UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

MANCHESTER
1824

The University of Manchester

A COMPUTATIONAL APPROACH TO STUDYING THE EVOLUTION OF STREPTOCOCCAL QUORUM SENSING SYSTEMS

A thesis submitted to The University of Manchester for the degree of

DOCTOR OF PHILOSOPHY (PhD)

in the Faculty of Life Sciences

2014

RAJA FARHANA RAJA KHAIRUDDIN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

TABLES OF CONTENTS

LIST OF FIGURES 6

LIST OF TABLES 8

ABSTRACT 9

DECLARATION 10

COPYRIGHT 11

ACKNOWLEDGEMENT 12

CHAPTER 1: Introduction 13

1.1 Quorum sensing systems in bacteria 14

1.1.1 Regulation of *S. pneumoniae* Com QS systems 16

1.1.2 DNA uptake mechanism and transformation 20

1.1.3 Regulation of competence via QS systems in other *Streptococcus* species 22

1.1.4 The bacteriocin QS system: a closely related competence QS system in *S. pneumoniae* 23

1.1.5 The importance of the Com QS system 23

1.2 Computational approaches in molecular evolution 25

1.2.1 Sources of genomes 26

1.2.2 Multiple sequence alignment 27

1.2.3 Phylogenetic inference methods 28

1.2.4 The reliability of phylogenetic inference 33

1.3 Molecular evolution and phylogenomics 35

1.3.1 Molecular evolution mechanisms 35

1.3.2 Gene trees and species trees 37

1.3.3 Phylogenomics 39

1.4 Understanding the evolution of quorum sensing systems in streptococcal species 42

1.4.1 Problem statement/ motivation of work 42

1.4.2 Experimental aims and subprojects 42

1.5 Alternative format 45

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

CHAPTER 2: Using the competence regulon structure to identify quorum sensing homologs in streptococcal species 46

2.1 Introduction 47

2.2 Methods and Input 49

2.2.1 Search Using Competence Regulon (SUCRE) methodology 49

2.2.2 Benchmarking dataset 55

2.2.3 The 86 genomes dataset 57

2.3 Results 57

2.3.1 Identification of Quality Putative Genes 57

2.3.2 Classification of QPGs based on the competence regulon structure 59

2.3.3 Identification of QS homologs 60

2.3.4 The 86 *Streptococcus* genomes 64

2.4 Discussion 67

CHAPTER 3: Assessing SUCRE performance in comparison with other orthology detection methods 70

3.1 Introduction 71

3.2 Materials and Methods 75

3.2.1 Protein query and subject genome sequences 75

3.2.2 The approaches and programs used for the comparison 75

3.2.3 Testing the performances of the compared methods in detecting QS genes 77

3.3 Results 79

3.3.1 Identification of the QS benchmark genes 79

3.3.2 Performance in detecting the QS genes 83

3.4 Discussion 86

3.4.1 Challenges in finding QS genes 86

3.4.2 SUCRE performance in detecting the QS genes 87

CHAPTER 4: *Streptococcus* species phylogeny 89

4.1 Introduction 90

4.2 Materials and Methods 91

4.2.1 The gene dataset 91

4.2.2 Supermatrix analysis 93

4.2.3 Supertree analysis 94

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

4.2.4 Constructing the *Streptococcus* species tree 95
4.3 Results 95
4.3.1 Evolutionary signals between the gene families 95
4.3.2 Finding the species tree using the supertree approach 100
4.3.3 A strict consensus species phylogeny 102
4.4 Discussion 104

CHAPTER 5: The evolution of streptococcal quorum sensing systems 106
5.1 Introduction 107
5.2 Materials and Methods 110
5.2.1 The QS homologs dataset from the 86 genomes of *Streptococcus* species 110
5.2.2 Mapping the regulon structures of the QS homologs 110
5.2.3 Estimating the maximum likelihood gene trees 112
5.2.4 Inferring QS orthologs 112
5.2.5 Inferring the presence and absence of the QS regulon to understand the evolution of QS systems in *Streptococcus* species 115
5.3 Results 116
5.3.1 The regulon structure of the QS homologous genes across *Streptococcus* species 116
5.3.2 The evolution of QS systems as one regulon unit 119
5.3.3 Contributions of duplications to the functional divergence of QS systems in *Streptococcus* species 123
5.3.4 Horizontal gene transfer contributes to the evolution of QS systems in *Streptococcus* species 125
5.4 Discussion 128

CHAPTER 6: Conclusion 133
6.1 The specificity of QS systems in streptococcal 134
6.2 The functional divergence of QS systems through duplication 135
6.3 The origin of streptococcal QS systems 137

BIBLIOGRAPHY 140

APPENDICES 168
APPENDIX A 169
APPENDIX B 171

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

APPENDIX C 174
APPENDIX D 175
APPENDIX E 177
APPENDIX F 183
APPENDIX G 184
APPENDIX H 186
APPENDIX I 192
APPENDIX J 195

Word count: 45,713

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

LIST OF FIGURES

Figure 1.1: The regulation of the S. pneumoniae Com QS system 17

Figure 1.2: The structure of an ABC transporter 18

Figure 1.3: The DNA uptake and transformation mechanism 21

Figure 1.4: The QS regulon structure of the bacteriocin-like peptide (Blp) and competence (Com) system of *S. pneumoniae* 24

Figure 1.5: The concept of orthologs, paralogs and xenologs. 35

Figure 1.6: The *Streptococcus* phylogeny from 16S ribosomal genes 37

Figure 1.7: Gene trees and species tree 38

Figure 1.8: Phylogenomics approach 41

Figure 2.1: An overview of SUCRE methodology 51

Figure 2.2: The process of finding Quality Putative Genes (QPGs) in SUCRE 53

Figure 2.3: The quality putative genes in Streptococcus species categorised into Prime, Lone, Unusual, and Wobbly categories. 59

Figure 2.4: The ABC exporter (*A* gene) maximum likelihood gene tree of 10 *Streptococcus* species 61

Figure 2.5: The subclade of the ABC exporter gene tree 62

Figure 2.6: Overview of QS homologs mapped across Streptococcus species 63

Figure 2.7: The ABC exporter (*A* gene) ML gene tree of 21 *Streptococcus* species 65

Figure 2.8: The overview of the presence and absence of QS homologs (*ABCDEX*) in 86 genomes across 21 *Streptococcus* species 66

Figure 3.1: The ortholog clustering used in COG 74

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

Figure 3.2: The detection of QS benchmark genes (gene ID) by the compared methods .. 80

Figure 3.3: Positive predictive values (PPVs) are plotted against the true positive values (TPVs) for all methods 84

Figure 3.4: Venn diagram of benchmark true positive (TP) genes detected by OrthoMCL and SUCRE in 100 testing sets 85

Figure 4.1: A heat map based on the SH test P-value calculated from 29 supermatrices in the supermatrix approach. 99

Figure 4.2: A heat map based on the SH test P-value calculated from the 27 gene trees (treeID) of the 19 *Streptococcus* species (45 genomes) 100

Figure 4.3: The unrooted COG0533 and COG0172 gene trees with the bootstrap values 101

Figure 4.4: A rooted, strict consensus phylogeny of the 19 streptococcal species 103

Figure 5.1: Methods used to map the QS regulon structure and infer orthologs for streptococcal species 111

Figure 5.2: The regulon structure of Streptococcus QS homologs mapped across 21 Streptococcus species next to the Streptococcus species tree 117

Figure 5.3: The rooted ML gene trees 122

Figure 5.4: Lone, duplicated genes (circled arrows) mapped in the genomes 123

Figure 5.5: The 19 streptococcal early QS homologs and the streptococcal species tree 124

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS VERSITI F

LIST OF TABLES

Table 2.1: The regulon categories: Prime, Lone, Unusual and Wobbly 54

Table 2.2: The list of the protein queries and subject genomes used in the analysis..... 56

Table 2.3: The BLAST hits in each iteration for each ABCDEX gene 58

Table 3.1: Summary of the characteristics of SBH, RBH, SUCRE, GH, COG and OrthoMCL..... 72

Table 3.2: The subject genomes of six *Streptococcus* species..... 75

Table 3.3: The QS genes detection performance of SBH, RBH, SUCRE, GH, COG and OrthoMCL..... 82

Table 4.1: The list of 31 orthologs from ribosomal and tRNA synthetase genes in the *Streptococcus* species..... 92

Table 4.2: The number of gene families in each supermatrix subset..... 96

Table 4.3: The distribution of the gene families (COG ID) in the supermatrices. 98

Table 4.4: The bootstrap values support the root of the gene trees of other gene families. 102

Table 5.1: Statistical support (P-value) for the SH test results 120

Table 5.2: GC content of the AB transporter genes of streptococcal species. 127

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

ABSTRACT

For many years, researchers have studied the social lives of bacteria to understand intra- and inter-species interactions. Cell-cell communication, also known as quorum sensing (QS), is used by bacteria to coordinate their behaviour in response to environmental conditions. The QS system in *Streptococcus* species is well known to regulate competence. Studies show that *Streptococcus pneumoniae* has two homologous QS systems: 1) the competence (Com) system that regulates competence; and 2) a bacteriocin-like peptide (Blp) system that regulates the production of bacteriocins. Both functions are widespread in the genus. In *S. pneumoniae*, the Blp QS system shares a common ancestor and has similar features to the Com QS system. However, the evolutionary relationship between these QS systems remains obscure.

SUCRE methodology was developed to identify the QS homologous genes in the streptococcal species. SUCRE uses four complementary approaches: homology search, putative gene finding, regulon construction, and evolutionary analysis. The performance of SUCRE was assessed in comparison with other orthology detection methods. SUCRE is precise in identifying the QS homologous genes and has similar performance to OrthoMCL. The QS system structures are found to be conserved across the streptococcal species.

A streptococcal species phylogeny was constructed from the ribosomal and tRNA synthetase gene families. Using the QS genes identified from SUCRE and the streptococcal species phylogeny, the study infers the evolution of the QS systems in *Streptococcus* species. The study shows that the QS systems evolved as a regulon unit. The paralogous relationship between each of the QS systems suggests that duplication has a huge influence on functional divergence of the QS systems in the genus. Although, horizontal gene transfer (HGT) is commonly found in bacteria, little evidence is found to support that the effect of HGT on the functional divergence of the QS systems in this genus. However, the QS regulon genes of the same QS system are found to be non-vertically transferred across species that signifies that the HGT event promotes the sequence variation between these genes.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

DECLARATION

The University of Manchester
PhD by published work Candidate Declaration

Candidate Name: Raja Farhana RAJA KHAIRUDDIN

Faculty: Life Sciences

Thesis Title: A computational approach to studying the evolution of the streptococcal quorum sensing systems.

Declaration to be completed by the candidate:

I declare that no portion of this work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Signed: Date: 2 November, 2014

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

COPYRIGHT

The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester the right to use such Copyright, including for administrative, promotional purposes.

Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

The ownership of certain Copyright patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs ad tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on presentation of Theses.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude to my supervisor Dr. Simon Whelan for being such an amazing mentor, who with knowledge, patience and enthusiasm guided me from the beginning of the journey until this paper is structured. And a very special thanks to my co-supervisor Dr. Daniel Rozen for the advice and the constructive suggestions throughout the research. My experience learning from both of you has been priceless.

It is my pleasure to thank Prof. Simon Lovell, Dr. Christopher Knight, and my fellow colleagues who had been there to support me and made this journey remarkable. I would also like to thank Prof. David Robertson and Prof. Peter Young for making my defence an enjoyment moment and for the brilliant comments and positive insights.

My deepest thanks and love to my beloved family, who continually encouraged me, whom without, none of this would be possible. The thesis is especially dedicated to them. Thank you.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

CHAPTER 1

Introduction

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

1.1 Quorum sensing systems in bacteria

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

For the longest time, it was thought that a single bacterium lives alone. However, studies have shown that this is not the case. Bacteria actually live in communities where they talk to each other in order to find nutrients and reproduce. There are trillions of cells in the human body, and there are approximately ten times as many bacteria that live in and on our body. These bacteria can be either beneficial or harmful. A single bacterium would not be sufficient to affect an organ or tissue in our body. A bacterium takes in nutrients and multiplies to increase its number within a population. Communication between bacteria cells, which later become known as quorum sensing (QS) (Fuqua et al., 1994), is important in co-ordinating bacterial behaviour within a population. Once there are sufficient bacterial numbers to engage in a particular action, bacteria communicate with each other to synchronise their actions.

This control mechanism in both Gram-positive and Gram-negative bacteria regulates gene expression, which governs physiological functions, in response to the density of the signalling molecules produced by the bacteria (Bassler, 1999, Dunny and Winans, 1999). Numerous studies have examined the mechanism of QS, both within and between different bacterial species (Bassler, 1999, Dunny and Winans, 1999, Claverys and Havarstein, 2002).

The QS mechanism involves producing, detecting and responding to small signalling molecules. This mechanism allows bacteria: to coordinate their behaviour to environmental changes, such as nutrient availability (Finkel and Kolter, 2001); to engage in defence mechanisms against other bacteria (Høyland-Kroghsbo et al., 2013); to control virulence factor expression (Rutherford and Bassler, 2012) and biofilm production (Vuong et al., 2003, Rathsam et al., 2005, Joelsson et al., 2006); and adapt to stimuli at the population level (Baysse et al., 2005). These responses are crucial for increasing the chance of survival of a bacterial species in population.

Each bacterium has specific signalling molecules that regulate different bacteria functions. A study on QS systems between bacteria showed that most bacteria have the ability to produce and transmit at least one of the signalling molecules (Miller and Bassler, 2001). Although each QS system regulates a different function, in general, how the QS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

systems response to the molecular signals is similar (Bassler, 1999, Dubnau, 1999, Miller and Bassler, 2001). First, the molecular signals are produced in the cell. Then, these molecules are secreted into the extracellular environment. The signalling molecules increase as the number of bacteria cells in a population increases. When the accumulation of the molecular signals reaches a certain concentration threshold, the signalling molecules bind to a sensor protein, which becomes phosphorylated. The transfer of a phosphate group to a receptor protein triggers a signal transduction cascade that leads to changes in the expression of genes with particular functions. Because QS systems allow the cells to work together as a group, changes in gene expression happen collectively among the bacterial species in a population.

In streptococcal species, the QS mechanism is known to regulate bacterial competence. In fact, the regulation of bacterial competence in *S. pneumoniae* was the first process that was shown to involve a QS mechanism (Dawson and Sia, 1931). In 1964, an experiment by Tomasz and Hotchkiss (1964) inferred that the competence in *S. pneumoniae* species is controlled by the accumulation of a signalling molecule, called the Competence Stimulating Peptide (CSP) (Havarstein et al., 1995).

Out of a million microbial species, there are approximately a few hundred that cause diseases. *S. pneumoniae* is one of the most infectious microbial species that can grow, metabolize and have independent interactions with other organisms, including humans. *S. pneumoniae* is known to cause about 5 million fatal pneumonia infections per year in children in the United States, and around 7% of the 1 million cases per year in the elderly are fatal (University of Warwick, 2008). Over the last two decades, the incidence of pneumococcal resistance to penicillin and other antibiotics has alarmed the world and led to a growing concern regarding the control of pneumococcal disease. The natural competence of this species has been well documented since then, as it is the major determinant that contributes to the reservoir of penicillin-resistant bacteria (Appelbaum, 1987).

The competence (Com) QS in *S. pneumoniae* is used as a model of a QS cascade to study the evolution of QS systems in *Streptococcus* species. A homologous QS system, bacteriocin-like peptide (Blp), regulates the production of bacteriocins in *S. pneumoniae* (de Saizieu et al., 2000, Martin et al., 2006). Both functions are widespread

UNIVERSITI PENDIDIKAN SULTAN IDRIS

in the *Streptococcus* genus. Detailed descriptions of the Com QS and the homologous QS systems, Blp, in *S. pneumoniae* are presented in the following sections.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

1.1.1 Regulation of *S. pneumoniae* Com QS systems

S. pneumoniae has been one of the best-studied bacteria for more than 80 years, as it was first used to demonstrate transformation by Fred Griffith in 1928 (Griffith, 1928). The presence of the Com regulatory cascade has led to the categorisation of competence genes into two classes based on two main processes: i) the induction of competence in the bacterial cell; and ii) the mechanism of DNA uptake into the cell. Seven early competence genes (*comABCDE*, *comX* and *comW*), which are dependent on the phosphorylation of ComE (ComE~P) for their expression, are employed to induce competence (Pestova et al., 1996). Subsequently, 14 late competence genes that are dependent on the *comX* gene, which encodes an alternative sigma factor, are used to process the DNA that was taken up (Claverys and Havarstein, 2002).

An overview of the Com QS mechanism in *S. pneumoniae* is illustrated in Figure 1.1. In *S. pneumoniae*, the QS mechanism has been found to initiate the regulation of the competence system, which comprises an unmodified peptide pheromone, a secretion apparatus and a two-regulatory system (TCS) (Morrison and Lee, 2000, Claverys and Havarstein, 2002). Although *S. pneumoniae* has been reported to be competent only for a short period of time during its exponential growth (Palmen et al., 1994), this cell-cell signalling triggers a global change in gene expression, which leads to the transcription of the late genes. The late genes then encode proteins involved in the binding, processing, uptake and recombination of DNA.

Pakula and Walczak (1963), together with Tomasz and Hotchkiss (1964), discovered that streptococci use QS to measure the external concentration of their secreted signalling molecules, CSPs, to induce competence (Claverys and Havarstein, 2002). The *comC* gene was shown to encode CSP, which is a precursor peptide pheromone (Havarstein et al., 1995). CSP is an unmodified, 17-residue peptide containing a double-glycine (GG) type leader at its N-terminal end (Havarstein et al., 1995). CSP is exported and matured

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

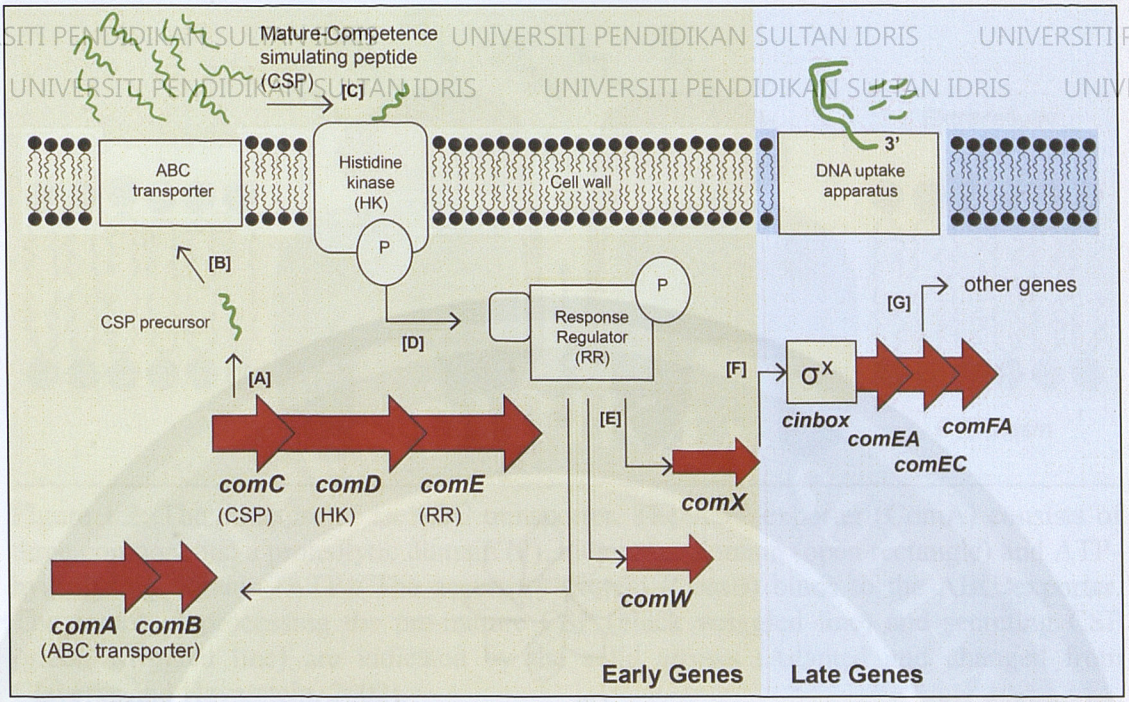


Figure 1.1: The regulation of the S. pneumoniae Com QS system. There are two main processes: i) competence induction (yellow box) involves early genes, which encode competence induction regulatory control components (comAB: an ABC transporter; comC: competence simulating peptide (CSP); comD: histidine kinase (HK); comE: response regulator (RR), ComW, ComX), respectively; ii) the DNA uptake mechanism (blue box) involves comEA, comEC, comEC, comFA and other 11 genes. The routes of the competence mechanisms are indicated by the black arrows in alphabetical order (Adapted and changed from Peterson et. al., 2006).

via the actions of ComAB, a dedicated ABC-transporter that consists of an ABC exporter (ComA) and an accessory protein (ComB) (Fath and Kolter, 1993).

ComA comprises three domains: a proteolytic domain, a membrane domain and an ATP-hydrolysing domain (Figure 1.2). An accessory protein, ComB, is required to assist the ABC exporter. Inactivation of either ComA or ComB affects the production of the peptide pheromone (Hui and Morrison, 1991, Hui et al., 1995). *comB* is located immediately downstream from *comA*. Although it is not clear how ComB assists ComA, ComB shares a similar feature with the accessory protein HlyD (Wang et al., 1991) at its N-terminus. The HlyD protein is co-expressed with HlyB (Hui and Morrison, 1991) to

UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

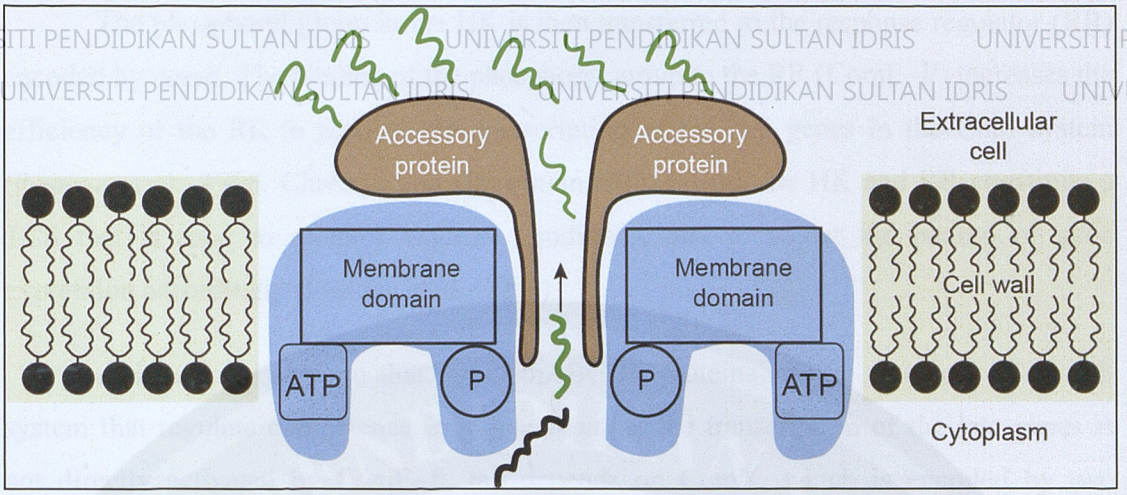


Figure 1.2: The structure of an ABC transporter. The ABC exporter (ComA) consists of three components: a proteolytic domain (P), membrane domain (open rectangle) and ATP-hydrolysing domain (ATP). The accessory protein (ComB) binds to the ABC exporter. The routes of processing the pre-mature CSP (black wriggled line) and secreting CSP (green wriggled line) are indicated by the solid arrows (Adapted and changed from Claverys and Havarstein, 2002).

provide a membrane structure for the export of α-haemolysin in the *Escherichia coli* haemolysis system that is involved in toxin production. ComA and ComB are most similar to LcnC and LcnD, which are the components of a bacteriocin ABC transporter in *Lactococcus lactis* (Hui et al., 1995).

Amino acid residues in the GG-leader of ComC are cleaved by the proteolytic domain at the N-terminus of the ABC transporter. The mature peptide pheromones are transported across the membrane, mediated by the membrane domain, with the energy provided by the ATP-hydrolysing domain. While the ABC transporter processes and exports CSP outside the cell, a dedicated TCS, the histidine kinase (HK) encoded by the *comD* gene (Pestova et al., 1996) detects the extracellular CSP. It was found that the HK (ComD) receptor belongs to the class of orthodox kinases that are composed of a highly variable membrane-spanning N-terminal sensor region and a more conserved C-terminal kinase that enables histidine to accept a phosphoryl group (Havarstein et al., 1996). The binding of CSP changes the conformation of the HK, which results in its dimerization and, subsequently, its autophosphorylation (Stock et al., 2000).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

The phosphoryl group in the HK is then transferred to the response regulator (RR) encoded by *comE*. The binding of the phosphoryl group to the RR (ComE~P) increases the efficiency of the RR to activate the transcription of the late genes in the Com system (Pestova et al., 1996, Claverys and Havarstein, 2002). Both the HK and RR constitute a TCS that is used to monitor external conditions, and to adjust the pattern of gene expression accordingly (Pestova et al., 1996).

Although it is known that the ComABCDE proteins are the elements of the QS system that regulate competence in *S. pneumoniae*, the transcription of the late genes is not directly activated by ComE~P, but depends on ComX, which is encoded by two identical genes, *comX*1 and *comX*2 (Lee and Morrison, 1999). They have been shown to have an imperfect direct repeat motif similar to the ComE binding site. This similar character suggests that ComE activates the expression of ComX1 and ComX2, which encode a specific alternative sigma factor required for the synthesis of competence proteins that are responsible for the DNA uptake process.

The late genes are expressed through a common regulatory element called the *cin*-box (Campbell et al., 1998, Morrison and Lee, 2000). The *cin*-box contains a conserved 8-mer motif (TACGAATA) that represents a typical binding site for the sigma factor. This conserved motif is usually centred approximately 10 base pairs upstream from the initiation site of transcription. Later, *comW* was shown to encode a protein that mediates the post-transcriptional control of *comX* (Luo et al., 2004, Sung and Morrison, 2005). ComW is required for the stabilization and activation of the sigma factor ComX. The activation of the sigma factor ComX results in the recognition of the motif conserved in the promoter regions of the late genes.

It has been shown by Lee and Morrison (1999) that the expression of the late genes is dependent on *comX*. Several independent studies (Martin et al., 1995, Campbell et al., 1998, Dubnau, 1999, Petersen and Scheie, 2000) have identified four major classes of ComX, based on the function of the proteins expressed by the late genes.

The first class of proteins encoded by the late genes has the ability to bind all types of double-stranded DNA present in the environment, and to process and transport one strand into the cytoplasm (Dubnau, 1999). The proteins in this class are known as DNA

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

uptake proteins. The second class of proteins is involved in chromosomal integration, e.g., RecA, which was shown to be essential for DNA recombination (Martin et al., 1995), and SsbB, which was found to have a major effect on the transformation rate of *S. pneumoniae* (Campbell et al., 1998). These proteins are classified as recombination proteins. The third class of proteins is dispensable for transformations, and includes LytA (Mortier-Barriere et al., 1998), and the last class consists of the proteins with unknown functions (Peterson et al., 2000).

1.1.2 DNA uptake mechanism and transformation

When a competent cell is in the 'transient' state of being competent, the *comX*-encoded sigma factor activates the transcription of the late genes that are essential for DNA uptake (Lee and Morrison, 1999). A double-stranded DNA (dsDNA) that is bound at the membrane of the competent cell is transported into the cytosol by this mechanism (Dubnau, 1999, Chen and Dubnau, 2004). The DNA uptake mechanism was characterized long before the uptake genes were identified. The characterisation of the DNA uptake mechanisms was mostly done by Dubnau and his colleagues, firstly in *Bacillus subtilis*, and subsequently on *S. pneumoniae* (Dubnau, 1999, Chen and Dubnau, 2004, Chen et al., 2005).

As illustrated in Figure 1.3, the uptake mechanism is divided into three steps. The first step involves the initial binding of the exogenous dsDNA to the cell surface. In this step, a transformation pseudopilus and a surface endonuclease are required to transport and process the dsDNA. The pseudopilus comprises a major pilin-like protein (ComGC), and minor pilin-like proteins (ComGD, ComGE and ComGG). The pseudopilus is evolutionarily related to Type IV pili (Chen and Dubnau, 2004, Chen et al., 2005). The binding of the dsDNA to the cell surface has been suggested to occur via the interaction between the negatively charged DNA backbone and positively charged residues in the pseudopilus proteins (Craig and Li, 2008).

In the next step, the retraction of the ComG proteins of the pseudopilus allows the dsDNA to pass through the peptidoglycan and contact the receptor protein, ComEA (Dubnau, 1999). The receptor then delivers the dsDNA to EndA. The EndA endonuclease

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

degrades one strand of the dsDNA, as only a single-stranded DNA (ssDNA) can traverse through the trans-membrane channel protein ComEC. The transportation of the ssDNA through ComEC is assisted by the ComFA, which is an ATP-binding protein that mediates the ssDNA internalisation (Dubnau, 1999).

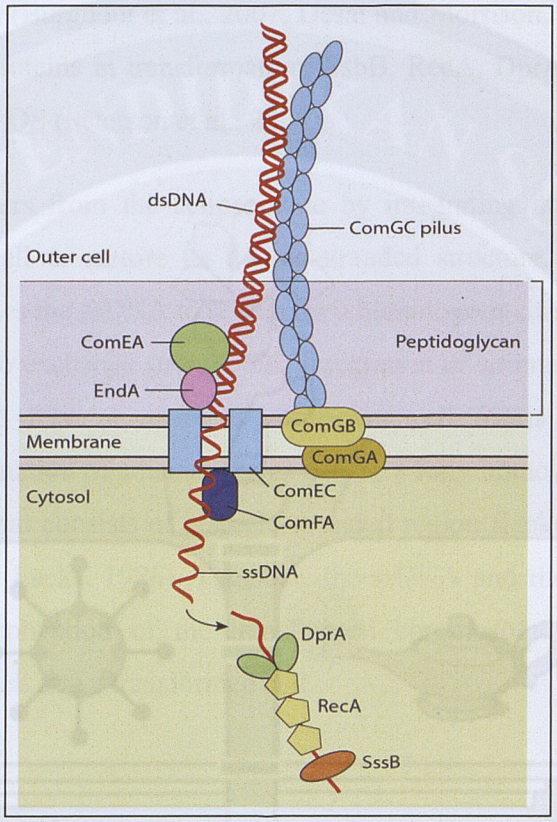


Figure 1.3: The DNA uptake and transformation mechanism. The recombination process by which the transported ssDNA is spliced into the chromosome of the recipient competent cell is shown.

Upon entry of the ssDNA into the cytosol of the competent cell, the ssDNA enters a transient state of eclipse. During this state, the ssDNA loses its ability to transform (Fox, 1957). Later, the ssDNA is bound by a set of proteins to form an eclipse complex (EC) (Morrison, 1977, Morrison, 1978). One of the major protein components of an EC is the SsbB protein, and the other components include SsbA, an SsbB paralog (Thanassi et al., 2002), DprA and RecA, which are all important for the DNA transformation process in

UNIVERSITI PENDIDIKAN SULTAN IDRIS

S. pneumoniae (Peterson et al., 2004). SsbB (probably together with SsbA) protects the ssDNA from degradation by endonucleases and assists with the recombination process (Morrison et al., 2007). DprA, a recombinase loader, binds to the ssDNA and promotes the interactions of the ssDNA with RecA (Mortier-Barriere et al., 2007) which is an essential DNA-strand exchange protein (Martin et al., 1992). Other protein components found in the EC are CoiA and RadA, which are crucial contributors to the optimisation of the recombination process (Burghout et al., 2007, Desai and Morrison, 2007). The expression of the five essential proteins in transformation: SsbB, RecA, DprA, RadA, and CoiA, is dependent on the Com QS (Peterson et al., 2004).

The EC recovers from the eclipse state by integrating into the genome of the recipient competent cell to restore its double-stranded structure. The incorporation of RecA into the EC allows the ssDNA to search for a homologous site in the chromosome of the recipient cell and to exchange strands. The integration of an internalised ssDNA into a complementary strand from a recipient competent cell forms a structure called a heteroduplex. The formation of a heteroduplex can be a fully homologous or heterologous recombination event that consists of a single-stranded region flanked by the homologous regions (Claverys and Lacks, 1986). Together, the ssDNA and the flanking regions are covalently joined. Replication of the transformed chromosome during cell division produces a wild-type cell and a transformed cell.

1.1.3 Regulation of competence via QS systems in other *Streptococcus* species

The first discovery of natural transformation occurred in 1928, when Griffith observed that non-virulent strains of *S. pneumoniae* were able to escape the immune system and cause pneumonia (Griffith, 1928). Later, the finding contributed to the discovery of DNA as the genetic material (Avery et al., 1944). Not all bacteria are naturally competent. However, the ability of a competent cell to naturally transform is unpredictably sporadic in the bacterial population, especially in the *Streptococcus* genus.

Most of the species, which are previously found to be naturally competent only in Mitis group, have also been observed in other streptococcal group: Anginosus group, Mutans group, Pyogenic group, and recently in *S. suis* (Leblanc et al., 1978, Gaustad,

1979, Perry and Kuramitsu, 1981, Havarstein et al., 1997) To date, approximately 67 prokaryotes, including 17 streptococci (Appendix C), are known to be naturally transformable (Lorenz and Wackernagel, 1994, Johnsborg et al., 2007).

1.1.4 The bacteriocin QS system: a closely related competence QS system in *S. pneumoniae*

S. pneumoniae contains a second QS regulation system (ComABCDE), known as BlpABCHR (de Saizieu et al., 2000). Both systems consist of: a peptide pheromone, encoded by *comC* and *blpC*; secretion apparatuses, *comAB* and *blpAB*; and also TCSs, *comDE* and *blpHR* (see Figure 1.4). Studies revealed that the function of BlpABCHR is to regulate the production of bacteriocin-like peptides (hence the name Blp) (de Saizieu et al., 2000, Reichmann and Hakenbeck, 2000). It has also been shown that BlpA and BlpB are highly related to ComA and ComB (ABC exporter), which have been shown to have an important role in exporting and processing the ComC peptide (de Saizieu et al., 2000). Both ComC and BlpC have the same responsibility to synthesise precursor peptide pheromones with double-glycine leaders at their N- termini.

Both QS systems appear to be regulated similarly; however, a study by de Seizieu et al. (de Saizieu et al., 2000) reported that these two QS systems appear to operate at different cell densities. The autoinduction of the Com QS system occurs in the early logarithmic phase, whereas the autoinduction of the Blp QS system occurs in the late logarithmic phase. The difference between these QS systems suggests that they operate independently of each other, probably without cross-communication at the pheromone and HK receptor levels (de Saizieu et al., 2000).

1.1.5 The importance of the Com QS system

Competent bacteria cells are able to uptake extracellular DNA fragments. The fates of the incoming DNAs are either for: i) nutritional purposes (Finkel and Kolter, 2001); ii) to facilitate the adaptation of the population through DNA repair (Michod et al., 1988, Wojciechowski et al., 1989, Hoelzer and Michod, 1991); or even iii) to increase the

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

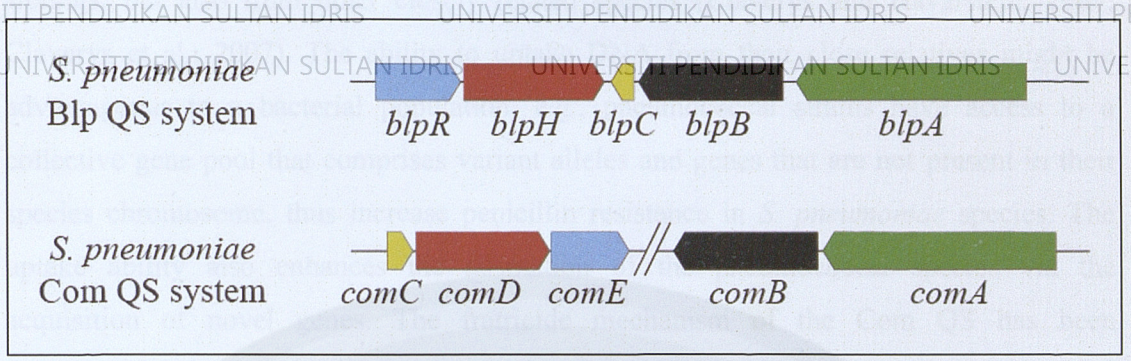


Figure 1.4: The QS regulon structure of the bacteriocin-like peptide (Blp) and competence (Com) system of *S. pneumoniae*. The homologous genes in both QS systems are coloured accordingly: ABC exporter (blpA/comA), accessory protein (blpB/comB), peptide pheromones (blpC/comC), histidine kinase (blpH/comD) and response regulator (blpR/ comE).

genetic diversity in the bacterial population (Tortosa and Dubnau, 1999, Morrison and Lee, 2000). However, the benefit of the competent bacteria cells remains unclear.

The first argument is that the incoming DNA is a source of food. Considering the complex and tightly regulated DNA uptake mechanism, as previously discussed in section 1.1.2, it is a bit crude to assume that the DNA is used as food for the bacteria. The fact that only a single strand of the extracellular dsDNA is translocated into the cytosol, while the other strand is degraded (Chen et al., 2005), would render this an inefficient food-uptake system and a metabolically costly behaviour.

Next, transformation is argued to provide advantages to the competent cells. The DNA is being taken up to repair damaged DNA and improve the population's adaptation to deleterious environmental changes. However, Redfield (Redfield, 1993) argues that the successful rate of repairing damaged DNA is low because only a very small fraction of extracellular DNA is being taken up. Additionally, bacteria cells do not usually uptake homologous DNAs that match the damaged sequence (Mongold, 1992).

The last argument seems to fit the purpose of transformation, whereby transformation promotes genetic diversity in a bacteria population (Tortosa and Dubnau, 1999, Morrison and Lee, 2000). Interestingly, a fratricide mechanism has been discovered in the *Streptococcus* species, whereby the streptococci not only uptake DNA from their

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

siblings, but also from other closely related species (Claverys and Havarstein, 2007, Claverys et al., 2007). The ability to uptake DNA from their close relatives might be advantageous to a bacterial population, e.g., pneumococcal strains have access to a collective gene pool that comprises variant alleles and genes that are not present in their species chromosome, thus increase penicillin resistance in *S. pneumoniae* species. The uptake ability also enhances the adaptation of the pneumococcal species via the acquisition of novel genes. The fratricide mechanism of the Com QS has been demonstrated to promote horizontal gene transfer (HGT) within and between streptococcal species (Johnsborg et al., 2008). However, competence only occurs in some species in certain habitats (Dubnau, 1999, Claverys et al., 2006).

1.2 Computational approaches in molecular evolution

There has been continual development of evolutionary theories since July 1, 1858, when Charles Darwin proposed an evolutionary theory by means of natural selection at the Linnaean Society of London. In 1859, Darwin shared his theory of evolution by publishing "The origin of species by means of natural selection." Although it was clear to him that natural selection was extensively defined in term of heritable traits, he did not describe how biological information was transmitted from one generation to the next. The nature of heredity was poorly understood at that time.

In 1866, Gregor Mendel tried to fill the gap in Darwin's theory by breeding and cross-breeding 30 000 green pea plants. The successful outcome of his experiment resulted in Mendel's laws of inheritance. In 1930s, the integration of Mendelian genetics with natural selection theory, by Seawall Wright, J. B. S. Haldane and R. A. Fisher, marked the starts of Modern evolutionary synthesis, population genetics, to describe the process of evolution.

A breakthrough resulted from a series of experiments (Griffith, 1928, Avery et al., 1944, Hershey and Chase, 1952) has changed the common perspective during the time that DNA is an important hereditary material. Soon after, Watson and Crick (1953) discovered the first DNA molecular structure, a double helix, which evolution can be taught of as changes in DNA structure. DNA sequences have become valuable since then as they

provided most detailed anatomy and the regulation of how each part should be assemble and operate. In 1955, a study by Fred Sanger and his colleagues, revealed how species differed at the molecular level by comparing the amino acid sequences of protein insulin from cattle, sheep and pig (Brown et al., 1955). All of the events are responsible for the beginning of molecular evolution where DNA are now used as the primary data sources.

1.2.1 Sources of genomes

The abundant access to sequenced genomes has led the discovery of many genes and the functions of their encoded proteins. Each of these genomes, along with its annotation information, literature references, identifiers of related sequences, is stored and organised in a sequence database, such as the National Center for Biotechnology Information (NCBI) Genbank.

Genbank is intended to be an archive of primary sequence data, where some quality control checks are applied to each of the submitted sequences (Benson et al., 1998). However, Genbank contains sequences of loci from multiple submissions that resulted from similar experiments, each conducted under slightly different conditions. Thus, NCBI built a Reference Sequence (RefSeq) database, which is a curated collection of DNA, RNA and protein sequences. RefSeq aims to provide separate and linked records for genomic DNAs, gene transcripts and encoded protein sequences (Pruitt et al., 2006). It is also limited to major organisms, with sufficient available data ranging from viruses and bacteria to eukaryotes. The advantage of RefSeq records is that the best loci representation of a sequence was selected from existing primary GenBank data based on the best available information (Pontius et al., 2003). This would further prevent any contradicting records that share the same common loci.

The easiest way to access these sequence databases is by using the Entrez tool, a medium to search across many databases that include DNA sequences, protein sequences, protein structures and genomes. Another popular tool is the Basic Local Alignment Search Tools (BLAST), which rapidly searches related sequences in the database based on their query sequence similarities (Altschul et al., 1990, Altschul et al., 1997).

UNIVERSITI PENDIDIKAN SULTAN IDRIS

1.2.2 Multiple sequence alignment

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

An accurate assembly of multiple sequence alignment (MSA) is needed to predict the evolution history of a species, a gene profile, or a protein structure (Kemena and Notredame, 2009). However, the alignment programs available today have yet been able to produce a biologically impeccable MSA, which is mostly due to the limitation on the computer ability. A heuristic approach is often used to infer the best assembly of MSA by optimising the similarity between sequences.

One of the most popular heuristic approaches is the progressive alignment (Feng and Doolittle, 1987). This approach progressively built the MSA by calculating the similarity scores of the pairwise alignment between each possible pair of sequences using methods such as Needle-Wunsch algorithm (Needleman and Wunsch, 1970), Smith-waterman algorithm (Smith and Waterman, 1981). The similarity scores represent the evolutionary distances between the sequences, which then be converted into a score matrix. A guide tree will be estimated from these scores using a simple tree building method e.g., Unweighted Pair Group Method and Arithmetic Mean, UPGMA. The guide tree is then used as a reference to assemble the sequences one by one to the alignment. The most similar sequence is added first to the alignment and gradually followed by the sequences that are more distant. The final alignment resulted in the process is therefore influence by the order of the sequences that are being added into the alignment. As the approach depends on the initial pairwise, any mistakes made in the initial stage are fixed and cannot be corrected. This mistake could get worse as the number of sequences increase, thus producing a final alignment that is far from optimal.

Iterative alignment approach was developed to alleviate the problems occurs in the progressive alignment approach by employing similar steps in the progressive alignment approach but allowing iterative realignment at the initial stage and while adding the sequence into the final alignment. The realignment process does help to correct the initial mistakes and improve the overall alignment (Edgar, 2004, Mount, 2009, Sievers et al., 2011). In this study, MUSCLE alignment program (Edgar, 2004), which is an iterative progressive alignment approach, was used to align the homologous sequences for the phylogenetic analysis.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

1.2.3 Phylogenetic inference methods

A phylogenetic tree is a graphical representation of the evolutionary relationships among a set of sequences or specific genes from different organisms. A phylogenetic tree is composed of nodes that are connected by branches. A terminal node at the tip of a tree represents a descendent taxon, which often refers to a species.

The length of each branch in a phylogenetic tree represents the amount of change estimated between the sequences of the examined species. To construct a phylogenetic tree, the nucleotide or protein sequences of the examined species need to be aligned using alignment methods, as previously described.

Because of a rapid growth in the number of available sequences and continuous improvements in alignment methods, it is now possible to compute a large phylogeny of thousands of organisms. However, the task to compute the tree of life, which includes all organisms on earth, remains the biggest challenge.

A fundamental problem faced in computing a phylogeny is the enormous number of possible trees, which represent the evolutionary histories of organisms. The number of possible trees is exponential to the number of the sequences. The number of possible trees for 60 organisms is 5.86×10^{96}, which is larger than the number of atoms in the universe ($\approx10^{80}$).

There are two categories for inferring phylogeny. The first category is distance methods and the second category is character-based methods. Distance methods construct a phylogenetic tree based on the pairwise distance matrix values that were computed from the alignment of homologous sequences. The distance method assumes the distances between two taxa are equall to the sum of branches that connected them. The Neighbor-Joining (NJ) method, which is a distance method, uses a star tree as the initial tree. An internal node is constructed between the two most similar sequences with the shortest distances. This step repeats until the tree is fully resolved. The distance method is a popular choice of inferring a phylogenetic tree due to its computationally fast performance and its capability to handle large datasets. However, the NJ method does not consider other possible trees. This leads to a problem when there is more than one pair of sequences with the same shortest distances, which suggest that there is more than one possible tree to

UNIVERSITI PENDIDIKAN SULTAN IDRIS

represent the evolutionary relationship between these sequences. Ignoring all the tree options may result in a suboptimal solution. Improvements were made on the NJ method by incorporating a substitution model to represent sequence divergence. However, the information regarding the sequence divergence is lost when the sequence variation is reduced to a single distance value.

In contrast, methods in the character-based category use sequence characteristics to infer the evolutionary history between sequences. The evolutionary information is embedded into the building process of the phylogenetic trees. Two of the most popular approaches are the Maximum Parsimony (MP) and Maximum Likelihood (ML) methods.

A tree with the fewest number of changes, or which has the shortest total branch lengths, is chosen as the most parsimonious tree in the MP approach. The approach assumes the tree with the minimum number of changes is the best explanation of the evolutionary relationship between the taxa under study. This assumption is supported by the fact that evolutionary change is a relatively a rare event within a reasonably short time frame. By minimising the number of change, the approach minimises the amount of phylogenetic noise, such as homoplasy. Therefore, the MP approach is likely to be a good estimate of the true tree.

The MP approach searches all possible trees based on the number of taxa, and reconstructs the ancestral sequences that minimise the number of changes needed to evolve the current sequences. Because the ancestral state is not known, it is possible to have multiple solutions. Using the parsimony principle, the character state that requires the fewest number of substitutions is chosen when multiple possible solutions of the ancestral state exist. The determination of the ancestral state needs to done carefully. There is a possibility that two tree topologies will have equally parsimonious solutions. Thus, a consensus tree needs to be built to represent all parsimonious solutions.

One advantage of the MP approach is that the method is easily understood. The approach also provides evolutionary information and incorporates this information into the tree construction. Thus, the tree inferred by the MP approach is more accurate than those achieved using a distance-based method. However, the parsimony assumption is not suitable to be employed for inferring highly divergent sequences, or sequences with high

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

amounts of homoplasy. The MP method recognises most of the sites from these types of sequences as non-informative sites. Thus, some (important) evolutionary signals may be lost because MP ignores the non-informative sites. The MP method is also sensitive to the "long-branch attraction" (LBA), in which two long branches are prone to be located together despite their true positions on a tree. This is also probably due to the parsimony assumptions that all mutations are equally distributed to the branch lengths and all lineages evolve at the same rate.

ML, which is a more rigorous character-based method, searches all possible tree topologies, and considers all sites of sequences in an alignment. ML is a probabilistic approach in phylogenetic inference, in which stochastic models are used to infer a tree from sets of sequences (Felsenstein, 1981). ML is the probability of observing the evolutionary relationship between a set of sequences under a specific model on a tree. It is not used to determine whether a model is correct or not, but to examine whether a tree under a specified evolutionary model is sufficient to describe the evolutionary history of the observed sequences.

To implement the ML method, three main elements are required, which are a molecular evolutionary model, sequence alignment data and trees. The likelihood value, L, is given by the probability of observing the data under the chosen model on a tree, $L = P$ (data | model, tree). The likelihood is presented as a log-likelihood. The larger (less negative) the log-likelihood value of a tree, the more likely the tree describes the observed data. The strengths of the ML approach are that it uses all sequence information, employs substitution models and is not sensitive to the LBA. However, these strengths could also be seen as weaknesses. Because ML is dependent on the chosen substitution model, a wrong choice of a substitution model could jeopardise the construction of the ML tree. Because of the exhaustive nature of ML, it is computational costly; thus, it may not be feasible to use it for inferring a tree for a large dataset.

The substitution models describe the probability of change in nucleotides or amino acids in a transition rate matrix. Substitution models make some specific assumptions based on the homogeneity, reversibility of nucleotide frequency and rate of substitutions (Lio and Goldman, 1998, Whelan et al., 2001). Amino acid replacement models were developed using the empirical approach. The first amino acids model is the Dayhoff

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

model, which was produced by Dayhoff and co-workers (Dayhoff and Schwartz, 1978). The prediction of this model is based on the frequency of amino acid changes in other related proteins, with the assumption that no multiple substitutions have occurred between two sequences of high identity via parsimony. Dayhoff and co-workers documented the matrices of mutation probabilities, known as the probability of accepted mutation (PAM), which were inferred from 34 protein families. Much later, Jones, Taylor and Thornton (1992) produced a Jones-Taylor-Thornton (JTT) model, which is an upgraded version of the PAM matrices, using a much larger database.

The success of these models in predicting evolutionary relationships cannot be denied, but they do make oversimplifying assumptions. The matrices may not work on highly divergent sequences due to their nature to do pairwise comparison between two sequences. These matrices also discard sequence regions that have low identity, which may throw away importance evolutionary information. The limitation of the counting method via parsimony, such as not being able to exploit multiple sequence alignments, may underestimate the phylogeny inferences.

The incorporation of an explicit substitution model into the ML approach was first demonstrated by Whelan and Goldman (2001). The WAG model incorporates an approximate method of phylogenetic inference by allowing multiple substitutions in all sites. Whelan and Goldman address the shortcomings of previous models through their WAG matrix, which is estimated from 180 protein models using the likelihood approach. This incorporation would model evolutionary processes and improve the likelihood score of the inferred phylogeny in comparison to other, older amino acid replacement matrices. The WAG matrix was found to be robust and tended to perform well in many cases (Ren et al., 2005).

However, substitution models assume that all sites have evolved equally and are under the same evolutionary pressure, which may not hold up in reality. Different sites are exposed to different environments and may perform different functions. For protein sequences, some amino acids change more slowly than others due to functional or structural constraints. Ignoring variations in the evolutionary rates among sites can cause artefacts in the phylogeny construction. The common practice to account for variations among sites is to incorporate the gamma (Γ) discrete distribution in the tree inference

UNIVERSITI PENDIDIKAN SULTAN IDRIS

process. Using the evolutionary rate heterogeneity has greatly improved the likelihood of the tree inference (Yang, 1994, Yang, 1996a).

The next challenge in inferring phylogenies is best tree searching. The scores produced from the optimal parameters can be used to rank the tree topologies from best to worst. There are a few approaches that have been used to search for the best tree topology in a tree space. The size of the tree space depends on the number of taxa, as the number of possible tree topologies increases exponentially with the number of taxa. We can look through all possible tree topologies or build up a tree using the optimal characterisation along the way. However, to look at every possible tree may not be feasible, especially on a large number (more than 20) of taxa, as it is very computationally costly and time consuming. Thus, a heuristic search method has to be used to look into the set of trees that appear to be most promising, based on the calculated scores. A heuristic search is an approximation strategy that generates a quick answer, but certainly not the best answer to a complicated problem.

