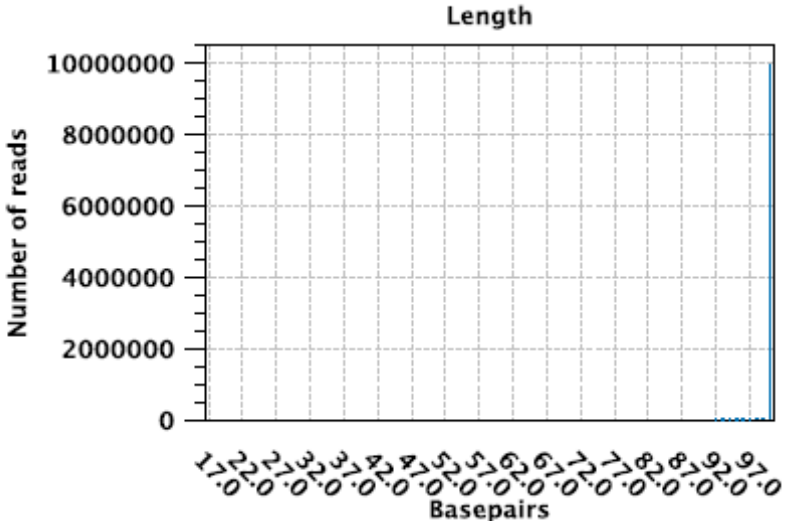
1.4 Summary statistics

	Count	Average length	Total bases
Reads	11,202,628	96.56	1,081,740,441
Matched	11,104,636	97.07	1,077,947,730
Not matched	97,992	38.7	3,792,711
Contigs	27	90,844	2,452,806
Reads in pairs	10,048,550	292.78	
Broken paired reads	1,056,086	94.53	

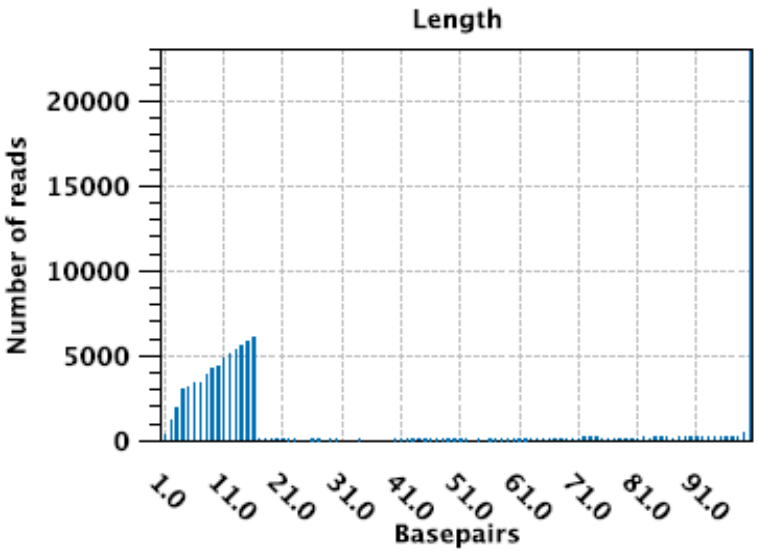
1.5 Distribution of read length



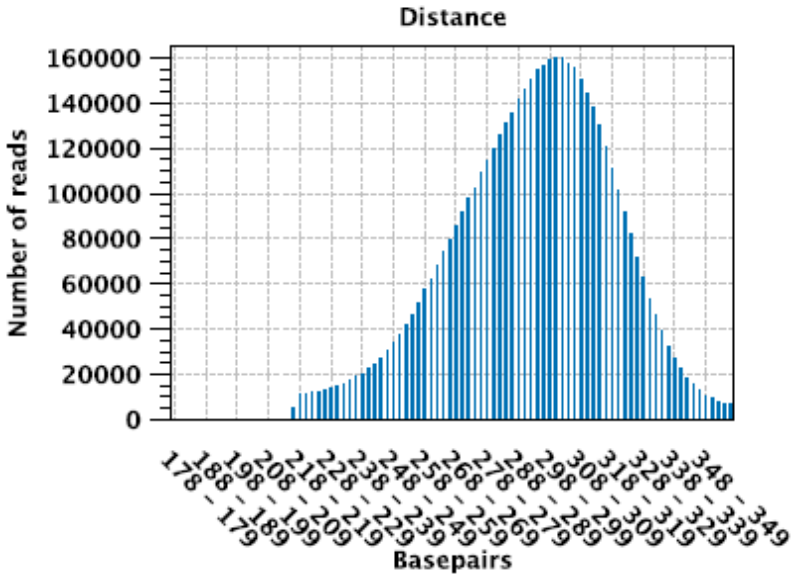
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix S: The CLC Genomics Workbench de novo assembly summary report of *S. sanguinis* MB451 with Phred score 20.

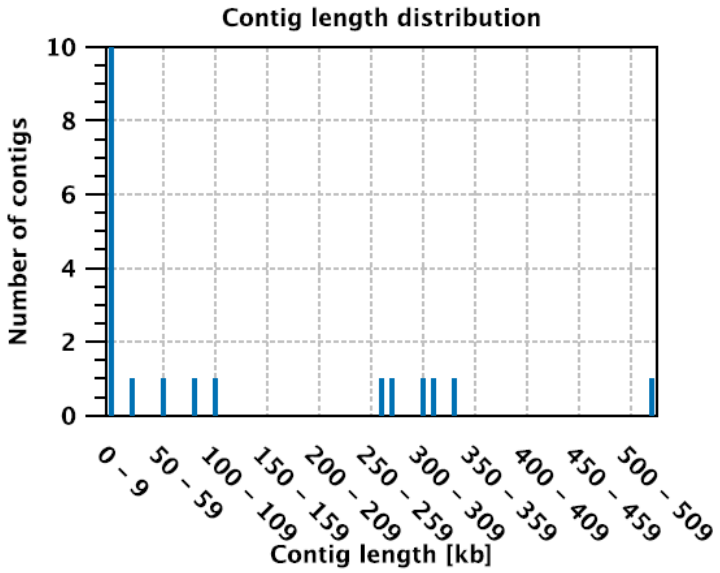
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	684,419	29.7%
Cytosine (C)	471,251	20.4%
Guanine (G)	458,982	19.9%
Thymine (T)	693,490	30.0%

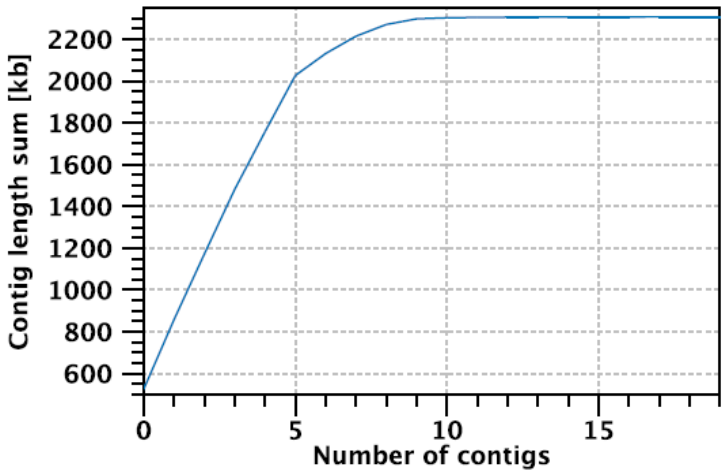
1.2 Contig measurements

N75	276,166
N50	313,888
N25	333,281
Minimum	229
Maximum	525,549
Average	115,407
Count	20

Total	2,308,142



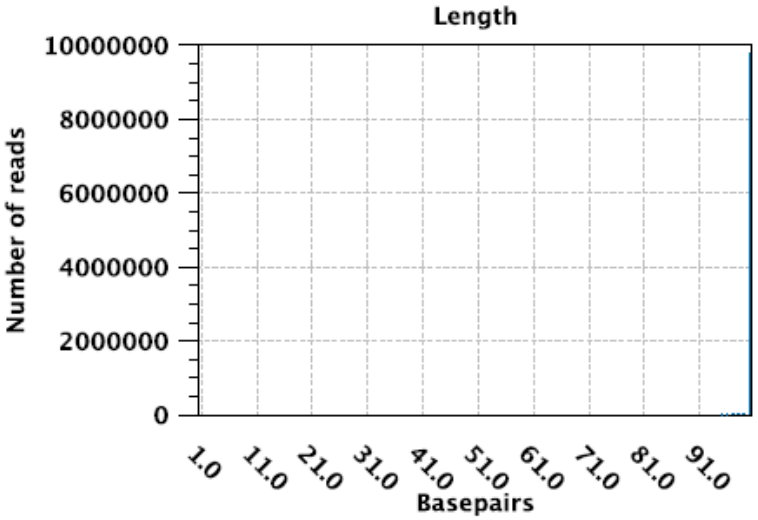
1.3 Accumulated contig lengths



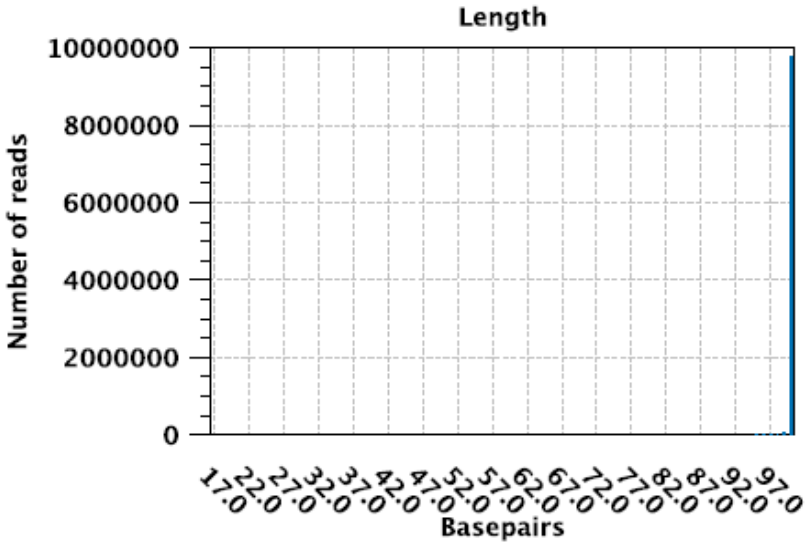
1.4 Summary statistics

	Count	Average length	Total bases
Reads	10,889,270	96.8	1,054,071,024
Matched	10,786,335	97.3	1,049,519,924
Not matched	102,935	44.21	4,551,100
Contigs	20	115,407	2,308,142
Reads in pairs	9,902,422	295.29	
Broken paired reads	883,913	94.52	