A heuristic search algorithm works like hill climbers, whereby it only climbs up and never takes a step down. It starts by building an initial approximation tree, which is usually an NJ tree, and modifies the tree into a different topology by cutting a branch or subtree and regrafting it onto the remaining tree to see whether that leads to a better tree. If the score of the modified tree is better, then the tree will be used as the initial point in the next tree search iteration. The process iterates until no next tree with a better score is found.

Three common branch-swapping algorithms in heuristic tree searches are Neighbour-Nearest Interchange (NNI), Subtree Pruning-Regrafting (SPR) and Tree Bisection-Reconnection (TBR). In NNI, the subtrees on a tree swap their internal connections between each other to produce other alternative trees. Alternatively, in SPR, an interior or exterior branch is pruned and then reinserted into a branch of the remaining tree to create a new node. While in TBR, a subtree is detached from the remaining tree. Then, TBR considers all possible connections between all the branches in the detached subtree and the remaining tree. The TBR and SPR approaches are similar in that both approaches detach a subtree, but only SPR keeps the original subtree rooting.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

RAxML (Stamatakis, 2006b) is a program that uses the ML method to estimate phylogenetic trees. The program is heavily used as a tool to construct phylogenetic trees throughout the studies in this thesis. The RAxML program uses a 'lazy' SPR approach to identify the best ML tree. The approach is similar approach to SPR, but it only optimises the adjacent branches where the pruned tree is attached. The program certainly reduced the time needed to find the ML tree (Stamatakis and Alachiotis, 2010). However, the approach does not ensure the provision of an optimised global tree, which is a classic problem in tree search.

1.2.4 The reliability of phylogenetic inference

To test the reliability of the inferred phylogeny, analytical re-sampling strategies, such as bootstrapping, are required. Bootstrapping is a statistical technique to test the robustness of the inferred phylogenetic tree by repeatedly sampling trees from a sequence alignment. Bootstrapping methods are used to evaluate whether the best tree produced could not have occurred by chance or be affected by sampling errors, such as sites that have low phylogenetic signals. Rearranging the sites of an alignment for sampling provides an indication of the robustness of a tree. There are two types of bootstrapping methods: i) parametric bootstrapping and ii) non-parametric bootstrapping. The difference between these two tests is the way in which the sampling is performed. Non-parametric bootstrapping randomly picks the alignment site (alignment column) to generate a new MSA, which has different nucleotide or amino acids composition, of a same length. As a result, some sites may appear multiple times and some may not appear at all in the new MSA. This process is repeated depending on the number of samples (n) to create n new alignments that are used to construct trees via the same methods for inferring the 'original' tree. Because the new alignments have different rates of heterogeneity, each may tend to produce a different tree topology. A major rule of consensus is usually used to summarise the supported branching patterns of the best tree by showing the bootstrap values. The bootstrap value represents the percentage of a branching pattern appearing in the sampling trees. A bootstrap value of approximately 70% corresponds to a 95% statistical confidence (Hillis and Bull, 1993).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

In contrast, in parametric bootstrapping, new MSA samples of the same length are generated from the distribution of the best tree under a given substitution model. The best tree inferred from the MSA is assumed to have an estimation close to the true tree, and that the model used to infer the tree is a close estimation of a true model. A tree is inferred from each of the simulated new alignment samples. All resulting trees are analysed using the majority consensus method, similar to the non-parametric bootstrap approach. The advantage of using a parametric bootstrap is that the MSA samples are drawn from the desired distribution, rather than randomly generated. However, a wrong choice of model may greatly affect the results, as the approach is highly dependent on the chosen model.

The reliability of the phylogenetic inference can also be tested by determining the degree of difference of two tree topologies and which tree topology better explains the evolutionary history of an organism. The suitability of the Kishino-Hasegawa (KH) (1989) test that is used to compare ML trees has been argued (Goldman et al., 2000). The KH-test hypothesis is that there is no significant difference in the log-likelihood between tree topologies:

H0: $L1 - L2 = 0$

H1: $L1 - L2 \neq 0$

which is not suitable to compare ML trees. This is because the ML tree will always have the best likelihood score of any compared trees; therefore, the difference is always greater than zero ($L1 - L2 > 0$). This may lead to an incorrect rejection of non-ML trees and overconfidence in the wrong trees. The KH-test is also not suitable to compare more than two trees because of the multiple comparisons problem.

A statistical method that is usually used to compare all of the ML trees is the Shimodaira-Hasegawa (SH) test (1999). The SH test allows all possible tree topologies, including the best ML tree, to be used in multiple comparisons. The null hypothesis of the SH test is that there is no significant difference between all possible trees, which implies that all trees are good explanations of the data.

H0: All compared trees are equally good explanations of the sequence dataset.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

H1: Some or all compared trees are not equally good explanations of the sequence dataset.

Nonparametric bootstrapping is used to replicate the dataset and to obtain bootstrap replicate statistics. Each alternative tree topology is tested to see whether it is significantly different from the ML tree topology. For example, an alternative tree is significantly different from the ML tree if the P-value of the alternative tree is less than 0.05.

1.3 Molecular evolution and phylogenomics

1.3.1 Molecular evolution mechanisms

Today, more than a thousand genome sequences are available, and, yet, making sense of the sequence data remains a challenge. Inferring evolutionary histories is one reliable approach to handling and manipulating these abundant gene sequences. The homology concept is fundamental in understanding the evolutionary process. The similarities between homologous sequences indicate their divergence from a common

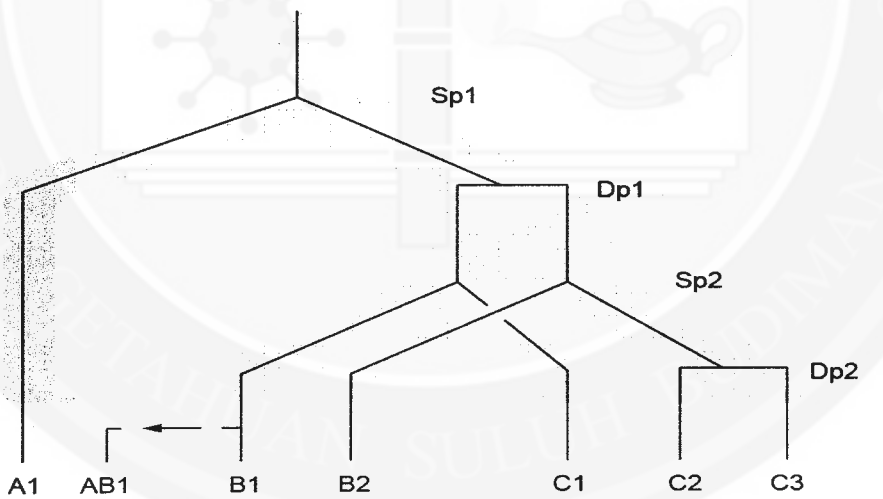


Figure 1.5: The concept of orthologs, paralogs and xenologs. The diagram taken from (Fitch, 2000). The black lines represent the history phylogeny of a gene in three species A, B, and C. The thick grey lines show the species phylogeny. An inversed Y (Sp1, Sp2) marks a speciation event, and a horizontal line (Dp2, Dp1) marks a duplication event. C2 and C3 that derived from a duplication event are paralogs. However, both (C2, C3) are orthologous to B2 as they derived from a speciation event of a common ancestor. AB1 is xenolog to all other genes indicated by the red arrow.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

ancestor. All of the homologous genes can be further classified into orthologs, paralogs, and xenologs (Figure 1.5) (Fitch, 2000). Homologs are described as orthologs if the set of homologous sequences is derived from the last common ancestor (LCA) via speciation events, while the homologous sequences that were derived from LCA through duplication events are classified as paralogs. Homologous sequences that have non-vertical history, horizontal gene transfer (HGT), are described as xenologs.

In prokaryote (bacteria and archaea), studies have found that HGT and gene loss are the two major evolutionary process, which contribute to the intensive 'gene flux' that seems to have shaped the genomes of these organisms (Lorenz and Wackernagel, 1994, Jain et al., 1999, Hao and Golding, 2004). This scenario applies to some of the *Streptococcus* species, e.g., *S. pneumoniae*, which has the ability to uptake foreign DNA into the cell. HGT is a common mechanism in the species and it is suggested to promote the emergence of a new virulence and the widespread of antibiotic resistance (Claverys et al., 2007). However, little experimental work that has been done to tackle: the HGT effects on the efficiency of transformation, and species adaptation and fitness; and to find out how specific condition and environment assist HGT mechanism in this species.

The limitation for the competent cell to uptake homologous DNA to be recombined to its DNA does not stop the competent cell to uptake a gene from other organism, like plant or fungi. (Ochman et al., 2000, Claverys et al., 2007). The availability of the complete genomes has offered a source to study HGT and drawn the attention to how HGT affect the construction of evolutionary tree e.g. the tree of life. The difficulty to detect horizontally transferred genes using the available methods remain a challenge.

One way to detect xenologous genes is using the phylogenetics analysis. Therefore, gathering the information from molecular data and phylogenetic approaches are essential in understanding the HGT mechanism in the *Streptococcus* genus. The first *Streptococcus* tree was published by Kawamura et. al, (1995) (Figure 1.6). The tree was estimated from 16S ribosomal RNA sequences using the Neighbor-joining method. The species in the *Streptococcus* genus were found to be clustered into six following major clusters: Pyogenic group, Mutans group, Bovis group, Salivarius group, Mitis group and Anginosus group (Kawamura et al., 1995).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

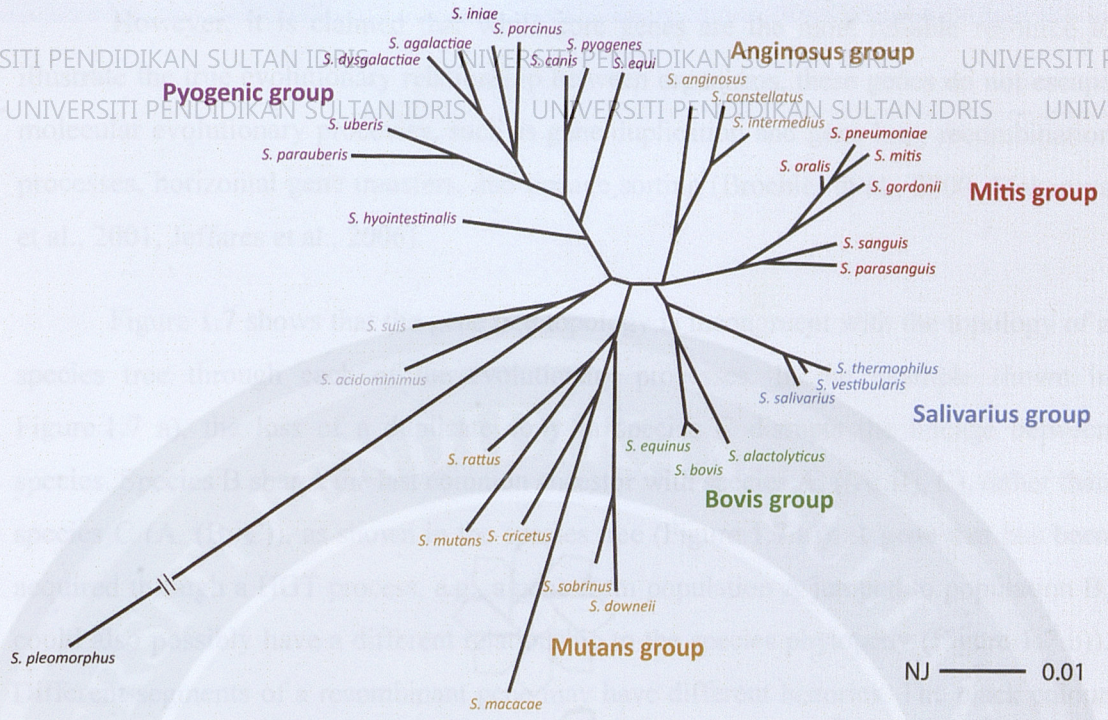


Figure 1.6: The *Streptococcus* phylogeny from 16S ribosomal genes (Kawamura et al., 1995). The tree was inferred using NJ method for 34 Streptococcus species.

1.3.2 Gene trees and species trees

A gene tree depicts the relationship between homologous genes or the encoded proteins of organisms in which they have resided, whereas a species tree represents the historical relationship between organisms. Sets of genes from these organisms are expected to have the same histories as the organisms' lineages. Therefore, it is possible to assume that gene trees have the same tree topology as species trees.

A common approach in phylogenetic study, a species tree is estimated from the gene trees of a set of individual genes. There are many arguments regarding whether a species tree exists, and what is the ideal sample size of the genes that should be used to represent the evolution of organisms (Martin, 1996, Maddison, 1997, Dagan and Martin, 2006, Slatkin and Pollack, 2006). A species tree is usually built using core genes, e.g., ribosomal genes, which are conserved, functionally important and slowly evolving, from the genome of an organism. Using core genes as a source to build a species tree would retrieve the 'true' evolutionary relationship between the organisms.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

However, it is claimed that while core genes are the most reliable resource to illustrate the true evolutionary relationship between organisms, these genes do not escape molecular evolutionary processes, such as gene duplication and gene loss, recombination processes, horizontal gene transfers, and lineage sorting (Brochier et al., 2000, Makarova et al., 2001, Jeffares et al., 2006).

Figure 1.7 shows that the gene tree topology is incongruent with the topology of a species tree through each of the evolutionary processes. In the example shown in Figure 1.7 a), the loss of a duplicate copy in species C disrupts the lineage between species. Species B shares the last common ancestor with species A, ((A, B), C), rather than species C (A, (B, C)), as shown in the species tree (Figure 1.7 a)). A gene that has been acquired through a HGT process, e.g., a gene from population A jumped to population B, could also possibly have a different relationship to the species phylogeny (Figure 1.7 b)). Different segments of a recombinant gene may have different histories. The black colour segment of the C gene has (A, (B, C)) relationship, while the white segment of the C gene has ((A, C), B) relationship (Figure 1.7 c)). Another complex event is lineage sorting, in which a speciation event in a gene tree occurs before the speciation event in the species tree.

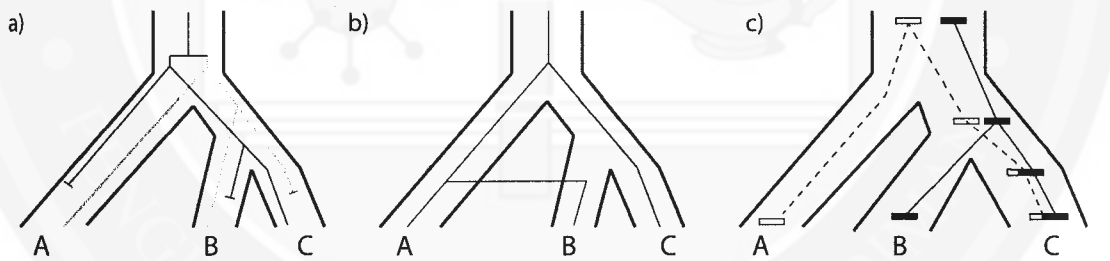


Figure 1.7: Gene trees and species tree. The molecular evolutionary processes that have caused discordance between the species tree and the gene trees. The causes are: a) duplication and gene loss; b) horizontal gene transfer; and c) gene recombination.

The discordance between species trees and gene trees can also be caused by estimation errors that occur during the phylogenetic analysis, e.g., sampling errors, stochastic errors (Phillips et al., 2004). As a result, the relationship between a species tree and a gene tree is more complex than a one-to-one correspondence.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

1.3.3 Phylogenomics

The construction of phylogenetic trees was historically based on the characteristics of morphological structures or organellar structures viewed under a microscope. The process of classification using such characteristics seems to be feasible for classifying the main groups of animals and plants. However, it is implausible to use such characteristics to classify microorganisms, and it is troublesome for complex organisms. In the 1960s, the emergence of molecular data as a source to describe evolutionary histories was a significant contribution to the development of phylogenetics (Zuckerka and Pauling, 1965). DNA and protein sequencing, which were introduced in the 1970s and 1950s, respectively, progressively increased the number of homologous sequences that are accessible, and can be used as reference markers for unknown sequences. Small subunit ribosomal RNA sequences were extensively used in phylogenetic analyses, which later led to the discovery of the Archaea as another domain of life (Woese and Fox, 1977).

The increase in the number of genes being analysed added to phylogenetic knowledge led to conflicts between the phylogenies of individual genes. Rapid progress in generating sequence information has led to a new approach, called phylogenomics (a portmanteau of phylogenetics and genomics). Phylogenomics used the phylogenetics approach to make sense of genome sequences. The phylogenomics approach was initially used to improve the prediction of protein functions (Eisen, 1998). Later, the approach was used to construct the evolutionary histories of organisms from genomic data (O'Brien and Stanyon, 1999). No matter how phylogenomics is used, it comprises two essential steps: the identification of homologous genes and the tree inference method.

A phylogenomics approach was introduced to shift the process to predict molecular functions, since a top-hit-search approach was usually used to identify the putative function of a novel sequence. The sequence similarity approach was excessively used for identifying and characterising unknown genes from genes with known functions. An unknown gene that is similar to a characterised gene in a database is assumed to share a similar function, which is not always the case. The similarity of two sequences does not necessarily indicate that these two sequences are homologous, as the sequence similarity approach does not take into account information regarding how the gene and the gene function evolved (Swofford, 1996), nor does it account for the evolutionary relatedness

UNIVERSITI PENDIDIKAN SULTAN IDRIS

between the sequences. The molecular evolution processes that are involved in shaping the histories of organisms, such as mutations, horizontal gene transfers and duplications, could affect the similarity between sequences; hence, to infer specific protein function based only on the aspect of sequence similarity is inadequate.

The general steps involved in predicting a molecular function are: i) select homologs; ii) align homologous sequences; iii) construct a phylogenetic tree; iv) overlay an annotation on the tree topology; v) distinguish homologs into orthologs and paralogs; and, lastly, vi) infer protein function. The process of inferring protein function is definitely not straightforward. Every computational task involved in accommodating each step in this process is far more complex and more prone to error when large and divergent protein families are used in the analysis.

Phylogenomics utilises genome characteristics and structures in the phylogenetic construction (Philippe and Laurent, 1998, Rokas and Holland, 2000). There are two categories of phylogenomics, which are the sequence-based approach and the whole genome features approach (Figure 1.8).

Two alternative approaches for sequence-based phylogenomics are the supermatrix and supertree approaches. There has been a debate regarding how to analyse multiple datasets. Opinions differ regarding how to accommodate this problem (Chippindale and Wiens, 1994), whether by: i) combining all partitions (Barrett et al., 1991), which is known as 'total evidence' (Kluge, 1989); ii) to never combine them (Swofford, 1991) in a separate analysis (Miyamoto and Fitch, 1995) and later combine the resulting topologies using consensus or supertree methods; or iii) conditional combination, known as the supermatrix approach, which only combines congruent datasets (Bull et al., 1993). Among all of the approaches, the supermatrix seems to be preferable (Wolf et al., 2001a, Wolf et al., 2001b, Rokas et al., 2003, Hao and Golding, 2006, Anisimova et al., 2007) to the supertree method (Daubin et al., 2001, Daubin et al., 2002).

Methods that are based on whole-genome features use any genome characteristics that are comparable between organisms to infer phylogenies. The method is divided into three classes of genome features, depending on what they are examined for. For the first two classes: gene content and gene order require the assessment of the homology and

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

orthology of genes. Neither class relies on the multiple sequence alignment procedure, but rely on changes in gene content and gene order within genomes. Although both classes use different genome features, they employ the same standard tree-construction algorithms for inferring phylogeny (Wolf et al., 2002). The final class examines the distribution of a sequence string, based on the distribution of oligonucleotides (Lin and Gerstein, 2000).

Many studies of phylogenomics have shown that there are three domains of life: Eukaryota, Bacteria and Archaea (Hui and Morrison, 1991, Baldauf et al., 2000, Murphy et al., 2001, Wolf et al., 2001b, Bapteste et al., 2002, Daubin et al., 2002, Pride et al., 2003, Rokas et al., 2003). The phylogenomics approach has the ability to reduce sampling errors, as it uses many characteristics of sequenced genomes to estimate the phylogenetic tree. However, systematic errors can still occur in phylogenomics, which is dependent on data quality and tree inference methods. Phylogenomics has provided important perspectives on the molecular processes of evolution between the genomes of organisms. The development of phylogenomics will further increase our knowledge of the evolutionary relationships of organisms.

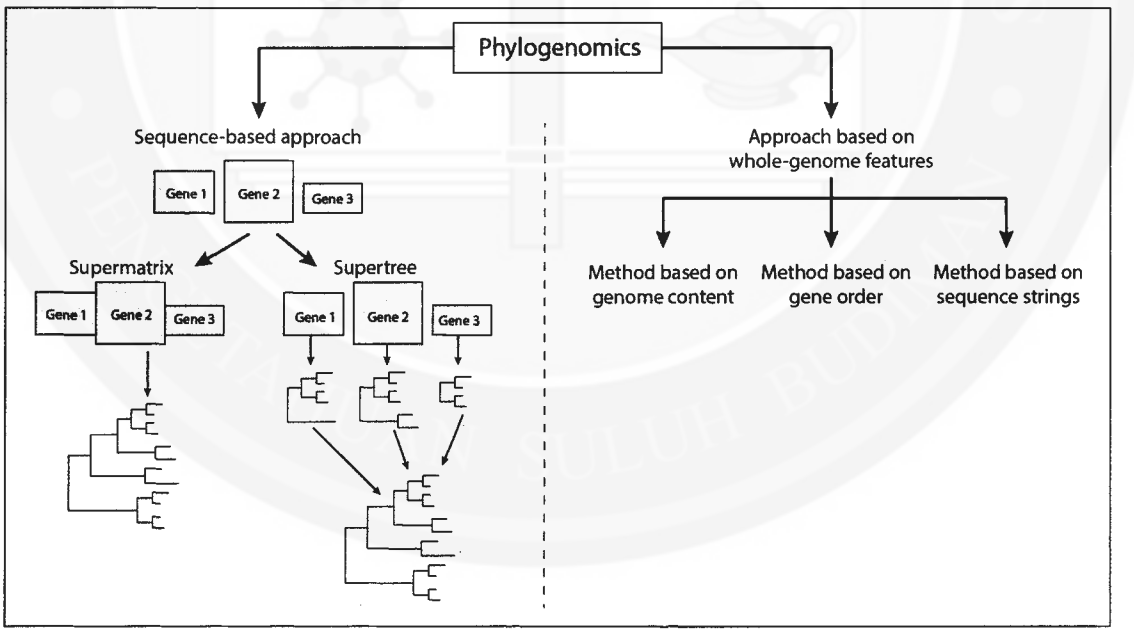


Figure 1.8: Phylogenomics approach. Two categories of phylogenomics approach: sequence-based approach (left) and approach based on the whole-genome features (right). (Adapted from Delsuc et al., 2005).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

1.4 Understanding the evolution of quorum sensing systems in streptococcal species

1.4.1 Problem statement/ motivation of work

Two homologous QS systems, Blp QS and Com QS, have been identified in the *S. pneumoniae* species (de Saizieu et al., 2000, Claverys and Havarstein, 2002, Martin et al., 2006). The regulation and the regulon genes involved in both QS systems have many similar features, e.g., both regulation are cell-density dependent, and the conserved region of the binding TCS of Blp QS is similar to the TCS of Com QS (de Saizieu et al., 2000). Although, it is known that each of the QS system has a specific function and tightly regulated, it is believed that both Blp QS and Com QS have evolved from an ancestor.

The QS regulon genes that are responsible to control competence in *S. mutans* were found to have more similar features to the regulon genes in the Blp QS in *S. pneumoniae* (van der Ploeg, 2005, Martin et al., 2006). This suggests the Blp QS regulon genes are also primarily associated with other functions, i.e., competence.

Investigating the homologous regulon genes of *S. pneumoniae* Com QS and Blp QS in other streptococcal species would help to elucidate the relationship between these two QS systems. This will further help us to answer the interesting questions: What is the ancestor of these QS systems in *S. pneumoniae*? Are these two QS systems independently regulated or dependently regulated? What are the selective forces maintaining these two QS systems in the streptococcal species? What caused the functional divergence of the QS systems in the *Streptococcus* species?

1.4.2 Experimental aims and subprojects

The main objective of this thesis is to understand the evolution of the QS systems in *Streptococcus* species. Specifically, there are some interesting questions as mentioned above that this thesis tried to answer upon addressing the objective of this study. In this study, a computational approach was used to examine the history of these systems, and to find out what are the selective forces that lead to their maintenance and diversification.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

To find out the answers to these questions, my PhD research is divided into four projects:

i. To develop a bioinformatics approach, Search Using Competence Regulon (SUCRE), to find the QS homologs in *Streptococcus* species.

ii. To validate the performance of SUCRE in comparison with other orthology detection methods.

iii. To construct the *Streptococcus* species phylogeny.

iv. To examine the evolution of the QS systems by using the dataset found in the first project and the streptococcal species tree constructed in the third project.

Each project is written as a chapter to represent the amount of work that I have done during my study. The chapters are complementary, but each has a distinctive aims and different methodologies. All studies are unpublished work between Simon Whelan, Daniel Rozen and myself.

Below is an overview of each chapter contents in my PhD thesis:

Chapter 2: Using the competence regulon structure to identify quorum sensing homologs in streptococcal species

Chapter 2 describes the development of SUCRE methodology. SUCRE was used to detect the QS homologs in 86 genomes of 21 *Streptococcus* species. Four steps were involved in detecting these QS homologs: i) homology search, ii) putative gene finding, iii) regulon construction, and iv) phylogenetic analysis. The first three phases were automated using a purposely written Perl script. The QS homologs that were found by SUCRE were later used in Chapter 5 as a dataset to understand the evolution of the QS systems in *Streptococcus* species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

Chapter 3: Assessing SUCRE performance in comparison with other orthology detection methods

Chapter 3 describes the validation of SUCRE approach in comparison with other established orthology detection methods. The study examines the performance and outlines the differences and similarity between SUCRE and orthology detections methods.

Chapter 4: *Streptococcus* species phylogeny

Chapter 4 describes the process of building the streptococcal species phylogeny. In order to investigate the origin of the QS systems in the streptococcal species, a reliable *Streptococcus* species phylogeny is required. The chapter explains and discusses the approach used to construct a *Streptococcus* species phylogeny from 31 genes, which encode the ribosomal protein and tRNA synthetase. The streptococcal phylogeny was later used in Chapter 5 to study the evolution of QS systems in Streptococci.

Chapter 5: The evolution of streptococcal quorum sensing systems

Chapter 5 describes the function and the evolution of QS systems in *Streptococcus* species. The chapter gathers the results from the previous chapters: the QS homologous genes found in Chapter 2 and the *Streptococcus* phylogeny constructed in Chapter 4. Here, an approach to assign the orthology and paralogy was constructed to examine the evolutionary process involved in shaping the history of QS systems. This chapter also discusses the mechanism and the factors contribute to the evolution of QS systems in streptococcal species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

1.5 Alternative format

This thesis is presented in the alternative format to give a better representation of the works that have been done throughout my PhD. The last results chapter is intended for journal submission.

Chapter 5: The evolution of streptococcal quorum sensing systems

Authors: Raja Farhana Raja Khairuddin, Daniel Rozen and Simon Whelan

Intended journal: Genome Biology and Evolution

Author contribution: This manuscript is representative of the work of which I have majorly contributed. My two supervisors, Dr. Simon Whelan provided advice and guidance mainly on the computational methods and the evolutionary approaches, while Dr. Daniel Rozen provided advice and guidance through in understanding the quorum sensing systems in *Streptococcus* species. As the first author of this paper, I was fully responsible for writing the text of the manuscript. The first draft was produced by myself, which was later reviewed by my supervisors. Their input comments were then synthesised by me into the final version as presented in the thesis.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

CHAPTER 2

Using the competence regulon structure to identify quorum sensing homologs in *Streptococcus* species

Abstract

One of the challenges when identifying quorum sensing (QS) homologs is the lack of reliably annotated genomes from which to draw data. Only a limited number of annotated *Streptococcus* genomes are available, and the reliable extraction of the QS genes is difficult because many of these genomes have poor annotations of QS genes. In this chapter, I develop the computational, SUCRE pipeline to reliably identify QS homologs. SUCRE uses gene properties, such as identity scores, to process BLAST hits output through an iterative search. A refinement procedure to obtain putative homologs was performed by assessing the quality of the matches to reduce redundancies in the BLAST hits. Putative homologs were further analysed by filtering false positives using evolutionary information to improve our confidence in the homologous QS genes that were found. The reliable QS genes that were detected using SUCRE will be used to study the evolution of the QS systems in streptococcal species in a subsequent chapter.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

2.1 Introduction

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

The first important step when studying the origin of quorum sensing (QS) systems in *Streptococcus* species is to obtain a reliable set of homologous QS genes from these species. Homologous genes are derived from a common ancestor through duplication or speciation events, or both (Fitch, 2000). Homologous genes can also be acquired through horizontal gene transfers from other organisms (Ochman, 2001). Finding and understanding the history of homologous genes may help to predict the functions and structures of unknown homologs.

The two most popular approaches for detecting homologous genes are pairwise methods that employ a dynamic or heuristic approach. In the dynamic approach, a global alignment is used to identify homologous genes (Gelfand et al., 1996, Birney et al., 2004), which provides relatively sensitive and accurate annotation. The drawback of this approach is that it is computationally intensive and slow.

Heuristics approaches, such as the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) or FASTA, provide an approximation of the optimal solution in the homology search. This approach improves the homology detection by more than 50% versus the dynamic approach (Altschul et al., 1990) and is preferred for a large-scale searching operation (Gish and States, 1993). Position-specific iterated BLAST (psi-BLAST) method (Altschul et al., 1997) improves on the original BLAST algorithm by using an iterative profile search derived from an alignment of a set of proteins, which gives more sensitivity when detecting distant homologs. The disadvantage of psi-BLAST is that the approach is dependent on how the profile is constructed.

An alternative approach to homology searches for gene prediction is to use gene properties, such as the exon, intron or gene compositions. The programs that use gene properties to find genes are GENSCAN (Burge and Karlin, 1997, Burge and Karlin, 1998), GeneMark (Borodovsky et al., 1994, Borodovsky et al., 1995) and GRAIL (Uberbacher et al., 1996). All of these programs claim to be accurate in predicting a gene based on gene composition, but may not perform well if multiple homologous genes appear in the genomes.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Both of these approaches, homology searches and gene predictions, are highly dependent on sequence quality, e.g., the GeneMark program (Borodovsky et al., 1995) that is used to annotate bacterial genomes by predicting the start codons of open reading frames (ORFs). Therefore, well-annotated genome sequences are required to identify homologous genes and to obtain reliable gene predictions. Typically, a gene is annotated through either de novo identification or homology matching, although both of these methods may lead to incorrect gene predictions. Errors can be categorised into two forms: i) false positives, in which a fraction of a sequence is incorrectly inferred to be part of a protein coding sequence; and ii) false negatives, in which a faction of a protein coding sequence is inferred to be non-coding. Any individual gene prediction may suffer both types of errors along its length.

Many thousands of prokaryote genomes are being sequenced, but at the time of this project, there were only 86 fully sequenced *Streptococcus* genomes publically accessible in the NCBI Genome database (June 2011). Only 60 of these genomes are well-annotated, and they cover a total of 21 out of the 91 *Streptococcus* species. Additionally, the promoter and terminator boundaries of the sequence motifs in *Streptococcus* species are yet to be well characterised (Whatmore et al., 1999).

Therefore, rather than predicting genes, other methods further analyse the genomes to predict a set of genes as a functional system. These approaches differentiate the homologs into orthologs and paralogs based on whether the homologous genes were derived from their last common ancestor through speciation or duplication events. I divide these approaches into three categories: i) graphical-clustering approaches, e.g., Cluster of Orthologs Groups (COG) (Tatusov et al., 1997, Tatusov et al., 2000), OrthoMCL (Li et al., 2003); ii) phylogenetic tree and graphical clustering, e.g., Ensemble Compara (Flicek et al., 2008); and iii) those that use, the sequence composition e.g., GenomeHistory (Conant and Wagner, 2002). See the next chapter (Chapter 3), where each category is discussed in more detail.

Based on the limited annotations available for *Streptococcus* genomes and the lack of out-of-the-box tools, I have developed the SUCRE (Search Using Competence Regulon) pipeline to process BLAST hits and identify QS homologs. SUCRE uses

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

extrinsic gene information, such as gene order and regulon structure, to improve sensitivity when detecting QS homologs.

The competence (Com) QS regulon in *Streptococcus pneumoniae* is one of the best-characterised QS systems in Gram-positive bacteria. The genes involved in *S. pneumoniae* Com QS are well-annotated and have been experimentally proven to regulate competence (Tomasz and Hotchkiss, 1964, Tomasz, 1965, Claverys and Havarstein, 2002). The Com QS genes involve an ABC transporter, encoded by *comA* and *comB*, which are always adjacent to each other (Hui and Morrison, 1991, Hui et al., 1995); peptide pheromones (*comC*); a two-component system encoded by *comD* and *comE* (Pestova et al., 1996, Cheng et al., 1997), which are also adjacent; as well as a sigma factor (*comX*). Bacteriocin-like peptide (Blp) QS systems are known to be homologous to Com QS systems. All of the QS Blp genes share similar features with the Com QS genes. Refer to the introduction section 1.1 for detailed descriptions of the regulation of Com QS and Blp QS systems.

QS regulon structures are broadly conserved across *Streptococcus* species. This study uses the structure of the regulon as a basis for developing the SUCRE methodology that assists in the detection of homologous QS genes across all streptococcal species. The identification of Com QS homologs is the first important step to explore the origin of Com QS systems in *Streptococcus* species and their relationship to the Blp QS system. The SUCRE methodology consists of four distinct phases: i) homology search; ii) putative gene finding; iii) regulon construction; and iv) evolutionary analysis, which help to find these homologs.

2.2 Methods and Input

2.2.1 Search Using Competence Regulon (SUCRE) methodology

SUCRE is an iterative pipeline inspired by psi-BLAST that seeks to identify QS genes in streptococcal species by taking a pre-defined dataset of query proteins and using it to search for a set of query genomes. Each iteration takes the current set of query proteins and searches the database for homologs of these proteins using BLAST. The

UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

results of this search are then used as the query sequences in the next iteration. Using this approach, SUCRE continues to expand the number of identified QS genes in the query genomes and stops when no new genes are identified.

An overview of this study is illustrated in Figure 2.1. The set of query genes are analysed in four phases: i) homology search; ii) putative gene finding; iii) regulon construction; and iv) evolutionary analysis. The first three phases constitute the iterative part of the pipeline and were automated using a purpose-written Perl script and BioPerl module (Stajich et al., 2002). The final phase takes all of the putative QS genes identified by the iterative search, and examines their evolutionary relationships to ensure true homology and to identify cases of orthology and paralogy.

I still consider the SUCRE Perl script to be experimental, as it has only been tested on a limited number of QS genes in streptococcal species. Further rigorous testing needs to be performed before it can be applied to other QS systems in different species.

The following sections discuss each phase in detail:

a. Homology search

To identify the homologs of the query proteins across *Streptococcus* species, BLAST (Altschul et al., 1990) was used to perform a sequence similarity comparison between the protein query and subject genomes sequences. Because the query sequences of the genomes are rather distantly related, I performed this comparison at the protein level using TBLASTN (Altschul et al., 1997), which translates all six possible reading frames in the genomes and compares them to the query sequences. This type of search helps to identify similar or homologous protein sequences from the unannotated nucleotide subject genomes. To allow short homologous sequence to be identified, the maximum E-value was set to 1e-5, with a minimum percentage of identity of 30%. All the BLAST hits were ordered according to their synteny in the genomes. This step is illustrated in Figure 2.1.

b. Putative gene finding

The BLAST hits identified in step **a.** were processed further to determine Quality Putative Genes (QPGs) in each subject genome. QPGs refer to two overlapping hits of the same query that are next to each other and less than 200 base pairs (bp) apart.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

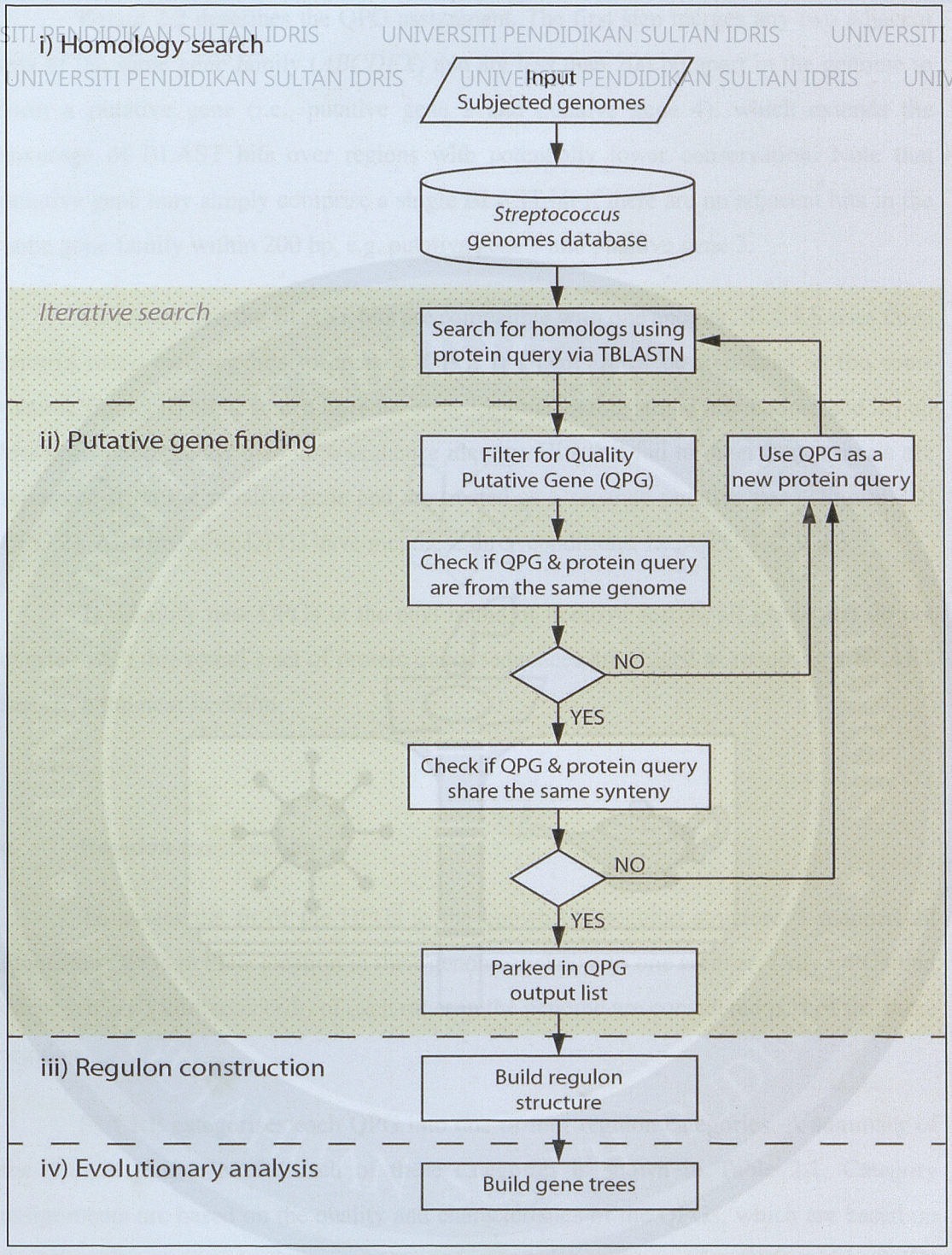


Figure 2.1: An overview of SUCRE methodology. Four phases: i) homology search; ii) putative gene finding; iii) regulon construction; and iv) evolutionary analysis used to identify homologous QS genes in streptococcal species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

Figure 2.2 describes the QPG assignment. The first step merges any two adjacent hits of the same gene family (*ABCDEX*) that are less than 200 bp apart in the genome to form a putative gene (i.e., putative gene 2 and putative gene 4), which extends the coverage of BLAST hits over regions with potentially lower conservation. Note that putative gene may simply comprise a single BLAST hit if there are no adjacent hits in the same gene family within 200 bp, e.g. putative gene 1 and putative gene 3.

To ensure that the BLAST hits are compatible with one another, SUCRE uses three criteria when merging hits to create a QPG: i) all hits must be translated in the same reading frame; ii) all hits must have the same lagging/leading strand orientation, and iii) all hits must have greater than 40% sequence identity. Hits that fail to meet these criteria are separated from the putative gene and are treated as a separate putative gene (Figure 2.2: QPG 2). All remaining QPGs have met these three conditions, i.e., QPG 1, 2 and 3.

To identify new QPGs in the next round of iterative search, all genes that do not overlap with this round's set of protein query sequences are tagged as requiring a BLAST search in the next iteration.

c. Regulon construction

To assess the fit of the QPGs to the known QS regulon structure, I mapped all identified QPGs to their location in their genomes relative to one another. Only QPGs that occur within 2 kilobases (Kb) of each other in the genome are considered part of the same regulon.

SUCRE categorises each QPG into one of four regulon categories. A summary of the protein properties for each of these categories is shown in Table 2.1. Category assignments are based on the quality and characteristics of the QPGs, which are based on their length coverage to their query sequence and the presence of a QPG within 2 Kb. SUCRE considers a QPG that has a coverage of >80% of bases and an appropriate neighbour within 2 Kb upstream or downstream in the genome as a 'Prime' gene. QPGs with >80% coverage, but no near neighbour, are classified as 'Lone' homologs. A QPG

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

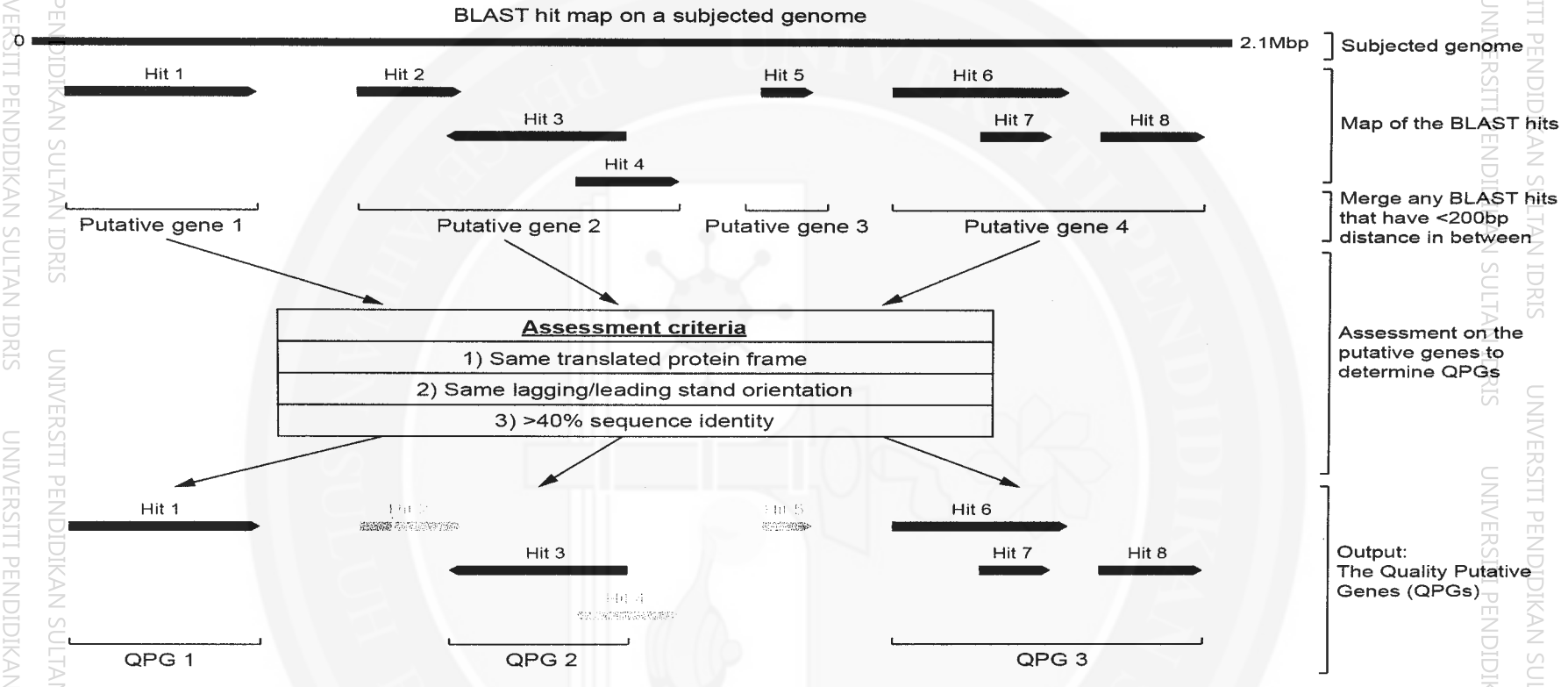


Figure 2.2: The process of finding Quality Putative Genes (QPGs) in SUCRE. All the BLAST hits were mapped to their location on the genome and numbered from 1 to 8 accordingly. Adjacent hits that are closer than 200 bp were merged as a putative gene (i.e., putative gene 1 (hit 1), putative gene 2 (hit 2, 3, 4), putative gene 3 (hit 5), putative gene 4 (hit 6, 7, 8)). Each putative gene will be assessed on three criteria: i) the same translated protein frame, ii) the same lagging/leading strand orientation, and iii) more than 40% of sequence identity. Only the QPGs that met all criteria (QPG 1 and QPG2) will be used to construct the regulon structure for the *Streptococcus* species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

with poor coverage (<80%), but which has an adjacent QPG within 2 Kb with >80% coverage is identified as an 'Unusual' gene. In this category, the sequence length of the QPG was expanded to the maximum length of its query sequence. This process will shorten the location distance between the QPG and its neighbour. Meanwhile, a poor coverage QPG with only poor coverage neighbours is considered as 'Wobbly'. The SUCRE methodology ignores QPGs that have poor coverage and no neighbour within 2 Kb because these QPGs may be inferred by chance and are likely to be false positives. Note that the *Streptococcus* genomes are in a circular form, so SUCRE accounted for the possibility that the first and the last QPGs in a genome belong to the same regulon.

Table 2.1: The regulon categories: Prime, Lone, Unusual and Wobbly. Each category uses three criteria to assess QPGs: i) the QPG coverage; ii) the presence of the adjacent QPG(s); and iii) the coverage of the present QPG(s).

Regulon category	Criteria		
	QPG Coverage (>80%)	Any right or left adjacent QPG(s) (<2 Kb)	Any right or left adjacent QPG(s) coverage (>80%)
Prime	true	true	true
	true	true	false
Lone	true	false	true
	true	false	false
Unusual	false	true	true
Wobbly	false	true	false

d. Evolutionary analysis

To enhance the confidence in homologous QS genes that were found in the previous phase, I further filtered any spurious homologs using evolution-related information. First, I obtained the amino acid sequences of the QPGs and used EMBOSS Transeq (Williams, 2002) to translate each nucleotide sequences in each of the six possible

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

reading frames, (three forward and three backward frames), mimicking the search done by TBLASTN. I aligned the translated amino acid sequences from all six frames to the query amino acid sequence using the Smith-Waterman pairwise sequence alignment via Water tool Water (Rice et al., 2000) from EMBOSS. The amino acid sequence with the highest alignment score was taken to be the best translated frame and it was used in a subsequent analysis of evolution relationships. A multiple sequence alignment of the amino acid sequences of all members of each gene family, *ABCDEX*, was computed using Muscle (Edgar, 2004).

A maximum likelihood (ML) gene tree and 100 bootstrap trees were estimated from each gene family alignment under the WAG+Γ substitution model (Whelan and Goldman, 2001) using the RAxML program (Stamatakis, 2006b). RAxML was chosen because it is one of the fastest ML tree-search programs, and studies suggest that it produces good estimates of tree topologies (Money and Whelan, 2012), especially for larger datasets. I performed ML-tree search using the 'lazy' subtree pruning and re-grafting (SPR) heuristic, which only optimises the adjacent branches of the inserted subtree (Stamatakis and Alachiotis, 2010).

Knowing that the ABC transporter (*AB* genes) and TCS (*DE* genes) regulon structures are conserved, it was expected that a similar total number of homologs should be found for each gene within a regulon. For example, the total number of *A* homologs found should be comparable to the total number of *B* homologs found. A QPG gene, in a gene family with outnumbered QPGs, is removed if it is found to be a 'lone' gene and clustered distantly from other QPGs that are paired with other gene families in a regulon. This step provides a rigorous procedure to eliminate any distantly related sequences.

2.2.2 Benchmarking dataset

To benchmark the performance of SUCRE, the 12 QS protein sequences from *S. pneumoniae* shown in Table 2.2 were used as queries to search for homologs in 10 other *Streptococcus* species. The *S. pneumoniae* query sequences are well characterised and studied and consist of the five early *ABCDE* Com QS genes and the homologous *ABCDE* Blp QS genes, as well as two *comX* genes that encode a sigma factor for competence

regulation. The amino acid sequences of these *S. pneumoniae* genes were retrieved from the NCBI Reference sequence database (RefSeq).

I searched for homologous genes in the genomes of the 10 different *Streptococcus* species shown in Table 2.2 b) which were downloaded from the Genome NCBI database (October 2010) using a scripted BioPerl module (Stajich et al., 2002). The majority of these species were chosen because there is experimental data indicating the presence of the *com* and *blp* genes (Martin et al., 2006, Johnsborg et al., 2007), which allowed us to assess the sensitivity and specificity of the SUCRE pipeline.

Table 2.2: The list of the protein queries and subject genomes used in the analysis. Table (a) shows a list of 12 Com QS and Blp QS protein queries in *S. pneumoniae*, with the NCBI accession number. Table (b) shows a list of 10 *Streptococcus* genomes of 10 *Streptococcus* species, with the NCBI accession number.

a)

Protein query	Accession number	Protein query	Accession number
ComA	YP_002036763	BlpA	YP_002037195
ComB	YP_002036764	BlpB	YP_002037194
ComC	YP_002038833	BlpC	YP_002037193
ComD	YP_002038832	BlpH	YP_002037192
ComE	YP_002038831	BlpR	YP_002037191
ComX1	YP_002036735	ComX2	YP_002038589

b)

Subject genome	Strain	Accession number
S. agalactiae	A909	NC_007432
S. equi	MGCS105665	NC_011134
S. gallolyticus	UCN34	NC_013798
S. gordonii	CHALISCH1	NC_009785
S. mutans	UA159	NC_004350
S. pyogenes	MGA8232	NC_007297
S. sanguinis	SK36	NC_009009
S. suis	pBM407	NC_012926
S. thermophilus	CNRZ1066	NC_006449
S. uberis	0140J	NC_012004

UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

2.2.3 The 86 genomes dataset

To exhaustively search for QS genes in streptococcal species. I obtained the 86 genomes of 21 streptococcal species from the NCBI Genome database (June 2011). These 21 species were considered due to the SUCRE approach of using competence regulon structure as a parameter to identify QS genes; thus, only streptococcal species with available genomes sequences were considered. Of the 21 species, 10 species have multiple genome sequences, while 11 species each had a single genome sequence (see Appendix B for details).

To identify QS homologs in these genomes, we took 44 amino acid sequences from the NCBI RefSeq database, which were used as the initial queries in the SUCRE pipeline. A set of *ABCDEX* homologs was chosen from the 44 query proteins, which are from six *Streptococcus* species: *S. pneumoniae, S. mutans*, *S. thermophilus*, *S. equi*, *S. gallolyticus* and *S. suis*. These six species were chosen as being representative of the six different phylogenetic groups previously defined by Kawamura et al. (1995), and their distribution across the streptococcal species tree should help to identify low sequence similarities of QS genes in the SUCRE pipeline.

A pan-genome was used to represent the presence and to mark the synteny of the QS homologs identified in each of the 10 species with multiple genome sequences. The remaining 11 *Streptococcus* species each has a single genome sequence available to be used in this study. The union set of identified *ABCDEX* homologs in the genome of a species characterises the homologous QS genes present in that species.

2.3 Results

2.3.1 Identification of Quality Putative Genes

In the first phase of SUCRE, I identified homologs of the query sequences in the subject genomes using TBLASTN. In the sample benchmarking dataset, 12 QS protein queries representing *ABCDEX* genes in the Com QS and Blp QS from *S. pneumoniae* were used to search the genomes of the 10 *Streptococcus* species shown in Table 2.2. Table 2.3 shows the distribution of 84,081 homologs found during BLAST searches in all iterations

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

of SUCRE. The majority 82,632/84,076 (98%), of these were from the *A* gene family (Table 2.3). This high number may be because the ABC transporter gene family is one of the largest and most ancient protein families, and is widely found across bacteria (Tatusov et al., 1996, Tomii and Kanehisa, 1998).

In contrast, only two *C* homologs were found during the homology search. The short length of the peptide pheromones encoded by the *C* gene family and their high sequence variation are required to maintain a wide range of signalling specificities of the peptide pheromones, (Whatmore et al., 1999, Allan et al., 2007, Cornejo et al., 2010), which may explain the difficulty in identifying *C* homologs. Allocating a short peptide in a genome using BLAST is a tricky process, as the approach is likely to discard the match between the peptide sequence and the genome sequence. The high variations in the peptide sequences suggest that this peptide is encoded by a fast-evolving gene; thus, the detection of peptide homologs through sequence similarity comparisons may not be feasible.

These BLAST hits were then processed by SUCRE to produce QPGs based on the three criteria shown in Figure 2.2. The procedure in constructing QPGs helps to eliminate any redundant hits obtained from the iteration process in the homology search. This step constructed 692 QPGs, which comprised 623 *A* genes, 7 *B* genes, 3 *C* genes, 22 *D* genes, 20 *E* genes and 17 *X* genes from the 84,081 BLAST hits.

Table 2.3: The BLAST hits in each iteration for each ABCDEX gene.

Total hits of each iteration	Iteration	*A*	*B*	*C*	*D*	*E*	*X*
1635	0	1478	30	2	47	44	34
9464	1	8407	105	1	283	405	263
3337	2	3155	0	0	164	18	0
15464	3	15416	0	0	48	0	0
5708	4	5708	0	0	0	0	0
19195	5	19195	0	0	0	0	0
25145	6	25145	0	0	0	0	0
4128	7	4128	0	0	0	0	0
	Total hits	82632	135	3	542	467	297

2.3.2 Classification of QPGs based on the competence regulon structure

The coverage of the QPGs and the presence of adjacent QPGs were used to categorise QPGs into the regulon categories: Prime, Unusual, Wobbly and Lone (see Table 2.1). Out of 692 QPGs constructed in the last phase, only 180 QPGs met the criteria of one of the categories, while the rest were rejected from being further analysed. The majority of the QPGs were categorised in the Prime category, followed by Lone, Unusual, and Wobbly category (Figure 2.3). The filtration with the competence regulon structure helps to refine the collection of QPGs of QS homologs found in *Streptococcus* species.

It is known that *AB* and *DE* genes tend to be conserved in pairs as part of a regulon structure, suggesting that the number of *A* and *B* QPGs should be broadly similar. Figure 2.3 shows there was a large difference between the number of *A* QPGs and *B* QPGs, suggesting there were false positives in the *A* QPGs results. The *A* QPGs outnumbered *B* QPGs as shown in Figure 2.3. The total number of *D* QPGs was slightly higher than the number of *E* QPGs.

The genomic locations of all of the QPGs were analysed using an in-house Perl script and carried forward to the evolutionary analysis to help reduce the number of false positives.

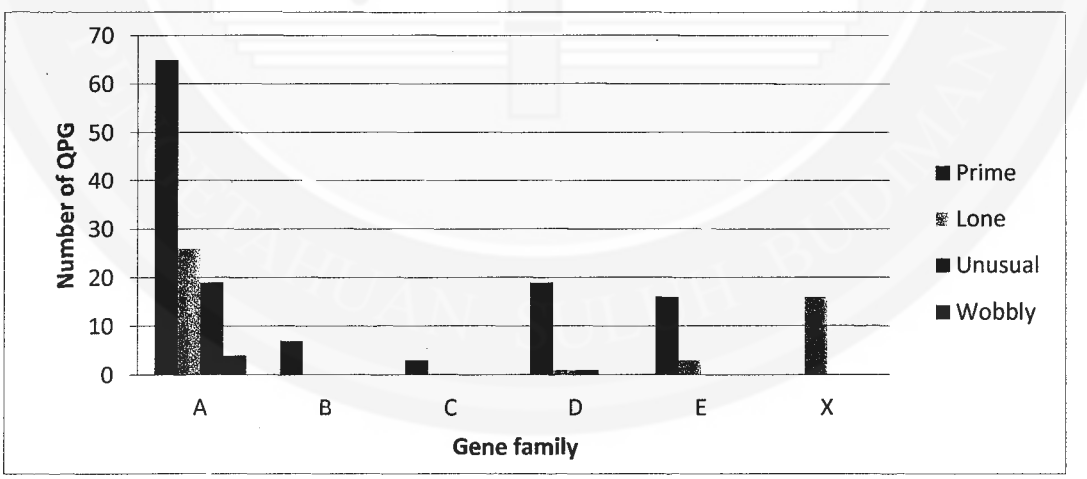


Figure 2.3: The quality putative genes in Streptococcus species categorised into Prime, Lone, Unusual, and Wobbly categories.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

2.3.3 Identification of QS homologs

To reduce the number of false positives produced by the SUCRE pipeline, the genomic mapping of the regulon structures was combined together with the trees estimated from the gene families. Figure 2.4 shows the tree of QS *A* genes, with a clade of 11 genes highlighted in red. This clade has 100% bootstrap support, and the majority of its members are closely linked to the *B* genes. This linkage is demonstrated in Figure 2.5, where lines between the subclade *A* gene tree and *B* gene tree indicate an AB regulon pairing found by SUCRE.

The *A* homologs in *S. pyogenes*, *S. thermophilus* and *S. uberis* were found to be paired with *D* and *E* homologous genes in a regulon, while *B* homologs were absent. These *A* homologs are the unlinked genes observed in Figure 2.5: *S. pyogenes* (pyog7), *S. thermophilus* (ther4) and *S. uberis* (uber5).

The final, homologous QS genes were mapped relative to one another in the genome across 10 *Streptococcus* species (Figure 2.6). The *A* and *B* homologs were observed to be adjacent to each other and to share the same transcription direction across all 10 *Streptococcus* species. The same relationship was observed between the *D* and *E* homologs.

In some species, homologous QS genes appeared in multiple copies. An extra duplicate copy(ies) of the *D* gene was found adjacent to the *DE* regulon genes in *S. uberis*, *S. pyogenes* and *S. equi*. Apart from *S. pneumoniae*, which is known to have two homologous QS systems, Com QS and Blp QS, a pair of TCS (*DE*) is present in several streptococcal genomes, *S. gordonii*, *S. uberis*, *S. pyogenes*, *S. equi* and *S. gallolyticus*. In contrast, most of the *X* homologs, and a few *E* and *A* genes, appear to be lone genes in the genomes.

C homologs were identified only in *S. pneumoniae* and *S. equi*. The complete *ABCDE* homologs in *S. equi* are recognised to have a similar regulon structure to that of *S. pneumoniae*. However, the *ABDE* homologs in *S. equi* are found in a regulon without a *C* homolog. The *C* homolog is located in another regulon with a *B* homolog. Two *Streptococcus* species, *S. thermophilus* and *S. uberis*, have a regulon that is composed of *ADE* homologs, without the *B* and *C* homologs.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

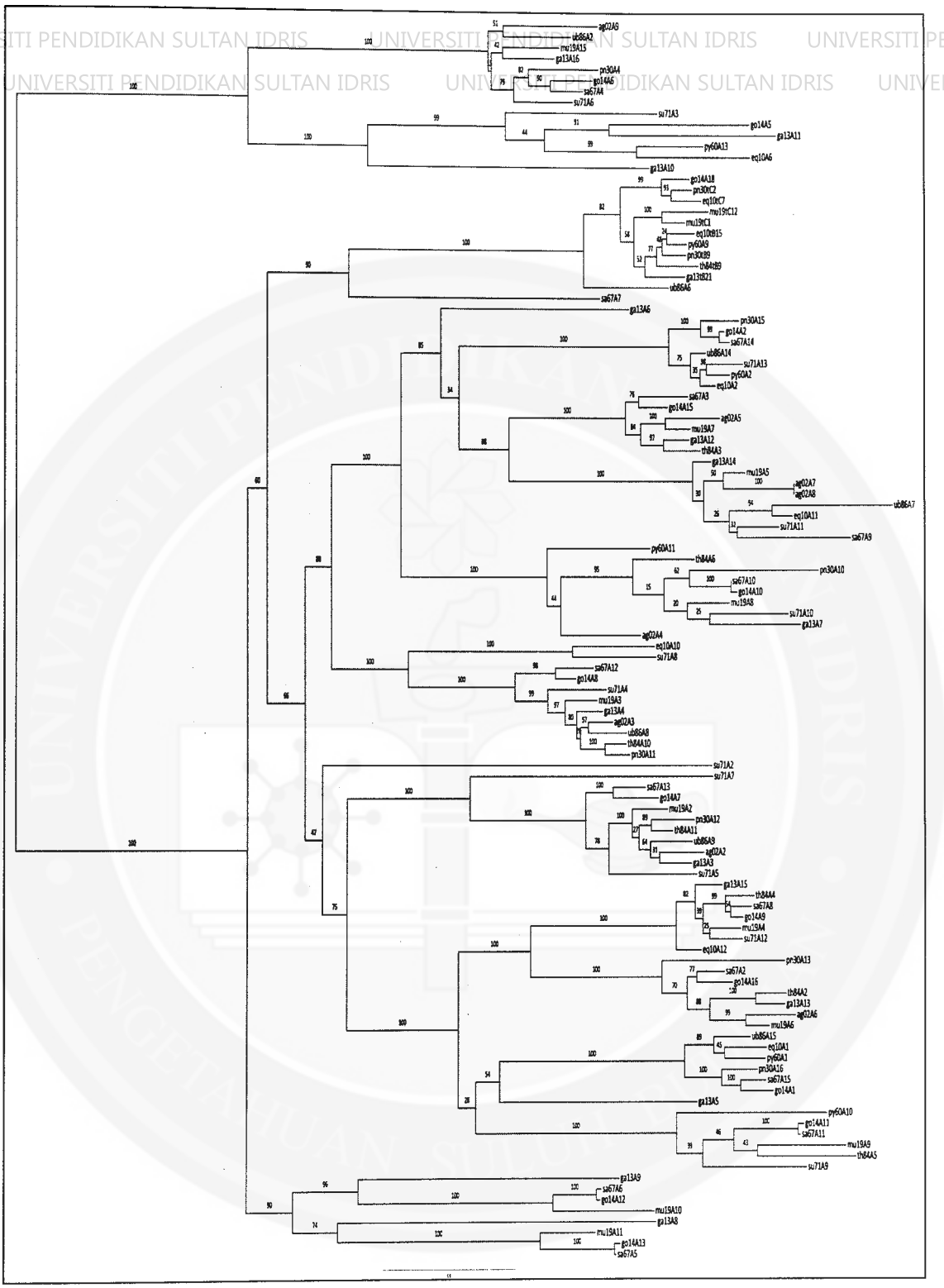


Figure 2.4: The ABC exporter (*A* gene) maximum likelihood gene tree of 10 *Streptococcus* species. Phylogenetic analysis in the last step in SUCRE helps to eliminate distant homologs. Only the taxa within the red clade will be considered as the final output of the ABC exporter homologs.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

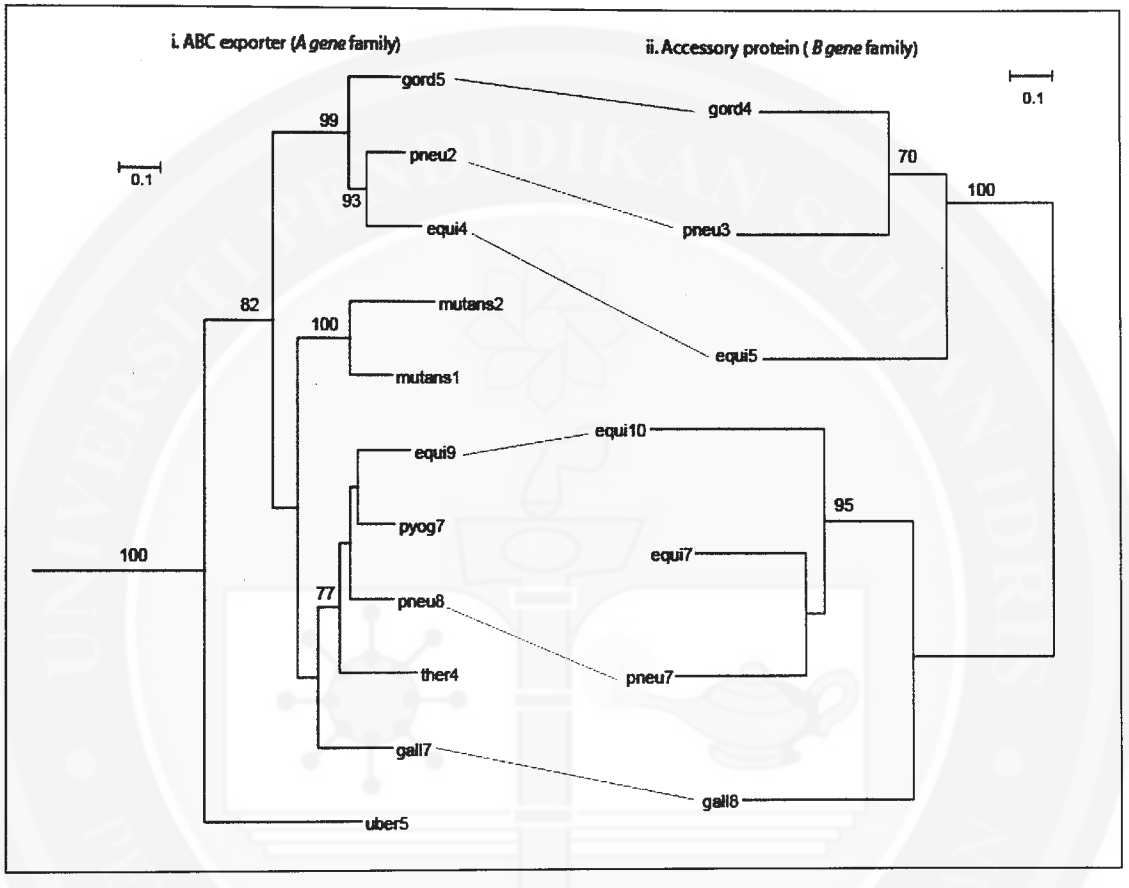


Figure 2.5: The subclade of the ABC exporter gene tree: A genes and the unrooted gene tree of the accessory protein family: B genes. Branches with a bootstrap support value greater than 50% are shown. The lines between the trees indicate the AB homologs pairing in a regulon. Each taxon refers to the species name, followed by the gene order in the genome, which can be used as a reference for the genome mapping of the regulon structure in Figure 2.6.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

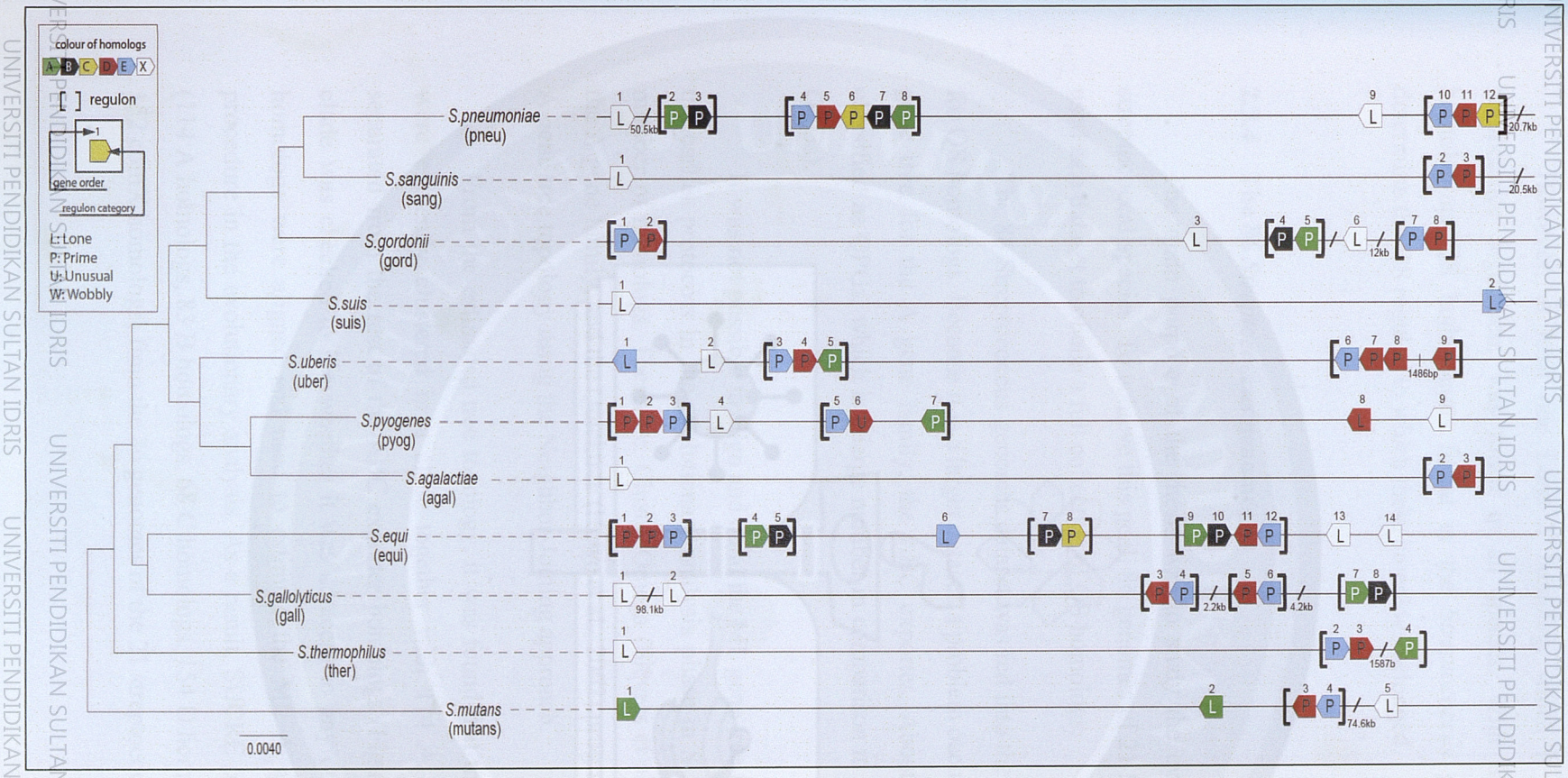


Figure 2.6: Overview of QS homologs mapped across Streptococcus species. All of the homologs were mapped based on their gene order (number on top of the arrow) in the genomes of each species. Homologous genes are coloured based on the gene family: A (green), B (black), C (yellow), D (red), E (blue) and X (white), with the regulon category tagged (Prime-P, Lone-L, Unusual-U and Wobbly-W). Homologs within the black bracket are within 2 Kb of each other. The 16sRNA Neighbor-Joining *Streptococcus* species tree (on the left) was obtained from the Ribosomal Database Project (Cole et al., 2005) for a reference to the evolutionary relationships between the species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

However, at least one Lone *X* homolog, which is an important element in competence regulation, was found in all species. With the incomplete set of the homologous QS homologous found in the *Streptococcus* genomes, it was hard to determine the QS regulon to which these homologs belonged.

2.3.4 The 86 *Streptococcus* genomes

The main purpose of the thesis was to study the evolution of QS systems in streptococcal species. To achieve this goal, 86 genomes of 21 Streptococcus species were used as a dataset to obtain a full collection of QS homologs.

The 86 Streptococcus genomes dataset was fed into the SUCRE pipeline to search for QS homologs. Because it is a larger dataset, a problem occurred when building the ML gene tree for the A gene family, the ABC exporter. About 6,269 A homologs were classified as QPGs, which is a large number in comparison to the paired B homologs (83 QPGs).

It was not possible to align and build the ML gene tree for the A gene family using the earlier protocols in the evolutionary analysis phase in SUCRE because of the large number of homologs. Therefore, Clustal Omega (Sievers et al., 2011) was used to align the A gene family, and the Quicktree (Howe et al., 2002) program was used to estimate the A gene tree topology using the Neighbor-Joining approach.

From the estimated tree topology, it was found that the known A query proteins were evidently clustered in a clade together with 114 A homologs, and they were separated from the rest of the ABC exporter homologs (Figure 2.7). Each taxon in this clade was checked to see whether it was adjacent to any other homologs. The 114 A homologs were aligned and used to estimate the ML A gene tree using the earlier procedure in the evolutionary analysis. As a result, SUCRE produced 708 QS homologs (114 A homologs, 83 B homologs, 68 C homologs, 154 D homologs, 137 E homologs and 152 comX homologs) from the 86 genomes of the 21 streptococcal species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

A pan-genome was used to illustrate the presence of QS homologs in a species with multiple genome sequences. The collection of the QS homologs in these species includes the homologs shared by all (strains) genomes of a species and the homologs that are present in some genomes of that species. Out of the 86 genome sequences, 75 belong to 10 different species. Because there was not much variation in the homologs that are uniquely present in the genomes of a species, it was adequate to determine the presence and the absence of these homologs using a pan-genome (Appendix E). An overview of the presence and absence of the representative *ABCDEX* homologs in each 21 *Streptococcus* species is shown in Figure 2.8.

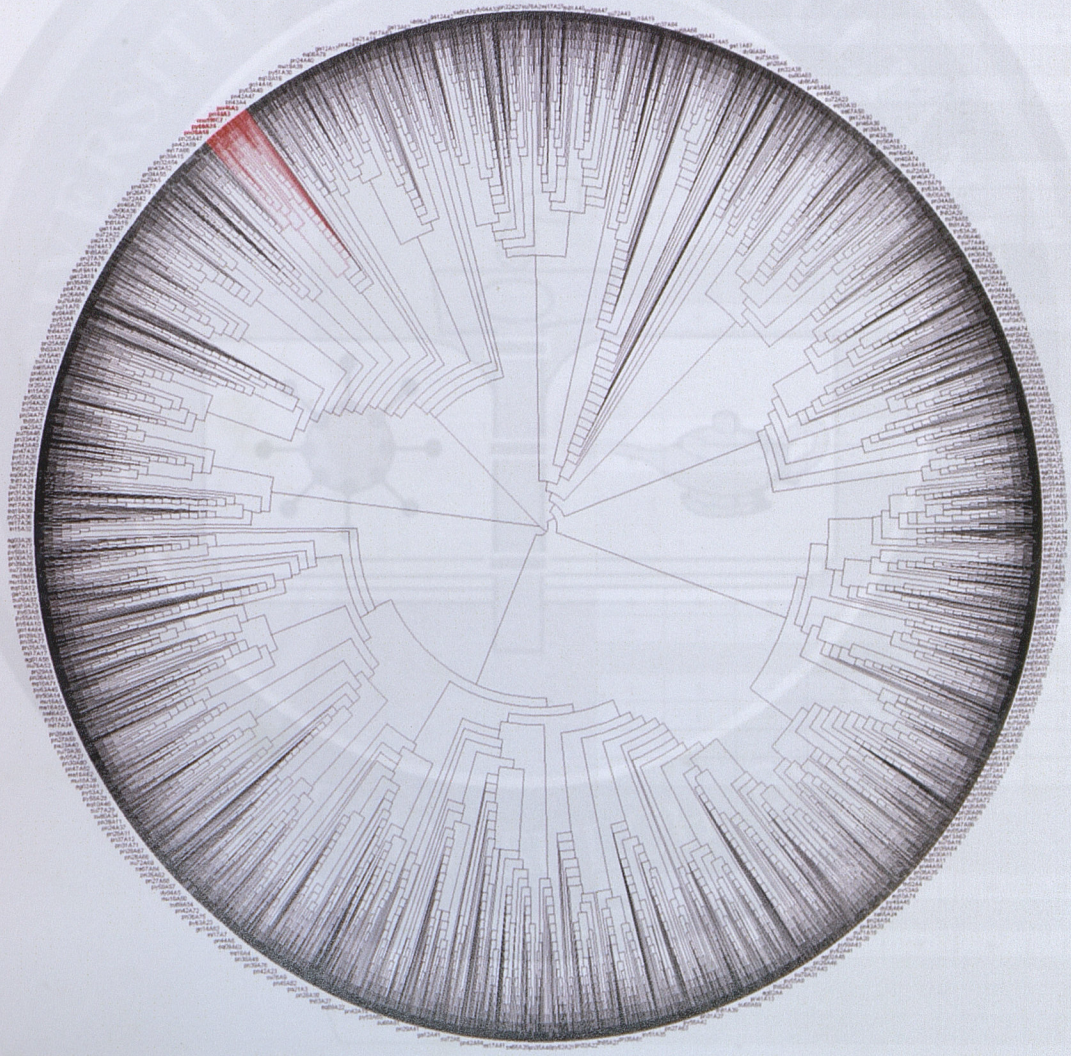
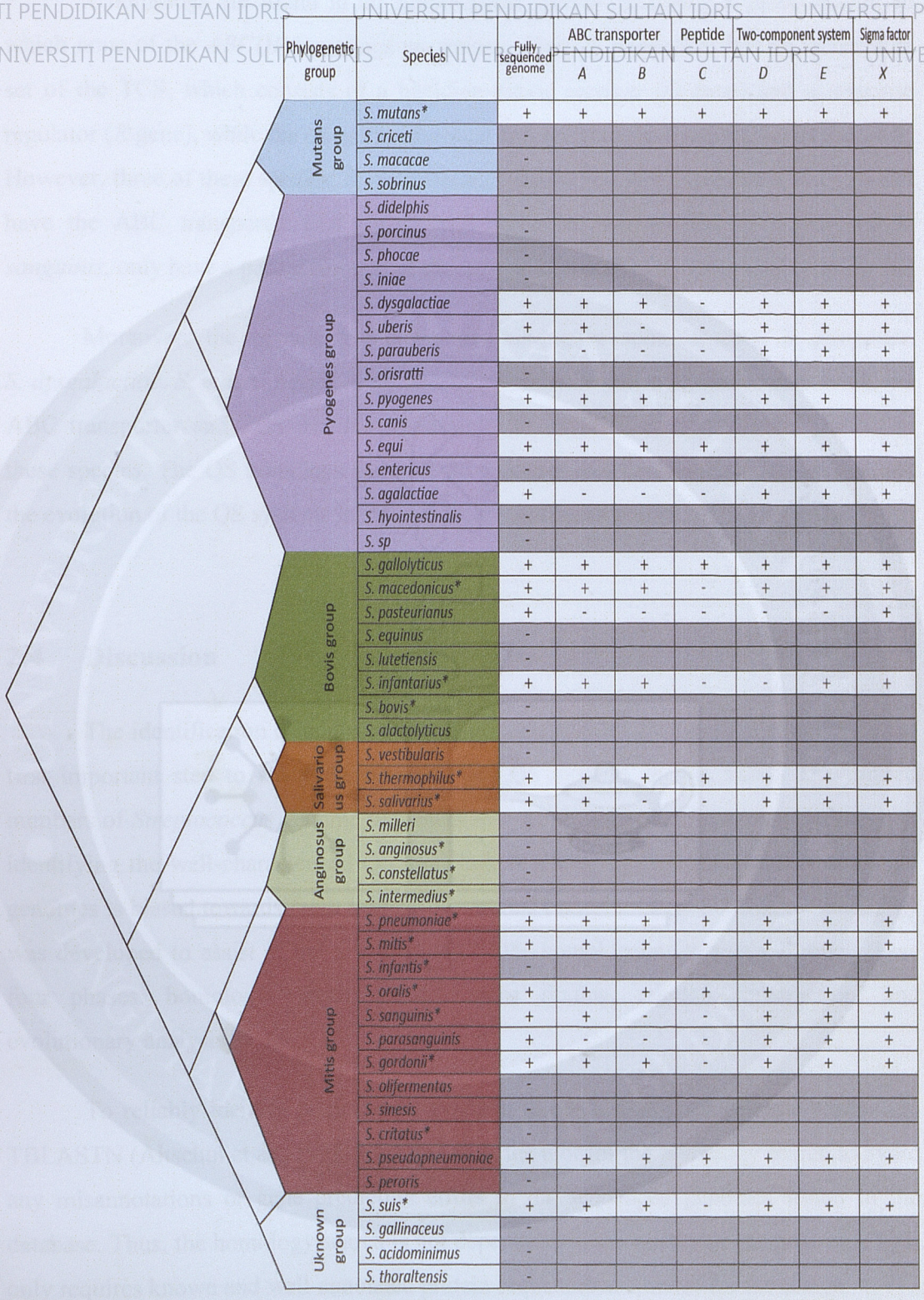


Figure 2.7: The ABC exporter (*A* gene) ML gene tree of 21 *Streptococcus* species. Phylogenetic analysis helps to eliminate distantly related *A* homologs. Only the taxa coloured in red will be considered as the final output of the ABC exporter homologs.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F



Phylogenetic group	Species	Fully sequenced genome	ABC transporter A	ABC transporter B	Peptide C	Two-component system D	Two-component system E	Sigma factor X
Mutans group	*S. mutans**	+	+	+	+	+	+	+
	S. criceti	-						
	S. macacae	-						
	S. sobrinus	-						
Pyogenes group	*S. didelphis*	-						
	S. porcinus	-						
	S. phocae	-						
	S. iniae	-						
	S. dysgalactiae	+	+	+	-	+	+	+
	S. uberis	+	+	-	-	+	+	+
	S. parauberis		-	-	-	+	+	+
	S. orisratti	-						
	S. pyogenes	+	+	+	-	+	+	+
	S. canis	-						
	S. equi	+	+	+	+	+	+	+
	S. entericus	-						
	S. agalactiae	+	-	-	-	+	+	+
	S. hyointestinalis	-						
	S. sp	-						
Bovis group	*S. gallolyticus*	+	+	+	+	+	+	+
	*S. macedonicus**	+	+	+	-	+	+	+
	S. pasteurianus	+	-	-	-	-	-	+
	S. equinus	-						
	S. lutetiensis	-						
	*S. infantarius**	+	+	+	-	-	+	+
	*S. bovis**	-						
	S. alactolyticus	-						
Salivarius group	*S. vestibularis*	-						
	*S. thermophilus**	+	+	+	+	+	+	+
	*S. salivarius**	+	+	-	-	+	+	+
Anginosus group	*S. milleri*	-						
	*S. anginosus**	-						
	*S. constellatus**	-						
	*S. intermedius**	-						
Mitis group	*S. pneumoniae**	+	+	+	+	+	+	+
	*S. mitis**	+	+	+	-	+	+	+
	*S. infantis**	-						
	*S. oralis**	+	+	+	+	+	+	+
	*S. sanguinis**	+	+	-	-	+	+	+
	S. parasanguinis	+	-	-	-	+	+	+
	*S. gordonii**	+	+	+	-	+	+	+
	S. oligofermentas	-						
	S. sinesis	-						
	*S. criatus**	-						
	S. pseudopneumoniae	+	+	+	+	+	+	+
	S. peroris	-						
Uknown group	*S. suis**	+	+	+	-	+	+	+
	S. gallinaceus	-						
	S. acidominimus	-						
	S. thoraltensis	-						

Figure 2.8: The overview of the presence and absence of QS homologs (*ABCDEX*) in 86 genomes across 21 *Streptococcus* species. The presence and absence are indicated by '+' and '-', respectively. The transformable species are indicated by an asterisk (*). Species without fully sequenced genomes are marked with a hyphen (-) in grey boxes.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

X homologs are found in all streptococcal species, including *S. pasteurianus*, in which none of the *ABCDE* homologs are present. Only *S. infantarius* has an incomplete set of the TCS, which consists of a histidine kinase receiver (*D* gene) and a response regulator (*E* gene), while the other streptococcal species have the complete set of the TCS. However, three of these species, *S. parauberis*; *S. agalactiae* and *S. parasanguinis*, do not have the ABC transporter (*AB* genes), and the other two species, *S. uberis* and *S. sanguinis*, only have a partial element of the ABC transporter.

Moreover, the peptide homologs are missing in some species, *S. pyogenes*, *S. dysgalactiae*, *S. macedonicus*, *S. infantarius*, *S. mitis*, *S. gordonii*, and *S. suis*, while the ABC transporter and TCS that are responsible for inducing competence are present in these species. The QS homologs from the 86 genomes will be used as a dataset to study the evolution of the QS systems in streptococcal species in Chapter 5.

2.4 Discussion

The identification of homologous QS genes across all *Streptococcus* species is the first important step to study the origin of the QS systems in this study. The limited numbers of *Streptococcus* genome sequences that are available has become a challenge in identifying the well-characterised QS homologous genes. The set of publically available genomes is biased towards more extensively studied bacteria. Thus, SUCRE methodology was developed to assist in the identification of the homologous QS genes through these four phases: homology search, putative gene finding, regulon construction, and evolutionary analysis.

To reliably identify homologous genes in the poor annotated genome sequences, TBLASTN (Altschul et al., 1997) was chosen as the tool for the homology search to avoid any misannotations or gene prediction errors in the sequenced genomes stored in the database. Thus, the homology search is not dependent on the quality of the genomes, as it only requires known and well-annotated protein sequences as queries for the search.

During the putative gene finding phase, SUCRE reconciled any overlapping hits or adjacent hits as a putative gene to refine the hits from the homology search in the first

phase. From the 10 *Streptococcus* species, 692 putative genes have been constructed from the 84,081 BLAST hits. The reconciliation step ensures the quality of the putative genes identified in the search and removed any redundant hits resulting in the BLAST search that is likely to produce short homologous hits. The assessment criteria: translated in the same protein frame, same lagging/leading strand, and have >40% sequence identity, were used to reassess the quality of the hits resulting in the homology search. These criteria are useful to recognise any sequence frameshift that may be caused by annotation error and any sequence insertion that potentially causing the loss of function in the locus.

The iteration step during the first two phases, homology search and putative gene finding, allow SUCRE to yield high number of QS homologous genes. Since HGT as well as polymorphism are well found in the competence genes in *Streptococcus* species (Cornejo et al., 2010, Whatmore et al., 1999, Claverys and Harvarstein, 2007), the approach takes the divergent QS sequences across *Streptococcus* species into account. The iteration step expands the search by allowing the results of the search to be used as the query sequences in the next iteration. The iteration process is advantageous when attempting to find any divergent sequences that could have been missed in the BLAST search. Nevertheless, this iteration affects the running time of SUCRE. A large dataset requires a longer running time, as more query sequences are used in each iteration. Therefore, the use of a sequence profile on the query using the Hidden Markov Model approach provides higher sensitivity and better search performance in handling sequence divergence, which could be considered as one of the future developments of SUCRE.

One important aspect about SUCRE is that it uses the Com QS regulon structure to search for QS homologs. The methodology incorporates the Com QS regulon structure of *S. pneumoniae*, which one of a well-studied Com QS system (Tomasz and Hotchkiss, 1964, Tomasz, 1965, Claverys and Havarstein, 2002), as a QS regulon structure model. The conserved QS regulon structures of the ABC-transporter (*AB* genes) (Hui and Morrison, 1991, Hui et al., 1995) and the TCS (*DE* genes) (Pestova et al., 1996, Cheng et al., 1997) help narrow the homologous QS genes found and filter away any false positive results. In the analysis of the benchmark dataset, the ABC exporter genes (*A* genes) that have been identified in streptococcal species are outnumbered the accessory protein (*B* genes) with more than 80%. SUCRE eliminates spurious resulting *A* genes based on the regulon structure (*AB* genes) and the evolutionary information from the estimated gene

trees. These criteria are useful for identifying genes from a large gene family, e.g., ABC transporter that is one of the most ancient protein families and widely found in bacteria (Tatusov et al., 1996, Tomii and Kanehisa, 1998), with a known regulon structure. SUCRE may not perform very well on a dataset in which the regulon structure is not conserved. However, SUCRE has a set of parameters that can be adjusted to suit different types of regulon structures (Appendix J). The performance of SUCRE is assessed in the next chapter with other orthology detection methods.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

CHAPTER 3

Assessing SUCRE performance in comparison with other orthology detection methods

Abstract

SUCRE's ability to detect quorum sensing (QS) genes is compared with other orthology detection methods: OrthoMCL, GenomeHistory (GH); Clusters of Orthologous Group (COG); Reciprocal BLAST Hits (RBH) and Simple BLAST Hits (SBH). All of the orthology detection methods under comparison rely on the output from BLAST searches to be used as inputs. From the comparison, SUCRE and OrthoMCL produced similar results in detecting true positive results. However, because of the different strategies and parameters employed, the results show that SUCRE's performance was slightly better, with an average of 3.3% higher true positive value and 4% higher in positive predictive value scores than OrthoMCL in detecting QS genes in streptococcal species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

3.1 Introduction

The homology concept is fundamental to understanding the evolutionary process that is involved in shaping the history of genes. Homologous genes share a common ancestor. Homologs can be described as orthologs if the homologs arise through duplication events of an ancestor, and as paralogs if they derived from an ancestor via speciation events (Fitch, 1970). Several strategies have been adopted to detect homologous genes, and one popular method is to use the Basic Local Alignment Search Tool (BLAST) search approach (Altschul et al., 1997). A BLAST search finds homologous genes based on sequence similarities. A homologous gene with a known protein function could indicate the protein functions of the unknown sequences. However, some of the protein functions are computationally estimated based on sequence similarity, which does not reflect the evolutionary history of the gene.

The Com QS regulon structure was used to complement the BLAST search approach to improve the sensitivity and precision of the homology search. Polymorphisms and horizontal gene transfers are common in bacteria (Nelson et al., 1999, Ochman et al., 2000, Thomas and Nielsen, 2005). The Search Using Competence Regulon (SUCRE) methodology takes the sequence diversity of quorum sensing (QS) homologs into account (see Chapter 2 for details). To improve the efficiency of detecting sequence divergences, SUCRE enables the homology search to run iteratively. As a BLAST search uses the word seeding approach to find homologous sequences, the resulting hits from the BLAST search are more likely to be redundant. Therefore, in SUCRE, adjacent hits with less than 200 bp between them are merged. Three assessment criteria i) translated protein frame, ii) leading/lagging strand, and iii) sequence identity percentage are used to assess the quality of the merged hits of a gene family. Afterwards, evolutionary analyses are used to filter any distantly related hits. The remaining hits are the homologous QS genes.

To assess the performance of SUCRE in detecting 'true' QS genes, SUCRE was compared with other established orthology detection methods: Reciprocal BLAST Hits (RBH); GenomeHistory (GH); Clusters of Orthologous Group (COG); OrthoMCL and Simple BLAST Hits (SBH) search. All of the orthology detection methods under comparison in this study used the BLAST output as an input. Details on the criteria of

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Table 3.1: Summary of the characteristics of SBH, RBH, SUCRE, GH, COG and OrthoMCL.

Approach	Simple BLAST Hits (SBH)	Reciprocal BLAST Hits (RBH)	Search Using Competence Regulon (SUCRE)	GenomeHistory (GH)	Cluster of Orthologous Group (COG)	OrthoMCL
Homology /orthology detection method	TBLASTN	BLASTP	TBLASTN	Wu-BLASTP	BLASTP	BLASTP
Data type	Protein and nucleotide	protein	Protein and nucleotide	Protein and nucleotide	Protein	Protein
Group/cluster capability	No	No	Regulon annotation	No	Yes	Yes
Setting/Parameters analysed	E-value: 1e-5	E-value: 1e-5 (orthoMCL output)	E-value: 1e-5 Percentage of identity: >30% Parameter in the reconciliation phase: genetic distance between adjacent hits (<200bp), same translation frame, same translation direction, coverage (> 40%), initial query proportion (>80%) Parameter in the regulon construction: Coverage (>80%), right and left distance(<2kb)	(Value by default) E-value: 0.01 Percentage of identity: >50% Min number of aligned residues: ≥40 amino acid Length of translated ORF: ≥40	(Value by default) E-value: 1e-5 COG stringency number:3	(Value by default) E-value: 1e-5 Percentage of identity : >50% Filter the sequence with stop codon length less than 10
Phase involved in the programs	1.BLAST protein queries to translated nucleotide genome in all 6 frames (3 forwards and 3 backwards frames)		1. TBLASTN: homology search 2. Filter blast result based on set of parameters 3. Regulon construction based on set parameters	1. wu-BLASTP : Proteomes pairs search 2. CLUSTALW: align pairs 3. calculate Ks/Ka	1. BLASTP: Proteome pairs search 2. SPROCKET (ccpccf): COGs number of stringency to build graph	1. Filter protein input file 2. BLASTP: Proteome pairs search 3. Calculate the percentage match length 4. Find pairwise relationship and assign weight 5. Group pairs based on weight using Markov Clustering algorithm

the comparisons are shown in Table 3.1. These orthology detection methods are classified into two categories: i) mutual BLAST approaches; and ii) graph-based approaches. The methods are categorised based on how each method processes the BLAST output.

The first category consists of the RBH and the GH (Conant and Wagner, 2002), which both use mutual BLAST hits between genomes to group the genes into orthologs and paralogs. RBH is the simplest method used to find orthologous genes. In this approach, two sequences from two different genomes that are the best hits of each other are likely to be putative orthologs (Altenhoff and Dessimoz, 2009, Salichos and Rokas, 2011). However, RBH does not always guarantee that the reciprocal best pair sequences are orthologous, as RBH relies only on the collection of genes that are being compared in each genome (Fuchsman and Rocap, 2006). Meanwhile, the GH program identifies duplicate gene copies in genomes, and examines their evolutionary relationship by calculating the non-synonymous (Ka) and synonymous (Ks) nucleotide substitution rates. The program uses Washington University BLASTP (wu-BLASTP) (Advanced Biocomputing, 1998) to search for all proteome pairs, and aligns them using CLUSTALW with the default BLOSUM 62 matrices. The quality of the sequence pair alignment is assessed based on four criteria: E-value, percentage of identity, the number of aligned residues, and the length of the open reading frame (ORF). Maximum likelihood (ML) estimation is used to calculate the Ka and Ks values of the alignment pairs based on the four criteria.

The extended version of the previous category is the graph-based approach. In this category, two detection approaches are compared in this study, OrthoMCL (Li et al., 2003) and COG (Tatusov et al., 1997). OrthoMCL is a pipeline that uses sequence clustering and the Markov Clustering algorithm (MCL) to cluster ortholog groups. The main orthologs, which are identified from the protein pair with the highest symmetric BLAST score, are used as a 'seed' to find all in-paralogs for each species. All possible pairs are mapped onto a graph in which the nodes represent genes, and the edges represent orthologous relationships. OrthoMCL uses the MCL algorithm to determine the probabilities among the nodes, and partitions the nodes into different ortholog groups. In contrast, COG identifies protein families based on their functions. Each protein sequence in a genome is compared with other protein sequences in another genome. The proteins from different genomes that are the best BLAST hits (BeT) of each other are determined to have a

UNIVERSITI PENDIDIKAN SULTAN IDRIS

congruent BeT score. The COG Stringency Number (CSN) is used to construct the COGs; by default, a congruent BeT score has a CSN of 3. A triangle is constructed from three congruent BeTs (CSN = 3) between three proteins of different genomes. These triangles are merged on their common vertices (Figure 3.1). All of these proteins and the information on their evolutionary relationships of orthology and paralogy are then stored in a database.

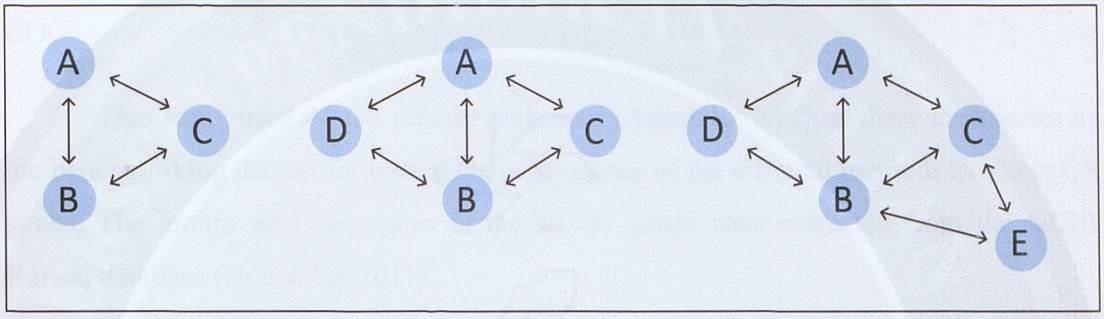


Figure 3.1: The ortholog clustering used in COG. Each circle represents a gene from a different genome (alphabets). The double arrow represents congruent BeTs. The left-most triangle is the minimal COG stringency (CSN=3). Triangles with common vertices are merged together to build COGs (from left to right).

In this study, a simple test was performed to validate the performance of the SUCRE methodology in detecting QS genes across the *Streptococcus* species. The test was also applied to four other orthology detection approaches (RBH, GH, COG and OrthoMCL) for comparison purposes. The Simple BLAST Hit (SBH) search, TBLASTN (Altschul et al., 1997), is also included in the analysis as a control method for searching for QS genes. The 44 QS genes from six *Streptococcus* species with well-annotated genomes were used as true positive (TP) benchmark genes. These approaches were compared based on their ability to detect the QS benchmarking genes. This comparison was useful to enhance the confidence in SUCRE's performance, and perhaps to help indicate any systematic errors that occurred during the homology detection (i.e., false positives).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

3.2 Materials and Methods

3.2.1 Protein query and subject genome sequences

Six *Streptococcus* genomes (Table 3.2) were retrieved from the NCBI Genome database (June 2011) and stored in a local BLAST database as an input for the SUCRE and SBH approaches. 12,010 protein and nucleotide sequences from these six genomes were extracted from the NCBI RefSeq database (November 2011). These sequences were used as an input to the other orthology detection methods, RBH, GH, OrthoMCL and COG.

This study used 44 QS protein sequences (Appendix A) from these six species as the benchmarking dataset for testing the performance of the different methods to detect QS genes. The amino acid sequences of the 44 QS genes were extracted from the NCBI Refseq database (November 2011).

Table 3.2: The subject genomes of six *Streptococcus* species.

Species	Strain	Accession number	Number of RefSeq protein
S. pneumoniae	G45	NC_011072	2114
S. equi	MGC10565	NC_011134	1893
S. mutans	UA159	NC_004350	1960
S. gallolyticus	UCN34	NC_013798	2223
S. thermophilus	LMG18311	NC_006448	1888
S. suis	BM407	NC_012926	1932

3.2.2 The approaches and programs used for the comparison

Given the above benchmark dataset, a validation test was conducted to assess the performance of SUCRE in comparison to the other orthology detection methods: OrthoMCL (Li et al., 2003), GH (Conant and Wagner, 2002), COG (Tatusov et al., 2000) and RBH.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

The following are the procedures and requirements of each method to run the performance validation testing:

a) Search Using Competence Regulon (SUCRE) methodology

SUCRE employs the QS regulon structure to identify homologous QS genes, as previously described in Chapter 2. SUCRE consists of four phases: i) homology search; ii) putative gene finding; iii) regulon construction; and iv) phylogenetics analysis. The protein queries were used to search for the QS genes in each dataset. Gene tree topologies and the regulon structure of these homologs were used to exclude any spurious homologs that resulted from the search.

b) Clusters of Ortholog Group (COG)

There are only two *Streptococcus* species available in the COG database (April 2012). However, these species are not included in the benchmarking dataset. I also noted that the QS genes in both genomes are not fully characterised by the COG database. Thus, a similar COG procedure was conducted to build the COGs for the six *Streptococcus* genomes. The all-against-all protein sequences in all six benchmark genomes (Table 3.2) were analysed using BLASTP to produce a list of best mutual protein hits. A C++ clique COG finder (cppccf), which is part of the system for protein recognition using ORF COGs – a knowledge extraction tool (SPROCKET) program (Powell and Hutchison, 2006), was applied to produced COGs at various levels of stringency (CSNs). The list of best mutual protein hits was used as an input to the cppccf program, with a specified CSN of 3 to employ the COG stringency.

c) OrthoMCL

The program was downloaded from http://orthomcl.org/common/downloads/software/, and the Markov Cluster Algorithm (MCL) approach was retrieved from http://micans.org/mcl/. The program was executed using the default parameters: the percentage of stop codons (20%) and protein sequence

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

length (10) to filter ambiguous protein sequences. The E-value (1e-5) and the percentage of identity (50%) parameters were set to the default cut-off value in the all-against-all BLASTP search.

d) GenomeHistory (GH)

The program was downloaded from the GenomeHistory website: http://www.ieu.uzh.ch/wagner/software/GenomeHistory/. The parameters were set to the default values: the E-value (0.01), the percentage of identity (50%), the minimum number of aligned residue (40) and the minimum translated ORF length (40).

e) Reciprocal best BLAST Hits (RBH)

In this study, the reciprocal results were obtained from the all-against-all BLASTP of the OrthoMCL output.

f) Simple BLAST Hits (SBH)

BLAST is an alignment tool that searches for identical nucleotide and amino acids sequences in a database of sequences (Altschul et al., 1990). In this study, the identification of homologous QS genes was performed at the amino acid level to increase BLAST search specificity and sensitivity. A cut-off E-value of 1e-5 was used to filter insignificant BLAST output hits.

3.2.3 Testing the performances of the compared methods in detecting QS genes

A simple test was done to validate SUCRE's performance in comparison with other orthology detection methods. A Perl script was written to generate 100 testing datasets. Each of the datasets had an input comprising six randomly picked genes (*ABCDEX*) from the 44-gene benchmarking dataset (Appendix A). For unbiased

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

comparisons and evaluations of QS genes detection, the same input sets were used for all the methods in the 100 tests.