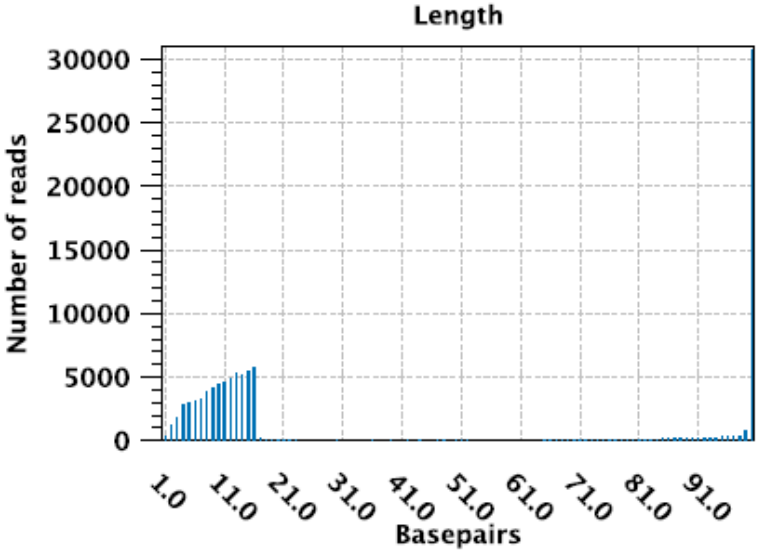
1.5 Distribution of read length



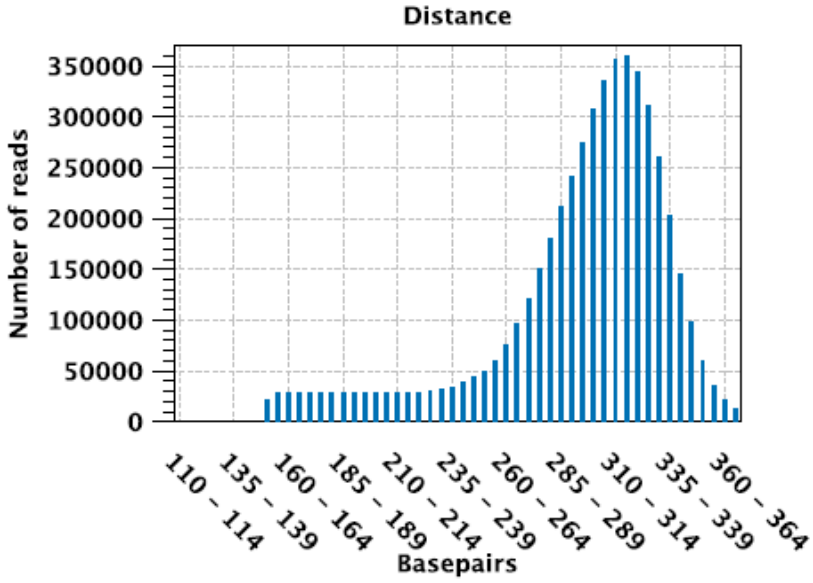
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix T: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* MB666 with Phred score 20.

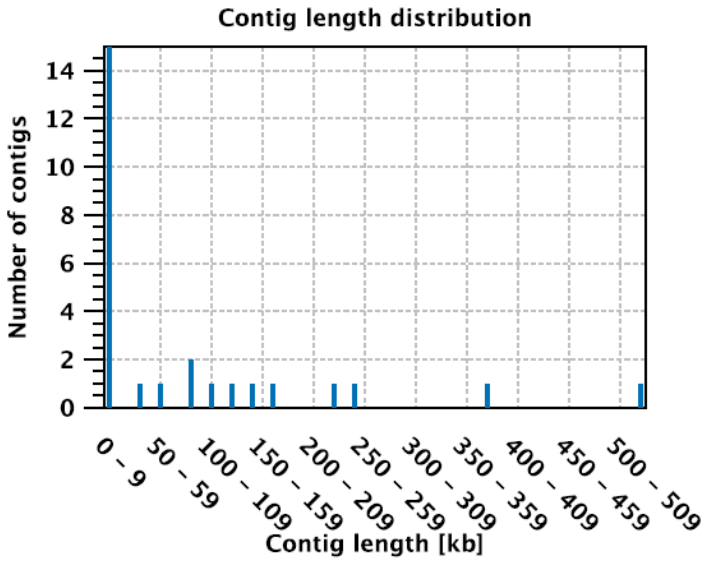
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	652,045	29.8%
Cytosine (C)	420,079	19.2%
Guanine (G)	464,812	21.3%
Thymine (T)	649,177	29.7%

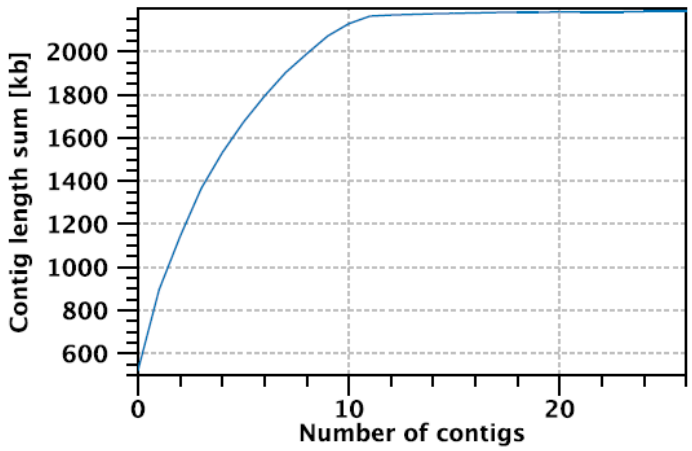
1.2 Contig measurements

N75	140,861
N50	247,835
N25	374,186
Minimum	212
Maximum	521,607
Average	80,967
Count	27

Total	2,186,113



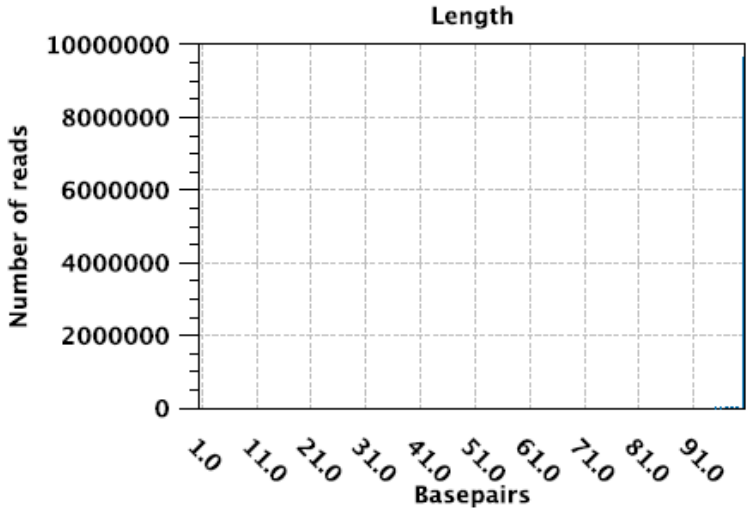
1.3 Accumulated contig lengths



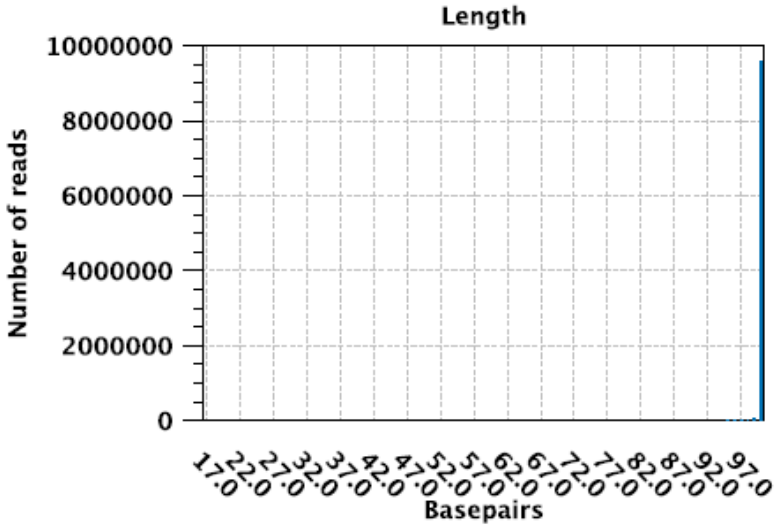
1.4 Summary statistics

	Count	Average length	Total bases
Reads	10,648,224	97.2	1,034,994,839
Matched	10,557,925	97.59	1,030,301,190
Not matched	90,299	51.98	4,693,649
Contigs	27	80,967	2,186,113
Reads in pairs	9,169,552	323.37	
Broken paired reads	1,388,373	96.43	

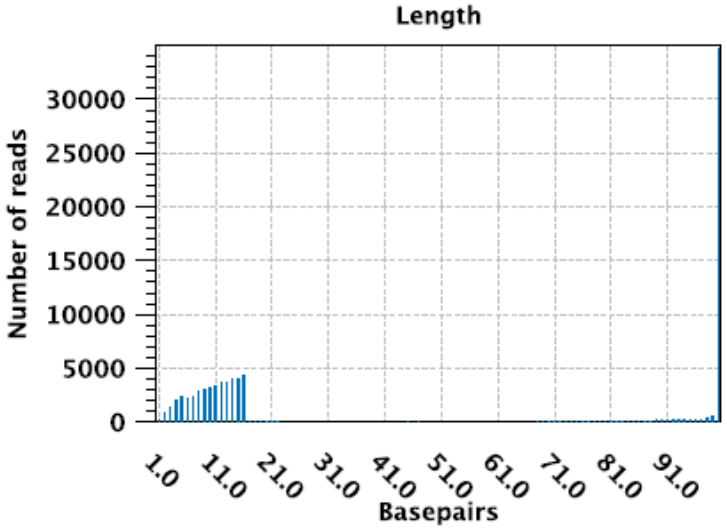
1.5 Distribution of read length



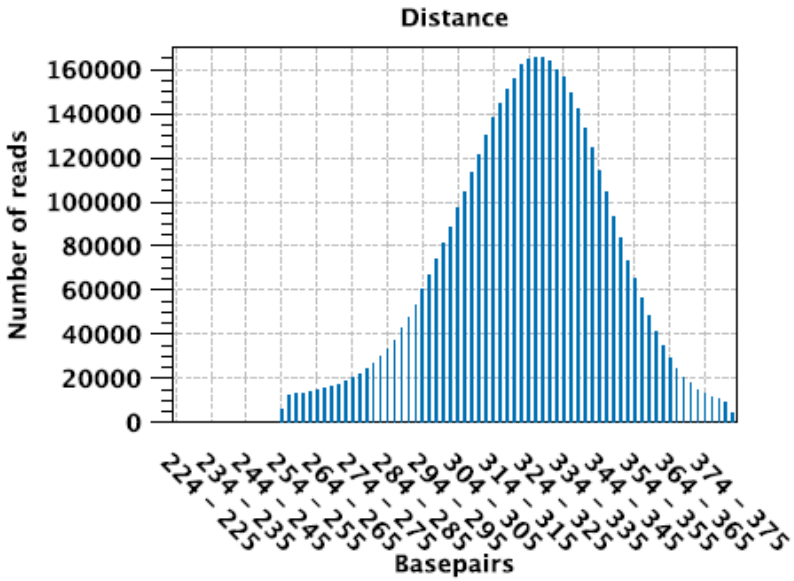
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix U: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* MW10 with Phred score 20.