To evaluate the performance of each method in detecting QS genes in the six *Streptococcus* genomes, the true positive rate (TPR) and the positive predictive value (PPV) were calculated for each output of the dataset. To calculate the TPR and PPV, I needed to determine the true positive (TP), false negative (FN) and false positive (FP) values. These three values were calculated for each of the 100 tests executed by each method. The input benchmark genes (Q_i) that were used in each test were excluded from these calculations. Thus, the expected number of the true benchmark genes detected in each test (Q_o) is the total number of the benchmark genes (Q_b) detected, excluding Q_i, as shown in the equation (Eq. 1):

$$Q_o = Q_b - Q_i \qquad \text{(Eq. 1)}$$

The TP value (Eq. 2) is calculated from the number of the actual QS benchmark genes detected (Q_o) from the inferred homologous QS genes (S_i):

$$TP = (S_i \cap Qo) \qquad \text{(Eq. 2)}$$

While, the remaining inferred homologous QS genes are accumulated as the FP (Eq. 3):

$$FP = S_i - TP \qquad \text{(Eq. 3)}$$

The QS benchmark genes that failed to be detected by the methods were counted as the FN (Eq. 4):

$$FN = Qo - TP \qquad \text{(Eq. 4)}$$

The TPR (Eq. 5), also known as recall, is the proportion of the QS benchmark genes that were successfully detected in the analysis:

$$TPR = TP/ (TP + FN) \qquad \text{(Eq. 5)}$$

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

While the PPV (Eq. 6), also known as precision, is the proportion of the detected homologous QS genes that are the actual QS benchmark genes:

$$PPV = TP/ (TP + FP) \qquad \text{(Eq. 6)}$$

I used the PPV values in comparison to the TPR values due to the difficulty in assigning the negative values. Although the genome is annotated, assigning all the non-benchmark genes in a genome as the true negative value can be misleading.

The outputs of each method were sorted and formatted accordingly before calculating the TP, FN and FP values. The BLAST outputs for the 100 datasets were checked for redundancy and overlapping hits. In each output dataset, any overlapping hits were merged together, given that they share the same direction, are translated in the same frame, and are predicted to be the same gene family as the final BLAST results. As for the RBH approach, only the mutual best hits of the input genes were considered as the final RBH results for each dataset. A similar process was applied in processing the final output of the GH program. Only the genes paired with the input genes were considered as the GH final results. As for the OrthoMCL and COGs that use the graph-clustering approach, the final result for each 100 tests was formed by picking up the ortholog clusters that contained an input benchmark gene, Q_o. The inferred homologous QS genes in these chosen clusters were counted as the results of the methods.

Here, 1's and 0's were used to represent the success and failure, respectively, of each detection method in detecting the QS benchmark genes.

3.3 Results

3.3.1 Identification of the QS benchmark genes

To visualise the distribution of the 44 QS benchmark genes detected in the 100 testing dataset, the frequency (percentage) of each benchmark gene detected by the six methods was calculated (Figure 3.2). The *A* gene family had the highest frequency of

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

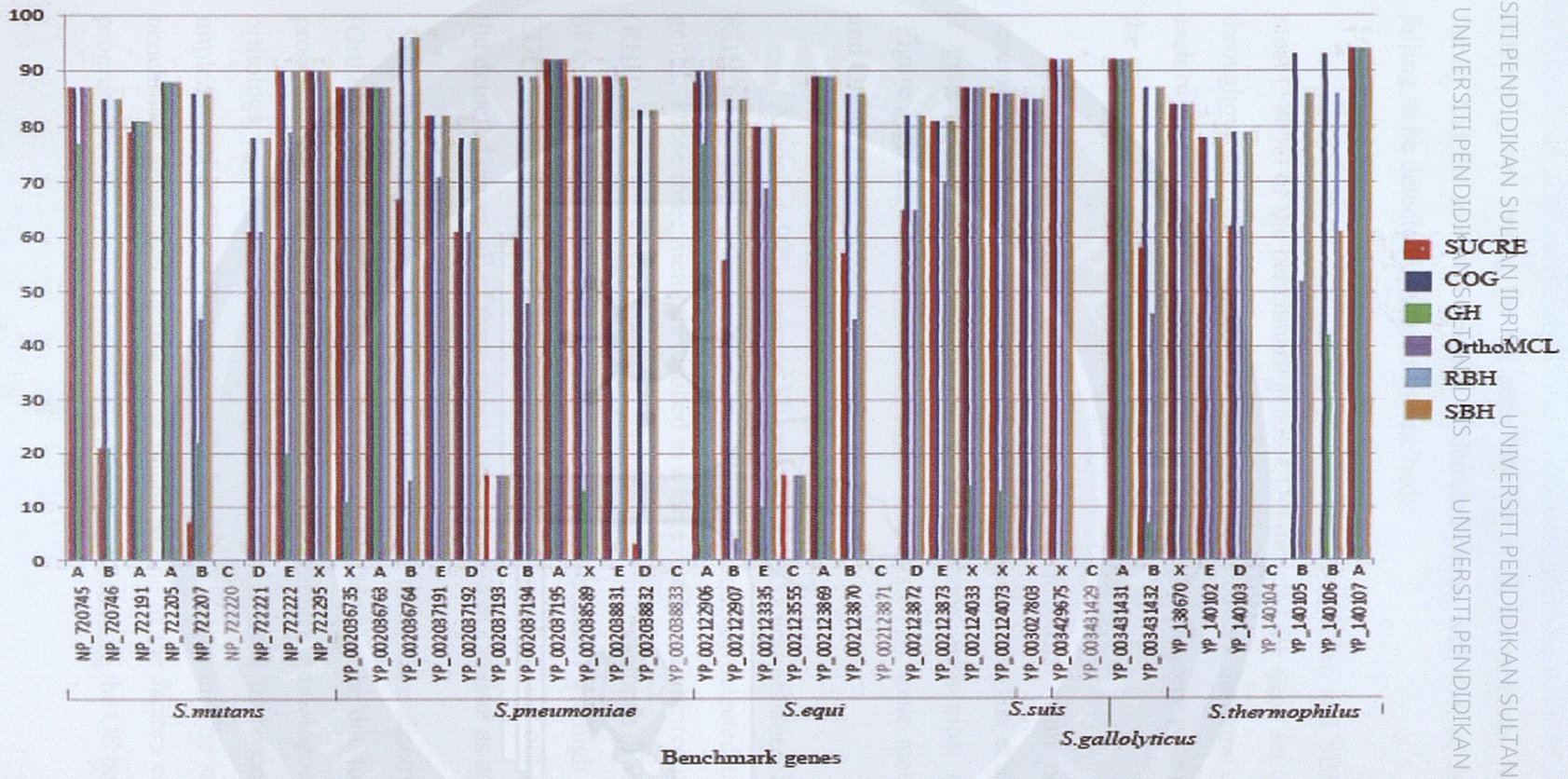


Figure 3.2: The detection of QS benchmark genes (gene ID) by the compared methods. The height of the bars shows the frequency of the detected benchmark genes in 100 tests. The bars are coloured according to each compared method: SBH (orange), RBH (light blue), SUCRE (red), GH (green), COG (dark blue) and OrthoMCL (purple). The geneID, in grey, is the gene that cannot be detected by any of the methods.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

detected genes among the six gene families in all the methods. Meanwhile, the *C* gene family had the lowest frequency of detected genes, with four of the *C* benchmark genes failing to be detected by any of the methods.

Based on the results (Figure 3.2), the COG, RBH, and SBH approaches detected most (~80%) of the benchmark genes. COG detected the same set of benchmarking genes throughout the 100 tests. This is because COG produces the same orthologous groups for each of *ABCDEX* genes, despite the fact that a different query gene set was used to initiate the searching in each test.

It is known that ABC transporters, which consist of an exporter *A* gene and an accessory *B* gene, are usually found adjacent to each other (Hui et al., 1995). Although the *A* gene was reliably detected by all six methods, with most methods scoring 80% (Figure 3.2), the detection of the *B* gene was missed by some methods, such as SUCRE and OrthoMCL.

Around a 90% average FP rate (Table 3.3) was obtained in the COG, SBH and RBH tests, despite the fact that these three approaches detected most of the benchmark genes. These three methods resulted in a large number of inferred genes: 593 (COG), 446 (RBH), and 410 (SBH), compared with the other three methods (Table 3.3). The simplicity of the SBH and RBH approaches could be the cause of the high FP values. As for the COG approach, as no differentiation between in-paralogs and out-paralogs was made for the detected genes, all of these genes were gathered in a cluster as an ortholog.

In contrast, the other three methods had lower false discovery percentages: 25% (OrthoMCL), ~24% (SUCRE) and 14% (GH). This may be due to the complexity of the program, e.g., the OrthoMCL program differentiates homologous genes into putative orthologs and paralogs. The stringent parameters used to process the detected genes implemented in SUCRE may affect the homology detection. The small sample size of the benchmark genes used in this study might be one of the factors explaining why the GH program was the poorest performing program in detecting the QS benchmark genes.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Table 3.3: The QS genes detection performance of SBH, RBH, SUCRE, GH, COG and OrthoMCL.

Approach	Simple BLAST Hits (SBH)	Reciprocal BLAST Hits (RBH)	Search Using Competence Regulon (SUCRE)	GenomeHistory (GH)	Clusters of Orthologous Group (COG)	OrthoMCL
Average fraction of total true positive genes of the benchmark dataset	31.93 ± 0.07	34.77 ± 0.20	25.18 ± 0.21	9.5 ±0.1298	32 ± 0	23.92 ±0.25
Inferred gene (Average fraction of the total of positively predicted (TP+FP) genes from the100 datasets)	410.93 ± 1.67	446.4 ± 1.53	31.71 ± 0.29	11.2 ± 0.19	593 ± 0	32.11 ± 0.4
% false positive (FP) rate	92.22 ± 0.04	92.20 ± 0.05	17.18 ± 0.10	14.38 ± 0.81	94.60 ± 1.56	25.15 ± 0.47
Average PPV of the 100 testing sets	0.08 ± 3.73E-4	0.08 ± 5.18E-4	0.79 ± 1.89E-3	0.85 ± 8.07E-3	0.05 ± 4.88E-18	0.75 ± 4.71E-3

3.3.2 Performance in detecting the QS genes

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID

The performance of all the approaches in identifying the QS benchmark genes was assessed through the recall (TPR) and the precision (PPV) values. The best method would have a score of '1' on recall and a score of '1' on precision. A score on recall indicates the proportion of the correct genes that are detected, and a score on the precision indicates how precise the detection method is in detecting actual genes.

The TPR and PPV values of the six methods were plotted for 100 testing datasets (Figure 3.3). The COG, SBH and RBH approaches had the highest recall rates, but the low precision rates, in predicting true positive genes. In contrast, the GH program missed nearly all of the true positive genes, but had a very high precision rate, with some of the tests scoring 100% precision.

From the QS benchmark genes detection, SUCRE performed slightly better than OrthoMCL, with an average of 3.3% higher in TPR and 4% higher in PPV. Therefore, it is safe to deduce that OrthoMCL and SUCRE are better approaches in comparison to the other detection methods.

The mapping pattern of the 44 benchmark genes detected by OrthoMCL and SUCRE is illustrated in a Venn diagram (Figure 3.4). Comparisons were made between each of the 100 testing datasets. The three patterns shown in the diagram of the QS genes were: i) those successfully detected by both methods, ii) those detected by one method, and iii) those that were not detected by either method. The percentage of the QS benchmark genes found through each pattern are calculated.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

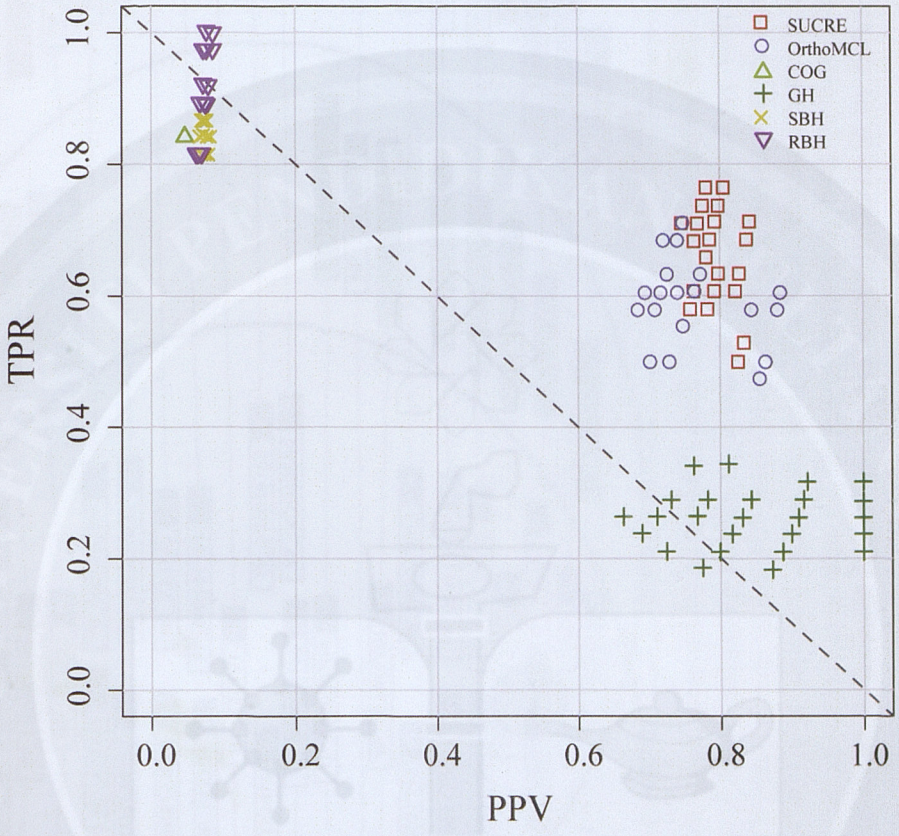


Figure 3.3: Positive predictive values (PPVs) are plotted against the true positive values (TPVs) for all methods: Search using Competence Regulon (SUCRE), OrthoMCL, Clusters of Orthologous Group (COG), GenomeHistory (GH), Simple BLAST Hits (SBH) and Reciprocal BLAST Hits (RBH). The best detection method is the method that has values of 1.0 for PPV and 1.0 for TPR (the upper right corner).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

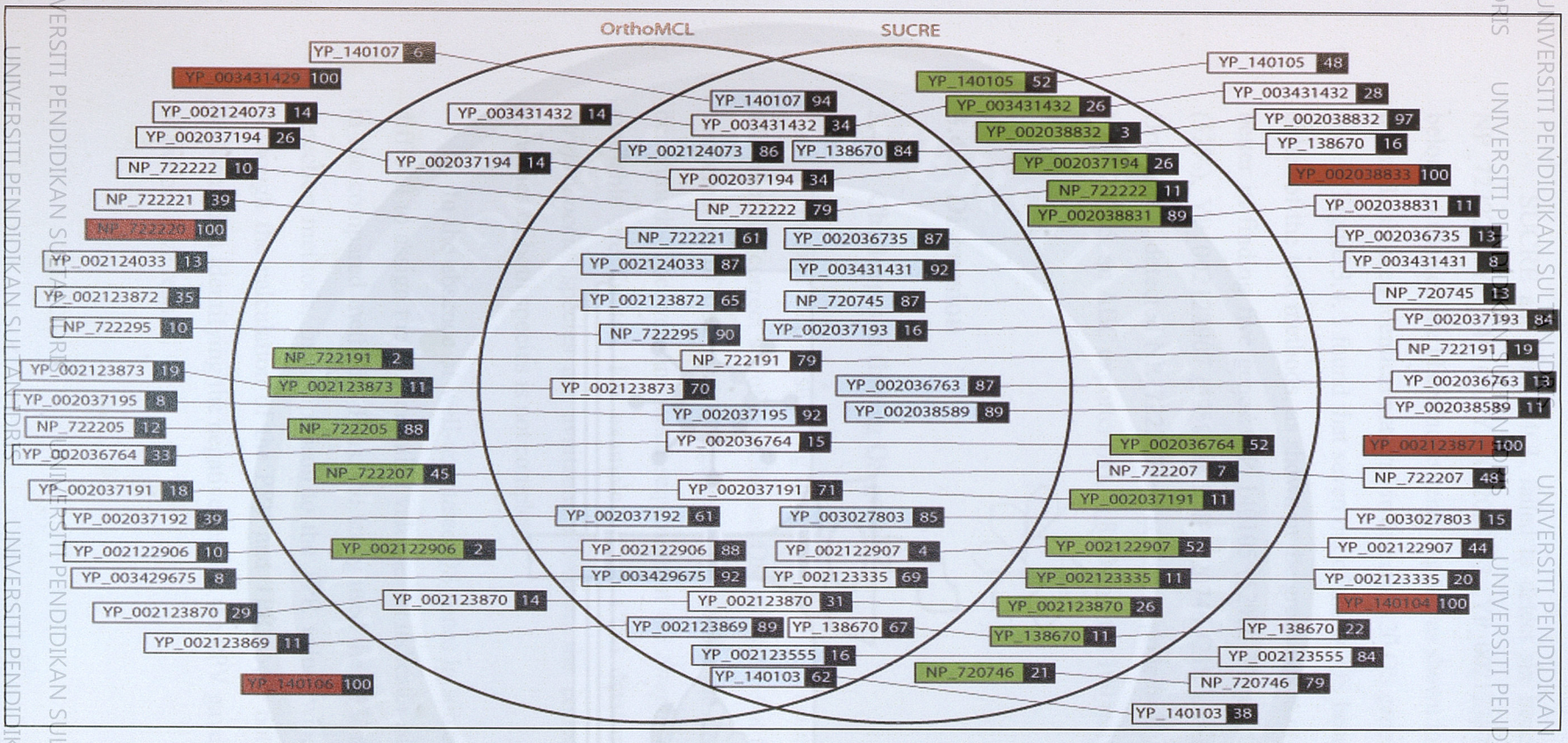


Figure 3.4: Venn diagram of benchmark true positive (TP) genes detected by OrthoMCL and SUCRE in 100 testing sets. The circles represent the set of TP genes detected by each method. The area of the overlapping circles indicates the genes detected by both methods. Each box indicates a TP gene (gene ID) with a percentage of the TP genes found in the 100 tests (black box). The boxes that are connected by lines are referring to the same TP gene to show the distribution of the gene detected during the tests. The best detection patterns of the TP genes are highlighted by the box's colour. Light blue boxes indicate an identical percentage of TP genes detected by both methods. Green boxes indicate the TP genes that were better detected by either method. Red boxes indicate the TP genes that were not detected by either method. The white boxes indicate the remaining genes detected or missed in the 100 tests.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

SUCRE and OrthoMCL failed to detect six benchmark genes, YP_003431429, NP_722220, YP_140106, YP_002028833, YP_002123871 and YP_140104, which all belong to the peptide pheromone gene, *C* gene, as shown in the red box in Figure 3.4. Both methods showed identical patterns in detecting 20 QS genes, as illustrated in the light blue box in Figure 3.4. I found that seven out of the 44 QS benchmark genes were detected by either of the two methods, as shown in the green boxes in Figure 3.4. SUCRE performed better in finding five genes: YP_140105 (52%), YP_002038831 (89%), YP_002036764 (52%), YP_002122907 (52%) and NP_720746 (21%). It is also worthwhile to note that OrthoMCL detected NP_722207, an accessory protein (B), in 45 out of the 100 tests and NP_722205, an ABC exporter (A), in 88 out of the 100 tests.

3.4 Discussion

3.4.1 Challenges in finding QS genes

The first step in studying the evolution of QS systems in *Streptococcus* species is to detect the correct set of homologous QS genes. One of the challenges in validating the performance of detecting QS genes is to prepare the dataset that could be used as a benchmark. The well-annotated sequences of the *Streptococcus* species that are publicly available are limited. For example, in the COG database, only two genomes of two *Streptococcus* species are available. However, the collection of the QS genes in the genomes of both species is not complete.

In the absence of well-annotated QS genes in some streptococcal genomes, it was difficult to assign the true negative value as an indicator of the detection of the QS genes in the compared methods. Thus, the study used the PPV to measure the accuracy of the detection methods, in comparison to the TPR value, which was used as an alternative. Although the calculation of the PPV and TPR could not determine how well SUCRE performed in identifying the negative results, the PPV gave an estimation of the reliability of the positive results.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

3.4.2 SUCRE performance in detecting the QS genes

The study shows SUCRE's performance in comparison to other orthology detection methods. SUCRE used a different strategy in refining the BLAST output to detect the QS genes. Stringent parameters were used in two phases: i) finding putative genes, and ii) constructing the regulon structure, as previously described in Chapter 2. These two phases were used to filter the BLAST hits output, which seemed to work well in reducing the number of false positives (Table 3.3). SUCRE reduced the false positive rate by more than 70% compared with the SBH and RBH approaches.

Although the SBH approach identified most of the TP genes, it produced the highest FP rate. The SBH's high FP rate is probably due to the direct approach, without stringent parameters, which was used in attaining the output. The SBH extended approach, RBH, which only considers two sequences that are mutual best hits of each other, increased the number of TP genes detected, but slightly reduced the high FP rate.

A different COG strategy, which applies the 'triangle' graph approach to classifying the identified BLAST output of QS homologous genes into orthologs and paralogs, did not help to reduce the FP rate, although the COG strategy successfully identified more than 80% of the TP results (Figure 3.3). The GH program, which calculates the Ka/Ks value of the pairwise sequences to eliminate unrelated sequences, produced the reverse results. GH had a very low FP rate, but lost many TP results. SUCRE, together with the OrthoMCL approach, had an advantage over the other competing methods in detecting QS genes.

Although SUCRE produced similar results to OrthoMCL in detecting the 44 benchmark QS genes, these two methods employed different strategies and parameters. However, SUCRE performed slightly better than OrthoMCL in detecting QS genes. SUCRE reduced the FP results by merging adjacent hits (those <200 bp apart) to find putative homologous genes and construct the regulon structure based on the genes' locations in the genome to eliminate any redundant hits. While in OrthoMCL, the ortholog groups are constructed from a complex graph that is based on a weighted pairwise comparison using the Markov graph algorithm.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

A direct comparison cannot be made between the SUCRE and OrthoMCL approaches, as OrthoMCL differentiates homologous genes into orthologs, in-paralogs and co-orthologs. However, the performance of OrthoMCL could be used as an indication to show how well SUCRE identifies homologous QS genes. The advantage of the SUCRE methodology is the addition of the QS regulon synteny that is used as a criterion to identify the homologous QS genes. There are still potential improvements that can be made to SUCRE, such as using a conserved region (e.g., a promoter) to enhance the detection of a short peptide pheromone genes.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

CHAPTER 4

Streptococcus species phylogeny

Abstract

A species phylogeny provides vital information on the evolutionary history of species; however, the estimation of a species phylogeny is not direct. Issues regarding the generation of a species tree, such as conflicts between a species tree and gene trees, and whether genes should be combined or analysed separately, are still debated. In this chapter, a *Streptococcus* species phylogeny was generated from ribosomal protein and tRNA synthetase. Because the genes were found to have different evolutionary signals, it was not possible to combine the genes into a supermatrix. Therefore, the genes were analysed separately. The Shimodaira-Hasegawa (SH) test was used to calculate any discordance between the gene trees and to identify the best tree(s) to represent the evolutionary history of the *Streptococcus* species. This study represents a crucial step in studying the evolution of quorum sensing in streptococcal species, as an incorrect, inferred species tree could negatively affect the 'true' explanations of how these genes evolved in each species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

4.1 Introduction

Two of the challenges in building a species tree are deciding which genes should be used, as well as the most appropriate approach. The discordance between gene trees and a species tree, which are due to horizontal gene transfers (HGTs), recombination events, duplications and lineage sorting, has made it more difficult to directly identify the best approach (Philippe and Douady, 2003, Beiko et al., 2005, Degnan and Rosenberg, 2006). Stochastic errors, e.g., model misspecification during the phylogenetic analysis, could also affect the congruency between a species tree and gene trees. Two popular methods used to build species trees are: i) the supermatrix approach, in which all of the genes are combined into a supermatrix, based on the assumption that these genes have similar evolutionary histories and; ii) the supertree approach, in which a consensus tree is determined from the gene trees, resulting in the separate phylogenetic analysis of each gene. There have been on-going arguments for and against combining genes (de Queiroz and Gatesy, 2007) or performing a separate analysis (Bininda-Emonds et al., 2002, Bininda-Emonds, 2004) to construct a reliable species tree.

The statistical approach, such as the Shimodaira-Hasegawa test (SH test) (Shimodaira and Hasegawa, 1999) and the likelihood ratio test (Huelsenbeck and Bull, 1996) help to assess the congruence of the evolutionary signals between genes in a phylogenetic analysis. The SH test was performed by running a statistical analysis on the competitive evolutionary hypothesis based on the tree topologies. The approach can provide an indication of whether the set of genes should be combined or separately treated, based on the trees' properties, i.e., tree topology.

To understand the evolution of quorum sensing (QS) systems in *Streptococcus* species, a reliable *Streptococcus* species tree is needed. A *Streptococcus* species tree is built using the Neighbor-Joining (NJ) approach (Saitou and Nei, 1987) from the 16S ribosomal protein (Kawamura et al., 1995, Martin et al., 2006). The divergence of 16S ribosomal proteins, in which the protein family was duplicated and some ribosomal proteins horizontally transferred, is already known (Brochier et al., 2000, Pei et al., 2010). The problem associated with the NJ approach (Farris, 1985, Farris et al., 1996), e.g., the estimation of branch length using the pairwise distance, may affect the topological structure of the inferred tree (Huelsenbeck and Bull, 1996, Ogden and Rosenberg, 2006).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

The divergence of the ribosomal protein sequence properties due to the evolutionary process, e.g., HGTs and duplications (Nelson et al., 1999, Gogarten et al., 2002, Sunita et al., 2007, Pei et al., 2010) can complicate the process of inferring phylogeny. In addition, the systematic errors and the accuracy of the alignment produced may affect the phylogenetic construction to attain the same species phylogeny. This triggered this study to use a different approach to construct a *Streptococcus* species phylogeny to better understand the evolutionary relationships between each *Streptococcus* species.

4.2 Materials and Methods

4.2.1 The gene dataset

An attempt to obtain the collection of 16S ribosomal genes in the Ribosomal Database Project II (used in previous study Martin et al., 2006) was unsuccessful, as most of the ribosomal genes were found to be present in multiple copies, and some of the gene sequences were absent in some species. The dissimilarity number of 16S ribosomal gene copies with only some species have the common 16S RNA genes may affect the estimation of *Streptococcus* species phylogeny and limit the representation of streptococcal taxa. The great variability of 16S RNA sequence is probably due to the HGT that was proposed to be a potential cause (Bentley and Leigh, 1995, Schouls et al., 2003, Bodilis et al., 2012).

Rather than extracting the 16S RNA genes, this study adopted 31 ribosomal and tRNA synthetase genes. These 31 gene families (Table 4.1), which comprise ribosomal proteins and tRNA synthetase, present in 58 genomes of 21 *Streptococcus* species (Appendix G) were used to generate the streptococcal species tree. The 31 orthologous genes were used in a previous study (Ciccarelli et al., 2006) to construct the tree of life, which included four *Streptococcus* species (*S. mutans, S. pneumoniae, S. pyogenes* and *S. agalactiae*). These gene families are universally present in prokaryotes and eukaryotes, which are most likely to be conserved in *Streptococcus* species. Good species coverage would increase robustness of the species phylogeny estimation.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Table 4.1: The list of 31 orthologs from ribosomal and tRNA synthetase genes in the *Streptococcus* species.

Orthologous group	Annotation
COG0012	Predicted GTPase, probable translation factor
COG0016	Phenylalanine-tRNA synthetase alpha subunit
COG0048	Ribosomal protein S12
COG0049	Ribosomal protein S7
COG0052	Ribosomal protein S2
COG0080	Ribosomal protein L11
COG0081	Ribosomal protein L1
COG0087	Ribosomal protein L3
COG0091	Ribosomal protein L22
COG0092	Ribosomal protein S3
COG0093	Ribosomal protein L14
COG0094	Ribosomal protein L5
COG0096	Ribosomal protein S8
COG0097	Ribosomal protein L6P/L9E
COG0098	Ribosomal protein S5
COG0099	Ribosomal protein S13
COG0100	Ribosomal protein S11
COG0102	Ribosomal protein L13
COG0103	Ribosomal protein S9
COG0172	Seryl-tRNA synthetase
COG0184	Ribosomal protein S15P/S13E
COG0186	Ribosomal protein S17
COG0197	Ribosomal protein L16/L10E
COG0200	Ribosomal protein L15
COG0201	Preprotein translocase subunit SecY
COG0202	DNA-directed RNA polymerase, alpha subunit
COG0256	Ribosomal protein L18
COG0495	Leucyl-tRNA synthetase
COG0522	Ribosomal protein S4 and related proteins
COG0525	Valyl-tRNA synthetase
COG0533	Metal-dependent protease with chaperone activity

UNIVERSITI PENDIDIKAN SULTAN IDRIS

The 31 orthologous genes have been carefully identified and analysed for the occurrence of HGT, which is a common mechanism in streptococcal species. The 21 *Streptococcus* species were chosen based on the availability of the genome sequences, as previously mentioned in Chapter 2. However, none of these orthologous genes was found in the *S. infantarius* genome. Thus, the protein sequences of these 31 orthologous genes in the remaining 20 *Streptococcus* species were extracted from the Protein Cluster database in the NCBI using the COG ID (August 2012).

4.2.2 Supermatrix analysis

To decide whether these genes should be analysed using the supermatrix approach or supertree approach, I checked the congruent evolutionary signals between these 31 genes using the Concaterpillar program (Leigh et al., 2008). Each gene family was aligned by Muscle (Edgar, 2004) and used as an input to the Concaterpillar program. The gene trees were estimated via RAxML (Stamatakis, 2006b), which was embedded in the Concaterpillar program. The program assesses the topologies (as well as the branch-lengths), which are in agreement between the estimated trees of each gene family, using hierarchical clustering and likelihood ratio testing (LRT) (Huelsenbeck and Bull, 1996). The program calculates the P-value from the log-likelihood score for all possible pairs between the 31 gene families using a non-parametric bootstrap. Gene pairs with a P-value larger than the corrected alpha level (default $\alpha = 0.05$) were combined in a supermatrix. The program iterates this assessment process until there are no possible gene to combine into any of the pre-computed supermatrices. By default, the program runs the congruent test on the branch length of the largest supermatrix.

In the end, the Concaterpillar program produces supermatrices subset(s), which consist of a concatenated alignment of the gene families that share congruent evolutionary signals. I compared the largest two supermatrix trees and manually removed the species that were assumed to affect these two supermatrices. *S. agalactiae* was found to be clustered differently in both supermatrix trees. To enhance the confidence of the new supermatrix, I removed *S. oralis* and *S. pseudopneumoniae*, as these two species were also found to be in different part of the trees. To ensure that removing these species improved the evolutionary signals between the remaining species, five sets of analyses, which consisted of 20 species, 19 species (after removing *S. agalactiae*), 17

UNIVERSITI PENDIDIKAN SULTAN IDRIS

species (after removing the *S. oralis* and *S. pseudopneumoniae*), as well as two analyses of 18 species (in which either *S. oralis* or *S. pseudopneumoniae* was removed), were conducted.

To check for significant differences between each of the estimated subset trees inferred in each analysis, the SH test (Shimodaira and Hasegawa, 1999, Goldman et al., 2000) via the CONSEL program (Shimodaira and Hasegawa, 2001) was used. Because the COG525 gene family has been reported to have undergone HGT (Woese et al., 2000), the gene was excluded from the analysis. The trees with P-values equal to or greater than 0.05 were considered to be non-significantly different than the ML subset trees. The distances between each of the supermatrix subset trees were calculated using the Robinson-Fould distance (Robinson and Foulds, 1981) in the EMBOSS Treedist program (Felsenstein, 2005).

4.2.3 Supertree analysis

In the supertree analysis, a ML gene tree with 100 nonparametric bootstrap re-sampling was built from each gene family alignment using RAxML (Stamatakis, 2006a) via the WAG+Γ substitution model (Whelan and Goldman, 2001). From the supermatrix analysis, a better coverage of supermatrices was obtained by removing the three *S. agalactiae* genomes. Thus, in this supertree analysis, 31 gene families from 55 genomes of 19 *Streptococcus* species were employed. This analysis also includes the COG0525 and COG0495 orthologs, which were found to be individually clustered from other gene family supermatrices, for the comparison purposes.

To test whether there were significant differences between the 31 trees of the orthologous genes, the SH test (Shimodaira and Hasegawa, 1999) via the PAML program (Yang, 2007) was used. All 31 compared gene trees needed to have the same number of taxa for the topological congruence test. Some of the 31 genes were found to be absent in 15 genomes. To maintain the species coverage, five species with a single genome representative *(S. parasanguinis, S. parauberis, S. salivarius, S. macedonicus* and *S. oralis)* were kept, and four gene families (COG0093, COG0184, COG0200 and COG0052) were excluded. While the other 10 genomes of three *Streptococcus* species (*S. pneumoniae*, *S. pyogenes* and *S .suis*.) were excluded in the analysis to maximise the number of gene families involved. As a result, only 27 gene families across 19

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Streptococcus species remained in the analysis. A 5% significance value was used to reject a tree that was significantly different than the ML gene tree.

The outgroups of four species, *Lactococcus lactis*, *Enterococcus faecalis*, *Lactobacillus johnsonii* and *Staphylococcus aureus*, were rooted to the best-scored tree(s). The root of the phylogeny was used to determine the direction of evolution in the *Streptococcus* species. The outgroup species were chosen based on the availability of the 27 orthologous genes. All of the amino acids sequences of the outgroup species were retrieved from the Protein Cluster database in the NCBI (Dec 2012).

4.2.4 Constructing the *Streptococcus* species tree

The best-scored tree(s) resulting from the supertree analysis was used to generate the *Streptococcus* species tree. A strict consensus species phylogeny was constructed from the two trees that best represented the history of the gene families using the Consense program in the PHYLIP package (Felsenstein, 2005). The strict consensus method builds a consensus tree by only considering the tree topological agreement found in all of the participating trees. Any conflict branch in the topology between these trees collapses into an unresolved polytomy.

4.3 Results

4.3.1 Evolutionary signals between the gene families

The species tree could also be produced by concatenating all the genes of interest into one supermatrix (de Queiroz and Gatesy, 2007). However, all of the genes were tested beforehand to ensure that they had similar evolutionary signals prior to their being combined. The evolutionary signals of the 31 orthologous genes in the 20 Streptococcus species were tested using the Concaterpillar program (Leigh et al., 2008). These genes clustered into five supermatrix subsets (Analysis 20sp. in Table 4.2). The largest supermatrix subset consisted of 13 genes, followed by a supermatrix with nine genes and another two supermatrix subsets composed of three and four genes each. Two of the supermatrix subsets consisted of a single gene family: COG0495, a leucyl-tRNA synthetase, and COG0525, a valyl-tRNA synthetase. However, only COG0525 was

reported to have been horizontally transferred (Doolittle and Handy, 1998, Woese et al., 2000). The two largest supermatrix subset (20sp.Subset0 and 20sp.Subset2, Table 4.2) trees had a Robinson-Foulds distance of 60, which indicates the unique split between these two trees topologies.

To improve the confidence in the supermatrix subset, I removed some species that had different evolutionary histories between these supermatrix trees. These species were assumed to affect the congruent signals between the supermatrix subsets. Removing *S. agalactiae* from the analysis (Analysis 19sp. in Table 4.2) produced a similar pattern of supermatrix clustering, with the two largest supermatrices comprising 10 and 13 gene families (Analysis 19sp. in Table 4.2). However, the supermatrices had a smaller number of gene families when another streptococcal species, e.g., S. oralis, S. pseudopneumoniae, was removed (Analysis 18sp.; Analysis 17sp. in Table 4.2).

Table 4.2: The number of gene families in each supermatrix subset. Five supermatrix analyses: i) Analysis 20sp. (20 streptococcal species); ii) Analysis 19sp. (19 streptococcal species); iii) Analysis 18sp_S.ps (18 streptococcal species, *S. oralis* excluded); iv) Analysis 18sp_S.or (18 streptococcal species, *S. pseudopneumoniae* excluded); v) Analysis 17sp. (17 streptococcal). The bold numbers represent the two largest supermatrices in each analysis.

Supermatrix subset	Analysis				
	20sp.	19sp.	18sp_S.ps	18sp_S.or	17sp.
Subset0	**9**	**10**	4	3	5
Subset1	3	2	4	**11**	4
Subset2	**13**	**13**	**8**	2	6
Subset3	4	4	**10**	**9**	**9**
Subset4	1	1	3	4	4
Subset5	1	1	1	1	1
Subset6	0	0	1	1	1
Subset7	0	0	0	0	1

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

Although the Analysis 20sp. and Analysis 19sp. produced a similar pattern of supermatrices, each supermatrix between these two analyses was composed of a different set of genes (Table 4.2). To analyse the consistency of the combined genes in the supermatrices, I compared the gene families in each supermatrices subset. As shown in Table 4.3, each analysis was observed to have supermatrices that consisted of different combination of gene families. Only two pairs of gene families, COG0012, COG0099 and COG0049, COG0087, were consistently combined in the same supermatrix, as shown in Table 4.3 (bold gene family ID). Additionally, the COG0525 and COG0495 gene families were consistently isolated from other gene families. The inconsistency of the 31 gene families being combined into the supermatrices made it impossible to choose which gene families were the ideal dataset with which to build the Streptococcus species tree. This may suggest that the Concaterpillar program is too conservative and largely dependent on the congruent evolutionary signals between species.

The SH test was used to evaluate the significant differences between the maximum likelihood (ML) trees resulting from all of the supermatrix subsets (Figure 4.1). The insignificant differences between the supermatrix subset trees indicated whether the gene families shared the same evolutionary histories. However, most of the compared supermatrix subset trees had different evolutionary signals (white area in Figure 4.1). Thus, it was not suitable for the 31 gene families to be combined into a supermatrix.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Table 4.3: The distribution of the gene families (COG ID) in the supermatrices. Each column represents a supermatrix analysis, and each row within a column represents a supermatrix subset. The 31 gene families (COG ID) are mapped into the supermatrices. The colour represents the set of genes in each supermatrix subset resulting in the Analysis 20sp. Bold COG IDs highlights the gene families that are consistently combined in a supermatrix.

	Analysis				
	20sp.	**19sp.**	**18sp_S.ps**	**18sp_S.or**	**17sp.**
Supermatrix subset0	**cog0012** **cog0049** **cog0087** cog0091 cog0092 cog0093 **cog0099** cog0186 cog0197	**cog0012** **cog0049** **cog0087** cog0091 **cog0099** cog0016 cog0202 cog0048 cog0052 cog0200	**cog0012** **cog0099** cog0197 cog0102	**cog0012** **cog0099** cog0081	**cog0012** cog0092 **cog0099** cog0197 cog0102
Supermatrix subset1	cog0016 cog0172 cog0522	cog0080 cog0081	cog0091 cog0016 cog0172 cog0096	cog0091 cog0092 cog0186 cog0016 cog0172 cog0052 cog0096 cog0097 cog0100 cog0184	cog0016 cog0172 cog0096 cog0184
Supermatrix subset2	cog0048 cog0052 cog0094 cog0096 cog0097 cog0098 cog0100 cog0102 cog0184 cog0200 cog0201 cog0202 cog0533	cog0096 cog0097 cog0098 cog0100 cog0184 cog0201 cog0092 cog0186 cog0197 cog0172 cog0522 cog0103 cog0256	cog0186 cog0522 cog0048 cog0098 cog0100 cog0201 cog0080 cog0256	cog0048 cog0201	cog0186 cog0048 cog0100 cog0200 cog0202 cog0533
Supermatrix subset3	cog0080 cog0081 cog0103 cog0256	cog0093 cog0094 cog0102 cog0533	**cog0049** **cog0087** cog0093 cog0052 cog0094 cog0097 cog0184 cog0200 cog0202 cog0533	**cog0049** **cog0087** cog0093 cog0094 cog0102 cog0200 cog0202 cog0533 cog0103	**cog0049** **cog0087** cog0091 cog0093 cog0052 cog0094 cog0097 cog0201 cog0103
Supermatrix subset4	cog0495	cog0495	cog0092 cog0081 cog0103	cog0522 cog0098 cog0080 cog0256	cog0522 cog0098 cog0080 cog0256
Supermatrix subset5	cog0525	cog0525	cog0495	cog0495	cog0081
Supermatrix subset6			cog0525	cog0525	cog0495

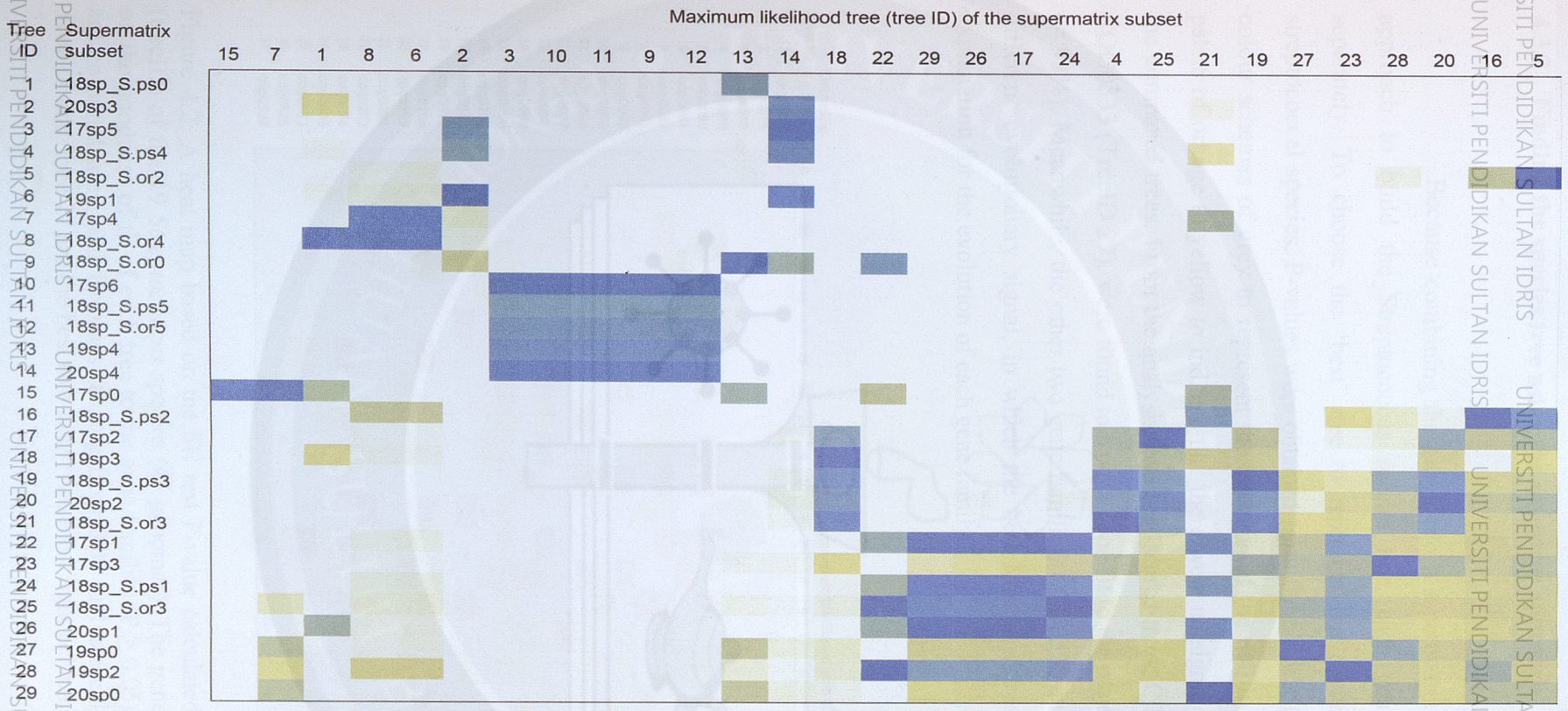


Figure 4.1: A heat map based on the SH test P-value calculated from 29 supermatrices in the supermatrix approach. Each supermatrix subset trees is marked with a Tree ID. The Supermatrix subsets are named according to the supermatrix analysis, followed by the subset number, e.g.; 18sp_S.ps0 refers to Analysis 18sp_S.ps supermatrix subset 0. The gene trees are sorted based on the goodness of fit of the trees to the set of genes in a supermatrix ($P \geq 0.05$). A P-value less than 0.05 is coloured white, while a P-value ranging from 0.05 to 1 is coloured light brown to blue. The tree in the most right order is the most accepted tree topology.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

4.3.2 Finding the species tree using the supertree approach

Because combining the genes into a supermatrix was likely not the best approach to build the Streptococcus species tree, each gene family was analysed separately. To choose the 'best' tree to represent the evolutionary histories of the streptococcal species, P-values were obtained from the SH test (Figure 4.2). I have used colour schemes of grey to represent the rejection of the null hypothesis, and the colour palate of orange to yellow to indicate that there were no significant differences between the compared trees. From the analysis, two best-scored trees, COG0172 (TreeID 18) and COG0533 (TreeID 27), were found to have a good fit to the majority of the gene families (≈ 93%). Meanwhile, the other two gene families, COG0525 and CO0495, had a totally different evolutionary signal, in which the best ML gene tree was the only possible explanation for the evolution of each gene family.

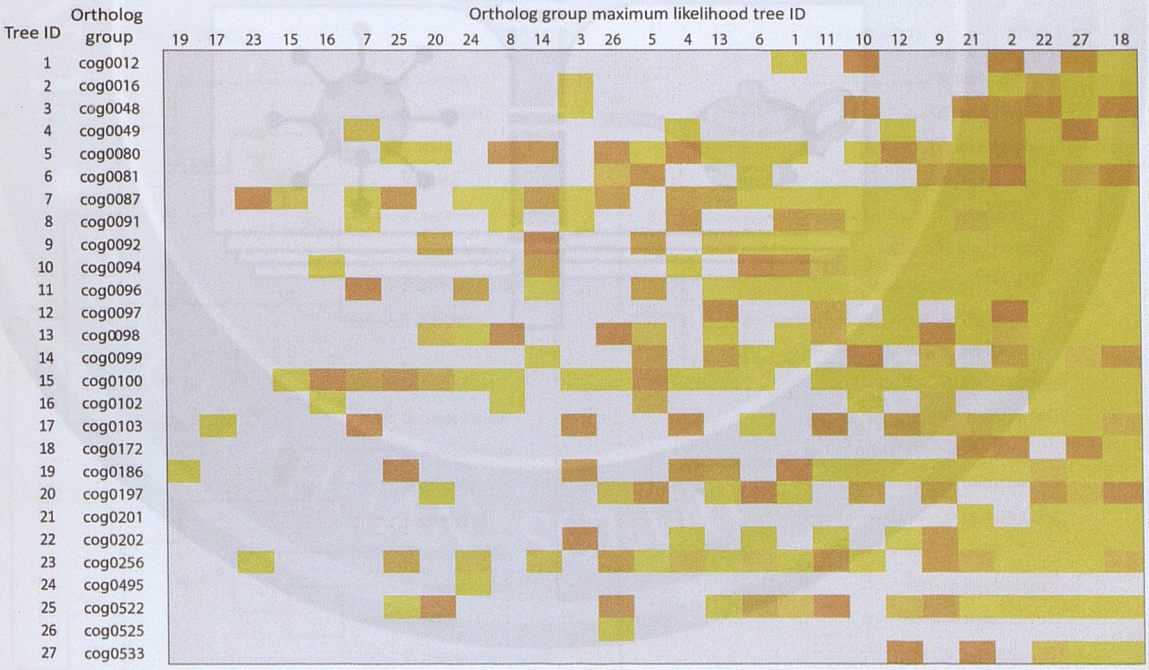


Figure 4.2: A heat map based on the SH test P-value calculated from the 27 gene trees (treeID) of the 19 *Streptococcus* species (45 genomes). The gene trees are ordered based on the goodness of fit of each tree to the gene families ($P \geq 0.05$). P-values less than 0.05 are coloured grey, and P-values ranging from 0.05 to 1 are coloured orange to yellow. Tree 18 and Tree 27 are the best two trees (two most right columns).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

The two unrooted gene trees, COG0172 (TreeID 18) and COG0533 (TreeID 27), are shown in Figure 4.3. The topological incongruences between these two trees are marked by the placement of eight species: *S. uberis*, *S. parauberis*, *S. mutans*, *S. parasanguinis*, *S. pneumoniae*, *S. oralis*, *S. mitis* and *S. pseudopneumoniae*. The conflicting branches and nodes between these two trees are depicted as red branches and black nodes in Figure 4.3. Although there is a different arrangement of *S. pneumoniae*, *S. oralis*, *S. mitis*, and *S. pseudopneumoniae* in these two trees, these four species are grouped together in a clade on both trees.

The two best trees resulting from the supertree method were rooted using outgroups. The outgroups from four distant species, *S. aureus*, *L. johnsonii*, *E. faecalis* and *L. lactis*, were used to root the phylogeny. The rooted phylogeny produced two primary clades. One clade consisted of only two streptococcal phylogenetic groups,

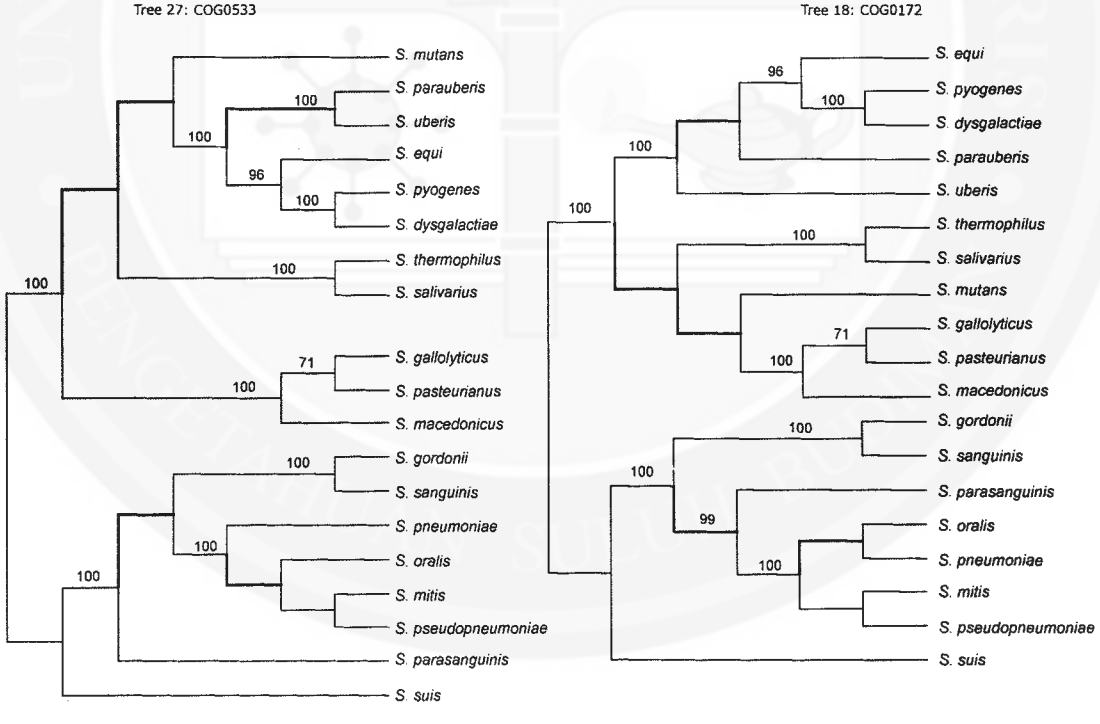


Figure 4.3: The unrooted COG0533 and COG0172 gene trees with the bootstrap values. The topological differences between these two trees are illustrated in red to indicate the conflicting branches.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

the Mitis group and the unknown *S. suis* group, which were separated from the other *Streptococcus* species. However, one of the two best ML trees had a low bootstrap value, which supports the root of the phylogeny. Therefore, to increase the confidence in the rooting, the next best-scored gene trees: COG0202 (TreeID 21), COG0016 (TreeID 2) and COG0201 (TreeID 22), from the SH test (Figure 4.2) were rooted using the same outgroups. These trees had high bootstrap values (> 85%), which supports the same rooting of the two best trees, as shown in Table 4.4.