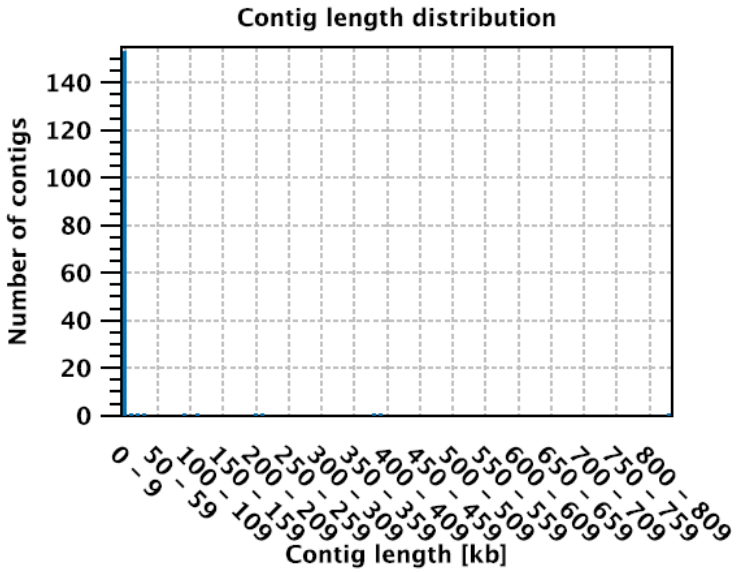
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	671,278	28.3%
Cytosine (C)	517,935	21.9%
Guanine (G)	504,924	21.3%
Thymine (T)	674,144	28.5%

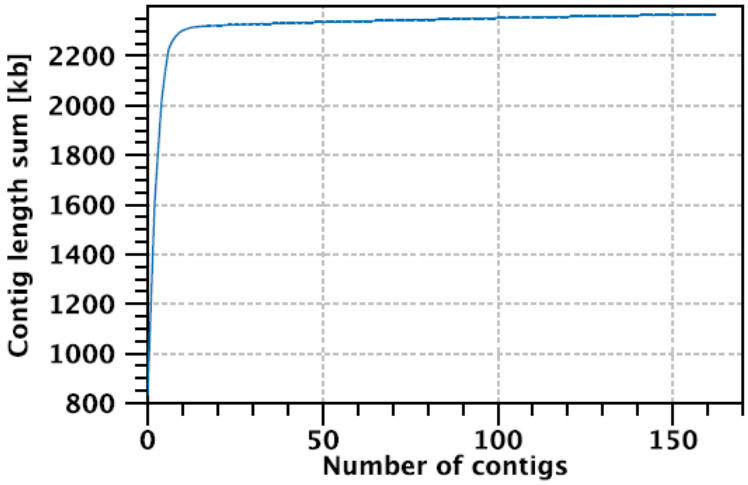
1.2 Contig measurements

N75	212,205
N50	396,031
N25	830,973
Minimum	201
Maximum	830,973
Average	14,529
Count	163

Total	2,368,281



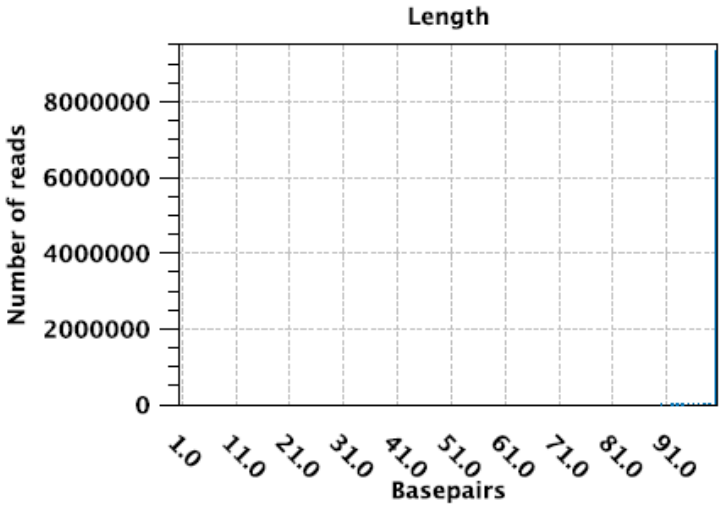
1.3 Accumulated contig lengths



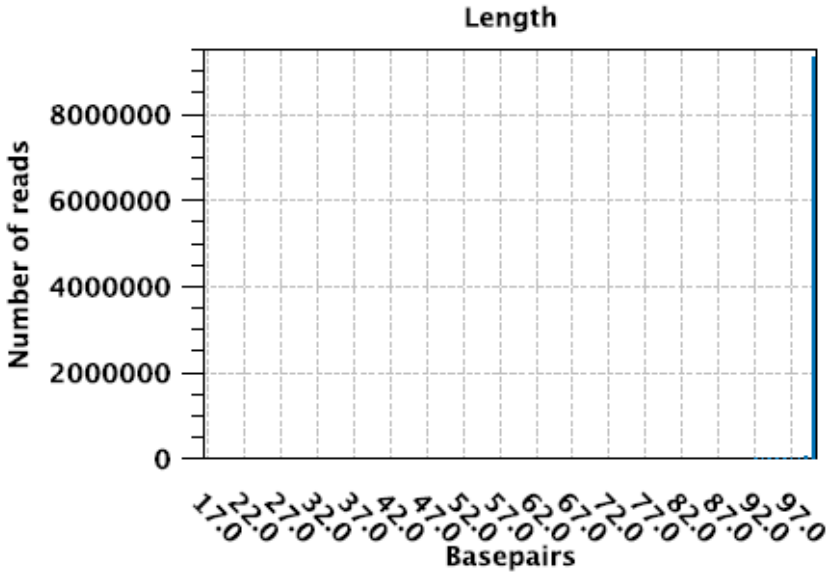
1.4 Summary statistics

	Count	Average length	Total bases
Reads	10,479,594	96.71	1,013,514,884
Matched	10,366,799	97.2	1,007,673,313
Not matched	112,795	51.79	5,841,571
Contigs	163	14,529	2,368,281
Reads in pairs	9,959,474	242.88	
Broken paired reads	407,325	91.18	

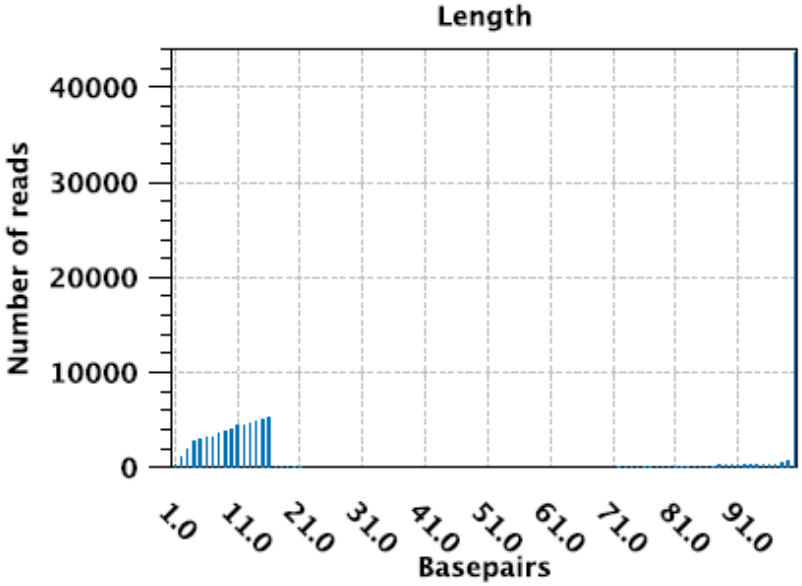
1.5 Distribution of read length



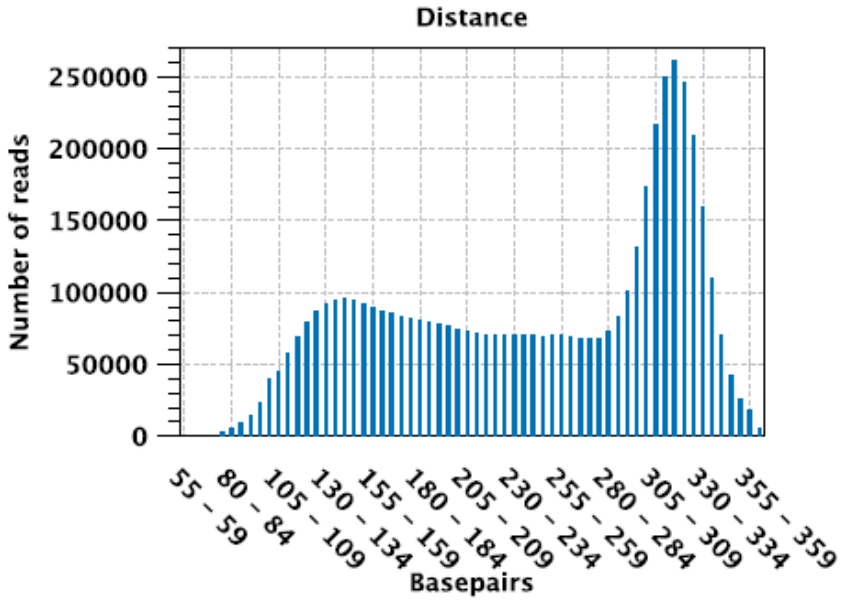
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix V: The CLC Genomics Workbench de novo assembly summary report of *S. sanguinis* PJM8 with Phred score 20.

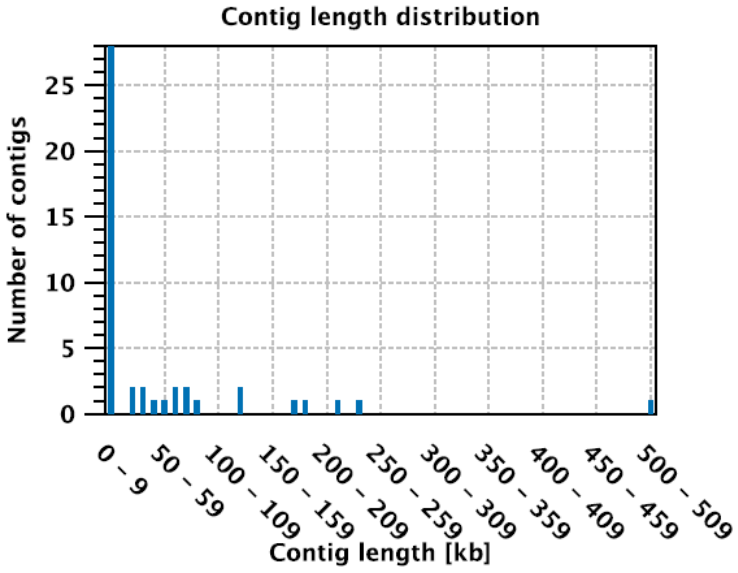
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	653,870	29.7%
Cytosine (C)	464,536	21.1%
Guanine (G)	425,437	19.3%
Thymine (T)	658,865	29.9%

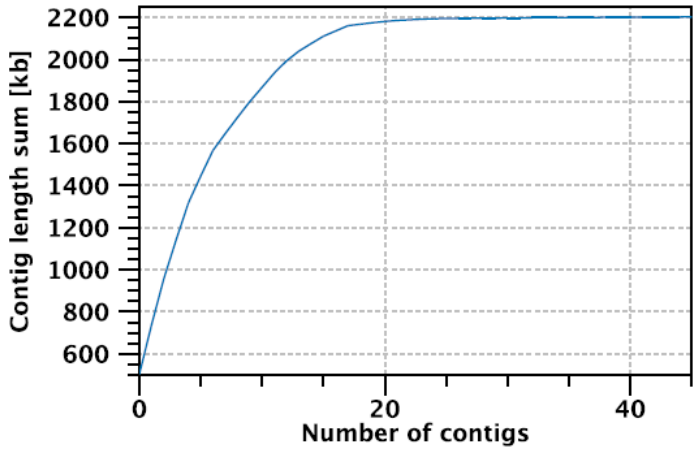
1.2 Contig measurements

N75	76,872
N50	183,297
N25	235,705
Minimum	233
Maximum	504,341
Average	47,885
Count	46

Total	2,202,708



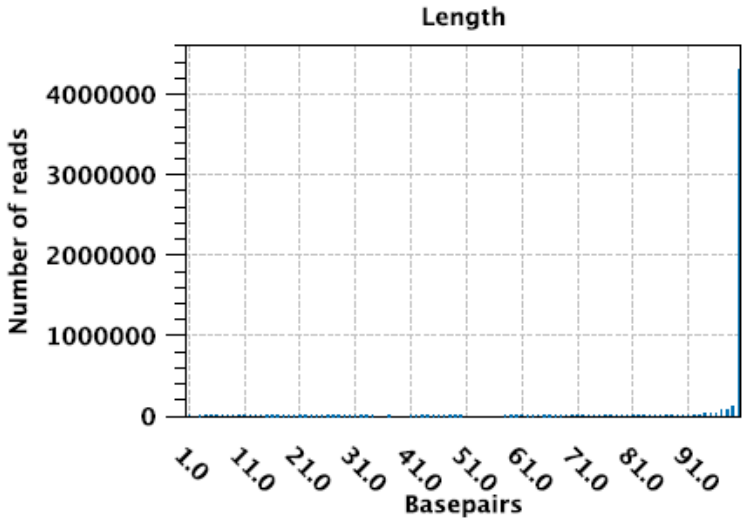
1.3 Accumulated contig lengths



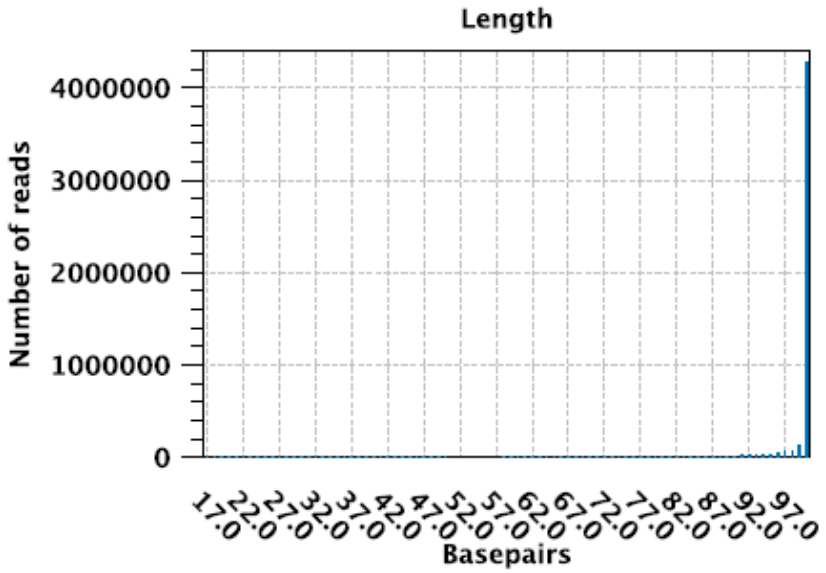
1.4 Summary statistics

	Count	Average length	Total bases
Reads	6,171,258	88.6	546,749,994
Matched	5,856,719	91.92	538,338,293
Not matched	314,539	26.74	8,411,701
Contigs	46	47,884	2,202,708
Reads in pairs	4,619,084	380.72	
Broken paired reads	1,237,635	86.26	

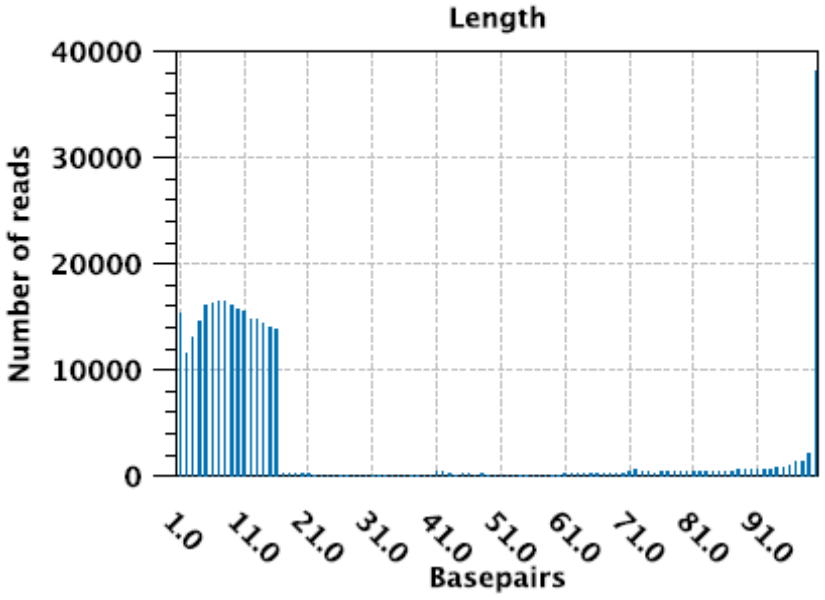
1.5 Distribution of read length



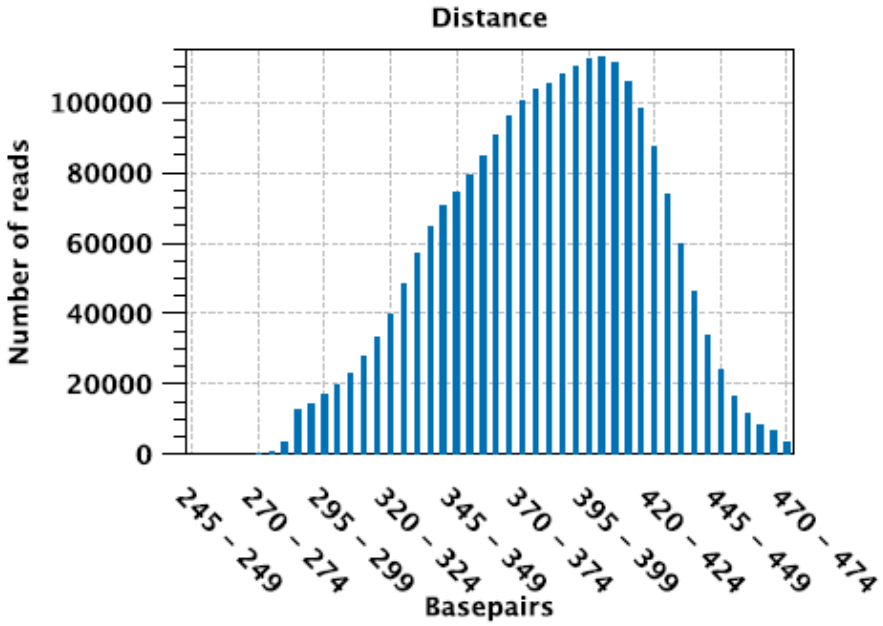
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix W: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* PK488 with Phred score 20.

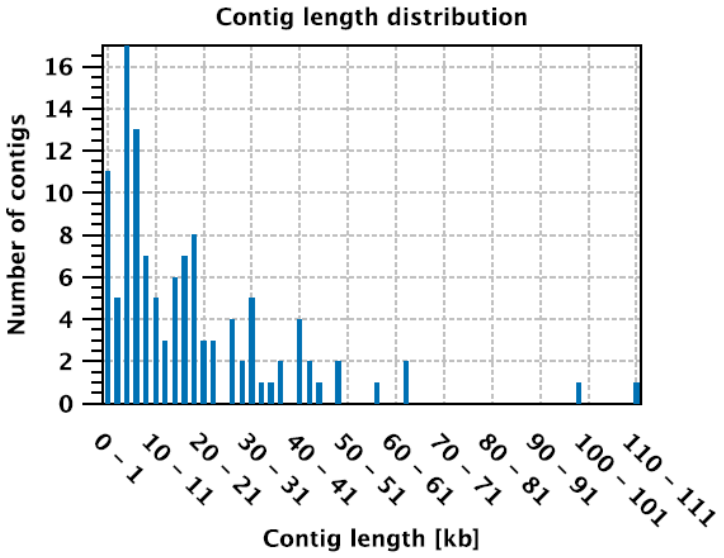
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	594,204	29.1%
Cytosine (C)	430,525	21.1%
Guanine (G)	425,893	20.9%
Thymine (T)	591,896	29.0%

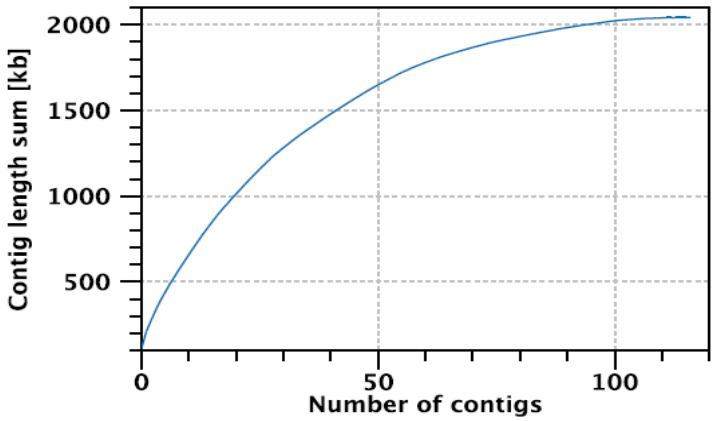
1.2 Contig measurements

N75	17,422
N50	30,074
N25	43,860
Minimum	218
Maximum	110,059
Average	17,457
Count	117