Table 4.4: The bootstrap values support the root of the gene trees of other gene families.

Gene family	Tree ID	Bootstrap value
COG0016	2	99
COG0201	21	97
COG0202	22	86

4.3.3 A strict consensus species phylogeny

A strict consensus phylogeny (Figure 4.4) from the two best-rooted ML gene trees resulting from the supertree approach (TreeID 18 and TreeID 27). *Streptococcus* species from the same phylogenetic groups, as defined by Kawamura et al. (1995), were clustered together. All 19 *Streptococcus* species fell into one of the five phylogenetic groups, the Pyogenic, Salivarius, Bovis, Mutans, Mitis, and *S. suis,* which does not belong to any phylogenetic group and is referred to as the Unknown group. The *Streptococcus* species from the Anginosus group were not included in the analysis due to the unavailability of genome sequences.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

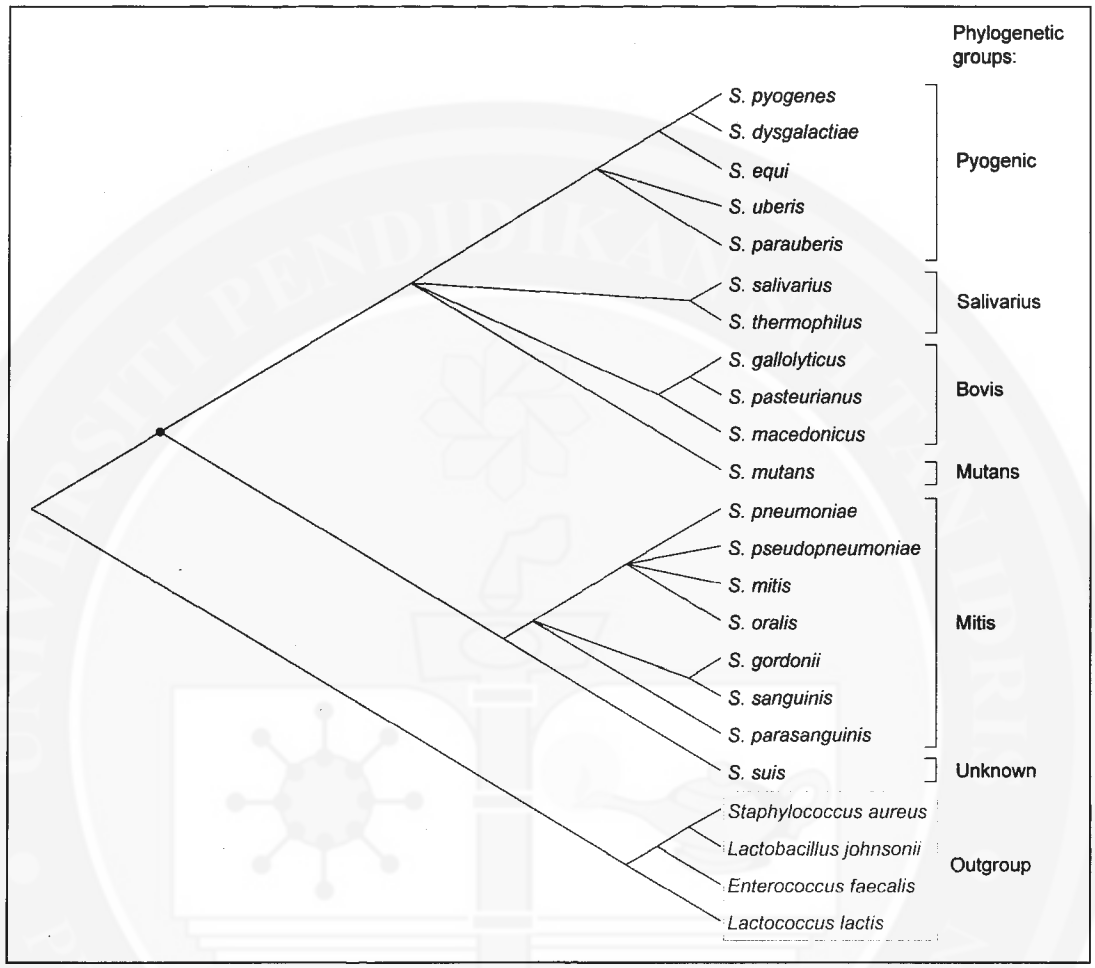


Figure 4.4: A rooted, strict consensus phylogeny of the 19 streptococcal species. The five phylogenetic groups are: the Pyogenic, Salivarius, Bovis, Mutans, Mitis, and *S. suis* from the unknown group. Four outgroup species (grey): *Staphylococcus aureus, Lactobacillus johnsonii*, *Enterococcus faecalis* and *Lactococcus lactis*, were used to root the phylogeny (black dot).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

4.4 Discussion

In the supermatrix approach, genes are concatenated and combined in a supermatrix to produce a species tree. However, it is crucial to ensure that all of the genes share similar evolutionary signals before combining them. This step can prevent any genes that were affected by duplications, HGTs, or incomplete linage sorting to be incorporated into the supermatrix (Philippe and Douady, 2003, Beiko et al., 2005, Degnan and Rosenberg, 2006). Including such 'heterogeneous' genes into the supermatrix would produce a misleading relationship between the species (Bull et al., 1993, Malia Jr et al., 2003, Kubatko and Degnan, 2007). The supermatrix approach hypothesises that all of the participating genes share one evolutionary tree (Kluge, 1989). In regard to the hypothesis, each of the gene datasets used in this study had a different evolutionary signal (Figure 4.1). Therefore, combining them may not be the best approach to construct the species tree.

In Table 4.2, COG0525, leucyl-tRNA synthetase, and COG0495, valyl-tRNA synthetase, were clearly shown to have different evolutionary signals. The Concaterpillar program detected and separated these gene families into a single supermatrix, signifying that each of the gene families cannot be analysed together with other genes. The SH test of separate analysis in the supertree approach also produced the same result (Figure 4.2), in which these two gene families have a completely different evolutionary history. The evolutionary histories of these two gene families are not supported by the other ML gene trees.

A classic approach to the supertree method is to combine all of the gene trees to produce a consensus supertree to represent a species tree (Sanderson et al., 1998). However, careful assessments are needed when deciding which gene trees should contribute to the construction of the species tree. Acknowledging the different evolutionary signals from the 27 gene families, I used the SH test to choose which trees had the best-accepted topologies to represent the evolutionary histories of the 19 *Streptococcus* species. From the test, two ML gene trees showed no significant differences from the 25 ML gene trees. This implies that these two trees offer possible explanations regarding the species histories of the 25 gene families (excluding COG0525 and

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P
COG0495). A *Streptococcus* phylogeny (Figure 4.4) was constructed from these two rooted ML gene trees using the strict consensus approach.

Although the *Streptococcus* phylogeny is not fully resolved (Figure 4.4), it consists of clades that are congruent to the *Streptococcus* phylogenetic group, as defined in a previous study (Kawamura et al., 1995). In addition, the species phylogeny is rooted to indicate the direction of the evolution between the *Streptococcus* species. The species phylogeny and the direction of the evolution could indicate how the QS systems have evolved in the *Streptococcus* species.

Some improvements can be made to produce a fully resolved species tree, e.g., by integrating multiple substitution models in the supermatrix to build a species tree (Sullivan, 1996, Yang, 1996b, Felsenstein, 2004) . However, the approach requires intense computation, and it is too complex to be applied in this study.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

CHAPTER 5

The evolution of streptococcal quorum sensing systems

Abstract

Quorum sensing (QS) is an important mechanism for controlling the behaviour of a bacterial population. Until recently, the phenomenon of bacteria working in a group like multicellular organisms was unknown. The discoveries of two homologous systems, each with a particular function in regulating bacteriocin and competence, in *S. pneumoniae* further support the complexity of QS regulation. This chapter discusses the analysis of the evolution of QS systems in *Streptococcus* species. This information could allow the discovery of the potential QS ancestor; whether QS systems were gained or lost throughout the genus; and the degree to which these shared QS regulatory genes have diverged from each other. The results show that QS systems evolved as a regulon unit through duplication events. Duplication may possibly play a major role in the functional divergence of the systems. The horizontal gene transfer was found to occur between regulon genes that regulate the same functions across the streptococcal species. This finding suggests that horizontal gene transfer promotes sequence variations in the QS regulon genes.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

5.1 Introduction

In a well-studied Com quorum sensing (QS) system in *S. pneumoniae*, two groups of genes are involved in the QS system which are called the early genes and the late genes. The early genes are responsible for inducing the bacteria cell to become competent, and the late genes will then regulate the uptake of the DNA from the extracellular environment. Competence is induced by a small peptide pheromone (≈17 amino acids) called the Competence Simulating Peptide (CSP), which is encoded by the *comC* gene (Havarstein et al., 1995). The peptides are transported and matured through an ABC transporter comprising an ABC-exporter (*comA*) and an accessory protein (*comB*). A two-component system (TCS) comprising a histidine kinase (*comD*) and a response regulator (*comE*), detects the matured CSPs and transmits signals to *comX* (Pestova et al., 1996, Lee and Morrison, 1999). The sigma factor protein, ComX, binds at the *cin*-box promoter (TACGAATA) of the late genes, which leads to the uptake and processing of extracellular DNA (Lee and Morrison, 1999).

In *S. mutans*, unlike *S. pneumoniae*, recent studies found that the activation of *comX* does not directly result from the TCS signals (Martin et al., 2006, Mashburn-Warren et al., 2010). The naturally transformable *S. mutans* uses another effector, *comRS*, to activate the transcription of *comX* (Perry and Kuramitsu, 1981, Mashburn-Warren et al., 2010). The *comRS* gene is also known as a Rgg regulator. Species in the Salivarius and Bovis groups were later discovered to be naturally competent via the actions of the same type of regulator (Ibrahim et al., 2007, Fontaine et al., 2010, Morrison et al., 2013). The ability of a competent cell to naturally transform is unpredictably sporadic in the *Streptococcus* genus (Johnsborg et al., 2007).

Despite the different pathways of *comX* activation, the regulon needs an ABC transporter, the peptide pheromones, and a TCS to induce the cell to become competent. A study showed that *S. pneumoniae* has two homologous QS regulons: a Com QS regulon that regulates competence, and a bacteriocin-like peptide (Blp) regulon that regulates the production of bacteriocin (de Saizieu et al., 2000). Both functions are widespread in the *Streptococcus* genus. The Com QS genes found in *S. mutans* have more similar features, i.e., sequences similarities and similar genes organisation, to the *S. pneumoniae* Blp system (Martin et al., 2006). However, the history of these regulons remains obscure.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

If these QS regulons share an ancestor, gene duplication may be a driving force behind the functional divergence of these gene families. It could be infered that the QS ancestor of the *Streptococcus* species has the competence function, but the function was duplicated and diverged into a new/sub-function, or was subsequently lost in the lineage.

Three possible outcomes of duplicated gene have been proposed by Ohno (Ohno, 1970, Taylor and Raes, 2004, Hahn, 2009): 1) the original copy is maintained with the ancestral function and the duplicate copy acquired a new function (neofunctionalisation (Force et al., 1999)); 2) the ancestral function splits, where each copy attained parts of the function (subfunctionalisation); and 3) both copies are maintained with the ancestral function (gene conservation), e.g., to increase gene activity.

Many models attempt to apply these outcomes based on their relation to genetic frequency and natural selection factors. The challenges lie upon how to characterise the experimental results based on these outcomes, without knowing the state of the pre-duplicated gene.

Recent co-option models, such as the Innovation, Amplification, Divergence (IAD) model (Bergthorsson et al., 2007) and the Escape Adaptive Conflict (EAC) model (Hughes, 1994, Des Marais and Rausher, 2008), suggest a blurred distinction between sub- and neofunctionalisation on the outcomes of duplicated genes. In the IAD model, the selective advantage of a minor function (likely to be initially neutral) in the ancestral gene arises as a novel function (neofunctionalisation). The frequency of the ancestral gene increases as the minor function becomes significant in the population. Divergence and mutation occur to improve the performance of the minor function. However, the original function of the ancestor is possibly maintained in the population when the minor function has reached optimisation.

In contrast, in the EAC model, a new function is assumed to arise in the ancestral gene. The multifunctional gene was duplicated, and each copy evolved and specialised to improve these functions. A well-known example of EAC, a single copy of the lens crystallins in vertebrate eyes, also functions as metabolic enzyme-gene sharing (Piatigorsky, 1998). To escape such an adaptive conflict, the ancestral gene duplicated and

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

evolved, with each of the duplicate copies specialising in one function (subfunctionalisation).

In the Duplication, Degeneration, Complementation model (Force et al., 1999), such selective adaption is unnecessary. The duplicated copies of the ancestral gene are altered through degenerative mutations (genetic drift). The mutations cause each paralog to lose function. Thus, both copies complement each other to function properly (subfunctionalisation) (Force et al., 1999).

The uptake of extracellular DNA is regarded as a horizontal gene transfer (HGT) mechanism. Therefore, it is commonly assumed that the spreading of competence ability in the streptococcal species is gained in the independent lineage through the HGT. Therefore, it is commonly assumed that the spreading of competence in streptococcal species was acquired in an independent lineage through HGT. Many studies in prokaryotes have found that HGT and gene loss are the two major evolutionary processes, with HGT known to occur among *Streptococcus* species (Lorenz and Wackernagel, 1994, Balsalobre et al., 2003, Hao and Golding, 2006). Within each host, streptococci are found to colonize diverse habitats, including the pharynx and respiratory tract, and intestinal and urogenital tracts (Madigan et al., 2008).

The lack of consistent nomenclatures in naming QS genes has given the wrong impression about the paralogous systems. In *S. pneumoniae*, the two homologous QS systems each regulate the Com and Blp systems (de Saizieu et al., 2000, Martin et al., 2006). While in *S. mutans*, one QS system seems to regulate both functions (Perry and Kuramitsu, 1981, Kreth et al., 2005, van der Giezen et al., 2005). Studies showed that *S. mutans*, which is a naturally transformable species, has Com QS genes that are more similar to the Blp QS genes in *S. pneumoniae* (Kreth et al., 2005, Martin et al., 2006).

Using the QS homologs from *Streptococcus* species collected in Chapter 2, and the *Streptococcus* species phylogeny derived in Chapter 4, I examined the history of these homologous QS systems and the selective forces that led to their maintenance and diversification. More specifically, the study was interested in determining whether related streptococcal species use the same QS regulatory cascade to regulate these functions. This information will tell us which QS system is the ancestral; whether the Com QS systems

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

were gained or lost throughout the genus; and the degree to which these shared QS regulatory elements have diverged from each other.

5.2 Materials and Methods

5.2.1 The QS homologs dataset from the 86 genomes of *Streptococcus* species

This study employed 708 QS homologs, described in Chapter 2, as the datasets. These 708 QS homologs were detected in 86 fully sequenced genomes of 21 *Streptococcus* species using the SUCRE methodology. Forty-four protein queries from six *Streptococcus* species were used as inputs in SUCRE to initiate the homology search. All of the protein queries (Appendix A) and genome sequences (Appendix B) were retrieved from the NCBI Genome database (June 2011).

5.2.2 Mapping the regulon structures of the QS homologs

I mapped the locations of all the QS homologs in the 86 genomes (Appendix E). For a better mapping visualisation, I only took the representative homologs from all genomes of each species, as illustrated in Figure 5.1(a).

As shown in Figure 5.1(a), the representative homologs of the S1 species regulon structure (rep S1) is from the union set of two genomes (S' and S''). A complete synteny map of the QS homologs present in 86 genomes is shown in Appendix E. Since there was not much variation in the synteny of the present homologs in a species, the average location distances between the neighbouring homologs in all genomes of the species were calculated. The average location distances were then used to map the regulon gene structures of QS homologs in the 21 *Streptococcus* species.

Homologs that have less than 2Kb location distance between each other were classified to be within a regulon (Figure 5.1(a)). Using the regulon structure, I manually identified the missing homologs, which are mostly the peptide pheromone (*C* gene). Due to the short lengths of the *C* genes, they are not easy to identify with an automated

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

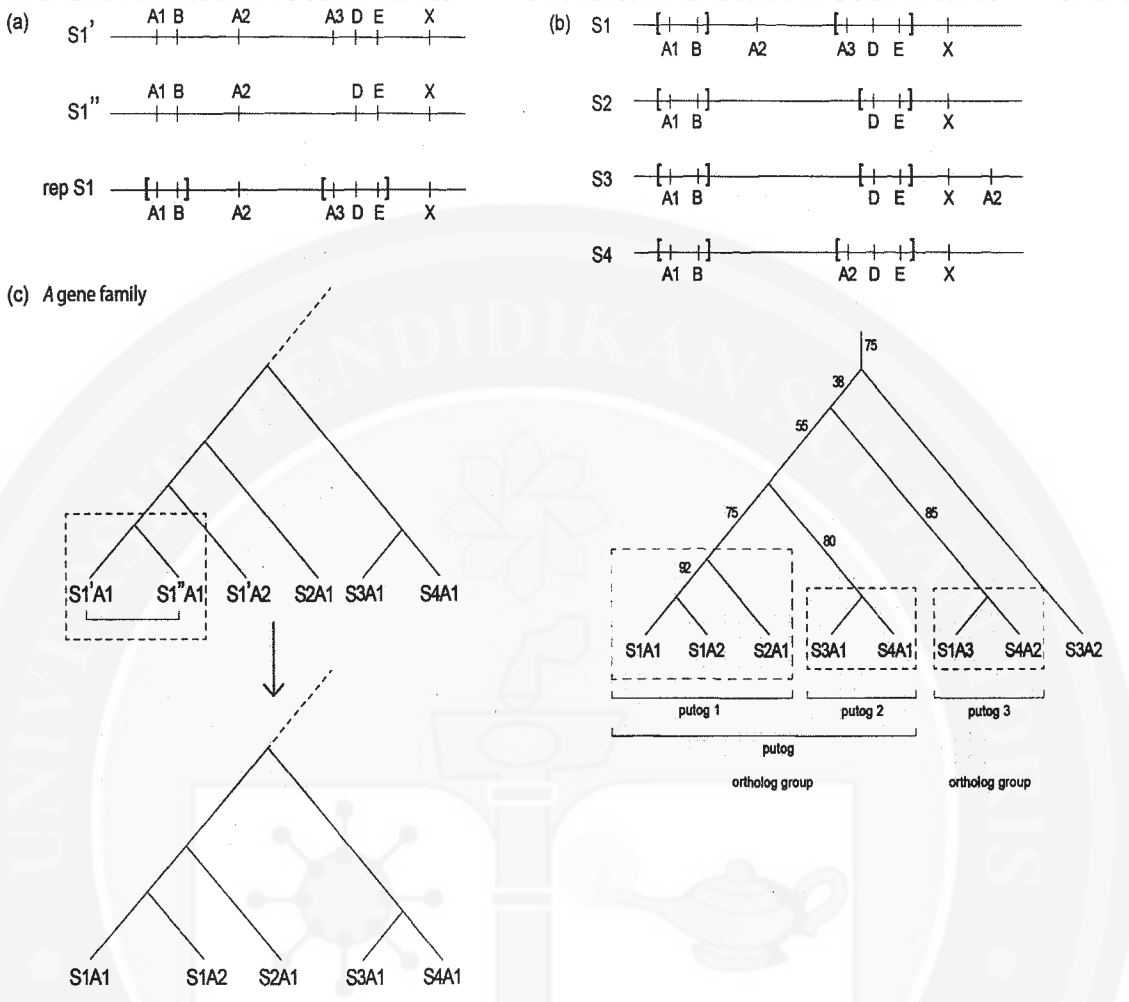


Figure 5.1: Methods used to map the QS regulon structure and infer orthologs for streptococcal species. (a) Mapping the regulon structure. The union set of regulon homologs from each strain's genome (S', S'') were mapped as the representative (rep) regulon structure of S1 species; (b) The regulon structure of each species. The parentheses [ ] indicate the regulon structure of homologs' that are less than 2 Kb apart in the rep genome (S1, S2, S3, S4) species; (c) Clustering species to form orthologous groups. From the left, a node that consists of a species with multiple copies of homologs with the same regulon structure (refer 5.1 (a): S1'A1, S1''A1) is collapsed into the species rep (S1A1). The tree on the right shows how a putative orthologous group (putog) is formed based on the clade(s) with high bootstrap support (>70%) that shares the same regulon structure. Putogs with the same the regulon gene structure in an ancestor are merged. Each finalised putog is considered as an orthologous group.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

procedure. However, they are usually associated with the TCS (*DE* genes) and can be manually identified using this information. These homologs were then mapped with the other homologs in the genomes. The synteny all of these homologs was later used to differentiate the homologs into orthologs and paralogs.

5.2.3 Estimating the maximum likelihood gene trees

Each of the *ABCDEX* gene family sequences was aligned using the Muscle program (Edgar, 2004). A best Maximum Likelihood (ML) gene tree was constructed for each gene family using the RAxML program (Stamatakis, 2006b) under the WAG+ Γ substitution model (Whelan and Goldman, 2001). The same set of parameters was used in the non-parametric bootstrap re-sampling for each gene family to produce one thousand ML trees. The purpose of bootstrap testing is to assess confidence in the gene trees.

5.2.4 Inferring QS orthologs

One approach that is commonly used in inferring orthologs is to reconcile a species tree and a gene tree to find incongruent topology that will help to explain the evolutionary process that had taken place. The most important limitation of the standard reconciliation programs is that they require a fully resolved and absolute confident species and gene trees. Although some of the reconciliation programs (e.g., Nutong) allow either of the trees to be an unresolved tree, the program does not consider the HGT process in inferring the orthologs and paralogs. To overcome this limitation, I developed the following steps to infer orthologs and paralogs. Note that in this analysis the *Streptococcus* species phylogeny obtained in Chapter 4 was used. Because *S. agalactiae* and *S. infantarius* were not present in the species phylogeny, I removed the QS homologs of both taxa from the gene trees.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

The following steps involved in inferring QS orthologs are described below:

a) Testing the congruence between the gene trees and the species tree in explaining the evolutionary relationships between Streptococcus species

A ML species tree-like was constructed for each gene family using RAxML (Stamatakis, 2006b), with the tree search options constraint to the species phylogeny topology. I used the SH-test implemented in the CONSEL program (Shimodaira and Hasegawa, 2001) to analyse the disagreement between each gene tree and the ML species tree. The per-site log-likelihoods were computed via RAxML (model optimisation precision [-e] was set to 10e-6) for all testing trees of each gene family for the CONSEL input. The SH test was used to measure whether there was a significant difference between each gene tree and its species tree-like. A gene family that failed to reject the null hypothesis (P-value <0.01) is more likely to consist of orthologous genes and assumed to be consistent with the species phylogeny. Otherwise, the gene family was further analysed in the next step to infer the orthologous groups.

b) Clustering species to form ortholog groups based on tree topology and regulon structure

To form the ortholog groups, gene families for which the species trees was rejected (in the previous step) were rooted using the outgroups. In this study, an ortholog group is defined as a group of species with the same regulon gene synteny information that was derived from a common ancestor through a series of speciation events. I used four species, *Lactobacillus lactis*, *Enterococcus faecalis*, *Lactobacillus johnsonii* and *Staphylococcus aureus*, as outgroups. These outgroups species fall in the same taxonomic class, Bacilli, as the *Streptococcus* species. The homolog sequences of the gene families, for which the species tree was rejected, were obtained from the NCBI Genbank. Note that any homolog falling outside the outgroups was removed.

For each gene tree, I collapsed nodes that had multiple copies of a species to represent the homolog of the species itself (Figure5.1 (b)). I then identified the clades of species with greater than 70% bootstrap values as the cut-off point to mark high bootstrap

support (Hillis and Bull, 1993). I took the union set of species in the overlapping clades (>70%) that had the same regulon gene synteny information to form a putative ortholog group (putog). Putogs that share an ancestor, given that they have the same regulon structure information, were merged into a single putog. This step automatically assigned the species with no regulon information, as well as species that do not belong to any putogs due to the low bootstrap support, into the merged putogs of the same ancestor. Each of these putogs was considered as an orthologous group. Species that could not be assigned to any of the orthologous groups were categorised as a dubious genes (dg) cluster.

c) Testing the goodness of fit of the ortholog groups' subtrees to support evolutionary relationships between Streptococcus species

To determine whether there was a significant difference between each subtree and the species tree topology, the subtrees were processed using the same method as previously described in the congruency-testing step 5.2.4 a). The Ape package (Paradis et al., 2004) in the R language (R Development Core Team, 2007) was used to trim and manipulate the subtrees from the ML gene trees for this analysis. A ML species subtree-like was generated for each orthologous group using the WAG+ Γ substitution model (Whelan and Goldman, 2001) in the RAxML program (Stamatakis, 2006b), with the constraint of the streptococcal species phylogeny. The congruence between a subtree and its species subtree-like was conducted using the SH test via CONSEL program (Shimodaira and Hasegawa, 2001). If there was a significant difference between these two trees (P-value < 0.01), I assumed that a duplication, HGT or recombination event may have occurred somewhere along the lineage after the speciation event to produce an ortholog group.

d) Testing gene trees for the occurrence of duplications, horizontal gene transfers and recombination events

All the gene trees from the previous step were further analysed for the occurrence of duplications, HGTs and recombination events. Each gene tree was reconciled to the

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

species tree using the Notung program (Chen et al., 2000) to infer the duplications and loss events. The duplication events between ortholog groups (out-paralog) and within an ortholog group (in-paralog) that were inferred by Nutong were marked onto the gene tree (Appendix H). I identified the discordance between the gene tree branches and the species tree topology. As illustrated in Figure 5.3, only the discordant branches with high bootstrap support (>70%) are depicted in thick black branches. All the trees were viewed using Dendroscope (Huson et al., 2007) and Figtree (http://tree.bio.ed.ac.uk/software/figtree/) programs.

5.2.5 Inferring the presence and absence of the QS regulon to understand the evolution of QS systems in *Streptococcus* species

To infer the evolutionary history of streptococcal QS systems, the presence and absence of the QS system were recorded across the *Streptococcus* species. The Dollo parsimony method (Farris, 1977) was used to infer the gain and loss of a QS system by allowing the QS system to be derived only once in the lineage and by minimising the number of losses of the QS system. However, the method does not include the chance of independent gain via HGT, which could have affected the inferred evolutionary scenario. HGT is known to be a common mechanism in bacteria. Thus, I incorporated the evidence of non-vertical descendent, obtained from the previous reconciliation approach in 5.2.4 d), in the most parsimonious evolutionary scenario of the QS systems inferred from the Dollo method. The GC content of the HGT homologs was calculated to provide evidence for unusual GC content from the genome's GC content and to support the non-vertical descendent.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

5.3 Results

5.3.1 The regulon structure of the QS homologous genes across *Streptococcus* species

The ABCDEX homologs were mapped to their synteny in the genomes. A pan-genome composed of the QS homologs was used to represent each species with multiple genomes (Appendix E). Most of the genomes of these species have similar synteny of the identified QS homologs, especially the group of homologous QS regulon.

As illustrated in Figure 5.2, the homologs are coloured according to their gene family: A (green), B (black), C (yellow), D (red), E (blue) and X (white). Three QS regulon structures patterns were discovered: i) Com QS (AB-EDC), ii) Blp QS (EDCBA) and iii) Fas QS (DDE). The homologs within these regulon structures were located in less than 100bp adjacently (parentheses in Figure 5.2). The colours of the parentheses indicate the different types of regulon structure based on the ortholog sets in each of the regulon structures, which will be discussed in detail in the next section.

The Com QS and Blp QS systems were found present in the *S. pneumoniae* genome (Figure 5.2). Both QS regulons have similar structure to the QS systems that were discovered in the previous studies (de Saizieu et al., 2000, Martin et al., 2006). The Blp QS regulon structure was found to be widespread in the six *Streptococcus* species, *S. pyogenes, S. dysgalactiae, S. equi, S. thermophilus, S. pneumoniae, S. pseudopneumoniae* and *S. suis*. While the Com QS regulon pattern was observed only in the Mitis group species.

Another regulon structure (DDE) that is similar to the Fas system, which regulates the expression of Streptokinase (SK) genes (Malke and Steiner, 2004), was present in the Pyogenic group and the *S. salivarius* genome. The Fas system requires the *fasX* gene to effectively transcript the virulence factor SK (Kreikemeyer et al., 2001, Ramirez-Pena et al., 2010). The *fasX* gene is commonly found adjacent to the DDE regulon.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

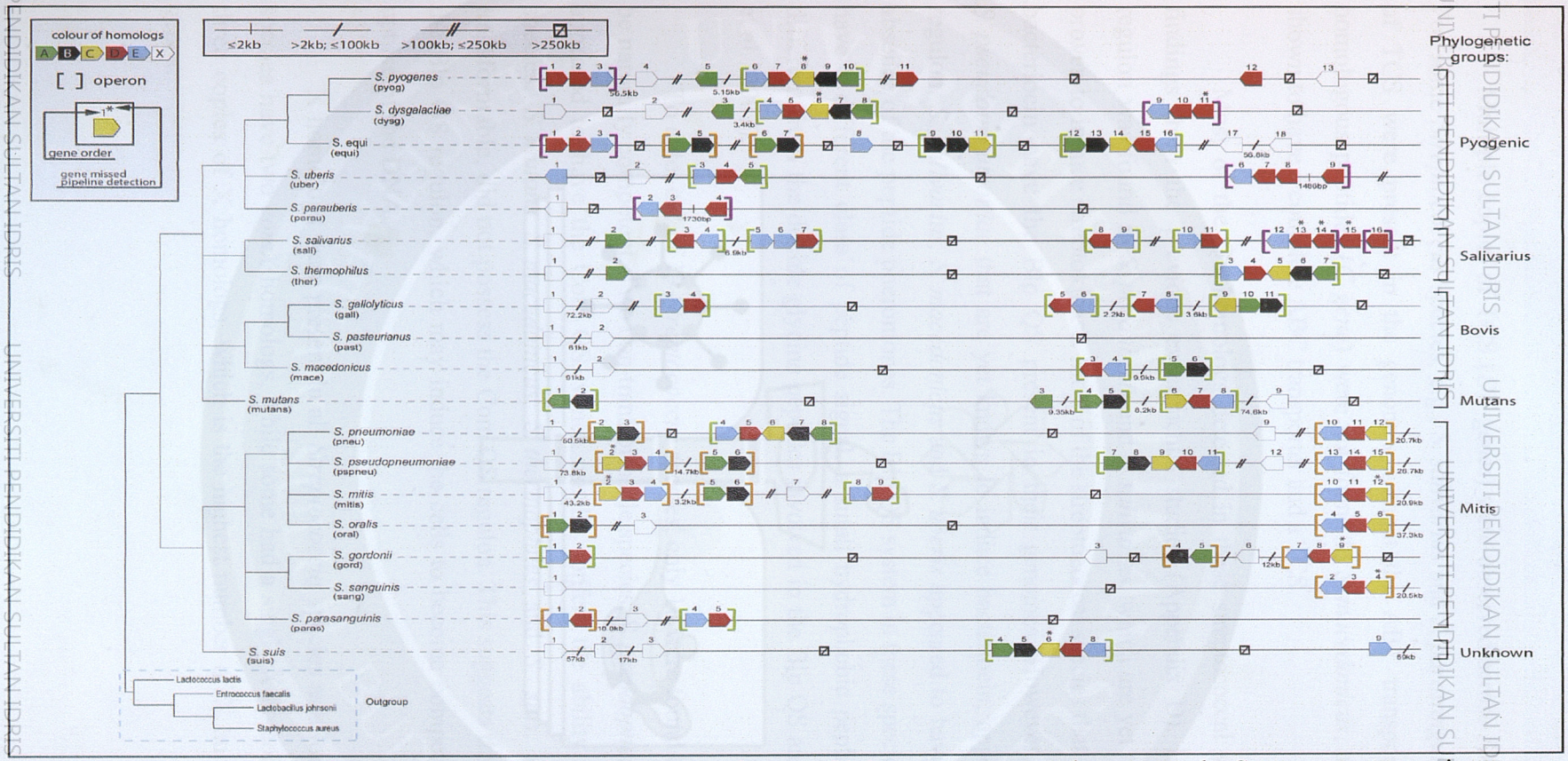


Figure 5.2: The regulon structure of Streptococcus QS homologs mapped across 21 Streptococcus species next to the Streptococcus species tree. Each QS homolog, which is indicated by a coloured arrow: A (green), B (black), C (yellow), D (red), E (blue) and X (white), is mapped to its synteny in the genome. The homologs within coloured parentheses are separated by less than 2 Kb. Parentheses are coloured based on the unique orthologs set, which indicate Blp (green), Com (orange) and Fas (purple).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

Multiple copies of TCS (*DE* genes) were observed across *Streptococcus* species. For some *Streptococcus* species, e.g., *S. salivarius* and *S. parasanguinis*, multiple copies of TCS were present in the genomes without the ABC transporter homologs. The homologous peptides (*C gene*) were not found in *S. macedonicus*, although its QS was shown to have a complete ABC transporter and a TCS.

Most of the QS homologs detected in these streptococcal species matched the findings from the previous experimental studies (Appendix D). However, a Fas QS regulon was found in *S. parauberis* and *S. salivarius*, which are yet to be experimentally proven to induce streptokinase. Most of the streptococcal species that are known to induce bacteriocin have the Blp QS present in their genomes. However, *S. gallolyticus* and *S. pseudopneumoniae* that are yet known to induce bacteriocin, appear to have Blp QS regulon. *S. salivarius, S. macedonicus* and *S. uberis* appeared to have a Blp QS with the absence of the peptide pheromones. The Blp QS system in these species was discovered to have a different class of peptide signal, a small hydrophobic peptide (SHP), which is different from the double glycine peptides involved in the Blp QS and Com QS (Ibrahim et al., 2007).

To date, out of 19 *Streptococcus* species, only 10 were found to have the ability to be naturally competent. Two non-transformable species, *S. pseudopneumoniae* and *S. equi*, appeared to have the Com QS, which usually confers the ability to transform. Some species, *S. mutans S. macedonicus*, *S. thermophiles* and *S. suis*, which are naturally transformable, did not possess the Com QS regulon. This suggests that these species have the Blp QS regulon, which may induce the cells to become competent and regulate the uptake of the DNA.

X homologs were present in all *Streptococcus* species as a lone gene. Most of the species have two *comX* homologs, while some had a single X homolog copy. *S. suis* has three copies of X homologs, which is the highest number of X homolog copies in this genus.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

5.3.2 The evolution of QS systems as one regulon unit

To infer the function of QS homologs, a clear distinction between orthologs and paralogs is crucial. This was done by inferring duplication and speciation events, where gene trees were compared to a species tree. If there was no significant difference between the gene tree and the species tree, then the gene tree had no statistical support for an ancient duplication. In other words, the genes evolved through a vertical descendent and are orthologous to one another.

Table 5.1 (a) shows the relative statistical support for the ML gene trees and the species phylogeny for QS genes (ABCDEX). Only the C and X gene trees were not significantly different from the *Streptococcus* species tree, which probably due to short gene length. Meanwhile, I found that the ABDE gene trees were significantly different and were deviated from the species phylogeny tree with a P-value of 0.001. This has led me to infer that the evolution of these gene families was probably shaped by the occurrence of gene duplication, HGT or recombination event, which complicates the evolutionary relationship between the streptococcal species. The gene families for which the species tree was strongly rejected were further analysed by rooting each of these family gene trees to identify orthologous groups (for details, see Materials and Methods).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

(a)

Testing trees	Gene family					
	A	*B*	*C*	*D*	*E*	*X*
Gene tree	0.999	0.999	0.983	0.999	0.999	0.506
Species tree_like	0.001	0.001	0.017	0.001	0.001	0.494

(b)

Gene family	Testing sub-tree	Ortholog group			
A		1	2		
	Gene sub-tree	0.957	0.981		
	Species sub-tree_like	0.043	0.091		
B		1	2		
	Gene sub-tree	0.477	0.831		
	Species sub-tree_like	0.523	0.169		
D		1	2	3	4
	Gene sub-tree	0.018	0.516	0.999	0.724
	Species sub-tree_like	0.982	0.484	0.001	0.276
E		1	2	3	
	Gene sub-tree	0.910	0.999	0.842	
	Species sub-tree_like	0.090	0.001	0.158	

Table 5.1: Statistical support (P-value) for the SH test results. (a) the ML gene (*ABCDEX*) trees and the species phylogeny, (b) the sub-tree of the orthologous groups to the species phylogeny. Black P-values indicate gene trees and sub-trees that are not significantly different, at a level of 0.01, to the species phylogeny; otherwise the P-values are in grey. The colours of the ortholog group numbers in (b) represents the regulon structure: Fas (purple), Blp (green) and Com (orange), to which each ortholog belongs to as mapped in Figure 5.2. The same colour of the orthologous group label (b) in each gene family indicates that these orthologs are found within a regulon.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

As shown in Table 5.1 (b), only two of the orthologous groups in the *D* and *E* gene families have a significantly different evolutionary signal from the species tree (P = 0.999). Both ortholog groups have different phylogenies than the species tree, with most of the branches having low bootstrap supports. The rest of the orthologous groups have similar evolutionary relationships to the *Streptococcus* species phylogeny. The All of the nodes in the gene trees of these ortholog groups were inferred to be duplication events, suggesting an out paralogy relationship. This further supports the ortholog group clustering as the species within the ortholog clustering are derived from the last common ancestor through speciation.

A similar pattern of orthologous groups has been discovered in the ABC transporter regulator, *A* and *B* gene families, and in the TCS regulator between the *D* and *E* gene families. Only three regulon sets of the *ABCDE* orthologs were found across the 19 *Streptococcus* species. The mapping distribution of these three regulons across the *Streptococcus* species is illustrated in Figure 5.2. These three QS regulon structures were found to belong to the Blp QS (*ABCDE*: green parentheses), Com QS (*AB-EDC*: orange parentheses) and Fas QS (*DDE*: purple parentheses) system. Some of these QS regulons consist partial of the ortholog members.

The ML gene trees of the *ABDE* gene family are shown in Figure 5.3. All of the orthologous genes in each QS regulon were derived from speciation event of an ancestor. The statistical support, as shown in Table 5.1 (b), for the *ABDE* orthologs set of each QS regulon suggests that these QS regulons were functionally diverged and adapted as a regulon unit.

However, the study found some evidence for the duplication of individual genes, which are the lone genes. Six duplicated lone genes belong to the Blp QS regulon in *S. pyogenes, S. dysgalactiae*, *S. equi* and *S. uberis* (black taxa name shown in the green boxes of the gene trees in Figure 5.3). These lone genes were not tandemly duplicated and are present in different parts of the genome (Figure 5.4).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

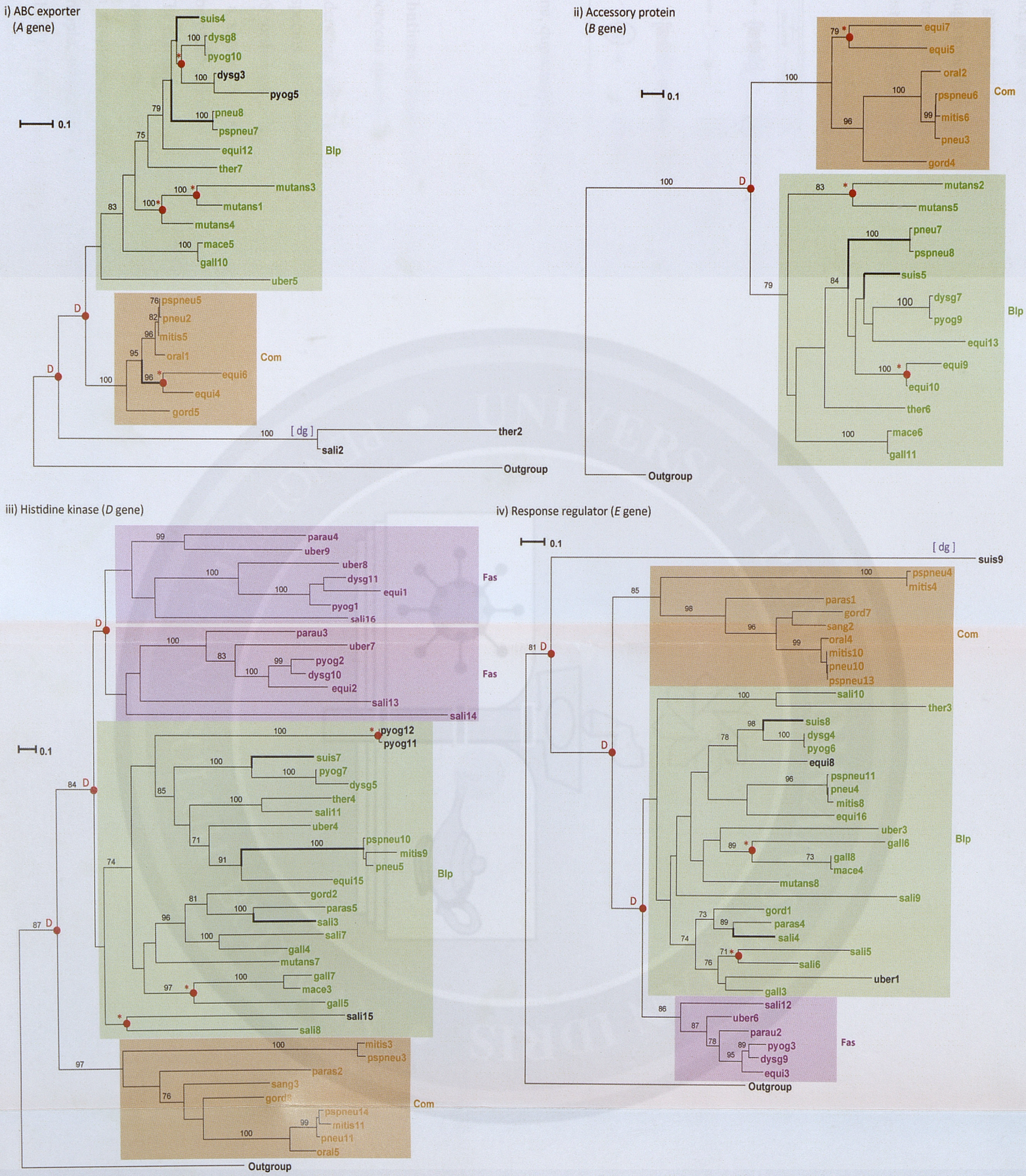


Figure 5.3: The rooted ML gene trees of: i) ABC exporter (A gene), ii) accessory protein (B gene), iii) histidine kinase (D gene), and iv) response regulator (E gene). Bootstrap support values > 70% are presented. The orthologs of each QS regulon are coloured and boxed accordingly: Blp (green), Com (orange) and Fas (purple). Duplication events are marked with a red dot, and with a capital D (ancient duplication) or an asterisk (recent duplication). Thick black lines represent incongruent gene tree branches to the species phylogeny with > 70% bootstrap support. All taxa are named according to their species (species rep name as shown in the species phylogeny in Figure 5.3), followed by the gene order in the genome (refer to the genome map in Figure 5.3). Taxa names in black indicate lone genes. Lone genes that do not belong to any of the regulon are marked'[dg]'.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

In the ML gene trees in Figure 5.3, there were also a few of the *blpABDE* genes and two *comA* genes that have different evolutionary histories (taxa with black thick line branch). Although these QS regulons were found to have evolved as a single regulon unit, the evidence for duplicated lone genes and the observation of some genes with different histories suggest some flexibility in the QS regulon systems.

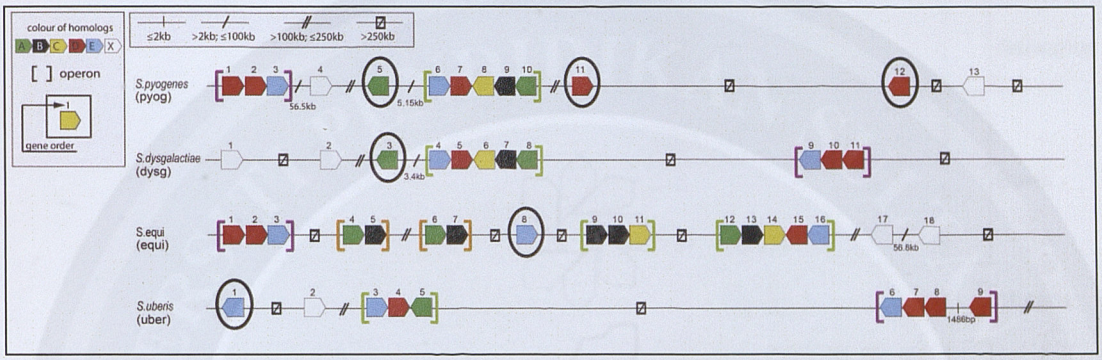


Figure 5.4: Lone, duplicated genes (circled arrows) mapped in the genomes.

5.3.3 Contributions of duplications to the functional divergence of QS systems in *Streptococcus* species

To understand the evolution of the QS systems (Blp, Com and Fas) in streptococcal species, as well as their relation to each other, I mapped the presence and absence of QS orthologs across 19 *Streptococcus* species next to the streptococcal phylogeny. If a species has at least one ortholog of a QS regulon, the QS system is considered to be present. This is made upon the evidence that the QS system evolved as a regulon unit. The presence and absence of the Blp, Com and Fas QS regulons were used as the inputs to construct the most parsimonious evolutionary scenario for the evolution of QS systems in streptococcal species (Figure 5.5). For this construction, I applied Dollo parsimony principle, which allows a QS regulon to be acquired once, but lost multiple times (Farris, 1977). The parsimonious scenario is the one with the minimum number of losses.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

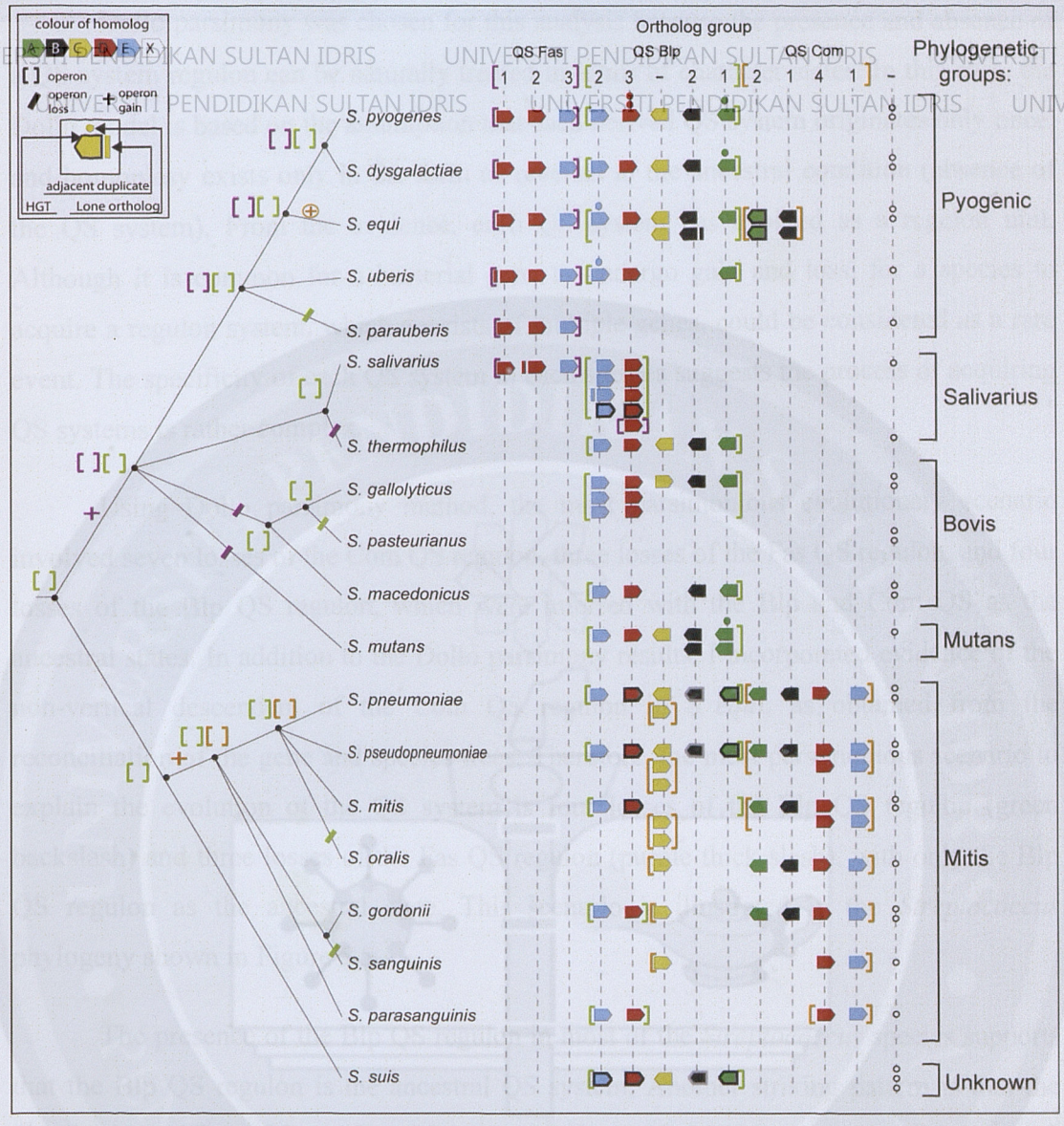


Figure 5.5: The 19 streptococcal early QS homologs and the streptococcal species tree (left). Each QS homolog, indicated by coloured arrow, A (green), B (black), C (yellow), D (red), E (light blue) and X (white), is mapped to its orthologous group (dotted lines with the orthologous group number on the top). The parentheses indicate a QS regulon structure that consists of homologs found to be next to each other (<2 Kb apart). The colour of each parenthesis represents the unique set of orthologs commonly found across 19 *Streptococcus* species. Only one orthologous group was found in the C and X gene families, which is marked by "-". The small coloured circle above the arrows (of the orthologs) shows the lone orthologs that do not fall within any of the regulon structures. The rectangular box at the tail of the arrow shows duplicated adjacent orthologs. The arrows with the thick black/grey borders show the orthologs that have undergone HGTs. Each arrow represents the direction of the ortholog within a regulon. The Dollo parsimony of gain (+) and loss (thick slash) of the regulon structure is illustrated on the species tree. The symbol '+' within a circle shows independent gain via HGT. Each node (black dot) on the species tree shows the inferred regulon structure of the ancestral state.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

Dollo parsimony was chosen for this analysis because the presence and absence of a QS system regulon can be naturally treated in terms of character states. In this case, the Dollo model is based on the assumption that each derived QS system originates only once, and homoplasy exists only in the form of reversal to the ancestral condition (absence of the QS system). From the evidence, each QS system has evolved as a regulon unit. Although it is common for a bacterial gene to undergo gain and loss, for a species to acquire a regulon system, which consists of multiple genes, could be considered as a rare event. The specificity of each QS system in each species suggests the process of acquiring QS systems is rather complex.

Using Dollo parsimony method, the most parsimonious evolutionary scenario involved seven losses of the Com QS regulon, three losses of the Fas QS regulon, and four losses of the Blp QS regulon, which were inferred with the Blp and Com QS as the ancestral states. In addition to the Dollo parsimony results, I incorporated evidence of the non-vertical descendent of the Com QS regulon in *S. equi*, as obtained from the reconciliation of the gene and species trees. Therefore, the most parsimonious scenario to explain the evolution of the QS system is four losses of the Blp QS regulon (green backslash) and three losses of the Fas QS regulon (purple thick slash), with only the Blp QS regulon as the ancestral state. This scenario is illustrated in the *Streptococcus* phylogeny shown in Figure 5.5.

The presence of the Blp QS regulon in most of the *Streptococcus* species supports that the Blp QS regulon is the ancestral QS system. Another striking pattern is that the distribution of Com QS and Fas QS were cleanly divided into two clades. The Com QS was only found in species of the Mitis group, while the Fas QS was distributed among species in the Pyogenic group and *S. salivarius* species.

5.3.4 Horizontal gene transfer contributes to the evolution of QS systems in *Streptococcus* species

Identifying HGTs can be done through a phylogenetic approach such as the reconciliation between gene trees and the species tree. Conflict between these compared trees could indicate the gene that have derived non-vertically descendent. Additionally,

UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

HGTs can be directly assessed using the genomic context. A gene that has unusual genomic properties, e.g., GC content, for its genome is suspected to derive from a different origin.

From the inferred parsimonious evolution of the QS systems in Figure 5.5, only the Com QS regulon was obtained through HGT in *S. equi*. From the reconciliation between the gene trees and the species phylogeny, only the *comA* gene families of the ABC transporter show evidence of non-vertically descent (Figure 5.3). In the *A* gene tree, a strong bootstrap support of 95% that cluster *S. equi* within the group of Mitis species was found. Although there is no evidence for non-vertical descent in the *B* gene tree, the subtree of *comB* has the same taxa that paired together with the subtree of *comA*, which shows that these genes work and evolve as a regulon unit.

Further comparisons were made regarding the GC contents between each ABC transporter gene and the genome GC content. Both the *comA* and *comB* homologs showed a low GC content compared with the GC content of the *S. equi* genome (Table 5.2). The GC content of *comA* (35.20%) and c*omB* (32.00%), that were thought to have arisen from HGT, differ from that of the *S. equi* genome (41.55%). The difference between these genes and the genome GC content supports the view that these genes were possibly obtained via HGTs from species in the Mitis group. The GC contents of the *comA* and *comB* homologs in *S. equi* were also found to be much lower than that of *comA* (≈40–44%) and *comB* (≈41–45%) in other streptococcal species. Thus, HGT can affect GC content and promotes the sequence diversity in QS regulon genes.

From the reconciliation of the gene trees and the *Streptococcus* species tree, there are some evidence for non-vertical descents, which may also be due to HGT events. This evidence was found to occur between the regulon genes of the Blp QS regulon, as shown on the conflict gene tree branches with >70% bootstrap support. As shown in Figure 5.3, the Blp QS ABC transporter (*blpAB*) of *S. pneumoniae* and *S. suis* were found to have a closer relationship to the species in the Pyogenic group. A similar case was observed in the TCS (*blpHR*) of the Blp QS in both species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

Table 5.2: GC content of the AB transporter genes of streptococcal species.

Species	Genome %GC	%GC ABC exporter (*A* gene)		%GC Accessory protein (*B gene*)	
		Com	Blp	Com	Blp
S. dysgalactiae	39.55	-	35.88	-	32.25
S. equi	41.55	35.20	39.88	32.00	35.66
S. gallolyticus	37.57	-	36.73	-	36.60
S. gordonii	40.50	40.20	-	40.80	-
S. macedonicus	37.60	-	37.10	-	36.50
S. mitis	40.00	43.40	-	44.90	-
S. mutans	36.80	-	37.06	-	32.25
S. oralis	41.10	44.50	-	41.90	-
S. pneumoniae	39.68	43.52	37.09	44.60	36.37
S. pseudopneumoniae	39.80	43.80	36.60	40.25	35.70
S. pyogenes	38.47	-	34.98	-	32.67
S. suis	41.30	-	37.30	-	35.30
S. thermophilus	39.00	-	37.60	-	40.13
S. uberis	36.60	-	27.70	-	-

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

5.4 Discussion

The results of this chapter confirm three major findings: i) QS systems evolved as a regulon unit; ii) duplication plays a major role in the functional divergence of the QS systems; and iii) HGT contributes to the sequence variations in the QS genes.

For the first finding, it was shown that three homologous QS regulon structures, i) the Blp QS regulon structure (*ABCDE*), ii) the Com QS regulon structure (*AB-EDC*), with the Blp QS regulon found within this structure, and iii) the Fas QS regulon structure (*DDE*), are mostly conserved across the *Streptococcus* species. This shows that the *Streptococcus* QS regulon structures are composed of well-conserved features under a strong positive selection, and were maintained despite the fact that any evolutionary event could have disrupted the regulon structure. This finding allows the evolution of the QS systems to be inferred with confidence, knowing that most of the gene order in prokaryotes is poorly conserved and is likely to have numerous rearrangements (Mushegian and Koonin, 1996, Watanabe et al., 1997, Itoh et al., 1999).

The three conserved regulon QS structures are similar to those reported in previous experimental studies (de Saizieu et al., 2000, Kreikemeyer et al., 2001, Berge et al., 2002). In addition, having all the genes orders conserved in a regulon may also indicate that the genes within a regulon are linked functionally. The study shows that the phylogenies of each QS regulon gene are not significantly different than the species phylogeny. This paralogous relationship between each of the regulon genes implies that the QS system evolved as a regulon unit, rather than as individual genes.

The given evidence also provides a clear indication that duplications are responsible for the functional divergence of the QS systems in *Streptococcus* species. It is found that the Blp QS regulon structure is present in most of the *Streptococcus* species. The inferred parsimonious scenario (Figure 5.5) of QS system evolution suggests that the Blp QS has an ancient origin that predates the divergence of the Com QS and the Fas QS. As shown in the inferred parsimonious scenario (Figure 5.5), a clear separation pattern is present between the Com QS (found only in the species in the Mitis group), and the Fas QS (found distributed among the pyogenic species). This information is important to assist in discovering the origin of the QS systems.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

For the third finding, some evidence was found that HGT events have occurred across the QS systems between *Streptococcus* species. Only the ABC transporter of the Com QS regulon in *S. equi* was observed to be non-vertically transferred from a species in the Mitis group (Figure 5.3). The unusual GC contents of the *comAB* genes in the *S. equi* genome further support this observation (Table 5.3).

The non-vertical descent of the QS genes was mostly observed to occur among the QS regulon genes of QS systems with the same function (Figure 5.3). The occurrence of HGTs contributes to the variability in GC contents and further promotes the sequence variation of the QS regulon genes from the same QS regulon structure, but may not significantly affect the functional divergence of QS systems. The information on HGT events was integrated into the parsimonious scenario to determine the QS ancestor to elucidate the evolutionary relationship between the QS systems in streptococcal species.

The lone ComX homologs were observed to be present in all 19 *Streptococcus* species. Studies showed that ComE of the TCS in Com QS activates *comX*, which encodes a sigma factor that facilitates the expression of the late genes that are required for DNA uptake (Lee and Morrison, 1999). Thus, it is possible that *Streptococcus* species with ComX homolog(s) are naturally competent. Despite the fact that some of the species are non-transformable or have yet to be shown to be naturally transformable, the Com QS is present in their genomes, e.g., *S. pseudopneumoniae*. Some species also have an incomplete set of Com QS regulon genes, but have been shown to be naturally transformable. *S. gordonii* regulates competence differently from *S. pneumoniae*, as the ABC transporter of the Com QS is absent in the genome. Thus, it is expected that some species without a complete set of *ABCDE* homologs will have a different QS system to regulate competence.

Further investigation of the duplication finding led to three observations regarding the gain and loss of the QS systems: i) four species (*S. mutans*, *S. thermophilus*, *S. salivarius* and *S. macedonicus*) that are known to be naturally transformable and able to produce bacteriocin have only a Blp QS regulon to regulate both functions; ii) species in the Mitis group have a specialised QS regulon (Blp, Com) to regulate bacteriocin and competence, respectively; iii) *S. salivarius* and species in the Pyogenic group have duplicated Fas QS regulons that are composed of pseudogenes.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

I consider two possible models that could explain the fate of duplicated genes in the aforementioned observations: the neofunctionalisation and subfunctionalisation models.

In the first observation, the ability of these species to regulate competence without the Com QS regulon suggests that the homologous Com genes in the Blp QS regulon act as the Com QS regulator (Knutsen et al., 2004, Kreth et al., 2005, van der Ploeg, 2005, Fontaine et al., 2010, Morrison et al., 2013). Three signal pathways: a two-component system (TCS), a Rgg regulator (*comRS*), or both were recently suggested to activate *comX* in *Streptococcus* species (Perry and Kuramitsu, 1981, Johnsborg et al., 2007, Mashburn-Warren et al., 2010). Despite the different pathways of *comX* activation in naturally transformable *Streptococcus* species (Mashburn-Warren et al., 2010, Berg et al., 2012, Mashburn-Warren et al., 2012, Morrison et al., 2013), each species shares common elements to induce competence, which are an ABC transporter, CSP and TCS.

This also suggests that the bacteriocin which is a minor function in the ancestral QS regulon has become a novel function due to positive selection This scenario fit to the 'Innovation, amplification, and divergence' (IAD) model (Nasvall et al., 2012) in the neofunctionalisation model. The ancestral gene amplifies as bacteriocin becomes important in the population. To improve bacteriocin function, some of the duplicate copies mutated and diverged until bacteriocin functioned proficiently, while some duplicated copies with the original function (competence) were maintained in the population. The ≈2-hour delay in Com regulon expression (Kreth et al., 2005) in *S. mutans* may indicate that the bacteriocin function has replaced competence as the main function.

In the second observation, the *ABCDE* regulon paralogs in the Mitis group seem to support the 'Escape adaptive conflict' (EAC) model (Hughes, 1994, Des Marais and Rausher, 2008) in the subfunctionalisation model. The multifunctional ancestor of the *ABCDE* QS regulon evolved and maintained the bacteriocin and Com functions. Because both functions are important to the fitness of the *Streptococcus* species in Mitis group, e.g., fratricide mechanisms such as the killing of non-competent cell (Claverys et al., 2007), the multifunctional ancestral genes were duplicated (subfunctionalisation). To optimise both functions, the duplicated genes evolved and each specialised in one function: bacteriocin or competence. The regulation of QS is tightly controlled and the

UNIVERSITI PENDIDIKAN SULTAN IDRIS

mechanisms are very specific to the function in each species. This is not only applied to the species in the different genera, e.g., the Com QS systems in two well-studied Gram-positive bacteria, *Bacillus subtilis* (Dubnau, 1999) and *S. pneumoniae* (Berge et al., 2002), but also between species in the same niche, e.g., the oral bacteria *S. mutans* and *S. gordonii* (Wang and Kuramitsu, 2005, Vickerman et al., 2007).

In the third observation, the TCSs in *S. salivarius* and Pyogenic group species were duplicated into a novel QS regulon (the DDE regulon structure in Fas) that controls the regulation of streptokinase activity (Kreikemeyer et al., 2001, Steiner and Malke, 2001). The HK (*D* gene) was found to be duplicated tandemly next to the TCS in the Fas QS regulon. These extra duplicated copies of the *D* gene were shown to be non-functional, and they do not benefit the streptokinase activity (Kreikemeyer et al., 2001, Steiner and Malke, 2002),which fits the criteria of the IAD model (Nasvall et al., 2012) (neofunctionalisation).

Another way of looking at QS regulon gain and loss is to consider the HGT mechanism that contributes to the acquisition of new characteristics in bacteria (Lorenz and Wackernagel, 1994). As shown in Figure 5.4, *S. equi*, a non-transformable species, has a Com ABC transporter that was shown to be horizontally transferred from a species in the Mitis group (Figure 5.3). Each species' ability to be naturally competent was gained in an independent lineage, possibly through the HGT mechanism. However, there is little evidence to suggest that HGT is the main mechanism to promote the functional divergence of the QS systems in *Streptococcus* species.

It is a challenge to classify the above observations under one of the many duplication models, as there is no clear indication of the function of the ancestral states. In future studies, this could be explored through experimental analyses of the gene expression of the QS systems to further infer the function of the ancestral states. Nevertheless, based on the analyses in this study, it is suggested that the functional divergence of the QS systems in *Streptococcus* species is best explained by the IAD model.

One of the limitations of this study is the absence of a fully resolved and absolutely confident streptococcal species tree to use when inferring the QS orthologs. This is due to

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

the complex dataset and the approach used when constructing the species tree. Nevertheless, a rooted, ML streptococcal phylogeny signifies the direction of the evolution between these species. Streptococcal species that share the same phylogenetic group, as previously defined by Kawamura et al. (1995), were show to cluster in a clade.

One of the subjects that remains to be explored is what happened to other species that have not yet been shown to be competent? Are they naturally incompetent? It may be that the competence function has been selected against in the population and lost along the lineage, or it has become a novel (minor) function that has yet to be discovered. The challenge to prepare the exact growth conditions, such as those that occur in the natural environment, which are needed for a species to transform may make it difficult for competence to be discovered in these species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

CHAPTER 6

Conclusion

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

6.1 The specificity of QS systems in streptococcal

The discovery on the presence of QS mechanism in bacteria gives a new perspective on how bacteria regulate their functions and act together in a population. QS system is a mechanism that allows the bacteria to control the behaviour of its population such as in nutrient uptake (Finkel and Kolter, 2001), defence mechanism (Høyland-Kroghsbo et al., 2013), etc. Efficient control of these responses through QS system is very important to the bacteria's chance of survival.

In this study, we are looking at the QS system in *Streptococcus pneumoniae*, which is well-known to regulate competence (Tomasz and Hotchkiss, 1964). A homologous QS system, bacteriocin-like-peptide (Blp), which responsible to regulate the production of bacteriocin, is present in *S. pneumoniae* genomes (de Saizieu et al., 2000, Martin et al., 2006). Although these two systems are homologous, each QS system is highly specific and tightly regulated to a function in the species (Havarstein et al., 1995, Reichmann and Hakenbeck, 2000). However, studies show that Com QS systems in *S. mutans* shares many features with the Blp QS system in *S. pneumoniae* (de Saizieu et al., 2000, Martin et al., 2006), where the QS system regulates both functions in *S. mutans* (Kreth et al., 2005, van der Ploeg, 2005).

Using the well-annotated Blp and Com QS regulon genes in *S. pneumoniae*, I identified the QS homologous genes in 86 *Streptococcus* genomes from the Genome NCBI database through SUCRE methodology (Chapter 2). This represents the first comprehensive study to identify QS systems across streptococcal species using a computational approach. I have shown that 3 homologous QS systems: the Com QS; the Blp QS; and the Fas QS (regulate Streptokinase), are conserved throughout the genus (Figure 5.2, Chapter 5). This suggests that strong selection has maintained the QS regulon structures against any mutation that could disrupt the regulon structures during evolution. Bacteria are known to have random gene rearrangement in their genomes, which is mostly due to the horizontal gene transfer (HGT) (Lorenz and Wackernagel, 1994, Koonin et al., 2001). The well-conserved QS regulon structures across *Streptococcus* species imply that the genes within a regulon are linked functionally. Paralogous relationships between the

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

genes in different QS systems strongly support that these genes evolved as a regulon unit rather than as an individual gene (Figure 5.3, Chapter 5).

I have also shown that the presence of QS systems varied in each species across *Streptococcus* genus in regulating competence and bacteriocin. From the comparative analysis (Figure 5.2, Chapter 5) and the experimental evidence (Appendix D) suggest that the competence and bacteriocin induction are controlled either by a QS system specific for a function as observed in *S.* pneumoniae, or a QS system that regulates both functions in *S. mutans*. With most of the streptococcal species seem to have one 'global' QS system (Figure 5.5) that regulates bacteriocin and competence, signify that the regulation of these functions are based on the switch on and off of the regulon genes in the QS system. In *S. mutans*, the delay in competence induction for almost ≈ 2 hours explains the dual functionality of the Blp QS system in the species (Kreth et al., 2005). This scenario shows that bacteriocin plays a role in the expression of competence, which also shows the regulation of bacteriocin should be considered in determining whether a species is naturally competent.

In addition on the variation in the way QS systems being regulated in each species, the QS systems have a remarkable ability to differentiate signals from other streptococcal, which implies that there is an explicit bacterial language to governing their population's behaviour. Each species has a 'secret code' to alarm its own species to defend themselves or to attack other bacterial species. However, the QS system cannot escape from the evolutionary mechanisms that may affect its regulation in the bacterial population.

6.2 The functional divergence of QS systems through duplication

The widespread of competence and bacteriocin function in the *Streptococcus* genus have gained a lot of attention in studying the divergence of two homologous QS systems that induce these functions. Although HGT is a common mechanism in streptococcal, e.g. the ability to uptake extracellular DNA into a competent cell (Lorenz and Wackernagel, 1994, Balsalobre et al., 2003, Hao and Golding, 2006), the spread of these two functions in this genus could be caused by another evolutionary event i.e., duplication, which is also often found to occur in bacterial (Makarova et al., 2001, Hooper and Berg, 2003).

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

The key indication of duplication is the presence of multiple gene copies or a gene with multiple functions in a genome. The new function of duplicated genes can be derived through subfunctionalisation and neofunctionalisation (Taylor and Raes, 2004, Bergthorsson et al., 2007, Hahn, 2009). The work presented here finds that duplication plays a major role in the functional divergence of the QS systems in the *Streptococcus* genus. I have shown that the phylogeny of each regulon gene is not significantly different from *Streptococcus* species phylogeny (Table 5.1, Chapter 5). The paralogous relationships between each regulon genes of the QS systems further support that these QS systems are being duplicated as a reguion unit.

Duplication models are used to postulate the fate of duplicated genes from the observations on the gain and loss of a QS regulon. It is evident that one QS system (Figure 5.5, chapter 5) regulates both functions in Mutans species (Appendix D) which may indicate that positive selection has acted on the QS system to maintain its duality functions. The specialisation of QS systems for each function in the species in Mitis group (Figure 5.5, Chapter 5), may indicate that the positive selection has acted on both functions equally. I have presented two duplication models, Innovation-Amplification-Duplication, IAD, and Escape-Adaptive-Conflict, EAC, which could be used to explain how the QS systems are being maintained in the population. In order to determine which model best explain the duplication process, further experimentation and analysis are required. The results suggest that the role of neofunctionalisation and subfunctionalization through the duplication process promote the functional divergence of these QS systems. The divergence of QS systems in this genus could represent the gene innovation through duplication in response to the selection.

The well-conserved QS regulon structures for each function, which are found evolving as a regulon unit, also support the functional divergence of the QS systems through duplication rather than HGT. However, it is undeniable that HGT promotes sequence variation to the QS regulon genes of the same function (Figure 5.3, Chapter 5). Therefore, the non-vertical evidence between the QS regulon genes could be one of the factor that lead to the increase of streptococcal infection and antibiotic resistance genes in this genus.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

6.3 The origin of streptococcal QS systems

The induction process of competence and bacteriocin by the QS mechanisms (Chandler and Morrison, 1988, de Saizieu et al., 2000) was discovered much later than the mechanism of bacteriocin production and transformation in competent cells(Dawson and Sia, 1931, Paul and Slade, 1975). Hence, the determination on whether a bacteria species is capable to release bacteriocin and competence, is still lacking, which leaves the study of the origin of QS system in *Streptococcus* inconclusive. The work presented here attempts to infer the origin of QS systems using a computational approach that identifies QS genes in *Streptococcus* genomes.

In this study, the first important step is finding the QS homologous genes in streptococcal species. Despite the abundance of genomes data available, the well-annotated *Streptococcus* genomes are still limited. To accommodate this limitation, I have developed SUCRE, a methodology that used competence regulon as one of the characteristics in identifying QS homologous genes across *Streptococcus* species (Chapter 2). SUCRE performed well along with OrthoMCL program (Li et al., 2003) in identifying QS homologous gene (Figure 3.3, Chapter 3).

For this study, having a good dataset alone is not enough without a reliable species phylogeny. A reliable species phylogeny is crucial as it provides important evolutionary information on the relationship between species (Maddison, 1997, Murphy et al., 2001). An incorrect inferred species tree could jeopardise the understanding on the evolution of the species (Rosenberg, 2002, Pride et al., 2003, Degnan and Rosenberg, 2006, Slatkin and Pollack, 2006). Using the supermatrix and supertree approaches, followed by the Shimodaira-Hasegawa statistical test (Shimodaira and Hasegawa, 1999), the work produced Streptococcus species phylogeny. Although the species phylogeny is not fully resolved, the streptococcal from the same phylogenetic groups as defined by Kawamura and colleagues (1995) were found to be within the same clade. The rooted species phylogeny (Figure 4.4, Chapter 4) provides the evolutionary direction of streptococcal species. This could give an insight to understand the evolution of streptococcal species in this genus, as well as other mechanisms and behaviour employed by this species.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Using the QS homologs identified by SUCRE (Chapter 2) and the *Streptococcus* species phylogeny (Chapter 4), I inferred the evolution of the QS systems (Figure 5.5, Chapter 5). By incorporating the evidences of non-vertical descendent into the Dollo parsimony approach, the gain and loss of the QS systems were deduced to infer the origin of QS systems. The work also describes the distribution of the well-conserved QS systems in the *Streptococcus* genus. The results show that the Com QS regulon was found only in the Mitis species, while the Fas QS regulon was found mostly present in the Pyogenic species, and the Blp QS regulon was found widespread across *Streptococcus* species. A clear distribution of the QS systems across streptococcal phylogenetic group provides important information in finding the origin of QS systems in this species.

The presence of Blp QS regulon in the majority of *Streptococcus* species indicates that the regulon has an ancient origin predating the divergence of Com QS and Fas QS regulon. It may well be that the ancestral QS system has Blp regulon structure with a dual-functions, either distributed equally in the genes or has a minor/major function which later evolved to the present day QS system in each species. This scenario fits *S. mutans* and *S. thermophilus* that are known to regulate competence and bacteriocin using the same QS system. The match between the identified two component-system (TCS) (Figure 5.5, Chapter 5) and the experimental evidence of the functional QS system (Appendix D) illustrate that TCS is the core element and tightly regulated to a specific function. This further suggests that the ABC transporter is relatively flexible in transporting the stimulating-peptide pheromones either for competence or bacteriocin.

The evidence presented in this study show that the QS systems in cell-cell communication in unicellular organisms are universally important, which were thought to occur indirectly e.g., sensing nutrient availability, or perhaps implausibly absent in these species. QS system controls the cell-cell communications to express particular behaviour while growing in the social community. Hence, it is not surprising that the core QS elements are evolutionary conserved between the bacterial species. This work gives an insight on the evolution of the QS systems for better understanding on the differences on the function and the characteristics of the QS systems in the streptococcal genus.

For future work this study could be used as a model to explore the evolution of QS systems in other bacteria species e.g., *Bacillus subtilis*, using the regulon structure and the

UNIVERSITI PENDIDIKAN SULTAN IDRIS

gene order. This study could also be a preliminary step to understand the interaction between bacteria and the habitat. For instance, we could explore more on how bacteria communicate in different habitats and see if the adaptation to the habitats plays a role in shaping the language of the bacteria. Therefore, the understanding of the QS system may elucidate the complex regulatory network of pathogenic bacteria during infection. This new insights will create opportunities for new development of the treatments for bacterial diseases.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

BIBLIOGRAPHY

ADVANCED BIOCOMPUTING, L. 1998. *WU-BLAST* [Online]. Available: http://blast.wustl.edu/.

ALLAN, E., HUSSAIN, H., CRAWFORD, K., MIAH, S., ASCOTT, Z., KHWAJA, M. & HOSIE, A. 2007. Genetic variation in comC, the gene encoding competence stimulating peptide (CSP) in Streptococcus mutans. *FEMS microbiology letters,* 268, 47-51.

ALTENHOFF, A. M. & DESSIMOZ, C. 2009. Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLoS Comput Biol,* 5, e1000262.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment tool. *Journal of Molecular Biology,* 215, 403-410.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J. H., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25, 3389-3402.

ANISIMOVA, M., BIELAWSKI, J., DUNN, K. & YANG, Z. 2007. Phylogenomic analysis of natural selection pressure in Streptococcus genomes. *Bmc Evolutionary Biology,* 7.

APPELBAUM, P. C. 1987. Worldwide development of antibiotic-resistance in pneumococci. *European Journal of Clinical Microbiology & Infectious Diseases,* 6, 367-377.

AVERY, O. T., MACLEOD, C. M. & MCCARTY, M. 1944. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J Exp Med,* 79, 137-58.

BALAKRISHNAN, M., SIMMONDS, R. S., KILIAN, M. & TAGG, J. R. 2002. Different bacteriocin activities of Streptococcus mutans reflect distinct phylogenetic lineages. *J Med Microbiol,* 51, 941-8.

BALDAUF, S. L., ROGER, A. J., WENK-SIEFERT, I. & DOOLITTLE, W. F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science,* 290, 972-7.

BALSALOBRE, L., FERRANDIZ, M. J., LINARES, J., TUBAU, F. & DE LA CARNPA, A. G. 2003. Viridans group streptococci are donors in horizontal transfer of topoisomerase IV genes to Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy,* 47, 2072-2081.

BAPTESTE, E., BRINKMANN, H., LEE, J. A., MOORE, D. V., SENSEN, C. W., GORDON, P., DURUFLE, L., GAASTERLAND, T., LOPEZ, P., MULLER, M. & PHILIPPE, H. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proc Natl Acad Sci U S A,* 99, 1414-9.

BARRETT, M., DONOGHUE, M. J. & SOBER, E. 1991. Against consensus. *Systematic Zoology,* 40, 486-493.

BASSLER, B. L. 1999. How bacteria talk to each other: regulation of gene expression by quorum sensing. *Current Opinion in Microbiology,* 2, 582-587.

BAYSSE, C., CULLINANE, M., DÉNERVAUD, V., BURROWES, E., DOW, J. M., MORRISSEY, J. P., TAM, L., TREVORS, J. T. & O'GARA, F. 2005. Modulation of quorum sensing in Pseudomonas aeruginosa through alteration of membrane properties. *Microbiology,* 151, 2529-2542.

BEIKO, R. G., HARLOW, T. J. & RAGAN, M. A. 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America,* 102, 14332-14337.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

BENSING, B. A., RUBENS, C. E. & SULLAM, P. M. 2001. Genetic loci of Streptococcus mitis that mediate binding to human platelets. *Infect Immun,* 69, 1373-80.

BENSON, D., BOGUSKI, M., LIPMAN, D., OSTELL, J. & OUELLETTE, B. 1998. GenBank. Nucleic acids research, 26, 1.

BENTLEY, R. W. & LEIGH, J. A. 1995. Determination of 16S ribosomal RNA gene copy number in Streptococcus uberis, S. agalactiae, S. dysgalactiae and S. parauberis. FEMS Immunol Med Microbiol, 12, 1-7.

BERG, K. H., OHNSTAD, H. S. & HAVARSTEIN, L. S. 2012. LytF, a novel competence-regulated murein hydrolase in the genus Streptococcus. *J Bacteriol,* 194, 627-35.

BERGE, M., MOSCOSO, M., PRUDHOMME, M., MARTIN, B. & CLAVERYS, J. P. 2002. Uptake of transforming DNA in Gram-positive bacteria: a view from Streptococcus pneumoniae. *Molecular Microbiology,* 45, 411-421.

BERGTHORSSON, U., ANDERSSON, D. I. & ROTH, J. R. 2007. Ohno's dilemma: evolution of new genes under continuous selection. *Proc Natl Acad Sci U S A,* 104, 17004-9.

BININDA-EMONDS, O. R. 2004. The evolution of supertrees. *Trends Ecol Evol,* 19, 315-22.

BININDA-EMONDS, O. R. P., GITTLEMAN, J. L. & STEEL, M. A. 2002. The (super)tree of life: Procedures, problems, and prospects. *Annual Review of Ecology and Systematics,* 33, 265-289.

BIRNEY, E., CLAMP, M. & DURBIN, R. 2004. GeneWise and Genomewise. *Genome Res,* 14, 988-95.

BLOMQVIST, T., STEINMOEN, H. & HAVARSTEIN, L. S. 2006. Natural genetic transformation: A novel tool for efficient genetic engineering of the dairy bacterium Streptococcus thermophilus. *Appl Environ Microbiol,* 72, 6751-6.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

BODILIS, J., NSIGUE-MEILO, S., BESAURY, L. & QUILLET, L. 2012. Variable Copy Number, Intra-Genomic Heterogeneities and Lateral Transfers of the 16S rRNA Gene in <italic>Pseudomonas</italic>. *PLoS ONE,* 7, e35647.

BORODOVSKY, M., MCININCH, J. D., KOONIN, E. V., RUDD, K. E., MEDIGUE, C. & DANCHIN, A. 1995. Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res,* 23, 3554-62.

BORODOVSKY, M., RUDD, K. E. & KOONIN, E. V. 1994. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res,* 22, 4756-67.

BROCHIER, C., PHILIPPE, H. & MOREIRA, D. 2000. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet,* 16, 529 - 533.

BROWN, H., SANGER, F. & KITAI, R. 1955. The structure of pig and sheep insulins. *Biochemical Journal,* 60, 556-565.

BULL, J. J., HUELSENBECK, J. P., CUNNINGHAM, C. W., SWOFFORD, D. L. & WADDELL, P. J. 1993. Partitioning and combining data in phylogenetic analysis. *Systematic Biology,* 42, 384-397.

BURGE, C. & KARLIN, S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol,* 268, 78-94.

BURGE, C. B. & KARLIN, S. 1998. Finding the genes in genomic DNA. *Curr Opin Struct Biol,* 8, 346-54.

BURGHOUT, P., BOOTSMA, H. J., KLOOSTERMAN, T. G., BIJLSMA, J. J., DE JONGH, C. E., KUIPERS, O. P. & HERMANS, P. W. 2007. Search for genes essential for pneumococcal transformation: the RADA DNA repair protein plays a role in genomic recombination of donor DNA. *J Bacteriol,* 189, 6540-50.

CAMPBELL, E. A., CHOI, S. Y. & MASURE, H. R. 1998. A competence regulon in Streptococcus pneumoniae revealed by genomic analysis. *Molecular Microbiology,* 27, 929-939.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

CHANDLER, M. S. & MORRISON, D. A. 1988. Identification of two proteins encoded by com, a competence control locus of Streptococcus pneumoniae. *J Bacteriol,* 170, 3136-41.

CHAUSSEE, M. S., SYLVA, G. L., STURDEVANT, D. E., SMOOT, L. M., GRAHAM, M. R., WATSON, R. O. & MUSSER, J. M. 2002. Rgg influences the expression of multiple regulatory loci to coregulate virulence factor expression in Streptococcus pyogenes. *Infect Immun,* 70, 762-70.

CHEN, I., CHRISTIE, P. J. & DUBNAU, D. 2005. The ins and outs of DNA transfer in bacteria. *Science,* 310, 1456-60.

CHEN, I. & DUBNAU, D. 2004. DNA uptake during bacterial transformation. *Nature Reviews Microbiology,* 2, 241-249.

CHEN, K., DURAND, D. & FARACH-COLTON, M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol,* 7, 429-47.

CHENG, Q., CAMPBELL, E. A., NAUGHTON, A. M., JOHNSON, S. & MASURE, H. R. 1997. The com locus controls genetic transformation in Streptococcus pneumoniae. *Molecular Microbiology,* 23, 683-692.

CHIPPINDALE, P. T. & WIENS, J. J. 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. *Systematic Biology,* 43, 278-287.

CICCARELLI, F. D., DOERKS, T., CHRISTIAN VON, M., CREEVEY, C. J., SNEL, B. & BORK, P. 2006. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science,* 311, 1283-1287.

CLAVERYS, J. P. & HAVARSTEIN, L. S. 2002. Extracellular-peptide control of competence for genetic transformation in Streptococcus pneumoniae. *Frontiers in Bioscience,* 7, D1798-D1814.

CLAVERYS, J. P. & HAVARSTEIN, L. S. 2007. Cannibalism and fratricide: mechanisms and raisons d'etre. *Nature Reviews Microbiology,* 5, 219-229.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

CLAVERYS, J. P. & LACKS, S. A. 1986. Heteroduplex deoxyribonucleic acid base mismatch repair in bacteria. *Microbiological Reviews,* 50, 133-165.

CLAVERYS, J. P., MARTIN, B. & HAVARSTEIN, L. S. 2007. Competence-induced fratricide in streptococci (vol 64, pg 1423, 2007). *Molecular Microbiology,* 65, 230-230.

CLAVERYS, J. P., PRUDHOMME, M. & MARTIN, B. 2006. Induction of competence regulons as a general response to stress in gram-positive bacteria. *Annual Review of Microbiology,* 60, 451-475.

COLE, J., CHAI, B., FARRIS, R., WANG, Q., KULAM, S., MCGARRELL, D., GARRITY, G. & TIEDJE, J. 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic acids research,* 33, D294.

CONANT, G. C. & WAGNER, A. 2002. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Research,* 30, 3378-3386.

COOKSON, A. L., NOEL, S. J., KELLY, W. J. & ATTWOOD, G. T. 2004. The use of PCR for the identification and characterisation of bacteriocin genes from bacterial strains isolated from rumen or caecal contents of cattle and sheep. *FEMS Microbiology Ecology,* 48, 199-207.

CORNEJO, O., MCGEE, L. & ROZEN, D. 2010. Polymorphic competence peptides do not restrict recombination in Streptococcus pneumoniae. *Molecular Biology and Evolution,* 27, 694.

CORREIA, F. F., MCKAY, T. L., FARROW, M. F., ROSAN, B. & DIRIENZO, J. M. 1996. Natural transformation of Streptococcus crista. *FEMS Microbiol Lett,* 143, 13-8.

CRAIG, L. & LI, J. 2008. Type IV pili: paradoxes in form and function. *Curr Opin Struct Biol,* 18, 267-77.

DAGAN, T. & MARTIN, W. 2006. The tree of one percent. *Genome Biology,* 7, 118.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

DAUBIN, V., GOUY, M. & PERRIERE, G. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Inform,* 12, 155-64.

DAUBIN, V., GOUY, M. & PERRIERE, G. 2002. A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Research,* 12, 1080-1090.

DAVIES, M. R., MCMILLAN, D. J., VAN DOMSELAAR, G. H., JONES, M. K. & SRIPRAKASH, K. S. 2007. Phage 3396 from a Streptococcus dysgalactiae subsp. equisimilis Pathovar May Have Its Origins in Streptococcus pyogenes. *Journal of Bacteriology,* 189, 2646-2652.

DAWSON, M. H. & SIA, R. H. 1931. IN VITRO TRANSFORMATION OF PNEUMOCOCCAL TYPES : I. A TECHNIQUE FOR INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES IN VITRO. *J Exp Med,* 54, 681-99.

DAYHOFF, M. & SCHWARTZ, R. A model of evolutionary change in proteins. 1978. Citeseer.

DE QUEIROZ, A. & GATESY, J. 2007. The supermatrix approach to systematics. *Trends in Ecology and Evolution,* 22, 34-41.

DE SAIZIEU, A., GARDES, C., FLINT, N., WAGNER, C., KAMBER, M., MITCHELL, T., KECK, W., AMREIN, K. & LANGE, R. 2000. Microarray-based identification of a novel Streptococcus pneumoniae regulon controlled by an autoinduced peptide. *Journal of Bacteriology,* 182, 4696.

DEGNAN, J. H. & ROSENBERG, N. A. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. *PLoS Genet,* 2, e68.

DELSUC, F., BRINKMANN, H. & PHILIPPE, H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet,* 6, 361-75.

DES MARAIS, D. L. & RAUSHER, M. D. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature,* 454, 762-5.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

DESAI, B. V. & MORRISON, D. A. 2007. Transformation in Streptococcus pneumoniae: formation of eclipse complex in a coiA mutant implicates CoiA in genetic recombination. *Mol Microbiol,* 63, 1107-17.

DOOLITTLE, R. F. & HANDY, J. 1998. Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Current Opinion in Genetics & Development,* 8, 630-636.

DUBNAU, D. 1999. DNA uptake in bacteria. *Annual Review of Microbiology,* 53, 217-244.

DUNNY, G. & WINANS, S. 1999. Bacterial life: neither lonely nor boring. *Cell–Cell Signaling in Bacteria*, 1–8.

EDGAR, R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research,* 32, 1792.

EISEN, J. A. 1998. Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Research,* 8, 163-167.

FARRIS, J. S. 1977. Phylogenetic Analysis Under Dollo's Law. *Systematic Biology,* 26, 77-88.

FARRIS, J. S. 1985. DISTANCE DATA REVISITED. *Cladistics,* 1, 67-86.

FARRIS, J. S., ALBERT, V. A., KÄLLERSJÖ, M., LIPSCOMB, D. & KLUGE, A. G. 1996. PARSIMONY JACKKNIFING OUTPERFORMS NEIGHBOR-JOINING. *Cladistics,* 12, 99-124.

FATH, M. J. & KOLTER, R. 1993. ABC transporters: bacterial exporters. *Microbiol Rev,* 57, 995-1017.

FELSENSTEIN, J. 1981. Evolutionary trees from dna-sequences - a maximum-likelihood approach. *Journal of Molecular Evolution,* 17, 368-376.

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

FELSENSTEIN, J. 2004. *Inferring phylogenies*, Sinauer Associates Sunderland.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

FENG, D. F. & DOOLITTLE, R. F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol,* 25, 351-60.

FINKEL, S. E. & KOLTER, R. 2001. DNA as a nutrient novel role for bacterial competence gene homologs. *Journal of Bacteriology,* 183, 6288-6293.

FITCH, W. M. 1970. Distinguishing Homologous from Analogous Proteins. *Systematic Biology,* 19, 99-113.

FITCH, W. M. 2000. Homology a personal view on some of the problems. *Trends Genet,* 16, 227-31.

FLICEK, P., AKEN, B. L., BEAL, K., BALLESTER, B., CACCAMO, M., CHEN, Y., CLARKE, L., COATES, G., CUNNINGHAM, F., CUTTS, T., DOWN, T., DYER, S. C., EYRE, T., FITZGERALD, S., FERNANDEZ-BANET, J., GRAF, S., HAIDER, S., HAMMOND, M., HOLLAND, R., HOWE, K. L., HOWE, K., JOHNSON, N., JENKINSON, A., KAHARI, A., KEEFE, D., KOKOCINSKI, F., KULESHA, E., LAWSON, D., LONGDEN, I., MEGY, K., MEIDL, P., OVERDUIN, B., PARKER, A., PRITCHARD, B., PRLIC, A., RICE, S., RIOS, D., SCHUSTER, M., SEALY, I., SLATER, G., SMEDLEY, D., SPUDICH, G., TREVANION, S., VILELLA, A. J., VOGEL, J., WHITE, S., WOOD, M., BIRNEY, E., COX, T., CURWEN, V., DURBIN, R., FERNANDEZ-SUAREZ, X. M., HERRERO, J., HUBBARD, T. J., KASPRZYK, A., PROCTOR, G., SMITH, J., URETA-VIDAL, A. & SEARLE, S. 2008. Ensembl 2008. *Nucleic Acids Res,* 36, D707-14.

FONTAINE, L., BOUTRY, C., DE FRAHAN, M. H., DELPLACE, B., FREMAUX, C., HORVATH, P., BOYAVAL, P. & HOLS, P. 2010. A novel pheromone quorum-sensing system controls the development of natural competence in Streptococcus thermophilus and Streptococcus salivarius. *J Bacteriol,* 192, 1444-54.

FONTAINE, L., BOUTRY, C., GUÉDON, E., GUILLOT, A., IBRAHIM, M., GROSSIORD, B. & HOLS, P. 2007. Quorum-Sensing Regulation of the Production of Blp Bacteriocins in Streptococcus thermophilus. *Journal of Bacteriology,* 189, 7195-7205.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P

FORCE, A., LYNCH, M., PICKETT, F. B., AMORES, A., YAN, Y. L. & POSTLETHWAIT, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics,* 151, 1531-45.

FOX, M. S. 1957. Deoxyribonucleic acid incorporation by transformed bacteria. *Biochimica et Biophysica Acta,* 26, 83-85.

FUCHSMAN, C. A. & ROCAP, G. 2006. Whole-Genome Reciprocal BLAST Analysis Reveals that Planctomycetes Do Not Share an Unusually Large Number of Genes with Eukarya and Archaea. *Applied and Environmental Microbiology,* 72, 6841-6844.

FUQUA, W. C., WINANS, S. C. & GREENBERG, E. P. 1994. Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. *Journal of bacteriology,* 176, 269-275.

GARGIS, A., O'ROURKE, A.-L., SLOAN, G. & SIMMONDS, R. 2009. Prevalence and Acquisition of the Genes for Zoocin A and Zoocin A Resistance in Streptococcus equi subsp. zooepidemicus. *Journal of Molecular Evolution,* 68, 498-505.

GAUSTAD, P. 1979. Genetic-transformation in streptococcus-sanguis - distribution of competence factors in a collection of strains. *Acta Pathologica Et Microbiologica Scandinavica Section B-Microbiology,* 87, 123-128.

GELFAND, M. S., MIRONOV, A. A. & PEVZNER, P. A. 1996. Gene recognition via spliced sequence alignment. *Proc Natl Acad Sci U S A,* 93, 9061-6.

GISH, W. & STATES, D. J. 1993. Identification of protein coding regions by database similarity search. *Nat Genet,* 3, 266-72.

GOGARTEN, J. P., DOOLITTLE, W. F. & LAWRENCE, J. G. 2002. Prokaryotic Evolution in Light of Gene Transfer. *Molecular Biology and Evolution,* 19, 2226-2238.

GOLDMAN, N., ANDERSON, J. P. & RODRIGO, A. G. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst Biol,* 49, 652-70.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

GRIFFITH, F. 1928. The Significance of Pneumococcal Types. *J Hyg (Lond),* 27, 113-59.

HAHN, M. W. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered,* 100, 605-17.

HAO, W. & GOLDING, G. 2004. Patterns of bacterial gene movement. *Mol Biol Evol,* 21, 1294 - 1307.

HAO, W. L. & GOLDING, G. B. 2006. The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research,* 16, 636-643.

HAVARSTEIN, L. S., COOMARASWAMY, G. & MORRISON, D. A. 1995. An unmodified heptadecapeptide pheromone induces competence for genetic-transformation in streptococcus-pneumoniae. *Proceedings of the National Academy of Sciences of the United States of America,* 92, 11140-11144.

HAVARSTEIN, L. S., GAUSTAD, P., NES, I. F. & MORRISON, D. A. 1996. Identification of the streptococcal competence-pheromone receptor. *Molecular Microbiology,* 21, 863-869.

HAVARSTEIN, L. S., HAKENBECK, R. & GAUSTAD, P. 1997. Natural competence in the genus Streptococcus: Evidence that streptococci can change pherotype by interspecies recombinational exchanges. *Journal of Bacteriology,* 179, 6589-6594.

HENG, N. C., RAGLAND, N. L., SWE, P. M., BAIRD, H. J., INGLIS, M. A., TAGG, J. R. & JACK, R. W. 2006. Dysgalacticin: a novel, plasmid-encoded antimicrobial protein (bacteriocin) produced by Streptococcus dysgalactiae subsp. equisimilis. *Microbiology,* 152, 1991-2001.

HERSHEY, A. D. & CHASE, M. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol,* 36, 39-56.

HILLIS, D. M. & BULL, J. J. 1993. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology,* 42, 182-192.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

HOELZER, M. A. & MICHOD, R. E. 1991. DNA-repair and the evolution of transformation in bacillus-subtilis .3. Sex with damaged DNA. *Genetics,* 128, 215-223.

HOLDEN, M. T., HEATHER, Z., PAILLOT, R., STEWARD, K. F., WEBB, K., AINSLIE, F., JOURDAN, T., BASON, N. C., HOLROYD, N. E., MUNGALL, K., QUAIL, M. A., SANDERS, M., SIMMONDS, M., WILLEY, D., BROOKS, K., AANENSEN, D. M., SPRATT, B. G., JOLLEY, K. A., MAIDEN, M. C., KEHOE, M., CHANTER, N., BENTLEY, S. D., ROBINSON, C., MASKELL, D. J., PARKHILL, J. & WALLER, A. S. 2009. Genomic evidence for the evolution of Streptococcus equi: host restriction, increased virulence, and genetic exchange with human pathogens. *PLoS Pathog,* 5, e1000346.

HOOPER, S. D. & BERG, O. G. 2003. Duplication is more common among laterally transferred genes than among indigenous genes. *Genome Biol,* 4, R48.

HOWE, K., BATEMAN, A. & DURBIN, R. 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics,* 18, 1546-1547.

HØYLAND-KROGHSBO, N. M., MÆRKEDAHL, R. B. & SVENNINGSEN, S. L. 2013. A Quorum-Sensing-Induced Bacteriophage Defense Mechanism. *mBio,* 4.

HUELSENBECK, J. P. & BULL, J. J. 1996. A Likelihood Ratio Test to Detect Conflicting Phylogenetic Signal. *Systematic Biology,* 45, 92-98.

HUGHES, A. L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci,* 256, 119-24.

HUI, F. M. & MORRISON, D. A. 1991. Genetic transformation in Streptococcus pneumoniae: nucleotide sequence analysis shows comA, a gene required for competence induction, to be a member of the bacterial ATP-dependent transport protein family. *J Bacteriol,* 173, 372-81.

HUI, F. M., ZHOU, L. & MORRISON, D. A. 1995. Competence for genetic transformation in Streptococcus pneumoniae: organization of a regulatory locus with homology to two lactococcin A secretion genes. *Gene,* 153, 25-31.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

HUSON, D., RICHTER, D., RAUSCH, C., DEZULIAN, T., FRANZ, M. & RUPP, R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. *Bmc Bioinformatics,* 8, 460.

IBRAHIM, M., GUILLOT, A., WESSNER, F., ALGARON, F., BESSET, C., COURTIN, P., GARDAN, R. & MONNET, V. 2007. Control of the transcription of a short gene encoding a cyclic peptide in Streptococcus thermophilus: a new quorum-sensing system? *J Bacteriol,* 189, 8844-54.

ITOH, T., TAKEMOTO, K., MORI, H. & GOJOBORI, T. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol,* 16, 332 - 346.

JAIN, R., RIVERA, M. & LAKE, J. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA,* 96, 3801 - 3806.

JEFFARES, D., MOURIER, T. & PENNY, D. 2006. The biology of intron gain and loss. *Trends Genet,* 22, 16 - 22.

JOELSSON, A., LIU, Z. & ZHU, J. 2006. Genetic and Phenotypic Diversity of Quorum-Sensing Systems in Clinical and Environmental Isolates of Vibrio cholerae. *Infection and Immunity,* 74, 1141-1147.

JOHNSBORG, O., ELDHOLM, V., BJORNSTAD, M. L. & HAVARSTEIN, L. S. 2008. A predatory mechanism dramatically increases the efficiency of lateral gene transfer in Streptococcus pneumoniae and related commensal species. *Molecular Microbiology,* 69, 245-253.

JOHNSBORG, O., ELDHOLM, V. & HAVARSTEIN, L. S. 2007. Natural genetic transformation: prevalence, mechanisms and function. *Research in Microbiology,* 158, 767-778.

KAWAMURA, Y., HOU, X., SULTANA, F., MIURA, H. & EZAKI, T. 1995. Determination of 16S rRNA sequences of Streptococcus mitis and Streptococcus gordonii and phylogenetic relationships among members of the genus

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Streptococcus. *International Journal of Systematic and Evolutionary Microbiology,* 45, 406.

KEMENA, C. & NOTREDAME, C. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics,* 25, 2455-2465.

KHAN, I. U., HASSAN, A. A., ABDULMAWJOOD, A., LAMMLER, C., WOLTER, W. & ZSCHOCK, M. 2003. Identification and epidemiological characterization of Streptococcus uberis isolated from bovine mastitis using conventional and molecular methods. *J Vet Sci,* 4, 213-24.

KLUGE, A. G. 1989. A Concern for Evidence and a Phylogenetic Hypothesis of Relationships Among Epicrates (Boidae, Serpentes). *Systematic Zoology,* 38, 7-25.

KNUTSEN, E., WEEN, O. & HAVARSTEIN, L. S. 2004. Two separate quorum-sensing systems upregulate transcription of the same ABC transporter in Streptococcus pneumoniae. *J Bacteriol,* 186, 3078-85.

KOONIN, E., MAKAROVA, K. & ARAVIND, L. 2001. Horizontal gene transfer in prokaryotes - quantification and classification. *Annu Rev Microbiol,* 55, 709 - 742.

KREIKEMEYER, B., BOYLE, M. D., BUTTARO, B. A., HEINEMANN, M. & PODBIELSKI, A. 2001. Group A streptococcal growth phase-associated virulence factor regulation by a novel operon (Fas) with homologies to two-component-type regulators requires a small RNA molecule. *Mol Microbiol,* 39, 392-406.

KRETH, J., MERRITT, J., SHI, W. & QI, F. 2005. Co-ordinated bacteriocin production and competence development: a possible mechanism for taking up DNA from neighbouring species. *Mol Microbiol,* 57, 392-404.

KUBATKO, L. S. & DEGNAN, J. H. 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Systematic Biology,* 56, 17-24.

LEBEL, G., VAILLANCOURT, K., FRENETTE, M., GOTTSCHALK, M. & GRENIER, D. 2014. Suicin 90-1330 from a nonvirulent strain of Streptococcus suis: a nisin-related lantibiotic active on gram-positive swine pathogens. *Appl Environ Microbiol,* 80, 5484-92.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

LEBLANC, D. J., HAWLEY, R. J., LEE, L. N. & STMARTIN, E. J. 1978. Conjugal transfer of plasmid dna among oral streptococci. *Proceedings of the National Academy of Sciences of the United States of America,* 75, 3484-3487.

LEE, M. S. & MORRISON, D. A. 1999. Identification of a new regulator in Streptococcus pneumoniae linking quorum sensing to competence for genetic transformation. *Journal of Bacteriology,* 181, 5004-5016.

LEIGH, J. W., SUSKO, E., BAUMGARTNER, M. & ROGER, A. J. 2008. Testing Congruence in Phylogenomic Analysis. *Systematic Biology,* 57, 104-115.

LI, L., STOECKERT, C. J. & ROOS, D. S. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research,* 13, 2178-2189.

LIN, J. & GERSTEIN, M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels. *Genome Research,* 10, 808-818.

LIO, P. & GOLDMAN, N. 1998. Models of molecular evolution and phylogeny. *Genome Research,* 8, 1233-1244.

LORENZ, M. G. & WACKERNAGEL, W. 1994. Bacterial gene-transfer by natural genetic-transformation in the environment. *Microbiological Reviews,* 58, 563-602.

LUO, P., LI, H. Y. & MORRISON, D. A. 2004. Identification of ComW as a new component in the regulation of genetic transformation in Streptococcus pneumoniae. *Molecular Microbiology,* 54, 172-183.