Total	2,042,518



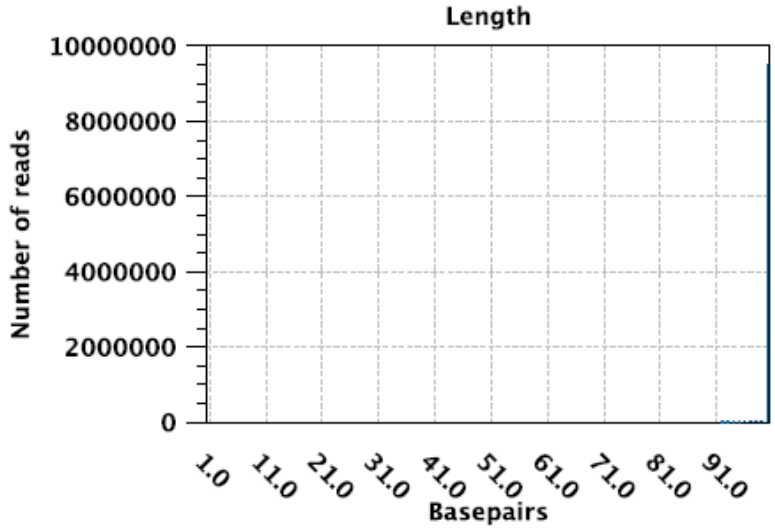
1.3 Accumulated contig lengths



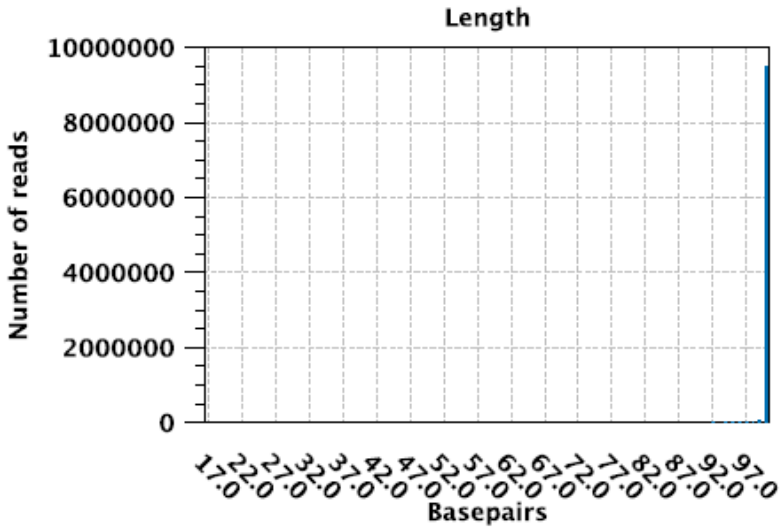
1.4 Summary statistics

	Count	Average length	Total bases
Reads	10,660,070	96.61	1,029,825,685
Matched	10,545,036	97.15	1,024,498,044
Not matched	115,034	46.31	5,327,641
Contigs	117	17,457	2,042,518
Reads in pairs	9,142,328	307.34	
Broken paired reads	1,402,708	95.97	

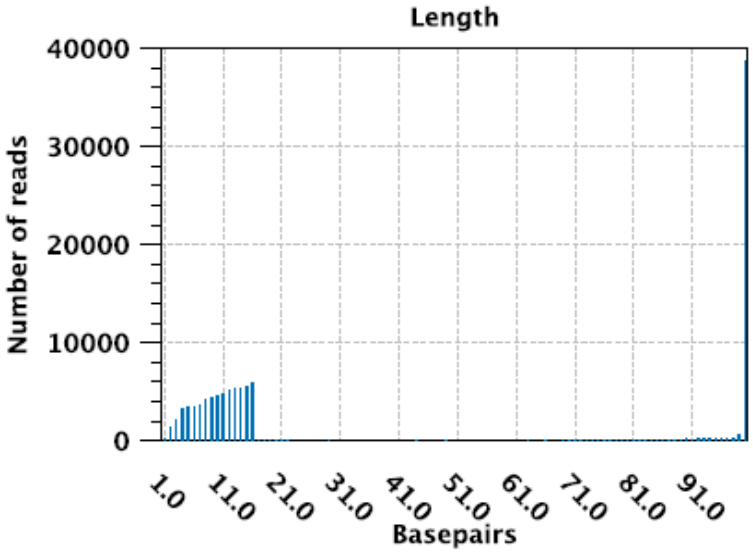
1.5 Distribution of read length



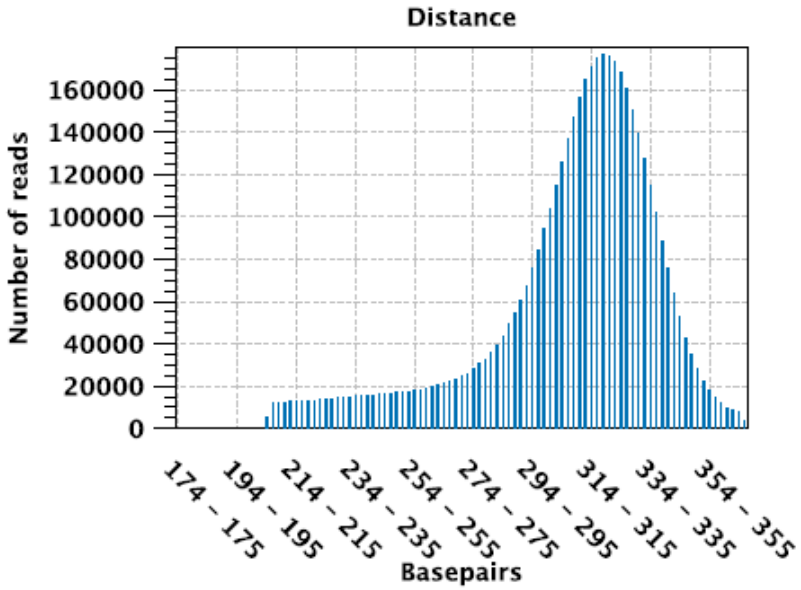
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix X: The CLC Genomics Workbench de novo assembly summary report of *S. parasanguinis* POW10 with Phred score 20.

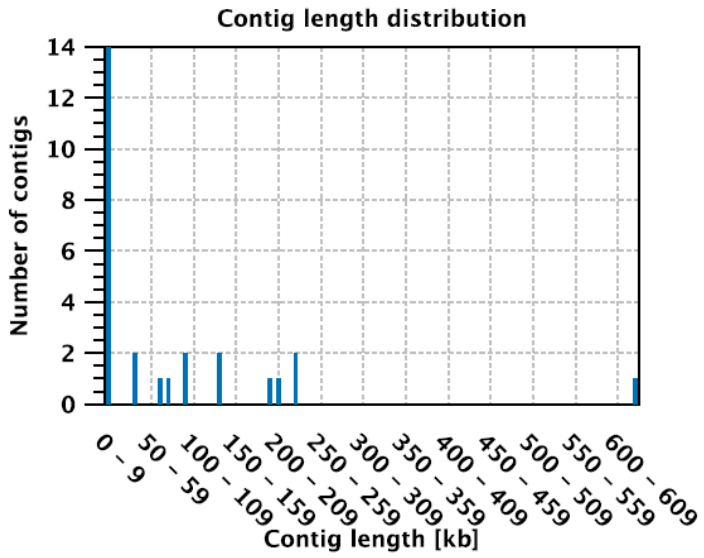
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	635,366	29.6%
Cytosine (C)	451,627	21.0%
Guanine (G)	420,307	19.6%
Thymine (T)	638,551	29.8%

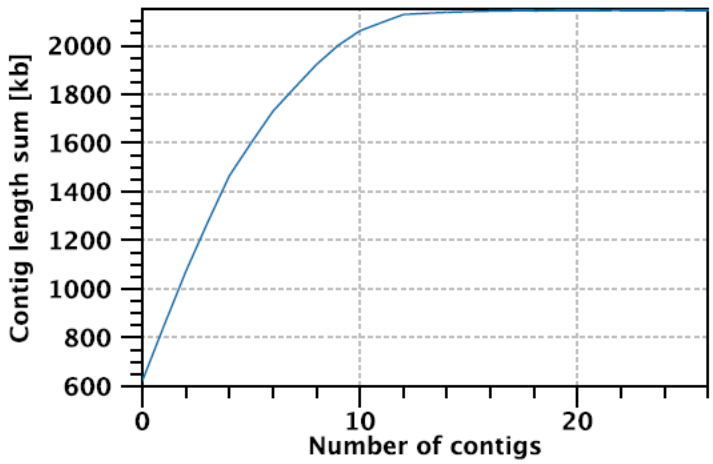
1.2 Contig measurements

N75	129,921
N50	200,167
N25	622,421
Minimum	245
Maximum	622,421
Average	79,476
Count	27

Total	2,145,851



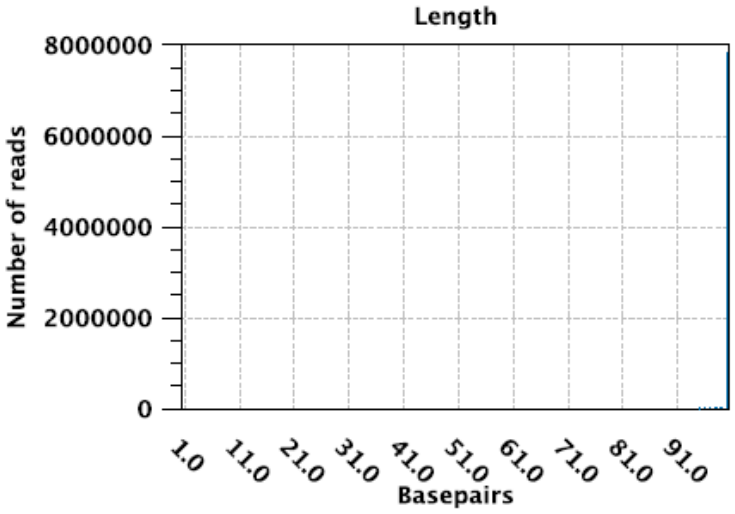
1.3 Accumulated contig lengths



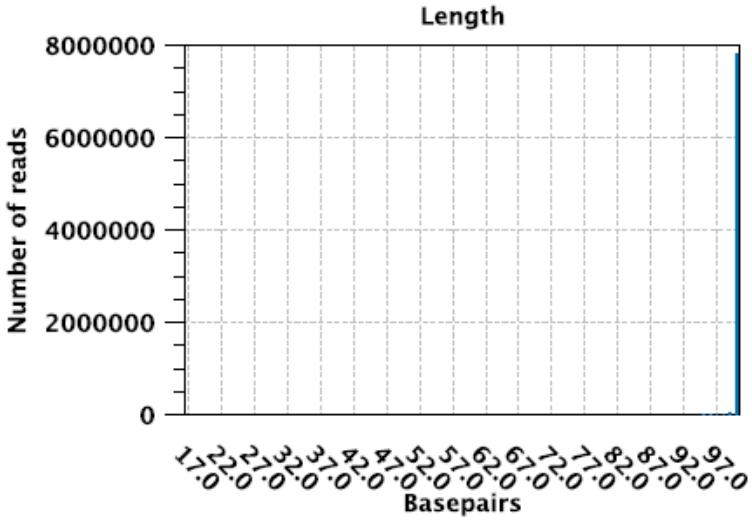
1.4 Summary statistics

	Count	Average length	Total bases
Reads	8,734,240	96.84	845,792,461
Matched	8,667,185	97.31	843,370,485
Not matched	67,055	36.12	2,421,976
Contigs	27	79,475	2,145,851
Reads in pairs	8,040,944	311.19	
Broken paired reads	626,241	94.37	

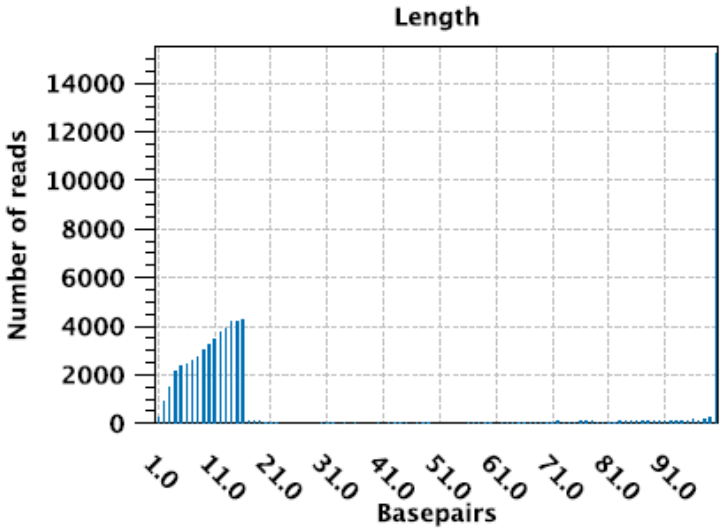
1.5 Distribution of read length



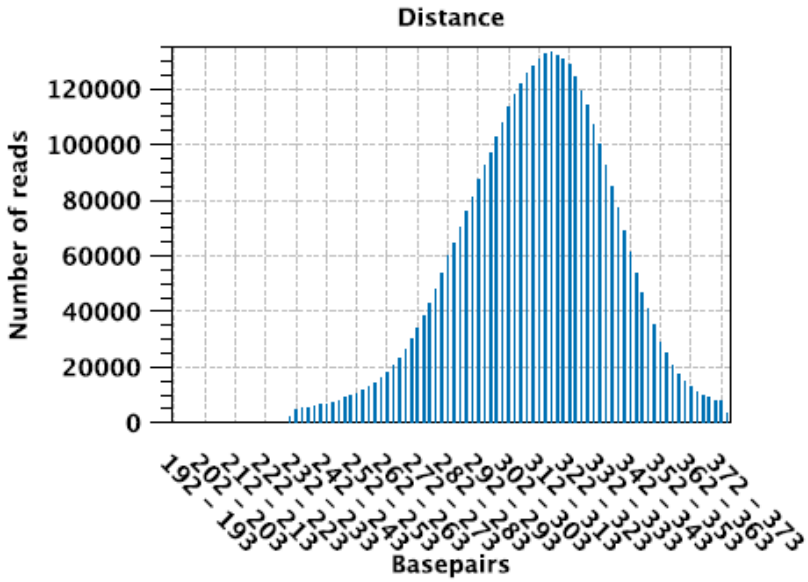
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix Y: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* SK12 with Phred score 20.

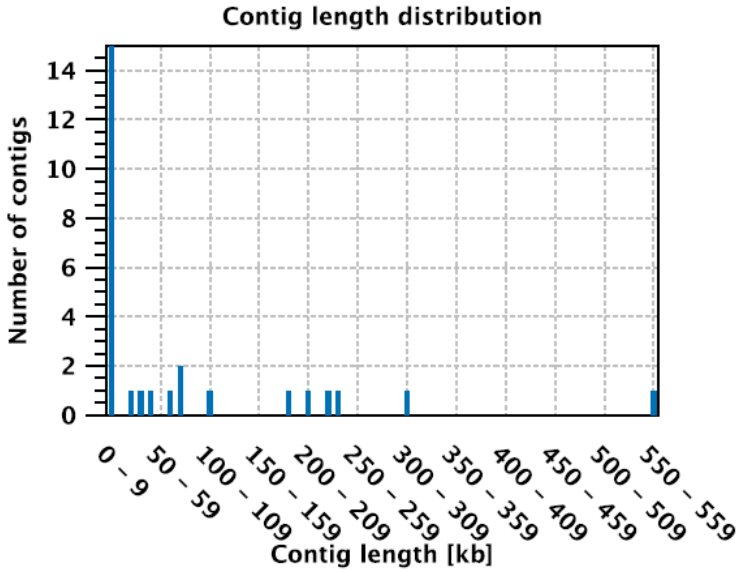
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	648,137	29.9%
Cytosine (C)	431,291	19.9%
Guanine (G)	443,112	20.5%
Thymine (T)	642,220	29.7%

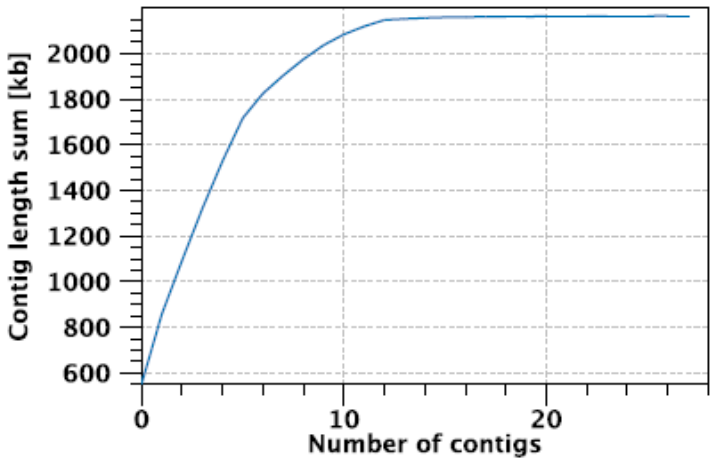
1.2 Contig measurements

Length	
N75	188,481
N50	235,294
N25	551,939
Minimum	215
Maximum	551,939
Average	77,313
Count	28

Length	
Total	2,164,760



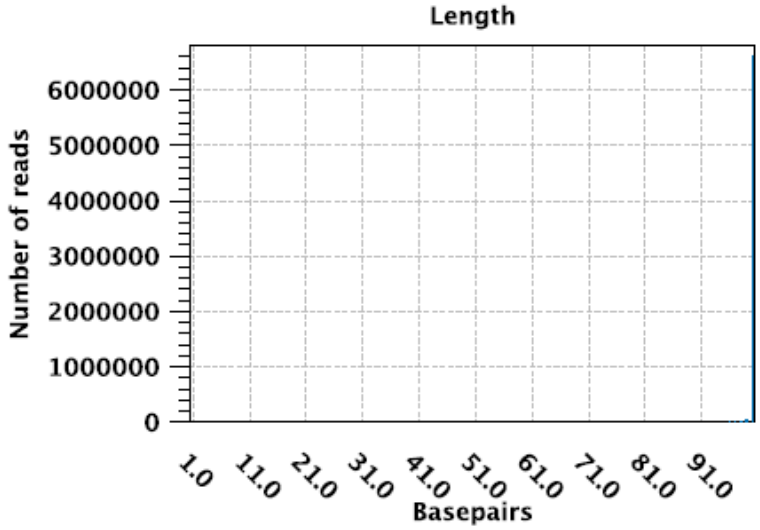
1.3 Accumulated contig lengths



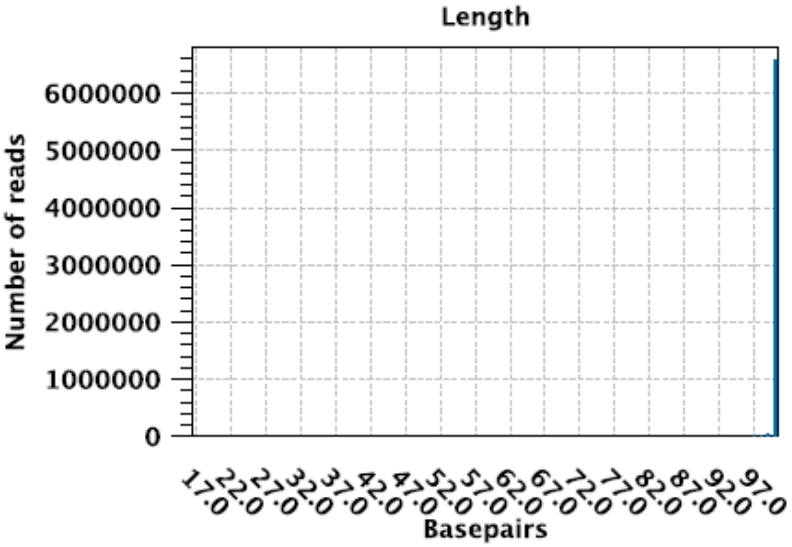
1.4 Summary statistics

	Count	Average length	Total bases
Reads	7,290,928	97.13	708,150,407
Matched	7,207,855	97.52	702,915,217
Not matched	83,073	63.02	5,235,190
Contigs	28	77,312	2,164,760
Reads in pairs	6,366,608	299.22	
Broken paired reads	841,247	94.81	

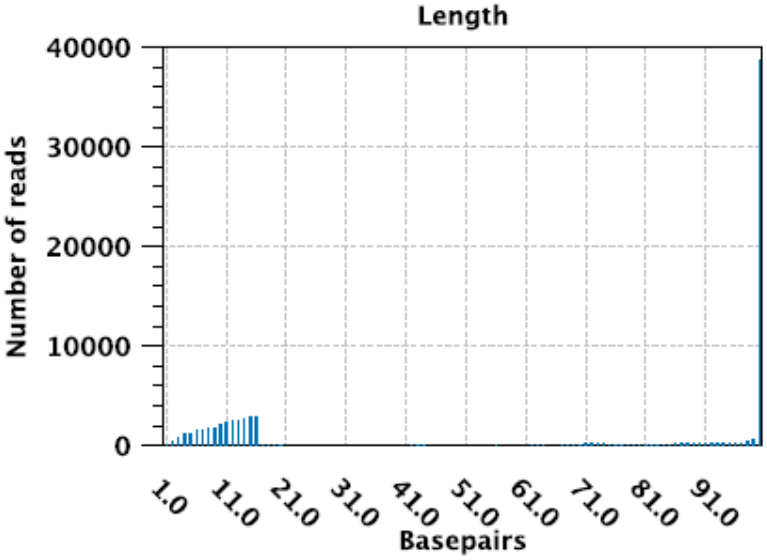
1.5 Distribution of read length



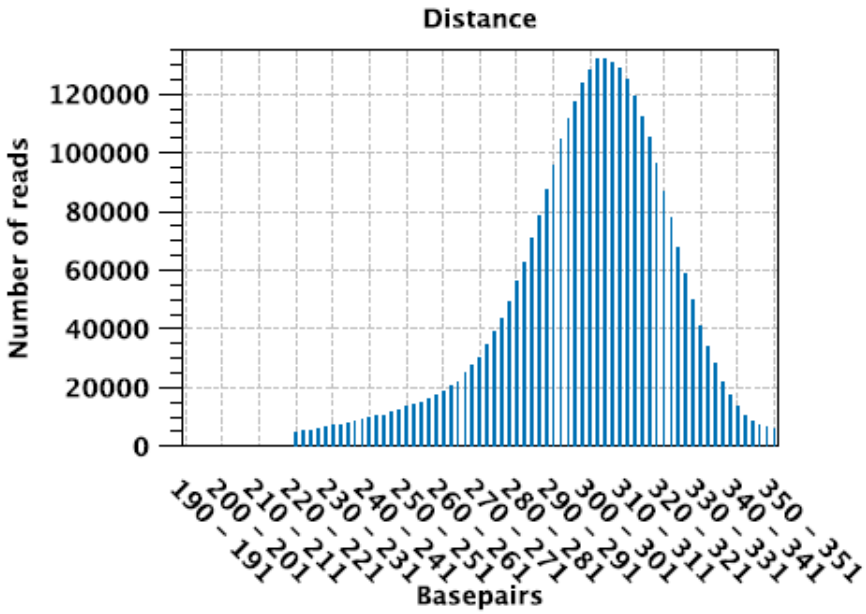
1.6 Distribution of matched read length



1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



Appendix Z: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* SK120 with Phred score 20.