MACRINA, F. L., WOOD, P. H. & JONES, K. R. 1980. Genetic transformation of Streptococcus sanguis (Challis) with cryptic plasmids from Streptococcus ferus. *Infect Immun,* 28, 692-9.

MADDISON, W. P. 1997. Gene Trees in Species Trees. *Systematic Biology,* 46, 523-536.

MADIGAN, M. T., MARTINKO, J. M., DUNLAP, P. V., CLARK, D. P. & BROCK, T. 2008. *Brock biology of microorganisms,* San Francisco, Pearson Benjamin-Cummings.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

MAKAROVA, K., PONOMAREV, V. & KOONIN, E. 2001. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol,* 2, research0033.1 - 0033.14.

MALIA JR, M. J., LIPSCOMB, D. L. & ALLARD, M. W. 2003. The misleading effects of composite taxa in supermatrices. *Molecular Phylogenetics and Evolution,* 27, 522-527.

MALKE, H. & STEINER, K. 2004. Control of streptokinase gene expression in group A & C streptococci by two-component regulators. *Indian J Med Res,* 119 Suppl, 48-56.

MARAGKOUDAKIS, P. A., PAPADELLI, M., GEORGALAKI, M., PANAYOTOPOULOU, E. G., MARTINEZ-GONZALEZ, B., MENTIS, A. F., PETRAKI, K., SGOURAS, D. N. & TSAKALIDOU, E. 2009. In vitro and in vivo safety evaluation of the bacteriocin producer Streptococcus macedonicus ACA-DC 198. *International Journal of Food Microbiology,* 133, 141-147.

MARTIN, B., GARCIA, P., CASTANIE, M. P. & CLAVERYS, J. P. 1995. The recA gene of streptococcus-pneumoniae is part of a competence-induced operon and controls lysogenic induction. *Molecular Microbiology,* 15, 367-379.

MARTIN, B., QUENTIN, Y., FICHANT, G. & CLAVERYS, J. P. 2006. Independent evolution of competence regulatory cascades in streptococci? *Trends in Microbiology,* 14, 339-345.

MARTIN, B., RUELLAN, J. M., ANGULO, J. F., DEVORET, R. & CLAVERYS, J. P. 1992. Identification of the recA gene of Streptococcus pneumoniae. *Nucleic Acids Res,* 20, 6412.

MARTIN, W. 1996. Is something wrong with the tree of life? *BioEssays,* 18, 523 - 527.

MASHBURN-WARREN, L., MORRISON, D. A. & FEDERLE, M. J. 2010. A novel double-tryptophan peptide pheromone controls competence in Streptococcus spp. via an Rgg regulator. *Mol Microbiol,* 78, 589-606.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

MASHBURN-WARREN, L., MORRISON, D. A. & FEDERLE, M. J. 2012. The cryptic competence pathway in Streptococcus pyogenes is controlled by a peptide pheromone. *J Bacteriol,* 194, 4589-600.

MERCER, D. K., MELVILLE, C. M., SCOTT, K. P. & FLINT, H. J. 1999. Natural genetic transformation in the rumen bacterium Streptococcus bovis JB1. *FEMS Microbiol Lett,* 179, 485-90.

MICHOD, R. E., WOJCIECHOWSKI, M. F. & HOELZER, M. A. 1988. DNA-repair and the evolution of transformation in the bacterium bacillus-subtilis. *Genetics,* 118, 31-39.

MILLER, M. B. & BASSLER, B. L. 2001. Quorum sensing in bacteria. *Annual Review of Microbiology,* 55, 165-199.

MIYAMOTO, M. M. & FITCH, W. M. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology,* 44, 64-76.

MONEY, D. & WHELAN, S. 2012. Characterizing the Phylogenetic Tree-Search Problem. *Systematic Biology,* 61, 228-239.

MONGOLD, J. A. 1992. DNA-repair and the evolution of transformation in haemophilus-influenzae. *Genetics,* 132, 893-898.

MORRISON, D. A. 1977. Transformation in pneumococcus: existence and properties of a complex involving donor deoxyribonucleate single strands in eclipse. *J Bacteriol,* 132, 576-83.

MORRISON, D. A. 1978. Transformation in pneumococcus: protein content of eclipse complex. *J Bacteriol,* 136, 548-57.

MORRISON, D. A., GUEDON, E. & RENAULT, P. 2013. Competence for natural genetic transformation in the Streptococcus bovis group streptococci S. infantarius and S. macedonicus. *J Bacteriol,* 195, 2612-20.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

MORRISON, D. A. & LEE, M. S. 2000. Regulation of competence for genetic transformation in Streptococcus pneumoniae: a link between quorum sensing and DNA processing genes. *Research in Microbiology,* 151, 445-451.

MORRISON, D. A., MORTIER-BARRIÈRE, I., ATTAIECH, L. & CLAVERYS, J.-P. 2007. Identification of the Major Protein Component of the Pneumococcal Eclipse Complex. *Journal of bacteriology,* 189, 6497-6500.

MORTIER-BARRIERE, I., DE SAIZIEU, A., CLAVERYS, J. P. & MARTIN, B. 1998. Competence-specific induction of recA is required for full recombination proficiency during transformation in Streptococcus pneumoniae. *Molecular Microbiology,* 27, 159-170.

MORTIER-BARRIERE, I., VELTEN, M., DUPAIGNE, P., MIROUZE, N., PIETREMENT, O., MCGOVERN, S., FICHANT, G., MARTIN, B., NOIROT, P., LE CAM, E., POLARD, P. & CLAVERYS, J. P. 2007. A key presynaptic role in transformation for a widespread bacterial protein: DprA conveys incoming ssDNA to RecA. *Cell,* 130, 824-36.

MOUNT, D. W. 2009. Using iterative methods for global multiple sequence alignment. *Cold Spring Harb Protoc,* 2009, pdb top44.

MURPHY, W. J., EIZIRIK, E., JOHNSON, W. E., ZHANG, Y. P., RYDER, O. A. & O'BRIEN, S. J. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature,* 409, 614-8.

MUSHEGIAN, A. R. & KOONIN, E. V. 1996. Gene order is not conserved in bacterial evolution. *Trends Genet,* 12, 289-90.

NASVALL, J., SUN, L., ROTH, J. R. & ANDERSSON, D. I. 2012. Real-time evolution of new genes by innovation, amplification, and divergence. *Science,* 338, 384-7.

NEEDLEMAN, S. B. & WUNSCH, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol,* 48, 443-53.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

NELSON, K. E., CLAYTON, R. A., GILL, S. R., GWINN, M. L., DODSON, R. J., HAFT, D. H., HICKEY, E. K., PETERSON, J. D., NELSON, W. C., KETCHUM, K. A., MCDONALD, L., UTTERBACK, T. R., MALEK, J. A., LINHER, K. D., GARRETT, M. M., STEWART, A. M., COTTON, M. D., PRATT, M. S., PHILLIPS, C. A., RICHARDSON, D., HEIDELBERG, J., SUTTON, G. G., FLEISCHMANN, R. D., EISEN, J. A., WHITE, O., SALZBERG, S. L., SMITH, H. O., VENTER, J. C. & FRASER, C. M. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature,* 399, 323-9.

O'BRIEN, S. J. & STANYON, R. 1999. Phylogenomics: Ancestral primate viewed. *Nature,* 402, 365-366.

OCHMAN, H. 2001. Lateral and oblique gene transfer. *Current Opinion in Genetics & Development,* 11, 616-619.

OCHMAN, H., LAWRENCE, J. G. & GROISMAN, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature,* 405, 299-304.

OGDEN, T. & ROSENBERG, M. 2006. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic biology,* 55, 314.

OHNO, S. 1970. *Evolution by gene duplication,* London, Allen & Unwin.

PAKULA, R. & WALCZAK, W. 1963. On the Nature of Competence of Transformable Streptococci. *Journal of General Microbiology,* 31, 125-133.

PALMEN, R., BUIJSMAN, P. & HELLINGWERF, K. J. 1994. Physiological regulation of competence induction for natural transformation in acinetobacter-calcoaceticus. *Archives of Microbiology,* 162, 344-351.

PARADIS, E., CLAUDE, J. & STRIMMER, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics,* 20, 289-290.

PAUL, D. & SLADE, H. D. 1975. Production and properties of an extracellular bacteriocin from Streptococcus mutans bacteriocidal for group A and other streptococci. *Infect Immun,* 12, 1375-85.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

PEI, A. Y., OBERDORF, W. E., NOSSA, C. W., AGARWAL, A., CHOKSHI, P., GERZ, E. A., JIN, Z., LEE, P., YANG, L., POLES, M., BROWN, S. M., SOTERO, S., DESANTIS, T., BRODIE, E., NELSON, K. & PEI, Z. 2010. Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl Environ Microbiol,* 76, 3886-97.

PERRY, D. & KURAMITSU, H. K. 1981. Genetic-transformation of streptococcus-mutans. *Infection and Immunity,* 32, 1295-1297.

PESTOVA, E. V., HAVARSTEIN, L. S. & MORRISON, D. A. 1996. Regulation of competence for genetic transformation in Streptococcus pneumoniae by an auto-induced peptide pheromone and a two-component regulatory system. *Molecular Microbiology,* 21, 853-862.

PETERSEN, F. C. & SCHEIE, A. A. 2000. Genetic transformation in Streptococcus mutans requires a peptide secretion–like apparatus. *Oral Microbiology and Immunology,* 15, 329-334.

PETERSON, S., SUNG, C., CLINE, R., DESAI, B., SNESRUD, E., LUO, P., WALLING, J., LI, H., MINTZ, M. & TSEGAYE, G. 2004. Identification of competence pheromone responsive genes in Streptococcus pneumoniae by use of DNA microarrays. *Molecular Microbiology,* 51, 1051-1070.

PHILIPPE, H. & DOUADY, C. J. 2003. Horizontal gene transfer and phylogenetics. *Current Opinion in Microbiology,* 6, 498-505.

PHILIPPE, H. & LAURENT, J. 1998. How good are deep phylogenetic trees? *Curr Opin Genet Dev,* 8, 616-23.

PHILLIPS, M. J., DELSUC, F. & PENNY, D. 2004. Genome-Scale Phylogeny and the Detection of Systematic Biases. *Molecular Biology and Evolution,* 21, 1455-1458.

PIATIGORSKY, J. 1998. Gene sharing in lens and cornea: facts and implications. *Prog Retin Eye Res,* 17, 145-74.

PONTIUS, J., WAGNER, L. & SCHULER, G. 2003. The NCBI handbook. *Bethesda, MD: National Center for Biotechnology Information.*

UNIVERSITI PENDIDIKAN SULTAN IDRIS

POWELL, B. C. & HUTCHISON, C. A. 2006. Similarity-based gene detection: using COGs to find evolutionarily-conserved ORFs. *Bmc Bioinformatics,* 7.

PRIDE, D. T., MEINERSMANN, R. J., WASSENAAR, T. M. & BLASER, M. J. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res,* 13, 145-58.

PRUITT, K., TATUSOVA, T. & MAGLOTT, D. 2006. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research.*

RAMIREZ-PENA, E., TREVINO, J., LIU, Z., PEREZ, N. & SUMBY, P. 2010. The group A Streptococcus small regulatory RNA FasX enhances streptokinase activity by increasing the stability of the ska mRNA transcript. *Mol Microbiol,* 78, 1332-47.

RATHSAM, C., EATON, R. E., SIMPSON, C. L., BROWNE, G. V., BERG, T., HARTY, D. W. S. & JACQUES, N. A. 2005. Up-regulation of competence- but not stress-responsive proteins accompanies an altered metabolic phenotype in Streptococcus mutans biofilms. *Microbiology,* 151, 1823-1837.

REDFIELD, R. J. 1993. Genes for Breakfast: The Have-Your-Cake and-Eat-lt-Too of Bacterial Transformation. *Journal of Heredity,* 5, 400-404.

REICHMANN, P. & HAKENBECK, R. 2000. Allelic variation in a peptide-inducible two-component system of Streptococcus pneumoniae. *FEMS Microbiol Lett,* 190, 231-6.

REN, F., TANAKA, H. & YANG, Z. 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Syst Biol,* 54, 808-18.

RICE, P., LONGDEN, I. & BLEASBY, A. 2000. EMBOSS: the European molecular biology open software suite. *Trends in genetics,* 16, 276-277.

ROBERTO SÁNCHEZ, U., FRANCISCO MELO, N. & MARC, A. 2000. Protein structure modeling for structural genomics. *Nature Structural & Molecular Biology,* 7, 986-990.

UNIVERSITI PENDIDIKAN SULTAN IDRIS

ROBINSON, D. F. & FOULDS, L. R. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences,* 53, 131-147.

ROKAS, A. & HOLLAND, P. W. H. 2000. Rave genomic changes as a tool for phylogenetics. *Trends in Ecology & Evolution,* 15, 454-459.

ROKAS, A., WILLIAMS, B. L., KING, N. & CARROLL, S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature,* 425, 798-804.

RONDA, C., GARCIA, J. L. & LOPEZ, R. 1988. Characterization of genetic transformation in Streptococcus oralis NCTC 11427: expression of the pneumococcal amidase in S. oralis using a new shuttle vector. *Mol Gen Genet,* 215, 53-7.

ROSENBERG, N. A. 2002. The probability of topological concordance of gene trees and species trees. *Theor Popul Biol,* 61, 225-47.

ROST, B. 1999. Twilight zone of protein sequence alignments. *Protein Engineering,* 12, 85.

RUTHERFORD, S. T. & BASSLER, B. L. 2012. Bacterial Quorum Sensing: Its Role in Virulence and Possibilities for Its Control. *Cold Spring Harbor Perspectives in Medicine,* 2.

SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol,* 4, 406-25.

SALICHOS, L. & ROKAS, A. 2011. Evaluating Ortholog Prediction Algorithms in a Yeast Model Clade. *PLoS ONE,* 6, e18755.

SANDERSON, M. J., PURVIS, A. & HENZE, C. 1998. Phylogenetic supertrees: Assembling the trees of life. *Trends in Ecology and Evolution,* 13, 105-109.

SCHOULS, L. M., SCHOT, C. S. & JACOBS, J. A. 2003. Horizontal Transfer of Segments of the 16S rRNA Genes between Species of the Streptococcus anginosus Group. *Journal of Bacteriology,* 185, 7241-7246.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

SHIMODAIRA, H. & HASEGAWA, M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution,* 16, 1114.

SHIMODAIRA, H. & HASEGAWA, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics,* 17, 1246-7.

SIEVERS, F., WILM, A., DINEEN, D., GIBSON, T. J., KARPLUS, K., LI, W., LOPEZ, R., MCWILLIAM, H., REMMERT, M., SODING, J., THOMPSON, J. D. & HIGGINS, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol,* 7, 539.

SLATKIN, M. & POLLACK, J. L. 2006. The Concordance of Gene Trees and Species Trees at Two Linked Loci. *Genetics,* 172, 1979-1984.

SMITH, T. F. & WATERMAN, M. S. 1981. Identification of common molecular subsequences. *J Mol Biol,* 147, 195-7.

STAJICH, J., BLOCK, D., BOULEZ, K., BRENNER, S., CHERVITZ, S., DAGDIGIAN, C., FUELLEN, G., GILBERT, J., KORF, I. & LAPP, H. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome research,* 12, 1611.

STAMATAKIS, A. 2006a. Phylogenetic Models of Rate Heterogeneity: A High Performance Computing Perspective. *Proc. of IPDPS2006, HICOMB Workshop, Proceedings on CD, Rhodos, Greece*.

STAMATAKIS, A. 2006b. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics,* 22, 2688-2690.

STAMATAKIS, A. & ALACHIOTIS, N. 2010. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics,* 26, i132-i139.

STEINER, K. & MALKE, H. 2001. relA-Independent Amino Acid Starvation Response Network of Streptococcus pyogenes. *Journal of bacteriology,* 183, 7354-7364.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

STEINER, K. & MALKE, H. 2002. Dual Control of Streptokinase and Streptolysin S Production by the covRS and fasCAX Two-Component Regulators in Streptococcus dysgalactiae subsp. equisimilis. *Infection and Immunity,* 70, 3627-3636.

STOCK, A. M., ROBINSON, V. L. & GOUDREAU, P. N. 2000. Two-component signal transduction. *Annu Rev Biochem,* 69, 183-215.

SULLIVAN, J. 1996. Combining Data with Different Distributions of Among-Site Rate Variation. *Systematic Biology,* 45, 375-380.

SUNG, C. K. & MORRISON, D. A. 2005. Two distinct functions of ComW in stabilization and activation of the alternative sigma factor ComX in Streptococcus pneumoniae. *Journal of Bacteriology,* 187, 3052-3061.

SUNITA, S., PURTA, E., DURAWA, M., TKACZUK, K. L., SWAATHI, J., BUJNICKI, J. M. & SIVARAMAN, J. 2007. Functional specialization of domains tandemly duplicated within 16S rRNA methyltransferase RsmC. *Nucleic Acids Research,* 35, 4264-4274.

SWOFFORD, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? *Miyamoto, M. M. And J. Cracraft (Ed.). Phylogenetic Analysis of DNA Sequences; Symposium on Recent Advances in Phylogenetic Studies of DNA Sequences, Boston, Massachusetts, USA, December 26-30, 1989. X+358p. Oxford University Press, Inc.: New York, New York, USA; Oxford, England, Uk. Illus*, 295-333.

TAGG, J., WESCOMBE, P. & BURTON, J. 2006. Oral streptococcal BLIS: Heterogeneity of the effector molecules and potential role in the prevention of streptococcal infections. *International Congress Series,* 1289, 347-350.

TAMURA, S., YONEZAWA, H., MOTEGI, M., NAKAO, R., YONEDA, S., WATANABE, H., YAMAZAKI, T. & SENPUKU, H. 2009. Inhibiting effects of Streptococcus salivarius on competence-stimulating peptide-dependent biofilm formation by Streptococcus mutans. *Oral Microbiol Immunol,* 24, 152-61.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

TATUSOV, R. L., GALPERIN, M. Y., NATALE, D. A. & KOONIN, E. V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research,* 28, 33-36.

TATUSOV, R. L., KOONIN, E. V. & LIPMAN, D. J. 1997. A Genomic Perspective on Protein Families. *Science,* 278, 631-637.

TATUSOV, R. L., MUSHEGIAN, A. R., BORK, P., BROWN, N. P., HAYES, W. S., BORODOVSKY, M., RUDD, K. E. & KOONIN, E. V. 1996. Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. *Current Biology,* 6, 279-291.

TAYLOR, J. S. & RAES, J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet,* 38, 615-43.

THANASSI, J. A., HARTMAN-NEUMANN, S. L., DOUGHERTY, T. J., DOUGHERTY, B. A. & PUCCI, M. J. 2002. Identification of 113 conserved essential genes using a high-throughput gene disruption system in Streptococcus pneumoniae. *Nucleic Acids Res,* 30, 3152-62.

THOMAS, C. M. & NIELSEN, K. M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol,* 3, 711-21.

TOMASZ, A. 1965. Control of the competent state in Pneumococcus by a hormone-like cell product: an example for a new type of regulatory mechanism in bacteria. *Nature,* 208, 155-159.

TOMASZ, A. & HOTCHKISS, R. D. 1964. REGULATION OF THE TRANSFORMABILITY OF PNEUMOCOCCAL CULTURES BY MACROMOLECULAR CELL PRODUCTS. *Proceedings of the National Academy of Sciences,* 51, 480-487.

TOMII, K. & KANEHISA, M. 1998. A Comparative Analysis of ABC Transporters in Complete Microbial Genomes. *Genome Research,* 8, 1048-1059.

TORTOSA, P. & DUBNAU, D. 1999. Competence for transformation: a matter of taste. *Current Opinion in Microbiology,* 2, 588-592.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

UBERBACHER, E. C., XU, Y. & MURAL, R. J. 1996. Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol,* 266, 259-81.

UNIVERSITY OF WARWICK 2008. How Antibiotic Resistant Bugs Became Resistant To Penicillin, And How Penicillin Could Work Again. *ScienceDaily.*

VAN DER GIEZEN, M., TOVAR, J. & CLARK, C. 2005. Mitochondrion-derived organelles in protists and fungi. *Int Rev Cytol,* 244, 175 - 225.

VAN DER PLOEG, J. R. 2005. Regulation of bacteriocin production in Streptococcus mutans by the quorum-sensing system required for development of genetic competence. *J Bacteriol,* 187, 3980-9.

VICKERMAN, M. M., IOBST, S., JESIONOWSKI, A. M. & GILL, S. R. 2007. Genome-wide transcriptional changes in Streptococcus gordonii in response to competence signaling peptide. *J Bacteriol,* 189, 7799-807.

VUONG, C., GERKE, C., SOMERVILLE, G. A., FISCHER, E. R. & OTTO, M. 2003. Quorum-sensing control of biofilm factors in Staphylococcus epidermidis. *J Infect Dis,* 188, 706-18.

WANG, B. Y. & KURAMITSU, H. K. 2005. Interactions between oral bacteria: inhibition of Streptococcus mutans bacteriocin production by Streptococcus gordonii. *Appl Environ Microbiol,* 71, 354-62.

WANG, R. C., SEROR, S. J., BLIGHT, M., PRATT, J. M., BROOME-SMITH, J. K. & HOLLAND, I. B. 1991. Analysis of the membrane organization of an Escherichia coli protein translocator, HlyB, a member of a large family of prokaryote and eukaryote surface transport proteins. *J Mol Biol,* 217, 441-54.

WARD, P. N. & LEIGH, J. A. 2004. Genetic analysis of Streptococcus uberis plasminogen activators. *Indian J Med Res,* 119 Suppl, 136-40.

WATANABE, H., MORI, H., ITOH, T. & GOJOBORI, T. 1997. Genome plasticity as a paradigm of eubacteria evolution. *J Mol Evol,* 44, S57 - S64.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

WATSON, J. D. & CRICK, F. H. C. 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature,* 171, 737-738.

WEEN, O., TEIGEN, S., GAUSTAD, P., KILIAN, M. & HÅVARSTEIN, L. S. 2002. Competence without a Competence Pheromone in a Natural Isolate of Streptococcus infantis. *Journal of Bacteriology,* 184, 3426-3432.

WHATMORE, A., BARCUS, V. & DOWSON, C. 1999. Genetic diversity of the streptococcal competence (com) gene locus. *Journal of Bacteriology,* 181, 3144.

WHELAN, S. & GOLDMAN, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution,* 18, 691.

WHELAN, S., LIO, P. & GOLDMAN, N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics,* 17, 262-272.

WHITFORD, M. F., MCPHERSON, M. A., FORSTER, R. J. & TEATHER, R. M. 2001. Identification of bacteriocin-like inhibitors from rumen Streptococcus spp. and isolation and characterization of bovicin 255. *Appl Environ Microbiol,* 67, 569-74.

WILLIAMS, G. 2002. EMBOSS Transeq.

WIRAWAN, R. E., JACK, R. W. & TAGG, J. R. 2006. A novel combination of bacteriocins produced by Streptococcus uberis strain 42. *International Congress Series,* 1289, 355-358.

WOESE, C. R. & FOX, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A,* 74, 5088-90.

WOESE, C. R., OLSEN, G. J., IBBA, M. & SÖLL, D. 2000. Aminoacyl-tRNA Synthetases, the Genetic Code, and the Evolutionary Process. *Microbiology and Molecular Biology Reviews,* 64, 202-236.

WOJCIECHOWSKI, M. F., HOELZER, M. A. & MICHOD, R. E. 1989. DNA-repair and the evolution of transformation in bacillus-subtilis .2. Role of inducible repair. *Genetics,* 121, 411-422.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

WOLF, Y., ROGOZIN, I., KONDRASHOV, A. & KOONIN, E. 2001a. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res,* 11, 356 - 372.

WOLF, Y. I., ROGOZIN, I. B., GRISHIN, N. V. & KOONIN, E. V. 2002. Genome trees and the Tree of Life. *Trends in Genetics,* 18, 472-479.

WOLF, Y. I., ROGOZIN, I. B., GRISHIN, N. V., TATUSOV, R. L. & KOONIN, E. V. 2001b. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol,* 1, 8.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol,* 39, 306-14.

YANG, Z. 1996a. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol,* 11, 367-72.

YANG, Z. 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution,* 42, 587-596.

YANG, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution,* 24, 1586-1591.

ZACCARIA, E., VAN BAARLEN, P., DE GREEFF, A., MORRISON, D. A., SMITH, H. & WELLS, J. M. 2014. Control of Competence for DNA Transformation in <italic>Streptococcus suis</italic> by Genetically Transferable Pherotypes. *PLoS ONE,* 9, e99394.

ZUCKERKA, E. & PAULING, L. 1965. Molecules as documents of evolutionary history. *Journal of Theoretical Biology,* 8, 357-&.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

APPENDICES

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

APPENDIX A

The list of the early quorum sensing Com homologs in the Streptococcus species used as queries in the analysis.

Species	Genome accession number	RefSeq accession number	Gene
Streptococcus mutans	NC_004350	NP_720745	A
		NP_722191	A
		NP_722205	A
		NP_720746	B
		NP_722207	B
		NP_722220	C
		NP_722221	D
		NP_722222	E
		NP_722295	X
Streptococcus thermophilus	NC_006448	YP_140107	A
		YP_140105	B
		YP_140106	B
		YP_140104	C
		YP_140103	D
		YP_140102	E
		YP_138670	X
Streptococcus pneumoniae	NC_011072	YP_002036763	A
		YP_002037195	A
		YP_002036764	B
		YP_002037194	B
		YP_002037193	C
		YP_002038833	C
		YP_002037192	D
		YP_002038832	D
		YP_002037191	E
		YP_002038831	E
		YP_002036735	X
		YP_002038589	X

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Appendix A – Continue from previous page

Streptococcus equi	NC_011134	YP_002122906	A
		YP_002123869	A
		YP_002122907	B
		YP_002123870	B
		YP_002123555	C
		YP_002123871	C
		YP_002123872	D
		YP_002123335	E
		YP_002123873	E
		YP_002124033	X
		YP_002124073	X
Streptococcus suis	NC_012926	YP_003027803	X
Streptococcus gallolyticus	NC_013798	YP_003431431	A
		YP_003431432	B
		YP_003431429	C
		YP_003429675	X

UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID

APPENDIX B

The 86 subjected fully sequenced genomes of the 21 Streptococcus species retrieved from the NCBI Genome database.

	Species	Strain	Accession number
1	*S. agalactiae*	2603V/R	NC_004116
2	*S. agalactiae*	A909	NC_007432
3	*S. agalactiae*	NEM316	NC_004368
4	*S. dysgalactiae*	ATCC27957	CM_001076
5	*S. dysgalactiae*	ATCC12394	CP_002215
6	*S. dysgalactiae*	GGS_124	NC_012891
7	*S. equi*	4047	NC_012471
8	*S. equi*	zooepidemicus	NC_012470
9	*S. equi*	ATCC35246	CP_002904
10	*S. equi*	MGCS10565	NC_011134
11	*S. gallolyticus*	ATCC43143	AP_012053
12	*S. gallolyticus*	ATCCBAA2069	NC_015215
13	*S. gallolyticus*	UCN34	NC_013798
14	*S. gordonii*	CH1	NC_009785
15	*S. infantarius*	CJ18	NC_016826
16	*S. macedonicus*	ACA-DC198	NC_016749
17	*S. mitis*	B6	NC_013853
18	*S. mutans*	NN2025	NC_013928
19	*S. mutans*	UA159	NC_004350
20	*S. oralis*	Uo5	NC_015291
21	*S. parasanguinis*	ATCC15912	NC_015678
22	*S. parauberis*	KCTC11537	NC_015558
23	*S. pasteurianus*	ATCC43144	NC_015600
24	*S. pneumoniae*	670-6B	NC_014498
25	*S. pneumoniae*	70585	NC_012468
26	*S. pneumoniae*	AP200	NC_014494
27	*S. pneumoniae*	ATCC700669	NC_011900
28	*S. pneumoniae*	CGSP14	NC_010582
29	*S. pneumoniae*	D39	NC_008533
30	*S. pneumoniae*	G54	NC_011072
31	*S. pneumoniae*	Hungary19A-6	NC_010380
32	*S. pneumoniae*	INV104	FQ_312030
33	*S. pneumoniae*	INV200	FQ_312029
34	*S. pneumoniae*	JJA	NC_012466
35	*S. pneumoniae*	OXC141	FQ_312027
36	*S. pneumoniae*	P1031	NC_012467

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

Appendix B – Continue from previous page

37	*S. pneumoniae*	R6	NC_003098
38	*S. pneumoniae*	SPN032672	FQ_312039
39	*S. pneumoniae*	SPN033038	FQ_312042
40	*S. pneumoniae*	SPN034156	FQ_312045
41	*S. pneumoniae*	SPN034183	FQ_312043
42	*S. pneumoniae*	SPN994038	FQ_312041
43	*S. pneumoniae*	SPN994039	FQ_312044
44	*S. pneumoniae*	ST556	CP_003357
45	*S. pneumoniae*	Taiwan19F-14	NC_012469
46	*S. pneumoniae*	TCH8431/19A	NC_014251
47	*S. pneumoniae*	TIGR4	NC_003028
48	*S. pseudopneumoniae*	IS7493	NC_015875
49	*S. pyogenes*	Alab49	CP_003068
50	*S. pyogenes*	M1GAS	NC_002737
51	*S. pyogenes*	MGAS10270	NC_008022
52	*S. pyogenes*	MGAS10394	NC_006086
53	*S. pyogenes*	MGAS10750	NC_008024
54	*S. pyogenes*	MGAS15252	NC_017040
55	*S. pyogenes*	MGAS1882	CP_003121
56	*S. pyogenes*	MGAS2096	NC_008023
57	*S. pyogenes*	MGAS315	NC_004070
58	*S. pyogenes*	MGAS5005	NC_007297
59	*S. pyogenes*	MGAS6180	NC_007296
60	*S. pyogenes*	MGAS8232	NC_003485
61	*S. pyogenes*	MGAS9429	NC_008021
62	*S. pyogenes*	NZ131	NC_011375
63	*S. pyogenes*	SSI-1	NC_004606
64	*S. pyogenes*	Manfredo	NC_009332
65	*S. salivarius*	JIM8777	FR_873482
66	*S. salivarius*	JIM8780	NC_015760
67	*S. sanguinis*	SK36	NC_009009
68	*S. suis*	05ZYH33	NC_009442
69	*S. suis*	98HAH33	NC_009443
70	*S. suis*	A7	CP_002570
71	*S. suis*	BM407	NC_012926
72	*S. suis*	D12	CP_002644
73	*S. suis*	D9	CP_002641
74	*S. suis*	GZ1	CP_000837
75	*S. suis*	JS14	CP_002465
76	*S. suis*	P1/7	NC_012925
77	*S. suis*	SC84	NC_012924

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Appendix B – Continue from previous page

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

78	*S. suis*	SS12	CP_002640
79	*S. suis*	ST1	CP_002651
80	*S. suis*	ST3	NC_015433
81	*S. thermophilus*	CNRZ1066	NC_006449
82	*S. thermophilus*	JIM8232	FR_875178
83	*S. thermophilus*	LMD-9	NC_008532
84	*S. thermophilus*	LMG18311	NC_006448
85	*S. thermophilus*	ND03	CP_002340
86	*S. uberis*	0140J	NC_012004

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

Natural transformable streptococcal species.

	Species	Phylum	References
1	*Streptococcus pneumoniae*	Firmicutes	(Dawson and Sia, 1931)
2	*Streptococcus mitis*		(Bensing et al., 2001)
3	*Streptococcus oralis*		(Ronda et al., 1988)
4	*Streptococcus crista*		(Correia et al., 1996)
5	*Streptococcus infantis*		(Ween et al., 2002)
6	*Streptococcus gordonii*		(Pakula and Walczak, 1963)
7	*Streptococcus sanguinis*		(Macrina et al., 1980, Havarstein et al., 1997)
8	*Streptococcus anginosus*		(Havarstein et al., 1997)
9	*Streptococcus intermedius*		(Havarstein et al., 1997)
10	*Streptococcus constellatus*		(Havarstein et al., 1997)
11	*Streptococcus themophilus*		(Blomqvist et al., 2006)
12	*Streptococcus bovis*		(Mercer et al., 1999)
13	*Streptococcus mutans*		(Perry and Kuramitsu, 1981)
14	*Streptococcus salivarius*		(Fontaine et al., 2010)
15	*Streptococcus macedonicus*		(Morrison et al., 2013)
16	*Streptococcus suis*		(Zaccaria et al., 2014)
17	*Streptococcus infantarius*		(Morrison et al., 2013)

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

APPENDIX D

The presence and absence of the quorum sensing (QS) systems: Fibronectin/fibrinogen-binding/haemolytic-activity/streptokinase-regulator (Fas) QS, Bacteriocin-like peptide (Blp) QS and Competence (Com) QS, in *Streptococcus* species. The present of a QS system in a species is based from the experimental study(s) that is indicated by a reference number in a bracket. The mark of '-' indicates that the QS system is absent or has yet to be found in the species.

Species	Fas QS	Blp QS	Com QS
S. pyogenes	[3]	[14]	-
S. dysgalactiae	[4]	[6]	-
S. equi	[20]	[15]	-
S. uberis	[18, 19]	[10]	-
S. parauberis	-	-	-
S. salivarius	-	[9]	[16]
S. thermophilus	-	[17]	[1, 16]
S. gallolyticus	-	[21, 22]	-
S. pasteurianus	-	-	-
S. macedonicus	-	[8]	[11]
S. mutans	-	[2]	[1]
S. pneumoniae	-	[5]	[1]
S. pseudopneumoniae	-	-	-
S. mitis	-	-	[1]
S. oralis	-	-	[1]
S. gordonii	-	[12]	[1]
S. sanguinis	-	-	[1]
S. parasanguinis	-	-	-
S. suis	-	[7]	[13]
S. agalactiae	-	-	-
S. infantarius	-	-	[11]

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

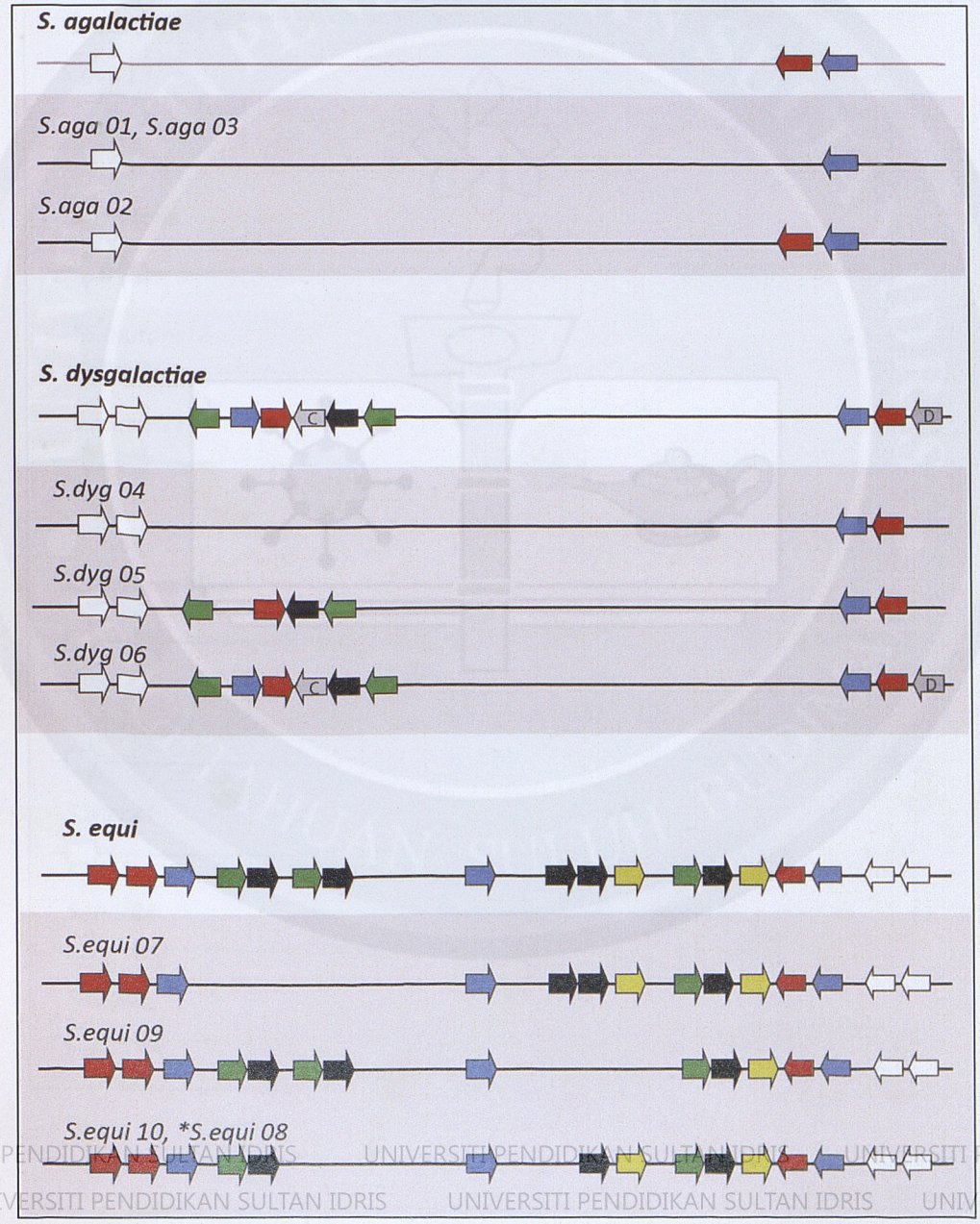
Appendix D – Continue from previous page

Source of references:

[1] (Johnsborg et al., 2007)
[2] (Balakrishnan et al., 2002)
[3] (Chaussee et al., 2002)
[4] (Davies et al., 2007)
[6] (Heng et al., 2006)
[7] (LeBel et al., 2014)
[8] (Maragkoudakis et al., 2009)
[9] (Tamura et al., 2009)
[10] (Wirawan et al., 2006)
[11] (Morrison et al., 2013)
[13] (Zaccaria et al., 2014)
[14] (Tagg et al., 2006)
[15] (Gargis et al., 2009)
[16] (Fontaine et al., 2010)
[17] (Fontaine et al., 2007)
[18] (Ward and Leigh, 2004)
[19] (Khan et al., 2003)
[20] (Holden et al., 2009)
[21] (Whitford et al., 2001)
[22] (Cookson et al., 2004)

UNIVERSITI PENDIDIKAN SULTAN IDRIS

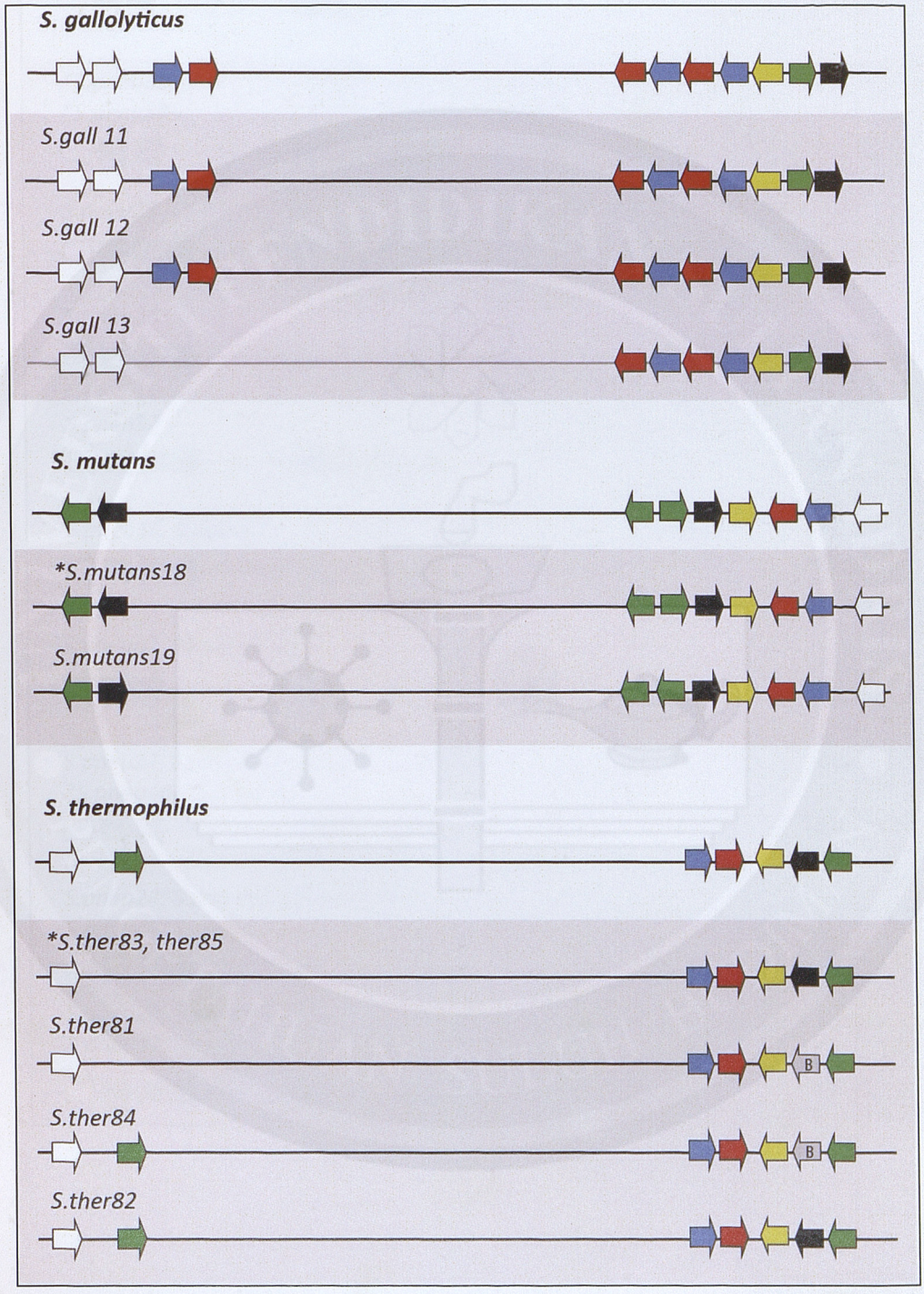
Quorum sensing genes in 86 genomes of 21 *Streptococcus* species. Each gene is mapped to their synteny in the genome. A pan genome is used to represent the QS genes present in the genomes of each species (white box). The genes are coloured according to their gene families: ABC exporter – *A* gene (green); Accessory protein – *B* gene (black); Competence simulating peptide – *C* gene (yellow); Histidine kinase – *D* gene (red); Response regulator – *E* gene (blue); ComX – *X* gene (white). The genes that have missed SUCRE detection are coloured in grey.



UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

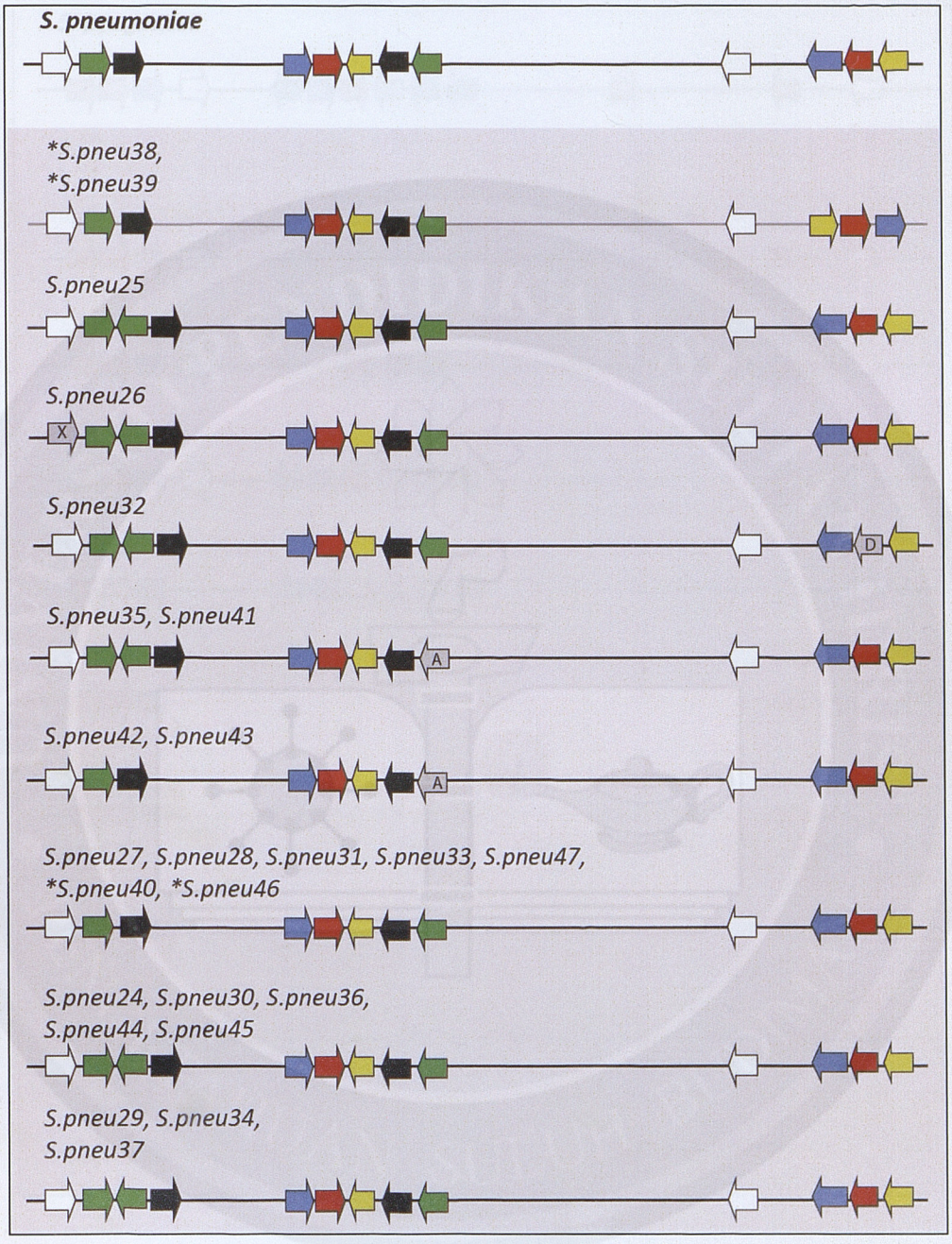
Appendix E – Continue from previous page



UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
ORIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Appendix E – Continue from previous page

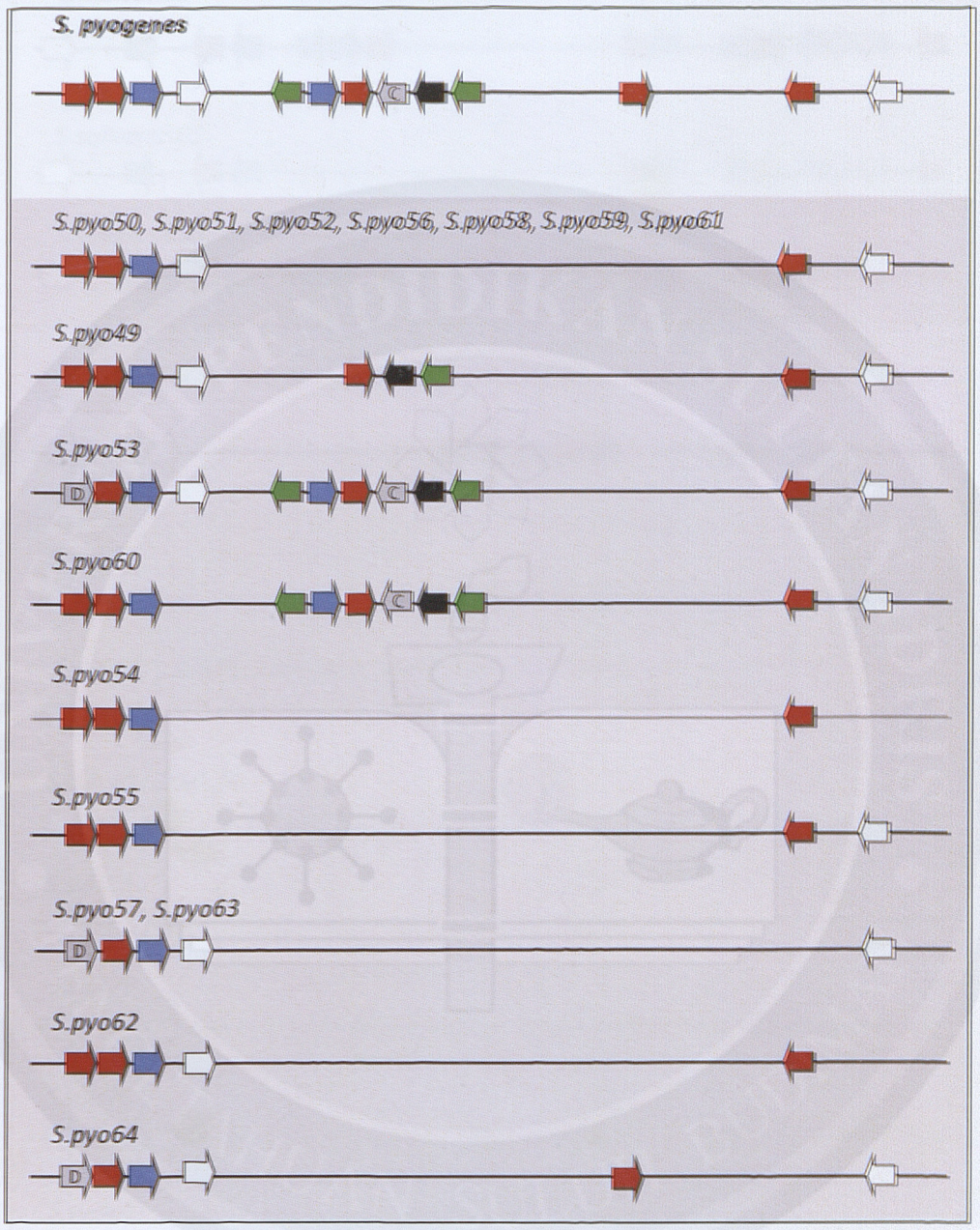
UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI P



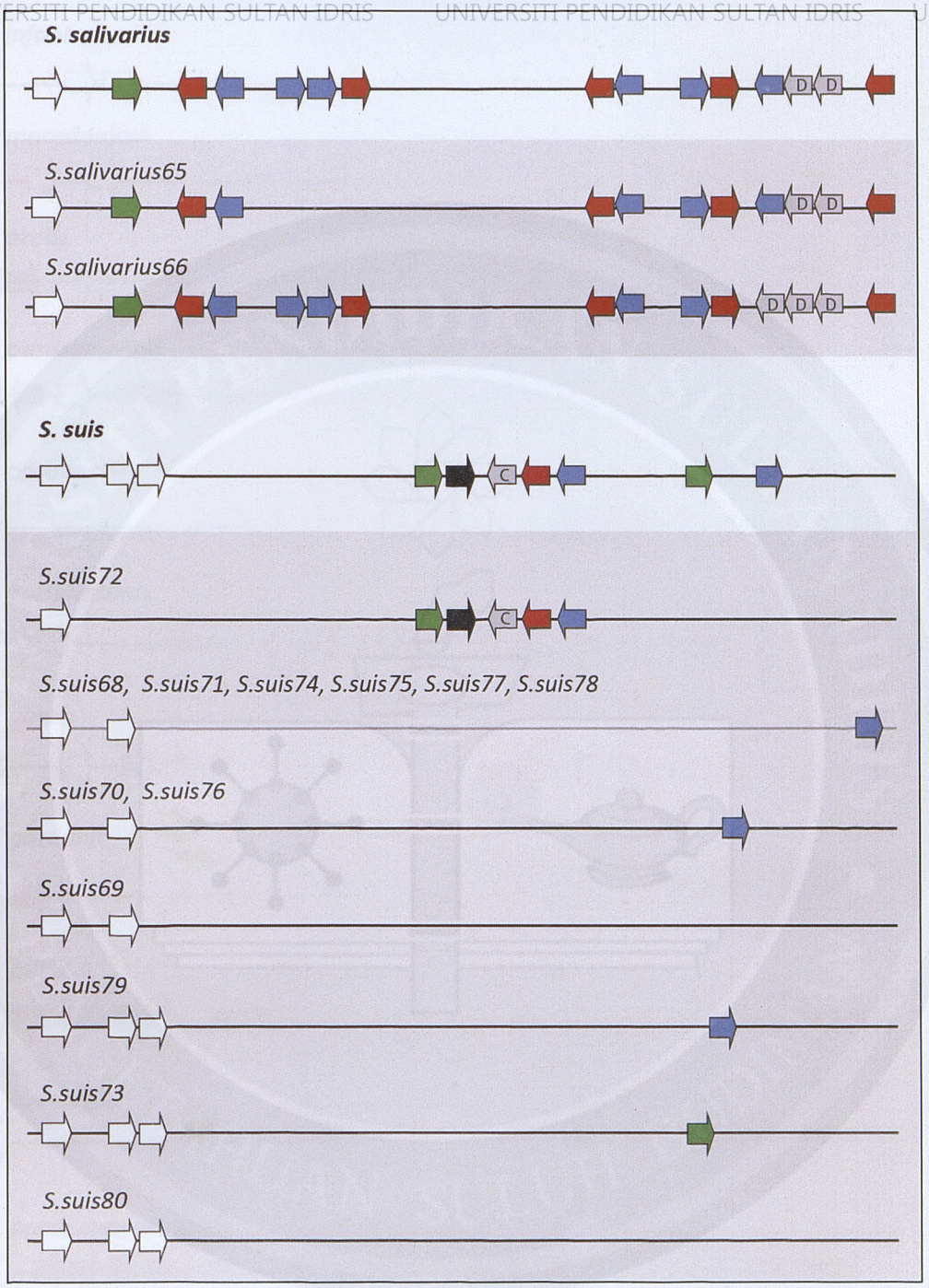
UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

Appendix E – Continue from previous page

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

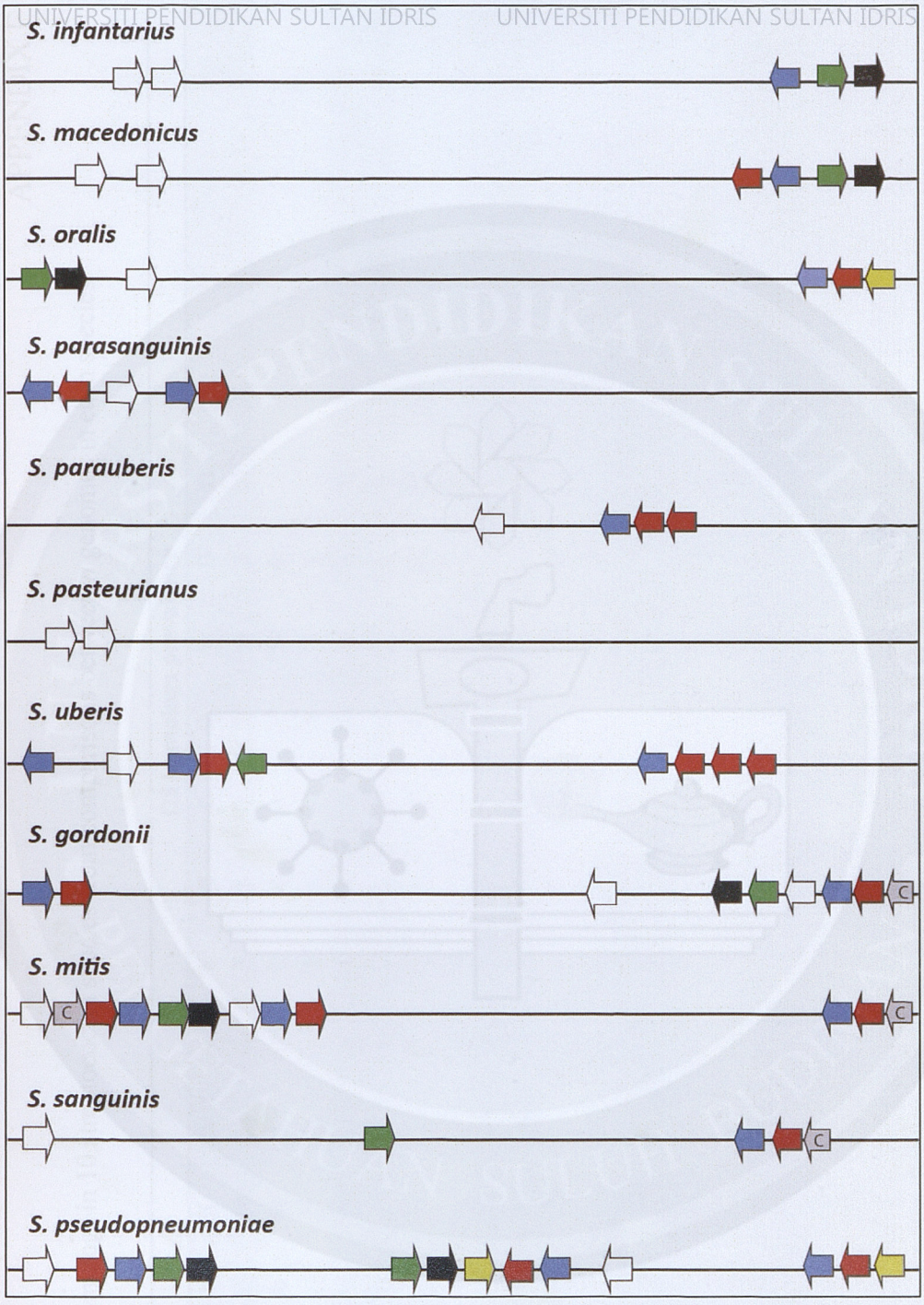


UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN



UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F



UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

APPENDIX F

The prevalence of QS homologs in 10 streptococcal species from comparisons between *n* genomes in each species.