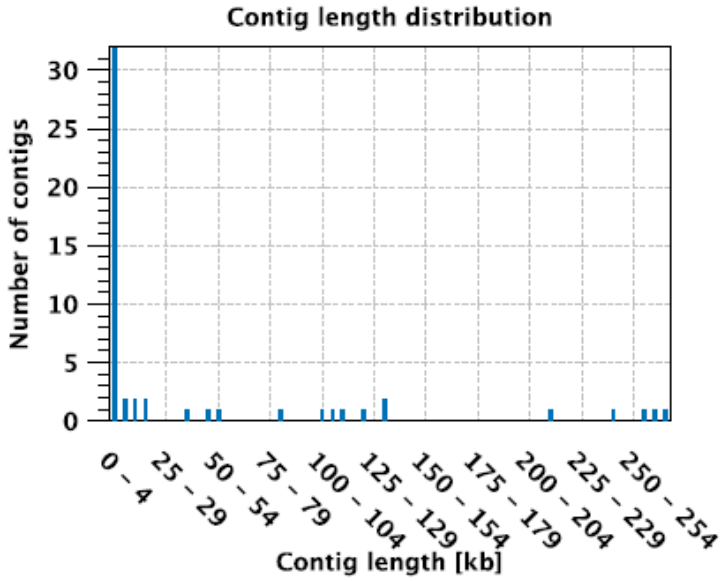
1.1 Nucleotide distribution

Nucleotide	Count	Frequency
Adenine (A)	671,491	29.8%
Cytosine (C)	439,624	19.5%
Guanine (G)	473,936	21.0%
Thymine (T)	670,070	29.7%

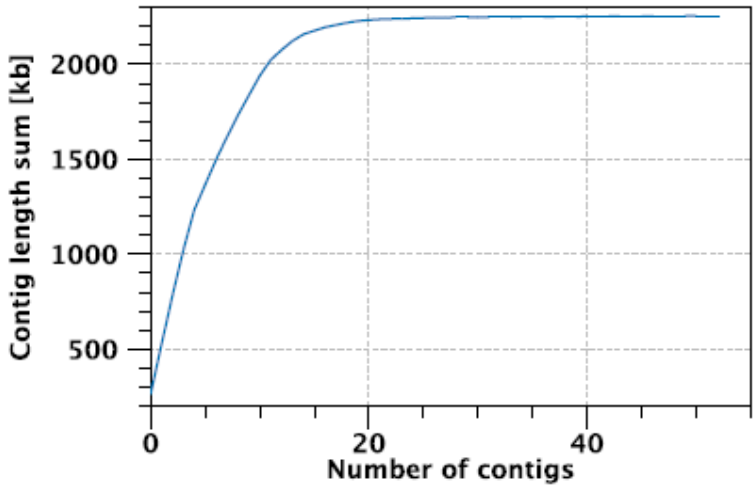
1.2 Contig measurements

N75	113,712
N50	210,865
N25	255,433
Minimum	212
Maximum	264,792
Average	42,549
Count	53

Total	2,255,121



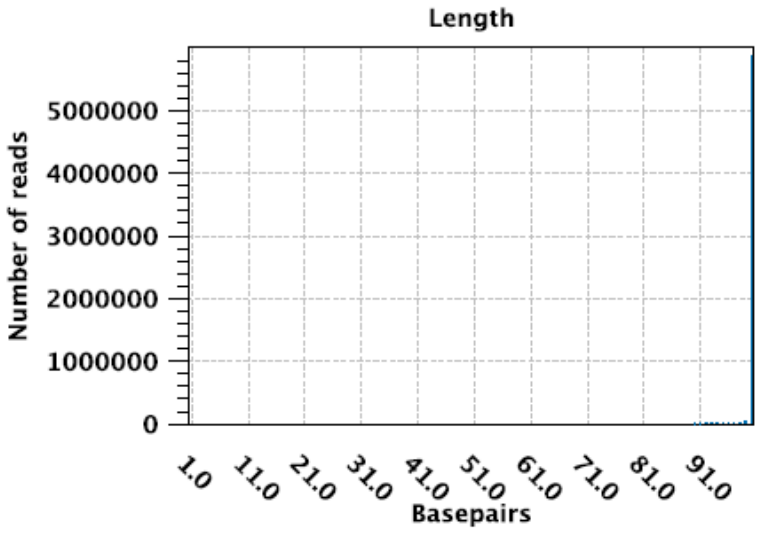
1.3 Accumulated contig lengths



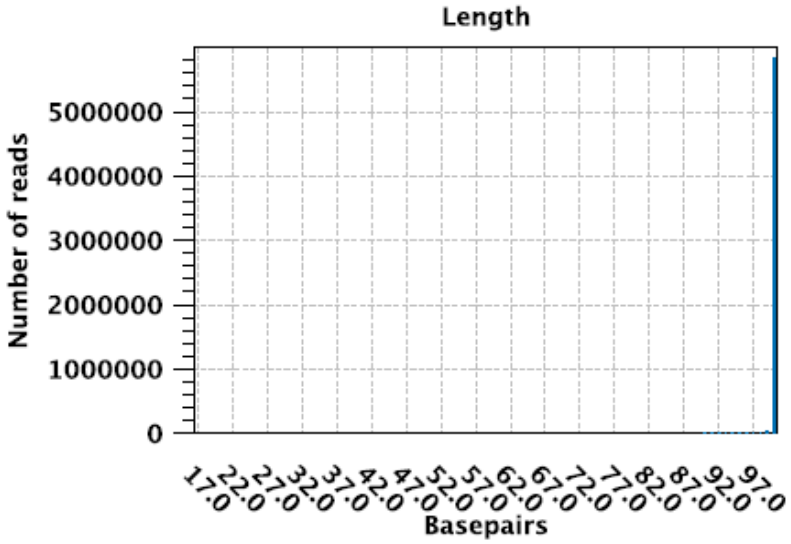
1.4 Summary statistics

	Count	Average length	Total bases
Reads	6,745,058	95.58	644,664,703
Matched	6,636,141	96.35	639,383,243
Not matched	108,917	48.49	5,281,460
Contigs	53	42,549	2,255,121
Reads in pairs	5,852,352	352.69	
Broken paired reads	783,789	90.53	