Species Name	QS homologs prevalence																	
S. pyogenes	D	D	E	X	A	E	D	C	B	A	D	D	X					
(*n*= 16)	16	16	16	14	2	2	3	2	3	3	2	11	14					
S. dysgalactiae	X	X	A	E	D	C	B	A	E	D	D							
(*n*= 3)	3	3	2	1	2	2	2	2	3	3	1							
S. equi	D	D	E	A	B	A	B	E	B	B	C	A	B	C	D	E	X	X
(*n*= 4)	4	4	4	3	3	1	1	4	1	1	2	4	4	4	4	4	4	4
S. salivarius	X	A	D	E	E	E	D	D	E	E	D	E	D	D				
(*n*= 2)	2	2	2	2	1	1	1	2	2	2	2	1	2	2				
S. thermophilus	X	A	E	D	C	B	A											
(*n*= 5)	5	2	5	5	5	5	5											
S. gallolyticus	X	X	E	D	D	E	D	E	C	A	B							
(*n*= 3)	3	3	2	2	3	3	3	3	3	3	3							
S. mutans	A	B	A	A	B	C	D	E	X									
(*n*= 2)	2	2	2	2	2	2	2	2	2									
S. pneumoniae	X	A	B	E	D	C	B	A	X	E	D	C						
(*n*= 24)	24	24	24	24	24	24	24	24	24	24	24	24						
S. suis	X	X	X	A	B	C	D	E	E									
(*n*= 13)	13	11	3	1	1	1	1	1	10									
S. agalactiae	X	D	E															
(*n* = 3)	3	1	3															

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

APPENDIX G

Table G.1: List of 58 genomes in 21 Streptococcus species.

No	SpID	Species and strain name	TaxaID
1	ag06	*Streptococcus agalactiae* NEM316	211110
2	ag29	*Streptococcus agalactiae* 2603V/R	208435
3	ag44	*Streptococcus agalactiae* A909	205921
4	dy19	*Streptococcus dysgalactiae subsp. equisimilis* GGS_124	486410
5	eq07	*Streptococcus equi subsp. equi* 4047	553482
6	eq18	*Streptococcus equi subsp. zooepidemicus*	40041
7	eq51	*Streptococcus equi subsp. zooepidemicus* MGCS10565	552526
8	ga02	*Streptococcus gallolyticus* UCN34	637909
9	ga56	*Streptococcus gallolyticus subsp. gallolyticus* ATCC BAA-2069	990317
10	go35	*Streptococcus gordonii str. Challis substr.* CH1	467705
11	ma42	*Streptococcus macedonicus* ACA-DC 198	1116231
12	mi41	*Streptococcus mitis* B6	365659
13	mu03	*Streptococcus mutans* UA159	210007
14	mu57	*Streptococcus mutans* NN2025	511691
15	or37	*Streptococcus oralis* Uo5	927666
16	pa04	*Streptococcus parasanguinis* ATCC 15912	760570
17	pa31	*Streptococcus parauberis* KCTC 11537	936154
18	pa52	*Streptococcus pasteurianus* ATCC 43144	981540
19	pn01	*Streptococcus pneumoniae* 70585	488221
20	pn08	*Streptococcus pneumoniae* Taiwan19F-14	487213
21	pn09	*Streptococcus pneumoniae* TIGR4	170187
22	pn10	*Streptococcus pneumoniae* D39	373153
23	pn12	*Streptococcus pneumoniae* R6	171101
24	pn16	*Streptococcus pneumoniae* 670-6B	189423
25	pn17	*Streptococcus pneumoniae* TCH8431/19A	525381
26	pn30	*Streptococcus pneumoniae* JJA	488222
27	pn32	*Streptococcus pneumoniae* AP200	574093
28	pn38	*Streptococcus pneumoniae* CGSP14	516950
29	pn43	*Streptococcus pneumoniae* P1031	488223
30	pn46	*Streptococcus pneumoniae* Hungary19A-6	487214
31	pn47	*Streptococcus pneumoniae* G54	512566
32	pn50	*Streptococcus pneumoniae* ATCC 700669	561276
33	ps15	*Streptococcus pseudopneumoniae* IS7493	1054460
34	py11	*Streptococcus pyogenes str.* Manfredo	160491
35	py13	*Streptococcus pyogenes* SSI-1	193567
36	py22	*Streptococcus pyogenes* MGAS10750	370554
37	py24	*Streptococcus pyogenes* M1 GAS	160490
38	py25	*Streptococcus pyogenes* MGAS10394	286636
39	py26	*Streptococcus pyogenes* MGAS2096	370553

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS

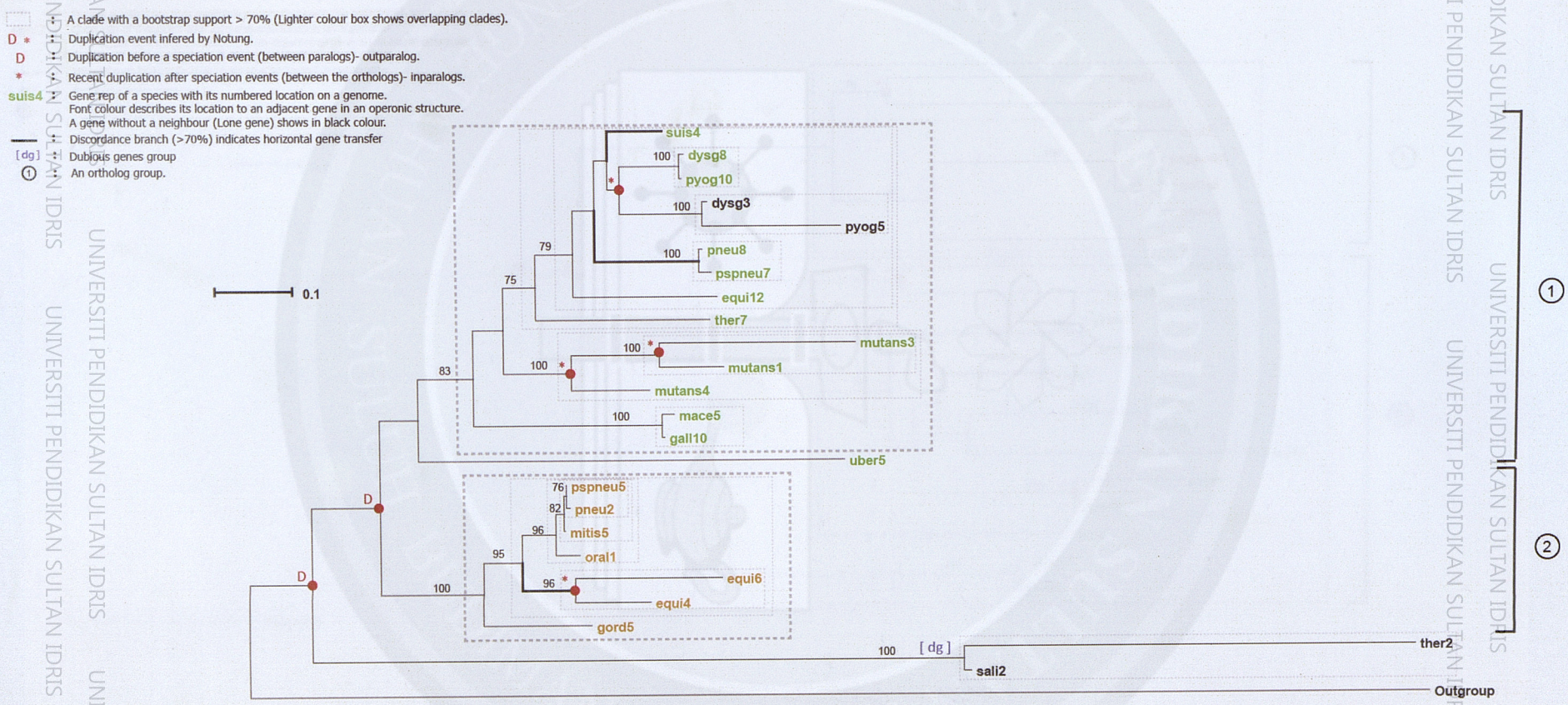
UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

Table G.1 – Continue from previous page

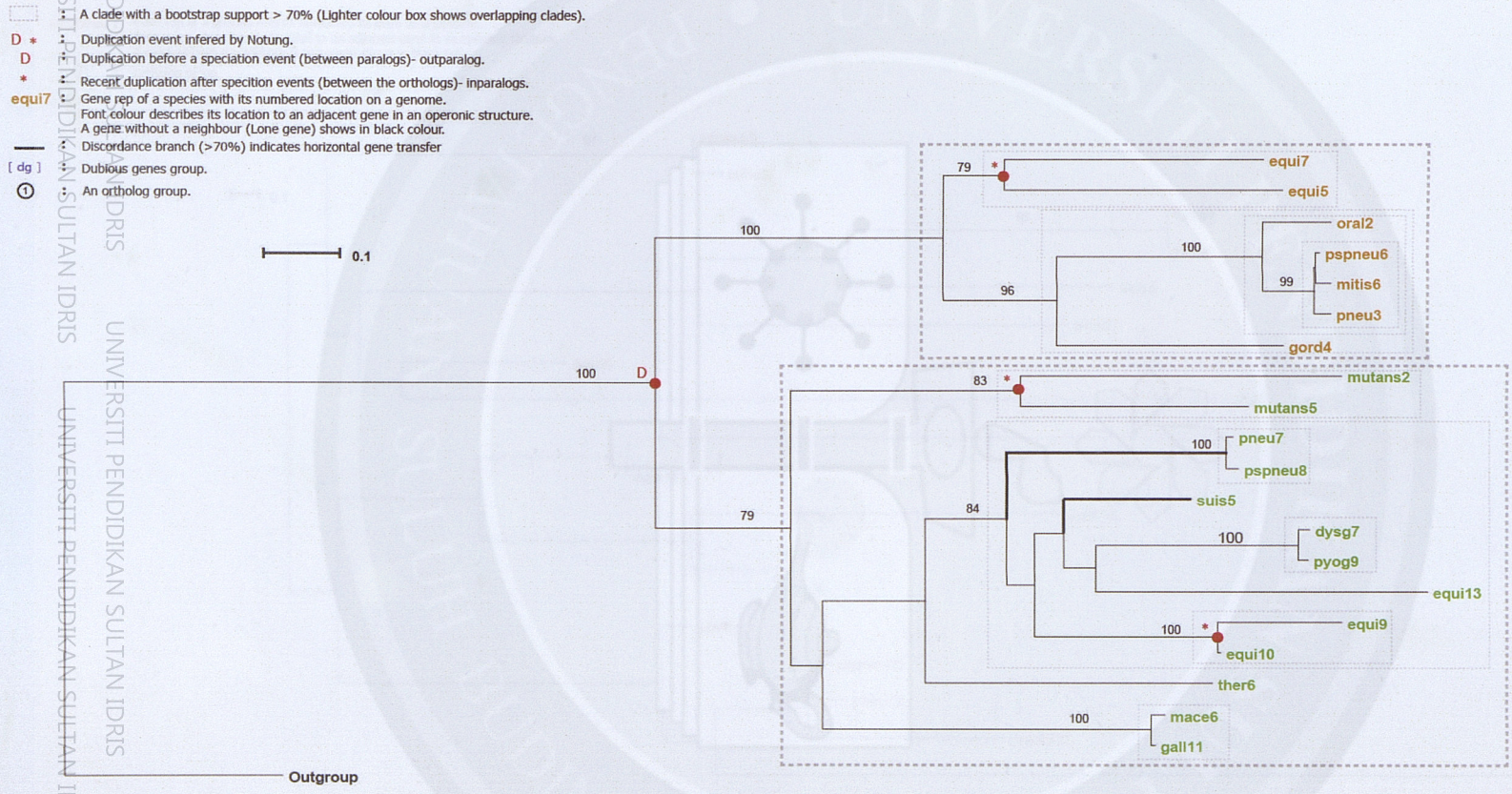
40	py27	*Streptococcus pyogenes* MGAS10270	370552
41	py33	*Streptococcus pyogenes* MGAS315	198466
42	py34	*Streptococcus pyogenes* MGAS8232	186103
43	py45	*Streptococcus pyogenes* MGAS5005	293653
44	py48	*Streptococcus pyogenes* NZ131	471876
45	py49	*Streptococcus pyogenes* MGAS9429	370551
46	py58	*Streptococcus pyogenes* MGAS6180	319701
47	sa39	*Streptococcus sanguinis* SK36	388919
48	sa54	*Streptococcus salivarius* CCHSS3	1048332
49	su05	*Streptococcus suis* 98HAH33	391296
50	su14	*Streptococcus suis* BM407	568814
51	su20	*Streptococcus suis* SC84	568813
52	su21	*Streptococcus suis* ST3	1007064
53	su36	*Streptococcus suis* P1/7	218494
54	su53	*Streptococcus suis* 05ZYH33	391295
55	th23	*Streptococcus thermophilus* CNRZ1066	299768
56	th40	*Streptococcus thermophilus* LMD-9	322159
57	th55	*Streptococcus thermophilus* LMG 18311	264199
58	ub28	*Streptococcus uberis* 0140J	218495

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

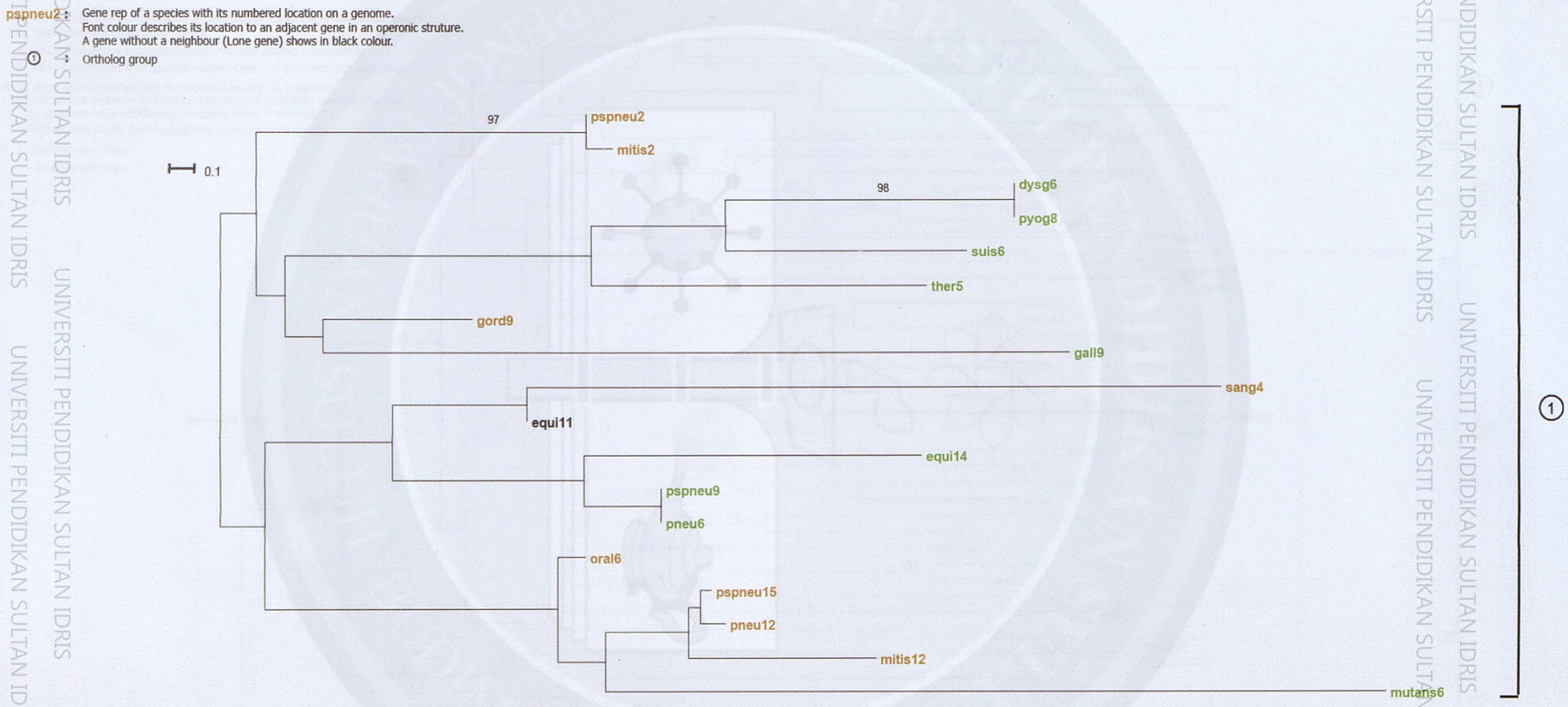
Appendix H.1: The ABC exporter (*A* gene) maximum likelihood gene tree.



Appendix H.2: The accessory protein (*B* gene) maximum likelihood gene tree.

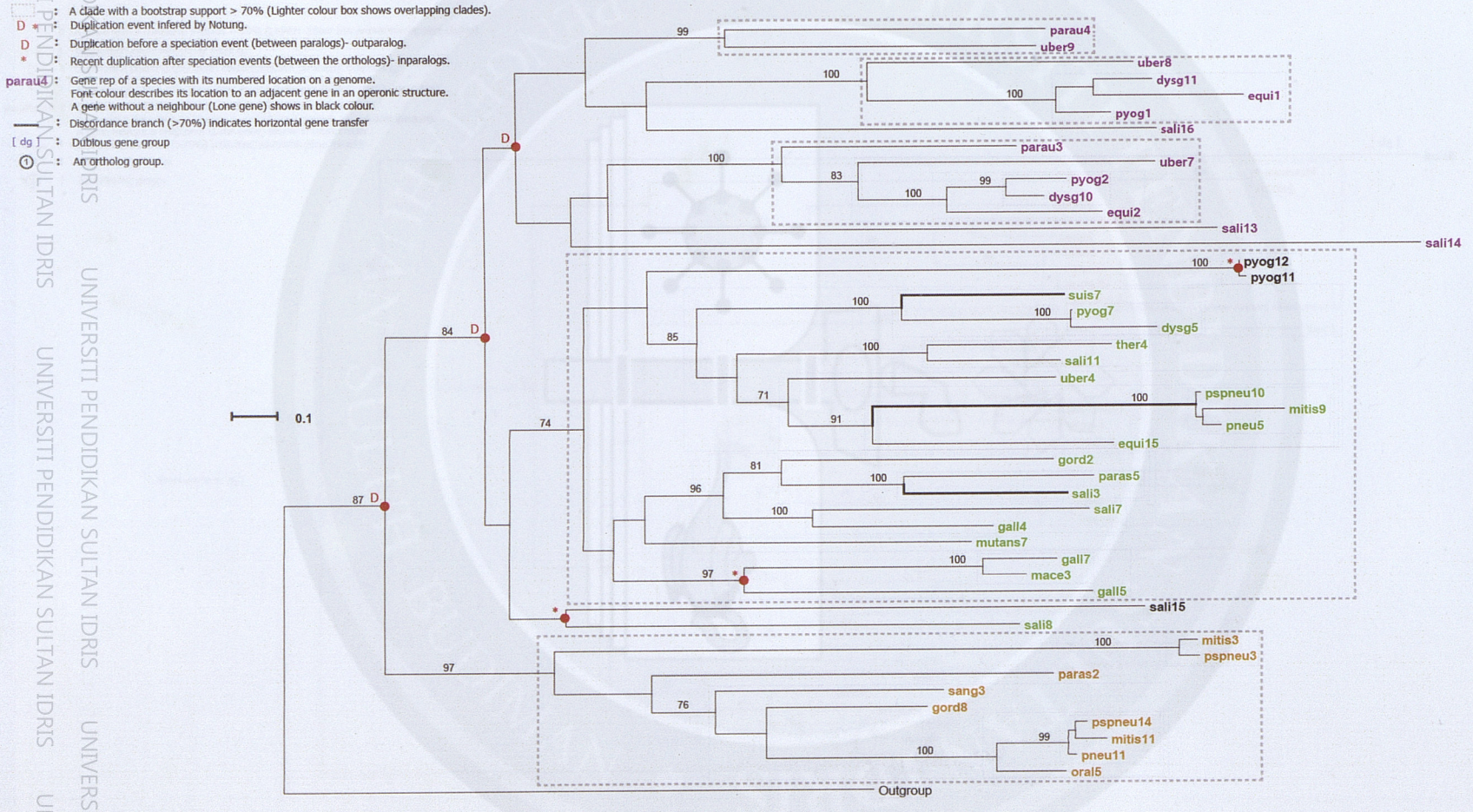


Appendix H.3: The accessory protein (*C* gene) maximum likelihood gene tree.

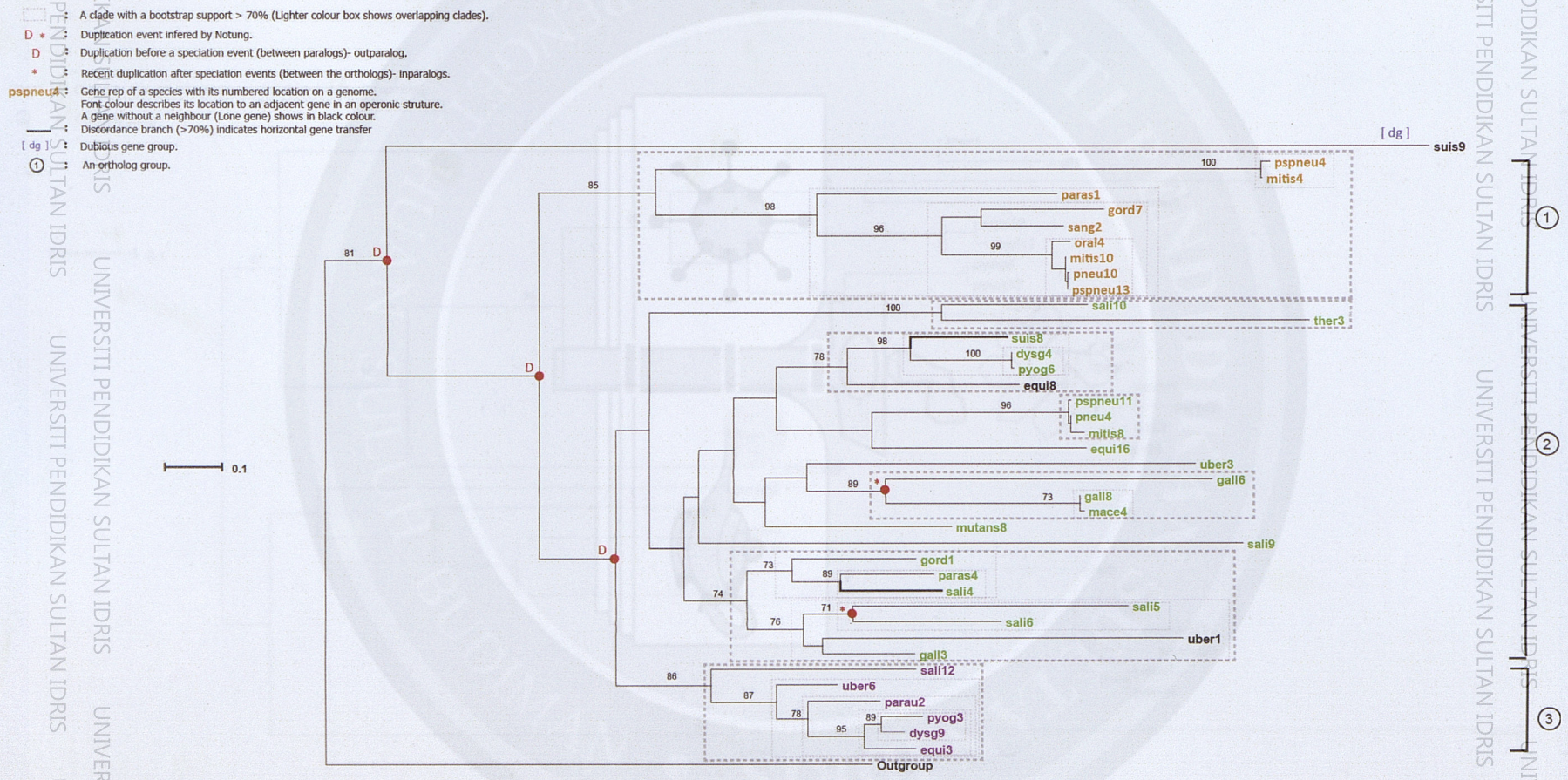


UNIVERSITI PENDIDIKAN SULTAN IDRIS

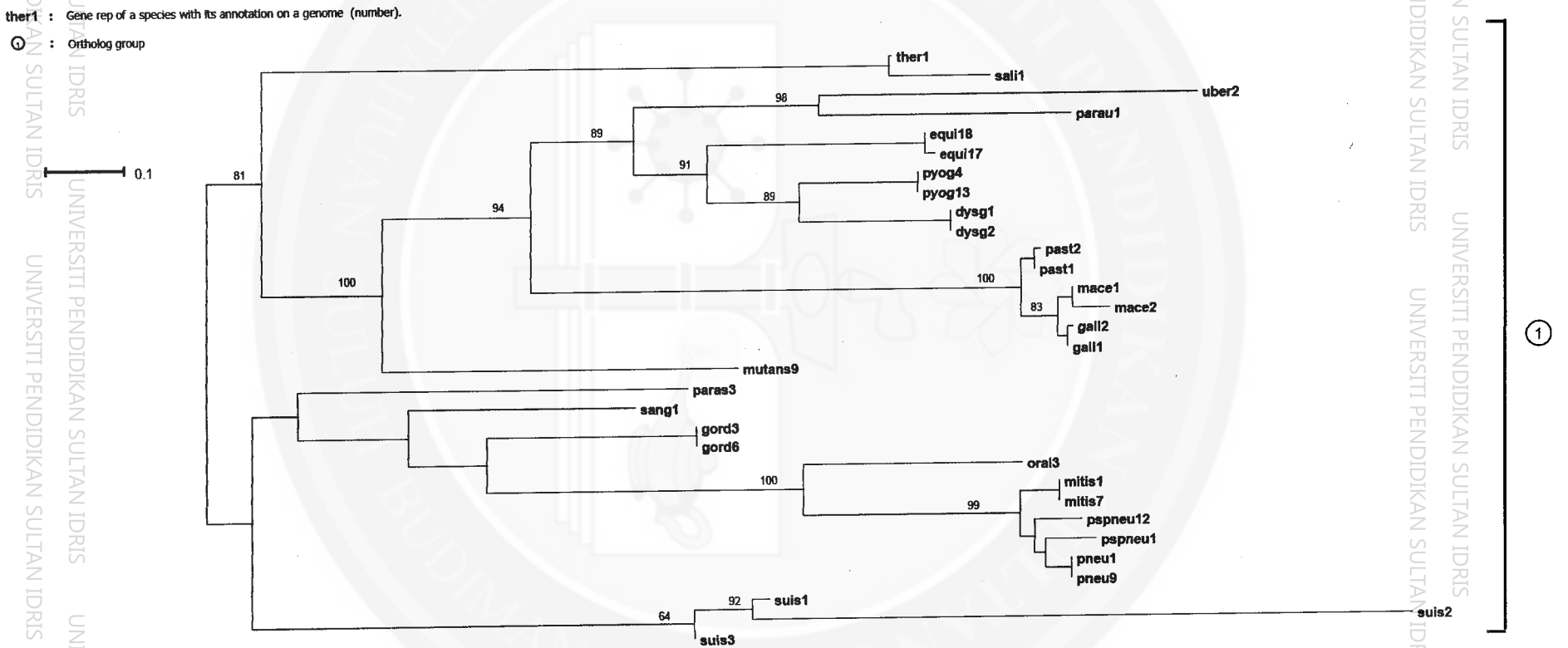
Appendix H.4: The accessory protein (*D* gene) maximum likelihood gene tree.



Appendix H.5: The accessory protein (*E* gene) maximum likelihood gene tree

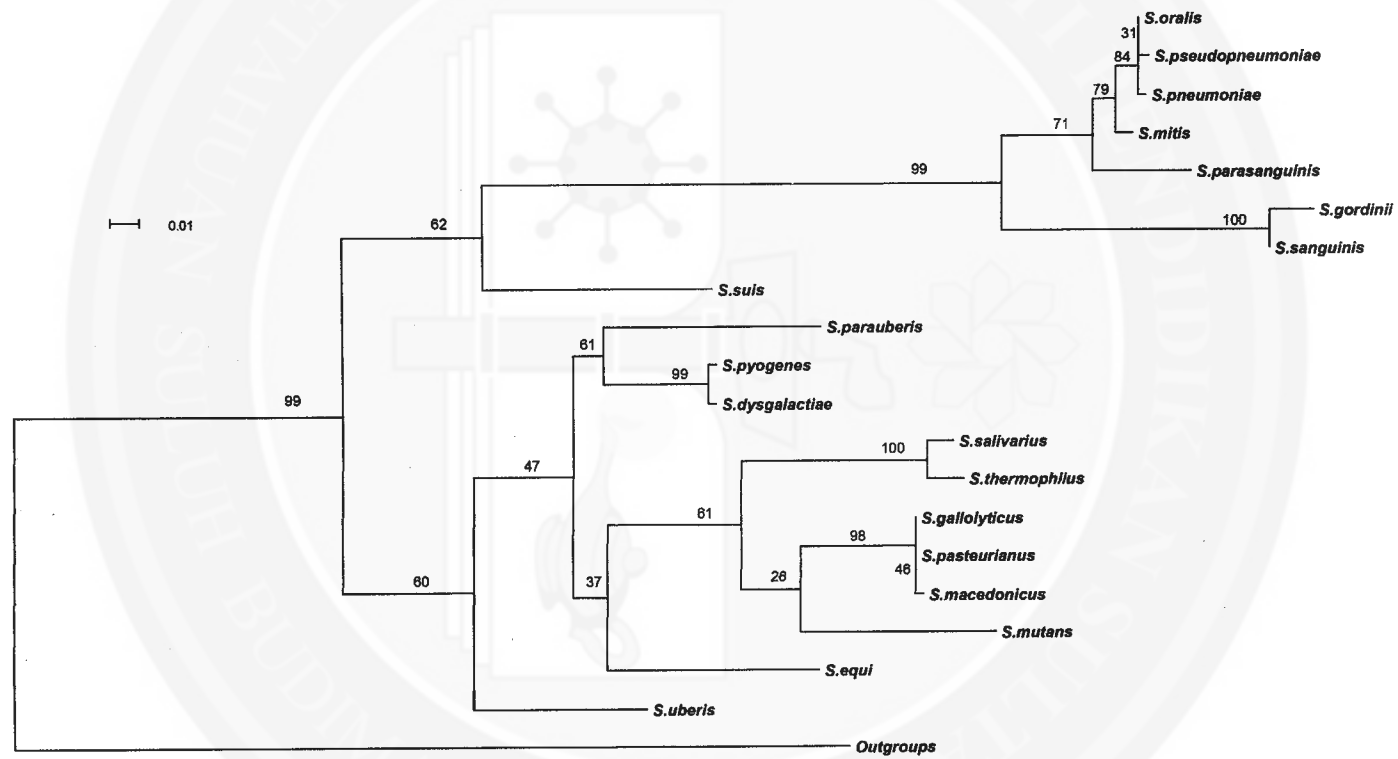


Appendix H.6: The accessory protein (*X* gene) maximum likelihood gene tree



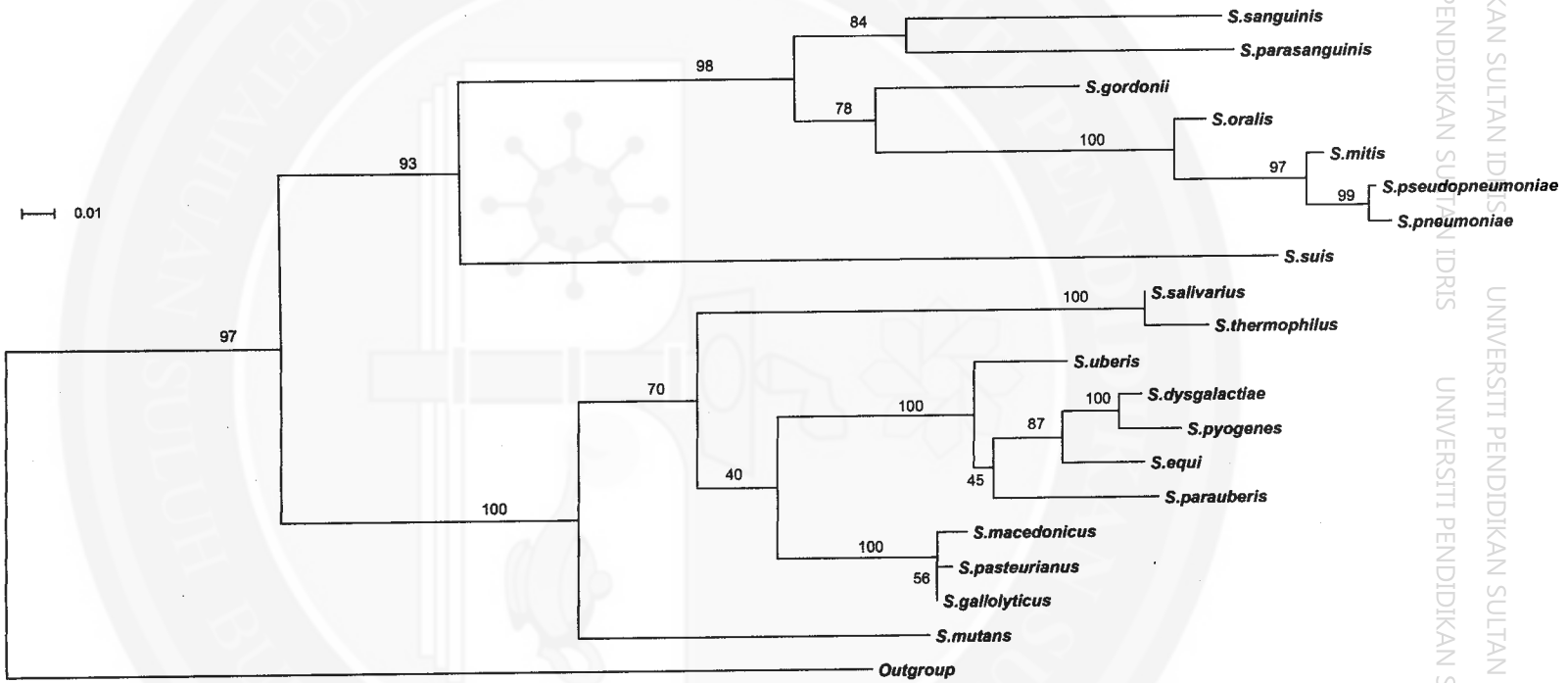
APPENDIX I

Appendix I.1: The rooted gene tree of COG0016, Phenylalanine-tRNA synthetase alpha subunit.

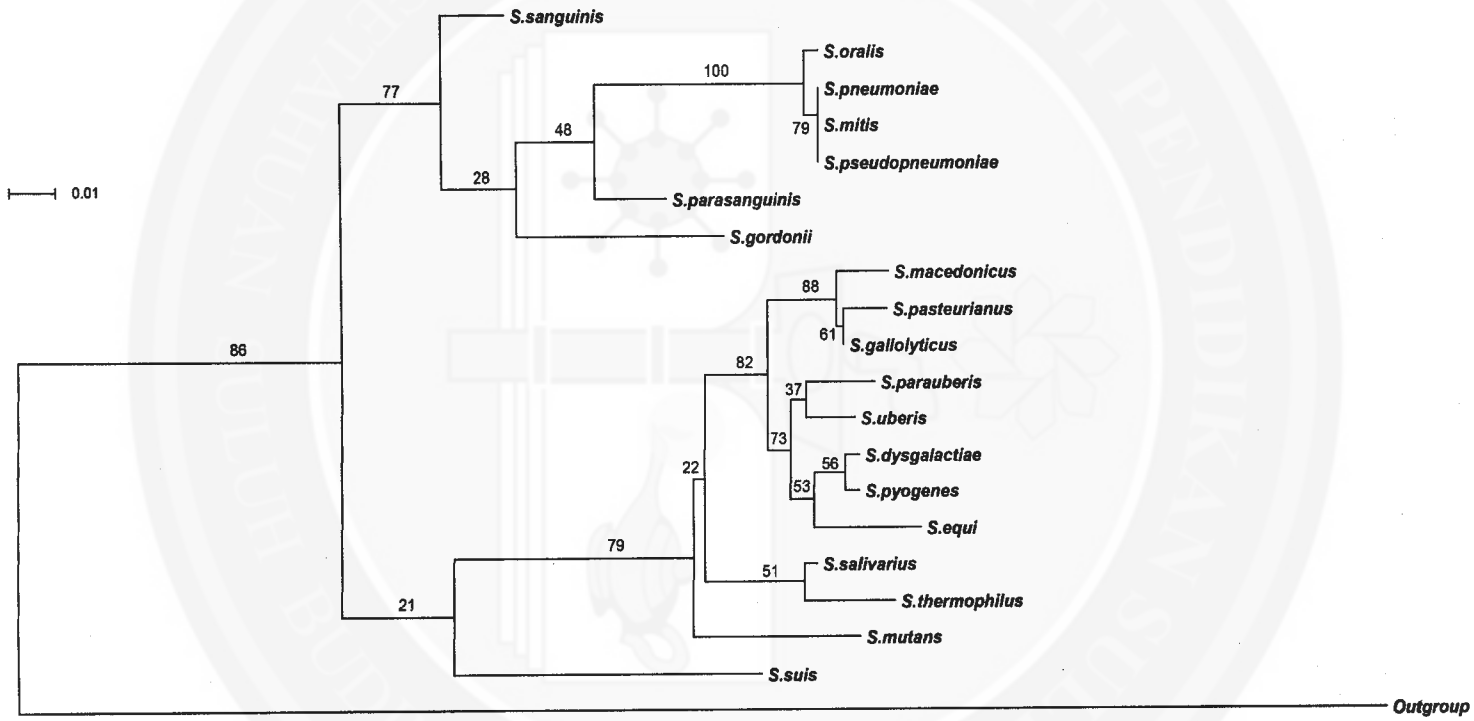


UNIVERSITI PENDIDIKAN SULTAN IDRIS

Appendix I.2: The rooted gene tree of COG0201, Preprotein translocase subunit SecY.



Appendix I.3: The rooted gene tree of COG0202, DNA-directed RNA polymerase, alpha subunit.



UNIVERSITI PENDIDIKAN SULTAN IDRIS

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

APPENDIX J

SUCRE methodology that was developed to identify the quorum sensing (QS) homologous genes in Streptococcus species contain several adjustable parameters (see Chapter 2 for details). These parameters are used to filter false positive hits result from BLAST homology search and use to construct the quality putative gene (QPG). The parameters are: i) V1-Distance between hits to merge as QPG (bp); ii) V2 -QPG sequence identity (%); iii) V3-QPG operonic distance (bp); iv) V4-QPG coverage (%); v) V5-Operonic distance of poor QPG coverage (bp).

The study optimised these parameters based on the SUCRE performance in detecting the QS benchmark genes dataset (Appendix B) in 100 testing dataset. The protein sequences of well-annotated QS genes from *S. pneumoniae* G54 (see Appendix B for full protein list in each species), are used as queries for the 100 testing dataset. We varying the values for each tested parameter and keeping others at fixed default value: V1 = 200bp; V2 = 40%; V3 = 2000bp; V4 =80%; V5 = 2000bp shown in the Table J.1.

To assess the overall homology detection performance, I have plotted the percentage of the TP score against each of the parameter value for the window size tested (Figure J.1). The optimisations of V1, V3, V4, and V5 parameters seem to be much robust with all of the default value maximised the performance in detecting the genes. As for the sequence identity, V2, it is expected that the higher the cut-off values of percentage of identity for the hits resulted from the BLAST homology search, the lower detection of the true positive genes. Thus, using the rule of thumb, I used the 40% as the cut off value as studies suggested that the sequence with lower than 35% identity may lose its protein structure thus alter the protein function (Rost, 1999, Roberto Sánchez et al., 2000). While the other parameters, the default values are used in SUCRE for optimal QS genes detection.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

Table J.1: The parameters value for testing dataset for SUCRE parameters.

SUCRE phase	Parameter	Test V1	Test V2	Test V3	Test V4	Test V5
Phase II Putative gene finding	V1 Distance between hits to merge as QPG (bp)	[1 ..10000]	200	200	200	200
	V2 QPG sequence identity (%)	40	[1..100]	40	40	40
Phase III Regulon construction	V3 QPG operonic distance (bp)	2000	2000	[1..4000]	2000	2000
	V4 QPG coverage (%)	80	80	80	[1..100]	80
	V5 Operonic distance of poor QPG coverage (bp)	2000	2000	2000	2000	[1..10000]

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDID
N IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI F

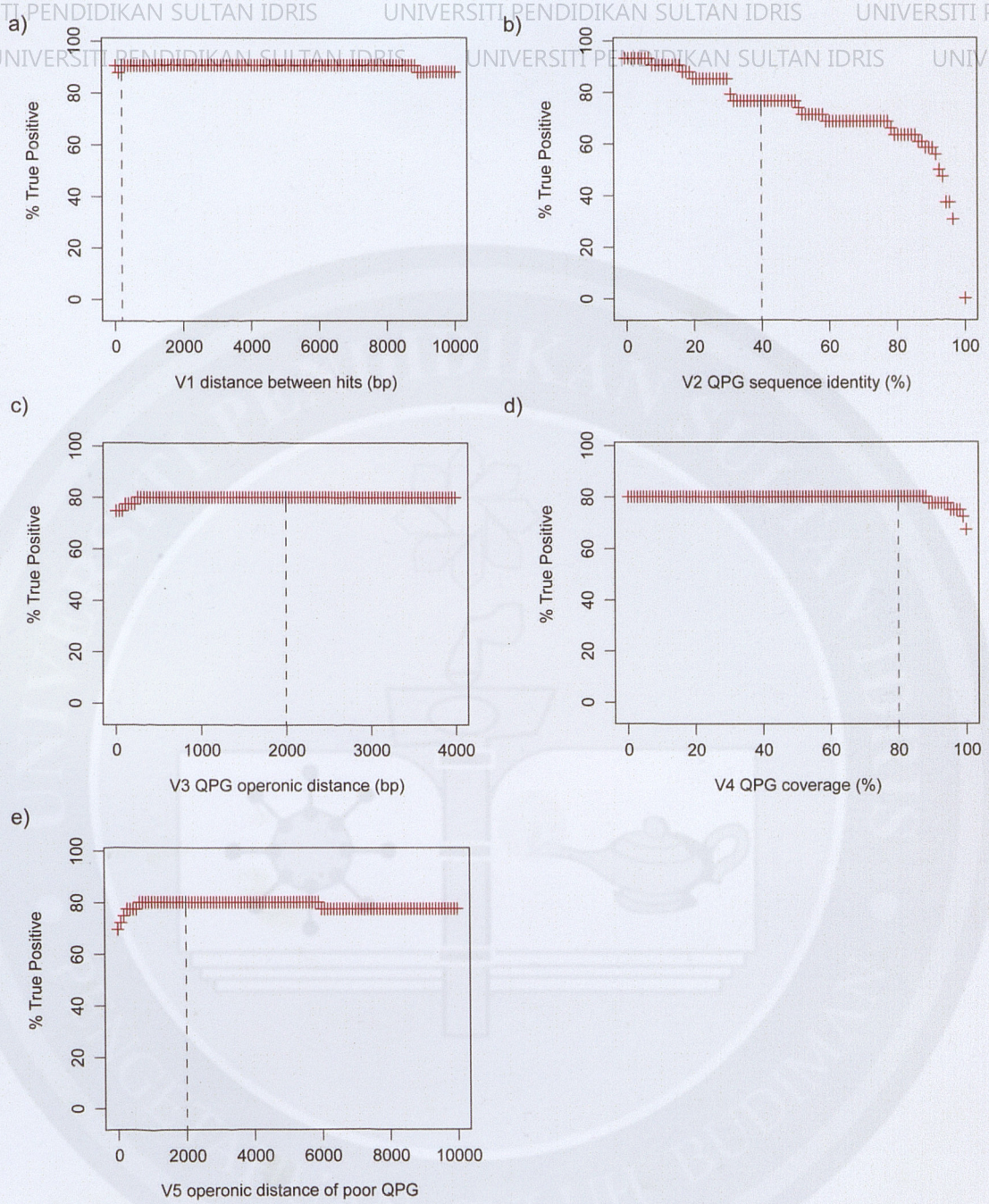


Figure J.1: The true positve value (%) plotted against each parameter: i) V1-Distance between hits to merge as QPG (bp); ii) V2 -QPG sequence identity (%); iii) V3-QPG operonic distance (bp); iv) V4-QPG coverage (%); v) V5-Operonic distance of poor QPG coverage (bp). The black line indicate the default value of each parameter.

UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKA
DRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PENDIDIKAN SULTAN IDRIS UNIVERSITI PEN