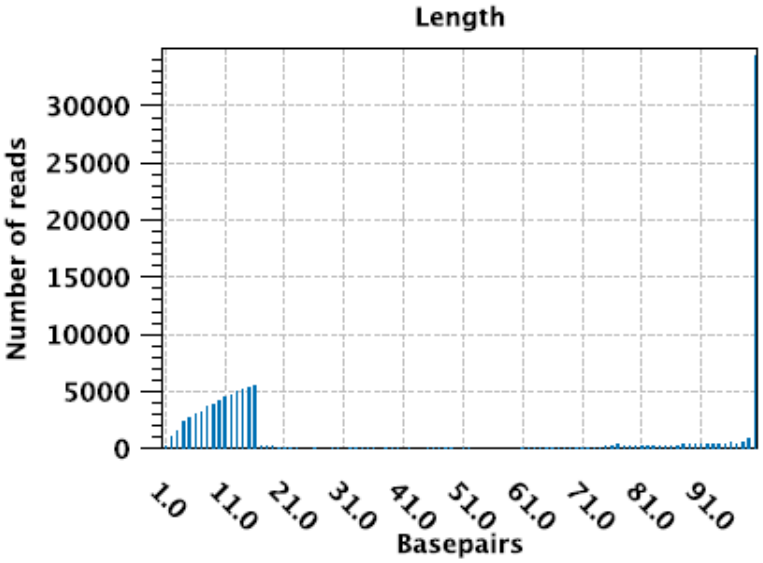
1.5 Distribution of read length



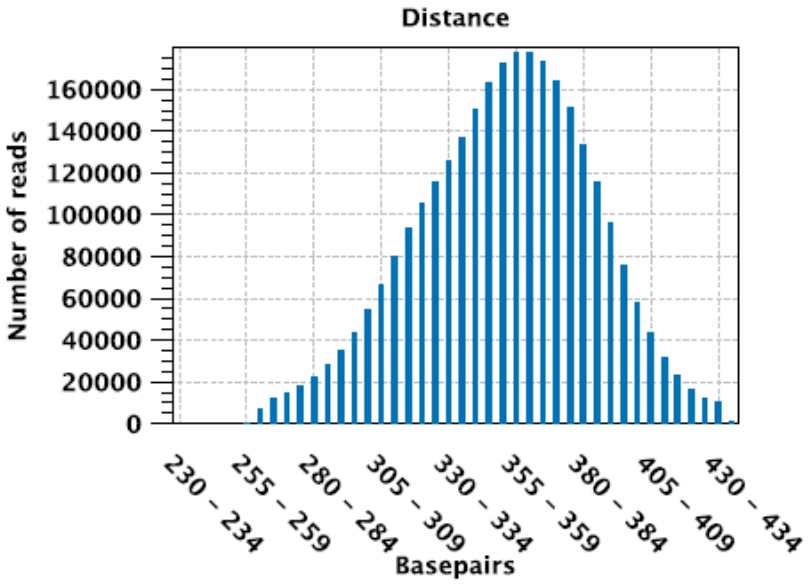
1.6 Distribution of matched read length



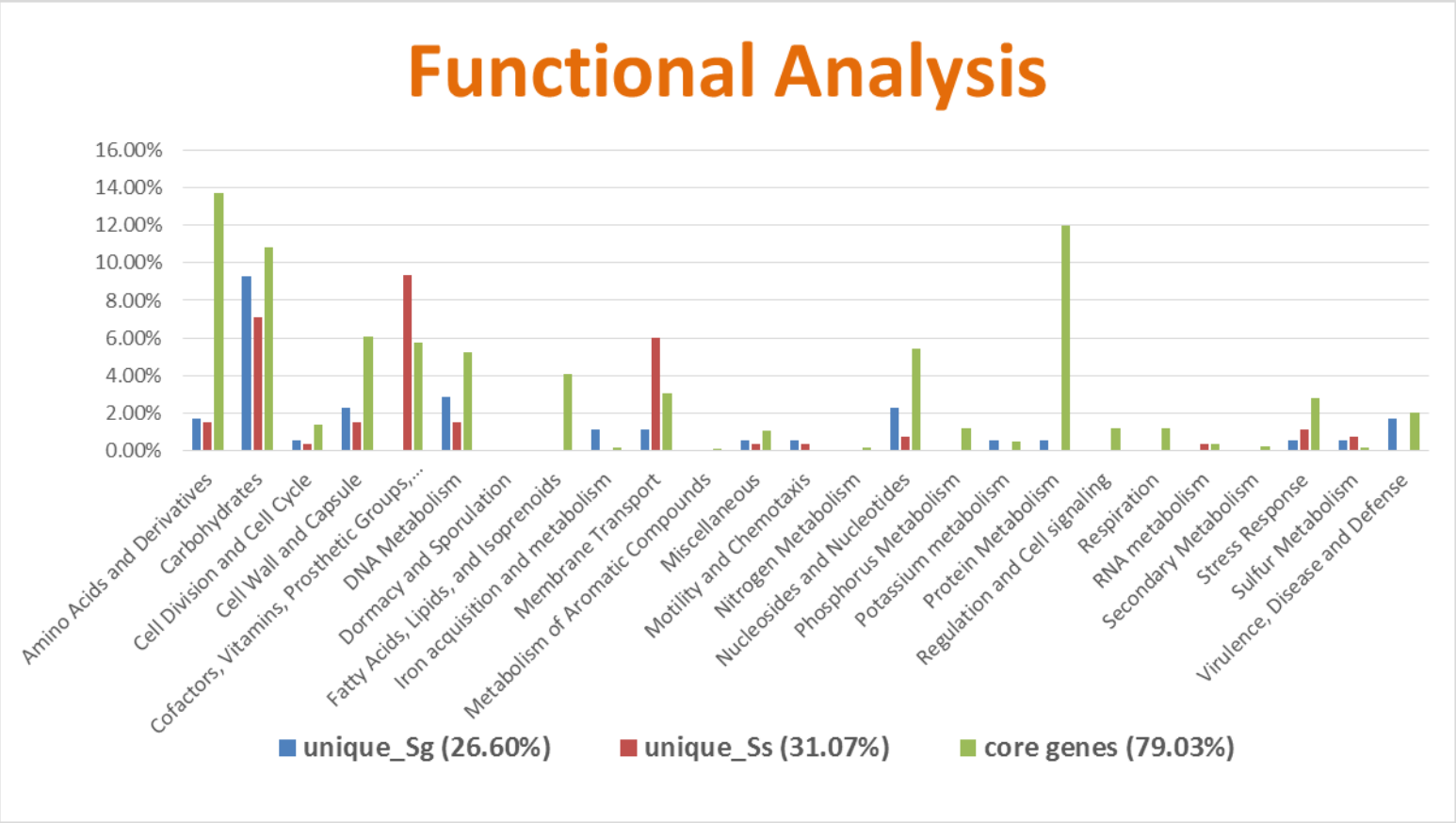
1.7 Distribution of non-matched read length



1.8 Paired reads distance distribution



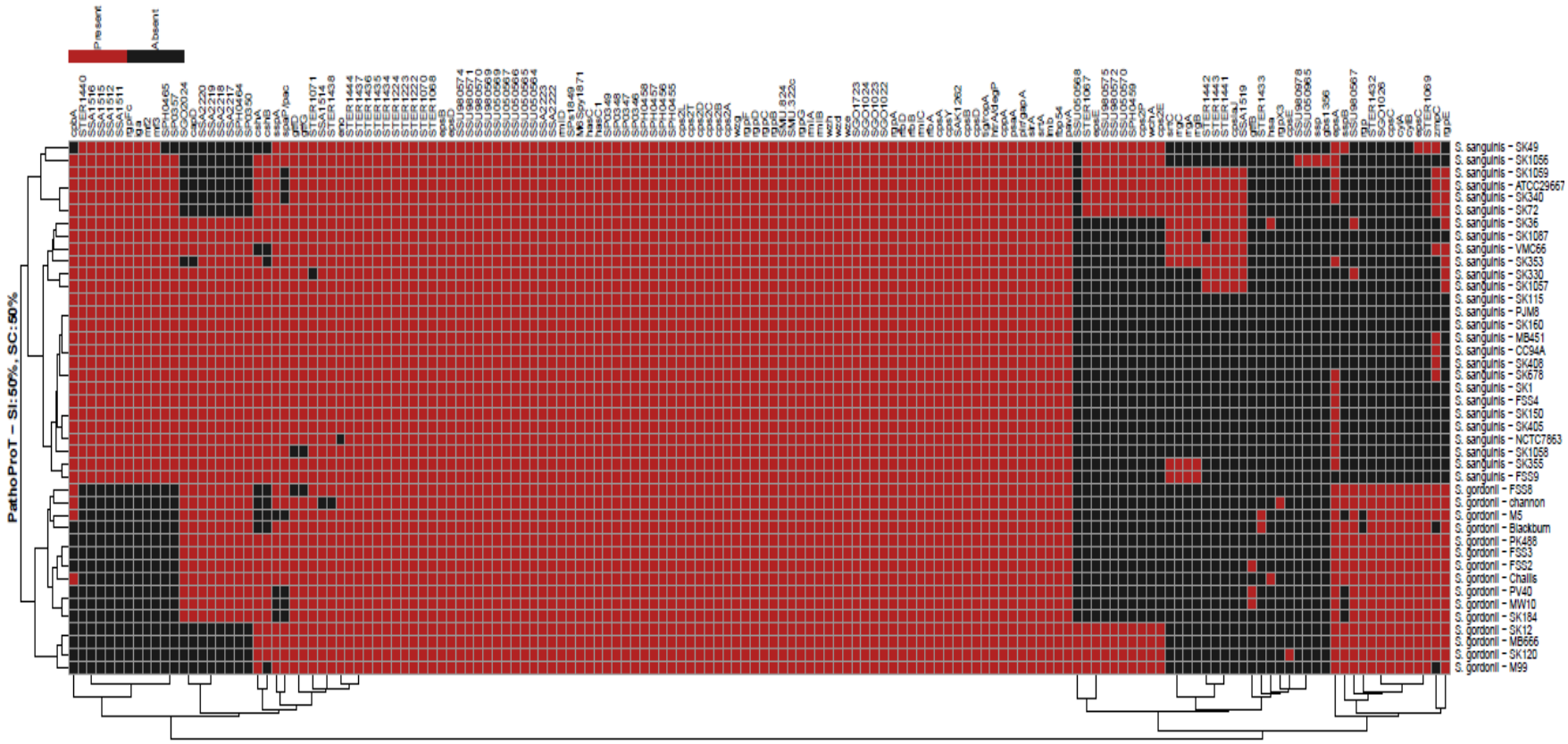
Appendix AA: The CLC Genomics Workbench de novo assembly summary report of *S. gordonii* SK184 with Phred score 20.



Appendix BB: Functional annotation analyses of *S. gordonii* unique core genes and *S. sanguinis* unique core genes. The two prominent functional groups which contributed to the *S. gordonii* unique core genes are cofactor, prosthetic group (9.36%) and pigments and membrane transport group (5.99%).

Enriched biological process	Rast_ID	Enriched Sg unique core genes
Porphyrin-containing compound biosynthetic process	SK36.peg.479	Glutamate-1-semialdehyde aminotransferase (EC 5.4.3.8)
	SK36.peg.476	Porphobilinogen deaminase (EC 2.5.1.61)
	SK36.peg.475	Glutamyl-tRNA reductase (EC 1.2.1.70)
	SK36.peg.474	Siroheme synthase / Precorrin-2 oxidase (EC 1.3.1.76)
	SK36.peg.465	Uroporphyrinogen-III methyltransferase (EC 2.1.1.107) / Uroporphyrinogen-III synthase (EC 4.2.1.75)
	SK36.peg.461	Cobalt-precorrin-4 C11-methyltransferase (EC 2.1.1.133) /cobM
	SK36.peg.2040	FIG01117915: hypothetical protein
	SK36.peg.2038	FIG01118726: hypothetical protein
Cobalamin biosynthetic process	SK36.peg.481	Cobalamin synthase
	SK36.peg.480	Adenosylcobinamide-phosphate guanylyltransferase (EC 2.7.7.62) / cobU
	SK36.peg.472	Cobyric acid synthase
	SK36.peg.502	Nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase (EC 2.4.2.21) / cobT
	SK36.peg.469	Additional substrate-specific component CbiN of cobalt ECF transporter
	SK36.peg.500	L-threonine 3-O-phosphate decarboxylase (EC 4.1.1.81)
	SK36.peg.468	Substrate-specific component CbiM of cobalt ECF transporter
	SK36.peg.467	Cobalt-precorrin-2 C20-methyltransferase (EC 2.1.1.130)
	SK36.peg.466	Sirohydrochlorin cobaltochelatase CbiK (EC 4.99.1.3)
	SK36.peg.464	Cobalt-precorrin-6x reductase (EC 1.3.1.54)
	SK36.peg.463	Cobalt-precorrin-3b C17-methyltransferase /cbiH
	SK36.peg.462	Cobalamin biosynthesis protein CbiG
	SK36.peg.461	Cobalt-precorrin-4 C11-methyltransferase (EC 2.1.1.133)
	SK36.peg.460	Cobalt-precorrin-6y C15-methyltransferase [decarboxylating] (EC 2.1.1.-)
	SK36.peg.459	Cobalt-precorrin-6y C5-methyltransferase (EC 2.1.1.-)
	SK36.peg.458	Cobalt-precorrin-6 synthase, anaerobic
	SK36.peg.457	Cobalt-precorrin-8x methylmutase (EC 5.4.1.2)
	SK36.peg.456	Cobalt-precorrin-8x methylmutase (EC 5.4.1.2)
	SK36.peg.455	Adenosylcobinamide-phosphate synthase /cbiP
	SK36.peg.454	Cobyrinic acid A, C-diamide synthase /cobB/cbiA

Appendix CC: Functional enrichment analyses of *S. sanguinis* unique core genes. These genes were statistically (FDR<0.05) enriched in compound biosynthetic process (8) and cobalamin biosynthesis process (20)



Appendix DD: A heat map generated by PathoProT in StreptoBase. This comparative virulence genes analysis was performed using 15 strains of *S. gordonii* and 27 strains of *S. sanguinis* based on the threshold of 50% sequence identity and 50% sequence coverage.

Appendix EE: A screenshot of the StreptoBase homepage. The main page of the website features the brief introduction of Mitis group oral streptococci, database summary, news and conferences as well as organization and blogs associated with Mitis group oral streptococci species.